Identification and quantitative analysis of the major determinants of translation elongation rate variation

Khanh Dao Duc[1] and Yun S. Song[1,2,*]

1 Department of Biology and Department of Mathematics, University of Pennsylvania
2 Computer Science Division, Department of Statistics, and Department of Integrative Biology, University of California, Berkeley

* To whom correspondence should be addressed: yss@berkeley.edu

Abstract

Ribosome profiling provides a detailed view into the complex dynamics of translation. Although the precise relation between the observed ribosome footprint densities and the actual translation elongation rates remains elusive, the data clearly suggest that elongation speed is quite heterogeneous along the transcript. Previous studies have shown that elongation is locally regulated by multiple factors, but the observed heterogeneity remains only partially explained. To dissect quantitatively the different determinants of translation speed, we here use probabilistic modeling of the translation dynamics to estimate transcript-specific initiation and local elongation rates from ribosome profiling data. Using this model-based approach, we estimate the fraction of ribosomes ($\sim 9\%$) undetected by the current ribosome profiling protocol. These missing ribosomes come from regions harboring two or more closely-stacked ribosomes, and not accounting for them leads to a substantial underestimation of translation efficiency for highly occupied transcripts. We further quantify the extent of transcript- and position-specific interference between ribosomes on the same transcript, and infer that the movement of $\sim 2.5\%$ of ribosomes is obstructed on average, with substantial variation across different genes. The extent of interference also varies noticeably along the transcript sequence, with a moderately elevated level near the start site and a significantly pronounced amount near the termination site. However, we show that neither ribosomal interference nor the distribution of slow codons is sufficient to explain the observed level of variation in the mean elongation rate across the transcript sequence. Surprisingly, by optimizing the fit of statistical linear models, we find that the hydropathy of the nascent polypeptide segment within the ribosome plays a major role in governing the variation of the mean elongation rate along the transcript. In addition, we find that positively and negatively charged amino acid residues near the beginning and end of the ribosomal exit tunnel, respectively, are important determinants of translation speed, and we argue that this result is consistent with the known biophysical properties of the exit tunnel.

Introduction

Ribosome profiling [1–3] is a powerful transcriptome-wide experimental protocol that utilizes high-throughput sequencing technology to provide detailed positional information of ribosomes on translated mRNA transcripts. As a useful tool to probe post-transcriptional regulations of gene expression, ribosome profiling has notably been used to identify translated sequences within transcriptomes, to monitor the process of translation and the maturation of nascent polypeptides *in vivo*, and to study limiting determinants of protein synthesis (see recent reviews [4–6] for an overview of diverse applications of the technique). In addition, since the ribosome occupancy at a given position reflects the relative duration of time spent at that position, ribosome profiling provides an unprecedented opportunity to study the local translational dynamics [7]. However, the precise relation between the observed footprint densities and the corresponding translation elongation rates remains elusive [6], thus making it difficult to interpret ribosome profiling data.

One factor which affects the translation elongation speed is ribosomal interference, which occurs when slow translocation of a ribosome at a certain site blocks another one preceding it. Because the information provided by ribosome profiling is marginal probability density (in the sense that it does not capture the joint occupancy probability of multiple ribosomes on the same transcript), it is not possible to observe ribosomal interference directly from data and therefore quantifying the role of interference in limiting the elongation speed has remained challenging. A potential analytical issue arises from the omission of polysomes (i.e., multiple adjacent ribosomes) in the current ribosome profiling protocol [8–10]. In most analyses, ribosome positional distributions along the open reading frame (ORF) are inferred from protected mRNA fragments which presumably reflect the size of the 60S ribosomal subunit (28-29 nt in *S. Cerevisiae* or 30-31 nt in mammalian cells). However, gradient footprint profile also shows other larger protected fragments of 40-65 nt which can be attributed to two closely stacked ribosomes that accumulate when the leading ribosome is stalled [10, 11]. Not taking these fragments into account in the ribosome profile may thus produce biased estimates of ribosome densities, and, as a consequence, of elongation rates.

Over the past few years, multiple studies have tried to utilize ribosome profiling data to identify the key determinants of the protein production and translation rates, but have arrived at contradictory results [12–19]. Due to the vast complexity of the different biophysical mechanisms involved in the decoding and translocation of the ribosome along the mRNA, it is indeed a challenging problem to disentangle the composite factors that can modulate the elongation speed for a given transcript sequence. Several studies have shown that elongation speed is locally regulated by multiple factors, including tRNA availability and decoding time [19, 20], mRNA secondary structure [21], peptide bond formation at the P-site [22], and the presence of specific amino acid residues [15, 23] in the nascent polypeptide that interact with the ribosomal exit tunnel [24]. However, the observed heterogeneity in elongation rates along the transcript, notably the so-called $5'$ "translational ramp" [1], remains only partially explained [14].

Here, we provide new insights into the major determinants of the translation dynamics, by identifying features that can explain a large portion of the variation in the mean elongation rate along the transcript, particularly the $5'$ translational ramp. We also present a new statistical method that can be used to obtain accurate estimates of initiation and local elongation rates from ribosome profiling and RNA-seq data. Our approach is based on a probabilistic model that takes into account the principal features of the translation dynamics, and it allowed us to quantify the extent of ribosomal interference (not directly observable from data) along the transcript.

Results

Inference of initiation rates and local elongation rates

We developed an inference procedure based on an extended version of the biophysical TASEP model [25] (see Fig. 1A, Fig. S1, **Materials and Methods**, and Supplementary Information) to estimate transcript-specific initiation and local elongation rates from ribosome profiling and RNA-seq data. In our model, the initiation rate is the exponential rate at which the A-site of a ribosome enters the start codon, while the elongation rate at a given codon position is the rate at which the A-site of a ribosome occupying that position translocates to the next downstream codon. Here, both events are conditioned on there being no other ribosomes in front obstructing the movement.

For the main part of our analysis, we used flash-freeze ribosome profiling and RNA-seq data of *S. Cerevisiae* generated by Weinberg *et al.* [14]. We ran our inference method on a subset of 850 genes selected based on length and footprint coverage (see **Materials and Methods**), and tested its accuracy (detailed in Supplementary Information). Fig. 1B is an example illustrating the excellent agreement between the actual ribosome footprint distribution for a specific gene from experiment and the distribution of monosomes obtained from simulation under the extended TASEP model with our inferred initiation and elongation rates.

We first used our estimates to see whether our method could recover what is known in the literature. For the set of genes we considered, we found that the mean time between initiation events varied from 5.5 s (5th percentile) to 20 s (95th percentile), with median = 10 s. These times are of similar order but shorter than the times found previously [13] (4 s to 233 s, with median = 40 s), which is explained by the fact that the set of genes we considered does not include lowly expressed genes (i.e., with low ribosomal density). In agreement with previous findings [13, 14], our inferred initiation rates were positively correlated (Pearson's correlation coefficient $r = 0.2646$, p-value $< 10^{-10}$) with the $5'$-cap folding energy (see **Materials and Methods**) and negatively correlated ($r = -0.4$, p-value $< 10^{-15}$) with the ORF length (these results are detailed in Fig. 2A).

To verify that our method effectively captured the dynamics associated with a specific codon at the A-site, we separated the inferred elongation rates according to their corresponding codon (the resulting distributions are shown in Fig. 2B). We observed that codon-specific mean elongation rate (MER) was positively correlated with the inverse of the codon-specific A-site decoding time estimated from Gardin *et al.* [19] ($r = 0.7$, p-value $< 4 \times 10^{-10}$, see Fig. 2C), supporting that different codons are decoded at different rates at the A-site. We then compared these MER with the ones estimated by applying our method to another flash-freeze dataset, generated by Williams *et al.* [26]. Because of lower sequencing depth compared to Weinberg *et al.*'s data, the number of genes passing our selection criteria decreased to 625 genes (see **Materials and Methods**). We obtained an excellent correlation between our MER estimates for the two datasets ($r = 0.92$, p-value $< 4 \times 10^{-25}$, see Fig. 2C).

Finally, since the differences in MER at different sites could be associated with tRNA availability variations [20], we further compared the MER and the codon tAI value [12, 27], which reflects the codon usage bias towards the more abundant tRNAs in the organism, and found a positive correlation ($r = 0.49$, p-value $< 5 \times 10^{-5}$, see Fig. 2C). Altogether, these results suggested that our estimates of the local elongation rates reflect tRNA-dependent regulation of translation speed and that our estimates are consistent across different ribosome profile datasets.

Isolated and stacked ribosome distributions across genes and codon positions

The standard ribosome profiling protocol selects for isolated ribosomes occupying 27 and 31 nt, so larger mRNA fragments protected by closely-stacked ribosomes (separated by ≤ 2 codons) are typically not included in the experimental data, thus making the ribosome footprint distribution inaccurate in regions of high traffic [8, 10, 11]. To assess how much of this information is possibly missing, we simulated under the extended TASEP model with our inferred initiation and elongation rates to generate total footprint distributions that include closely stacked ribosomes, which would be undetected by ribosome profiling experiments (a precise mathematical definition for our model is provided in **Materials and Methods**).

We first verified that experimental ribosome profiling data did not capture closely-stacked ribosomes, by comparing the translation efficiency (TE) in Weinberg's dataset to the measurement of per-gene ribosome density from polysome profiling carried out by Arava *et al.* [28] (for 588 genes common to both datasets). TE is the ratio of the RPKM measurement for ribosomal footprint to the RPKM measurement for mRNA [1], where RPKM corresponds to the number of mapped reads per length of transcript in kilo base per million mapped reads. In other words, it quantifies for each gene the average number of detected ribosomes per single transcript, up to a normalization constant. When the total ribosome density of a gene is low, it coincides with the TE. We therefore determined the normalization constant (0.83) by linearly fitting the TE to the total ribosome densities measured by Arava *et al.* for values less than 1 ribosome per 100 codons (see Fig. 3A). Interestingly, as shown in Fig. 3B, we found that the normalization constant found by fitting larger densities was lower (0.61), suggesting that for highly occupied transcripts, the density of ribosome inferred from TE underestimates the actual total ribosomal density. To see if our method could accurately capture this difference, we simulated ribosomal footprint densities using our inferred rates for a subset of high-density transcripts (195 genes with > 1 ribosome per 100 codons) contained in Arava's dataset, and distinguished the densities of detected ribosomes (which are not closely stacked) and the total densities (which include all ribosomes). Performing the same linear fitting against ribosomal densities from Arava *et al.* (Fig. 3C), we found a normalization constant of 0.80 for the simulated total ribosome density, which agrees very well with the aforementioned normalization constant (0.83) for low-density transcripts (with < 1 ribosome per 100 codons). In contrast, when the detected-ribosome density was fitted against Arava *et al.*'s data (Fig. 3D), we found the normalization constant to be lower (0.67), consistent with the decrease we saw when the raw TE values for high-density transcripts were fitted against Arava *et al.*'s data. We therefore conclude that closely-stacked ribosomes comprise a large fraction of undetected ribosomes, and that our method allows us to correct the TE value to get close to the actual total ribosome density.

After finding that highly occupied transcripts tend to have a significant amount of undetected ribosomes, we then examined the overall proportion of ribosomes appearing as strictly stacked (with no gap between them) or closely stacked (with ≤ 2 codons between them), by computing their proportion for each gene and averaging these fractions over all genes. We found that on average 12% of ribosomes were closely stacked (Fig. 4A) and hence hidden from the experimental data. Among these undetected ribosomes, we found that strictly-stacked ribosomes made up 58%, representing 7% of all ribosomes. Furthermore, obstructed ribosomes (a ribosome is said to be obstructed if there is another ribosome immediately in front of it with no gap between them) comprised about a half of the strictly-stacked ribosomes, which suggests that long ribosomal queues with three or more ribosomes are rare. Because the 850 genes we selected were more highly occupied than average, the fraction of closely-stacked and strictly-stacked ribosomes should be lower for the entire set of genes. By extrapolating our estimates to a larger sample of 3941 genes (using their TE, see Fig. 4A), we found that the average fractions of closely-stacked and strictly-stacked ribosomes should respectively

be 9% and 5%; that is more than a half of undetected ribosomes are strictly stacked.

