Global transcription regulation revealed from dynamical correlations in time-resolved single-cell RNA-sequencing

Dimitris Volteras, Vahid Shahrezaei†, Philipp Thomas†
Department of Mathematics, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK
† Correspondence to: v.shahrezaei@imperial.ac.uk and p.thomas@imperial.ac.uk

October 24, 2023

Abstract

Single-cell transcriptomics reveals significant variations in the transcriptional activity across cells. Yet, it remains challenging to identify mechanisms of transcription dynamics from static snapshots. It is thus still unknown what drives global transcription dynamics in single cells. We present a stochastic model of gene expression in growing and dividing cells that harnesses temporal dimensions of single-cell RNA-sequencing through metabolic labelling protocols and cell cycle reporters. We develop a parallel and highly scalable Approximate Bayesian Computation method that corrects for technical variation and accurately quantifies absolute burst frequency, burst size and degradation rate at a transcriptome-wide scale. Using Bayesian model selection, we reveal scaling between transcription rates and cell size and unveil waves of gene regulation across the cell cycle-dependent transcriptome. Our study shows that stochastic modelling of dynamical correlations identifies global mechanisms of transcription regulation. We illustrate how mechanistic modelling and amortised inference can efficiently integrate distributions across time, modalities and conditions. Our framework thus provides a versatile tool to uncover how global transcriptional dynamics is orchestrated genome-wide in single cells.

1 Introduction

Transcription occurs in episodic bursts of activity, causing fluctuations in the level of transcripts produced by a single gene [1–3]. The frequency and size of these bursts lead to diverse temporal patterns of gene activity [4, 5], often referred to as cellular noise, that can have a functional role in gene regulatory networks and cell fate decisions [6–14]. The cell cycle has been shown to contribute significantly to transcript variability [15–18], either through random partitioning of molecules at cell division [19–21], modulation of transcription rates [22–26] or through coupling to cell size as a mechanism of transcript homeostasis [27–30]. Yet, it remains elusive what the global drivers of transcription dynamics are on a transcriptome-wide scale.

Single-cell RNA-sequencing (scRNA-seq) has become the de facto standard in profiling transcription patterns. The method allows quantifying distributions of transcript levels across cells that can be harnessed to learn about gene regulation. A few studies have started to investigate how transcriptional processes are regulated through synergising stochastic modelling and inference [31–35]. To this end, stochastic models of reaction kinetics inside living cells are simulated using the Gillespie algorithm or solved analytically and matched with experimental mRNA counts using statistical inference [24, 36–40]. However, a challenge is that not one model fits it all at a genome-wide scale, and identifying transcriptional patterns requires weighing competing model hypotheses against each other that need to be tested against experiments.

In principle, this aim can be achieved using Bayesian inference by computing the posterior distribution over models and their corresponding parameters. The posterior model probability weighs the model likelihood by its prior distribution. Several problems with this approach appear in practice: (i) the likelihood of mechanistic models is intractable or is computationally expensive to optimise [37], (ii) scRNA-seq data is heavily corrupted by technical noise including dropout [41], (iii) absolute kinetic rates are experimentally inaccessible from static snapshots [42].

In the literature, these problems are typically dealt with separately. The first issue is addressed by replacing likelihoods with moment- or simulation-based methods [33, 35, 37, 43]. Yet, these methods remain computationally challenging as they need to be performed for several competing models and many genes. The second issue has been addressed by rescaling transcript counts [41, 44], but Bayesian normalisation has been shown to capture distributions more truthfully [45, 46]. The third issue remains a major obstacle for single-cell analysis, as traditional sequencing protocols can only capture a static snapshot of cells in time, thus prohibiting absolute quantification of transcription and mRNA turnover rates.

Recently, metabolic RNA labelling has become available to temporally resolve conventional single-cell transcriptomics [47]. These protocols allow monitoring of mRNA synthesis and degradation in single cells, thus capturing temporal transcription dynamics [25, 26, 48–53]. This provides new possibilities for modelling as it enables absolute rate quantification, as demonstrated recently for the inference of transcriptional burst kinetics through maximum likelihood methods [54] or MCMC sampling from the posterior distribution [55]. These approaches have either neglected technical noise of sequencing protocols or have been limited to relatively simple models and have not explored the landscape of the model posterior. It thus remains elusive which models describe gene expression on a transcriptome-wide scale.

Here, we develop a scalable approach to Bayesian inference, normalisation and model selection of bursty gene expression dynamics along the cell cycle. A stochastic model enables absolute quantification of transcription dynamics along the cell cycle, resolving mRNA synthesis, bursting and

degradation processes from scRNA-seq data with metabolic labelling and cell cycle reporters [25]. We demonstrate that correlations between unlabelled and labelled mRNA levels can delineate different modes of transcriptional regulation and quantify the contributions from the cell cycle and mRNA synthesis, bursting, and degradation kinetics to the overall cell-to-cell variation. The framework allows us to discover which of these processes are modulated by the cell cycle and identify genes with distinct transcription regulation. Our study thus reveals, for the first time, how genome-wide single-cell transcriptional dynamics is orchestrated with global factors such as cell size and cell cycle.

2 Results

Stochastic modelling captures distinct patterns of bursty transcription dynamics along the cell cycle

We develop a stochastic model to describe the dependence of bursty transcription dynamics on the cell cycle as well as the effects of RNA labelling, according to the scEU-seq protocol (Figure 1 (A)). We assume reaction kinetics based on a two-stage telegraph model of transcription, capturing random gene promoter switching between active and inactive states, followed by synthesis and degradation of labelled and unlabelled mRNA (Supplementary Methods 1.2.1). The rates driving these reactions are assumed to be cell cycle-dependent, and we additionally assume a linear scaling of mRNA synthesis rate with cell size (Figure 1 (B), Methods 1.2.2) [27, 56, 57]. We formulate moment equations that describe the temporal dependence of mean and covariance of labelled and unlabelled mRNA on cell cycle progression and cell division dynamics as well as on pulse-chase labelling conditions. These equations track the mRNA statistics over time along a single-cell lineage, following consecutive cell divisions until convergence to a steady state, capturing the random partitioning of molecules at cell division as well as the effect of pulse-chase labelling (Figure 1 (C), Supplementary Methods 1.2.3, 1.2.4, 1.2.5). The cell cycle-dependent changes of mRNA statistics and kinetic rates are described in five distinct cell cycle phases (Supplementary Figure S1, Supplementary Methods 1.1.3).

Our modelling framework makes it possible to test multiple hypotheses on the cell cycle-dependence of kinetic rate parameters and identify genes characterised by distinct underlying kinetics. To this end, we develop an Approximate Bayesian Computation framework [58] to infer the mode of regulation that best describes each gene (Figure 1 (C)), as we explain in more detail in the next section. We define five distinct model classes - hypotheses, each corresponding to a specific cell cycle-dependent regulation of kinetic rates (Figure 2 (A)). Two of the model classes, which we refer to as *constant* models, assume that the synthesis rate either scales linearly with cell size or stays constant, while the rest of the kinetic rates are assumed to be constant along the cell cycle. The other three model classes, which we refer to as *non-constant* models, assume scaling of the synthesis rate with cell size and additionally that one of the kinetic rates, either burst frequency or synthesis rate or decay rate, varies with cell cycle progression. We refer to the three non-constant models as the *burst frequency* model, the *burst size* model and the *decay rate* model, respectively. The two constant models identify genes for which the kinetics are conserved through the cell cycle and distinguish between size-scaling and size-independent expression. The three non-constant models identify genes for which there is a kinetic rate being dominantly modulated through the cell cycle.

We find that these five models yield diverse summary statistics (defined in Tables 1 (A) and 2) and can identify distinct dynamic features of transcription. Example model realisations (Supplementary

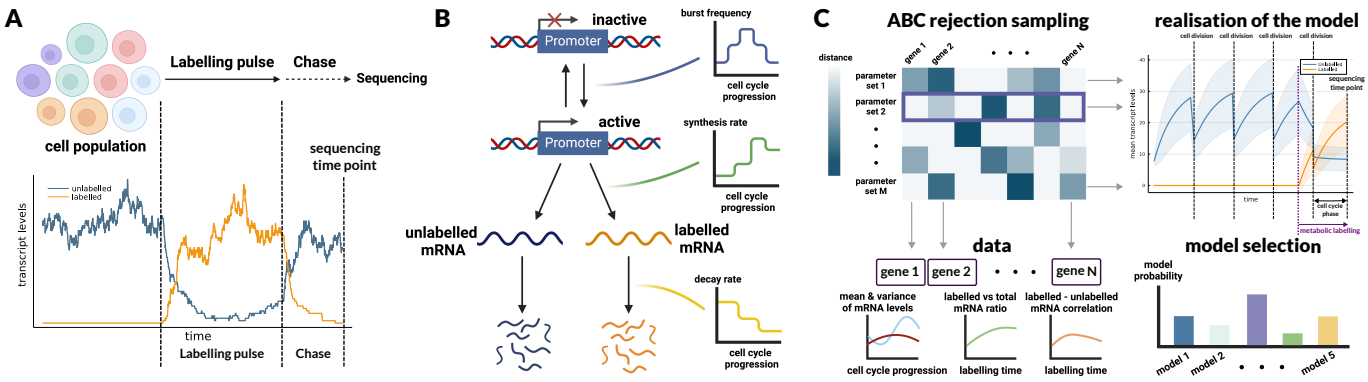


Figure 1: **Overview of modelling and inference framework.** (A) Illustration of single-cell EU RNA-sequencing (scEU-seq) protocol, developed by [25]. A heterogeneous cell population is pulse-chase labelled prior to sequencing. An example of stochastic transcription of a gene at the single-cell level is shown below, where labelled mRNA is produced and degraded following the timing of pulse-chase labelling. The pulse and chase experiments with multiple durations of the scEU-seq protocol are informative of the timescales of mRNA turnover. (B) Illustration of the reaction kinetics model describing bursty transcription of unlabelled and labelled mRNA with cell cycle-dependent kinetic rates. The promoter activation rate (burst frequency), synthesis rate and decay rate are cell cycle-dependent. (C) Overview of the inference scheme: Each parameter set yields a model realisation which is tested against summary statistics of all genes in the data through ABC rejection sampling. Model selection is used to determine the best-fitting model(s) for each gene.

Figure S2) suggest that noise statistics have a crucial role in model identifiability. For instance, even though the varying burst frequency and burst size models can give rise to overall similar summary statistics, the cell cycle-dependent noise statistics can be used to distinguish between the two models. We also find that the varying decay rate model yields very different patterns compared to the other two non-constant models, suggesting that our model can distinguish regulatory mechanisms of transcription from mechanisms regulating mRNA degradation. Swapping linear for exponential growth did not affect our results significantly as exponential and linear growth are nearly indistinguishable over one cell size doubling (Supplementary Figure S1). Additionally, we observe that correlation coefficients have an essential role in distinguishing constant versus cell cycle-modulated kinetics.

Approximate Bayesian inference reveals global patterns of transcription kinetics regulation

Motivated by the ability of the model to produce distinct summary statistics between different underlying modes of kinetics regulation, we develop an inference framework to infer these modes of regulation along with gene-specific kinetic parameters from scEU-seq data. A main challenge here is that inference requires optimising over a large parameter space across all five modelling hypotheses for each gene in the data separately. Therefore, using a maximum likelihood, optimisation-based or Markov Chain Monte Carlo approach would be computationally inefficient. For that reason, we develop a simulation-based inference framework based on ABC rejection sampling, which allows to test multiple modelling hypotheses and millions of parameter sets against the entire set of genes in a highly parallel manner (Figure 1 (C)). Given a parameter set sampled from a prior distribution and a realisation of the model, we generate a collection of summary statistics (see Tables 1, 2 and Supplementary Methods 1.4) and simultaneously compare them to summary statistics of all genes in the data (Supplementary Methods 1.5.1). Additionally, using an Approximate Bayesian framework

allows to perform model selection across several models with different levels of complexity (number of parameters) [59], *i.e.* to assign probabilities to the various hypotheses, and decide which kinetic model is most likely to describe each gene. A significant challenge for inference is that scEU-seq data are characterised by low mRNA capture efficiency [25], meaning that summary statistics of the data and the model are not directly comparable. We address this issue by modelling cell-specific capture efficiencies and downsampling the moments to enable reliable inference [35, 46, 60], as we describe in detail in the next sections.

In Figure 2 (B) we show example posterior fits over the different summary statistics (see Tables 1 and 2) for a constant and a non-constant model. We observe that the mean expression of the gene *RNH1* shows a monotonous, approximately linear increase, while its Fano Factor stays relatively constant over the cell cycle (Figures 2 (B i-ii)). Even though the constant model shows a linear increase in both mean expression and Fano factor, the posterior model fit agrees well with the data. The ratio of mean labelled vs total expression of *RNH1* increases with pulse labelling time and decreases along chase conditions as the chase duration increases (Figure 2 (B iii)). The model behaves similarly and we see that the data lies well within the posterior model fit. The mean (cell cycle-averaged) correlation of labelled-unlabelled expression follows a similar pattern with the ratios and even though the data exhibits a more noisy behaviour, its error bars lie within the posterior model fit (Figure 2 (B iv)). The correlation of cell cycle-dependent means is almost zero across labelling conditions for both model and data (Figure 2 (B v)).

The gene *KIF22* shows a large variation in mean and Fano factor of expression over the cell cycle (Figures 2 (B vii-viii)) as we see a strong increase towards the late cell cycle stages. The non-constant decay rate model explains these patterns well as the error bars of the data lie within the posterior fit. The ratio of mean labelled vs total expression and the mean correlation of labelled-unlabelled expression of *KIF22* show similar behaviour with the respective statistics of *RNH1*, and even though they appear more noisy than the model, we obtain a good overall posterior fit (Figures 2 (B ix-x)). The correlation of means is significantly higher compared to the *RNH1* case, and even though it is also very noisy, it is captured well by the non-constant posterior fit (Figure 2 (B xi)). These model fits show that *RNH1* is explained by constant kinetics along the cell cycle and its synthesis rate scales linearly with cell size (Figure 2 (B vi)). On the other hand, the gene *KIF22* is explained by non-constant kinetics as its decay rate varies with cell cycle progression (Figure 2 (B xii)). Overall, our inference framework provides us with approximate posterior distributions of gene-specific kinetic parameters (Supplementary Figure S3). From this example we see that synthesis and decay rate are better identified compared to burst size and burst frequency. Interestingly, we observe that burst size and frequency are negatively correlated as well as that decay rate estimates are independent from the rest of the parameters.

After performing parameter inference over all 3419 genes that we selected for the analysis (Supplementary Methods 1.1.4), we find that the majority (87%) are explained by constant models, and about 7% of genes are assigned to non-constant models (Figure 2 (C i)). Within the subset of genes that are assigned to constant models (*constant* genes), we find that the majority (62%) are assigned to the size-scaling synthesis rate model (*scaling* genes) and a smaller fraction of genes (16%) are assigned to the non-scaling synthesis rate model (*non-scaling* genes), while for the rest of constant genes we get an uncertain model choice outcome (Figure 2 (C ii)). Among the subset of genes assigned to non-constant models (*non-constant* genes), we find two major groups assigned to the burst frequency model (41%) and the decay rate model (27%), respectively, while a small fraction of genes are assigned to the burst size model (Figure 2 (C iii)). For the rest of non-constant genes, we find two noteworthy groups of genes, one that is described well by both the burst frequency and the burst size model and one that is described well by both the burst frequency and the decay rate model. For the remaining genes (6%) we either get an uncertain model selection outcome or all the

