Accurate inference of stochastic gene expression from nascent transcript heterogeneity

Xiaoming Fu [*][1,2], Heta P. Patel [†][3], Stefano Coppola[3], Libin Xu[1], Zhixing Cao[‡][1], Tineke L. Lenstra[§][3], and Ramon Grima[¶][2]

[1]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China
[2]School of Biological Sciences, University of Edinburgh, United Kingdom
[3]The Netherlands Cancer Institute, Oncode Institute, Division of Gene Regulation, Amsterdam, The Netherlands

November 9, 2021

Abstract

Transcriptional rates are often estimated by fitting the distribution of mature mRNA numbers measured using smFISH (single molecule fluorescence *in situ* hybridization) with the distribution predicted by the telegraph model of gene expression, which defines two promoter states of activity and inactivity. However, fluctuations in mature mRNA numbers are strongly affected by processes downstream of transcription. In addition, the telegraph model assumes one gene copy, but in experiments cells may have two gene copies as cells replicate their genome during the cell cycle. It is thus unclear how accurately the inferred parameters reflect transcription. To address these issues, here we measure both mature and nascent mRNA distributions of *GAL10* in yeast cells using smFISH and classify each cell according to its cell cycle stage. We infer transcriptional parameters from mature and nascent mRNA distributions, with and without accounting for cell cycle stage and compare the results to live-cell transcription measurements of the same gene. We conclude that: (i) not accounting for cell cycle dynamics in nascent mRNA data overestimates the magnitude of promoter switching rates and the initiation rate, and underestimates the fraction of time spent in the active state and the burst size. (ii) use of mature mRNA data, instead of nascent data, significantly increases the errors in parameter estimation and can mistakenly classify a gene as non-bursting. Furthermore, we show how to correctly adjust for measurement noise in smFISH at low nascent transcript numbers. Simulations with parameters estimated from nascent smFISH data corrected for cell cycle phases and measurement noise leads to autocorrelation functions that agree with those obtained from live-cell imaging. Therefore, our novel data curation method yields a quantitatively accurate picture of gene expression.

1 Introduction

Transcription in single cells occurs in stochastic bursts [1, 2]. Although the first observation of bursting occurred more than 40 years ago [3], the precise mechanisms behind this phenomenon are still under active investigation [4, 5]. The direct measurement of the dynamic properties of bursting employs live-cell imaging approaches, which allow visualization of bursts as they occur in living cells [6]. However, in practice such live-cell measurements are challenging because they are low-throughput and require genome-editing [7, 8]. To circumvent this, one can exploit the fact that

[*] Joint first author
[†] Joint first author
[‡] Email: zcao@ecust.edu.cn
[§] Email: t.lenstra@nki.nl
[¶] Email: ramon.grima@ed.ac.uk

bursting creates heterogeneity in a population. In this case, it is relatively straightforward to obtain s steady-state distributions of the number of mRNAs per cell from smFISH or single-cell sequencing experiments. These distributions have been used to infer dynamics by comparison to theoretical models. The simplest mathematical model describing bursting is the telegraph (or two-state) model [9, 10]. In this model, promoters switch between an active and inactive state, where initiation occurs during the active promoter state. The model makes the further simplifying assumption that the gene copy number is one and that all the reactions are effectively first-order. The mRNA in this model can be interpreted as cellular (mature) mRNA since its removal via various decay pathways in the cytoplasm is known to follow single-exponential (first-order) decay kinetics for eukaryotic cells [11,12]. The solution of the telegraph model for the steady-state distribution of mRNA numbers have been fitted to experimental mature mRNA number distributions to estimate the transcriptional parameters [1, 2, 10, 13].

However, the reliability of the estimates of transcriptional parameters from mRNA distributions is questionable because the noise in mature mRNA (and consequently the shape of the mRNA distribution) is affected by a wide variety of factors. Recent extensions of the telegraph model have carefully investigated how mRNA fluctuations are influenced by the number of promoter states [14, 15], polymerase dynamics [16], cell-to-cell variability in the rate parameter values [17,18], replication and binomial partitioning due to cell division [19], nuclear export [20] and cell cycle duration variability [21]. A way to avoid noise from various post-transcriptional sources is to measure distributions of nascent mRNA rather than mature mRNA, and then fit these to the distributions predicted by an appropriate mathematical model. Nascent mRNA [22,23] is mRNA that is being actively transcribed, i.e. it is still tethered to an RNA polymerase II (Pol II) moving along a gene during transcriptional elongation. Fluctuation in nascent mRNA numbers thus directly reflects the process of transcription. Because nascent mRNA removal is not first-order, an extension of the telegraph model has been developed (the delay telegraph model) [24]. Fitting the distribution predicted by the delay telegraph model to the experimental distribution is therefore expected to improve the estimates of transcriptional parameters [25–29].

However, nascent mRNA data still suffers from extrinsic sources of noise due to cell-to-cell variability. For example in a asynchronous population of dividing cells, cells can have either one or two gene copies. In the absence of a molecular mechanism that compensates for the increase in gene copy number upon replication, cells with two gene copies which cannot be spatially resolved will have a different distribution of nascent mRNA numbers (one with higher mean) than cells with one gene copy. The importance of the cell cyle is illustrated by the finding [30] that noisy transcription from the synthetic TetO promoter in *S. cerevisiae* is dominated by its dependence on the cell cycle. Although it is possible to simultaneously measure mature and nascent mRNA as a function of the cell cycle position [26], cell cycle stage is generally not taken into account when fitting smFISH distributions. Since estimation of all transcriptional parameters (switching rates and initiation rate) from nascent data as a function of the cell cycle phase has not been reported, it is unclear how the cell cycle affects transcriptional parameter inference. Additionally, there are no studies which compare transcriptional parameters estimated from nascent mRNA data and those estimated from cellular (mature) mRNA data, and hence the reliability of the latter (which is the standard and commonest procedure) remains an open question.

In this paper, we seek to understand the impact of post-transcriptional noise and cell-to-cell variability on the accuracy of transcriptional parameters inferred from mature mRNA data. The fitting algorithms (for mature and nascent mRNA data) are first tested on simulated data, where some limitations of the algorithms are uncovered in accurately estimating a subset of the transcriptional parameters in certain regions of parameter space. The algorithms are then applied to four independent experimental data sets, each measuring *GAL10* mature and nascent mRNA data from smFISH in galactose-induced budding yeast, conditional on the stage of the cell cycle (G1 or G2) for thousands of cells. Comparison of the transcriptional parameter estimates allows us to separate the influence of ignoring cell cycle variability from that of post-transcriptional noise (mature vs nascent mRNA data). We find that only fitting of nascent cell-cycle data, corrected for measurement noise, provides good agreement with measurements from live-cell data. Cell-cycle specific analysis also reveals that upon transition from G1 to G2, yeast cells show dosage compensation by reducing burst frequency, similar to mammalian cells [31]. Our systematic comparison highlights the challenges of obtaining kinetic information from static data, and provides insight into potential biases when inferring transcriptional parameters from smFISH distributions.

2 Methods

2.1 Mature mRNA inference

2.1.1 Mathematical model

The steady-state solution of the telegraph model of gene expression [9] gives mature mRNA distributions. The reaction steps in this model are illustrated in Fig. 1a. Next we describe the generation of synthetic mature mRNA data and the algorithm used to infer parameters from this data.

2.1.2 Generation of synthetic mature mRNA data

We generate a set of synthetic mature mRNA data by randomly choosing parameters in the space:

$$(\sigma_{\text{off}}, \sigma_{\text{on}}, \rho) \in [\text{Uniform}(0.1, 150), \text{Uniform}(0.1, 50), \text{Uniform}(10, 250)] \tag{2.1}$$

The unit of these parameters is inverse minute. To avoid the possibility that the inference algorithm does not consider parameter sets that are in the physiological range, we choose the upperbound of the parameter ratios in (2.1) to be much larger than those ratios estimated for thousands of genes in eukaryotic cells (see Table I of [19]). Once a set of parameters is chosen, we use the stochastic simulation algorithm (SSA [32]) to simulate the telegraph model reactions in Fig. 1a and generate 10^4 samples of synthetic data.

2.1.3 Steps of the algorithm to estimate parameters from mature mRNA data

The procedure consists of the following steps: (i) select a set of random transcriptional parameters; (ii) use the solution of the telegraph model to calculate the probability of observing the measured number of mature mRNA from each cell; (iii) evaluate the likelihood function for the observed data; (iv) iterate the procedure until the negative log-likelihood is minimized; (v) the set of parameters that accomplishes the latter provides the best point-estimate of the parameters of the model used to generate the synthetic data.

For step (i), we restrict the search for optimal parameters in the following region of parameter space

$$(\sigma_{\text{off}}, \sigma_{\text{on}}, \rho) \in [\text{Uniform}(0, 150), \text{Uniform}(0, 50), \text{Uniform}(0, 300)] \ (\text{min}^{-1}) =: \Theta. \tag{2.2}$$

Step (ii) can be obtained either by computing the distribution from the analytical solution [9] or by using the finite state projection (FSP) method [33]. Here, for the sake of computational efficiency, we use the FSP method to compute the probability distribution of mature mRNA numbers.

For step (iii) we calculate the likelihood of observing the data given a chosen parameter set θ

$$\mathcal{L}(\theta) = \prod_{j=1}^{N_{\text{cell}}} P(n_j; \theta), \tag{2.3}$$

where $P(n_j; \theta)$ is the probability distribution of mature mRNA numbers obtained from step (ii) given a parameter set θ, n_j is the total number of mature mRNA from cell j and N_{cell} is the total number of cells.

Steps (i) and (iv) involve an optimization problem. Specifically we use a gradient-free optimization algorithm, namely *adaptive differential evolution optimizer* (ADE optimizer) using *BlackBoxOptim.jl* within the *Julia* programming language to find the optimal parameters

$$\theta^* = \arg\min_{\theta \in \Theta} \left(- \sum_{j=1}^{N_{\text{cell}}} \log P(n_j; \theta) \right). \tag{2.4}$$

The minimization of the negative log-likelihood is equivalent to maximizing the likelihood. Note the optimization algorithm is terminated when the number of iterations is larger than 10^4; this number is chosen because we have found that invariably after this number of iterations, the likelihood has converged to some maximal value. Note that the inference algorithm is particularly low cost computationally, with the optimal parameter values estimated in at most a few minutes.

Once the best parameter set θ^* is found, we calculate the mean relative error which is defined as

$$\text{Mean relative error } = \frac{1}{M} \sum_{i=1}^{M} \frac{|\theta_i^* - \theta_{\text{true},i}|}{|\theta_{\text{true},i}|}, \tag{2.5}$$

where θ_i^* and $\theta_{\text{true},i}$ represent the i-th estimated and true parameters respectively, and M denotes the number of the estimated parameters. Thus, the mean relative error reflects the deviation of the estimated parameters from the true parameters.

