Transcriptome-wide analysis of cell cycle-dependent bursty gene expression from single-cell RNA-seq data using mechanistic model-based inference

Augustinas Sukys[1,2] and Ramon Grima*[1]

[1]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, UK
[2]The Alan Turing Institute, London NW1 2DB, UK

Abstract

Bursty gene expression is quantified by two intuitive parameters: the burst frequency and the burst size. While these parameters are known to be cell-cycle dependent for some genes, a transcriptome-wide picture remains missing. Here we address this question by fitting a suite of mechanistic models of gene expression to mRNA count data for thousands of mouse genes, obtained by sequencing of single cells for which the cell-cycle position has been inferred using a deep-learning approach. This leads to the estimation of the burst frequency and size per allele in the G1 and G2/M cell-cycle phases, hence providing insight into the global patterns of transcriptional regulation. In particular, we identify an interesting balancing mechanism: on average, upon DNA replication, the burst frequency decreases by $\approx 50\%$, while the burst size increases by the same amount. We also show that for accurate estimation of the ratio of burst parameters in the G1 and G2/M phases, mechanistic models must explicitly account for gene copy number differences between cells but, surprisingly, additional corrections for extrinsic noise due to the coupling of transcription to cell age within the cell cycle or technical noise due to imperfect capture of RNA molecules in sequencing experiments are unnecessary.

Introduction

Gene expression occurs in random bursts of transcription, associated with alternating active and inactive promoter states [1, 2]. By tagging mRNA molecules using fluorescent reporters, live-cell imaging reveals a train of pulses of fluorescent intensity whose mean height is proportional to the mean number of transcripts produced when the gene is actively transcribing (the mean burst size) and whose frequency reflects how often transcription occurs (the burst frequency) [3]. These two burst parameters offer a simple, intuitive and practical quantitative description of gene expression and hence their estimation has been the focus of many studies (for reviews see [4–6]).

While live-cell measurements are ideal to understand transcription, they are challenging because they are low-throughput and require genome editing [7, 8]. To overcome this challenge, one can leverage the inherent heterogeneity generated by bursting within a population of cells. Acquiring a distribution of mRNA counts per cell is nowadays relatively straightforward through techniques such as single-molecule fluorescence in situ hybrization (smFISH) and single-cell sequencing. These measured distributions can then be used to infer the burst size and frequency by comparing them to the distributions predicted by mechanistic models of gene expression (where the rate parameters determine the speed of molecular processes). Common examples of such models are the two-state telegraph model [1, 9] which predicts both unimodal and bimodal mRNA count distributions, and the simpler bursty model that predicts unimodal negative binomial distributions [10].

The vast majority of studies reporting estimates of the burst size and frequency have used smFISH data [11–14] because it is widely considered to be the gold standard for the accurate measurement of the number of RNA molecules for a specific gene in single cells. However many of these studies suffer from a major limitation: they implicitly assume that the burst parameters are the same for each cell in the sample. Equivalently, it is assumed that the observed inter-cell differences in the transcript numbers are generated by the uncertainty in the timing of biochemical events (intrinsic noise) [15, 16]. However generally this is not the case—burst parameters vary from one cell to another because of differences in the number of cellular components such as RNA polymerases, transcription factors and other molecules playing key roles in transcription. These parameter fluctuations are commonly referred to as extrinsic noise since they arise independently of a gene of interest but act on it. Thus the

*Correspondence: asukys@turing.ac.uk, ramon.grima@ed.ac.uk

differences in transcript numbers between cells are due to both intrinsic and extrinsic noise, although the latter has been suggested to be the dominant source of heterogeneity [17, 18].

One might suppose that the burst parameters estimated using models that assume perfectly identical cells can be approximately interpreted as the parameter values averaged over all cells. However a recent study conclusively showed this is not the case. In Ref. [19] it was shown that there exist systematic biases in parameters inferred using the telegraph model when its distribution or moments are fitted to those from simulated single-allele mRNA data from a population of cells with extrinsic noise in one of the parameters. Whether the burst parameters are over- or underestimated was found to depend on the source of extrinsic noise (which parameter is the most variable amongst cells) and the mode of transcriptional activity. In reality, these systematic biases are likely to be larger when non-allele-specific mRNA count data is used for inference because the cell-to-cell differences in the gene copy numbers can be a strong determinant of gene expression noise [20].

Together, these studies show that it is impossible to obtain a reliable quantitative picture of gene expression by fitting models to single-cell data unless the inference procedure appropriately corrects for the changes of gene copy number with the cell-cycle phase and for extrinsic noise due to other exogenous factors that vary within each phase. A number of smFISH-based studies have suggested that the burst frequency is principally affected by the cell-cycle phase [20–22] while the burst size scales with the cell volume [21, 23], which leads to concentration homeostasis [24]. Note that cell volume varies considerably within a cell-cycle phase and hence it is possibly a dominant source of extrinsic noise within each phase; other extrinsic noise sources include mitochondrial variability [25] and Ca2+ signalling [26]. However, even if sophisticated modelling frameworks that include all the principal sources of extrinsic noise can be developed and used to fit the mRNA count distributions measured using smFISH, this would not lead to a transcriptome-wide quantitative understanding of gene expression as such data is typically limited to a relatively small number of genes of interest.

A possible way of circumventing this problem involves the use of data generated by single-cell RNA sequencing (scRNA-seq). This allows the comprehensive profiling of large numbers of cells on a genome-wide level thus providing an unprecedented insight into cell-to-cell variability [27–29]. While it has primarily been used to distinguish between different cell types in a sample [30–33], to provide insights into the dynamics of cell state switching [34–36] and to reconstruct gene-regulatory networks [37–41], in recent years it has also been used to study the inherent stochasticity of gene expression [19, 42–48]. Despite its potential for transcriptome-wide inference, scRNA-seq data has one major disadvantage compared to smFISH data, namely the high amount of technical (non-biological) noise [49, 50]. In droplet-based sequencing this can arise from the imperfect capture of all RNA molecules (especially from lowly expressed genes), amplification errors and non-uniquely identifiable reads. While technical noise is reduced by the use of data that incorporates unique molecular identifiers (UMIs) [51, 52]—random nucleotide sequences that tag an individual mRNA molecule and thus enable separation between the original molecules and amplification duplicates derived from the cDNA library amplification [53–55]—this is not a panacea. A comparison of the probability distributions of mRNA counts for five genes measured using both smFISH and a UMI-based scRNA-seq platform implies there is a non-negligible fraction of transcripts in each cell that are not captured by the sequencing technology (Extended Fig. 7 of [43]). Considering the limitations of inference already identified using smFISH data, it is clear that the deduction of quantitative models of bursty gene expression using scRNA-seq data would need an inference procedure that corrects for the cell-cycle phase, extrinsic noise within each phase and technical noise. This trifarious correction is missing from the literature. Current methods of inferring burst parameters from scRNA-seq data either ignore/cannot properly account for extrinsic and technical noise [42, 43] or correct only for extrinsic noise [19] or correct for certain types of extrinsic noise and technical noise but not for the variation in gene copy number during the cell-cycle [44, 45, 48, 56].

In this paper, we develop an inference framework that takes into account the cell-cycle phase, extrinsic noise due to factors varying within each phase and technical noise, and use it to investigate how the burst frequency and size are modulated by the position of the cell cycle in eukaryotic genes. While previously other studies sought to answer this question using smFISH-based approaches for a few tens of genes, we use previously published scRNA-seq data [57] to understand the cell-cycle dependence of the bursty expression of about 1500 mouse genes. The overall approach is illustrated in Fig. 1. Sequencing data provides spliced and unspliced read counts for over 10^4 genes in undifferentiated mouse embryonic stem cells. DeepCycle [57] uses a deep-learning approach to map the cycling patterns in the unspliced-spliced RNA space of cell cycle-related genes to a parameter θ for each cell, which gives the cell age: $\theta = 0$ means a newborn cell, while $\theta = 1$ indicates a cell that is about to divide via mitosis (hence θ is a rough proxy for cell size). From this information, it is also possible to determine the cell-cycle phase which agrees with the phases according to the DNA content given by Fluorescence-Activated Cell Sorting (FACS). Given this joint information of mRNA counts, cell age θ and cell-cycle phase for each cell, we fit and select the best of six different mechanistic models of stochastic transcription (accounting for extrinsic, intrinsic and technical noise) that

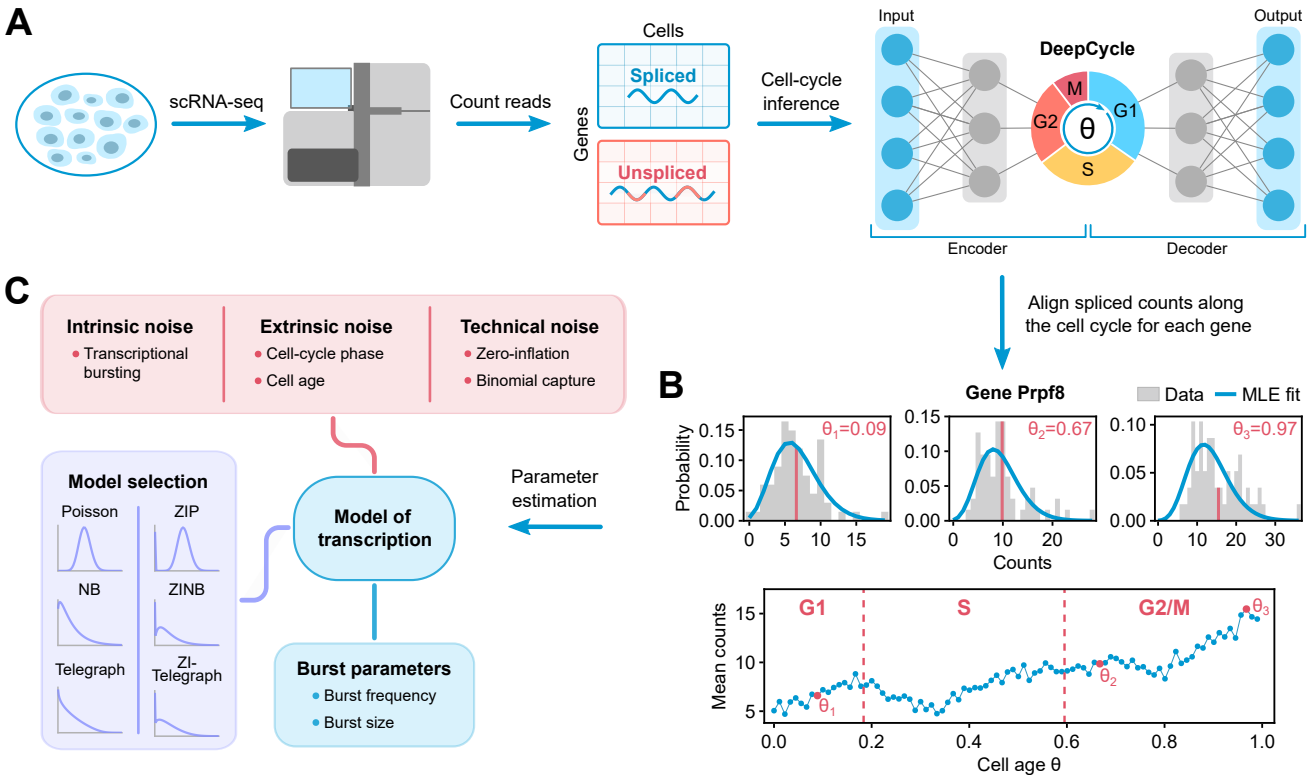

Figure 1: Schematic of the single-cell data acquisition, processing and parameter inference pipeline. **(A)** Our mechanistic analysis is built upon the study of Riba *et al.* [57], which generated a scRNA-seq dataset of over five thousand mouse embryonic stem cells and developed a deep-learning approach called DeepCycle to estimate the cell age and cell-cycle phase for each cell in the population. Specifically, DeepCycle fits the gene expression patterns in the spliced-unspliced space to estimate the cell age $\theta \in [0, 1)$ associated with the relative position of a cell in its progression along the cell cycle from G1 to mitosis. In addition, certain values of θ can be linked to transitions between the main cell-cycle phases, and hence all cells in the dataset can be grouped by their cell-cycle phase (G1, S or G2/M). **(B)** An example showing a subset of the processed data for the Prpf8 gene. The bottom graph shows how the mean spliced mRNA counts (average over cells falling in each bin of θ values) evolve as the cells go through G1, S, and G2/M phases of the cell cycle. The histograms above show the spliced count distributions for three different values of the cell age θ; the vertical red lines denote the mean values (also indicated by the red circles on the bottom plot). **(C)** In this paper, we use the processed data and the associated cell-cycle information to perform maximum likelihood-based model inference on the experimentally observed (spliced) mRNA count distributions, taking into account intrinsic, extrinsic and technical noise sources. We perform model selection from six different stochastic models of transcription: Poisson, zero-inflated Poisson (ZIP), negative binomial (NB), zero-inflated negative binomial (ZINB), telegraph and zero-inflated telegraph (ZI-Telegraph) models. As an example, maximum likelihood (MLE) fits for the optimal model describing the gene Prpf8 are shown by blue lines in the upper graphs in (B). The burst parameters (burst size and frequency) are then computed from the optimal model parameters. For technical details see Methods.

can explain the data; this also enables us to estimate the rate parameters of the reactions described by the model, from which the burst parameters can be calculated. We show that the most robust inference strategy involves the estimation of the ratios of the burst parameters between G1 and G2/M rather than their values in each phase, which are highly sensitive to the set of models used for inference. From the burst parameter ratios, we identify the different ways in which the burst frequency and size vary with cell-cycle phase, hence providing an unprecedented quantitative view of the dynamics of gene expression in mammalian cells.

Results

For most genes, both burst frequency and burst size vary with cell-cycle phase

To understand whether the burst parameters vary across the cell cycle, we calculate the mean mRNA counts of each gene using data from all cells that are in G1 ($\langle m_{\mathrm{G1}} \rangle$) and similarly using data from all cells in G2/M ($\langle m_{\mathrm{G2/M}} \rangle$). In Fig. 2A, we show the distribution of the ratio of the means ($\Lambda = \langle m_{\mathrm{G2/M}} \rangle / \langle m_{\mathrm{G1}} \rangle$). Note that the mass of the distribution is concentrated between $\Lambda = 1$ and $\Lambda = 2$ with a median $\Lambda = 1.40$. The case $\Lambda = 2$ is expected if there is no change in the rate parameters from G1 to G2/M, since the doubling of the gene copies would simply imply a corresponding doubling of the mean mRNA count. In contrast, the case $\Lambda = 1$ is what can be described as

perfect gene dosage compensation, in the sense that the rate parameters have changed in such a way that the mean mRNA count remains unaltered as the cell cycle progresses. For the vast majority of genes, Λ is neither 1 nor 2, implying there that there is a change in the rate parameters between the two cell-cycle phases but it is insufficient to perfectly compensate for the doubling of the gene copy number during DNA replication. Hence we conclude that partial dosage compensation is the norm across the transcriptome.

