ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Cavallaro M, Wang Y, Hebenstreit D, Dutta R. 2023 Bayesian inference of polymerase dynamics over the exclusion process. *R. Soc. Open Sci.* **10**: 221469. https://doi.org/10.1098/rsos.221469

Received: 14 November 2022
Accepted: 12 July 2023

Subject Category:
Physics and biophysics

Subject Areas:
systems biology/computational biology/biophysics

Keywords:
gene expression, non-equilbrium physics, Bayesian statistics, particle transport

Author for correspondence:
Massimo Cavallaro
e-mail: m.cavallaro@warwick.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6760029.

THE ROYAL SOCIETY PUBLISHING

Bayesian inference of polymerase dynamics over the exclusion process

Massimo Cavallaro[1,2,3], Yuexuan Wang[5], Daniel Hebenstreit[2] and Ritabrata Dutta[4]

[1]Mathematics Institute, [2]School of Life Sciences, [3]Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, and [4]Department of Statistics, University of Warwick, Coventry, UK
[5]Institute of Applied Statistics, Johannes Kepler Universität, Linz, Austria

MC, 0000-0002-2365-6024; YW, 0000-0002-0728-921X

Transcription is a complex phenomenon that permits the conversion of genetic information into phenotype by means of an enzyme called RNA polymerase, which erratically moves along and scans the DNA template. We perform Bayesian inference over a paradigmatic mechanistic model of non-equilibrium statistical physics, i.e. the asymmetric exclusion processes in the hydrodynamic limit, assuming a Gaussian process prior for the polymerase progression rate as a latent variable. Our framework allows us to infer the speed of polymerases during transcription given their spatial distribution, while avoiding the explicit inversion of the system's dynamics. The results, which show processing rates strongly varying with genomic position and minor role of traffic-like congestion, may have strong implications for the understanding of gene expression.

1. Introduction

DNA is a long polymeric molecule that encodes information as a sequence of nucleotides (Nts). Turning this information into a phenotype is a complex phenomenon hinged upon transcription, the molecular process in which particular segments of DNA (i.e. the genes) are scanned and their information is copied into mRNA by the enzyme RNA polymerase II (PolII). The transcription itself consists of several steps which can be differentially regulated to alter the timing and the output of the mRNA production [1,2].

The transcription can also be seen as a non-equilibrium process, where the PolIIs are being transported as particles on a one-dimensional lattice, the lattice being the DNA template which the PolIIs bind to. We can further consider this process having left and right boundaries, representing the transcription start site (TSS) and the transcription end site (TES), respectively (figure 1*a*).

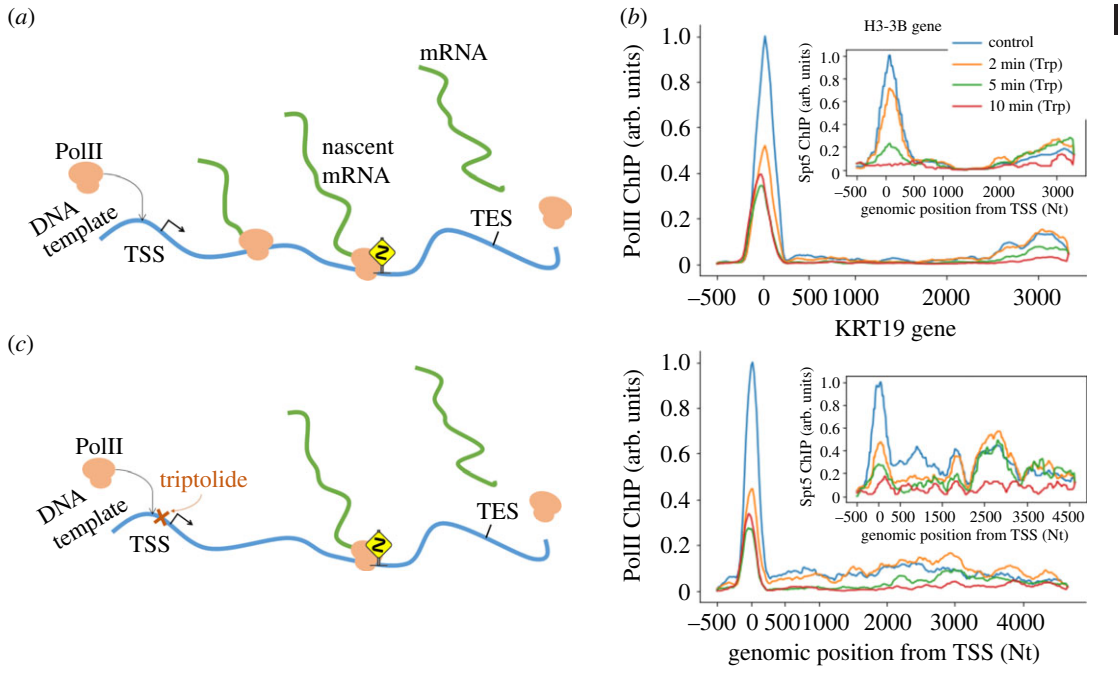

Figure 1. Biological processes and data. (*a*) Simplified diagram of mRNA synthesis. PolII molecules bind to the DNA upstream of TSS and moves downstream towards the TES, where it is released along with the synthesized mRNA. In certain genomic regions (indicated by a dangerous bend sign), PolIIs slow down. (*b*) ChIP-seq experiments yield the relative abundance of PolII at each genomic position, here illustrated for *H3-3B* (top) and *KR19* (bottom) genes; insets show the Spt5-bound PolII abundances for the same genes. (*c*) In the presence of triptolide (Trp), transport is blocked upstream of TSS, while transcriptional engaged PolII are allowed to complete elongation; this is reflected in the ChIP-seq profiles obtained 2, 5 and 10 min after treatment (also in (*b*)).

Within the gene body, the PolIIs erratically travel along the template and their abrupt slowing down in certain genomic regions is known as *pausing* dynamics [3,4]. While the pausing is an essential part of the transcriptional machinery and contributes to the regulation of genes' expression levels, a comprehensive quantitative understanding of its dynamics is still missing [5,6].

We present a modelling framework to help understand gene regulation and quantitatively study the pausing dynamics given real-world data. In the literature, a number of different mechanistic models have been introduced to elucidate transcription, starting from the simple telegraph model [7] to more complicated multi-state models that account for many interactions [8–11], with each model reflecting determinate aspects of the whole biological system complexity. Here, we are primarily interested in the pausing and employ a generalization of a paradigmatic model of particle transport, the asymmetric simple exclusion process (ASEP, [12–14]) in the hydrodynamic limit [15]. The ASEP is a class of models of particles on a one-dimensional lattice, whose behaviour is chiefly determined by the rates at which the particles hop on the lattice. More specifically we require the rate profile function, which we refer to as $\tilde{p}$, to be spatially varying yet smooth as in [16,17], see also [18], thus making it possible to model this function by a Gaussian process (GP) [19]. Noticing the analogy between the PolII transport in the gene body and the particle hopping in the exclusion process, learning $\tilde{p}$ allows the study of the pausing dynamics in a gene. Importantly, we provide an inferential scheme to learn this rate function by Bayesian inference given real molecular biology data, assuming a prior on the profile function induced by a GP prior on a latent variable. In other words, integrating the dynamics defined by the rate $\tilde{p}$ generates transient time-course density profiles; we estimate $\tilde{p}$ given observed density profiles without explicitly inverting the system's dynamics. Other models of PolII dynamics also leverage GPs for inference from biological data, with GPs representing transcriptional activity over time [20,21]. By contrast, the GP here describes a function of genomic position, with its minima corresponding to pausing regions. Due to its generality, our framework can be deployed to estimate the rate profiles of any one-dimensional transport problem.

The manuscript is organized as follows. Section 2.1 describes the biology of pausing and the next-generation sequencing (NGS) data types which are available to study it. Sections 2.2 and 2.3,

respectively, discuss the ASEP as a mathematical model for transcription with pausing and a Bayesian inferential framework for model fitting. We present the results in §3 and conclude with a discussion in §4.