We looked at each gene separately and found that the proportion of the detected ribosomes varied substantially between different genes, ranging from 76% (5th percentile) to 96% (95th percentile), with mean and standard deviation equal to 88% and 6%, respectively (Fig. 4A). Such heterogeneity can be explained by differences in the total ribosome occupancy. In Fig. 4B, we observed that the difference between the total ribosome density (including undetected ribosomes) and the density of detected ribosomes increased super-linearly as a function of the total ribosome density. As a consequence, highly occupied transcripts have relatively more undetected ribosomes. On average, the density of detected ribosomes was 6.5% lower than the total density for lowly occupied genes (density < 1 ribosomes per 100 codons, 13% of the dataset), compared with 30% for highly occupied genes (density > 2 ribosomes per 100 codons, 13 % of the dataset), and 14% for the rest.

At least two factors contribute to closely-stacked ribosomes in a transcript. First is the initiation rate, which directly determines the average number of ribosomes occupying an mRNA transcript. The second is the heterogeneity of translation speed along the ORF, which could result in ribosomal interference. To examine their influence, we first plotted (Fig. 4C) the inferred initiation rate against the polysome proportion and found a positive correlation ($r = 0.56$, p-value $< 10^{-15}$). As this only partially explained the heterogeneity of closely-stacked ribosome proportions across different genes, we then looked at the local fraction of the obstructed ribosomes along the transcript sequences (Fig. 4D). Upon aligning the transcript sequences with respect to the start codon, we estimated the average amount of interference generated at each position. We observed a global increase of interference from the start to a peak located around the 30th codon (with the extent of interference being 2.45 times higher than the gene-specific mean). This peak was followed by a slow decrease to a plateau where no significant change in interference is observed. Since ribosomes in a high density region are more likely to interfere, this result is consistent with the experimentally observed pattern of average footprint distribution along the transcript [14], in particular with the trend of decreasing ribosome-footprint density forming the $5'$ translation ramp [1,12] (see Fig. S2).

Aligning the transcript sequences with respect to the stop codon position, we detected a significantly large peak of interference fraction located at 10 codons preceding the stop codon (showing 14 times more interference than the gene-specific mean, Fig. 4D), with the corresponding amount of ribosomes representing on average 3.5% of all obstructed ribosomes. Note that the length of 10 codons corresponds to the footprint size of a single ribosome, and hence the distance between the A-sites of two abutting ribosomes. Therefore, our result suggests that slow termination process (in agreement with previous observations of ribosomal pausing during translation termination [29,30]) also affects the neighboring ribosome densities and causes more frequent stalling (supported by another smaller peak present at position -20) at the end of translation.

The impact of ribosomal interference on translation dynamics

The differences in the amount of ribosome interference between different genes could lead to significant biases when using the TE as a proxy for protein production rate. Using our results, we could quantify the production rate precisely, and thus relate it to the detected or total ribosome density. Simulating under the model with our inferred parameters, we first computed the protein production rate, defined for each gene as the rate at which a single ribosome reaches the end of the ORF and unbinds, leading to protein production.

We examined the distribution of protein production rates (Fig. 5A) and observed a range between 0.042 s^{-1} (5th percentile) and 0.12 s^{-1} (95th percentile), with median and standard deviation equal to 0.075 s^{-1} and 0.025 s^{-1}, respectively. The protein production rate of a gene was generally lower than the corresponding translation initiation rate, due to an additional waiting time (~ 3 s

on average) caused by ribosomal interference. Comparing the protein production rate with the detected-ribosome density (Fig. 5A) gave a high correlation (Pearson's $r = 0.91$). However, we also observed a super-linear increase of the production rate as the detected-ribosome density increased, suggesting that because closely-stacked ribosomes are not included, the translational expression of large TE genes could be underestimated. Using the total density of ribosomes (Fig. 5A) instead of the detected-ribosome density improved the correlation ($r = 0.94$), but also led to a slight sub-linear trend, due to some saturation appearing when the initiation rate gets so high that elongation rates become limiting factors of translation.

To study how ribosomal interference affects the local ribosome dynamics, we examined the difference between the inferred elongation rates of our mathematical model (we call them *unobstructed* rates) and the effective rates given by the inverse of the average time spent at a particular position (we call them *observed* rates). Upon aligning all transcripts with respect to the start codon and averaging across the transcripts, we compared the average unobstructed rate at each position with the corresponding average observed rate (Fig. 5B). Both curves showed an initial decrease to a trough located at codon position around 40, followed by a slow increase to a plateau. These variations were vertical reflections of the polysome proportion curve in Fig. 4D (with a shift such that the peak is observed 10 codons downstream) and the $5'$ ramp obtained for ribosomal normalized density (Fig. S2). Both unobstructed and observed rates initially increased from a very low rate (~ 3 codons/s) to a peak of 11.5 and 10 codons/s, respectively, located at position 10. They then decreased to a local minimum of 9 and 7.9 codons/s, respectively, before increasing again to a plateau around 11.5 and 10.9 codons/s, respectively. Furthermore, the gap between the unobstructed and observed rates generally decreased (Fig. 5B, bottom plot) from 1.6 to 0.4 codons/s along the transcript, suggesting a decreasing impact of ribosomal interference on the translation dynamics. The reduction in the observed speed from the unobstructed elongation rate ranged from 5% (at the plateau) to 15% (between codon positions 10 and 20).

Aligning the transcript sequences with respect to the stop codon position and applying the same procedure, we observed a significant difference between the unobstructed and observed rates at codon position -10. The gap size is 3 codons/s, which amounts to 30% reduction from the unobstructed speed, while nearby sites have a regular level of 0.4 codons/s. This enhanced gap is likely induced by stalling at the stop codon. A smaller bump (1.3 codons/s) was also observed at codon position -20, reflecting the formation of a queue of three ribosomes.

Variation of codon-specific mean elongation rates along the transcript

After studying the local dynamics of translation and quantifying the increase of elongation rates corresponding to the $5'$ ramp of decreasing ribosome density, we investigated the possible determinants of such variation. The $5'$ ramp of ribosome density has previously been attributed to slower elongation due to more frequent use of codons with low-abundance cognate tRNAs near the $5'$-end [12]. However, this explanation has been recently argued to be insufficient [14], suggesting other mechanisms to cause the ramp.

To study whether the preferential use of slow codons can explain the variation of elongation rates along the transcript, we analyzed the positional distribution of different codons. To do so, we first grouped the codons (except stop codons) into five groups according to their mean elongation rates, and then plotted (Fig. 6A) their frequency of appearance at each position in the set of genes we considered. At almost all positions, we found that the higher the mean elongation rate of a group, the higher the frequency of its appearance (the average frequency of appearance per codon type was 0.25%, 0.9%, 1.6%, 1.9% and 2.25% for the five groups in increasing order of the mean elongation rate).

Looking more closely at how these frequencies changed along the transcript between positions 50 and 200 (Fig. S3A), we observed an increase in frequency for the fastest codons, while the opposite was true for slow codons. However, when we examined the associated positional variation in elongation speed by setting the elongation rate of each codon type at all positions to its corresponding average speed, we obtained an increase of 0.3 codons/s (Fig. S4A). This increase was not large enough to explain the total variation observed at the $5'$-ramp (approximately 2 codons/s). This result thus suggested the existence of other major factors influencing the codon translation speed along the first 200 codons.

To confirm this hypothesis, we plotted the variation of average elongation speed for each codon type along the transcript sequence (Fig. 6B), which displayed a range between approximately 2 and 14 codons/s. Also, for each position, we computed the mean deviation of each codon's elongation rate from the codon-specific mean elongation rate. Fig. S3B shows the results, which groups the codons according to their mean elongation rates, as done above.

Interestingly, we observed a general increase of the position-specific mean elongation rate from position 40 to 200 (corresponding to the ramp region). Weighting these variations by position-specific codon frequencies (Fig. S4B), we found that the mean elongation rate from position 40 to 200 increases from approximately 9.5 to 11.5 codons/s, which gives an increase of 2 codons/s, comparable to what we previously observed in Fig. 5B. We thus concluded that the major determinant of the $5'$ translational ramp was not the codon distribution, but an overall increase of translational speed along the ORF.

The major role of hydropathy and charge distributions of nascent polypeptides in explaining the positional variation of mean elongation rates

The above analyses suggested the existence of additional determinants that modulate local elongation rates and may explain the observed pattern of elongation rates along the transcript. We sought out to find these determinants using a statistical method.

Using molecular biology techniques, it has been demonstrated previously that electrostatic interactions between nascent polypeptides and the ribosomal exit tunnel can modulate elongation rates [31]. Motivated by this observation, we employed statistical linear models to identify specific features of the nascent polypeptide that affect elongation rates and to quantify the extent of their influence. We first analyzed the data from Weinberg *et al.* [14]. In order to eliminate potential additional complications near stop codons due to ribosomal pausing, we focused on the genes of length at least 300 (codons) among the set of 850 genes considered hitherto. In total 640 genes were used for this study. The dependent variable in each linear model was the position-specific mean deviation of elongation rates from codon-type-specific average elongation rates (the latter was obtained by averaging over all transcripts and positions).

About 40 or so amino acid residues can be accommodated within the ribosome [24], so we first considered codon positions 6 to 44, in order to focus on the dynamics as the nascent polypeptide chain makes its initial pass through the peptidyl transferase center (PTC) and the ribosomal exit tunnel. By optimizing the fit of linear models, we found that the PARS score (**Materials and Methods**), which reflects the existence of mRNA secondary structure, in the window $[9:19]$ downstream of the A-site is a statistically significant explanatory feature that is negatively correlated with the position-specific mean elongation rate in this region. This result is consistent with previous findings [32] that mRNA secondary structure inhibits elongation near the $5'$-end. This feature was generally more important for longer transcripts. We also found important regulatory features of the nascent polypeptide segment within the PTC and near the beginning of the exit tunnel. Specifically, when we scanned linear models with different feature windows to obtain the best fit, we found

that the number of positively charged residues in the window $[1:9]$ and the number of negatively charged residues in the window $[6:26]$ are important features with opposite effects; the former facilitates elongation, while the latter slows down elongation. These two charge features together with the PARS score explain 93% of the positional variation (Fig. 7A) in the mean deviation of elongation rates in this region.

We then tried to construct a linear model for codon positions 45 to 300. We could not obtain a good fit only using explanatory features based on the PARS score and the number of charged residues. Surprisingly, we found that the hydropathy of the nascent polypeptide chain in the window $[1:42]$ upstream of the A-site can alone explain 84% of the positional variation in the mean deviation of elongation rates in this region. This window $[1:42]$ was determined by optimizing the fit of a linear model with hydropathy as the sole feature; the resulting fit is shown Fig. 7B. This result implies that the more hydrophobic the nascent polypeptide segment is, the higher the mean elongation rate.

Finally, we tried to obtain a single linear model for the combined region between positions 6 and 300. When we limited the number of features to three, we found the following quantities to be particularly important: the mean hydropathy in the window $[1:42]$, the mean number of positively charged residues in the window $[1:9]$, and the mean number of negatively charged residues in the window $[27:45]$, all upstream of the A-site. The PARS score did not contribute significantly to improving the fit, suggesting that the mRNA secondary structure is not a notable determinant of translation speed for the whole transcript sequence but rather only in the beginning of translation, before the nascent polypeptide emerges from the ribosome tunnel. Using these features, we obtained a fit (Fig. 7C) with the coefficient of determination $R^2 = 0.82$; all three features had positive regression coefficients.

We then took the above-mentioned features that we learned from analyzing the data from Weinberg *et al.* [14] and used them to fit the previously-mentioned ribosome profiling data for 625 genes from Williams *et al.* [26]. This led to fits with goodness comparable to the ones mentioned above: $R^2 = 0.86$ for the region $[6:44]$, $R^2 = 0.74$ for the region $[45:300]$, and $R^2 = 0.74$ for the region $[6:300]$. A few factors potentially contributed to slightly lower coefficients of determination for Williams *et al.*'s data. First, 167 out of 625 genes in the dataset were shorter than 300 codons, while we excluded such genes when we analyzed Weinberg *et al.*'s data to eliminate the effects of ribosomal pausing near stop codons. Second, there are no RNA-seq data associated with the ribosome profiling from Williams *et al.*, so we could not refine the "naive" estimates of elongation rates for this dataset (see **Materials and Methods**).