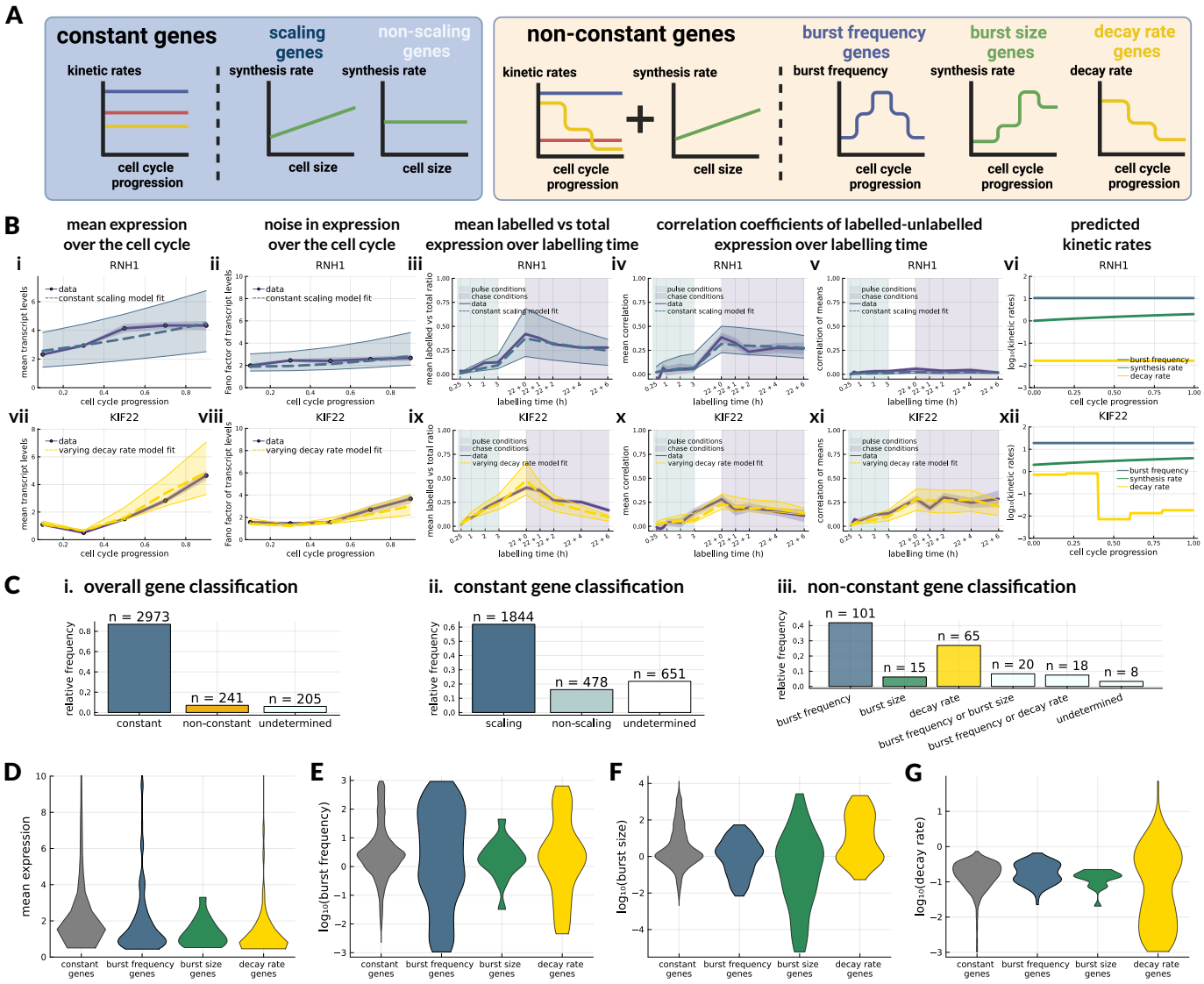


Figure 2: **Gene classification with respect to models of kinetics regulation** (A) Cartoon overview of the different model - gene classes. Constant models (blue box) assume that synthesis rate either scales with cell size (scaling genes) or stays constant (non-scaling genes), while all other kinetic rates stay constant along the cell cycle. Non-constant models (yellow box) assume that synthesis rate scales with cell size and additionally one kinetic rate varies with the cell cycle, while the rest of kinetic rates stay constant. (B) Summary statistics used for comparing models and data for an example constant gene (top row) and an example non-constant gene (bottom row), together with estimated kinetic rates for each gene. (C) Overall statistics of gene classification obtained from the ABC model selection. (D-G) Comparison of mean expression and kinetic rate magnitudes across different classes of genes. For each model class (constant and non-constant), proportions of specific sub-classes are shown. For constant, varying burst frequency, varying burst size and varying decay rate genes, we show distributions of (D) mean expression, (E) burst frequency , (F) burst size and (G) decay rate.

models performed poorly.

Furthermore, our inference reveals general patterns of mRNA kinetics across the different groups of genes. We observe that genes in all groups span a wide range of mean expression levels, while we observe that constant genes have a slightly higher median expression level compared to non-constant genes (Figure 2 (D)). We find that burst frequency is highly variable across all constant and non-constant genes, while burst frequency genes exhibit the strongest variation (Figure 2 (E)). Additionally, cell cycle-regulated burst size shows the largest variation, reaching very low levels compared to genes with constant burst size (Figure 2 (F)). We similarly see large variation in decay rate among decay rate genes, while for the rest of the genes the decay rate is constrained in similar levels (Figure 2 (G)).

We also find that cell cycle-dependent regulation of decay rate has an important impact on mRNA half-life. The majority of genes have median half-lives of approximately 4-5 hours, while the median mRNA half-life of decay rate genes is longer than 10 hours (Supplementary Figure S4 (A) and (D)). After comparing our estimates of mRNA half-life with two other metabolic labelling studies (sci-fate [51] and scSLAM-seq [48]), we find that the estimated half-lives of constant genes agree well with the sci-fate predictions (Spearman's $\rho \approx 0.9$, Supplementary Figure S4 (B)), however a lower correlation is observed in the case of non-constant genes (Supplementary Figure S4 (E)). We find good agreement between our constant gene half-lives and the scSLAM-seq predictions (Spearman's $\rho \approx 0.7$, Supplementary Figure S4 (C)), while there is a poor agreement in the case of non-constant genes (Supplementary Figure S4 (F)). These discrepancies in the case of non-constant genes indicate that modelling cell cycle-dependence improves the estimation of mRNA half-life. The overall discrepancies could in addition be attributed to differences between the scEU-seq, sci-fate and scSLAM-seq metabolic labelling protocols [47] or to the fact that these estimations are based on different cell lines.

Modelling capture and labelling efficiency allows to uncouple technical and biological noise

The technical biases associated with sequencing protocols make the task of modelling scRNA-seq data challenging. Building a stochastic model that accurately explains the observed variability in the data, requires that the various biological and technical contributions to the total noise are carefully accounted for. We here demonstrate that the relationship between unlabelled and labelled mRNA levels of a single gene can reveal how different noise components contribute to the total transcript variability. In the absence of technical noise, the total variance in transcript levels $x = u + l$ of a gene (u: unlabelled, l: labelled) in a time-resolved single-cell snapshot can be partitioned as follows (see details in Supplementary Methods 1.6):

$$\mathrm{Var}(x) = \underbrace{\mathbb{E}(\mathrm{Var}(u \mid \tau, t) + \mathrm{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2\,\mathbb{E}(\mathrm{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} + \underbrace{\mathrm{Var}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}}$$

where τ denotes the cell cycle phase and t the pulse-chase labelling time (condition). Averaging the cell cycle- and labelling time-dependent transcript variances gives a measure of the noise due to mRNA synthesis and degradation kinetics (1st term). This term includes variation due to both active degradation and dilution, where the latter stems from stochastic partitioning at cell division. The upstream noise due to bursty gene promoter can be measured by averaging the cell cycle- and labelling time-dependent covariances of unlabelled and labelled transcripts (2nd term). The extrinsic noise due to cell cycle variation can be measured by the variance of the cell cycle-dependent mean transcript levels (3rd term). The bursty promoter noise and the cell cycle variation contribute

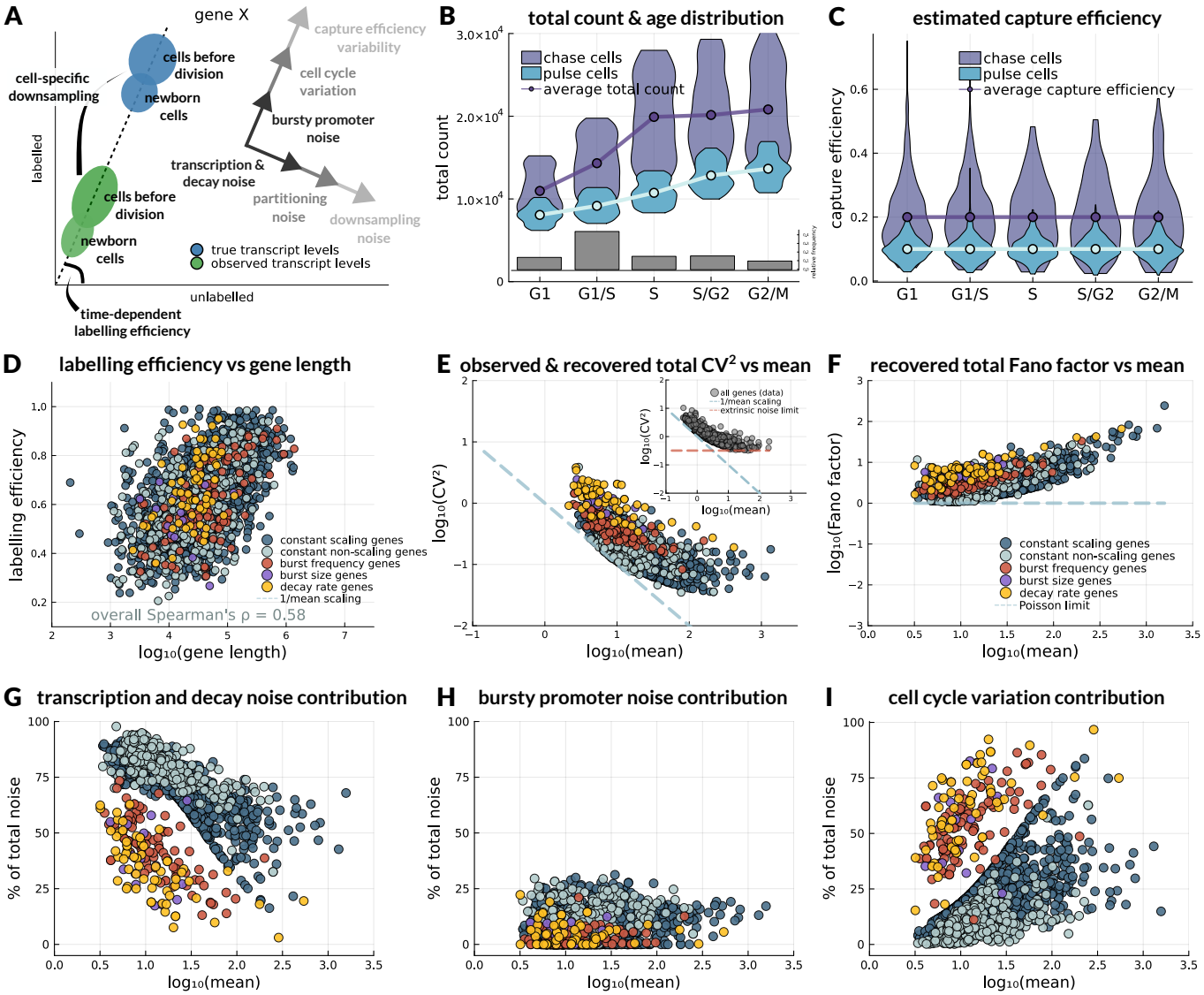


Figure 3: **Biological and technical effects on transcriptional noise and biological noise decomposition.** (A) Cartoon scatter plot of unlabelled-labelled mRNA, where the observed transcript levels (green ellipses) is a noisy version of the true transcript levels (blue ellipses) due to cell-specific downsampling. The vectors describe different noise components that contribute to the total transcript variability by correlating or decorrelating the unlabelled-labelled mRNA levels, illustrated by lengthening or widening the ellipses, respectively. Bursty promoter noise, cell cycle variation (biological) and capture efficiency variability (technical) contribute to correlating unlabelled and labelled transcript levels, while transcription & mRNA degradation noise, partitioning noise due to cell division (biological), and cell-specific downsampling (technical) contribute to decorrelating transcript levels. (B - C) Uncoupling biological and technical effects on transcript variability: (B) Total mRNA count distributions per cell cycle phase and age distribution contribute to biological variability. The average total count is shown to approximately double over the course of the cell cycle. Calculation of the empirical age distribution is described in Supplementary Methods 1.1.3. (C) Estimated capture efficiency distributions per cell cycle phase. The distributions are shown to be independent of cell cycle progression. (D) Dependence of inferred gene-specific labelling efficiency on gene length. (E-F) Observed and recovered noise-mean relationships: (E) shows the total CV^2 as observed in the data (inset scatter plot in grey) and as recovered by the different models. (F) shows the recovered total Fano factor as given by the model after removing technical noise. (G-I) Model-based biological noise decomposition: proportional contributions of the different biological noise sources to the total noise (Supplementary Methods 1.6).

to correlating the unlabelled-labelled transcript levels, while transcription and degradation kinetics noise reduces these correlations (Figure 3 (A)). In the presence of technical noise the above formula does not hold as there are additional non-biological contributions to the total transcript variability: cell-specific downsampling noise and cell-cell variability attributed to capture efficiency (Figure 3 (A)).

Before we are able to decompose and quantify the different noise components, technical noise effects need to be removed from the data. We therefore devise an approach to account for technical noise in our modelling framework. This allows us to construct reliable summary statistics that are directly comparable with summary statistics of the noisy data. We first observe that the average total mRNA count increases and almost doubles over the course of the cell cycle, which suggests that it is representative of the average cell size (Figure 3 (B)). At the same time, we note that the total count within each cell cycle phase can exhibit larger than 2-fold variation. Significant differences are also observed between the total counts of pulse-treated and chase-treated cells. Additionally, the age distribution (proportion of cells across cell cycle phases) introduces another source of variability that can affect the shapes of the total count distributions.

Nonetheless, it is unclear what part of this total count variation comes from technical noise effects. To quantify these effects requires to have access to the capture efficiency of the sequencing experiment, however this information is not provided by the scEU-seq protocol. We therefore estimate the cell-specific capture efficiencies directly from the data, based on the idea that the total mRNA count of a cell is proportional to cell size so it can be used as a cell-specific scaling factor. We then postulate that cells that are in the same cell cycle phase have on average the same cell size - total count. The deviation from this average total count provides an estimate for the relative capture efficiency of the cell within its cell cycle phase. We show that our estimated capture efficiency distributions per cell cycle phase do not scale with cell cycle progression (Figure 3 (C)). This indicates that our approach decouples cell size and age variability from technical noise effects on total mRNA count. Moreover, it normalises the different sample sizes of pulse and chase conditions (see details in Supplementary Methods 1.3.1). Having obtained the capture efficiency estimates, we are able to correct the model outputs by computing downsampled versions of moments (see Supplementary Methods 1.3.2).

An important additional technical noise source at the single-gene level is the labelling efficiency of the pulse-chase experiments, as RNA labelling might not impact genes uniformly. To address this, we model labelling efficiency as a gene-specific parameter that is fitted simultaneously with the kinetic rates. We find that the inferred labelling efficiency shows significant variation across genes as well as that it correlates with gene length, demonstrating the importance of estimating it as a gene-specific parameter (Figure 3 (D)). It is also evident that this trend is independent of the selected gene-specific models, thus validating that it does not influence model selection.

Bayesian inference reconstructs the transcriptional noise landscape

Our approach of uncoupling technical and biological noise makes it possible to estimate the true mean expression levels and the true levels of transcriptional noise. The observed total noise (CV^2) for all 3419 genes decreases with mean expression but stays above the Poisson limit (Figure 3 (E) inset), and deviates from the inverse mean scaling for highly expressed genes. This deviation at high expression levels reflects extrinsic noise attributed to both biological and technical effects. Informing the model with the point estimates from the inferred kinetics posteriors, we recover and visualise the true noise-mean relationship (Figure 3 (E)). The dependence of noise on the mean recovered by

the model follows a similar pattern to the data, but the noise is an order of magnitude lower and deviates less from the inverse mean scaling law as these deviations are purely due to biological noise. We observe distinct noise patterns among groups of genes explained by different models, with cell cycle-dependent genes commonly displaying higher noise levels than constant genes. Furthermore, using the Fano factor to quantify the noise-mean relationship, we observe that the increase in the mean is associated with increasing deviation from the Poisson limit due to extrinsic noise (Figure 3 (F)).

Next, we use the noise decomposition formula to quantify how the different biological noise sources contribute to the total transcript variability (Figure 3 (G-I)). We find that transcription and decay kinetics and cell cycle variation are the most prominent sources of noise. Interestingly, for most constant genes the largest proportion of noise is explained by mRNA synthesis and decay kinetics (Figure 3 (G)) that contribute to more than 50% of the total noise. On the other hand, cell cycle variation is the dominant source of noise for most non-constant genes (Figure 3 (I)). The bursty promoter noise is less prominent across genes, however not negligible as in a considerable number of genes it explains more than 20% of the noise.

Further, we investigate how the different noise sources depend on the inferred kinetic rates of constant genes. We find that low relative burst frequency (burst frequency divided by decay rate) gives rise to higher levels of transcription and decay noise while high relative burst frequency leads to Poisson noise (Supplementary Figure S5 (A)). On the other hand, we do not observe a clear dependence of transcription and decay noise on burst size (Supplementary Figure S5 (B)). Interestingly, we find that the interplay between burst size and burst frequency strongly determines the bursty promoter noise levels (Supplementary Figure S5 (C)). We observe an overall increase of bursty promoter noise with burst size, however in the limit of high burst frequency, noise decreases with burst size. In the limit of high burst size and low burst frequency, bursty promoter noise reaches its highest levels.

We find that decay rate is a key determinant of all components of noise. High decay rates give rise to high levels of transcription and decay noise while more stable transcripts with lower decay rate show a decrease in the noise as the mean increases (Supplementary Figure S6 (A)). Interestingly, high decay rates also cause high levels of bursty promoter noise in the limit of low mean expression (Supplementary Figure S6 (B)). Finally, the cell cycle variation of size-scaling genes is relatively constant and hence does not depend on decay rate, however the cell cycle variation of non-scaling genes decreases as decay rate increases (Supplementary Figure S6 (C)). These observations are in agreement with model predictions reported in [20] and [28].

The model identifies scaling properties of gene expression with cell size

We next focus on genes with constant kinetics and we find that the model reveals key differences between size-dependent and size-independent gene expression. As an example, *RAB7A*, a gene that is involved in intracellular vesicle trafficking and growth-factor-mediated signalling, is explained by a constant scaling model and *MGST1*, a gene responsible for protecting endoplasmic membranes from oxidative stress, is explained by a non-scaling model (Figure 4 (A)). *RAB7A* shows an increase in the mean and Fano factor of transcript levels along the cell cycle (Figure 4 (A i-ii)), which is explained by a size-scaling mRNA synthesis rate (Figure 4 (A iii)), while the mean transcript levels and Fano factor of *MGST1* stay relatively constant during cell cycle progression (Figure 4 (A iv-v)), which is explained by a constant synthesis rate (Figure 4 (A vi)).

