Inferring efficiency of translation initiation and elongation from ribosome profiling

Juraj Szavits-Nossan[1,*] **and Luca Ciandrini**[2,+]

[1]SUPA, School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom
[2]Centre de Biochimie Structurale (CBS), CNRS, INSERM, Univ Montpellier, Montpellier, France
[*]jszavits@staffmail.ed.ac.uk
[+]luca.ciandrini@umontpellier.fr

ABSTRACT

One of the main goals of ribosome profiling is to quantify the rate of protein synthesis at the level of translation. Here, we develop a method for inferring translation elongation kinetics from ribosome profiling data using recent advances in the mathematical modelling of mRNA translation. Our method distinguishes between the elongation rate intrinsic to the ribosome's stepping cycle and the actual elongation rate that takes into account ribosome interference. This distinction allows us to quantify the extent of ribosomal collisions along the transcript and identify individual codons where ribosomal collisions are likely. When examining ribosome profiling in yeast, we observe that translation initiation and elongation are close to their optima, and traffic is minimised at the beginning of the transcript to favour ribosome recruitment. However, we find many individual sites of congestion along the mRNAs where the probability of ribosome interference can reach 50%. Our work provides new measures of translation initiation and elongation efficiencies, emphasising the importance of rating these two stages of translation separately.

Introduction

Understanding the rationale behind codon usage bias and the role of synonymous codons in regulating protein synthesis are among the main open questions in molecular biology. Despite the fact that mRNA translation is a pivotal stage of gene expression, its sequence determinants are in fact still largely elusive[1]. Recent advances in sequencing, such as ribosome profiling[2], have made it possible to probe translation dynamics at codon resolution allowing for quantitative studies of translational efficiency.

Ribosome profiling (Ribo-seq or ribosome footprinting as it is often called), is an experimental technique delivering a snapshot of ribosome positions along all transcripts in the cell at a given condition. Its archetypal version has been developed at the end of the 1960s to study translation initiation[3,4], and has been extended in the 1980s to investigate the role of slow codons and ribosome pausing[5]. Recently, Ingolia et *al.* (2009)[2] revamped this technique to exploit the Next Generation Sequencing, and since then it is considered to be the state-of-the-art technique for studying gene expression at the level of translation.

In short, the method consists in isolating mRNA fragments (called "reads") covered by a ribosome engaged in translation ($\sim$ 30 nucleotides), which are then sequenced and aligned in order to build histograms of ribosome occupancy at codon resolution on each transcript. This technique has provided an unprecedented view on translation leading to many new discoveries[6]. Examples include detecting novel translation initiation sites[7], identifying actively translated open reading frames[8], quantifying the extent of stop codon readthrough[9] and elucidating the translation of long non-coding RNAs[10].

Translational activity on a given transcript is typically assessed by the number of read counts per kilobase of transcript per million mapped reads of the sample (RPKM), which takes into account the length of the transcript and the size of the sample. The RPKM is proportional to the ribosome density, which in turn is assumed to be proportional to the rate of translation –the more ribosomes on a transcript, the more efficient is protein synthesis. However, a large body of work based on mathematical modelling of ribosome dynamics suggests that the protein synthesis rate is negatively affected by increased ribosome density due to ribosome collisions[11–13]. To which extent ribosome collisions can be found using ribosome profiling has been an active topic of research[14–18]

One of the goals of ribosome profiling is to understand how the elongation rate along the transcript depends on the choice of codons. Codon elongation rates are usually estimated assuming that the ribosome density at codon i is proportional to the ribosome's dwell time on that particular codon [15,16,19–24]; this assumption follows from the conservation of the ribosome current assuming no ribosome drop-off. Our estimate of the drop-off rate of $\sim 10^{-3}$ s^{-1} (obtained from the probability of premature termination estimated to 10^{-4} per codon[25,26] and the elongation rate of the order of magnitude of ≈ 10 codon/s)

justifies the hypothesis. The inferred elongation rates are then checked against mRNA codon sequence features such as codon usage bias, tRNA availability and mRNA secondary structures.

If ribosome collisions are not rare, then the elongation rates proxied by the inverse ribosome densities do not depend only on the molecular details of the elongation cycle, but also on the extent of slowing down due to ribosome traffic. The crux of the matter is that it is difficult to distinguish from the ribosome density alone whether the ribosome spent long time on a particular codon because of the long decoding time or because it had to wait for the downstream ribosome to move away. This distinction between the *actual* elongation rates that account for ribosome traffic and the *intrinsic* ones in the absence of other ribosomes has been well accounted for in the standard model for mRNA translation known as the totally asymmetric simple exclusion process (TASEP), which considers ribosomes moving along the transcript in a stochastic manner[11]. Yet, very few of the existing studies use the TASEP to infer elongation rates from Ribo-seq; ones that do either do not infer the intrinsic rates[24] or use time consuming stochastic simulations to fit the Ribo-seq data[16,22].

In this work we develop an efficient method for inferring both actual and intrinsic codon-specific elongation rates from the ribosome profiling data based on the mathematical solution of the TASEP that we recently developed[12,27]. Using the TASEP, we argue that the ribosome density alone is not sufficient to estimate the *absolute* elongation rates from the ribosome profiling data. Instead,our method infers elongation rates of an mRNA *relative* to the initiation rate of that transcript. Moreover, we propose new measures of translation efficiency that quantify the amount of ribosome traffic around the START codon and along the transcript. We apply our method to several Ribo-seq datasets in *Saccharomyces cerevisiae* and show evidence of local queuing *in vivo*.

1 Results

We base our method on a well-established stochastic model for mRNA translation, the *totally asymmetric simple exclusion process* (TASEP)[11,28]. Over the years, the model has been improved in many ways to better match real translation [29,30] and has been repeatedly used to interpret experimental data[16,18,24,31–34].

1.1 The model

The TASEP describes ribosome dynamics on a discrete one-dimensional lattice representing the coding part of the mRNA molecule. Ribosomes do not interact apart from steric interactions, i.e. they cannot overtake each other. Each lattice site corresponds to a codon, and particles cover $\ell = 10$ sites, as the ribosome footprint covers ~ 30 nt or equivalently ~ 10 codons. A ribosome initiates translation at rate α, whereby its A-site is positioned at codon 2; this happens only if the codons $2,\ldots,\ell+1$ are not occupied by another ribosome's A-site. The ribosome then advances from codon i to codon $i+1$ at rate k_i, provided that codon $i+\ell$ is not covered by the downstream ribosome (see top right drawing of the model in Fig. 1); these are the intrinsic elongation rates at which the ribosome advances in the absence of other ribosomes. Eventually, when the A-site of the ribosome is at the STOP codon (the L-th site), it detaches the mRNA at rate $k_L = \beta$. A coding sequence is then characterised by a set of L rates α and $\{k_i\} = \{k_2,\ldots,k_L\}$. Further information about the model can be found in Methods and Materials.

The knowledge of initiation, elongation and termination rates allows one to compute simulated ribosome density profiles and protein production rates that can be compared to experimental outcomes. However, there is an open debate regarding the estimates of these rates, and no direct experimental method to measure them exists. For example, codon-specific translation elongation rates k_i are often assumed to be proportional to the tRNA gene copy number (GCN) or to the local tRNA adaptation index (tAI) [35–37].

Here, we use the model to quantitatively determine codon elongation rates from ribosome profiling data. This is an *inverse problem* because we need to optimise the inputs (parameters α and $\{k_i\}$) in order to match the outputs (Ribo-seq data). There are are several reasons why this problem is computationally difficult to solve, which we discuss below.

1. The parameter space is extremely vast: a typical protein consists of a several hundreds of amino acids, meaning that one generally needs to optimise a comparable number of parameters.

2. A change in a single k_i may affect a large part of the density profile, i.e there is a complex non-linear relation between the set of rates $\{k_i\}$ and the ribosome density profile.

3. The stochastic model predicts that the ribosome density profile depends on the ratios between the elongation rates and the initiation rate, meaning that it is not possible to estimate *absolute* rates without integrating more information.

We now explain how our method tackles these problems and how it compares to existing methods that have been proposed to infer ribosome dynamics from ribosome profiling data[16,20,22,24]. Our method uses an analytic expression for the ribosome density profile in terms of translation initiation, elongation and termination rates that we recently derived by solving for the steady state of the TASEP in the initiation-limited regime[12,27]. This relationship allows for a quick computation of the ribosome

density profile from the kinetic parameters of the model instead of running costly stochastic simulations for every iteration of the optimisation process[16,22]. In addition, our method searches for optimal elongation rates at each codon position and separately for every transcript, i.e. we do not reduce the parameter space by assuming equal elongation rates for every instance of the same codon[20,22]. Furthermore, we emphasise that our method infers *intrinsic* elongation rates (relative to the initiation rate) related to the ribosomal elongation cycle, and not only the *actual* elongation rates that also take into account slowing down due to ribosome traffic[24]. Thus, we are able to detect separately the mean decoding time for a particular codon and the mean time that the ribosome spends waiting for a ribosome downstream of it to move away. This distinction is central to our method.

Before we present further details of our method, we first discuss the problem of estimating absolute elongation rates, which is inherent to all methods and limits the amount of information that can be inferred from ribosome profiling data alone.

1.2 Ribosome profiles alone cannot estimate absolute elongation rates

We remind that the ribosome density ρ_i measures the probability of detecting a ribosome's A-site at codon position i (see Methods and Materials for further details). In Appendix A we show that ρ_i depends only on the ratios between the elongation/termination rates $\{k_j\}$ and the initiation rate α–we will refer to these ratios as $\{\kappa_j\}$. Thus given the ribosome densities $\{\rho_j\}$, one can only infer the ratios $\{\kappa_j\}$, but not the absolute rates $\{k_j\}$ and α. Since the initiation rate α is highly gene-dependent, it is not possible to compare the elongation-to-initiation ratios $\{\kappa_j\}$ from different genes without the knowledge of α for each gene. We demonstrate this point in Fig. 2(a), which shows the outcome of simulations of translation having different absolute rates $\{k_i\}$ but same relative speed profile $\{\kappa_i\}$.

We now examine two approaches that have been proposed to deal with this problem. The first approach is to fix a unique timescale shared by all mRNAs, for instance the average ribosome speed[38] or the average codon decoding time[24], which in turn allows one to estimate the initiation rate for each gene. The second approach is to normalise ρ_i by the average ribosome density

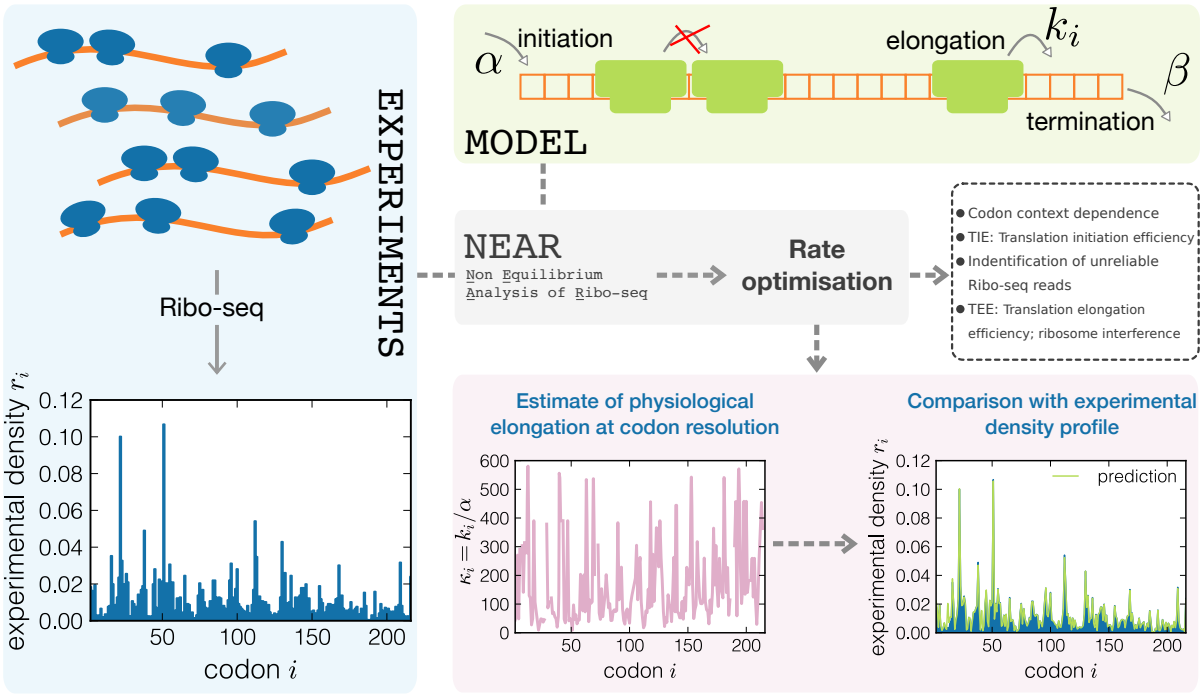

Figure 1. Sketch of the NEAR workflow for an individual gene (YAL007C). Experimental Ribo-seq profiles are first normalised and then analysed using the stochastic model. The normalised ribosome density profile $\{r_i\}$ is represented in the bottom left panel. The model is shown in the top right green box: ribosomes covering ℓ sites are added to the lattice with an initiation rate α, provided that the first ℓ sites are not occupied by the A-site of another ribosome. Ribosomes then move from site i to site $i+1$ at rate k_i provided that the A-site of the neighbouring ribosome downstream is at least ℓ sites away. Eventually, ribosomes leave the lattice at rate β ($k_L = \beta$) when their A-site is on the last site. In this drawing $\ell = 3$ for clarity, while in our analysis we used $\ell = 10$. NEAR searches for the optimal elongation rates k_i (relative to initiation rate α) for which the stochastic model reproduces the experimental ribosome profile. Once we find the optimal rates, we examine the extent of ribosome traffic using the translation initiation and elongation efficiencies (TIE and TEE), analyse the context dependence of elongation rates, and identify problematic transcript regions in which Ribo-seq data are not consistent with the model.