Discussion

We used probabilistic modeling of the translation dynamics to dissect the different determinants of translation speed, and developed an efficient, simulation-based inference algorithm to estimate transcript-specific initiation and local elongation rates from ribosome profiling data. Recently, an alternate method was developed to estimate average codon elongation rates, by down-sampling data to independently analyze many selected regions (windows) where the effects of codon usage are particularly easy to analyze [19]. One advantage of our approach is that it enabled us to quantify the extent of transcript- and position-specific ribosomal interference, which is hidden from the experimental data because of the filtering of larger ribosomal footprints. Our analysis suggests that these undetected ribosomes represent a non-negligible fraction (9%) of the total amount and that more than a half of these ribosomes are involved in interference. A consequence of not accounting for closely-stacked ribosomes is that the standard TE measure underestimates the total ribosome

density for highly occupied genes. We showed that this bias can be corrected by employing our inference method. Further, it allowed us to capture local variations of elongation rates that are actually necessary to explain the observed translational ramp. Our results revealed that the extent of ribosomal interference varies substantially across different genes and different positions. This caused significant variations in the difference between the theoretical unobstructed rate and the observed rate, suggesting that a more detailed experimental quantification of disomes is notably needed to fully characterize ribosome occupancy [10].

Other approaches based on simulating translation dynamics have been proposed to explain the experimentally observed ribosome profiles [9, 12, 13, 17, 18, 33], but with contradictory results. One source of complication comes from technical artifacts in the data. In particular, cycloheximide pre-treatment, used to immobilize ribosomes, can lead to substantial codon-specific biases [14, 34, 35]. The use of flash-freeze technique alleviates some of these problems, and allows one to obtain ribosome-footprint profiles and mRNA abundances that more faithfully reflect the translation dynamics [14]. Our inferred rates from flash-freeze data showed that the elongation rate is indeed modulated by the decoded codon located at the A-site of the ribosome and the corresponding tRNA availability. The positive correlation between the codon-specific mean elongation rate and the translation adaptation index supports the hypothesis that tRNA abundance and codon usage co-evolved to optimize translation rates [12, 36].

However, our refined analysis of the distribution of codon-specific elongation rates showed that tRNA availability is not sufficient to fully explain the observed translational speed variation. In particular, the $5'$ translational ramp variation cannot be sufficiently explained by the change of frequencies of slow and fast codons across the transcript sequence, contrary to what was previously suggested by Tuller *et al.* [12, 37]. An earlier study [31] proposed that electrostatic interactions of nascent polypeptides with the charged walls of the ribosomal exit tunnel could be one of the possible mechanisms of modulation of translation speed. Indeed, subsequent studies showed that specific configurations of amino acids along the nascent polypeptide segment within the exit tunnel can contribute to a slowdown or arrest of translation [15, 23, 24, 38, 39]. One of the major findings of our work is that the mean variation of elongation rates averaged over all the genes can be well explained by a linear model with only few features, namely the amount of charged amino acid residues in the nascent polypeptide at the beginning and end of the tunnel, and the hydropathy of the nascent polypeptide segment within the ribosome. These features were selected by statistically optimizing the fit of the linear model to position-specific mean elongation rates in a large window, which included the $5'$ ramp.

We note that Tuller *et al.* [37] also employed linear regression to fit the $5'$-ramp, but their model and results are quite different from ours. First, they fitted a smoothed version of the normalized average ribosome density variation across the transcript (see Fig. S2), whereas our model fits the deviation from the mean codon elongation rate without any smoothing. Second, the features used in their model — the tAI value, the total charge approximately covered by the ribosome (13 codons), and the $5'$ folding energy down the A-site — were different from ours. They hand-picked their features, while we learned them from the data. Lastly, while our fit is globally good over the first 300 codons (Fig. 7C), the fit obtained by Tuller *et al.* mainly explains the (smoothed) variation of ribosome density in the first 50 codons (Spearman correlation of their fit was only 0.33 when the first 50 codons were removed).

There are reasonable biophysical explanations for the particular set of features selected by our statistical analyses. In order for the ribosome to translocate from one site to the next, the nascent polypeptide has to be displaced to liberate enough space for the chain to incorporate the next amino acid. The associated force needed to achieve this process is constrained by the biophysics of the tunnel, which is known to be charged, aqueous, and narrow [24, 31, 40]. Our statistical analysis

selected the number of positively charged residues near the beginning of the tunnel and the number of negatively charged residues near the end of the tunnel, and both features had positive regression coefficients, which implies that they both facilitate elongation. This finding is consistent with the known electrostatic properties of the tunnel. Specifically, previous measurements of the electrostatic potential inside the tunnel [31] shows that it is non-uniform (Fig. S5A). Performing a cubic spline fit and taking its negative gradient suggests that the electric field induced by the potential points outward (i.e., away from the PTC) near the beginning of the tunnel, while it points inward near the end of the tunnel (Fig. S5B). Hence, both positively charged residue near the beginning of the tunnel and negatively charged residue near the end of the tunnel will experience a force pointing outward, thereby facilitating the movement of the polypeptide chain through the tunnel.

Another important feature, which to our knowledge has not been previously noted as a major determinant of translation speed, is the hydropathy of the polypeptide segment within the PTC and the exit tunnel. A possible explanation for the impact of hydropathy on the elongation rate is that since the tunnel is aqueous [24] and wide enough to allow the formation of α-helical structure [40], the hydrophobicity (which is an important factor driving compactness and rigidity [41]) of the polypeptide segment inside the ribosome consequently drives the amount of force needed to push the chain up the tunnel. While variation in translation rates could play a functional role in regulating co-translational folding of the nascent polypeptide chain [42], our results on the impact of hydropathy suggest that this link is more complex in that the folding (or pre-folding) in turn can actually alter the rate of translation.

In summary, our results show how the time spent by the ribosome decoding and translocating at a particular codon site is governed by three major determinants: ribosome interference, tRNA abundance, and biophysical properties of the nascent polypeptide within the PTC and the ribosome exit tunnel. It is quite remarkable that using a linear model with only few features allowed us to fully and robustly capture the variations of the average elongation rate across the transcript sequence. In addition to these overall determinants, our study also demonstrated the importance of mRNA secondary structure in the first 40 codons and a pausing of the ribosome at or near the stop codon, suggesting that additional local mechanisms may play a role in modulating translation in specific parts of a transcript sequence.

A natural extension of our work is to investigate in more detail, based on the above findings, the determinants of translation at the individual transcript level. To do so, a more detailed analysis and modeling of the nascent polypeptide within and immediately outside the exit tunnel is needed, to reveal how a specific amino acid sequence can affect the translation rate through possible interactions or co-translational folding [42, 43].

Materials and Methods

Experimental dataset

We used publicly available data in our analysis. The flash-freeze ribosome profiling data from Weinberg *et al.* [14] can be accessed from the Gene Expression Omnibus (GEO) database with the accession number GSE75897. The accession number for the flash-freeze data from Williams *et al.* [26] is GSM1495503. To be able to determine normalization constants (detailed below) without being biased by the heterogeneity of translational speed along the $5'$ ramp and to obtain robust estimates of the steady-state distribution, we selected among the pool of 5887 genes the ones longer than 200 codons and for which the average ribosome density was greater than 10 per site. For the Weinberg *et al.* dataset this led to a set of 894 genes, to which we applied the first step of our inference procedure (described below) to produce an estimate of the initiation rate. The algorithm

converged for 850 genes, and the main results presented in this paper are based on those genes. For the Williams *et al.* dataset, the same procedure gave 625 genes.

Mapping of the A-site from raw ribosome profile data

To map the A-sites from the raw short-read data, we used the following procedure: We selected the reads of lengths 28, 29 and 30 nt, and, for each read, we looked at its first nucleotide and determined how shifted ($0, +1$, or -1) it was from the closest codon's first nucleotide. For the reads of length 28, we assigned the A-site to the codon located at position 15 for shift equal to $+1$, at position 16 for shift equal to 0, and removed the ones with shift -1 from our dataset, since there is ambiguity as to which codon to select. For the reads of length 29, we assigned the A-site to the codon located at position 16 for shift equal to $+0$, and removed the rest. For the reads of length 30, we assigned the A-site to the codon located at position 16 for shift equal to 0, at position 17 for shift equal to -1, and removed the reads with shift $+1$.

Estimation of detected-ribosome densities from translation efficiency measurements

Translation efficiency measurements were used to compute the monosome average density. Since translation efficiency is given by the ratio of the RPKM measurement for ribosomal footprint to the RPKM measurement for mRNA, it is proportional to the monosome average density. To estimate the associated constant for each gene of our dataset, we used the measurements of ribosome density from Arava *et al.* [28]. For genes with a ribosome density of less than 1 ribosome per 100 codons, we fitted the translation efficiency as a function of the density to a linear function and divided all the TEs by the coefficient of this fit to obtain estimates of the detected-ribosome density (Fig. 3).

Estimation of $5'$-cap folding energy

The $5'$-cap folding energy associated with each gene of our dataset was taken from Weinberg *et al.* [14], who used sequences of length 70 nt from the $5'$ end of the mRNA transcript and calculated the folding energies at 37°C using RNAfold algorithm from Vienna RNA package [44].

Estimation of RNA secondary structure (PARS score)

To quantify RNA secondary structure at specific sites, we used the parallel analysis of RNA structure (PARS) scores from Kertesz *et al.* [45]. It is based on deep sequencing of RNA fragments, providing simultaneous *in vitro* profiling of the secondary structure of RNA species at single nucleotide resolution in *S. Cerevisiae* (GEO accession number: GSE22393). We defined the PARS score of a codon by averaging the PARS scores of the nucleotides in that codon.

Mathematical modeling of translation

To simulate ribosome profiles, we used a mathematical model based on the totally asymmetric simple exclusion process (TASEP) [25]. Compared with the original TASEP, our model included additional features accounting for the heterogeneity of elongation rates and physical size of the ribosome. We assumed that each ribosome has a footprint size of 30 nucleotides (i.e., 10 codons) and that the A-site is located at nucleotide positions 16-18 (from the $5'$ end) [46]. Protein production consists of three phases: First, a ribosome arrives at the start codon, with exponentially distributed rate (defined as the initiation rate). Subsequently, a ribosome with its A-site located at position i is

allowed to move forward one codon position if this movement is not obstructed by the presence of another ribosome located downstream. The associated conditional hopping time at each site is exponentially distributed with a certain rate (defined as its elongation rate). When a ribosome eventually reaches a stop codon, it unbinds at an exponential rate (for simplicity, we also refer to this as an elongation rate), which eventually leads to protein production. By simulating under this model with given initiation and position-specific elongation rates, we can sample ribosome positions at different times and thereby approximate the marginal steady state distribution of ribosome positions.

Definition of undetected ribosome and undetected ribosome fraction

During simulation we monitor the distance between consecutive ribosomes along the transcript. Since experimental disome fragments were shown [10] to protect a broad range of sizes below ~ 65 nt, in our simulations we defined an undetection of ribosomes to occur when the distance between the A-sites of consecutive ribosomes is ≤ 12 codons (i.e., free space between the ribosomes is ≤ 2 codons).

Inference procedure

A detailed description of our inference procedure is provided in Supplementary Information. Briefly, for given experimental ribosome profile and monosome density (average number of monosomes occupying a single mRNA copy), our inference procedure for estimating transcript-specific initiation and local elongation rates of the assumed TASEP model consists of two steps (Fig. S1). 1) First, we approximate the position-specific elongation rate by taking the inverse of the observed footprint number (such approximation is valid when is no ribosomal interference), and then use simulation to search over the initiation rate that minimizes the difference between the experimental detected-ribosome density and the one obtained from simulation. 2) Then, simulating under these naive estimates, we compare the simulated ribosome profile with the experimental one and detect positions, called "error-sites", where the absolute density difference is larger than a fixed threshold. If error-sites are detected, we first consider the one closest to the $5'$-end. We jointly optimize the elongation rates in a neighborhood of this error-site and the initiation rate to minimize the error between the simulated and the observed profile. With these new parameters, we then re-detect possible error-sites located downstream and repeat the procedure until there are no more error-sites to correct.