2. Model definition

2.1. Biological processes and data

RNA polymerases have a central role in the biology of transcription. We distinguish different classes of RNA polymerases, each having different structure and control mechanism. Bacteria and Archaea only have one RNA polymerase type. Eukaryotes have multiple types, of which RNA PolII is known to catalyse synthesis of protein-encoding RNA (messenger RNA or mRNA). In this paper, we describe mRNA transcription by PolII, but the inferential framework we present is general and can be extended to other transport phenomena. PolII binds to DNA upstream of the TSS, initiates the mRNA synthesis and then traverses the DNA downstream (elongation) until it pauses at a certain gene location, ready to respond to a developmental or environmental signal that instructs to resume the elongation. PolIIs are also found proximal to TSS in a so-called 'poised' state, which has not initiated synthesis of the mRNA chain. Poised and paused PolIIs can be differentiated as only paused PolIIs have a tail of nascent mRNA and are bound to transcription factor Spt5 [22]. The process terminates when the PolII reaches the TES and the transcribed mRNA is released. As a result of these steps, the output is modulated in both timing and intensity. However, many details, such as the pausing, are not well understood [5]. The presence of transcriptional pausing in eukaryotes is revealed by several assays based on NGS, which is widely used in molecular biology to study molecules involved in genic processes. In the PolII *ChIP-seq* assay, PolII-bound DNA is isolated by chromatin immunoprecipitation with a PolII antibody and is then subject to high-throughput sequencing. This provides a genome-wide view of the PolII binding sites for all forms of PolII, including both those poised or transcriptionally engaged and those which are bound to DNA and static. In ChIP-seq experiments, DNA fragments extracted from cells and associated with a specific protein (here polymerase) are amplified, sequenced and mapped to the reference genome, with fragments generally in the 150–300 Nt range [23] (while transcribing PolII covers less than 50 Nts of DNA [24]). This means that the precise locations of the individual proteins are not known and the assay only returns the overlap of reads from many different cells. For each genomic position, PolII ChIP-seq returns a signal as a proxy of polymerase occupancy.

For this study, we binned ChIP-seq reads from genomic ranges of selected genes (from cultured human cell lines, *Material and methods*) into 20 Nt bins, thus yielding coarse-grained read profiles (which we refer to as y) such as those illustrated in figure 1*b*. The number of these reads at a position x is proportional to the occupation probability $\varrho(x)$. The proportionality factor, which depends on the number of cells used in the experiment and on further signal amplification intrinsic to the sequencing procedure, cannot be directly accessed with precision and is only known with substantial uncertainty [25].

Other methods available to study the pausing include but are not limited to *NET-seq*, where nascent mRNA chunks associated with immunoprecipitated PolII complexes are isolated and sequenced [26], *GRO-seq*, where RNAs recently transcribed only by transcriptionally engaged PolIIs are sequenced [27], and *PRO-seq*, which is similar to GRO-seq but reaches single-nucleotide resolution [28]. The evidences of PolII transport are particularly clear in time-course experiments, where sequencing data are collected over time following a perturbation. As an example, time-variant PRO-seq has been suggested to estimate pausing times in key peak regions [29]. A classical way to perturb these molecular dynamics is inhibiting the initiation by treating the cells with triptolide (Trp), which is a highly specific drug that blocks initiation [22,30]. This permits the PolII already engaged in transcription to progress further downstream the gene while new PolIIs are prevented from attaching, thus freeing upstream genomic regions as the run-on time progresses (figure 1*c*). Our approach consists of using the read profiles y as functions of x, collected at fixed run-on times t_1, t_2, t_3 and t_4 after treatment, to infer the dynamics. While Trp inhibits new initiation, poised PolII upstream of the TSS can still pass through it, enter the gene template and perform elongation immediately after Trp treatment [22,30]. To account for this, we also perform inference over Spt5 ChIP-seq data, where the poised polymerases are masked while those bound are detected [22].

These types of experiments reveal the presence of a flux of PolIIs, which is the signature of the non-equilibrium physics involved in the elongation process. The profile y^* observed prior to the treatment corresponds to a non-equilibrium stationary state (NESS). Disrupting initiation with Trp yields a transient state, which evolves from y^* until it settles down to a new NESS.

2.2. Mathematical model

The transport of particles on a one-dimensional lattice is a well-studied problem in mathematics and physics. Its basic features are captured by the ASEP [14], which defines the stochastic dynamics of interacting particles on a discrete lattice, which we take here to be a one-dimensional chain with open boundary conditions. Let the total number of lattice sites be N. The state of each site i, $1 \le i \le N$, is characterized by the occupation number n_i such that $n_i = 0$ if the site is empty and $n_i = 1$ if it is occupied by a particle. The evolution proceeds in continuous time. A particle on site $i < N$ hops rightward into the site $i+1$ with rate p_i, the transition being successful only if the site $i+1$ is empty. Similarly, a particle on site $i > 1$ hops leftward into $i-1$ with rate q_i, if the site $i-1$ is empty. Further, particles on the left (right) boundary site $i = 1$ ($i = N$) leave the lattice at rate q_1 (p_N), while particles are injected in the same boundary site at rate p_0 (q_{N+1}) if the site is empty. The constraint that a jump can occur only if the target is empty prevents the accumulation of more than one particle on a site and is generically referred to as the exclusion rule. This rule allows particle collision, which causes congestion when the particle density is sufficiently high and permits phase transitions between a low density, high density and a maximum current phase, even if the systems is one-dimensional [31]. Interestingly, based on theoretical considerations, it has been suggested that traffic-like congestion of PolIIs is important in transcription [32–34].

While the ASEP was originally proposed to model biopolymerization on nucleic acid templates [12,13], this and related models have been more recently applied to diverse problems, including protein translation [35–37], but also e.g. molecular motors [38] and pedestrian and vehicle traffic [39]. Applications to transcription incorporating disordered dynamics and obstacles (e.g. [40,41]) were also proposed. ASEP's theoretical appeal is due to its analytical results representative of a large class of models [42,43] and a convenient mean-field treatment that yields the exact stationary solution [44]. In the context of transcription, particles entering site 1, moving along the chain and exiting from site N correspond to initiation, elongation and termination, respectively. In our setting, the lowest values of p_i correspond to genomic locations were elongation slows down.

The dynamics of the expected occupation of a single site i in the bulk are governed by the lattice continuity equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}(n_i(t)) = J^{\mathrm{left}}(t) - J^{\mathrm{right}}(t), \tag{2.1}$$

$0 < i < N$, where $\mathbb{E}$ denotes expectation value and $J^{\mathrm{left}}(t)$ and $J^{\mathrm{right}}(t)$ are the average flux of particles from site $i-1$ to site i and from site i to site $i+1$, respectively. These are subject to the exclusion rule and therefore obey

$$\begin{aligned} J^{\mathrm{left}}(t) &= p_{i-1}\mathbb{E}(n_{i-1}(t)(1-n_i(t))) - q_i\mathbb{E}(n_i(t)(1-n_{i-1}(t))) \\ \text{and} \qquad J^{\mathrm{right}}(t) &= p_i\mathbb{E}(n_i(t)(1-n_{i+1}(t))) - q_{i+1}\mathbb{E}(n_{i+1}(t)(1-n_i(t))). \end{aligned} \tag{2.2}$$

In order to exactly solve these dynamics, second-order moments such as $\mathbb{E}(n_i(t)n_{i+1}(t))$ need to be known. Under independence assumption, these moments are factorized, which in our case amounts to replacing equations (2.1) and (2.2) with

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi_i(t) = {} & p_{i-1}\phi_{i-1}(t)(1-\phi_i(t)) - p_i\phi_i(t)(1-\phi_{i+1}(t)) \\ & + q_{i+1}\phi_{i+1}(t)(1-\phi_i(t)) - q_i\phi_i(t)(1-\phi_{i-1}(t)), \end{aligned} \tag{2.3}$$

where we used $\phi_i(t) := \mathbb{E}(n_i(t))$ to lighten the notation. In other words, equations (2.3) define the so-called mean-field dynamics of the asymmetric exclusion process, which are known to approximate well the true dynamics in many contexts, predict crucial features such as dynamical phase-transitions, and ease mathematical treatment [31,44,45]. With open boundaries,