2.2 Nascent mRNA inference

2.2.1 Mathematical model

The steady-state solution of the delay telegraph model [24] gives the distribution of the number of bound Pol II. In SI Section 1, we present an alternative approach to derive the steady-state solution. The reaction steps are illustrated in Fig. 2a.

The position of a Pol II molecule on the gene determines the fluorescence intensity of the mRNA attached to it. In particular for fluorescence data acquired from smFISH PP7-*GAL10*, the fluorescence intensity of a single mRNA on the DNA locus looks like a trapezoidal pulse (see Fig. 2b for an illustration). This presents a problem because using the delay telegraph model, we can predict the distribution of the number of bound Pol II however we do not have any specific information on their spatial distribution along the gene. However since the delay telegraph model implicitly assumes that a Pol II molecule has fixed velocity and that Pol II molecules do not interact with each other (via volume exclusion), it sounds reasonable to assume that in steady-state the bound Pol II molecules are uniformly distributed along the gene. This hypothesis is confirmed by stochastic simulations of the delay telegraph model where the position of a Pol II molecule is calculated as the product of the constant Pol II velocity and the time since its production.

By the uniform distribution assumption and the measured trapezoidal fluorescence intensity profile, it follows that the signal intensity of each bound Pol II has the density function g defined by

$$g(s) = \frac{L_1}{L} \mathbb{1}_{[0,1]}(s) + \frac{L_2}{L} \delta_1(s), \; s \in [0,1],$$

where $L_1 = 862$ bp (base pairs), $L_2 = 2200$ bp, $L = L_1 + L_2$ as defined in Fig. 2b. The indicator function $\mathbb{1}_{[0,1]}(s) = 1$ if and only if $s \in [0,1]$ and $\delta_1(s)$ is the Dirac function at 1. The probability of the signal s being between 0 and 1 is due to the first part of the trapezoid function and hence is multiplied by L_1/L which is the probability of being in this region if Pol II is uniformly distributed. Similarly, the probability of s being 1 is due to the L_2 part of the trapezoid and hence the probability is L_2/L by the uniform distribution assumption. Note that the signal s from each Pol II is at most 1 because in practice, the signal intensity from the transcription site is normalized by the median intensity of single cytoplasmic mRNAs [22].

The total signal is the sum of the signals from each bound Pol II. Hence, the density function of the sum is given by the convolution of the signal densities from each bound Pol II. Defining $p(s|k)$ as the density function of the signal given there are k bound Pol II molecules, we have that $p(s|k)$ is the k–th convolution power of g, i.e.

$$p(s|k) = (g * g \cdots * g)(s) = g^{*k}(s), \quad g^{*0}(s) = \delta_0(s), \tag{2.6}$$

where $\delta_0(s)$ is the Dirac function at 0. Finally we can write the total fluorescent signal density function as

$$p(s;\theta) = \sum_{k=0}^{\infty} p(s|k)P(k;\theta), \tag{2.7}$$

where $P(k;\theta)$ is the steady-state solution of the delay telegraph model giving the probability of observing k bound Pol II molecules for the parameter set θ. Hence Eq. (2.7) represents the extension of the delay telegraph model to predict the smFISH fluorescent signal of the transcription site.

Next we describe the generation of synthetic nascent mRNA data and the algorithm used to infer parameters from this data.

2.2.2 Generation of synthetic nascent mRNA data

We generate synthetic smFISH signal data by using the SSA modified to include delay to simulate the delay telegraph model; specifically we use Algorithm 2 described in [34]. One realization of the algorithm simulates the fluctuating number of bound Pol II molecules in a single cell.

The total fluorescence intensity (mimicking smFISH) is obtained as follows. When a particular bound Pol II is produced by a firing of the transcription reaction $G \to G + N$, we record this production time; since the elongation rate is assumed to be constant, given the production time we can calculate the position of the Pol II molecule on the gene at any later time and hence using Fig. 2b we can deduce the fluorescent signal due to this Pol II molecule.

Specifically we normalize each transcribing Pol II's position to $[0, 1]$ and map the position to its normalized signal by

$$q(x) = \begin{cases} x\frac{L}{L_1} & x \in \left[0, \frac{L_1}{L}\right), \\ 1 & x \in \left[\frac{L_1}{L}, 1\right], \end{cases}$$

where x is the normalized position on the gene. Thus at a given time, the total fluorescent signal from the n-th cell (the n-th realization of the SSA) equals

$$q_n = \sum_{j=1}^{J_n} q(x_j),$$

where J_n is the number of bound Pol II molecules in the n-th cell, and $\{x_j\}$ with $j = 1, \ldots, J_n$ is the vector of all Pol II positions on the gene. The total signal from each cell is a real number but it is discretized into an integer.

The kinetic parameters are randomly chosen from the following region of parameter space

$$\begin{aligned} &(\sigma_{\text{off}}, \sigma_{\text{on}}, \rho) \in [\text{Uniform}(0.1, 4), \text{Uniform}(1, 10), \text{Uniform}(1, 80)]\ (\text{min}^{-1}) \\ &5 < \rho/\sigma_{\text{off}} < 100, \quad \tau = 0.5\ (\text{min}). \end{aligned} \tag{2.8}$$

Note that τ is fixed because this parameter is often directly measurable in experiments and hence there is no need to infer it. Note also that each synthetic signal histogram is computed from 10^4 samples (each sample represents a single cell measurement).

2.2.3 Steps of the algorithm to estimate parameters from nascent mRNA data

The inference procedure is essentially the same as steps (i)-(v) described in mature mRNA inference except for the following points.

In step (ii), the probability of observing a total signal of intensity i from a single cell is obtained by integrating $p(s;\theta)$ in Eq. (2.7) on an interval $[i-1,i]$ for $i \in \mathbb{N}$ which, in our numerical scheme, means

$$S(i;\theta) := \sum_{k=0}^{K} P(k;\theta) \int_{i-1}^{i} g^{*k}(x)\mathrm{d}x, \quad i = 1,2,\ldots \tag{2.9}$$

Note that the integration over the interval of length 1 is to match the discretization of the synthetic data and $\theta \in \Theta$. Intuitively, one can always choose a positive integer K such that $P(k) = 0$ for any $k \geq K$. The computation of the solution of the delay telegraph model $P(k)$ can be done either using the analytical solution (evaluated using high precision) or using the finite state projection algorithm (FSP) [33]. In SI Fig. 1 and SI Table 1, we show that the two methods yield comparable accuracy and CPU time.

For step (iii) we calculate the likelihood of observing the data given a chosen parameter set θ

$$\mathcal{L}(\theta) = \prod_{j=1}^{N_{\text{cell}}} S(q_j;\theta), \tag{2.10}$$

where q_j is the discretized total signal intensity from cell j and N_{cell} is the total number of cells. In the optimization, we aim to find

$$\theta^* = \arg\min_{\theta\in\Theta} \left(-\sum_{j=1}^{N_{\text{cell}}} \log S(q_j;\theta) \right).$$

The whole procedure is summarized by a flow-chart in Fig. 2c.

2.3 Experimental data acquisition and processing

Yeast cultures were grown to early mid-log, fixed with paraformaldehyde (PFA), permeabilized with lyticase and hybridized with 7.5pmol each of four PP7 probes labeled with Cy3 (Integrated DNA Technologies) as described in Trcek et al. [35] and Lenstra et al. [36, 37]. The PP7 probe sequences are: atatcgtctgctcctttcta, atatgctctgctggtttcta, gcaattaggtaccttaggat, aatgaacccgggaatactgc. Coverslips were mounted on microscope slides using mounting media with DAPI (ProLong Gold, Life Technologies).

The coverslips were imaged on a Zeiss AxioObserver (Zeiss, USA) widefield microscope. Light sources for imaging DAPI, Cy3 were 440/20nm and 550/15nm, respectively from SpectaX Lumencor light engine (Lumencor, USA). The signal was detected on a Hamamatsu ORCA-Flash4.0 V3 Digital CMOS camera (Hamamatsu Photonics, Japan). The imaging was performed using MicroManager (UCSF). Fields of view were selected based on the DAPI channel and stacks of 13 images with a z-step of 0.5 um were acquired in two colors.

A custom python pipeline was used for analysis (https://github.com/Lenstralab/smFISH). Maximum intensity projected images were used for cell and nucleus segmentation, and the diffraction-limited Cy3 spots were detected per z-slice using band-pass filtering and refined using iterative Gaussian mask localization procedure (Crocker and Grier [38]; Thompson et al. [39]; Larson et al. [40, 41] and Coulon et al. [42]).

Spots were classified as nuclear or cytoplasmic and the brightest nuclear spots were classified as transcription sites. The intensity of the brightest nuclear spot in a cell was normalized with the median intensity of all the cytoplasmic spots in the same cell. This is due to the fact that most of cytoplasmic mRNAs are isolated, thus the median of the fluorescence signal of cytoplasmic mRNAs can be considered as the normalizing value. The distribution of the normalised intensity of the brightest nuclear spot, calculated over the cell population, is the experimental equivalent of the total fluorescent signal density function as given by the solution of the modified delay telegraph model, Eq. (2.7).

The number of mature mRNA in each cell is given by counting the number of spots in the entire cell, i.e. nuclear plus cytoplasmic. The transcription site is counted as 1 mRNA, regardless of its intensity, but this has negligible influence since the mean number of mature mRNA is much greater than 1. The distribution of the number of spots is the experimental equivalent of the solution of the telegraph model, i.e. the marginal distribution of mature mRNA numbers in steady-state conditions.

The integrated nuclear intensity of each cell was calculated by summing the DNA content intensity (DAPI) of all the pixels within the nucleus mask. The distribution of the intensities was fit with a bimodal Gaussian distribution. Those cells whose intensity was within a standard deviation of the mean of the first (second) Gaussian peak was classified as G1 (G2) (see SI Fig. 2). This gave similar results to a different cell cycle classification method using the Fried/Baisch model [43] which was recently employed in [26]. See SI Fig. 3 for a comparison of the two methods.

We did four independent experiments with a total number of cells equal to 2510, 6411, 4592, 3181 respectively. After classification, the numbers of G1 cells are 766, 2111, 1495, 904 and the number of G2 cells are 683, 1657, 1209, 1143, whereas the rest were classified as undetermined.