Given the foregoing conclusion that for most genes there is a change in the rate parameters as the cell cycle progresses, we next infer the rate parameter values in each phase, from which we can also calculate the burst parameters. We assume that in each of the two phases, the distribution of spliced mRNA counts is well described by the steady-state distribution of a stochastic model of gene expression describing transcription from N independent promoters, where $N = 2$ in the G1 phase and $N = 4$ in the G2/M phase. The mRNA count distribution of this model is obtained by convolving the steady-state mRNA distribution of a stochastic model of gene expression for one promoter with itself N times; this is equivalent to assuming that expression from individual gene copies is independent of each other, a property of many eukaryotic genes [22, 34, 43]. Using the aforementioned model count distribution and the transcript count data for each cell, we use the method of maximum likelihood to estimate the rate parameters (normalised by the mRNA degradation rate) in each cell-cycle phase. We fit six different gene expression models (telegraph model, negative binomial model, Poisson model and their equivalents with zero-inflation) and then select the optimal model using the Bayesian Information Criterion—we note that these models are all age-independent, in the sense that the rate parameters in each cell-cycle phase are constant and do not depend on the cell age θ. The telegraph model and its use to fit the experimental distributions of mRNA counts in the two cell-cycle phases is illustrated in Fig. 2B (note that the model selection step is not shown). In this model, the normalised burst frequency per allele in G1 is the ratio of the activation rate parameter and the mRNA degradation rate, $\sigma_1^{\mathrm{G1}}/d_{\mathrm{G1}}$, and the burst size per allele is the ratio of the transcription and inactivation rate parameters, $\rho_{\mathrm{G1}}/\sigma_0^{\mathrm{G1}}$; in G2/M the expressions are the same, except the subscripts and superscripts of the rate parameters are changed to G2/M. Note that we cannot directly infer the true burst frequencies (σ_1^{G1} and $\sigma_1^{\mathrm{G2/M}}$) without separate measurements of the degradation rate for each transcript in each cell-cycle phase. Hence, in all that follows, what we mean by burst frequency is the normalised burst frequency. For more details see Methods.

For each gene that is not best fit by a Poisson or a zero-inflated Poisson model, i.e. bursty genes, of which there are 1669, we calculate $Q_f = f_{\mathrm{G2/M}}/f_{\mathrm{G1}}$ and $Q_b = b_{\mathrm{G2/M}}/b_{\mathrm{G1}}$. These are the ratios of the burst frequency and burst size (for each gene copy) in the two phases, respectively. In Fig. 2C we show these ratios for 5 genes. Note that there appears to be a tendency for the burst frequency in the G1 phase to be greater than that in the G2/M phase and the opposite tendency is observed for the burst size. In Fig. 2D we show the distributions of Q_f and Q_b for all genes, confirming that for most genes there is an attenuation of the burst frequency and a corresponding amplification of the burst size as the cell cycle progresses. We note that the median estimates of these distributions are in good agreement with what is expected from theory. Since for most bursty genes the model selection procedure selects the negative binomial model, the mean mRNA count equals the product of the number of gene copies, the burst frequency and the burst size for each gene copy, and hence the fold-change of the mean mRNA count can be simply expressed as:

$$\Lambda = \frac{\langle m_{\mathrm{G2/M}} \rangle}{\langle m_{\mathrm{G1}} \rangle} = 2 Q_f Q_b. \tag{1}$$

From Fig. 2A, the median of Λ is approximately 1.40 while from Fig. 2D the medians of the two ratios are $Q_f \approx 0.49$ and $Q_b \approx 1.42$, which perfectly agrees with the theoretical result given by Eq. 1 (the result also holds if one uses the average values as these are very similar to the median ones). Since for most genes the changes in the burst frequency and size are anticorrelated, the bursty nature of gene expression will be more noticeable in the G2/M phase, characterised by bursts that are larger and more separated in time than in the G1 phase.

We next seek to understand which of the burst parameters changes the most as the cell cycle progresses. By comparing the relative change in the inferred burst frequency and size for each of the 1669 bursty genes in our dataset (taking into account the confidence intervals for the burst parameter estimates) we found that for most genes (79%) the relative change in the burst size and frequency were comparable, for 16% of the genes the burst frequency changes more than the burst size and for 5% of the genes the change in the burst size was larger than that in frequency (Fig. 2E). As we now show, these three gene categories can be related to the range of values of the ratio of the mean mRNA counts in the G2/M and G1 phases (Λ). First we note that the relative changes in the burst frequency Δf and burst size Δb can be written as $\Delta f = Q_f - 1$ and $\Delta b = Q_b - 1$, from which it follows that Eq. (1) can be written as $\Lambda = 2(1 + \Delta f)(1 + \Delta b)$. Let Λ_1 be a typical value of Λ associated with genes characterised by $|\Delta f| \approx |\Delta b|$, Λ_2 be a typical value of Λ associated with genes characterised by $|\Delta f| < |\Delta b|$ and Λ_3 be a typical value of Λ associated with genes characterised by $|\Delta f| > |\Delta b|$. If the change in burst frequency

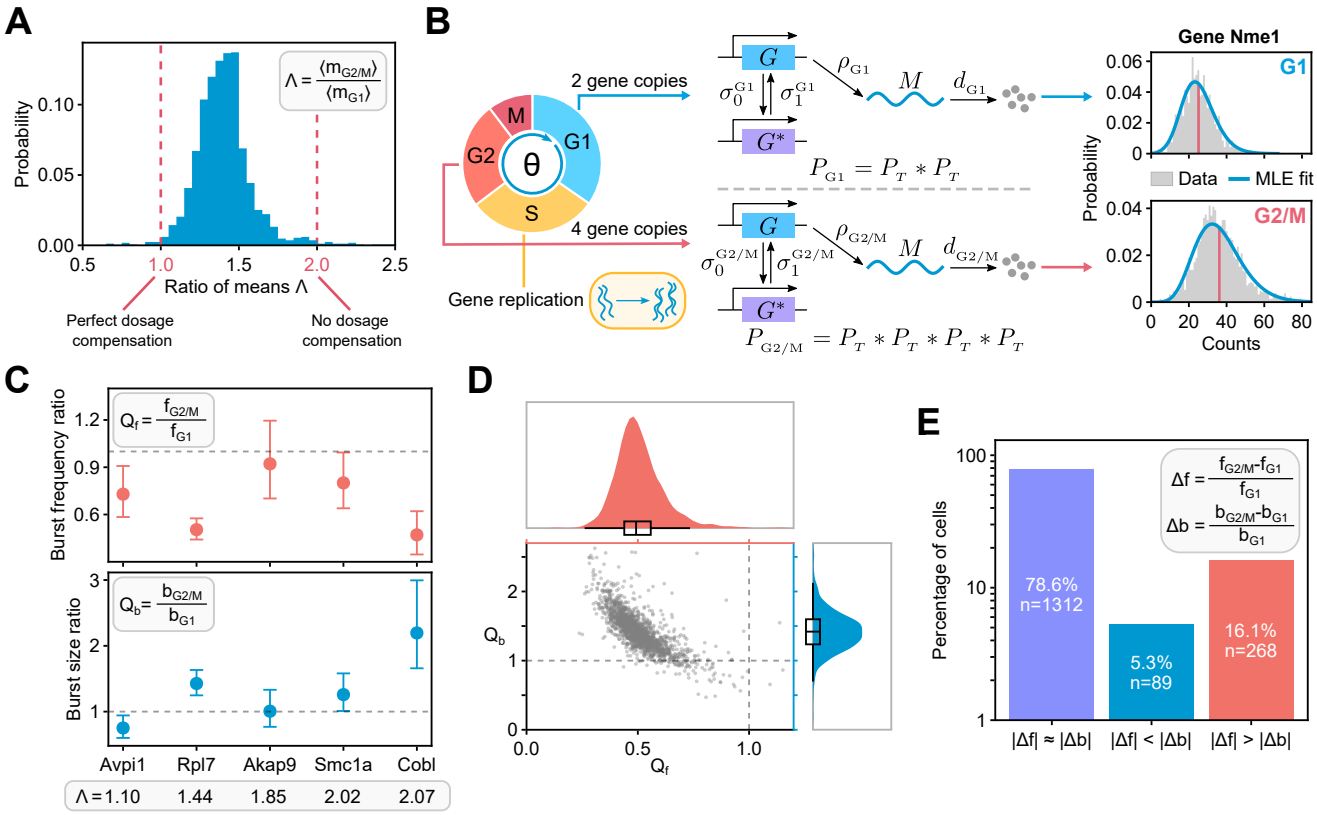

Figure 2: The relationship between cell-cycle phases and the burst parameters. **(A)** Histogram of the ratio of mean mRNA counts in G2/M and G1 phases, $\Lambda = \langle m_{G2/M} \rangle / \langle m_{G1} \rangle$, for all 1669 bursty genes in the dataset. Ratio of 1 implies a mechanism that perfectly compensates for the doubling of the mRNA number due to the doubling of the gene dosage during replication, while a ratio of 2 indicates the opposite. **(B)** Illustration of one of the 6 models used in the model fitting and selection algorithm—the telegraph model. The rate parameters (normalised by the mRNA degradation rate) are inferred per gene copy. The mRNA distribution in G1, P_{G1}, is given by a convolution of two standard telegraph models P_T (with rate parameters having a subscript G1). In the G2/M phase, due to the doubling of gene copy number in the S phase, the mRNA distribution, $P_{\mathrm{G2/M}}$, becomes a convolution of four standard telegraph models (with rate parameters subscripted by G2/M). Plots to the right show the mRNA distributions of gene Nme1 in G1 and G2/M phases and the corresponding best fits using maximum likelihood estimation (MLE fit). **(C)** Ratios of the burst frequency per allele (top) and burst size per allele (bottom) in the G2/M and G1 phases, denoted by Q_f and Q_b respectively, for five different genes in the dataset. Error bars indicate the 95% confidence intervals. The grey dashed lines indicate no change in the burst parameter values upon cell-cycle progression. **(D)** Scatter plot of the burst size ratio between G2/M and G1 phases versus the burst frequency ratio for all 1669 genes (each grey dot indicates a single gene). The smoothed histogram on the right (in blue) represents the distribution of burst size ratios. Similarly, the distribution of burst frequency ratios is given on the top (in red). **(E)** Bar chart comparing the absolute value of relative change in burst frequency between the G2/M and G1 phases, $|\Delta f|$, to the corresponding relative change in burst size, $|\Delta b|$, across all genes. The plot shows the percentages of genes (on a logarithmic scale) for which the relative change in burst frequency is comparable to the relative change in burst size ($|\Delta f| \approx |\Delta b|$), the relative change in burst size is greater than the change in burst frequency ($|\Delta f| < |\Delta b|$) or vice-versa ($|\Delta f| > |\Delta b|$). The categorisation is performed by computing the 95% confidence interval of the ratio $|\Delta f / \Delta b|$ for each gene and checking for overlap with 1 (which indicates no change). The number n indicates the number of genes falling into each category, where the total number of genes is $n = 1669$.

Δf is usually negative and the change in burst size Δb is positive (as suggested by Fig. 2D), then it follows that $\Lambda_2 > \Lambda_1 > \Lambda_3$. The first, second and third quartiles $(\mathrm{Q}_1, \mathrm{Q}_2, \mathrm{Q}_3)$ of the distributions of Λ values for the genes in each category as given in Fig 2E are found to be approximately $(1.31, 1.41, 1.49)$ for Λ_1, $(1.47, 1.59, 1.95)$ for Λ_2 and $(1.20, 1.29, 1.37)$ for Λ_3—on average, the results are in agreement with our theoretical observations. Hence we have shown that the strength of dosage compensation for a gene determines to some extent which of the three categories it falls in.

Mean gene expression level tends tends to scale with cell age

Next, we seek to uncover how the cell age influences the mean expression level within each cell-cycle phase—we undertake this to understand whether we need a further refinement of the mechanistic models of gene expression in Fig. 2 to incorporate a cell age dependence of the rate parameters. Accounting for significant extrinsic noise sources

within a cell-cycle phase is crucial, as otherwise it might bias the burst frequency and burst size estimates [19].

As previously mentioned, DeepCycle assigns a normalised cell age θ to each cell, which varies between 0 and 1. For a cell in generation i, θ can be understood as equal to t/T_i, where T_i is the cell-cycle duration in this generation and t is the true cell age (which varies between 0 and T_i); hence in this method there is no implicit assumption of a fixed cell-cycle duration as in other inference methods [56]. This is important because cell-cycle duration variability is a significant contributor to the magnitude of gene expression noise [58].

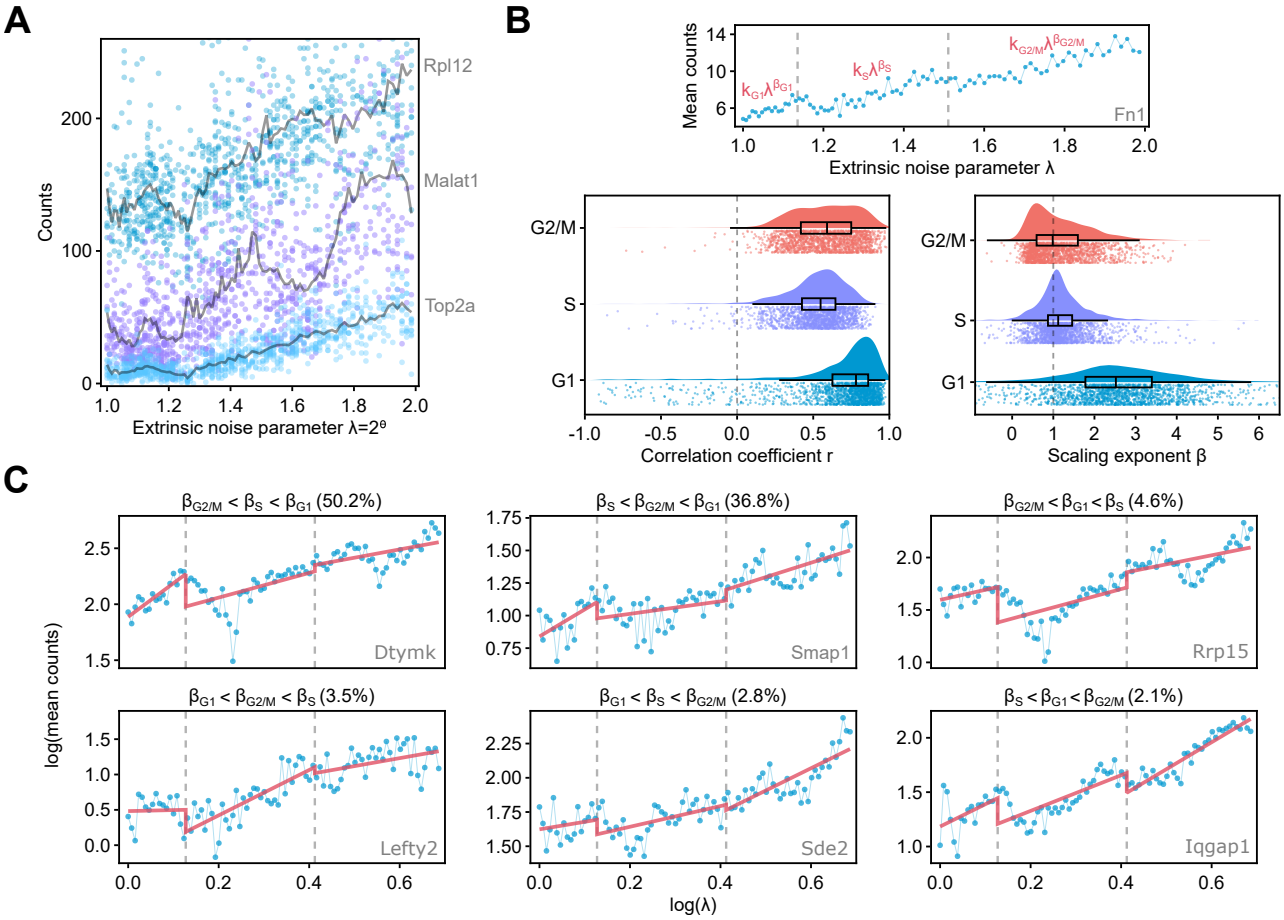

Figure 3: Effect of cell age on mean gene expression levels. **(A)** Spliced mRNA counts versus the effective extrinsic noise parameter λ (given by 2^θ) for Rpl12, Malat1 and Top2a genes. The scatter plots display a sample of 1000 cells (only around 20% of the population) for visual clarity. Each black line corresponds to the mean mRNA counts (an average over cells falling into each bin of 2^θ values). **(B)** Top: plot of the mean mRNA counts versus the extrinsic noise parameter λ for the Fn1 gene. The two dashed grey vertical lines indicate the transitions from G1 to S and from S to G2/M respectively. We denote the nonlinear relationship between the mean counts and λ by a power law with a cell-cycle phase-specific proportionality constant k and exponent β (shown in red). Bottom left: raincloud plots [59] of the Pearson correlation coefficient r computed for all genes in each cell-cycle phase. Each raincloud plot combines a smoothed histogram, a box plot and the jittered raw r values. Bottom right: raincloud plots of β for all genes, determined separately for cells in the G1, S and G2/M phases by linear regression in the log-log space of mean counts versus λ. The grey dashed line at $\beta = 1$ indicates the genes whose expression depends linearly on λ. **(C)** Mean mRNA counts versus λ in log-log space for six sample genes corresponding to the six different possible orderings of β_{G1}, β_{S} and $\beta_{\mathrm{G2/M}}$ values. The red line indicates the best discontinuous piecewise linear fit over the three cell-cycle phases, where the slope corresponds to β_{G1}, β_{S} and $\beta_{\mathrm{G2/M}}$ respectively. The percentage of all genes falling into each category is shown at the top of each plot.

