Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions

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

August 5, 2022

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. Whilst it is often presumed that post-transcriptional noise and gene copy number variation affect transcriptional parameter estimation, the size of the error introduced remains unclear. To address this issue, 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 phase. We infer transcriptional parameters from mature and nascent mRNA distributions, with and without accounting for cell cycle phase and compare the results to live-cell transcription measurements of the same gene. We find that: (i) correcting for cell cycle dynamics decreases the promoter switching rates and the initiation rate, and increases the fraction of time spent in the active state, as well as the burst size; (ii) additional correction for post-transcriptional noise leads to further increases in the burst size and to a large reduction in the errors in parameter estimation. Furthermore, we outline how to correctly adjust for measurement noise in smFISH due to uncertainty in transcription site localisation when introns cannot be labelled. Simulations with parameters estimated from nascent smFISH data, which is corrected for cell cycle phases and measurement noise, leads to autocorrelation functions that agree with those obtained from live-cell imaging.

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

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

low-throughput and require genome-editing [7, 8]. To circumvent this, one can exploit the fact that bursting creates heterogeneity in a population. In this case, it is relatively straightforward to obtain a steady-state distribution 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 in eukaryotic cells [11, 12]. The solution of the telegraph model for the steady-state distribution of mRNA numbers has 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]. One 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. A nascent mRNA [22, 23] is an 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. Fluctuations in nascent mRNA numbers thus directly reflect 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].

However, nascent mRNA data still suffers from other sources of noise due to cell-to-cell variability. For example in an 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 cycle is illustrated by the finding [25] that noisy transcription from the synthetic TetO promoter in *S. cerevisiae* is dominated by its dependence on the cell cycle. The estimation of transcriptional parameters from nascent mRNA data for pre- and post-replication phases of the cell cycle has, to the best of our knowledge, only been reported in [26].

Interestingly, all of the studies that estimate transcriptional parameters from nascent mRNA data [26–30] do not compare them with transcriptional parameters estimated from cellular (mature) mRNA data measured in the same experiment. Similarly, a quantitative comparison between inference from cell-cycle specific data and data which contains information from all cell cycle phases is lacking. Likely, this is because it is considered evident that quantifying fluctuations earlier in the gene expression process and adjusted for the cell-cycle will improve estimates. However, nascent mRNA distributions are technically more challenging to acquire than mature mRNA distributions; and inference from nascent mRNA distributions is substantially more complex [24]. Thus, it still needs to be shown that acquiring nascent mRNA data is a necessity from a parameter inference point of view, i.e. that it leads to significantly different and more robust estimates than using mature mRNA data. We also note that current studies report parameter inference from organisms where it is possible to label introns to identify mRNA located at the transcription site. This is not possible in many yeast genes and other microorganisms, and in these cases it is unclear how to correct parameter estimates for uncertainty in the transcription site location.

In this paper, we sought to understand the precise 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) were first tested on simulated data, where limitations of the algorithms were uncovered in accurately estimating the transcriptional parameters in certain regions of parameter space. The algorithms were 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 allowed us to separate the influence of ignoring cell cycle variability from that of post-transcriptional noise (mature vs nascent mRNA data). We found that only fitting of nascent cell-cycle data, corrected for measurement noise (due to uncertainty in the transcription site location), provided good agreement with measurements from live-cell data. Cell-cycle specific analysis also revealed 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 Results

2.1 Inference from mature mRNA data vs inference from nascent mRNA data: testing inference accuracy using synthetic data

To understand the accuracy of the inference algorithms from nascent and mature mRNA data, in various regions of parameter space, (i) we generated synthetic data using stochastic simulations with certain known values of the parameters; (ii) applied the inference algorithms to estimate the parameters from the synthetic data; (iii) compared the true and inferred kinetic parameter values.

The generation of synthetic mature mRNA data (mature mRNA measurements in each of 10^4 cells) using stochastic simulations of the telegraph model (Fig. 1a) is described in Methods Sections 4.1.1 and 4.1.2. The inference algorithm is described in detail in Methods Section 4.1.3. 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 of mature mRNA numbers from a given cell is given by evaluating the telegraph model's steady-state mature mRNA count probability distribution.

For nascent RNA data, we used stochastic simulations of the delay telegraph model (Fig. 1b) to generate the position of bound Pol II molecules from which we constructed the synthetic smFISH signal in each of 10^4 cells (Methods Section 4.2.2). An inference algorithm estimates the parameters, based on a maximization of the likelihood of observing the single cell total fluorescence intensity measured in a population of cells (Methods Section 4.2.3). Note that the likelihood of observing a certain fluorescence signal intensity from a cell is given by extension of the delay telegraph model (but not directly by the delay telegraph model itself) to account for the smFISH probe positions.

This extension takes into account that the experimental fluorescence data used in this manuscript was acquired from smFISH of *PP7-GAL10* in budding yeast, where probes were hybridized to the PP7 sequences. Because the PP7 sequences are positioned at the 5′ of the *GAL10* gene, the fluorescence intensity of a single mRNA on the DNA locus resembles a trapezoidal pulse (see Fig. 1c for an illustration). As the Pol II molecule travels through the 14 repeats of the PP7 loops, the fluorescence intensity increases as the fluorescent probes binds to the nascent mRNA (this is the linear part of the trapezoidal pulse). However, once all 14 loops on the nascent mRNA are bound by the fluorescent probes, the intensity of a single mRNA reaches maximal intensity and the plot plateaus as the RNA elongates through the *GAL10* gene body before termination and release. The total fluorescent signal density function is hence given by

$$p(s;\theta) = \sum_{k=0}^{\infty} p(s|k)P(k;\theta), \tag{2.1}$$

where $p(s|k)$ is the density function of the signal s 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 4.2.1 we show how $p(s|k)$ can be approximately calculated for the trapezoidal pulse. Hence Eq. (2.1) represents the extension of the delay telegraph model to predict the smFISH fluorescent signal of the transcription site.

Note that both of these inference algorithms were used to infer the promoter switching and initiation rate parameters. The degradation rate and the elongation time were not estimated but assumed to be known. The inference and synthetic data generation procedures are summarised and illustrated in Fig. 1d.