3 Results

3.1 Inference from mature mRNA data: testing inference accuracy using synthetic data

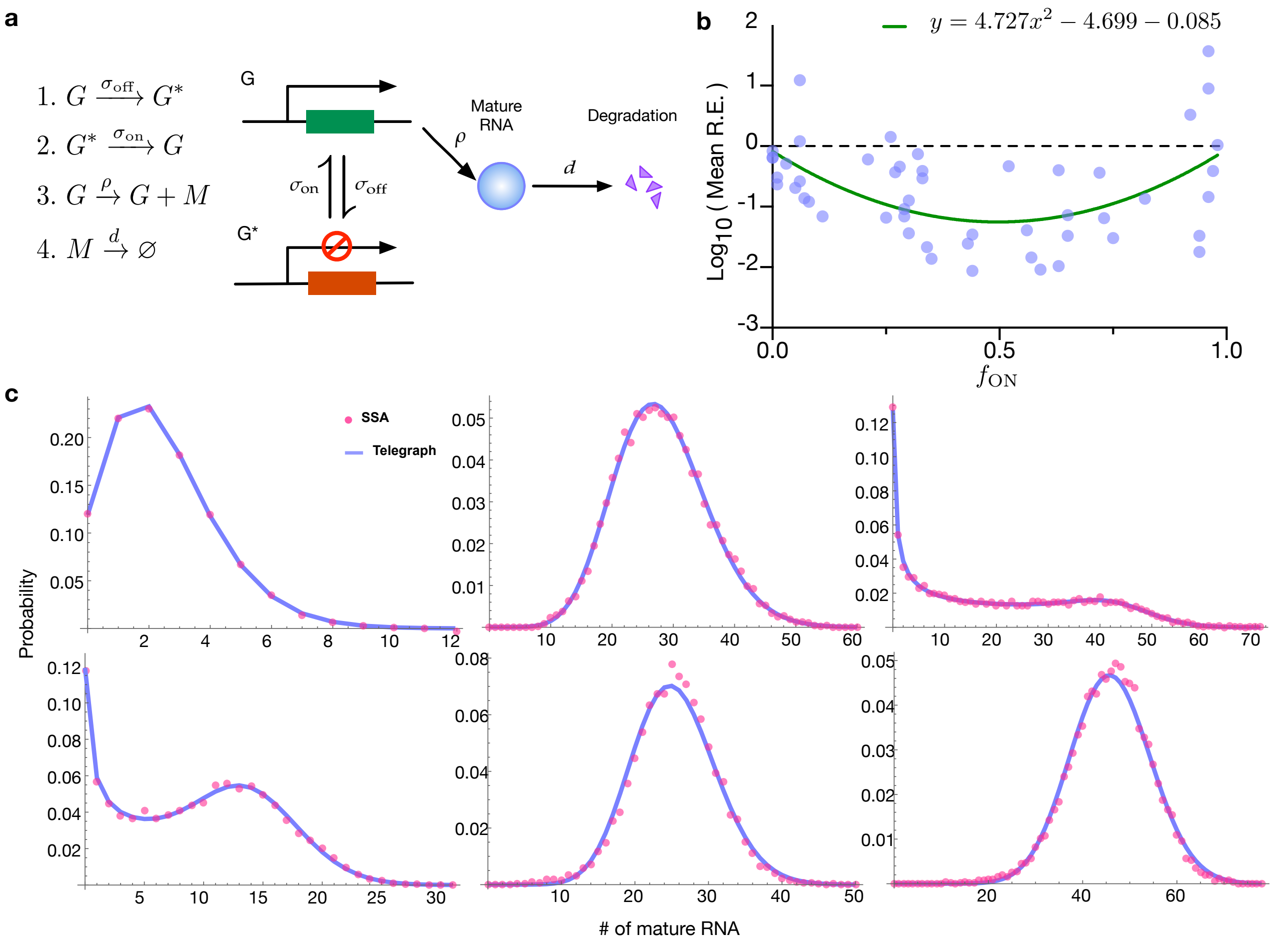
Transcriptional bursting can be mathematically described with a two-state model, where we define the rate of switching from the ON state (active state) to the OFF state (inactive state) as σ_{off}, the rate of switching from the OFF state to the ON state as σ_{on} and the production rate of mRNAs in the ON state as ρ. The first-order decay rate of mature mRNA in the telegraph model is given by d (Fig. 1a).

To understand the accuracy of the algorithm in various regions of the parameter space, we used stochastic simulations of the telegraph model to generate synthetic data and then used an inference algorithm to infer transcriptional parameters by matching the distribution of synthetic mature mRNA data to the analytical distribution of the conventional telegraph model.

We note that the steady-state solution of the telegraph model is a function of the non-dimensional parameter ratios $\rho/d, \sigma_{\text{off}}/d$ and σ_{on}/d [10]. Hence without a direct experimental measurement of the degradation rate d, only these three ratios can be inferred [2]. In these simulations, we fix the degradation rate $d = 1$ min^{-1} and aim to estimate $\rho, \sigma_{\text{off}}$ and σ_{on} from synthetic data.

The generation of synthetic data is described in Methods Section 2.1.2. The inference algorithm is described in detail in Methods Section 2.1.3. It is based on a maximization of the likelihood of observing the single cell mature mRNA numbers measured in a population of cells. The likelihood of observing a certain number n of mature mRNA numbers from a given cell is given by the telegraph model's steady-state probability distribution of mature mRNA numbers evaluated for n copy numbers.

To test the accuracy of the inference algorithm, we generated 50 independent sets of synthetic mature mRNA data (each having 10^4 cells and a unique set of parameters) and for each, we calculated the mean relative error in the parameters (for its definition see Methods, Eq. (2.5)). Fig. 1b shows the mean relative error as a function of the fraction of ON time (this is defined as $f_{\text{ON}} = \sigma_{\text{on}}/(\sigma_{\text{off}} + \sigma_{\text{on}})$). We find that the error reaches a minimum when the promoter is ON half of the time which occurs when σ_{off} is equal to σ_{on}. Fig. 1c shows the best fit distributions for 6 different parameter sets. In Fig. 1d we show the corresponding estimated parameters compared to the true ones. In agreement with Fig. 1b, we find that only for parameters sets 3 and 4 (where the switching rates are similar and the fraction of ON time is not far from 0.5) the estimates of the three transcriptional parameter rates ($\sigma_{\rm off}, \sigma_{\rm on}, \rho$) and the burst size (the ratio $\rho/\sigma_{\rm off}$) are accurate. For parameter sets 1 and 2 (where the promoter spends a large fraction of time in the OFF state), only the burst size and the burst frequency $\sigma_{\rm on}$ are reliable. This finding agrees with the theoretical proof that with long OFF periods, the distribution solution of the master equation of the telegraph model is well approximated by a negative binomial distribution with only two free parameters (the burst size and burst frequency) [19]. For parameter sets 5 and 6 (where the promoter spends a large fraction of time in the ON state), errors are large in all parameters; while the error is smallest in the production rate in the ON state, this is still sizeable. We note that for all data sets, the effective rate of transcription given by $\hat{\rho} = \rho f_{\rm ON}$ is reliably inferred . This is because the inference algorithm tries to match the mean mature mRNA number, which equals $\hat{\rho}/d$ and since d is fixed, $\hat{\rho}$ is generally well inferred.



d

Table: the inferred parameters using synthetic mature RNA data

	True						Telegraph					
Parameter	$\sigma_{\rm off}$	$\sigma_{\rm on}$	ρ	Burst Size	$f_{\rm ON}$	$\hat{\rho}$	$\sigma_{\rm off}$	$\sigma_{\rm on}$	ρ	Burst Size	$f_{\rm ON}$	$\hat{\rho}$
Set 1	120.38	7.70	40.85	0.34	0.06	2.45	73.32	7.13	27.66	0.38	0.09	2.45
Set 2	35.48	17.39	85.05	2.40	0.33	27.98	59.95	20.30	111.05	1.85	0.25	28.09
Set 3	0.52	0.42	46.21	88.36	0.44	20.50	0.57	0.42	46.65	82.43	0.43	19.93
Set 4	0.32	0.47	15.71	48.65	0.59	9.29	0.33	0.47	15.71	47.91	0.59	9.28
Set 5	0.53	5.85	49.95	94.72	0.92	45.82	4.65	16.62	58.86	12.65	0.78	45.99
Set 6	0.16	3.89	26.45	169.81	0.96	25.44	16.26	31.26	38.64	2.38	0.66	25.42

Figure 1: Inference of transcriptional parameters using as input synthetic mature mRNA data generated by SSA simulations. **a.** A schematic illustration of the telegraph model. **b.** Plot of the Log mean relative error (RE) against fraction of ON time ($f_{\rm ON}$) for 50 numerical experiments, each consisting of 10^4 SSA samples generated for a unique (random) set of parameters. Note that the Log(Mean R.E.) is base 10. **c.** Distributions of mature mRNA from synthetic data (red dots) fit using the inference algorithm with telegraph model (blue) for six different parameter sets. **d.** Estimates using the inference algorithm with the telegraph model for the six parameter sets in **c.** For both the ground truth and the estimated parameters, we fix the degradation rate $d = 1$ min^{-1}.

3.2 Inference from nascent mRNA data: testing inference accuracy using synthetic data

While it is common to infer transcriptional parameters using mature mRNA data, the upstream process of splicing, export and nuclear mRNA degradation adds additional extrinsic noise and changes the distributions of transcript copy numbers [44]. One way to avoid this type of noise is to directly infer the parameters from distributions of nascent mRNA numbers since this is free from post-transcriptional noise sources.