We investigate global size-scaling properties of gene expression. We find that the mean expression of

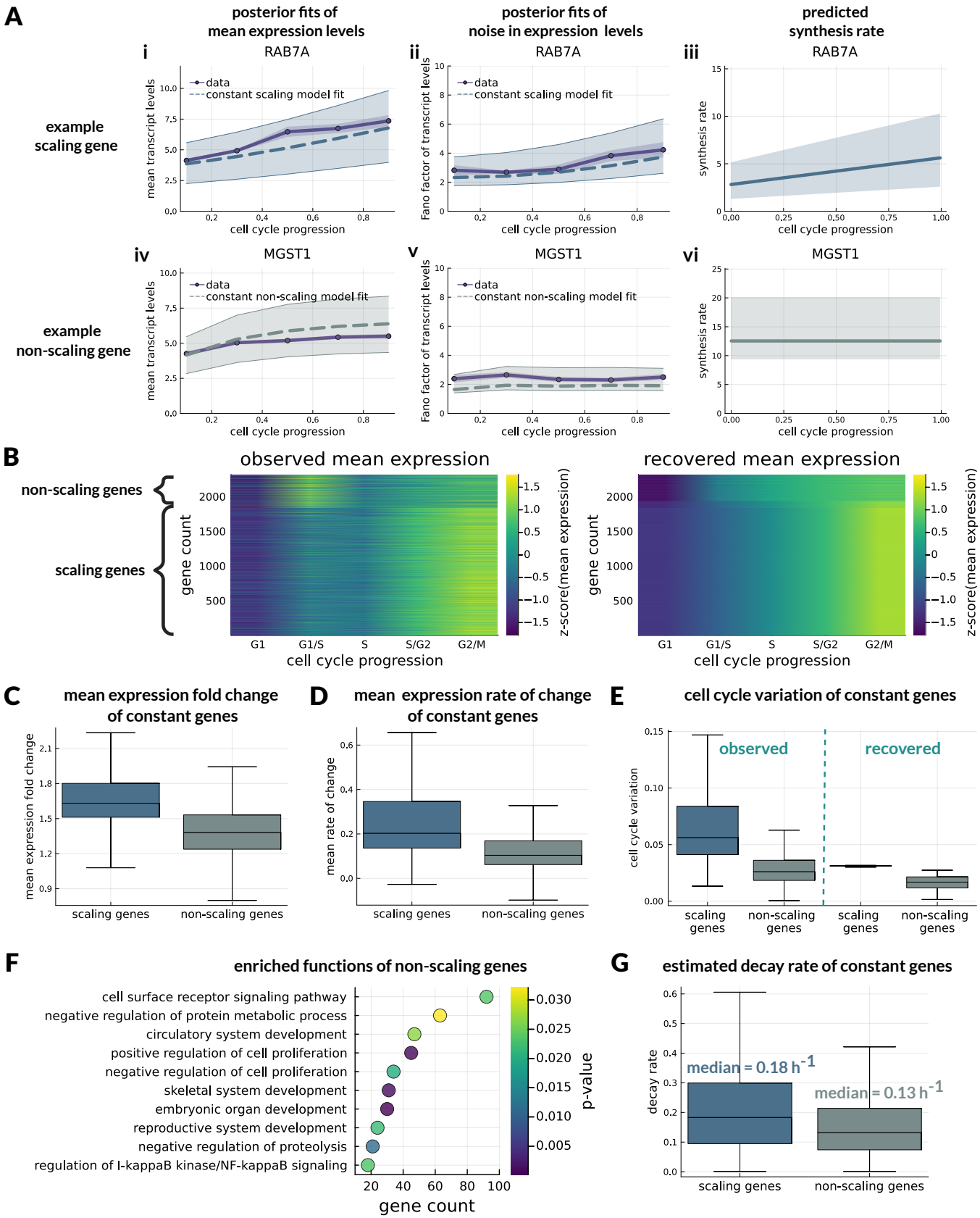


Figure 4: **Constant genes and size-scaling properties of gene expression.** (A) Posterior model fits for a scaling and a non-scaling gene. Panels (i)-(ii) and (iv)-(v) show cell cycle-dependent mean and Fano factor of transcript levels, and panels (iii) and (vi) show posterior estimates of scaling and non-scaling synthesis rate. (B) Clustered heatmaps of normalised mean expression along the cell cycle for constant genes, as observed in the data (left) and as recovered by the model (right), after removing technical noise effects. (C-E) Differences between scaling and non-scaling genes in cell cycle-dependent expression: (C) shows differences in the observed fold change of mean expression between G1 and G2/M phase, (D) shows differences in mean rates of change of expression along consecutive cell cycle phases and (E) shows differences in variability due to cell cycle-dependence, both observed in the data and recovered by the model. (F) Enriched GO terms of non-scaling genes. (G) Differences in decay rate distributions of scaling and non-scaling genes.

scaling genes increases monotonically during the cell cycle, while the mean expression of non-scaling genes reaches a steady state towards the end of cell cycle progression. This can be observed in clustered heatmaps of both observed and recovered (after technical noise removal) mean expression of all constant genes (Figure 4 (B)). To validate this observation, we quantify changes in mean expression of constant genes along the cell cycle. We find that both the fold change of expression between early and late cell cycle and the mean rate of change of expression along cell cycle progression are significantly higher for scaling compared to non-scaling genes (Figures 4 (C-D)). Additionally, we quantify variation due to cell cycle-dependence, including variation observed in the data and variation recovered from the model after technical noise correction (see Supplementary Methods 1.6). We find that in both cases, cell cycle variation of size-scaling expression is higher compared to non-scaling (Figure 4 (E)). This suggests that the combined signatures of mean expression and noise levels identify the scaling with cell size.

We analysed which gene functions were associated with non-scaling gene expression using the enrichment analysis provided by the DAVID bioinformatics database [61]. We find that the majority of non-scaling genes are involved in cell surface receptor signalling. Other important enriched sets include protein metabolic processes, regulation of cell proliferation, tissue and system development and proteolysis. Finally, we asked how scaling of transcription with cell size relates to mRNA degradation. We find that scaling genes have overall higher decay rates compared to non-scaling genes (Figure 4(G)). As a consequence, scaling genes are characterised by shorter mRNA half-life, with their median mRNA half-life being shorter by more than one hour compared to non-scaling genes (Supplementary Figure S4 (A)).

The model predicts patterns of cell cycle-dependent kinetics regulation

Next, we focus on genes with non-constant kinetics and we find that the model identifies three groups of genes, each characterised by a distinct mode of kinetics regulation during the cell cycle. For these groups of genes, either burst frequency, burst size or degradation rate are primarily modulated over the cell cycle, while the rest of kinetic rates stay constant. Posterior model fits for *CDK4*, *PCNA* and *CDK1*, which are genes with key functional roles in cell cycle control, display good agreement with the data (Figure 5 (A)). The variations in mean and Fano factor of *CDK4*, *PCNA* and *CDK1* transcript levels along the cell cycle (Figure 5 (A i-ii, iv-v, vii-vii)) are explained by a particular modulation of their kinetic rates (Figure 5 (A iii, vi, ix)). Interestingly, in all three cases the predicted kinetics show a peak in a particular cell cycle phase. The increase in mean and noise of expression of these genes in the late part of the cell cycle is achieved by a sharp increase in burst frequency or burst size about S phase, or by a sharp decrease in decay rate after G1 phase, respectively.

To obtain comprehensive insights into the global regulation of kinetics during the cell cycle, we perform hierarchical clustering of each set of varying kinetic rates based on cosine similarity (Figure 5 (B)). Strikingly, all emerging clusters are characterised by a unique peak in a particular cell cycle phase. Combining all clusters together, we observe a wave-like pattern in the overall kinetics regulation throughout the cell cycle. We find that varying decay rate shows peaks in the early stages of the cell cycle, while varying burst size and frequency show peaks in S, G2 and M phase. This suggests distinct regulatory strategies of transcription and degradation during the cell cycle. This non-trivial clustering of genes based on their kinetics would not be possible to extract from mean expression dynamics alone, as the different gene groups show similar expression profiles, with an increase towards the late cell cycle stages (Supplementary Figure S7).

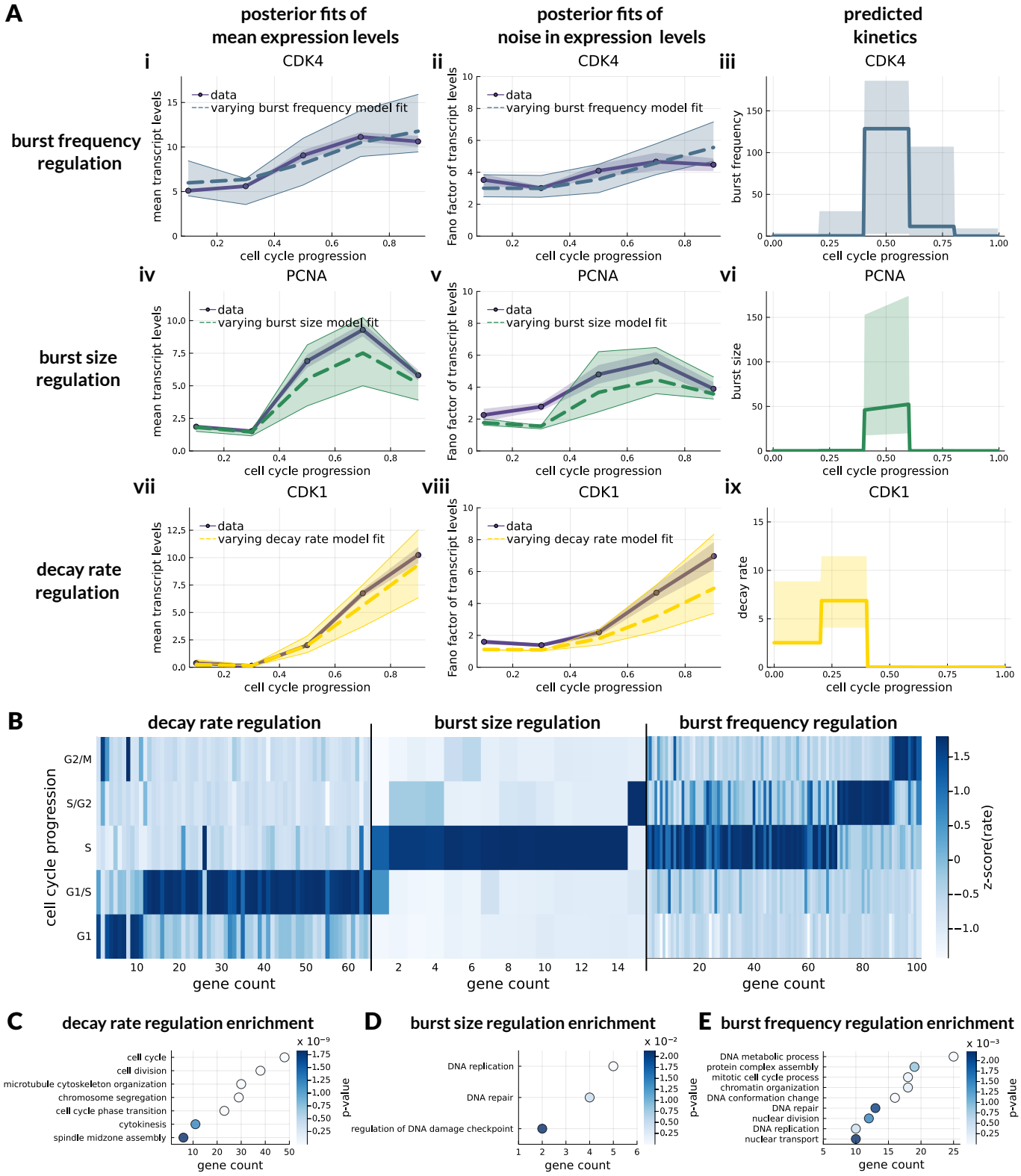


Figure 5: **Non-constant genes and cell cycle-dependent regulation of kinetic rates.** (A) Posterior model fits for three example non-constant genes. Panels (i)-(ii),(iv)-(v) and (vii)-(viii) show cell cycle-dependent mean and Fano factor of transcript levels, and panels (iii), (vi) and (ix) show posterior estimates of cell cycle-modulated kinetic rates. (B) Clustered heatmap of cell cycle-modulated kinetic rates: the different sections of the heatmap correspond to decay rate genes, burst size genes and burst frequency genes. (C-E) Enriched GO terms associated with non-constant genes: (C) shows GO terms linked with decay rate regulation, (D) shows GO terms linked with burst size regulation and (E) shows GO terms linked with burst frequency regulation.

We also find, using gene enrichment analysis, that each mode of kinetics regulation is associated with a particular set of functions related to cell cycle control and cell division (Figures 5 (C-E)). Genes with varying decay rate are found to be involved in the regulation of G2/M phase transition and cell division, including chromosome segregation, spindle assembly and cytokinesis (Figure 5 (C)). Genes with varying burst size are primarily involved in S phase regulation and DNA replication, including DNA repair and damage checkpoint control (Figure 5 (D)). Varying burst frequency genes are associated with a variety of cell cycle control functions, including DNA replication, DNA repair and nuclear division (Figure 5 (E)). Interestingly, burst frequency regulation is also associated with other cellular functions such as protein complex assembly, chromatin organisation and DNA conformation changes, suggesting a potential link of these processes with the cell cycle. Overall, these results indicate that biological functions responsible for cell cycle regulation could be achieved by the coordinated activation of kinetic rates in particular cell cycle phases.

3 Discussion

We developed a comprehensive modelling framework to reveal global regulatory mechanisms of mRNA transcription and degradation during the cell cycle from time-resolved transcriptomics data. Our stochastic model integrates cell cycle progression and pulse-chase labelling to reveal different modes of dynamic regulation in mRNA control. Modelling capture and labelling efficiency allowed us to deconvolve technical and biological noise of the data and reveal the true sources of cellular variation that stem from distinct mechanisms of transcription regulation.

A central result of our analysis is that patterns of cell cycle- and cell size-dependent regulation of transcription dynamics are distinguished through dynamic correlations of metabolic labels. In particular, we showed that the correlations of unlabelled and labelled mRNAs quantify transcriptional bursting and can be distinguished from noise due to mRNA synthesis and degradation kinetics, which we exploit for inference and model selection. Existing models of metabolic labelling protocols are limited to averages across cells as they employ deterministic approaches [25, 26, 48–52, 62] or they are restricted to labelled transcripts [54, 55] and hence cannot make use of dynamic correlations. Our approach, which is based on the ABC rejection method, effectively samples from the full posterior distribution across parameters and models by comparing millions of candidate simulations to any number of genes in parallel. The method results in an efficient inference that scales favourably with the number of genes as we perform all the simulations once for each model, independent of the specific data relevant to any gene. Our method is thus ideally suited for the analysis of genome-wide data [63]. In simulation-based inference literature, these kind of approaches are known as amortised, incurring an up-front computational cost, but are applicable to new data points efficiently [64].

A distinctive feature of our approach is that it simultaneously corrects for technical variation through cell-specific capture efficiency and gene-specific metabolic labelling efficiency. Previous inference methods utilising metabolic labelling data relied on specific distribution assumptions [25, 55] or on RNA spike-ins [55] for modelling capture efficiency, while other studies modelled labelling efficiency using either a global rescaling or mixture models [26, 48–50, 53, 62, 65]. We verified that capture efficiency was cell cycle-independent, thus validating that we can distinguish the effects of cell cycle and size. We also showed that labelling efficiency varied dramatically and correlated well with gene length but did not bias our model selection, thus demonstrating that one needs to calibrate labelling efficiency for parameter estimation. The observed correlation between labelling efficiency and gene length may provide insights into the mechanistic basis of metabolic labelling.

By removing technical noise effects, we were able to dissect sources of biological noise contributing to the total transcript variability. We found that mRNA synthesis and degradation kinetics and cell cycle-dependent variation are the primary genome-wide sources of noise, while only a small fraction of the total noise stems from transcriptional bursting. The low levels of bursting noise were explained by overall low burst sizes and relatively high burst frequencies. These observations indicate that transcript abundance is characterised by minimal intrinsic stochasticity, as demonstrated in [11].

We found that most genes scale transcription with cell size. Genes that did not scale were either involved in growth-independent processes, such as cell surface receptor signalling, or in other processes such as proliferation and development. Cell surface and DNA binding proteins do not scale with cell volume and have been previously identified as potential non-scaling genes [66–68]. Size-scaling genes typically rise more rapidly during the cell cycle, have higher degradation rates, and have higher cell cycle variation than non-scaling genes. This suggests that increased synthesis rate with cell size requires higher degradation for cells to achieve homeostasis [29, 69].

Our model predicts waves of regulation across the cell cycle-dependent transcriptome, with kinetic rates peaking at specific cell cycle phases. We observed no step-wise transcriptional changes, suggesting that gene dosage effects mattered less than specific cell cycle regulation of kinetic rates. Genes with kinetics modulating decay rate peaked first in G1, then burst size peaked in S and S/G2, then burst frequency peaked in S phase and the later stages of the cell cycle. Our predictions agree qualitatively with a similar sharp decrease in burst frequency following S phase observed using smFISH [57], and the increased synthesis rates and decreased decay rates following G1 reported in the original scEU-seq study [25]. However, this study could not distinguish burst size and frequency modulations.

