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 and Department of Statistics, 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 elongation speed, we here use probabilistic modeling 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 elongation rate. Instead, we find that electrostatic interaction between the ribosomal exit tunnel and specific parts of the nascent polypeptide governs the elongation rate variation as the polypeptide makes its initial pass through the tunnel. Once the N-terminus has escaped the tunnel, the hydropathy of the nascent polypeptide within the ribosome plays a major role in modulating the elongation speed. We provide evidence that our results are consistent with evolutionary signals and the known biophysical properties of the exit tunnel.

INTRODUCTION

Ribosome profiling (1, 2, 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, 5, 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 stacked ribosomes (i.e., multiple ribosomes that are less than a few codons apart) in the current ribosome profiling protocol (8, 9, 10, 11). 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, 12). 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 (13, 14, 15, 16, 17, 18, 19, 20, 21). 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 (20, 22, 23), mRNA secondary structure (24), peptide bond formation at the P-site (25), and the presence of specific amino acid residues (16, 26) in the nascent polypeptide that interact with the ribosomal exit tunnel (27). However, the observed heterogeneity in elongation rates along the transcript, notably the so-called $5'$ "translational ramp" (1), remains only partially explained (15).

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.

MATERIALS AND METHODS

Experimental dataset

We used publicly available data in our analysis. The flash-freeze ribosome profiling data from Weinberg *et al.* (15) 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.* (28) is GSM1495503 and the one from Pop *et al.*(18) are GSM1557442 (RNA-seq) and GSM1557447 (ribo-seq). The method used to map ribosome footprint reads is described in Weinberg *et al.* (15). 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. For Pop *et al.* it gave 212 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 average density of detected ribosomes. 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 average density of detected ribosomes.To estimate the associated constant for each gene of our dataset, we used the measurements of ribosome density from Arava *et al.* (29). 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.

Estimation of $5'$-cap folding energy

The $5'$-cap folding energy associated with each gene of our dataset was taken from Weinberg *et al.* (15), 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 (30).

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.* (31). 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) (32, 33). 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) (34). Protein production consists of three phases: First, a ribosome enters the ORF with its empty A-site at the second codon position; the waiting time follows an exponential distribution and we define its rate as the initiation rate. Subsequently, a ribosome is allowed to move forward one codon position if this movement is not obstructed by the presence of another ribosome located downstream. As the dynamics of a ribosome along an mRNA transcript can be seen as a Markov jump process, the associated conditional hopping time at each site is exponentially distributed, with its rate defined as the elongation rate at the site. 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 (for further details, see **Supplementary Methods**).

Definition of closely stacked ribosomes

During simulation we monitor the distance between consecutive ribosomes along the transcript. Since experimental "disome" fragments (i.e., footprints covering two stacked ribosomes) were shown (10) to protect a broad range of sizes below ~ 65 nt, in our simulations we defined closely stacked 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 with examples is provided in **Supplementary Methods**. Briefly, for given experimental ribosome profile and detected density (average number of detected ribosomes 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 (Supplementary Figure S1). 1) First, we approximate the position-specific elongation rate by taking the inverse of the observed footprint number (such approximation is valid when there is no ribosomal interference, see **Supplementary Methods**), 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 located downstream 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 (35, 36, 37). However, since translation is generally limited by initiation, not by elongation, under realistic physiological conditions (14, 38), typically only a few sites were affected by interference (Supplementary Figure S2A). 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 (Supplementary Figure S3).

Ribosome cryo-EM data and exit tunnel extraction

The ribosome crystallographic structure from Schmidt *et al.* (39) (Protein Data Bank ID 5GAK, resolution ~ 3.88Å) was used to study the ribosomal exit tunnel, and the structure was visualized using Pymol. We extracted the tunnel coordinates using MOLE 2.0 software (40), and used custom python and Matlab scripts to compute the radius and charge properties.

Software implementation

Simulation of translation and our inference algorithm were implemented in Matlab. We simulated the model using the next reaction method (41) 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 (see **Supplementary Methods**). To simulate a ribosome profile of size N, we first simulated $\sim 10^4$ 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.

Software and dataset availability

A software package for simulating ribosome footprint profiles and inferring rates will be made freely available.

RESULTS

Inference of initiation rates and local elongation rates

We developed an inference procedure based on an extended version of the biophysical TASEP model (32) (see Figure 1A, Supplementary Figure S1, **Materials and Methods**, and **Supplementary**

Methods) 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 P-site of a lower ribosomal subunit arrives at the start codon and the upper ribosomal subunit gets assembled, 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.* (15). These data have been shown to have substantial improvements over previous datasets, alleviating protocol-specific biases (15) that can influence interpretation of ribosome-profiling experiments (11). 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 Methods** and Supplementary Figure S3). Of these, 383 genes (45%) did not require any corrections after the first step of our inference procedure, which means that only the initiation rate was fitted to match the data (the inverse of the observed density was used to estimate the elongation rate, see **Materials and Methods**). For the remaining 467 genes, the number of inconsistent sites that required a correction procedure was on average 1.57 per gene (std = 0.925) (see **Supplementary Methods** and Supplementary Figure S2A). Figure 1B is an example illustrating the excellent agreement between the actual ribosome footprint distribution for a specific gene from experiment and the distribution of detected ribosomes obtained from simulation under the extended TASEP model with our inferred initiation and elongation rates. Other comparisons of experimental and simulated profiles are provided in Supplementary Figure S4.

Stacked ribosomes are missing from ribosome profiles

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 possibly not included in the experimental data, thus making the ribosome footprint distribution inaccurate in regions of high traffic (8, 10, 11, 12). 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.* (29) (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 Figure 2A). Interestingly, when we looked at a subset of higher-density transcripts ($>$ 1 ribosome per 100 codons) contained in Arava *et al.*'s dataset, we found that the normalization constant obtained by fitting these higher densities was lower (0.61), as shown in Figure 2B. This suggested that for highly occupied transcripts, the density of ribosome inferred from TE underestimates the actual total ribosomal density. Using Pop *et al.*'s data and performing the same comparisons led to similar results (Supplementary Figure S5A-B).

To see if our method could accurately capture this difference, we compared the experimentally measured densities (in Arava *et al.*) with our simulated average densities. Specifically, we simulated average densities using the rates inferred under three different scenarios of undetected ribosomes (Supplementary Figure S6): 1) when all closely-stacked ribosomes are detected, 2) when each detection is only partially successful with probability 0.5, and 3) with no detection of the closely-

stacked ribosomes. For each scenario, we produced a linear fit of the total ribosome density (obtained by combining detected and undetected ribosomes) against Arava *et al.*'s data. Our goal here was to see which scenario produces a linear-fit coefficient that is the closest to the aforementioned normalization constant (0.83) for low-density transcripts (with < 1 ribosome per 100 codons). The linear-fit coefficient was 0.63 for the first scenario (complete detection of closely-stacked ribosomes), 0.7 for the second (partial detection), and 0.80 for the last (no detection). The last scenario agrees the best with the normalization constant (0.83) for low-density transcripts (Figure 2C). Furthermore, when we used this model to fit the density of only *detected* ribosomes against Arava *et al.*'s data (Figure 2D), we found the normalization constant to be lower (0.67), consistent with the decrease we observed in the fit of the raw TE values for high-density transcripts (Figure 2B). Applying our inference procedure to Pop *et al.*'s data yielded similar results (see Supplementary Figure S5C-D). 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.

Inferred rates are consistent with existing results and across different datasets

Upon selecting a model of ribosome profiling with no detection of closely-stacked ribosomes, we used the corresponding 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 (14) (4 s to 233 s for the 5th to 95th interpercentile range, 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 (14, 15, 42), our inferred initiation rates were also 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 Figure 3A). We also compared our estimated initiation rates with the ones inferred by Ciandrini *et al.* (42), who develop a simpler approach to infer initiation rates from polysome profile data (43). The mean initiation times from Ciandrini *et al.* were close to ours (4.6 s to 15 s for the 5th to 95th interpercentile range, with median $= 8$ s, Supplementary Figure S7A), and a direct comparison between the two sets of initiation rates showed a positive correlation ($R^2 = 0.4$, see Supplementary Figure S7B).

