bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Theoretical analysis of the distribution of isolated particles in the TASEP: Application to mRNA translation rate estimation

Khanh Dao Duc
Computer Science Division, University of California, Berkeley, CA 94720, USA

Zain H. Saleem
Department of Mathematics, University of Pennsylvania, PA 19104, USA

Yun S. Song*
Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA †

(Dated: October 21, 2017)

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. Although this model has been widely studied in the past, the extent of collision between particles and the average distance between a particle to its nearest neighbor have not been quantified explicitly. We provide here a theoretical analysis of such quantities via the distribution of isolated particles. In the classical form of the model in which each particle occupies only a single site, we obtain an exact analytic solution 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 ~ 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 associated with collisions can be robustly estimated, and show that approximately 1% of the translating ribosomes get obstructed.

INTRODUCTION

The Totally Asymmetric Exclusion Process (TASEP) 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. [1] 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 [2–11], and TASEP-based models have been recently applied to infer the translation rate from experimental measurements [9], in particular ribosome profiling data [12–14]. Ribosome profiling (also known as Ribo-Seq) is an experimental technique developed to examine position-specific densities of ribosomes along each mRNA [15], and thus captures the dynamics of mRNA translation to some extent. However, analytical tools for interpreting ribosome profiling data are still much in need of development [16], as relating the observed footprint density to the corresponding protein production rate 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 [6, 14, 17–19]. 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 [2], 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 ~ 3 codons apart. Using a measurement of ribosome density called "translation efficiency" (TE), we provide estimates of the interference rate associated with traffic

* Also at Chan Zuckerberg Biohub, San Francisco, CA 94158
† yss@berkeley.edu

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

collision, 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

In this section, we briefly introduce the TASEP model, and present our main theoretical results on the classical and generalized versions of model. Appendices A and B summarize some previously known results used in our analysis.

A. 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 Fig. 1a). 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 ith 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 αdt if $\tau_1 = 0$. If $\tau_N = 1$, the particle at site N exits the lattice with probability βdt. The parameters α and β 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,

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

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 Appendix A), 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 Appendix B). 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 (α, β)-space.

In our work, we employed the aforementioned matrix formulation to derive analytic expressions for the average

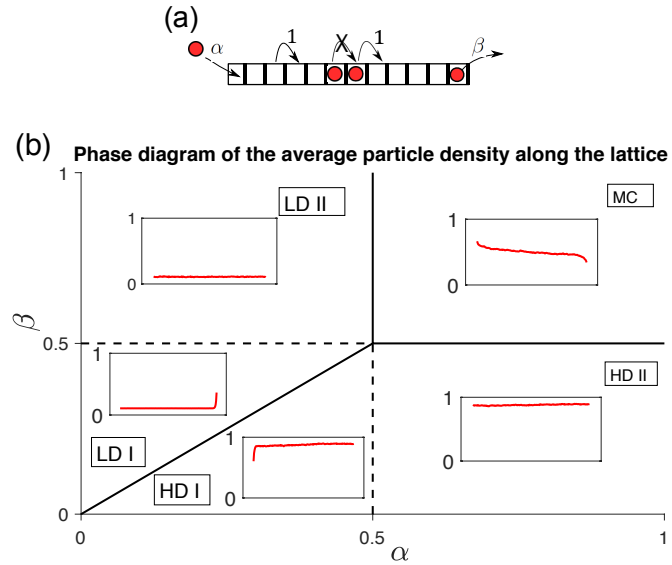

FIG. 1. **Illustration of the TASEP with open boundaries. (a):** A schematic representation of the TASEP model. Particles are introduced at the start of the lattice with exponential rate α and move along with exponential rate 1, provided that there is no particle occupying the next site. At the end of the lattice, they exit with exponential rate β. **(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 (α, β)-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.

density of isolated particles. Specifically, consider the random variable τ_i' defined as

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

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

$$\langle \tau_i' \rangle_N = \langle \tau_i \rangle_N - \langle \tau_{i-1}\tau_i \rangle_N - \langle \tau_i \tau_{i+1} \rangle_N + \langle \tau_{i-1}\tau_i\tau_{i+1} \rangle_N. \tag{3}$$

As detailed in Appendix C, by analyzing the terms on the right hand side of (3), we obtained, for $2 \le i \le N-1$,

$$\langle \tau_i' \rangle_N = D_0(\alpha, \beta, N) - D_1(\alpha, \beta, N)\langle \tau_{i-1} \rangle_{N-1}, \tag{4}$$

where

$$D_0(\alpha, \beta, N) = \alpha\left[1 - \langle \tau_2 \rangle_N + (1 - \langle \tau_1 \rangle_N)(\langle \tau_1 \rangle_{N-1} - \alpha)\right] \tag{5}$$

$$D_1(\alpha, \beta, N) = \alpha(1 - \langle \tau_1 \rangle_N). \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the boundaries, we obtained

$$\langle \tau_1' \rangle_N = \alpha(1 - \langle \tau_1 \rangle_N), \tag{7}$$

$$\langle \tau_N' \rangle_N = \langle \tau_N \rangle_N (1 + \beta) - \langle \tau_{N-1} \rangle_N. \tag{8}$$

As mentioned earlier, exact formulas for $\langle \tau_i \rangle_N$ are known [22] (see Appendix B), so plugging them into (4)–(8) leads to exact results for the average densities of isolated particles along the lattice.

B. 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 (α, β) (see Fig. 1b and (9) below). 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. Using the asymptotics of the particle densities in the three phases [22], we found that $D_0(\alpha, \beta, N)$ and $D_1(\alpha, \beta, N)$ in (5) and (6), respectively, are both asymptotically equivalent to the asymptotics of J in the large-N limit, given by

$$J \sim \begin{cases} \frac{1}{4}, & \text{if } \alpha > \frac{1}{2},\ \beta > \frac{1}{2} \text{ (MC regime)}, \\ \alpha(1-\alpha), & \text{if } \alpha < \frac{1}{2},\ \beta > \alpha \text{ (LD regime)}, \\ \beta(1-\beta), & \text{if } \beta < \frac{1}{2},\ \beta < \alpha \text{ (HD regime)}. \end{cases} \tag{9}$$

Hence, it turns out that the asymptotics of $\langle \tau_i' \rangle$ for $2 \leq i \leq N-1$ are correctly given by using in (3) 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

$$\langle \tau_i' \rangle \sim \begin{cases} J, & \text{for } i = 1, \\ J(1 - \langle \tau_{i-1} \rangle), & \text{for } 2 \leq i \leq N-1, \\ \frac{J}{\beta}\left(1 - \frac{J}{\beta}\right), & \text{for } i = N. \end{cases} \tag{10}$$