Because the profile and average density are invariant to a global scaling of the initiation and elongation rates, the parameters obtained needed to be normalized to get the rates in appropriate units. We normalized the rates such that the global average speed measured by simulations between position 150 and the stop codon is 5.6 codons/s, as measured experimentally [7]. We restricted our analysis to genes longer than 200 codons so that this normalization procedure is not biased by the heterogeneity of translational speed along the $5'$ ramp.

For given initiation and position-specific elongation rates along the transcript, obtaining an analytic formula for the protein synthesis flux is often difficult, if not impossible, due to potential interference between ribosomes occupying the same transcript [47–49]. However, since translation is generally limited by initiation, not by elongation, under realistic physiological conditions [13], [50], typically only a few sites were affected by interference. This allowed us to cope with the high dimensionality of the model space and obtain estimates of rate parameters that produced excellent fit to the experimental data.

Software implementation

Simulation of translation and our inference algorithm were implemented in Matlab. We simulated the model using the next reaction method [51] derived from the Gillespie algorithm, which at each step samples the next event (initiation, elongation, or termination) and the associated time based on the current ribosome occupancy. To simulate a ribosome profile of size N, we first simulated 10^5 steps for burn-in. Then, after a fixed interval of subsequent time steps, we randomly picked one occupied A-site (if there is one) and recorded it as a footprint location; this sampling scheme was iterated until we obtained N footprints. Protein production flux was obtained by computing the ratio between the number of ribosomes going through termination and the total time.

Acknowledgments

We thank Oana Carja, Joshua Plotkin, Premal Shah, and David Weinberg for useful discussions and for providing us with the data analyzed in this paper. This research is supported in part by National Science Foundation (NSF) CAREER Grant DBI-0846015, a Math+X Research Grant from the Simons Foundation, and a Packard Fellowship for Science and Engineering.

References

1. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–223.

2. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature Protocols. 2012;7(8):1534–1550.

3. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. Genome-Wide Annotation and Quantitation of Translation by Ribosome Profiling. Current Protocols in Molecular Biology. 2013; p. 4–18.

4. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nature Reviews Genetics. 2014;15(3):205–213.

5. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics. 2012;13(4):227–232.

6. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nature Reviews Molecular Cell Biology. 2015;.

7. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789–802.

8. Subramaniam AR, Zid BM, O?Shea EK. An integrated approach reveals regulatory controls on bacterial translation elongation. Cell. 2014;159(5):1200–1211.

9. Dana A, Tuller T. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. PLoS Comput Biol. 2012;8(11):e1002755.

10. Guydosh NR, Green R. Dom34 rescues ribosomes in 3' untranslated regions. Cell. 2014;156(5):950–962.

11. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, et al. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell. 2011;147(6):1295–1308.

12. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141(2):344–354.

13. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-limiting steps in yeast protein translation. Cell. 2013;153(7):1589–1601.

14. Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Reports. 2016;14(7):1787–1799.

15. Charneski CA, Hurst LD. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol. 2013;11(3):e1001508.

16. Artieri CG, Fraser HB. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Research. 2014;24(12):2011–2021.

17. Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular Systems Biology. 2014;10(12):770.

18. Gritsenko AA, Hulsman M, Reinders MJ, de Ridder D. Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. PLoS Comput Biol. 2015;11(8):e1004336.

19. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. Measurement of average decoding rates of the 61 sense codons in vivo. Elife. 2014;3:e03735.

20. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Research. 2014;42(14):9171–9181.

21. Qu X, Wen JD, Lancaster L, Noller HF, Bustamante C, Tinoco I. The ribosome uses two active mechanisms to unwind messenger RNA during translation. Nature. 2011;475(7354):118–121.

22. Chevance FF, Le Guyon S, Hughes KT. The effects of codon context on in vivo translation speed. PLoS Genet. 2014;10(6):e1004392.

23. Sabi R, Tuller T. A comparative genomics study on the effect of individual amino acids on ribosome stalling. BMC Genomics. 2015;16(Suppl 10):S5.

24. Ito K, editor. Regulatory Nascent Polypeptides. Springer; 2014.

25. Spitzer F. Interaction of Markov processes. Advances in Mathematics. 1970;5(2):246–290.

26. Williams CC, Jan CH, Weissman JS. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. Science. 2014;346(6210):748–751.

27. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research. 2004;32(17):5036–5044.

28. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences. 2003;100(7):3889–3894. doi:10.1073/pnas.0635171100.

29. Yu X, Willmann MR, Anderson SJ, Gregory BD. Genome-wide mapping of uncapped and cleaved transcripts reveals a role for the nuclear mRNA cap-binding complex in cotranslational RNA decay in Arabidopsis. The Plant Cell. 2016;28(10):2385–2397.

30. Pelechano V, Wei W, Steinmetz LM. Widespread co-translational RNA decay reveals ribosome dynamics. Cell. 2015;161(6):1400–1412.

31. Lu J, Deutsch C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. Journal of Molecular Biology. 2008;384(1):73–86.

32. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. 2016;529(7586):358–363.

33. Ciandrini L, Stansfield I, Romano MC. Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. PLoS Comput Biol. 2013;9(1):e1002866.

34. Gerashchenko MV, Gladyshev VN. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Research. 2014;42(17):e134–e134.

35. Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. PLoS Genet. 2015;11(12):e1005732.

36. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009;324(5924):255–258.

37. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biology. 2011;12(11):1.

38. Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proceedings of the National Academy of Sciences. 2009;106(1):50–54.

39. Chiba S, Ito K. Multisite ribosomal stalling: a unique mode of regulatory nascent chain action revealed for MifM. Molecular Cell. 2012;47(6):863–872.

40. Voss N, Gerstein M, Steitz T, Moore P. The geometry of the ribosomal polypeptide exit tunnel. Journal of Molecular Biology. 2006;360(4):893–906.

41. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annual Review of Biophysics and Biomolecular Structure. 1999;28(1):319–365.

42. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Molecular Cell. 2015;59(5):744–754.

43. Pechmann S, Chartron JW, Frydman J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nature Structural & Molecular Biology. 2014;21(12):1100–1105.

44. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly. 1994;125(2):167–188.

45. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010;467(7311):103–107.

46. Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife. 2014;3:e01257.

47. Shaw LB, Kolomeisky AB, Lee KH. Local inhomogeneity in asymmetric simple exclusion processes with extended objects. Journal of Physics A: Mathematical and General. 2004;37(6):2105.

48. Chou T, Mallick K, Zia R. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. Reports on Progress in Physics. 2011;74(11):116601.

49. Derrida B, Evans MR, Hakim V, Pasquier V. Exact solution of a 1D asymmetric exclusion model using a matrix formulation. Journal of Physics A: Mathematical and General. 1993;26(7):1493.

50. Chu D, Kazana E, Bellanger N, Singh T, Tuite MF, von der Haar T. Translation elongation can control translation initiation on eukaryotic mRNAs. The EMBO journal. 2014;33(1):21–34.

51. Gibson MA, Bruck J. Efficient exact stochastic simulation of chemical systems with many species and many channels. The Journal of Physical Chemistry A. 2000;104(9):1876–1889.

52. Zhang S, Zubay G, Goldman E. Low-usage codons in Escherichia coli, yeast, fruit fly and primates. Gene. 1991;105(1):61 – 72. doi:http://dx.doi.org/10.1016/0378-1119(91)90514-C.

Figures

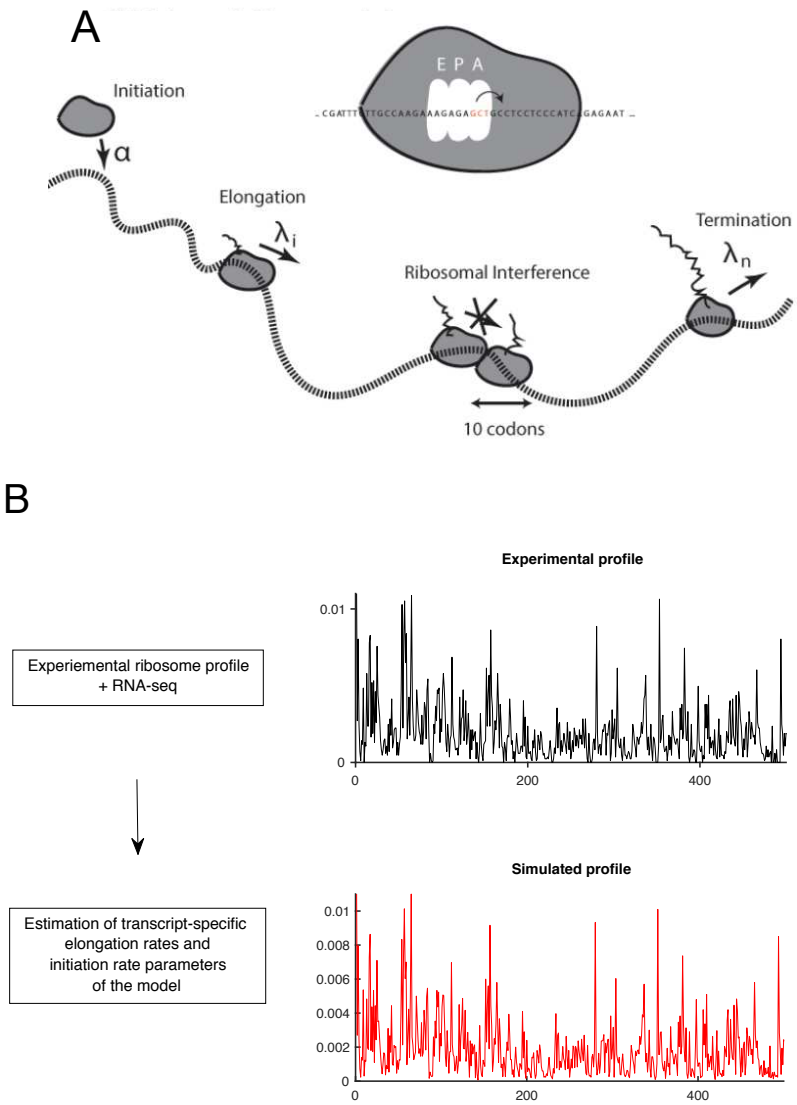

Figure 1. Illustration of translation dynamics and inference from experimental data. A. A schematic representation of the mathematical model of translation considered in this paper. Each ribosome is assumed to occupy 10 codons. Initiation corresponds to an event where the A-site of a ribosome enters the start codon, while elongation corresponds to a movement of the ribosome such that its A-site moves to the next downstream codon. Both events are conditioned on there being no other ribosomes in front obstructing the movement. The ribosome eventually reaches a stop codon and subsequently unbinds from the transcript, leading to protein production. All these stochastic events occur (conditioned on there being no obstruction) at some specific exponential rates, which we try to infer from experimental data (see **Materials and Methods**). In our simulations, we say that a ribosome is undetected when the distance between the A-sites of consecutive ribosomes is ≤ 12 codons (i.e., free space between the ribosomes is ≤ 2 codons).. **B.** A comparison between the actual experimental profile of detected ribosomes for a particular gene and the distribution of detected ribosomes obtained from simulation under the mathematical model with our inferred initiation and elongation rates.