Although we expect our inference method to generalise to more complex models and parametric assumptions, it also has several limitations. One limitation is that ABC rejection sampling can lead to weaker accuracy in parameter estimation and broader posterior distributions compared to optimisation-based schemes (see e.g. Supplementary Figure S3). A hybrid method based on ABC that utilises parallelisation while improving accuracy could provide more accurate parameter inference [63]. Another limitation is that we assumed deterministic cell cycles across cells, thus a potential improvement would be to account for cell cycle length variability in our moment estimates.

Other studies use modelling to better understand transcription dynamics through nascent mRNA [62, 70–73] including cell cycle-dependent effects [74] and during development [53, 71]. Therefore, splicing data could be used instead of cell cycle reporters in our inference method. Similarly, splicing kinetics could extend our analysis to understand transcription variation along developmental trajectories [53, 70, 71] or identify hidden sources of cellular variation [39].

Our study illustrates how mechanistic modelling and amortised inference can efficiently integrate data across labelled, unlabelled transcripts and cell cycle reporters and also across different pulse and chase protocols. We have demonstrated that mechanistic models can be used to overcome the challenge of technical noise across modalities commonly suffered by multi-view statistical methods [75, 76]. Data integration across conditions, time and modalities is an important challenge for single-cell multi-omics, including epigenome and proteome [77–79], and mechanistic models like ours could provide a versatile tool for integrating these data [80].

In summary, our study reveals global mechanisms of transcription regulation through stochastic modelling of time-resolved single-cell data. Our inference methodology is highly scalable and translates seemingly to complex models of gene regulation. It is thus well suited to uncover complex mechanisms of gene expression control from single-cell transcriptomics. We thus expect that our framework could contribute to uncover regulatory mechanisms that drive cell fate decisions or their dysregulation in disease.

Acknowledgements

This work was supported by UKRI through a Future Leaders Fellowship [MR/T018429/1 to P.T.]; and the Department of Mathematics, Imperial College London through a Roth scholarship [to D.V.]. We would like to thank Nicolas Battich for sharing and discussing their published experimental data from the scEU-seq study. Also, we would like to acknowledge Wenhao Tang, Fern Hughes, Francesco Puccioni, Hannah Dewhurst, Paul Piho and Johannes Pausch for useful conversations.

References

[1] William J. Blake et al. "Noise in eukaryotic gene expression". *Nature* 422 (2003), pp. 633–637. DOI: 10.1038/nature01546.

[2] Arjun Raj et al. "Stochastic mRNA Synthesis in Mammalian Cells". *PLOS Biology* 4 (2006), pp. 1–13. DOI: 10.1371/journal.pbio.0040309.

[3] Jonathan R. Chubb et al. "Transcriptional Pulsing of a Developmental Gene". *Current Biology* 16 (2006), pp. 1018–1025. DOI: 10.1016/j.cub.2006.03.092.

[4] David M. Suter et al. "Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics". *Science* 332 (2011), pp. 472–474. DOI: 10.1126/science.1198817.

[5] Roy D. Dar et al. "Transcriptional burst frequency and burst size are equally modulated across the human genome". *Proceedings of the National Academy of Sciences* 109 (2012), pp. 17454–17459. DOI: 10.1073/pnas.1213530109.

[6] Mads Kærn et al. "Stochasticity in gene expression: from theories to phenotypes". *Nature Reviews Genetics* 6 (2005), pp. 451–464. DOI: 10.1038/nrg1615.

[7] Vahid Shahrezaei and Peter S Swain. "The stochastic nature of biochemical networks". *Current Opinion in Biotechnology* 19 (2008), pp. 369–374. DOI: 10.1016/j.copbio.2008.06.011.

[8] Arjun Raj and Alexander van Oudenaarden. "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences". *Cell* 135 (2008), pp. 216–226. DOI: 10.1016/j.cell.2008.09.050.

[9] Avigdor Eldar and Michael B. Elowitz. "Functional roles for noise in genetic circuits". *Nature* 467 (2010), pp. 167–173. DOI: 10.1038/nature09326.

[10] P. Thomas, N. Popovic, and R. Grima. "Phenotypic switching in gene regulatory networks". *Proceedings of the National Academy of Sciences* 111 (2014), pp. 6994–6999. DOI: 10.1073/pnas.1400049111.

[11] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. "Control of Transcript Variability in Single Mammalian Cells". *Cell* 163 (2015), pp. 1596–1610. DOI: 10.1016/j.cell.2015.11.018.

[12] Orsolya Symmons and Arjun Raj. "What's Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism". *Molecular Cell* 62 (2016), pp. 788–802. DOI: https://doi.org/10.1016/j.molcel.2016.05.023.

[13] Richard Losick and Claude Desplan. "Stochasticity and Cell Fate". *Science* 320 (2008), pp. 65–68. DOI: 10.1126/science.1147888.

[14] Megan A. Coomer, Lucy Ham, and Michael P.H. Stumpf. "Noise distorts the epigenetic landscape and shapes cell-fate decisions". *Cell Systems* 13 (2022), pp. 83–102. DOI: https://doi.org/10.1016/j.cels.2021.09.002.

[15] Natalie A. Cookson et al. "Cell cycle-dependent variations in protein concentration". *Nucleic Acids Research* 38 (2009), pp. 2676–2681. DOI: 10.1093/nar/gkp1069.

[16] Veronika Bierbaum and Stefan Klumpp. "Impact of the cell division cycle on gene circuits". *Physical Biology* 12 (2015), e066003. DOI: 10.1088/1478-3975/12/6/066003.

[17] Philipp Thomas. "Making sense of snapshot data: ergodic principle for clonal cell populations". *Journal of The Royal Society Interface* 14.136 (2017), p. 20170467. DOI: 10.1098/rsif.2017.0467.

[18] Noreen Walker, Philippe Nghe, and Sander J. Tans. "Generation and filtering of gene expression noise by the bacterial cell cycle". *BMC Biology* 14 (2016), p. 11. DOI: 10.1186/s12915-016-0231-z.

[19] Dann Huh and Johan Paulsson. "Non-genetic heterogeneity from stochastic partitioning at cell division". *Nature Genetics* 43 (2010), pp. 95–100. DOI: 10.1038/ng.729.

[20] Philipp Thomas. "Intrinsic and extrinsic noise of gene expression in lineage trees". *Scientific Reports* 9 (2019), p. 474. DOI: 10.1038/s41598-018-35927-x.

[21] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. "Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance". *Journal of The Royal Society Interface* 17 (2020), p. 20200360. DOI: 10.1098/rsif.2020.0360.

[22] C. J. Zopf et al. "Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression". *PLoS Computational Biology* 9 (2013), e1003161. DOI: 10.1371/journal.pcbi.1003161.

[23] Samuel O Skinner et al. "Single-cell analysis of transcription kinetics across the cell cycle". *eLife* 5 (2016), e12175. DOI: 10.7554/elife.12175.

[24] Zhixing Cao and Ramon Grima. "Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells". *Proceedings of the National Academy of Sciences* 117 (2020), pp. 4682–4692. DOI: 10.1073/pnas.1910888117.

[25] Nico Battich et al. "Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies". *Science* 367 (2020), pp. 1151–1156. DOI: 10.1126/science.aax3072.

[26] Haiyue Liu et al. "SLAM-Drop-seq reveals mRNA kinetic rates throughout the cell cycle". *Molecular Systems Biology* (2023), e11427. DOI: 10.15252/msb.202211427.

[27] Xi-Ming Sun et al. "Size-Dependent Increase in RNA Polymerase II Initiation Rates Mediates Gene Expression Scaling with Cell Size". *Current Biology* 30 (2020), pp. 1217–1230. DOI: 10.1016/j.cub.2020.01.053.

[28] Philipp Thomas and Vahid Shahrezaei. "Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations". *Journal of The Royal Society Interface* 18 (2021), p. 20210274. DOI: 10.1098/rsif.2021.0274.

[29] Scott Berry et al. "Feedback from nuclear RNA on transcription promotes robust RNA concentration homeostasis in human cells". *Cell Systems* 13 (2022), pp. 454–470. DOI: 10.1016/j.cels.2022.04.005.

[30] Chen Jia and Ramon Grima. "Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model". *iScience* 26 (2023), p. 105746. DOI: 10.1016/j.isci.2022.105746.

[31] Jong Kim and John C Marioni. "Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data". *Genome Biology* 14 (2013), R7. DOI: 10.1186/gb-2013-14-1-r7.

[32] Anton J. M. Larsson et al. "Genomic encoding of transcriptional burst kinetics". *Nature* 565 (2019), pp. 251–254. DOI: 10.1038/s41586-018-0836-1.

[33] Songhao Luo et al. "Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model". *Royal Society Open Science* 10 (2023), p. 221057. DOI: 10.1098/rsos.221057.

[34] Songhao Luo et al. "Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics". *Nucleic Acids Research* 51 (2022), pp. 68–83. DOI: 10.1093/nar/gkac1204.

[35] Wenhao Tang et al. "Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics". *Bioinformatics* 39 (2023), btad395. DOI: 10.1093/bioinformatics/btad395.

[36] Vahid Shahrezaei and Peter S. Swain. "Analytical distributions for stochastic gene expression". *Proceedings of the National Academy of Sciences* 105 (2008), pp. 17256–17261. DOI: 10.1073/pnas.0803850105.

[37] Fabian Fröhlich et al. "Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion". *PLOS Computational Biology* 12 (2016), e1005030. DOI: 10.1371/journal.pcbi.1005030.

[38] Lucy Ham et al. "Exactly solvable models of stochastic gene expression". *The Journal of Chemical Physics* 152 (2020), p. 144106. DOI: 10.1063/1.5143540.

[39] Lucy Ham, Marcel Jackson, and Michael PH Stumpf. "Pathway dynamics can delineate the sources of transcriptional noise in gene expression". *eLife* 10 (2021), e69324. DOI: 10.7554/elife.69324.

[40] Gennady Gorin et al. "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". *Nature Communications* 13 (2022), p. 7620. DOI: 10.1038/s41467-022-34857-7.

[41] Catalina A Vallejos et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities". *Nature methods* 14 (2017), pp. 565–571. DOI: 10.1038/nmeth.4292.

[42] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. "Computational and analytical challenges in single-cell transcriptomics". *Nature Reviews Genetics* 16 (2015), pp. 133–145. DOI: 10.1038/nrg3833.

[43] Gennady Gorin et al. "Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics". *PLOS ONE* 15 (2020), e0230736. DOI: 10.1371/journal.pone.0230736.

[44] Philip Brennecke et al. "Accounting for technical noise in single-cell RNA-seq experiments". *Nature Methods* 10 (2013), pp. 1093–1095. DOI: 10.1038/nmeth.2645.

[45] Mo Huang et al. "SAVER: gene expression recovery for single-cell RNA sequencing". *Nature methods* 15 (2018), pp. 539–542. DOI: 10.1038/s41592-018-0033-z.

[46] Wenhao Tang et al. "bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data". *Bioinformatics* 36 (2019), pp. 1174–1181. DOI: 10.1093/bioinformatics/btz726.

[47] Florian Erhard et al. "Time-resolved single-cell RNA-seq using metabolic RNA labelling". *Nature Reviews Methods Primers* 2 (2022), p. 77. DOI: 10.1038/s43586-022-00157-z.

[48] Florian Erhard et al. "scSLAM-seq reveals core features of transcription dynamics in single cells". *Nature* 571 (2019), pp. 419–423. DOI: 10.1038/s41586-019-1369-y.

[49] Gert-Jan Hendriks et al. "NASC-seq monitors RNA synthesis in single cells". *Nature Communications* 10 (2019), p. 3138. DOI: 10.1038/s41467-019-11028-9.

[50] Qi Qiu et al. "Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq". *Nature Methods* 17 (2020), pp. 991–1001. DOI: 10.1038/s41592-020-0935-4.

[51] Junyue Cao et al. "Sci-fate characterizes the dynamics of gene expression in single cells". *Nature Biotechnology* 38 (2020), pp. 980–988. DOI: 10.1038/s41587-020-0480-9.

[52] Jingyi Ren et al. "Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape". *Nature Methods* 20 (2023), pp. 695–705. DOI: 10.1038/s41592-023-01829-8.

[53] Rory J. Maizels, Daniel M. Snell, and James Briscoe. "Deep dynamical modelling of developmental trajectories with temporal transcriptomics". *bioRxiv* (2023). DOI: 10.1101/2023.07.06.547989.

[54] Qiangwei Peng, Xiaojie Qiu, and Tiejun Li. "Storm: Incorporating transient dynamics to infer the RNA velocity with metabolic labeling information". *bioRxiv* (2023). DOI: 10.1101/2023.06.21.545990.

[55] David M. Edwards, Philip Davies, and Daniel Hebenstreit. "Synergising single-cell resolution and 4sU labelling boosts inference of transcriptional bursting". *Genome Biology* 24 (2023), p. 138. DOI: 10.1186/s13059-023-02977-y.

[56] Hermannus Kempe et al. "The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise". *Molecular Biology of the Cell* 26 (2015), pp. 797–804. DOI: 10.1091/mbc.e14-08-1296.

[57] Olivia Padovan-Merhar et al. "Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms". *Molecular Cell* 58 (2015), pp. 339–352. DOI: 10.1016/j.molcel.2015.03.005.

[58] Tina Toni et al. "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". *Journal of The Royal Society Interface* 6 (2008), pp. 187–202. DOI: 10.1098/rsif.2008.0172.

[59] Juliane Liepe et al. "A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation". *Nature Protocols* 9 (2014), pp. 439–456. DOI: 10.1038/nprot.2014.025.

[60] Ramon Grima and Pierre-Marie Esmenjaud. "Quantifying and correcting bias in transcriptional parameter inference from single-cell data". *bioRxiv* (2023). DOI: 10.1101/2023.06.19.545536.

[61] Brad T Sherman et al. "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)". *Nucleic Acids Research* 50 (2022), pp. 216–221. DOI: 10.1093/nar/gkac194.

[62] Xiaojie Qiu et al. "Mapping transcriptomic vector fields of single cells". *Cell* 185 (2022), pp. 690–711. DOI: 10.1016/j.cell.2021.12.045.

[63] Simon Martina Perez, Heba Sailem, and Ruth E. Baker. "Efficient Bayesian inference for mechanistic modelling with high-throughput data". *PLOS Computational Biology* 18 (2022), p. 1010191. DOI: 10.1371/journal.pcbi.1010191.

[64] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference". *Proceedings of the National Academy of Sciences* 117 (2020), pp. 30055–30062. DOI: 10.1073/pnas.1912789117.

[65] Christopher Jürges, Lars Dölken, and Florian Erhard. "Dissecting newly transcribed and old RNA using GRAND-SLAM". *Bioinformatics* 34 (2018), pp. 218–226. DOI: 10.1093/bioinformatics/bty256.

[66] Samuel Marguerat and Jürg Bähler. "Coordinating genome expression with cell size". *Trends in Genetics* 28 (2012), pp. 560–565. DOI: 10.1016/j.tig.2012.07.003.

[67] Malika Saint et al. "Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation". *Nature Microbiology* 4 (2019), pp. 480–491. DOI: 10.1038/s41564-018-0330-4.

[68] Kora-Lee Claude et al. "Transcription coordinates histone amounts and genome content". *Nature Communications* 12 (2021), p. 4202. DOI: 10.1038/s41467-021-24451-8.

[69] Scott Berry and Lucas Pelkmans. "Mechanisms of cellular mRNA transcript homeostasis". *Trends in Cell Biology* 32 (2022), pp. 655–668. DOI: 10.1016/j.tcb.2022.05.003.

[70] Amit Zeisel et al. "Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli". *Molecular Systems Biology* 7 (2011), p. 529. DOI: 10.1038/msb.2011.62.

[71] Gioele La Manno et al. "RNA velocity of single cells". *Nature* 560 (2018), pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

[72] Volker Bergen et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". *Nature Biotechnology* 38 (2020), pp. 1408–1414. DOI: 10.1038/s41587-020-0591-3.

[73] Gennady Gorin et al. "RNA velocity unraveled". *PLOS Computational Biology* 18 (2022), e1010492. DOI: 10.1371/journal.pcbi.1010492.

[74] Andrea Riba et al. "Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning". *Nature Communications* 13 (2022). DOI: 10.1038/s41467-022-30545-8.

[75] Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. "Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study". *Bioinformatics* 36 (2020), pp. 4616–4625. DOI: 10.1093/bioinformatics/btaa530.

[76] Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. *Multi-view Data Visualisation via Manifold Learning*. 2021. DOI: 10.48550/arxiv.2101.06763.

[77] Ricard Argelaguet et al. "Computational principles and challenges in single-cell data integration". *Nature Biotechnology* 39 (2021), pp. 1202–1215. DOI: 10.1038/s41587-021-00895-7.

[78] Britta Velten and Oliver Stegle. "Principles and challenges of modeling temporal and spatial omics data". *Nature Methods* (2023). DOI: 10.1038/s41592-023-01992-y.

[79] Alev Baysoy et al. "The technological landscape and applications of single-cell multi-omics". *Nature Reviews Molecular Cell Biology* 24 (2023), pp. 695–713. DOI: 10.1038/s41580-023-00615-w.

[80] Maria Carilli et al. "Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data". *bioRxiv* (2023). DOI: 10.1101/2023.01.13.523995.

