Accuracy and limitations of extrinsic noise models to describe gene expression in growing cells

Chen Jia[1], Ramon Grima[2,*]

[1]Applied and Computational Mathematics Division, Beijing Computational Science Research Center, Beijing, 100193, China

[2] School of Biological Sciences, University of Edinburgh, EH9 3JH, U.K.

[*] Correspondence: chenjia@csrc.ac.cn; ramon.grima@ed.ac.uk

Abstract

The standard model describing the fluctuations of mRNA numbers in single cells is the telegraph model which includes synthesis and degradation of mRNA, and switching of the gene between active and inactive states. While commonly used, this model does not describe how fluctuations are influenced by the cell cycle phase, cellular growth and division, and other crucial aspects of cellular biology. Here we derive the analytical time-dependent solution of a stochastic model that explicitly considers various sources of intrinsic and extrinsic noise: switching between inactive and active states, doubling of gene copy numbers upon DNA replication, dependence of the mRNA synthesis rate on cellular volume, gene dosage compensation, partitioning of molecules during cell division, cell-cycle duration variability, and cell-size control strategies. We show that generally the analytical distribution of transcript numbers in steady-state growth cannot be accurately approximated by the steady-state solution of extrinsic noise models, i.e. a telegraph model with parameters drawn from probability distributions. This is because the mRNA lifetime is often not small enough compared to the cell cycle duration to erase the memory of division and replication. Accurate approximations are possible when this memory is weak, e.g. for genes with bursty expression and for which there is sufficient gene dosage compensation when replication occurs.

Introduction

Experiments have revealed a large cell-to-cell variation in the number of mRNA molecules in isogenic populations [1–3]. This can in part be explained by stochastic effects in gene expression due to the low copy numbers of many components, including DNA and important regulatory molecules [4]. Live-cell imaging approaches allow a direct visualization of stochastic bursts of gene expression in living cells [5]. However these experiments are challenging and hence more commonly one measures the mRNA expression per cell from single-molecule fluorescence *in situ* hybridization (smFISH) [5] or single-cell RNA sequencing (scRNA-seq) experiments [6].

The experimental distributions of mRNA numbers are fitted to the predictions of mathematical models, by which one can obtain estimates of the rates of several important transcriptional processes [7–9]. The most common model of this type is the so-called two-state or random telegraph model of gene expression [10, 11]. This is composed of four (effective) reactions

$$G \xrightarrow{\sigma_1} G^*, \quad G^* \xrightarrow{\sigma_0} G, \quad G^* \xrightarrow{\rho} G^* + M, \quad M \xrightarrow{d} \varnothing, \tag{1}$$

where the first two reactions describe the switching of the gene between an active state G^* and an inactive state G, the third reaction describes transcription while the gene is in the active state, and the fourth reaction describes the degradation of the mRNA M. The chemical master equation (CME) describing the telegraph model can be exactly solved in steady-state, as well as in time [11–13]. Extensions of this model to include more than two gene states have also been considered [14–16].

A substantial number of genes are inactive most of the time and in the brief time that they are active, a large number of mRNA molecules are transcribed but not degraded [17]. This leads to bursty expression. The probability of r new mRNA molecules being transcribed before the gene switches off, i.e. a burst of size r, is $P(r) = p^r(1-p)$, where $p = \rho/(\rho + \sigma_0)$ is the probability that the gene synthesizes an mRNA molecule, conditional on it being in the active state [18]. This distribution is geometric with mean ρ/σ_0. The average time between two consecutive bursts is $1/\sigma_0 + 1/\sigma_1 \approx 1/\sigma_1$ since the gene spends most of its time off ($\sigma_0 \gg \sigma_1$); in other words the rate of burst production is approximately σ_1. It follows that the reaction scheme given in Eq. (1) can be reduced to an effective one-state model composed of only two reactions

$$G \xrightarrow{\sigma_1} G + kM, \quad M \xrightarrow{d} \varnothing, \tag{2}$$

where k is the transcriptional burst size which is geometrically distributed with mean ρ/σ_0. The geometric burst size distribution has been validated experimentally [1]. The CME for this model can be solved exactly in steady-state leading to the well-known negative binomial distribution of mRNA numbers [19–21], which is also widely used in scRNA-seq analysis [22]. Because of the unimodality of this distribution, this simplified model cannot explain bimodality in gene expression [23, 24], a feature that can be explained by the two-state model.

However, the conventional one-state and two-state models are very limited in their predictive power because they lack a description of many cellular processes that are known to have a profound impact on the distribution of mRNA numbers in single cells, e.g. the doubling of gene copy numbers upon DNA replication [25], partitioning of molecules during cell division [26], scaling of the mRNA synthesis rate with cell volume [27–31], and stochasticity in the cell cycle duration and growth rate that is related to cell-size control strategies [32–38]. Recently, numerous efforts have been made to extend the conventional one-state and two-state models to include some description of these processes and yet retain analytical tractability. Some studies focused on the moment statistics (mean and variance) of mRNA and protein numbers [39–43], while other studies additionally obtain the analytical distributions of molecule numbers [20, 44–48]. Please refer to Table 1 for a summary of exactly solvable extensions of the one-state and two-state models that explicitly capture cell birth, growth, and division.

Effective one-state models		
	gene replication not considered	gene replication considered
volume-independent transcription	Ref. [44]	Refs. [45–47]
volume-dependent transcription	Ref. [48]	Refs. [20, 47]
Two-state models		
	gene replication not considered	gene replication considered
volume-independent transcription	×	×
volume-dependent transcription	Ref. [48]	×

Table 1: **Exactly solvable gene expression models that explicitly describe cell birth, growth, and division.**

Due to mathematical complexity, most previous work is limited to the effective one-state model with the gene product (mRNA or protein) produced in a constitutive or bursty manner [20, 44–47]. Some of these models incorporate the scaling of transcription activity with cell volume [20, 47], while the rest do not. We note that the latter case is not to be seen as unphysical since while the scaling of transcription with volume is commonly observed, it is by no means a universal phenomenon (in both prokaryotic [49, 50] and eukaryotic cells [51–54] there are examples where there is no such scaling). As for the conventional one-state model shown in Eq. (2), the main limitation is the assumption of instantaneous bursts, while in reality there is a finite time for the bursts to occur. A distinct advantage of the extended one-state models over the conventional one is that those which describe gene replication [46] are able to produce bimodal distributions.

The exact solution of extended two-state models that incorporate cell birth, growth, and division has not received much attention. A recent study [48] made progress in this direction. In particular, the two-state telegraph

model in growing and dividing cells was shown to be exactly solvable when (i) the mRNA synthesis rate scales linearly with cell volume and (ii) there is no variation of gene copy numbers across the cell cycle, i.e. gene replication is not taken into account. As previously mentioned, while (i) is common, it is not universal. The assumption behind (ii) is of course a means to simplify the model but clearly unrealistic. Relaxing any one of these two properties means that within the theoretical framework presented in [48], it is not possible to obtain an exact solution for the distribution of mRNA numbers.

While the aforementioned literature summarised in Table 1 has sought to fix the biological limitations of the conventional one-state and two-state models by directly introducing more processes and solving the master equation of the resulting complex models, a different indirect approach has also been proposed. This approach takes the point of view that biological processes not explicitly modelled by the conventional models can be incorporated by considering the model parameters themselves to vary between cells, and therefore to be drawn from probability distributions [4, 55–57] — we call this an extrinsic noise model (ENM). This model can be solved exactly in steady-state for various distributions of parameter values (see Table I of [57]). It is expected that such an approach produces meaningful results provided the parameters controlling cell-to-cell variability change very slowly. Under certain conditions, the solution of the ENM might even exactly match that of complex models of stochastic gene expression. For example, it has recently been shown that the exact solution of the two-state telegraph model in growing and dividing cells where gene replication is ignored and where the mRNA synthesis rate scales with cell volume is precisely the same as that of the ENM with the mRNA synthesis rate sampled from the distribution of cellular volume and with the mRNA degradation rate being replaced by an effective rate that also incorporates the dilution of molecules at cell division [48]. A natural question is, if in a two-state telegraph model we introduce gene replication and allow the mRNA synthesis rate potentially to be volume-dependent, then does the ENM still provide an exact or at least an accurate approximation of this model?

In this paper, we answer this question. We first exactly solve an extension of the telegraph model that explicitly describes cell birth, growth, division, replication, and an mRNA synthesis rate that can be either independent of cell volume or else that linearly scales with it. Many of the known exact solutions of the one-state and two-state models to-date can be shown to be special cases of the present theory. The analytical distribution of transcript numbers is subsequently used to study the accuracy of the ENM. We show that the transcript number distribution in steady-state growth is generally not well approximated by the steady-state distribution of the ENM. Conditions under which the ENM provides an accurate approximation are derived and verified using simulations.

Results

Model

We consider an extension of the telegraph model which takes into account cell growth, cell division, gene replication, gene dosage compensation, and volume-dependent transcription (see Fig. 1 for an illustration). The specific meaning of all model parameters can be found in Table 2. The model has the following properties.

1) Let T denote the cell cycle duration and let $V(t)$ denote the cell volume at time t. We assume that cell volume grows exponentially within each cell cycle, i.e. $V(t) = V_b e^{gt}$ for any $0 \leq t \leq T$, where V_b is the cell volume at birth and g is the growth rate. The exponential growth of cell volume is commonly observed for various types of cells [36, 38, 58, 59]. For simplicity, we assume that the doubling time T and the growth rate g do not involve any stochasticity [20]. Generalization of the model to stochastic cell volume dynamics will be discussed at the end of the paper.

2) In each cell cycle, we use a two-state model to describe the gene expression dynamics. Let G and G^* denote the inactive and active states of the gene, respectively, and let M denote the corresponding mRNA. Consider a gene expression model described by the effective reactions

$$G \xrightarrow{\sigma_1} G^*, \quad G^* \xrightarrow{\sigma_0} G, \quad G^* \xrightarrow{\rho V(t)^\beta} G^* + M, \quad M \xrightarrow{d} \varnothing,$$

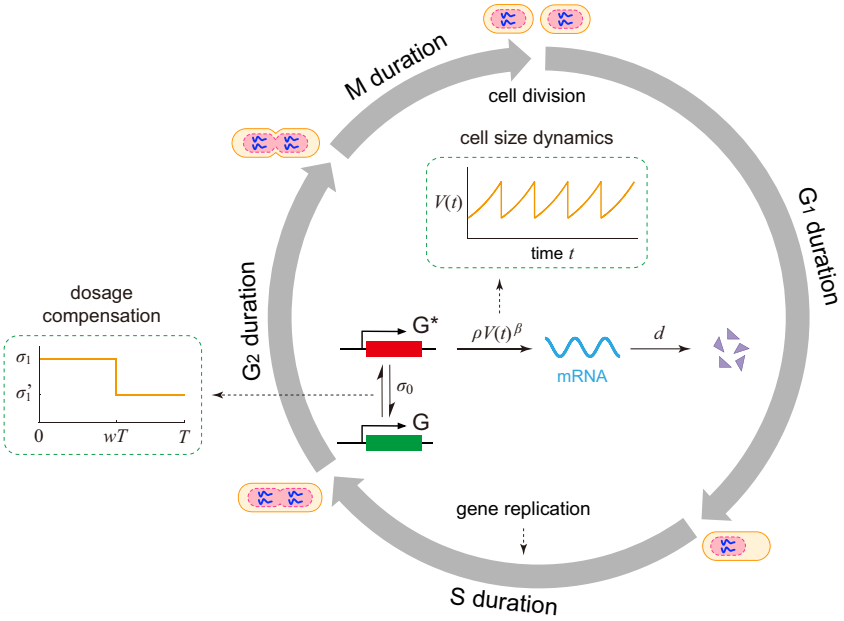

Figure 1: **Model.** Schematic of an extension of the telegraph model of gene expression in growing and dividing cells. The volume $V(t)$ of a cell grows exponentially with constant growth rate g and doubling time T. The gene expression dynamics is characterized by a two-state model with volume-dependent transcription and volume-independent degradation. Specifically, the gene can switch between an active state G^* and an inactive state G. Transcription occurs when the gene is active. The synthesis rate of mRNA depends on cell volume $V(t)$ via a power law form with power $\beta \in [0, 1]$, and the degradation rate of mRNA is a constant. Gene replication occurs at a time T_0 where $w = T_0/T \in (0, 1)$ is some fixed proportion of the cell cycle. Upon replication, the activation rate for each gene copy decreases from σ_1 to σ_1' due to gene dosage compensation.