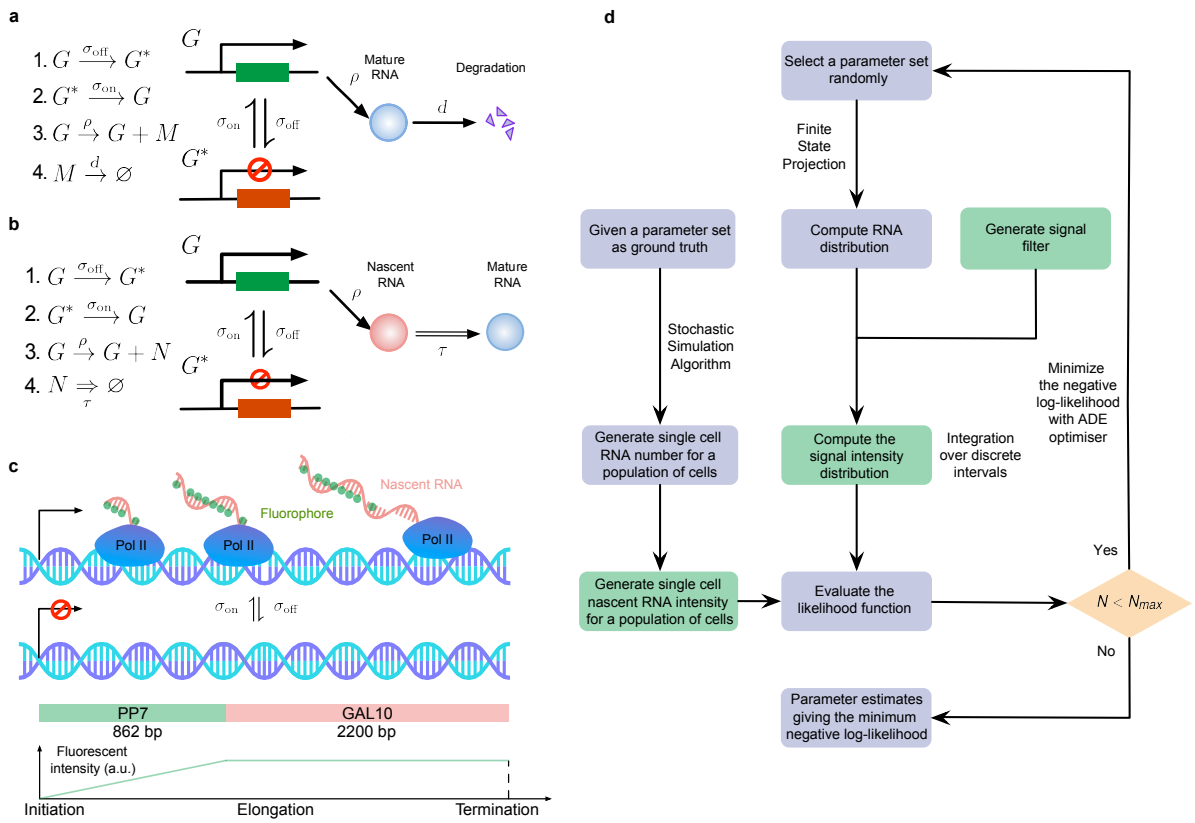


Figure 1: a. A schematic illustration of the telegraph model. **b.** A schematic of the delay telegraph model. The double horizontal line for nascent mRNA removal indicates this is a delayed reaction. **c.** Illustration showing promoter switching between two states, Pol II binding to the promoter in the ON state and subsequently undergoing productive elongation. Note that the length of the nascent mRNA tail increases until Pol II terminates at the end of the gene. As Pol II travels through the 14 repeats of the PP7 loops, the intensity of the mRNA increases due to fluorescent probe binding to the mRNA; intensity saturates as Pol II enters the *GAL10* gene body. **d.** Illustration of the algorithms to generate synthetic data and to perform inference from mature and nascent mRNA data. The green boxes are only applicable for the inference of the fluorescence signal intensity of nascent mRNAs; note that in nascent mRNA inference, the "RNA number" in the flow chart should be interpreted as the number of bound Pol II molecules on the gene. A large iteration step $N_{\max}$ ($\geq 10^4$) is chosen as the termination condition for the optimizer.

The accuracy of inference was first calculated as the mean of the relative error in the estimated parameters $\sigma_{\text{off}}, \sigma_{\text{on}}$, and ρ (for its definition see Methods, Eq. (4.5)); note that this error measures deviations from the known ground truth values. Fig. 2a shows, by means of a 3D scatter plot, the ratio of the mean relative error from nascent mRNA data (using delay telegraph model) and the mean relative error from mature mRNA data (using the telegraph model) for 789 independent parameter sets sampled on a grid (for each of these sets, we simulated 10^4 cells). The overall bluish hue of the plot suggested that the mean relative error from nascent mRNA data was typically less than the error from mature mRNA data. This was confirmed in Fig. 2b where the same data was plotted but now as a function of the fraction of ON time (defined as $f_{\text{ON}} = \sigma_{\text{on}}/(\sigma_{\text{off}} + \sigma_{\text{on}})$). Out of 789 parameter sets, for 483 of them ($\approx 61\%$) the inference accuracy was higher when using nascent mRNA data.

Thus far we have implicitly assumed that fluctuations in both nascent and mature mRNA are due to transcriptional bursting. However, it is clear that mature mRNA data exhibit a higher degree of noise due to post-transcriptional processing. For example, it has been shown that transcriptional noise is typically amplified during mRNA nuclear export [32]. In addition, cell-to-cell variation in the number of nuclear pore complexes has recently been identified as the source of heterogeneity in nuclear export rates within isogenic yeast populations [33]. To take into account these additional noise sources, which we call external noise, we added noise to the initiation rate ρ in the telegraph model since this rate implicitly models all processes between the synthesis of the transcript and the appearance of mature mRNA in the cytoplasm. Specifically, for each of the 789 parameter sets previously used, we changed ρ to ρ' where the latter is a log-normal distributed random variable such that its mean is ρ and its standard deviation is equal to 0.05 of the mean (5% external noise). Note that this implies that at the time of measurement, each cell in the population had a different value of the initiation rate. Simulations with this perturbed set of parameters led to a synthetic mature mRNA data set from which we re-inferred parameters using the telegraph model. In Fig. 2c we show the ratio of mean relative error from nascent mRNA data and the mean relative error from perturbed mature mRNA data as a function of the fraction of ON time, f_{ON}. The percentage of parameters where nascent mRNA is more accurate is slightly increased compared to the data without noise (64% versus 61% of the parameters) (compare Fig. 2c and Fig. 2b). However, the addition of even more noise (10% external noise added to the initiation rate) increases the inference accuracy for 91% of the parameter sets when the nascent mRNA data is used (SI Section 1.1 and SI Fig. 1).