Using the asymptotics of $\langle \tau_i \rangle$ in different phases [22], 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 I. 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]

$$\langle \tau_{N+1-i} \rangle_N (\alpha, \beta) = 1 - \langle \tau_i \rangle_N (\beta, \alpha). \tag{11}$$

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

$$\frac{\langle \tau_i' \rangle}{\langle \tau_i \rangle} \sim \begin{cases} \dfrac{\alpha J}{\alpha - J}, & \text{for } i = 1, \\ \dfrac{J(1 - \langle \tau_{i-1} \rangle)}{\langle \tau_i \rangle}, & \text{for } 2 \leq i \leq N-1, \\ 1 - \dfrac{J}{\beta}, & \text{for } i = N. \end{cases} \tag{12}$$

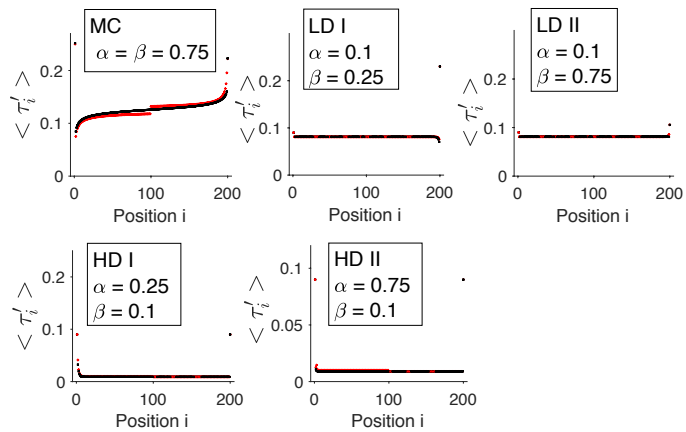

FIG. 2. **The density of isolated particles in different regions of the TASEP phase diagram.** For the different regimes of the TASEP (see also Fig. 1), the asymptotic formulas from Table I (red points) are compared with the exact densities (black points) of isolated particles given by (4)–(8).

As shown in Fig. 2, there is good agreement between our asymptotic formulas and the exact results obtained from using the exact $\langle \tau_i \rangle_N$ in equations (4)–(8). 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. In this regime, some mismatches can be observed at the boundaries and in the middle of the lattice. The apparent discontinuity in the middle of the lattice is due to the fact that we respectively employed in left and right parts of the lattice the asymptotics of the densities far from left and far from right of the boundaries, obtained from Table I. The resulting order of magnitude of the discontinuity gap in the middle is $\frac{1}{\sqrt{N}}$, and it thus vanishes as the lattice length increases. Interestingly, this gap can also be reduced to any arbitrary size by considering larger order approximations of the exact formulas for the densities found in [22]. Close to the boundaries, the formulas we used in Fig. 2 also lead to a mismatch with Monte Carlo simulations, which can be attributed to using asymptotics for positions far from the boundaries. This mismatch can be easily corrected by using the exact formulas for the densities at positions $N-1$ and 2 (Equation 77 in [22]).

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE I. **Asymptotics of $\langle\tau'\rangle$ in the different phases of the classical 1-TASEP.** These are obtained by combining equations (4), (7) and (8) with asymptotics given in [22]. The asymptotics far from the left boundary ($\langle\tau_j\rangle$, $1 \ll j \ll N$) can be derived using the "particle-hole symmetry" (11).

	$\langle\tau'_1\rangle$ (Eq.(7))	$\langle\tau'_{N-j}\rangle$ $(1 \ll j \ll N)$ (Eq.(4))	$\langle\tau'_N\rangle$ (Eq.(8))
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	$\beta(1-\beta)$	$\beta^2(1-\beta)$	$\beta(1-\beta)$

C. The ℓ-TASEP with extended particles

During translation, ribosomes move along mRNAs by decoding one codon at a time, but occupy an extended space of ~ 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 ℓ-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 [2]. 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 (Appendix D) that the density of isolated particles far from the boundaries, simply denoted $\langle\tau'\rangle$, is given by

$$\langle\tau'\rangle = \langle\tau\rangle\left[\frac{1-\ell\langle\tau\rangle}{1-(\ell-1)\langle\tau\rangle}\right]^2. \tag{13}$$

Using the asymptotic densities and currents found by Lakatos and Chou [2], 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 (13) with the asymptotic density $\langle\tau\rangle$ in the large-N limit [2], we obtained the following density of isolated particles in the bulk:

$$\langle\tau'\rangle = \begin{cases} \frac{\sqrt{\ell}}{(1+\sqrt{\ell})^3}, \\ \qquad \text{if } \alpha > \alpha^*,\ \beta > \beta^* \text{ (MC regime)}, \\ \frac{1}{2}J\left[(\ell-1)J+1-\sqrt{((\ell-1)J+1)^2-4\ell J}\right], \\ \qquad \text{if } \alpha < \alpha^*,\ \beta > \alpha \text{ (LD regime)}, \\ \frac{1}{2}J\left[(\ell-1)J+1+\sqrt{((\ell-1)J+1)^2-4\ell J}\right], \\ \qquad \text{if } \beta < \beta^*,\ \beta < \alpha \text{ (HD regime)}, \end{cases} \tag{14}$$

where J is the particle flux given by [2]

$$J \sim \begin{cases} \dfrac{1}{(1+\sqrt{\ell})^2}, & \text{in MC regime}, \\ \dfrac{\alpha(1-\alpha)}{1+(\ell-1)\alpha}, & \text{in LD regime}, \\ \dfrac{\beta(1-\beta)}{1+(\ell-1)\beta}, & \text{in HD regime}. \end{cases} \tag{15}$$

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

$$\langle\tau'_i\rangle = \mathbb{P}(t_i = 1\ ,\ t_{i+\ell} = 0) = J. \tag{16}$$

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

$$\langle\tau'_i\rangle = \mathbb{P}(t_i = 1\ ,\ t_{i-\ell} = 0). \tag{17}$$

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

$$\langle\tau'_i\rangle = J+\mathbb{P}(t_i = 1)-\mathbb{P}(t_{i-\ell} = 1) = J+\langle\tau_i\rangle-\langle\tau_{i-\ell}\rangle. \tag{18}$$

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

$$\langle\tau'_i\rangle = \begin{cases} 2J-\langle\tau_{i-\ell}\rangle, & \text{for } N-\ell < i < N, \\ J\left(1+\frac{1}{\beta}\right)-\langle\tau_{N-\ell}\rangle, & \text{for } i = N. \end{cases} \tag{19}$$

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

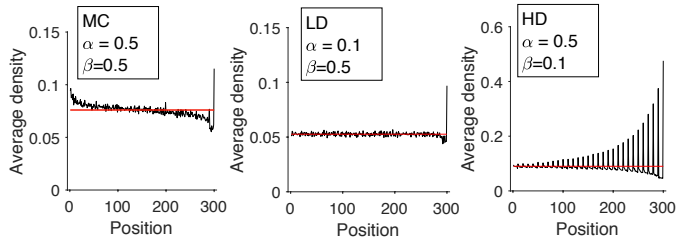

FIG. 3. **The density of particles in the ℓ-TASEP model**. We simulated and plot (in black) the density of particles of the ℓ-TASEP ($\ell = 10$) in the different regimes LD, HD and MC. In red, we plot the estimates of the density in the bulk from Lakatos and Chou [2].