In Fig. 3A we show for 3 genes that the number of spliced mRNA counts tends to increase with the parameter $\lambda = 2^\theta$. While at first sight the relationship appears to be roughly linear, averaging of the spliced counts over cells falling into each bin of the parameter λ (see black lines in Fig. 3A) suggests that the relationship can be nonlinear over some ranges of λ. Since potentially the expression level of all genes in a cell depends on a single global parameter λ, henceforth we shall call it the extrinsic noise parameter.

We quantify the nonlinearity of the coupling of transcription to cell age for all 1669 bursty genes by assuming that the mean mRNA count has a power-law dependence on the extrinsic noise parameter λ and that the exponent β can vary between the cell-cycle phases (Fig. 3B top). We then use linear regression on the log-log plots of mean counts versus λ to obtain the cell-cycle phase-dependent value of β for each gene. Since we are interested

in deducing the strength and direction of the linear relationship in log-log space, for each plot we computed the Pearson correlation coefficient r; the distribution of this coefficient (over all genes) is shown in Fig. 3B bottom left. For the large majority of genes (approximately 92%), the correlation coefficient is positive, with medians ranging between 0.55 and 0.78 (depending on the cell-cycle phase), thus providing strong evidence that the mean expression level increases with the extrinsic noise parameter via a power law. In Fig. 3B bottom right we show the distributions of the power-law exponent β; the medians are $\beta \approx 2.5$ in G1 and $\beta \approx 1$ in the S and the G2/M phases. Interestingly, this implies a nonlinear relationship between the mean counts and the extrinsic noise parameter λ in the G1 phase ($\approx 18\%$ of cells) and a linear one in the other two cell-cycle phases ($\approx 82\%$ of cells).

An analysis comparing β values in the different cell-cycle phases for each gene allows us to classify all genes into 6 categories—examples of genes for each of these categories and the discontinuous piecewise linear approximation obtained by joining the power laws for the different cell-cycle phases in log-log space are shown in Fig. 3C. In agreement with the median analysis above, the majority of genes (87%) fall in the first two classes where β in the G1 phase is larger than in the other two phases. We note that in this analysis we did not consider a small percentage of genes (8%) for which the correlation coefficient r in Fig. 3B bottom left was negative, or equivalently for which β in Fig. 3B bottom right was negative. In these cases, the expression level tends to reach a minimum (maximum) in the S phase, which implies a downward trend of the mean counts with λ in the G1 (G2/M) phase.

We remark that the linear relationships between the mean counts and λ found in the S and G2/M phases are reminiscent of a similar relationship between the mean counts and cell volume, in particular for cells maintaining concentration homeostasis [21, 23, 24, 60, 61]. This suggests that for most cells, the extrinsic noise parameter λ is roughly proportional to the cell volume. In fact, if we were to assume exponential growth [62] and doubling of cell size from birth to cell division (supported by flow cytometry analysis [57]), it immediately follows that the cell volume V can be roughly estimated as $V/V_b = \lambda = 2^\theta$ where V_b is the volume at birth. Nevertheless, since the median of β is considerably larger than 1 for genes in G1 and since the distributions of β are quite wide for genes in the other two cell-cycle phases, it is clear that in general λ does not simply linearly scale with the cell volume and likely depends on other cell-level factors. Hence it is best to simply interpret λ as an extrinsic parameter that simultaneously affects the expression of many genes via the modulation of one or more transcriptional parameters.

Ratios of burst parameters are robust to extrinsic and technical noise sources

Given the significant degree of extrinsic noise within each cell-cycle phase due to the coordination of transcription with cell age, it is natural to ask how does this affect the results of Fig. 2, where we used models that only took into account the change in transcription parameters and gene copy numbers with cell-cycle phase.

Hence, next we modify the six models used in the model fitting and selection algorithm to include extrinsic noise due to cell age. This is accomplished by assuming that the transcription rate ρ has a power-law dependence on the extrinsic noise parameter λ. This choice is motivated by the fact that it naturally implies that the mean number of counts also has a power-law dependence on λ, in agreement with our results in Fig. 3. This also implies that the burst size (which linearly scales with the transcription rate) is an increasing function of λ^β. Moreover, if λ is taken as a proxy for cell volume and $\beta = 1$, this implies a linear scaling of the burst size with volume, as suggested by smFISH studies [21, 23].

To determine β and the rest of the transcriptional parameters for each gene in a self-consistent manner, we proceed as follows. For each phase of the cell-cycle, we maximise the likelihood for each of the six models with age-dependent transcription and then select the optimal model using the Bayesian Information Criterion. Note that the likelihood is now maximised using not only the mRNA count for each cell but also its value of θ. One of the six models, the telegraph model augmented with a cell age-dependent transcription rate, is illustrated in Fig. 4A bottom (the top of the figure shows, for contrast, the age-independent telegraph model used in the data analysis illustrated in Fig. 2). We analyse only the 1537 genes out of 1669 bursty genes for which the optimal model is not Poisson or zero-inflated Poisson, and for which the regression analysis in Fig. 3B bottom left did not return a negative Pearson correlation coefficient. The histograms of the maximum likelihood estimated values of β in the G1 and G2/M phases for each gene are very similar to those previously determined using linear regression of the log-log plots of the counts versus the extrinsic noise parameter in Fig. 3B. Note that using the likelihood maximisation procedure we cannot determine β values for the S phase because we only analyse cells in G1 and G2/M for which the number of gene copies is clear—this number is necessary to properly account within the modelling framework (Fig. 4A right) for the change in noise magnitude when DNA replication occurs [63].

As previously mentioned, the procedure used to determine β simultaneously determines the burst frequency and size for each bursty gene in each of the two cell-cycle phases. Note that since the burst size is a function of cell age θ, to calculate the mean burst size for a gene in a certain cell-cycle phase, we average the burst size over the distribution of cell age for cells in this phase, as described in Methods. In Fig. 4B, we contrast the burst

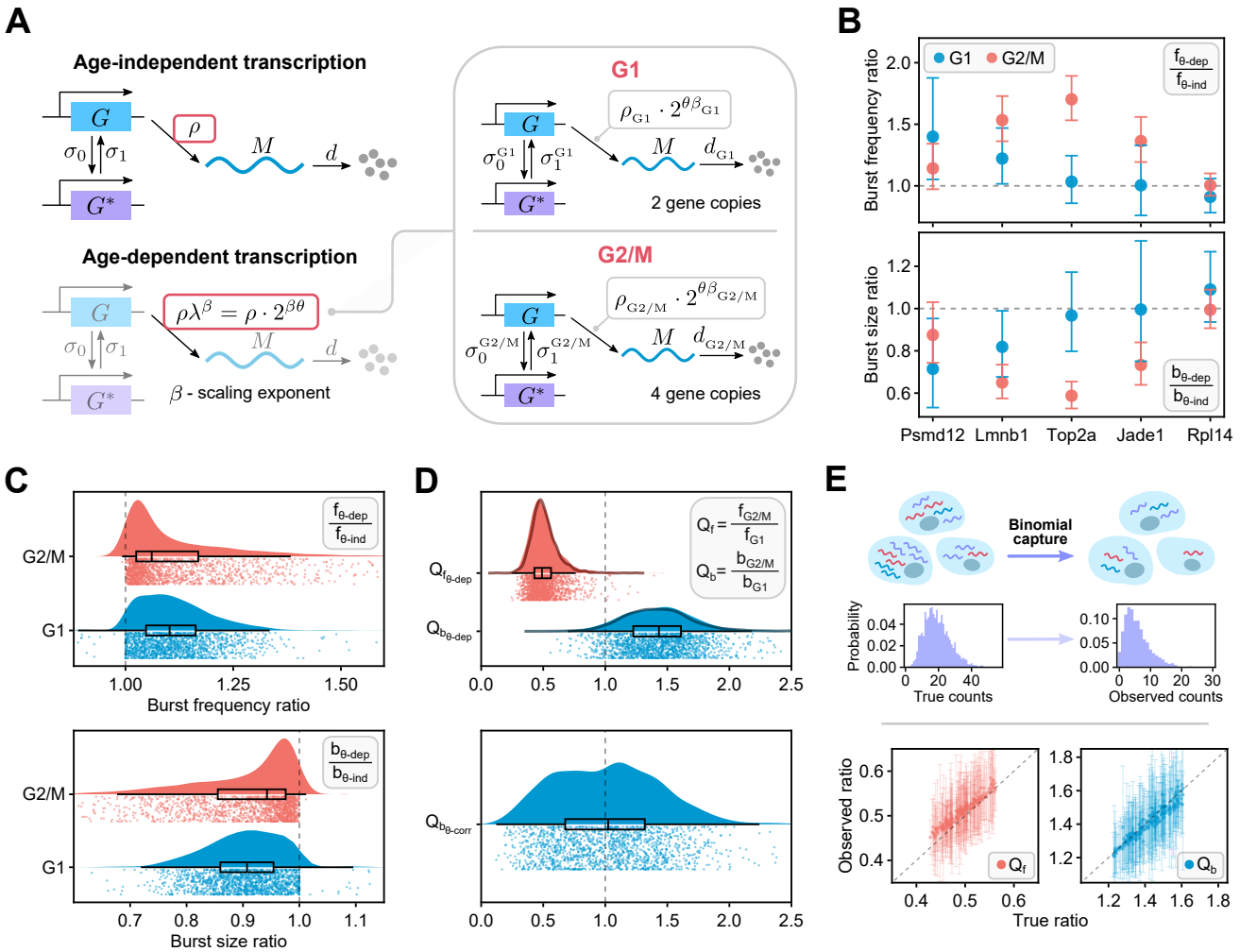

Figure 4: Effect of extrinsic noise due to age-dependent transcription rates and technical noise on the burst parameter inference. **(A)** Schematic of the telegraph models incorporating only intrinsic noise within a cell-cycle phase (θ-independent model) and both intrinsic and extrinsic noise within a cell-cycle phase (θ-dependent model). In the former model, mRNA synthesis is simply governed by a constant transcription rate ρ. In the θ-dependent model, we introduce a scaling of transcription with the extrinsic noise parameter $\lambda = 2^{\theta}$ via a power law with constant ρ and power $\beta \in [0, \infty)$, where β defines the gene-specific scaling exponent. The cell-cycle phase is accounted for in the same way as for the θ-independent model illustrated in Fig. 2B, but now using the extrinsic noise-adjusted transcription rates that are shown in the inset to the right. Namely, the mRNA count distribution in G1 is given by a convolution of two θ-dependent telegraph models with G1 phase-specific rate parameters, and similarly, the distribution in G2/M is a convolution of four corresponding θ-dependent telegraph models. **(B)** Burst frequency (top) and burst size (bottom) ratios between θ-dependent (θ-dep) and θ-independent (θ-ind) model estimates for five different genes in the dataset. The ratios are inferred both for G1 (blue) and G2/M (red) phases. The error bars show 95% confidence intervals. The grey dashed lines indicate no change in the parameter values between the two cell-cycle phases. **(C)** Raincloud plots of burst frequency (top) and burst size (bottom) ratios between θ-dependent and θ-independent model estimates for all genes, inferred for G1 (blue) and G2/M (red) phases separately. **(D)** Top: raincloud plots of burst frequency (red) and burst size (blue) ratios between G2/M and G1 phases for all genes estimated with the θ-dependent model, denoted as $Q_{f_{\theta\text{-dep}}}$ and $Q_{b_{\theta\text{-dep}}}$ respectively. The equivalent smoothed histograms of burst frequency and burst size ratios between G2/M and G1 obtained using the θ-independent model (shown in Fig 2D), i.e. $Q_{f_{\theta\text{-ind}}}$ and $Q_{b_{\theta\text{-ind}}}$, are superimposed as transparent black lines for comparison. Bottom: raincloud plot of the age-corrected burst size ratio between G2/M and G1, $Q_{b_{\theta\text{-corr}}}$, for all genes. The age-corrected burst size is obtained with the θ-dependent model after setting θ to zero (equivalent to ignoring the transcription scaling factor $2^{\beta\theta}$). **(E)** Top: Illustration of the binomial model of mRNA capture. In a sequencing experiment, only a fraction of the original mRNA counts will be detected due to the finite capture probability p that varies between cells. Bottom: scatter plot of the burst frequency (red) and burst size (blue) ratios between the G2/M and G1 phases estimated using the true count data and the corresponding downsampled (observed) data, obtained by simulating binomial transcript capture (see text for details).

parameters computed using the age-independent and the age-dependent models (Fig. 4A and Fig. 4B, respectively) for 5 genes—this allows us to assess the importance of including extrinsic noise within a cell-cycle phase due to the coupling of transcription to cell age. Note that for these 5 genes, the burst frequency estimated from the age-dependent models is larger than from those which have no age dependence, while the opposite is true for the burst size. In Fig. 4C we confirm for the vast majority of bursty genes that ignoring cell age in the stochastic model leads to a systematic overestimation of the burst size and an underestimation of the burst frequency—this is in agreement with a recent moment-based study [19].

Given these differences in parameter estimates between the age-independent and age-dependent models, the results for the ratios Q_f and Q_b of the burst frequency and size in G1 and G2/M shown in Fig. 2D and E are questionable. However, interestingly, when we compute Q_f and Q_b using the optimal age-dependent model (denoted by $Q_{f_{\theta\text{-dep}}}$ and $Q_{b_{\theta\text{-dep}}}$), the distributions of these ratios across the transcriptome are very similar to those computed using age-independent models (now denoted by $Q_{f_{\theta\text{-ind}}}$ and $Q_{b_{\theta\text{-ind}}}$, and shown by the black lines in Fig. 4D top). The implicit reason for this agreement between the two models is that the age-induced change in a parameter occurs in the same direction in both cell-cycle phases (for example the burst frequency in both phases is increased by taking the age fluctuations within a cell-cycle phase into account, as shown in Fig. 4C top) and hence the ratio of the parameters in the G2/M and G1 phases is not much affected by taking into account extrinsic noise due to cell age.