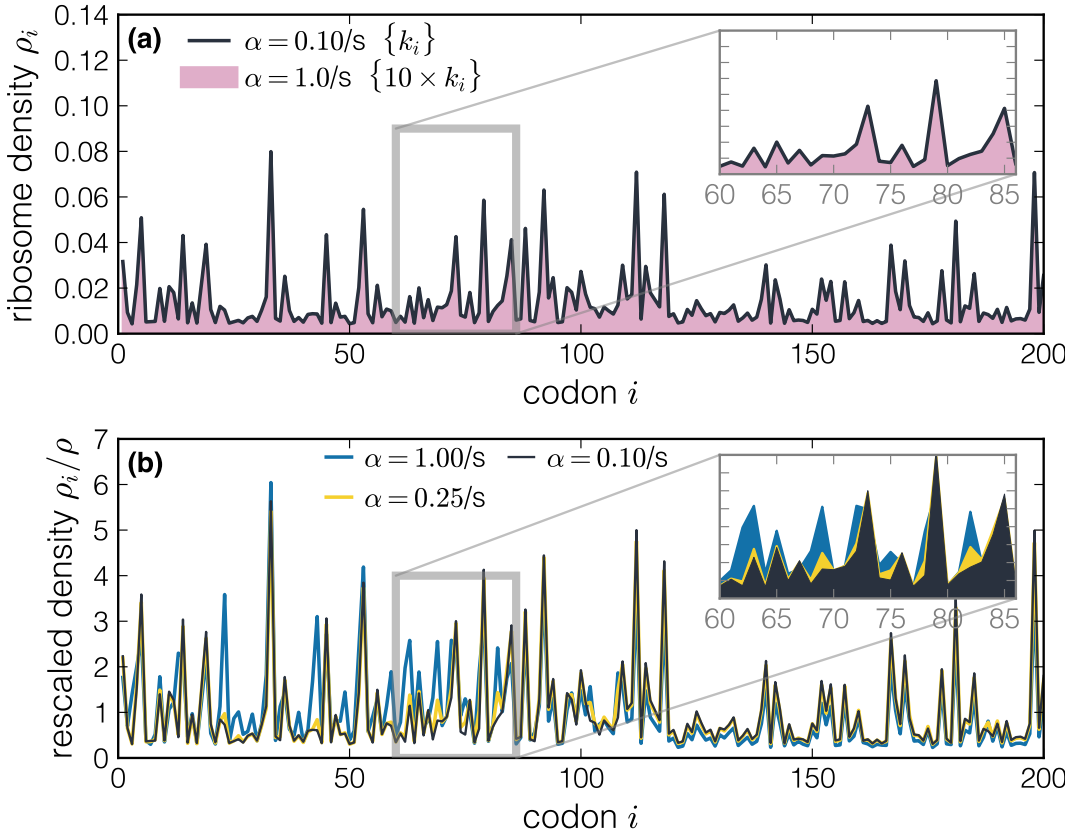

Figure 2. *in silico* density profiles. In panel (a), black line, we show the density profile obtained from the stochastic simulation of a transcript with a speed profile $\{k_i\}$ and initiation rate $\alpha = 0.1$/s. The pink region corresponds to the profile of a transcript simulated with a tenfold larger initiation rate, but keeping $\{\kappa_i\}$ constant (i.e. also increasing the elongation rates by a factor 10). This shows that densities obtained with the same elongation-to-initiation ratios $\{\kappa_i\}$ are indistinguishable. In panel (b) we fix the speed profile $\{k_i\}$ for three different values of the initiation rate α and we plot the rescaled profiles ρ_i/ρ. As expected, by increasing the initiation rate we obtain different profiles with increasing density and traffic effects.

ρ for that gene. This is a common practice in the analysis of Ribo-seq data, whereby the ribosome footprint read densities on individual codons are normalised by the average ribosome footprint density for that gene. The scaled read density is then assumed to be independent of the initiation rate, allowing for different genes to be compared. We argue that both of these approaches are problematic. In the first approach, the average elongation rate or the average decoding time could be highly variable from transcript to transcript, which in turn would introduce a bias when comparing absolute elongation rates between different genes. In the second approach, the normalisation of ρ_i by ρ does not necessarily mean that genes with different initiation rates can be directly compared. We show that explicitly by computing ρ_i/ρ in our model for different initiation rates but keeping the elongation speed profile $\{k_i\}$ fixed. As shown in Fig. 2(b), we find quantitatively different profiles for different initiation rates. This observation is further supported by the analytic expression for ρ_i, which predicts a non-linear α-dependent correction to the linear expression $\rho_i \approx \alpha/k_i$ (see Appendix A).

Instead, our approach is to scale κ_i by the termination-to-initiation ratio κ_L which removes dependence on the initiation rate since $\kappa_i/\kappa_L = k_i/k_L$. Later we show that the values of κ_L inferred from ribosome profiling data in *S. cerevisiae* have among the least variation of all codons, which supports our choice for κ_L as the scaling factor. In addition, we introduce new measures of translation efficiency and ribosome traffic that take values between 0 and 1 and can be compared between different genes.

1.3 Non-Equilibrium Analysis of Ribo-seq (NEAR)

Having shown that the ribosome density profile alone can inform us only on the ratios $\{\kappa_i\}$ between the elongation rates and the initiation rate, we now turn to the method for inferring $\{\kappa_i\}$ from Ribo-seq. We call the method Non-Equilibrium Analysis of Ribo-seq data (NEAR) in order to emphasise that we use the model that is far from the thermal equilibrium.

NEAR infers $\{\kappa_i\}$ with an optimisation procedure that aims to find a model-predicted density profile $\{\rho_i\}$ which is a close

match to the experimental one $\{r_i\}$ (see Fig. 1). This is possible since we have recently found a mathematical expression for the ribosome density profile in terms of translation initiation, elongation and termination rates. This expression was obtained under the assumption of a limiting initiation rate α[12,27], which is supposed to hold for most of the mRNAs under physiological conditions (see Appendix A). However, we emphasise that the initiation-limited approximation does not assume that ribosome collisions are absent. Instead, our analytic solution takes ribosome collisions into account and is applicable to a wide range of initiation rates as long as they are smaller than the elongation and termination rates (see Appendix A).

We have applied our method to ribosome profiling data of *Saccharomyces cerevisiae* obtained by Guydosh et al. (2014)[14], Pop et al. (2014)[20] and Weinberg et al. (2016)[23]. These datasets were selected for their lack of using cycloheximide to inhibit translation elongation, which is known to distort ribosome coverage profiles[39,40]. The raw data was processed by the Riboviz software[41] and mapped to A-site positions following the table provided in Ahmed et *al.* (2019)[42]. After obtaining the A-site read density profiles the method proceeded in four steps, which we summarise below.

1. We first normalised the number of A-site read density on each codon by the total number of reads. This number was then multiplied by the absolute ribosome density for that particular gene obtained by polysome profiling experiments in MacKay et *al.* (2004)[43]. The end result is a normalised ribosome density profile $\{r_i\}$ that reveals how likely is to find a ribosome at codon i.

2. Next, we solved a least-squares optimisation problem which consisted of finding $\{\kappa_i\}$ that minimise the objective function

$$S = \sum_{i=2}^{L} (\rho_i(\{\kappa_i\}) - r_i)^2 . \tag{1}$$

   This step ensured that the model-predicted ribosome density profile $\{\rho_i\}$ was a close match to r_i.

3. Once we found the best estimate of $\{\kappa_i\}$, we then computed the density profile $\{\rho_i^{\text{sim}}\}$ from stochastic simulations using the estimated $\{\kappa_i\}$.

4. In the last step we performed a quality check on each κ_i, and verified that:

   (a) the initiation-limited approximation was applicable by comparing the analytic density profiles with the simulated ones. This step is necessary because our solution of the model is approximate, see Refs. [12,27] and also the Appendix A.

   (b) the estimated κ_i reproduced the experimental density r_i (within 5% tolerance).

These are the main steps of NEAR, and we provide further mathematical details in the Appendix B.

We emphasise the importance of optimising the absolute ribosome densities $\{\rho_i\}$ (step 2), rather than the scaled ones, $\{\rho_i/\rho\}$, as in other methods that analyse ribosome profiling data[16,20]. The problem is that $\{\rho_i/\rho\}$ remains the same if we multiply all ρ_i by a constant factor, which in turn means that different density profiles $\{\rho_i\}$ can result in the same scaled profile $\{\rho_i/\rho\}$. Since each ρ_i is uniquely determined by the set of elongation-to-initiation ratios $\{\kappa_i\}$, we conclude that the scaled density profile $\{\rho_i/\rho\}$ does not uniquely determine $\{\kappa_i\}$, see for instance Figure 3 –figure supplement 2. Thus, we reiterate that ribosome profiling data is not sufficient to infer ribosome dynamics and in turn the extent of ribosome traffic without the additional information on the mean number of ribosomes bound per mRNA (step 1).

Our quality check is step 4 is also able to reject codons whose κ_i cannot be trusted, and identify why the inference of elongation rates for those codons is problematic. We are in fact able to say whether our analytic solution is satisfied or not, or if the problem is due to the model being inconsistent with the experimental data.

Before moving on to real sequences of *S. cerevisiae*, we tested NEAR on a mock sequence with known $\{k_i\}$, and checked that it can accurately infer the original elongation rates (Figure 3–figure supplement 1). We also remark that the quality check allows to push the analysis to relatively high elongation rates.

1.3.1 Using NEAR to study translation of individual genes

We demonstate our method on a particular gene (YLR301W) using Ribo-seq data from the Weinberg dataset. We first compute the normalised experimental density profile $\{r_i\}$ using the experimental absolute density r from MacKay et *al.*[43] (in units of ribosomes/codon). This profile is then analysed following the method explained in the previous section. A set of elongation-to-initiation ratios $\{\kappa_i\}$ is obtained by optimising the match between the model-predicted density profile and the experimental one. Each κ_i is then examined to see whether it provides a good prediction for that particular codon position and to check for

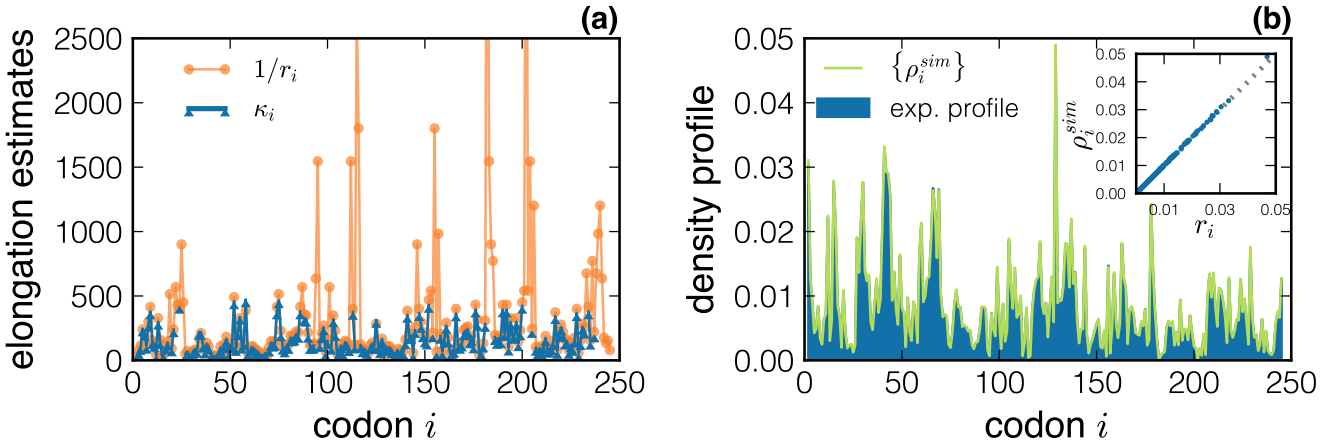

Figure 3. Results of NEAR applied to the YLR301W gene. (a) The optimised profile $\{\kappa_i\}$ is plotted (blue line, triangle markers) as a function of the codon position i, and compared to the naive estimates $\{1/r_i\}$ (orange line, round marker). In panel (b) we compare the model-predicted density profile obtained using the inferred $\{\kappa_i\}$ (green line) with the experimental normalised profile $\{r_i\}$. The inset shows the scatter plot between the two densities (for each codon i) demonstrating an excellent agreement between theory and experiments.

inconsistencies in the method as previously explained. As a result, a few values of κ_i that do not pass this quality check are rejected and are not included in the final analysis.

In Fig. 3(a) we plot the the optimised $\{\kappa_i\}$ profile that passed the quality check (blue line, triangle markers) compared to the naive estimate $1/r_i$ (orange line, round marker) that ignores ribosome interference but it is usually judged as a good estimator of the elongation rate. We find many codon positions where the two profiles $\{\kappa_i\}$ and $\{1/r_i\}$ significantly differ from each other. Moreover, we identify values of κ_i that are not consistent with the model, while this cannot be done when using $\{1/r_i\}$ as a proxy for elongation determinants.

The result of a stochastic simulation of ribosome dynamics performed with the optimised elongation ratios $\{\kappa_i\}$ is shown in Fig. 3(b). The agreement between the simulated density profile $\{\rho_i^{\rm sim}\}$ (green line) and the experimental one (in blue) is excellent for most of the codons. The inset shows the scatter plot between the values (for each codon) of the simulated and experimental ribosome density.

1.4 Estimate of elongation-to-initiation ratios at codon resolution in yeast

We analysed three different datasets and gathered the NEAR elongation-to-initiation ratios $\{\kappa_i\}$ for each gene. The estimated elongation-to-initiation ratios passing the quality check are plotted in Fig. 4 and compared to the naive estimates $1/r_i$. In particular, we find many instances where $1/r_i$ deviates from κ_i obtained by NEAR. The model predicts that $\kappa_i \approx 1/r_i$ if there are very few ribosomes on the transcript so that ribosome collisions are rare. Our findings in Fig. 4 thus suggest that the effect of ribosome interference is not negligible. We will discuss this point later when we introduce better measures for detecting ribosome interference.

Next, we wan to understand if each codon type has a characteristic decoding time and verify or reject a common hypothesis that elongation rates are determined by the availability of aminoacyl-tRNAs. By definition, κ_i is equal to the ratio k_i/α between the elongation rate of codon i and the initiation rate α of the gene. Because the initiation rates are likely to be gene-specific, we do not know if the observed variation in elongation-to-initiation ratios of the same codon types (see Fig. 4–figure supplement 1) is due to variation in the elongation or initiation rates.