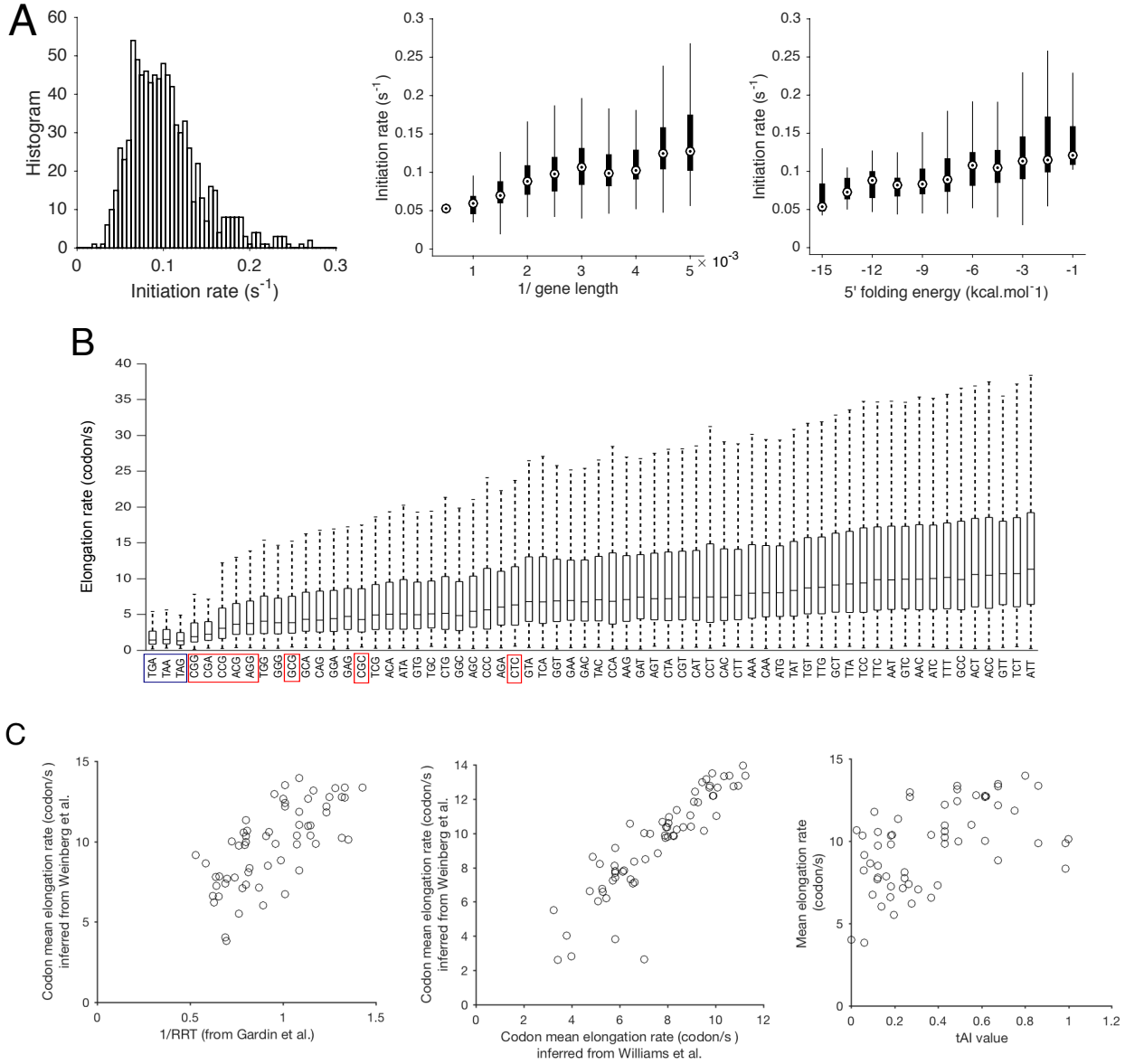

Figure 2. Analysis and comparison of the inferred rates. A. (Left) A histogram of inferred initiation rates. (Middle) Comparison between the inferred initiation rates and the inverse of the ORF length of the gene, showing a positive correlation ($r = 0.44$, p-value $< 10^{-15}$, computed for unbinned data). (Right) Comparison between the inferred initiation rates and the $5'$-cap folding energy computed in Weinberg *et al.* [14], showing a positive correlation (Pearson's correlation coefficient $r = 0.2646$, p-value $< 10^{-10}$, computed for unbinned data). **B**. Distribution of codon-specific elongation rates. Stop codons are boxed in blue, while the eight low-usage codons reported by Zhang *et al.* [52] are boxed in red. **C**. Comparison between the codon-specific mean elongation rates computed from **B** and (Left) the inverse of the codon mean "ribosome residence time" (RRT) estimated by Gardin *et al.* [19], (Middle) the codon-specific mean elongation rates computed from running our method on the Williams *et al.* dataset [26], and (Right) the tAI value, computed by Tuller *et al.* [12].

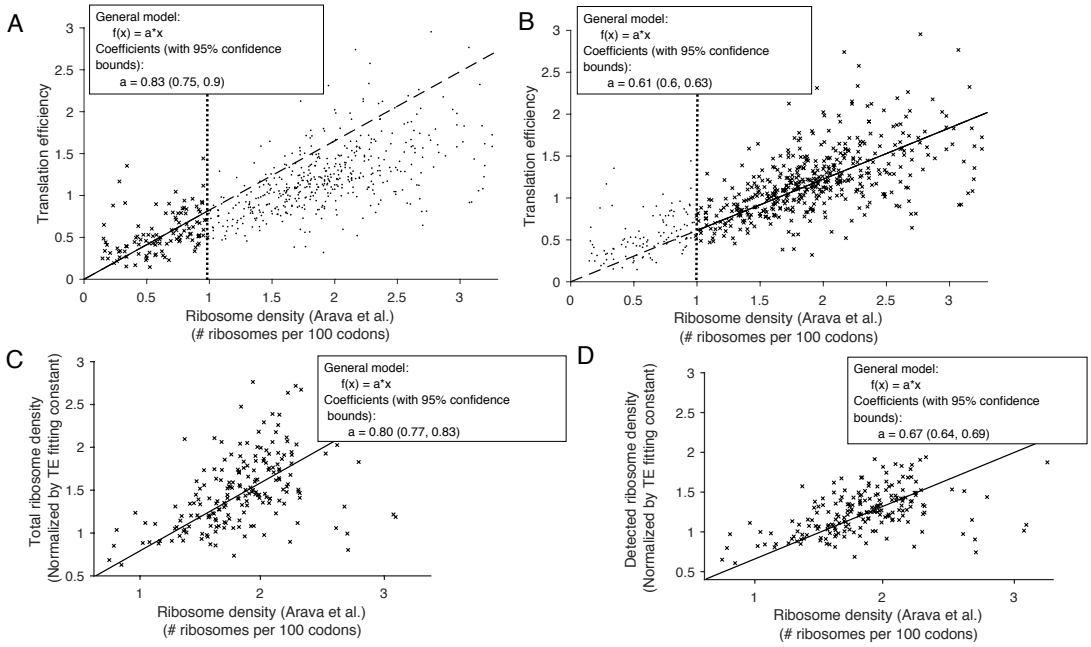

Figure 3. Comparison between translation efficiency (TE) and total ribosome density. All linear fit results are shown in the inset. **A.** The gene-specific TE for 588 genes from Weinberg *et al.*'s data [14] (see **Materials and Methods**) against the corresponding total ribosome density (average number of ribosomes per 100 codons) from Arava *et al.* [28]. We performed a linear fit of the points for which the corresponding ribosome density was less than 1 ribosome per 100 codons. **B.** Similar fit as in **A** in the range of ribosome density larger than 1 ribosome per 100 codons. **C.** Simulated total densities for a subset of 195 genes obtained using our inferred rates, against the ribosome density from Arava *et al.* **D.** Simulated detected-ribosome densities for the same 195 genes against the ribosome density from Arava *et al.* These results suggest that closely-stacked ribosomes comprise a large fraction of undetected ribosomes, and that our method allows us to correct the TE value to get close to the actual total ribosome density.

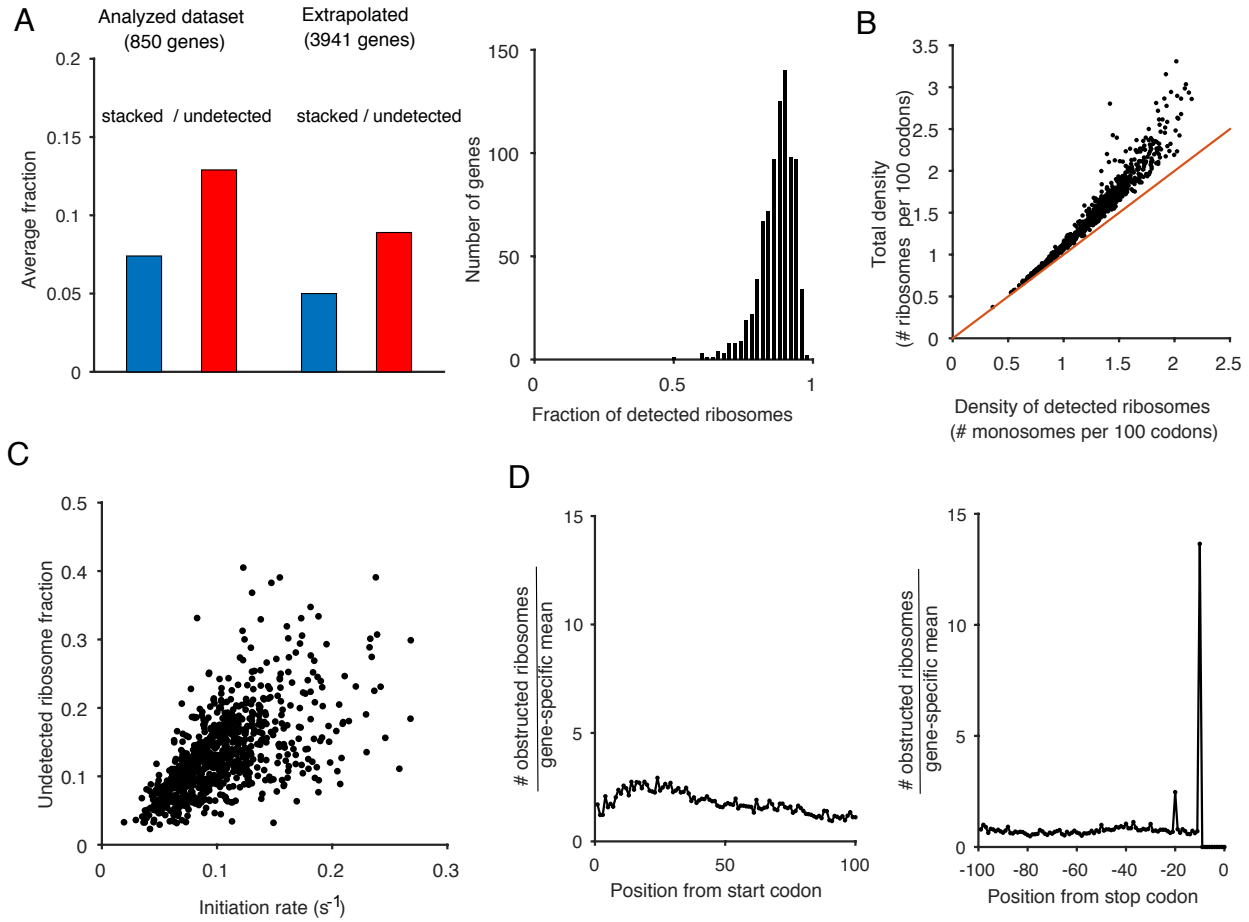

Figure 4. Stacked and undetected ribosome distributions. A. (Left) We first estimated the average proportions of strictly-stacked (blue) and undetected ribosomes (red) across all genes considered (850 genes). We then extrapolated these proportions to 3941 genes by binning the 850 genes by their detected-ribosome density (bin width = 0.1) and computing for each bin the average proportion of strictly-stacked and undetected ribosomes. For any given gene of the extended dataset (3941 genes), we assigned the bin-specific average proportions of strictly-stacked and undetected ribosomes. (Right) Histograms of gene-specific proportions of detected ribosomes for the 850 genes dataset. **B.** The difference between the total ribosome density and the detected-ribosome density as a function of the total ribosome density. The plot shows a super-linear behavior. **C.** Undetected ribosome fraction against the initiation rate. Pearson's correlation coefficient $r = 0.61$. **D.** (Left) Relative amount of interference along the first 200 codons. After filtering the obstructed ribosomes in our simulations for each transcript profile, we normalized the resulting profiles by the average number of obstructed ribosomes over the whole sequence. Upon aligning the transcript sequences with respect to the start codon, we then averaged these different normalized profiles at each site. (Right) Relative amount of site specific interference when the transcript sequences are aligned with respect to the stop codon.