As an aside, we note that if we replot the distribution of the ratio of burst sizes in G2/M and G1 from the age-dependent models by forcing $\theta = 0$, we find that the distribution Q_b changes shape with a corresponding change of the median from 1.43 to 1.02 (compare the blue histograms in Fig. 4D top and bottom). This confirms that the increase in the burst size from G1 to G2/M can be largely explained by the coupling of the transcription rate to cell age within the cell cycle, a relationship akin to the coupling of the transcription rate to cell size [24].

In the foregoing estimation using both age-dependent and age-independent models, technical noise was to some extent taken into account by fitting models that allow for zero-inflation. However, for the vast majority of genes the zero-inflated models were rarely chosen by the model selection algorithm. Whilst this implies that zero-inflation is likely not an issue for the 10x sequencing dataset we have analysed (in agreement with Ref. [51]), technical noise introduced by the inherent downsampling of mRNA counts may still be affecting the analysis we have presented. For example, the real transcript count for a particular gene in a cell might be 10 molecules but only 6 of these are captured—in this case there is no inflation of the zero count but clearly technical noise has impacted the number of measured counts. This implies that downsampling could have possibly introduced a systematic bias in our estimation of the ratio of burst parameters in the two cell-cycle phases.

To address this concern, using a binomial mRNA capture model [64, 65], we developed a stochastic simulation-based strategy to estimate the bias introduced by ignoring this important source of technical noise in the sequencing data. We consider 200 genes randomly chosen from a subset of all bursty genes that are best fit by the age-dependent negative binomial model with the Q_f and Q_b values in their respective interquartile ranges, as shown in Fig. 4D top. The capture efficiency p for each cell is a random number drawn from Beta$(30, 70)$ distribution (with the mean $\langle p \rangle = 0.3$ and the coefficient of variation $\text{CV}_p \approx 0.15$). For each gene in each cell-cycle phase, we use the best fit model parameters and rescale the transcription rate (burst size) by the inverse of the average capture efficiency $\langle p \rangle$. The observed mRNA count x in each cell is obtained by sampling the true count y from the rescaled model and subsequently downsampling y, so that $x \sim \text{Binomial}(y, p)$. We fit the age-dependent negative binomial model to the downsampled data to estimate the detected Q_f and Q_b. In Fig. 4E bottom we show a scatter plot of the burst parameter ratios between the G2/M and G1 phases estimated using the true count data and the corresponding downsampled (observed) data. Each point in the plot is the average over 100 repeats of the procedure, where the error bars indicate the standard deviation. Remarkably our results show that downsampling due to a finite capture probability has very little effect on the estimated ratio of burst parameters in the two cell-cycle phases.

In Appendix A we provide a theoretical justification for this surprising resilience to technical noise. Assuming a population of an infinite number of cells, cell-to-cell variation in the burst parameters and that counts from each gene copy are distributed according to a negative binomial distribution, we show that the ratio of the burst frequency and size in G1 and G2/M estimated using the measured count data (denoted by the subscript x) is approximately related to the true ratio of the burst frequency and size in G1 and G2/M (estimated using the actual data before downsampling, which we denote by the subscript y) by the formula:

$$\frac{f_{x,\text{G2/M}}}{f_{x,\text{G1}}} = \frac{f_{y,\text{G2/M}}}{f_{y,\text{G1}}}\eta, \quad \frac{b_{x,\text{G2/M}}}{b_{x,\text{G1}}} = \frac{b_{y,\text{G2/M}}}{b_{y,\text{G1}}}\eta^{-1}, \tag{2}$$

where

$$\eta = \frac{1 + \mathrm{CV}_p^2(1 + 2f_{y,\mathrm{G1}})}{1 + \mathrm{CV}_p^2(1 + 4f_{y,\mathrm{G2/M}})}. \tag{3}$$

Note that CV_p^2 is the coefficient of variation squared of the distribution of the capture probability p across the cell population. Simulations indicate the theory is accurate provided CV_p^2 is not very large. From Eq. (3) it follows that if all cells have the same capture probability, i.e. $\mathrm{CV}_p^2 = 0$, $\eta = 1$; if the distribution of capture probability is narrow, i.e. $\mathrm{CV}_p^2 \ll 1$, then $\eta \approx 1$. Another particularly interesting observation is that if the true burst frequency in G2/M is half that in G1, then independent of the value of CV_p^2, $\eta = 1$. This last property is likely the main reason explaining the agreement of the ratio of burst parameters using true and downsampled data in Fig. 4E, since the burst frequencies chosen in the simulations were consistent with the distribution of Q_f in Fig. 4D, which displays a peak at 0.49.

Ignoring gene copy number variations leads to inaccurate parameter estimation

Thus far we have addressed how the burst parameters vary with the cell-cycle phase. However in the literature there has often been an interest in simply calculating a single set of parameters for each gene, in the hope that by correlating these parameters with certain gene features, one can dissect the molecular determinants of stochastic gene expression. For example, in Ref. [43] burst sizes obtained by fitting the telegraph model to scRNA-seq data were found to be significantly larger for genes with TATA elements in their core promoters.

The problem with this approach is how to properly infer a single set of parameters for each gene because, as shown in Ref. [19], the ones inferred from the telegraph model are generally not equal to the population-averaged parameters, even if the mRNA count data is single-allele specific and from one cell-cycle phase. This is because when there is extrinsic noise due to cell-to-cell differences in the rate parameters, the moments of mRNA numbers do not simply depend on the population average of the single-cell parameters (as implicitly assumed by the standard telegraph model) but rather are a function of the higher moments of these parameters.

To shed light on how the choice of model assumptions affects the estimated population-averaged parameters for each gene when using non-allele-specific mRNA count data from cells in different cell-cycle phases (the most common type of data), we determine these parameters for the mouse embryonic stem cell scRNA-seq dataset using four distinct approaches:

1. Model I (no extrinsic noise due to cell-cycle phase or cell age): given the mRNA count from each cell (independent of its cell-cycle phase) we maximise the likelihood of a stochastic model (for two gene copies). This leads to a single set of allele burst parameters.

2. Model II (extrinsic noise due to cell age but not due to cell-cycle phase): given the mRNA count and the cell age parameter θ for each cell (independent of its cell-cycle phase) we maximise the likelihood of a stochastic model (for two gene copies) with the transcription rate having a power-law dependence on the extrinsic noise parameter λ, which is a function of the cell age θ. This leads to an estimate for the allele burst frequency and to estimates of the cell-dependent allele burst size, which can be averaged over the population to give the mean allele burst size, as discussed in Methods.

3. Model III (extrinsic noise due to cell-cycle phase but not due to cell age): the procedure in Model I is performed for cells only in the G1 phase and then separately repeated for cells in the G2/M phase, with the difference that in the latter case the number of gene copies is 4. This leads to two sets of allele burst parameters, one for each cell-cycle phase; the mean burst parameters are then computed from the weighted average of the two sets of parameters, with the weights w_1 and w_2 defined as

$$w_1 = \frac{2n_{\mathrm{G1}}}{2n_{\mathrm{G1}} + 4n_{\mathrm{G2/M}}}, \quad w_2 = \frac{4n_{\mathrm{G2/M}}}{2n_{\mathrm{G1}} + 4n_{\mathrm{G2/M}}}, \tag{4}$$

   where n_{G1} and $n_{\mathrm{G2/M}}$ denote the number of cells in each phase, and the factors 2 and 4 correspond to the number of gene copies per cell in G1 and G2/M.

4. Model IV (extrinsic noise due to both cell age and cell-cycle phase) the procedure in Model II is performed for cells in G1 (2 gene copies) and then separately repeated for cells in G2/M (4 gene copies), giving two allele burst parameter sets, each averaged over all cells in that particular cell-cycle phase; equivalently to Model III, the mean burst parameters of the whole population are computed from the weighted average of the two sets of parameters, with the weights given by Eq. (4).

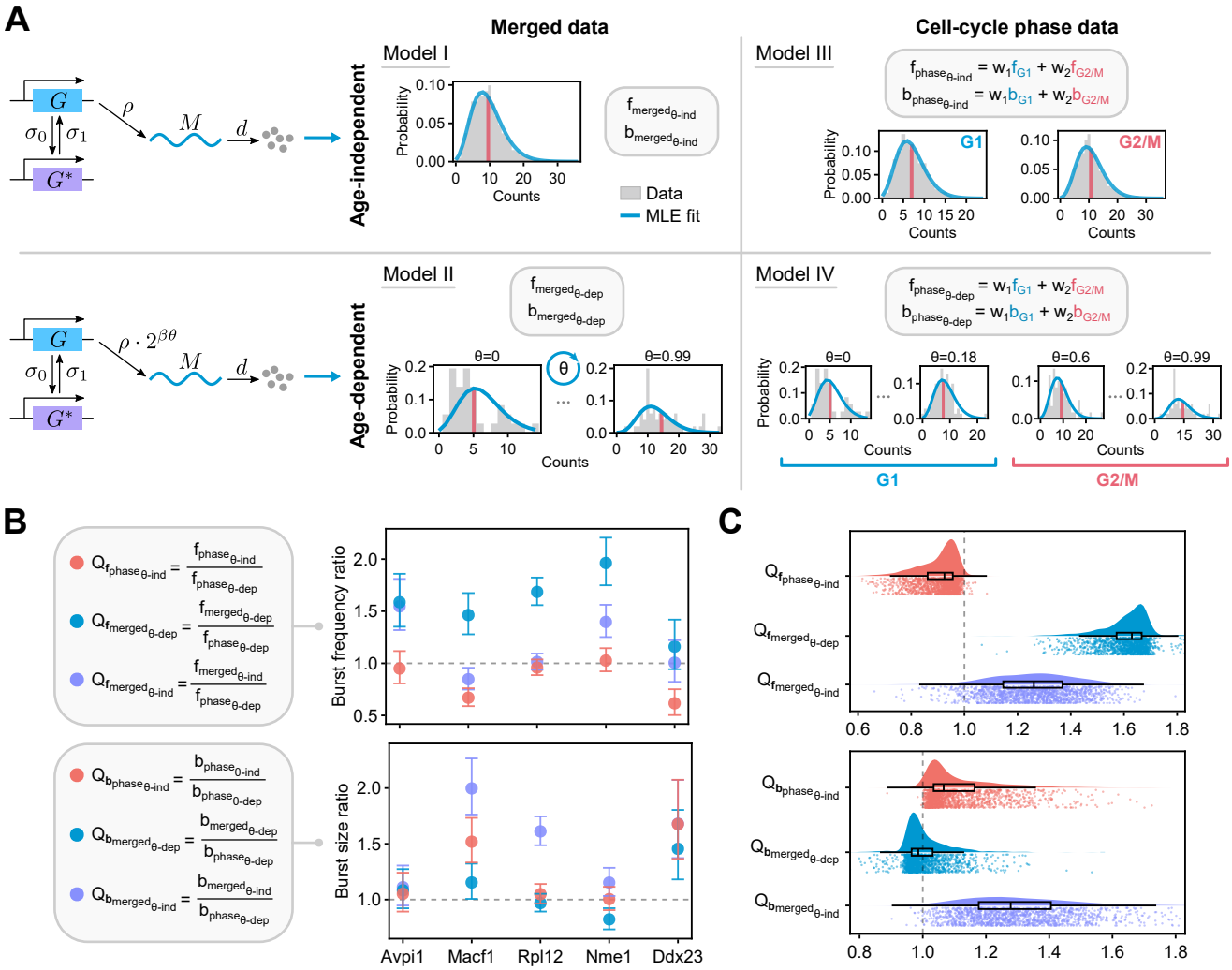

Figure 5: Relationship between the model type and the estimated population-averaged burst parameters. **(A)** Schematic of four types of modelling approaches and their corresponding population-averaged parameters. Namely, we consider an age-independent model (no extrinsic noise due to cell age) fit either on the merged mRNA count distribution of all cells in G1 and G2/M phases ('merged θ-ind' or model I) or specifically on the distributions of cells either in G1 or G2/M ('phase θ-ind' or model III). Similarly, we also fit an age-dependent model (taking into account the extrinsic noise due to cell age) on merged data ('merged θ-dep' or model II) and cell-cycle phase-specific data ('phase θ-dep' or model IV). For models fit on the cell-cycle phase-specific data, i.e. that account for the extrinsic noise due to cell-cycle phase, we consider population-averaged parameters given by a weighted sum over the G1 and G2/M parameter estimates, with weights w_1 and w_2 defined by Eq. (4), in order to enable a direct comparison to the parameter estimates for merged data. The telegraph model diagrams on the left illustrate the difference in the transcriptional rates between the age-independent and age-dependent models. The histograms show the corresponding data distributions for Prpf8 gene in each modelling scenario, where the red vertical lines indicate the mean counts and the blue lines are the maximum likelihood (MLE) fits. **(B)** Burst frequency (top) and burst size (bottom) ratios between the different model estimates for five sample genes from the scRNA-seq dataset. The ratios are computed between the parameters of models I, II and III to the ground truth parameters of model IV. The corresponding ratios and the colour coding are indicated in the legend on the left. The error bars show 95% confidence intervals. The grey dashed lines indicate no change in the parameter values between the two compared models. **(C)** Raincloud plots of burst frequency (top) and burst size (bottom) ratios between the different model estimates for all genes, as outlined in (B).

The models are illustrated in Fig. 5A. Clearly, Model IV is the most realistic and hence we shall consider the population-averaged allele burst parameters of this model as the ground truth parameters. To assess the systematic bias introduced by the lack of a description of the variation of gene copy numbers and age between cells, we computed ratios of the inferred parameters from Models I, II and III and the ground truth parameters from Model IV. These ratios, defined in the left-hand panel of Fig. 5B, are shown for 5 genes on the right-hand panel of the same figure. Distributions of the 3 ratios for all 1537 bursty genes are shown in Fig. 5C. The overall systematic tendencies of each model are clear. In line with Ref. [19], which compares inference from two models with the same gene copy number but in which only one takes into account extrinsic noise on the rate parameters, Model III underestimates the mean burst frequencies and overestimates the mean burst sizes (red histogram). However, Models I and II (purple and blue histograms, respectively), which do not take into account the variation of gene copies amongst cells, systematically overestimate the burst frequency for the majority of genes. Of particular curiosity is the contrast between the predictions of the two models: while including extrinsic noise due to cell age improves the prediction of burst size (median relative error decreased from 28% for Model I to 1% for Model II), it reduces the prediction accuracy for the burst frequency (median relative error increased from 26% for Model I to 63% for Model II). These results make sense in the light that the true burst size (not the burst frequency) depends on cell age, a relationship that is accounted for in Model II.