However, we observe that STOP codons show the least variability of all the elongation-to-initiation ratios $\{\kappa_i\}$ in the Guydosh and Weinberg datasets (Figure 4–figure supplement 1). This result is consistent with the expectation of a context-independent termination. Thanks to this observation, we then compute the elongation-to-termination ratio $\kappa_i/\kappa_L = k_i/\beta$, i.e. the elongation rate of codon i with respect to the termination of the gene under investigation (Fig. 4–figure supplement 2). This quantity does not depend on the initiation rate α that is likely to be context-dependent and different from gene to gene. Indeed, the variation in κ_i/κ_L linked to the same codon type is now more uniform across 61 codon types, especially in the Guydosh dataset (Fig. 4–figure supplement 2). We have also compared median values of κ_i/κ_L for each codon type against two common measures of tRNA availability: a codon-dependent rate of translation based on the tRNA gene copy number (GCN) corrected for the wobble base pairing from Weinberg et *al.* (2016)[23], and the tRNA adaptation index (tAI)[44]. We find a moderate correlation between the median of the κ_i/κ_L distributions and the corresponding tRNA gene copy numbers (Figure

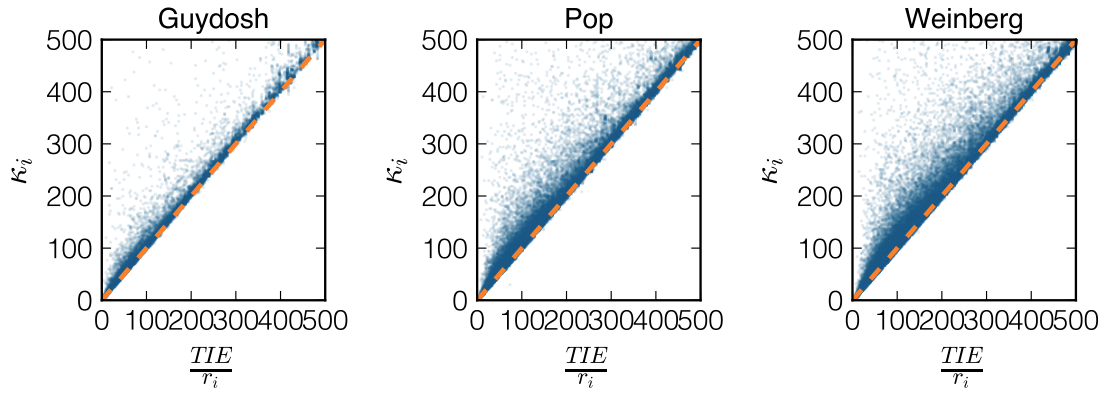

Figure 4. Scatter plot of the elongation-to-initiation κ_i for each codon that passed the quality check versus the inverse of the experimental density $1/r_i$ for the Guydosh, Pop and Weinberg datasets. The dashed line corresponds to the bisect.

4–figure supplement 3). This result suggests that the elongation speed of individual codons is only partially determined by their codon type.

We now turn to ribosome traffic and its impact on translation efficiency. In the following sections we will define quantities that, contrary to the κ_i, can be used to compare translation efficiency of different genes. Those quantities, which we name the translation initiation efficiency (TIE) and the translation elongation efficiency (TEE), can be used to rank initiation of different transcripts and quantify the impact of ribosome interference along a mRNA.

1.5 Translation Initiation Efficiency (TIE)

By running stochastic simulations with the inferred κ_i we can measure the ribosomal current J divided by the initiation rate α, which is a quantity dependent on $\{\kappa_i\}$ only; the current J can be used as a proxy for protein synthesis rate per mRNA.

In the biological literature translation initiation is often identified with protein synthesis rate, i.e. $J = \alpha$. However this is true only if initiation is much slower than elongation so that essentially only one ribosome is translating a transcript at a given time. Yet, this approximation is too crude to quantitatively describe translation[12]. Instead, when more than one ribosome is engaged in translation, J becomes a function of α and the elongation rates $\{k_i\}$; the current J can be thought of as the intrinsic initiation rate α multiplied by the probability that the first codons of the mRNA are not occupied by another ribosome (which would otherwise obstruct initiation).

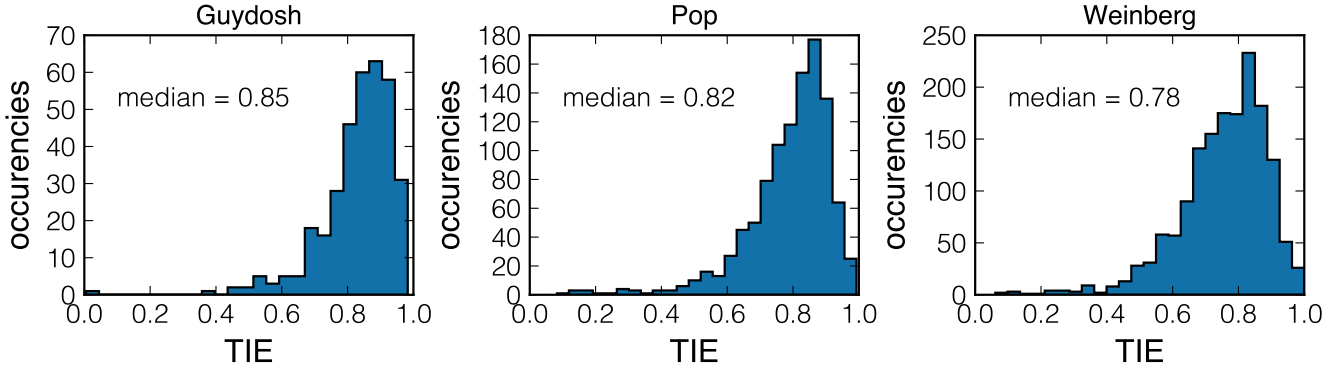

Figure 5. Histogram of estimated Translation Initiation Efficiency (TIE) for all *Saccharomyces cerevisiae* genes included in our study for the Guydosh, Pop and Weinberg datasets. The TIE gives the probability that the first codons are unoccupied.

Therefore we propose to use J/α as a measure of the Translation Initiation Efficiency (TIE), which takes values between 0 and 1. The TIE would be equal to 1 in the optimal case in which initiation is not hindered by ribosome traffic ($J = \alpha$, hence TIE = 1). Otherwise, the TIE gives the probability that the first codons, potentially interfering with ribosome recruitment and initiation, are unoccupied. A TIE smaller than 0.5 means that more than half of the times a new ribosome tries to initiate, it fails because of another ribosome whose A-site is located within the first ten codons. In Fig. 5 we plot the histograms of TIE for all the genes and datasets included in our study. We find that almost all genes show TIE > 0.5 with a median value around

0.8 for all the datasets. These values suggest that the first codons are mainly free from ribosomes that are already engaged in translation.

Our previous theoretical work on the exclusion process showed that if translation is rate-limited by initiation, then TIE predominately depends on the κ_i of the first $\ell \approx 10$ codons, which is the ribosome footprint on the mRNA[12]. Based on that prediction, TIE > 0.5 is a strong signature that the codon sequences, and in particular the first codons of *Saccharomyces cerevisiae* genes might have been selected to optimise translation initiation.

1.6 Translation Elongation Efficiency (TEE) shows congestion of ribosomes *in vivo*

In contrast to the TIE, we define an efficiency index for translation elongation that identifies local ribosome interference along the transcript and not only around the initiation region. In order to do that, we emphasise that the total time $t_i(\text{total})$ that a ribosome spends with its A-site on a given codon i can be seen as the sum of two contributions: the time $t_i(\text{intrinsic})$ needed to decode this codon and incorporate the new amino acid to the growing peptide chain, plus the time $t_i(\text{collision})$ spent in a queue waiting for the downstream ribosome to move. For each codon i, $t_i(\text{total}) = t_i(\text{intrinsic}) + t_i(\text{collision})$. The inverse of $t_i(\text{total})$ is the actual elongation rate, while the inverse of $t_i(\text{intrinsic})$ is the intrinsic elongation rate k_i, i.e. the elongation rate in the absence of other ribosomes. The distinction between these two rates is important because the actual elongation rates may be much smaller than the intrinsic ones in genes with higher initiation rates in which ribosomal collisions are more likely. Thus analysing the actual instead of intrinsic elongation rates could obscure our search for the molecular determinants of the translation speed.

We consider a codon as *efficient* if a ribosome attempting to translate it is not blocked by other ribosomes. We thus define the Translation Elongation Efficiency at codon i (TEE_i) as the ratio of intrinsic and total time: $\text{TEE}_i = t_i(\text{intrinsic})/t_i(\text{total})$, or put differently, as the ration between the actual and intrinsic elongation rate. The TEE_i is a measure of the local mRNA congestion seen by a ribosome translating the codon i and it depends on the context at which the codon is placed. Mathematically, it is equivalent to the probability that the $i+1 \ldots i+\ell$ codons are not occupied, given that a ribosome's A-site is at site i. If the intrinsic decoding time of the ribosome is equal to the total time dwelt on the codon, then the ribosome experiences on average no interference with other ribosomes and $\text{TEE}_i = 1$. Otherwise, $0 < \text{TEE} < 1$. In the extreme case of the completely jammed ribosome one would get $\text{TEE}_i \approx 0$, i.e. the ribosome is ready to advance but it is not allowed to move forward because the transcript is overcrowded. Furthermore there is a relationship between the TEE_i and TIE given by $\text{TEE}_i = \text{TIE}/(\kappa_i \rho_i)$. Further details are given in Appendix B.

We note that the TEE_i is a function of $\{\kappa_i\}$ only, meaning that ribosome interference is governed by the balance between initiation and elongation rates. A TEE profile that is close to 1 means that initiation is not frequent enough to cause ribosome congestion along the transcript. Inferring TEE profile from ribosome profiling data is thus a convenient method for testing whether translation is limited by initiation. We also stress that both the TIE and TEE_i are dimensionless quantities that take values between 0 and 1. Therefore it is possible to compare the TIE and TEE profiles between different genes.

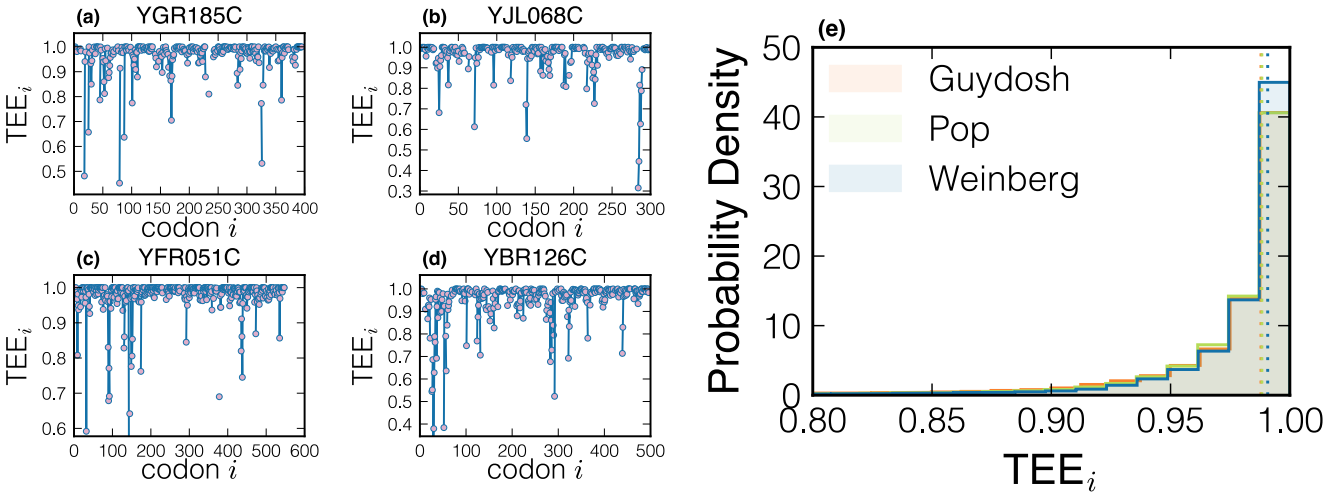

Figure 6. Panels (a)-(b)-(c)-(d) show single-gene TEE profiles (randomly selected genes from Weinberg dataset). In panel (e) we plot the distributions of the TEE collected on all codons of the three datasets analysed. The dashed vertical lines represent the median of those distributions (Guydosh: 0.988, Pop: 0.988, Weinberg: 0.990).

In Figure 6(a)-(b)-(c)-(d) we plot the TEE profile of four randomly selected genes. We observe that the TEE is typically close to 1 indicating that traffic is negligible for most of codons. We also identify particular codons where ribosome interference

is significant and TEE_i drops to 0.6. These examples demonstrate that NEAR can locate, at codon resolution and excluding unreliable estimates, particular regions on the transcript that are affected by ribosome interference.

After analysing TEE profiles of all genes included in our study, we observe that the distribution of TEE_i for all the codons that passed the NEAR quality check is peaked at 1, with the median at about 0.99, as shown in Fig. 6(e). This result is consistently found in all three datasets that we analysed, suggesting that ribosome interference is present only locally on a few codons, and is generally absent *in vivo*.

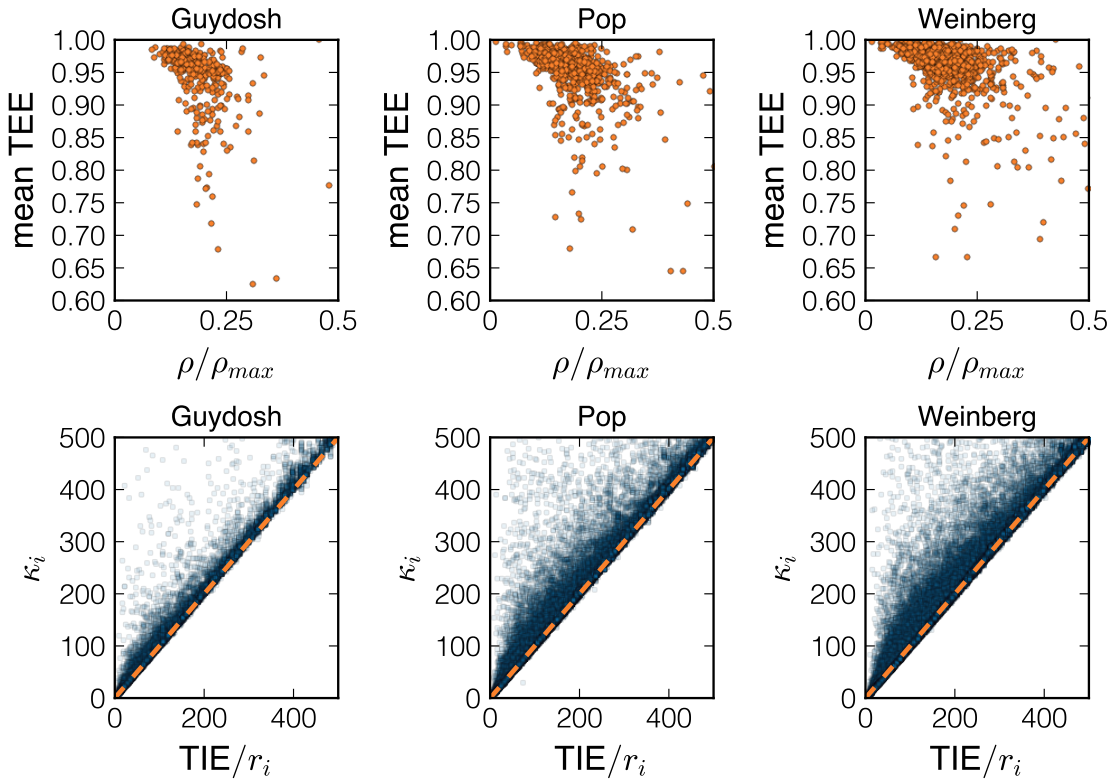

Figure 7. In the first row we show the ribosome density ρ normalised with the maximal density $\rho_{\max}$ versus the mean value of TEE along the transcript. In the second raw we plot TIE/r_i versus κ_i to emphasise the extent of traffic in determining the ribosome's dwelling time. The orange dashed line is the bisect $TIE/r_i = \kappa_i$ (no traffic). Codons far from that line are the ones more impacted by ribosome interference.