Supplementary Methods

In this section, we outline the details of our modelling framework, including data pre-processing and selection of summary statistics, construction of the model, parameter inference and model selection.

1.1 Data description, data pre-processing and summary statistics

1.1.1 Unlabelled and labelled mRNA count matrices

The scEU-seq dataset [25] includes count matrices for 4 transcriptomic modalities: unlabelled unspliced (**UU**), unlabelled spliced (**US**), labelled unspliced (**LU**) and labelled spliced (**LS**) mRNA. Our study is focused on unlabelled and labelled mRNA, so we initially add the unspliced and spliced counts to create the (total) unlabelled (**U**) and (total) labelled (**L**) mRNA count matrices: $\mathbf{U} = \mathbf{UU} + \mathbf{US}$, $\mathbf{L} = \mathbf{LU} + \mathbf{LS}$.

1.1.2 Grouping cells with respect to pulse-chase labelling condition

The dataset contains distinct subpopulations of cells that are treated with different metabolic labelling conditions of various durations (t_{pulse}) and various sequencing times after labelling (t_{chase}). These include the *pulse* conditions (c_{pulse}) that involve pulse labelling for time t_{pulse} as well as the *chase* conditions (c_{chase}) that involve pulse labelling for $t_{pulse} = 22$ hours followed by a chase experiment for time t_{chase}. The various combinations of (t_{pulse}, t_{chase}) for each case are outlined in the following tables (time in hours):

$$c_{pulse} = \begin{bmatrix} t_{pulse}: & 0.25 & 0.5 & 0.75 & 1 & 2 & 3 \\ t_{chase}: & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{1a}$$

$$c_{chase} = \begin{bmatrix} t_{pulse}: & 22 & 22 & 22 & 22 & 22 \\ t_{chase}: & 0 & 1 & 2 & 4 & 6 \end{bmatrix} \tag{1b}$$

Part of our summary statistics are computed based on pulse-treated and chase-treated cells separately, as there are significant differences between the two groups in the data (see for example Figure 3 (B)). Also, to simplify the notation, we use t to denote the dependence of statistics on a specific labelling condition (t_{pulse}, t_{chase}), which we refer to as "labelling time".

1.1.3 Grouping cells with respect to the cell cycle

The original scEU-seq study provides an estimated cell cycle ordering that assigns each cell to a relative cell cycle position along the interval $(0, 1]$ based on the cell cycle reporters. We rely on these estimates and divide the 5422 cells into 5 groups by partitioning this ordering trajectory into 5 sub-intervals $(0, 1/5], (1/5, 2/5], \ldots, (4/5, 1]$. According to the fluorescence of cell cycle reporters (Supplementary Figure S1), we assume that the first two intervals correspond to G1-G1/S phase, the next two intervals correspond to S-S/G2 phase and the last interval corresponds to G2/M phase. We then use the proportion of cells across cell cycle phases to define an empirical *age distribution* (see Figure 3 (B)):

$$\text{age distribution} = \left\{ \frac{N_\theta}{N} \; : \; \theta = \text{G1}, \text{G1/S}, \text{S}, \text{S/G2}, \text{G2/M} \right\} \tag{2}$$

where N_θ denotes the number of cells in cell cycle phase θ and $N = 5422$. Additionally, each cell cycle phase includes cells from different labelling conditions (Supplementary Methods 1.1.2), and therefore we also define age distributions per labelling time, $\left\{\frac{N_{\theta,t}}{N_t}\right\}$, where N_t is the number of cells in condition t and $N_{\theta,t}$ is the number of cells in condition t and cell cycle phase θ.

1.1.4 Gene filtering

We impose certain criteria to select genes for parameter inference and downstream analysis. Starting from 11848 genes of the scEU-seq dataset, we first select genes with at least 500 labelled mRNA counts across all cells (similarly to the original study [25]), which leaves us with a subset of 6087 genes. Then, we impose an additional criterion to filter out genes with very low expression. We require that each gene has at least 1 count on average in at least one cell cycle phase. That is, if $\mathbf{X} = (x_{ij}) = \mathbf{U}+\mathbf{L}$ is the total (unlabelled + labelled) mRNA count matrix (where cells are indexed with i and genes are indexed with j), then the gene-specific condition is that

$$\frac{1}{N_\theta}\sum_{i\in\theta} x_{ij} \geq 1 \text{ for at least one cell cycle phase } \theta \in \{\text{G1}, \text{G1/S}, \text{S}, \text{S/G2}, \text{G2/M}\} \tag{3}$$

This criterion gives us the final set of 3419 genes used for further analysis.

1.1.5 Summary statistics

We now describe how we calculate summary statistics of data. We consider the count matrices $\mathbf{U} = (u_{ij})$, $\mathbf{L} = (l_{ij})$ of unlabelled and labelled mRNA, as well as the count matrix $\mathbf{X} = (x_{ij}) = \mathbf{U}+\mathbf{L}$ of total mRNA. The summary statistics of a gene j in the data depend on the cell cycle phase $\theta \in \{\text{G1, G1/S, S, S/G2, G2/M}\}$ or on the labelling condition $c \in \{c_{pulse}, c_{chase}\}$. The definitions of gene-specific summary statistics are outlined in Table 1. In these definitions, the notation $i \in c$ refers to either all *pulse*-treated cells or all *chase*-treated cells, while the notation $i \in t$ refers to a specific labelling condition (t_{pulse}, t_{chase}), where $t = t_{pulse} + t_{chase}$ denotes the total duration of an experiment (see paragraph 1.1.2).

1.1.6 Standard errors of summary statistics

Aside from computing summary statistics from data, we also estimate standard errors for each statistic to quantify the measurement bias. If $s_{\text{obs}} = s_{\text{obs}}(y)$ is a summary statistic of the data y, we estimate the standard error of s_{obs} with

$$\hat{\text{se}}(s_{\text{obs}}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(s_{\text{obs}}^{(b)} - \frac{1}{B}\sum_{j=1}^{B} s_{\text{obs}}^{(j)}\right)^2} \tag{4}$$

where $s_{\text{obs}}^{(b)}$ is the value of the summary statistic that we obtain from each bootstrapped dataset $y^{(b)}$, $b = 1, \ldots, B = 100$. These standard errors are used during parameter inference to normalise the magnitudes of different summary statistics (see Supplementary Methods 1.5.1).

Table 1: **Definitions of summary statistics.** (A) Theoretical model-based definitions. Dependence on the cell cycle is denoted by τ and dependence on labelling condition is denoted by t. Unlabelled, labelled and total mRNA levels of a single gene are denoted by u, l and $x = u + l$, respectively. (B) Data-based definitions. Notation is explained in paragraph 1.1.5.

	A. Model	**B. Data**
(1)	$\mathbb{E}(x \mid \tau)$	$\bar{x}_j(\theta, c) = \frac{1}{N_{\theta,c}} \sum_{i \in \theta \cap c} x_{ij}$
(2)	$\frac{\mathrm{Var}(x \mid \tau)}{\mathbb{E}(x \mid \tau)}$	$\frac{\frac{1}{N_{\theta,c}-1} \sum_{i \in \theta \cap c} (x_{ij} - \bar{x}_j(\theta, c))^2}{\bar{x}_j(\theta, c)}$
(3)	$\frac{\mathbb{E}_\tau(\mathbb{E}(l \mid \tau, t))}{\mathbb{E}_\tau(\mathbb{E}(u \mid \tau, t) + \mathbb{E}(l \mid \tau, t))}$	$\frac{\frac{1}{N_t} \sum_{i \in t} l_{ij}}{\frac{1}{N_t} \sum_{i \in t} x_{ij}}$
(4)	$\frac{\mathbb{E}_\tau(\mathrm{Cov}(u, l \mid \tau, t))}{\sqrt{\mathrm{Var}(u \mid t)\, \mathrm{Var}(l \mid t)}}$	$\frac{\sum_\theta \frac{N_{\theta,t}}{N_t} \sum_{i \in \theta \cap t} (u_{ij} - \bar{u}_j(\theta, t))(l_{ij} - \bar{l}_j(\theta, t))}{\left(\frac{1}{N_t} \sum_{i \in t} (u_{ij} - \bar{u}_j(t))^2 \frac{1}{N_t} \sum_{i \in t} (l_{ij} - \bar{l}_j(t))^2\right)^{1/2}}$
(5)	$\frac{\mathrm{Cov}_\tau(\mathbb{E}(u \mid \tau, t), \mathbb{E}(l \mid \tau, t))}{\sqrt{\mathrm{Var}(u \mid t)\, \mathrm{Var}(l \mid t)}}$	$\frac{\sum_\theta \frac{N_{\theta,t}}{N_t} \left(\bar{u}_j(\theta, t) - \sum_\theta \frac{N_{\theta,t}}{N_t} \bar{u}_j(\theta, t)\right) \left(\bar{l}_j(\theta, t) - \sum_\theta \frac{N_{\theta,t}}{N_t} \bar{l}_j(\theta, t)\right)}{\left(\frac{1}{N_t} \sum_{i \in t} (u_{ij} - \bar{u}_j(t))^2 \frac{1}{N_t} \sum_{i \in t} (l_{ij} - \bar{l}_j(t))^2\right)^{1/2}}$

Table 2: **Interpretation of summary statistics.**

	Interpretation
(1)	cell cycle-dependent mean of mRNA levels
(2)	cell cycle-dependent Fano factor of mRNA levels
(3)	ratio of mean labelled vs total mRNA levels over labelling time
(4)	cell cycle-averaged correlation of labelled-unlabelled mRNA levels over labelling time
(5)	correlation of cell cycle-dependent labelled & unlabelled means over labelling time

1.2 Stochastic modelling framework

We develop a stochastic model that effectively describes the underlying generative process of mRNA transcription in cells growing and dividing and going through the scEU-seq labelling protocols.

1.2.1 Model of bursty transcription of unlabelled and labelled mRNA

We assume that mRNA molecules are produced according to a telegraph model of gene expression, capturing random switching of the gene promoter between active and inactive states with rates k_{on} and k_{off} respectively, mRNA synthesis with rate α and mRNA decay with rate γ. During pulse-chase labelling, labelled and unlabelled mRNA molecules are synthesised with probability $\lambda\,\alpha$ and $(1-\lambda)\,\alpha$ respectively, where λ denotes the gene-specific labelling efficiency. These are encoded in the reaction scheme:

$$\begin{gathered} G^{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} G^{\text{on}} \\ G^{\text{on}} \xrightarrow{(1-\lambda)\alpha} G^{\text{on}} + U \quad , \quad G^{\text{on}} \xrightarrow{\lambda\alpha} G^{\text{on}} + L \\ U \xrightarrow{\gamma} \varnothing \quad , \quad L \xrightarrow{\gamma} \varnothing \end{gathered} \tag{5}$$

1.2.2 Cell cycle-dependent kinetic rates

We further allow for three of these kinetic rates to vary over the cell cycle. As a result, they are time-dependent periodic functions $k_{\text{on}} = k_{\text{on}}(t), \alpha = \alpha(t), \gamma = \gamma(t)$, for $t^{(n)}_{\text{birth}} \leq t \leq t^{(n)}_{\text{division}}$, $n = 1, 2, \ldots$ (with n representing the nth cell generation in a single-cell lineage). As they are periodic functions, their value does not depend on n and their cell cycle-dependence can be equivalently written as

$$k_{\text{on}} = k_{\text{on}}(\tau), \alpha = \alpha(\tau), \gamma = \gamma(\tau)\,, \quad 0 \leq \tau \leq T_d \tag{6}$$

where τ is the cell age (time along cell cycle progression) and $T_d = t^{(n)}_{\text{division}} - t^{(n)}_{\text{birth}}$ is the interdivision time (cell cycle length). The rate k_{off} is always assumed to be constant. An additional assumption on the kinetics is that the synthesis rate α scales linearly with cell size (except for the constant non-scaling model). This is in accordance with observations from the data S1, which allows us to assume that cells follow linear growth. This implies that the cell size s at age τ is:

$$s(\tau) = s_0 \left(1 + \frac{\tau}{T_d}\right) \tag{7}$$

where s_0 is the cell size at birth. Accordingly, we model the dependence of synthesis rate on cell size as:

$$\alpha(s(\tau), \tau) = \alpha(\tau) \left(1 + \frac{\tau}{T_d}\right) \tag{8}$$

where $\alpha(\tau)$ describes the dependence of the synthesis rate on cell cycle progression without size-scaling. Additionally, we define the burst kinetics as follows:

$$\begin{aligned} \text{burst frequency} &= k_{\text{on}} \\ \text{burst size} &= \frac{\alpha}{k_{\text{off}}} \end{aligned} \tag{9}$$

As a consequence, both burst frequency and burst size can be cell cycle-dependent and additionally the burst size could be size-dependent (except for the constant non-scaling model), in accordance

to the synthesis rate. The unit of all kinetic parameters is h^{-1}, apart from burst size which is unitless. Finally, we define the mRNA half-life $t_{1/2}$ (measured in hours) that corresponds to a constant degradation rate γ as:

$$t_{1/2} = \frac{\ln(2)}{\gamma} \tag{10}$$

1.2.3 Moment equations with time-dependent kinetic rates

The joint probability distribution of molecule numbers $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$ of the respective species $\{G^{\mathrm{on}}, U, L\}$ in the reaction scheme (5) obeys the Chemical Master Equation (CME):

$$\frac{\mathrm{d}}{\mathrm{d}t} P(\mathbf{x}, t) = \sum_{r=1}^{6} \Big(w_r(\mathbf{x} - \mathbf{S}_r, t)\, P(\mathbf{x} - \mathbf{S}_r, t) - w_r(\mathbf{x}, t)\, P(\mathbf{x}, t) \Big) \tag{11}$$

where $\mathbf{S}_r = (\mathbf{S}_{1r}, \mathbf{S}_{2r}, \mathbf{S}_{3r})^{\mathrm{T}}$ denotes the r-th column of the stoichiometric matrix

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \tag{12}$$

of the reaction system (with $\mathbf{S}_{ir}$ being the change in molecule number of species i by reaction r) and $w_r(\mathbf{x}, t)$ denotes the propensity of reaction r, which is the probability per unit time for the r-th reaction to occur. The propensities of the network are

$$\begin{aligned} &w_1(\mathbf{x}, t) = k_{\mathrm{on}}(t)(1 - x_1),\ w_2(\mathbf{x}, t) = k_{\mathrm{off}}(t) x_1,\ w_3(\mathbf{x}, t) = (1 - \lambda(t))\alpha(t) x_1, \\ &w_4(\mathbf{x}, t) = \lambda(t)\alpha(t) x_1,\ w_5(\mathbf{x}, t) = \gamma(t) x_2,\ w_6(\mathbf{x}, t) = \gamma(t) x_3 \end{aligned} \tag{13}$$

Note that the gene promoter state x_1 is either 0 or 1 so it is sufficient to only describe the state of G^{on} in the equations. In our model, we are interested in the time-evolution of the 1st and 2nd order central moments of the molecule abundances, which can be derived using a moment-closure approximation of the CME [37]:

$$\frac{\mathrm{d}}{\mathrm{d}t} \mu_i(t) = \sum_{r=1}^{6} \mathbf{S}_{ir} w_r(\boldsymbol{\mu}, t) \tag{14}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \Sigma_{ij}(t) = \sum_{\substack{r=1,\dots,6 \\ k=1,\dots,4}} \left(\mathbf{S}_{ir} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{kj}(t) + \mathbf{S}_{jr} \frac{\partial w_r(\boldsymbol{\mu}, t)}{\partial \mu_k} \Sigma_{ki}(t) \right) + \sum_{r=1}^{6} \mathbf{S}_{ir} \mathbf{S}_{jr} w_r(\boldsymbol{\mu}, t) \tag{15}$$

where $\mu_i = \mathbb{E}(x_i)$ is the mean molecule number of species i and $\boldsymbol{\Sigma} = (\Sigma_{ij})$ is the covariance matrix with $\Sigma_{ij} = \mathbb{E}\Big((x_i - \mu_i)(x_j - \mu_j)\Big)$ being the covariance of molecule numbers of species i and j and the diagonal elements $\Sigma_{ii} = \mathrm{Var}(x_i)$ being the variance of molecule number of species i. Note that since the reactions in (5) are linear, the propensities are linear functions of the abundances x_i, that is $\mathbb{E}(w_r(\mathbf{x}, t)) = w_r(\mathbb{E}(\mathbf{x}), t) = w_r(\boldsymbol{\mu}, t)$, and therefore the system of moment equations is closed (does not depend on higher order moments). Given our assumption about cell cycle-dependence of the reaction rate constants, the propensities are periodic functions of time. As a consequence, the moment equations are constrained by periodic boundary conditions.