D. Comparison with Monte Carlo simulations and estimation of interference rate

Combining (14), (16) and (19) leads to approximate densities of isolated particles along the lattice in the ℓ-TASEP. The isolated particle densities in the bulk (14) and near the entrance (16) depend only on the flux J, whereas near the exit the result (19) also depends on the density of particles located ℓ sites behind. In the LD regime, this density can be approximated by the density in the bulk [2]. In the other regimes, the density varies near the boundary, so using this approximation might be inaccurate (see Fig. 3). As Fig. 4a shows, however, our theoretical results agree well with the empirical densities of isolated particles obtained from Monte Carlo simulations, for specific values of (α, β) in the LD, HD and MC regimes (for a lattice of length 300, typical of the mRNA sequences we studied next). 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 yield a better match with simulated data than when the same is done for the total density.

More generally, we studied in Fig. 4b how the density and proportion of isolated particles vary as a function of α, for fixed values of β. Overall, our theoretical results were in good agreement with Monte Carlo simulations. Interestingly, whereas the total density (Fig. 4b) 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 β 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 $\frac{\langle \tau' \rangle}{\langle \tau \rangle}$ (Fig. 4b, right panels), showing a linear decrease from $\alpha = 0$ to $\alpha = \beta$, while the total density gets sublinear as α gets closer to β. The first-order Taylor expansion in α 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

$$\frac{\langle \tau' \rangle}{\langle \tau \rangle} = 1 - 2\alpha + O\left(\alpha^2\right). \tag{20}$$

Interestingly, this formula does not depend on ℓ 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

$$I = \frac{1}{2}\left(1 - \frac{\langle \tau' \rangle}{\langle \tau \rangle}\right). \tag{21}$$

Using equation (20), we obtained that the interference rate is close to α in the LD regime.

E. 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

$$\langle \tau_i^{(d)} \rangle = \mathbb{P}(\tau_i = 1,\ x_i^- \leq i - \ell - d,\ x_i^+ \geq i + \ell + d), \tag{22}$$

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 ℓ-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:

$$\langle \tau^{(d)} \rangle \sim \langle \tau \rangle \left[\frac{1 - \ell\langle \tau \rangle}{1 - (\ell - 1)\langle \tau \rangle}\right]^{2d}. \tag{23}$$

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

$$\left(\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d}} = \left(\frac{\langle \tau^{(d')} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d'}}. \tag{24}$$

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

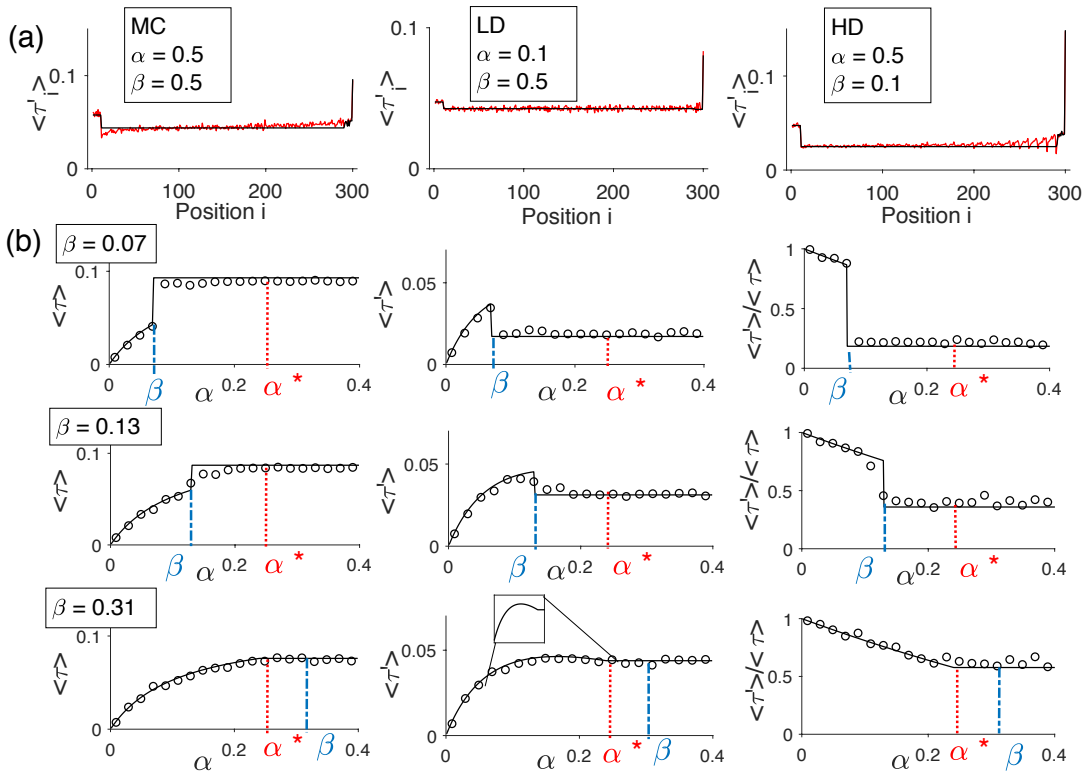

FIG. 4. **Comparison of the results from the refined mean field approach with Monte Carlo simulations. (a):** We simulated the TASEP with extended particles (size $\ell = 10$, sample size $= 10^9$) and plotted (in red) the densities of isolate particles in the three different regimes of the phase diagram. We compared these simulation results with the asymptotics obtain from (13), (16) and (19) (in black). **(b):** For fixed values of β, these plots show how the total density, the density of isolated particles, and their ratio vary as a function of α. 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).