If the ribosome density on a given transcript is high, one would expect to see an increased number of ribosomal collisions resulting in the TEE profile that clearly deviates from 1. In Figure 7 we present the mean of the TEE profile for each gene that we analysed compared to the ratio of the ribosome density for that gene and the maximum achievable density $\rho_{\max} = 1/\ell$, where $\ell \approx 10$ codons is the ribosome footprint length. The results across all three Ribo-seq datasets clearly show that genes with low ribosome density have the mean TEE very close to 1 (few collisions). On the other hand, the mean TEE of genes with high ribosome density deviates significantly from 1 (many collisions).

Another way to demonstrate the importance of ribosome collisions is to directly estimate t_i(collisions). Since the total time spent on a codon is given by r_i divided by the ribosomal current, we obtain

$$\alpha\, t_i(\text{collisions}) = \frac{r_i}{\text{TIE}} - \frac{1}{\kappa_i}\,. \tag{2}$$

The time spent in traffic on codon i is then larger than zero if $r_i/\text{TIE} > 1/\kappa_i$ and equal to zero only if there is no ribosome interference. In the second row of Figure 7 we show that many of the codons analysed deviate from the bisect. This is another quantitative evidence that, according to experimental data, it is not that rare to observe ribosomes queuing *in vivo*.

1.7 Initiation and elongation interdependence

After observing that translation elongation efficiency is generally close to its optimum value of 1, we now look for spatial distribution of the TEE_i along the transcripts. To this end we compute a metagene TEE profile by aligning the genes at their

START codon and computing the distributions of the TEE_i at each position i. We then take the median of the distribution on each codon. The results are plotted in Figure 8(a). This genome-averaged profile confirms our earlier observation that TEE is close to 1. However, we also observe that the first ~ 10 sites have a larger elongation efficiency. A large value of TEE around the START codon helps to clear this region from queueing ribosomes and thus facilitates ribosome recruitment (see also Relationship between TIE and TEE in Appendix B). This result is consistent with a large value of TIE previously observed in Figure 5, and it confirms the importance of the beginning of the coding sequence in controlling translation.

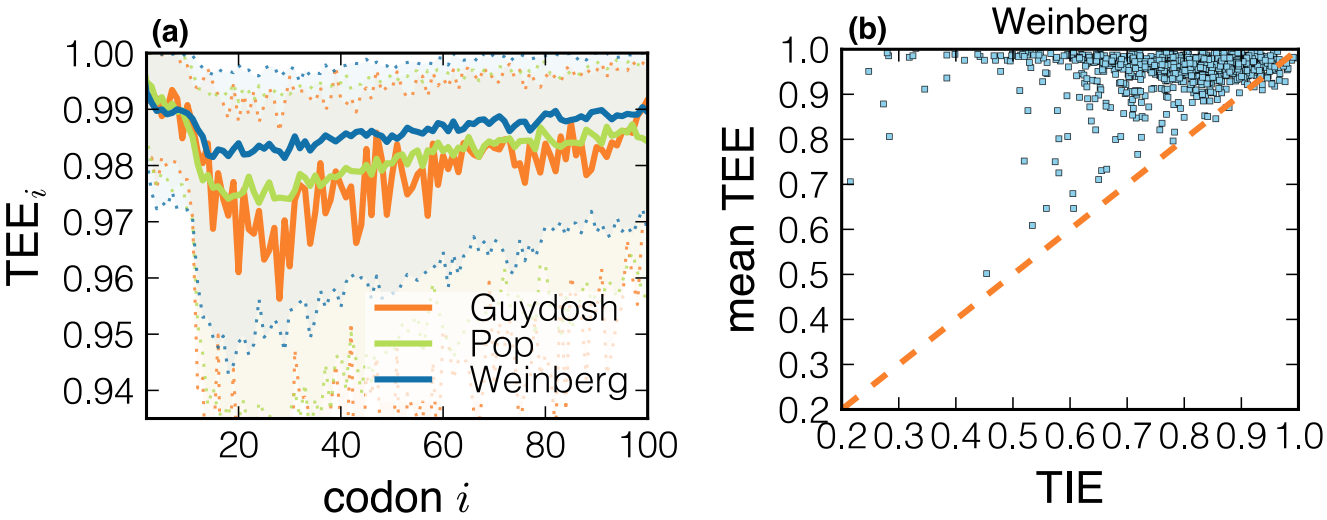

Figure 8. (a) Translation Elongation Efficiency metagene profile. The full line is the median TEE profile for three different datasets included in our study. Dotted lines delimiting the shadow area correspond to the first and third quartiles of the distribution. In panel (b) we show the scatter plot of the TIE and mean TEE of each gene in the Weinberg dataset. The orange dashed line represents mean TEE = TIE. The analysis of the other dataset can be found in Figure 8 Supplement 1.

Our results seem to suggest that TIE and TEE should be strongly related. On the one hand, if translation is efficient but elongation is inefficient, ribosome interference would dominate and ribosomal resources would be wasted. On the other hand, effective elongation and weak initiation would still finely tune the overall protein production without harming cellular fitness. Following these considerations, from the evolutionary point of view there should exist a constraint between the relative weights of initiation and elongation, and a situation with strong initiation and weak elongation should be avoided.

We can roughly evaluate the overall elongation efficiency as the mean of the TEE_i profile of each gene, and thus associate a couple of values (TIE, mean TEE) to each gene analysed. In Figure 8(b) we observe that the constraint TIE < mean TEE is satisfied for most of the genes analysed (only a few exceed the TIE = mean TEE dashed line, and for very initiation efficient genes); thus the data analysed are consistent with the hypothesis explained above. We also notice that transcript with inefficient initiation might also present a less optimised elongation, suggesting that initiation and elongation are interdependent.

2 Discussion

In this work we introduce NEAR, a Non-Equilibrium Analysis of Ribo-seq, which is based on a well-studied model borrowed from statistical physics. The model tracks individual ribosomes engaged in translation and predicts their spatial distribution on the mRNA and the rate of protein synthesis using initiation, elongation and termination rates as input parameters. Here we do the opposite—we develop a method that infers elongation-to-initiation ratios at codon resolution directly from ribosome profiling data.

We first emphasise that Ribo-seq profiles, being an averaged snapshot of the translatome, do not contain information on the absolute timescales of the process, and that thus it is possible to estimate relative rates only. These rates uniquely determine the density profile and allow us to evaluate the extent of ribosome traffic along the transcript and show a possible interplay between initiation and elongation. To this end we introduce new measures of translation efficiency that we named translation initiation and elongation efficiencies (TIE and TEE, respectively). Importantly, both TIE and TEE are dimensionless scores taking values between 0 and 1, which allows us to compare ribosome traffic between different genes.

Translation initiation efficiency (TIE) is defined as the probability that a ribosome attempting to initiate translation is not obstructed by another ribosome on the coding sequence. The distribution of TIE for the three datasets that we used in this study show that ribosomes can easily access most transcripts, with the median value of 0.8 for the probability to find the initiation

region unobstructed (Figure 6). Yet, we find genes with low TIE suggesting that the first codons can exert control over protein synthesis through ribosome traffic interfering with translation initiation. These results are in line with recent experimental evidence on ribosome stalling and traffic in the initiation region[17,45,46].

Similarly, translation elongation efficiency (TEE_i) is defined as the probability that a ribosome at codon position i is not blocked by another ribosome downstream of i. The distribution of TEE_i across all transcripts shows that TEE_i is generally close to 1 suggesting that ribosome interference is negligible for most codons (Figure 6(e)). However, when looking at the individual gene TEE profiles, we observe that it is not so rare to find the probability of ribosome interference as high as 50% (Figures 6(a)-(d)). In accordance with these results, we find more evidence of ribosome interference (lower TEE) in genes with higher ribosome density (Figure 7). We also compute the average time t_i(collisions) that each ribosome spends on a codon due to the blockage of downstream ribosomes. If no traffic is present then $t_i(\text{collisions}) = 0$. Instead, we observe many codons for which $t_i(\text{collisions}) > 0$ (Figure 7).

The fact that the value of TEE at each codon must fall between 0 and 1 allows us to agglomerate all values of TEE into a "metagene" profile (Figure 8). Interestingly, the median TEE shows slightly higher values at the first 10 codons, suggesting that queuing is avoided in order to allow for efficient ribosome recruitment at the start codon. This result is consistent with a recent study in which replacing the first 8 codons with their slower synonymous variants significantly reduced protein expression without affecting mRNA levels[45]. Furthermore, the first codons have been recognised as critical in determining protein synthesis both theoretically[12,13,47] and experimentally[46,48–50]. Beyond the first 10 codons, the metagene profile of TEE further reveals a small but noticeable drop between codons 10 and 20, followed by a slow increase between codons 20 and 100. These results are consistent with the "ramp hypothesis" proposing that rare codons are more frequent at the beginning of genes in order to avoid ribosome traffic further along the transcript[51].

All together, our results indicate that translation initiation is slow compared to elongation (all $\kappa_i = k_i/\alpha < 1$) and the coding sequence interfering with initiation is cleared efficiently (median TIE at 0.8). We also find that translation elongation is largely optimised to avoid traffic (median TEE at 0.99), although one can locally observe high levels of ribosome interference. Interestingly, despite variations in translation initiation efficiency between genes (Figure 5), elongation remains consistently more efficient than initiation (mean TEE > TIE, Figure 8). It is possible that the relative role of elongation and initiation is under evolutionary pressure to allow for an efficient ribosome recruitment and to avoid ribosome interference for efficient transition from initiation to elongation[51].

Perhaps the most surprising result of our study is the variability of the inferred elongation-to-initiation ratios κ_i. We can affirm that there is a correlation between common indices of codon optimality, such as the local tAI, and the estimated elongation-to-initiation ratios (see supplements of Figure 4). However, the large variability of the estimated rates of each individual codon type implies that using those indices for protein synthesis optimisation or other synthetic applications will probably not lead to the expected results. Instead, our findings indicate that codon context in the sequence is as relevant as the particular codon used, and further studies should focus on the discovery of mechanisms giving rise to the codon context dependence. For instance, mRNA secondary structures might be relevant, particularly around the initiation region[48–50,52] or the amino-acid charge at the beginning of the coding sequence[38].

Our method has detected many codons at which the model is incompatible with the ribosome profiling data, particularly for genes for which we estimated high level of ribosome interference (see for instance Figure 6–figure supplement 1). One possibility is that our results are affected by known biases during the bioinformatic analysis[53]. Another source of inconsistency between the model and the data is possibly hidden in the nature of the ribosome profiling technique. Queuing ribosomes generate large footprints[14] that are usually discarded in the experimental pipeline. Intuitively, one would expect that ribosome profiling discarding large footprints would be insensitive to ribosome interference. However, we note that the model is able to capture correlations between ribosomes that are not immediately adjacent to each other. A recent theoretical study by Scott and Szavits-Nossan (2019)[47] showed that a slow codon affects ribosome density over multiple codons, although the effect subsides with the distance from the slow codon. Indeed, NEAR finds evidence of local jamming despite the experimental bias that discards jammed ribosomes. We remark that the high TIE and TEE values at the first 10 codons could also be attributed to the nature of Ribo-seq that exclude disome footprints; a recent study by Diament et *al.* (2018)[17] in fact showed that the largest concentration of disomes in *S. cerevisiae* is at the first 10 codons.

To summarise, we have developed a model-based method for inferring codon-specific elongation rates (relative to the initiation rate) from ribosome profiling data. In addition, we have proposed new measures of translation initiation and elongation efficiencies that quantify the extent of ribosome traffic *in vivo* and can be used to compare different genes and experimental conditions. We believe these new scores will complement the standard indices of translation efficiency and will contribute to the understanding of this complex biological process.

Despite the tremendous importance and potential of ribosome profiling, our work emphasises its limitations when deciphering translation dynamics such as the lack of quantification in physical units and the lack of absolute time scales. These challenges have been recognised and steps have been made recently to combine Ribo-seq with other methods for absolute

quantification such as RNA-seq with spike-ins[1] and pulsed stable isotope labelling of amino acids[18]. Future developments of NEAR will include these data to obtain a more detailed view on translation dynamics. Another key question that quantitative studies using ribosome profiling should address in the future is the role of density normalisation in order to better compare the outcome of different genes and of different organisms.

3 Methods and Materials

3.1 Datasets used

We have analysed ribosome profiling data of *Saccharomyces cerevisiae* from Guydosh et al. (2014)[14], Pop et al. (2014)[20] and Weinberg et al. (2016)[23]: NCBI GEO accession numbers GSM1279568, GSM1557447 and GSE75897 respectively. We downloaded the HDF5 files from Riboviz (https://riboviz.org)[41] and mapped to A-site positions according the following table[42]:

Fragment size	Frame 0	Frame 1	Frame 2
27	15	15	18
28	15	15	18
29	15	X	18
30	15	18	18
31	15	18	18
32	X	18	18
33	18	18	18

After obtaining the A-site read density profiles, our method successfully optimised 345 of the total 346 genes from the Guydosh dataset and for which the experimental ribosome density necessary for normalisation was known from MacKay et *al.* (2004)[43] Analogously, the optimisation was successful for 1051/1053 genes of the Pop dataset and for all 1589 genes in the Weinberg dataset. For the omitted genes the normalisation was not possible because it resulted in ribosome density larger than 1.

3.2 Notations

In this section we summarise the notations used in the paper. The main symbols for densities, rates, and rates relative to initiation are given in Table 1. When the quantity is codon specific we use the suffix $i = 2, \ldots, L$ to identify the codon number (the first codon after the START codon is at $i = 2$, the STOP codon is at $i = L$). Brackets $\{\cdot\}$ indicate a set of values: for instance $\{a_i\}$ is the set of all the values a_i for $i = 2, \ldots, L$.

We emphasise that we use *normalised* densities, in units of ribosomes (A-sites) per codon. The total density is thus the averaged ribosome profile $r = \sum_{i=2}^{L} r_i/(L-1)$, and the number of ribosomes translating an mRNA is $N = r(L-1)$.

Symbol	Meaning
L	length of the mRNA (in codons, including START)
ℓ	length of the ribosome (in codons)
α	initiation rate $[s^{-1}]$
k_i	elongation rate $[s^{-1}]$ of codon i, with $i = 2, \ldots, L-1$
k_L (or β)	termination rate $[s^{-1}]$
$\{k_i\}$	speed profile (elongation) of a given transcript
$\kappa_i = k_i/\alpha$	relative (to initiation) elongation rate of codon i, or *elongation-to-initiation ratio*
$\{\kappa_i\}$	relative (to initiation) elongation profile
r_i	experimental (normalised) density of codon i
$\{r_i\}$	experimental (normalised) density profile
$r = \sum_{i=2}^{L} r_i/(L-1)$	mean density of a given gene
ρ_i	theoretical (normalised) density of codon i
$\{\rho_i\}$	theoretical (normalised) density profile
ρ_i^{sim}	simulated (normalised) density of codon i
$\{\rho_i^{\text{sim}}\}$	simulated (normalised) density profile

Table 1. Summary of the symbols used and their meaning.