1.2.4 Coupling to cell division dynamics

To determine the boundary conditions, we model the random partitioning of molecules at cell division as a binomial process, which gives us the following conditions for the mean, variance and covariance of unlabelled and labelled mRNA at the n-th cell division:

$$\mu_i\left(t_{\text{birth}}^{(n)}\right) = \frac{\mu_i\left(t_{\text{division}}^{(n-1)}\right)}{2} \quad , \quad i = 2,3 \tag{16}$$

$$\Sigma_{ij}\left(t_{\text{birth}}^{(n)}\right) = \frac{\Sigma_{ij}\left(t_{\text{division}}^{(n-1)}\right)}{4} + \delta_{ij}\frac{\mu_i\left(t_{\text{division}}^{(n-1)}\right)}{4} \quad , \quad i,j = 2,3 \tag{17}$$

The gene promoter state $g^{(\text{birth})}$ in a daughter cell at birth is identical to the promoter state $g^{(\text{division})}$ in the mother cell at division, and therefore the mean and variance of the promoter state remain constant before and after division:

$$\mu_1\left(t_{\text{birth}}^{(n)}\right) = \mu_1\left(t_{\text{division}}^{(n-1)}\right) \quad , \quad \Sigma_{11}\left(t_{\text{birth}}^{(n)}\right) = \Sigma_{11}\left(t_{\text{division}}^{(n-1)}\right) \tag{18}$$

However, the covariance of the promoter state with the mRNA decreases due to partitioning of mRNA molecules:

$$\Sigma_{12}\left(t_{\text{birth}}^{(n)}\right) = \frac{\Sigma_{12}\left(t_{\text{division}}^{(n-1)}\right)}{2} \quad , \quad \Sigma_{13}\left(t_{\text{birth}}^{(n)}\right) = \frac{\Sigma_{13}\left(t_{\text{division}}^{(n-1)}\right)}{2} \tag{19}$$

To prove the above, for notation simplicity, m denotes mRNA (unlabelled or labelled) and g denotes the gene promoter state. Also, m_0, g_0 denote the mRNA and gene states in a daughter cell at birth and m_T, g_T denote the mRNA and gene states in the mother cell at cell division. At birth, a daughter cell inherits m_0 mRNA molecules out of total m_T molecules of the mother cell according to a Binomial distribution:

$$P(m_0 \mid m_T) = \binom{m_T}{m_0} p^{m_0}(1-p)^{m_T - m_0} \tag{20}$$

with $p = 1/2$. Also, the gene promoter state is not altered at cell division and therefore:

$$g_0 \mid g_T \sim \delta(g_T) \tag{21}$$

Using the law of total expectation, the expected number of mRNA at cell birth is:

$$\mathbb{E}(m_0) = \mathbb{E}(\mathbb{E}(m_0 \mid m_T)) = \frac{\mathbb{E}(m_T)}{2} \tag{22}$$

Using the law of total variance, the variance of mRNA molecules at birth is:

$$\begin{aligned}
\text{Var}(m_0) &= \mathbb{E}(\text{Var}(m_0 \mid m_T)) + \text{Var}(\mathbb{E}(m_0 \mid m_T)) && \text{(23a)}\\
&= \mathbb{E}\left(\frac{m_T}{4}\right) + \text{Var}\left(\frac{m_T}{2}\right) && \text{(23b)}\\
&= \frac{1}{4}(\mathbb{E}(m_T) + \text{Var}(m_T)) && \text{(23c)}
\end{aligned}$$

Denoting with u, l the unlabelled and labelled mRNA and using the law of total covariance, we have that the covariance of unlabelled and labelled mRNA at birth is:

$$\begin{aligned}
\mathrm{Cov}(u_0, l_0) &= \mathbb{E}(\mathrm{Cov}(u_0, l_0 \mid u_T, l_T)) + \mathrm{Cov}(\mathbb{E}(u_0 \mid u_T, l_T), \mathbb{E}(l_0 \mid u_T, l_T)) && \text{(24a)}\\
&= 0 + \mathrm{Cov}(\mathbb{E}(u_0 \mid u_T), \mathbb{E}(l_0 \mid l_T)) && \text{(24b)}\\
&= \mathrm{Cov}\left(\frac{u_T}{2}, \frac{l_T}{2}\right) && \text{(24c)}\\
&= \frac{1}{4}\mathrm{Cov}(u_T, l_T) && \text{(24d)}
\end{aligned}$$

where in (24b) we use that the conditional variables $u_0 \mid u_T$ and $l_0 \mid l_T$ are independent.

The covariance of the mRNA and gene states at birth is given by:

$$\begin{aligned}
\mathrm{Cov}(m_0, g_0) &= \mathbb{E}(\mathrm{Cov}(m_0, g_0 \mid m_T, g_T)) + \mathrm{Cov}(\mathbb{E}(m_0 \mid m_T, g_T), \mathbb{E}(g_0 \mid m_T, g_T)) && \text{(25a)}\\
&= 0 + \mathrm{Cov}(\mathbb{E}(m_0 \mid m_T), \mathbb{E}(g_0 \mid g_T)) && \text{(25b)}\\
&= \mathrm{Cov}\left(\frac{m_T}{2}, g_T\right) && \text{(25c)}\\
&= \mathbb{E}\left(\frac{m_T}{2} g_T\right) - \mathbb{E}\left(\frac{m_T}{2}\right)\mathbb{E}(g_T) && \text{(25d)}\\
&= \frac{1}{2}\mathrm{Cov}(m_T, g_T) && \text{(25e)}
\end{aligned}$$

where in (25b) we use that the conditional variables $m_0 \mid m_T$ and $g_0 \mid g_T$ are independent.

1.2.5 Description of a single realisation of the model

A single model realisation depends on a gene-specific parameter set $\{k_{\text{off}}, k_{\text{on}}(t), \alpha(t), \gamma(t), \lambda\}$, a cell age τ and a labelling condition (t_{pulse}, t_{chase}). We solve the moment equations numerically over time for multiple consecutive cell divisions, following a single-cell lineage up until the sequencing time-point t_{seq} (see example model realisations in Figure 1 (C) and Supplementary Figure S2). We additionally assume that cells have a deterministic cell cycle length, $T_d = 20$ hours. First, the equations are solved for multiple cell cycles until convergence to a steady state. The convergence criterion is that the relative change of all trajectories at the end of two consecutive cell cycles needs to be smaller than 0.01%. After reaching convergence, we solve the model for additional time equal to $3\,T_d + \tau$, up until the final time-point t_{seq} (observation or sequencing time-point). To model the effect of pulse labelling on mRNA synthesis, the labelling efficiency $\lambda(t)$ is set to its constant parameter value λ for time t_{pulse}:

$$\lambda(t) = \begin{cases} \lambda & \text{if } t_{seq} - t_{pulse} - t_{chase} \leq t < t_{seq} - t_{chase} \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

The endpoints $\mathbb{E}(u), \mathbb{E}(l), \mathrm{Var}(u), \mathrm{Var}(l), \mathrm{Cov}(u, l)$ of the trajectories at time t_{seq} comprise the model estimates, which are conditional on cell age τ and pulse-chase labelling time (t_{pulse}, t_{chase}).

1.3 Capture efficiency estimation and downsampling of moments

To take into account technical effects of sequencing in our modelling framework, we correct the model outputs using estimates of cell-specific capture efficiencies. Consider the count matrices

$\mathbf{U} = (u_{ij})$ and $\mathbf{L} = (l_{ij})$ of unlabelled and labelled mRNA, as well as the total mRNA count matrix $\mathbf{X} = (x_{ij}) = \mathbf{U} + \mathbf{L}$, where the indices $i = 1, ..., n$ represent cells and the indices $j = 1, ..., d$ represent genes. If $u = u_{ij}$ and $l = l_{ij}$ are a pair of observations for a cell i and gene j, then due to low transcript capture of the sequencing protocol, these observations are random samples drawn from Binomial distributions, $u \sim \text{Binomial}(\tilde{u}, \beta_i)$ and $l \sim \text{Binomial}(\tilde{l}, \beta_i)$, where $\tilde{u}, \tilde{l}$ are the mRNA counts we would observe in a sequencing experiment with 100% capture efficiency and β_i is the capture efficiency of cell i. Below, we outline the method for estimating the capture efficiency distribution directly from the data and provide formulas for the corrected (downsampled) moments of u and l.

1.3.1 Estimation of cell-specific capture efficiencies from scEU-seq data

The estimation of β_i is based on the assumption that cells in the same cell cycle phase have on average the same cell size. We use the total mRNA abundance $s_i = \sum_{j=1}^{d} x_{ij}$ of cell i of as a proxy for its size, while the average total mRNA count $s_\theta = \frac{1}{N_\theta} \sum_{\substack{i \in \theta \\ j=1,...,d}} x_{ij}$ of cells in cell cycle phase θ represents the characteristic cell size in θ (N_θ is the number of cells in θ). Based on the above, we can estimate:

$$\hat{\beta}_i = \bar{\beta} \frac{s_i}{s_\theta} \tag{27}$$

where $\bar{\beta}$ is an estimate of the average capture efficiency of the sequencing experiment.

An additional factor that we take into account in the above estimation is that cells treated in pulse conditions have a different capture efficiency than cells treated in chase conditions (see Figure 3 (B) and Supplementary Methods 1.1.2). For that reason, we consider different average capture efficiencies $\bar{\beta}_{\text{pulse}}, \bar{\beta}_{\text{chase}}$ for pulse-treated and chase-treated cells, and the normalising factors s_θ for each cell cycle phase are evaluated independently for pulse and chase cells. The estimated capture efficiency $\hat{\beta}_i$ of cell i is then given by:

$$\hat{\beta}_i = \bar{\beta}_c \frac{s_i}{s_{c,\theta}} \tag{28}$$

where $\bar{\beta}_c = \begin{cases} 0.1 & \text{if } c = c_{pulse} \\ 0.2 & \text{if } c = c_{chase} \end{cases}$ is the average capture efficiency of pulse/chase experiments and $s_{c,\theta} = \frac{1}{N_{c,\theta}} \sum_{\substack{i \in c \cap \theta \\ j=1,...,d}} x_{ij}$ is the average total count of cells in condition $c \in \{c_{pulse}, c_{chase}\}$ cell cycle phase θ. The ratio of the values we have assumed for $\bar{\beta}_c$ (capture efficiency of chase being twice the capture efficiency of pulse) is inferred from the data, while the absolute scale of 10% capture efficiency is an estimate that is reasonable for scRNA-seq protocols [46], but it is not typically directly measured.

1.3.2 Downsampling the mean, variance and covariance

We can now use the estimated capture efficiencies to incorporate the effect of Binomial downsampling of data in the moments. For a cell with capture efficiency β, the downsampled mean mRNA level is given by:

$$\mathbb{E}(x) = \mathbb{E}(\beta)\, \mathbb{E}(\tilde{x}) \tag{29}$$

where $\tilde{x}$ is the mRNA levels in the absence of technical noise. The downsampled variance is given by:

$$\mathrm{Var}(x) = \mathbb{E}(\beta(1-\beta))\,\mathbb{E}(\tilde{x}) + \mathrm{Var}(\beta)\,\mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2\,\mathrm{Var}(\tilde{x}) \tag{30}$$

The formulas (29),(30) hold for both unlabelled and labelled mRNA.

Finally, the downsampled covariance of unlabelled and labelled mRNA is given by:

$$\mathrm{Cov}(u,l) = \ \mathrm{Var}(\beta)\,\mathbb{E}(\,\tilde{u}\,\tilde{l}\,) + \mathbb{E}(\beta)^2\mathrm{Cov}(\,\tilde{u},\tilde{l}\,) \tag{31}$$

To prove the above, we use that an observed mRNA count $x_{ij} = x$ is a random variable $x \sim \mathrm{Binomial}(\tilde{x},\beta)$ where $\beta = \beta_i$ is the capture efficiency of cell i and $\tilde{x}$ is the observation in the absence of technical errors. The downsampled mean is given by:

$$\mathbb{E}(x) = \mathbb{E}\left(\mathbb{E}\left(x \mid \tilde{x},\beta\right)\right) = \mathbb{E}\left(\beta\,\tilde{x}\right) = \mathbb{E}(\beta)\,\mathbb{E}(\tilde{x}) \tag{32}$$

and the downsampled variance is given by:

$$\begin{aligned}
\mathrm{Var}(x) &= \mathbb{E}\left(\mathrm{Var}\left(x \mid \tilde{x},\beta\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(x \mid \tilde{x},\beta\right)\right) && \text{(33a)}\\
&= \mathbb{E}(\beta(1-\beta)\,\tilde{x}) \ + \mathbb{E}\left(\beta^2\tilde{x}^2\right) - \mathbb{E}\left(\beta\tilde{x}\right)^2 && \text{(33b)}\\
&= \mathbb{E}(\beta(1-\beta))\,\mathbb{E}(\tilde{x}) + \mathbb{E}(\beta^2)\,\mathbb{E}(\tilde{x}^2) - \mathbb{E}(\beta)^2\,\mathbb{E}(\tilde{x})^2 && \text{(33c)}\\
&= \mathbb{E}(\beta(1-\beta))\,\mathbb{E}(\tilde{x}) + \mathbb{E}(\beta^2)\,\mathbb{E}(\tilde{x}^2) - \mathbb{E}(\beta)^2\,\mathbb{E}(\tilde{x})^2 - \mathbb{E}(\beta)^2\,\mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2\,\mathbb{E}(\tilde{x}^2) && \text{(33d)}\\
&= \mathbb{E}(\beta(1-\beta))\,\mathbb{E}(\tilde{x}) + \mathrm{Var}(\beta)\,\mathbb{E}(\tilde{x}^2) + \mathbb{E}(\beta)^2\,\mathrm{Var}(\tilde{x}) && \text{(33e)}
\end{aligned}$$

Finally, the downsampled covariance of unlabelled and labelled mRNA is given by:

$$\begin{aligned}
\mathrm{Cov}(u,l) &= \mathbb{E}\left(\mathrm{Cov}(u,l \mid \tilde{u},\tilde{l},\beta)\right) + \mathrm{Cov}\left(\mathbb{E}(u \mid \tilde{u},\beta),\ \mathbb{E}(\,l \mid \tilde{l},\beta\,)\right) && \text{(34a)}\\
&= 0 + \mathrm{Cov}\left(\beta\,\tilde{u},\ \beta\,\tilde{l}\right) && \text{(34b)}\\
&= \mathrm{Var}(\beta)\,\mathbb{E}(\,\tilde{u}\,\tilde{l}\,) + \mathbb{E}(\beta)^2\mathrm{Cov}(\,\tilde{u},\tilde{l}\,) && \text{(34c)}
\end{aligned}$$

1.4 Time-dependent summary statistics generated from the model

We have demonstrated how our model generates 1st and 2nd order moments that are dependent on cell age τ and labelling time t (Supplementary Methods 1.2.5), as well as how we correct them to account for technical noise (Supplementary Methods 1.3.2). Below, we describe how we further use these moments to construct a collection of summary statistics and extract interpretable information about bursty transcription dynamics.

1.4.1 Cell cycle-dependent mean and noise statistics

First, we construct summary statistics that describe cell cycle-dependence of mean and noise in transcript levels. Given the means of unlabelled and labelled mRNA $\mathbb{E}(u \mid \tau,t), \mathbb{E}(l \mid \tau,t)$, we compute the mean of total transcript levels $x = u + l$:

$$\mathbb{E}(x \mid \tau) = \mathbb{E}(u \mid \tau,t) + \mathbb{E}(l \mid \tau,t) \tag{35}$$

Note that this mean does not depend on labelling condition t. Similarly, we compute the variance of total transcripts 2nd order moments of unlabelled and labelled mRNA:

$$\mathrm{Var}(x \mid \tau) = \mathrm{Var}(u \mid \tau, t) + \mathrm{Var}(l \mid \tau, t) + 2\,\mathrm{Cov}(u, l \mid \tau, t) \tag{36}$$

as well as the Fano factor:

$$\mathrm{FF}(x \mid \tau) = \frac{\mathrm{Var}(x \mid \tau)}{\mathbb{E}(x \mid \tau)} \tag{37}$$

These summary statistics describe temporal changes of transcription dynamics along the cell cycle that are independent of variations due to labelling.

1.4.2 Labelled vs total mRNA ratios capture the timescales of RNA kinetics

Next, we generate summary statistics that aim to extract information about the timescales of bursty transcription and degradation from the various pulse and chase labelling conditions. We first construct a statistic that describes how the proportion of labelled mRNA changes with labelling time:

$$\mathrm{Ratio}(t) = \frac{\mathbb{E}(l \mid t)}{\mathbb{E}(u + l \mid t)} \tag{38}$$

To integrate out the cell cycle-dependence we use empirical age distributions from the data, which depend on the labelling condition (Supplementary Methods 1.1.3):

$$\mathbb{E}(u \mid t) = \mathbb{E}_\tau(\mathbb{E}(u \mid \tau, t)) \quad , \quad \mathbb{E}(l \mid t) = \mathbb{E}_\tau(\mathbb{E}(l \mid \tau, t)) \tag{39}$$

1.4.3 Labelled-unlabelled mRNA correlations capture sources of transcriptional noise

Further, we construct summary statistics that describe the covariation of labelled and unlabelled mRNA. The law of total covariance gives:

$$\underbrace{\mathrm{Cov}(u, l)}_{\text{total noise}} = \underbrace{\mathbb{E}_\tau(\mathrm{Cov}(u, l) \mid \tau)}_{\text{upstream bursty promoter noise}} + \underbrace{\mathrm{Cov}_\tau\left(\mathbb{E}(u \mid \tau), \mathbb{E}(l \mid \tau)\right)}_{\text{cell cycle variation}} \tag{40}$$

which shows that the unlabelled-labelled RNA modalities and the cell cycle reporter provided by the scEU-seq protocol can be used to decompose the overall noise into two contributions: upstream noise induced by the bursty gene promoter and noise due to dependence on the cell cycle (see also Figure 3 (A)). To exploit this information, our summary statistics include two correlation coefficients - corresponding to the two noise components - that contribute to the total correlation of unlabelled-labelled mRNA:

$$\underbrace{\rho_{u,l}(t)}_{\text{total correlation}} = \underbrace{\frac{\mathbb{E}_\tau(\mathrm{Cov}(u, l) \mid \tau, t)}{\sqrt{\mathrm{Var}(u \mid t)\,\mathrm{Var}(l \mid t)}}}_{\text{correlation coefficient 1}} + \underbrace{\frac{\mathrm{Cov}_\tau\left(\mathbb{E}(u \mid \tau, t), \mathbb{E}(l \mid \tau, t)\right)}{\sqrt{\mathrm{Var}(u \mid t)\,\mathrm{Var}(l \mid t)}}}_{\text{correlation coefficient 2}} \tag{41}$$

Similarly to the ratios (Supplementary Methods 1.4.2), we compute labelling-specific variances and covariances by integrating over the labelling-specific age distributions. Note that the total variances in the denominators of (41) are given by the law of total variance:

$$\mathrm{Var}(x \mid t) = \mathbb{E}_\tau(\mathrm{Var}(x) \mid \tau, t) + \mathrm{Var}_\tau(\mathbb{E}(x \mid \tau, t)) \tag{42}$$

1.5 Approximate Bayesian inference framework

In this section, we outline the details of our inference framework. As our moment-based model does not provide us with an analytical likelihood, relying on numerical solutions of our model, we employ an Approximate Bayesian approach for estimating parameters, where generated summary statistics from the model are compared with summary statistics of the data.

