# Theoretical quantification of interference in the TASEP: application to mRNA translation shows near-optimality of termination rates

Khanh Dao Duc[1,2], Zain H. Saleem[1], and Yun S. Song[1,2,3,4,*]

[1] Department of Mathematics, University of Pennsylvania, PA 19104, USA
[2] Department of Biology, University of Pennsylvania, PA 19104, USA
[3] Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA
[4] Chan Zuckerberg Biohub, San Francisco, CA 94158

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

**Abstract**

The Totally Asymmetric Exclusion Process (TASEP) is a classical stochastic model for describing the transport of interacting particles, such as ribosomes moving along the mRNA during translation. Applying this model to quantify translation dynamics from ribosome profiling data is not straightforward, however, and it requires characterizing the extent of interference, since the experimental protocol may be biased against nearby ribosomes. To evaluate and correct for this potential bias, we provide here a theoretical analysis of the distribution of isolated particles in the TASEP. In the classical form of the model in which each particle occupies only a single site, we obtain exact analytic solutions using the Matrix Ansatz. We then employ a refined mean field approach to extend the analysis to a generalized TASEP with particles of an arbitrary size. Our theoretical study has direct applications in mRNA translation and the interpretation of experimental ribosome profiling data. In particular, our analysis of data from *S. cerevisiae* suggests a potential bias against the detection of nearby ribosomes with gap distance less than $\sim 3$ codons, which leads to some ambiguity in estimating the initiation rate and protein production flux for a substantial fraction of genes. Despite such ambiguity, however, we demonstrate theoretically that the interference rate can be robustly estimated, and show that approximately 1% of the translating ribosomes get obstructed. Lastly, we find that, on average, the termination rate is near optimal in that it is close to the minimum value needed to not limit the ribosome flux.

# Introduction

Translation of mRNAs into proteins is one of the most essential biological processes underlying cellular function. To understand its complex dynamics, ribosome profiling (also known as Ribo-Seq) has been developed to examine position-specific densities of ribosomes along each mRNA [1]. However, although this powerful experimental technique captures the dynamics of mRNA translation to some extent, analytical tools for interpreting ribosome profiling data and relating the observed footprint density to the corresponding protein production rate are still much in need of development [2].

In this context, a natural tool to use is the Totally Asymmetric Exclusion Process (TASEP), which is a classical stochastic model for transport phenomena in a non-equilibrium particle system. Although it has been widely studied by mathematicians and physicists, the TASEP was first introduced in a biological context by McDonald et al. [3] to model mRNA translation and describe the dynamics of ribosomes moving along the mRNA. Over the past fifteen years, the TASEP and its extensions have been used for this purpose [4–13], and TASEP-based models have been used to infer the translation rate from experimental data [11, 14–16]. Yet, developing such inference method for ribosome profiling data is not straightforward and still remains challenging for several reasons [17]. One notable issue comes from the experimental protocol used to generate the ribosome profile. In general, long mRNA fragments that may account for stacked ribosomes are not sequenced. As a result, the observed density may only include well-isolated ribosomes, thus leading to a bias that needs to be corrected when evaluating the ribosome density [8, 16–19]. Although the TASEP has been broadly studied under different conditions and using various approaches [20, 21], to our knowledge, the density of isolated particles has not been studied previously.

These theoretical and technical issues motivate us to study the extent of isolated particles in the TASEP, in order to quantify the relation between the mRNA translation dynamics and the observed densities in ribosome profiling data. To do so, we first employ the matrix formulation of Derrida *et al.* [22] to derive exact formulas for the density of isolated particles in the classical TASEP model, in which each particle is pointlike and occupies a single site. For the case when the number $N$ of sites is large, we obtain simple asymptotic formulas. We then extend our study to the general case with particles of an arbitrary size. Using a refined mean field approach introduced by Lakatos and Chou [4], we derive new asymptotic formulas that agree well with Monte Carlo simulations.

We obtain new results regarding the translation dynamics by applying our theory to ribosome profiling data. In particular, our analysis of undetected ribosomes suggests a potential bias against consecutive ribosomes less than $\sim 3$ codons apart. Furthermore, we characterize the variation of the termination rate, which is a key parameter of the TASEP, across different transcripts. We show that the termination rate is on average close to the minimum value needed to achieve maximal current, and hence does not limit the translation speed. Combining these estimates with a measurement of ribosome density called "translation efficiency" (TE), we provide estimates of the detection and interference rates, and find that, for a significant fraction of genes, there is some ambiguity in identifying the initiation rate and the flux from TE. Although the TE has been widely used as a proxy for protein production rate [23], these results suggest that more refined methods and estimates should be used to properly quantify gene expression at the translation level.

# Theoretical Results

We first present our main theoretical results on the classical and generalized TASEP models.

**The density of isolated particles in the classical TASEP model.** We first studied the density of isolated particles in the context of the classical TASEP model with open boundaries [24]. Briefly, the dynamics of this stochastic process can be described as follows (see **Figure 1**A). On a one-dimensional lattice of $N$ sites, the classical TASEP describes the configuration of pointlike particles, described by a vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$ such that $\tau_i = 0$ if the $i$-th site is empty and $\tau_i = 1$ if it is occupied. During every infinitesimal time interval $dt$, each particle at site $i \in \{1, \ldots, N-1\}$ has probability $dt$ of hopping to the next site to its right, provided that the site is empty. Additionally, a new particle enters site 1 with probability $\alpha dt$ if $\tau_1 = 0$. If $\tau_N = 1$, the particle at site $N$ exits the lattice with probability $\beta dt$. The parameters $\alpha$ and $\beta$ are respectively called the initiation and termination rates. In the long time limit, the system reaches steady state and the corresponding expected marginal density of particles at position $i$ on a lattice of size $N$, denoted $\langle \tau_i \rangle_N$, is defined as,

$$\langle \tau_i \rangle_N = \sum_{\tau \in \{0,1\}} \tau \mathbb{P}(\tau_i = \tau) = \mathbb{P}(\tau_i = 1).$$

Averaging the process over the events that may occur between $t$ and $t + dt$ leads to a system of equations relating one-point correlators to two-point correlators [25]. Similarly, two-point correlators can be related to three-point correlators (see Section 1 of Supporting Information), and so on. To derive analytic expressions for the average densities, Derrida *et al.* [22] showed that the steady state probability of a given configuration can be derived using a matrix formulation satisfying a set of algebraic rules (see Section 2 of Supporting Information). Using these rules, they obtained an exact formula for $\langle \tau_i \rangle_N$, and showed that, in the large-$N$ limit, the TASEP follows different dynamics according to a phase diagram in $(\alpha, \beta)$-space.

In our work, we employed the aforementioned matrix formulation to derive analytic expressions for the average density of isolated particles. Specifically, consider the random variable $\tau_i'$ defined as

$$\tau_i' = \begin{cases} \tau_1(1 - \tau_2), & \text{for } i = 1, \\ \tau_i(1 - \tau_{i-1})(1 - \tau_{i+1}), & \text{for } 2 \leq i \leq N-1, \\ \tau_N(1 - \tau_{N-1}), & \text{for } i = N. \end{cases} \tag{1}$$

Note that $\tau_i' = 1$ if there is an isolated particle at position $i$, and $\tau_i' = 0$ otherwise. From (1), we see that the average density $\langle \tau_i' \rangle_N$ of isolated particles at an interior site $i$, where $2 \leq i \leq N-1$, is given by

$$\langle \tau_i' \rangle_N = \langle \tau_i \rangle_N - \langle \tau_{i-1}\tau_i \rangle_N - \langle \tau_i \tau_{i+1} \rangle_N + \langle \tau_{i-1}\tau_i\tau_{i+1} \rangle_N. \tag{2}$$

As detailed in the Supporting Information, by analyzing the terms on the right hand side of (2), we obtained

$$\langle \tau_i' \rangle_N = D_0(\alpha, \beta, N) - D_1(\alpha, \beta, N) \langle \tau_{i-1} \rangle_{N-1}, \quad 2 \leq i \leq N-1, \tag{3}$$

where

$$D_0(\alpha, \beta, N) = \alpha \left[1 - \langle \tau_2 \rangle_N + (1 - \langle \tau_1 \rangle_N)(\langle \tau_1 \rangle_{N-1} - \alpha)\right] \tag{4}$$

$$D_1(\alpha, \beta, N) = \alpha(1 - \langle \tau_1 \rangle_N). \tag{5}$$

For the boundaries, we obtained

$$\langle \tau_1' \rangle_N = \alpha(1 - \langle \tau_1 \rangle_N), \tag{6}$$

$$\langle \tau_N' \rangle_N = \langle \tau_N \rangle_N (1 + \beta) - \langle \tau_{N-1} \rangle_N. \tag{7}$$

As mentioned earlier, exact formulas for $\langle \tau_i \rangle_N$ are known [22] (Supporting Information), so plugging them into (3)–(7) leads to exact results for the average densities of isolated particles along the lattice.