Therefore, we can generalize (21) to obtain a formula for the interference rate for an arbitrary isolation range $d \geq 1$:

$$I = \frac{1}{2}\left[1 - \left(\frac{\langle \tau^{(d)} \rangle}{\langle \tau \rangle}\right)^{\frac{1}{d}}\right]. \tag{25}$$

APPLICATION

We applied our theoretical results to analyze ribosome profiling data and mRNA translation. Ribosomes are complex molecular machines (corresponding to particles in the TASEP) that move along mRNA (the lattice) to translate its associated sequence of codons into proteins. Once bound to the mRNA, ribosomes occupy a space of ~ 10 codons ($\ell = 10$ in the TASEP). For the reader who is new to biology, more basics on how proteins are synthesized in a cell can be found in [4]. Briefly, the ribosome profiling procedure consists of using nuclease to digest translating ribosomes and get ribosome-protected mRNA fragments [15]. These fragments are then aligned to the mRNA sequence to produce a positional distribution of ribosomes along the mRNA. Assuming that there is no bias in ribosome detection and that sufficiently many fragments are observed, this distribution can be associated with the stationary average density of particles in the ℓ-TASEP. However, it is possible that the nuclease may fail to cleave stacked ribosomes [6, 17–19], so only the density of "isolated" ribosomes gets measured. Hence, the profile of ribosome counts along the mRNA produced by the experimental procedure might be different from the true profile (see Fig. 5). 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].

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

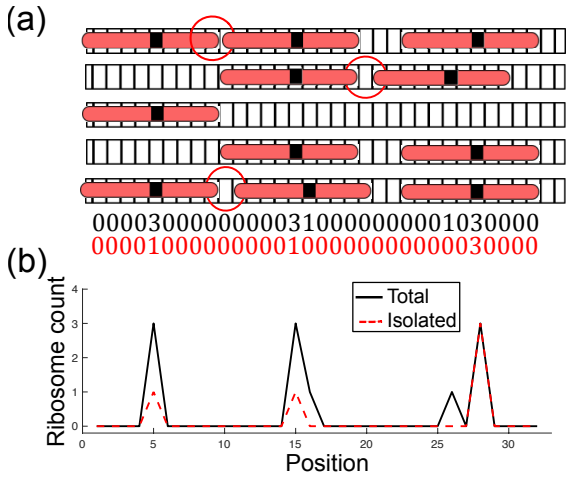

FIG. 5. **A schematic representation of ribosome profiling. (a):** Positions of ribosomes along the mRNA are obtained by nuclease digestion and allow to count the number of ribosomes found at a specific position. However, it is possible that the nuclease cannot cleave stacked ribosomes [6, 14, 17–19]. **(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).

F. 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 Appendix E). 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.

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 (23),

$$TE = aD \left(\frac{1 - 10D}{1 - 9D} \right)^{2d}, \tag{26}$$

where a is the rescaling factor (specifically, $TE = a \cdot \langle \tau^{(d)} \rangle$), that we estimate in practice in Fig. 6a, 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 ℓ in (23) is set to 10. In Fig. 6a, we plotted (26) 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 Fig. 6b, we plotted the root mean square error between (26) and the experimental data, as a function of d and for the value of a corresponding to the linear fit to genes with total density less than 1 ribosome per 100 codons. 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 (26) decreases (Fig. 6a), 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 β for each gene). In Fig. 6c, 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 ~ 65 nucleotides [19], suggesting that $d = 3$ (2 times 30 nucleotides plus 2 other codons).

G. Identifiability of initiation rates and flux from TE measurements

Under the ℓ-TASEP model in the LD regime, the TE is related (as shown in Fig. 7a) to the initiation rate α through equation (23) and the asymptotics of $\langle \tau \rangle$ and J (given in [2]). Assuming that translation occurs in the LD regime (since translation is generally limited by initiation under realistic physiological conditions [31, 32]), we studied whether we could infer the gene-specific initiation rate α using our theoretical results. The detected density is bounded by ~ 0.02 ribosomes per codon in our dataset. From the plotted curves in Fig. 7a, this suggests

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

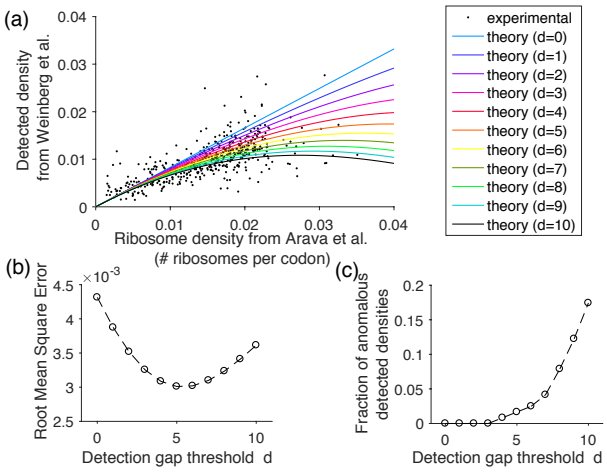

FIG. 6. **Estimation of undetected ribosomes from ribosome profiling experiment. (a):** This plot shows experimental ribosome profiling data of *S. cerevisiae* from Weinberg *et al.* [29] against the total ribosome density obtained from polysome profiling by Arava *et al.* [30] (482 genes). Also shown 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 ℓ-TASEP (see (26)) with various isolation range $d = 0, \ldots, 10$. We set $a = 0.82$, obtained by linear fit to genes with total density less than 1 ribosome per 100 codons. **(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 (26), used in **(a)**.

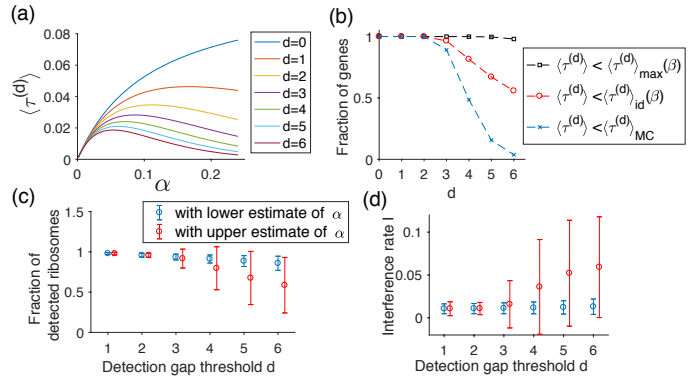

FIG. 7. **Analysis of initiation and interference. (a):** For different values of isolation range d, we plot the density of isolated particles (see (23)) as a function of the initiation rate α in the LD regime. **(b):** For different ranges d of isolation, we studied the identifiability of the initiation rate. Black line: we estimated the fraction of genes for which there exists a corresponding value for the initiation rate α, 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 (27)), where β is the inferred termination rate Red line: we estimated the fraction of genes 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 (28)). **(c):** From our dataset of 3712 genes, we used (23) 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 α (see **(b)**), we considered two possible estimates: a lower estimate and an upper one (see also Fig. 8a). The left 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 α. **(d):** The same as in **(c)** for interference rates, using (25) (see also Fig. 8b).

that for $d \leq 5$ and for all the experimental detected densities, there exists a value for the initiation rate satisfying (23). However, for $d \geq 3$, the identifiability of α (i.e., the uniqueness of α) 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 β, as

$$\langle\tau^{(d)}\rangle_{\max}(\beta) = \sup(\langle\tau^{(d)}\rangle(\alpha),\ \alpha \in [0, \beta]). \tag{27}$$

After estimating the termination rates from our ribosome profiling data (see Appendix F), in Fig. 7b 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 (26)). 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 α 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

$$\langle\tau^{(d)}\rangle_{\mathrm{id}}(\beta) = \begin{cases} \langle\tau^{(d)}\rangle(\beta), & \text{if } \beta \leq \beta^*, \\ \langle\tau^{(d)}\rangle_{\mathrm{MC}}, & \text{otherwise,} \end{cases} \tag{28}$$

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 Fig. 7b, 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 Fig. 6, 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 α_2 is larger than for α_1, there are also more closely stacked ribosomes that are not detected. Hence, the density of isolated particles

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is the same for both. As the flux is an increasing function of the initiation rate, such ambiguity also applies for inferring the flux.

