Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics

Casper H. L. Beentjes ⬤
Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom

Ruben Perez-Carrasco ⬤
Department of Mathematics, University College London, London WC1H 0AY, United Kingdom

Ramon Grima[*]
School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

The bulk of stochastic gene expression models in the literature do not have an explicit description of the age of a cell within a generation and hence they cannot capture events such as cell division and DNA replication. Instead, many models incorporate the cell cycle implicitly by assuming that dilution due to cell division can be described by an effective decay reaction with first-order kinetics. If it is further assumed that protein production occurs in bursts, then the stationary protein distribution is a negative binomial. Here we seek to understand how accurate these implicit models are when compared with more detailed models of stochastic gene expression. We derive the exact stationary solution of the chemical master equation describing bursty protein dynamics, binomial partitioning at mitosis, age-dependent transcription dynamics including replication, and random interdivision times sampled from Erlang or more general distributions; the solution is different for single lineage and population snapshot settings. We show that protein distributions are well approximated by the solution of implicit models (a negative binomial) when the mean number of mRNAs produced per cycle is low and the cell cycle length variability is large. When these conditions are not met, the distributions are either almost bimodal or else display very flat regions near the mode and cannot be described by implicit models. We also show that for genes with low transcription rates, the size of protein noise has a strong dependence on the replication time, it is almost independent of cell cycle variability for lineage measurements, and increases with cell cycle variability for population snapshot measurements. In contrast for large transcription rates, the size of protein noise is independent of replication time and increases with cell cycle variability for both lineage and population measurements.

I. INTRODUCTION

It is well known that gene expression is stochastic [1]. The randomness in the time at which each reaction occurs leads to fluctuations in the molecule number of gene products such as mRNA and proteins. Hence, over the past two decades there has been considerable effort devoted to constructing and solving stochastic models of gene expression [2,3]. The exact solution of the chemical master equation (CME) describing the standard models of stochastic gene expression is currently unknown except in certain limiting cases such as when mRNA degrades much faster than protein [4].

The majority of gene expression models in the literature do not have a description of cellular age and hence do not explicitly describe the cell cycle [4–12]. Rather, it is assumed, following [13,14], that protein dilution effects due to cell division can be implicitly included via an effective first-order decay reaction. The rate of this reaction is chosen such that the

half-life of protein numbers corresponds to the mean cell cycle length. This approximation is thought to be reasonable since active protein degradation timescales are considerably longer than the cell cycle time [15,16] and hence dilution occurring during cell division is the dominant means of protein removal.

Since these effective models do not have a description of the cell age within a cell cycle, they also cannot take into account events which happen at specific points during the cycle, e.g., the replication of the genome which leads to an increase of the transcription rate. The main advantage of these models is the relative ease with which they can be analytically solved, approximated, and simulated. In particular, the chemical master equation of the most commonly used model of this type, which describes proteins produced in bursts whose size is sampled from the geometric distribution and protein decay via an effective first-order reaction (modeling dilution as described above), can be solved exactly, leading to a negative binomial distribution (or a gamma distribution, its continuous analog) of protein numbers [4,14].

In contrast to this implicit model, more sophisticated models have been developed during the past few years that include

---
[*]ramon.grima@ed.ac.uk

an explicit description of the cell cycle. Curiously, results for the case of periodic cell division were first obtained by Berg [17], about 20 years before the explosion of interest in stochastic gene expression [1]. More recently, Johnston and Jones obtained the distribution of protein numbers assuming nonbursty production, binomial partitioning at cell division, and regularly spaced (periodic) cell division events [18]. Since experimental data clearly show that the time between two successive cell division effects is a random variable [19,20], models were also devised to study how dynamics are influenced by this extra source of randomness. Antunes and Singh [21] obtained the moments of mRNA and protein numbers in a simplified model of gene expression which ignores intrinsic noise (due to the stochastic birth-death of individual molecules) and that due to binomial partitioning but takes into account noise stemming from the random timing of cell division events. Soltani *et al.* [22,23] obtained the mean and variance of protein numbers in a considerably more detailed model for stable protein (one that is degraded only by dilution) that includes intrinsic noise, stochastic partitioning of molecules at cell division, a cell cycle that is divided in a number of phases whose duration is exponentially distributed and also can include replication. In these studies, the formulas are obtained assuming single lineage measurements, i.e., upon cell division, one of the daughter cells is followed (the other discarded) such that one obtains information about the stochastic dynamics of a cell's protein contents along an arbitrarily chosen lineage (measurements done using a mother machine such as in Ref. [24]). The two major disadvantages of the latter two papers are that they do not derive results for the protein number distributions and also they do not calculate statistics in a growing population of cells, the most common experimental scenario (measurements done using flow cytometry such as [25]). Jędrak *et al.* [26] derived an explicit expression for the protein distribution solution of a stochastic model where protein fluctuations are treated continuously, there are no gene duplication effects, and where the cell cycle is assumed to be exponentially distributed. These results analyze the behavior of a single lineage (as previous papers), and also for a whole proliferating population, i.e., both daughter cells are followed upon cell division such that one obtains information about the stochastic dynamics of a cell's protein contents across a growing population. Note that lineage and population statistics are not generally equivalent, unlike what one may assume based on the ergodic hypothesis [27]. The major limitations of the model in Ref. [26] are the large protein approximation implicit in the continuous approximation, the lack of DNA replication, and the assumption that cell cycle duration is exponentially distributed, which is contrary to experimental evidence that reveals distributions comparable to Erlang, gamma, or lognormal distributions or variations thereof [20,28].

Given these two different approaches including implicit and explicit description of the cell cycle, a question arises: How well can the negative binomial distribution of implicit models describe the protein distribution of more detailed models of gene expression? This question remains unanswered because, as discussed above, none of the current literature derives the protein distribution in a model that explicitly includes dilution due to stochastic partitioning of molecules at

cell division, random interdivision times, and age-dependent transcription. In this paper, we answer this question by deriving expressions for the distributions of proteins in models that incorporate explicit descriptions of the cell cycle. For the sake of clarity, instead of starting from the most general model, our presentation considers a set of simpler models which gradually build up to it. The three models that we study, in order of complexity, have the following properties: (i) no replication and a cell cycle of fixed duration; (ii) no replication and an Erlang distributed cell cycle duration; (iii) age-dependent transcriptional dynamics including replication and a cell cycle described by a number of phases, each of which has an exponentially distributed duration (hypoexponential distribution). All the models consider proteins that are produced in bursts [29], degraded only via dilution (stable proteins [14]) and assume binomial partitioning of proteins at cell division [17]. We study the relationships between the solutions of all models for both single lineage and population snapshot statistics, and identify conditions under which the protein distributions can be well approximated by the negative binomial solution of the conventional model of gene expression with an implicit cell cycle description.

II. MODEL I: STOCHASTIC GENE EXPRESSION WITH AN IMPLICIT DESCRIPTION OF THE CELL CYCLE

It is well known that under the assumptions that mRNA degrades much faster than protein and that promoter switching is also much faster than protein decay, the stochastic dynamics of protein P can be effectively described by the reaction scheme [4]

$$G \xrightarrow{r} G + mP, \quad P \xrightarrow{d} \emptyset, \tag{1}$$
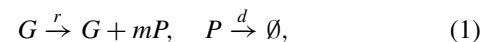
where G denotes a single gene copy, r is the effective burst production rate, and d is the protein degradation rate. Note that protein is produced in bursts of size m, which follows a geometric distribution (in accordance with experiments [29]), i.e., $m \sim \text{Geom}(p)$ with a mean burst size $\alpha = (1 - p)/p = h/d_m$ where h is the protein translation rate and d_m is the mRNA degradation rate. This burstiness implicitly describes the mRNA dynamics since a burst in protein expression occurs due to rapid translation of proteins by a single short-lived mRNA. Hence, within this context, the parameter r is also the same as the effective mRNA transcription rate which is given by $r = \rho_u \sigma_u / (\sigma_b + \sigma_u)$ where ρ_u is the mRNA transcription rate, σ_b is the rate of switching from the active state to the inactive state, and σ_u is the rate of switching from the inactive state to the active state.

Note that this model has been rigorously derived from a three-stage model of gene expression that does not take the cell cycle explicitly into account [4]. However, it is commonly assumed that the protein degradation reaction effectively models the dilution which occurs due to binomial partitioning at cell division. The question then is how we should choose the effective protein degradation rate. A simple argument is as follows. Since the protein decays exponentially via an effective first-order reaction, its half-life is $t_d = \log(2)/d$; we also know that the protein concentration is on average halved at cell division due to binomial partitioning and hence $t_d = T$, where T is the cell cycle length. Hence, it follows

that the effective protein degradation rate should be chosen as $d = \log(2)/T$. The effective model given by reaction scheme (1) and d chosen as aforementioned is one of the standard models of gene expression in the literature [14].

In Appendix A 1, we show that the effective model described above provides an accurate description of the mean number of proteins in a three-stage model of gene expression with explicit mRNA and protein dynamics, binomial partitioning, and fixed cell cycle length T, provided (i) the mRNA degrades much faster than protein; (ii) promoter switching is much faster than protein decay; (iii) mRNA degrades on a much shorter timescale than the cell cycle length T; (iv) the mean number of proteins is calculated from population measurements. Note that if instead the mean number of proteins is calculated from single lineage measurements then we arrive at model (1) but with effective protein degradation rate given by $d = (\frac{2}{3})/T$ (see Appendix A). Note that effective degradation rates derived for lineage data are slightly smaller than their population snapshot equivalent because $\frac{2}{3} < \log 2$. This discrepancy stems from the fact that the mean number of proteins calculated from population measurements is smaller than the mean number of proteins calculated in single lineage measurements. This is because in population snapshots, all cells are tracked and hence due to a doubling of the number of cells at cell division, there is a bias toward observing young cells with small protein counts. In contrast, in lineage measurements since only one cell is tracked, the probability of observing a cell of any age is the same and hence the protein counts on average are higher than in population measurements. A detailed discussion of the difference between these two types of measurement can be found in [27].

The CME of model I whose reactions are given by (1) is straightforward to solve using the method of generating functions. In steady state, its solution reads as

$$G(z) = \left[\frac{1}{1 - \alpha(z - 1)} \right]^{\beta}, \qquad (2)$$

where $G(z) = \sum_n z^n P(n)$ is the probability generating function (PGF), $P(n)$ is the steady-state protein distribution of protein number n, and $\beta = r/d$. Note that β is the average number of mRNA molecules produced in the protein lifetime. In the case of stable proteins (which degrade only by dilution) it follows from our previous results for effective degradation rates that $\beta = 3y/2$ for lineage measurements and $\beta = y/\log 2$ for population snapshot measurements, where we have defined $y = rT$ as the average number of mRNA produced in a cell cycle. The distribution can be obtained using $P(n) = (1/n!)dG^n/dz^n|_{z=0}$ which leads to a negative binomial $NB(\beta, \alpha/(1 + \alpha))$.

III. MODEL II: STOCHASTIC GENE EXPRESSION WITH EXPLICIT MODELING OF A FIXED LENGTH CELL CYCLE

Next, we consider a model where protein production occurs in bursts as in model I but there is no effective first-order reaction modeling protein degradation. Instead, we explicitly model binomial partitioning of the proteins at cell division. The major assumption of this model is that cell division occurs

at regular time intervals of length T. This is often referred to as a "timer" mechanism and has been found in certain types of cells, e.g., early frog embryos [30]. In what follows, we will find an exact steady-state solution of the CME for this model and compare it with that of model I.

Let $t \in [0, T]$ be the age of a given cell, namely, $t = 0$ corresponds to its birth and $t = T$ corresponds to the time at which it divides into two. The CME describing bursty protein expression and no active degradation in a cell is given by

$$\frac{dP^j(n, t)}{dt} = r \sum_{m=0}^{\infty} P^j(n - m, t)Q(m) - rP^j(n, t) \sum_{m=0}^{\infty} Q(m), \qquad (3)$$

where $P^j(n, t)$ is the probability that at cell age t in generation j there are n proteins observed. Here, $Q(m) = p(1 - p)^m$ with $\alpha = (1 - p)/p$ is the geometric distribution with mean α. Note that each time cell division occurs, the generation number j is increased by one. The PGF equation corresponding to the CME is

$$\frac{\partial G^j(z, t)}{\partial t} = -rG^j(z, t)\left[1 - \frac{1}{1 + \alpha(1 - z)} \right], \qquad (4)$$

which has a time-dependent solution

$$G^j(z, t) = F^j(z) \exp\left[\frac{-\alpha rt(1 - z)}{1 + \alpha(1 - z)} \right]. \qquad (5)$$

Note that $F^j(z) = \sum_n z^n P^j(n, 0)$, namely, the PGF corresponding to the protein distribution at cell birth in generation j.