**Large-$N$ asymptotics in three different phases.** We next derived the large-$N$ asymptotics of $\langle \tau_i' \rangle_N$ from those of $\langle \tau_i \rangle_N$. In this section, we drop the dependence on $N$ and write $\langle \tau_i \rangle$ instead of $\langle \tau_i \rangle_N$. In the large $N$ limit, the dynamics of the TASEP can be separated into three different phases—namely, maximal current (MC), low density (LD), and high density (HD)—depending on the values of $(\alpha, \beta)$ (see **Figure 1**B and (8) below). Using the asymptotics of the particle densities in these three phases (Section 4 of Supporting Information), we found that $D_0(\alpha, \beta, N)$ and $D_1(\alpha, \beta, N)$ in (4) and (5), respectively, are both asymptotically equivalent to $D_{\alpha,\beta}$, given by

$$D_{\alpha,\beta} = \begin{cases} \frac{1}{4}, & \text{if } \alpha > \frac{1}{2},\ \beta > \frac{1}{2} \text{ (MC regime)}, \\ \alpha(1-\alpha), & \text{if } \alpha < \frac{1}{2},\ \beta > \alpha \text{ (LD regime)}, \\ \beta(1-\beta), & \text{if } \beta < \frac{1}{2},\ \beta < \alpha \text{ (HD regime)}. \end{cases} \tag{8}$$

At steady state, $\langle \tau_i(1 - \tau_{i+1}) \rangle$ is the same for all $i = 1, \ldots, N-1$. This quantity is defined as the current (or flux) and is denoted by $J$. We note that $D_{\alpha,\beta}$ in (8) is in fact identical to the asymptotics of $J$ in the large-$N$ limit. Hence, it turns out that the asymptotics of $\langle \tau_i' \rangle$ for $2 \le i \le N-1$ are correctly given by using in (2) the mean-field approximation $\langle \tau_{i-1}\tau_i(1-\tau_{i+1}) \rangle \sim \langle \tau_{i-1} \rangle \langle \tau_i(1-\tau_{i+1}) \rangle = J \langle \tau_{i-1} \rangle$. Finally, noting that $\langle \tau_1' \rangle = \langle \tau_1(1-\tau_2) \rangle = J$ and $\beta \langle \tau_N \rangle = J$ at steady state, while $\langle \tau_{N-1} \rangle \sim J + (J/\beta)^2$ asymptotically, we obtain that $\langle \tau_i' \rangle$ is asymptotically given by

$$\langle \tau_i' \rangle \sim \begin{cases} D_{\alpha,\beta}, & \text{for } i = 1, \\ D_{\alpha,\beta}(1 - \langle \tau_{i-1} \rangle), & \text{for } 2 \le i \le N-1, \\ \frac{D_{\alpha,\beta}}{\beta}\left(1 - \frac{D_{\alpha,\beta}}{\beta}\right), & \text{for } i = N. \end{cases}$$

Using the asymptotics of $\langle \tau_i \rangle$ in different phases (Section 4 of Supporting Information), the resulting densities at the boundaries and far from the right boundary ($\langle \tau_{N-j} \rangle$, $1 \ll j \ll N$) can be computed, as summarized in **Table 1**. The asymptotics far from the left boundary ($\langle \tau_j \rangle$, $1 \ll j \ll N$) can be derived using the "particle-hole symmetry" [22]

$$\langle \tau_{N+1-i} \rangle_N(\alpha, \beta) = 1 - \langle \tau_i \rangle_N(\beta, \alpha). \tag{9}$$

The fraction of isolated particles $\frac{\langle \tau_i' \rangle}{\langle \tau_i \rangle}$ is given by

$$\frac{\langle \tau_i' \rangle}{\langle \tau_i \rangle} \sim \begin{cases} \dfrac{\alpha D_{\alpha,\beta}}{\alpha - D_{\alpha,\beta}}, & \text{for } i = 1, \\ \dfrac{D_{\alpha,\beta}(1 - \langle \tau_{i-1} \rangle)}{\langle \tau_i \rangle}, & \text{for } 2 \leq i \leq N-1, \\ 1 - \dfrac{D_{\alpha,\beta}}{\beta}, & \text{for } i = N. \end{cases}$$

As shown in **Figure 2**, there is good agreement between our asymptotic formulas and the exact results obtained from using the exact $\langle \tau_i \rangle_N$ [22] in equations (3)–(7). We observed some large boundary effects, as the density of isolated particles at the boundaries is always larger than in the bulk. In the LD I regime ($\beta < \frac{1}{2}$), slow termination creates queuing so that the density of isolated particles decreases close to the end, in contrast to the total density. In the HD regime, high density creates a lot of stacked particles so the proportion of isolated particles is very small. In the MC regime, stacked particles are present more in the beginning of the lattice. As a result, the density of isolated particles in the bulk increases along the lattice, in contrast to the total density.

**The $\ell$-TASEP with extended particles.** During translation, ribosomes move along mRNAs by decoding one codon at a time, but occupy an extended space of $\sim 10$ codons. For that reason, it is also of interest to generalize our theoretical results to a process where particles occupy a certain size $\ell \geq 1$ (this process is usually called the $\ell$-TASEP [26]). In this general case, using a matrix product to represent the steady-state solution leads to equations that are more complex, making the method employed above inapplicable (see Discussion). To cope with this complexity, we used a refined mean field approach introduced by Lakatos and Chou [4]. Although this approach cannot capture the variation of densities along the lattice as in the previous section, it well approximates the global average density and the current of particles. The key idea is to approximate the distribution of particles in the large-$N$ limit by an equilibrium ensemble in which particles get uniformly distributed. Using such approximation, we obtained (Section 5 of Supporting Information) that the density of isolated particles far from the boundaries, simply denoted $\langle \tau' \rangle$, is given by

$$\langle \tau' \rangle = \langle \tau \rangle \left[ \frac{1 - \ell \langle \tau \rangle}{1 - (\ell - 1)\langle \tau \rangle} \right]^2. \tag{10}$$

Using the asymptotic densities and currents found by Lakatos and Chou [4], we derived the asymptotics of $\langle \tau' \rangle$. As for the $\ell = 1$ case, the phase diagram can be decomposed into three parts (MC, HD, LD), separated by critical values $\alpha^* = \beta^* = \frac{1}{1+\sqrt{\ell}}$. For $\ell = 1$, we have $\alpha^* = \beta^* = \frac{1}{2}$, in agreement with the previous section. Combining (10) with the asymptotic density $\langle \tau \rangle$ in the large-$N$ limit (Section 6 of Supporting Information), we obtained the following density of isolated particles in the bulk:

$$\langle \tau' \rangle = \begin{cases} \frac{\sqrt{\ell}}{(1+\sqrt{\ell})^3}, & \text{if } \alpha > \alpha^*,\ \beta > \beta^* \text{ (MC regime)}, \\ \frac{1}{2} J \left[ (\ell - 1)J + 1 - \sqrt{((\ell-1)J+1)^2 - 4\ell J} \right], & \text{if } \alpha < \alpha^*,\ \beta > \alpha \text{ (LD regime)}, \\ \frac{1}{2} J \left[ (\ell - 1)J + 1 + \sqrt{((\ell-1)J+1)^2 - 4\ell J} \right], & \text{if } \beta < \beta^*,\ \beta < \alpha \text{ (HD regime)}, \end{cases} \tag{11}$$

where $J$ is the particle flux given by [4]

$$J \sim \begin{cases} \dfrac{1}{(1+\sqrt{\ell})^2}, & \text{if } \alpha > \alpha^*,\ \beta > \beta^* \text{ (MC regime)}, \\ \dfrac{\alpha(1-\alpha)}{1+(\ell-1)\alpha}, & \text{if } \alpha < \alpha^*,\ \beta > \alpha \text{ (LD regime)}, \\ \dfrac{\beta(1-\beta)}{1+(\ell-1)\beta}, & \text{if } \beta < \beta^*,\ \beta < \alpha \text{ (HD regime)}. \end{cases}$$

Near the entrance and exit, particles potentially get stacked on one side only. At the entrance, the density of isolated particles is, for $i < \ell$,

$$\langle \tau_i' \rangle = \mathbb{P}(t_i = 1\ ,\ t_{i+\ell} = 0) = J. \tag{12}$$

Hence, $\langle \tau_i' \rangle$ at the entrance is exactly given by the current flux $J$, as in the case of $\ell = 1$. Near the exit, for $i > N - \ell$, $\langle \tau_i' \rangle$ satisfies

$$\langle \tau_i' \rangle = \mathbb{P}(t_i = 1\ ,\ t_{i-\ell} = 0).$$

Using $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) + \mathbb{P}(A^c \cap B^c)$ and $\mathbb{P}(t_{i-\ell} = 1\ ,\ t_i = 0) = J$ yields

$$\langle \tau_i' \rangle = J + \mathbb{P}(t_i = 1) - \mathbb{P}(t_{i-\ell} = 1) = J + \langle \tau_i \rangle - \langle \tau_{i-\ell} \rangle.$$

As the flux satisfies $J = \beta \langle \tau_N \rangle = \langle \tau_{N-1} \rangle = \cdots = \langle \tau_{N-\ell+1} \rangle$, we obtained

$$\langle \tau_i' \rangle = \begin{cases} 2J - \langle \tau_{i-\ell} \rangle, & \text{for } N - \ell < i < N, \\ J\left(1 + \frac{1}{\beta}\right) - \langle \tau_{N-\ell} \rangle, & \text{for } i = N. \end{cases} \tag{13}$$

**Comparison with Monte Carlo simulations and estimation of interference rate.** Combining (11), (12) and (13) leads to approximate densities of isolated particles along the lattice in the $\ell$-TASEP. The isolated particle densities in the bulk (11) and near the entrance (12) depend only on the flux $J$, whereas near the exit the result (13) also depends on the density of particles located $\ell$ sites behind. In the LD regime, this density can be approximated by the density in the bulk (Section 6 of Supporting Information). However, in the other regimes, the density varies near the boundary, so using this approximation might be inaccurate (see **Figure S1**). As **Figure 3**A shows, however, our theoretical results agree well with the empirical densities of isolated particles obtained from Monte Carlo simulations, for specific values of $(\alpha, \beta)$ in the LD, HD and MC regimes. Contrary to the matrix method for the classical 1-TASEP model, the refined mean field approximation does not capture the variation of isolated particle densities across the lattice. However, this variation is much smaller than that of the total density, especially in regions of high traffic. Thus, assuming the density of isolated particles to be constant turns out to be a fine approximation for the majority of the lattice.