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 Figure 3B). 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.* (20) ($r = 0.7$, p-value $< 4 \times 10^{-10}$, see Figure 3C), 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.* (28) and Pop *et al.* (18). Because of lower sequencing depth compared to Weinberg *et al.*'s data, the number of genes passing our selection criteria decreased to 625 genes for Williams and 212 for Pop (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 Supplementary Figure S8A. We obtained a positive but less good correlation with Pop's dataset ($r = 0.58$, p-value $< 4 \times 10^{-6}$, see Supplementary Figure S8B). We explained this less good correlation by the decrease in sequencing depth (which creates more sites with no footprints). By selecting the 30 codons with largest sample size used to compute the associated averaged codon elongation rate, the r coefficient indeed increased

to 0.68.

Finally, since the differences in MER at different sites could be associated with tRNA availability variations (23), we further compared the MER and the codon tAI value (13, 44), 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 Figure 3C). Altogether, these results suggested that our estimates of the local elongation rates reflect tRNA-dependent regulation of elongation speed and that our estimates are consistent across different ribosome profile datasets.

Isolated and stacked ribosome distributions across genes and codon positions

As discussed earlier, we found that most stacked ribosomes are not present in our dataset and that highly occupied transcripts tend to have a significant amount of undetected ribosomes. To study this behavior more closely, we 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. (Note that, by definition, strictly-stacked ribosomes are also closely-stacked, so the proportion of the latter is at least as large as the former.) We found that on average 12% of ribosomes were closely stacked (Figure 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 on average be lower for the entire set of genes. By extrapolating our estimates to a larger sample of 3941 genes (using their TE, see Figure 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 and the movement of about 2.5% of ribosomes is obstructed on average.

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 (Figure 4A). Such heterogeneity can be explained by differences in the total ribosome occupancy. In Figure 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 elongation speed along the ORF, which could result in ribosomal interference. To examine their influence, we first plotted (Figure 4C) the inferred initiation rate against the fraction of stacked ribosomes 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 (Figure 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 (15), in particular with the trend of decreasing ribosome-footprint density forming the $5'$ translation ramp (1, 13) (see Supplementary Figure S9A).

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, Figure 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 (45, 46)) 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. Interestingly, while the high level of interference near the termination site is consistent with there being a larger amount of footprint reads at the stop codon (see Supplementary Figure S9B), we did not observe in the experimental ribosome footprint data any residual peak at position -10 that would indicate the presence of queuing. This absence of a peak at position -10 was consistent (see Supplementary Figure S10A and Carja *et al.* (47)) across other flash-freeze datasets from yeast (10, 18, 20, 28, 34, 48, 49, 50), suggesting that stacked ribosomes (more likely to be present at the end due to slow termination) do not get detected in yeast under the standard ribosome profiling protocol (in contrast, such peaks have been detected for other organisms and different protocols (51, 52)). To confirm the failure of detection, we simulated metagene profiles using the rates inferred under the models of complete and partial detection of stacked ribosomes, as done in Supplementary Figure S6. These simulations led to peaks at position -10 from the stop codon, which was inconsistent with the experimental data (see Supplementary Figure S10B).

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 our model and inferred parameters, we estimated the protein production rate using the particle flux: for each gene, we defined it 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 (Figure 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 (Figure 5A) gave a high correlation (Pearson's $r = 0.91$). However, we observed a super-linear increase of the production rate as the detected-ribosome density increased. Since our simulated detected-densities match the experimental TE measurements up to a normalization constant, this suggests that, because closely-stacked ribosomes are not included, the standard TE measure tends to underestimate the true protein production rate for large-TE genes. Using the total density of ribosomes (Figure 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 (Figure 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 curve representing the fraction of stacked ribosomes in Figure 4D (with a shift such that the peak is observed 10 codons downstream) and the $5'$ ramp obtained for ribosomal normalized density (Supplementary Figure S9 A). 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 (Figure 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 (13). However, this explanation has been recently argued to be insufficient (15), 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 (Figure 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 (Supplementary Figure S11A), 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 (Supplementary Figure S12A). 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 elongation 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 (Figure 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. Supplementary Figure S11B 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 (Supplementary Figure S12B), 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 Figure 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 (53). 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 (see **Supplementary Methods** for a detailed description). We first analyzed Weinberg *et al.*'s (15) data discussed above (850 genes). 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). We used various features in our linear models. One feature was the average PARS score, which reflects the existence of mRNA secondary structure (see **Materials and Methods**), over a window $[a : b]$ located downstream of the A-site: for a given A-site i, the PARS score window $[a : b]$ refers to positions from $i + a$ to $i + b$. The other features were related to the nascent polypeptide properties, namely total positive charges, negative charges, and hydropathy scores (54). For these features, we used different windows located upstream of the A-site; in this case, a window $[a : b]$ for a given A-site i refers to positions from $i - b$ to $i - a$. A more detailed description of the features is given in **Supplementary Methods** and Supplementary Figure S13.

About 40 or so amino acid residues can be accommodated within the ribosome (27), so we first considered codon positions 6 to 44 from the start codon in order to focus on the dynamics as the N-terminus of the nascent polypeptide chain makes its 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 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 (55) 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 used as additional features the mean number of charged amino acid residues and scanned linear models with different feature windows to

obtain the best fit, we found that the number of positively charged residues in the window $[1:11]$ and the number of negatively charged residues in the window $[6:14]$ upstream of the A-site 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 91% of the positional variation (Figure 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 Figure 7B. This result implies that the more hydrophobic the nascent polypeptide segment is, the higher the mean elongation rate.

We then took the above-mentioned features that we learned from analyzing the data from Weinberg *et al.* (15) and used them to fit the previously-mentioned ribosome profiling data for 625 genes from Williams *et al.* (28). This led to fits with goodness comparable to the ones mentioned above: $R^2 = 0.86$ for codon positions 6 to 44, $R^2 = 0.74$ for positions 45 to 300, and $R^2 = 0.74$ for the entire region between positions 6 and 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**).

The charge features that modulate elongation rates are consistent with the electrostatic properties of the ribosome exit tunnel

To explain why the aforementioned windows of charge features got selected by our linear model of elongation rate variation, we studied the properties of the ribosome exit tunnel. To this end, we first extracted the ribosome tunnel coordinates and composition from cryo-EM data (39) using a tunnel detection algorithm (40) (see **Materials and Methods** and Figure 8). The tunnel spans more than 80 Å and is composed of three regions: the upper region connected to the PTC, the constriction region (where two ribosomal proteins L4 and L22 reduce the width of the tunnel (27)), and the lower region connected to the exit (Figure 8A). Since our statistical analysis suggested that the presence of positive and negative charges in the upper region may respectively facilitate and inhibit elongation as the nascent polypeptide makes its initial pass through the tunnel (i.e., when the ribosome is translating the first ~ 40 codons), we aimed to study the longitudinal direction of the force that a charged particle would experience along the tunnel. Having charges from its constituent RNAs and proteins, the ribosome creates an electrostatic potential within the tunnel. The spatial variation of this potential creates an electric field, and therefore a charged particle inside the tunnel experiences a force, the direction of which is determined by the orientation of the electric field and the charge of the particle.

Upon extracting the coordinates of the tunnel, we found its radius to vary in the range of 2–8 Å (Figure 8B). In such a three-dimensional structure with a varying cross-section size, the diffusion of a particle can be treated as a one-dimensional process with an entropy barrier, described by the so-called Fick-Jacobs diffusion equation (56). The Fick-Jacobs equation has the same structure as the Smoluchowski equation for diffusion in a one-dimensional potential (56), where the potential

$S(x)$ arises from the entropy along the tunnel, determined by the cross-sectional area as

$$S(x) = \ln(\pi r(x)^2), \tag{1}$$

where $r(x)$ is the radius of the cross-section at position x along the tunnel. Thus, the movement of a charged particle across the tunnel is driven by two potentials: one associated with the entropy of the system and the second with the electrostatic potential along the tunnel. The local signs of the gradients of these potentials determine whether the movement of a charged particle towards the exit of the tunnel is facilitated or inhibited.