Introducing binomial partitioning at mitosis leads to a simple relationship between the protein distribution at cell division of a cell in generation j and the distribution observed at the birth of the daughter cell in generation $j + 1$,

$$P^{j+1}(n, 0) = \sum_{i=0}^{\infty} \binom{i}{n} 2^{-i} P^j(i, T), \qquad (6)$$

which implies for the PGF

$$F^{j+1}(z) = G^{j+1}(z, 0) = G^j\left(\frac{1 + z}{2}, T \right). \qquad (7)$$

Note that in Eq. (6) we used the convention that i choose n equals zero when $n > i$. In this case, we cannot impose steady state as in model I because cell division occurs at regular time intervals. Rather, we consider cyclostationary conditions which are achieved when the probability that a cell of age t has a given number of proteins is independent of which generation it belongs to, i.e., the superscript j in Eqs. (3)–(7) can be ignored. Hence, substituting Eq. (7) in Eq. (5) we obtain

$$G(z, t) = G\left(\frac{1 + z}{2}, T \right) \exp\left[\frac{-\alpha rt(1 - z)}{1 + \alpha(1 - z)} \right]. \qquad (8)$$

Next, we proceed to solve Eq. (8) by substituting $t = T$ in this equation to obtain

$$G(z, T) = G\left(\frac{1 + z}{2}, T \right) f(z), \qquad (9)$$

where $f(z) = \exp\{-\alpha rT(1-z)/[1 + \alpha(1-z)]\}$. Equation (9) can be solved by iteration as follows:

$$
\begin{aligned}
G(z,T) &= f(z)f\left(\frac{1+z}{2}\right)G\left(\frac{3+z}{4},T\right) \\
&= f(z)f\left(\frac{1+z}{2}\right)f\left(\frac{3+z}{4}\right)G\left(\frac{7+z}{8},T\right) \\
&= G(1,T)\prod_{s=0}^{\infty} f\left(\frac{2^s + z - 1}{2^s}\right) = \prod_{s=0}^{\infty} f\left(\frac{2^s + z - 1}{2^s}\right).
\end{aligned}
\tag{10}
$$

Note that here we used $G(1,T) = 1$ which follows from the normalization of the distribution. Substituting Eq. (10) in Eq. (9) we obtain $G((1+z)/2, T)$ which after substituting in Eq. (8) leads us to an explicit solution of the generating function for model II in cyclostationary conditions:

$$
\begin{aligned}
G(z,t) &= \left[\prod_{s=0}^{\infty} f\left(1 + \frac{z-1}{2^{s+1}}\right)\right]\exp\left[\frac{-\alpha rt(1-z)}{1 + \alpha(1-z)}\right] \\
&= \exp\left(\frac{rxt}{1-x} + rxT\sum_{s=0}^{\infty}\frac{1}{2^{1+s} - x}\right),
\end{aligned}
\tag{11}
$$

where in the last line we used the definition of the function f and the definition $x = \alpha(z-1)$. Note that the sum over s in the argument of the exponent can be written in terms of the q-digamma function.

The solution we have computed corresponds to the PGF of the protein distribution computed from an ensemble of identical cells, all of which are at the same cell age t. However, distributions are often calculated from experimental measurements of time traces of the fluorescent protein molecules along a cell lineage or else from population snapshots. Considering the single lineage case, the corresponding PGF is given by a time average of the generating function calculated earlier:

$$
\begin{aligned}
G_s(z) &= \int_0^T \frac{1}{T}G(z,t)\,dt \\
&= \frac{x-1}{xy}\left[\exp\left(\frac{xy}{x-1}\right) - 1\right]\exp\left(xy\sum_{s=0}^{\infty}\frac{1}{2^s - x}\right),
\end{aligned}
\tag{12}
$$

where we have used $y = rT$ as we did for model I. Note that the subscript s will henceforth be used to denote single lineage measurement. Note also that the time average is computed since the probability of observing cells of any age is uniform in lineage measurements. Comparing this PGF to that of model I, i.e., Eq. (2), it is clear that in model II the protein distribution is generally not equal to the negative binomial distribution of model I.

To understand the differences between these two distributions, we next use the PGF's given by Eq. (2) (with $\beta = 3y/2$) and Eq. (12) to compute the mean $\langle n \rangle$ and variance σ^2 in stationary conditions

$$
\langle n \rangle_{\text{NB},s} = \frac{3}{2}\alpha y,
\tag{13}
$$

$$
\sigma_{\text{NB},s}^2 = \frac{3}{2}\alpha y + \frac{3}{2}\alpha^2 y = 1.5\alpha y + 1.5\alpha^2 y,
\tag{14}
$$

$$
\langle n \rangle_s = \frac{3}{2}\alpha y,
\tag{15}
$$

$$
\sigma_s^2 = \frac{3}{2}\alpha y + \alpha^2\left(\frac{5y}{3} + \frac{y^2}{12}\right)
$$

$$
\approx 1.5\alpha y + 1.67\alpha^2 y + 0.08\alpha^2 y^2.
\tag{16}
$$

It is clear that while the mean of the two distributions is the same, the variances are generally different. In Appendix A, we clarify the origin of this discrepancy. In particular, model I, under certain conditions described earlier, can match the mean number of proteins in a three-stage model of gene expression with explicit mRNA and protein dynamics, binomial partitioning, and fixed cell cycle length T whereas model II can match the full PGF of the three-stage model under the same conditions (see in particular Appendix A 2). The variance of model II is always greater than that of model I ($\sigma_s^2 > \sigma_{\text{NB},s}^2$). Furthermore, while the two variances are both quadratic in α, $\sigma_{\text{NB},s}^2$ is linear in y while σ_s^2 is quadratic in y. The relative error between the two variances computed as $(\sigma_s^2 - \sigma_{\text{NB},s}^2)/\sigma_s^2$ increases monotonically with α and y but is mostly determined by the value of y, i.e., the average number of mRNA produced in a cell cycle, as illustrated in Fig. 1(a). Expressions for the skewness squared can also be easily derived from the PGF

$$
S_{\text{NB},s}^2 = \frac{2(2\alpha + 1)^2}{3\alpha(\alpha + 1)y},
\tag{17}
$$

$$
S_s^2 = \frac{108[2\alpha^2(7y + 54) + 7\alpha(y + 20) + 42]^2}{49\alpha y[\alpha(y + 20) + 18]^3}.
\tag{18}
$$

For small y, both of these expressions are proportional to $1/y$ while for large y, we have $S_{\text{NB},s}^2 \propto 1/y$ while $S_s^2 \propto 1/y^2$, i.e., for large enough y model II will predict a less skewed distribution than model I. Generally, the skewness of the distribution of model II can be larger or smaller than that of model I depending on the values of α and y [Fig. 1(b)].

To get a fuller picture of the differences between the two models (assuming lineage measurements), we plot in Fig. 2 the distributions for various values of y while keeping the value of α fixed. Note that the distribution of protein numbers for model II is constructed by first expanding the generating function (12) as a Taylor series using a symbolic computation software and then $P(n)$ is simply given by the nth coefficient of this series. The theoretical predictions for model II are also verified by means of the stochastic simulation algorithm (SSA, see Appendix B for a full description of the algorithm). The negative binomial distribution of model I (red line) is a good approximation of the distribution of model II (black line) for small y but clearly is inappropriate for large y; it can also be shown that the difference between distributions becomes more pronounced if α is increased.

A visual inspection of the distribution of model II (black line) leads one to believe that one can likely fit well an effective negative binomial for the cases shown in Figs. 2(a)–2(c) but not for 2(d). This intuition is verified in Fig. 2 by plotting an effective negative binomial of the same mean and variance (green open circles) as the distribution of model II; in Fig. 2(d), the distribution is considerably flatter
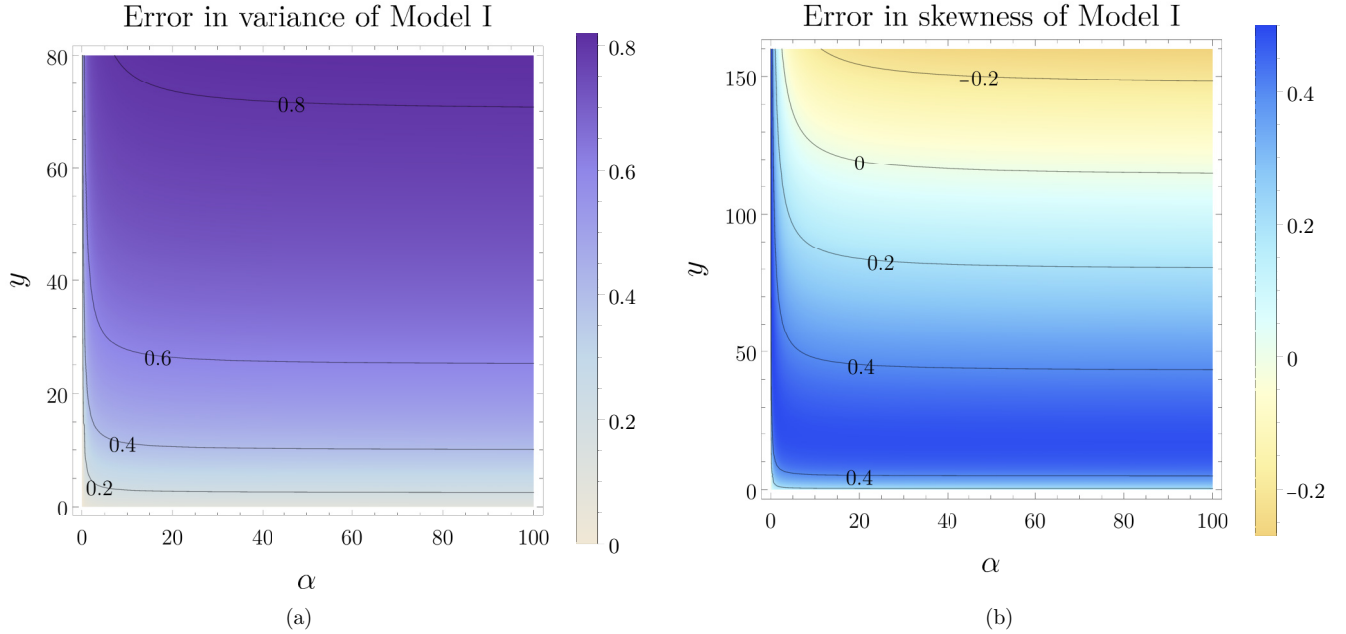
FIG. 1. (a) Plot of the relative error between the steady-state variance of model I and the cyclostationary variance of model II as a function of the mean burst size (α) and the mean number of mRNAs produced in one cell cycle (y), assuming lineage measurements. (b) Plot of the relative error between the skewness of models I and II as a function of α and y. Note that for burst sizes $\alpha \gtrsim 10$, the relative errors are a strong function of y, with only a weak dependence on α. Model I in all cases underestimates the variance of model II. In contrast, model I underestimates the skewness of model II for small y and overestimates it for large y. The relative error in the variance is defined as $(\sigma_s^2 - \sigma_{\mathrm{NB},s}^2)/\sigma_s^2$ and the relative error in the skewness is defined as $(S_s - S_{\mathrm{NB},s})/S_s$.

near the mode than the fitted negative binomial. Note that the effective negative binomial is given by NB(z_1, z_2) where

$$z_1 = \frac{27y}{y + 20}, \tag{19}$$

$$z_2 = \frac{\alpha(y + 20)}{\alpha(y + 20) + 18}. \tag{20}$$

Since model I has a negative binomial solution, it follows that through renormalization of its parameters, it can be matched to the effective negative binomial for model II. If the renormalized parameters for model I are y_e and α_e, then its solution is NB(z_1, z_2) where $z_1 = 3y_e/2$, $z_2 = \alpha_e/(1 + \alpha_e)$. Equating z_1, z_2 to those in Eqs. (19) and (20), we obtain

$$\alpha_e = \alpha\left(\frac{y + 20}{18}\right), \tag{21}$$

$$y_e = y\left(\frac{18}{y + 20}\right). \tag{22}$$

Note that renormalized model I is the same as the green open circles shown in Fig. 2 which is a much better approximation to model II (black line) than the original model I (red line). From these equations, it is also clear that if parameters had to be estimated from experimental data (with low cell cycle duration variability) using model I, then the estimated mean burst size α_e overestimates the true value α, while the estimated mean number of mRNA per cell cycle y_e underestimates the true value y.