More generally, we studied in **Figure 3**B and **Figure S2** how the flux, density, and proportion of isolated particles vary as a function of $\alpha$, for fixed values of $\beta$. Overall, our theoretical results were in good agreement with Monte Carlo simulations. Interestingly, whereas the flux (**Figure S2**) and total density (**Figure 3**B) increase and reach a plateau after transitioning to the HD (when $\beta < \beta^*$) or the MC (when $\beta > \beta^*$) regime, the density of isolated particles follows a more complex

pattern: First, there is a drop in the density of isolated particles when transition occurs from LD to HD. In contrast, we observed an increase in the total density, showing that most particles contributing to the density are stacked. Second, as $\beta$ increases, the amplitude of the drop decreases until it becomes 0, when the MC regime replaces the HD regime. However, the maximum of $\langle \tau' \rangle$ is not reached in the MC regime but in the LD regime before phase transition occurs. In other words, as the initiation rate increases, the level of interference increases faster than the global density. This was confirmed when we plotted the ratio $\frac{\langle \tau' \rangle}{\langle \tau \rangle}$(**Figure 3**B, right panels), showing a linear decrease from $\alpha = 0$ to $\alpha = \beta$, while the total density gets sublinear as $\alpha$ gets closer to $\beta$. The first-order Taylor expansion in $\alpha$ of $\frac{\langle \tau' \rangle}{\langle \tau \rangle} = \left[\frac{1-\ell\langle \tau \rangle}{1-(\ell-1)\langle \tau \rangle}\right]^2$ in the LD regime gives

$$\frac{\langle \tau' \rangle}{\langle \tau \rangle} = 1 - 2\alpha + O\left(\alpha^2\right). \tag{14}$$

Interestingly, this formula does not depend on $\ell$ and using the formula obtained for the classical 1-TASEP model leads to the same result. To estimate the amount of interference associated with the dynamics of particles, we approximated the interference rate $I$, defined as the probability for a particle to get obstructed, as

$$I = \frac{1}{2}\left(1 - \frac{\langle \tau' \rangle}{\langle \tau \rangle}\right). \tag{15}$$

Using equation (14), we obtained that the interference rate is close to $\alpha$ in the LD regime.

**Generalization to larger isolation range.** In the next section, one of our goals will be to determine whether stacked particles are detected in ribosome profiling experimental protocols. A problem is that we do not know *a priori* what is the exact range between two ribosomes that may prevent them from being detected. For this reason, we considered the density associated with isolation range $d$, denoted $\langle \tau_i^{(d)} \rangle$, as

$$\langle \tau_i^{(d)} \rangle = \mathbb{P}(\tau_i = 1,\ x_i^- \leq i - \ell - d,\ x_i^+ \geq i + \ell + d),$$

where $x_i^-$ and $x_i^+$ are the positions of the closest particles located before and after site $i$, respectively. In other words, $\langle \tau_i^{(d)} \rangle$ gives the steady-state density of particles under the $\ell$-TASEP at position $i$ such that the distance to their closest neighbor is at least $d + \ell$. In particular $\langle \tau_i^{(0)} \rangle$ gives the total density of all particles, while $\langle \tau_i^{(1)} \rangle$ is equal to $\langle \tau_i' \rangle$, the density of isolated particles computed above. Following the same method as in the previous section, we obtained the following expression for particles in the bulk in the large-$N$ limit:

$$\langle \tau^{(d)} \rangle \sim \langle \tau \rangle \left[\frac{1 - \ell\langle \tau \rangle}{1 - (\ell - 1)\langle \tau \rangle}\right]^{2d}. \tag{16}$$

Hence, for two given isolation ranges $d$ and $d'$, the associated fractions of isolated particles satisfy

$$\left(\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d}} = \left(\frac{\langle \tau^{(d')} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d'}}.$$

Therefore, we can generalize (15) to obtain a formula for the interference rate for an arbitrary

isolation range $d \geq 1$:

$$I = \frac{1}{2}\left[1 - \left(\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d}}\right]. \tag{17}$$

## Application

We applied our theoretical results to analyze ribosome profiling data and mRNA translation. Briefly, the ribosome profiling procedure consists of using nuclease to digest translating ribosomes and get ribosome-protected mRNA fragments [1]. Fragments of length $\sim 10$ codons, corresponding to the size occupied by a ribosome, are extracted by the procedure and aligned to produce a positional distribution of ribosomes along the mRNA. However, it is possible that the nuclease may fail to cleave stacked ribosomes [8, 17–19]. Hence, the profile of ribosome counts along the mRNA produced by the experimental procedure might be different from the true profile (see **Figure 4**). Whether the nuclease can cleave two nearby ribosomes is still in debate, as the digestion and its efficiency vary depending on the organism and the protocols which are used [27, 28].

**Estimating the isolation range associated with non-detection of ribosomes.** To assess the extent of non-detection of stacked ribosomes in an actual ribosome profiling dataset, we used publicly available data of *S. cerevisiae* from Weinberg *et al.* [29] (more details in Materials and Methods). The experimental protocol used for these data minimizes some of the other biases known to affect the ribosome profiling, 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 [29]. For a given gene, a measure of the average density of detected ribosomes is given by the so-called translation efficiency ($TE$) [23]. More precisely, the $TE$ is given by the ratio of the RPKM measurement for ribosomal footprint to the RPKM measurement for mRNA, where RPKM denotes the number of Reads Per Kilobase of transcript per Million mapped reads. Hence, the $TE$ is proportional to the average density of detected ribosomes per site of a single mRNA; in our notation, $TE \propto \langle \tau^{(d)} \rangle$. To get the total density of ribosomes, we used another dataset from Arava *et al.* [30], obtained by polysome profiling, which is another technique giving, for a specific gene, the distribution of the number of ribosomes located on a single mRNA (and forming polysomes). While polysome profiling data is not biased by the possible omission of stacked ribosomes, the advantage of ribosome profiling is that it gives some local information about the ribosome occupancy.

When the density is low, ribosomes translating on the same mRNA are well isolated, so the $TE$ and the total ribosome density coincide. In order to get the scaling factor associated with the $TE$, we thus performed a linear fit between the two datasets in the region where the total density is less than 1 ribosome per 100 codons (**Figure S3**). Depending on the gap between two ribosomes that prevents them from being detected, the relation between the $TE$ and the total average density $D = \langle \tau \rangle$ is, according to equation (16),

$$TE = aD\left(\frac{1 - 10D}{1 - 9D}\right)^{2d}, \tag{18}$$

where $a$ is the rescaling factor (specifically, $TE = a \cdot \langle \tau^{(d)} \rangle$) obtained from fitting and $d$ denotes the detection gap-threshold mentioned previously (if the gap between a ribosome and its closest neighbor is larger or equal to $d$, then it gets detected). Since a ribosome occupies 10 codons, the parameter $\ell$ in (16) is set to 10. In **Figure 5**A, we plotted (18) for different values of $d$ and

compared it with the experimental data from Weinberg *et al.* and Arava *et al.* Our goal was then to determine which value of $d$ leads to the best match with the experimental data. In **Figure 5**B, we plotted the root mean square error between (18) and the experimental data, as a function of $d$ and for the value of $a$ corresponding to the fit obtained in **Figure S3**. We found that the minimum error is obtained when $d$ is between 4 and 6. On the other hand, as $d$ increases, the maximum value of $TE$ that can be obtained using (18) decreases (**Figure 5**A), potentially leading to some detected densities from experiment to be greater than the theoretical maximum of $TE$; we call such detected densities "anomalous" (as we shall see below, we can obtain a more refined estimate of the maximum possible detected density using an estimate of the termination rate $\beta$ for each gene). In **Figure 5**C, we plotted for each $d$ the fraction of genes with anomalous detected densities. For $d \leq 3$, no anomalous detected density was found, while the fraction becomes positive for $d \geq 4$ (less than 1% for $d = 4$, $\sim 2.5\%$ for $d = 6$, and $\sim 8\%$ for $d = 8$). We concluded that the best values of $d$ that both minimize the error and the fraction of anomalous detected density were obtained for $d = 3$ or 4. In agreement with our estimate, previous ribosome profiling experiments found disome fragments (accounting for the mapping of two ribosomes) of length $\sim 65$ nucleotides [19], suggesting that $d = 3$ (2 times 30 nucleotides plus 2 other codons).