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_1(t) = p_0(1-\phi_1(t)) - p_1\phi_1(t)(1-\phi_2(t)) - q_1\phi_1(t) + q_2\phi_2(t)(1-\phi_1(t)), \tag{2.4}$$

and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi_N(t) = {} & p_{N-1}\phi_{N-1}(t)(1-\phi_N(t)) \\ & - p_N\phi_N(t) + q_{N+1}(1-\phi_N(t)) - q_N\phi_N(1-\phi_{N-1}). \end{aligned} \tag{2.5}$$

To match the available data that is coarse grained (figure 1*b*), instead of considering particles individually we rely on their hydrodynamics description, which is obtained as follows. We assume Euler scaling with constant a and let $a \to 0$, $N \to \infty$, with $L := N\,a$ held finite. We define the functions $\varrho : \mathbb{R}^2 \to \mathbb{R}_0^+$, $\tilde{p} : \mathbb{R} \to \mathbb{R}_0^+$ and $\tilde{q} : \mathbb{R} \to \mathbb{R}_0^+$ such that they are analytic and bounded on $]0, L[\times]0, \infty[$, $]0, L[$ and $]0, L[$, respectively, and

$$\begin{aligned} \phi_i(t) &= \varrho((i-1)a, t), \\ ap_i &= \tilde{p}((i-1)a) \\ \text{and} \qquad aq_i &= \tilde{q}((i-1)a). \end{aligned} \tag{2.6}$$

We further assume that the left and right jump rates satisfy $\tilde{q}(x) = b\,\tilde{p}(x)$, $\forall x \in [0, L]$, with $0 \le b < 1$, where b governs the relative strength of the non-equilibrium driving forces. The case $b = 0$ corresponds to a *totally asymmetric exclusion process* (TASEP), while the limit case $b = 1$ corresponds to the *symmetric exclusion process*. Intermediate values $0 < b < 1$ correspond to settings where the particles can jump in both directions, but are driven rightwards on average. A continuum-limit counterpart of equations (2.2), as derived in [16,18], is

$$\begin{aligned} J(x, t) = \tilde{p}(x)\varrho\Big(x - \frac{a}{2}, t\Big)\Big(1 - \varrho\Big(x + \frac{a}{2}, t\Big)\Big) \\ - \tilde{q}(x)\varrho\Big(x + \frac{a}{2}, t\Big)\Big(1 - \varrho\Big(x - \frac{a}{2}, t\Big)\Big), \end{aligned} \tag{2.7}$$

which, using first-order Taylor expansion, yields

$$J(x, t) \approx (\tilde{p}(x) - \tilde{q}(x))\varrho(x, t)(1 - \varrho(x, t)) - \frac{a}{2}(\tilde{p}(x) + \tilde{q}(x))\frac{\partial}{\partial x}\varrho(x, t). \tag{2.8}$$

To lighten the mathematical notation, we define the two quantities

$$\begin{aligned} \lambda(x) &:= (\tilde{p}(x) - \tilde{q}(x)) = \tilde{p}(x)(1 - b) \\ \text{and} \qquad \nu(x) &:= \frac{a}{2}(\tilde{p}(x) + \tilde{q}(x)) = \frac{a}{2}\tilde{p}(x)(1 + b) \end{aligned} \tag{2.9}$$

their ratio is constant in x, viz., $\nu(x)/\lambda(x) = a/2\,(1 + b)/(1 - b)$, which equals $a/2$ in the totally asymmetric case.

Substituting (2.8)–(2.9) into the continuity equation

$$\frac{\partial}{\partial t}\varrho(x, t) = -\frac{\partial}{\partial x}J(x, t), \tag{2.10}$$

which is the hydrodynamics limit of equation (2.1), gives the nonlinear partial differential equation

$$\frac{\partial}{\partial t}\varrho(x, t) = -\frac{\partial}{\partial x}\Big\{\lambda(x)\varrho(x, t)(1 - \varrho(x, t)) - \nu(x)\frac{\partial}{\partial x}\varrho(x, t)\Big\}, \tag{2.11}$$

which can be linearized to

$$\frac{\partial}{\partial t}u(x, t) = \frac{\tilde{p}(x)}{2}\left\{a^2(1 + b)\frac{\partial^2}{\partial^2 x}u(x, t) - \frac{(1 - b)^2}{1 + b}u(x, t)\right\} \tag{2.12}$$

by means of a generalization of the Cole–Hopf transform (electronic supplementary material, appendix A and [18,46,47]).

In transcription, the particle flux is left to right. While PolIIs can backtrack few Nts under certain circumstances [48–50], this phenomenon is overall minor and is not observable at our ChIP-seq resolution. Therefore, we assume $b = 0$ and focus on the inference of the net forward rate profile $\tilde{p}(x)$. For simplicity we also set $a = 1$, arguing that our considerations remain valid with such a choice. The required boundary values $\varrho(0, t)$, $\varrho(L, t)$ and $\varrho(x, 0)$, and the numerical scheme used to integrate equation (2.12) are detailed in electronic supplementary material, appendices A and B.

Integrating equation (2.11) with boundary conditions analogous to equations (2.4) and (2.5) and initial density $\varrho(x, 0) > 0$ yields a NESS for large t, characterized by a non-vanishing average flux and a density profile $\varrho^*(x)$ which is invariant in time. Setting the latter as initial condition and further integrating with no inward particle flux ($p_0 = q_N = 0$) produces a transient state that mimics the evolution of the PolII profile after Trp treatment until the density profile vanishes. This is illustrated,