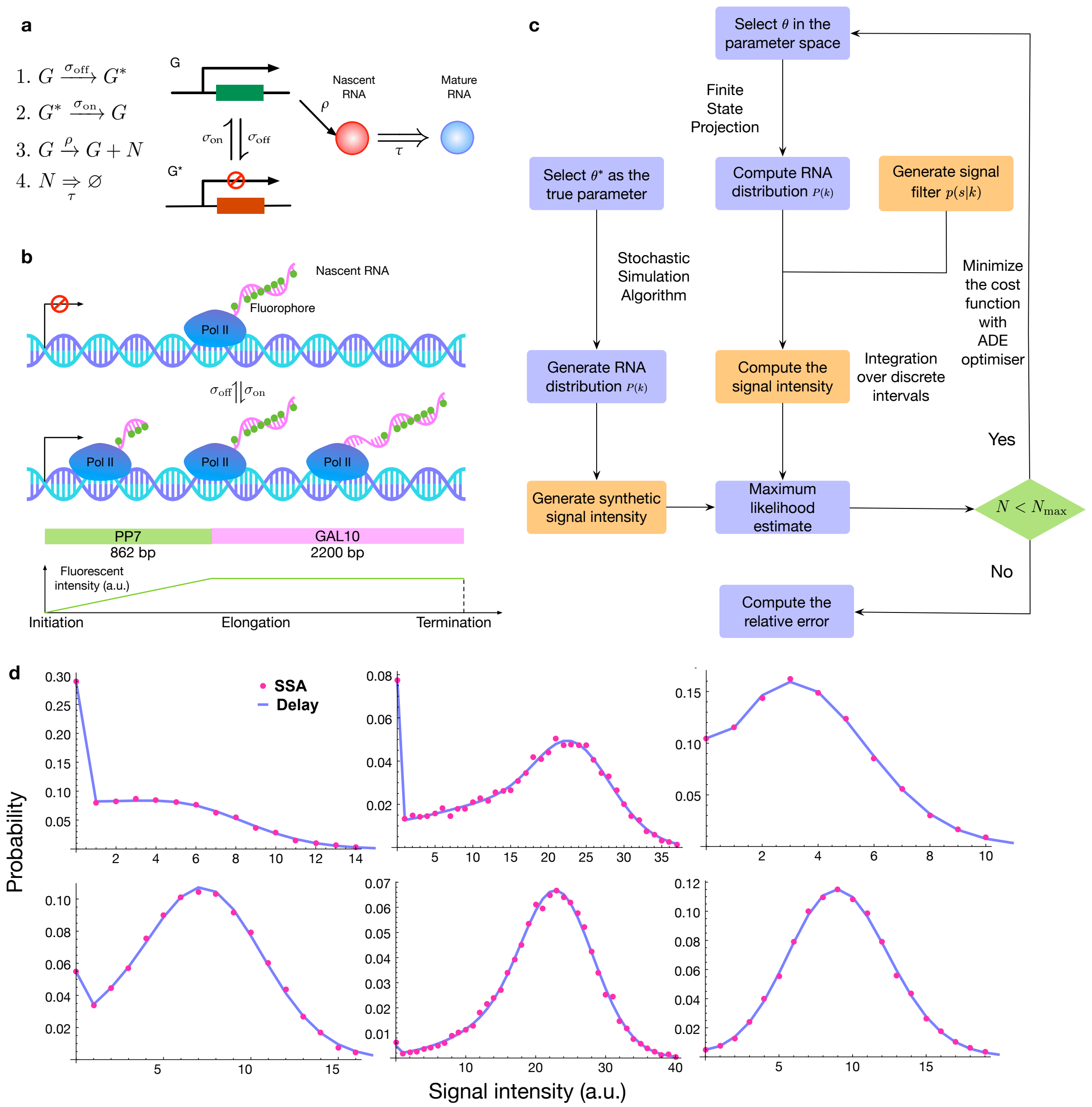
While the telegraph model could in principle be used to describe nascent mRNA, it is not appropriate because nascent mRNA does not follow a first-order decay. Once a Pol II molecule is bound, it travels along the gene with an approximately constant velocity for a fixed time after which it unbinds and the nascent mRNA tail dissociates to form a mature mRNA. Hence the time between a nascent mRNA production event (the binding of a Pol II to the gene) and its removal (the dissociation of the Pol II molecule from the gene) is not exponentially distributed and cannot be simply modelled by an effective first-order reaction as for mature mRNA. To model nascent mRNA dynamics, it is more appropriate to use a delay telegraph model, where the first-order degradation reaction is replaced by a reaction that removes a nascent transcript after a fixed elongation time τ (Fig. 2a). This non-Markovian model was studied by Xu et al. [24] who found an exact steady-state solution for the distribution of nascent mRNAs; the first two moments for this distribution were also reported in [45].

The delay telegraph model predicts the distribution of the number of bound Pol II molecules. However, the latter does not directly translate into the numbers of nascent RNA molecules. In the case of smFISH [46–48], a method that is commonly used for mRNA detection, a fluorescent signal is emitted by oligonucleotide probes bound to the nascent mRNA. Since as a bound Pol II travels along the gene, its nascent mRNA tail grows, we expect the fluorescent signal intensity to increase as well.

Specifically in this manuscript, we use fluorescence data acquired from smFISH of PP7-*GAL10* in budding yeast, where probes were hybridized to the PP7 sequences. In this case the fluorescence intensity of a single mRNA on the DNA locus looks like a trapezoidal pulse (see Fig. 2b for an illustration). As the Pol II molecule travels through the 14 repeats of the PP7 loops, the intensity of the mRNA increases as the fluorescent protein binds to the mRNA (this is the linear part of the trapezoidal pulse). However once all 14 loops on the mRNA are bound by the fluorescent proteins, the plot starts to plateau because the mRNA is as bright as it can be but still needs to elongate through the *GAL10* gene body before it is released (hence the flat part of the trapezoidal pulse). The total fluorescent signal density function is hence given by

$$p(s;\theta) = \sum_{k=0}^{\infty} p(s|k)P(k;\theta), \tag{3.1}$$

where $p(s|k)$ is the density function of the signal given there are k bound Pol II molecules and $P(k;\theta)$ is the steady-state solution of the delay telegraph model giving the probability of observing k bound Pol II molecules for the parameter set θ. In Methods Section 2.2.1 we show how $p(s|k)$ can be approximately calculated for the trapezoidal pulse. Hence Eq. (3.1) represents the extension of the delay telegraph model to predict the smFISH fluorescent signal of the transcription site.



e

Table: the inferred parameters using synthetic nascent RNA data

	True						Delay					
Parameter	$\sigma_{\rm off}$	$\sigma_{\rm on}$	ρ	Burst size	$f_{\rm ON}$	$\hat{\rho}$	$\sigma_{\rm off}$	$\sigma_{\rm on}$	ρ	Burst size	$f_{\rm ON}$	$\hat{\rho}$
Set 1	2.26	2.14	18.64	8.26	0.49	9.08	1.77	1.76	18.28	10.36	0.50	9.13
Set 2	1.27	3.14	58.17	45.69	0.71	41.38	1.13	2.91	58.01	51.16	0.72	41.73
Set 3	2.27	5.80	12.55	5.53	0.72	9.02	1.60	4.48	12.21	7.63	0.74	9.00
Set 4	1.38	4.77	21.74	15.79	0.78	16.87	1.08	4.00	21.50	19.94	0.79	16.94
Set 5	1.05	8.20	57.99	55.09	0.89	51.39	0.94	7.19	57.92	61.87	0.88	51.25
Set 6	1.18	9.94	24.92	21.15	0.89	22.28	0.94	8.74	24.74	26.37	0.90	22.34

Figure 2: Inference of transcriptional parameters using as input synthetic fluorescent signal data generated by SSA simulations of transcription and fluorescent tagging for 10^4 cells. **a.** Illustration of the delay telegraph model. The double horizontal line for nascent mRNA removal indicates this is a delayed reaction. **b.** Illustration showing promoter switching between two states, Pol II binding to the promoter in the ON state and subsequently undergoing productive elongation. Note that the nascent mRNA tail increases as Pol II approaches a stop (termination) sequence in the gene. As Pol II travels through the 14 repeats of the PP7 loops, the intensity of the mRNA increases due to fluorescent protein binding to the mRNA; intensity saturates as Pol II enters the *GAL10* region. **c.** Illustration of the inference algorithm from mature and nascent mRNA data. The orange boxes are only applicable for the inference of the fluorescence signal intensity of nascent mRNAs. A large iteration step $N_{\rm max}$ ($\geq 10^4$) is chosen as the termination condition for the optimizer. **d.** Distributions of total fluorescence intensity from synthetic data (red dots) fit using the inference algorithm with the delay telegraph model (blue) for 6 different parameter sets. **e.** Estimates using the inference algorithm with delay telegraph model for the six parameter sets in **d.** The unit of $\sigma_{\rm off}, \sigma_{\rm on}$ and ρ is min^{-1}. Here the effective transcription rate $\hat{\rho}$ is defined by $\rho f_{\rm ON}$ and the delay $\tau = 0.5$ min.

We use this extension to generate synthetic data that mimicks the smFISH signal (Methods Section 2.2.2). Specifically we use the SSA to generate histograms of the total fluorescence intensity for kinetic parameters that are randomly selected. An inference algorithm is constructed to infer the parameters, which is based on a maximization of the likelihood of observing the single cell total fluorescence intensity measured in a population of cells (Methods Section 2.2.3). Note that the likelihood of observing a certain fluorescence signal intensity from a cell is given by Eq. (3.1). This algorithm is used to infer the promoter switching and initiation rate parameters. The elongation time is not estimated but assumed to be known since this can often be measured experimentally.

The mean relative errors (see Eq. (2.5) for definition) obtained from 20 independent numerical experiments (averaged over all parameter sets) is about 0.1, which indicates that the inference is accurate. The algorithm fits the distributions of signal intensity very accurately, as shown in Fig. 2d for 6 parameter sets. A direct comparison of the estimated and true parameters for these parameter sets is shown in Fig. 2e. Note that these cases describe moderate to highly frequent gene expression, i.e. $f_{\mathrm{ON}} \gtrapprox 0.5$.

A more extensive analysis using sets drawn over a wider range of parameter space and for which f_{ON} spans the whole range from 0 to 1 is reported in SI Section 3, SI Fig. 4, and SI Tables 2 and 3. Therein we show that the mean relative error versus the fraction of ON time displays the same trend as Fig. 1b with the errors being largest when the fraction of ON time is close to 0 or 1.

Additionally, if one utilizes the conventional telegraph model to fit the nascent data generated by delay SSA, it is possible to obtain a distribution fitting as good as the delay telegraph model but with low-fidelity parameter estimation (SI Section 4, SI Fig. 5 and SI Table 4). Analytically, the telegraph model is only an accurate approximation of the delay telegraph model when: (i) the promoter switching timescales are much longer than the time spent by Pol II on a gene or (ii) the off switching rates are very small such that gene expression is practically constitutive.

Overall, from these synthetic data, we learn that parameter interference is more accurate from nascent than from mature mRNA distributions, and the reliability is highest when the fraction of ON time is roughly 0.5.

3.3 Applications to experimental yeast mRNA data

Now that we have introduced the inference algorithms and tested them thoroughly using synthetic data, we applied the algorithms to experimental data (see Method Section 2.3 for details of data acquisition).