In Appendix C, we develop a simple theory to more precisely quantify the errors in the burst parameters estimated using models that do not properly account for the gene copy number in a specific cell-cycle phase. We show that if a model which assumes R independent gene copies and accounts for the extrinsic noise in the burst size is fit to non-allele-specific data from an infinite population of cells, each with S gene copies and in which there are cell-to-cell differences in the burst size, then the ratio of the estimated and true burst frequency is approximately S/R, while the ratio of the estimated and true burst size is approximately 1; this is provided the extrinsic noise is very small. As the magnitude of extrinsic noise in the burst size increases, differences between the estimated and true burst size become more apparent, but generally the error in the burst frequency exceeds that in the burst size if $S \neq R$, i.e. if the model incorrectly accounts for the number of gene copies. For instance, errors of this type are made when fitting a model with one gene copy to non-allele-specific data in eukaryotic cells, such as in Ref. [56], or when we previously fit Model II to the scRNA-seq mouse embryonic stem cell data used in this paper. In particular, the errors due to inference using Model II stem from the fact that it assumes 2 gene copies but many cells in the population are in the G2/M phase and instead have 4 gene copies. The theory with $S = 4, R = 2$ predicts the ratio of the estimated and true burst frequency to be ≈ 2, and we measured a median of 1.63. The theory also predicts that the burst size is relatively accurately estimated and indeed we measured a median of 0.99. Moreover, the distribution of the ratio between the true and estimated burst sizes is rather wide (see Fig. 5C), in agreement with the theoretical prediction that as the size of extrinsic noise increases, errors in the estimated burst size also increase.

In summary, we have shown that using a model that takes into account only extrinsic noise due to cell age but without accounting for the variation of gene-copy number with cell-cycle phase is generally not a useful inference strategy. This suggests that incorporating gene copy number fluctuations is more important for predictive accuracy than extrinsic noise due to cell age, a hypothesis that is reinforced by our previous results in Fig. 4D and by the fact that the red-coloured histograms in Fig. 5C (comparing the predictions of Models III and IV) have medians that are quite close to 1 (median relative errors of 8% and 7% for the burst frequency and size, respectively).

Discussion

In this paper, by fitting stochastic models of gene expression to single-cell sequencing data that is cell-cycle specific, we have estimated the dependence of the burst parameters (the burst frequency and size) on the position of the cell cycle for about 1500 bursty genes in undifferentiated mouse embryonic stem cells. Our results reveal that on average, as the cell-cycle progresses from G1 to G2/M, the burst frequency decreases by about 50% while the burst size increases by approximately the same amount. This balancing principle between the burst parameters implies that partial, rather than complete, gene dosage compensation is the norm in the murine transcriptome. We have also shown that the increase in the burst size is due to a coupling of the transcription rate to cell age within the cell cycle, a relationship reminiscent of the coupling of the transcription rate to cell size reported in previous studies [24], which is thought to be an important factor contributing to concentration homeostasis [21, 23, 66]. Importantly, our study has also revealed that while the values of the burst parameters are quite sensitive to the choice of the mechanistic model used for inference, the ratios of the burst parameters in the G1 and G2/M phases are robust estimates, i.e. the ratios are largely independent of whether the models take into account or ignore various sources of extrinsic and technical noise, as long as the model explicitly takes into account the gene copy number variation

between cells if the scRNA-seq data is not allele-specific.

We note that whilst the burst parameters have been previously estimated using model-based inference applied to smFISH and scRNA-seq data, in the vast majority of cases, these were not estimated for individual cell-cycle phases [11–14, 19, 42, 44, 45, 48]. However there are a few published studies which have reported this. An scRNA-seq-based study reported a comparison of single-allele burst parameters in G1, S and G2/M phases estimated using the telegraph model (see Extended Data Figure 10 in [43]); because of the small numbers of cells in each phase (165, 28 and 31 in G1, S and G2/M, respectively) the errors in these parameters are necessarily large and it is hence difficult to come to any reasonable conclusion on how the burst parameters vary across the cell cycle. Two smFISH-based studies [20, 22] used model-based inference to estimate the burst parameters before and after replication but these were limited to only three eukaryotic genes and did not account for extrinsic noise due to factors varying within a cell-cycle phase, such as cell size. In Ref. [21] the burst parameter dependence with cell-cycle phase and cell size for about 25 mammalian genes were roughly estimated directly from the number of active transcription sites and their fluorescent intensity in smFISH measurements (without the use of model-based inference). Therefore, our study is the first to our knowledge that uses mechanistic model-based inference to reliably estimate the ratio of the burst parameters pre- and post-replication for over a thousand bursty eukaryotic genes.

Our results quantitatively agree with the findings of other studies and extend their predictions to a far larger number of genes. We found partial dosage compensation to be the norm across the transcriptome. This is because the distribution of the ratio of the mean counts in G2/M and G1 had a median close to 1.40, whereas perfect gene dosage compensation would be consistent with 1 and no dosage compensation with 2. This is in agreement with the findings of previous smFISH studies investigating gene expression in a few eukaryotic genes: in Ref. [22] it was found that for the Oct4 and Nanog genes, the fold change in the nascent mRNA level between the G1 and G2/M phases was 1.28 and 1.51 respectively, while in Ref. [20] for the GAL10 gene the fold change was measured to be 1.7. We inferred the transcriptome-wide distribution of Q_f, the ratio of the estimated burst frequencies per allele in G2/M and G1, and found that its median is equal to 0.49—this distribution is fairly tight as the first and third quartiles are 0.43 and 0.56, values close to the median. This is in good agreement with the few cases in which these were estimated in previous smFISH studies of eukaryotic gene expression. In Ref. [22] for the Oct4 and Nanog genes in mouse embryonic stem cells, it was found that the burst frequency ratio was $Q_f = 0.63 \pm 0.06$, while in Ref. [20] for the GAL10 gene in yeast, it was found that the frequency ratio was $Q_f = 0.66 \pm 0.17$. In Ref. [21] from smFISH measurements of the ratio of the number of active transcription sites in a cell and the total number of gene copies it was estimated that $Q_f \approx 0.5$ for a few tens of genes in human primary fibroblast and lung cancer lines. We also found that the transcriptome-wide distribution of Q_b, the ratio of the burst size per allele in G2/M and G1, has a median equal to 1.43 with first and third quartiles equal to 1.23 and 1.61. This is in contrast to that estimated in Ref [20] for the GAL10 gene, which found no dependence of the burst size on the cell-cycle phase, with $Q_b = 1.07 \pm 0.31$. In Ref. [21] the burst size estimated for a few tens of genes was also not found to change with the cell-cycle phase but rather it increased with the cell size. We note that in this case, the cell size and the cell-cycle phase were not very clearly correlated ("cell volume varied as much for cells in individual phases of the cell cycle as the population overall, with a shift in the distribution toward G2 cells being larger" [21]). This observation provides a simple way to explain our results for mouse embryonic stem cells. If the burst size increases with the cell size and, additionally, the cell size clearly correlates with the cell-cycle phase, the burst size will increase with cell age and with the cell-cycle progression from G1 to G2/M—this hypothesis is supported by the fact that when we account for the burst size dependence on cell age, the median of the distribution of Q_b shifts to from 1.43 to 1.02.

Our results also emphasise that the common practice of reporting the burst parameter values is not particularly useful or insightful because these are highly sensitive to the assumptions of the mechanistic model used for inference. While this point has recently been made when inferring burst parameters from allele-specific scRNA-seq data in the G1 phase [19], our study goes further by comparing the inference results using four different types of models applied to mRNA count data that is not allele-specific. In particular, we found that the inference of burst parameters from this type of data, using models that correct for extrinsic noise due to coupling of the transcription rate to an exogenous factor that varies within each cell-cycle phase (such as size or age) but without correcting for the variation of the gene copy number between different cell-cycle phases, leads to a significant overestimation of the burst frequency with median relative errors of approximately 60%, although the error in the burst size is significantly smaller (typically less than 20%). The error in the burst frequency was actually larger than if we used a model that did not at all account for extrinsic noise in the transcription rate. Hence, our results suggest that several recently developed approaches using models that account for extrinsic noise in the transcription rate but assume a fixed number of gene copies per cell (typically one) to infer the burst parameters from non-allele-specific single-cell sequencing data [44, 45, 56] should be used with great care. Instead, the estimation of ratios of the burst parameters in G1 and G2/M from cell-cycle phase specific scRNA-seq data, using mechanistic models that account for gene

copy number variation across the cell cycle, is generally a far more robust inference strategy.

An apparent limitation of our approach is that for each cell-cycle phase, we used the steady-state solution of the mechanistic models for the maximum likelihood fitting to the scRNA-seq data. This means that we can only estimate the ratio of the burst frequency and the degradation rate, and not the burst frequency per se, and hence we cannot distinguish between changes in the burst frequency and the degradation rate across the cell cycle. Furthermore, since the mean expression increases with cell age, one could instead argue for the use of time-dependent solutions of mechanistic models. The primary reason why we adopted the steady-state approach is its practicality: these solutions are known in analytical form and, because of their simplicity, they can be straightforwardly used for parameter inference. In addition, there is compelling evidence that our results are unlikely to be significantly modified by the use of the time-dependent solutions of more complex mechanistic models of gene expression. In Ref. [67] the exact analytical time-dependent solution of an extended telegraph model that explicitly considers the doubling of gene copy numbers upon DNA replication, dependence of the mRNA synthesis rate on cellular volume, exponential increase of the cell volume within the cell cycle, gene dosage compensation, partitioning of molecules during cell division and fixed cell-cycle duration T was derived. This solution is complex but it was shown that in the special case that the transcription rate scales linearly with the cell volume, gene expression is bursty and the burst frequency in G2/M is 1/2 that in G1 (a condition approximately satisfied by many genes in our dataset as shown in Fig. 2D and Fig. 4D top), then the time-dependent distribution of mRNA numbers at a certain time $t \in [0, T]$ within the cell cycle is approximately the same as the steady-state solution of the standard telegraph model with the transcription rate set equal to a constant multiplied by the cell volume at time t (see Fig. 3C in Ref. [67]). Note that this applies even if mRNA degradation is not very fast, as is the case for many mammalian genes [68]. Our inference strategy is based on this equivalence, since in our case the exogenous parameter is the cell age instead of the cell volume, and the steady-state solution of the telegraph model with a transcription rate that depends on the cell age (what we have called an age-dependent model) was fit to the mRNA distributions measured at specific cell ages.

In conclusion, we have devised a simple, tractable and robust approach to estimate from scRNA-seq data the dependence of the transcriptional burst parameters on the position of the cell cycle. While the use of mechanistic models for the extraction of the parameters controlling gene expression has become commonplace in fluorescence-based single-cell transcriptomics, their use in the field of single-cell sequencing has severely lagged behind [69]. This is largely due to the significant technical noise inherent to the sequencing technology and the difficulty of quantifying the sources of extrinsic noise. However, as we have here shown, with a carefully designed model-based inference approach, these issues present no impediment to building quantitative models of gene expression at the transcriptome level.

Methods

Processing of single-cell RNA sequencing data

In our analysis, we use the scRNA-seq dataset of a population of mouse embryonic stem cells with inhibited differentiation generated by Riba *et al.* [57] using the 10x Genomics Chromium platform, containing spliced and unspliced transcript counts for over 5000 cells and more than 10000 genes. Each cell in the population has an associated cell age θ (originally referred to as the *transcriptional phase* in [57]) which characterises its position along the cell cycle and can also be used to determine the cell-cycle phase (G1, S or G2/M). The parameter $\theta \in [0, 1)$ is a continuous periodic variable where $\theta = 0$ indicates the beginning of the G1 phase of a newly formed cell and $\theta = 1$ corresponds to cell division. In the following, we reiterate the main steps of the data processing and downstream analysis published in [57] and discuss how it is reprocessed for the purposes of this paper.

In order to infer the age of each cell, Riba *et al.* developed a neural network-based method called DeepCycle. Their approach is motivated by the concept of RNA velocity, which can be used to estimate the future transcriptional state of a cell from unspliced/spliced mRNA counts [35, 70]. Namely, the authors proposed to order cells along the cell-cycle trajectory by fitting the circular expression patterns of cell cycle-related genes in the unspliced-spliced space using an autoencoder. The data is initially processed using scVelo [36]: the spliced/unspliced counts are normalised to the median of total counts per cell; a nearest-neighbour graph is constructed from principal component analysis on logarithmised counts (using the first 20 principal components); spliced/unspliced counts are smoothed out for each cell by averaging the counts over its 30 nearest neighbours. The next step in the DeepCycle procedure is to identify the cell cycle-related genes that exhibit clearly cycling transcriptional dynamics in the unspliced-spliced space. The z-scored spliced and unspliced counts of the selected genes for each cell comprise the training data for the autoencoder, which learns to denoise and reconstruct the expression patterns of the cycling

genes. This way, the single latent variable of the autoencoder, the cell age parameter θ, effectively characterises the relative position of a cell along the cell cycle. The variable is then discretised as $\theta \in [0, 0.01, \ldots, 0.99]$ and all cells are assigned to their corresponding bins based on the autoencoder predictions. Finally, specific values of θ can be associated with the transitions between the main cell-cycle phases based on the expression of cell-cycle marker genes and the total transcripts per cell—this allows mapping the cells to the G1, S and G2/M cell-cycle phases and to align θ so that cell division occurs at $\theta = 1$. Specifically, the S/G1 and G1/G2 transitions occur at θ values of 0.26–0.27 and 0.63–0.64 respectively. Although RNA velocity-based methods have various shortcomings [71], Riba *et al.* performed a number of self-consistency checks that establish the validity of DeepCycle for the considered datasets. For instance, the authors verified that the expected expression patterns of marker genes are recovered and also provide complementary validation using flow cytometry analysis.

As the processed mESC dataset published in [57] does not contain the discrete UMI counts needed for our inference and these cannot be easily retrieved, we start from the raw FASTQ data generated by the authors and reprocess it following the same methodology. The reads were aligned to the mouse reference transcriptome (mm10-2020-A) using CellRanger count pipeline (v7.1.0), and the spliced and unspliced transcripts were quantified using velocyto (v0.17.17) [35]. Low quality cells and genes were filtered out using scVelo (v0.2.5) [36]. More specifically, we select genes that are expressed in at least 200 cells and cells that have a total number of UMIs greater than 10000, express more than 3000 unique genes and have a low percentage of mitochondrial genes (under 15%)—the resulting dataset contains 5686 cells and 11544 genes. Next, as described in the paragraph above, we process the data further using scVelo and run DeepCycle to assign the transcriptional phase $\theta \in [0, 0.01, \ldots, 0.99]$ and the cell-cycle phase (G1, S or G2/M) for each cell. Although not exactly equivalent, the obtained phase characterisation is similar both quantitatively and qualitatively to the original processed data in Ref. [57].

In DeepCycle, the transition between mitosis and G1 is defined so that $\theta = 0.99$ corresponds to the peak in the average total RNA counts per cell, whereas $\theta = 0$ coincides with their immediate drop. Although one may expect to observe a sharp halving in the total counts upon division, this decline is spread over a wider range of θ bins assigned to the G1 phase (Fig. 3C of [57]) and it remains unclear how many of the cells in this range are misclassified by DeepCycle as belonging to the G1 phase. Therefore, as a precaution, we choose to remove all cells assigned to the $[0, 0.09]$ range of θ values and rescale the remaining $\theta \in [0.1, 0.11, \ldots, 0.99]$ into the same range (0–0.99) as $\theta \mapsto (\theta - 0.1) \times 0.99/0.89$. The truncated dataset contains 5294 cells in total, out of which 941 are assigned to the G1 phase, 2240 to the S phase and 2113 to the G2/M phase. Finally, we also remove genes with low expression, i.e., genes that in either G1 or G2/M phases have a mean transcript abundance below one. The parameter inference is then performed on the spliced mRNA counts of the remaining 3882 genes.