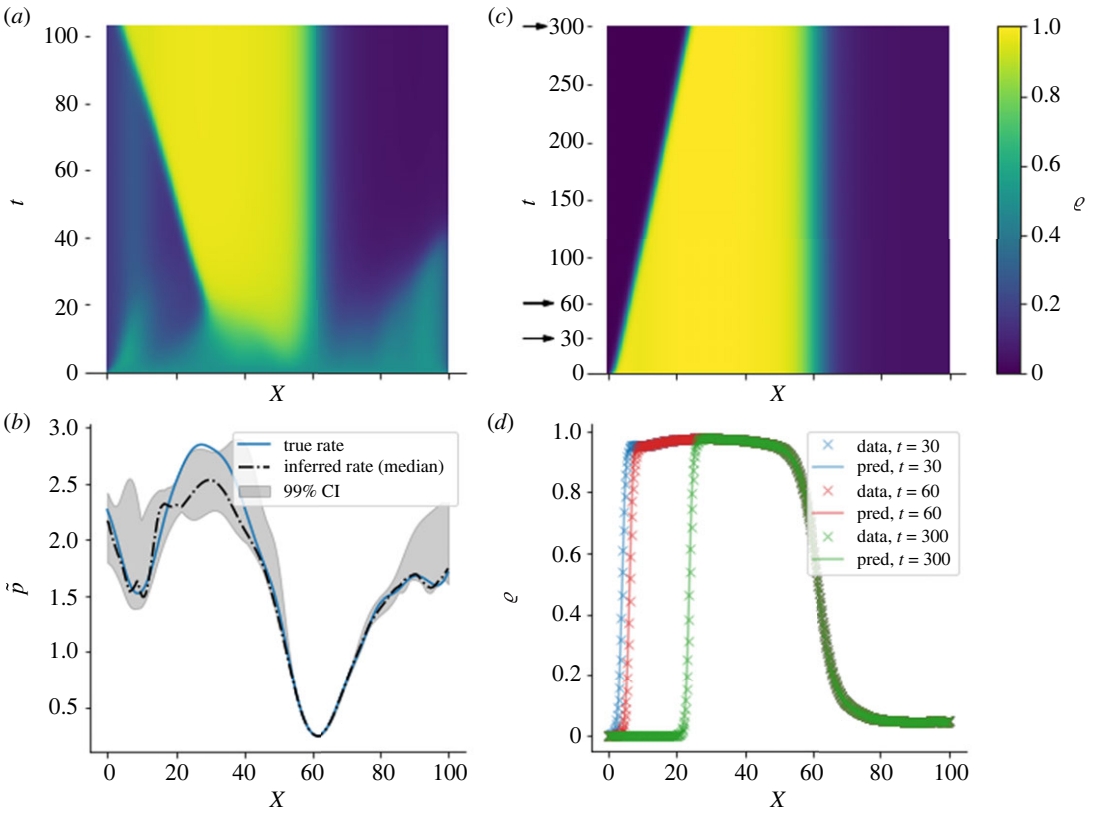

Figure 2. Simulation study. (*a*) A non-equilibrium stationary state (NESS) of profile $\varrho^*(x)$ is obtained integrating the hydrodynamic TASEP with open boundaries, initial density profile $\varrho(x, 0) = 0.5\ \forall x \in [0, L]$ and chosen rate profile (solid line in (*b*)). (*b*) True rate profile (solid line) and inferred rate profile (dash-dot line); the shaded area is 99% credible interval (CI). (*c*) Integrating the same dynamics with initial profile $\varrho^*(x)$ and no-influx boundary conditions shows that the density decreases in proximity of the left boundary, similar to ChIP-seq readings followed by Trp treatment; the density profiles corresponding to times 30, 60 and 300 and used for inference are marked by arrows. (*d*) The posterior predictive samples (solid lines) are in excellent agreement with the extracted density profiles (cross markers); the posterior predictive dispersion is of the order of the line width, see also electronic supplementary material, figure S3

for a choice of boundary values and jump rate profile, in figure 2, which also includes the result of the inference process described in the next sections.

2.3. Bayesian framework

We fit the model to real-world data by means of a Bayesian approach leveraging its ability to explicitly encode prior hypotheses about the quantities we wish to infer [51]. We are interested in the forward rate profile $\tilde{p}$. As this is required to be analytic and non-negative, it is convenient to assume a GP [19] functional prior on a latent variable f and induce a prior on $\tilde{p}$ using a sigmoid link function of f, such that $\tilde{p} = \tilde{p}_{\max}/(1 + \exp(-f))$, which further imposes an upper bound $\tilde{p}_{\max}$ to $\tilde{p}$. The GP prior here defines a distribution over real valued C^1 functions in $\mathbb{R}$, where any finite set of function evaluations $f(x)$ has multivariate normal distribution with mean m and covariance kernel $k(x, x'; \sigma_f^2, l) = \sigma_f^2 \exp(-(x - x')^2/(2l^2))$, $x, x' \in \mathbb{R}$. In practice, the GP is evaluated at the positions x_i, $i = 1, \ldots, n$, where it is equivalent by definition to $\mathbf{f} \sim \mathcal{N}(m, \mathbf{K})$, a multivariate normal random variable with mean m and covariance matrix $\mathbf{K}(\sigma_f, l)$ induced by the kernel.

The observations are organized into a collection of values $\mathbf{y} = \{y_{ij}\}_{i=1,2,\ldots,n, j=1,2,\ldots,t}$, where the subscripts indicate that an observation is taken at position x_i and time t_j. As the values of x_i do not necessarily coincide with the bin centres of ChIP-seq data, we used simple linear interpolation to estimate the data at intermediate coordinates. We assume that the observed values depend on a multiplicative factor and also include an additive error term $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. This can be written in terms of the equation $\kappa y_{ij} = \varrho(x_i, t_j) + \epsilon$, where κ is the inverse of the amplification factor. The

likelihood $P(\mathbf{y}|\mathbf{f}, \sigma_\epsilon, \kappa)$ satisfies

$$\log P(\mathbf{y}|\mathbf{f}, \sigma_\epsilon, \kappa) = -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\sum_{j=1}^{t}(\varrho(x_i, t_j; \mathbf{f}) - \kappa y_{ij}) - \frac{n}{2}\log(2\pi\sigma_\epsilon^2), \tag{2.13}$$

where we made explicit that ϱ depends on $\mathbf{f}$. For the hierarchical parameters $(m, \sigma_\epsilon, \kappa, \sigma_f, l) =: \theta$ we assume a scaled sigmoid Gaussian prior probability $P(m, \sigma_\epsilon, \kappa, \sigma_f, l)$ such that

$$\theta = \theta_{\min} + \frac{(\theta_{\max} - \theta_{\min})}{1 + \exp(-\xi)}, \quad \xi \sim \mathcal{N}(\mu_\xi, \mathbb{1}\sigma_\xi), \tag{2.14}$$

where $\theta_{\min} := (m_{\min}, \sigma_{\epsilon\min}, \kappa_{\min}, \sigma_{f\,\min}, l_{\min})$, $\theta_{\max} := (m_{\max}, \sigma_{\epsilon\max}, \kappa_{\max}, \sigma_{f\,\max}, l_{\max})$ and (μ_ξ, σ_ξ) are referred to as hyperparameters. Prior distributions are chosen to pull Markov chain Monte Carlo (MCMC) samples away from inappropriate results that are consistent with the likelihood but would not be consistent with domain knowledge [51]. By using scaled sigmoid Gaussian prior probability bounded by $\theta_{\min}$ and $\theta_{\max}$, we only search for solutions constrained in an appropriate interval [52]. By virtue of the Bayes theorem the joint posterior probability for θ and $\mathbf{f}$ satisfies

$$P(\mathbf{f}, m, \sigma_\epsilon, \kappa, \sigma_f, l|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{f}, \sigma_\epsilon, \kappa)P(\mathbf{f}|m, \sigma_f, l)P(m)P(\sigma_\epsilon)P(\kappa)P(\sigma_f)P(l), \tag{2.15}$$

which we draw random samples from by MCMC sampling, more specifically block Gibbs sampling with elliptical slice sampling at each block [52,53] (electronic supplementary material, appendix C). Equation (2.15) expresses the distribution of parameters given the observed data $\mathbf{y}$ and completes the definition of the model. It is worth noting that evaluating the likelihood also requires computing ϱ by integrating equation (2.11) with initial condition $\varrho(x, 0) = \kappa y^*(x)$, $\forall x \in [0, L]$.

3. Results

We first consider simulated data from a given profile of length $L = 100$ obtained from GP drawn with parameters $(l, \sigma_f, m, \tilde{p}_{\max}) = (7.32, 0.67, 0.29, 3)$. We integrate the dynamics with NESS initial profile (obtained by fixing the boundary conditions to $\varrho(0, t) = \varrho(L, t) = 0.5$, $\forall t$) and no-influx boundary conditions (figure 2*a*,*b*). The chosen rate profile shows a local minimum close to the left boundary, which yields a minor local perturbation in the density, and a global minimum around $x \approx 60$, whose effect propagates along the lattice and acts as a major bottleneck, which separates a low-density phase downstream from a high-density phase upstream. These minima correspond to regions where particles slow down or pause for an exponentially distributed amount of time. As the particles leave the system through the right boundary and are not replenished by the influx through the left boundary, the region upstream of the bottleneck is emptied by a reverse wavefront.

For the purpose of testing whether we are able to recover the rate profile from time-course observations, we extract density profiles $\mathbf{y}$ at times $(t_1, t_2, t_3) = (30, 60, 300)$ and set the hyperparameters $\theta_{\min}$, $\theta_{\max}$ and (μ_ξ, σ_ξ) to (0, 0, 0.8, 0, 0), (2, 10, 1.2, 1, 10) and (0, 1), respectively. With these settings and data, we generated 10^4 MCMC samples targeting the posterior (2.15), discarding the first 2×10^3 as burn-in, demonstrating that the fitting procedure is able to capture the location of both the major and minor minima of the generative model, as well as the overall elongation rate (figure 2*b*). It is worth noting that the integrated density profile in figure 2*c*,*d* displays a very small effect of the first local minimum (minor dip, captured only by time-course profiles at t_1 and t_2); this is reflected in relatively wide credible intervals for the inferred rate profile (grey ribbon in figure 2*b*). On the other hand, the rate at the bottleneck is inferred with very high confidence. The covariance hyperparameters l and σ_f control how quickly the rate changes over x; these were slightly misestimated to 6.86 (95% CI 4.63–7.16) and 0.76 (95% CI 0.75–0.82), respectively, thus suggesting that increased wobbling in the rate profile is tolerated; minor patterns in the rate profile are in fact smoothed out and are essentially not identifiable in the density profiles obtained by integration (see electronic supplementary material, figure S3). The difficulty of sampling covariance hyperparameters is also addressed, e.g. in [52]. The predicted transient density profiles at $t = 30$, 60, 300 also are in very good agreement with the input data (figure 2*d*; in fact, all sampled rate profiles yield similar time-course density profiles despite wide CIs in certain regions (see also electronic supplementary material, figure S3).