H. 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 α, 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 α, 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 interference rate 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 α when $d \geq 3$ (see Fig. 7b). Thus, when such ambiguity occurred, we considered both lower and upper estimates of α, and found their respective resulting fractions of detected ribosomes $\langle\tau^{(d)}\rangle/\langle\tau\rangle$ and interference rates (Fig. 7c and d). We obtained that for $d = 3$ or 4, suggested by Fig. 6b and c, the lower estimates of α lead to fractions of detected ribosomes lying between $91.2\pm5\%$ and $93.5\pm3.5\%$. The upper estimates of α lead to smaller mean and larger variability (between $80\pm26\%$ 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 α remain stable around 1% for all d, with only a slight increase of standard deviation from 0.5 to 0.9%. Somewhat larger variation is observed for the interference rates corresponding to the upper estimates of α, 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 Fig. 8. When plotting the fraction $\frac{\langle\tau^{(d)}\rangle}{\langle\tau\rangle}$ of detected ribosomes as a function of $\langle\tau^{(d)}\rangle$ (Fig. 8a), we observed that for large values of the fraction (associated with low α), 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 Fig. 8b). More generally, the formula (25) 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 ~ 0.2), explaining its smaller variation across our dataset and different values of d.

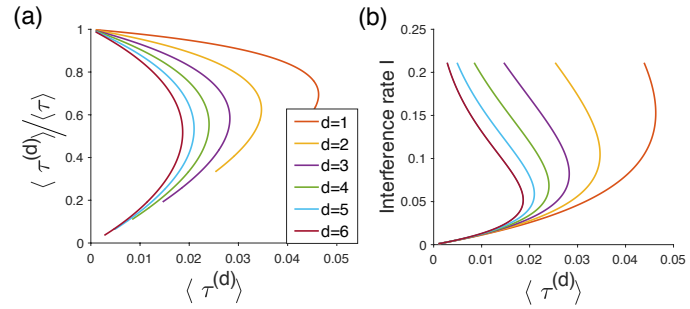

FIG. 8. **The fraction of isolated particles and interference rate as a function of $\langle\tau^{(d)}\rangle$.(a):** For different isolation ranges $d \in \{1, \ldots, 6\}$, we plot the fraction of isolated particles as a function of the average density of isolated particles $\langle\tau^{(d)}\rangle$, according to (23). Note that for given d, some values of $\langle\tau^{(d)}\rangle$ can lead to two possible fractions of isolated particles. **(b):** As in **(a)**, we plot the isolation rate as a function of the average density of isolated particles $\langle\tau^{(d)}\rangle$, according to (25). Note that for $\langle\tau^{(d)}\rangle \leq 0.02$ and all d, the initiation rates associated with the lower branch are very close.

DISCUSSION

I. Comparison with existing literature

In this article, we provided a complete analysis of the distribution of isolated particles in the TASEP model with open boundaries. This study was motivated by the possible non-detection of stacked ribosome in ribosome profiling, which is a recent experimental technique [23]. As shown in (3), the density of isolated particle is related to two- and three-point correlators, while most past analyses focused on computing the total density profile and the flux, which involve one- and two-point correlators. 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 [33, 34] and networks [35, 36], or with more complex dynamics associated with pausing [37, 38], random rates [33, 39, 40] or multiple species [22, 34, 40–42], 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

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cases, the solution of the associated master equation can be found using a matrix formulation [20, 21, 40, 42], suggesting that the work presented here could be generalized.

We further studied the ℓ-TASEP model with extended particles of size ℓ 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, 43]. In practice, however, the associated algebra is rather complex, making it challenging to derive analytic results [2, 20]. To cope with this complexity, several approaches using mean field approximation have been developed [2, 4, 37, 44, 45]. Although the mean-field treatment may inaccurately capture the full profile in some regimes [37], it provides a more accurate approximation when the profile is restricted to isolated particles. More generally, the "level" of "mean-field" can also impact the quality of the approximation. At the simplest level, assuming a uniform distribution of particles without anti-correlations due to local interactions and using (13), one may obtain a rather poor approximation of the density of isolated particle, as the different regimes are not even separated correctly (all the cases in Fig. 4 would, for example, be considered as being in HD). Unlike this simple mean-field approach, the refined analytic approximation proposed by Lakatos and Chou [2] leads to formulas that show good agreement with simulations for current and bulk density [37]. 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. Higher "levels" of "mean-field" [44, 45] can help to improve the accuracy of the local density, but at the cost of losing analytic expressions, and possible existence of numerical instabilities and imprecisions [45].

The choice of lattice length ($n = 300$) in our comparison with Monte Carlo simulations was motivated by the typical size of the mRNA found in our dataset. As the length of the lattice increases, we expect the accuracy to improve, especially in the bulk, as the density would vary less. For much longer lattices, it would also be natural to study the hydrodynamic limit of the ℓ-TASEP with open boundaries. Interestingly, while previous studies derived a general PDE satisfied by the density for the ℓ-TASEP in the hydrodynamic limit [46, 47], a rigorous derivation, notably including that of boundary conditions, and analysis of the PDE to determine the associated phase diagram are still missing. We are currently exploring this research direction.

J. Application to ribosome profiling data and comparison with other approaches

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 ~ 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 [48, 49]. In yeast (which is the organism studied in our dataset), no periodic peaks of density were detected in this region across multiple datasets [19, 50–57], suggesting non-detection of stacked ribosomes. In contrast, such peaks have been detected for other organisms and different protocols [58, 59].

Other methods have also been developed previously to infer the initiation rates associated with specific genes from polysome [9] or ribosome profiling [12]. 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. More precisely, there is evidence of variation of the elongation rate along the transcript, especially in the first ~ 200 codons, leading to the so-called "5' translational ramp" [23] (in another study [14], we quantified the extent of the interference created by this ramp). However, it has been shown that the average elongation speed along the transcript sequence is approximately constant around 5.6 codon/s [15], allowing the use of the homogeneous TASEP model as a first approximation of the translation dynamics.