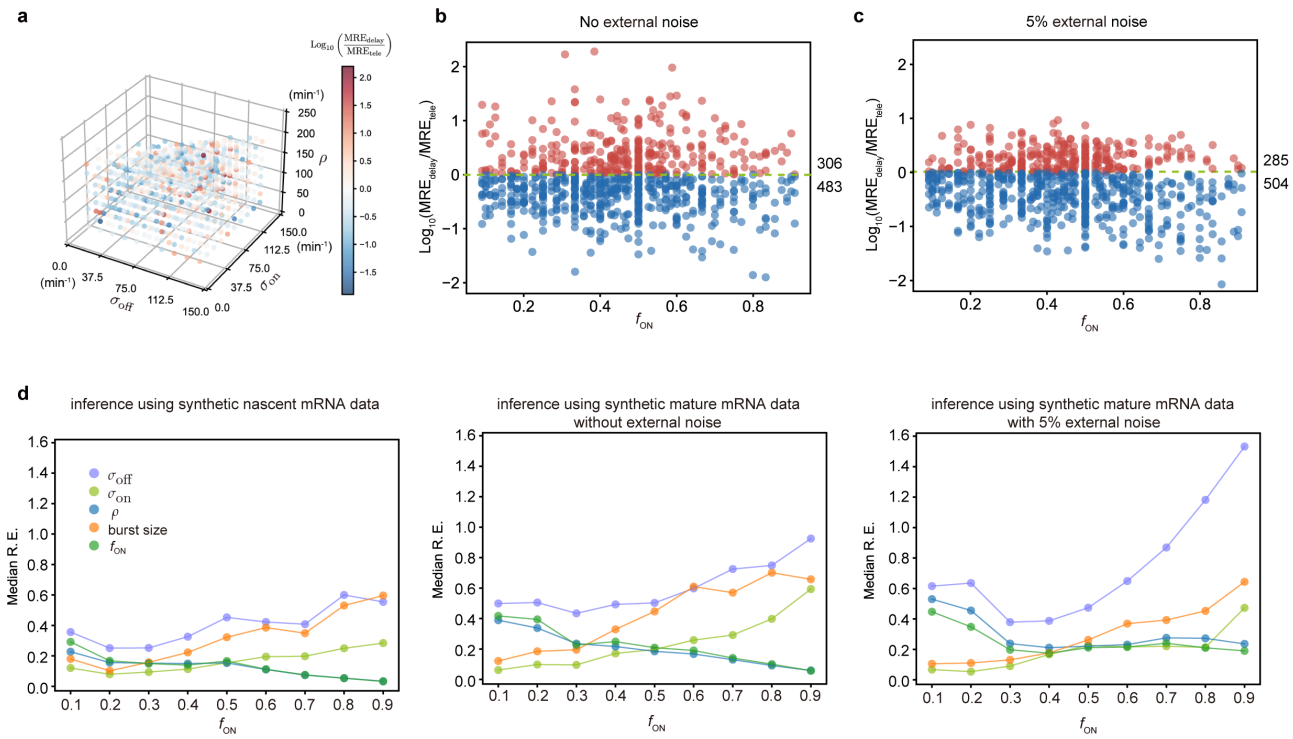


Figure 2: Accuracy of the inferred kinetic parameters from synthetic mature and nascent mRNA data using the telegraph and delay telegraph model, respectively. **a.** 3D scatter plot showing the ratio of the mean relative error from nascent mRNA data (using delay telegraph model $\mathrm{MRE}_{\mathrm{delay}}$) and the mean relative error from mature mRNA data (using the telegraph model $\mathrm{MRE}_{\mathrm{tele}}$) for 789 independent parameter sets sampled on a grid. Red datapoints indicate parameter sets with lower relative errors for mature data compared to nascent data, blue datapoints indicate parameter sets with lower relative error for nascent data compared to mature data **b.** Same data as **(a)** but shown as a function of the fraction of ON time, f_{ON}. For $\approx$ 61% of the parameters, the inference accuracy is higher when using nascent mRNA data. **c.** Sampling from the same parameter space, we then add log-normal distributed noise (size 5%) to the initiation rate ρ (see text for details) to mimic external noise due to post-transcriptional processing that is only present in mature mRNA. Log_{10} of the ratio of the median relative error (MRE) using perturbed mature mRNA data against Log_{10} MRE using nascent mRNA data is shown as a function of the true fraction of ON time, f_{ON}. For $\approx$ 64% of the parameters, the inference accuracy is higher when using nascent mRNA data. **d.** The median relative error of each transcriptional parameter as a function of the fraction of ON time, using synthetic nascent mRNA, synthetic mature mRNA data and synthetic mature mRNA with external noise. Inference from nascent data is generally more accurate than using mature mRNA data.

To obtain more insight into the accuracy of the individual parameters, we next plotted the median relative error of transcriptional parameters $\sigma_{\text{off}}, \sigma_{\text{on}}, \rho$, burst size and the inferred fraction of ON time, as a function of the true fraction of ON time (Fig. 2d). We compared the results using synthetic nascent mRNA, synthetic mature mRNA data and synthetic mature mRNA with 5% external noise. The median of the relative error for each transcriptional parameter (as given by the second equation of Eq. 4.5) was obtained for the subset of the 789 parameter sets for which the true fraction of ON time f_{ON} falls into the interval $[x - 0.05, x + 0.05]$ where $x = 0.1, 0.2, \ldots, 0.9$. From the plots, the following can be deduced: (i) the errors in σ_{on} (the burst frequency), σ_{off} and the burst size ρ/σ_{off} tend to increase with f_{ON} while the rest of the parameters (ρ and the estimated value of f_{ON}) decrease; (ii) for small f_{ON}, the best estimated parameters are the burst frequency and size while for large f_{ON}, it was ρ and the estimated value of f_{ON}. The worst estimated parameter was σ_{off}, independent of the value of f_{ON}; (iii) the addition of external noise to mature mRNA data had a small impact on inference for small f_{ON}; in contrast, for large f_{ON} the noise appreciably increased the relative error in σ_{off} and to a lesser extent the error in the other parameters too.

Additionally, in the SI we show that (i) independent of the accuracy of parameter estimation, the best fit distributions accurately matched the ground truth distributions (SI Section 1.2 and SI Fig. 2); (ii) the parameters ordered by relative error were in agreement with the parameters ordered by sample variability (SI Section 1.3 and SI Table 1) and by profile likelihood error (SI Section 1.4, SI Tables 2 and 3). Since from experimental data, only the sample variability and the profile likelihood error are available, it follows that the results of our synthetic data study in Fig. 2 based on relative error from the ground truth have wide practical applicability; (iii) stochastic perturbation of the mature or nascent mRNA data (due to errors in the measurement of the number of spots and the fluorescent intensity) had little effect on the inference quality, unless the gene spent a large proportion of time in the OFF state (SI Sections 1.5 and 1.6, SI Tables 4 and 5); (iv) if one utilized the conventional telegraph model to fit the nascent data generated by the delay telegraph model, it was possible to obtain a distribution fitting as good as the delay telegraph model but with low-fidelity parameter estimation (SI Section 2, SI Fig. 3 and SI Table 6). Analytically, the telegraph model is only an accurate approximation of the delay telegraph model when the promoter switching timescales are much longer than the time spent by Pol II on a gene or the off switching rates are very small such that gene expression is nearly constitutive.

In summary, by means of synthetic experiments, we have clarified how the accuracy of the parameter inference strongly depends on the type of data (nascent or mature mRNA) and the fraction of time spent in the ON state (which determines the mode of gene expression).

2.2 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 4.3 for details of the data acquisition). Note that in what follows, delay telegraph model refers to the extended delay telegraph model that accounts for the smFISH probe positions that was used to predict the smFISH fluorescent signal of the transcription site.

2.2.1 Inference from mature mRNA data: experimental artifacts

We have four independent datasets from which we determined mRNA count and nascent RNA distributions. Fig. 3a shows an example cell with mature single RNAs in the cytoplasm, and a bright nuclear spot representing the site of nascent transcription. Spots and cell outlines were identified using automated pipelines. Importantly, to obtain an accurate estimation of transcriptional parameters, the experimental input distributions of mRNA count and nascent RNAs require high accuracy. We therefore first determined how technical artifacts in the analysis affects the inference estimates.

First, if the number of mRNA transcripts per cell is high, accurate determination of the number of transcripts may be challenging, as transcripts may overlap. To determine if this occurred in our

datasets, we analyzed the distributions of intensities of the cytoplasmic spots, which revealed unimodal distributions where $\sim$ 90 percent of the detected spots fell in the range 0.5$\times$ median – 1.5$\times$ median (Fig. 4a). We therefore concluded that overlapping spots are not a large confounder in our data, likely because cytoplasmic PP7-*GAL10* RNAs have a high turnover from the addition of the PP7 loops. We note that such high turnover should aid transcriptional parameter estimates, as it closely reflects transcriptional activity.

A second possible source of error is cell segmentation. To test how cell segmentation errors contribute to the mature mRNA distribution and the transcriptional bursting estimates, we compared two independent segmentation tools, where segmentation 1 often resulted in missed spots (Fig. 3b), resulting in an underestimation of the mean mRNA count and of the variance (compare Fig. 3b,c). We inferred the transcriptional parameters using the algorithm described in Methods Section 4.1.3. In the absence of an experimental measurement of the degradation rate, we could only estimate the 3 transcriptional parameters normalised by d. The best fits of dataset 1 are shown in (Fig. 3b,c) and the transcriptional parameters (for all four datasets) are summarized in (Fig. 3e). Note that the estimated parameters for all four datasets, using both segmentations, are shown in SI Table 7 and the associated best fit distributions in SI Figure 4a. Notably the segmentation algorithms led to similar estimates for the burst frequency but considerably different estimates for the rest of the parameters. In particular segmentation 1 suggested that burst expression is infrequent ($\approx$20% of the time) whereas segmentation 2 was consistent with burst expression occurring half of the time. Given that accurate cell segmentation remains challenging, this analysis illustrates that parameter estimation from mature mRNA counts may be affected by technical errors. For the remainder of the mature mRNA analysis, we have used only segmentation 2 data.

Lastly, it may be challenging to distinguish the nascent transcription site from a mature RNA, especially if few nascent RNAs are being produced. Either one can decide to include all cellular spots in the total mRNA count, including the transcription site, with the result that the number of mature transcripts is overestimated with one RNA for cells which show a transcription site. Or conversely, one can decide to exclude the transcription site by subtracting one spot from each cell, with the result that the number of mature mRNAs may be underestimated by one RNA for cells that are transcriptionally silent. To understand how this choice affects the accuracy of parameter inference, we compared both options in (Fig. 3c,d,e), where seg2 included all spots, and seg2-TS excluded transcription sites (by subtracting 1 from each cell). The estimated parameters for all four datasets are shown in SI Table 7 and the associated best fit distributions in SI Figure 4a. Although the mean was lower when transcription sites were excluded, all the parameters except the burst frequency σ_{on} were within the error, indicating that the choice of whether or not to include the transcription site in the mature mRNA count had a small influence on parameter estimation. For the remainder of the analysis, we included all spots, and counted the transcription site as one RNA.

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

The above analysis was performed using the merged data from all cells, irrespective of their position in the cell cycle. The inferred parameters of all 4 datasets are shown in Fig. 3g (grey). To understand the effect of the cell cycle on these parameter estimates, we compared this inference with cell-cycle specific data. We used the integrated nuclear DAPI intensity as a measure for DNA content to classify cells into G1 or G2 cells (Fig. 3f (left)) to obtain separate mature mRNA distributions for G1 and G2 cells.