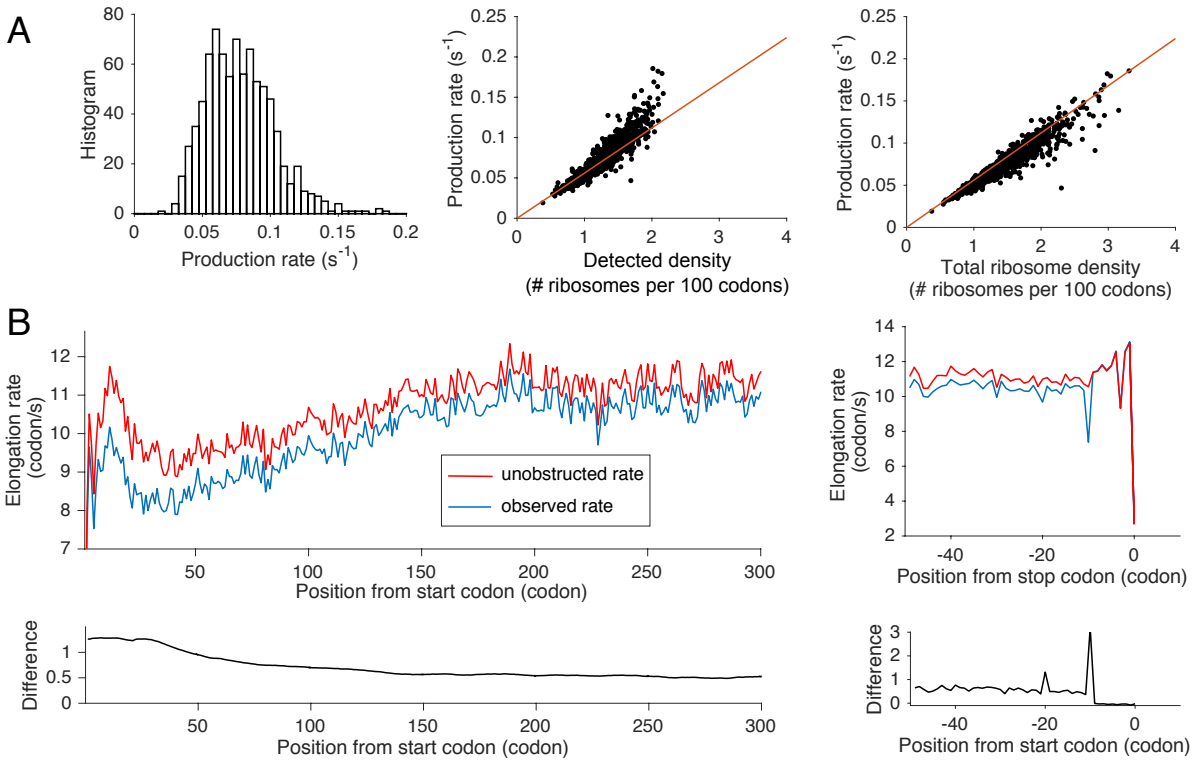

Figure 5. The impact of ribosomal interference on translation dynamics. A. Analysis of protein production. (Left) A histogram of protein production rates. (Middle) Comparison between the protein production rate and the detected-ribosome density obtained from simulations. In red, we plotted the simulated production rate as a function of ribosome density. The red line corresponds to the production rate when we assume no interference and a constant elongation speed of 5.6 codons/s, which was measured experimentally [7]. (Right) Comparison between the production rate and the total ribosome density density obtained from simulations. **B.** (Left) Position-specific elongation rates averaged over all transcript sequences, aligned with respect to the start codon. Plotted are the inferred unobstructed rate (in red) and the observed rate (in blue). The bottom plot shows the difference between the two curves. (Right) Similar plots as the ones on the left, when the transcript sequences are aligned with respect to the stop codon position.

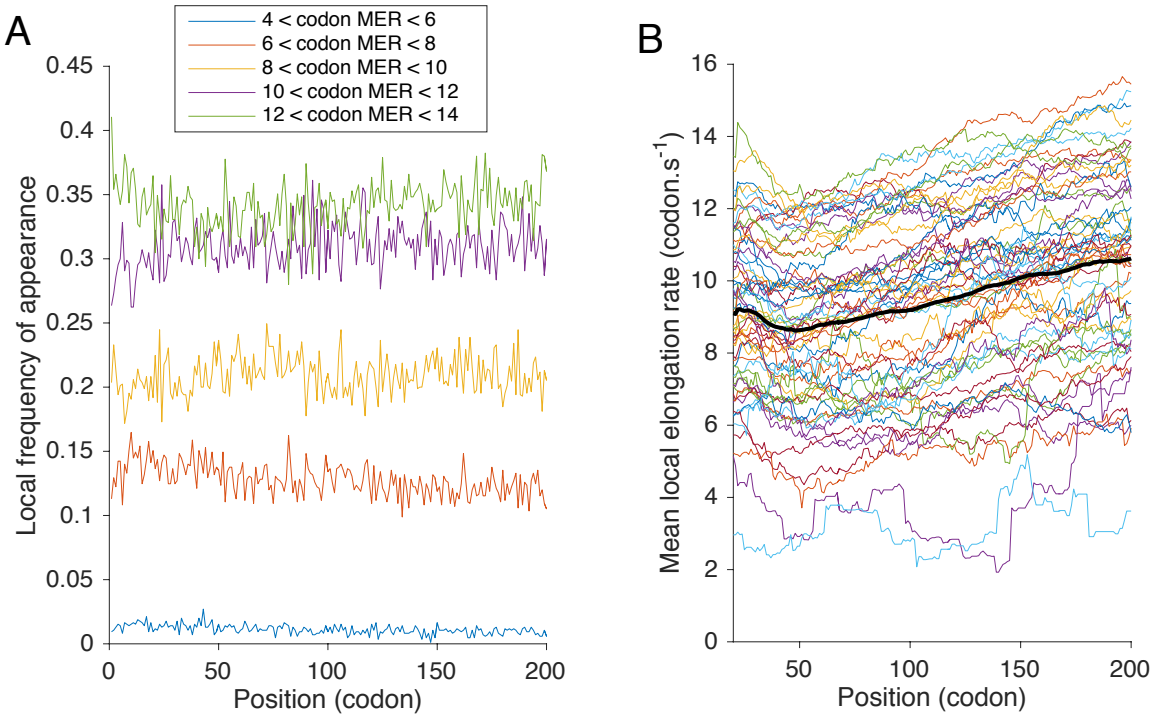

Figure 6. Heterogeneity of codon distributions and elongation speed along the transcript. A. Codon frequency metagene analysis. We grouped the codons (except stop codons) into five groups according to their mean elongation rates (MER) and plotted their frequency of appearance at each position in the set of genes we considered. The first group contained 4 codons with MER between 4 and 6 codons/s; the second group 13 codons with MER between 6 and 8; the third group 13 codons with MER between 8 and 10; the fourth group 16 codons with MER between 10 and 12; and the fifth group 15 codons with MER > 12. **B.** Smoothed mean elongation speed along the ORF for each codon type (stop codons are excluded). At each position i, we computed an average of codon-specific MER between positions $i - 20$ and $i + 20$. In black, we plot an average of the 61 curves.

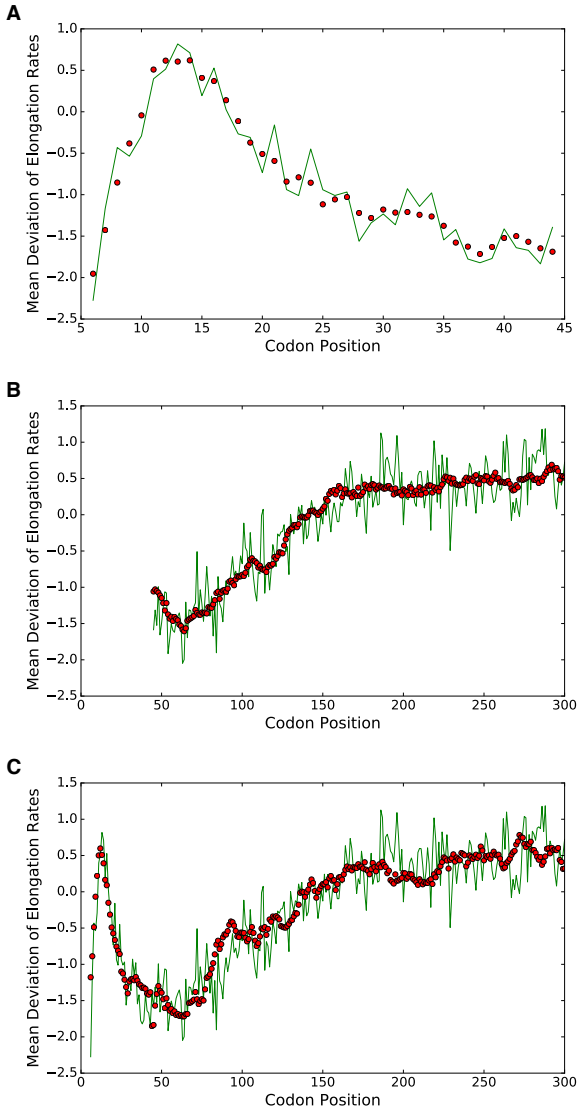

Figure 7. Linear model fits of the mean deviation of elongation rates for the data from Weinberg *et al.* [14]. The dependent variable is the mean deviation of elongation rates from codon-type-specific average elongation rates. Green lines correspond to the estimates from ribosome profiling data, while red dots correspond to our model fits based on a small (1 or 3) number of features. **A.** A fit for codon positions $[6:44]$ obtained using three features: the mean PARS score in the window $[9:19]$ downstream of the A-site, the mean number of negatively charged nascent amino acid residues in the window $[6:26]$ upstream of the A-site, and the mean number of positively charged residues in the window $[1:9]$ upstream of the A-site. The first two features had negative regression coefficients, while the last one had a positive regression coefficient. The coefficient of determination R^2 was 0.93 for this fit. **B.** A fit ($R^2 = 0.84$) for the region $[45:300]$ obtained using only a single feature: the mean hydropathy of the nascent peptide segment in the window $[1:42]$ upstream of the A-site. **C.** A fit ($R^2 = 0.82$) for the combined region $[6:300]$ obtained using three features upstream of the A-site: the mean hydropathy of the nascent peptide chain in the window $[1:42]$, the mean number of positively charged residues in the window $[1:9]$, and the mean number of negatively charged residues in the window $[27:45]$.

Supplementary Information

Identification and quantitative analysis of the major determinants of translation elongation rate variation

Khanh Dao Duc[1] and Yun S. Song[1,2,*]

1 Department of Biology and Department of Mathematics, University of Pennsylvania
2 Computer Science Division, Department of Statistics, and Department of Integrative Biology, University of California, Berkeley

* To whom correspondence should be addressed: yss@berkeley.edu

Content:

– Supporting Text

– Supporting Figures S1–S7

Inference of Initiation and Elongation Rates

For an experimental profile $P_{\text{exp}} = (P_{\text{exp}}(1), P_{\text{exp}}(2), \ldots, P_{\text{exp}}(L))$, where L is the gene length and $P_{\text{exp}}(i)$ the number of reads detected at position i, and a monosome density D_{exp} (number of monosomes per mRNA), we detail here the procedure to infer the associated elongation rates and initiation rate.

Naive estimates of elongation rates

To infer the different parameters of the model (namely, the initiation rate and the elongation rates), we first approximate the elongation rate at position i by

$$\lambda_0(i) = \begin{cases} \min\left(\lambda_{\max}, \dfrac{P_{\max}}{P_{\text{exp}}(i)}\right), & \text{if } P_{\text{exp}}(i) \neq 0, \\ \lambda_{\max}, & \text{else.} \end{cases},$$

where $P_{\max} = \max_i (P_{\text{exp}}(i))$ and $\lambda_{\max}$ is a fixed threshold value. These estimates well approximate the true elongation rates when the ribosomes encounter few interference. In this case, the elongation process of a single ribosome following the TASEP can indeed be approximated by a one dimensional irreversible Markov chain. In this case, for two sites indexed by position i and j, with respective elongation rates λ_i and λ_j, the ribosomal densities $p(i)$ and $p(j)$ satisfy at steady state the relation $\frac{p(i)}{p(j)} = \frac{\lambda_j}{\lambda_i}$. Using the estimates given by λ_0, we can simulate profiles for any initiation rate value α and obtain the corresponding monosomal average density $D(\alpha)$. Since D is an increasing function of α, we then estimate the initiation rate by computing $\operatorname{argmin}_\alpha |D_{\text{exp}} - D(\alpha)|$ by using a binary search algorithm.