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.

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

K. 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.

[1] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, Biopolymers **6**, 1 (1968).
[2] G. Lakatos and T. Chou, Journal of Physics A: Mathematical and General **36**, 2027 (2003).
[3] T. Chou, K. Mallick, and R. Zia, Reports on Progress in Physics **74**, 116601 (2011).
[4] R. K. Zia, J. Dong, and B. Schmittmann, Journal of Statistical Physics **144**, 405 (2011).
[5] D. Chowdhury, A. Schadschneider, and K. Nishinari, Physics of Life reviews **2**, 318 (2005).
[6] A. Dana and T. Tuller, PLoS Comput Biol **8**, e1002755 (2012).
[7] A. K. Sharma and D. Chowdhury, Journal of Theoretical Biology **289**, 36 (2011).
[8] T. Chou and G. Lakatos, Physical Review Letters **93**, 198101 (2004).
[9] L. Ciandrini, I. Stansfield, and M. C. Romano, PLoS Comput Biol **9**, e1002866 (2013).
[10] A. Basu and D. Chowdhury, Physical Review E **75**, 021902 (2007).
[11] T. von der Haar, Computational and Structural Biotechnology Journal **1**, 1 (2012).
[12] A. A. Gritsenko, M. Hulsman, M. J. Reinders, and D. de Ridder, PLoS Comput Biol **11**, e1004336 (2015).
[13] H. Zur and T. Tuller, Nucleic Acids Research **44**, 9031 (2016).
[14] K. Dao Duc and Y. S. Song, bioRxiv , 090837 (2016).
[15] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, Cell **147**, 789 (2011).
[16] G. A. Brar and J. S. Weissman, Nature Reviews Molecular Cell Biology **16**, 651 (2015).
[17] D. E. Andreev, P. B. O'Connor, G. Loughran, S. E. Dmitriev, P. V. Baranov, and I. N. Shatsky, Nucleic Acids Research **45**, 513 (2017).
[18] A. R. Subramaniam, B. M. Zid, and E. K. O'Shea, Cell **159**, 1200 (2014).
[19] N. R. Guydosh and R. Green, Cell **156**, 950 (2014).
[20] R. A. Blythe and M. R. Evans, Journal of Physics A: Mathematical and Theoretical **40**, R333 (2007).
[21] A. Schadschneider, D. Chowdhury, and K. Nishinari, *Stochastic Transport in Complex Systems: from Molecules to Vehicles* (Elsevier, 2010).
[22] B. Derrida, M. R. Evans, V. Hakim, and V. Pasquier, Journal of Physics A: Mathematical and General **26**, 1493 (1993).
[23] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, Science **324**, 218 (2009).
[24] F. Spitzer, Advances in Mathematics **5**, 246 (1970).
[25] V. Privman, *Nonequilibrium Statistical Mechanics in One Dimension* (Cambridge University Press, 2005).
[26] T. Sasamoto and M. Wadati, Journal of Physics A: Mathematical and General **31**, 6057 (1998).
[27] P. B. O'connor, D. E. Andreev, and P. V. Baranov, Nature Communications **7** (2016).
[28] M. V. Gerashchenko and V. N. Gladyshev, Nucleic Acids Research **45**, e6 (2017).
[29] D. E. Weinberg, P. Shah, S. W. Eichhorn, J. A. Hussmann, J. B. Plotkin, and D. P. Bartel, Cell Reports **14**, 1787 (2016).
[30] Y. Arava, Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag, Proceedings of the National Academy of Sciences **100**, 3889 (2003).
[31] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin, Cell **153**, 1589 (2013).
[32] D. Chu, E. Kazana, N. Bellanger, T. Singh, M. F. Tuite, and T. von der Haar, The EMBO Journal **33**, 21 (2014).
[33] S. de Queiroz and R. Stinchcombe, Physical Review E **78**, 031106 (2008).
[34] A. Ayyer and S. Linusson, Advances in Applied Mathematics **57**, 21 (2014).
[35] I. Neri, N. Kern, and A. Parmeggiani, Physical Review Letters **107**, 068702 (2011).
[36] S. Bittihn and A. Schadschneider, Physical Review E **94**, 062312 (2016).
[37] J. Dong, B. Schmittmann, and R. K. Zia, Physical Review E **76**, 051113 (2007).
[38] M. Sahoo and S. Klumpp, Journal of Physics A: Mathematical and Theoretical **49**, 315001 (2016).
[39] J. S. Nossan, Journal of Physics A: Mathematical and Theoretical **46**, 315001 (2013).
[40] C. Arita and K. Mallick, Journal of Physics A: Mathematical and Theoretical **46**, 085002 (2013).
[41] S. Prolhac, M. R. Evans, and K. Mallick, Journal of Physics A: Mathematical and Theoretical **42**, 165004 (2009).
[42] M. R. Evans, P. A. Ferrari, and K. Mallick, Journal of Statistical Physics **135**, 217 (2009).
[43] K. Klauck and A. Schadschneider, Physica A: Statistical Mechanics and its Applications **271**, 102 (1999).
[44] L. B. Shaw, R. Zia, and K. H. Lee, Physical Review E **68**, 021910 (2003).
[45] L. B. Shaw, J. P. Sethna, and K. H. Lee, Physical Review E **70**, 021901 (2004).
[46] G. Schönherr, Physical Review E **71**, 026122 (2005).
[47] G. Schönherr and G. Schütz, Journal of Physics A: Mathematical and General **37**, 8215 (2004).
[48] X. Yu, M. R. Willmann, S. J. Anderson, and B. D. Gregory, The Plant Cell **28**, 2385 (2016).
[49] V. Pelechano, W. Wei, and L. M. Steinmetz, Cell **161**, 1400 (2015).
[50] J. Gardin, R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, and B. Futcher, eLife **3**, e03735 (2014).
[51] M. V. Gerashchenko and V. N. Gladyshev, Nucleic Acids Research **42**, e134 (2014).
[52] C. C. Williams, C. H. Jan, and J. S. Weissman, Science **346**, 748 (2014).
[53] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, eLife **3**, e01257 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[54] C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller, Molecular Systems Biology **10**, 770 (2014).
[55] D. D. Nedialkova and S. A. Leidel, Cell **161**, 1606 (2015).
[56] C. H. Jan, C. C. Williams, and J. S. Weissman, Science **346**, 1257521 (2014).
[57] O. Carja, T. Xing, J. B. Plotkin, and P. Shah, bioRxiv , 100032 (2017).
[58] D. E. Andreev, P. B. O'Connor, A. V. Zhdanov, R. I. Dmitriev, I. N. Shatsky, D. B. Papkovsky, and P. V. Baranov, Genome Biology **16**, 90 (2015).
[59] A. V. Lobanov, S. M. Heaphy, A. A. Turanov, M. V. Gerashchenko, S. Pucciarelli, R. R. Devaraj, F. Xie, V. A. Petyuk, R. D. Smith, L. A. Klobutcher, *et al.*, Nature Structural & Molecular Biology **24**, 61 (2017).
[60] B. Derrida, E. Domany, and D. Mukamel, Journal of Statistical Physics **69**, 667 (1992).
[61] J. Buschle, P. Maass, and W. Dieterich, Journal of Physics A: Mathematical and General **33**, L41 (2000).