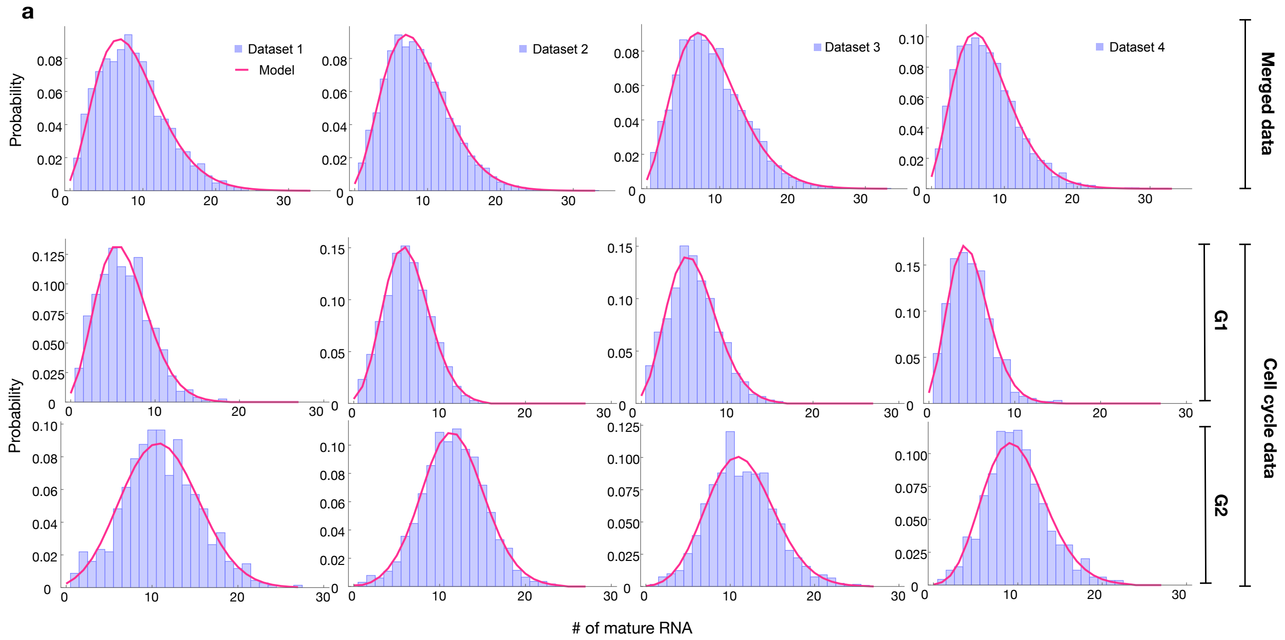
3.3.1 Inference from mature mRNA data: merged versus cell-cycle specific

We perform the inference in two different ways: (i) using the merged data from all cells, irrespective of their position in the cell cycle. (ii) using cell-cycle specific data.

The inference of transcriptional parameters using the merged data is done using the algorithm described in Methods Section 2.1.3 but with the experimental mature mRNA data replacing the synthetic data. For the cell-cycle specific data, for mature mRNA data of cells in G1, the inference protocol remains the same. However for cells in the G2 stage, this protocol needs change because now there are two gene copies whereas the solution of the telegraph model assumes one gene copy. Assuming the transcriptional activities of the two gene copies are independent, the distribution of the total molecule number is the convolution of the molecule number (obtained from the telegraph model) with itself for mature mRNA data. This convolved distribution is then used in steps (ii) and (iii) of the inference algorithm.

Note that the independence of gene copy transcription has been verified for genes in some eukaryotic cells [26] where the two copies can be easily resolved. For yeast data, as we are analyzing in this paper, it is not generally possible to distinguish the two copies of the allele in G2 because they are within the diffraction limit. This is because the yeast nucleus is a lot smaller than the mammalian nucleus and the two *GAL10* genes tend to be close together, so many cells do not have two clearly separate nuclear spots in fluorescence intensity. Hence in the absence of this data, the independence assumption is the simplest reasonable assumption that we can make.

We start by the inference of the merged mature mRNA data where we have four independent data sets. In the absence of an experimental measurement of the degradation rate, we can only estimate the 3 transcriptional parameters normalised by d. The best distribution fits and the inferred transcriptional parameters for the 4 experimental data sets are shown in Fig. 3a (top) and Fig. 3b, respectively. Since σ_{off} is much larger than σ_{on}, only the normalised burst frequency σ_{on}/d and the burst size are reliable, as shown for synthetic data. The small estimated values of burst size and fraction of ON time suggests small but infrequent bursty expression.



b Table: the inferred parameters using merged mature RNA data

Merged	σ_{off}	σ_{on}	ρ	Burst Size	f_{ON}	Fano factor
Set 1	21.18	4.73	47.15	2.23	0.18	2.46
Set 2	19.53	5.44	40.11	2.05	0.22	2.22
Set 3	16.53	4.77	39.48	2.39	0.22	2.39
Set 4	28.72	5.16	50.00	1.74	0.15	2.29
Mean	21.49	5.02	44.19	2.10	0.19	2.34
Std	5.19	0.34	5.21	0.28	0.06	0.11

c Table: the inferred parameters using G1 and G2 mature RNA data

G1	$\hat{\rho}/d$	Fano factor	G2	$\hat{\rho}/d$	Fano factor
Set 1	6.35	1.46	Set 1	5.55	1.77
Set 2	6.35	1.09	Set 2	5.81	1.16
Set 3	6.17	1.33	Set 3	5.68	1.39
Set 4	4.86	1.20	Set 4	4.94	1.41
Mean	5.93	1.27	Mean	5.49	1.43
Std	0.72	0.16	Std	0.38	0.25

Figure 3: Inference results using four mature mRNA data sets with sample sizes as 2333, 6366, 4550 and 3163 cells, respectively. **a.** Best fit distributions of mature mRNA data. **b.** The inferred transcriptional parameters (merged mature RNA data) and the burst size is computed as ρ/σ_{off}. Note that the transcriptional parameters in the first 3 columns are normalised by the degradation rate. **c.** Inferred effective (normalised) production rate per gene copy $\hat{\rho}$ and Fano factors (variance of molecule numbers divided by the mean of molecule numbers) for the G1 and G2 cell cycle phases.

Next, we performed inference for cell-cycle specific mature mRNA data, the results of which are shown in Fig. 3a (bottom) and Fig. 3c. As expected, the mean number of mRNAs in G2 cells is larger than that in G1 cells. However, inference of the transcriptional parameters for these cell cycle stages was found to be very unreliable, where very different sets of parameters lead to excellent fits of the data.

Note that the problem of estimating any of the parameters reliably is in line with the findings using synthetic data when the promoter spends most of its time in the ON state (see Section 3.1 and Fig. 1b). A measure that is frequently used for the burstiness of a gene is the Fano factor, which is defined as the variance of molecule numbers divided by the mean of molecule numbers. If the Fano factor is around 1 the gene follows non-bursting constitutive initiation, whereas larger Fano factors indicate larger burst sizes. For mature cell-cycle specific data, the Fano factor is close to 1 (Fig. 3c), which implies expression that is almost constitutive. As we showed using synthetic data, the only effective parameter that can be reliably estimated is the normalised effective rate of transcription $\hat{\rho}/d$.

Since for 3 out of 4 data sets, the effective normalised production rate per gene copy is less in G2 compared to G1, there is likely a mild form of gene dosage compensation at play, namely the transcriptional parameters are altered upon replication such that each copy of the gene has a reduced expression [26].

Hence, using cell-cycle fits suggests that expression in G1 and G2 is not bursty. This apparent lack of bursting does not agree with live-cell transcription measurements that clearly show transcriptional bursts of transcription for *GAL10* [6, 36].

What is particularly surprising in our analysis is the differences in the inference results using merged and cell-cycle specific data: the former suggests a higher degree of bursty expression than the latter (compare the Fano factors in Fig. 3b and Fig. 3c) [49]. Heterogeneity in the merged data, due to cells with one or two gene copies, could thus lead to fluctuations in mature mRNA abundance that can be mistakenly interpreted as arising from bursty expression. Conversely, a gene that displays bursting in other measurements, can mistakenly be classified as non-bursting, as in the case for *GAL10* [6, 36]. This analysis exemplifies that estimates of transcriptional parameters from mature mRNA distributions should be interpreted with caution.

3.4 Inference from nascent mRNA data: cell cycle effects, experimental artifacts and comparison with mature mRNA inference

3.4.1 cell-cycle specific versus merged data

The inference of transcriptional parameters using the merged data is done using the algorithm described in Methods Section 2.2.3 but with the experimental nascent mRNA data replacing the synthetic data.

Similar to above, to account for two gene copies in G2 cells, we assume that the transcriptional activities of the two gene copies are independent. The distribution of the total fluorescent signal from both gene copies is the convolution of the signal distribution (obtained from the extended delay telegraph model, i.e. Eq. (3.1)) with itself (for an illustration see SI Fig. 2). This convolved distribution is then used in steps (ii) and (iii) of the inference algorithm.

The inference of transcriptional parameters from nascent RNA data was done using a fixed elongation time, which was measured previously at a related galactose-responsive gene (*GAL3*) at 65 bp/s [6]. Since the total transcript length is 3062 bp (see Fig 2b), the elongation time (τ in our model) is $\approx 47.11\,s \approx 0.785$ min. The fixed elongation rate enables us to infer the absolute values of the three transcriptional parameters $\sigma_{\text{off}}, \sigma_{\text{on}}$ and ρ.

In Fig. 4a and Table 1, we show the estimates of the transcriptional parameters using both merged and cell-cycle specific data. In contrast to the mature estimates, for the nascent estimates the switching rates σ_{off} and σ_{on} are comparable, and we are in a parameter regime where the absolute values of all three transcriptional parameters can be reliably estimated and distributions are reasonable fit (Fig. 4b). However, merged and cell-cycle specific data produce different parameter estimates. To understand which of these estimates is correct, we compare these estimates to previous live-cell transcription measurements of the same gene [6]. Because live-cell traces and simulated traces with the estimated transcriptional parameters are difficult to compare directly, we instead compare their normalized autocorrelation functions (ACFs). Specifically we feed the parameter estimates to the SSA to generate synthetic live-cell data and then calculate the corresponding ACF (SI Section 6). We find that the estimates from cell-cycle specific data produce ACFs that match the live-cell data closer than that from the merged data – see Fig. 4c (left and middle). This is also clear from the sum of squared residuals which for each dataset is smaller for the ACF computed using the cell-cycle specific estimates rather than those from merged data – see Fig. 4c right.

By not taking into account the cell cycle, heterogeneity is increased, which artificially amplifies the Fano factor and burstiness of gene expression. In addition, it results in underestimation of the fraction of time spent in the ON state (f_{ON}) and the burst size. This comparison indicates that inference using merged data will lead to parameter estimates that are incorrect.

Comparing the results of inference from cell-cycle specific mature mRNA data (Fig. 3c) vs. cell-cycle specific nascent mRNA data (Table 1), we find that mature data does not allow the estimation of any of the transcriptional parameters (or their normalised values). However using cell-cycle specific nascent data we could estimate all parameters reliably. Interestingly, use of mature mRNA data drastically underestimates the burstiness of gene expression: the Fano factors are 1.27 in G1 and 1.43 in G2 from mature mRNA inference vs 4.12 in G1 and 4.61 in G2 from nascent mRNA inference.