parameters	meaning
V_b	cell volume at birth
g	growth rate of cell volume
$T = \log(2)/g$	cell cycle duration
β	strength of balanced mRNA synthesis
w	proportion of cell cycle before replication
σ_1	switching rate of the gene from OFF to ON before replication
σ_1'	switching rate of the gene from OFF to ON after replication
σ_0	switching rate of the gene from ON to OFF
ρ	proportionality constant of the mRNA synthesis rate
d	mRNA degradation rate
$d_{\text{eff}} = d + g$	effective mRNA decay rate
$\eta = d/g$	ratio of the degradation rate to the growth rate

Table 2: **Model parameters and their meaning.**

where σ_0 and σ_1 are the switching rates between the two gene states, and d is the mRNA degradation rate. For many genes in fission yeast [27, 28], mammalian cells [29, 30], and plant cells [31], there is evidence that the mRNA number scales linearly with cell volume in order to maintain approximately constant concentrations (concentration homeostasis; for a recent review see [60]). This is due to a coordination of the mRNA synthesis rate with cell volume — we shall refer to this mechanism as balanced mRNA synthesis. However, in both prokaryotic [49, 50] and eukaryotic cells [51–54] there are examples where there is no such scaling. Since each cell has a different volume, the mechanism of volume-dependent transcription is a source of extrinsic noise [56], potentially accounting for a significant amount of the observed cell-to-cell variation in mRNA numbers. To unify non-balanced and balanced mRNA synthesis, we assume that the mRNA synthesis rate depends on cell volume $V(t)$ via a power law form with

proportionality constant ρ and power $\beta \in [0, 1]$. Then $\beta = 1$ ($\beta = 0$) corresponds to the situation where the mRNA synthesis rate scales linearly with cell volume (does not depend on cell volume). It has recently been postulated that the non-linear scaling between gene expression levels and cellular volume is due to the heterogeneous recruitment abilities of promoters to RNA polymerases [61].

3) The replication of the gene of interest occurs at a fixed proportion $w \in (0, 1)$ of the cell cycle. This is known as a stretched cell cycle model, which is supported by experiments [62]. Under this assumption, the time before replication within a cell cycle is wT and the time after replication is $(1 - w)T$. We shall refer to the gene copy that is replicated as the mother copy and to the duplicated gene copies as the daughter copies. For haploid cells, there is only one mother copy before replication and two daughter copies after replication; for diploid cells, the number of gene copies varies from two to four upon replication. For diploid cells, we assume that the two alleles act independently of each other [63, 64].

4) At replication, the daughter copies inherit the gene state from the mother copy [20, 65]. The presence of specific histone marks dictate transcription permissiveness [66] and the landscape of histone modifications is copied during DNA replication [67]. An alternative case is the one where all daughter copies are reset to the inactive state upon replication — potentially a mechanism to avoid the risk of conflict between replication and transcription (and the resulting DNA damage) [20]. Here we only consider the former perfect state copying mechanism.

5) A doubling of gene copy numbers upon replication would be expected to also double the amount of mRNA molecules. However, experiments show that this is not always the case [25, 29, 68] principally due to a decrease of the gene activation rate upon replication, a phenomenon known as gene dosage compensation. We model this by choosing the gene activation rate before replication σ_1 to be potentially different than that after replication σ_1'. In the absence of dosage compensation, we have $\sigma_1' = \sigma_1$. Perfect dosage compensation occurs when $\sigma_1' = \sigma_1/2$; in this case, the total burst frequency (the sum is over all gene copies) is unaffected by replication.

6) At division, the mother cell is divided into two daughter cells. The volumes of the two daughter cells are assumed to be the same and exactly one half of the volume of the mother cell before division (of course there is some stochasticity in the partitioning of cell size [69, 70] which we are here ignoring). Moreover, we assume that each mRNA molecule has probability $1/2$ of being allocated to each daughter cell. With this assumption, the number of transcripts that are allocated to each daughter cell has a binomial distribution. We also assume that gene state is not changed upon cell division.

Time-dependent mRNA distribution within a cell cycle

Here we compute the time-dependent distribution of the mRNA number within a cell cycle under arbitrary initial conditions. We first consider the dynamics before replication for haploid cells. The microstate of the gene of interest can be described by an ordered pair (i, n), where i denotes the state of the gene with $i = 0, 1$ corresponding to the inactive and active states, respectively, and n denotes the number of mRNA molecules. Let $p_{i,n}(t)$ denote the probability of having n transcripts at time $t \in [0, wT]$ when the gene is in state i. Note that $t = 0$ corresponds to cell birth. Then the stochastic gene expression dynamics before replication is governed by the coupled set of CMEs

$$\begin{aligned}\dot{p}_{0,n} &= d[(n+1)p_{0,n+1} - np_{0,n}] + [\sigma_0 p_{1,n} - \sigma_1 p_{0,n}],\\ \dot{p}_{1,n} &= \rho V(t)^{\beta}[p_{1,n-1} - p_{1,n}] + d[(n+1)p_{1,n+1} - np_{1,n}] + [\sigma_1 p_{0,n} - \sigma_0 p_{1,n}],\end{aligned} \tag{3}$$

where $p_{1,-1} = 0$ by default, the term involving ρ represents mRNA synthesis, the terms involving d represent mRNA degradation, and the terms involving σ_0 and σ_1 represent gene switching. To solve these, we define a pair of generating functions $F_i(t, z) = \sum_{n=0}^{\infty} p_{i,n}(t)(z + 1)^n$ for $i = 0, 1$. Note that here we use $(z + 1)^n$ rather than the conventional z^n in the definition of the generating function — with this choice, the formulas given below are much more concise. In addition, let $p_n(t) = p_{0,n}(t) + p_{1,n}(t)$ denote the probability of having n transcripts at time t and let $F(t, z) = F_0(t, z) + F_1(t, z)$ be the corresponding generating function. In terms of the generating

functions, Eq. (3) can be converted into the first-order linear partial differential equations (PDEs)

$$\begin{aligned}\partial_t F_0 &= -dz\partial_z F_0 + \sigma_0 F_1 - \sigma_1 F_0,\\ \partial_t F_1 &= \rho V(t)^\beta z F_1 - dz\partial_z F_1 + \sigma_1 F_0 - \sigma_0 F_1.\end{aligned} \tag{4}$$

To solve these, we first convert them into a second-order parabolic PDE and then transform the second-order PDE into a hypergeometric differential equation through a change of variables. Complex computations show that for each $t \in [0, wT]$, the generating functions F_i, $i = 0, 1$ can be computed in closed form as (see Supplementary Section 1 for the proof)

$$\begin{aligned}F_0(t,z) &= K_{00}(t,z)F_0(0,e^{-dt}z) + K_{01}(t,z)F_1(0,e^{-dt}z),\\ F_1(t,z) &= K_{10}(t,z)F_0(0,e^{-dt}z) + K_{11}(t,z)F_1(0,e^{-dt}z).\end{aligned} \tag{5}$$

Here $F_i(0,z)$, $i = 0, 1$ are the generating functions at $t = 0$ which can be determined by the initial conditions, and the functions K_{ij}, $i, j = 0, 1$ are given by

$$\begin{aligned}K_{00}(t,z) &= \frac{b-a}{b}\big[M(1+a-b;1-b;ue^{-dt}z)M(a;1+b;ue^{\beta gt}z) + \frac{a}{b-a}e^{-(r+\beta g)t}\\ &\quad\times M(1+a;1+b;ue^{-dt}z)M(a-b;1-b;ue^{\beta gt}z)\big]e^{-ue^{-dt}z},\\ K_{01}(t,z) &= \frac{b-a}{b}\big[M(a-b;1-b;ue^{-dt}z)M(a;1+b;ue^{\beta gt}z) - e^{-(r+\beta g)t}\\ &\quad\times M(a;1+b;ue^{-dt}z)M(a-b;1-b;ue^{\beta gt}z)\big]e^{-ue^{-dt}z},\\ K_{10}(t,z) &= \frac{a}{b}\big[M(1+a-b;1-b;ue^{-dt}z)M(1+a;1+b;ue^{\beta gt}z) - e^{-(r+\beta g)t}\\ &\quad\times M(1+a;1+b;ue^{-dt}z)M(1+a-b;1-b;ue^{\beta gt}z)\big]e^{-ue^{-dt}z},\\ K_{11}(t,z) &= \frac{a}{b}\big[M(a-b;1-b;ue^{-dt}z)M(1+a;1+b;ue^{\beta gt}z) + \frac{b-a}{a}e^{-(r+\beta g)t}\\ &\quad\times M(a;1+b;ue^{-dt}z)M(1+a-b;1-b;ue^{\beta gt}z)\big]e^{-ue^{-dt}z},\end{aligned} \tag{6}$$

where the parameters r, a, b, and u are given by

$$r = \sigma_0 + \sigma_1 - \beta g, \quad a = \frac{\sigma_1}{d+\beta g}, \quad b = \frac{\sigma_0+\sigma_1}{d+\beta g}, \quad u = \frac{\rho V_b^\beta}{d+\beta g}. \tag{7}$$

Adding the two identities in Eq. (5) gives the explicit expression of the generating function F before replication, i.e.

$$F(t,z) = L_0(t,z)F_0(0,e^{-dt}z) + L_1(t,z)F_1(0,e^{-dt}z), \quad t \in [0, wT], \tag{8}$$

where the functions L_i, $i = 0, 1$ are given by

$$\begin{aligned}L_0(t,z) &= \Big[M(1+a-b;1-b;ue^{-dt}z)M(a;b;ue^{\beta gt}z) + \frac{auz}{b(b-1)}e^{-rt}\\ &\quad\times M(1+a;1+b;ue^{-dt}z)M(1+a-b;2-b;ue^{\beta gt}z)\Big]e^{-ue^{-dt}z},\\ L_1(t,z) &= \Big[M(a-b;1-b;ue^{-dt}z)M(a;b;ue^{\beta gt}z) - \frac{(b-a)uz}{b(b-1)}e^{-rt}\\ &\quad\times M(a;1+b;ue^{-dt}z)M(1+a-b;2-b;ue^{\beta gt}z)\Big]e^{-ue^{-dt}z}.\end{aligned} \tag{9}$$

When $b = 1$, the term $b - 1$ appears in the dominator of the above two equations and the equalities should be understood in the limiting sense. Note that when the mRNA synthesis rate is volume-independent ($\beta = 0$), the expression of F given in Eq. (8) coincides with the time-dependent solution of the standard telegraph model [13].

We next focus on the dynamics after replication for haploid cells. Since there are two daughter gene copies after replication, to distinguish them, we call them daughter copy A and daughter copy B. The dynamics of each

gene copy is governed by the CMEs given in Eq. (3) with σ_1 being replaced by σ_1'. Let $p_n(t)$ denote the probability of having n transcripts at time $t \in [wT, T]$ and let $F(t,z) = \sum_{n=0}^{\infty} p_n(t)(z+1)^n$ be the corresponding generating function. In Supplementary Section 2, we prove that the generating function F after replication can be computed in closed form as

$$F(t,z) = L_0'(t-wT,z)^2 F_0(wT, e^{-d(t-wT)}z) + L_1'(t-wT,z)^2 F_1(wT, e^{-d(t-wT)}z), \quad t \in [wT, T],$$

where L_0' and L_1' are functions obtained from L_0 and L_1 by replacing the parameters r, a, b and u with

$$r' = \sigma_0 + \sigma_1' - \beta g, \quad a' = \frac{\sigma_1'}{d+\beta g}, \quad b' = \frac{\sigma_0 + \sigma_1'}{d+\beta g}, \quad u' = 2^{\beta w} u.$$

In summary, we have derived the analytical expression of the generating function F at any time $t \in [0, T]$ within a cell cycle, which is given by

$$F(t,z) = \begin{cases} \sum_{i=0}^{1} L_i(t,z) F_i(0, e^{-dt}z), & t \in [0, wT], \\ \sum_{i=0}^{1} L_i'(t-wT,z)^2 F_i(wT, e^{-d(t-wT)}z), & t \in [wT, T], \end{cases} \tag{10}$$

where $F_i(wT, z)$, $i = 0, 1$ are determined by Eq. (5). The time-dependent distribution of the mRNA number can be recovered by taking the derivatives of the generating function F at $z = -1$, i.e.

$$p_n(t) = \frac{1}{n!} \frac{\partial^n}{\partial z^n} F(t,z)\Big|_{z=-1}. \tag{11}$$

Our analytical expression of the transient mRNA distribution is rather complicated. However, it can be simplified to a large extent in some special cases. In Supplementary Section 3, we show how the analytical solution can be simplified for two non-trivial special cases: (i) the gene switches rapidly between the active and inactive states ($\sigma_0, \sigma_1 \gg g$); (ii) the mRNA is produced in a bursty manner ($\sigma_0 \gg \sigma_1$), i.e. the gene is mostly inactive but synthesizes a large number of transcripts when it becomes active [71–73]. In the latter case, the mean burst size is ρ/σ_0; the burst frequency is σ_1 before replication and the total burst frequency for the two gene copies is $2\sigma_1'$ after replication.