3.3 The exclusion process

We model translation by a stochastic process called the totally asymmetric simple exclusion process (TASEP) introduced by MacDonald et *al.*[11,28]. Messenger RNAs are represented by a one-dimensional lattice consisting of L codons labelled by $1,\ldots,L$, whereby each ribosome takes $\ell = 10$ lattice sites. Ribosomes on the lattice are tracked according to the position of their A-site. A codon i that is occupied by the A-site of the ribosome is labelled by A_i and is otherwise labelled by $\emptyset_i$. The model accounts for translation initiation, elongation and termination of individual ribosomes described by the following first-order kinetic reactions:

$$\text{initiation:} \quad \emptyset_2\emptyset_3\ldots\emptyset_{\ell+1} \xrightarrow{\alpha} A_2\emptyset_3\ldots\emptyset_{\ell+1} \tag{3}$$

$$\text{elongation:} \quad A_i\emptyset_{i+1}\ldots\emptyset_{i+\ell+1} \xrightarrow{k_i} \emptyset_i A_{i+1}\emptyset_{i+2}\ldots\emptyset_{i+\ell+1} \tag{4}$$

$$\text{termination:} \quad A_L \xrightarrow{k_L} \emptyset_L. \tag{5}$$

The process is described by the probability density $P(C,t)$ to find a configuration C of ribosomes on an mRNA at a particular time t. By *configuration* we mean a particular arrangement of ribosomes described by the positions $\{A_i\}$ of their A-sites. The time evolution of $P(C,t)$ is governed by the master equation

$$\frac{\mathrm{d}P(C,t)}{\mathrm{d}t} = \sum_{C'\neq C} W(C'\to C)P(C') - \sum_{C'\neq C} W(C\to C')P(C), \tag{6}$$

where $W(C\to C')$ is the rate of transition from C to C'. We assume that translation takes place in the stationary limit so that $\mathrm{d}P(C,t)/\mathrm{d}t = 0$ in which case Eq. (6) becomes a system of linear equations,

$$0 = \sum_{C'\neq C} W(C'\to C)P(C') - \sum_{C'\neq C} W(C\to C')P(C). \tag{7}$$

The three main quantities of interest are the rate of translation J, which measures the amount of proteins produced per unit time, the local ribosome densities ρ_i, which measure the probability of detecting a ribosome at codon i, and total ribosome density ρ, which measures the average number of ribosomes per unit length of the transcript. In the stationary TASEP J, ρ_i and ρ are defined as

$$J = k_L\langle\tau_L\rangle, \quad \rho_i = \langle\tau_i\rangle, \quad \rho = \frac{1}{L-1}\sum_{i=2}^{L}\rho_i, \tag{8}$$

where averaging is performed with respect to the steady-state probability $P(C)$ and τ_i is an occupation number whose value is equal to 1 if codon i is occupied by the A-site of the ribosome and is 0 otherwise.If we ignore premature termination due to ribosome drop-off, then J is constant across the transcript and is equal to the actual rate at which ribosomes initiate translation,

$$J = \alpha P(\text{first } \ell \text{ codons free}) = \alpha\left(1 - \sum_{i=2}^{\ell+1}\rho_i\right), \tag{9}$$

where $P(\text{first } \ell \text{ codons free})$ is the steady-state probability that codons $2,\ldots,\ell+1$ are not occupied by an A-site.

Computing these quantities requires an exact knowledge of $P(C)$, which is known only in the biologically unrealistic case of $\ell = 1$[54]. Instead, we compute J, ρ_i and ρ using two approximation methods: the mean-field approximation developed in MacDonald et *al.* (1968, 1969)[11,28] and initiation-limited approximation developed in Svazits-Nossan et *al* (2018a,b)[12,27]. Details of these methods are presented in Appendix A.

4 Acknowledgements

We would like to thank Guillaume Cambray and Edward Wallace for useful discussions.

References

1. Gorochowski, T. E. & Ellis, T. Designing efficient translation. *Nat. Biotechnol.* **36**, 934–935, DOI: 10.1038/nbt.4257 (2018).
2. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Sci. (New York, N.Y.)* **324**, 218–23, DOI: 10.1126/science.1168978 (2009).

3. Steitz, J. A. Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature* **224**, 957–964, DOI: 10.1038/224957a0 (1969).

4. Kozak, M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* **9**, 5233–5252, DOI: 10.1093/nar/9.20.5233 (1981).

5. Wolin, S. L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *The EMBO J.* **7**, 3559–3569, DOI: 10.1002/j.1460-2075.1988.tb03233.x (1988).

6. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16**, 651–664, DOI: 10.1038/nrm4069 (2015).

7. Fritsch, C. *et al.* Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome research* **22**, 2208–18, DOI: 10.1101/gr.139568.112 (2012).

8. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170, DOI: 10.1038/nmeth.3688 (2016).

9. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife* **2**, DOI: 10.7554/eLife.01179 (2013).

10. Guttman, M., Russell, P., Ingolia, N., Weissman, J. & Lander, E. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**, 240–251, DOI: 10.1016/J.CELL.2013.06.009 (2013).

11. MacDonald, C. T., Gibbs, J. H. & Pipkin, A. C. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* **6**, 1–25, DOI: 10.1002/bip.1968.360060102 (1968).

12. Szavits-Nossan, J., Ciandrini, L. & Romano, M. C. Deciphering mRNA Sequence Determinants of Protein Production Rate. *Phys. Rev. Lett.* **120**, 128101 (2018).

13. Erdmann-Pham, D. D., Dao Duc, K. & Song, Y. S. The Key Parameters that Govern Translation Efficiency. *Cell Syst.* **10**, 183–192, DOI: 10.1016/j.cels.2019.12.003 (2020).

14. Guydosh, N. & Green, R. Dom34 Rescues Ribosomes in 3' Untranslated Regions. *Cell* **156**, 950–962, DOI: 10.1016/J.CELL.2014.02.006 (2014).

15. Li, G. W. How do bacteria tune translation efficiency? *Curr. Opin. Microbiol.* **24**, 66–71, DOI: 10.1016/j.mib.2015.01.001 (2015).

16. Dao Duc, K. & Song, Y. S. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genet.* **14**, 1–32, DOI: 10.1371/journal.pgen.1007166 (2018).

17. Diament, A. *et al.* The extent of ribosome queuing in budding yeast. *PLOS Comput. Biol.* **14**, e1005951, DOI: 10.1371/journal.pcbi.1005951 (2018).

18. Riba, A. *et al.* Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proc. Natl. Acad. Sci. United States Am.* 201817299, DOI: 10.1073/pnas.1817299116 (2019).

19. Siwiak, M. & Zielenkiewicz, P. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Comput. Biol.* **6**, 4, DOI: 10.1371/journal.pcbi.1000865 (2010).

20. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. systems biology* **10**, 770, DOI: 10.15252/msb.20145524 (2014).

21. Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3**, DOI: 10.7554/eLife.03735 (2014).

22. Gritsenko, A. A., Hulsman, M., Reinders, M. J. T. & de Ridder, D. Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLOS Comput. Biol.* **11**, e1004336, DOI: 10.1371/journal.pcbi.1004336 (2015).

23. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports* **14**, 1787–1799, DOI: 10.1016/j.celrep.2016.01.043 (2016).

24. Sharma, A. K. *et al.* A chemical kinetic basis for measuring translation initiation and elongation rates from ribosome profiling data. *PLOS Comput. Biol.* **15**, e1007070, DOI: 10.1371/journal.pcbi.1007070 (2019).

25. Sin, C., Chiarugi, D. & Valleriani, A. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Res.* **44**, 2528–2537, DOI: 10.1093/nar/gkw137 (2016).

26. Bonnin, P., Kern, N., Young, N. T., Stansfield, I. & Romano, M. C. Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile. *PLOS Comput. Biol.* **13**, e1005555, DOI: 10.1371/journal.pcbi.1005555 (2017).

27. Szavits-Nossan, J., Romano, M. C. & Ciandrini, L. Power series solution of the inhomogeneous exclusion process. *Phys. Rev. E* **97**, 1–13, DOI: 10.1103/PhysRevE.97.052139 (2018).

28. MacDonald, C. T. & Gibbs, J. H. Concerning the kinetics of polypeptide synthesis on polyribosomes. *Biopolymers* **7**, 707–725, DOI: 10.1002/bip.1969.360070508 (1969).

29. von der Haar, T. Mathematical and computational modelling of ribosomal movement and protein synthesis: an overview. *Comput. Struct. Biotechnol. J.* **1**, e201204002, DOI: 10.5936/CSBJ.201204002 (2012).

30. Zur, H. & Tuller, T. Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. *Nucleic Acids Res.* **44**, gkw764, DOI: 10.1093/nar/gkw764 (2016).

31. Mitarai, N., Sneppen, K. & Pedersen, S. Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization. *J. Mol. Bio.* **382**, 236–245, DOI: 10.1016/J.JMB.2008.06.068 (2008).

32. Reuveni, S., Meilijson, I., Kupiec, M., Ruppin, E. & Tuller, T. Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model. *PLoS Comput. Biol.* **7**, e1002127, DOI: 10.1371/journal.pcbi.1002127 (2011).

33. Ciandrini, L., Stansfield, I. & Romano, M. C. Ribosome Traffic on mRNAs Maps to Gene Ontology: Genome-wide Quantification of Translation Initiation Rates and Polysome Size Regulation. *PLoS Comput. Biol.* **9** (2013).

34. Sharma, A. K., Ahmed, N. & O'Brien, E. P. Determinants of translation speed are randomly distributed across transcripts resulting in a universal scaling of protein synthesis times. *Phys. Rev. E* **97**, 022409, DOI: 10.1103/PhysRevE.97.022409 (2018).

35. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–1601, DOI: 10.1016/j.cell.2013.05.049 (2013).

36. Rudorf, S. & Lipowsky, R. Protein synthesis in E. coli: Dependence of codon-specific elongation on tRNA concentration and codon usage. *PLoS ONE* **10**, 1–22, DOI: 10.1371/journal.pone.0134994 (2015).

37. Gorgoni, B., Ciandrini, L., McFarland, M. R., Romano, M. C. & Stansfield, I. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. *Nucleic Acids Res.* **44**, 9231–9244 (2016).

38. Dao Duc, K., Saleem, Z. H. & Song, Y. S. Theoretical analysis of the distribution of isolated particles in totally asymmetric exclusion processes: Application to mRNA translation rate estimation. *Phys. Rev. E* **97**, DOI: 10.1103/PhysRevE.97.012106 (2018).

39. Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S. & Press, W. H. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet.* **11**, 1–25, DOI: 10.1371/journal.pgen.1005732 (2015).

40. Duncan, C. D. S. & Mata, J. Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Sci. Reports* **7**, 10331, DOI: 10.1038/s41598-017-10650-1 (2017).

41. Carja, O., Xing, T., Wallace, E. W., Plotkin, J. B. & Shah, P. Riboviz: Analysis and visualization of ribosome profiling datasets. *BMC Bioinforma.* **18**, DOI: 10.1186/s12859-017-1873-8 (2017).

42. Ahmed, N. *et al.* Identifying A- and P-site locations on ribosome-protected mRNA fragments using Integer Programming. *Sci. Reports 2019 9:1* **9**, 6256, DOI: 10.1038/s41598-019-42348-x (2019).

43. MacKay, V. L. *et al.* Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Mol. & cellular proteomics : MCP* **3**, 478–89, DOI: 10.1074/mcp.M300129-MCP200 (2004).

44. Reis, M. d., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044, DOI: 10.1093/nar/gkh834 (2004).

45. Chu, D. *et al.* Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO J.* DOI: 10.1002/embj.201385651 (2014).

46. Verma, M. *et al.* Short translational ramp determines efficiency of protein synthesis. *bioRxiv* 571059, DOI: 10.1101/571059 (2019).

47. Scott, S. & Szavits-Nossan, J. Power series method for solving TASEP-based models of mRNA translation. *Phys. biology* **17**, DOI: 10.1088/1478-3975/ab57a0 (2019).

48. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258, DOI: 10.1126/science.1170160 (2009).

49. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Bluthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **9**, 675–675, DOI: 10.1038/msb.2013.32 (2014).

50. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005–1015, DOI: 10.1038/nbt.4238 (2018).

51. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354, DOI: 10.1016/j.cell.2010.03.031 (2010).

52. Li, J. J., Chew, G.-L. & Biggin, M. D. Quantitative Principles of cis-translational control by general mRNA sequence features in eukaryotes. *Genome Biol.* DOI: doi.org/10.1186/s13059-019-1761-9 (2019).

53. Bartholomäus, A., Del Campo, C. & Ignatova, Z. Mapping the non-standardized biases of ribosome profiling. *Biol. Chem.* **397**, 23–35, DOI: 10.1515/hsz-2015-0197 (2016).

54. Derrida, B., Evans, M. R., Hakim, V. & Pasquier, V. Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *J. Phys. A: Math. Gen.* **26**, 1493–1517, DOI: 10.1088/0305-4470/26/7/011 (1993).

55. Shaw, L. B., Sethna, J. P. & Lee, K. H. Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* DOI: 10.1103/PhysRevE.70.021901 (2004).

56. Johnson, S. G. The nlopt nonlinear-optimization package.

57. Powel, M. J. The bobyqa algorithm for bound constrained optimization without derivatives.Technical Report (2009).

A Analytic results in the steady state as a basis of NEAR

The model has been defined in the Materials and Methods section of the main text. Here we explain how to compute the ribosomal current J and the local ribosome density ρ_i using two approximate approaches: the mean-field approximation developed in MacDonald et al. (1968, 1969)[11,28] and Shaw et *al.* (2004)[55], and initiation-limited approximation developed in Svazits-Nossan et *al* (2018a,b)[12,27].

The theory summarised in this Appendix constitutes the basis of NEAR, whose procedure is given in the section Non-Equilibrium Analysis of Ribo-seq (NEAR) of the main text.