In this section we have so far focused on the distributions for single lineage measurements. The distribution of protein numbers for population snapshots can also be derived and

instead of Eq. (12) we then have

$$
\begin{aligned}
G_p(z) &= \int_0^T \frac{2^{1-t/T}\log(2)}{T} G(z, t)dt \\
&= \frac{(x-1)\log 2}{xy + (x-1)\log 2}\left\{\exp\left[\frac{xy + (x-1)\log 2}{x - 1}\right] - 1\right\} \\
&\quad \times \exp\left(xy\sum_{s=0}^{\infty}\frac{1}{2^s - x}\right),
\end{aligned} \tag{23}
$$

where the subscript p will be used to denote population snapshot measurements from hereon. Note that we used the fact that when interdivision times are regularly spaced in time, the probability of observing a cell of age $t \in [0, T]$ is $2^{1-t/T}\log 2/T$ for population measurements [17] (see also Appendix D for a derivation of the latter). The PGF of model I has already been calculated for the population scenario in Sec. II and was found to be given by Eq. (2) [with $\beta = y/\log(2)$]. For the parameters used in Fig. 2, the protein distributions for population data for models I and II are found to be close to those calculated for lineage data and, hence, we do not show them. The mean and variance of the protein numbers of models I and II are now given by

$$\langle n\rangle_{\mathrm{NB},p} = \frac{\alpha y}{\log(2)}, \tag{24}$$

$$\sigma_{\mathrm{NB},p}^2 = \frac{\alpha(1 + \alpha)y}{\log(2)} \approx 1.44\alpha y + 1.44\alpha^2 y, \tag{25}$$

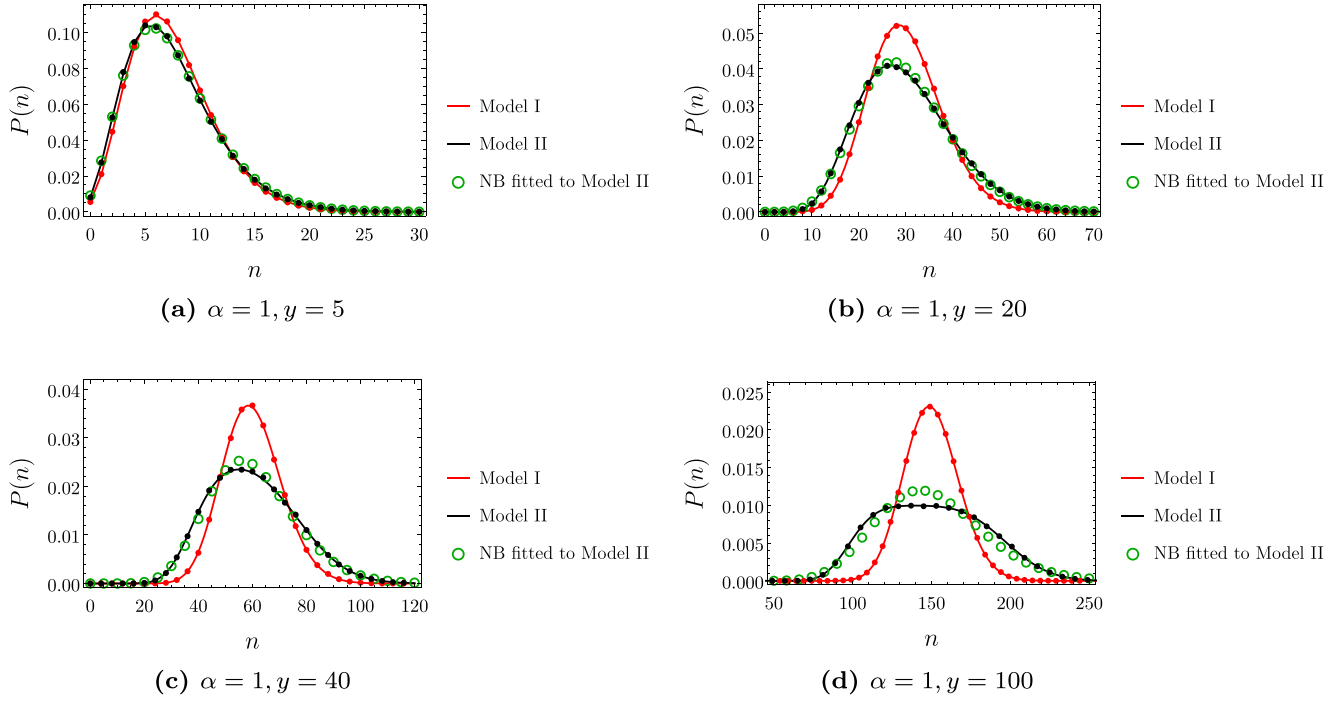$$\langle n\rangle_p = \frac{\alpha y}{\log(2)}, \tag{26}$$

FIG. 2. Comparison of the protein number distributions (assuming lineage measurements) of models I and II in steady-state and cyclostationary conditions, respectively. For model I (red line), the distribution is NB$(3y/2, \alpha/(1 + \alpha))$ and for model II (black line) it is given by $P(n) = (1/n!)dG_s^n/dz^n|_{z=0}$ where G_s is given by Eq. (12). Note that the numerical computation of the distribution from Eq. (12) can be greatly accelerated (while maintaining accuracy) if the sum over s is truncated to a few tens of terms. The dots show the distributions of models I and II obtained from stochastic simulations (simulations of model I are done using the conventional SSA and those of model II are done using a modified SSA, see Appendix B and also at the end of this caption). The open green circles show the negative binomial distribution which has the same first and second moments as the distribution of model II. The stochastic simulations of model II are performed as follows. Initially, we have a single cell with zero protein. We measure the protein content of the cell at intervals T/Z where $Z = 10\pi$. Each time a cell divides, we follow only one of the daughter cells. The simulation is run until 10^5 cycles have passed and a histogram is calculated from these data (we discard the first 10^3 cell cycles to ignore any possible transients). The cell cycle length is $T = 1$ in all cases. All lineage simulations in this article use this protocol, unless otherwise stated.

$$\sigma_p^2 = \frac{1}{\log 2}\alpha y + \frac{6 - 4\log 2}{3\log 2}\alpha^2 y + \frac{1 - \log^2 2}{\log^2 2}\alpha^2 y^2$$

$$\approx 1.44\alpha y + 1.55\alpha^2 y + 0.08\alpha^2 y^2. \qquad (27)$$

These equations are the population equivalent of Eqs. (13)–(16) and the same observations we made earlier regarding the comparison of the moments of models I and II for single lineage data are also seen to hold for population data. Note also that generally we can state $\langle n \rangle_p < \langle n \rangle_s$ and $\sigma_p^2 < \sigma_s^2$, which can be explained by the enhanced probability of observing younger cells (and hence having a smaller protein content) in population measurements (as mentioned in Sec. II).

Summarizing, our results in this section imply that the effective degradation reaction in model I cannot effectively account for dilution via binomial partitioning. Generally, the models agree on the mean number of proteins in stationary conditions but not on the higher-order moments. The discrepancies are particularly obvious whenever y is greater than a few tens. The variance of model I is always less than that of model II, but the skewness of model I can be greater or smaller than that of model II. We have also shown that the protein distribution of model II can be well approximated by an effective negative binomial distribution only if y, the mean number of mRNAs produced in a cell cycle, is small. In this case, it is possible to renormalize the parameters in model I so that its solution approximates that of model II well.

IV. MODEL III: STOCHASTIC GENE EXPRESSION WITH EXPLICIT MODELING OF AN ERLANG DISTRIBUTED CELL CYCLE LENGTH

Next, we consider a more complex and realistic model of bursty gene expression, namely, one that includes cell cycle length variability. Such variability could, for example, originate in cell types where the cell growth rate is stochastic and cell division is triggered when the volume of a cell exceeds its volume at birth by a certain fixed amount (also called an "adder" mechanism, see for example [31]). As we shall see, this requires a very different master equation description than the previous models. Specifically, the model has the following properties: (i) The cell cycle is divided in N phases where the duration of each phase is exponentially distributed with parameter k. It then follows that the cell cycle length distribution is Erlang, the mean cell cycle time is $T = N/k$ and the coefficient of variation of the cell cycle duration is $1/\sqrt{N}$. (ii) Proteins are produced at a rate r and in geometrically distributed burst sizes with mean α. (iii) Cell division occurs instantaneously after the end of the Nth phase

which leads to binomial partitioning of proteins between mother and daughter cells. Note that an Erlang distribution provides a good fit to some of the measured cell cycle length distributions [21,28].

Let $P_i(n, t)$ be the probability that the cell cycle is in phase i at time t and that there are n protein molecules. We shall here ignore the generation number since in steady-state conditions this does not matter. It then follows that the master equation describing the above model is given by

$$\frac{dP_1(n, t)}{dt} = -kP_1(n, t) + kP'_N(n, t)$$
$$+ r \sum_{m=0}^{\infty} P_1(n - m, t)Q(m)$$
$$- rP_1(n, t) \sum_{m=0}^{\infty} Q(m), \quad (28)$$

$$\frac{dP_i(n, t)}{dt} = -kP_i(n, t) + kP_{i-1}(n, t)$$
$$+ r \sum_{m=0}^{\infty} P_i(n - m, t)Q(m)$$
$$- rP_i(n, t) \sum_{m=0}^{\infty} Q(m), \quad i \in [2, N] \quad (29)$$

where $Q(m) = p(1 - p)^m$ is the geometric distribution with mean $\alpha = (1 - p)/p$. The first term in these equations models exits from the present cell cycle phase into the next phase, the second term models the entry into the present cell cycle phase from the previous one, and the third term models bursty protein production. Note that binomial partitioning during cell division is explicitly taken into account by the second term of Eq. (28). In particular, this process implies

$$P'_N(n, t) = \sum_{m=0}^{\infty} \binom{m}{n} 2^{-m} P_N(m, t), \quad (30)$$

where we take the convention m choose n equals zero when $n > m$. The PGF equations corresponding to the CME equations (28) and (29) are given by

$$\frac{\partial G_{1,s}(z)}{\partial t} = -kG_{1,s}(z) + kG_{N,s}\left(\frac{1 + z}{2}\right)$$
$$- rG_{1,s}(z)\left[1 - \frac{1}{1 + \alpha(1 - z)}\right], \quad (31)$$

$$\frac{\partial G_{i,s}(z)}{\partial t} = -kG_{i,s}(z) + kG_{i-1,s}(z)$$
$$- rG_{i,s}(z)\left[1 - \frac{1}{1 + \alpha(1 - z)}\right], \quad i \in [2, N] \quad (32)$$

where we have suppressed the time dependence for convenience. Note that while for model II, the cyclostationary condition meant that the protein distribution at a given cell age is independent of generation number, for model III the condition means that the protein number at a given cell cycle phase is independent of the generation number. Hence, we can

set Eq. (32) to zero and solve recursively for $G_{i,s}(z)$ to obtain

$$G_{i,s}(z) = \left[\frac{k(x - 1)}{k(x - 1) + rx}\right]^{i-1} G_{1,s}(z), \quad i \in [2, N]. \quad (33)$$

Substituting Eq. (33) with $i = N$ in Eq. (31) with the left hand side equal to zero, we obtain

$$G_{N,s}\left(\frac{1 + z}{2}\right) = \frac{k(x - 1) + rx}{k(x - 1)} G_{1,s}(z)$$
$$= \left[\frac{k(x - 1)}{k(x - 1) + rx}\right]^{-N} G_{N,s}(z). \quad (34)$$