Thus far, we have obtained the transient mRNA distribution for haploid cells. For diploid cells, since the two alleles act independently and since each allele has the mRNA distribution given in Eq. (11), the generating function for the total number of transcripts at any time $t \in [0, T]$ is given by $F_{\text{diploid}}(t,z) = F(t,z)^2$, where $F(t,z)$ is given by Eq. (10). Due to independence of the two alleles, when the rate parameters for each allele are fixed, the gene expression noise (measured by the coefficient of variation squared of mRNA numbers) in diploid cells is one half that in haploid cells. Without loss of generality, we always focus on haploid cells in what follows.

Time-dependent mRNA distribution across cell cycles

Thus far, we have derived the exact mRNA distribution at any time within a cell cycle. Here we focus on the full time-dependence of the mRNA distribution across cell cycles under arbitrary initial conditions. To this end, we not only need the expression of F at any time $t \in [0, T]$, but also need the expressions of F_i, $i = 0, 1$.

Recall that Eq. (5) gives the analytical expressions of the generating functions F_i, $i = 0, 1$ before replication under any initial conditions. In particular, at replication, we have

$$F_i(wT, z) = K_{i0}(wT, z) F_0(0, e^{-dwT}z) + K_{i1}(wT, z) F_1(0, e^{-dwT}z). \tag{12}$$

Now we focus on the dynamics of daughter copy A after replication. Let $p_{i,n}(t)$ denote the probability of having n transcripts at time $t \in [wT, T]$ when the daughter copy A is in state i and let $F_i(t,z) = \sum_{n=0}^{\infty} p_{i,n}(t)(z+1)^n$ be

the corresponding generating function. In Supplementary Section 2, we prove that the generating functions F_i, $i = 0, 1$ after replication can be computed exactly as

$$\begin{aligned} F_i(t,z) = {} & K'_{i0}(t-wT,z)L'_0(t-wT,z)F_0(wT, e^{-d(t-wT)}z) \\ & + K'_{i1}(t-wT,z)L'_1(t-wT,z)F_1(wT, e^{-d(t-wT)}z), \quad t \in [wT, T], \end{aligned} \tag{13}$$

where K'_{ij}, $i, j = 0, 1$ are functions obtained from K_{ij} by replacing the parameters r, a, b, and u with the parameters r', a', b', and u', respectively. Inserting Eq. (12) into Eq. (13) and taking $t = T$, we obtain

$$F_i(T,z) = \tilde{K}_{i0}(z)F_0(0, e^{-dT}z) + \tilde{K}_{i1}(z)F_1(0, e^{-dT}z), \tag{14}$$

where

$$\begin{aligned} \tilde{K}_{ij}(z) = {} & K'_{i0}((1-w)T,z)L'_0((1-w)T,z)K_{0j}(wT, e^{-d(1-w)T}z) \\ & + K'_{i1}((1-w)T,z)L'_1((1-w)T,z)K_{1j}(wT, e^{-d(1-w)T}z), \quad i,j = 0,1. \end{aligned} \tag{15}$$

Suppose that the daughter cell with daughter copy A is tracked after division. Since we have assumed binomial partitioning of molecules at division, the probability $p^{\text{next}}_{i,n}(0)$ at birth in the next generation is given by

$$p^{\text{next}}_{i,n}(0) = \sum_{m=n}^{\infty} p_{i,m}(T)\binom{m}{n}\left(\frac{1}{2}\right)^m. \tag{16}$$

In terms of the generating function, the above relation can be written as

$$F^{\text{next}}_i(0,z) = F_i\left(T, z/2\right). \tag{17}$$

This gives the initial conditions for the next generation and the time-dependent mRNA distribution within the next cell cycle can be computed via Eq. (10). Applying Eqs. (14), and (17) repeatedly, we are able to compute the full time-dependence of the F_i functions across cell cycles; substituting these in Eq. (10) gives the full-time dependence of the mRNA distribution across cell cycles.

As a check of our analytical solutions, we compare the exact distributions of the mRNA number with the numerical ones obtained from a modified version of the finite-state projection (FSP) [74] algorithm at three different time points (birth, replication, and division) across four cell cycles (Fig. 2). In this algorithm, we couple the standard FSP with cell cycle events; for details see Supplementary Section 4. Here we assume that initially there is no mRNA molecules in the cell and the gene is off. This mimics the situation where the gene has been silenced by some repressor over a period of time such that all transcripts have been removed via degradation; at time $t = 0$, the repressor is removed and we study how gene expression recovers. When using FSP, we truncate the state space (to exclude states that are visited very rarely) and solve the associated truncated master equation numerically using the MATLAB function ODE45 with the dynamics before and after replication solved separately. Note that while the FSP and the stochastic simulation algorithm (SSA) yield comparable distributions of molecule numbers, the computational time of the former is much less than of the latter provided the biochemical reaction networks are small enough — hence here we used the FSP. As expected, the analytical and simulated solutions coincide with each other completely at all times, and the mRNA distributions at birth, replication, and division reach a steady state within a few cell cycles.

Another interesting observation is that the time-dependent mRNA distributions for our detailed telegraph model may have more than two modes (Fig. 2) — this is the combined effect of gene replication and slow switching between gene states. This is different from the prediction of the conventional telegraph model [11] whose distribution has at most two modes.

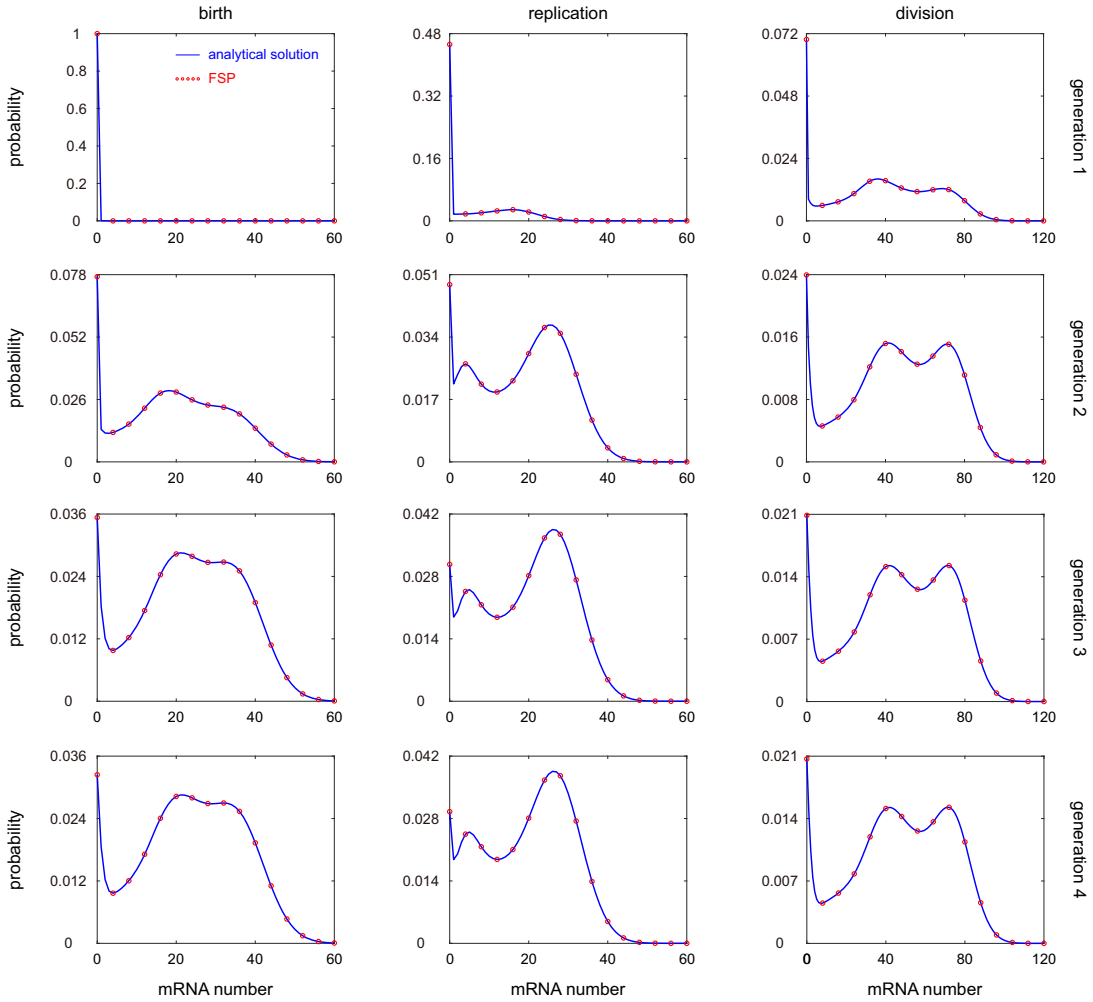

Figure 2: **Time-dependent mRNA distributions at birth, replication, and division across four cell cycles.** The blue curves show the analytical distributions computed by applying Eqs. (10), (14), and (17) repeatedly, and the red circles show the numerical ones obtained from FSP. The model parameters are chosen as $V_b = 1, g = 1, \beta = 1, w = 0.4, d = 5, \rho = 20 d_{\text{eff}}, \sigma_0 = 1.5, \sigma_1 = 3, \sigma_1' = 2.4$.

Time-dependent mRNA distribution under cyclo-stationary conditions

Thus far, we have obtained the full time-dependence of the mRNA distribution across cell cycles under arbitrary initial conditions. After several generations, the distribution at any fixed time within a cell cycle (such as the distributions at birth, replication, and division) becomes independent of the generation number. This is also called the cyclo-stationary condition in the literature [45] or steady-state growth [18]. Next we compute the time-dependent mRNA distribution within a cell cycle under cyclo-stationary conditions.