A.1 Ribosome density and current in the mean-field approximation

In the mean-field approximation, correlations between ribosomes are ignored, which leads to the following system of equations for J and ρ_i,

$$J = \alpha\left(1 - \sum_{s=1}^{\ell} \rho_{i+s}\right), \quad i = 1 \tag{10a}$$

$$J = k_i \rho_i \frac{1 - \sum_{s=1}^{\ell} \rho_{i+s}}{1 - \sum_{s=1}^{\ell} \rho_{i+s} + \rho_{i+\ell}}, \quad i = 2, \ldots, L-\ell \tag{10b}$$

$$J = k_i \rho_i, \quad i = L-\ell+1, \ldots, L. \tag{10c}$$

It is straightforward to invert these equations to get the ratio k_i/α as a function of the local densities $\rho_2, \ldots, \rho_L$:

$$\frac{k_i}{\alpha} = \frac{\left(1 - \sum_{s=1}^{\ell} \rho_{1+s}\right)\left(1 - \sum_{s=1}^{\ell} \rho_{i+s} + \rho_{i+\ell}\right)}{\rho_i \left(1 - \sum_{s=1}^{\ell} \rho_{i+s}\right)}, \quad i = 2, \ldots, L-\ell \tag{11a}$$

$$\frac{k_i}{\alpha} = \frac{1 - \sum_{s=1}^{\ell} \rho_{1+s}}{\rho_i}, \quad i = L-\ell+1, \ldots, L. \tag{11b}$$

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

A.2 Ribosome density and current in the initiation-limited approximation

Recently we introduced a power-series method for computing J, ρ_i and ρ by expanding steady-state probability $P^*(C)$ in the initiation rate α,

$$P^*(C) = \sum_{n=0}^{\infty} c_n(C)\alpha^n. \tag{12}$$

The method then delineates how to compute the coefficients $c_n(C)$ recursively starting from $n = 0$. Due to its recursive nature, the power-series method is much faster than a stochastic simulation. This is absolutely crucial in order to infer $\{k_i/\alpha\}$ from $\{\rho_i\}$, which is the main part of NEAR.

It turns out that most coefficients $c_n(C)$ for small n are in fact equal to zero, which considerably simplifies the calculation of $P(C)$. Specifically, if $N(C)$ denotes the number of ribosomes in a configuration C then

$$c_n(C) = 0 \text{ if } N(C) > n. \tag{13}$$

This non-trivial result follows from a graph-theoretical interpretation of Markov chains applied to the TASEP, see Szavits-Nossan et *al.* (2018a)[12] and Szavits-Nossan et *al.* (2018a)[27] for more details. Another useful relation is

$$\sum_C c_0(C) = 1, \quad \sum_C c_n(C) = 1,\ n \geq 1, \tag{14}$$

which follows from the fact that the sum of $P(C)$ over all C is equal to 1.

If initiation is slow we can keep the first K terms in the series (12) and ignore all the rest, which we call the initiation-limited approximation,

$$P^*(C) \approx \sum_{n=0}^{K} c_n(C)\alpha^n. \tag{15}$$

In this work we compute coefficients up to and including the third order ($K = 3$), which we present in detail below. According to Eq. (13), for $K = 3$ we only need to consider configurations with at most three ribosomes on the lattice. 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).

For $n = 0$ (zeroth order), $c_0(C) = 1$ if the lattice is empty ($C = \emptyset$) and is equal to 0 otherwise, which leads to $P(C) = 1$ if we ignore higher-order terms. 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 ribosome. The corresponding coefficients $c_1(\emptyset)$ and $c_1(A_i)$ are equal to

$$c_1(\emptyset) = -\sum_{i=2}^{L} \frac{1}{k_i}, \quad c_1(A_i) = \frac{1}{k_i},\ i = 2, \ldots, L. \tag{16}$$

These coefficients yield the following expressions for J, ρ_i and ρ in the first-order approximation

$$J = \alpha, \quad \rho_i = \frac{\alpha}{k_i}, \quad \rho = \frac{1}{L-1}\sum_{i=2}^{L}\frac{\alpha}{k_i}. \tag{17}$$

We note that the relationship $\rho_i \approx \alpha/k_i$ between ribosome densities and elongation rates that we derived above is often used in the interpretation of ribosome profiling data. Because the first order considers configurations with at most one ribosome per lattice, this result neglects the interference between ribosomes and is thus valid only for very small initiation rates.

For $n = 2$ (second order), $c_2(C) \neq 0$ only if C contains at most two ribosomes. The equations for $c_2(C)$ are more involved 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

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \theta[i] = \begin{cases} 1 & i \geq 0 \\ 0 & i < 0 \end{cases}. \tag{18}$$

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

$$c_2(A_iA_j) = \frac{\delta_{i,2}}{e_0(A_iA_j)}c_1(A_j) + \frac{\theta[i-3]k_{i-1}}{e_0(A_iA_j)}c_2(A_{i-1}A_j) + \frac{\theta[j-1-i-\ell]k_{j-1}}{e_0(A_iA_j)}c_2(A_iA_{j-1}), \tag{19}$$

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 $i = L$,

$$c_2(A_i) = \frac{\delta_{i,2}}{k_i}c_1(\emptyset) + \frac{\theta[i-3]k_{i-1}}{k_i}c_2(A_{i-1}) + \frac{\theta[L-\ell-i]k_L}{k_i}c_2(A_iA_L). \tag{20}$$

Finally, once all $c_2(A_iA_j)$ and $c_2(A_i)$ are computed, $c_2(\emptyset)$ follows from Eq. (14),

$$c_2(\emptyset) = -\sum_{i=2}^{L-\ell}\sum_{j=i+\ell}^{L} c_2(A_iA_j) - \sum_{i=2}^{L} c_2(A_i). \tag{21}$$

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

$$J = \alpha - k_L\left(\sum_{j=2}^{L-\ell} c_2(A_jA_L) + c_2(A_L)\right)\alpha^2 = \alpha - \sum_{j=2}^{\ell+1}\frac{\alpha^2}{k_i} \tag{22}$$

$$\rho_i = \frac{\alpha}{k_i} + \left(\sum_{j=2}^{i-\ell} c_2(A_jA_i) + \sum_{j=i+\ell}^{L} c_2(A_iA_j) + c_2(A_i)\right)\alpha^2. \tag{23}$$

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

$$\begin{aligned} c_3(A_iA_jA_m) = {} & \frac{\delta_{i,2}}{e_0(A_iA_jA_m)}c_2(A_jA_m) + \frac{\theta[i-3]k_{i-1}}{e_0(A_iA_jA_m)}c_3(A_{i-1}A_jA_m) \\ & + \frac{\theta[j-1-i-\ell]k_{j-1}}{e_0(A_iA_jA_m)}c_3(A_iA_{j-1}A_m) + \frac{\theta[m-1-j-\ell]k_{m-1}}{e_0(A_iA_jA_m)}c_3(A_iA_jA_{m-1}), \end{aligned} \tag{24}$$

where $e_0(A_iA_jA_m)$ is given by

$$e_0(A_iA_jA_m) = \theta[j-\ell-i-1]k_i + \theta[m-\ell-j-1]k_j + k_m. \tag{25}$$

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 $c_3(A_iA_jA_m)$ are found, the equations for two-particle coefficients $c_3(A_iA_j)$ are solved recursively starting from $i = 2$ and $j = \ell+2$,

$$\begin{aligned} c_3(A_iA_j) = {} & \frac{\delta_{i,2}}{e_0(A_iA_j)}c_2(A_j) + \frac{\theta[i-3]k_{i-1}}{e_0(A_iA_j)}c_3(A_{i-1}A_j) \\ & + \frac{\theta[j-1-i-\ell]k_{j-1}}{e_0(A_iA_j)}c_3(A_iA_{j-1}) + \frac{\theta[L-\ell-j]k_L}{e_0(A_iA_j)}c_3(A_iA_jA_L). \end{aligned} \tag{26}$$

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

$$c_3(A_i) = \frac{\delta_{i,2}}{k_i}c_2(\emptyset) + \frac{\theta[i-3]k_{i-1}}{k_i}c_3(A_{i-1}) + \frac{\theta[L-\ell-i]k_L}{k_i}c_3(A_iA_L). \tag{27}$$

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

$$c_3(\emptyset) = -\sum_{i=2}^{L-2\ell}\sum_{j=i+\ell}^{L-\ell}\sum_{m=j+\ell}^{L} c_3(A_iA_jA_m) - \sum_{i=2}^{L-\ell}\sum_{j=i+\ell}^{L} c_3(A_iA_j) - \sum_{i=2}^{L} c_3(A_i). \tag{28}$$

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

$$J = \alpha - \sum_{j=2}^{\ell+1}\frac{1}{k_i}\alpha^2 + k_L\left(\sum_{j=2}^{L-2\ell}\sum_{m=j+\ell}^{L-\ell} c_3(A_jA_mA_L) + \sum_{j=2}^{L-\ell} c_3(A_jA_L) + c_3(A_L)\right)\alpha^3, \tag{29}$$

$$\rho_i = f_i^{(1)}\alpha + f_i^{(2)}\alpha^2 + f_i^{(3)}\alpha^3, \tag{30}$$

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

$$f_i^{(1)} = \frac{1}{k_i}, \quad f_i^{(2)} = \sum_{j=2}^{i-\ell} c_2(A_jA_i) + \sum_{j=i+\ell}^{L} c_2(A_iA_j) + c_2(A_i),$$
$$f_i^{(3)} = \sum_{j=i+\ell}^{L-\ell}\sum_{m=j+\ell}^{L} c_3(A_iA_jA_m) + \sum_{j=2}^{i-\ell}\sum_{m=i+\ell}^{L} c_3(A_jA_iA_m) + \sum_{j=2}^{i-2\ell}\sum_{m=j+\ell}^{i-\ell} c_3(A_jA_mA_i)$$
$$+ \sum_{j=2}^{i-\ell} c_3(A_jA_i) + \sum_{j=i+\ell}^{L} c_3(A_iA_j) + c_3(A_i). \tag{31}$$

From the stationary master equation (7) in the Material and Methods section of the main text, it follows that $P(C)$ and thus ρ_i are functions of $\{\kappa_i\} = \{k_i/\alpha\}$ only. Using this fact we can absorb α^n into $f^{(n)}(C)$ and rewrite Eq. (30) as

$$\rho_i = g_i^{(1)} + g_i^{(2)} + g_i^{(3)}, \quad g_i^{(n)}(\kappa_2,\ldots,\kappa_L) = \alpha^n f_i^{(n)}(k_2\ldots,k_L),\ n = 1,2,3. \tag{32}$$

B Details of the NEAR method

The method of inferring elongation-to-initiation ratios from ribosome profiling data is summarised in the Section Non-equilibrium Analysis of Ribo-seq (NEAR) of the main text. Here we provide further details of the stochastic simulations, absolute normalisation of Ribo-seq data and non-linear optimisation for computing elongation-to-initiation ratios $\{\kappa_i\}$.

B.1 Stochastic 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

$$\rho_i^{\text{sim}} = \frac{1}{T}\sum_{n=1}^{M} \tau_i^{(n)} dt^{(n)}, \tag{33}$$

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.

B.2 Normalisation of Ribo-seq data

Genes from the three datasets analysed were selected to have at least 10 "reads" per codon on average.

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

$$r_i = \rho \frac{N_i}{\sum_{n=2}^{L} N_n}, \quad i = 2,\ldots,L \tag{34}$$

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. (34). The normalisation procedure failed for only three genes across all three datasets, because the normalised ribosome density r_i was larger than 1 at some codons.

B.3 Non-linear optimisation for computing elongation-to-initiation ratios

We looked for $\kappa_i = k_i/\alpha$ such that the theoretical ribosome density ρ_i predicted from the initiation-limited approximation in Eq. (30) matched the experimental density r_i,

$$\rho_i(\kappa_2,\ldots,\kappa_L) = r_i, \quad i = 2,\ldots,L. \tag{35}$$

The procedure of finding κ_i consisted of four steps:

1. First we estimated the values of κ_i^{MF} in the mean-field approximation according to Eq. (11).

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).

3. We then compared $\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\})$ to r_i for each $i = 2, \ldots, L$ and identified codons $x(1), \ldots, x(N)$ for which the percentage error between $\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\})$ and r_i was less than 5%.

4. Finally, we solved a nonlinear least squares optimisation problem using $\{\kappa_i^{\text{MF}}\}$ as a starting point, which consisted of finding $\kappa_{x(1)}, \ldots, \kappa_{x(N)}$ for which the sum of squares

$$S(\kappa_2, \ldots, \kappa_L) = \sum_{=2}^{L} (\rho_j(\kappa_2, \ldots, \kappa_L) - r_j)^2 \tag{36}$$

was minimal, where $\kappa_i = \kappa_i^{\text{MF}}$ for codons $i \neq x(1), \ldots, x(N)$ that were not selected for optimisation.

The nonlinear optimisation was preformed using NLopt library from Ref.[56]. We used a local derivative-free algorithm called BOBYQA which was developed in Ref.[57]. In order to prevent unrealistic values of κ_i and to speed up the optimisation procedure, we restricted the search to $10^{-3} \leq \kappa_{x(1)}, \ldots, \kappa_{x(N)} \leq 10^6$. The optimisation search was preformed until one of the following two stopping criteria was met: (1) the fractional error $\Delta S/S$ between two iterations was less than 10^{-8} and (2) the total run time exceeded 60 minutes. These numbers were chosen due to time constraints 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.

B.4 Quality check of the estimated elongation-to-initiation ratios

In this section we detail the quality check that we carried out for each value of the inferred $\kappa_i = k_i/\alpha$. For each gene for which the normalisation of Ribo-seq data was successful we preformed the following seven checks:

1. *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}

$$S_{\text{MF}} = \sum_{i=2}^{L} \left(\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\}) - r_i\right)^2, \tag{37}$$

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

$$S_{\text{opt}} = \sum_{i=2}^{L} \left(\rho_i^{\text{sim}}(\{\kappa_i\}) - r_i\right)^2, \tag{38}$$

where $\{\rho_i^{\text{sim}}(\{\kappa_i\})\}$ is the simulated density profile computed using the inferred elongation-to-initiation ratios $\{\kappa_i\}$. If $S_{\text{opt}} < S_{\text{MF}}$ 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 the initiation-limited approximation cannot be used).

3. *Applicability of the power series method.* We set a tolerance ε_{PSM} for the ribosome density ρ_i^{PSM} predicted by the power series method in Eq. (30). For each codon we check if $|\rho_i^{\text{PSM}} - \rho_i^{\text{sim}}(\{\kappa_i\})|/\rho_i^{\text{sim}} < \varepsilon_{\text{PSM}}$. If that is the case the power series approximation holds and the method is reliable. We set $\varepsilon_{\text{PSM}} = 0.1$.

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 ε_{EXP}) by checking if $|r_i - \rho_i^{\text{sim}}(\{\kappa_i\})|/r_i < \varepsilon_{\text{EXP}}$, where we set $\varepsilon_{\text{EXP}} = 0.05$. Codons that pass this check are kept and considered reliable only if $\kappa_i < \kappa_{\text{thr}}$. This last check is to discard values of κ_i that are deemed suspiciously large and are likely unrealistic. 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 $\kappa_{\text{thr}} = 500$ corresponding to an elongation rate k= 60/s (assuming $\alpha = 0.12$/s).