Detection of sites with significant errors

Using these first estimates, we simulate the associated profile and compare it to P_{exp}. When there is interference at a certain position, the previous estimates may indeed not be valid and thus lead to significant errors. To compare the experimental profile P_{exp} to another profile $P = (P(1), P(2), \ldots, P(L))$, we define the error at position i as

$$\varepsilon(i) = \left| \frac{P_{\text{exp}}(i)}{\sum_k P_{\text{exp}}(k)} - \frac{P(i)}{\sum_k P(k)} \right|.$$

We introduce a significant error threshold, given by

$$\varepsilon_0 = \frac{10 \sum_i \varepsilon(i)}{L}.$$

When the error at a site i_0 is larger than ε_0, we define the complementary set of sites to correct as

$$I_0 = \{\text{sites } i \mid i \in [i_0 - 30, i_0 - 10] \text{ and } P(i) < P_t\},$$

where P_t is a fixed threshold value.

Refinement step

In the next step of the inference procedure, we correct the elongation rates at i_0 and I_0 to minimize the global error between the corresponding simulated profile and P_{exp}. For a given sequence of rates $\lambda = (\lambda(1), \ldots, \lambda(L))$, error site i_0 and corresponding set I_{i_0}, we define the modified sequence

$$\lambda^{i_0, I_0}_{\beta_1, \beta_2}(i) = \begin{cases} \beta_1 \lambda(i) & \text{if } i = i_0 \\ \beta_2 \lambda(i) & \text{if } i \in I_0 \\ \lambda(i) & \text{else} \end{cases},$$

where $\beta_1, \beta_2 > 0$. For such a sequence, we can apply the same previous procedure to estimate the initiation rate α_{β_1,β_2} which fits the experimental monosomal density and get an associated profile P_{β_1,β_2}. By double golden section search method (for a fixed β_1, we find optimal β_2 by golden search and use this to also optimize β_1 by golden search), we compute

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \left(\sum_i |P_{\beta_1,\beta_2}(i) - P_{\text{exp}}(i)| \right).$$

If only one error site has been detected, the inference procedure ends with $\alpha_{\hat{\beta}_1,\hat{\beta}_2}$ and $\lambda^{i_0,I_0}_{\hat{\beta}_1,\hat{\beta}_2}$ as the final estimated initiation and elongation rates. For multiple error sites, we dynamically apply the procedure we described. We first find the optimal rates for the site closest to the $5'$ end (with the supplementary condition that the modified rates do not introduce significant errors upstream of the original error site). We then update the positions of error sites and repeat the procedure until no error sites are detected downstream of the last error site treated.

Normalization

The elongation and initiation rates obtained after running the inference procedure need to be normalized to get the translation dynamics in appropriate units. To get for each gene the normalization constant, we simulated 10000 ribosomal runs from position 150 and recorded the average time to reach the last codon. After computing the associated average speed v (dividing the length run by ribosomes by the corresponding average time), we normalized the rates by $5.6/v$ to match the experimental observations of Ingolia *et al.* [1], which found a consistent average speed of approximately 5.6 codon.s^{-1} for each gene.

Results of the Inference Procedure

We ran the inference procedure on a yeast ribosome profiling dataset of 850 genes (see **Methods**). The parameters used were $P_t = \frac{1}{0.7\lambda_{\max}} = 35$ and $\lambda_{\max} = 50$. We tested the accuracy of our method by first comparing the experimental ribosomal densities with the ones obtained by simulations (Figure S6). The simulated and experimental densities were in good agreement, showing a Pearson's correlation coefficient of 0.986. The individual profiles obtained by simulations also showed good agreement with the experiments, with Pearson's correlation coefficient of 0.975.

We then analyzed more precisely how the procedure performed on each gene. During the inference procedure, 383 genes over the 850 in the dataset (45%) did not require corrections, which means that the difference of profile observed between the simulation and the experiment was for these genes globally under the threshold error fixed by our procedure. For the remaining 467 genes, the number of error sites per gene above the threshold error was in average 1.57 (std = 0.925) (Figure S7A). Since the first step of the inference gives the right estimates when there is no interference, we measured if some interference influenced the translation at the error sites we detected. To do so, we numerically estimated the probability of a ribosome occupying a certain site to block a ribosome located 10 codons before. We call this probability the interference rate. We found that the interference rate of error sites was in average equal to 0.245 compared with an average rate of 0.011 over all the sites of our dataset (Figure S7B). This large difference showed, as we expected, that local profile errors between experimental and simulations after the first round of estimation are primarily due to ribosomal interference. We then studied the efficiency of the rates refinement step of the procedure. The decrease in error was in average of 57% (Figure S7C). For 65% of the sites the error after correction went under the initial threshold of error site detection. The reasons for correction failure can vary from too large initial error, configurations of error sites too close to allow separate correction and more generally possible missing reads or errors in the original data.

Supplementary References

[1] Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789–802.

[2] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–223.

[3] Lu J, Deutsch C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. Journal of Molecular Biology. 2008;384(1):73–86.

Supporting Figures

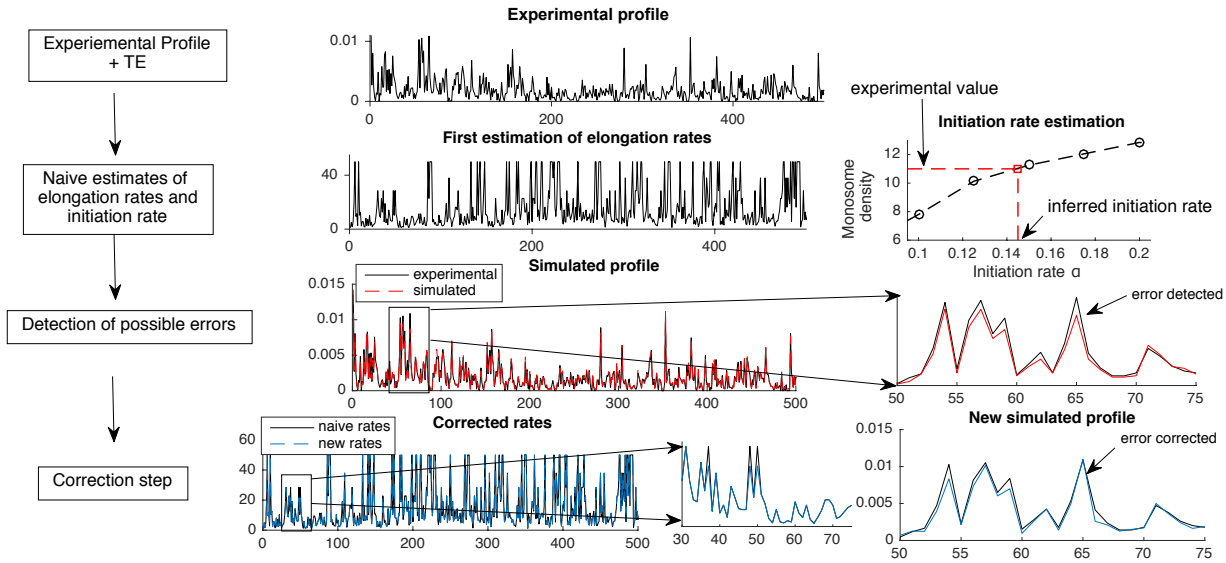

Figure S1: **A schematic description of our inference procedure.** Given a ribosome footprint profile and a measure of average detected-ribosome density ("translation efficiency"), we first approximate the position-specific elongation rate by taking the inverse of the observed footprint number. Then, with these approximate elongation rates, we use simulation to search over the initiation rate that minimizes the difference between the experimental density and the one obtained from simulation. We then refine these naive estimates using an iterative procedure as follows: Starting with the naive estimates, we compare the simulation result with the experimental ribosome profile and detect "error-sites" where the absolute density difference is larger than a chosen threshold. If error-sites are found, we start with the one closest to the $5'$-end, and jointly optimize the elongation rates in a neighborhood of this error-site and the initiation rate to minimize the error between the simulated and the observed profile. Using these new parameters in simulation, we then re-detect possible error-sites located downstream and repeat the procedure until there are no more error-sites to correct.

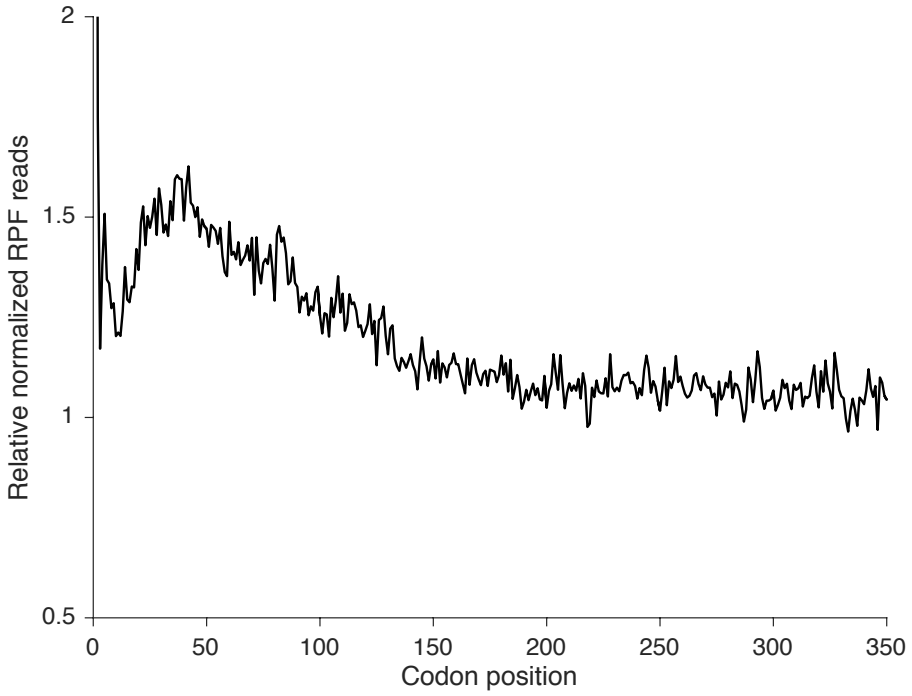

Figure S2: **Metagene relative normalized ribosome-footprint density as a function of codon position.** Ribosome profile footprint (RPF) reads in open reading frames (ORFs) were individually normalized by the mean RPF reads within the ORF, and then averaged with equal weight for each codon position across all ORFs, as in Ingolia *et al.* [2].