**Near optimality of termination rates.** We then studied how termination rates were distributed in the dataset we analyzed. We restricted to those genes with length $\geq 200$ codons, and then used a method developed previously [16] to estimate, for a given gene of length $N$, the termination rate $r_N$ and position-specific elongation rates $r_i$ along the transcripts, where $i \in \{1, \ldots, N-1\}$. We then estimated the scaled termination rate as

$$\beta = \frac{(N-1)r_N}{\sum_{i=1}^{N-1} r_i}. \tag{19}$$

Applying this to each of our 3712 filtered genes, we obtained a distribution of termination rates shown in **Figure 6**A. Interestingly, we found that the average termination rate was slightly larger than $\beta^* = \frac{1}{1+\sqrt{\ell}}$ for $\ell = 10$. (Recall that $\beta^*$ is the critical termination rate at which phase transition to the MC regime occurs.) This suggests that although the termination is known to be slow in order to complete the different steps associated with the completion of protein synthesis [31, 32], on average, it does not limit the protein synthesis rate even if the initiation rate increases, as it allows to achieve the maximal current.

We further studied the possible determinants of termination rate variation and studied the influence of the stop codon (TAA, TGA, or TAG) on $\beta$. As shown in **Figure 6**B, we found that, when looking at increasing termination rates, the fraction of genes with the TAA stop codon increases (from $\sim 35\%$ for the 500 genes with the lowest $\beta$ to $\sim 60\%$ for the 500 genes with the highest $\beta$), while the opposite applies to TGA (with a slighter decrease from $\sim 40\%$ to $\sim 25\%$) and TAG (from $\sim 25\%$ to $\sim 15\%$). As a consequence, the distribution of $\beta$ for genes with the TAA stop codon has a larger tail (**Figure 6**C).

**Identifiability of initiation rates and flux from TE measurements.** Under the $\ell$-TASEP model in the LD regime, the $TE$ is related (**Figure 7**A) to the initiation rate $\alpha$ through equation (16) and the asymptotics of $\langle \tau \rangle$ and $J$ (Section 6 of Supporting Information). Assuming that translation occurs in the LD regime (since translation is generally limited by initiation under realistic physiological conditions [33, 34]), we studied whether we could infer the gene-specific initiation rate $\alpha$ using our theoretical results. Illustrated in **Figure 7**B is a histogram of detected densities for our dataset, which shows that the detected density is bounded by $\sim 0.02$ ribosomes per codon. From the plotted curves in **Figure 7**A, this suggests that for $d \leq 5$ and for all the experimental

detected densities, there exists a value for the initiation rate satisfying (16). However, for $d \geq 3$, the identifiability of $\alpha$ (i.e., the uniqueness of $\alpha$) does not seem to be guaranteed.

More precisely, for a given gene and isolation range $d$, the theoretical maximal value of the $TE$, denoted $\langle \tau^{(d)} \rangle_{\max}$, is determined by the termination rate $\beta$, as

$$\langle \tau^{(d)} \rangle_{\max}(\beta) = \sup(\langle \tau^{(d)} \rangle(\alpha),\ \alpha \in [0, \beta]). \tag{20}$$

In **Figure 7**C, we computed for different values of $d$ the fraction of genes satisfying $TE' \leq \langle \tau^{(d)} \rangle_{\max}$, where $TE'$ is the $TE$ normalized by the scaling factor $a$ (see (18)). We found that all the genes satisfied this condition for $d \leq 5$, before observing a small decrease for $d = 6$ (98%).

We further looked at the fraction of genes for which we can identify a unique initiation rate that matches the associated detected density with the measured $TE$. As $\alpha$ increases to its critical value $\min(\beta, \beta^*)$ (leading to a transition from LD to the other regimes), the density of isolated particles either only increases, or increases then decreases, to $\langle \tau^{(d)} \rangle_{\mathrm{id}}(\beta)$, given by

$$\langle \tau^{(d)} \rangle_{\mathrm{id}}(\beta) = \begin{cases} \langle \tau^{(d)} \rangle(\beta), & \text{if } \beta \leq \beta^*, \\ \langle \tau^{(d)} \rangle_{\mathrm{MC}}, & \text{otherwise,} \end{cases} \tag{21}$$

where $\langle \tau^{(d)} \rangle_{\mathrm{MC}}$ is the density of isolated particles in the MC regime. As a consequence, there is only one identifiable initiation rate in the LD region when $TE' < \langle \tau^{(d)} \rangle_{\mathrm{id}}(\beta)$, and two when $\langle \tau^{(d)} \rangle_{\mathrm{id}}(\beta) \leq TE' \leq \langle \tau^{(d)} \rangle_{\max}$. In **Figure 7**C, we computed the fraction of genes satisfying $TE \leq \langle \tau^{(d)} \rangle_{\mathrm{id}}$. We found that all genes were then strictly identifiable for $d \leq 2$, before the fraction starts to decrease for $d = 3$ (96%). For $d \geq 4$, a significant fraction of genes (at least 19%) is not strictly identifiable. Thus, in the range of $d$ associated with non-detection found from **Figure 5**, the TE measurement may lead to some ambiguity in the initiation rates. In this case, two values of the initiation rate $\alpha_1 < \alpha_2$ lead to the same detected density: Although the total density for $\alpha_2$ is larger than for $\alpha_1$, there are also more closely stacked ribosomes that are not detected. Hence, the density of isolated particles is the same for both. As the flux is an increasing function of the initiation rate (**Figure S2**), such ambiguity also applies for inferring the flux. Under the $\ell$-TASEP model, a way to eliminate this ambiguity is to compare the variations of local densities. For example, we have previously shown that the density at the entrance is exactly the flux (see (12)). Thus, two profiles associated with different initiation rates in the LD regime also differ at the entrance.

**The fraction of detected ribosomes and interference rates.** Upon estimating the threshold of gap distance between consecutive ribosomes leading to their non-detection and studying the identifiability of the initiation rate $\alpha$, we then quantified the resulting fraction of detected ribosomes and the associated interference rate. As discussed above, for some values of $d$ and $\langle \tau^{(d)} \rangle$, there may be two distinct values of $\alpha$, and hence two distinct values of the total average density $\langle \tau \rangle$, corresponding to the same $\langle \tau^{(d)} \rangle$. This implies that the fraction $\langle \tau^{(d)} \rangle / \langle \tau \rangle$ of detected ribosomes and the the interference rate (17) may not be uniquely determined for some values of $d$ and $\langle \tau^{(d)} \rangle$. Indeed, for some of the experimentally observed $TE$ values from Weinberg *et al.* [29], we encountered ambiguity in estimating $\alpha$ when $d \geq 3$ (see **Figure 7**C). Thus, when such ambiguity occurred, we considered both lower and upper estimates of $\alpha$, and found their respective resulting fractions of detected ribosomes $\langle \tau^{(d)} \rangle / \langle \tau \rangle$ and interference rates (**Figure 8**). We obtained that for $d = 3$ or 4, suggested by **Figure 5**B and C, the lower estimates of $\alpha$ lead to fractions of detected ribosomes lying between $91.2 \pm 5\%$ and $93.5 \pm 3.5\%$ (**Figure 8**A). The upper estimates of $\alpha$ lead to smaller mean and larger variability (between $80 \pm 26\%$ and $91.6 \pm 11.7\%$). As expected, we observed no substantial difference between the lower and upper estimates for $d = 1$ or 2 (since no gene presents

any ambiguity). As $d$ increases, however, the fraction of detected ribosomes decreases (notably because of the increasing fraction of genes with ambiguity). Interestingly, in contrast to these important variations, we observed that the interference rates corresponding to the lower estimates of $\alpha$ remain stable around 1% for all $d$, with only a slight increase of standard deviation from 0.5 to 0.9% (**Figure 8**B). Somewhat larger variation is observed for the interference rates corresponding to the upper estimates of $\alpha$, with ranges $1.5 \pm 2.7\%$ and $3.6 \pm 5.5\%$ for $d = 3$ and 4, respectively.

This difference in the amplitude between the fraction of detected ribosomes and interference rate can be explained theoretically, as illustrated in **Figure S4**. When plotting the fraction $\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}$ of detected ribosomes as a function of $\langle \tau^{(d)} \rangle$ (**Figure S4**A), we observed that for large values of the fraction (associated with low $\alpha$), the curves for different values of $d$ were well separated, such that for $\langle \tau^{(d)} \rangle \sim 0.01$ (corresponding to the range of our dataset), the fraction of detected ribosomes can vary between 98% (for $d = 1$) and 85% (for $d = 6$). In contrast, the interference rate takes approximately the same value for all $d$ ($\sim 1\%$, see **Figure S4**B). More generally, the formula (17) for interference rate shows that, as $d$ increases, any observed decrease in the fraction $\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}$ is compensated by the power $\frac{1}{d}$. Furthermore, as $d$ increases, the range of the ratio $\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}$ also increases (from $60 \sim 100\%$ for $d = 0$ to $3 \sim 100\%$ for $d = 6$), leading to larger differences between the lower and upper estimates, and higher variability across genes. In contrast, the interference rate remains bounded (by $\sim 0.2$), explaining its smaller variation across our dataset and different values of $d$.