1.5.1 ABC rejection sampling for parameter inference

For each gene j in the data, we consider its associated collection of summary statistics $s_{\text{obs}}^{(j)}$ (as described in Supplementary Methods 1.1.5, 1.1.6). The scheme for estimating gene-specific parameters is the following:

For each model index $m = 1, \ldots, 5$ and for $i = 1, \ldots, M$,

- Sample a parameter set $\theta^{(i)} = \theta^{(i)}(m)$ from the prior $\pi(\theta)$.
- Generate synthetic summary statistics $s^{(i)} = s^{(i)}(\theta^{(i)})$ using the model.
- Compute distances $d(s^{(i)}, s_{\text{obs}}^{(j)})$ for all $j = 1, \ldots, N$.
- Accept parameter set $\theta^{(i)}$ for all $j = 1, \ldots, N$ that satisfy $d(s^{(i)}, s_{\text{obs}}^{(j)}) \leq \varepsilon$, where ε is a predetermined acceptance threshold.

For each $j = 1, \ldots, N$, the collection of accepted parameter sets $\{\Theta^{(j)}(m) : m = 1, \ldots, 5\}$ comprises an approximate (joint) posterior distribution over parameters and models for gene j. After model selection (see Supplementary Methods 1.5.2), we choose the best model $m^\star = m^\star(j)$ for each gene j and we use the particle $\theta^{(i)}(m^\star(j))$ that achieved the minimum distance $d(s^{(i)}, s_{\text{obs}}^{(j)})^\star$ between model and data as the gene-specific point estimate to inform the model for downstream analyses.

To determine the prior distribution $\pi(\theta)$, we sample parameters from the following log-uniform distributions: $\log_{10}(k_{\text{on}}) \sim \mathcal{U}(-3, 3), \log_{10}(k_{\text{off}}) \sim \mathcal{U}(-3, 3), \log_{10}(\alpha) \sim \mathcal{U}(-3, 3), \log_{10}(\gamma) \sim \mathcal{U}(-3, 2)$ and $\log_{10}(\lambda) \sim \mathcal{U}(-0.7, 0)$. We sample $M = 10^6$ parameter sets for each model m (Figure 2(A)) so we run in total $5 \cdot 10^6$ simulations (this is the computationally expensive step but is independent of the specific data for individual genes hence making our method scalable). For each sampled parameter set $\theta^{(i)} = \theta^{(i)}(m)$, we generate summary statistics $s^{(i)}(\theta^{(i)})$ from the model m according to the procedure described in Supplementary Methods 1.3, 1.4, which we then compare to all $N = 3419$ gene-specific sets of summary statistics. The distance between a synthetic and an observed collection of summary statistics, $s = \{s_1, \ldots, s_n\}$ and $s_{\text{obs}} = \{s_{\text{obs},1}, \ldots, s_{\text{obs},n}\}$, is determined by the following function:

$$d(s, s_{\text{obs}}) = \frac{1}{n} \sum_{k=1}^{n} \frac{(s_k - s_{\text{obs},k})^2}{\hat{\text{se}}(s_{\text{obs}})^2 + \sigma^2 s_{\text{obs}}^2} \tag{43}$$

where $n = 53$, $\hat{\text{se}}(s_{\text{obs}})$ is the standard error of the summary statistic due to sampling noise (Supplementary Methods 1.1.6) and $\sigma = 0.1$ models other technical variations.

1.5.2 ABC-based model selection

As we are testing multiple modelling hypotheses against all genes in the data (Figure 1 (C), Figure 2 (A)), we wish to assign probabilities to these hypotheses and decide which model is most likely to describe each gene. To do that, we use an ABC-based model selection approach [59], which involves two decision steps. The first one is to select whether a gene is constant or non-constant (see Figure 2 (A), Figure 2 (B)). Given this decision, the second step is to select a specific constant model (scaling, non-scaling) or a specific non-constant model (varying burst frequency, varying burst size or varying decay rate). Let H_1, H_2 denote the constant and non-constant model hypotheses and let $\mathcal{D}$ denote the data. As we have in total two constant models and three non-constant models, we set the prior probabilities of the two hypotheses to be $\frac{\pi(H_1)}{\pi(H_2)} = \frac{2}{3}$. The posterior probability of model H_i is approximated by:

$$p(H_i \mid \mathcal{D}) = \frac{N_{i,\varepsilon}}{N_\varepsilon} \tag{44}$$

where $N_{i,\varepsilon}$ is the number of particles accepted for model H_i and N_ε is the total number of particles accepted for all models, given the acceptance threshold ε. In this case, $N_{1,\varepsilon}$ includes the number of particles accepted for both constant models, while $N_{2,\varepsilon}$ includes the number of particles accepted for all three non-constant models. Model comparison is then based on the (approximate) Bayes factor

$$B_{1,2} = \frac{p(H_1 \mid \mathcal{D})\, \pi(H_2)}{p(H_2 \mid \mathcal{D})\, \pi(H_1)} \tag{45}$$

We then need a criterion for thresholding this Bayes factor to decide whether we have sufficient evidence in favour of H_1 over H_2, to a certain level of confidence. To do that we use bootstrapping of the accepted particles for each model, and estimate confidence intervals $(L_{H_i|\mathcal{D}}, U_{H_i|\mathcal{D}})$ for each posterior probability, so that

$$\mathbb{P}\left(L_{H_i|\mathcal{D}} \leq p(H_i \mid \mathcal{D}) \leq U_{H_i|\mathcal{D}}\right) = 1 - \alpha \tag{46}$$

with $\alpha = 0.05$. For model H_1 to be selected versus model H_2 it is required that

$$p(H_1 \mid \mathcal{D}) > \frac{\pi(H_1)}{\pi(H_2)}\, p(H_2 \mid \mathcal{D}) \tag{47}$$

or equivalently, since $p(H_2 \mid \mathcal{D}) = 1 - p(H_1 \mid \mathcal{D})$, we need

$$p(H_1 \mid \mathcal{D}) > \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \tag{48}$$

To strengthen this criterion, we require

$$L_{H_1|\mathcal{D}} > \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \tag{49}$$

Equivalently, model M_2 is selected if

$$U_{H_1|\mathcal{D}} \leq \frac{\frac{\pi(H_1)}{\pi(H_2)}}{1 + \frac{\pi(H_1)}{\pi(H_2)}} \tag{50}$$

After selecting between the constant and non-constant hypotheses, we proceed to the second step of model selection, where we follow a similar approach. If the constant hypothesis is selected,

we now decide between the scaling H_1 and non-scaling H_2 hypotheses, with probabilities given by (44), where now $N_{1,\varepsilon}, N_{2,\varepsilon}$ are the number of accepted particles of the scaling and the non-scaling model respectively and $N_\varepsilon = N_{1,\varepsilon} + N_{2,\varepsilon}$. If the non-constant hypothesis is selected, we decide between the varying burst frequency H_1, varying burst size H_2 and varying decay rate H_3 hypotheses. In that case, the choice criterion is formed as follows. The model probabilities $p(H_i \mid \mathcal{D})$ are given by (44), where now $N_{i,\varepsilon}$ is the number of accepted particles of each non-constant model and $N_\varepsilon = N_{1,\varepsilon} + N_{2,\varepsilon} + N_{3,\varepsilon}$. If $(L_{H_i|\mathcal{D}}, U_{H_i|\mathcal{D}})$ are the corresponding confidence intervals from bootstrapping, we select the non-constant model $i \in \{1, 2, 3\}$ if

$$L_{H_i|\mathcal{D}} \geq U_{H_j|\mathcal{D}} \quad , \quad \text{for all } j \neq i \tag{51}$$

1.6 Noise decomposition

Let x be the mRNA count of a gene and u, l the corresponding unlabelled and labelled modalities, so that $x = u + l$. Also, let τ denote the cell age and t the labelling time (based on experimental condition). Then, using the conditional moments of the model output, we compute the mean transcript levels as:

$$\mathbb{E}(x) = \mathbb{E}_{\tau,t}(\mathbb{E}(x \mid \tau, t)) \tag{52}$$

Then, the total variance of x can be decomposed as follows:

$$\mathrm{Var}(x) = \underbrace{\mathbb{E}_{\tau,t}(\mathrm{Var}(x \mid \tau, t))}_{\text{intrinsic - reaction kinetics noise}} + \underbrace{\mathrm{Var}_{\tau,t}(\mathbb{E}(x \mid \tau, t))}_{\text{extrinsic - temporal dependence noise}} \tag{53}$$

The intrinsic noise component can be further decomposed as follows:

$$\underbrace{\mathbb{E}_{\tau,t}(\mathrm{Var}(x \mid \tau, t))}_{\text{intrinsic noise}} = \underbrace{\mathbb{E}_{\tau,t}(\mathrm{Var}(u \mid \tau, t)) + \mathbb{E}_{\tau,t}(\mathrm{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2\,\mathbb{E}_{\tau,t}(\mathrm{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} \tag{54}$$

where the transcription and decay term includes noise due to random partitioning at cell division. The extrinsic noise component is explained by variations due to cell cycle-dependence and labelling time:

$$\underbrace{\mathrm{Var}_{\tau,t}(\mathbb{E}(x \mid \tau, t))}_{\text{extrinsic noise}} = \mathrm{Var}_{\tau,t}(\mathbb{E}(\mathbb{E}(x \mid \tau, t) \mid \tau)) + \mathbb{E}_{\tau,t}(\mathrm{Var}(\mathbb{E}(x \mid \tau, t) \mid \tau)) \tag{55a}$$

$$= \underbrace{\mathrm{Var}_{\tau}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}} + \underbrace{\mathbb{E}_{\tau,t}(\mathrm{Var}(\mathbb{E}(x \mid \tau, t) \mid \tau))}_{\text{labelling time-dependence noise}} \tag{55b}$$

Computing the labelling time-dependence component from the model yields a negligible, practically zero contribution and therefore we can neglect it from the formulas.

We finally obtain the following total variance decomposition:

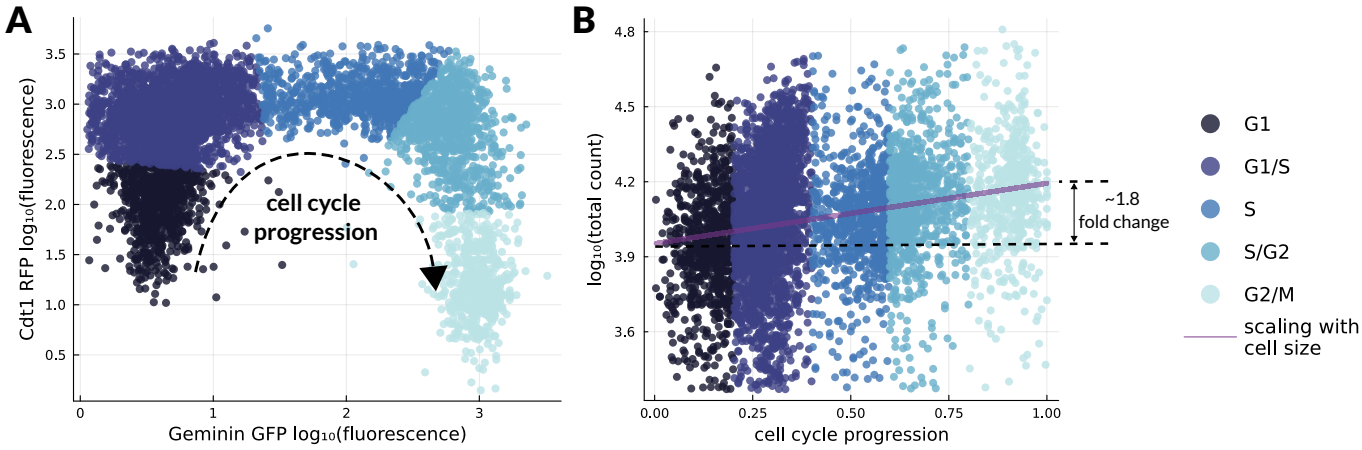
$$\mathrm{Var}(x) = \underbrace{\mathbb{E}_{\tau,t}(\mathrm{Var}(u \mid \tau, t) + \mathrm{Var}(l \mid \tau, t))}_{\text{transcription and decay noise}} + \underbrace{2\,\mathbb{E}_{\tau,t}(\mathrm{Cov}(u, l \mid \tau, t))}_{\text{bursty promoter noise}} + \underbrace{\mathrm{Var}_{\tau}(\mathbb{E}(x \mid \tau))}_{\text{cell cycle variation}} \tag{56}$$

Dividing the whole expression in (56) by the squared total mean $\mathbb{E}(x)^2$, we can obtain the formula for decomposing the total noise $\mathrm{CV}^2_{\text{total}}(x)$. It is important to note that we use the same formula for decomposing the observed total variance in the data and for quantifying cell cycle variation (Figure 4 (E)), however in that case the formula is not exact as it neglects technical noise contributions.

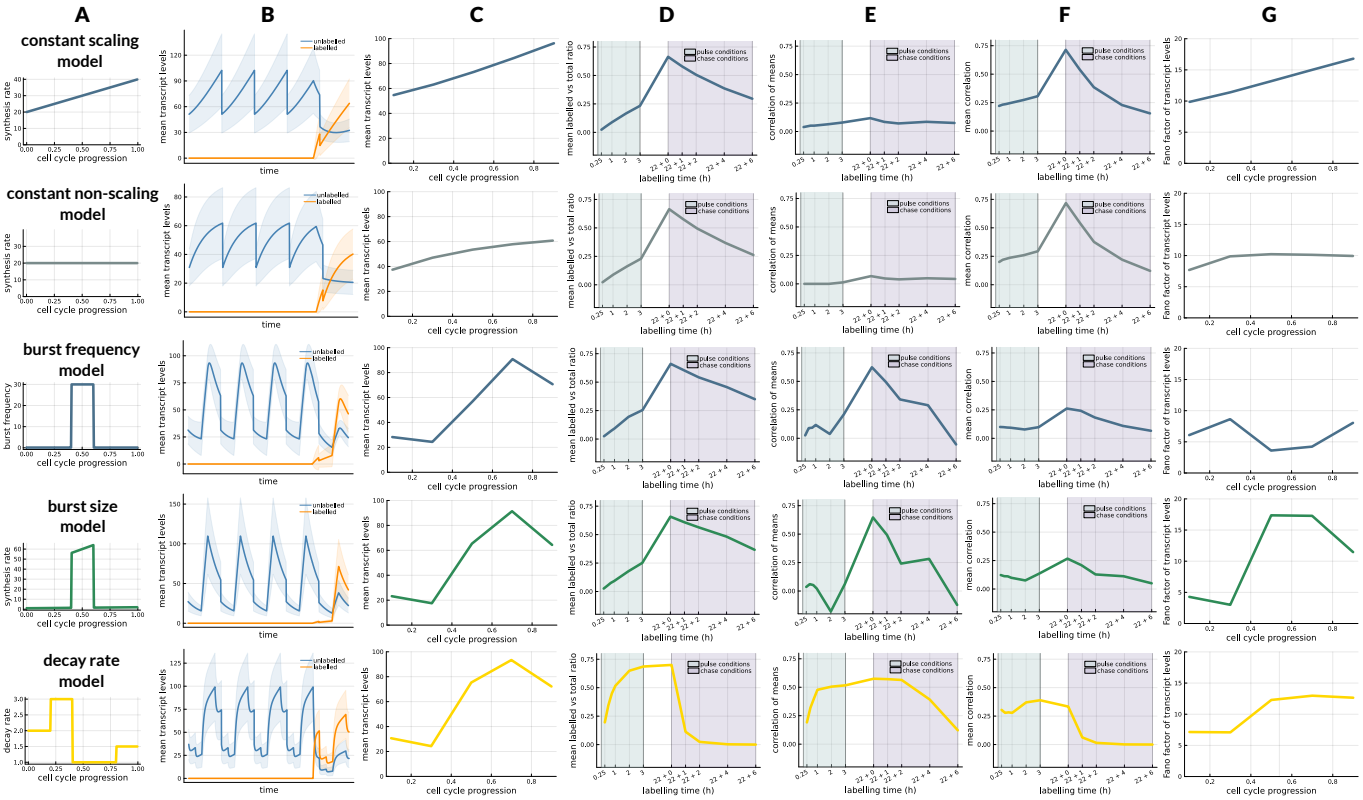
Code availability

Julia code used for modelling, inference and data analysis is available at https://github.com/pthomaslab/abc_inference_transcription.