Following the same method of solution as used for solving Eq. (9), we obtain

$$G_{N,s}(z) = \prod_{s=0}^{\infty} \left[1 - \frac{rx}{x(k + r) - 2^s k}\right]^N, \quad (35)$$

where we used the normalization condition for the conditional distribution in each phase, i.e., $G_i(1) = 1$. Using Eqs. (33) and (35) we obtain the PGF for the conditional protein distribution in cell phase i:

$$G_{i,s}(z) = \left[1 + \frac{rx}{k(x - 1)}\right]^{N-i} \prod_{s=0}^{\infty} \left[1 - \frac{rx}{x(k + r) - 2^s k}\right]^N,$$
$$i \in [1, N]. \quad (36)$$

Since we are considering the case of single lineage measurements, we must average the PGF over all cell phases by marginalizing out the phase in which a cell is at observation time. Note that in order to do this we need an expression for $\Pi_{i,s}$, the probability that the cell is in phase i at observation in a single lineage measurement. For our case of N identically exponentially distributed phases it can be easily shown that $\Pi_{i,s} = N^{-1}$, reflecting that every phase is equally likely to be observed. Using this we derive the PGF for the protein distribution for lineage measurements

$$G_s^E(z) = \sum_{i=1}^{N} \Pi_{i,s} G_{i,s}(z) = \frac{x - 1}{xy} \left\{\left[1 + \frac{xy}{N(x - 1)}\right]^N - 1\right\}$$
$$\times \prod_{s=0}^{\infty} \left[1 - \frac{xy}{x(N + y)x - 2^s N}\right]^N, \quad (37)$$

where we used that $y = rT$ and the mean cell cycle length $T = N/k$. Note that the superscript E denotes an Erlang distributed cell cycle length. This solution can be conveniently written in terms of exponential functions yielding

$$G_s^E(z) = \frac{x - 1}{xy} \left(\exp\left\{N \log\left[1 + \frac{xy}{N(x - 1)}\right]\right\} - 1\right)$$
$$\times \exp\left\{N \sum_{s=0}^{\infty} \log\left[1 - \frac{xy}{(N + y)x - 2^s N}\right]\right\}. \quad (38)$$

Note that the argument of the log is always positive because $x = \alpha(z - 1) \leqslant 0$ due to $z \leqslant 1$. Note also that in the limit of large N (at constant T), the Erlang distribution describing the cell cycle length tends to a delta function centered on T, i.e., a cell cycle of fixed length. It is straightforward to show by a series expansion in $1/N$ that in the limit of large N, Eq. (38)
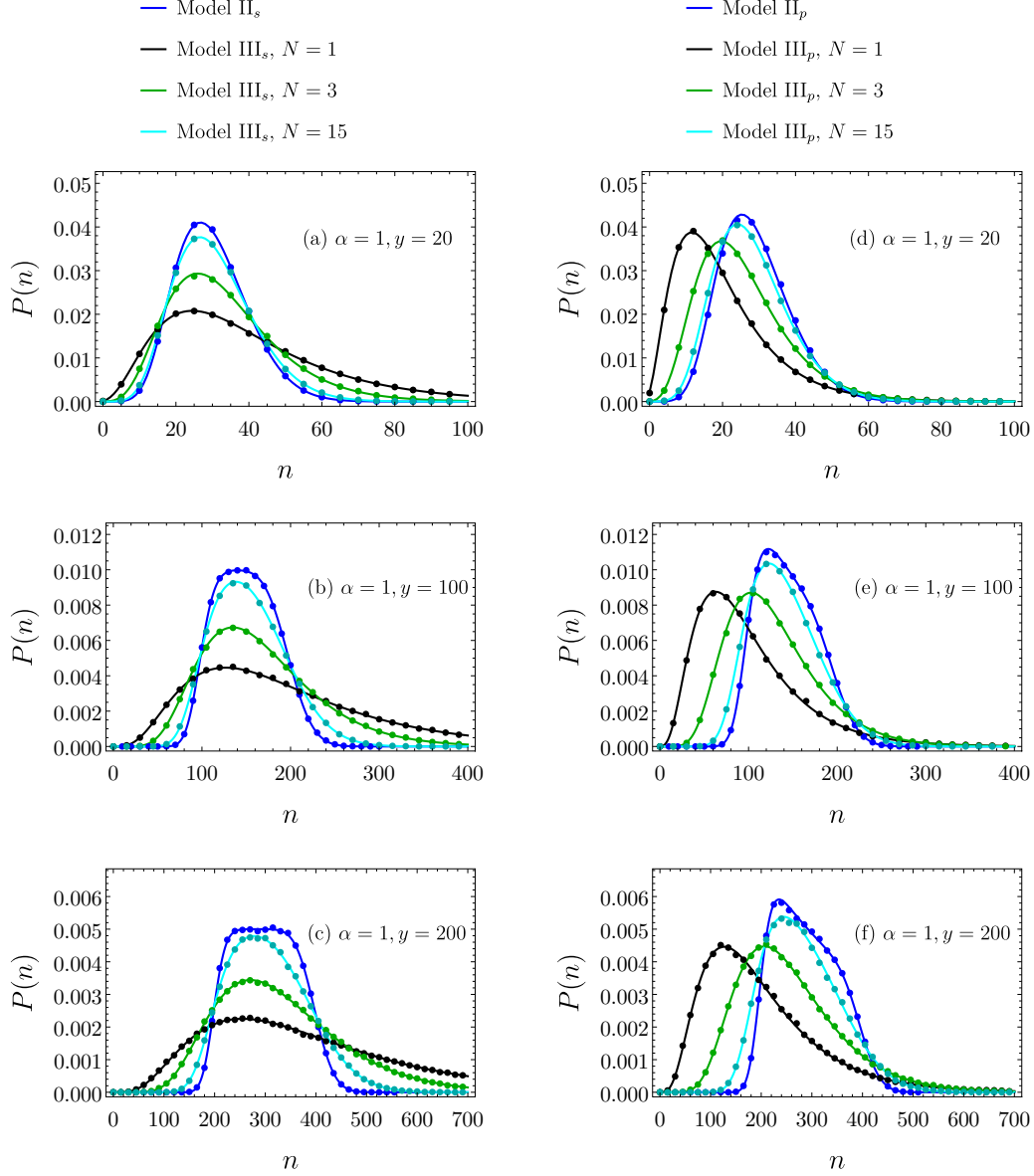
FIG. 3. Plots of the protein number distribution $P(n)$ for model II (blue line) and model III (black, green, cyan lines) with single lineage (a)–(c) and population observations (d)–(f). The solid dots show the distributions obtained from stochastic simulations using the SSA (see Appendix B) which agree with those from theory in all cases. Note that in the limit of large number of cycle cycle phases $N \gg 1$, model III approaches model II since the cell cycle length variability tends to zero. The non-negative binomial nature of model II for large y [the flat region near the mode in (b) and (c) and the right shoulder in (e) and (f)] is washed away as the cell cycle length variability increases, i.e., as N decreases in model III. Note that the numerical computation of the distribution of model III from Eqs. (37) and (45) can be greatly accelerated (while maintaining accuracy) if the infinite product is truncated to a few tens of terms.

converges to Eq. (12), i.e., in the limit of small cell cycle length variability, model III converges to model II. The mean and variance of the protein number distribution in steady-state conditions can be straightforwardly computed from the PGF:

$$\langle n \rangle_s^E = \frac{\alpha y (3N + 1)}{2N}, \tag{39}$$

$$\left(\sigma_s^E\right)^2 = \frac{\alpha y (3N + 1)}{2N}$$
$$+ \alpha^2 \left\{ \frac{[N(N + 10) + 5]y^2}{12N^2} + \frac{(5N + 3)y}{3N} \right\}. \tag{40}$$

As expected, the variance of model III is always larger than that of model II; the mean of model III is slightly larger than that of model II, but the difference can be ignored in most cases of interest. Both the mean and variance are monotonic decreasing functions of N and hence they are bounded from above by the moments evaluated for $N = 1$, i.e., an exponentially distributed cell cycle length.

The differences in the protein number distributions predicted by models II and III for lineage observations are illustrated in Figs. 3(a)–3(c). There we show the excellent agreement between the theoretical expressions and stochastic simulations for both small y [Fig. 3(a)] and large y [Figs. 3(b)

and 3(c)]. While the differences between models I and II are due to binomial partitioning, the differences between models II and III are due to cell cycle length variability. We find that typically for $N > 20$ the differences between models II and III are small; the differences are at their largest when the cell cycle length is exponentially distributed, i.e., $N = 1$. Previously, we saw how for model II the negative binomial was not a good fit for the distribution when y was large. However, as can be appreciated from Figs. 3(b) and 3(c), the fit becomes better when we consider cell cycle length variability: the deviations from negative binomial, which manifest in the flattish region around the mode, become less visible as N decreases. Note also that the deviations from negative binomial are largely unaffected by the value of the mean burst size (increasing α by 10 times in Fig. 3 causes the protein distributions to move right and their height to be rescaled but their shape remains practically unaltered).

Since the solution of model I is generally a negative binomial, it follows that we can renormalize the parameters of this model such that its protein distribution provides a good match to the distribution of model III when the cell cycle length variability is sufficiently high. Equating the mean and variance of a negative binomial NB(z_1, z_2) to Eqs. (39) and (40), we find

$$z_1 = \frac{3(3N+1)^2 y}{[N(N+10)+5]y + 4N(5N+3)}, \quad (41)$$

$$z_2 = \left\{ \frac{6N(3N+1)}{\alpha[N(N+10)+5]y + 4\alpha N(5N+3)} + 1 \right\}^{-1}. \quad (42)$$

Equating these two parameters (z_1, z_2) to those of model I with renormalized parameters $[3y_e/2, \alpha_e/(1+\alpha_e)]$, we obtain the relationship between the actual and renormalized parameters:

$$\alpha_e = \frac{z_2}{1-z_2} = \alpha \left\{ 1 + \frac{N+3}{3(3N+1)} + y \left[\frac{N^2 + 10N + 5}{6N(3N+1)} \right] \right\}, \quad (43)$$

$$y_e = \frac{2z_1}{3} = y \left[\frac{2(3N+1)^2}{(N^2 + 10N + 5)y + 4N(5N+3)} \right]. \quad (44)$$

It also follows from these formulas that if we had to fit a negative binomial to experimental data from cells with an Erlang distributed cell cycle length (data consistent with model III) and estimate the parameters using model I, then this will lead to an overestimate for the mean burst size and an underestimate for the mean number of mRNAs per cycle (and hence for the transcription rate). The errors increase with decreasing N and hence with increasing cell cycle duration variability.

We have here focused on the distributions for single lineage measurements. The distributions of protein numbers for population snapshots can also be derived. Due to the rather more complex analysis involved, the derivation is presented in Appendix D. Here, we simply state the equivalent of Eq. (37)

for population measurements:

$$G_p^E(z) = (2^{\frac{1}{N}} - 1)N(x-1) \frac{\left[\frac{xy}{N(x-1)} + 2^{\frac{1}{N}} \right]^N - 1}{(2^{\frac{1}{N}} - 1)N(x-1) + xy}$$

$$\times \prod_{s=0}^{\infty} \left[1 - \frac{xy}{x(N2^{1/N} + y) - N2^{1/N}2^s} \right]^N. \quad (45)$$

The protein distributions corresponding to this PGF are shown in Figs. 3(d)–3(f) where they are also compared with those of model II. Note that given the same parameters, the protein distributions for lineage and populations observations are considerably different. These differences become more appreciable with increasing y and decreasing N. While the increase in cell cycle length variability (through decreasing N) results in little changes to the mode of the lineage distribution, it causes the mode of the population distribution to shift to the left. However, there are also qualitative similarities, namely, that in both cases the deviations from negative binomial are maximal for small cell cycle length variability (large N) and large y. Similar to what we previously did for lineage observations [see Eqs. (43) and (44)], from the equations for the mean and variance for population snapshots (see Appendix D), it is also possible to calculate the renormalized parameters in model I such that it provides a good negative binomial approximation to the population distribution of model III when there is sufficient cell cycle length variability.