By studying the entropic force (Figure 8B), we found that the gradient was small in the constriction and lower regions but positive and much larger in the first half of the upper region (first $\sim$ 20Å). Thus, a particle crossing this region will experience a strong entropy barrier due to the increase of radius in the tunnel. For the electrostatic force, due to the heterogeneity of the solvent inside the tunnel (57), it is difficult to accurately compute the electrostatic potential inside the tunnel using classical numerical methods (58). Therefore, we simply computed along the tunnel the net charge contained within a radius of 20 Å from the central axis of the tunnel, and assumed that the variation of this charge distribution across the tunnel follows the variation of the electrostatic potential. Under this assumption, we found that the electric field is outward-pointing in the first half of the upper region (0–20 Å) (Figure 8C). As the tunnel can accommodate approximately 40 amino acids (27), we concluded that this is consistent with our results that the presence of positively charged residues in the first 11 amino acids of the nascent polypeptide tends to facilitate elongation, whereas negatively charged residues between positions 6 and 14 tend to slow it down (see Discussion below). Moreover, these results suggest that the presence of positively charged amino acids should help in the first steps of elongation to move the N-terminus of the nascent polypeptide across the entropy barrier and towards the tunnel exit.

Finally, we studied whether there is an evolutionary signature that is consistent with our finding regarding the specific role that electrostatic interaction plays in modulating elongation speed as the N-terminus of the nascent polypeptide makes its way through the tunnel. To this end, we looked at the position-specific frequency of positive and negative amino acid charges over an extended set of genes (all 2862 genes of length $\geq$ 200 codons) in the first 200 codons of the ORF. As shown in Figure 8D, the frequency of positive charges was much larger in the first 40 codons, followed by a gradual decrease to a plateau ($\sim$ 12% at position 50 and $\sim$ 11% at 200). In contrast, we observed a stark depletion in the frequency of negative charges in the beginning of the ORF, followed by an increase to a plateau starting around position 40 ($\sim$ 12.5%). Overall, these patterns are consistent with the results of our statistical analysis and our hypothesis that evolution has tried to optimize charge distributions in the beginning of the nascent polypeptide to facilitate elongation.

DISCUSSION

Difference of our method from previous methods

We used probabilistic modeling of the translation dynamics to dissect the different determinants of elongation speed, and developed an efficient, simulation-based inference algorithm to estimate transcript-specific initiation and local elongation rates from ribosome profiling data. The first step of our inference procedure is similar to the method introduced by Ciandrini *et al.* (42), which uses a TASEP-based model to infer gene-specific initiation rates. As Ciandrini *et al.*'s method uses only the ribosome density from polysome profile, it assumes that the elongation rate depends only on the codon identity at the A-site, neglecting other determinants that we notably found in our study

to explain the elongation rate variability. While our estimates of initiation rates were of similar order and positively correlated, it is interesting to note that the correlation decreases for genes with higher initiation rates. This could be due to the additional information (specifically, the positions of ribosomes on mRNAs) provided by ribosome profiling that are missing in their method. For example, Ciandrini *et al.* assumed that termination is a fast process, although our results show the contrary. This suggests that they may underestimate the impact of interference, causing them to overestimate the initiation rate necessary to reach a given density. This is consistent with the fact that their estimated initiation rates are in general slightly larger than our estimates.

Gritsenko *et al.* (19) proposed another TASEP-based approach to estimate initiation and elongation rates from ribosome profiling data. However, in their approach only 61 parameters of elongation are estimated by minimizing an objective function over all the profiles. The observed difference between tAI-based and their fitted elongation rates led them to conclude that additional unknown factors, possibly arising from larger sequence context, are shaping elongation rates. Our results illustrate this point more precisely and show that some specific properties of the nascent polypeptide can explain the variability of elongation rates that we observed across different transcripts and codon positions.

As detailed in a recent review by Zur and Tuller (59), there has been many previous studies using computational tools to infer and predict the dynamics of mRNA translation using biophysical modeling (9, 13, 14, 18, 19, 59), but with contradictory results. The models proposed to date have been developed under the assumption that elongation rates are not influenced by the sequence context surrounding the A-site, while our results suggest that elongation rates are modulated by the nascent polypeptide interaction with the exit tunnel that depend on the context of ~ 40 codons preceding the A-site.

Potential technical artifacts

When combining a biophysical modeling approach with ribosome profiling data, another 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 (15, 48, 60). 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 (15). Andreev *et al.* (11) recently described several important artifacts and biases associated with ribosome profiling data that affect the representation of translation dynamics. The experimental protocol used for the main flash-freeze data (15) considered in this paper minimizes some of the biases (such as sequence biases introduced during ribosome footprint library preparation and conversion to cDNA for subsequent sequencing, and mRNA-abundance measurement biases and other artifacts caused by poly(A) selection). In addition, our method allowed us to correct other important biases related to TE measurements and depletion of stacked ribosomes from selecting only ~ 30 nt fragments.

Undetected ribosomes and the extent of interference

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 the movement of $\sim 2.5\%$ of ribosomes is obstructed on average. 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 larger ribosome-protected fragments is notably needed to fully characterize ribosome occupancy (10).

The major determinants of the elongation rate

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, which has been used as a proxy for codon-specific decoding rate (9, 42, 44, 59), supports the hypothesis that tRNA abundance and codon usage co-evolved to optimize translation rates (13, 61).

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 (13, 62). An earlier study (53) 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 elongation 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 (16, 26, 27, 63, 64). One of the major findings of our work is that the variation of the mean elongation rate along the transcript can be well explained by linear models in two distinct regions of the transcript which correspond to two different regimes of translation. While the N-terminus of the nascent polypeptide has not yet escaped the tunnel, the possible presence of mRNA secondary structure downstream of the A-site and the amount of charged amino acid residues in the nascent polypeptide at the beginning of the tunnel modulate the elongation rate. Once the N-terminus has exited the tunnel, the hydropathy of the part of the nascent polypeptide within the ribosome plays a major role in governing the elongation rate variation. These features were selected by statistically optimizing the fit of the linear model to position-specific mean elongation rates in a large region, which included the $5'$ ramp.

We note that Tuller *et al.* (62) also employed linear regression to fit the $5'$ ramp. However, their model and results are quite different from ours, as they fitted a smoothed version of the normalized average ribosome footprint density along the transcript (see Supplementary Figure S9A), whereas we fitted the position-specific deviation from the mean elongation rate. By selecting some features by hand (tAI value, the total charge of the amino acid residues coded by the 13 codons upstream of the A-site, and the $5'$ folding energy downstream of the A-site), they could explain the variation of ribosome density in the first 50 codons, whereas we tried to optimize over features and our resulting fit is globally good over the first 300 codons (Figure 7C).

Possible biophysical explanations

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 (27, 53, 65). When the nascent

polypeptide has not yet exited the tunnel (i.e., the first 40 residues), our statistical analysis found that charged amino acid residues near the A-site play an important role in governing the elongation dynamics. Furthermore, we found that positive charges and negative charges have opposite effects: the former facilitates the elongation speed, while the latter inhibit it. This finding is consistent with the electrostatic properties of the tunnel. Specifically, our estimations of the net local charge across the tunnel suggests that the electric field induced by the potential points outward (i.e., away from the PTC) near the beginning of the tunnel (Figure 8C), in agreement with what our linear model predicts. Previous measurements of the electrostatic potential inside the tunnel (66) were also consistent with our findings, suggesting a decrease of the potential from the PTC along the upper region. Hence, positively charged residues near the beginning of the tunnel will experience an electrostatic force pointing outward, thereby facilitating the movement of the polypeptide chain through the tunnel. Moreover, we showed that in the context of studying particle diffusion inside the tunnel, the increase of radius across the tunnel in the upper region (Figure 8B) creates a strong entropic barrier. This barrier can be compensated by the electrostatic potential if the particle is positively charged, explaining the specific selection of positively charged amino acids in the upper region when the nascent polypeptide makes its initial pass through the tunnel. The opposite applies to negatively charged residues, with an effect of inhibiting the movement of the chain.