## Discussion

In this article, we provided a complete analysis of the distribution of isolated particles in the TASEP model with open boundaries. In the classical form of the model, we obtained exact analytic solutions using the matrix formulation originally developed by Derrida *et al.* [22]. We also obtained accurate asymptotic formulas in the limit of large $N$ for different regimes of the phase diagram. In the past, the classical 1-TASEP has been studied in various geometric settings [20, 21], such as rings [35, 36] and networks [37, 38], or with more complex dynamics associated with pausing [39, 40], random rates [35, 41, 42] or multiple species [22, 36, 42–44], to name a few. A possible extension of our work would be to investigate the behavior of isolated particles in these different contexts. In many cases, the solution of the associated master equation can be found using a matrix formulation [20,21,42,44], suggesting that the work presented here could be generalized.

We further studied the $\ell$-TASEP model with extended particles of size $\ell$ and derived asymptotic formulas for densities using a refined mean field approach. In this more general case, the steady-state solution of the associated master equation can, in principle, also be written in the form of a generic matrix product [20, 45]. In practice, however, the associated algebra is rather complex, making it challenging to derive analytic results [4, 20]. To cope with this complexity, several approaches using mean field approximation have been developed [4, 6, 33]. Unlike simpler mean field approaches, the refined analytic approximation proposed by Lakatos and Chou [4] leads to simple formulas that agree extremely well with Monte Carlo simulations. In our work, we employed a similar approach to obtain a simple, accurate formula for the density of isolated particles with a given minimum distance to the closest neighbor.

We applied our theoretical results to study mRNA translation using ribosome profiling data. In particular, our analysis suggests that the representation of the ribosome density may be biased by the non-detection of ribosomes with gap distance less than $\sim 3$ codons. In general, different protocols applied to different organisms can affect the nuclease action and in particular its ability to cleave ribosomes [28]. Hence, it would be interesting to apply our method to other datasets

and other organisms to find possible differences in the detection gap distance. In particular, such differences could be visible near the terminal end of the transcript sequence, where slow termination can cause interference [31,32]. In yeast (which is the organism studied in our dataset), no periodic peaks of density were detected in this region across multiple datasets [19, 46–53], suggesting non-detection of stacked ribosomes. In contrast, such peaks have been detected for other organisms and different protocols [54, 55].

As mentioned previously, termination has been shown to be significantly slower than elongation along the transcript. We confirmed this result by estimating the termination rate from ribosome profiling data. More precisely, we found that the average termination rate is only slightly larger than the minimum value $\beta^*$ that ensures, for large enough initiation rate, being in the maximum current regime. In other words, our study suggests that, on average, termination is as slow as possible, while not being a possible limiting factor of translation. Related to this perspective, it would be interesting to investigate in more detail the determinants of the variability of the termination rate. In our study, we notably found that genes with faster termination rates are more likely to have the TAA stop codon. A possible explanation for such a preference may be codon readthrough (which happens when the ribosome moves through the stop codon). Interestingly, it has been found [56] that in yeast a second stop codon is significantly more likely to be present after the TAA stop codon, forming a tandem that reduces the chance of readthrough. This suggests that TAA is less effective in stopping the ribosome, and our observation of faster termination at TAA could result from a relatively higher fraction of ribosomes (compared to TGA and TAG) continuing to move along the mRNA, instead of being stopped and waiting for release at TAA.

Other methods have also been developed previously to infer the initiation rates associated with specific genes from polysome [11] or ribosome profiling [14]. These approaches used Monte Carlo simulations that can be computationally expensive. Using our theoretical results, it is possible to infer the initiation rate directly from the observed average detected density. Interestingly, we found that for our typical detection gap distance, some initiation rates were not uniquely identifiable (i.e., two initiation rates can lead to the same observed $TE$ arising from isolated ribosomes), as having a higher initiation rate also creates higher interference that decreases the detected density. As a result, our work suggests that, for some genes, there could be ambiguity in identifying the initiation rate and the flux from $TE$, although this measurement has been widely used as a proxy for protein production [23].

We also provided robust estimates of the average rate of interference that ribosomes experience during translation. These estimates implicitly depend on the initiation rate and homogeneous elongation rate, but do not include other possible sources of interference due to local heterogeneities. While it has been shown that the average elongation speed along the transcript sequence is approximately constant around 5.6 codon/s [1], there is evidence of variation of the elongation rate along the transcript, especially in the first $\sim 200$ codons, leading to the so-called "5' translational ramp" [23] (in another study [16], we quantified the extent of the interference created by this ramp).

Overall, our work shows how studying the interaction range of particles in exclusion process can help to get a better understanding of the process, and that it can be applied to problems where the data available are biased against this range. Similarly, while we focused here on isolated particles, our methods can be applied to situations where only aggregated particles following a transport process get detected.

# Materials and Methods

**Experimental dataset.** The flash-freeze ribosome profiling data from Weinberg *et al.* [29] can be accessed from the Gene Expression Omnibus (GEO) database with the accession number GSE75897. 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 termination rates.** For a given profile $(P_1, \ldots, P_N)$ containing the number of footprints with A-site detected at each position, we estimate the associated elongation rates $(r_1, \ldots r_n)$ as

$$r_i = \begin{cases} \min\left(r_{\max}, \dfrac{P_{\max}}{P_i}\right), & \text{if } P_i \neq 0, \\ r_{\max}, & \text{else,} \end{cases}$$

where $P_{\max} = \max_i (P_i)$ and $r_{\max}$ is a fixed threshold value (in practice, we set it to 40). Such approximation is valid when there is little ribosomal interference [16]. In another study [16], we developed a more refined inference procedure that uses these rates as first estimates (this method applies for genes with high footprint coverage), leading to excellent agreement between the observed and simulated profiles for the same dataset used here. As in average, our refined procedure lead to correction for $\sim 1.57$ site per gene, these "naive" estimates are valid over a large majority of the sites.

**ACKNOWLEDGMENTS.** This research is supported in part by a Math+X Research Grant from the Simons Foundation and a Packard Fellowship for Science and Engineering. YSS is a Chan Zuckerberg Biohub investigator.

# References

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

[2] Brar GA, Weissman JS (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology* 16:651–664.

[3] MacDonald CT, Gibbs JH, Pipkin AC (1968) Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* 6(1):1–25.

[4] Lakatos G, Chou T (2003) Totally asymmetric exclusion processes with particles of arbitrary size. *Journal of Physics A: Mathematical and General* 36(8):2027.

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

[6] Zia RK, Dong J, Schmittmann B (2011) Modeling translation in protein synthesis with tasep: A tutorial and recent developments. *Journal of Statistical Physics* 144(2):405–428.

[7] Chowdhury D, Schadschneider A, Nishinari K (2005) Physics of transport and traffic phenomena in biology: from molecular motors and cells to organisms. *Physics of Life reviews* 2(4):318–352.

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

[9] Sharma AK, Chowdhury D (2011) Stochastic theory of protein synthesis and polysome: Ribosome profile on a single mrna transcript. *Journal of Theoretical Biology* 289:36–46.

[10] Chou T, Lakatos G (2004) Clustered bottlenecks in mrna translation and protein synthesis. *Physical Review Letters* 93(19):198101.

[11] Ciandrini L, Stansfield I, Romano MC (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.

[12] Basu A, Chowdhury D (2007) Traffic of interacting ribosomes: effects of single-machine mechanochemistry on protein synthesis. *Physical Review E* 75(2):021902.

[13] von der Haar T (2012) Mathematical and computational modelling of ribosomal movement and protein synthesis: an overview. *Computational and Structural Biotechnology Journal* 1(1):1–7.

[14] Gritsenko AA, Hulsman M, Reinders MJ, de Ridder D (2015) Unbiased quantitative models of protein translation derived from ribosome profiling data. *PLoS Comput Biol* 11(8):e1004336.

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

[16] Dao Duc K, Song YS (2016) Identification and quantitative analysis of the major determinants of translation elongation rate variation. *bioRxiv* p. 090837.

[17] Andreev DE, et al. (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Research* 45(2):513–526.

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

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

[20] Blythe RA, Evans MR (2007) Nonequilibrium steady states of matrix-product form: a solver's guide. *Journal of Physics A: Mathematical and Theoretical* 40(46):R333.

[21] Schadschneider A, Chowdhury D, Nishinari K (2010) *Stochastic transport in complex systems: from molecules to vehicles*. (Elsevier).

[22] Derrida B, Evans MR, Hakim V, 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.

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

[24] Spitzer F (1970) Interaction of markov processes. *Advances in Mathematics* 5(2):246–290.

[25] Privman V (2005) *Nonequilibrium statistical mechanics in one dimension.* (Cambridge University Press).

[26] Sasamoto T, Wadati M (1998) Exact results for one-dimensional totally asymmetric diffusion models. *Journal of Physics A: Mathematical and General* 31(28):6057.

[27] O'connor PB, Andreev DE, Baranov PV (2016) Comparative survey of the relative impact of mrna features on local ribosome profiling read density. *Nature Communications* 7.