Summarizing in this section we have studied a model (model III) of bursty gene expression with an Erlang distributed cell cycle length. This model recovers model II in the limit of small cell cycle length variability. Also, the presence of sufficient cell cycle length variability is found to lift the deviations from negative binomial observed for model II; in this case, by an appropriate renormalization of the parameters, model I can describe the distribution predicted by model III well. The mean and variance of protein numbers calculated from a similar model (assuming lineage observations) have been reported previously [22].

V. MODEL IV: STOCHASTIC GENE EXPRESSION WITH HYPOEXPONENTIAL CELL CYCLE LENGTH DISTRIBUTION AND AGE-DEPENDENT TRANSCRIPTION

We next consider a more general version of model III: (i) the time spent in phase i of the cell cycle is exponentially distributed with parameter k_i, which implies that the cell cycle length distribution is hypoexponential; (ii) the transcription rate and burst size are age dependent, i.e., they are r_i and α_i in phase i, respectively. Note that the Erlang distribution is a special case of the hypoexponential distribution and hence the use of this distribution, in principle, allows more flexibility in fitting experimental cell cycle distributions. Note also that modeling the transcription rate as age dependent enables us to capture replication, hence considerably extending the realism of our model. The master equation describing the above model is a generalization of that for model III and is

given by

$$\frac{dP_1(n,t)}{dt} = -k_1 P_1(n,t) + k_N P'_N(n,t)$$

$$+ r_1 \sum_{m=0}^{\infty} P_1(n-m,t) Q_1(m)$$

$$- r_1 P_1(n,t) \sum_{m=0}^{\infty} Q_1(m), \qquad (46)$$

$$\frac{dP_i(n,t)}{dt} = -k_i P_i(n,t) + k_{i-1} P_{i-1}(n,t)$$

$$+ \sum_{m=0}^{\infty} r_i P_i(n-m,t) Q_i(m)$$

$$- r_i P_i(n,t) \sum_{m=0}^{\infty} Q_i(m), \quad i \in [2,N] \quad (47)$$

where $Q_i(m) = p_i(1-p_i)^m$ is the geometric distribution with mean $\alpha_i = (1-p_i)/p_i$. Note that $P'_N(n,t)$ is defined as before using Eq. (30). The corresponding PGF equations are given by

$$\frac{\partial G_{1,s}(z)}{\partial t} = -k_1 G_{1,s}(z) + k_N G_{N,s}\left(\frac{1+z}{2}\right)$$

$$- r_1 G_{1,s}(z)\left[1 - \frac{1}{1+\alpha_1(1-z)}\right], \quad (48)$$

$$\frac{\partial G_{i,s}(z)}{\partial t} = -k_i G_{i,s}(z) + k_{i-1} G_{i-1,s}(z)$$

$$- r_i G_{i,s}(z)\left[1 - \frac{1}{1+\alpha_i(1-z)}\right], \quad i \in [2,N] \quad (49)$$

where we have suppressed the time dependence for convenience. Setting Eq. (49) to zero (steady-state conditions) and solving recursively for $G_{i,s}(z)$ we obtain

$$G_{i,s}(z) = w_{i,s}(z) G_{1,s}(z), \qquad (50)$$

where

$$w_{i,s}(z) = \begin{cases} 1, & i=1 \\ \prod_{j=2}^{i} \frac{k_{j-1}(x_j-1)}{k_j(x_j-1)+r_j x_j}, & i \in [2,N]. \end{cases} \quad (51)$$

Note that here we defined $x_j = \alpha_j(z-1)$. Substituting Eq. (50) with $i = N$ in Eq. (48) with the left hand side equal to zero, we obtain

$$G_{N,s}\left(\frac{1+z}{2}\right) = G_{N,s}(z)\left[\frac{k_1(x_1-1)+r_1 x_1}{k_N(x_1-1)w_N(z)}\right]. \quad (52)$$

Following the same method of solution as used for solving Eq. (9), we obtain

$$G_{N,s}(z) = \prod_{s=0}^{\infty} \prod_{j=1}^{N}\left[1 - \frac{r_j x_j}{x_j(k_j+r_j)-2^s k_j}\right], \quad (53)$$

where again we used the normalization condition for the conditional distribution in each phase, i.e., $G_{i,s}(1) = 1$. Hence, finally by means of Eqs. (50) and (53), we can write an equation for the PGF of the distribution assuming single lineage measurements

$$G_{i,s}(z) = \frac{w_{i,s}(z)}{w_{N,s}(z)} G_{N,s}(z), \qquad (54)$$

$$G_s^H(z) = \sum_{i=1}^{N} \Pi_{i,s} G_{i,s}(z), \qquad (55)$$

where again we let $\Pi_{i,s}$ denote the probability of observing the cell in phase i in a lineage measurement, which can be shown to be given by $\Pi_{i,s} = k_i^{-1}/(\sum_j k_j^{-1})$. Note the superscript H (standing for hypoexponential) is to distinguish this PGF from the one calculated for model III using the Erlang distribution.

A. Modeling replication

We next use this theory to understand the effects of gene replication on stochastic gene expression. One of the simplest models of this process assumes (i) the transcription rate to be a constant r before replication, doubling to $2r$ right after replication (the doubling in transcription rate is due to the doubling in gene copy number during replication); (ii) the cell cycle length and the replication time are Erlang distributed. This implies the special case where $k_i = k$ and $\alpha_i = \alpha$ for all i, and $r_i = r$, $i \in [1,M]$, and $r_i = 2r$, $i \in [M+1,N]$, where M is the cell cycle phase after which replication occurs. Substituting these values in Eqs. (54) and (55), we obtain

$$G_s^H(z) = \frac{(x-1)}{2xy}\left(\left[\frac{2xy}{N(x-1)}+1\right]^{N-M}\left\{2\left[\frac{xy}{N(x-1)}+1\right]^M - 1\right\} - 1\right)$$

$$\times \prod_{s=0}^{\infty}\left[1 - \frac{2xy}{x(N+2y)-N2^s}\right]^{N-M}\left[1 - \frac{xy}{x(N+y)-N2^s}\right]^M. \quad (56)$$

This can also be written in exponential form as we have previously done for other expressions. The mean and variance of protein fluctuations are given by

$$\langle n \rangle_s^H = \frac{\alpha y[M(M-1)+2N(3N-2M+1)]}{2N^2} = \langle n \rangle_s^E\left[\frac{M+2(N-M)}{N}\right] - \frac{\alpha y M(N-M)}{2N^2}, \quad (57)$$

$$\left(\sigma_s^H\right)^2 = \frac{\alpha y[M(M-1)+2N(3N-2M+1)]}{2N^2} + \frac{\alpha^2 y[10N^2+2N(3-4M)+3M(M-1)]}{3N^2}$$

$$+ \alpha^2 y^2\left\{\frac{4N^4+40N^3-4[3M(M+3)-5]N^2+12M(M^2-1)N-3M^2(M-1)^2}{12N^4}\right\}. \quad (58)$$

Using the notation $(\sigma_s^E)^2\big|_{vy}$ to denote the variance in model III, i.e., Eq. (40), when the transcription rate is vr we find

$$(\sigma_s^H)^2 = \left[(\sigma_s^E)^2\big|_y \left(\frac{M}{N}\right) + (\sigma_s^E)^2\big|_{2y} \left(\frac{N-M}{N}\right)\right] - \frac{\alpha y M(N-M)}{2N^2}\left\{1 + 2\alpha - \frac{\alpha y[(M-1)^2 + N(N-3M-2)]}{2N^2}\right\}. \quad (59)$$

Note that for the special cases $M = 0$ and N, these two equations are the same as the mean and variance of model III [given by Eqs. (39) and (40)] with y replaced by $2y$ and y, respectively; this is since in this case the transcription rate is the same in all phases of the cell cycle and equal to $2r$ and r, respectively. It also follows that model II is obtained by setting $M = N$ and taking the limit $N \to \infty$. Hence, model IV contains as special cases the previous models II and III. Note that while we have considered the cases $M = 0$ and N to see the relationships between the various models, when we want to explicitly model replication we need $0 < M < N$ and $N \geqslant 2$ since there is always a prereplication and postreplication phase of the cell cycle. In Fig. 4 we show that the theoretical protein distributions for each phase of the cell cycle accurately match those obtained from stochastic simulations.

The coefficient of variation squared $[CV_s^2 = (\sigma_s^H/\langle n\rangle_s^H)^2]$ can be shown to decrease monotonically with increasing α and y. However, the dependence on N and M (the replication phase of the cell cycle) is nonmonotonic (see Appendix C). In Fig. 5 we show plots of the CV_s^2 as a function of all four parameters which numerically verifies the aforementioned properties. We also observe that the size of noise as measured by CV_s^2 is almost independent of the replication phase M for large y but increases monotonically with M for small y.

There is an interesting relationship between models III and IV, as follows. If we renormalize the parameter y in model III by changing it to $\Lambda_s y$ where

$$\Lambda_s = 2 - \frac{M}{N} - \frac{M(N-M)}{N(1+3N)}, \quad (60)$$

then $\langle n\rangle_s^H = \langle n\rangle_s^E$ and $(\sigma_s^H)^2 \approx (\sigma_s^E)^2$. Note that while the relationship between the means is exact, this is not true for the variances; the accuracy of the latter, however, is shown for six different parameter sets in Fig. 6(a). Note also that since the mean and variance of model IV and the renormalized model III match, it follows that if the distribution in both models is well approximated by a negative binomial (a two parameter distribution), then we expect the two distributions to also match. This is indeed the case for models with small y and including sufficient cell cycle length variability, i.e., moderate N, as can be seen in Fig. 6(b).