The transcriptional estimates of the G1 and G2 populations show that the burst frequency (σ_{on}) is considerably less in G2 compared to G1 (a 41% reduction on average)(Table 1); the other two parameters σ_{off} and ρ show smaller differences between the two cell cycle phases (reductions of 27% and 8% on average, respectively). This decrease of the burst frequency σ_{on} after replication has also been reported for some genes in mammalian cells [26, 31], indicating that this could be a general mechanism for gene dosage compensation. Our results are consistent with a ChIP-Seq study [50] which showed that an increase in DNA dosage after replication does not increase gene expression in budding yeast.

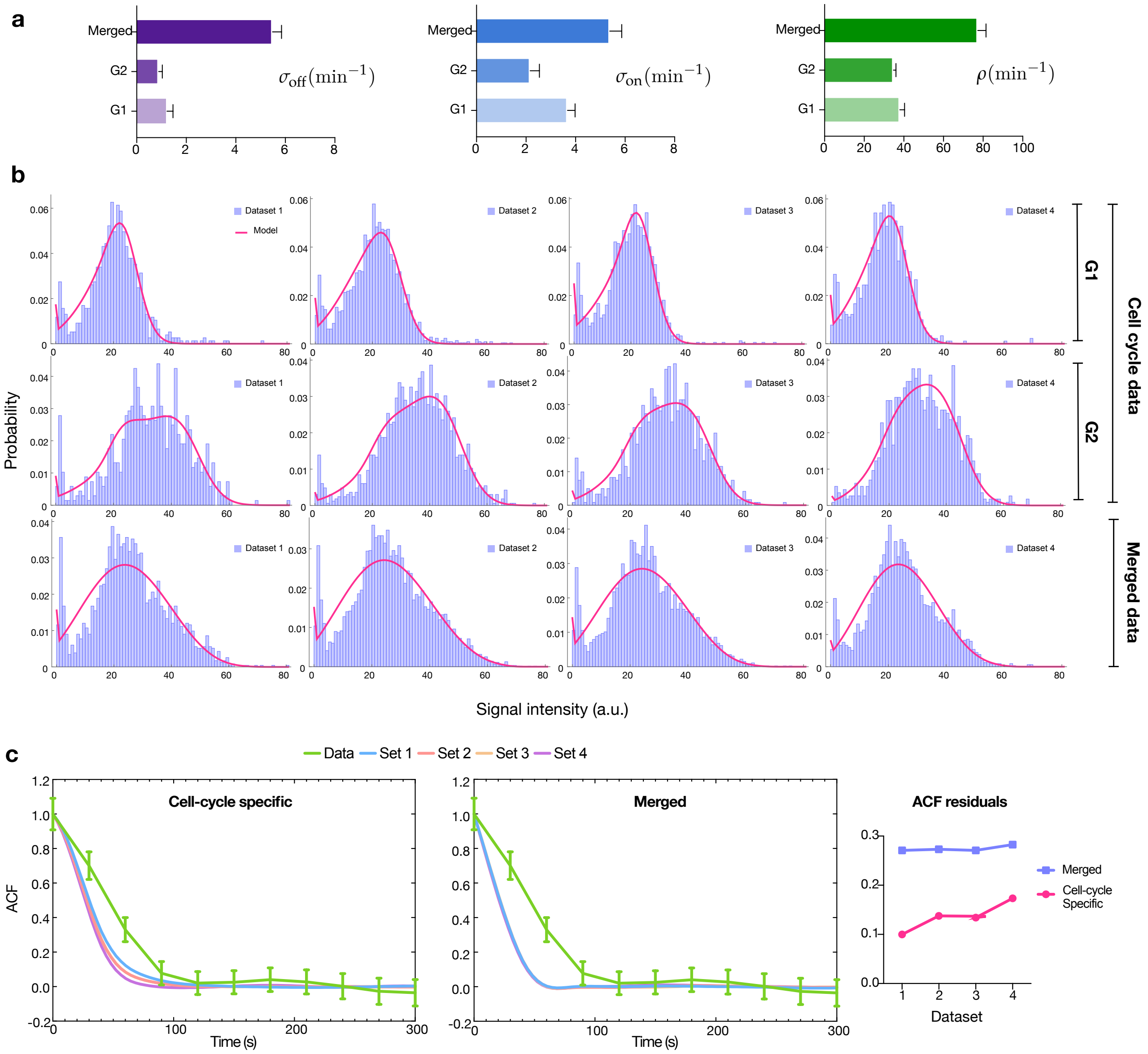


Figure 4: Inference from the distribution of the normalised intensity of the brightest nuclear spot (nascent mRNA data) constructed by merging all data or else specific to the cell cycle phases G1 and G2. **a.** Comparison of the three inferred parameter values using merged and cell-cycle specific data. The bar graph shows the mean and the standard deviation of the estimates calculated over the four data sets. **b.** Best fit distributions for G1 and G2 (top row and middle row respectively) and merged data (bottom row) for data sets 1-4 (from left to right). **c.** Normalised ACF plots of merged and cell-cycle specific data (middle and left) and their residuals (right). The ACF plots are generated by stochastic simulations using estimated parameters from merged and cell-cycle specific nascent mRNA data for each of the four data sets; these are compared with the ACF measured directly using live-cell data in [6] (green line). We also compare the sum of squared ACF residuals of merged and cell-cycle specific data from each dataset (this is the sum of squared deviations between the measured and estimated normalised ACF where the sum is calculated over all time points).

Nascent		σ_{off}	σ_{on}	ρ	Burst size	f_{ON}	Fano Factor
G1	Set 1	1.11	3.76	37.83	34.10	0.77	4.15
	Set 2	1.53	3.94	41.33	27.06	0.72	4.50
	Set 3	0.95	3.23	36.79	38.56	0.77	3.94
	Set 4	1.28	3.76	36.04	28.09	0.75	3.88
	Mean	1.22	3.67	38.00	31.95	0.75	4.12
	Std	0.25	0.31	2.34	5.39	0.02	0.28
G1	Set 1	0.74	1.69	35.00	47.30	0.70	5.32
	Set 2	0.82	2.18	36.30	44.37	0.73	4.51
	Set 3	0.91	2.19	34.54	37.90	0.71	4.57
	Set 4	1.08	2.61	33.27	30.76	0.71	4.04
	Mean	0.89	2.17	34.78	40.08	0.71	4.61
	Std	0.15	0.38	1.25	7.35	0.01	0.53
Merged	Set 1	5.23	5.07	77.38	14.81	0.49	6.72
	Set 2	5.71	5.17	82.86	14.51	0.47	6.95
	Set 3	5.12	5.18	76.15	14.87	0.50	6.45
	Set 4	5.87	6.10	72.83	12.41	0.51	5.54
	Mean	5.48	5.38	77.30	14.15	0.49	6.41
	Std	0.36	0.49	4.17	1.17	0.02	0.62

Table 1: Estimated parameters from the distribution of the normalised intensity of the brightest nuclear spot (nascent mRNA data) constructed by merging all data or else specific to the cell cycle phases G1 and G2. The elongation time τ is estimated to be 0.785 mins, based on measurements of the elongation speed.

3.4.2 Correcting for experimental artefacts

Although inference on cell cycle seperated data outperformed inference on merged data, we noticed that the corresponding best fit distributions did not match well to the experimental signal distributions in the lower bins (Fig. 4b). In all cases, the experimental distributions show high intensities in bins 1, 2, and 3, which is likely an artifact of the experimental data acquisition system. Since we define the transcription site as the brightest spot, that means that if in reality there is no transcription site, we confuse a mature transcript with a nascent transcript. We therefore investigated two methods to correct for this, the "fusion" method and the "rejection" method.

The rejection method removes all data associated with the first k bins of the experimentally obtained histogram of fluorescent intensities (and renormalises afterwards). We find that the parameter estimates vary strongly when the number of bins from which data is rejected (k) is changed (Fig. 5a). Although the distributions fit well to the experimental histograms (Fig. 5b), comparison with the live-cell normalized ACF indicates that the estimates actually become worse than non-curated estimates, with a higher sum of squared residuals (Fig. 5c). The rejection method therefore does not produce reliable estimates.