[28] Gerashchenko MV, Gladyshev VN (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Research* 45(2):e6.

[29] Weinberg DE, et al. (2016) Improved ribosome-footprint and mrna measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports* 14(7):1787–1799.

[30] Arava Y, et al. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* 100(7):3889–3894.

[31] Yu X, Willmann MR, Anderson SJ, Gregory BD (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.

[32] Pelechano V, Wei W, Steinmetz LM (2015) Widespread co-translational rna decay reveals ribosome dynamics. *Cell* 161(6):1400–1412.

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

[34] Chu D, et al. (2014) Translation elongation can control translation initiation on eukaryotic mrnas. *The EMBO Journal* 33(1):21–34.

[35] de Queiroz S, Stinchcombe R (2008) Nonequilibrium processes: Driven lattice gases, interface dynamics, and quenched-disorder effects on density profiles and currents. *Physical Review E* 78(3):031106.

[36] Ayyer A, Linusson S (2014) An inhomogeneous multispecies tasep on a ring. *Advances in Applied Mathematics* 57:21–43.

[37] Neri I, Kern N, Parmeggiani A (2011) Totally asymmetric simple exclusion process on networks. *Physical Review Letters* 107(6):068702.

[38] Bittihn S, Schadschneider A (2016) Braess paradox in a network of totally asymmetric exclusion processes. *Physical Review E* 94(6):062312.

[39] Dong J, Schmittmann B, Zia RK (2007) Inhomogeneous exclusion processes with extended objects: The effect of defect locations. *Physical Review E* 76(5):051113.

[40] Sahoo M, Klumpp S (2016) Asymmetric exclusion process with a dynamic roadblock and open boundaries. *Journal of Physics A: Mathematical and Theoretical* 49(31):315001.

[41] Nossan JS (2013) Disordered exclusion process revisited: some exact results in the low-current regime. *Journal of Physics A: Mathematical and Theoretical* 46(31):315001.

[42] Arita C, Mallick K (2013) Matrix product solution of an inhomogeneous multi-species tasep. *Journal of Physics A: Mathematical and Theoretical* 46(8):085002.

[43] Prolhac S, Evans MR, Mallick K (2009) The matrix product solution of the multispecies partially asymmetric exclusion process. *Journal of Physics A: Mathematical and Theoretical* 42(16):165004.

[44] Evans MR, Ferrari PA, Mallick K (2009) Matrix representation of the stationary measure for the multispecies tasep. *Journal of Statistical Physics* 135(2):217–239.

[45] Klauck K, Schadschneider A (1999) On the ubiquity of matrix-product states in one-dimensional stochastic processes with boundary interactions. *Physica A: Statistical Mechanics and its Applications* 271(1):102–117.

[46] Gardin J, et al. (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* 3:e03735.

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

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

[49] Lareau LF, Hite DH, Hogan GJ, Brown PO (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments. *eLife* 3:e01257.

[50] Pop C, et al. (2014) Causal signals between codon bias, mrna structure, and the efficiency of translation and elongation. *Molecular Systems Biology* 10(12):770.

[51] Nedialkova DD, Leidel SA (2015) Optimization of codon translation rates via trna modifications maintains proteome integrity. *Cell* 161(7):1606–1618.

[52] Jan CH, Williams CC, Weissman JS (2014) Principles of er cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* 346(6210):1257521.

[53] Carja O, Xing T, Plotkin JB, Shah P (2017) riboviz: analysis and visualization of ribosome profiling datasets. *bioRxiv* p. 100032.

[54] Andreev DE, et al. (2015) Oxygen and glucose deprivation induces widespread alterations in mrna translation within 20 minutes. *Genome Biology* 16(1):90.

[55] Lobanov AV, et al. (2017) Position-dependent termination and widespread obligatory frameshifting in euplotes translation. *Nature Structural & Molecular Biology* 24(1):61–68.

[56] Liang H, Cavalcanti AR, Landweber LF (2005) Conservation of tandem stop codons in yeasts. *Genome Biology* 6(4):R31.

**Table 1**: **Asymptotics of $\langle\tau'\rangle$ in the different phases of the classical 1-TASEP.** These are obtained by combining equations (3), (6) and (7) with asymptotics given in Section 4 of Supporting Information. The asymptotics far from the left boundary ($\langle\tau_j\rangle$, $1 \ll j \ll N$) can be derived using the "particle-hole symmetry" (9).

| | $\langle\tau'_1\rangle$ (Eq.(6)) | $\langle\tau'_{N-j}\rangle$ $(1 \ll j \ll N)$ (Eq.(3)) | $\langle\tau'_N\rangle$ (Eq.(7)) |
|---|---|---|---|
| MC | $\frac{1}{4}$ | $\frac{1}{8}\left[1+\frac{1}{\sqrt{\pi(j+1)}}\right]$ | $\frac{1}{4\beta}\left(1-\frac{1}{4\beta}\right)$ |
| LD I ($\beta<\frac{1}{2}$) | $\alpha(1-\alpha)$ | $\alpha(1-\alpha)^2\left[1+\frac{2\beta-1}{1-\alpha}\left(\frac{\alpha(1-\alpha)}{\beta(1-\beta)}\right)^{j+2}\right]$ | $\frac{\alpha(1-\alpha)}{\beta}\left[1-\frac{\alpha(1-\alpha)}{\beta}\right]$ |
| LD II ($\beta>\frac{1}{2}$) | $\alpha(1-\alpha)$ | $\alpha(1-\alpha)^2\left[1+\frac{\left[\frac{1}{(2\alpha-1)^2}-\frac{1}{(2\beta-1)^2}\right]\alpha(4\alpha(1-\alpha))^{j+1}}{\sqrt{\pi}(j+1)^{3/2}}\right]$ | $\frac{\alpha(1-\alpha)}{\beta}\left(1-\frac{\alpha(1-\alpha)}{\beta}\right)$ |
| HD | $\beta(1-\beta)$ | $\beta^2(1-\beta)$ | $\beta(1-\beta)$ |

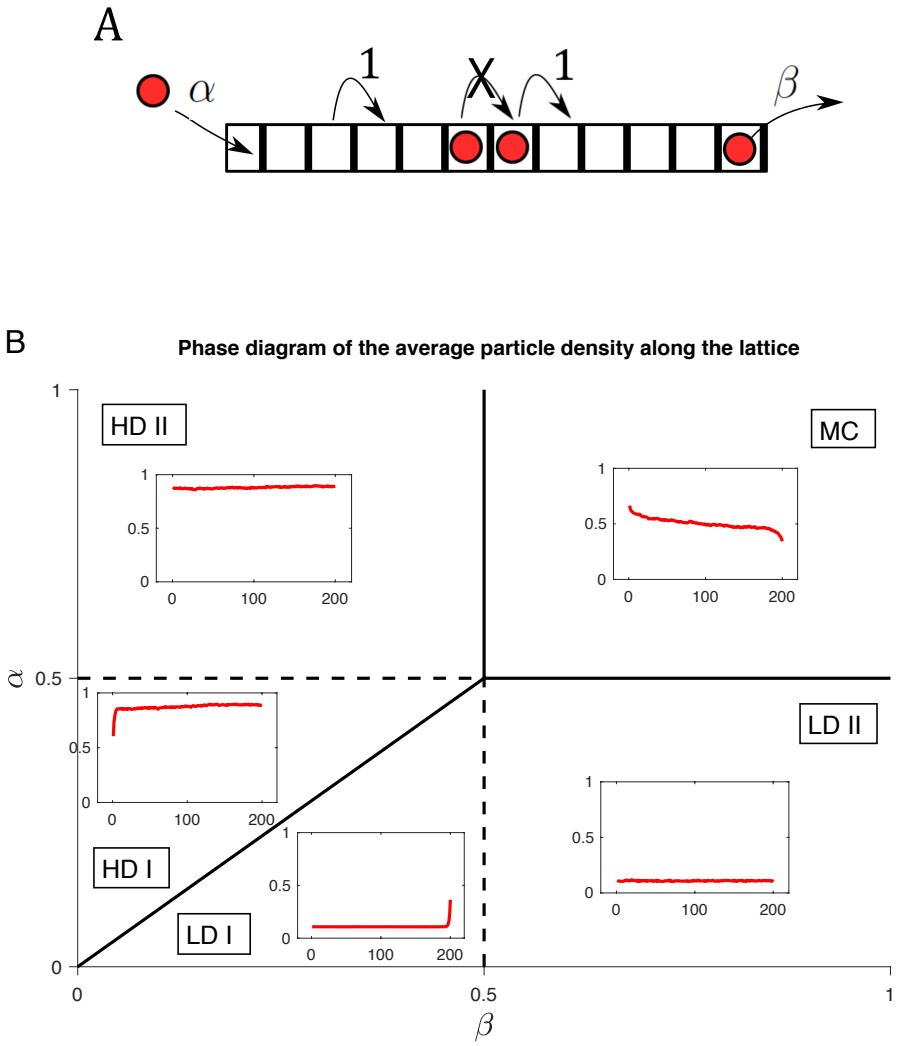


**Figure 1**: **Illustration of the TASEP with open boundaries. A:** A schematic representation of the TASEP model. Particles are introduced at the start of the lattice with exponential rate $\alpha$ and move along with exponential rate 1, provided that there is no particle occupying the next site. At the end of the lattice, they exit with exponential rate $\beta$. **B:** Phase diagram of the average particle density along the lattice. The profile of average density of particles along the lattice can be classified according to a phase diagram in $(\alpha, \beta)$-space, separating different regions: the maximal current regime (MC), the low density regime (LD), and the high density regime (HD). The LD and HD regions can also be decomposed into two separate ones: LD I/II, and HD I/II, respectively.