5. *Problematic codons.* If $|r_i - \rho_i^{\text{sim}}(\{\kappa_i\})|/r_i > \varepsilon_{\text{EXP}}$ 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 $\{\rho_i^{\text{sim}}(\{\kappa_i\})\}$.

   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 Files 1,2,3.

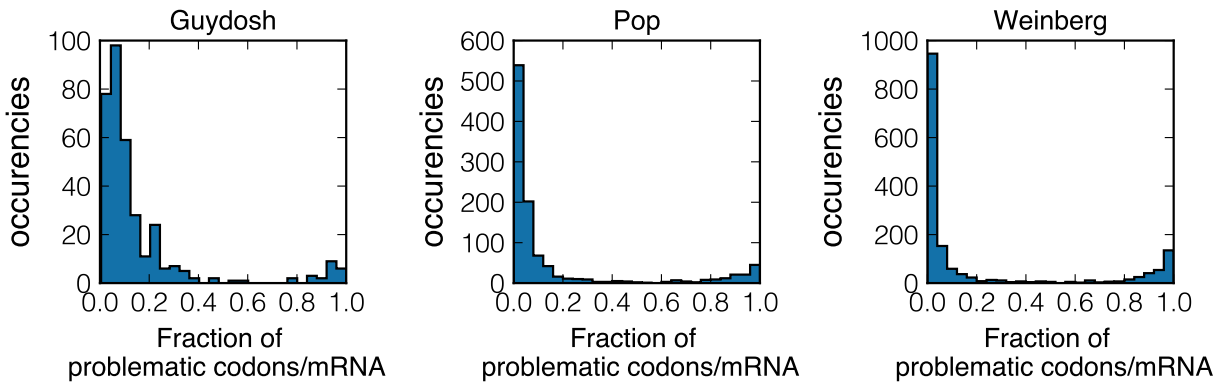

Figure 9. Percentage of codons (per mRNA transcript) that are excluded from the analysis because of the inconsistency between the underlying model and experimental densities (codons that did not pass the point 5 of the quality check). This is a signature of local ribosome interference that cannot be detected by ribosome profiling, and is highlighted by NEAR.

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 $\{\kappa_i^{\text{MF}}\}$ will be considered. If $|r_i - \{\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\})\}|/r_i < \varepsilon_{\text{EXP}}$, then κ_i^{MF} 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.

B.5 Computing TEE profiles

We remind the following relation from the main text: $t_i(\text{total}) = t_i(\text{intrinsic}) + t_i(\text{collision})$. By definition of TEE_i,

$$\frac{1}{t_i(\text{total})} = \frac{1}{t_i(\text{intrinsic})}\,\text{TEE}_i\,. \tag{39}$$

The ribosomal current is given by

$$J = k_i \rho_i P(\ell \text{ free codons } |i)\ , \tag{40}$$

where $P(\ell \text{ free codons } |i)$ is the probability that the $i+1 \ldots i+\ell$ codons are not occupied (given that a ribosome's A-site is at site i). If we identify the elongation rate k_i with $1/t_i(\text{intrinsic})$ and J/ρ_i with the effective elongation rate $\frac{1}{t_i(\text{total})}$, then by comparing Eqs. (39) and (40) we find that the TEE_i is equivalent to $P(\ell \text{ free codons } |i)$. Based on the definition of the TIE=J/α, we obtain the TEE profile by computing $\text{TEE}_i = \text{TIE}/(\kappa_i \rho_i^{\text{sim}})$ at each codon i at which κ_i passed the quality check. Therefore the profile is "broken" when NEAR cannot find a reliable estimate for the elongation-to-initiation ratio κ_i. 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 ρ_i^{sim} comes with its own statistical error, TEE may occasionally be larger than 1. In those cases we capped the value to 1.

B.6 Relationship between TIE and TEE

Translation initiation index measures the unoccupancy of the first $\ell = 10$ codons and is thus related to κ_i and TEE_i in that region. The TIE is defined as J/α, where J is the ribosome current i.e. the number of successful translation initiations per unit time. The ribosome current is given by $J = \alpha P(\text{first } \ell \text{ codons free})$, where $P(\text{first } \ell \text{ codons free})$ is the probability that the first $\ell = 10$ codons are free. The latter is given by

$$P(\text{first } \ell \text{ codons free}) = 1 - \sum_{i=2}^{\ell+1} \rho_i, \tag{41}$$

From there we conclude that

$$\mathrm{TIE} = P(\text{first}\ell\text{codons free}) = 1 - \sum_{i=2}^{\ell+1} \rho_i = 1 - \sum_{i=2}^{\ell+1} \frac{\mathrm{TIE}}{\kappa_i \mathrm{TEE}_i}, \tag{42}$$

where in the last step we used that $\rho_i = \mathrm{TIE}/(\kappa_i \mathrm{TEE}_i)$. Finally, the relationship between TIE, κ_i and TEE_i of the first ℓ codons is given by

$$\mathrm{TIE} = \frac{1}{1 + \sum_{i=2}^{\ell+1} \frac{1}{\kappa_i \mathrm{TEE}_i}}. \tag{43}$$

This expression reveals that efficient translation initiation requires both fast elongation ($\kappa_i > 1$) and no ribosome queuing (TEE_i close to 1) within the first $\ell = 10$ codons.

C Figure Supplements

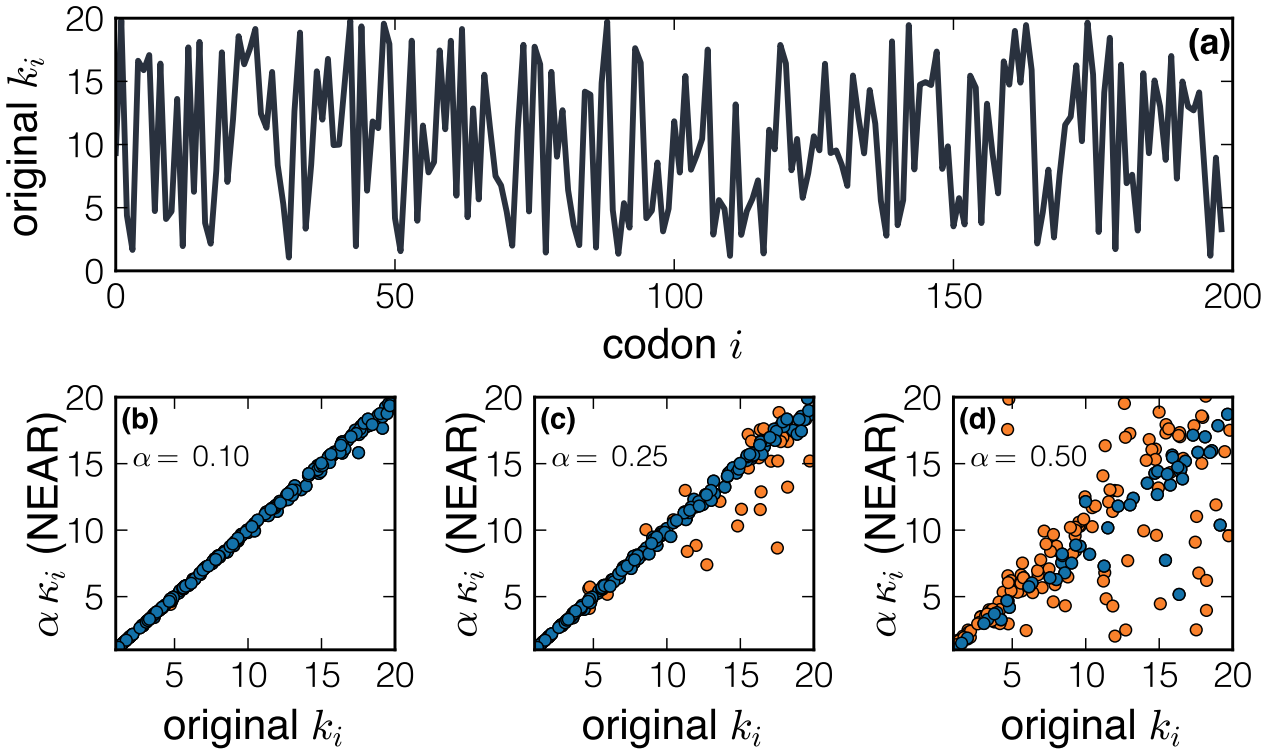

Figure 3–figure supplement 1: Testing NEAR on a random sequence of rates $\{k_i\}$. The $\{k_i\}$ profile is shown in panel (a). Panels (b)-(c)-(d) show the scatter plot between the rates $\alpha\,\kappa_i$ estimated by NEAR (y-axis) and the original rates $\{k_i\}$ (x-axis) with increasing values of 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[33].

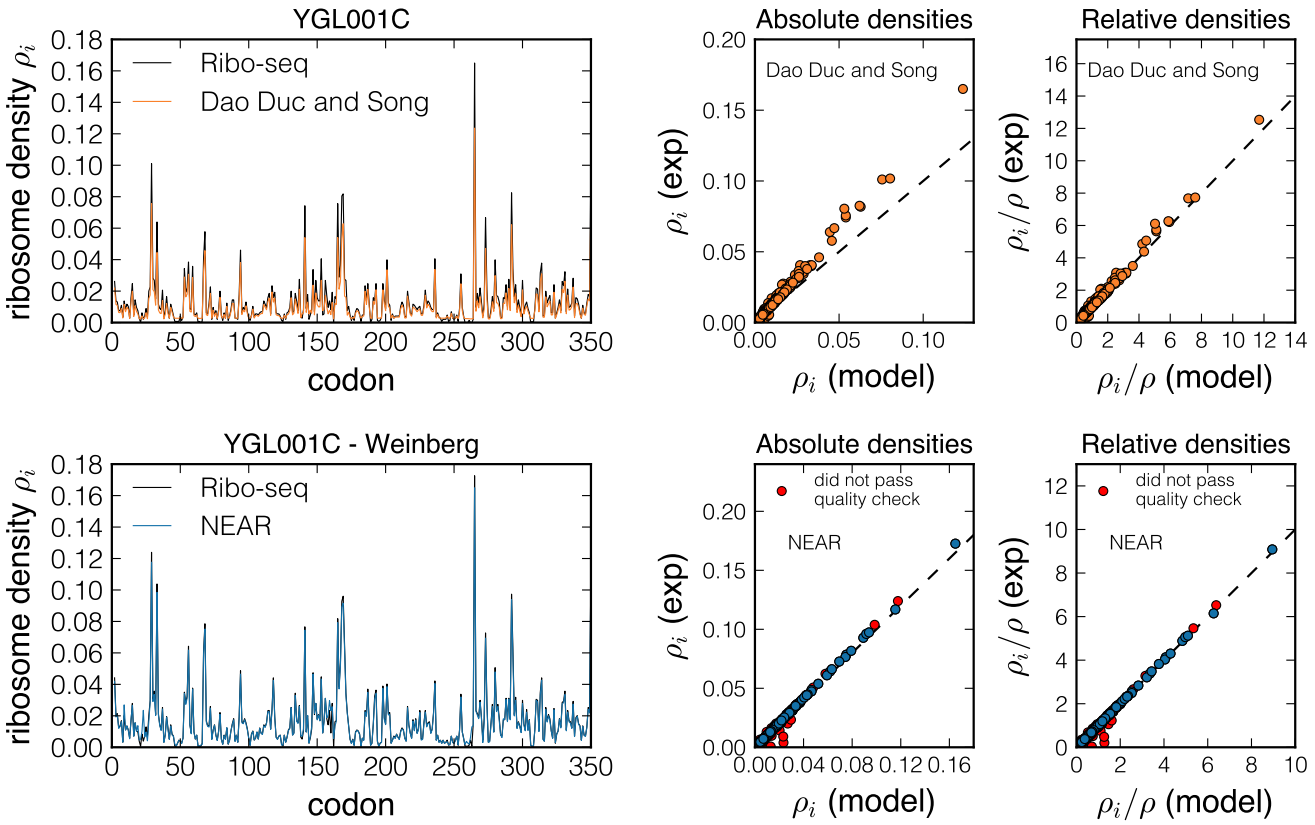

Figure 3–figure supplement 2: We compare results from Dao Duc and Song (2018)[16] (first row) and NEAR (second row). Although the density profile from simulations with rates obtained by optimising the relative density profile $\{\rho_i/\rho\}$ looks comparable with the experimental one (top-left panel), there is a bias between absolute densities (top-middle panel). This follows from the fact that $\{\rho_i/\rho\}$ remains unchanged if we multiply all ρ_i by a constant factor, which in turn means different densities ρ_i and therefore different κ_i. In the bottom-left panel, ribosome density for the same gene predicted by NEAR (blue line) is compared to the experimental data (black line). NEAR is founded on the optimisation of *absolute* densities, thus both absolute and relative densities match the experimental data (bottom-middle and right panel). In the top row we computed absolute densities as $(TE/100*0.83)*L*R_i/R$, where TE is the translation efficiency, L is the length of the mRNA in codons, R_i is the number of reads for codon i and R is the total number of reads of that gene; data from Dao Duc and Song (2018)[16]. In the bottom row the experimental densities are the r_i introduced in the main text. We highlight in red (bottom-middle and right panel), the codons that did not pass the quality check of our method.