To infer the transcriptional parameters from mature mRNA data of cells in G1, the inference protocol remained the same. However for cells in the G2 stage, this protocol needed to be altered since G2 cells have 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 was used in steps (ii) and (iii) of the inference algorithm in Methods Section 4.1.3. A difference between our method of estimating

parameters in G2 from that in the literature [26] is that we do not assume that the burst frequency is the only parameter that changes upon replication, and we estimated all transcription parameters simultaneously.

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 generally not possible to resolve the two copies of the allele in G2 because they are within the diffraction limit. However, in the absence of experimental evidence, the independence assumption is the simplest reasonable assumption that we could make (see later for a relaxation of this assumption).

For both G1 and G2 cells, we performed inference for cell-cycle specific mature mRNA data, the results of which are shown in Fig. 3e (right) and Fig. 3f (centre and right) – see SI Table 8 for the confidence intervals of the estimates calculated using profile likelihood. As expected, the mean number of mRNAs in G2 cells was larger than that in G1 cells. For both merged and cell-cycle specific data, the parameters ordered by increasing variability of the estimates from independent samples (the standard deviation divided by the mean) were: ρ, f_{ON}, σ_{ON}, burst size and σ_{OFF}, and the same order was predicted by the relative error (from ground truth values) from our synthetic experiments (compare with $f_{\mathrm{ON}} = 0.50$ and $f_{\mathrm{ON}} = 0.80$ in the middle and right panels of Fig. 2d) and by sample variability (SI Section 1.3). In SI Section 3.3 and SI Table 9 we show that the relaxation of the assumption of independence between the allele copies in G2 (by instead assuming perfect state correlation of the two alleles) had practically no influence on the inference of the two best estimated parameters (ρ, f_{ON}).

A comparison of the two types of data predicted different behaviour (Fig. 3g bottom): merged data indicated behaviour consistent with the gene being ON half of the time and small burst sizes, while cell-cycle specific data implied the gene is ON $\approx$80% of the time with large burst sizes. We note that the burst sizes have considerable sample variability, exemplifying burst size estimates of transcriptional parameters from mature mRNA distributions have to be treated with caution. Nevertheless, in line with this high fraction ON and large burst size, which start to approach constitutive expression, the variation introduced by the transcription kinetics is relatively modest with Fano factors not far from one: 2.43 ± 0.21 for merged data and 1.75 ± 0.45 for cell-cycle data (the slightly higher value for merged data likely was due to heterogeneity stemming from varying gene copy numbers per cell).

Comparing the mean rates between the G1 and G2 phases, we found that σ_{off}, σ_{on}, ρ decreased while f_{ON} and the burst size increased upon replication. However, taking into account the variability in estimates across the four datasets, the only two parameters which were well-separated between the two phases were σ_{on} and ρ. These two parameters decreased by 65% and 21%, respectively, which suggests that upon replication, there are mechanisms at play which reduce the expression of each copy to partially compensate for the doubling of the gene copy number (gene dosage compensation) [26].

In conclusion, what is particularly surprising in our analysis is the differences in the inference results using merged and cell-cycle specific data: the former suggests the gene spends only half of its time in the ON state while the latter implies the gene is mostly in its ON state.

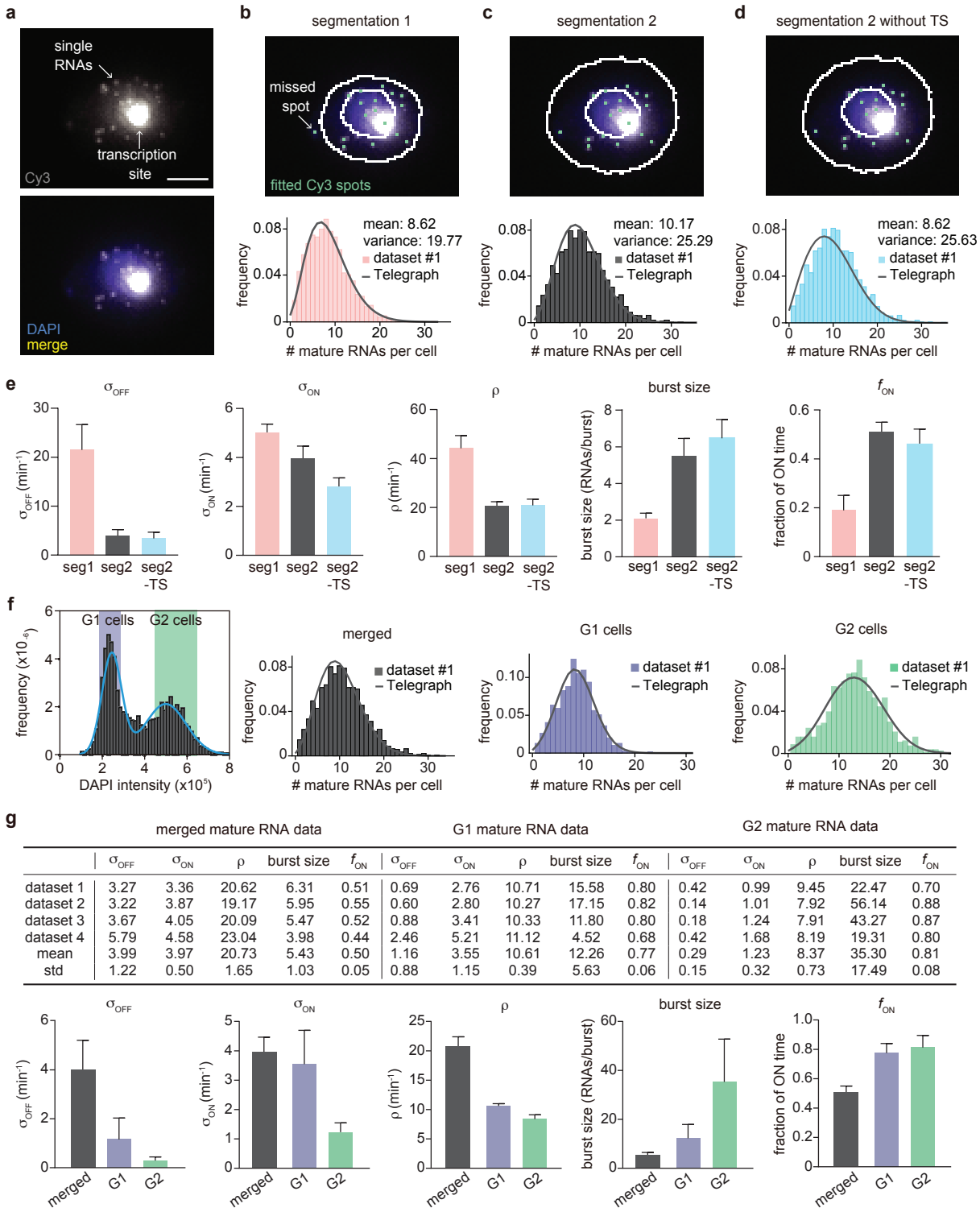


Figure 3: Inference results using four mature mRNA data sets with sample sizes of 2333, 6366, 4550 and 3163 cells, respectively. **a.** Representative smFISH image of a yeast cell with *PP7-GAL10* RNAs labeled with Cy3 and the nucleus labeled with DAPI. **b.** The DAPI and Cy3 signals were used to determine the nuclear and cellular mask, respectively. Detected and fitted spots are indicated in green. Mature RNA count distribution (pink) for segmentation method 1 with a best fit obtained from the telegraph model (gray curve). **c-d.** The DAPI and Cy3 signals were used to determine the nuclear and cellular mask using a second independent segmentation tool (segmentation 2). Mature RNA count distribution (gray and cyan) with/without counting the transcription site (TS) for segmentation method 2 with a best fit obtained from the telegraph model (gray curves). **e.** Bar graphs of inferred transcriptional parameters (merged mature RNA data) from fitting the distributions of the two segmentation methods (”seg1” and ”seg2”) as well as the distribution of mature RNAs only (”seg2 -TS” which indicates the exclusion of one spot in each cell that represents the transcription site). The burst size was computed as ρ/σ_{off} and the fraction of ON time as $\sigma_{\text{on}}/(\sigma_{\text{on}}+\sigma_{\text{off}})$. Error bars indicate standard deviation computed over the four datasets. **f.** Distribution of the integrated DAPI intensity for each cell. Cyan line represents a Gaussian bimodal fit with highlighted regions indicating the intensity-based classification of G1 and G2 cells. Distributions of the mature RNA count for all cells (merged) and cell-cycle classified cells (G1 cells and G2 cells). **g.** Tables and bar graphs of inferred parameters for merged and cell-cycle specific data. Note that the transcriptional parameters $\sigma_{\text{on}}, \sigma_{\text{off}}, \rho$ are normalised by the degradation rate and hence dimensionless. For the cell-cycle specific data, parameters were inferred per gene copy.