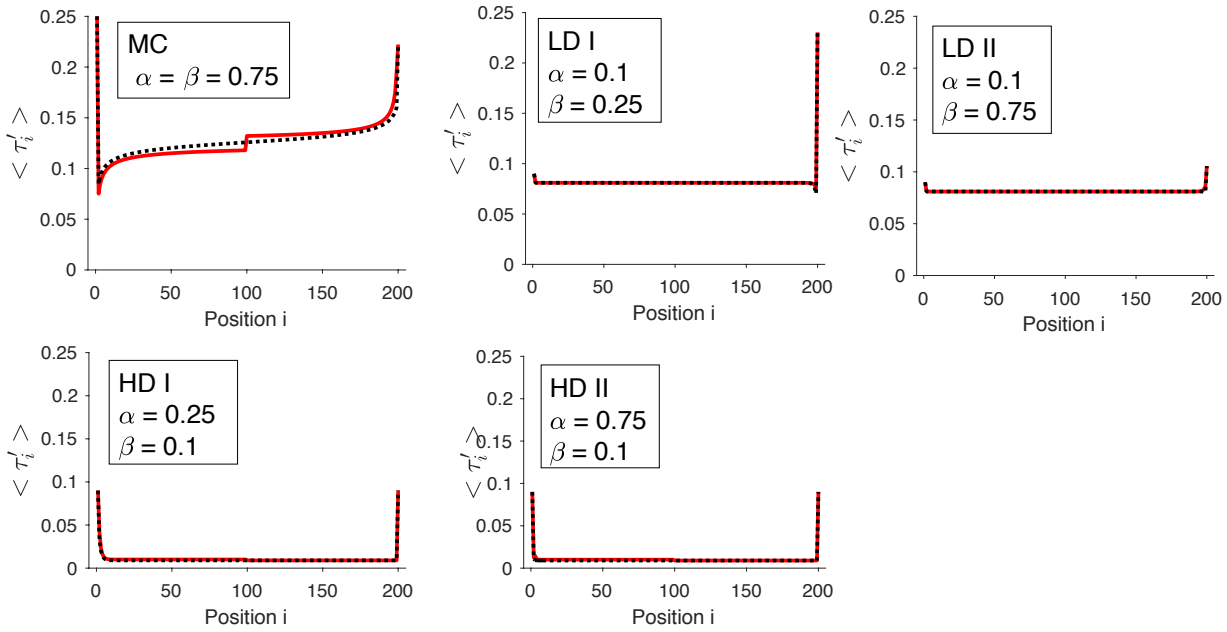


**Figure 2**: **The density of isolated particles in different regions of the TASEP phase diagram.** For the different regimes of the TASEP (**Figure 1**), the asymptotic formulas from **Table 1** (solid red lines) are compared with the exact densities (dotted black lines) of isolated particles given by (3)–(7).

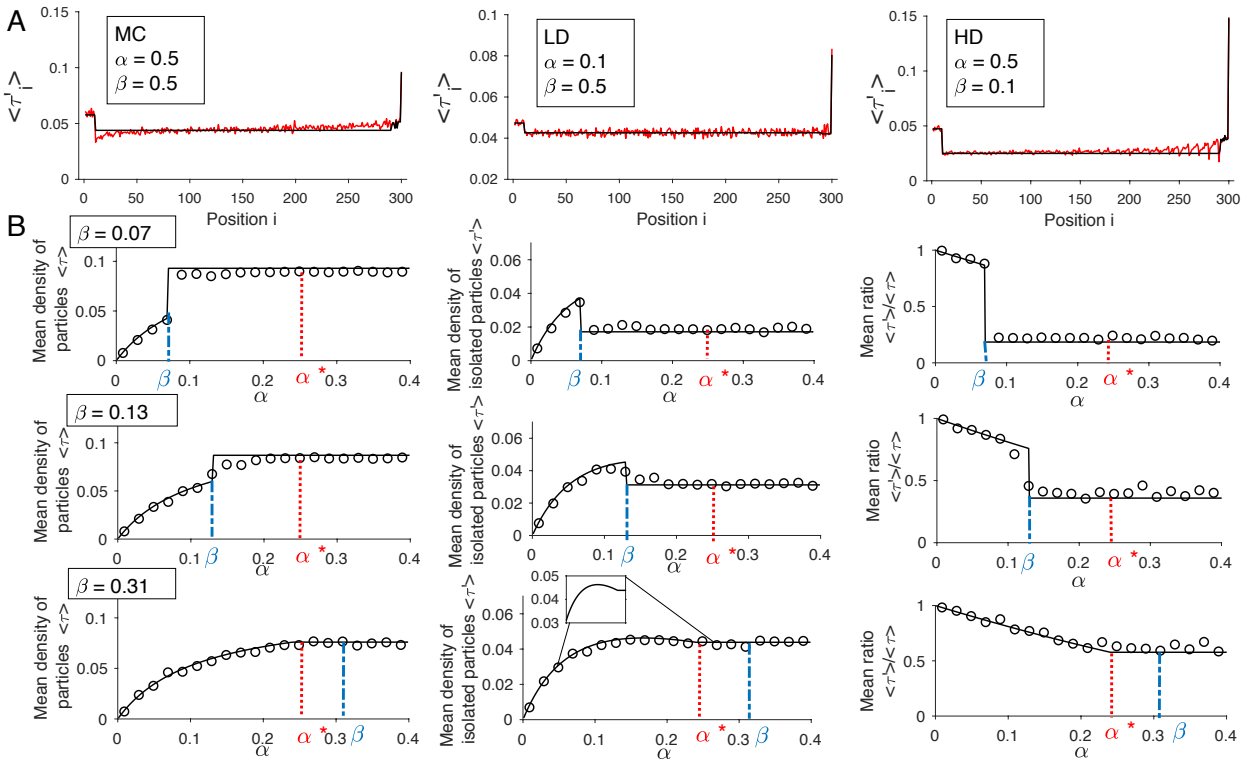


**Figure 3**: **Comparison of the results from the refined mean field approach with Monte Carlo simulations. A:** We simulated the TASEP with extended particles (size $\ell = 10$, sample size $= 10^9$) and plotted (in red) the densities of isolate particles in the three different regimes of the phase diagram. We compared these simulation results with the asymptotics obtain from (10), (12) and (13) (in black). **B:** For fixed values of $\beta$, these plots show how the total density, the density of isolated particles, and their ratio vary as a function of $\alpha$. The results obtained using Monte Carlo simulations (open circles) of the TASEP with extended particles (size of particles $\ell = 10$, sample size of isolated particles $10^4$, lattice size $= 400$) are compared with the results obtained from the refined mean field approach (solid lines). Note that there are discontinuities when transitioning from LD to HD regime (first and second rows).

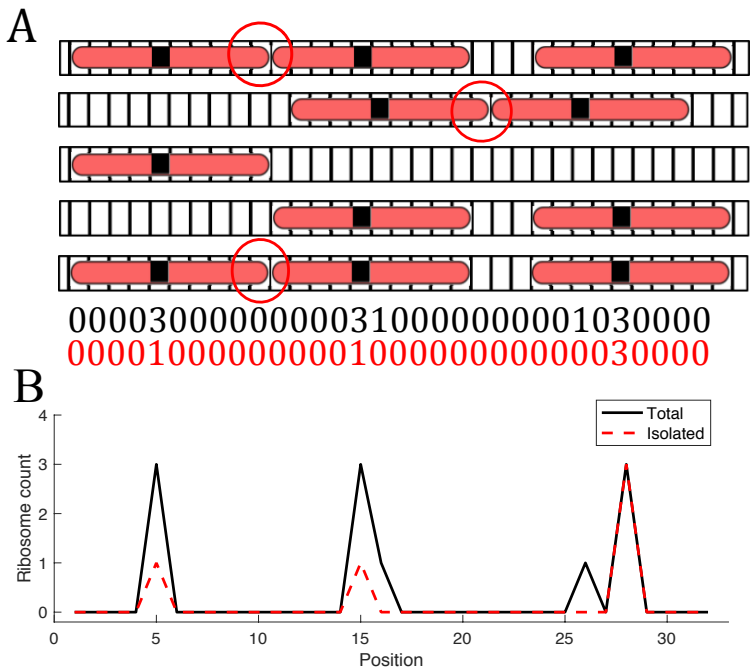


**Figure 4**: **A schematic representation of ribosome profiling. A:** Positions of ribosomes along the mRNA are obtained by nuclease digestion and allow to count the number of ribosomes found at a specific position. However, it is possible that the nuclease cannot cleave stacked ribosomes [8,16–19]. **B:** As a result, the profile of ribosome count along the mRNA recorded from isolated ribosomes (plotted in red) might be different from the true profile (plotted in black).