In seeming contrast to our finding, it was previously suggested that positively charged amino acids slow down translation (16, 53). We note, however, that these studies did not focus on the initial stage of elongation. Our results from statistical linear modeling and analyzing the structure of the ribosome exit tunnel suggest that while the N-terminus of the nascent polypeptide has not exited from the tunnel, positively charged amino acids in specific parts of the polypeptide actually facilitate the elongation speed. Furthermore, in contrast to Charneski and Hurst's results (16), we saw an opposite effect for negatively charged amino acids. Averaging the charged amino acid frequency over all transcripts of length ≥ 200 codons, we found that there is a starkly elevated amount of positively charged amino acids in the first 25 codons, while the opposite is true for negatively charged amino acids (see Figure 8D). These patterns are consistent with our proposed role of positive and negative charges, and suggest that evolution has tried to optimize charge distributions to facilitate the translation dynamics as the nascent polypeptide makes its initial pass through the exit tunnel.

Another important feature, which to our knowledge has not been previously noted as a major determinant of elongation 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 (27) and wide enough to allow the formation of α-helical structure (65), the hydrophobicity (which is an important factor driving compactness and rigidity (67)) 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 (68), 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. Interestingly, the observed variation in the mean hydropathy score along the transcript (Supplementary Figure S14) suggests that the elongation speed is regulated at different stages of the polypeptide assembly and folding. The selection of different determinant features for different stages of translation also suggests that the movement of the polypeptide inside the tunnel is driven by two distinct biophysical mechanisms: First, when the polypeptide chain has not yet exited the tunnel, electrostatic interactions in the upper part of the tunnel play a major role in regulating the movement of the chain down the exit tunnel. Second, when the polypeptide has reached a certain length and its N-terminus has exited the tunnel, it is the structure of the chain itself (which we captured through the hydropathy) that determines its movement through the tunnel.

Possible limitations

One of the possible limitations of our present approach is that it does not take into account possible translational regulatory events which could be specific to some genes (such as co-translational translocation (69)) or sequence motifs (such as arrest sequences inducing mRNA cleavage (70)). Another regulatory mechanism is ribosomal drop-off, which has been shown to occur for some sequences that lead to pausing or under specific stress conditions (21). Under a non-stress environment, it has been hypothesized that there exists a "basal" drop-off rate and it has been estimated to be on the order of 10^{-4} per codon in *E. Coli*, assuming sequence-specific features to be well averaged out. This has led to the hypothesis that ribosomal drop-off could explain the ramp variations (21). Interestingly, using the same method as in Sin *et al.* (21) (see **Supplementary Methods**), we came to different estimates of the drop-off rate for the ramp region and the rest of the transcript (Supplementary Figure S15A-B). More precisely, while the drop-off rate we estimated for the region outside the ramp was consistent with the drop-off rate estimate (3.7×10^{-4}) from Sin *et al.*, the estimation procedure led to an unrealistically large drop-off rate in the ramp region (0.002, which leads to a survival probability of only 0.67 after 200 translated codons). Incorporating the basal drop-off rate of 3.7×10^{-4} per codon into our inference procedure did not significantly change our rate estimates (see **Supplementary Methods** and Supplementary Figure S15C). Another possible drawback of our work is that our main analysis was carried out on a subset of only 850 genes, which were selected to assure sufficiently high local coverage-depth of footprints. To confirm that our results and conclusions did not suffer from any potential biases due to such filtering, we analyzed the ramp pattern and the codon-specific elongation rates obtained from a larger dataset (2862 genes) consisting of all genes of length ≥ 200 codons. Comparison with our original results (see Supplementary Figure S16) did not show any significant difference, suggesting that our overall conclusions are robust and applicable to a broader level.

CONCLUSION

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 along the transcript sequence. The results from our statistical analysis suggest that the translation elongation dynamics while the nascent polypeptide is initially passing through the ribosome exit tunnel is rather different from the elongation dynamics after the N-terminus has escaped the tunnel, and that different biophysical mechanisms modulate the elongation speed in the two stages. 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.

Since the ribosome structure is highly conserved (71, 72), we believe that our results can be generalized to other organisms. However, applying the present method to other datasets to estimate translation rates may not be straightforward. Differences in experimental protocols and nuclease digestion conditions could affect the data substantially (73), and in particular make the fraction of detected stacked-ribosomes to vary across different datasets and organisms. However, combining polysome profiling, TE measurements, and different profile simulation models (with different probabilities of detecting stacked-ribosomes) can allow one to estimate the proportion of undetected stacked-ribosomes (as done in Figure 2 and Supplementary Figure S6), and also enable

the estimation of translation rates upon selecting the best model. Finally, 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 (68, 74, 75).

ACKNOWLEDGEMENTS

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, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science,* **324**(5924), 218–223.

2. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols,* **7**(8), 1534–1550.

3. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2013) Genome-Wide Annotation and Quantitation of Translation by Ribosome Profiling. *Current Protocols in Molecular Biology,* pp. 4–18.

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

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

6. Brar, G. A. and Weissman, J. S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology,* **16**, 651–664.

7. Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell,* **147**(4), 789–802.

8. Subramaniam, A. R., Zid, B. M., and O'Shea, E. K. (2014) An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell,* **159**(5), 1200–1211.

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

10. Guydosh, N. R. and Green, R. (2014) Dom34 rescues ribosomes in 3' untranslated regions. *Cell,* **156**(5), 950–962.

11. Andreev, D. E., O'Connor, P. B., Loughran, G., Dmitriev, S. E., Baranov, P. V., and Shatsky, I. N. (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Research,* **45**(2), 513–526.

12. Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. A., Kramer, G., et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell,* **147**(6), 1295–1308.

13. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell,* **141**(2), 344–354.

14. Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. B. (2013) Rate-limiting steps in yeast protein translation. *Cell,* **153**(7), 1589–1601.

15. Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. (2016) Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports,* **14**(7), 1787–1799.

16. Charneski, C. A. and Hurst, L. D. (2013) Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol,* **11**(3), e1001508.

17. Artieri, C. G. and Fraser, H. B. (2014) Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Research,* **24**(12), 2011–2021.

18. Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., and Koller, D. (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular Systems Biology,* **10**(12), 770.

19. Gritsenko, A. A., Hulsman, M., Reinders, M. J., and de Ridder, D. (2015) Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLoS Comput Biol,* **11**(8), e1004336.

20. Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *Elife,* **3**, e03735.

21. Sin, C., Chiarugi, D., and Valleriani, A. (2016) Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Research,* pp. 2528–2537.

22. Zhang, G., Fedyunin, I., Miekley, O., Valleriani, A., Moura, A., and Ignatova, Z. (2010) Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Research,* **38**(14), 4778–4787.

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

24. Qu, X., Wen, J.-D., Lancaster, L., Noller, H. F., Bustamante, C., and Tinoco, I. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature,* **475**(7354), 118–121.

25. Chevance, F. F., Le Guyon, S., and Hughes, K. T. (2014) The effects of codon context on in vivo translation speed. *PLoS Genet,* **10**(6), e1004392.

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

27. Ito, K., (ed.) (2014) Regulatory Nascent Polypeptides, Springer, .

28. Williams, C. C., Jan, C. H., and Weissman, J. S. (2014) Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science,* **346**(6210), 748–751.

29. Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. (April, 2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences,* **100**(7), 3889–3894.

30. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly,* **125**(2), 167–188.

31. Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature,* **467**(7311), 103–107.

32. Spitzer, F. (1970) Interaction of Markov processes. *Advances in Mathematics,* **5**(2), 246–290.

33. MacDonald, C. T., Gibbs, J. H., and Pipkin, A. C. (1968) Kinetics of biopolymerization on nucleic acid templates. *Biopolymers,* **6**(1), 1–25.

34. Lareau, L. F., Hite, D. H., Hogan, G. J., and Brown, P. O. (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife,* **3**, e01257.

35. Shaw, L. B., Kolomeisky, A. B., and Lee, K. H. (2004) Local inhomogeneity in asymmetric simple exclusion processes with extended objects. *Journal of Physics A: Mathematical and General,* **37**(6), 2105.