Before computing the mRNA distribution, we first derive the probabilities of the gene being in the active and inactive states at any time within a cell cycle under cyclo-stationary conditions. Let $p_{\text{on}}(t)$ denote the probability of each gene copy being in the active state at time $t \in [0, T]$. Before replication, the dynamics of the active probability satisfies the differential equation $\dot{p}_{\text{on}} = \sigma_1(1 - p_{\text{on}}) - \sigma_0 p_{\text{on}}$. Solving this equation gives rise to

$$p_{\text{on}}(t) = \frac{a}{b} + \left[p_{\text{on}}(0) - \frac{a}{b}\right] e^{-(r+\beta g)t}, \quad t \in [0, wT], \tag{18}$$

where we have used the fact that $a/b = \sigma_1/(\sigma_0 + \sigma_1)$ and $r + \beta g = \sigma_0 + \sigma_1$ (see Eq. (7)). Recall that the gene activation rate decreases from σ_1 to σ_1' upon replication. After replication, the dynamics of the active probability satisfies the differential equation $\dot{p}_{\text{on}} = \sigma_1'(1 - p_{\text{on}}) - \sigma_0 p_{\text{on}}$. Solving this equation yields

$$p_{\text{on}}(t) = \frac{a'}{b'} + \left[p_{\text{on}}(wT) - \frac{a'}{b'}\right] e^{-(r'+\beta g)(t-wT)}, \quad t \in [wT, T], \tag{19}$$

where $p_{\rm on}(wT)$ is determined by Eq. (18). Combining Eqs. (18) and (19), we obtain the active probability of the gene at division, i.e.

$$p_{\rm on}(T) = \frac{a'}{b'} + \left[\frac{a}{b} - \frac{a'}{b'}\right] e^{-(r'+\beta g)(1-w)T} + \left[p_{\rm on}(0) - \frac{a}{b}\right] e^{-(r+\beta g)wT-(r'+\beta g)(1-w)T}.$$

Under cyclo-stationary conditions, the active probabilities at cell birth in two successive generations must be the same, i.e. $p^{ss}_{\rm on}(0) = p^{ss}_{\rm on}(T)$. Then the steady-state active probability of the gene at birth is given by

$$p^b_{\rm on} = p^{ss}_{\rm on}(0) = \frac{\frac{a'}{b'}\left[1 - e^{-(r'+\beta g)(1-w)T}\right] + \frac{a}{b}e^{-(r'+\beta g)(1-w)T}\left[1 - e^{-(r+\beta g)wT}\right]}{1 - e^{-(r+\beta g)wT-(r'+\beta g)(1-w)T}}, \tag{20}$$

and thus the steady-state inactive probability at birth is given by $p^b_{\rm off} = 1 - p^b_{\rm on}$. It then follows from Eq. (18) that the steady-state active probability of the gene at replication is given by

$$p^r_{\rm on} = p^{ss}_{\rm on}(wT) = \frac{\frac{a}{b}\left[1 - e^{-(r+\beta g)wT}\right] + \frac{a'}{b'}e^{-(r+\beta g)wT}\left[1 - e^{-(r'+\beta g)(1-w)T}\right]}{1 - e^{-(r+\beta g)wT-(r'+\beta g)(1-w)T}}, \tag{21}$$

and thus the steady-state inactive probability at replication is given by $p^r_{\rm off} = 1 - p^r_{\rm on}$.

Next we focus on the time-dependent mRNA distributions under cyclo-stationary conditions. Recall that we have obtained the time-dependent mRNA distributions within a cell cycle, whose generating function $F(t,z)$ is given by Eq. (10), provided that the initial conditions $F_i(0,z)$, $i = 0, 1$ are known. Under cyclo-stationary conditions, the values of $F_i(0,z)$ in two successive generations must be the same, i.e. $F_i(0,z) = F^{\rm next}_i(0,z)$, where $F^{\rm next}_i(0,z)$ has been derived in Eqs. (14) and (17). It then follows that the steady-state values of $F_i(0,z)$ should satisfy

$$\begin{pmatrix} F^{ss}_0(0,z) \\ F^{ss}_1(0,z) \end{pmatrix} = R(z) \begin{pmatrix} F^{ss}_0(0,e^{-dT}z/2) \\ F^{ss}_1(0,e^{-dT}z/2) \end{pmatrix}, \tag{22}$$

where

$$R(z) = \begin{pmatrix} \tilde{K}_{00}(z/2) & \tilde{K}_{01}(z/2) \\ \tilde{K}_{10}(z/2) & \tilde{K}_{11}(z/2) \end{pmatrix}$$

is a matrix-valued function with $\tilde{K}_{ij}$, $i, j = 0, 1$ being given in Eq. (15). Applying Eq. (22) repeatedly, we obtain

$$\begin{pmatrix} F^{ss}_0(0,z) \\ F^{ss}_1(0,z) \end{pmatrix} = \prod_{k=0}^{n-1} R((e^{-dT}/2)^k z) \begin{pmatrix} F^{ss}_0(0,(e^{-dT}/2)^n z) \\ F^{ss}_1(0,(e^{-dT}/2)^n z) \end{pmatrix}.$$

Taking $n \to \infty$ in the above equation yields

$$\begin{pmatrix} F^{ss}_0(0,z) \\ F^{ss}_1(0,z) \end{pmatrix} = \prod_{k=0}^{\infty} R((e^{-dT}/2)^k z) \begin{pmatrix} p^b_{\rm off} \\ p^b_{\rm on} \end{pmatrix}, \tag{23}$$

where we have used the fact that

$$\lim_{n\to\infty} F^{ss}_0(0,(e^{-dT}/2)^n z) = F^{ss}_0(0,0) = p^b_{\rm off},$$
$$\lim_{n\to\infty} F^{ss}_1(0,(e^{-dT}/2)^n z) = F^{ss}_1(0,0) = p^b_{\rm on}.$$

Once we have derived the steady-state values of $F_i(0,z)$, $i = 0, 1$, it immediately follows from Eq. (10) that the time-dependent generating function F under cyclo-stationary conditions is given by

$$F^{ss}(t,z) = \begin{cases} \sum_{i=0}^{1} L_i(t,z) F^{ss}_i(0,e^{-dt}z), & t \in [0, wT], \\ \sum_{i=0}^{1} L'_i(t-wT,z)^2 F^{ss}_i(wT, e^{-d(t-wT)}z), & t \in [wT, T]. \end{cases} \tag{24}$$

Comparison with the effective dilution model

Special case 1. Consider the case where gene replication is not taken into account ($w = 1$) and when the mRNA synthesis rate scales with cell volume ($\beta = 1$) [48]. In this case, the functions $\tilde{K}_{ij}(z)$ given in Eq. (15) reduce to $\tilde{K}_{ij}(z) = K_{ij}(T, z)$, and it is not difficult to see that Eq. (22) can be solved analytically as

$$F_0^{ss}(0,z) = \frac{b-a}{b} M(a; b+1; uz), \quad F_1^{ss}(0,z) = \frac{a}{b} M(a+1; b+1; uz).$$

Inserting these equations into Eq. (24) yields

$$F^{ss}(t,z) = L_0(t,z) F_0^{ss}(0, e^{-dt}z) + L_1(t,z) F_1^{ss}(0, e^{-dt}z) = M(a; b; ue^{gt}z), \quad t \in [0, T].$$

For a given cell of volume V, its age is given by $t = \log(V/V_b)/g$. Taking $t = \log(V/V_b)/g$ in the above equation shows that the steady-state generating function for a cell of constant volume V is given by $F_V(z) = M(a; b; \tilde{u}z)$, where $a = \sigma_1/(d+g)$, $b = (\sigma_0 + \sigma_1)/(d+g)$, and $\tilde{u} = \rho V/(d+g)$. We make a crucial observation that this is exactly the steady-state generating function of the mRNA distribution for the conventional telegraph model [11]

$$G \xrightarrow{\sigma_1} G^*, \quad G^* \xrightarrow{\sigma_0} G, \quad G^* \xrightarrow{\rho V} G^* + M, \quad M \xrightarrow{d+g} \varnothing. \tag{25}$$

This result has been found in [48], which states that when $w = \beta = 1$, the steady-state mRNA distribution for a cell of constant volume V of the detailed telegraph model is the same as that of the conventional telegraph model with effective decay rate $d_{\text{eff}} = d + g$. Note that the two terms in this rate capture the fact that transcripts are lost both by active degradation (with rate d) and by dilution at cell division (with rate g) — hence a model of this type is known as an effective dilution model (EDM) [75]. Intuitively, the EDM considers a population of cells with synchronised cell cycles so that at each time, all cells have the same volume.

Special case 2. Experiments have shown that in bacteria, most mRNAs have a half-life that is much shorter than the cell cycle duration, i.e. $d \gg g$ (see Supplementary Section 5 for the typical values of d and g in various cell types), and thus are very unstable. The value of $\eta = d/g$ can be used to measure the stability of mRNA. For unstable mRNAs ($\eta \gg 1$), the terms e^{-dt} and $e^{-d(t-wT)}$ in Eq. (10) are very small and thus can be approximated by zero (whenever t is not very close to 0 and wT). In this case, the time-dependent generating function F under cyclo-stationary conditions reduces to

$$F^{ss}(t,z) = \begin{cases} p_{\text{off}}^b L_0(t,z) + p_{\text{on}}^b L_1(t,z), & t \in (0, wT], \\ p_{\text{off}}^r L_0'(t-wT, z)^2 + p_{\text{on}}^r L_1'(t-wT, z)^2, & t \in (wT, T], \end{cases} \tag{26}$$

where we have used the fact that $F_i(0,0) = \sum_{n=0}^{\infty} p_{i,n}(0)$ and $F_i(wT, 0) = \sum_{n=0}^{\infty} p_{i,n}(wT)$ are the probabilities of the gene being in state i at birth and at replication, respectively. Imposing the term e^{-dt} as zero in Eq. (9) yields

$$\begin{aligned} L_0(t,z) &= M(a; b; ue^{\beta g t}z) + \frac{auz}{b(b-1)} e^{-rt} M(1+a-b; 2-b; ue^{\beta g t}z), \\ L_1(t,z) &= M(a; b; ue^{\beta g t}z) - \frac{(b-a)uz}{b(b-1)} e^{-rt} M(1+a-b; 2-b; ue^{\beta g t}z). \end{aligned} \tag{27}$$

When one of the gene switching rates σ_0 and σ_1 is very large, we have $r = \sigma_0 + \sigma_1 - \beta g \gg g$ and thus the second term on the right-hand side of Eq. (27) can be neglected. This may occur when (i) the gene switches rapidly between the two states ($\sigma_0, \sigma_1 \gg g$), or (ii) the mRNA is produced in a constitutive manner ($\sigma_1 \gg \sigma_0, g$), or (iii) the mRNA is produced in a bursty manner ($\sigma_0 \gg \sigma_1, g$). In this case, the cyclo-stationary generating function F^{ss} can be simplified significantly as

$$F^{ss}(t,z) = \begin{cases} M(a; b; ue^{\beta g t}z), & t \in (0, wT], \\ M(a'; b'; ue^{\beta g t}z)^2, & t \in (wT, T]. \end{cases} \tag{28}$$

This contains much information. For a given cell of volume $V < 2^w V_b$, its age is given by $t = \log(V/V_b)/g < wT$ and hence there is only one gene copy in the cell. Taking $t = \log(V/V_b)/g$ in the above equation shows that the steady-state generating function for a cell of constant volume V is given by $F_V(z) \approx M(a; b; \tilde{u}z)$, where

$$a = \frac{\sigma_1}{d+\beta g} \approx \frac{\sigma_1}{d_{\text{eff}}}, \quad b = \frac{\sigma_0+\sigma_1}{d+\beta g} \approx \frac{\sigma_0+\sigma_1}{d_{\text{eff}}}, \tilde{u} = \frac{\rho V^\beta}{d+\beta g} \approx \frac{\rho V^\beta}{d_{\text{eff}}}.$$

Here we have used the fact that $d_{\text{eff}}/(d+g) \approx 1$ when mRNA is very unstable. Note that $F_V(z)$ is exactly the steady-state generating function of the mRNA distribution for the EDM

$$G \xrightarrow{\sigma_1} G^*, \quad G^* \xrightarrow{\sigma_0} G, \quad G^* \xrightarrow{\rho V^\beta} G^* + M, \quad M \xrightarrow{d_{\text{eff}}} \varnothing. \tag{29}$$

On the other hand, for a given cell of volume $V > 2^w V_b$, its age is given by $t = \log(V/V_b)/g > wT$ and hence there are two gene copies in the cell. In this case, the EDM should be modified as

$$\begin{aligned} &G_A \xrightarrow{\sigma_1'} G_A^*, \quad G_A^* \xrightarrow{\sigma_0} G_A, \quad G_B \xrightarrow{\sigma_1'} G_B^*, \quad G_B^* \xrightarrow{\sigma_0} G_B, \\ &G_A^* \xrightarrow{\rho V^\beta} G_A^* + M, \quad G_B^* \xrightarrow{\rho V^\beta} G_B^* + M, \quad M \xrightarrow{d_{\text{eff}}} \varnothing, \end{aligned} \tag{30}$$

where G_A and G_B denote the two daughter copies whose dynamics are both governed by the conventional telegraph model. Taking $t = \log(V/V_b)/g$ in Eq. (28) shows that the steady-state generating function for a cell of constant volume V is given by $F_V(z) = M(a; b; \tilde{u}z)^2$, where $a' \approx \sigma_1'/d_{\text{eff}}$ and $b' \approx (\sigma_0 + \sigma_1')/d_{\text{eff}}$. Note that $F_V(z)$ is exactly the steady-state generating function of the mRNA distribution for the EDM given in Eq. (30) since the two gene copies are independent of each other.