Applying this method to real-world data requires setting the value of $\tilde{p}_{\max}$ to an upper limit of prior expectations on the elongation rate. As this has been estimated at around 2×10^3 Nt min^{-1} in previous studies [30], we set $\tilde{p}_{\max} = 6 \times 10^3$ Nt min^{-1} as an arguably safe upper bound. Literature results can

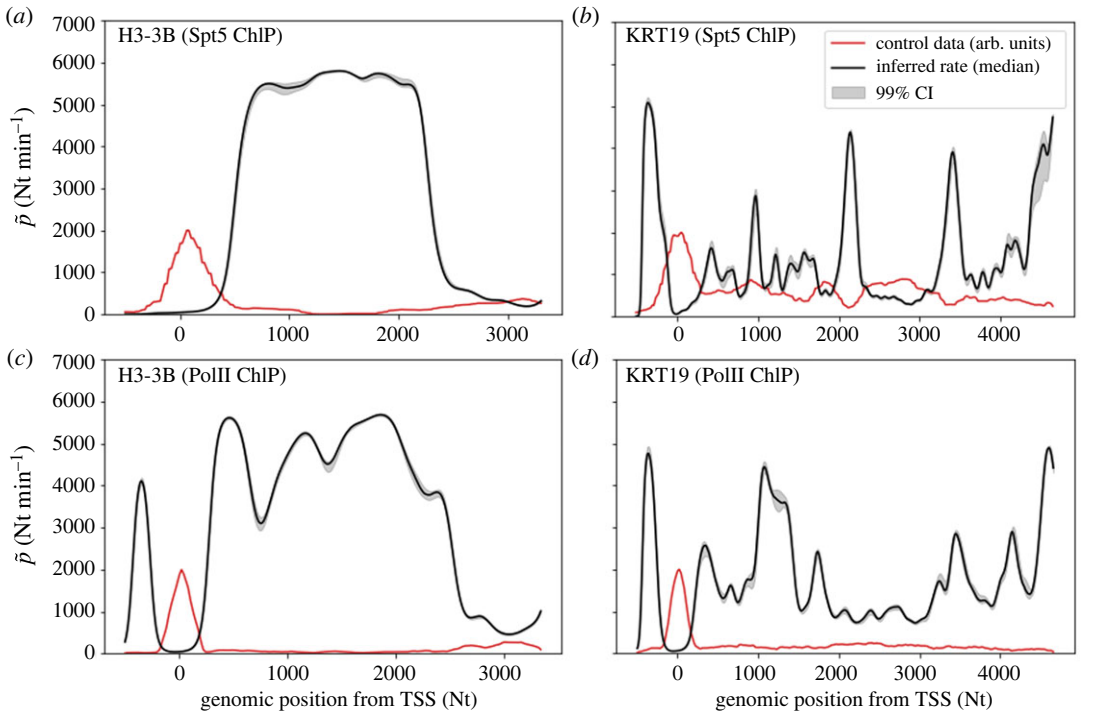

Figure 3. Inferred rate profiles from Spt5 ChIP-seq (*a*,*b*) and PolII ChIP-seq (*c*,*d*) for genes *H3-B3* (*a*,*c*) and *KRT19* (*b*,*d*) (black lines are posterior medians, shaded areas are 99% credible intervals), with the latter gene showing a distinctive jagged profiles. Both genes show the lowest rates in proximity of the transcription starting site (TSS). Red lines are unperturbed ChIP-seq signals in arbitrary units (arb. units).

be also used to set bounds on the prior for κ, which regularizes the estimation problem [19]. From cultured human cell lines, the total number P of bound PolII molecules per cell is estimated to be between $P_{min} = 11 \times 10^5$ and $P_{max} = 18 \times 10^5$ [54]. This is related to the total number Y of ChIP-seq counts by $P = \kappa Y$. Based on these heuristic considerations, we set $\kappa_{min} = P_{min}/Y$ and $\kappa_{max} = P_{max}/Y$. All remaining hyperparameters were set identical to the previous simulation experiment.

The results from different genes show a variety of rate profiles which share similar patterns (figure 3). The most important observation is that, in all genes considered, the rates vary strongly with the genomic position, with local minima corresponding to regions where PolIIs slow down or pause. In order to look for average patterns, it is desirable to aggregate data from all genes. As genes have different lengths (which in our sample range from 16 680 to 59 880 Nts), we stretch all the rate profiles in the region from TSS + 1000 to TES – 1000 Nts to the same support length and then average over the genes at each position. This yields the summaries illustrated in figure 4, which we refer to as *metagene* rates and are akin to the so-called metagene profiles [55]. Rates are typically lower near the TSS than in the gene body, where elongation approaches its highest rate. The behaviour in proximity of the TES is less definite, with rates varying several fold among the different genes. At the TSS the rate typically dips down consistently with the presence of strong and widespread pausing in this region. Further downstream in the gene body the rate increases to its highest average value. While the dip is evident in both Spt5 and PolII results, it is worth noting that upstream of the TSS the average rate inferred from PolII data is higher than that from Spt5. We argue that this difference is due to the fact the former also include poised PolIIs which are not strongly bound to the template and can quickly move towards the TSS before being engaged in transcription. A by-product of the fitting procedure is the estimate of the occupation density $\varrho(x, t) = \kappa y(x, t)$, as illustrated in figure 2*d* for the simulation experiment and figure 5 and electronic supplementary material, figures S4–S6 for selected genes. The predicted densities are typically very low (total predicted number of PolIIs in a gene is of the order of 10^{-1}), thus suggesting that crowding and congestion of PolIIs into a gene might not be substantial even proximal to rate minima.

4. Discussion

We developed a general Bayesian framework to study the dynamics of a one-dimensional transport model given time-resolved density profiles. The general problem addressed here is the identification

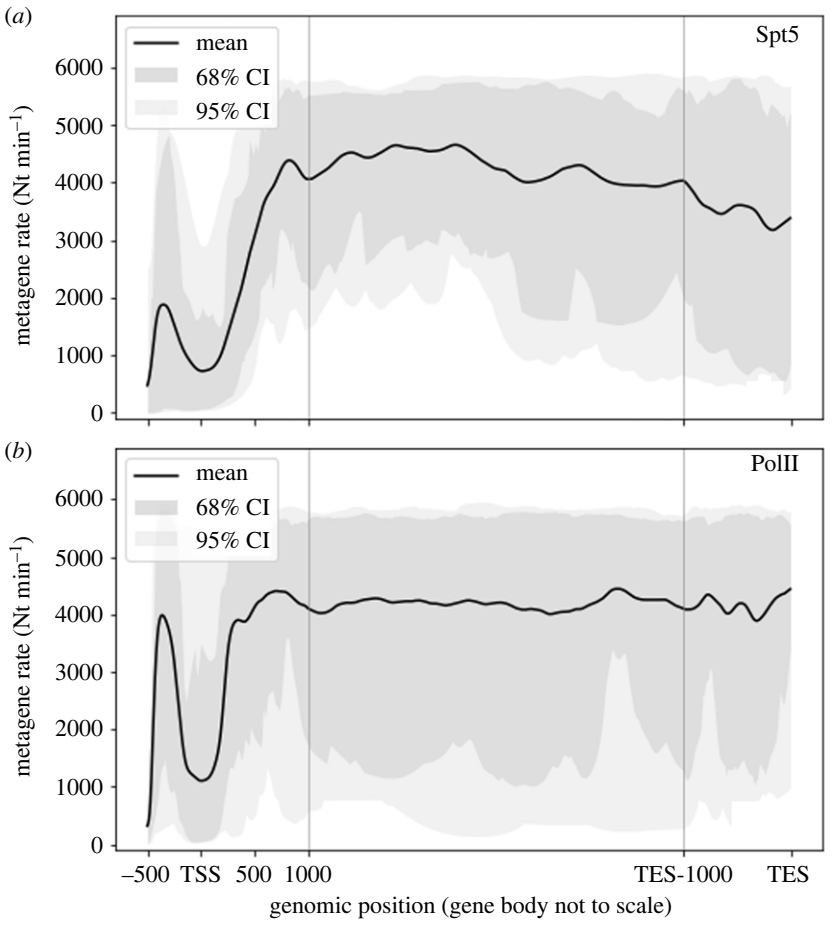

Figure 4. Metagene rates from PolII ChIP-seq (*a*) and Spt5 ChIP-seq (*b*) data. By construction, the metagene analysis conserves the length scale only in proximity of TSS and TES. Upstream of TSS the Spt5 ChIP-seq yields lower average rate than PolII ChIP-seq, as this assay does not detect PolII poised to move downstream.