Mechanistic models of stochastic gene expression

Age-independent telegraph, negative binomial and Poisson models

In this paper, we consider the standard model of stochastic gene expression (the telegraph model) [9] and its variants. The telegraph model is defined in terms of the following effective reactions:

$$G \underset{\sigma_1}{\overset{\sigma_0}{\rightleftharpoons}} G^*, \quad G \xrightarrow{\rho} G + M, \quad M \xrightarrow{d} \emptyset\,, \tag{5}$$

where a gene switches between the active (G) and the inactive (G^*) promoter states with rates σ_0 and σ_1, and in the active state produces spliced mRNA (M) with rate ρ, which subsequently degrades with rate d. Transcriptional bursting is characterised by the burst frequency f and burst size b, which for the telegraph model are given by $f = \sigma_1$ and $b = \rho/\sigma_0$. Note that the burst size is the mean number of mRNA molecules produced when the gene is active.

The chemical master equation (CME) describing the telegraph model can be solved exactly in the steady state for the probability distribution of mRNA counts m [1, 9]:

$$P(m) = \frac{1}{m!}\left(\frac{\rho}{d}\right)^m \frac{\left(\frac{\sigma_1}{d}\right)_m}{\left(\frac{\sigma_1}{d}+\frac{\sigma_0}{d}\right)_m}\, {}_1F_1\left(\frac{\sigma_1}{d}+m, \frac{\sigma_1}{d}+\frac{\sigma_0}{d}+m, -\frac{\rho}{d}\right)\,, \tag{6}$$

where $(x)_m = \Gamma(x+m)/\Gamma(x)$ is the Pochhammer symbol and ${}_1F_1$ is Kummer's (confluent hypergeometric) function. Alternatively, Eq. (6) can be represented by a Beta-Poisson distribution [42, 72]. The rate parameters $(\sigma_0, \sigma_1, \rho)$ as they appear in the function are divided by the mRNA degradation rate d, and hence their values cannot be determined using only static snapshot data without explicit measurements of the degradation rate. Since such measurements are not available, we can only infer the ratios of the rate parameters and the mRNA degradation rate, and hence without loss of generality, in what follows we set $d = 1$.

In the limit $\sigma_0 \gg \sigma_1$, the gene is mostly inactive and transcription only occurs in short bursts, and the steady-state solution of the telegraph model reduces to a negative binomial distribution:

$$P(m) = \mathrm{NB}\left(m; \sigma_1, \frac{\sigma_0}{\rho + \sigma_0}\right) = \mathrm{NB}\left(m; f, (1+b)^{-1}\right), \tag{7}$$

parameterised by the burst frequency $f = \sigma_1$ and the burst size $b = \rho/\sigma_0$. Note that this corresponds to the steady-state CME solution of the following reaction network [10]:

$$G \xrightarrow{\sigma_1} G + rM, \quad M \xrightarrow{1} \emptyset, \tag{8}$$

where r is a geometrically distributed random variable with mean b.

In the limit $\sigma_1 \gg \sigma_0$, the gene is mostly active and the model reduces to a simple birth-death process:

$$G \xrightarrow{\rho} G + M, \quad M \xrightarrow{1} \emptyset, \tag{9}$$

which in steady-state conditions is described by the Poisson distribution $P(m) = \mathrm{Pois}\,(m; \rho)$. The burst frequency and burst size cannot be defined in this case as the distribution is characterised by a single parameter, the transcription rate ρ, which equals the mean mRNA count.

Extending the models to include age dependence

As discussed in the Results section, the telegraph model needs to be extended to incorporate the power-law dependence of the mean expression level with the extrinsic noise parameter $\lambda = 2^{\theta}$, where θ is the cell age given by DeepCycle.

As visualised in Fig 4A bottom, following Ref. [67] we define the *age-dependent* telegraph model with a transcription rate $\rho\lambda^{\beta}$, where the power $\beta \in [0, \infty)$ is a gene-specific age scaling exponent (an extra parameter to be estimated). Note that the proportionality constant ρ in this case has a different interpretation from the transcription rate ρ of the standard telegraph model. The mRNA count distribution for a specific value of θ is given by a simple modification of Eq. (6) (with $d = 1$):

$$P_{\theta}(m) = \frac{1}{m!}\left(\rho \cdot 2^{\beta\theta}\right)^{m} \frac{(\sigma_1)_m}{(\sigma_1 + \sigma_0)_m}\, {}_1F_1\left(\sigma_1 + m, \sigma_1 + \sigma_0 + m, -\rho \cdot 2^{\beta\theta}\right). \tag{10}$$

Although the burst frequency remains the same as for the age-independent model, the burst size at each θ is now given by $b_{\theta} = \rho \cdot 2^{\beta\theta}/\sigma_0$. The burst size averaged over all cells is given by

$$b = \frac{\rho}{\sigma_0} \sum_{\theta=\theta_i}^{\theta_f} g(\theta) \cdot 2^{\beta\theta}, \tag{11}$$

which is defined for a specific cell-cycle phase where θ varies from θ_i to θ_f and $g(\theta)$ is the distribution of θ bins over all cells assigned to that phase (the age distribution in the cell-cycle phase).

Similarly we can obtain the age-dependent negative binomial distribution for the specified θ as:

$$P_{\theta}(m) = \mathrm{NB}\left(m; f, (1 + b' \cdot 2^{\beta\theta})^{-1}\right). \tag{12}$$

Note that the distribution is now parameterised by three parameters (f, b' and β). The average burst size over all cells in a certain cell-cycle phase is given by $b = b' \sum_{\theta=\theta_i}^{\theta_f} g(\theta) \cdot 2^{\beta\theta}$.

Finally, the age-dependent Poisson model has two parameters (ρ and β) and can be simply defined as $P_{\theta}(m) = \mathrm{Pois}\left(m; \rho \cdot 2^{\beta\theta}\right)$.

Extending the models to account for zero inflation of scRNA-seq data

Substantial technical noise in scRNA-seq experiments [49, 73] can lead to the presence of excess zero counts in the data. The treatment of such non-biological zeros has been a subject of active debate [50, 74]; although it is commonly taken into account by considering zero-inflated models, a number of recent studies have argued that zero-inflation is largely avoided in UMI count experiments [50–52, 75, 76]. Nevertheless, to allow for more flexibility in our inference procedure, for age-independent models we model zero-inflation explicitly as follows:

$$P_{ZI}(m) = \begin{cases} p_0 + (1 - p_0)P(m), & \text{if } m = 0, \\ (1 - p_0)P(m), & \text{if } m \geq 1, \end{cases} \tag{13}$$

where the parameter $p_0 \in [0,1]$ describes the probability of zero-inflation (fraction of zeros that are non-biological). For age-dependent models, the formula is the same except $P(m)$ is replaced by $P_\theta(m)$. For parameter inference we specifically consider (age-independent or age-dependent) telegraph, negative binomial and Poisson models, and their zero-inflated counterparts, i.e. 6 models as illustrated in Fig. 1C.

Accounting for multiple gene copy numbers

The models considered above describe the transcription from a single allele. However, the mESC dataset that we are analysing in this paper is non-allele-specific, and mouse embryonic stem cells possess two alleles which are replicated during the S phase of the cell cycle. Therefore, we need to consider the combined expression of two gene copies in the G1 phase and four gene copies in the G2/M phase. By assuming that the expression of the alleles is identical and independent from each other, we can model the total count distribution in the G1 phase by convolving the allele-specific count distribution P with itself:

$$P_{\mathrm{G1}}(m) = (P * P)\,(m) = \sum_{k=0}^{\infty} P(k)P(m-k). \tag{14}$$

Similarly, the total count distribution in the G2/M phase will be given by the convolution of P with itself 4 times:

$$P_{\mathrm{G2/M}}(m) = (P * P * P * P)\,(m). \tag{15}$$

The allele-specific distribution $P(m)$ is given by the telegraph, negative binomial or Poisson models, or their zero-inflated counterparts, with cell-cycle phase-specific parameters—this is schematically illustrated in Fig. 2B. For age-dependent models, the convolution is done on $P_\theta(m)$. Note that closed-form expressions are known for the convolutions of Poisson and negative binomial distributions, e.g., $\mathrm{Pois}(\rho) * \mathrm{Pois}(\rho) = \mathrm{Pois}(2\rho)$ and $\mathrm{NB}(f,p) * \mathrm{NB}(f,p) = \mathrm{NB}(2f,p)$, which considerably simplifies the numerics.

Parameter inference

We use Julia [77] for the parameter inference and the subsequent analysis, utilising Distributions.jl [78] in the implementation and Makie.jl [79] for the visualisation of results, together with a number of numerical optimisation packages listed below.

Maximum likelihood estimation

Let $\boldsymbol{m} = (m_1, \ldots, m_n)$ denote a vector of spliced mRNA counts for one gene over n cells in a specific cell-cycle phase. The likelihood of observing the data $\boldsymbol{m}$ for a model with a count distribution $P(\cdot\,|\,\boldsymbol{\xi})$ given a set of parameters $\boldsymbol{\xi}$ is defined by

$$L(\boldsymbol{\xi}\,|\,\boldsymbol{m}) = \prod_{i=1}^{n} P(m_i\,|\,\boldsymbol{\xi}). \tag{16}$$

In practice, to find the maximum likelihood estimate (MLE) of the model parameters, $\boldsymbol{\xi}^*$, we minimise the negative log-likelihood:

$$\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}} \sum_{i=1}^{n} -\log P(m_i\,|\,\boldsymbol{\xi}). \tag{17}$$

Note that for the age-dependent models we have to maximise the likelihood by explicitly taking into account the age θ of each cell. For a given cell-cycle phase with the associated θ bins in the range from θ_i to θ_f the optimal parameters will be given by:

$$\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}} \sum_{\theta=\theta_i}^{\theta_f} \sum_{i=1}^{n_\theta} -\log P(m_{i\theta}\,|\,\boldsymbol{\xi}), \tag{18}$$

where n_θ is the number of cells in a given θ bin and $m_{i\theta}$ is the mRNA count in cell i with the same θ value.

We fit the gene expression models to the scRNA-seq data for all 3882 genes using maximum likelihood, considering both the age-independent and the age-dependent models. Note that each model here refers to the total count

distribution given by a k-fold convolution of the allele-specific count distribution with itself, where k is the number of gene copies in a given cell-cycle phase—throughout the paper we report the burst parameters per one gene copy. We compute the count distributions explicitly (as given in the section on Mechanistic models of stochastic gene expression) to obtain the likelihood of each model. Although some studies have highlighted the computational challenges of evaluating the steady-state solution of the telegraph model involving a confluent hypergeometric function [42], we found our direct implementation using log-probabilities to be sufficiently numerically stable and more efficient than the corresponding Beta-Poisson formulation using Gaussian quadratures [43, 80].

To minimise the negative log-likelihood, we initially perform a global optimisation using the adaptive differential evolution optimiser with radius limited sampling, as implemented in BlackBoxOptim.jl [81]. To ensure convergence, we then use the candidate solution as the initial value for a local search algorithm—here we employ the BFGS optimiser implemented in Optim.jl [82]. The optimisation is performed in log- or logit-transformed parameter space with model-specific parameter bounds. Note that for the Poisson model the solution is simply given by the sample mean. For the age-dependent Poisson, we set $\rho \in [-10, 10]$ and $\beta \in [-20, 5]$ on the log-scale. The search space for the negative binomial model, $\mathrm{NB}(f, p = (1 + b)^{-1})$, is constrained to $f \in [-10, 7]$ on the log-scale and $p \in [-20, 20]$ on the logit-scale (as p is a probability from zero to one). For the age-dependent negative binomial given by Eq. (12), we set $f, b' \in [-10, 7]$ and $\beta \in [-20, 5]$ (all on the log-scale). For the telegraph model, we fix $\sigma_0, \sigma_1, \rho \in [-10, 7]$ on the log-scale. The parameter constraints for the age-dependent telegraph model are identical, and the scaling parameter as before is fixed to $\beta \in [-20, 5]$ on the log-scale. The zero-inflated counterparts of the mentioned models have exactly the same parameter constraints, and we set the zero-inflation probability to $p_0 \in [-20, 20]$ on the logit-scale. The global and local optimisation steps are both terminated after 60 seconds, although they generally converge much faster for the majority of genes. Although the optimiser may get stuck in a bad local minimum and fail to converge in the designated time given highly bursty genes for the telegraph-like models due to parameter unidentifiability, we found that simply restarting the inference procedure for the problematic genes typically resolved the issue, albeit in such cases a simpler model is preferable.

Model selection

We determine the optimal model for each gene using the Bayesian information criterion (BIC):

$$\mathrm{BIC} = -2\log(L(\boldsymbol{\xi}^* \mid \boldsymbol{m})) + k\log(n), \tag{19}$$

where k is the number of model parameters, n is the number of data points (cells) and $L(\boldsymbol{\xi}^* \mid \boldsymbol{m})$ denotes the maximised likelihood function for the data vector $\boldsymbol{m}$. Given two models M_1 and M_2, the difference between their respective BIC values BIC_1 and BIC_2, $\Delta = \mathrm{BIC}_1 - \mathrm{BIC}_2$, reflects the strength of evidence in favour of M_2 over M_1, with $\Delta > 10$ constituting "very strong" evidence for model M_2 over M_1 [83, 84]. We utilise this criterion to iteratively compare six different models (from the simplest to the most complex): Poisson, zero-inflated Poisson (ZIP), negative binomial, zero-inflated negative binomial, telegraph and zero-inflated telegraph (ZI-telegraph) models.

We start by computing the difference in the BIC values between the Poisson model and all the other candidates, i.e. $\Delta_1 = \mathrm{BIC}_{\mathrm{Poisson}} - \mathrm{BIC}_{\mathrm{ZIP}}, \ldots, \Delta_5 = \mathrm{BIC}_{\mathrm{Poisson}} - \mathrm{BIC}_{\mathrm{ZI\text{-}Telegraph}}$. If all five Δ values are smaller than 10, we pick the Poisson model as the optimal choice; otherwise, we proceed to compute the difference in the BIC values between the zero-inflated Poisson and the remaining (more complex) models, repeating the procedure until the best candidate is found. This allows us to filter out the models with practically unidentifiable parameters and select the simplest gene expression model that still provides a good fit to the mRNA count data.

The model selection is performed for all 3882 genes in the dataset in both G1 and G2/M phases (done separately for the age-independent and age-dependent models). As the burst frequency and size parameters cannot be extracted from the Poisson and zero-inflated Poisson models, all genes that either in G1 or in G2/M are best fit by the Poisson or zero-inflated Poisson models are excluded from further analysis. After this filtering step applied to both age-independent and age-dependent models, we are left with 1669 genes that in both cell-cycle phases are best fit by either the negative binomial, zero-inflated negative binomial or telegraph models. Namely, for the age-independent case, we find that the overwhelming majority of the remaining genes are best fit by the negative binomial model; the zero-inflated negative binomial is the optimal model for 22 genes in G1 and 15 genes in G2/M; the telegraph model fits only 1 gene in G1 and 2 genes in G2/M; and the zero-inflated telegraph model was not found to be optimal for any genes. We note that the distribution of the best fit models is very similar in the age-dependent case.