In summary, our analysis shows that for mRNAs with short lifetimes, the EDM makes a good approximation when one of the gene switching rates σ_0 and σ_1 is large (here the cell age t cannot be very close to 0 and wT, i.e. newborn cells and cells that have just finished gene replication should be excluded). This can be understood as follows. Previous studies [76] have shown that the relaxation speed of the EDM to the steady state is governed by both the mRNA degradation rate d and the total gene switching rate $\sigma_{\text{tot}} = \sigma_0 + \sigma_1$. When d and σ_{tot} are both large, any memory at birth from the previous cycle (due to binomial partitioning of molecules at division and to the gene state prior to division) and any memory at replication (due to gene state copying of the two daughter copies) will be rapidly erased. Each time that the volume changes, the mRNA distribution instantaneously equilibrates and hence the EDM works. Note that when the cell age t is close to 0 and wT, the memory at birth and at replication cannot be erased, which leads to the failure of the EDM. Relatively slow mRNA degradation and relative slow gene switching will both result in a deviation of the EDM from the full model.

Testing the accuracy of the EDM approximation. In Fig. 3 we compare the exact mRNA distributions with the numerical ones obtained from FSP at three different time points (birth, replication, and division) across the cell cycle under cyclo-stationary conditions. The truncated master equations are solved across several (usually less than five) cell cycles until the Hellinger distance between mRNA distributions at birth in two successive generations is less than 10^{-4}. This guarantees that cyclo-stationary conditions are reached. When gene replication is not taken into account ($w = 1$) and when the mRNA synthesis rate scales with cell size ($\beta = 1$), the distributions of the full model agree perfectly with those of the EDM given in Eq. (25) (Fig. 3(a)). This coincides with our theoretical predictions. When gene replication is taken into account, the EDMs before and after replication are given by Eqs. (29) and (30), respectively. In this case, the EDM may deviate remarkably from the full model with the deviation being much larger at early stages of the cell cycle (Fig. 3(b)), especially when mRNA degradation and gene switching are relatively slow. This can be understood as follows. According to the steady-state properties of the conventional telegraph model, in the presence of gene replication, the mean and the Fano factor of the mRNA number at birth for the EDM are given by

$$\langle n \rangle_{\text{EDM}}(0) = \frac{au}{b}, \quad \text{Fano}_{\text{EDM}}(0) = 1 + \frac{(a+1)u}{b+1},$$

and the mean and the Fano factor of the mRNA number at division are given by

$$\langle n\rangle_{\text{EDM}}(T) = \frac{2^{\beta+1}a'u}{b'}, \quad \text{Fano}_{\text{EDM}}(T) = 1 + \frac{2^{\beta}(a'+1)u}{b'+1} + \frac{1}{2}\langle n\rangle_{\text{EDM}}(T).$$

Under cyclo-stationary conditions, it follows from Eq. (16) that the mean mRNA numbers at birth and at division for the full model should satisfy $\langle n\rangle(T) = 2\langle n\rangle(0)$ and $\text{Fano}\,(T) = 2\text{Fano}\,(0) - 1$. However, these two restrictions in general do not hold for the EDM — the EDM satisfies these two restrictions only when

$$2^{\beta}\frac{a'}{b'} = \frac{a}{b}, \quad 2^{\beta-1}\left[\frac{a'+1}{b'+1} + \frac{a'}{b'}\right] = \frac{a+1}{b+1}.$$

Note that when mRNA synthesis is balanced ($\beta = 1$) and bursty ($\sigma_0 \gg \sigma_1$), the above restrictions are satisfied when dosage compensation is perfect ($\sigma_1' = \sigma_1/2$), i.e. when the total burst frequency does not change when replication occurs. When these three conditions are satisfied, the EDM makes accurate predictions and the mRNA number follows a negative binomial distribution (Fig. 3(c)). The breakdown of the above restrictions will give rise to the deviation of the EDM from the full model, as observed in Fig. 3(b). Intuitively, this is because the mRNA distribution at birth is affected by the fluctuations of the two gene copies at division and thus in general it cannot be captured solely by an EDM with only one gene copy. Note that special case 2 discussed above may not satisfy the above moment equalities since in this special case, the EDM fails for newborn cells.

Comparison with the extrinsic noise model

We next compute the steady-state distributions of transcript numbers measured over a cell lineage or from a population snapshot . In lineage measurements, the mRNA number from an individual cell is tracked at any point in time, i.e. once the cell divides, only one of the two daughter cells is tracked. Clearly, the probability of observing a cell of age $t \in [0, T]$ is $1/T$ for lineage measurements. As a result, the generating function of the steady-state distribution along a cell lineage is given by

$$F_{\text{lin}}(z) = \frac{1}{T}\int_0^T F^{ss}(t, z)dt. \tag{31}$$

In contrast, in population measurements, the mRNA numbers in a population of isogenic cells are observed at a particular time. Previous studies [18] have shown that the probability of observing a cell of age $t \in [0, T]$ is $2^{(1-t/T)}(\log 2)/T = 2ge^{-gt}$ for population measurements. Thus the generating function of the steady-state distribution in a population of cells is given by

$$F_{\text{pop}}(z) = 2g\int_0^T F^{ss}(t, z)e^{-gt}dt. \tag{32}$$

Our analytical expression of the steady-state distribution is rather complicated since we have to integrate the time-dependent distribution over time which involves complex confluent hypergeometric functions. However, it can be simplified to a large extent in some special cases. In Supplementary Section 3, we show how the analytical solution can be simplified in two non-trivial special cases: (i) the mRNA is unstable and the gene switches rapidly between the two states; (ii) the mRNA is unstable and the gene switches slowly between the two states. In Supplementary Section 6 and Fig. S1, we also compute the time-dependent mean of the mRNA number, as well as the steady-state means of lineage and population measurements. We find that the lineage mean is always greater than the population mean, and the difference between them is at most 10%.

Our detailed telegraph model involves the coupling between gene expression dynamics, cell volume dynamics, and cell cycle events. In Supplementary Section 7 and Fig. S2, we show that the steady-state distribution of the detailed model cannot be captured by the steady-state solution of the conventional telegraph model given in Eq. (1) with volume-independent rates, even when gene replication is not taken into account ($w = 1$). In previous studies, the lineage and population distributions for the detailed model have often been approximated by the distributions

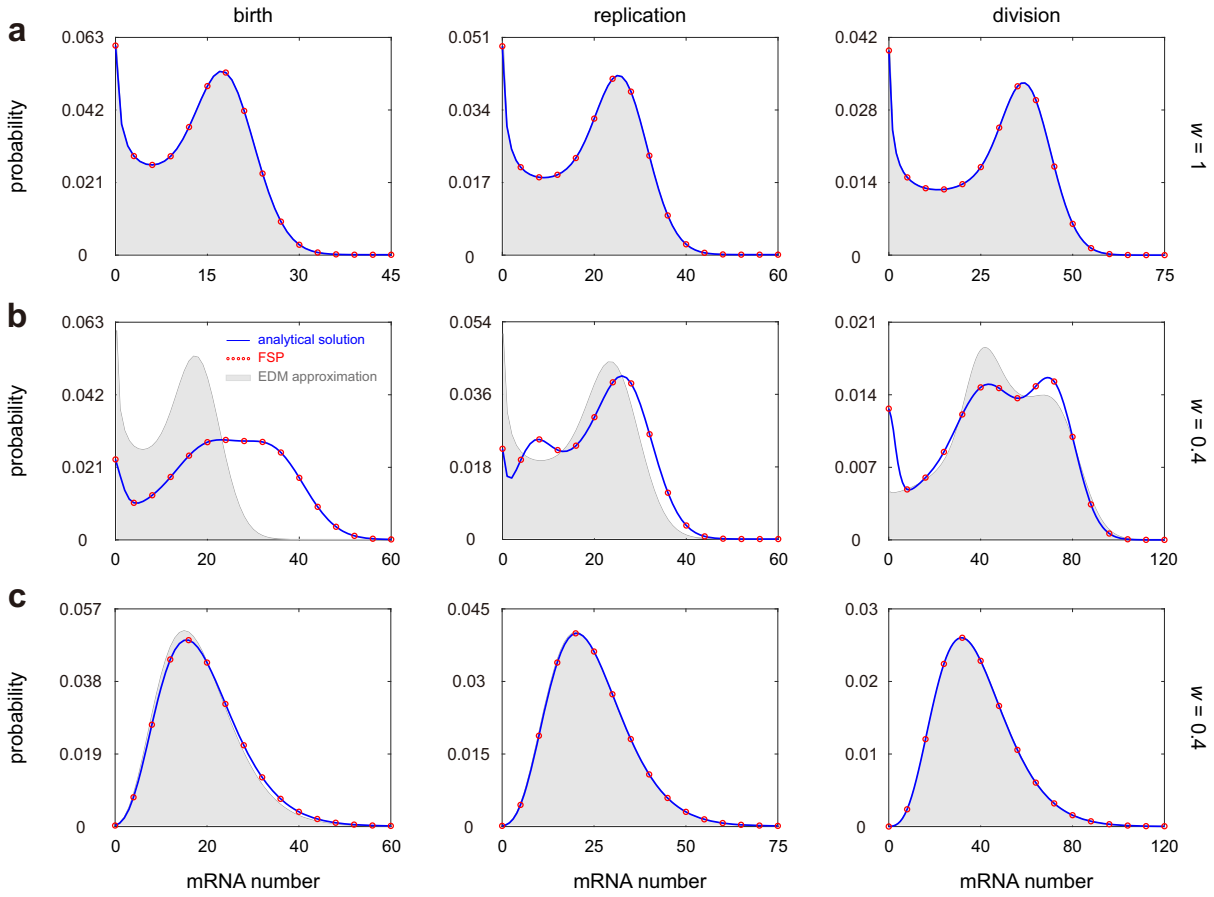

Figure 3: **Comparison between the full model and the EDM.** (**a**) Steady-state mRNA distributions at birth, replication, and division for the full model and the EDM when gene replication is not taken into account. The blue curves show the analytical distributions given in Eqs. (23) and (24), the red circles show the numerical ones obtained from FSP, and the grey regions show the distributions of the EDM. (**b**) Same as (a) but when gene replication is taken into account. In (a),(b), the model parameters are chosen as $V_b = 1, g = 1, \beta = 1, d = 4, \rho = 20 d_{\rm eff}, \sigma_0 = 1.5, \sigma_1 = 3, \sigma_1' = 2.4$. The parameter w is chosen as $w = 1$ in (a) and $w = 0.4$ in (b). (**c**) Same as (b) but in the special case where mRNA synthesis is balanced and bursty, and dosage compensation is perfect. The model parameters are chosen as $V_b = 1, g = 1, \beta = 1, w = 0.4, d = 4, \rho = 200 d_{\rm eff}, \sigma_0 = 300, \sigma_1 = 30, \sigma_1' = 15$.

for the ENM [48]. In the ENM, the mRNA distribution for a cell of constant volume V is exactly the one predicted by the EDM, and the fluctuations of cell volume V are regarded as extrinsic noise [56, 57]. In other words, the mRNA distribution for the ENM is given by

$$p^{\rm ENM}(n) = \int_0^\infty p^{\rm EDM}(n|V)\Pi(V)dV, \tag{33}$$

where $\Pi(V)$ is the distribution of cell volume. We emphasize here that the EDM varies depending on the number of gene copies and thus also depending on cell volume. For a cell of volume $V < 2^w V_b$, there is only one gene copy and the EDM is given by Eq. (29); for a cell of volume $V \geq 2^w V_b$, there are two gene copies and the EDM is given by Eq. (30). In addition, note that the distribution of cell volume is different for lineage and population measurements. Since cell volume $V(t)$ and cell age t are related by $V(t) = V_b e^{gt}$, the cell volume distribution can be obtained from the cell age distribution which has already been given above (see the paragraphs before Eqs. (31) and Eq. (32)). Specifically, the volume distribution for lineage measurements is given by [69]

$$\Pi(V) = \frac{1}{(\log 2)V}, \quad V_b \leq V \leq 2V_b,$$

and the volume distribution for population measurements is given by

$$\Pi(V) = \frac{2V_b}{V^2}, \quad V_b \leq V \leq 2V_b.$$

Inserting the above two equations into Eq. (33) gives the mRNA distribution for the ENM.