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

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

38. Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M. F., and von der Haar, T. (2014) Translation elongation can control translation initiation on eukaryotic mRNAs. *The EMBO journal,* **33**(1), 21–34.

39. Schmidt, C., Becker, T., Heuer, A., Braunger, K., Shanmuganathan, V., Pech, M., Berninghausen, O., Wilson, D. N., and Beckmann, R. (2016) Structure of the hypusinylated eukaryotic translation factor eIF-5A bound to the ribosome. *Nucleic Acids Research,* **44**(4), 1944.

40. Sehnal, D., Vařeková, R. S., Berka, K., Pravda, L., Navrátilová, V., Banáš, P., Ionescu, C.-M., Otyepka, M., and Koča, J. (2013) MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of cheminformatics,* **5**(1), 1.

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

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

43. MacKay, V. L., Li, X., Flory, M. R., Turcott, E., Law, G. L., Serikawa, K. A., Xu, X., Lee, H., Goodlett, D. R., Aebersold, R., et al. (2004) Gene expression analyzed by high-resolution state array analysis and quantitative proteomics response of yeast to mating pheromone. *Molecular & Cellular Proteomics,* **3**(5), 478–489.

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

45. Yu, X., Willmann, M. R., Anderson, S. J., and Gregory, B. D. (2016) 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,* **28**(10), 2385–2397.

46. Pelechano, V., Wei, W., and Steinmetz, L. M. (2015) Widespread co-translational RNA decay reveals ribosome dynamics. *Cell,* **161**(6), 1400–1412.

47. Carja, O., Xing, T., Plotkin, J. B., and Shah, P. (2017) riboviz: analysis and visualization of ribosome profiling datasets. *bioRxiv,* p. 100032.

48. Gerashchenko, M. V. and Gladyshev, V. N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research,* **42**(17), e134–e134.

49. Nedialkova, D. D. and Leidel, S. A. (2015) Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell,* **161**(7), 1606–1618.

50. Jan, C. H., Williams, C. C., and Weissman, J. S. (2014) Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science,* **346**(6210), 1257521.

51. Andreev, D. E., O'Connor, P. B., Zhdanov, A. V., Dmitriev, R. I., Shatsky, I. N., Papkovsky, D. B., and Baranov, P. V. (2015) Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biology,* **16**(1), 90.

52. Lobanov, A. V., Heaphy, S. M., Turanov, A. A., Gerashchenko, M. V., Pucciarelli, S., Devaraj, R. R., Xie, F., Petyuk, V. A., Smith, R. D., Klobutcher, L. A., et al. (2017) Position-dependent termination and widespread obligatory frameshifting in Euplotes translation. *Nature Structural & Molecular Biology,* **24**(1), 61–68.

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

54. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology,* **157**(1), 105–132.

55. Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., et al. (2016) Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature,* **529**(7586), 358–363.

56. Zwanzig, R. (1992) Diffusion past an entropy barrier. *The Journal of Physical Chemistry,* **96**(10), 3926–3930.

57. Lucent, D., Snow, C. D., Aitken, C. E., and Pande, V. S. (2010) Non-bulk-like solvent behavior in the ribosome exit tunnel. *PLoS Comput Biol,* **6**(10), e1000963.

58. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences,* **98**(18), 10037–10041.

59. Zur, H. and Tuller, T. (2016) Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. *Nucleic Acids Research,* **44**(19), 9031–9049.

60. Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., and Press, W. H. (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet,* **11**(12), e1005732.

61. Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science,* **324**(5924), 255–258.

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

63. Pavlov, M. Y., Watts, R. E., Tan, Z., Cornish, V. W., Ehrenberg, M., and Forster, A. C. (2009) Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proceedings of the National Academy of Sciences,* **106**(1), 50–54.

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

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

66. Lu, J., Kobertz, W. R., and Deutsch, C. (2007) Mapping the electrostatic potential within the ribosomal exit tunnel. *Journal of Molecular Biology,* **371**(5), 1378–1391.

67. White, S. H. and Wimley, W. C. (1999) Membrane protein folding and stability: physical principles. *Annual Review of Biophysics and Biomolecular Structure,* **28**(1), 319–365.

68. Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., and Liu, Y. (2015) Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular Cell,* **59**(5), 744–754.

69. Nyathi, Y., Wilkinson, B. M., and Pool, M. R. (2013) Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research,* **1833**(11), 2392–2402.

70. Hayes, C. S. and Sauer, R. T. (2003) Cleavage of the A site mRNA codon during ribosome pausing provides a mechanism for translational quality control. *Molecular cell,* **12**(4), 903–911.

71. Isenbarger, T. A., Carr, C. E., Johnson, S. S., Finney, M., Church, G. M., Gilbert, W., Zuber, M. T., and Ruvkun, G. (2008) The most conserved genome segments for life detection on Earth and other planets. *Origins of Life and Evolution of Biospheres,* **38**(6), 517–533.

72. Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., and Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Research,* **30**(24), 5382–5390.

73. Gerashchenko, M. V. and Gladyshev, V. N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Research,* **45**(2), e6.

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

75. Nissley, D. A., Sharma, A. K., Ahmed, N., Friedrich, U. A., Kramer, G., Bukau, B., and O'Brien, E. P. (2016) Accurate prediction of cellular co-translational folding indicates proteins can switch from post-to co-translational folding. *Nature Communications,* **7**.

76. Zhang, S., Zubay, G., and Goldman, E. (1991) Low-usage codons in Escherichia coli, yeast, fruit fly and primates. *Gene,* **105**(1), 61 – 72.

FIGURES

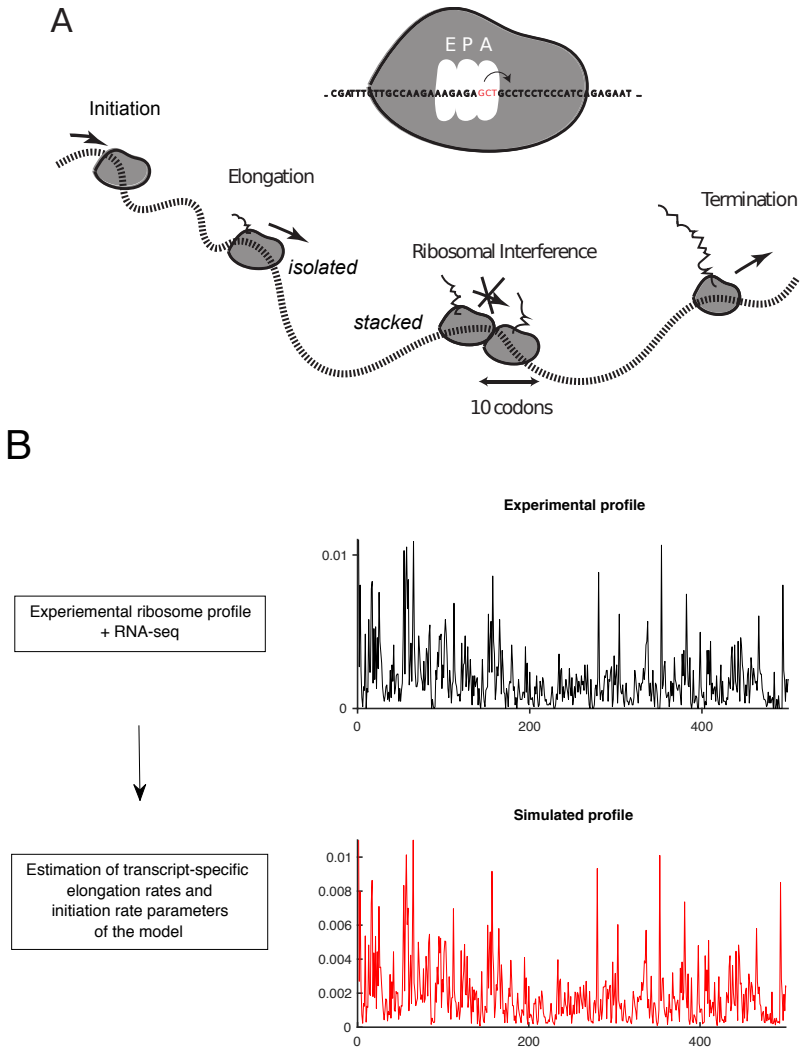

Figure 1. Illustration of translation dynamics and inference from experimental data. **A.** A schematic representation of the mathematical model of translation considered in this paper. Each ribosome is assumed to occupy 10 codons. Initiation corresponds to an event where the A-site of a ribosome enters the second codon position, 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 main 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. More examples of comparisons between experimental and simulated profiles are shown in Supplementary Figure S4.