However, when the cell cycle length variability decreases as $N \to \infty$, we see that the distribution of model IV starts to deviate from the renormalized model III when y grows large as we show in Fig. 6(c). This can be understood from the observation that the prereplication and postreplication phase of the cell for large y have distinct protein distributions, which was
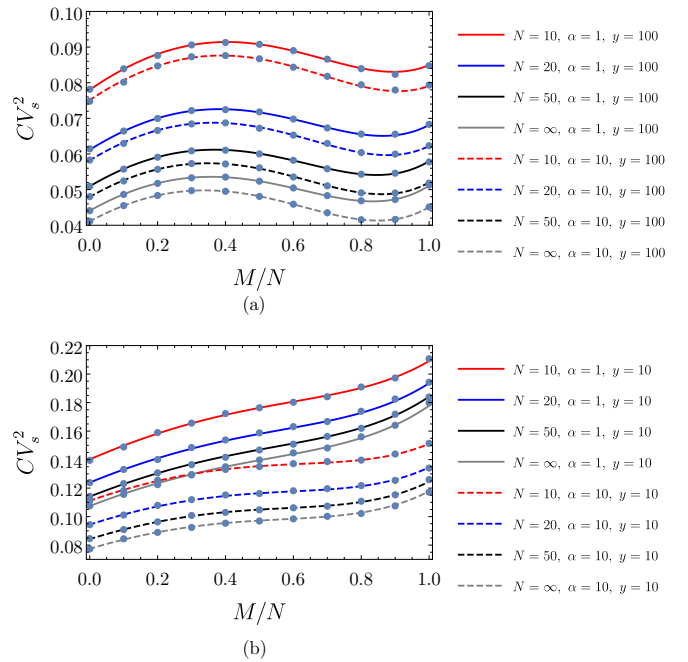
FIG. 5. Effects of the position of the replication point in the cell cycle on the size of protein fluctuations for lineage observations. Plots of the coefficient of variation squared CV_s^2 as a function of the mean burst size α, the mean number of mRNAs produced in a cell cycle y, and the fraction of the cell cycle in the prereplication stage M/N for model IV (same setup as previous figure). The mean and variance for the computation of the CV_s^2 are given by Eqs. (57) and (58). A comparison of (a) and (b) shows that the CV_s^2 decreases with increasing α and y. However, the CV_s^2 has a complex dependence on M: for large y, CV_s^2 is roughly independent of M while it increases with M for small y. Solid dots show the results of the SSA and solid lines show the theory.
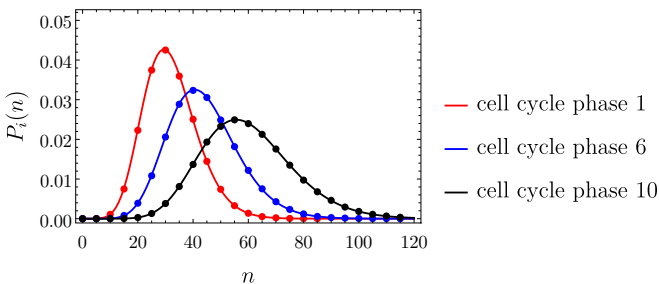
FIG. 4. Plots of the protein distribution at different points in the cell cycle for model IV with an Erlang distributed cell cycle (10 phases) and replication occurring at the start of phase 6. The distributions (solid lines) are calculated from the theory for lineage observations while the dots show the same calculated from a single trajectory generated by the SSA for 10^6 cell cycles. Note that the theory is for model IV Eq. (54) with $i = 1, 6, 10$, $k_i = k = N/T$, and $\alpha_i = \alpha$ for all i, and $r_i = r$, $i \in [1, M]$, and $r_i = 2r$, $i \in [M + 1, N]$, where M is the cell cycle stage in which replication occurs. The parameters are $N = 10$, $M = 5$, $\alpha = 1$, $y = 20$, $T = 1$.
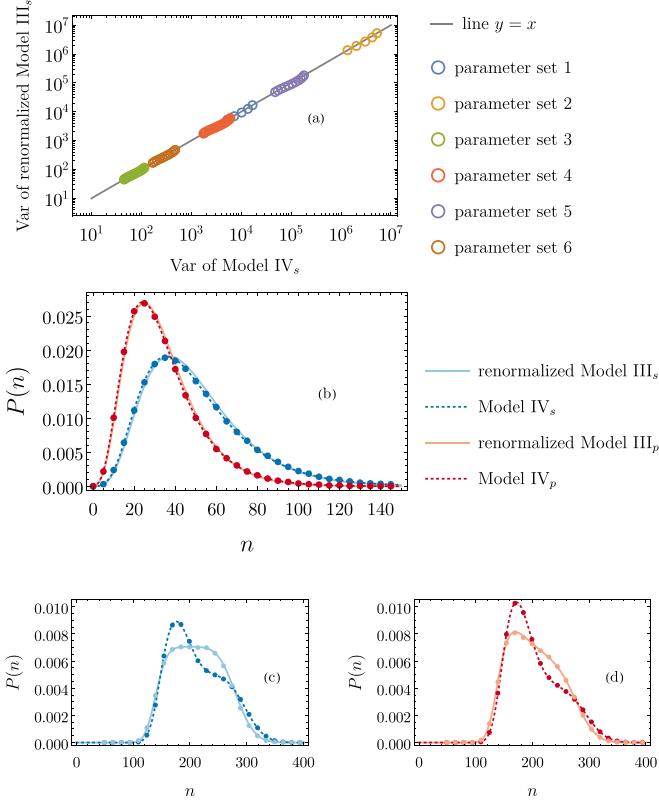
FIG. 6. Relationship between models III and IV for lineage (denoted by subscript s in figure) and population observations (denoted by subscript p in figure). (a) Renormalizing the parameter y in model III (lineage) by changing it to $\Lambda_s y$, we find that the variance agrees to a good degree of approximation with the variance of model IV (lineage) for all parameter sets tested. The mean of the two models is exactly the same under this renormalization. The same applies for population observations (not shown). (b) The protein distribution of model IV and of the renormalized model III are also in excellent agreement (these are obtained from the PGF solutions of both models) for small y and N. Note the renormalization factors are given by Eq. (60) for lineage and $\Lambda_p = 2^{1-M/N}$ for population observations. (c) In contrast, the distributions of model IV (lineage) and of renormalized model III (lineage) are very different when y and N are large. In (d) we show that the same conclusion holds for population observations. In all cases, dots show the distribution obtained from the SSA is in good agreement with the theory (solid or dashed lines) for models III and IV. Parameter sets are as follows. The six parameter sets for model IV in (a) are as follows: (1) $N = 4, \alpha = 10, y = 10$, (2) $N = 4, \alpha = 20, y = 100$, (3) $N = 20, \alpha = 1, y = 10$, (4) $N = 20, \alpha = 2, y = 50$, (5) $N = 10, \alpha = 5, y = 100$, and (6) $N = 10, \alpha = 2, y = 10$; for renormalized model III, the parameters are the same except that y is renormalized. For each parameter set, we take $M = 0, \ldots, N$. Parameters for (b) are $N = 2, \alpha = 1.0, y = 20, M = 1$ for model IV and same but with $y = 20\Lambda_{s,p}$ for renormalized model III. Parameters for (c) and (d) are $N \to \infty, \alpha = 1.0, y = 100, M = 0.5N$ for model IV and same for renormalized model III, but with $y = 100\Lambda_{s,p}$. Population snapshot SSA data for model IV consist of $\sim 10^6$ cells starting from a single cell with zero protein content (see Appendix B for details of the simulations).

visible in Fig. 4 as well, and the fact that the total observation distribution is a sum of both contributions. For moderate N, i.e., significant cell cycle length variability, the prereplication

and postreplication protein distributions overlap since they are wide, which yields good correspondence with model III as can be seen in Fig. 6(b) (compare solid and dashed blue lines). However, for N large enough (low cell cycle length variability) and large y, the overlap between the two contributions becomes smaller, resulting in the almost bimodal nature of the distribution in Fig. 6(c). Hence, it follows that replication effects in model IV can be described well by an appropriately scaled model III provided cell cycle length variability is not too small. It also follows that it can also be well described by model I with renormalized parameters since the latter we have shown in Sec. IV to be in good agreement with model III provided y and N are not large.

Finally, we mention that similar conclusions hold for population snapshot observations as we have seen for lineage data. A derivation of the snapshot distribution for model IV can be found in Appendix D. The population equivalent of Eq. (56) is given by $G_p^H(z) = G_{N,p}(z)W(z)$, where

$$G_{N,p}(z) = \prod_{s=0}^{\infty} \left[1 - \frac{2xy}{x(N2^{1/N} + 2y) - N2^{1/N}2^s} \right]^{N-M}$$

$$\times \left[1 - \frac{xy}{x(N2^{1/N} + y) - N2^{1/N}2^s} \right]^{M}, \quad (61)$$

$$W(z) = (2^{\frac{1}{N}} - 1)N(x - 1)$$

$$\times \left(\frac{\left[\frac{2xy}{N(x-1)} + 2^{\frac{1}{N}} \right]^{N-M} \left\{ \left[\frac{xy}{N(x-1)} + 2^{\frac{1}{N}} \right]^{M} - 1 \right\}}{(2^{\frac{1}{N}} - 1)N(x - 1) + xy} \right.$$

$$\left. + \frac{\left[\frac{2xy}{N(x-1)} + 2^{\frac{1}{N}} \right]^{N-M} - 1}{(2^{\frac{1}{N}} - 1)N(x - 1) + 2xy} \right). \quad (62)$$

As shown in Fig. 6(b), the difference between the population snapshot distribution and the lineage distribution can be significant (compare solid red and blue lines). The mean of the former is less than that of the latter which is due to a preponderance of young cells in population measurements (as discussed in Sec. II). Considering the appropriate rescaling of the parameter y by changing it to $\Lambda_p y$, where $\Lambda_p = 2^{1-(M/N)}$, we also get good agreement between models IV and III for population snapshots when the cell cycle length variability is moderate (small N) and y is small [compare dashed and solid red lines in Fig. 6(b)]. This, however, does not carry through for the case of large y and N [Fig. 6(d)]. Hence, while the distributions are appreciably different for population and lineage cases, nevertheless, qualitatively the results for the two are similar.

VI. EXTRINSIC NOISE FLOOR

Plots of the coefficient of variation squared versus mean protein number have been used in the literature to separate intrinsic noise from extrinsic noise (see in particular Fig. 2B of Ref. [32]). Specifically, intrinsic noise is associated with the term proportional to the inverse mean since its contribution decreases with the mean protein number, while extrinsic noise is associated with the term which is independent of the mean. Within this interpretation, using the expressions previously derived for the mean and variance for lineage observations,

it is clear that model I predicts no extrinsic noise (CV^2 is inversely proportional to mean) while models II, III, and IV predict extrinsic noise stemming from binomial partitioning and cell cycle length variability. Since models II and III are special cases of model IV, we shall only consider the latter. Using Eqs. (57) and (58) for lineage distributions, we can write the coefficient of variation squared in terms of the mean protein expression level

$$CV_s^2 = \frac{\mathcal{E}_{s,\text{int}}(M, N, \alpha)}{\langle n \rangle_s^H} + \mathcal{E}_{s,\text{ext}}(M, N), \qquad (63)$$

where the positive functions $\mathcal{E}_{s,\text{int}}$ and $\mathcal{E}_{s,\text{ext}}$ are given by

$$\mathcal{E}_{s,\text{int}}(M, N, \alpha) = \alpha \left[\frac{6M(M-1) + 4N(5N + 3 - 8M)}{3M(M-1) + 6N(3N + 1 - 2M)} \right] + 1, \qquad (64)$$

$$\mathcal{E}_{s,\text{ext}}(M, N) = \frac{4N^4 + 40N^3 - 4[3M(M+3) - 5]N^2 + 12(M^2 - 1)MN - 3M^2(M-1)^2}{3[M(M-1) + 2N(3N + 1 - 2M)]^2}. \qquad (65)$$

This shows that in the limit of abundant proteins, the protein noise as measured by the coefficient of variation squared, tends to a constant $\mathcal{E}_{s,\text{ext}}(N, M)$ which is independent of the intrinsic protein dynamics. This limiting value is only controlled by the cell cycle length variability via M and N. The intrinsic noise, on the other hand, is governed by both cell cycle length variability and protein burst size.

To study the dependencies on cell cycle length variability we let $M = uN$, where $u \in [0, 1]$ represents the fraction of the cell cycle spent in the nonreplicated phase. It can be shown that if u and α are fixed, then the internal component of noise given by the first term in Eq. (63) increases with N while the external component given by the second term decreases with N (see Appendix C). Hence, unexpectedly, CV_s^2 can increase or decrease with cell cycle length variability. More specifically in Appendix C [see Eq. (C9)] we show that if y is above a threshold, then protein noise increases with increasing cell cycle variability. This condition is met in mammalian cells and yeast cells [33,34] since the range of y and α is greater than one. In contrast in bacteria, due to the very low mean mRNA produced per cell cycle (as low as 0.1 [32]), the condition is not necessarily met and hence it is possible to have CV_s^2 decrease with increasing cell cycle length variability [the dependence is, however, very weak as can be seen in Fig. 7(a)]. Note that one can derive population expressions equivalent to Eqs. (64) and (65). It can be proved (see Appendix C) that unlike for lineages, CV_p^2 computed from population snapshots always increases with increasing cell cycle length variability (see Fig. 7).

Next, we determine bounds on the external noise floor for $N \geqslant 1$. First of all, we note that since the extrinsic noise floor $\mathcal{E}_{s,\text{ext}}$ decreases monotonically with N, its upper bound must be given by $N = 1$ while its lower bound by $N \to \infty$. For the case of purely exponential cell cycle length variability, i.e., $N = 1$, we then find that

$$\mathcal{E}_{s,\text{int}}(u, \alpha) = \left(\frac{4\alpha}{3} + 1 \right) - \frac{2\alpha u(1-u)}{3(u^2 - 5u + 8)}, \qquad (66)$$

$$\mathcal{E}_{s,\text{ext}}(u) = \frac{1}{3} + \frac{4u(1-u)(2-u)(4-u)}{3(u^2 - 5u + 8)^2}, \qquad (67)$$

which is maximal when $u \approx 0.55$ at 0.39. On the other hand, in the opposite limit of deterministic cell cycle times, i.e.,

$N \to \infty$, the noise contributions become

$$\mathcal{E}_{s,\text{int}}(u, \alpha) = \left(\frac{10\alpha}{9} + 1 \right) - \frac{8\alpha u(1-u)}{9(u^2 - 4u + 6)}, \qquad (68)$$

$$\mathcal{E}_{s,\text{ext}}(u) = \frac{1}{27} + \frac{4u(1-u)(7u^2 - 22u + 12)}{27(u^2 - 4u + 6)^2}, \qquad (69)$$

which is minimal when $u \approx 0.87$ at 0.034. Hence, it follows that the external noise floor for $N \geqslant 1$ is contained in the approximate interval (0.034,0.39).