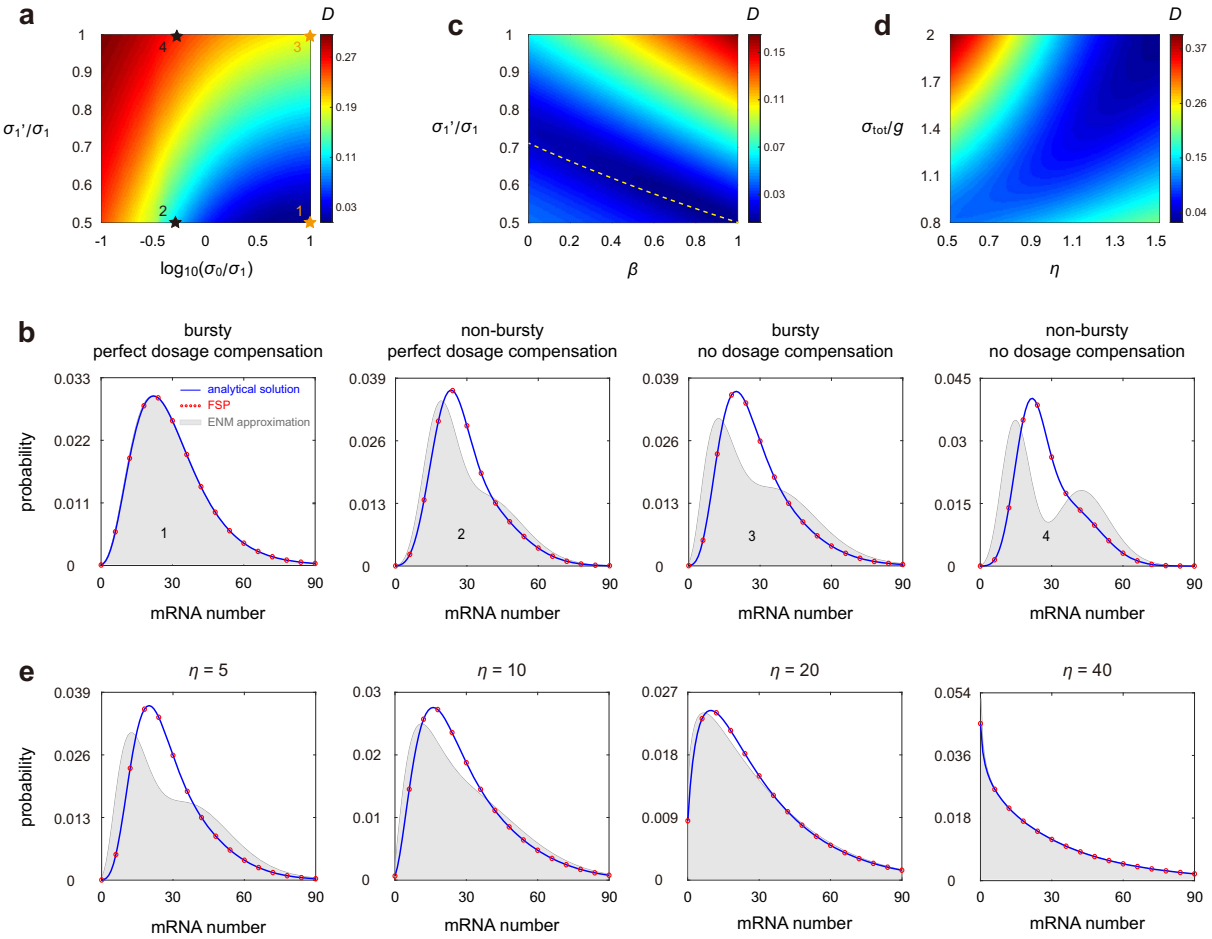

Figure 4: **Comparison between the full model and the ENM.** (**a**) Heat plot of the Hellinger distance D between lineage distributions for the full model and the ENM as σ_0/σ_1 and σ_1'/σ_1 vary. The model parameters are chosen as $V_b = 1, g = 1, \beta = 1, w = 0.5, d = 5, \sigma_1 = 30$. (**b**) Comparison between the lineage distributions for the full model and the ENM as σ_0/σ_1 and σ_1'/σ_1 vary. The blue curves show the analytical distributions for the full model given in Eq. (31), the red circles show the numerical ones obtained from FSP, and the grey regions show the distributions for the ENM. The model parameters are chosen as in (a). The parameter σ_0 is chosen as $\sigma_0 = 10\sigma_1$ (bursty case) and $\sigma_0 = 0.5\sigma_1$ (non-bursty case). The parameter σ_1' is chosen as $\sigma_1' = \sigma_1/2$ (perfect dosage compensation) and $\sigma_0 = \sigma_1$ (no dosage compensation). The parameters associated with the four panels are marked in (a) by stars. (**c**) Heat plot of D as β and σ_1'/σ_1 vary. The model parameters are chosen as $V_b = 1, g = 1, w = 0.5, d = 5, \sigma_0 = 300, \sigma_1 = 30$. (**d**) Heat plot of D as η and $\sigma_{\rm tot}/g$ vary. The model parameters are chosen as $V_b = 1, g = 1, \beta = 1, w = 0.5, \sigma_0 = 2.5\sigma_1, \sigma_1' = \sigma_1$. (**e**) The model parameters are chosen to be the same as in the third panel of (b) but η is varied. In (a)-(e) the parameter ρ is chosen so that $\langle n \rangle_{\rm lin} = 30$.

To evaluate the performance of the ENM approximation, we first illustrate the Hellinger distance D between the lineage distributions of the full model and the ENM as a function of σ_0/σ_1 and σ_1'/σ_1 when mRNA synthesis is balanced, i.e. $\beta = 1$ (Fig. 4(a)). It can be seen that the ENM serves as a good approximation when gene expression is bursty ($\sigma_0 \gg \sigma_1$) and when dosage compensation is perfect ($\sigma_1' = \sigma_1/2$). This is indeed a sufficient condition for mRNA to display concentration homeostasis when gene replication is taken into account [47]. A proof of this condition can be found in Supplementary Section 6. The breaking of either dosage compensation or bursty expression will lead to a significant deviation of the ENM from the full model (Fig. 4(b)). In particular, the distribution of the ENM model can show bimodality whereas that of the full model is unimodal.

It is still unclear how the ENM performs when mRNA synthesis is not balanced ($\beta < 1$). To see this, we further illustrate D as a function of β and σ_1'/σ_1 when gene expression is bursty (Fig. 4(c)). Interestingly, there is a region of parameter space (shown in dark blue) where D is minimised. In particular, when the mRNA synthesis rate is volume-independent ($\beta = 0$), the ENM works well when σ_1'/σ_1 is between 0.65 and 0.8. This shows

that to maintain the effectiveness of the ENM approximation, a lack of balanced mRNA synthesis requires also a lower degree of dosage compensation. Recent studies have shown that even when $\beta < 1$, strong concentration homeostasis (characterised by a small coefficient of variation of the mean concentrations across the cell cycle) can still be obtained when $\sigma_1'/\sigma_1 \approx 1/\sqrt{2}^{\beta+1}$ (shown by the yellow dashed line) and when replication occurs halfway through the cell cycle $(w = 0.5)$ [47]. Note that the region where D is minimized is exactly around the yellow dashed line. This shows that the effectiveness of the ENM approximation is closely related to concentration homeostasis even when $\beta < 1$.

To further confirm our results, we use the transcriptional parameters inferred in [25]. In this case, the mRNA distributions for two bursty genes *Oct4* and *Nanog* in mouse embryonic stem cells were measured as a function of time in the cell cycle from which all the rate parameters involved in our model were estimated. Since the cell-to-cell variability in volume within each cell-cycle phase was quite small, it was assumed that $\beta = 0$, i.e. the mRNA synthesis rate is volume-independent. Dosage compensation was found to be apparent for both genes, with σ_1'/σ_1 estimated to be 0.63 for *Oct4* and 0.71 for *Nanog*. Based on the inferred parameters, we compare the mRNA distributions of population measurements for the full model and the ENM (Supplementary Fig. S3(a)). We find that the ENM performs well for both genes. This agrees with our prediction that the ENM is valid when $\beta = 0$ and when σ_1'/σ_1 is around 0.7 (Fig. 4(c)). However, if we keep all rate parameters the same but reset σ_1'/σ_1 to 1 (no dosage compensation), then the EDM approximation will become significantly less satisfying (Supplementary Fig. S3(b)). This also coincides with the simulations shown in Fig. 4(c).

When mRNA synthesis is balanced and bursty, we have seen that the ENM approximation is accurate when dosage compensation is strong. However, in bacteria and budding yeast, there has been some evidence that dosage compensation is not widespread [49, 77]. It is unclear under what conditions the ENM is still valid when dosage compensation is weak. To see this, we also depict D as a function of $\eta = d/g$ and $\sigma_{\rm tot}/g$ when there is no dosage compensation, i.e. $\sigma_1' = \sigma_1$ (Fig. 4(d)). In this case, we find that the ENM still works well when the mRNA is very unstable $(d \gg g)$ and when the total gene switching rate is very large $(\sigma_{\rm tot} \gg g)$. This is fully consistent with our earlier theoretical predictions for the accuracy of the EDM, on which the ENM depends. In particular, when gene expression is bursty $(\sigma_0 \gg \sigma_1, g)$, increasing the mRNA degradation rate will give rise to a better ENM approximation (Fig. 4(e)). Note that this is not true when the total gene switching rate is slow (Fig. 4(d)). We emphasize that while Figs. 4(b),(e) show the mRNA distributions for lineage measurements, the same results are applicable for population measurements (Supplementary Fig. S4).

The value of $\eta = d/g$ can be determined experimentally since both d and g can be measured. In bacteria, η is typically between $6 - 30$, depending strongly on the strain and the growth condition; in yeast, it is typically between $3 - 8$; in mammalian cells, it is typically between $2 - 4$ (see Supplementary Section 5 for the median and range of η in various cell types). This suggests that the ENM approximation may be generally most useful in bacteria and less useful in yeast and mammalian cells.

Including stochasticity in cell cycle duration and cell size dynamics

Thus far, we have considered a detailed telegraph model of gene expression with a cell cycle description when the cell volume dynamics and the cell cycle duration are deterministic. However, in naturally occurring systems, the cell cycle duration is appreciably stochastic (see Fig. 1(c) of [46] for experimental distributions of cell cycle durations in eight different cell types). Moreover, there has been ample evidence [32–38] that the amount of growth produced during the cell cycle must be controlled such that, on average, larger cells at birth have shorter cell cycle durations than smaller ones. This mechanism maintains size homeostasis.

To model cell-cycle duration variability and size homeostasis, we use the size-additive autoregressive model of stochastic cell volume dynamics [35, 78]. The model assumes that the volume at birth V_b and the volume at division V_d are connected by the relation

$$V_d = \alpha V_b + (2 - \alpha)\,\bar{v} + \epsilon, \tag{34}$$

where $0 \leq \alpha \leq 2$ is the strength of size control, $\bar{v} > 0$ is the typical (average over generations) birth volume which is a time-independent constant, and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is a Gaussian noise term independent of V_b. The idea behind the model is as follows: upon being born with volume V_b, the cell attempts to grow for a period of time such that its target volume at division is $f(V_b) = \alpha V_b + (2-\alpha)\,\bar{v}$, but due to stochasticity, the actual volume at division may deviate from the target volume. Due to exponential cell growth, the cell cycle duration T is given by

$$T = \frac{1}{g} \log \frac{V_d}{V_b} = \frac{1}{g} \log \left[\alpha + \frac{(2-\alpha)\bar{v} + \epsilon}{V_b} \right], \tag{35}$$

where for simplicity we have assumed constant growth rate g across generations. This implies that on average, larger cells at birth have shorter cell cycle durations than smaller ones. Different size control strategies correspond to different values of α. When $\alpha = 0$, the target division volume $f(V_b) = 2\bar{v}$ is constant; this corresponds to the sizer strategy, where cells have to reach a certain size before division occurs. When $\alpha = 1$, the cell attempts to add a constant volume $f(V_b) - V_b = \bar{v}$ to its newborn size; this corresponds to the adder strategy. Since the growth is exponential, attempting to grow for a constant time is the same as having $f(V_b) = 2V_b$; hence $\alpha = 2$ corresponds to the timer strategy. The adder or near-adder behavior has been observed in bacteria, budding yeast, and mammalian cells [34, 36, 38], while fission yeast exhibits a near-sizer behavior [32].