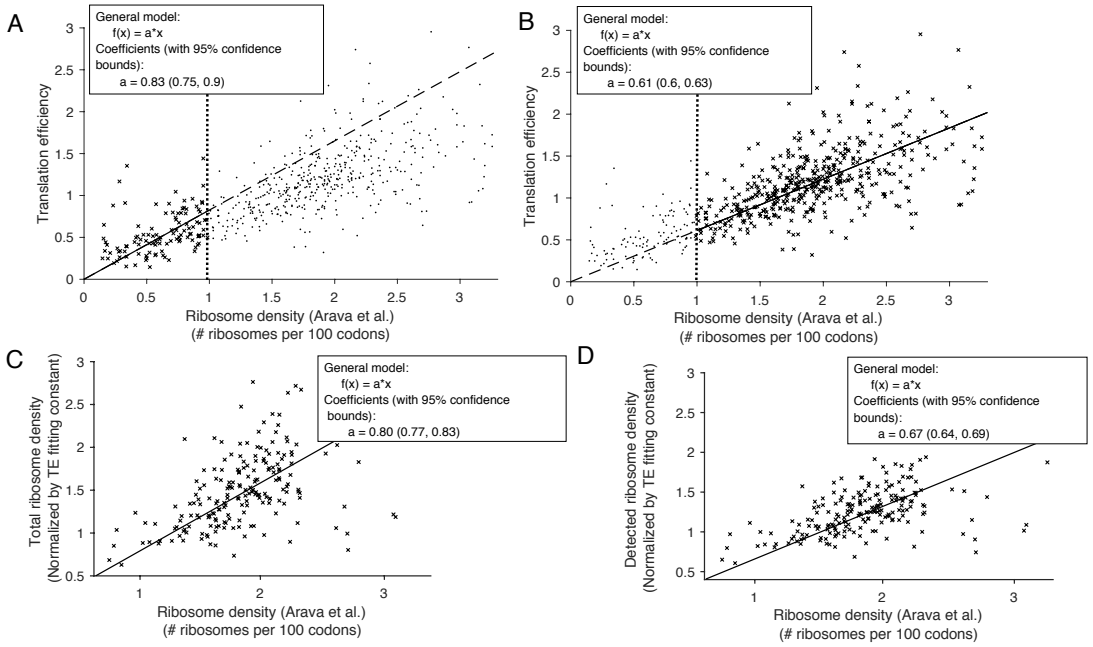

Figure 2. Comparison between translation efficiency (TE) and total ribosome density. All linear fit results are shown in the inset. **A.** The gene-specific TE for 588 genes from Weinberg *et al.*'s data (15) (see **Materials and Methods**) against the corresponding total ribosome density (average number of ribosomes per 100 codons) from Arava *et al.* (29). 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.

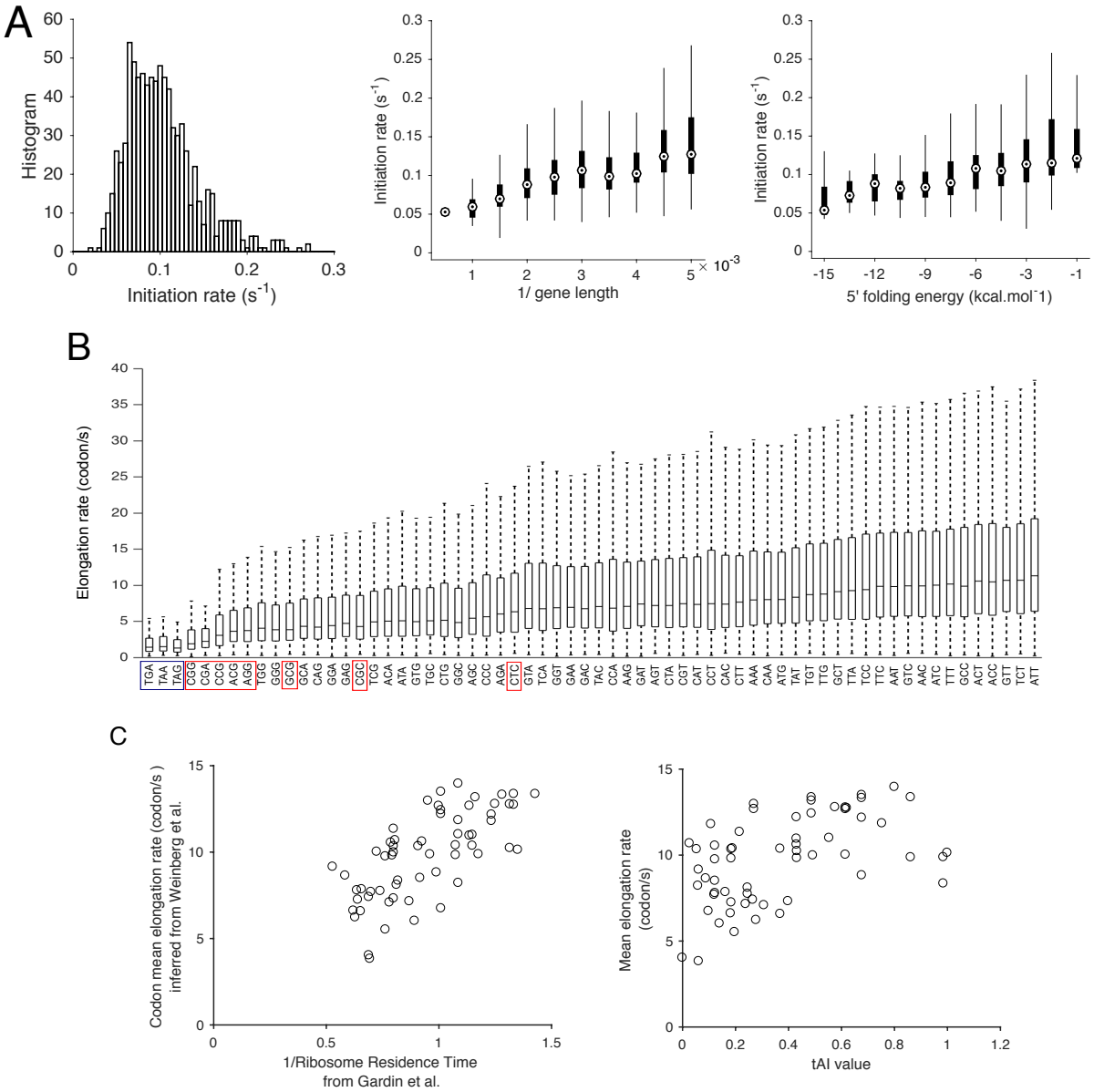

Figure 3. Analysis and comparison of the inferred rates. A. (Left) A histogram of inferred initiation rates. (Middle) Comparison between the inferred initiation rates and the inverse of the ORF length of the gene, showing a positive correlation ($r = 0.44$, p-value $< 10^{-15}$, computed for unbinned data). (Right) Comparison between the inferred initiation rates and the $5'$-cap folding energy computed in Weinberg *et al.* (15), 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.* (76) 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.* (20), and (Right) the tAI value, computed by Tuller *et al.* (13).