Appendix A: Equations satisfied by the correlators in the TASEP

Averaging the master equation associated with the TASEP, the particle densities satisfy the following relations [60]:

$$0 = \langle\tau_1\rangle_N - \langle\tau_1\tau_2\rangle_N - \alpha(1-\langle\tau_1\rangle_N), \tag{A1}$$
$$0 = \langle\tau_i\tau_{i+1}\rangle_N - \langle\tau_{i-1}\tau_i\rangle_N - \langle\tau_i\rangle_N + \langle\tau_{i-1}\rangle_N \quad \text{, for } 2\le i\le N-1, \tag{A2}$$
$$0 = \beta\langle\tau_N\rangle_N - \langle\tau_{N-1}\rangle_N + \langle\tau_{N-1}\tau_N\rangle_N. \tag{A3}$$

Note that (A2) implies $\langle\tau_i(1-\tau_{i+1})\rangle_N = \langle\tau_{i-1}(1-\tau_i)\rangle_N$ for all $i = 2,\dots,N-1$. This translation-invariant quantity is called the current (or flux) and is denoted by J. One can also relate the two-point correlators with the three-point correlators as

$$0 = \langle\tau_1\tau_2\tau_3\rangle_N - \langle\tau_1\tau_2\rangle_N(1+\alpha) + \alpha\langle\tau_2\rangle_N, \tag{A4}$$
$$0 = \langle\tau_{i-1}\tau_i\tau_{i+1}\rangle_N - \langle\tau_{i-2}\tau_{i-1}\tau_i\rangle_N - \langle\tau_{i-1}\tau_i\rangle_N + \langle\tau_{i-2}\tau_i\rangle_N \quad \text{, for } 3\le i\le N-1, \tag{A5}$$
$$0 = \langle\tau_{N-2}\tau_{N-1}\tau_N\rangle_N - \langle\tau_{N-2}\tau_N\rangle_N + \beta\langle\tau_{N-1}\tau_N\rangle_N. \tag{A6}$$

Appendix B: Description of the matrix Ansatz used in the simple TASEP

To derive analytical expressions for the average densities of the TASEP, Derrida *et al.* [22] showed that the steady state probability of a given configuration can be derived using a matrix formulation as

$$\mathbb{P}(t_1,\dots,t_N) = \frac{f_N(t_1,\dots,t_N)}{\sum_{\theta_1=0,1}\cdots\sum_{\theta_N=0,1} f_N(\theta_1,\dots,\theta_N)}, \tag{B1}$$

where

$$f_N(t_1,\dots,t_N) = \langle W|\prod_{i=1}^{N}(t_iD + (1-t_i)E|V\rangle. \tag{B2}$$

Here, D and E are infinite dimensional square matrices and $|V\rangle$ and $\langle W|$ are column and row vectors respectively satisfying

$$DE = D + E, \tag{B3}$$
$$D\,|V\rangle = \frac{1}{\beta}\,|V\rangle, \tag{B4}$$
$$\langle W|\,E = \frac{1}{\alpha}\,\langle W|. \tag{B5}$$

Using this formulation, the particle density can be derived as

$$\langle\tau_i\rangle_N = \frac{\langle W|C^{i-1}DC^{N-i}|V\rangle}{\langle W|C^N|V\rangle}, \tag{B6}$$

where $C = D + E$. More generally, for any given index set $i_1, i_2, \dots, i_k$ such that $1 \le i_1 < \cdots < i_k \le N$, we get

$$\langle\tau_{i_1}\cdots\tau_{i_k}\rangle_N = \frac{\langle W|C^{i_1-1}DC^{i_2-i_1-1}\cdots C^{i_k-i_{k-1}-1}DC^{N-i_k}|V\rangle}{\langle W|C^N|V\rangle}. \tag{B7}$$

Appendix C: Computing the density of isolated particles

Using the matrix Ansatz, we derive here an analytical expression for the average density of isolated particles $\langle\tau'_j\rangle_N$. Our goal is to get $\langle\tau'_j\rangle_N$ as a function of the average densities $\langle\tau_j\rangle_N$. The density of isolated particles inside the lattice ($2 \le i \le N-1$) is given by (see equation (3))

$$\langle\tau'_i\rangle_N = \langle\tau_i\rangle_N - \langle\tau_{i-1}\tau_i\rangle_N - \langle\tau_i\tau_{i+1}\rangle_N + \langle\tau_{i-1}\tau_i\tau_{i+1}\rangle_N. \tag{C1}$$

For $2 \le j \le N-1$, we first derive the expression of the two point correlators $\langle\tau_j\tau_{j+1}\rangle_N$ by summing equation (A2) over $i \in \{2,\dots,j\}$ and using the boundary equation (A1)

$$\langle\tau_j\tau_{j+1}\rangle_N = \langle\tau_j\rangle_N - \alpha\left(1-\langle\tau_1\rangle_N\right). \tag{C2}$$

Similarly, for $3 \le j \le N-1$, summing equation (A5) from $i = 3$ to j and using boundary equations (A1) and (A4) gives

$$\langle\tau_{j-1}\tau_j\tau_{j+1}\rangle_N = \langle\tau_1\tau_2\tau_3\rangle_N + \sum_{p=3}^{j}\langle\tau_{p-1}\tau_p\rangle_N - \langle\tau_{p-2}\tau_p\rangle_N \tag{C3}$$
$$= (1+\alpha)^2\langle\tau_1\rangle_N - \alpha(1+\alpha+\langle\tau_2\rangle_N) + \sum_{p=3}^{j}\langle\tau_{p-1}\tau_p\rangle_N - \langle\tau_{p-2}\tau_p\rangle_N. \tag{C4}$$