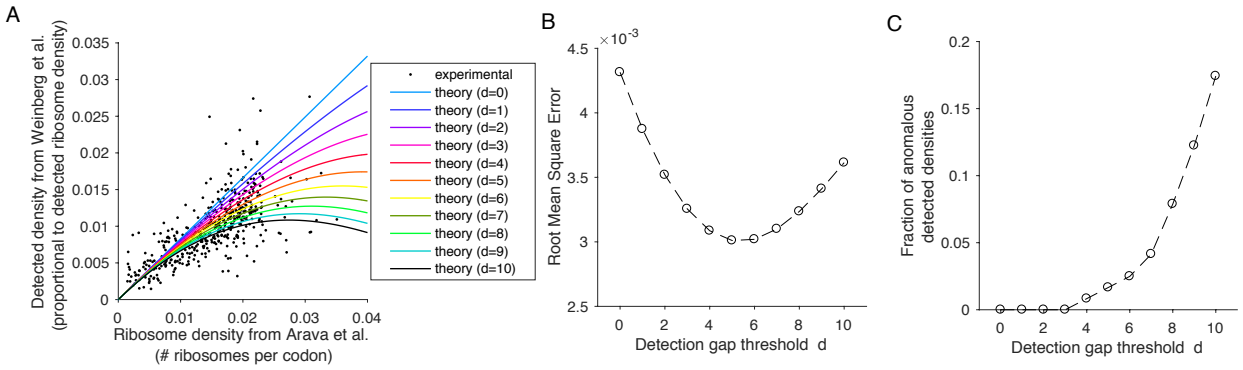


**Figure 5**: **Estimation of undetected ribosomes from ribosome profiling experiment. A:** This plot shows experimental ribosome profiling data of *S. cerevisiae* from Weinberg *et al.* [29] against the total ribosome density obtained from polysome profiling by Arava *et al.* [30] (482 genes). Also shown are plots of $y = ax\left(\frac{1-10x}{1-9x}\right)^{2d}$, obtained from computing the density of detected particles of size $\ell = 10$ as a function of the total density in the $\ell$-TASEP (see (18)) with various isolation range $d = 0, \ldots, 10$. We set $a = 0.82$ (see **Figure S3**). **B:** For values of $d \in \{0, \ldots, 10\}$, we plot the root mean square error obtained from comparing experimental data to the theoretical plots in **A**. **C:** For $d \in \{0, \ldots, 10\}$, we plot the corresponding fraction of genes with anomalous detected densities, where a detected density said to be anomalous if it is larger than the theoretical maximum value implied by (18), used in **A**.

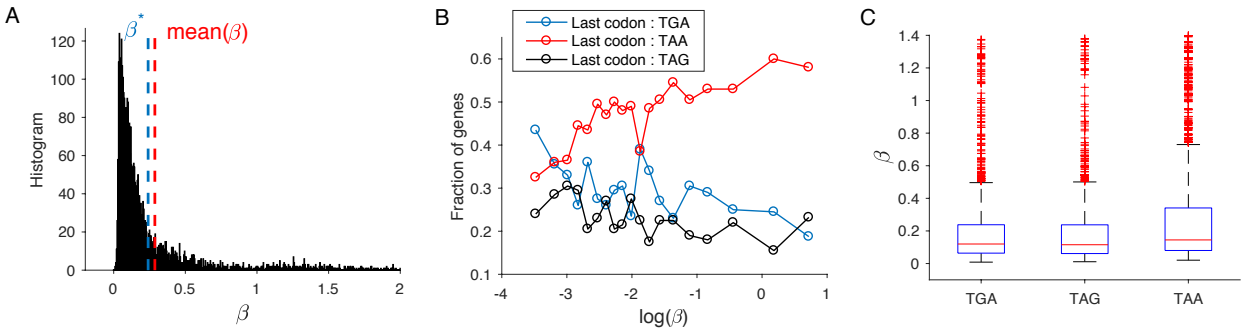


**Figure 6**: **Analysis of termination rates from ribosome profiling data. A:** A histogram of the termination rates $\beta$ for 3712 genes (longer than 200 codons) estimated (see (19)) from ribosome profiling data of Weinberg *et al.* [29]. Note that the average value (0.29) of $\beta$ is close to $\beta^* = 0.24$. **B:** We binned the genes into subsets of 200 genes according to $\beta$, and computed for each subset the fraction of genes corresponding to each stop codon (TAA, TGA and TAG). For each stop codon, we plotted the resulting fractions as a function of the logarithm of the mean value of $\beta$. **C:** Box plots of $\beta$ for each stop codon.

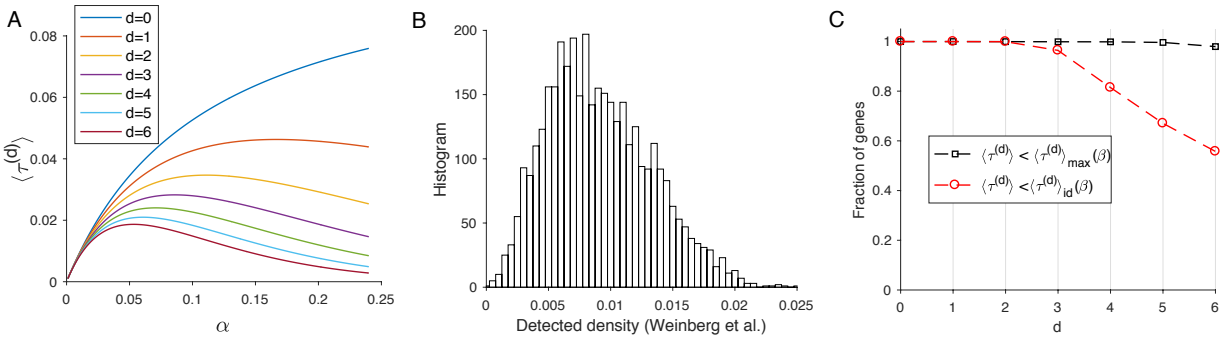


**Figure 7**: **Identifiability of initiation rates from TE measurements.** **A:** For different values of isolation range $d$, we plot the density of isolated particles (see (16)) as a function of the initiation rate $\alpha$ in the LD regime. **B:** A histogram of average detected density, for the 3712 genes in our dataset (see Materials and Methods). **C:** For different ranges $d$ of isolation, we studied the identifiability of the initiation rate from the experimentally observed density shown in B. Black line: we estimated the fraction of genes for which there exists a corresponding value for the initiation rate $\alpha$, such that the associated density of isolated particles is equal to the detected density. This happens when the detected density is less than $\langle \tau^{(d)} \rangle_{\max}(\beta)$ (see (20)), where $\beta$ is the inferred termination rate shown in **Figure 6**. Red line: we estimated the fraction of genes from B for which the initiation rate can be inferred without ambiguity from the plotted curves in A, which happens when the detected density is less than $\langle \tau^{(d)} \rangle_{\mathrm{id}}(\beta)$ (see (21)).

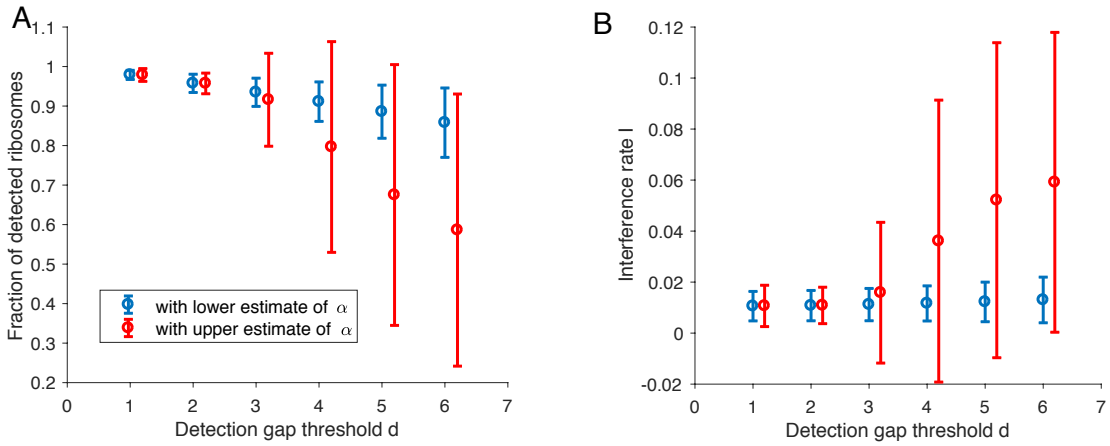


**Figure 8**: **The fraction of detected ribosomes and interference rates. A:** From our dataset of 3712 genes, we used (16) to estimate the fraction of detected ribosomes for different values of the detection gap threshold $d \in \{1, \ldots, 6\}$. To compute these fractions when there is an ambiguity in identifying the initiation rate $\alpha$ (see **Figure 7**C), we considered two possible estimates: a lower estimate and an upper one (see also **Figure S4**A). The plot represents the average fraction of detected ribosomes, with error bars indicating the standard deviation, using lower estimates (in blue) and upper estimates (in red) of $\alpha$. **B:** Interference rates estimated using (17) (see also **Figure S4**B).