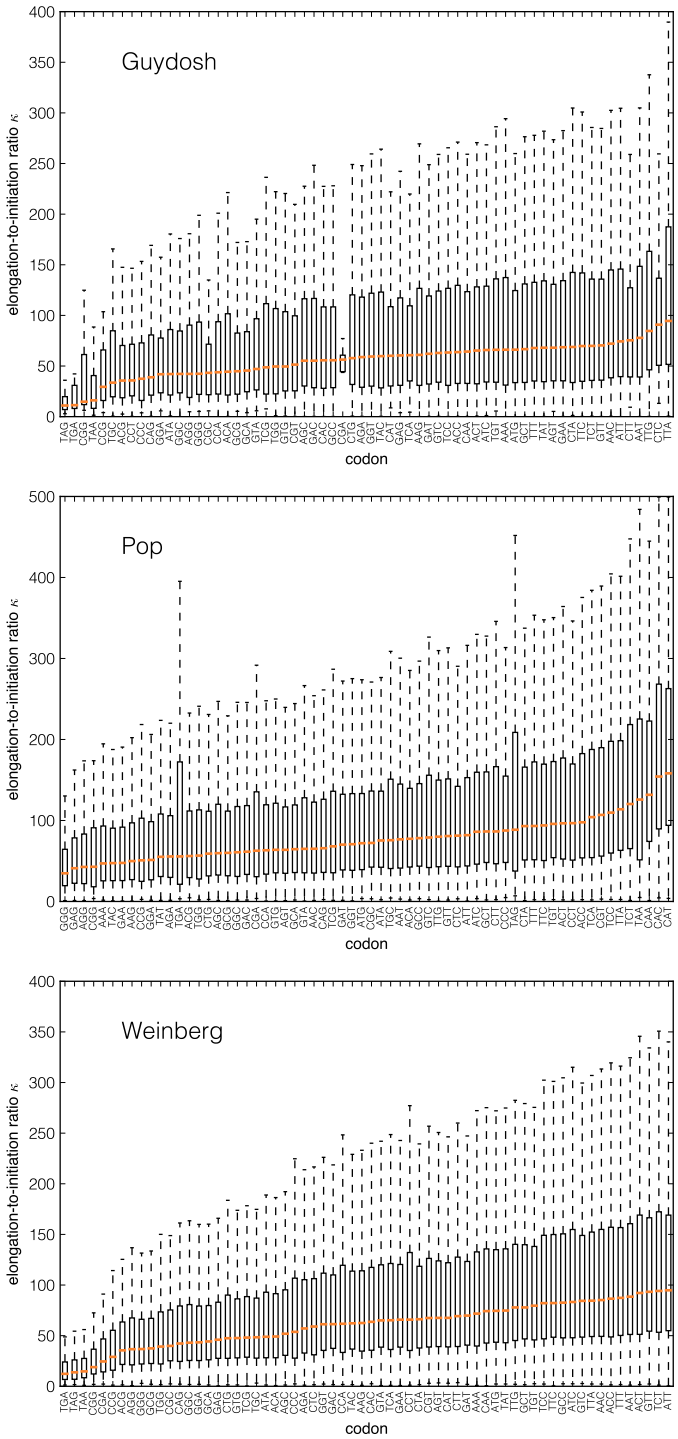

Figure 4–figure supplement 1: Distributions of elongation-to-initiation ratios κ for all codon types. We gather the ratios $\kappa_i = k_i/\alpha$ for each codon-type, and plot their distributions. In principle codons from different genes cannot be compared because they have a different initiation rate α. However, we notice a small variability in the estimates of STOP codons (UAA, UAG, UAA) for the Guydosh and Weinberg datasets. This is reasonable since STOP codons are supposed to be less context dependent. We also remark that, as opposed to what is generally believed, STOP codons are the slowest, but still ~ 10 times faster compared to initiation. The boxplots show the quartiles $Q1$ and $Q3$ ($\Delta Q = Q3 - Q1$) of the distribution, the median is the orange horizontal line; the whiskers extend to the most extreme, non-outlier data points (i.e. to the last data point less than $Q3 + 1.5\Delta Q$ and greater than $Q1 - 1.5\Delta Q$). The Guydosh dataset is the smallest one, and only 10 CGA codons passed the quality check.

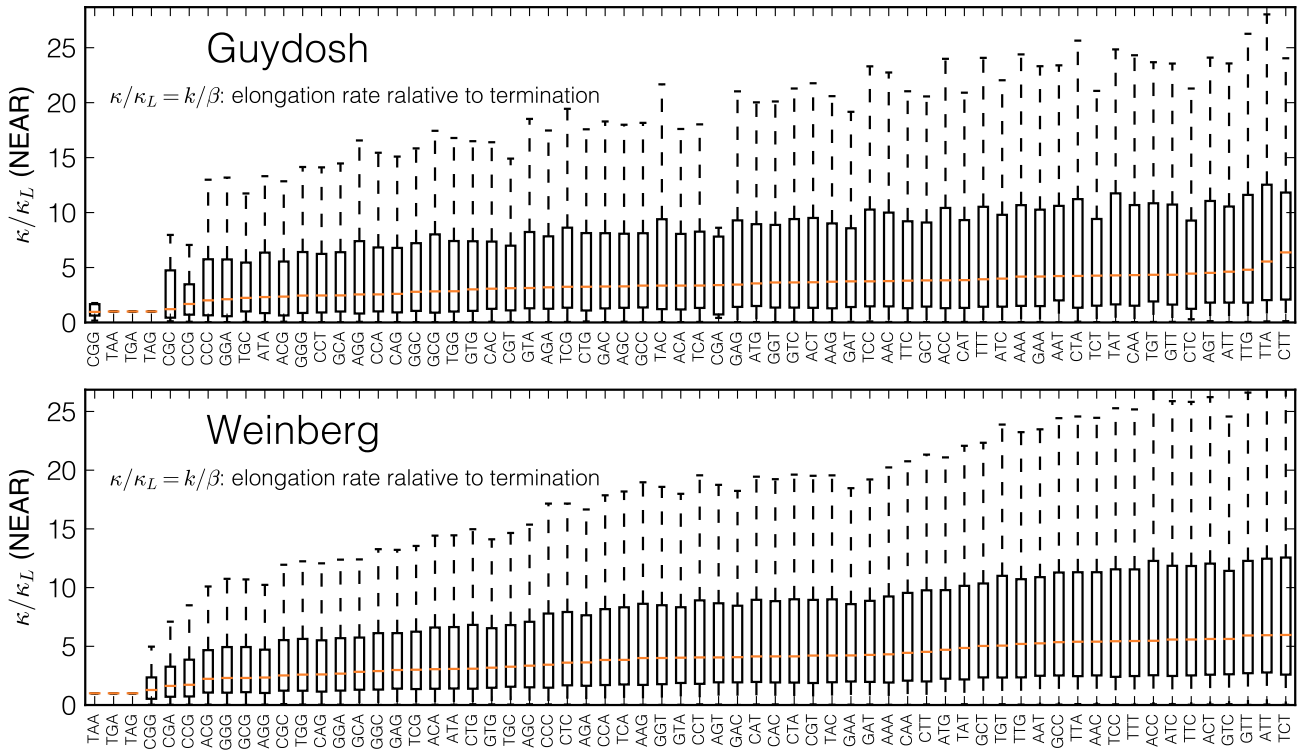

Figure 4–figure supplement 2: Boxplot of the distribution of the elongation-to-termination ratios κ/κ_L for each codon type for the Guydosh and Weinberg dataset. There is a ≈5-fold difference in the median value of κ/κ_L between the slowest and the fastest codon, in accordance with a common view that translation elongation is a non-uniform process. However, in we also find a large variability in the values of κ/κ_L belonging to the same codon type. This variability suggests that the elongation speed of individual codons is only partially determined by their codon type. The boxplots show the quartiles $Q1$ and $Q3$ ($\Delta Q = Q3 - Q1$) of the distribution, the median is the orange horizontal line; the whiskers extend to the most extreme, non-outlier data points (i.e. to the last data point less than $Q3 + 1.5\Delta Q$ and greater than $Q1 - 1.5\Delta Q$). Only 11 CGG codons of the Guydosh dataset can be used to plot these distributions (both CGG and STOP codon passed the quality check).

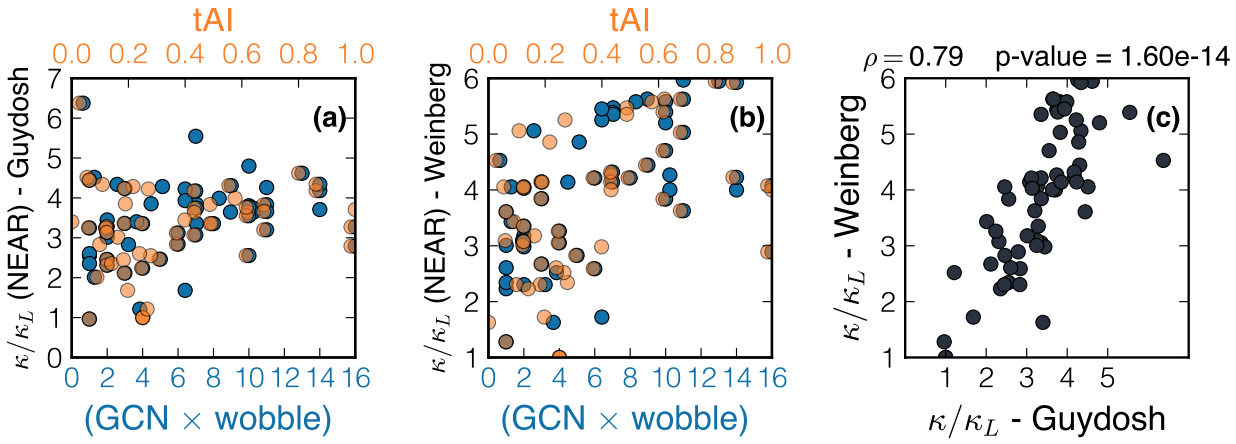

Figure 4–figure supplement 3: We compare median values of κ/κ_L against two common measures of tRNA availability, a codon-dependent rate of translation based on the tRNA gene copy number (GCN) corrected for the wobble base pairing from Weinberg et *al.* (2016)[23], and the tRNA adaptation index (tAI)[44] for Guydosh (a) and Weinberg (b) datasets. We find a moderate correlation between the median of the κ/κ_L distributions and both measures of tRNA availability (GCN $\times$ wobble: Spearman $\rho = 0.34$, $p = 0.008$ - Guydosh; $\rho = 0.57$, $p = 1.4 \times 10^{-6}$ - Weinberg; tAI: $\rho = 0.22$, $p = 0.1$ - Guydosh; $\rho = 0.46$, $p = 3 \times 10^{-4}$); this suggests that the elongation speed of individual codons is only partially determined by their codon type. Panel (c) shows the comparison between the two datasets.

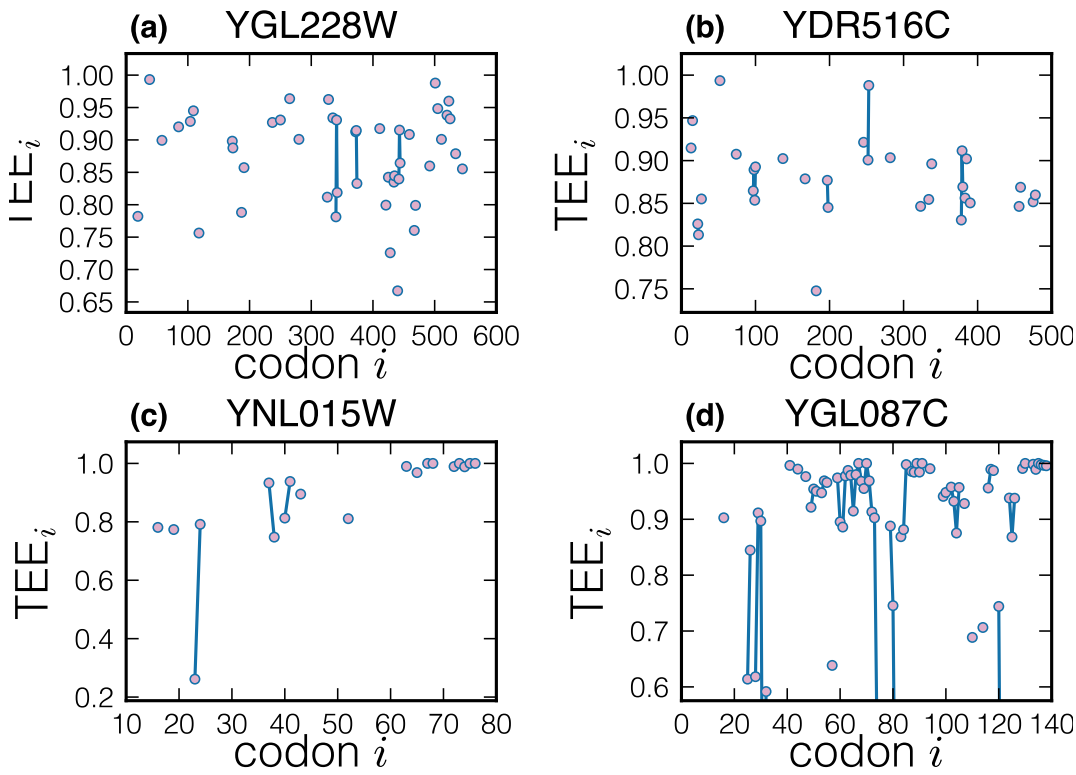

Figure 6–figure supplement 1: Profiles of TEE for four genes amongst the ones with the smallest average TEE. When the TEE is low, many κ_i do not pass the quality check, as it can be seen from the many points missing in the TEE profiles. This means that the experimental read counts are inconsistent with 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 (small TEE).

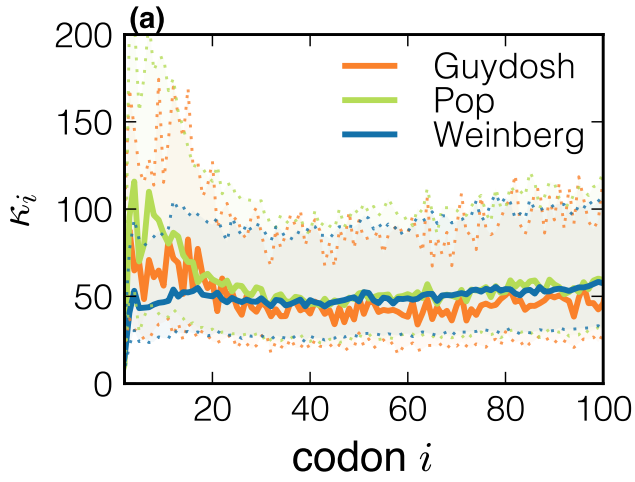

Figure 6–figure supplement 2: $\{\kappa_i\}$ profile obtained by averaging the κ_i values of each gene. The coloured regions delimits the first and third quartiles of the distributions.

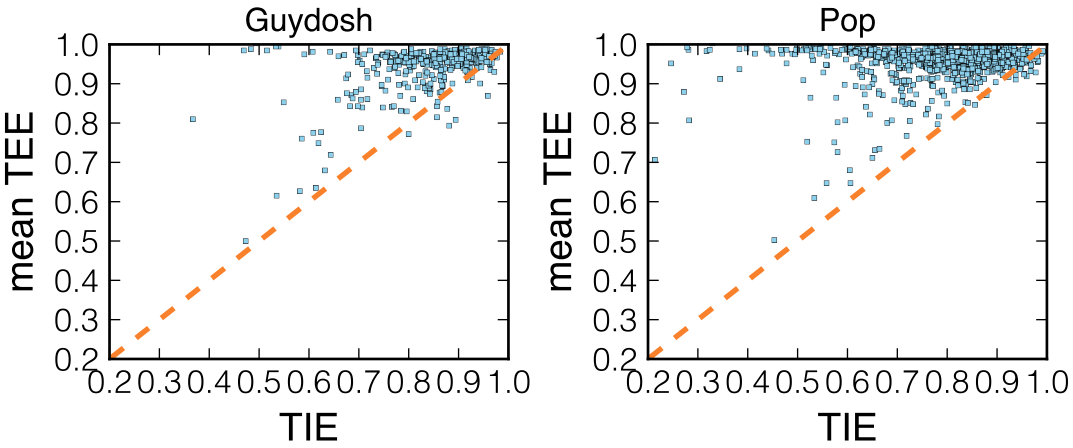

Figure 8–figure supplement 1: TIE vs mean TEE for the two datasets not shown in the main text.