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Using the matrix formulation and the identities $DCD = D(DC - DE + ED) = DDC - DC + CD$, we get

$$\langle \tau_{p-2}\tau_p\rangle_N = \frac{\langle W|C^{p-3}DCDC^{N-p}|V\rangle}{\langle W|C^N|V\rangle} = \langle \tau_{p-2}\tau_{p-1}\rangle_N + J_N\left(\langle \tau_{p-1}\rangle_{N-1} - \langle \tau_{p-2}\rangle_{N-1}\right), \tag{C5}$$

where $J_N = \frac{\langle W|C^{N-1}|V\rangle}{\langle W|C^N|V\rangle} = \alpha\left(1 - \langle \tau_1\rangle_N\right)$ is the particle current at steady state [22]. Combining (C5) with (C4) and using (C1) and (A1) yield the result for the three-point correlator

$$\langle \tau_{j-1}\tau_j\tau_{j+1}\rangle_N = \langle \tau_{j-1}\rangle_N - \alpha\left[1 + \alpha + \langle \tau_2\rangle_N - \cdots (2+\alpha)\langle \tau_1\rangle_N\right] - J_N\left(\langle \tau_{j-1}\rangle_{N-1} - \langle \tau_1\rangle_{N-1}\right), \tag{C6}$$

for $3 \le j \le N-1$. Using (A1) and (A4), this equation is also true for $j = 2$. Using (C6), (C2) and (3) gives us the formula for the density of isolated particles, for $2 \le i \le N-1$

$$\langle \tau_i'\rangle_N = \alpha\left[1 - \langle \tau_2\rangle_N + \alpha(\langle \tau_1\rangle_N - 1)\right] - \cdots J_N\left(\langle \tau_{i-1}\rangle_{N-1} - \langle \tau_1\rangle_{N-1}\right). \tag{C7}$$

Finally we can use $J_N = \alpha\left(1 - \langle \tau_1\rangle_N\right)$ to write the above formula in a more compact notation, as

$$\langle \tau_i'\rangle_N = D_0(\alpha,\beta,N) - D_1(\alpha,\beta,N)\langle \tau_{i-1}\rangle_{N-1}, \tag{C8}$$

where

$$D_0(\alpha,\beta,N) = \alpha\left[1 - \langle \tau_2\rangle_N + \left(1 - \langle \tau_1\rangle_N\right)\left(\langle \tau_1\rangle_{N-1} - \alpha\right)\right], \tag{C9}$$

$$D_1(\alpha,\beta,N) = \alpha(1 - \langle \tau_1\rangle_N). \tag{C10}$$

Similarly, using equations (A1) and (A3) at the boundaries yields

$$\langle \tau_1'\rangle_N = \alpha(1 - \langle \tau_1\rangle_N), \tag{C11}$$

$$\langle \tau_N'\rangle_N = \langle \tau_N\rangle_N(1+\beta) - \langle \tau_{N-1}\rangle_N. \tag{C12}$$

Appendix D: Density of isolated particles in the bulk for the ℓ-TASEP

We compute here an estimate of the density of isolated particles of size ℓ in the bulk ($\langle \tau_i\rangle$, $1 \ll i \ll N-l$). To do so, we use an approximation from Lakatos and Chou [2], assuming that the number of states of n particles of length l, confined to a length of $N' \ge n\ell$ lattice sites, is given by the partition function [61]

$$Z(n,N') = \binom{N' - (\ell-1)n}{n}. \tag{D1}$$

For a given position $i \in \{1,\ldots,\le N-l\}$, we introduce x_i^- and x_i^+ as the positions of the closest particles to the left and the right of i, respectively, so we get

$$\langle \tau_i'\rangle = \mathbb{P}(\tau_i = 1,\ x_i^- < i-\ell,\ x_i^+ > i+\ell) \tag{D2}$$

$$= \mathbb{P}(\tau_i = 1)\mathbb{P}(x_i^- < i-\ell,\ x_i^+ > i+\ell \mid \tau_i = 1). \tag{D3}$$

Assuming x_i^- and x_i^+ being independent yields

$$\langle \tau_i'\rangle = \mathbb{P}(\tau_i = 1)\mathbb{P}(x_i^- < i-\ell \mid \tau_i = 1)\mathbb{P}(x_i^+ > i+\ell \mid \tau_i = 1). \tag{D4}$$

Using (D1), the probability $p^+_{n,N'}$ that $x_i^+ > i+\ell$, conditioned on $\tau_i = 1$ and there being n particles in the window $[i+\ell : i+\ell+N'-1]$ is

$$p^+_{n,N'} = \frac{Z(n,N'-1)}{Z(n,N')} = \frac{1-\rho\ell}{1-\rho(\ell-1)}, \tag{D5}$$

where $\rho = \frac{n}{N'}$. When n and N' get large and assuming the density of particles in the bulk of the lattice to be approximately constant (denoted $\langle\tau\rangle$), we can replace $p^+_{n,N'}$ and ρ in equation (D5) by $\mathbb{P}(x_i^+ > i+\ell \mid \tau_i = 1)$ and $\langle\tau\rangle$, respectively, which gives

$$\mathbb{P}(x_i^+ > i+\ell \mid \tau_i = 1) = \frac{1-\langle\tau\rangle\ell}{1-\langle\tau\rangle(\ell-1)}. \tag{D6}$$

Similarly, we obtain $\mathbb{P}(x_i^- < i-\ell \mid \tau_i = 1) = \frac{1-\langle\tau\rangle\ell}{1-\langle\tau\rangle(\ell-1)}$. Combining these relations and replacing $\mathbb{P}(\tau_i = 1)$ by $\langle\tau\rangle$ in equation (D3), we obtain that the density of isolated particles in the bulk, simply denoted $\langle\tau'\rangle$, is given by

$$\langle\tau'\rangle = \langle\tau\rangle\left(\frac{1-\ell\langle\tau\rangle}{1-(\ell-1)\langle\tau\rangle}\right)^2. \tag{D7}$$

Similarly, for isolation range d, we obtain

$$\langle\tau_i^{(d)}\rangle \sim \mathbb{P}(\tau_i = 1)\left[\frac{Z(n,N'-d)}{Z(n,N')}\right]^2, \tag{D8}$$

which simplifies to the following expression in the large-N limit:

$$\langle\tau^{(d)}\rangle \sim \langle\tau\rangle\left[\frac{1-\ell\langle\tau\rangle}{1-(\ell-1)\langle\tau\rangle}\right]^{2d}. \tag{D9}$$

Appendix E: 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

bioRxiv preprint doi: https://doi.org/10.1101/147017; this version posted October 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $+1$.

Appendix F: 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 scaled termination rate as

$$\beta = \frac{(N-1)}{P_N \sum_{i=1}^{N-1} \frac{1}{P_i}}. \tag{F1}$$

Such estimation is valid when there is little ribosomal interference, such that the elongation rate can be approximated by the inverse of the profile [14]. In another study [14], 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 ~ 1.57 site per gene, these "naive" estimates are valid over a large majority of the sites.