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

2.3.1 Cell-cycle specific versus merged data

To determine the number of nascent transcripts at the transcription site, we selected the brightest spot from each nucleus and normalized its intensity to the median intensity of the cytoplasmic spots. As the distribution of intensities of the cytoplasmic mRNAs followed a narrow unimodal distribution, its median likely represents the intensity of a single RNA (orange distribution in the central panel of Fig. 4a). The inference of transcriptional parameters using the merged data was done using the algorithm described in Methods Section 4.2.3.

Similar to above, to account for two gene copies in G2 cells, we assumed that the transcriptional activities of the two gene copies are independent. The distribution of the total fluorescent signal from both gene copies was the convolution of the signal distribution (obtained from the extended delay telegraph model, i.e. Eq. (2.1)) with itself. This convolved distribution was 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 1c), the elongation time (τ in our model) is $\approx 47.1\,s \approx 0.785$ min. The fixed elongation rate enabled us to infer the absolute values of the three transcriptional parameters $\sigma_{\text{off}}, \sigma_{\text{on}}$ and ρ.

Best fits of the extended delay telegraph model to the distribution of signal intensity of nascent mRNAs at the transcription site are shown in Fig. 4a, b for dataset 1; for the other datasets see SI Fig 5. The corresponding estimates of the transcriptional parameters are shown in SI Table 10 and also illustrated by bar charts in Fig. 4c. The confidence intervals of the transcriptional parameters (computed using the profile likelihood method) are shown in SI Table 11.

Comparing this estimation with that from mature mRNA, we observed that in both cases $f_{\text{ON}} \approx 0.5$ for merged data and in the range $0.7 - 0.8$ for cell-cycle specific data. Also in both cases, the Fano factors of merged data were larger than those of cell-cycle specific data. Hence, we are confident that not accounting for the cell cycle phase leads to an over-estimation of the time spent in the OFF state and of the Fano factor. In addition, comparing the burst sizes in Fig 3g and SI Table 10, we found that not taking into account post-transcriptional noise (by using mature mRNA data) led to an lower estimation of the burst size (2.6 fold, 2.6 fold and 1.1 fold lower for inference from merged, G1 and G2 data, respectively). We note that it would be useful to directly compare the absolute estimates of the other transcriptional parameters from mature and nascent mRNA data. However, this was not possible because the telegraph model only estimates the switching rates and the initiation rate scaled by the degradation rate, and the latter is unknown. On the other hand, the estimates from nascent data were rates multiplied by the average elongation time, which is known and hence the absolute rates can be estimated from nascent mRNA data only. The only quantities that could be directly compared were the burst size and the fraction of ON time, since these are both non-dimensional.

Comparing the variability of the parameter estimates, we found that ρ and f_{ON} were the parameters with the smallest variability across samples for the nascent data, as for inference from mature data. However, the inferred parameter variability across samples was on average about 2.5-fold lower for nascent data compared to mature mRNA data (this was obtained by computing the standard deviation divided by the mean for each parameter and then averaging over all parameters and over merged, G1 and G2 data). Likely this is because nascent data does not suffer from post-transcriptional noise. Indeed, synthetic experiments suggested that the errors in parameter inference using nascent data are often less than those in mature data when $f_{\text{ON}} \approx 0.80$ (Fig. 2d). In summary, we have more confidence in the parameter estimates from nascent data, in particular those from cell-cycle separated data.

To further investigate the hypothesis that estimates from cell-cycle specific data are more accurate

than merged data, we compared the estimates from merged and cell-cycle specific data 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 compared their normalized autocorrelation functions (ACFs). Specifically we fed the parameter estimates to the SSA to generate synthetic live-cell data and then calculated the corresponding ACF (SI Section 5). We found that the estimates from cell-cycle specific data produced ACFs that match the live-cell data closer than that from the merged data (Fig. 4d). This was also clear from the sum of squared residuals which for each dataset was smaller for the ACF computed using the cell-cycle specific estimates rather than those from merged data (Fig. 4e).

Using nascent data, we also reinvestigated the hypothesis that the gene exhibits dosage compensation. Comparing the mean rates between the G1 and G2 phases, we found that σ_{off}, σ_{on}, ρ, f_{ON} decreased while the burst size increased upon replication. However, taking into account the variability in estimates across the four datasets, the only two parameters which were cleanly separated between the two phases were σ_{on} and f_{ON}. These two decreased by 41% and 5% respectively. These results had some similarity to those deduced from cell-cycle separated mature mRNA data (the decrease of σ_{on}) but they also displayed differences. Namely, from mature mRNA data it was predicted that ρ decreased upon replication while from nascent data we predicted that ρ did not change and it was rather f_{ON} that decreased by a small degree. The 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 population-based ChIP-seq study [34] that showed DNA dosage compensation after replication in budding yeast. We note that our single-cell analysis only revealed partial dosage compensation, where the mean signal intensity of nascent mRNAs in G2 is not the same as in G1, but 1.7-fold higher in G2 than in G1 (Fig. 4c).