When $\sigma_\epsilon = 0$, the model reduces to deterministic (previously considered) cell volume dynamics, in which case the timer, adder, and sizer strategies are exactly the same since $V_b = \bar{v}$ is a constant. As σ_ϵ increases, the time series of cell volume becomes much more noisy; however, it is difficult to identify whether there is a change in the magnitude of fluctuations solely from the time series of the mRNA number (Fig. 5(a)). Note that when σ_ϵ is small, the model produces a steady-state cell size distribution (from lineage simulations) characterized by three features: a fast increase in the size count for small cells, a slow decay for moderately large cells, and a fast decay for large cells (Fig. 5(b)). This is consistent with the cell size distribution in *E. coli* [69]. A natural question is what are the values of σ_ϵ in naturally occurring systems. To see this, we examined the publicly available lineage data of cell size in *E. coli* and fission yeast [35, 79] and found that the typical value of σ_ϵ is between $0.2\bar{v}$ and $0.3\bar{v}$ (see Supplementary Section 8 for a discussion about the inference of σ_ϵ and the estimated values of σ_ϵ in *E. coli* and fission yeast under different growth conditions).

To compute the mRNA distribution for stochastic cell volume dynamics, note that the evolution of the system within a cell cycle is controlled by four random variables: (i) the gene state at birth α_b, (ii) the mRNA number at birth N_b, (iii) the birth volume V_b, and (iv) the cell cycle duration T. Once the values of the four variables are fixed, the generating function F at any time $t \in [0, T]$ within a cell cycle is given by Eq. (10), i.e.

$$F(t, z|\alpha_b, N_b, V_b, T) = \begin{cases} \sum_{i=0}^{1} L_i(t, z|V_b) F_i(0, e^{-dt} z|\alpha_b, N_b), & t \in [0, wT], \\ \sum_{i=0}^{1} L_i'(t - wT, z|V_b)^2 F_i(wT, e^{-d(t-wT)} z|\alpha_b, N_b, V_b, T), & t \in [wT, T]. \end{cases} \tag{36}$$

Here the initial conditions $F_i(0, z)$, $i = 0, 1$ are determined by α_b and N_b as

$$F_{\alpha_b}(0, z|\alpha_b, N_b) = z^{N_b}, \quad F_{1-\alpha_b}(0, z|\alpha_b, N_b) = 0.$$

The functions L_i and L_i', $i = 0, 1$ given in Eq. (9) depend on V_b since the parameters u and u' are functions of V_b; the replication time wT depends on T. Hence the generating function F depends on all the four variables. Once we know the joint distribution of the four variables in some generation, we can use Eq. (34) to compute their joint distribution in the next generation. In this way, we obtain the full time-dependence of the mRNA distribution cross cell cycles. In Supplementary Section 8, we have generalized the analytical results obtained previously to the model with stochastic cell volume dynamics. Specifically, we have derived the exact time-dependent mRNA distribution for a cell of any age in any generation, as well as the exact steady-state distribution for lineage measurements.

To reveal the influence of cell-cycle duration variability and size homeostasis on gene expression, we compare the lineage distributions for the model with deterministic cell size dynamics and the model with stochastic cell size

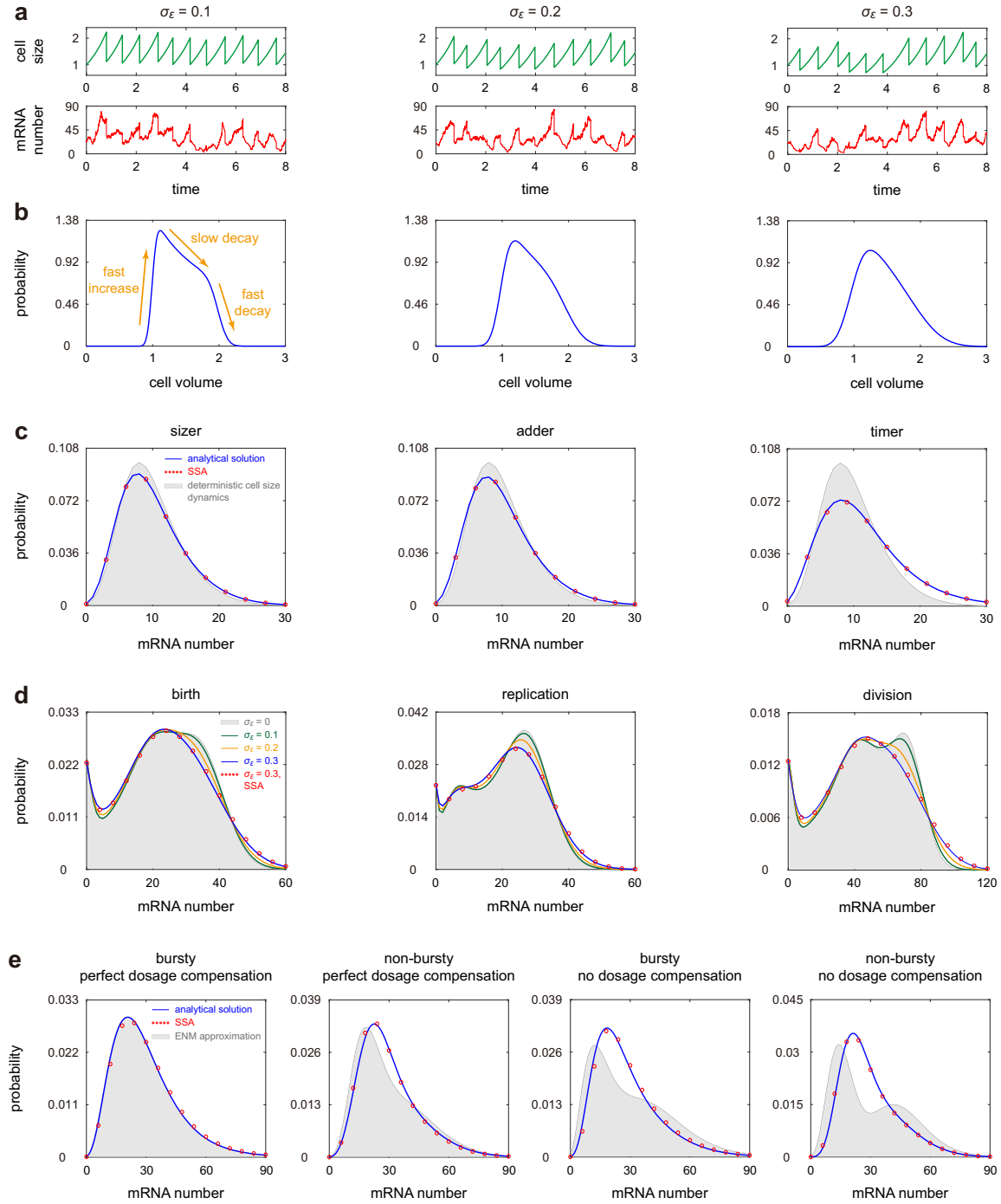

Figure 5: **Effects of stochastic cell volume dynamics on mRNA fluctuations.** (**a**) Typical trajectories of cell size and mRNA number as σ_ϵ increases. (**b**) Cell volume distribution of lineage measurements as σ_ϵ increases. In (a),(b), the model parameters are chosen as $\bar{v} = 1, g = 1, \beta = 1, w = 0.4, d = 4, \rho = 20 d_{\rm eff}, \sigma_0 = 1.5, \sigma_1 = 3, \sigma_1' = 2.4, \alpha = 1$. (**c**) Comparison between the steady-state mRNA distributions of lineage measurements for deterministic and stochastic cell size dynamics under different size control strategies. The blue curves show the analytical distributions for stochastic cell size dynamics, the red circles show the numerical ones obtained from the SSA, and the grey regions show the distributions for deterministic cell size dynamics. The model parameters are chosen as $\bar{v} = 1, g = 1, \beta = 1, w = 0.5, d = 5, \sigma_0 = \sigma_1 = 100, \sigma_1' = 50, \sigma_\epsilon = 0.4$. The parameter ρ is chosen so that $\langle n \rangle_{\rm lin} = 10$ for deterministic cell size dynamics. Previous studies [80] have shown that the timer strategy with $\alpha = 2$ is not stable since it cannot produce a finite and nonzero mean of cell volume. Hence we choose $\alpha = 1.8$ for the timer strategy here. (**d**) Steady-state mRNA distributions at birth, replication, and division as σ_ϵ increases. The model parameters are the same as in (a),(b). (**e**) Comparison between the lineage distributions of the full model and the ENM for stochastic cell size dynamics. The grey regions show the distributions for the ENM. The model parameters are chosen to be the same as in Fig. 4(b).

dynamics under different size control strategies (Fig. 5(c)). The two distributions deviate remarkably from each other for the timer strategy, but the deviation is much smaller for the adder and sizer strategies. This demonstrates the advantage of the adder and sizer strategies in reducing gene expression noise. In addition, Fig. 5(d) illustrates the steady-state mRNA distribution at three different points (birth, replication, and division) across the cell cycle as noise in cell size dynamics, characterized by σ_ϵ, varies. Clearly, larger noise in cell size results in larger noise in gene expression, as expected. A sufficiently large σ_ϵ may even change the number of modes of the mRNA distribution. Interestingly, we find that when the mRNA distribution exhibits multimodality, increasing σ_ϵ will not change the height of the zero peak but may affect the height and position of non-zero peaks (Fig. 5(d)).

Finally we investigate the accuracy of the ENM approximation for stochastic cell volume dynamics. Note that we can no longer use the EDM to approximate the mRNA distributions at birth, replication, and division, since the cell volumes are stochastic. We compare the steady-state mRNA distributions at birth, replication, and division for the full model with their ENM approximations in Supplementary Fig. S5 and also compare the lineage distribution for the full model with its ENM approximation in Fig. 5(e) (see Supplementary Section 8 for the analytical expressions of the ENM approximations). The model parameters in the two figures are chosen to be the same as in Figs. 3(b),(c) and 4(b), respectively. We can see that in the presence of fluctuations in cell volume, the results of the present paper are still valid — the ENM does not work in general but performs well when mRNA synthesis is balanced and bursty and when dosage compensation is perfect. Comparing Supplementary Fig. S5 with Fig. 3(b),(c) and comparing Fig. 5(e) with Fig. 4(b), we also find that the differences between the mRNA distributions for the full model and the ENM are slightly diminished when the cell volume dynamics is stochastic.

Discussion

In this work, we analytically solved a detailed model of stochastic gene expression with cell cycle and cell volume descriptions including gene switching, cell growth, cell division, volume-dependent transcription, gene replication, and gene dosage compensation. We first considered the case where the cell volume dynamics is deterministic and then generalized the results to include cell-cycle duration variability and cell-size control strategies. Previous models of stochastic mRNA dynamics in growing and dividing cells [20, 46] can be seen as special cases of the present modelling framework. In addition, we emphasize that our model not only characterizes the mRNA dynamics, but can also be used to describe the protein dynamics. For example, when gene expression is bursty and when the degradation rate is taken to be zero, our model reduces to the effective one-state model of the protein dynamics proposed in [40, 41, 45]. If the intrinsic noise due to the random birth-death of transcripts is ignored, then our model reduces to the one-state model studied in [39]. Our work is also distinctive from recent related work [48] since our derivations of the distributions of mRNA numbers as a function of cell age and generation number, and of the distributions in steady-state growth do not need the assumption of stochastic concentration homeostasis (SCH); the relaxation of this assumption is crucial to model the variation of gene copy numbers across a cell cycle due to DNA replication. We have also investigated how well can the model be approximated by the effective dilution and extrinsic noise models (EDM/ENM). When gene replication is taken into account, we showed that the mRNA distributions of the full model may differ significantly from the predictions of the EDM/ENM. We elucidated three cases where the EDM/ENM makes accurate approximations.

The first case takes place when the mRNA is very unstable and the total gene switching rate (the sum of the gene activation and inactivation rates) is very large such that on the timescale of volume change, the mRNA distribution instantaneously equilibrates. This condition is intuitive and has been discussed in earlier work [56]. However as we showed using data from various cell types, the typical mRNA lifetime in eukaryotes (especially mammalian cells) is generally not small enough compared to the cell cycle duration to enforce instantaneous equilibrium; rather the fluctuations have memory of birth and replication events.