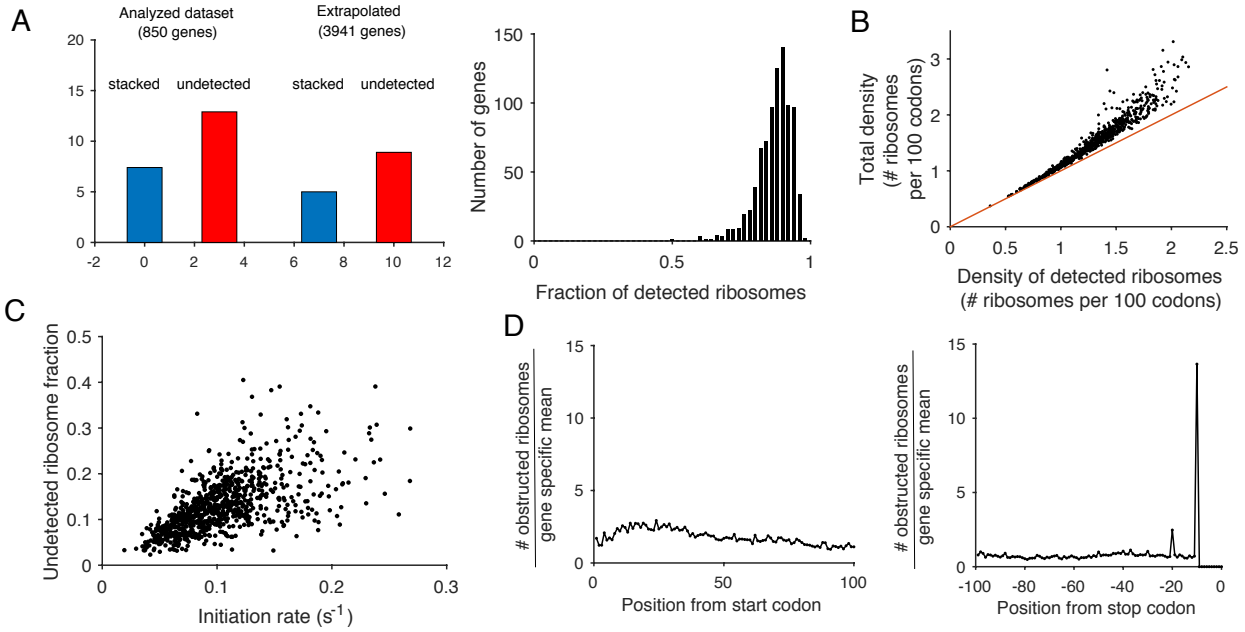

Figure 4. Stacked and undetected ribosome distributions. A. (Left) We first estimated the average proportions of strictly-stacked (blue) and undetected ribosomes (red) across all genes considered (850 genes). We then extrapolated these proportions to 3941 genes by binning the 850 genes by their detected-ribosome 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.

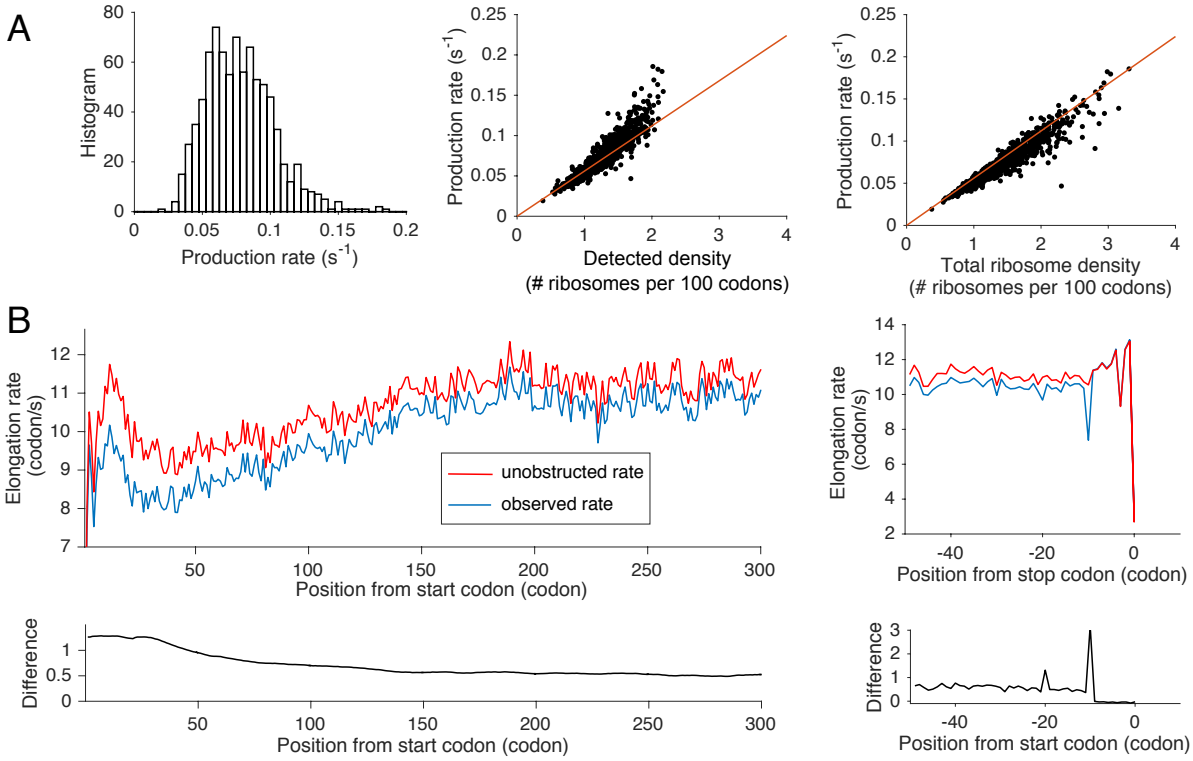

Figure 5. The impact of ribosomal interference on translation dynamics. A. Analysis of protein production. (Left) A histogram of protein production rates. (Middle) Comparison between the protein production rate and the detected-ribosome density obtained from simulations. In red, we plotted the simulated 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.

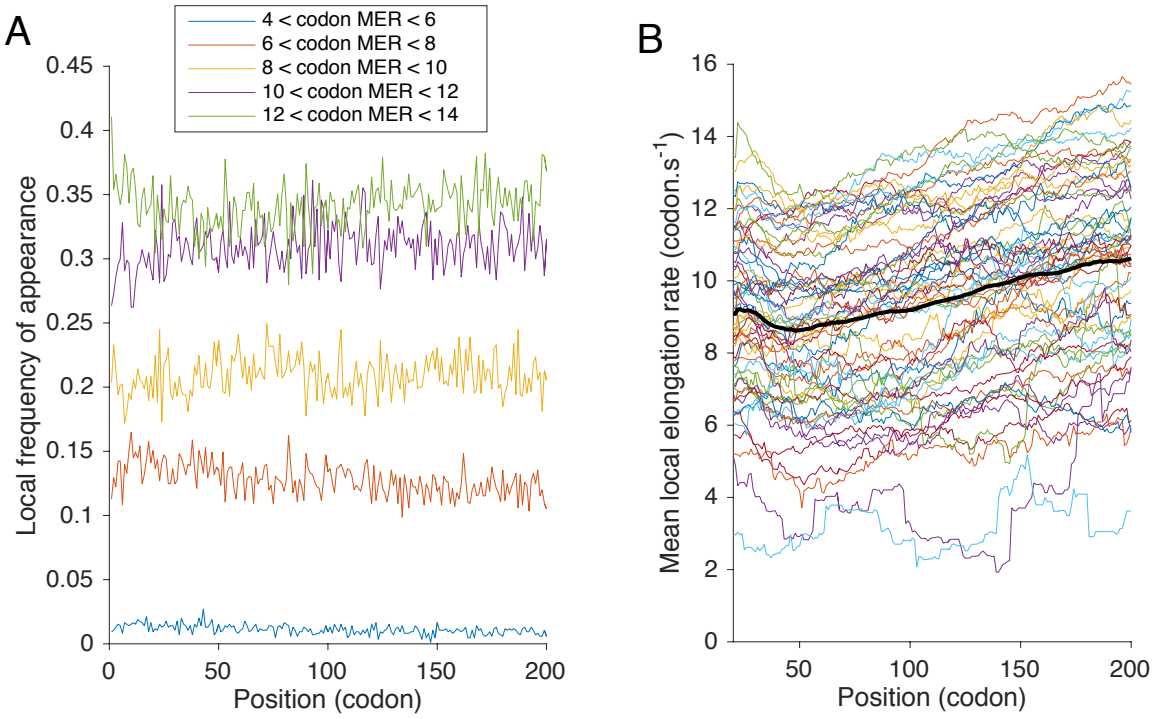

Figure 6. Heterogeneity of codon distributions and elongation speed along the transcript. A. Codon frequency metagene analysis. We grouped the codons (except stop codons) into five groups according to their mean elongation rates (MER) and plotted their frequency of appearance at each position in the 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.