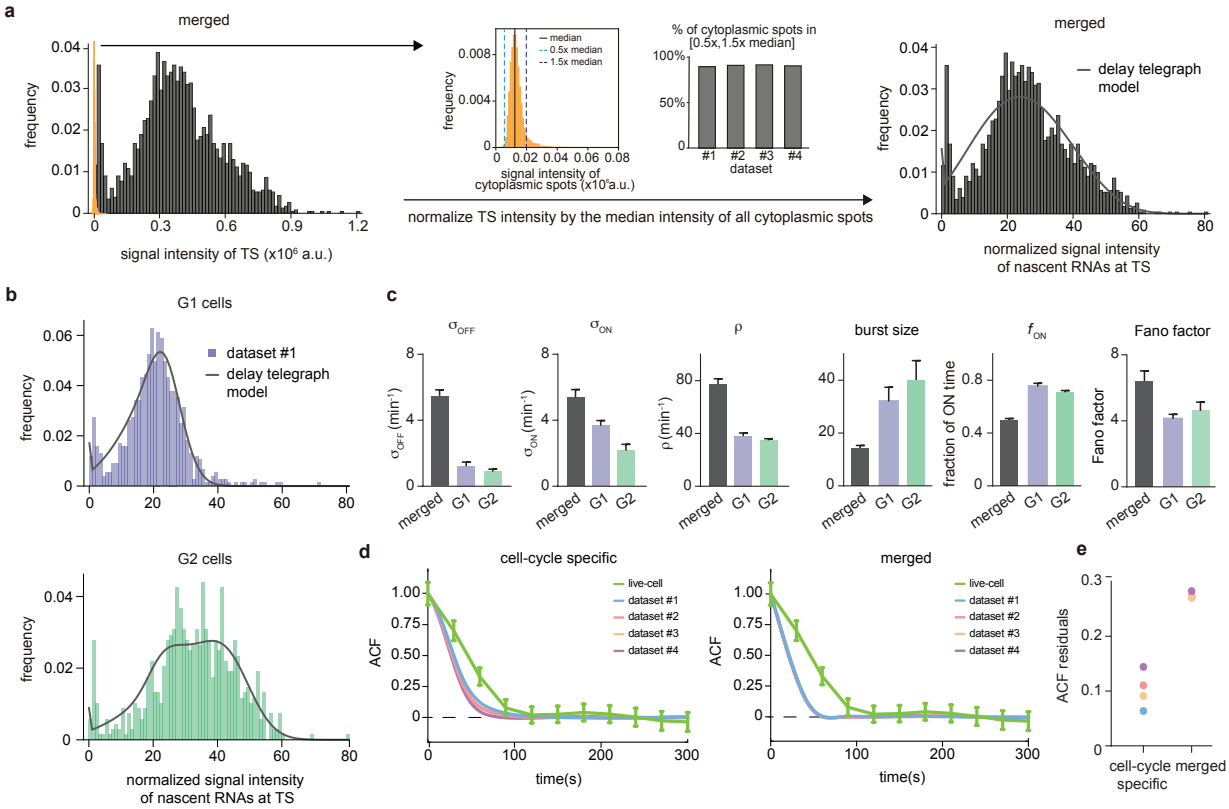


Figure 4: Inference from the normalized nascent mRNA distributions for merged and cell-cycle specific data **a.** Normalized nascent mRNA distributions of merged cell-cycle data were obtained by normalizing the signal intensity of the transcription site (defined as the brightest spot in the cell) by the median signal intensity of the cytoplasmic spots (shown in orange and zoom-in depicted in the inset). In all 4 datasets, approximately 90% of the detected cytoplasmic spots fell in the range 0.5× median – 1.5× median (grey bargraph) Black line in normalized distribution on the right represents best fit with delay telegraph model. **b.** Nascent RNA distributions for cell-cycle specific data. Black line represents best fit with delay telegraph model. **c.** Bar graphs comparing the transcriptional parameters, burst size, fraction of ON time and Fano factor for cell-cycle specific and merged data. Error bars indicate standard deviation. **d.** Normalized ACF plots of cell-cycle specific and merged data. 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 were compared with the ACF measured directly using live-cell data in [6] (green line). **e.** 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 was calculated over all time points).

2.3.2 Correcting for experimental artefacts

Although inference on cell cycle separated 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 and SI Fig 5). In all cases, the experimental distributions showed high intensities in bins 1, 2, and 3, which was likely an artifact of the experimental data acquisition system. Since we defined the transcription site as the brightest spot, this implies that in the absence of a transcription site, a mature transcript can be misclassified as a nascent transcript. We therefore investigated two methods to correct for this, the "rejection" method and the "fusion" method.

The rejection method removed all data associated with the first k bins of the experimentally obtained histogram of fluorescent intensities (Fig. 5a shows the fits for dataset 1; for the other datasets see SI Fig 6). We found that the parameter estimates varied strongly when the number of bins from which data was rejected (k) was changed (Fig. 5b; see also SI Table 12). Although the distributions fit well to the experimental histograms (Fig. 5a), comparison with the live-cell normalized ACF indicated that the estimates actually became worse than non-curated estimates, with a higher sum of squared residuals (Fig. 5c,d). The rejection method therefore does not produce reliable estimates.

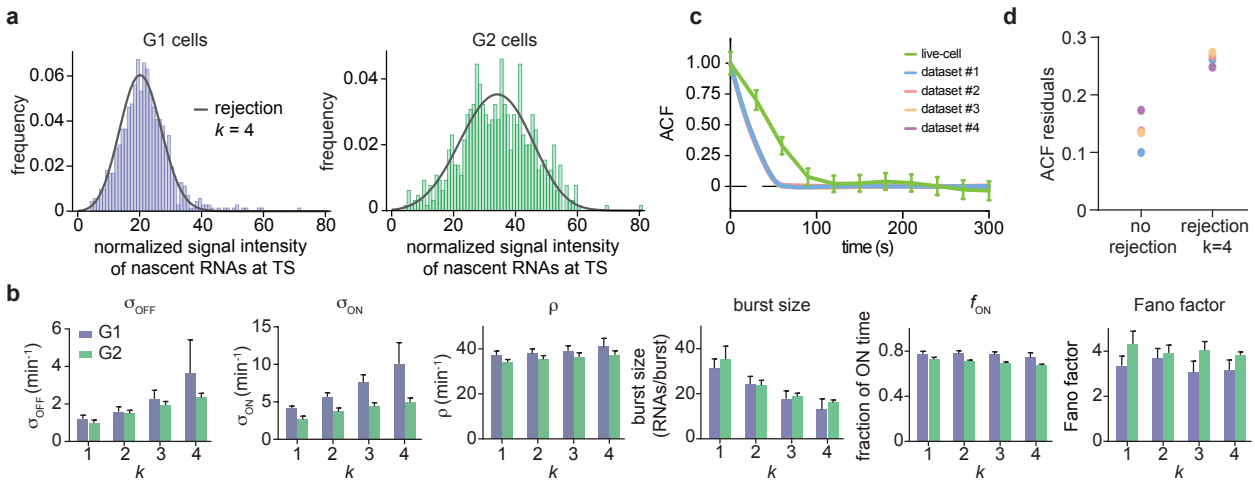
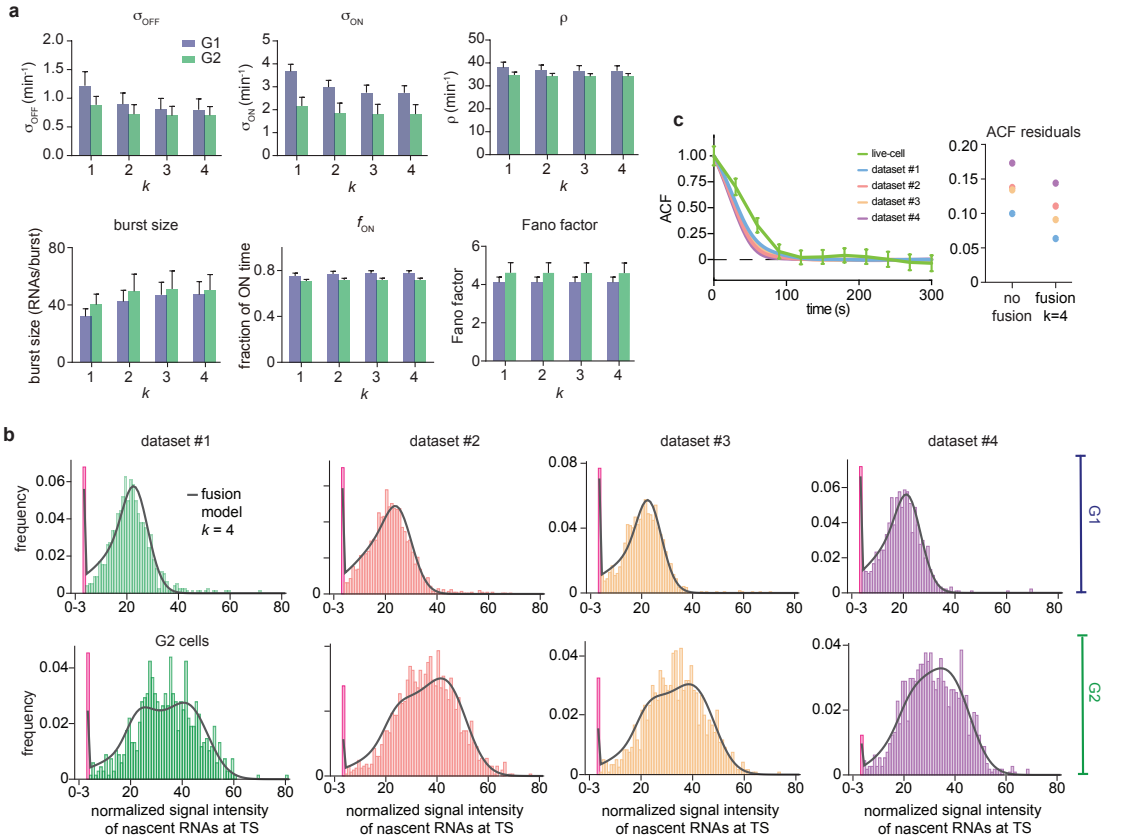


Figure 5: Inference results using the rejection method. **a.** Nascent RNA distributions for cell-cycle specific and merged data. Gray line represents best fit with delay telegraph model using the rejection method. (only the distributions for dataset #1 $k = 4$ are shown). **b.** Estimated transcriptional parameters, burst size, fraction of ON time and Fano factor (mean values and standard deviation error bars) by rejecting the first k bins with $k = 1, 2, 3, 4$. The estimated parameters are listed in SI Table 12. **c.** Normalized 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). **d.** The sum of squared residuals of the ACF of cell-cycle specific data from each dataset without/with rejection when $k = 4$.