Confidence intervals

To quantify the uncertainty in the best fit model estimates, we compute the 95% confidence intervals for each model parameter using profile likelihood [85, 86]. If we partition a vector of model parameters $\boldsymbol{\xi}$ into a parameter

of interest ϕ and a vector of nuisance parameters $\boldsymbol{\psi}$, i.e. $\boldsymbol{\xi} = (\phi, \boldsymbol{\psi})$, the profile likelihood of the parameter ϕ given a set of observations $\boldsymbol{m}$ can be defined as:

$$\mathrm{PL}(\phi \,|\, \boldsymbol{m}) = \max_{\boldsymbol{\psi}} L(\phi, \boldsymbol{\psi} \,|\, \boldsymbol{m}), \tag{20}$$

where we maximise the likelihood keeping ϕ fixed. It can be shown that the approximate $100(1-\alpha)\%$ confidence interval for ϕ is given by [85]:

$$\left\{ \phi : \log\left(\mathrm{PL}(\phi \,|\, \boldsymbol{m})\right) - \log(L(\boldsymbol{\xi}^* \,|\, \boldsymbol{m})) \geq -\frac{1}{2}\chi^2_{1,(1-\alpha)} \right\}, \tag{21}$$

where $L(\boldsymbol{\xi}^* \,|\, \boldsymbol{m})$ is the maximised likelihood (with all parameters allowed to vary) and $\chi^2_{1,(1-\alpha)}$ is the $(1-\alpha)$ quantile of the χ^2 distribution with one degree of freedom.

We can generalise the approach above to compute the confidence intervals for some function of the model parameters, $g(\boldsymbol{\xi})$, by defining the prediction profile likelihood [86, 87]:

$$\mathrm{PPL}_g(z \,|\, \boldsymbol{m}) = \max_{\boldsymbol{\xi} \,|\, g(\boldsymbol{\xi}) = z} L(\boldsymbol{\xi} \,|\, \boldsymbol{m}), \tag{22}$$

where we maximise the likelihood under the constraint that $g(\boldsymbol{\xi})$ equals the value z. Similarly to the confidence interval for a single model parameter, the $100(1-\alpha)\%$ confidence interval for the set of predictions $g(\boldsymbol{\xi}) = z$ is given by:

$$\left\{ z : \log\left(\mathrm{PPL}_g(z \,|\, \boldsymbol{m})\right) - \log(L(\boldsymbol{\xi}^* \,|\, \boldsymbol{m})) \geq -\frac{1}{2}\chi^2_{1,(1-\alpha)} \right\}. \tag{23}$$

We can also find the confidence intervals for functions of the parameters of multiple different models by considering their joint likelihood. For example, consider the ratio of the burst frequency between the G2/M and G1 phases, $Q_f = f_{\mathrm{G2/M}}/f_{\mathrm{G1}}$ (examined in Fig. 2C). Let $\boldsymbol{m}_{\mathrm{G1}}$ and $\boldsymbol{m}_{\mathrm{G2/M}}$ be the vectors of mRNA counts in G1 and G2/M respectively, and similarly define $\boldsymbol{\xi}_{\mathrm{G1}}$ and $\boldsymbol{\xi}_{\mathrm{G2/M}}$ as the phase-specific model parameters. The joint likelihood is then given by

$$L(\boldsymbol{\xi}_{\mathrm{G1}}, \boldsymbol{\xi}_{\mathrm{G2/M}} \,|\, \boldsymbol{m}_{\mathrm{G1}}, \boldsymbol{m}_{\mathrm{G2/M}}) = L(\boldsymbol{\xi}_{\mathrm{G1}} \,|\, \boldsymbol{m}_{\mathrm{G1}}) L(\boldsymbol{\xi}_{\mathrm{G2/M}} \,|\, \boldsymbol{m}_{\mathrm{G2/M}}), \tag{24}$$

and the prediction profile likelihood can be written as

$$\mathrm{PPL}(Q_f \,|\, \boldsymbol{m}_{\mathrm{G1}}, \boldsymbol{m}_{\mathrm{G2/M}}) = \max_{\boldsymbol{\xi}_{\mathrm{G1}}, \boldsymbol{\xi}_{\mathrm{G2/M}} \,\big|\, \frac{f_{\mathrm{G2/M}}}{f_{\mathrm{G1}}} = Q_f} L(\boldsymbol{\xi}_{\mathrm{G1}}, \boldsymbol{\xi}_{\mathrm{G2/M}} \,|\, \boldsymbol{m}_{\mathrm{G1}}, \boldsymbol{m}_{\mathrm{G2/M}}), \tag{25}$$

so that the confidence interval for Q_f is similar in form to Eq. (23).

The outlined approach is used throughout the paper to compute the confidence intervals for any considered quantity z. We implement this numerically by simply increasing (decreasing) z in small steps starting from its maximum likelihood value, and recomputing the prediction profile likelihood at each step until the threshold defining its upper (lower) confidence interval limit is reached. We perform the constrained optimisation using the Ipopt algorithm [88] through the MetaOptInterface.jl [89] and Optimization.jl packages. The parameter bounds for each model are the same as previously used for maximum likelihood estimation.

Data Availability Statement

The original scRNA-seq dataset can be obtained from Ref. [57]. The processed data has been deposited at `https://doi.org/10.5281/zenodo.10467234` and the code for this paper is available at `https://github.com/augustinas1/CellCycle-RNAseq`.

Acknowledgments

This work was supported by the Alan Turing Institute Doctoral Studentship (EPSRC grant EP/N510129/1) for A. S. and a Leverhulme Trust grant (Grant No. RPG-2020-327) for R. G. The authors would like to thank Kaan Öcal and Abhyudai Singh for useful discussions.

References

[1] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA synthesis in mammalian cells," *PLoS biology*, 4(10): e309, 2006.

[2] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, "Mammalian genes are transcribed with widely different bursting kinetics," *Science*, 332(6028): 472–474, 2011.

[3] B. T. Donovan, A. Huynh, D. A. Ball, H. P. Patel, M. G. Poirier, D. R. Larson, M. L. Ferguson, and T. L. Lenstra, "Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting," *The EMBO journal*, 38(12): e100809, 2019.

[4] J. Rodriguez and D. R. Larson, "Transcription in living cells: Molecular mechanisms of bursting," *Annual review of biochemistry*, 89: 189–212, 2020.

[5] E. Tunnacliffe and J. R. Chubb, "What is a transcriptional burst?" *Trends in Genetics*, 36(4): 288–297, 2020.

[6] A. Sanchez and I. Golding, "Genetic determinants and cellular constraints in noisy gene expression," *Science*, 342(6163): 1188–1193, 2013.

[7] I. Brouwer, H. P. Patel, J. V. W. Meeussen, W. Pomp, and T. L. Lenstra, "Single-molecule fluorescence imaging in living saccharomyces cerevisiae cells," *STAR protocols*, 1(3): 100142, 2020.

[8] T. L. Lenstra and D. R. Larson, "Single-molecule mRNA detection in live yeast," *Current protocols in molecular biology*, 113(1): 14–24, 2016.

[9] J. Peccoud and B. Ycart, "Markovian modeling of gene-product synthesis," *Theoretical population biology*, 48(2): 222–234, 1995.

[10] J. Paulsson, O. G. Berg, and M. Ehrenberg, "Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation," *Proceedings of the National Academy of Sciences*, 97(13): 7148–7153, 2000.

[11] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast," *Nature structural & molecular biology*, 15(12): 1263–1271, 2008.

[12] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq, "Transcription factors modulate c-Fos transcriptional bursts," *Cell reports*, 8(1): 75–83, 2014.

[13] H. Ochiai, T. Sugawara, T. Sakuma, and T. Yamamoto, "Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells," *Scientific reports*, 4(1): 7125, 2014.

[14] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, "Real-time kinetics of gene activity in individual bacteria," *Cell*, 123(6): 1025–1036, 2005.

[15] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz, "Gene regulation at the single-cell level," *Science*, 307(5717): 1962–1965, 2005.

[16] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annual Review of Physical Chemistry*, 58: 35–55, 2007.

[17] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, 297(5584): 1183–1186, 2002.

[18] J. M. Raser and E. K. O'Shea, "Control of stochasticity in eukaryotic gene expression," *Science*, 304(5678): 1811–1814, 2004.

[19] R. Grima and P.-M. Esmenjaud, "Quantifying and correcting bias in transcriptional parameter inference from single-cell data," *Biophysical Journal*, 123: 4–30, 2023.

[20] X. Fu, H. P. Patel, S. Coppola, L. Xu, Z. Cao, T. L. Lenstra, and R. Grima, "Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions," *Elife*, 11: e82493, 2022.

[21] O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj, "Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms," *Molecular Cell*, 58(2): 339–352, 2015.

[22] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, "Single-cell analysis of transcription kinetics across the cell cycle," *Elife*, 5: e12175, 2016.

[23] X.-M. Sun, A. Bowman, M. Priestman, F. Bertaux, A. Martinez-Segura, W. Tang, C. Whilding, D. Dormann, V. Shahrezaei, and S. Marguerat, "Size-dependent increase in RNA polymerase II initiation rates mediates gene expression scaling with cell size," *Current Biology*, 30(7): 1217–1230, 2020.

[24] S. Berry and L. Pelkmans, "Mechanisms of cellular mRNA transcript homeostasis," *Trends in Cell Biology* : 2022.

[25] I. G. Johnston, B. Gaal, R. P. d. Neves, T. Enver, F. J. Iborra, and N. S. Jones, "Mitochondrial variability as a source of extrinsic cellular noise," *PLoS computational biology*, 8(3): e1002416, 2012.

[26] R. Foreman and R. Wollman, "Mammalian gene expression variability is explained by underlying cell state," *Molecular systems biology*, 16(2): e9146, 2020.

[27] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature methods*, 6(5): 377–382, 2009.

[28] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade," *Nature protocols*, 13(4): 599–604, 2018.

[29] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, 8(1): 14049, 2017.

[30] Z. Yao, H. Liu, F. Xie, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. Van den Berge, *et al.*, "A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex," *Nature*, 598(7879): 103–110, 2021.

[31] S. Chen, P. Rivaud, J. H. Park, T. Tsou, E. Charles, J. R. Haliburton, F. Pichiorri, and M. Thomson, "Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign," *Proceedings of the National Academy of Sciences*, 117(46): 28784–28794, 2020.

[32] S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, *et al.*, "SCENIC: single-cell regulatory network inference and clustering," *Nature methods*, 14(11): 1083–1086, 2017.

[33] T. Chari, G. Gorin, and L. Pachter, "Biophysically Interpretable Inference of Cell Types from Multimodal Sequencing Data," *bioRxiv* : 2023–09, 2023.

[34] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, 343(6167): 193–196, 2014.

[35] G. La Manno *et al.*, "RNA velocity of single cells," *Nature*, 560: 494–498, 2018.

[36] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing RNA velocity to transient cell states through dynamical modeling," *Nature biotechnology*, 38(12): 1408–1414, 2020.

[37] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, *et al.*, "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types," *Science*, 343(6172): 776–779, 2014.

[38] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, *et al.*, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, 344(6190): 1396–1401, 2014.

[39] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, 32(4): 381–386, 2014.

[40] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq," *Nature*, 509(7500): 371–375, 2014.

[41] Z. Xue, K. Huang, C. Cai, L. Cai, C.-y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, *et al.*, "Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing," *Nature*, 500(7464): 593–597, 2013.

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

[43] A. J. M. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, "Genomic encoding of transcriptional burst kinetics," *Nature*, 565: 251–254, 2019.

[44] W. Tang, A. C. S. Jørgensen, S. Marguerat, P. Thomas, and V. Shahrezaei, "Modelling capture efficiency of single cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics," *Bioinformatics*, 39(7): btad395, 2023.

[45] K. Öcal, "Incorporating extrinsic noise into mechanistic modelling of single-cell transcriptomics," *bioRxiv* : 2023–09, 2023.

[46] S. Luo, Z. Wang, Z. Zhang, T. Zhou, and J. Zhang, "Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics," *Nucleic Acids Research*, 51(1): 68–83, 2023.

[47] G. Gorin, J. J. Vastola, M. Fang, and L. Pachter, "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments," *Nature Communications*, 13(1): 7620, 2022.

[48] Y. Jiang, N. R. Zhang, and M. Li, "SCALE: modeling allele-specific gene expression by single-cell RNA sequencing," *Genome biology*, 18(1): 1–15, 2017.

[49] J. K. Kim, A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni, "Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression," *Nature communications*, 6(1): 8687, 2015.

[50] R. Jiang, T. Sun, D. Song, and J. J. Li, "Statistics or biology: the zero-inflation controversy about scRNA-seq data," *Genome Biology*, 23(1): 1–24, 2022.

[51] V. Svensson, "Droplet scrna-seq is not zero-inflated," *Nature Biotechnology*, 38(2): 147–150, 2020.

[52] Y. Cao, S. Kitanovski, R. Küppers, and D. Hoffmann, "UMI or not UMI, that is the question for scRNA-seq zero-inflation," *Nature Biotechnology*, 39: 158–159, 2021.

[53] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature methods*, 9(1): 72–74, 2012.

[54] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of single-cell RNA sequencing methods," *Molecular cell*, 65(4): 631–643, 2017.

[55] D. Grün, L. Kester, and A. Van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature methods*, 11(6): 637–640, 2014.

[56] D. Volteras, V. Shahrezaei, and P. Thomas, "Global transcription regulation revealed from dynamical correlations in time-resolved single-cell RNA-sequencing," *bioRxiv* : 2023–10, 2023.

[57] A. Riba, A. Oravecz, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes, and N. Molina, "Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning," *Nature Communications*, 13(2865): 1–13, 2022.

[58] R. Perez-Carrasco, C. Beentjes, and R. Grima, "Effects of cell cycle variability on lineage and population measurements of messenger rna abundance," *Journal of the Royal Society Interface*, 17(168): 20200360, 2020.

[59] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, and R. A. Kievit, "Raincloud plots: a multi-platform tool for robust data visualization," *Wellcome Open Research*, 4:63.: 2021. eprint: `31069261`.

[60] S. Berry, M. Müller, A. Rai, and L. Pelkmans, "Feedback from nuclear RNA on transcription promotes robust RNA concentration homeostasis in human cells," *Cell Systems*, 13(6): 454–470, 2022.

[61] D. E. Weidemann, J. Holehouse, A. Singh, R. Grima, and S. Hauf, "The minimal intrinsic stochasticity of constitutively expressed eukaryotic genes is sub-Poissonian," *Science Advances*, 9(32): eadh5138, 2023.

[62] C. Cadart, S. Monnier, J. Grilli, P. J. Sáez, N. Srivastava, R. Attia, E. Terriac, B. Baum, M. Cosentino-Lagomarsino, and M. Piel, "Size control in mammalian cells involves modulation of both growth rate and cell cycle duration," *Nature communications*, 9(1): 3275, 2018.

[63] J. R. Peterson, J. A. Cole, J. Fei, T. Ha, and Z. A. Luthey-Schulten, "Effects of dna replication on mrna noise," *Proceedings of the National Academy of Sciences*, 112(52): 15886–15891, 2015.

[64] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, 161(5): 1187–1201, 2015.

[65] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, and V. Shahrezaei, "BayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data," *Bioinformatics*, 36(4): 1174–1181, 2020.

[66] C. Jia, A. Singh, and R. Grima, "Concentration fluctuations in growing and dividing cells: Insights into the emergence of concentration homeostasis," *PLoS Computational Biology*, 18(10): e1010574, 2022.

[67] C. Jia and R. Grima, "Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model," *iScience*, 26(1): 105746, 2023.

[68] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, 473(7347): 337–342, 2011.

[69] G. Gorin, J. J. Vastola, and L. Pachter, "Studying stochastic systems biology of the cell with single-cell genomics data," *Cell Systems*, 14(10): 822–843, 2023.

[70] V. Bergen, R. A. Soldatov, P. V. Kharchenko, and F. J. Theis, "RNA velocity—current challenges and future perspectives," *Molecular Systems Biology*, 17(8): e10282, 2021.