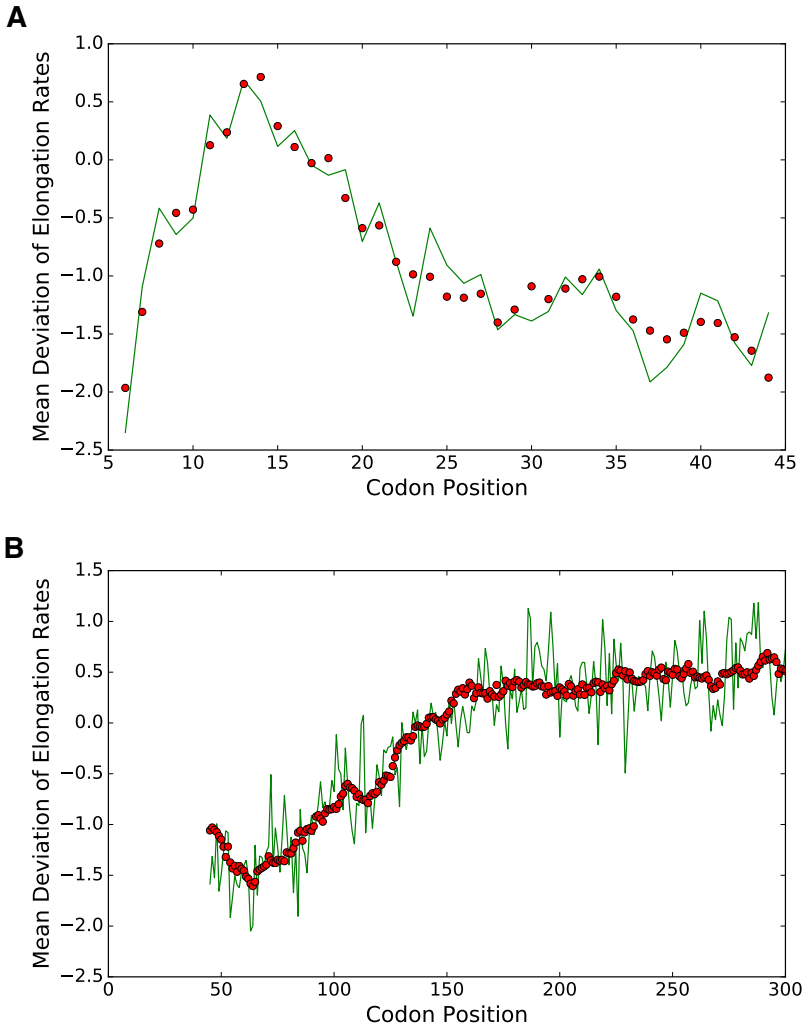

Figure 7. Linear model fits of the mean deviation of elongation rates for the data from Weinberg *et al.* (15). The dependent variable is the mean deviation of elongation rates from codon-type-specific average elongation rates. Green lines correspond to the estimates from ribosome profiling data, while 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 $[6:14]$ upstream of the A-site, and the mean number of positively charged residues in the window $[1:11]$ 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 0.91 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.

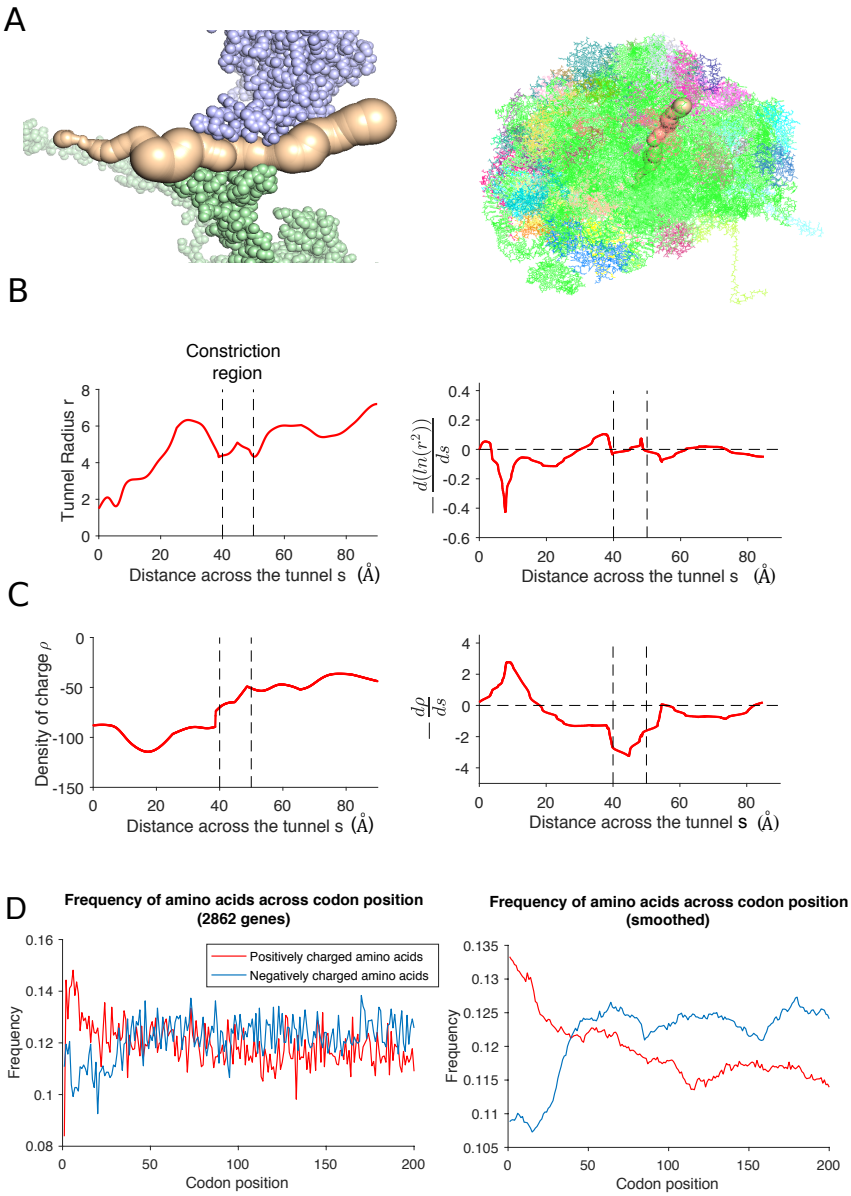

Figure 8. Biophysical properties of the ribosome exit tunnel. A. The figure on the left illustrates the exit tunnel and ribosomal proteins surrounding the constriction region. We extracted the tunnel geometry from the cryo-EM structure (39) illustrated on the right. **B.** (Left) The variation of the tunnel radius r along the tunnel. (Right) The negative gradient of $\ln(r^2)$ (smoothed over a 10 Å window). Note a region of large negative "entropic" potential (56) in the upper region of the exit tunnel. **C.** (Left) The variation of the net charge induced by ribosomal RNA and proteins within a radius of 20 Å from the center of the tunnel. (Right) The negative gradient of the net charge (smoothed over a 10 Å window), notably showing a region of positive electric field (pointing towards the exit) in the region 0-18 Å. **D.** (Left) Position-specific average frequencies of positively (red) and negatively (blue) charged amino acids averaged over all genes of length ≥ 200 codons. (Right) The frequency curves smoothed by averaging over a 10-codon window. Compared to other parts of the transcript, the frequency of positive (negative) amino acids is significantly higher (lower) in the first $\sim$25 codons.