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 were below a certain threshold. In other words, we fused or combined the first k bins of the experimentally obtained histogram of fluorescent intensities, thereby taking into account that the true intensity of bin 0 was artificially distributed over some of the first bins.

Fig. 6a and SI Table 13 show that the fusion method led to estimates that varied little with k which enhanced 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 was better captured using the fusion method than using non-curated data (compare Fig. 4b and SI Fig. 5, with Fig. 6b). Comparison to the autocorrelation function of the live-cell data shows that correction with the fusion method also led 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}	Fusion		σ_{OFF}	σ_{ON}	ρ	burst size	f_{ON}
G1	dataset # 1	0.69	2.73	36.30	52.85	0.80	G2	dataset # 1	0.56	1.34	34.45	61.05	0.70
	dataset # 2	1.05	3.09	39.69	37.71	0.75		dataset # 2	0.66	1.86	35.73	54.11	0.74
	dataset # 3	0.63	2.35	35.74	56.78	0.79		dataset # 3	0.67	1.72	33.70	50.56	0.72
	dataset # 4	0.83	2.79	34.68	41.57	0.77		dataset # 4	0.92	2.33	32.79	35.48	0.72
	Mean	0.80	2.74	36.60	47.23	0.78		Mean	0.70	1.81	34.17	50.30	0.72
	Std	0.19	0.31	2.17	9.04	0.02		Std	0.15	0.41	1.24	10.80	0.01

Figure 6: Inference results using the fusion method. **a.** Estimated burst size, fraction of ON and Fano factor (mean values and standard deviation error bars) by combining the first k bins with $k = 1, 2, 3, 4$. **b.** Corresponding fitted distributions for G1 (top row) and G2 (bottom row) using delay telegraph model with the fusion method (only the distributions for $k = 4$ are shown). Magenta 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). The sum of squared residuals of the ACF plots using cell-cycle specific data without/with fusion method when $k = 4$. **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 13 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 the fusion method. The optimally inferred parameters for the four data sets in our study are those given in Fig 6d. The profile likelihood estimates of the 95% confidence intervals of these parameters are shown in SI Table 14. Note that in line with our synthetic data study in Fig. 2, the parameters suffering from the least sample variability were f_{ON} and ρ. The rest of the parameters ($\sigma_{\text{off}}, \sigma_{\text{on}}$ and burst size) suffered more sample variability because the fraction of ON time was high; however since their standard deviation divided by the mean (computed over the four datasets) was not high (in the range of $10 - 20\%$), they still can be regarded as useful estimates. Note also that the previous prediction that gene dosage compensation involves regulation of the burst frequency did not change upon correction of the nascent data using the fusion method. All these results were deduced assuming that the two copies in G2 are independent from each other. Inferring rates under the opposite assumption of perfectly synchronized copies (SI Table 15) gave very similar estimates for ρ and f_{ON} (to be expected since according to the synthetic data study, these two are the most robustly estimated parameters for genes spending most of their time

in the active state) but different estimates for the rest of the parameters. While such perfect synchronization of alleles is unlikely, some degree of synchronization is plausible and further improvement of the transcriptional parameters in the G2 phase will require its precise experimental quantification.

3 Discussion

In this study, we compared the reliability of transcriptional parameter inference from mature and nascent mRNA distributions, with and without taking into account the cell cycle phase. Although these distributions come from the same experiment, we found that the different fits produced very different parameter estimates, ranging from small bursts to very large bursts. Comparison to live-cell data revealed 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. First of all, as we have shown, these fits are sensitive to the segmentation method which can lead to large errors in the estimates. Secondly, the most common method of parameter inference in the 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 (using merged data) can be significantly different from the optimal values (using nascent cell-cycle separated data corrected using the fusion method), with underestimation of the burst sizes of almost 10-fold and underestimation of the active fraction of more than 1.5-fold. These results indicate that parameter inference from merged mature mRNA data should be treated with caution. There were smaller differences between the burst size and the active fraction inferred from cell-cycle separated mature and nascent data (only these two can be directly compared because these are non-dimensional); however the relative errors in the estimates (computed over the four datasets) were more than 2-fold higher for mature data likely due to post-transcriptional noise which nascent data is free from.

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 [35], the only method available so far for nascent measurements is smFISH [36]. 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 [37]. 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 phase *in silico* [26, 38]. 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 not analyzed in smFISH. There could also be differences in the exact percentage of G1 and G2 cells, or other 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 accuracy of some parameters ($\sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}$ and the burst size) deteriorated as f_{ON} approached 1 (2c). 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. The promoter may switch between more than 2 states, or there may be sources of extrinsic noise other than the cell cycle that contribute to the heterogeneity. Previous studies have for example identified extrinsic noise on the elongation rate [30]. However, these more complex transcription models also have more parameters, which in practice often means that very few will be identifiable with the current set of experimental observables. To fit these models, one requires temporal data on the transcription kinetics [30,39], or simultaneous measurements of various sources of extrinsic noise, such as single cell transcription factor concentration and RNA polymerase number measurements, cellular volume, local cell crowding, etc, which are often not available in standard smFISH experiments [40,41]. 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 [42] 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 partial 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 [43]; 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 reasonable estimates that agree with live-cell measurements, if one infers parameters from nascent mRNA distributions that are accounted for cell cycle phase.

4 Methods

4.1 Inference from mature mRNA data

4.1.1 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.

4.1.2 Generation of synthetic mature mRNA data

We generate parameter sets on an equidistant mesh grid laid over the space:

$$(\sigma_{\text{off}}, \sigma_{\text{on}}, \rho) \in [\text{Uniform}(0, 150), \text{Uniform}(0, 150), \text{Uniform}(0, 250)], \tag{4.1}$$

where the units are inverse minute. Furthermore we apply a constraint on the effective transcription rate

$$\hat{\rho} = \frac{\rho\sigma_{\text{on}}}{\sigma_{\text{on}} + \sigma_{\text{off}}} < 100.$$

In each of the three dimensions of the parameter space, we take 10 points that are equidistant, leading to a total of 1000 parameter sets which reduce to 789 after the effective transcription rate constraint is enforced.

We additionally fix the degradation rate $d = 1$ min^{-1}. Note that we choose not to vary the degradation rate (as we did for the other three parameters) since it is not possible to infer all four rates simultaneously – this is because 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].

Once a set of parameters is chosen, we use the stochastic simulation algorithm (SSA [44]) to simulate the telegraph model reactions in Fig. 1a and generate 10^4 samples of synthetic data. Note that each sample mimicks a single cell measurement of mature mRNA.

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

The inference 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 number of mature mRNA measured for 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 telegraph model that describes the measured mature mRNA fluctuations.

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

$$(\sigma_{\text{off}}, \sigma_{\text{on}}, \rho) \in [\text{Uniform}(0, 250), \text{Uniform}(0, 250), \text{Uniform}(0, 300)]\ (\text{min}^{-1}) =: \Theta. \tag{4.2}$$

The degradation rate is fixed to $d = 1$ min^{-1}.

Step (ii) can be obtained either by computing the distribution from the analytical solution [9] or by using the finite state projection (FSP) method [45]. 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

$$\mathcal{L}(\theta) = \prod_{j=1}^{N_{\text{cell}}} P(n_j; \theta), \tag{4.3}$$

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* (https://github.com/robertfeldt/BlackBoxOptim.jl) within the *Julia* programming language to find the optimal parameters

$$\theta^* = \arg\min_{\theta \in \Theta} \left(-\sum_{j=1}^{N_{\text{cell}}} \log P(n_j; \theta) \right). \tag{4.4}$$

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 (MRE) which is defined as

$$\begin{aligned} \text{MRE} &= \frac{1}{M}\sum_{i=1}^{M} \text{Relative error}(\theta_i^*, \theta_{\text{true},i}), \\ \text{Relative error}(\theta_i^*, \theta_{\text{true},i}) &= \frac{|\theta_i^* - \theta_{\text{true},i}|}{|\theta_{\text{true},i}|} \end{aligned} \tag{4.5}$$

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.