Note that for $u = 0$ or 1, the special case without a replication phase, Eq. (63) together with Eqs. (66) and (67) almost perfectly recovers a recent result in the literature for the CV^2 of protein fluctuations in a model with exponential cell cycle length variability [see Eq. (7) of Ref. [26]], $CV^2 = (4/3)(\alpha + 1/2)/\langle n \rangle + 1/3)$. The same result is obtained for population snapshot calculations. The slight discrepancy is likely caused by the continuum approximation for protein levels in [26].

An experimental value of approximately 0.1 has been measured for the extrinsic noise floor in the bacterium *E. coli* (see Fig. 2B of [32]) which falls within the range of our theory (as described above). From the limited lineage data shown in the Taniguchi *et al.* paper (Fig. 2C shows three lineages with 9 cell division events), it is not clear which Erlang distribution would best fit their data. However, if we consider replication to occur in the middle of the cycle ($M = N/2$) and the cell cycle length to be well described by an Erlang distribution with $N = 5$ phases, then Eq. (65) predicts a value of $\mathcal{E}_{s,\text{ext}} \approx 0.12$ which is remarkably close to the experimental value of 0.1. Note that $N = 5$ is not unrealistic; it would imply a maximum difference of up to $1/\sqrt{N} \approx 40\%$ of the cell cycle length from its mean which is consistent with some recent experiments in *E. coli* [27,35]. The noise decomposition above can also be done for population snapshots using the equations for the mean and variance derived in Appendix D. It is found that the theory predicts the same value of $\mathcal{E}_{p,\text{ext}} \approx 0.1$ for extrinsic noise, and hence this result is insensitive to the type of observations made.

VII. CONCLUSION

In this paper we have derived the PGF corresponding to the stable protein number distributions in stochastic gene expression models with cell cycle length variability. Our work has the following special features: (i) the solution method used
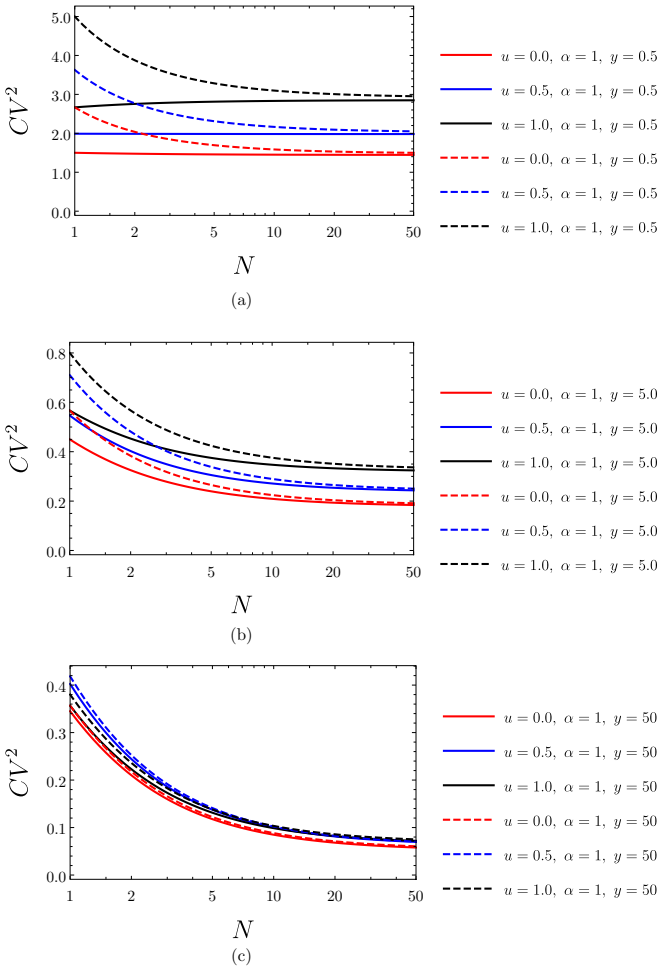
FIG. 7. Effects of the cell cycle variability on the size of protein fluctuations showing the difference between lineage (solid) and population snapshot (dotted) observations. Plots of the coefficient of variation squared CV^2 as a function of the mean burst size α, the mean number of mRNAs produced in a cell cycle y, and the fraction of the cell cycle in the prereplication stage $u = M/N$. The mean and variance for the computation of the CV_s^2 are given by Eqs. (57) and (58), whereas for CV_p^2 we have used Eqs. (D23) and (D26). A comparison of (a), (b), and (c) shows that the CV_s^2 can be both increasing and decreasing with increasing N, but CV_p^2 is always decreasing with increasing N, confirming theoretical predictions. For large values of y, we see that the protein noise only weakly depends on u, which can be understood from noting that this is the regime in which protein noise is dominated by the extrinsic noise floor.

allows the derivation of distributions rather than the mean and variance; (ii) the distributions for cell cycle length variability assumed by the model are of a very general form (hypoexponential) which fit the majority of experimentally measured distributions; (iii) the model allows the explicit description of the variation of transcription and burstiness with the position of the cell cycle (the cell age); and (iv) the calculations are done for both lineage and population snapshot observations which enhances the match between theory and experiments. A necessary underlying assumption of our approach to compute the PGF is that protein is stable, i.e., its decay occurs purely due to dilution by cell division. This is a good assumption when protein lifetimes are much longer than the mean cell cycle time such as in *E. coli* [15] and yeast [16]. In mammalian cells, about 70% percent of proteins are longer lived than the mean cell cycle time, and hence the approximation is also reasonable [33].

Special cases of our model can be found in the literature: (i) for a cell cycle composed of N phases, each of which is exponentially distributed in length, and assuming lineage observations, expressions for the mean and variance have been obtained in Refs. [22,23]; (ii) for a cycle whose length is exponentially distributed, an expression for the approximate protein number distribution (assuming large enough protein numbers) for both lineages and population snapshots has been derived in Ref. [26].

A major contribution of our study is the comparison of different models of gene expression including those with an implicit cell cycle description (model I) via effective protein degradation and models with an explicit cell cycle description with either regular (model II) or random interdivision times (models III and IV). We found that the protein distributions of models II–IV are well approximated by a negative binomial provided cell cycle length variability is large and y (the mean number of mRNA per cycle) is small (we shall henceforth call these the special conditions). In such cases, the implicit cell cycle model (model I) with renormalized parameters can describe the results of the explicit cell cycle models. When the special conditions are not met, the distributions show either a flat region near the mode or else have a right shoulder which in some cases can almost look like bimodality; of course, model I cannot capture these distributions. Such a case may be common for gene expression in mammalian cells and yeast where it is estimated that for many genes, y can take values in the range of 1 to about 600 [33,34]; in contrast in bacteria y has the range 0.1–10 [32] and hence deviations from negative binomial distributions are likely much less common. Also, we have shown that when the special conditions are not met, the distributions of models including replication or more complex age-dependent transcriptional dynamics cannot be described by models that assume constant transcription through the cell cycle such as those found in [26,27]. Our analysis shows that in a model assuming Erlang distributed cell cycle duration and replication time, for lineages, the coefficient of variation squared can either increase or decrease with cell cycle variability whereas for population snapshots, the coefficient of variation squared increases with cell cycle variability. We also show that the the coefficient of variation squared has a complex dependence on the replication time; it is practically independent of the replication time for large y but increases monotonically with the replication time for small y. Finally, we show that given experimental cell cycle length distributions for *E. coli* and assuming replication occurs halfway through the cell cycle, our theory predicts a value for extrinsic noise which is within a few percent of that measured in Ref. [32].

Despite its generality, our study has a number of limitations: (i) The analytical approach cannot be extended to derive mRNA distributions. This is since mRNA typically degrades faster than the mean cell cycle time [32] and hence is not stable, which is a necessary assumption to solve for the PGF. (ii) We have assumed binomial partitioning. While

this assumption presents the simplest reasonable model of stochastic partitioning, it likely fails for those proteins which are highly localized [36]. (iii) We have assumed that there is no correlation between the cell cycle duration of mother and daughter cells. Experiments show such a correlation exists [37,38]. Overcoming these limitations is key to expanding the realism of the model and is the subject of ongoing research.
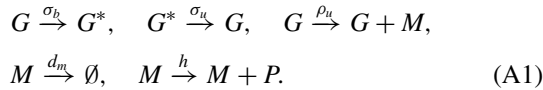
APPENDIX A: DERIVATION OF MODELS I AND II FROM A MODEL WITH AN EXPLICIT DESCRIPTION OF mRNA AND OF A CELL CYCLE OF FIXED LENGTH

1. Derivation of model I

We consider a three-stage gene expression model of the type

$$
G \xrightarrow{\sigma_b} G^*, \quad G^* \xrightarrow{\sigma_u} G, \quad G \xrightarrow{\rho_u} G + M,
$$

$$
M \xrightarrow{d_m} \emptyset, \quad M \xrightarrow{h} M + P. \tag{A1}
$$

Note that there is no active protein decay and instead we assume protein decay occurs only due to binomial partitioning at cell division. The latter is assumed to happen at regularly spaced time intervals of length T. If the rate of promoter switching is very fast compared to the mRNA and protein timescales, then there is no need to explicitly model G^*. Rather, it is sufficient to model expression from a single promoter state with an effective transcription rate equal to the true transcription rate multiplied by the fraction of time that the gene is on. In other words, the three-stage model (A1)
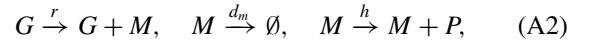
reduces to the two-stage model

$$
G \xrightarrow{r} G + M, \quad M \xrightarrow{d_m} \emptyset, \quad M \xrightarrow{h} M + P, \tag{A2}
$$

where $r = \rho_u \sigma_u / (\sigma_b + \sigma_u)$ is the effective mRNA transcription rate. Let the PGF be defined as

$$
G(z', z, t) = \sum_{m,n} (z')^m (z)^n P(m, n, t), \tag{A3}
$$

where $P(m, n, t)$ is the probability of observing m mRNAs and n proteins at time t. The PGF then satisfies a PDE which when nondimensionalized on the cell cycle timescale, $t = T\tau$, is given by

$$
\frac{\partial G}{\partial \tau} = rT(z' - 1)G + d_m T(1 - z') \frac{\partial G}{\partial z'} + hT z'(z - 1) \frac{\partial G}{\partial z'}. \tag{A4}
$$