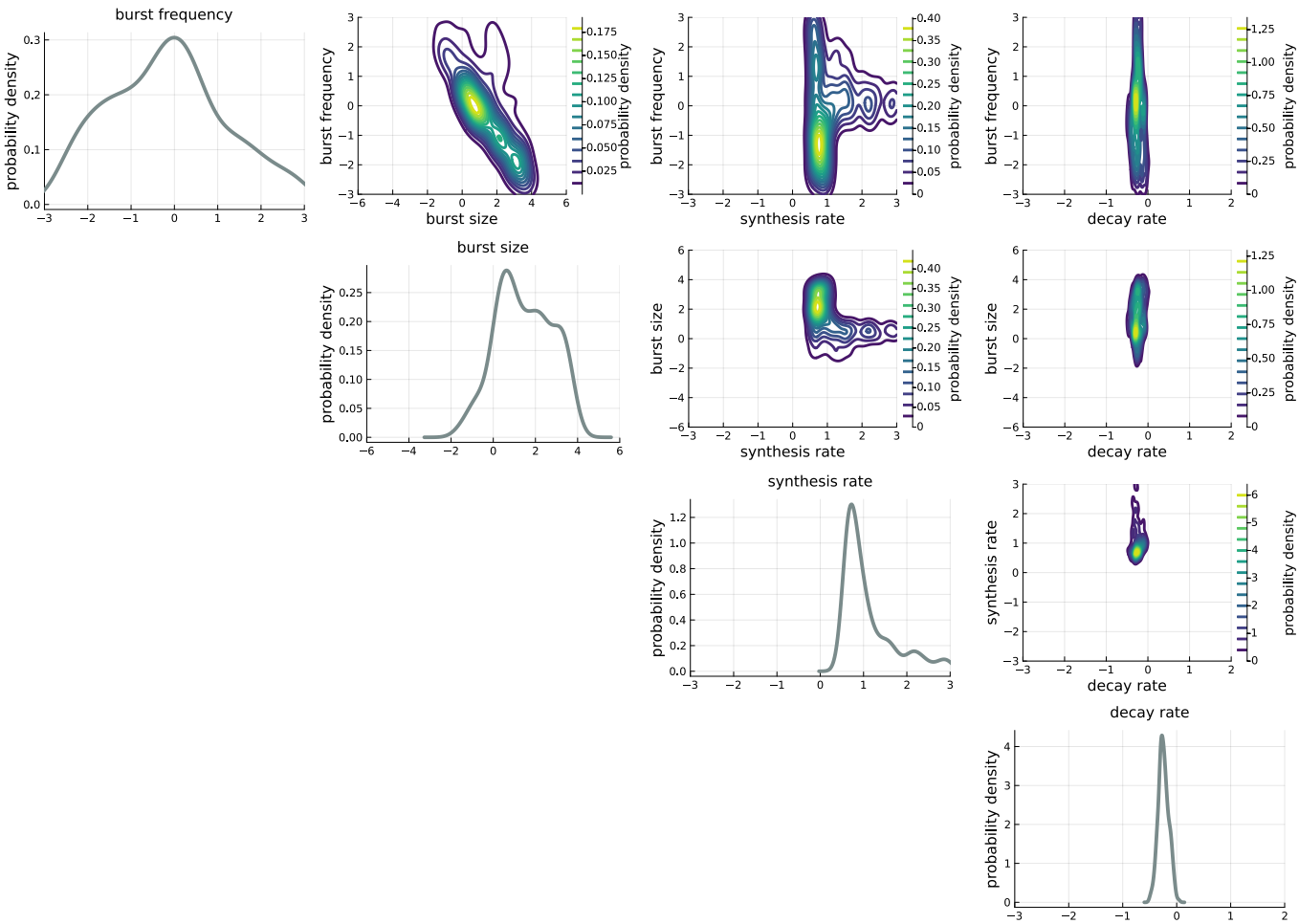
Supplementary Figures



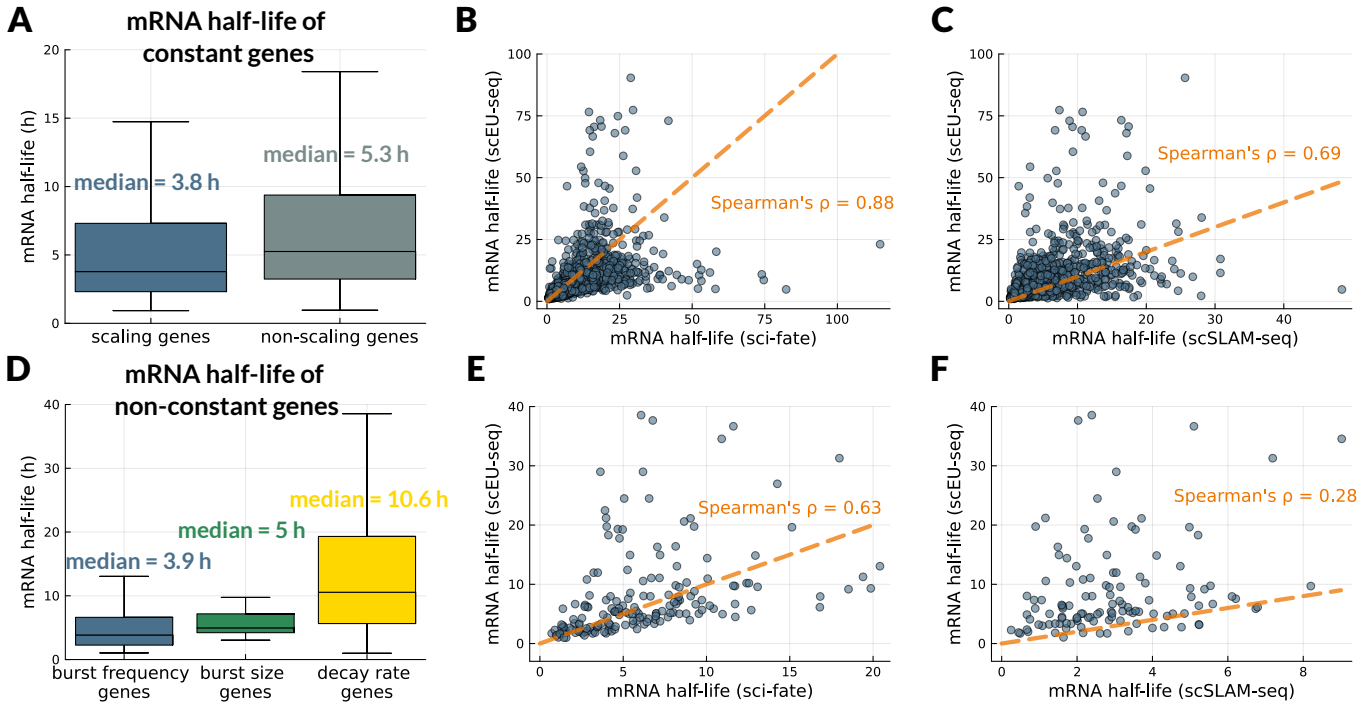
Supplementary Figure S1: **Definition of cell cycle phases and observed scaling of total count with cell size.** (A) Scatter plot of log-fluorescence of the cell cycle markers Geminin-GFP and Cdt1-RFP as reported in the original scEU-seq study [25]. The colours show our grouping of cells in cell cycle phases (see also Supplementary Methods 1.1.3). (B) Observed cell-specific total mRNA counts along cell cycle progression, with respect to the cell cycle trajectory reported in [25]. The purple line shows a linear regression fit and the almost 2-fold increase indicates the scaling of total count with cell size.



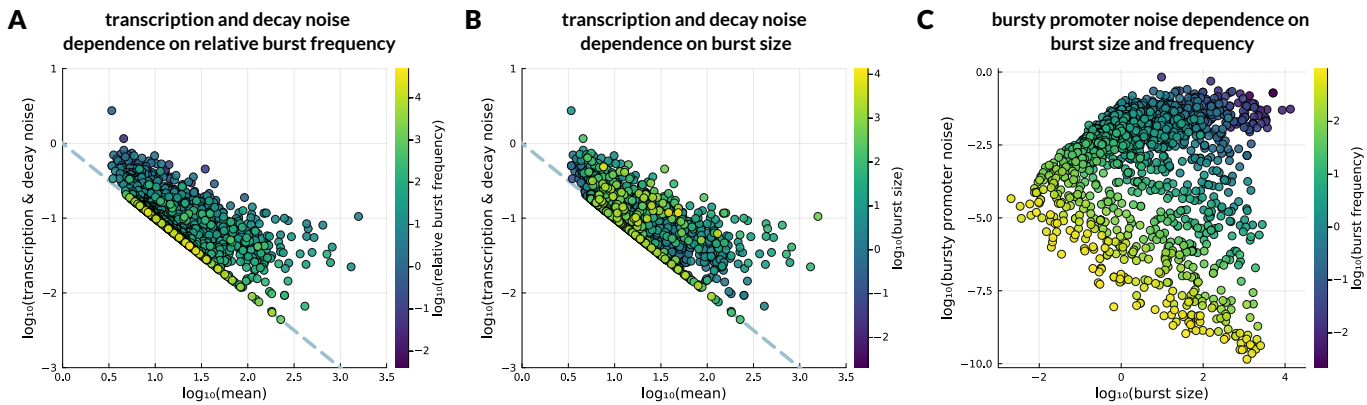
Supplementary Figure S2: **Comparison of summary statistics arising from the different models.** (A) Labels of the different models and associated regulation of kinetics (see also Figure 2 (A)). (B) Model realisations for each parameter setup: Mean unlabelled and labelled transcript levels across a single-cell lineage over time, where the shaded region corresponds to the standard deviation of transcript levels. The 1st realisation (row 1) assumes that all kinetic rates are constant along the cell cycle, while the synthesis rate scales linearly with cell size. The 2nd realisation (row 2) assumes identical parameters with the 1st realisation, with the only difference being that the synthesis rate does not depend on cell size. The next 3 realisations (rows 3-5) assume time-varying burst frequency, synthesis rate or decay rate and scaling of synthesis rate with cell size, while the rest of the kinetic rates are constant. The chosen parameter sets in rows 3-5 are such that we obtain almost identical patterns of cell cycle-dependent mean expression (column (C)). (C-G) Summary statistics generated by each model using the cell cycle-dependent and labelling time-dependent moments of unlabelled and labelled mRNA (see Supplementary Methods 1.4 and Tables 1 and 2). Columns (C-E) show statistics that can be computed from mean measurements and are defined in rows (1),(3) and (5) of Table 1. Columns (F-G) show statistics that can be computed from noise measurements and are defined in rows (2) and (4) of Table 1. The constant scaling and non-scaling models (rows 1-2) can be distinguished by the cell cycle-dependent mean and noise statistics (columns (C) and (G)). The varying burst frequency and burst size models (rows 3-4) yield almost identical statistics, but they have a striking difference in the cell cycle-dependent Fano factor. The varying decay rate model (row 5) yields substantially different statistics compared to the other two non-constant models. Distinguishing between constant and non-constant models can be based on correlation coefficients: The correlation of cell cycle-dependent means (column (E)), that measures noise due to cell cycle variation, is higher in non-constant models, while the mean (cell cycle-averaged) correlation, that measures upstream noise due to the bursty gene promoter, is higher in constant models.



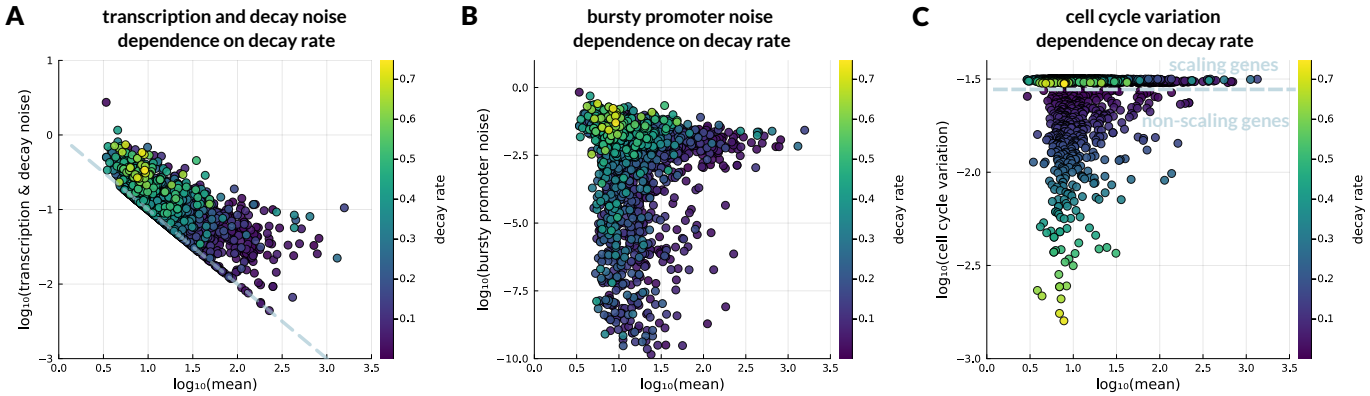
Supplementary Figure S3: **Approximate posterior distributions of parameters for an example gene *SGK1***. The diagonal plots show kernel density estimates of the posterior distributions of burst frequency k_{on}, burst size α/k_{off}, synthesis rate α and decay rate γ in $\log_{10}$ scale. The off-diagonal plots show kernel density estimates of the joint posterior distributions of pairs of parameters. All parameters are measured in h^{-1}, apart from burst size which is unitless.



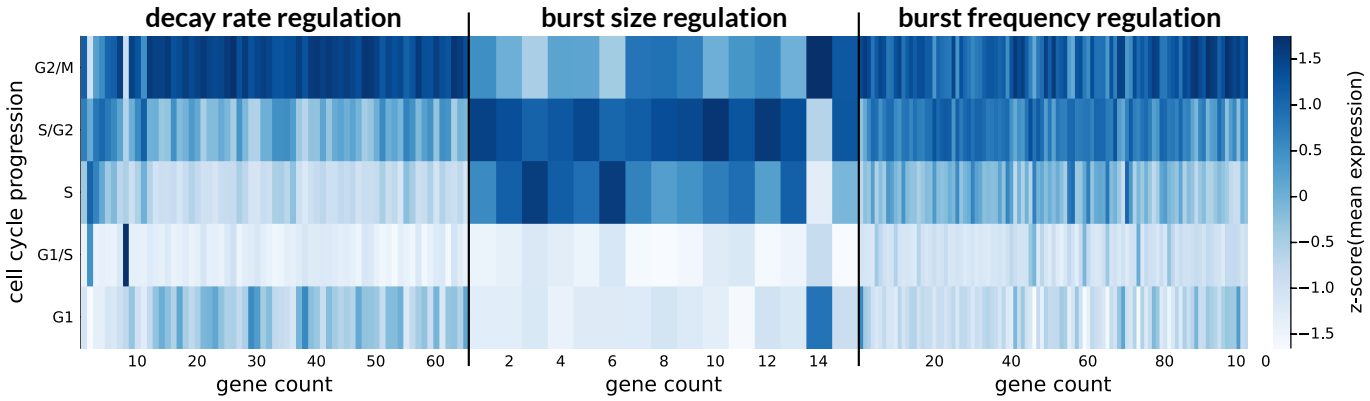
Supplementary Figure S4: **mRNA half-life distributions and comparison with previously reported estimates.** (A) Estimated mRNA half-lives of constant scaling and non-scaling genes (in hours), as defined in (10) using the point estimates of decay rate. (B-C) Comparison of half-lives of constant genes with n = 2209 common genes in A549 cells as measured in the sci-fate study [51] and with n = 1565 common genes in NIH-3T3 cells as measured in the scSLAM-seq study [48]. These comparisons exclude half-lives that are larger than 100 hours (less than 2% of constant genes). The orange dashed lines are identity lines. (D) Estimated mRNA half-lives of non-constant genes. To compute the half-lives of varying decay rate genes, we use the average decay rate over the cell cycle. (E-F) Comparison of half-lives of non-constant genes with n = 177 common genes in A549 cells as measured in the sci-fate study and with n = 119 common genes in NIH-3T3 cells as measured in the scSLAM-seq study. These comparisons exclude half-lives that are larger than 40 hours (1% of non-constant genes). The orange dashed lines are identity lines.



Supplementary Figure S5: **Dependence of intrinsic noise sources on burst parameters.** Here we show constant kinetic parameters associated to all constant (scaling and non-scaling) genes. (A) Dependence of transcription & decay noise and mean on relative burst frequency k_{on}/γ (unitless parameter). (B) Dependence of transcription & decay noise and mean on relative burst size α/k_{off} (unitless parameter). (C) Dependence of bursty promoter noise on burst size and absolute burst frequency k_{on} (burst frequency measured in h^{-1}). All parameters are shown in $\log_{10}$ scale.



Supplementary Figure S6: **Dependence of intrinsic and extrinsic noise sources on decay rate.** We only show constant decay rates of all constant (scaling and non-scaling) genes. Decay rate is measured in h^{-1}.



Supplementary Figure S7: **Mean expression profiles of cell cycle-dependent genes.** Heatmaps of mean expression z-scores of decay rate genes, burst size genes and burst frequency genes. The ordering of the genes corresponds exactly to the ordering of kinetic rates in the heatmap of Figure 5 (B).