of the PDE parameters that best describe data as a subset of the true PDE solution (see, e.g. [52,56–58] and references therein). We focused on the hydrodynamic TASEP with smoothly varying jump rates (which are the parameters to be inferred) as a paradigmatic and well-characterized model of transport. By means of its application to ChIP-seq time-course data, we inferred the rate of PolII elongation as a function of the genomic position in selected genes. This rate is not constant but varies within the gene body. It typically dips down nearby the TSS, confirming widespread pausing in this region, while in the bulk the rate also varies between genes. Low predicted densities suggest that the pausing did not cause congestion or crowding. This is an important observation, as factor crowding has been experimentally observed and associated with regulated gene expression in synthetic and mammalian cell systems [59–63]. Our analysis supports the view that this phenomenon does not happen between PolIIs bound to the gene but probably occurs in suspension in the nucleoplasm, as described e.g. in [61,63–67].

The inference here is complicated by the high dimensionality of the parameter space (which grows as the genes' length increases). We addressed this by assuming a GP latent prior for the jump-rate profile and using elliptic slice sampling as an appropriate MCMC algorithm. The sampling requires multiple evaluations of the likelihood of equation (2.13). This in turns requires numerically integrating equation (2.11), which is also slower in longer genes (require larger integration grids, see electronic supplementary material, appendix B).

This study of molecular dynamics is also subject to limitations. While ChIP-seq is a widely used assay to quantify the abundances of DNA-bound PolII, studies suggest that it has limited resolution (between 150 and 300 Nts) and might be subject to technical issues [23]. Most importantly ChIP-seq profiles are obtained from the aggregation of sequencing reads from many cells, which hides variation within the cell population. The transcription of mRNA is a very complex process and it may be interesting to include features not encoded in the model used for this study. Other TASEP variants, such as those incorporating non-Markovian jump dynamics [68,69] or Langmuir kinetics

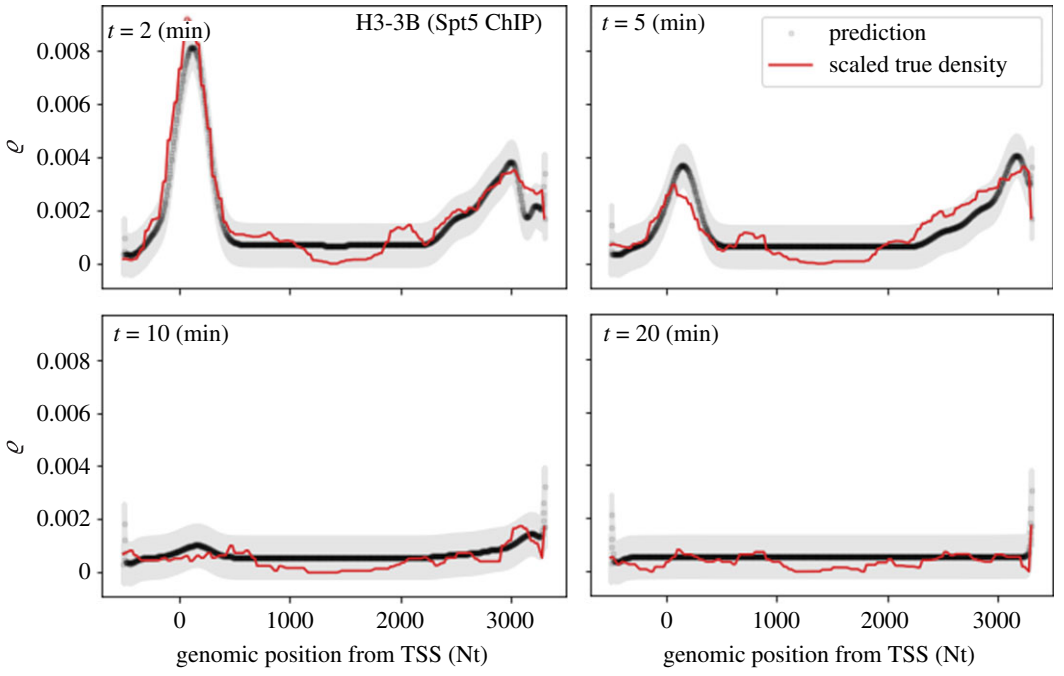

Figure 5. Predicted density profiles for gene *H3-3B* from Spt5 ChIP-seq 2, 5, 10 and 20 min after Trp treatment as samples from posterior predictive distribution. Grey area is 95% credible interval due to noise model.

[70], are relevant for the modelling of PolII recycling and its early detachment from DNA [63,71]. An assumption of the TASEP is that particles stay in a site for an exponentially distributed waiting time. Variants of TASEP in which defects appear and disappear randomly on any site (and thus slow down the movement of particles or even block it completely) have been introduced in physics literature and can account for occasionally long pausing times [40,41] (see also [68,69]), with defect dynamics representing the effects of pausing and elongation factors. Modelling advancements that combine the site-specific pausing with long pausing times and extended particle size, supplemented by an appropriate inference scheme such as the one presented here, would be an important additional potential area for research and application. Potential extensions of our work also include estimation of the parameters that encode the system's size and asymmetry (a and b, respectively) and the boundary values. Statistical mechanics literature is rich in quantitative studies of TASEPs with particles that occupy more than one lattice site [72–74], some generalized to include site-dependent elongation rates or localized defects [37,75–77]. These studies have been used to describe protein translation and could be useful to predict PolII-size effects in gene expression, although, with genes much longer than PolIIs, ChIP-seq limited resolution, and very low PolII coverage density, the observable correction would arguably be minor. In fact, including more features plausibly requires sequencing assays of higher resolution than ChIP-seq and comes at the cost of increased computational burden and decreased tractability. Conversely, the chosen TASEP with smoothly varying jump rates is simple and yet is able to reveal PolII elongation slowing down and speeding up at certain genomic locations. Due to its generality, our approach also serves as a template for future studies seeking to shed light on complex transport phenomena.

5. Material and methods

Spt5 and PolII ChIP-seq data mapped to the hg19 University of California at Santa Cruz human genome were downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo), accession number GSE117006. We filtered the list of genes from the reference genome to only contain those with unique gene symbols on chromosomes 1–22 and X, thus excluding alternatively spliced genes. Hg19 gene coordinates were flanked 500 Nts upstream of the TSS in order to include poised PolII. The 20 non-overlapping genes with the highest coverage of Spt5 ChIP-seq reads were selected. All simulation codes are written in c++ and Python (v. 3.7.1), with the PDE solver using Numba JIT compiler (v. 0.41.0) [78] (https://github.com/mcavallaro/dTASEP-fit).

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Supplementary material is available online [79].