[71] G. Gorin, M. Fang, T. Chari, and L. Pachter, "RNA velocity unraveled," *PLOS Computational Biology*, 18(9): e1010492, 2022.

[72] A. R. Stinchcombe, C. S. Peskin, and D. Tranchina, "Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression," *Physical Review E*, 85(6): 061919, 2012.

[73] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments," *Biostatistics*, 19(4): 562–578, 2018.

[74] K. Choi, Y. Chen, D. A. Skelly, and G. A. Churchill, "Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics," *Genome Biology*, 21(1): 1–16, 2020.

[75] W. Chen, Y. Li, J. Easton, D. Finkelstein, G. Wu, and X. Chen, "UMI-count modeling and differential expression analysis for single-cell RNA sequencing," *Genome Biology*, 19(1): 1–17, 2018.

[76] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model," *Genome Biology*, 20(1): 1–16, 2019.

[77] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review* : 2017.

[78] M. Besançon, T. Papamarkou, D. Anthoff, A. Arslan, S. Byrne, D. Lin, and J. Pearson, "Distributions.jl: Definition and Modeling of Probability Distributions in the JuliaStats Ecosystem," *Journal of Statistical Software*, 98(16): 1–30, 2021.

[79] S. Danisch and J. Krumbiegel, "Makie.jl: Flexible high-performance data visualization for Julia," *Journal of Open Source Software*, 6(65): 3349, 2021.

[80] T. N. Vu, Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, and Y. Pawitan, "Beta-Poisson model for single-cell RNA-seq data analyses," *Bioinformatics*, 32(14): 2128–2135, 2016.

[81] R. Feldt and A. Stukalov, *BlackBoxOptim.jl*, `https://github.com/robertfeldt/BlackBoxOptim.jl`, 2018.

[82] P. K. Mogensen and A. N. Riseth, "Optim: A mathematical optimization package for Julia," *Journal of Open Source Software*, 3(24): 615, 2018.

[83] A. E. Raftery, "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25: 111–163, 1995.

[84] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *WIREs Computational Statistics*, 4(2): 199–203, 2012.

[85] Y. Pawitan, *In All Likelihood*. Oxford, England, UK: Oxford University Press, 2013.

[86] C. Kreutz, A. Raue, D. Kaschek, and J. Timmer, "Profile likelihood in systems biology," *FEBS Journal*, 280(11): 2564–2571, 2013.

[87] C. Kreutz, A. Raue, and J. Timmer, "Likelihood based observability analysis and confidence intervals for predictions of dynamic models," *BMC Systems Biology*, 6(1): 1–9, 2012.

[88] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, 106(1): 25–57, 2006.

[89] B. Legat, O. Dowson, J. Dias Garcia, and M. Lubin, "MathOptInterface: A data structure for mathematical optimization problems," *INFORMS Journal on Computing*, 34(2): 672–689, 2021.

Appendix

A Correcting the ratio of burst parameters in G1 and G2/M for technical noise

Consider transcriptional bursting from each allele copy. This implies that the expression for each allele copy can be described by the following reaction scheme:

$$G \xrightarrow{\mathrm{f}} G + rM, \quad M \xrightarrow{1} \emptyset, \tag{26}$$

where r is an integer sampled from the geometric distribution with mean b (the mean burst size) and f is the burst frequency. Using the chemical master equation for this reaction scheme, the steady-state distribution of mRNA counts for an allele is found to be negative binomial with first and second factorial moments of the mRNA counts equal to bf and $b^2 f(1+f)$, respectively (here and in what follows in the Appendix, there is the implicit assumption that the number of cells is infinite).

Now if we assume that the burst frequency of each copy does not vary from cell to cell but that the mean burst size does, i.e. let the probability distribution of b be $P(b)$, then the population-averaged factorial moments of mRNA counts (from each allele) are:

$$\begin{aligned} \langle m \rangle &= \int_b fbP(b)db = f\langle b \rangle, \\ \langle m(m-1) \rangle &= \int_b b^2 f(1+f)P(b)db = \langle b^2 \rangle f(1+f). \end{aligned} \tag{27}$$

It then follows that the Fano factor FF of mRNA counts (defined as the variance of mRNA counts divided by the mean mRNA count) can be written as:

$$\mathrm{FF} = 1 + \langle b \rangle + \langle b \rangle \mathrm{CV}_b^2 + \mathrm{CV}_b^2 \langle m \rangle, \tag{28}$$

where CV_b^2 is the coefficient of variation squared for the burst size. Note that since in age-dependent models the burst size for a gene is proportional to λ^β, it follows that

$$\mathrm{CV}_b^2 = \mathrm{CV}_{\lambda^\beta}^2. \tag{29}$$

If we have N independent allele copies, the mean and the Fano Factor of the total mRNA counts for these copies is equal to:

$$\langle m \rangle_{tot} = N\langle m \rangle, \tag{30}$$

$$\mathrm{FF}_{tot} = \mathrm{FF}. \tag{31}$$

In Fig. 4C the maximisation of the negative binomial likelihood (this is the optimal model most commonly selected for the bursty genes by the model selection algorithm), given the mRNA counts from cells with N allele copies and the age for each cell, is used to find the burst size, burst frequency and the power-law scaling exponent for each gene in G1 ($N = 2$) or G2/M ($N = 4$). Since the parameters of a negative binomial distribution are completely determined by its mean and variance, it follows that in the limit of a large number of cells, the maximisation of the likelihood is equivalent to finding the values of $f, \langle b \rangle$ and β such that Eqs. (30) and (31) are satisfied.

This parameter estimation procedure can in principle be performed with two types of count data: the observed count data x and the true data y, i.e. the actual number of counts prior to the technical noise introduced by single-cell sequencing. We denote the corresponding parameter estimates as $f_x, \langle b \rangle_x, \beta_x$ and $f_y, \langle b \rangle_y, \beta_y$, respectively.

Next, we establish the relationship between these two sets of parameters. Given the results above, we can write:

$$\langle x \rangle = N f_x \langle b \rangle_x, \tag{32}$$

$$\mathrm{FF}_x = 1 + \langle b \rangle_x + \langle b \rangle_x \mathrm{CV}_{\lambda^\beta}^2 + \mathrm{CV}_{\lambda^\beta}^2 \langle x \rangle, \tag{33}$$

$$\langle y \rangle = N f_y \langle b \rangle_y, \tag{34}$$

$$\mathrm{FF}_y = 1 + \langle b \rangle_y + \langle b \rangle_y \mathrm{CV}_{\lambda^\beta}^2 + \mathrm{CV}_{\lambda^\beta}^2 \langle y \rangle. \tag{35}$$

Note that for clarity we have dropped the *tot* subscript. Note also that here we have assumed that the estimated noise in the cell age distribution is not affected by technical noise, i.e. $\mathrm{CV}_{\lambda^\beta}^2$ is the same if estimated using x or y data. This is a reasonable assumption because (i) the extrinsic noise parameter λ is a function of θ, which itself is

computed by DeepCycle from the cyclic patterns in the unspliced-spliced RNA space and the latter are qualitatively unaffected by technical noise; (ii) $\beta_x \approx \beta_y$ because x is a downsampled version of y and hence a least squares fit of a power law to a set of data points $(\lambda, \langle y \rangle)$ data would lead to the same exponent as the procedure applied to a set of data points $(\lambda, \langle x \rangle)$.

Next, we assume that the technical noise introduced by the sequencing method can be well described by a binomial capture model [64]. In other words, we assume that in each cell i, there is a finite probability p_i of scRNA-seq detecting each mRNA molecule in that cell. The variation of this detection probability from cell to cell is described by some distribution $H(p)$. This allows us to connect the statistics of the measured counts (x) with those of the true counts (y). Specifically, it can then be proved (Appendix B) that:

$$\langle x \rangle = \langle y \rangle \langle p \rangle, \tag{36}$$

$$\mathrm{FF}_x = 1 + \langle p \rangle (\mathrm{FF}_y - 1) + \mathrm{FF}_p [\mathrm{FF}_y - (1 - \langle y \rangle)], \tag{37}$$

where $\langle p \rangle$ and FF_p are the mean and Fano factor of the detection probability computed using $H(p)$.

Substituting Eqs. (36)-(37) in Eqs. (32)-(35), after some algebra we find

$$f_x = \frac{f_y \langle p \rangle}{\mathrm{FF}_p (1 + f_y N) + \langle p \rangle}, \tag{38}$$

$$b_x = b_y [\mathrm{FF}_p (1 + f_y N) + \langle p \rangle]. \tag{39}$$

This implies that the burst frequency estimated using the observed count data f_x underestimates the true burst frequency f_y according to the degree of cell-to-cell variation in the detection probability (as determined by FF_p), i.e. if all cells have the same detection probability then there is no underestimation of the burst frequency. In contrast, the burst size is generally overestimated; this occurs even when there is no cell-to-cell variation in the detection probability, i.e. $\mathrm{FF}_p = 0$.

Finally, we can see how the ratio of the burst frequency and size in G1 and G2/M change when technical noise is taken into account. Since in the G1 phase $N = 2$ and in the G2/M phase $N = 4$, using Eqs. (38) and (39) we find:

$$\frac{f_{x,\mathrm{G2/M}}}{f_{x,\mathrm{G1}}} = \frac{f_{y,\mathrm{G2/M}}}{f_{y,\mathrm{G1}}} \eta, \tag{40}$$

$$\frac{b_{x,\mathrm{G2/M}}}{b_{x,\mathrm{G1}}} = \frac{b_{y,\mathrm{G2/M}}}{b_{y,\mathrm{G1}}} \eta^{-1}, \tag{41}$$

where

$$\eta = \frac{1 + \mathrm{CV}_p^2 (1 + 2 f_{y,\mathrm{G1}})}{1 + \mathrm{CV}_p^2 (1 + 4 f_{y,\mathrm{G2/M}})}. \tag{42}$$

Note that in these calculations we have assumed that the distribution of the capture probability $H(p)$ does not vary with the cell-cycle phase, i.e. with the number of gene copies N. Evidence for this can be found in Ref. [56].

B Relationship between the Fano factors of observed and true counts

If the probability of detecting each mRNA molecule in a cell is p and the actual number of molecules is y, it follows that the number of measured molecules x is a random variable distributed according to the binomial distribution with the number of trials y, number of successes x and probability of success p. If each cell has the same detection probability p, the distribution of the observed counts $P(x)$ is given by:

$$P(x) = \sum_{y=0}^{\infty} P_{\mathrm{bin}}^{p}(y, x) Q(y), \tag{43}$$

where $Q(y)$ is the distribution of true counts and $P_{\mathrm{bin}}^{p}(y, x)$ is the binomial distribution:

$$P_{\mathrm{bin}}^{p}(y, x) = \binom{y}{x} p^x (1 - p)^{y - x}. \tag{44}$$

Generally each cell will have a different detection probability p, and hence, as in the previous Appendix, we define a distribution of detection probabilities $H(p)$. It then follows that Eq. (43) generalises to

$$P(x) = \int_{p=0}^{1} \sum_{y=0}^{\infty} P_{\text{bin}}^{p}(y,x) Q(y) H(p) dp. \tag{45}$$

By computing the moments we find:

$$\langle x \rangle = \sum_{x} x P(x) = \int_{p=0}^{1} \sum_{y} y p Q(y) H(p) dp = \langle y \rangle \langle p \rangle, \tag{46}$$

$$\langle x^2 \rangle = \sum_{x} x^2 P(x) = \int_{p=0}^{1} \sum_{y} y p (1 + (y-1)p) Q(y) H(p) dp = \langle y \rangle \langle p \rangle + \langle p^2 \rangle \langle y(y-1) \rangle. \tag{47}$$

Finally, using these results and some algebra, we can compute the Fano factor of the observed counts:

$$\text{FF}_x = \frac{\langle x^2 \rangle - \langle x \rangle^2}{\langle x \rangle} = 1 + \langle p \rangle (\text{FF}_y - 1) + \text{FF}_p [\text{FF}_y - (1 - \langle y \rangle)], \tag{48}$$

where $\text{FF}_y = (\langle y^2 \rangle - \langle y \rangle^2)/\langle y \rangle$ is the Fano factor of true counts and $\text{FF}_p = (\langle p^2 \rangle - \langle p \rangle^2)/\langle p \rangle$ is the Fano factor of the detection probability distribution. Note that $\text{FF}_x > 0$ since $0 < \langle p \rangle \leq 1$ and $\text{FF}_y \geq 1 - \langle y \rangle$ for any discrete probability distribution $Q(y)$.

C Estimation errors by models that assume an incorrect number of gene copies

Assume we have data from a large number of cells that are in the same cell-cycle phase, so that each cell has S independent gene copies. Furthermore, as in Appendix A, assume transcriptional bursting from each gene copy and that the burst size varies from cell to cell, e.g. due to a coupling of the transcription rate to cell age or cell size. If the burst frequency is f, the mean burst size is $\langle b \rangle$ and the coefficient of variation squared of the burst size in the population is CV_b^2, then it follows by Eqs. (27), (28), (30) and (31) that the mean mRNA count and the Fano factor of the mRNA distribution for the population of cells are given by:

$$\langle m \rangle_{tot} = S f \langle b \rangle, \tag{49}$$

$$\text{FF}_\text{tot} = 1 + \langle b \rangle + \langle b \rangle \text{CV}_b^2 + \text{CV}_b^2 f \langle b \rangle. \tag{50}$$

Next, assume that to this data we fit a negative binomial model considering R gene copies, so that the estimated burst frequency is f_{est}, the mean burst size is $\langle b \rangle_{est}$ and the coefficient of variation squared of the burst size in the population is $\text{CV}_{\text{b,est}}^2$. The mean mRNA count and the Fano factor predicted by the model are given by:

$$\langle m \rangle_{tot,est} = R f_{est} \langle b \rangle_{est}, \tag{51}$$

$$\text{FF}_{\text{tot,est}} = 1 + \langle b \rangle_{est} + \langle b \rangle_{est} \text{CV}_{b,est}^2 + \text{CV}_{b,est}^2 f_{est} \langle b \rangle_{est}. \tag{52}$$

Furthermore, for simplicity, assume that that the size of extrinsic noise can be determined, i.e. $\text{CV}_{\text{b,est}}^2 = \text{CV}_\text{b}^2$. For example, if the burst size is proportional to the cell volume, CV_b^2 is simply the coefficient of variation of the cell volume, and hence a measurement of the cell volumes would provide an estimate of extrinsic noise in the burst size. Alternatively, this can be estimated from the correlations in expression between different genes [19].

In any case, if the size of extrinsic noise is known, the only parameters left to be estimated are the burst frequency and the mean burst size. Since the negative binomial distribution is completely determined by its first two moments, parameter inference is exactly equivalent to equating Eq. (49) with Eq. (51) and Eq. (50) with Eq. (52) and solving for f_{est} and $\langle b \rangle_{est}$. By Taylor expanding the resulting equations for f_{est} and $\langle b \rangle_{est}$, assuming the size of extrinsic noise CV_b^2 is small, we finally obtain:

$$\frac{f_{est}}{f} \approx \frac{S}{R} + \frac{f S (S-R)}{R^2} \text{CV}_\text{b}^2, \tag{53}$$

$$\frac{\langle b \rangle_{est}}{\langle b \rangle} \approx 1 - f \text{CV}_\text{b}^2 \left(\frac{\text{S}}{\text{R}} - 1 \right). \tag{54}$$