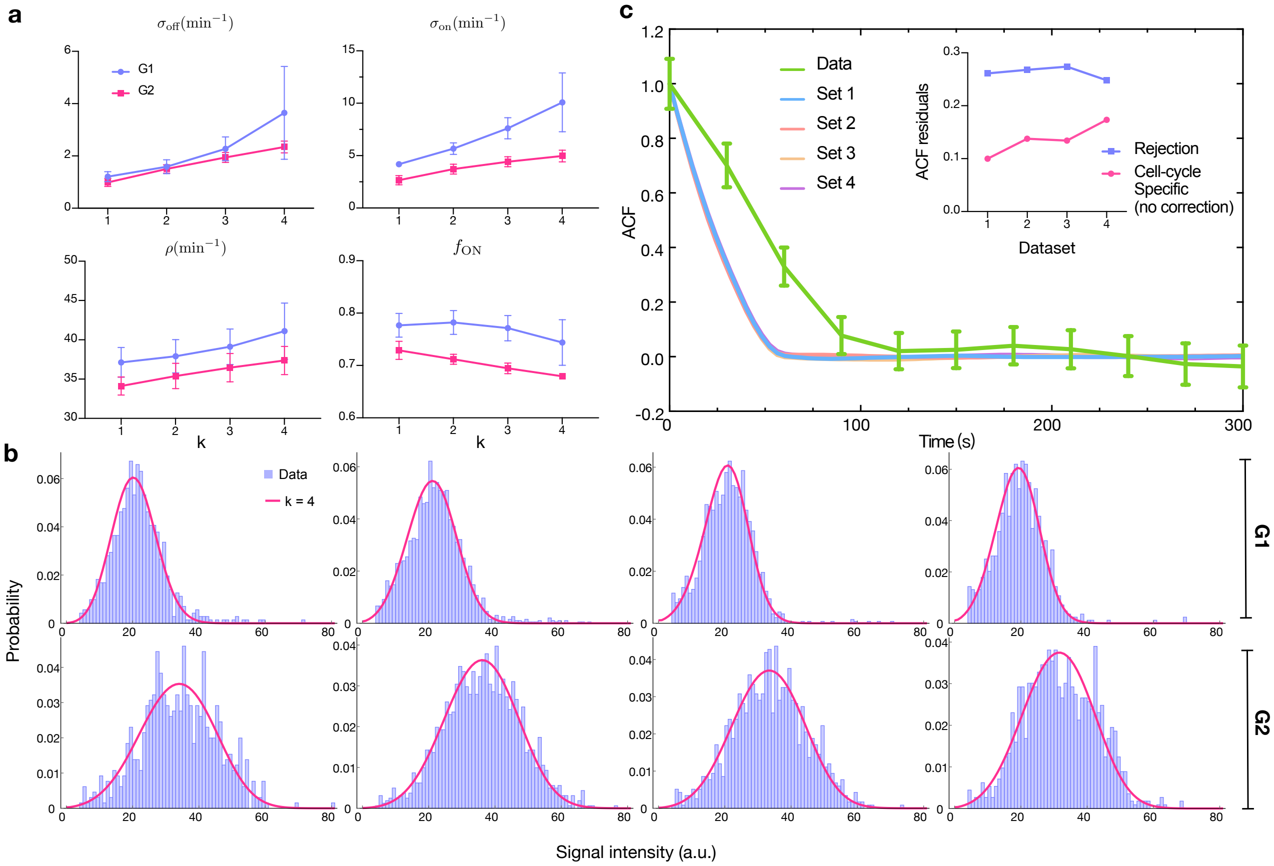
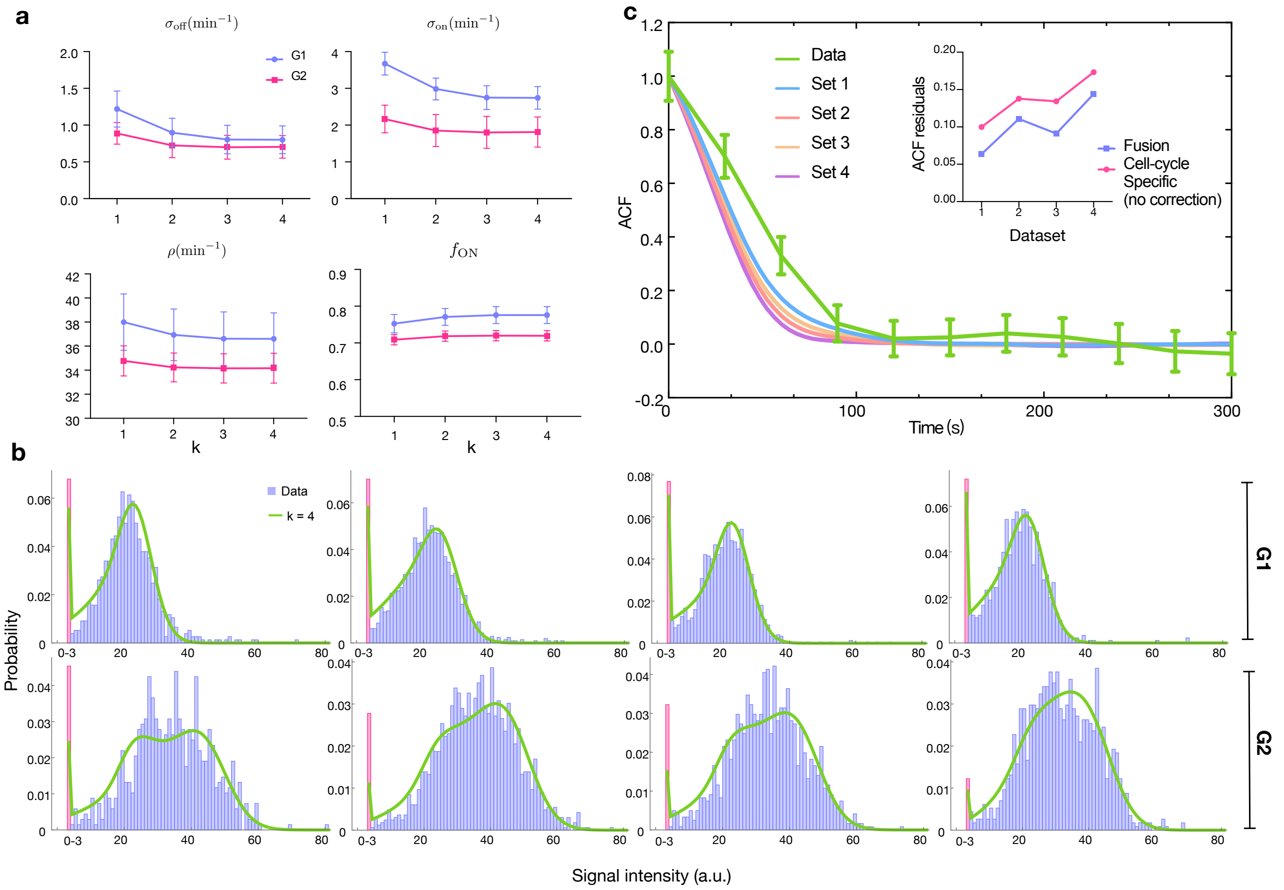


Figure 5: Inference results using rejection method. **a.** Estimated parameters (mean values and standard deviation error bars) by rejecting the first k bins with $k = 1, 2, 3, 4$ where the fraction of ON time $f_{\text{ON}} = \sigma_{\text{on}}/(\sigma_{\text{off}} + \sigma_{\text{on}})$. **b.** Corresponding distributions for G1 (top row) and G2 (bottom row) with the rejection method (only the distributions for $k = 4$ are shown). The estimated parameters are listed in in SI Table 5. **c.** Normalised autocorrelation function (ACF) predicted by stochastic simulations using the estimated parameters (for $k = 4$) for each of the four data sets versus that measured directly using live-cell data (green line). Inset shows the sum of squared residuals of the ACF plots.

Next we considered another data curation method which we call the fusion method. This works by setting to zero all fluorescent intensities in a cell population which are below a certain threshold. In other words, we fuse or combine the first k bins of the experimentally obtained histogram of fluorescent intensities, thereby taking into account that the true intensity of bin 0 is artificially distributed over some of the first bins.

Fig. 6a shows that the fusion method leads to estimates that vary little with k which enhances our degree of confidence in them (note that $k = 1$ is the same as the uncurated data). The peak at the zero bin for both G1 and G2 is better captured using the fusion method than using non-curated data (compare Fig. 4b and Fig. 6b). Comparison to the autocorrelation function of the live-cell data shows that correction with the fusion method also leads to improved transcriptional estimates, as indicated by a reduction in the sum of the squared residuals for all 4 data sets (Fig. 6c).



d **Table: the inferred parameters using nascent RNA data with fusion method**

Fusion		σ_{off}	σ_{on}	ρ	Burst Size	f_{ON}
G1	Set 1	0.69	2.73	36.30	52.85	0.80
	Set 2	1.05	3.09	39.69	37.71	0.75
	Set 3	0.63	2.35	35.74	56.78	0.79
	Set 4	0.83	2.79	34.68	41.57	0.77
	Mean	0.80	2.74	36.60	47.23	0.78
	Std	0.19	0.31	2.17	9.04	0.02
G2	Set 1	0.56	1.34	34.45	61.05	0.70
	Set 2	0.66	1.86	35.73	54.11	0.74
	Set 3	0.67	1.72	33.70	50.56	0.72
	Set 4	0.92	2.33	32.79	35.48	0.72
	Mean	0.70	1.81	34.17	50.30	0.72
	Std	0.15	0.41	1.24	10.80	0.01

Figure 6: Inference results using the fusion method. **a.** Estimated parameters (mean values and standard deviation error bars) by combining the first k bins with $k = 1,2,3,4$ where the fraction of ON time $f_{\text{ON}} = \sigma_{\text{on}}/(\sigma_{\text{off}} + \sigma_{\text{on}})$. **b.** Corresponding distributions for G1 (top row) and G2 (bottom row) with the fusion method (only the distributions for $k = 4$ are shown). The red bar represents the combined bin 0-3 when $k = 4$. **c.** Normalised autocorrelation function (ACF) predicted by stochastic simulations using the estimated parameters (for $k = 4$) for each of the four data sets versus that measured directly using live-cell data (green line). Inset shows the sum of squared residuals of the ACF plots. **d.** Estimated parameters of cell cycle specified data and merged data of nascent mRNAs with fusion method with k = 4 (fusing bins 0-3). These correspond to the fitted distributions in **b.** The elongation time τ is fixed to 0.785 mins. See the inferred parameters in SI Table 6 for all other values of k.

Overall, we conclude that for inferring parameters from the smFISH data, the optimal method is to use nascent cell-cycle specific data, corrected by fusion method. The optimally inferred parameters for the four data sets in our study are those given in Fig 6d. In SI Section 7, we use the profile likelihood method to obtain the 95% confidence intervals of each of the estimated parameters.

4 Discussion

In this study, we compare the reliability of transcriptional parameter interference from mature and nascent mRNA distributions, with and without taking into account the cell cycle stage. Although these distributions come from the same experiment, we find that the different fits produce very different parameter estimates, ranging from non-bursting to small bursts to very large bursts. Comparison to live-cell data reveals that the optimal inference method is to use nascent mRNA data that is separated by cell cycle.

Our findings illustrate the risk of inferring transcriptional parameters from fitting of mRNA distributions. The commonest method of parameter inference in literature is fitting of mature mRNA distributions that are not separated by cell cycle [2, 10, 22]. Obtaining such distributions is straightforward using methods such as smFISH, where one can directly count the number of mRNAs per cell. Additionally, with the advance of single-cell mRNA sequencing technologies, it is possible to obtain mRNA distributions for many genes simultaneously and it is tempting to use these to estimate bursting behaviour across the genome [2, 13]. However, our comparisons on the same dataset show that the values obtained from mature mRNA fits can be significantly different from the real values, with underestimation of the burst sizes of more than 25-fold and underestimation of the active fraction of more than 4-fold. Such large inaccuracies indicate that parameter inference of mature mRNA data should be treated with caution.

It is more common to fit mature distributions rather than nascent distributions because nascent distributions are technically more challenging to obtain. As nascent single-cell sequencing methods are still in the early phase [51], the only method available so far for nascent measurements is smFISH [37]. In such smFISH experiments, intronic probes can be used to specifically label nascent RNA, although there may be some effects of splicing kinetics on the distribution [52]. If introns are not present, like for most yeast genes, one can use exonic probes instead [22]. Since exonic probes label both nascent and mature mRNA transcript, it may be challenging to identify the nascent transcription site unambiguously, especially at lower transcription levels. We show in this manuscript that the fusion method can correct for this bias by combining bins below k RNAs, which results in an improvement of the parameter estimates.

Our analysis also emphasizes the importance of separately analyzing G1 and G2 cells [26]. It is important to note that for cell-cycle-specific analysis, experimental adjustments or cell-cycle synchronized cultures are not required. Although asynchronous cultures consist of a mix G1, S and G2 cells, the integrated DNA intensity of the nucleus of each cell, for example from a DAPI signal, can be used to separate these cells by cell cycle stage *in silico* [26, 53]. As most smFISH experiments already include a DNA-labelled channel, adding an extra analysis step should in principle not limit the incorporation of this step in future smFISH fitting procedures.