Authors' contributions. M.C.: conceptualization, data curation, formal analysis, investigation, software, visualization, writing—original draft, writing—review and editing; Y.W.: software; D.H.: conceptualization, data curation, writing—review and editing; R.D.: conceptualization, formal analysis, software, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research used WISB computational facilities (grant ref: no. BB/M017982/1) funded under the UK Research Councils' Synthetic Biology for Growth programme. R.D. is funded by EPSRC (grant nos. EP/V025899/1 and EP/T017112/1) and NERC (grant no. NE/T00973X/1). M.C. acknowledges support from Matt J. Keeling and Health Data Research UK, which is funded by the UK Medical Research Council, EPSRC, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and the Wellcome Trust. D.H. is funded by EPSRC (grant no. EP/T002794/1). We thank Carlo Albert and Jie Zhang for valuable comments and the Warwick Bioinformatics RTP for sharing computational resources.

References

1. Munsky B, Neuert G, van Oudenaarden A. 2012 Using gene expression noise to understand gene regulation. *Science* **336**, 183–187. (doi:10.1126/science.1216379)
2. Rajala T, Häkkinen A, Healy S, Yli-Harja O, Ribeiro AS. 2010 Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput. Biol.* **6**, 1–12. (doi:10.1371/journal.pcbi.1000704)
3. Jonkers I, Lis JT. 2015 Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177. (doi:10.1038/nrm3953)
4. Mayer A, Landry HM, Churchman LS. 2017 Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr. Opin. Cell Biol.* **46**, 72–80. (doi:10.1016/j.ceb.2017.03.002)
5. Adelman K, Lis JT. 2012 Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731. (doi:10.1038/nrg3293)
6. Liu X, Kraus WL, Bai X. 2015 Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem. Sci.* **40**, 516–525. (doi:10.1016/j.tibs.2015.07.003)
7. Peccoud J, Ycart B. 1995 Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234. (doi:10.1006/tpbi.1995.1027)
8. Tripathi T, Chowdhury D. 2008 Interacting RNA polymerase motors on a DNA track: effects of traffic congestion and intrinsic noise on RNA synthesis. *Phys. Rev. E* **77**, 011921. (doi:10.1103/PhysRevE.77.011921)
9. Dobrzynski M, Bruggeman FJ. 2009 Elongation dynamics shape bursty transcription and translation. *Proc. Natl Acad. Sci. USA* **106**, 2583–2588. (doi:10.1073/pnas.0803507106)
10. Cao Z, Filatova T, Oyarzún DA, Grima R. 2020 A stochastic model of gene expression with polymerase recruitment and pause release. *Biophys. J.* **119**, 1002–1014. (doi:10.1016/j.bpj.2020.07.020)
11. Szavits-Nossan J, Grima R. 2022 Steady-state distributions of nascent RNA for general initiation mechanisms. *bioRxiv*. (doi:10.1101/2022.03.30.486441).
12. MacDonald CT, Gibbs JH, Pipkin AC. 1968 Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* **6**, 1–25. (doi:10.1002/bip.1968.360060102)
13. MacDonald CT, Gibbs JH. 1969 Concerning the kinetics of polypeptide synthesis on polyribosomes. *Biopolymers* **7**, 707–725. (doi:10.1002/bip.1969.360070508)
14. Spitzer F. 1970 Interaction of Markov processes. *Adv. Math.* **5**, 246–290. (doi:10.1016/0001-8708(70)90034-4)
15. Benassi A, Fouque JP. 1987 Hydrodynamical limit for the asymmetric simple exclusion process. *Ann. Probab.* **15**, 546–560. (doi:10.1214/aop/1176992158)
16. Stinchcombe RB, De Queiroz SL. 2011 Smoothly varying hopping rates in driven flow with exclusion. *Phys. Rev. E* **83**, 1–12. (doi:10.1103/PhysRevE.83.061113)
17. Lakatos G, O'Brien J, Chou T. 2006 Hydrodynamic mean-field solutions of 1D exclusion processes with spatially varying hopping rates. *J. Phys. A Math. Gen.* **39**, 2253–2264. (doi:10.1088/0305-4470/39/10/002)
18. Harris RJ, Stinchcombe RB. 2004 Disordered asymmetric simple exclusion process: mean-field treatment. *Phys. Rev. E* **70**, 016108. (doi:10.1103/PhysRevE.70.016108)
19. Rasmussen CE, Williams CKI. 2006 *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
20. wa Maina C, Honkela A, Matarese F, Grote K, Stunnenberg HG, Reid G, Lawrence ND, Rattray M. 2014 Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data. *PLoS Comput. Biol.* **10**, 1–17. (doi:10.1371/journal.pcbi.1003598)
21. Honkela A. 2015 Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proc. Natl Acad. Sci. USA* **112**, 13 115–13 120. (doi:10.1073/pnas.1420404112)
22. Erickson B, Sheridan RM, Cortazar M, Bentley DL. 2018 Dynamic turnover of paused Pol II complexes at human promoters. *Genes Dev.* **32**, 1215–1225. (doi:10.1101/gad.316810.118)
23. Park PJ. 2009 ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680. (doi:10.1038/nrg2641)
24. Ehara H, Yokoyama T, Shigematsu H, Yokoyama S, Shirouzu M, Sekine S. 2017 Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science* **357**, 921–924. (doi:10.1126/science.aan8552)
25. Hu B, Petela N, Kurze A, Chan KL, Chapard C, Nasmyth K. 2015 Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. *Nucl. Acids Res.* **43**, e132. (doi:10.1093/nar/gkv670)
26. Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015 Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540. (doi:10.1016/j.cell.2015.03.027)
27. Core LJ, Waterfall JJ, Lis JT. 2008 Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848. (doi:10.1126/science.1162228)
28. Kwak H, Fuda NJ, Core LJ, Lis JT. 2013 Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953. (doi:10.1126/science.1229386)
29. Zhang J, Cavallaro M, Hebenstreit D. 2021 Timing RNA polymerase pausing with TV-PRO-seq. *Cell Rep. Methods* **1**, 100083. (doi:10.1016/j.crmeth.2021.100083)
30. Jonkers I, Kwak H, Lis JT. 2014 Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **2014**, 1–25. (doi:10.7554/eLife.02407)
31. Chou T, Mallick K, Zia RKP. 2011 Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Rep. Prog. Phys.* **74**, 116601. (doi:10.1088/0034-4885/74/11/116601)