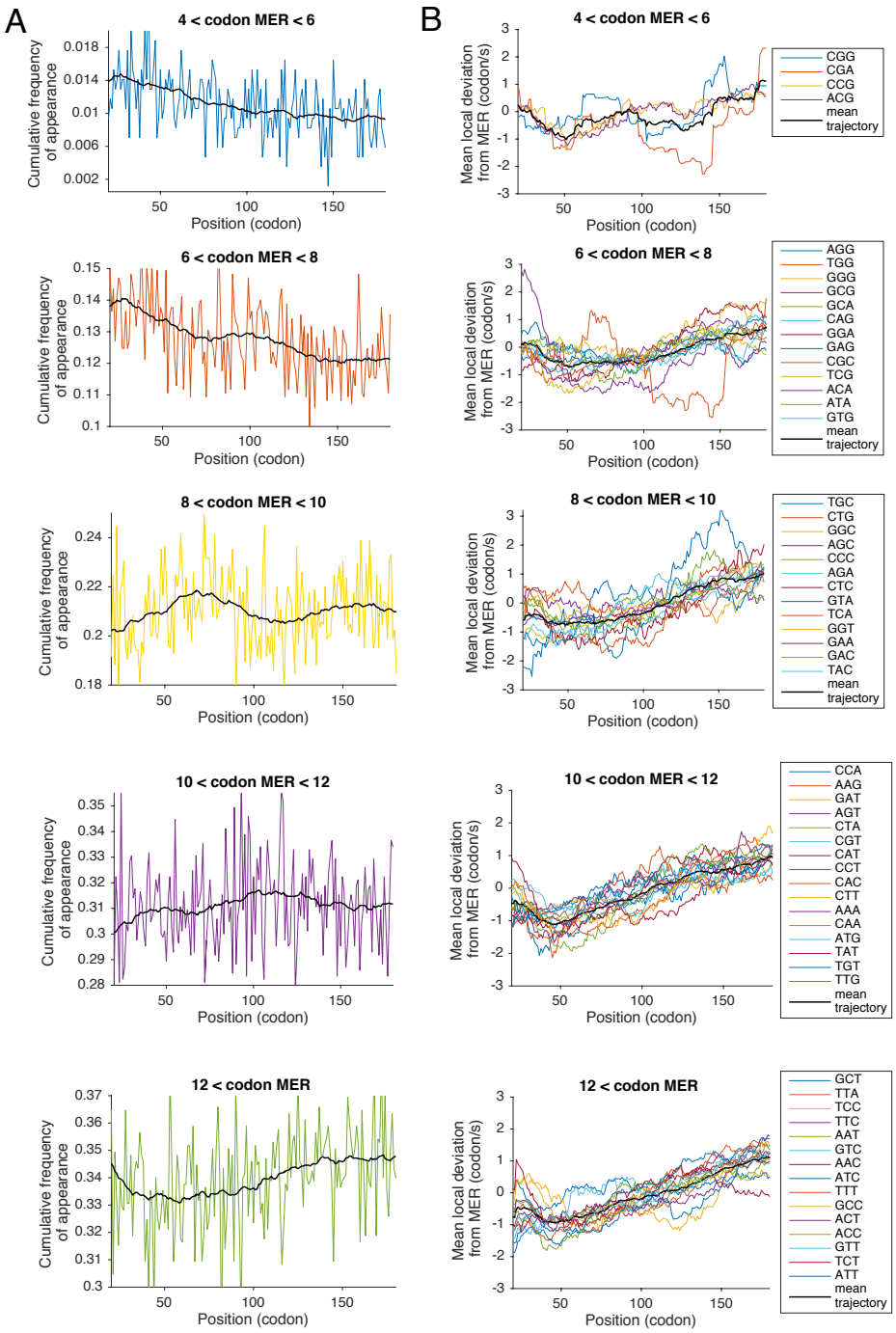

Figure S3: **Detailed codon frequency of appearance and elongation speed along the transcript. A.** Different panels show the frequency of appearance for each group of codons from Fig. 6A. The black curve in each panel corresponds to a smoothed version, for which the value at position i is obtained by averaging the values between positions $i - 20$ and $i + 20$. **B.** The difference between codon-specific local speed shown in Fig. 6B and the average of codon-specific speeds between position 20 and 180. Different codons are grouped as in **A**. For each panel, the black curve corresponds to an average of the curves in that panel.

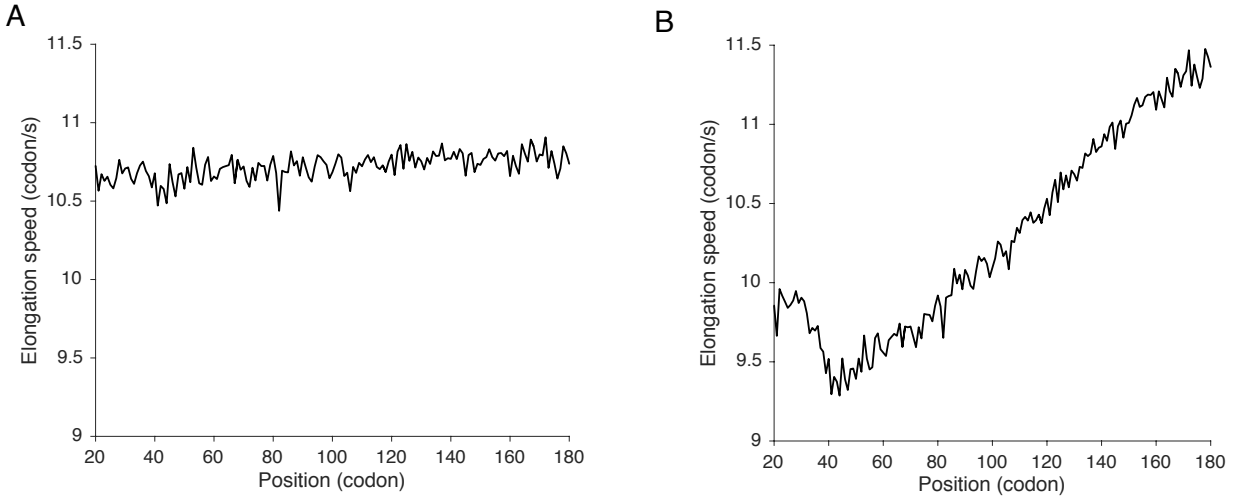

Figure S4: **A.** Average elongation speed along the transcript obtained by setting the elongation speed for each codon type at all positions to the corresponding mean elongation speed computed from Fig. 2B. This plot shows that the variation of codon frequency along the transcript is not sufficient to explain the 5′ translational ramp. **B.** Average elongation speed along the transcript obtained by setting the elongation speed for each codon to the position-specific mean elongation rate in Fig. 6B.

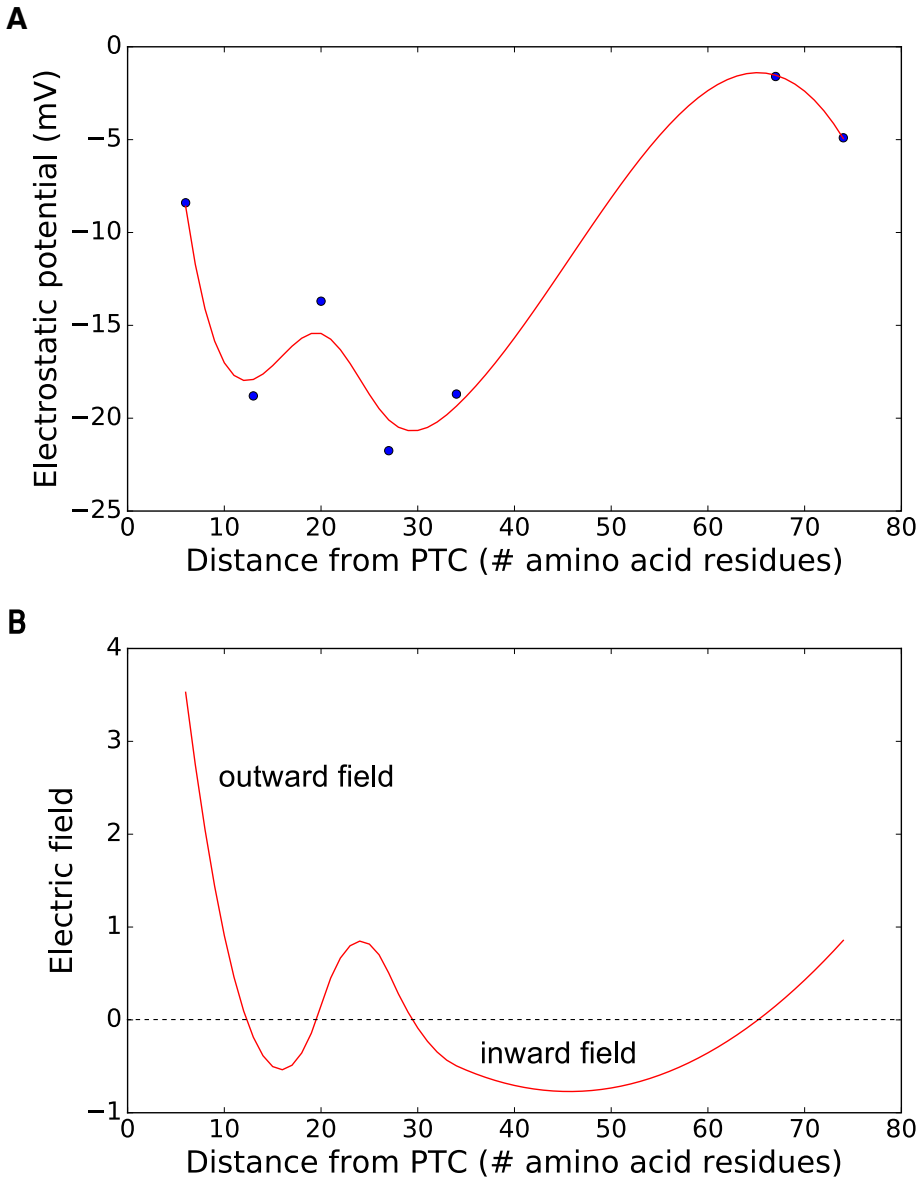

Figure S5: **A.** A cubic spline fit $V(z)$ (red curve) of experimentally measured electrostatic potential (blue dots) inside and immediately outside the ribosomal exit tunnel. The experimental potential measurements are from Lu and Deutsch [3]. UnivariateSpline function in scipy.interpolate module of Python was used with the default smoothing setting. **B.** Electric field obtained by $-dV(z)/dz$ where z is the direction along the tunnel. This plot suggests that the induced electric field points outward (i.e., away from the PTC) near the beginning of the tunnel, while it points inward near the end of the tunnel.

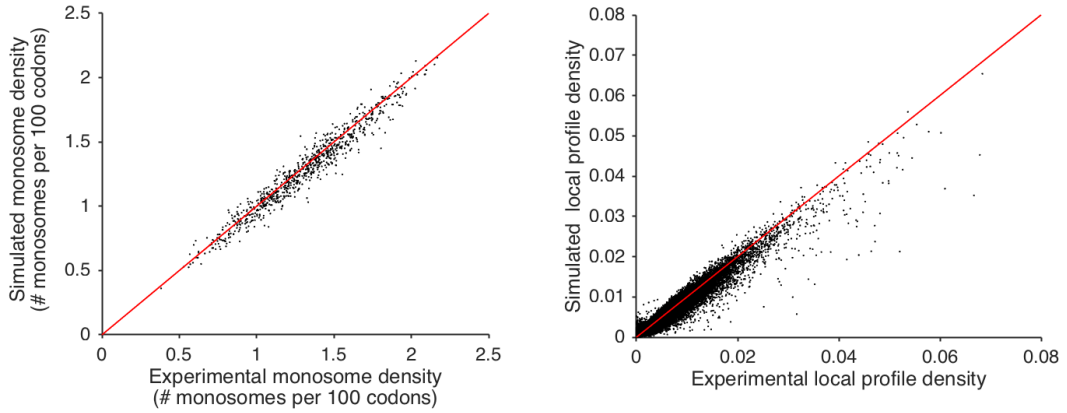

Figure S6: **Comparison between experimental data and numerical results from inference procedure.** We applied the inference method to a set of 850 genes in *S. Cerevisiae* (see **Methods**) and compared the total (left) and local (right) densities of the original dataset with the ones obtained by simulations of the model with the inferred parameters.

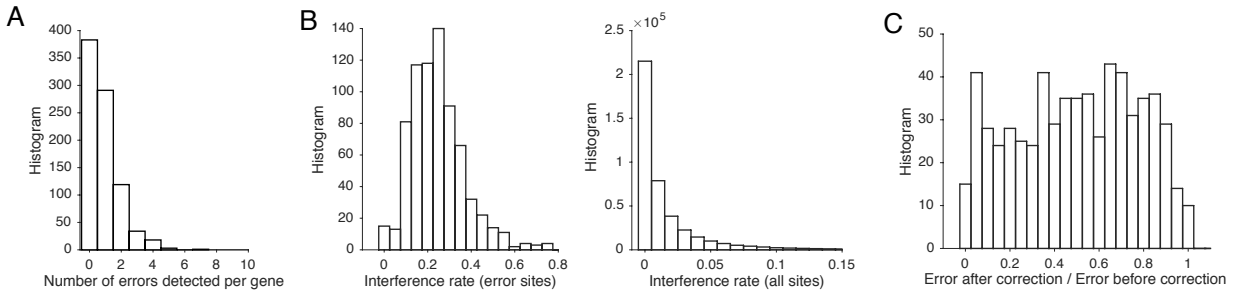

Figure S7: **Detailed results of the inference procedure. A.** Histogram of the number of significant errors detected for each gene. **B.** Histogram of interference rate for significant error sites (left) and for all sites (right). **C.** Histogram of error improvement after the refinement step, given by the ratio of error after correction over error before.