The second case occurs when mRNA synthesis is balanced and bursty, and when dosage compensation is perfect. While our model does not generally obey SCH due to gene copy number variation upon replication,

however in this case parameter conditions effectively enforce SCH. Note that if expression is balanced and it is bursty with weak dosage compensation or else it is constitutive with perfect dosage compensation, there is an apparent breakdown of the EDM/ENM's ability to accurately approximate the full model. This is since in these cases the dependence of the mean mRNA numbers with cellular volume is significantly influenced by the doubling of gene copy numbers at replication. Examples where expression is balanced but the effects of replication are not completely buffered by dosage compensation are starting to be uncovered, e.g. in human cells while the overall mRNA synthesis rates increase with cell volume, however S/G2-phase cells show increased synthesis rates compared to G1-phase cells of the same volume [81]. As pointed out in [60], this is reminiscent of a step-increase in RNA production during or after S phase which was previously observed in synchronized HeLa cell populations and other organisms [82] – this suggests that perfect dosage compensation in mammalian cells may not be common.

The third case is when mRNA synthesis is non-balanced and bursty, and when dosage compensation is of an intermediate strength such that concentration homeostasis is approximately maintained, i.e. there is only a small variation of the mean mRNA concentration throughout the cell cycle — note that this is a much weaker condition than SCH. We showed that this is indeed the case for two genes, *Oct4* and *Nanog* in mouse embryonic stem cells, whose parameters have been previously estimated before and after gene replication [25].

In summary, our work shows that caution is needed when the ENM is applied to explain data collected in growing and dividing cells and that the accuracy of this reduced model of gene expression cannot be *a priori* assumed genome-wide. Our model, though detailed, has some limitations. We have focused on models that explain cell-to-cell variability in the synthesis rates due to their dependence on cell volume. However, likely other descriptors of cell state (such as shape, local cell crowding, mitochondrial abundance, capacity to respond to Ca^{2+}) can explain a higher degree of cell-to-cell variability than size alone [83, 84]. Also it is the case that here we have considered the expression of unregulated genes but it is well known that many genes regulate each other resulting in complex gene regulatory networks [85]. Overcoming the latter limitation is particularly pressing but it is analytically challenging because such models have nonlinear propensities stemming from the modelling of bimolecular interactions between transcriptional factors and genes [86]. Progress in this direction will be reported in a separate paper.

Data and Code Availability

MATLAB codes of the FSP algorithm are found on GitHub https://github.com/chenjiacsrc/telegraph-model.

Acknowledgements

We thank Prof. Qiwen Sun and Prof. Feng Jiao for pointing us to useful references. C. J. acknowledges support from National Natural Science Foundation of China with grant No. U1930402 and grant No. 12131005. R. G. acknowledges support from the Leverhulme Trust (RPG-2020-327).

References

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

[2] Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).

[3] Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).

[4] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences* **99**, 12795–12800 (2002).

[5] Lenstra, T. L., Rodriguez, J., Chen, H. & Larson, D. R. Transcription dynamics in living cells. *Annual review of biophysics* **45**, 25–47 (2016).

[6] Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).

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

[8] Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).

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

[10] Ko, M. S. A stochastic model for gene induction. *J. Theor. Biol.* **153**, 181–194 (1991).

[11] Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).

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

[13] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).

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

[15] Chen, J. & Jiao, F. A novel approach for calculating exact forms of mRNA distribution in single-cell measurements. *Mathematics* **10**, 27 (2021).

[16] Jia, C. & Li, Y. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. *bioRxiv* (2022).

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

[18] Berg, O. G. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71**, 587–603 (1978).

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

[20] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. USA* **117**, 4682–4692 (2020).

[21] Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM J. Appl. Math.* **80**, 1336–1355 (2020).

[22] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

[23] Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

[24] Singer, Z. S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Molecular cell* **55**, 319–331 (2014).

[25] Skinner, S. O. *et al.* Single-cell analysis of transcription kinetics across the cell cycle. *Elife* **5**, e12175 (2016).

[26] Huh, D. & Paulsson, J. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* **43**, 95 (2011).

[27] Zhurinsky, J. *et al.* A coordinated global control over cellular transcription. *Curr. Biol.* **20** (2010).

[28] Sun, X.-M. *et al.* Size-dependent increase in RNA Polymerase II initiation rates mediates gene expression scaling with cell size. *Curr. Biol.* **30**, 1217–1230 (2020).

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

[30] Kempe, H., Schwabe, A., Crémazy, F., Verschure, P. J. & Bruggeman, F. J. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Mol. Biol. Cell* **26**, 797–804 (2015).

[31] Ietswaart, R., Rosa, S., Wu, Z., Dean, C. & Howard, M. Cell-size-dependent transcription of FLC and its antisense long non-coding RNA COOLAIR explain cell-to-cell expression variation. *Cell Syst.* **4**, 622–635 (2017).

[32] Fantes, P. & Nurse, P. Control of cell size at division in fission yeast by a growth-modulated size control over nuclear division. *Exp. Cell Res.* **107**, 377–386 (1977).

[33] Campos, M. *et al.* A constant size extension drives bacterial cell size homeostasis. *Cell* **159**, 1433–1446 (2014).

[34] Taheri-Araghi, S. *et al.* Cell-size control and homeostasis in bacteria. *Curr. Biol.* **25**, 385–391 (2015).

[35] Tanouchi, Y. *et al.* A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature* **523**, 357–360 (2015).

[36] Soifer, I., Robert, L. & Amir, A. Single-cell analysis of growth in budding yeast and bacteria reveals a common size regulation strategy. *Curr. Biol.* **26**, 356–361 (2016).

[37] Facchetti, G., Chang, F. & Howard, M. Controlling cell size through sizer mechanisms. *Curr. Opin. Syst. Biol.* **5**, 86–92 (2017).

[38] Cadart, C. *et al.* Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nat. Commun.* **9**, 1–15 (2018).

[39] Antunes, D. & Singh, A. Quantifying gene expression variability arising from randomness in cell division times. *J. Math. Biol.* **71**, 437–463 (2015).

[40] Soltani, M., Vargas-Garcia, C. A., Antunes, D. & Singh, A. Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLoS Comput. Biol.* **12**, e1004972 (2016).

[41] Soltani, M. & Singh, A. Effects of cell-cycle-dependent expression on random fluctuations in protein levels. *R. Soc. Open Sci.* **3**, 160578 (2016).

[42] Sun, Q., Jiao, F., Lin, G., Yu, J. & Tang, M. The nonlinear dynamics and fluctuations of mRNA levels in cell cycle coupled transcription. *PLoS Comput. Biol.* **15**, e1007017 (2019).

[43] Dessalles, R., Fromion, V. & Robert, P. Models of protein production along the cell cycle: An investigation of possible sources of noise. *PLoS one* **15**, e0226016 (2020).

[44] Jedrak, J., Kwiatkowski, M. & Ochab-Marcinek, A. Exactly solvable model of gene expression in a proliferating bacterial cell population with stochastic protein bursts and protein partitioning. *Phys. Rev. E* **99**, 042416 (2019).

[45] Beentjes, C. H., Perez-Carrasco, R. & Grima, R. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Phys. Rev. E* **101**, 032403 (2020).

[46] Perez-Carrasco, R., Beentjes, C. & Grima, R. Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. *J. R. Soc. Interface* **17**, 20200360 (2020).

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

[48] Thomas, P. & Shahrezaei, V. Coordination of gene expression noise with cell size: extrinsic noise versus agent-based models of growing cell populations. *J. R. Soc. Interface* **18**, 20210274 (2021).

[49] Wang, M., Zhang, J., Xu, H. & Golding, I. Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality. *Nat. Microbiol.* **4**, 2118–2127 (2019).

[50] Kalita, I., Iosub, I. A., Granneman, S. & El Karoui, M. Fine-tuning of RecBCD expression by post-transcriptional regulation is required for optimal DNA repair in Escherichia coli. *bioRxiv* (2021).

[51] Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends in Genetics* **28**, 560–565 (2012).

[52] Neurohr, G. E. *et al.* Excessive cell growth causes cytoplasm dilution and contributes to senescence. *cell* **176**, 1083–1097 (2019).

[53] Dolatabadi, S. *et al.* Cell cycle and cell size dependent gene expression reveals distinct subpopulations at single-cell level. *Frontiers in genetics* **8**, 1 (2017).

[54] Swaffer, M. P. *et al.* Transcriptional and chromatin-based partitioning mechanisms uncouple protein scaling from cell size. *Molecular Cell* **81**, 4861–4875 (2021).

[55] Sherman, M. S., Lorenz, K., Lanier, M. H. & Cohen, B. A. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell systems* **1**, 315–325 (2015).

[56] Ham, L., Brackston, R. D. & Stumpf, M. P. Extrinsic noise and heavy-tailed laws in gene expression. *Phys. Rev. Lett.* **124**, 108101 (2020).

[57] Ham, L., Jackson, M. & Stumpf, M. P. Pathway dynamics can delineate the sources of transcriptional noise in gene expression. *Elife* **10**, e69324 (2021).

[58] Wang, P. *et al.* Robust growth of Escherichia coli. *Curr. Biol.* **20**, 1099–1103 (2010).

[59] Eun, Y.-J. *et al.* Archaeal cells share common size control with bacteria despite noisier growth and division. *Nat. Microbiol.* **3**, 148–154 (2018).

[60] Berry, S. & Pelkmans, L. Mechanisms of cellular mRNA transcript homeostasis. *Trends in Cell Biology* (2022).

[61] Wang, Q. & Lin, J. Heterogeneous recruitment abilities to RNA polymerases generate nonlinear scaling of gene expression with cell volume. *Nature communications* **12**, 1–11 (2021).

[62] Dowling, M. R. *et al.* Stretched cell cycle model for proliferating lymphocytes. *Proc. Natl. Acad. Sci. USA* **111**, 6377–6382 (2014).

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

[64] Sepúlveda, L. A., Xu, H., Zhang, J., Wang, M. & Golding, I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science* **351**, 1218–1222 (2016).

[65] Jia, C. & Grima, R. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation. *Phys. Rev. X* **11**, 021032 (2021).

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

[67] Reverón-Gómez, N. *et al.* Accurate recycling of parental histones reproduces the histone modification landscape during DNA replication. *Mol. Cell* **72**, 239–249 (2018).

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

[69] Jia, C., Singh, A. & Grima, R. Cell size distribution of lineage data: analytic results and parameter inference. *iScience* **24**, 102220 (2021).

[70] Jia, C., Singh, A. & Grima, R. Characterizing non-exponential growth and bimodal cell size distributions in fission yeast: An analytical approach. *PLoS Comput. Biol.* **18**, e1009793 (2022).

[71] Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2**, 157–175 (2005).

[72] Jia, C., Zhang, M. Q. & Hong, Q. Emergent Lévy behavior in single-cell stochastic gene expression. *Phys. Rev. E* **96**, 040402(R) (2017).

[73] Jia, C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. *Phys. Rev. E* **96**, 032402 (2017).

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

[75] Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302 (2006).

[76] Jia, C., Qian, H., Chen, M. & Zhang, M. Q. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. *J. Chem. Phys.* **148**, 095102 (2018).

[77] Torres, E. M., Springer, M. & Amon, A. No current evidence for widespread dosage compensation in S. cerevisiae. *Elife* **5**, e10996 (2016).

[78] Amir, A. Cell size regulation in bacteria. *Phys. Rev. Lett.* **112**, 208102 (2014).

[79] Nakaoka, H. & Wakamoto, Y. Aging, mortality, and the fast growth trade-off of Schizosaccharomyces pombe. *PLoS Biol.* **15**, e2001109 (2017).

[80] Vargas-Garcia, C. A., Soltani, M. & Singh, A. Conditions for cell size homeostasis: a stochastic hybrid system approach. *IEEE Life Sci. Lett.* **2**, 47–50 (2016).

[81] Berry, S., Müller, M., Rai, A. & Pelkmans, L. Feedback from nuclear RNA on transcription promotes robust RNA concentration homeostasis in human cells. *Cell Systems* (2022).

[82] Mitchison, J. Growth during the cell cycle. *International review of cytology* 166–258 (2003).

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

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

[85] Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics* **2**, 268–279 (2001).

[86] Cao, Z. & Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 3305 (2018).