Even with our optimal fitting strategy, there is a residual error of the simulated ACF and the measured ACF from live-cell measurements. This difference may be the result of different experimental biases of the two measurements. For example, live-cell measurements have a detection threshold below which RNAs may not be detected. In addition, live-cell measurements include cells in S phase, which are excluded in smFISH. There could also be differences in the exact percentage of G1 and G2 cells, or other extrinsic noise sources between live-cell and smFISH experiments. Alternatively, the fit may be imperfect because there might be parameter sets, others than the ones which our inference algorithm found, which provide an accurate fit of the nascent mRNA distribution and perhaps an even better fit to the ACF than we found. We cannot exclude this possibility because we estimated f_{ON} to be $0.7 - 0.8$ and using synthetic data we showed that the inference algorithm performed best when f_{ON} was about 0.5, and its accuracy deteriorated as f_{ON} approached 0 or 1. Another factor which could explain the residual error between the simulated ACF and the measured ACF is that perhaps the two-state model may be too simplistic to cover the true promoter states in living cells and may therefore not be able to describe the true *in vivo* kinetics. Nevertheless, given that there is no explicit time component in smFISH data, the closeness of the simulated ACF to the measured

ACF provides confidence we are close to the real values.

The optimal parameter set (Fig 6d) indicates long ON promoter times of 75s, during which almost 50 RNAs are produced in a burst. Large burst sizes (> 70) have been previously reported for mouse embryonic stem cells [26], mouse hepatocytes [54] and human fibroblasts [2]. The large burst size and high active fraction of 0.78 suggests that *GAL10* expression is reaching its limit of maximal expression, which may not be surprising as it is already one of the most highly expressed genes in yeast. It is also interesting to note that the ON time of 75s is longer than the residence time of a single transcript (47s), which means that RNA polymerases in the beginning of a burst have already left the locus before the burst has finished.

The optimal parameter set (Fig 6d) also indicates gene dosage compensation. Specifically the burst frequency per gene copy (σ_{on}) in the G2 phase is 0.66 that in the G1 phase; the other transcriptional rates are not significantly different between the two cell cycle phases. The fold change in the burst frequency per gene copy was previously estimated for the *Oct*4 and *Nanog* genes to be 0.63 and 0.71 respectively, in mouse embryonic stem cells [26]. The similarity of our estimate of the fold change to those previously measured could be explained by the results of a recent study [55]; using a detailed model of gene expression, it was shown that in the absence of a dependence of the initiation rate on cell volume, gene dosage compensation optimally leads to approximate mRNA concentration homeostasis when the fold change in the burst frequency upon DNA replication is $\sqrt{2}/2 \approx 0.71$.

In conclusion, obtaining kinetic information from static distributions can introduce biases. However, we show that it is possible to obtain reliable estimates that agree with live-cell measurements, if one infers parameters from nascent mRNA distributions that are accounted for cell cycle stage.

Acknowledgments

Z.C., X.F. and L.X. acknowledge the support from Natural Science Foundation of China (NSFC No. 61988101, 62073137). R.G was supported by a Leverhulme Trust research award (RPG-2020-327). T.L.L was supported by the Netherlands Organization for Scientific Research (NWO, gravitation program CancerGenomiCs.nl), Oncode Institute, which is partly financed by the Dutch Cancer Society, and the European Research Council (ERC Starting Grant 755695 BURSTREG).

References

[1] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).

[2] Larsson, A. J. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

[3] McKnight, S. L. & Miller Jr, O. L. Electron microscopic analysis of chromatin replication in the cellular blastoderm drosophila melanogaster embryo. *Cell* **12**, 795–804 (1977).

[4] Nicolas, D., Phillips, N. E. & Naef, F. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems* **13**, 1280–1290 (2017).

[5] Tunnacliffe, E. & Chubb, J. R. What is a transcriptional burst? *Trends in Genetics* **36**, 288–297 (2020).

[6] Donovan, B. T. *et al.* Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *The EMBO journal* **38**, e100809 (2019).

[7] Brouwer, I., Patel, H. P., Meeussen, J. V. W., Pomp, W. & Lenstra, T. L. Single-molecule fluorescence imaging in living saccharomyces cerevisiae cells. *STAR protocols* **1**, 100142 (2020).

[8] Lenstra, T. L. & Larson, D. R. Single-molecule mrna detection in live yeast. *Current protocols in molecular biology* **113**, 14–24 (2016).

[9] Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theoretical Population Biology* **48**, 222–234 (1995).

[10] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mrna synthesis in mammalian cells. *PLoS Biol* **4**, e309 (2006).

[11] Wang, Y. *et al.* Precision and functional specificity in mrna decay. *Proceedings of the National Academy of Sciences* **99**, 5860–5865 (2002).

[12] Herzog, V. A. *et al.* Thiol-linked alkylation of rna to assess expression dynamics. *Nature methods* **14**, 1198–1204 (2017).

[13] Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome biology* **14**, 1–12 (2013).

[14] Zhou, T. & Zhang, J. Analytical results for a multistate gene model. *SIAM Journal on Applied Mathematics* **72**, 789–818 (2012).

[15] Ham, L., Schnoerr, D., Brackston, R. D. & Stumpf, M. P. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics* **152**, 144106 (2020).

[16] Cao, Z., Filatova, T., Oyarzún, D. A. & Grima, R. A stochastic model of gene expression with polymerase recruitment and pause release. *Biophysical Journal* **119**, 1002–1014 (2020).

[17] Dattani, J. & Barahona, M. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *Journal of The Royal Society Interface* **14**, 20160833 (2017).

[18] Ham, L., Brackston, R. D. & Stumpf, M. P. Extrinsic noise and heavy-tailed laws in gene expression. *Physical review letters* **124**, 108101 (2020).

[19] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences* **117**, 4682–4692 (2020).

[20] Singh, A. & Bokes, P. Consequences of mrna transport on stochastic variability in protein levels. *Biophysical journal* **103**, 1087–1096 (2012).

[21] Perez-Carrasco, R., Beentjes, C. & Grima, R. Effects of cell cycle variability on lineage and population measurements of messenger rna abundance. *Journal of the Royal Society Interface* **17**, 20200360 (2020).

[22] Zenklusen, D., Larson, D. R. & Singer, R. H. Single-rna counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology* **15**, 1263–1271 (2008).

[23] Larson, D. R., Singer, R. H. & Zenklusen, D. A single molecule view of gene expression. *Trends in cell biology* **19**, 630–637 (2009).

[24] Xu, H., Skinner, S. O., Sokac, A. M. & Golding, I. Stochastic kinetics of nascent rna. *Physical review letters* **117**, 128101 (2016).

[25] Jiang, Q. *et al.* Neural network aided approximation and parameter inference of non-markovian models of gene expression. *Nature communications* **12**, 1–12 (2021).

[26] Skinner, S. O. *et al.* Single-cell analysis of transcription kinetics across the cell cycle. *eLife* **5**, 1–24 (2016).

[27] Xu, H., Sepúlveda, L. A., Figard, L., Sokac, A. M. & Golding, I. Combining protein and mrna quantification to decipher transcriptional regulation. *Nature methods* **12**, 739–742 (2015).

[28] Zoller, B., Little, S. C. & Gregor, T. Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting. *Cell* **175**, 835–847 (2018).

[29] Senecal, A. *et al.* Transcription factors modulate c-fos transcriptional bursts. *Cell reports* **8**, 75–83 (2014).

[30] Zopf, C., Quinn, K., Zeidman, J. & Maheshri, N. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS computational biology* **9**, e1003161 (2013).

[31] Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. *Molecular cell* **58**, 339–352 (2015).

[32] Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).

[33] Munsky, B. & Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys* **124**, 044104 (2006).

[34] Barrio, M. *et al.* Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Computational Biology* **2**, 1017–1030 (2006).

[35] Trcek, T. *et al.* Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature Protocols* **7**, 408–419 (2012).

[36] Lenstra, T. L., Coulon, A., Chow, C. C. & Larson, D. R. Single-Molecule Imaging Reveals a Switch between Spurious and Functional ncRNA Transcription. *Molecular Cell* **60**, 597–610 (2015).

[37] Patel, H. P., Brouwer, I. & Lenstra, T. L. Optimized protocol for single-molecule rna fish to visualize gene expression in s. cerevisiae. *STAR protocols* **2**, 100647 (2021).

[38] Crocker, J. C. & Grier, D. G. Methods of digital video microscopy for colloidal studies. *Journal of Colloid and Interface Science* **179**, 298–310 (1996).

[39] Thompson, R. E., Larson, D. R. & Webb, W. W. Precise nanometer localization analysis for individual fluorescent probes. *Biophysical Journal* **82**, 2775–2783 (2002).

[40] Larson, D. R., Johnson, M. C., Webb, W. W. & Vogt, V. M. Visualization of retrovirus budding with correlated light and electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15453–15458 (2005).

[41] Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332**, 475–478 (2011).

[42] Coulon, A. *et al.* Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife* **3**, e03939 (2014).

[43] Johnston, D. A., White, R. A. & Barlogie, B. Automatic processing and interpretation of dna distributions: comparison of several techniques. *Computers and Biomedical Research* **11**, 393–404 (1978).

[44] Szavits-Nossan, J. & Grima, R. Mean-field theory accurately captures the variation of copy number distributions across the mrna's life cycle. *bioRxiv* (2021).

[45] Choubey, S., Kondev, J. & Sanchez, A. Deciphering transcriptional dynamics in vivo by counting nascent rna molecules. *PLoS computational biology* **11**, e1004345 (2015).

[46] Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single rna transcripts in situ. *Science* **280**, 585–590 (1998).

[47] Raj, A., Van Den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A. & Tyagi, S. Imaging individual mrna molecules using multiple singly labeled probes. *Nature methods* **5**, 877–879 (2008).

[48] Skinner, S. O., Sepúlveda, L. A., Xu, H. & Golding, I. Measuring mrna copy number in individual escherichia coli cells using single-molecule fluorescent in situ hybridization. *Nature protocols* **8**, 1100–1113 (2013).

[49] Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**, 1188–1193 (2013).

[50] Voichek, Y., Bar-Ziv, R. & Barkai, N. Expression homeostasis during dna replication. *Science* **351**, 1087–1090 (2016).

[51] Hendriks, G.-J. *et al.* Nasc-seq monitors rna synthesis in single cells. *Nature communications* **10**, 1–9 (2019).

[52] Wan, Y. *et al.* Dynamic imaging of nascent rna reveals general principles of transcription dynamics and stochastic splice site selection. *Cell* **184**, 2878–2895 (2021).

[53] Roukos, V., Pegoraro, G., Voss, T. C. & Misteli, T. Cell cycle staging of individual cells by fluorescence microscopy. *Nature protocols* **10**, 334–348 (2015).

[54] Halpern, K. B. *et al.* Bursty gene expression in the intact mammalian liver. *Molecular cell* **58**, 147–156 (2015).

[55] Jia, C., Singh, A. & Grima, R. Concentration fluctuations due to size-dependent gene expression and cell-size control mechanisms. *bioRxiv* (2021).