32. Klumpp S, Hwa T. 2008 Stochasticity and traffic jams in the transcription of ribosomal RNA: intriguing role of termination and antitermination. *Proc. Natl Acad. Sci. USA* **105**, 18 159–18 164. (doi:10.1073/PNAS.0806084105)
33. Cholewa-Waclaw J *et al.* 2019 Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proc. Natl Acad. Sci. USA* **116**, 14 995–15 000. (doi:10.1073/pnas.1903549116)
34. Tripathi T, Schütz GM, Chowdhury D. 2009 RNA polymerase motors: dwell time distribution, velocity and dynamical phases. *J. Stat. Mech. Theory Exp.* **2009**, P08018. (doi:10.1088/1742-5468/2009/08/P08018)
35. Zia RKP, Dong JJ, Schmittmann B. 2011 Modeling translation in protein synthesis with TASEP: a tutorial and recent developments. *J. Stat. Phys.* **144**, 405–428. (doi:10.1007/s10955-011-0183-1)
36. Szavits-Nossan J, Ciandrini L, Romano MC. 2018 Deciphering mRNA sequence determinants of protein production rate. *Phys. Rev. Lett.* **120**, 128101. (doi:10.1103/PhysRevLett.120.128101)
37. Erdmann-Pham DD, Dao Duc K, Song YS. 2020 The key parameters that govern translation efficiency. *Cell Syst.* **10**, 183–192.e6. (doi:10.1016/j.cels.2019.12.003)
38. Lipowsky R, Klumpp S, Nieuwenhuizen TM. 2001 Random walks of cytoskeletal motors in open and closed compartments. *Phys. Rev. Lett.* **87**, 108101. (doi:10.1103/PhysRevLett.87.108101)
39. Chowdhury D, Santen L, Schadschneider A. 2000 Statistical physics of vehicular traffic and some related systems. *Phys. Rep.* **329**, 199–329. (doi:10.1016/S0370-1573(99)00117-9)
40. Wang J, Pfeuty B, Thommen Q, Romano MC, Lefranc M. 2014 Minimal model of transcriptional elongation processes with pauses. *Phys. Rev. E* **90**, 050701. (doi:10.1103/PhysRevE.90.050701)
41. Waclaw B, Cholewa-Waclaw J, Greulich P. 2019 Totally asymmetric exclusion process with site-wise dynamic disorder. *J. Phys. A Math. Theor.* **52**, 065002. (doi:10.1088/1751-8121/aafb8a)
42. Kardar M, Parisi G, Zhang YC. 1986 Dynamic scaling of growing interfaces. *Phys. Rev. Lett.* **56**, 889–892. (doi:10.1103/PhysRevLett.56.889)
43. Bertini L, Giacomin G. 1997 Stochastic burgers and KPZ equations from particle systems. *Commun. Math. Phys.* **183**, 571–607. (doi:10.1007/s002200050044)
44. Derrida B, Evans MR, Hakim V, Pasquier V. 1993 Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *J. Phys. A Math. Gen.* **26**, 1493–1517. (doi:10.1088/0305-4470/26/7/011)
45. Lazarescu A. 2015 The physicist's companion to current fluctuations: one-dimensional bulk-driven lattice gases. *J. Phys. A Math. Theor.* **48**, 503001. (doi:10.1088/1751-8113/48/50/503001)
46. Hopf E. 1950 The partial differential equation $u_t + uu_x = \mu_{xx}$. *Commun. Pur. Appl. Math.* **3**, 201–230. (doi:10.1002/cpa.3160030302)
47. Cole JD. 1951 On a quasi-linear parabolic equation occurring in aerodynamics. *Q. Appl. Math.* **9**, 225–236. (doi:10.1090/qam/42889)
48. Nudler E, Mustaev A, Goldfarb A, Lukhtanov E. 1997 The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89**, 33–41. (doi:10.1016/S0092-8674(00)80180-4)
49. Jülicher F, Bruinsma R. 1998 Motion of RNA polymerase along DNA: a stochastic model. *Biophys. J.* **74**, 1169–1185. (doi:10.1016/S0006-3495(98)77833-6)
50. Wang D, Bushnell DA, Huang X, Westover KD, Levitt M, Kornberg RD. 2009 Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. *Science* **324**, 1203–1206. (doi:10.1126/science.1168729)
51. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013 *Bayesian data analysis*. New York, NY: CRC Press.
52. Tegnér M, Roberts S. 2019 A probabilistic approach to nonparametric local volatility. *arXiv*. (https://arxiv.org/abs/1901.06021)
53. Murray I, Adams R, MacKay D. 2010 Elliptical slice sampling. In *Proc. of the 13th Int. Conf. on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May*, pp. 541–548. PMLR.
54. Kimura H, Tao Y, Roeder RG, Cook PR. 1999 Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure. *Mol. Cell. Biol.* **19**, 5383–5392. (doi:10.1128/mcb.19.8.5383)
55. Joly Beauparlant C, Lamaze FC, Deschênes A, Samb R, Lemaçon A, Belleau P, Bilodeau S, Droit A. 2016 *metagene* profiles analyses reveal regulatory element's factor-specific recruitment patterns. *PLoS Comput. Biol.* **12**, 1–12. (doi:10.1371/journal.pcbi.1004751)
56. Rudy SH, Brunton SL, Proctor JL, Kutz JN. 2017 Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614. (doi:10.1126/sciadv.1602614)
57. Raissi M, Perdikaris P, Karniadakis G. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707. (doi:10.1016/J.JCP.2018.10.045)
58. Berg J, Nyström K. 2019 Data-driven discovery of PDEs in complex datasets. *J. Comput. Phys.* **384**, 239–252. (doi:10.1016/j.jcp.2019.01.036)
59. Tan C, Saurabh S, Bruchez MP, Schwartz R, Leduc P. 2013 Molecular crowding shapes gene expression in synthetic cellular nanosystems. *Nat. Nanotechnol.* **8**, 602–608. (doi:10.1038/nnano.2013.132)
60. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. 2017 A phase separation model for transcriptional control. *Cell* **169**, 13–23. (doi:10.1016/j.cell.2017.02.007)
61. Plys AJ, Kingston RE. 2018 Dynamic condensates activate transcription. *Science* **361**, 329–330. (doi:10.1126/science.aau4795)
62. Boija A *et al.* 2018 Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855.e16. (doi:10.1016/j.cell.2018.10.042)
63. Cavallaro M, Walsh MD, Jones M, Teahan J, Tiberi S, Finkenstädt B, Hebenstreit D. 2021 3′–5′ crosstalk contributes to transcriptional bursting. *Genome Biol.* **22**, 56. (doi:10.1186/s13059-020-02227-5)
64. Papantonis A, Cook PR. 2013 Transcription factories: genome organization and gene regulation. *Chem. Rev.* **113**, 8683–8705. (doi:10.1021/cr300513p)
65. Boehning M *et al.* 2018 RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.* **25**, 833–840. (doi:10.1038/s41594-018-0112-y)
66. Cramer P. 2019 Organization and regulation of gene transcription. *Nature* **573**, 45–54. (doi:10.1038/s41586-019-1517-4)
67. Wei MT, Chang YC, Shimobayashi SF, Shin Y, Strom AR, Brangwynne CP. 2020 Nucleated transcriptional condensates amplify gene expression. *Nat. Cell Biol.* **22**, 1187–1196. (doi:10.1038/s41556-020-00578-6)
68. Khoromskaia D, Harris RJ, Grosskinsky S. 2014 Dynamics of non-Markovian exclusion processes. *J. Stat. Mech. Theory Exp.* **2014**, P12013. (doi:10.1088/1742-5468/2014/12/P12013)
69. Concannon RJ, Blythe RA. 2014 Spatiotemporally complete condensation in a non-Poissonian exclusion process. *Phys. Rev. Lett.* **112**, 050603. (doi:10.1103/PhysRevLett.112.050603)
70. Parmeggiani A, Franosch T, Frey E. 2004 Totally asymmetric simple exclusion process with Langmuir kinetics. *Phys. Rev. E* **70**, 046101. (doi:10.1103/PhysRevE.70.046101)
71. Steurer B *et al.* 2018 Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA polymerase II. *Proc. Natl Acad. Sci. USA* **115**, E4368–E4376. (doi:10.1073/pnas.1717920115)
72. Shaw LB, Zia RKP, Lee KH. 2003 Totally asymmetric exclusion process with extended objects: a model for protein synthesis. *Phys. Rev. E* **68**, 021910. (doi:10.1103/PhysRevE.68.021910)
73. Schönherr G, Schütz GM. 2004 Exclusion process for particles of arbitrary extension: hydrodynamic limit and algebraic properties. *J. Phys. A: Math. Gen.* **37**, 8215–8231. (doi:10.1088/0305-4470/37/34/002)
74. Schönherr G. 2005 Hard rod gas with long-range interactions: exact predictions for hydrodynamic properties of continuum systems from discrete models. *Phys. Rev. E* **71**, 026122. (doi:10.1103/PhysRevE.71.026122)
75. Shaw LB, Sethna JP, Lee KH. 2004 Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles. *Phys. Rev. E* **70**, 021901. (doi:10.1103/PhysRevE.70.021901)
76. Shaw LB, Kolomeisky AB, Lee KH. 2004b Local inhomogeneity in asymmetric simple exclusion processes with extended objects. *J. Phys. A Math. Gen.* **37**, 2105–2113. (doi:10.1088/0305-4470/37/6/010)
77. Dong JJ, Schmittmann B, Zia RKP. 2007 Inhomogeneous exclusion processes with extended objects: the effect of defect locations. *Phys. Rev. E* **76**, 051113. (doi:10.1103/PhysRevE.76.051113)
78. Lam SK, Pitrou A, Seibert S. 2015 Numba: a LLVM-based Python JIT compiler. In *Proc. of the 2nd Workshop on the LLVM Compiler Infrastructure in HPC—LLVM'15, Austin, TX, 15 November*, pp. 1–6. New York, NY: ACM Press.
79. Cavallaro M, Wang Y, Hebenstreit D, Dutta R. 2023 Bayesian inference of polymerase dynamics over the exclusion process. Figshare. (doi:10.6084/m9.figshare.c.6760029)