4.2 Inference from nascent mRNA data

4.2.1 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 6, we present an alternative approach to derive the steady-state solution. The reaction steps are illustrated in Fig. 1a.

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. 1b for an illustration). This presents a problem because although we can predict the distribution of the number of bound Pol II using the delay telegraph model, 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 is 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. 1b. 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.

$$p(s|k) = (g * g \cdots * g)(s) = g^{*k}(s), \quad g^{*0}(s) = \delta_0(s), \tag{4.6}$$

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

$$p(s;\theta) = \sum_{k=0}^{\infty} p(s|k)P(k;\theta), \tag{4.7}$$

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. (4.7) represents the extension of the delay telegraph model to predict the smFISH fluorescent signal of the transcription site.

Comparison to the algorithm in [24]. Both algorithms take into account the fact that the signal intensity depends on the position of Pol II on the gene, albeit this is done in different ways. In [24] a master equation is written for the joint distribution of gene state and the number of nascent mRNA. In this case the number of nascent RNAs can have non-integer values since it represents the experimentally measured signal from the (incomplete) nascent RNA. Solution of this master equation proceeds by (a) a discretization of the continuous nascent mRNA signal into bins which are much smaller than one; (b) solution using finite state projection (FSP). This approach can lead to a large state space which incurs a large computational cost. In contrast, in our method, we use FSP to solve for the delay telegraph model, i.e. the distribution of the discrete number of bound Pol II from which we construct (using convolution) the approximate distribution of the continuous nascent mRNA signal by assuming the Pol II is uniformly distributed on the gene. Since the state space of bound Pol II is typically not large, our method will typically be more computationally efficient than the one described in [24].

4.2.2 Generation of synthetic nascent mRNA data

We generated synthetic smFISH signal data by using the SSA, modified to include delay to simulate the delay telegraph model [46]. Specifically, we use Algorithm 2 described in [47]. One run 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. 1b 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 chosen from the same region of parameter space as in (4.1), on the same equidistant mesh grid and with the same constraint on the effective transcription rate. Unlike the mature mRNA case, here there is no degradation rate; instead we have the elongation time, which we fix to $\tau = 0.5$ (min). Note that fixing this time is necessary since it is not possible to infer

the 3 transcriptional parameters rates and the elongation time simultaneously because the steady-state solution of the delay telegraph model is a function of the non-dimensional parameter ratios $\rho\tau, \sigma_{\text{off}}\tau$ and $\sigma_{\text{on}}\tau$.

Once a set of parameters is chosen, we use the modified SSA (as described above) to simulate the signal intensity in each of 10^4 cells.

4.2.3 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. (4.7) on an interval $[i-1, i]$ for $i \in \mathbb{N}$ which, in our numerical scheme, means

$$S(i;\theta) := \sum_{k=0}^{K} P(k;\theta) \int_{i-1}^{i} g^{*k}(x)\mathrm{d}x, \quad i = 1, 2, \ldots \tag{4.8}$$

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) [45]. In SI Fig. 8 and SI Table 16, 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

$$\mathcal{L}(\theta) = \prod_{j=1}^{N_{\text{cell}}} S(q_j;\theta), \tag{4.9}$$

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 (for both mature and nascent mRNA inference) is summarized by a flow-chart in Fig. 1c.

4.3 Experimental data acquisition and processing

A diploid yeast strain of BY4743 background with a single integration of 14xPP7 loops at the 5'UTR of *GAL10* was used in this study. Yeast cultures were grown in synthetic complete media with 2 % galactose to early mid-log (OD 0.5), fixed with 5% paraformaldehyde (PFA) for 20 min, permeabilized with 300 units of lyticase and hybridized with 7.5 pmol each of four PP7 probes labeled with Cy3 (Integrated DNA Technologies) as described in Trcek et al. [48] and Lenstra et al. [36, 49]. The PP7 probe sequences are: atatcgtctgctcctttcta, atatgctctgctggtttcta, gcaattaggtaccttaggat, aatgaaccgggaatactgc. 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 with a Plan-Apochromat 40x 1.4NA oil DIC UV objective and a 1.25x optovar. For Cy3, a 562 nm longpass dichroic (Chroma T562lpxr), 595/50 nm emission filter (Chroma ET595/50m) and 550/15 nm LED excitation at full power (Spectra X, Lumencor) were used. For DAPI, a 425 nm longpass dichroic (Chroma T425lpxr) and a 460/50 nm emission filter (Chroma ET460/50m) and LED excitation at 395/25 nm at 25% power (Spectra X, Lumencor) were used. The signal was detected on a Hamamatsu ORCA-Flash4.0 V3 Digital CMOS camera (Hamamatsu Photonics, Japan). For each sample

and each channel, we utilized the Micro-Manager software (UCSF) to acquire at least 20 fields-of-view based on the DAPI channel. Each field-of-view consisted of 13 z-stacks (with a z-step of 0.5 um) at 25 ms exposure for DAPI and 250 ms exposure for Cy3.

A custom python pipeline was used for analysis (https://github.com/Lenstralab/smFISH). Maximum intensity projected images were used to segment the cell and nucleus using Otsu thresholding and watershedding (segmentation 1). In addition, we segmented cells using CellProfiler (segmentation 2). 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 [50]; Thompson et al. [51]; Larson et al. [52, 53] and Coulon et al. [54]). Cells in which no spots were detected were excluded from further analysis since a visual inspection indicated that these cells were not properly segmented or were improperly permeabilized.

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 fluorescence intensity of all the cytoplasmic spots in all cells. This is due to the fact that 90% of cytoplasmic mRNAs are isolated (Fig. 4a), 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. (4.7).

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. We show in Fig. 3c that this has negligible influence on the estimated parameters 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 Fig. 3e left). This gave similar results to a different cell cycle classication method using the Fried/Baisch model [55] which was recently employed in [26]. See SI Fig. 9 for a comparison of the two methods. We note that cells in late G2 may contain two separate transcription sites, one in the mother and one in the bud. When the nucleus moves into the bud, buds often contain less DNA than G1 cells, and mothers contain more DNA than G1 cells, both of which are excluded from the analysis. When the DNA content of the mother and daughter is similar, both mother and daughter are counted separately as G1 cells. We note that this late G2 subpopulation is very small.

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.

4.4 Data availability

The 4 smFISH datasets are available from https://osf.io/d5nvj/. These datasets include the maximum intensity projected images, the spot localization results, the nuclear and cellular masks used for merged, G1 and G2 cells and the analyzed results of the mature and nascent data. The analysis code of the smFISH microscopy data is available at https://github.com/Lenstralab/smFISH. The code for the the synthetic simulations and the parameter inference is available at https://github.com/palmtree2013/RNAInferenceTool.jl.

Acknowledgments

Z.C., X.F. and L.X. acknowledge the support from Natural Science Foundation of China (NSFC No. 61988101, 62073137). X.F. acknowledge the support from Shanghai Sailing Program (22YF1410700). 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). R.G was supported by a Leverhulme Trust research award (RPG-2020-327).

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

[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] Fritzsch, C. *et al.* Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. *Molecular systems biology* **14**, e7678 (2018).

[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] Hansen, M. M., Desai, R. V., Simpson, M. L. & Weinberger, L. S. Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. *Cell systems* **7**, 384–397 (2018).

[33] Durrieu, L. *et al.* Characterization of cell-to-cell variation in nuclear transport rates and identification of its sources. *bioRxiv* (2022).

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

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

[36] 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).

[37] 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).

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

[39] Rodriguez, J. *et al.* Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell* **176**, 213–226 (2019).

[40] Battich, N., Stoeger, T. & Pelkmans, L. Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).

[41] Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by underlying cell state. *Molecular systems biology* **16**, e9146 (2020).

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

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

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

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

[46] Fu, X., Zhou, X., Gu, D., Cao, Z. & Grima, R. Delayssatoolkit. jl: stochastic simulation of reaction systems with time delays in julia. *bioRxiv* (2022).

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

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

[49] 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).

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

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

[52] 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).

[53] 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).

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

[55] 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).