Furthermore, binomial partitioning under cyclostationarity conditions [17] leads to the boundary condition (in nondimensional form)

$$
G(z', z, 0) = G\left(\frac{z' + 1}{2}, \frac{z + 1}{2}, 1\right). \tag{A5}
$$

By using the definitions of the mean numbers of proteins and mRNA

$$
\langle n \rangle(\tau) = \left. \frac{\partial G(z', z, \tau)}{\partial z} \right|_{z'=z=1}, \quad \langle m \rangle(\tau) = \left. \frac{\partial G(z', z, \tau)}{\partial z'} \right|_{z'=z=1}, \tag{A6}
$$

it is straightforward to show that the time evolution of the means is given by the coupled ordinary differential equations (ODEs)

$$
\frac{d\langle n \rangle(\tau)}{d\tau} = hT \langle m \rangle(\tau), \tag{A7}
$$

$$
\frac{d\langle m \rangle(\tau)}{d\tau} = rT - d_m T \langle m \rangle(\tau), \tag{A8}
$$

with the boundary conditions $2\langle n \rangle(0) = \langle n \rangle(1)$ and $2\langle m \rangle(0) = \langle m \rangle(1)$. Solving these ODEs one obtains

$$
\langle n \rangle(\tau) = \frac{hre^{-d_m T\tau}\{e^{\tau T d_m}[1 - (\tau + 1)T d_m] + 2e^{(\tau+1)T d_m}[(\tau + 1)T d_m - 1] + e^{T d_m}\}}{d_m^2(2e^{T d_m} - 1)}, \tag{A9}
$$

where τ is to be understood as the fractional cell age which equals zero when a cell is born and one just before a cell divides. Note that we do not show the equation for the mRNA since it is not relevant to our analysis. Hence, it follows that the mean number of proteins in single lineage and population measurements is given by

$$
\langle n \rangle_s = \int_0^1 \langle n \rangle(\tau) d\tau = \frac{\alpha y \left[d_m(3T d_m - 2) + \frac{1}{T - 2Te^{T d_m}} + \frac{1}{T} \right]}{2T d_m^2}, \tag{A10}
$$

$$
\langle n \rangle_p = \log(2) \int_0^1 2^{1-\tau} \langle n \rangle(\tau) d\tau = \frac{\alpha T y d_m}{T \log(2) d_m + \log^2(2)}, \tag{A11}
$$

where we used $y = rT$ and $\alpha = h/d_m$. Note that here we used the fact that when interdivision times are regularly spaced in time, the probability of observing a cell of age $t \in [0, T]$ is uniform for single lineage measurements and equal to $2^{1-t/T} \log 2/T$ for population measurements [17] (see also Appendix D for a derivation of the latter). Taking the limit that mRNA decays fast compared to the cell cycle length, i.e., $d_m T \to \infty$, we obtain

$$
\langle n \rangle_s \approx \frac{3\alpha y}{2}, \tag{A12}
$$

$$
\langle n \rangle_p \approx \frac{\alpha y}{\log(2)}. \tag{A13}
$$

Now, one may ask what constitutes an effective system of reactions that describes protein dynamics and which has the

same mean number of proteins as above. If we consider this effective system to be given by model I, that is the set of reactions

$$G \xrightarrow{r} G + mP, \quad P \xrightarrow{d'} \emptyset, \tag{A14}$$
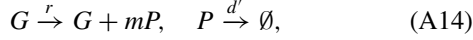
where m is a geometrically distributed random number with mean α, it follows that the mean number of proteins is given by

$$\langle n \rangle_I = \frac{r\alpha}{d'} = \frac{\alpha y}{d'T}. \tag{A15}$$

Equating Eq. (A15) with (A12), we obtain the effective protein degradation rates for single lineage measurements,

$$d' = \frac{2}{3T}. \tag{A16}$$

Similarly equating Eq. (A15) with (A13), we find that the effective protein degradation rates for population measurements are

$$d' = \frac{\log(2)}{T}. \tag{A17}$$

2. Derivation of model II

Starting from the same mRNA model as used in Appendix A 1 we now derive model II. Consider the PDE (A4) with boundary condition (A5). To solve this PDE, we employ the method of characteristics which yields the system of equations

$$\frac{d\tau}{ds} = 1, \tag{A18}$$

$$\frac{dz'}{ds} = d_m T(z' - 1) + hTz'(1 - z), \tag{A19}$$

$$\frac{dz}{ds} = 0, \tag{A20}$$

$$\frac{dG}{ds} = rT(z' - 1)G, \tag{A21}$$

where s is the characteristic variable. In particular, we see that $\tau = s$ and $z = z_0$ is constant.

To consider the case of unstable mRNA, we define $\varepsilon = 1/(d_m T)$ as the asymptotic variable. We see that to achieve a dominant balance as $\varepsilon \to 0$ (or $d_m T \to \infty$) we could choose $hT = O(d_m T)$. As before, we use $\alpha = \lim_{d_m T \to \infty} (hT/d_m T)$ and taking this scaling we see that as $\varepsilon \to 0$ we get

$$\varepsilon \frac{dz'}{ds} = z' - 1 + \alpha z'(1 - z) + O(\varepsilon), \tag{A22}$$

and therefore have to leading order $(z' - 1) - \alpha z'(z - 1) = 0$. This results in $z' \to 1/[1 - \alpha(z - 1)]$ quickly in the limit of $\varepsilon \to 0$ and, therefore, the following effective PDE for the PGF:

$$\frac{\partial G}{\partial \tau} = rT \frac{\alpha(z - 1)}{1 - \alpha(z - 1)} G. \tag{A23}$$

Note that we can then define $\tilde{G}(z, \tau) = \lim_{z' \to 1} G(z', z, \tau) = \sum_n y^n P(n, \tau)$, i.e., the PGF for the marginal protein distribu-

tion. It follows now that $\tilde{G}$ satisfies the following cyclostationary system in the limit of $\varepsilon \to 0$:

$$\frac{\partial \tilde{G}}{\partial \tau} = rT \frac{\alpha(z - 1)}{1 - \alpha(z - 1)} \tilde{G}, \tag{A24a}$$

$$\tilde{G}(z, 0) = \tilde{G}\left(\frac{z + 1}{2}, 1\right). \tag{A24b}$$

Note that this is the same system satisfied by $G(z, t)$ from model II, Eqs. (4) and (7) in the cyclostationary limit where we drop the superscript for the generation. Therefore, we have shown that the two-stage expression model (A2) with binomial partitioning converges to model II in the limit of fast mRNA decay compared to the cell cycle length, i.e., $d_m T \gg 1$.

APPENDIX B: STOCHASTIC SIMULATIONS

Stochastic simulations in this work were carried out using a method similar to the First Division Algorithm in [27], which effectively is the (modified) next reaction method with the addition of cell division and observation events. Below follows a description of the exact procedures used to calculate stochastic realizations of models II, III, and IV in this paper. The code used for this work can be found online in Ref. [39].

1. Lineage simulation

For the simulation of lineage data we consider a single cell in a mother machine that we continuously track. Measurements of the cell's contents are taken at intervals defined by some observation time distribution. For this paper we take measurements at regular intervals, i.e., a delta distribution for the observation interval distribution.

(1) *Initialization.* Start a cell in the mother machine at time $t = 0$ with initial molecule content n. Assign a phase j and age within the phase a.

(2) Generate waiting times τ_r, τ_p, and τ_o, the time until the next biochemical reaction, the next phase change, and the next observation, respectively.

(3) Until $t \geqslant t_{\text{final}}$. Pick $\Delta = \min(\tau_r, \tau_p, \tau_o)$, update all waiting times (τ_r, τ_p, τ_o) by $\tau \to \tau - \Delta$, and update phase age $a \to a + \Delta$. Based on the minimum found for the waiting times proceed to the following:

(a) *Biochemical reaction.* Select the biochemical reaction occurring, e.g., using Gillespie-type algorithm, and update cell molecule content. Generate a new time until next biochemical reaction τ_r.

(b) *Phase progression.* Based on the current phase of the cell proceed to the following:

(i) If phase of the cell is N, reset the new phase of the cell to the first phase, $j \to 1$. Binomially partition the cell contents across the cell in mother machine and daughter cell, discard daughter cell, and keep following cell in mother machine.

(ii) Otherwise, set the new phase of the cell, $j \to j + 1$.

Set the age of cell in the new phase to $a = 0$ and update the biochemical rate parameters according to the current phase j. Generate a new time until next phase progression τ_p and a new time until next biochemical reaction τ_r.

(c) *Observation.* Write the current cell contents and other quantities of interest to disk. Generate a new time until next observation τ_o.

Update simulation time $t \rightarrow t + \Delta$ and return to the start of step 3.

2. Population snapshot simulation

For the simulation of population snapshot data, we consider an initial batch of cells that we let grow and divide until some final time t_{final} at which point we measure the contents of all the cells present.

(1) *Initialization.* Start batch of cells at time $t_i = 0$ with initial molecule contents n_i, where i index rolls over the batch of present cells. Assign a phase j_i and age within the phase a_i to each cell in the batch. Start with the first cell in the batch, $l \rightarrow 1$.

(2) Until all cells in batch have reached final time t_{final} pick the next cell in the batch and proceed to step 3.

(3) Generate waiting times τ_r and τ_p the time until the next biochemical reaction and the next phase change, respectively. Set the time until observation as $\tau_o \rightarrow t_{\text{final}} - t_l$.

(4) Until $t_l = t_{\text{final}}$. Pick $\Delta = \min(\tau_r, \tau_p, \tau_o)$, update all waiting times (τ_r, τ_p, τ_o) by $\tau \rightarrow \tau - \Delta$, and update phase age $a \rightarrow a + \Delta$. Based on the minimum found for the waiting times proceed to the following:

(a) *Biochemical reaction.* Select biochemical reaction occurring, e.g., using Gillespie-type algorithm, and update cell molecule content. Generate a new time until next biochemical reaction τ_r.

(b) *Phase progression.* Based on the current phase of the cell proceed to the following:

(i) If phase of the cell is N, reset the new phase of the cell to the first phase, $j \rightarrow 1$. Binomially partition the cell contents across the current cell and a daughter cell. Add the daughter cell with its content to the batch of cells with start time $t_{i'} = t_l$, $a_{i'} = 0$, and $j_{i'} = 1$.

(ii) Otherwise, set the new phase of the cell, $j \rightarrow j + 1$.

Set the age of cell in the new phase to $a = 0$ and update the biochemical rate parameters according to the current phase j. Generate a new time until next phase

progression τ_p and a new time until next biochemical reaction τ_r.

(c) *Observation.* Write the current cell contents and other quantities of interest to disk. Let $l \rightarrow l + 1$ and return to step 2.

Update simulation time $t_l \rightarrow t_l + \Delta$ and return to the start of step 4.

APPENDIX C: COEFFICIENT OF VARIATION PROOFS FOR MODEL IV

1. Effect of burst size α and mRNA production per cycle y

First, we prove that CV^2 (and CV) monotonically decrease with increasing α and y. We note using Eqs. (57) and (58) that the mean and variance of protein numbers in model IV with replication can be written as

$$\left(\langle n \rangle_s^H\right)^2 = \alpha^2 y^2 A, \quad \left(\sigma_s^H\right)^2 = \alpha y B + \alpha^2 y C + \alpha^2 y^2 D,$$

where A, B, C, D are functions of M, N solely. First, we observe that $A > 0$ must hold by the observation that the mean of the model IV is necessarily positive if $\alpha > 0$ and $y > 0$. This means that we have for the squared coefficient of variation

$$CV^2 = \frac{1}{A}\left(\frac{B}{\alpha y} + \frac{C}{y} + D\right).$$

Note that $CV^2 \geqslant 0$ must hold for all $\alpha \geqslant 0$ and $y \geqslant 0$. By considering the limit $\alpha \downarrow 0$ we deduce that $B \geqslant 0$. Similarly, by considering the limit $y \downarrow 0$ we find that for all $\alpha \geqslant 0$ we must have $B + C\alpha \geqslant 0$ and therefore $C \geqslant 0$. Finally, from the limit $y \rightarrow \infty$ we deduce that $D \geqslant 0$. It then immediately follows that CV^2 and CV decrease monotonically with α and/or y increasing. Note that this argument applies equally well to the population snapshot case.

2. Effect of cell cycle variability N for lineages

Next, we show that the extrinsic noise floor, i.e., the limit of CV^2 for abundant proteins, is a monotonic decreasing function of N. Recalling Eq. (65) and rewriting the noise floor $\mathcal{E}_{s,\text{ext}}(M, N)$ using $M = uN$, where $u \in [0, 1]$, we have

$$\mathcal{E}_{s,\text{ext}}(u, N) = \frac{N^2(-3u^4 + 12u^3 - 12u^2 + 4) + N(6u^3 - 36u + 40) - 3u^2 - 12u + 20}{3[N(u^2 - 4u + 6) - u + 2]^2}. \tag{C1}$$

We then note that the derivative of the noise floor with respect to N is given by

$$\frac{d\mathcal{E}_{s,\text{ext}}}{dN} = -\frac{4[(-3u^3 + 16u^2 - 48u + 40) + N(3u^4 - 18u^3 + 58u^2 - 92u + 56)]}{3[N(u^2 - 4u + 6) + (2 - u)]^3} < 0. \tag{C2}$$

The inequality follows from considering every term in brackets for $N > 0$ and $u \in [0, 1]$ and showing it is positive. As a result, we see that the external noise floor is a monotonic decreasing function of increasing N, which can be interpreted as decreasing cell cycle length variability.

Next, we show that CV_s^2 is not monotonic in N for single-cell lineage measurements. Recalling Eqs. (57) and (64), we can rewrite the internal noise component and protein mean

using u and N as

$$\mathcal{E}_{s,\text{int}}(u, N) = 1 + 2\alpha - \frac{8\alpha N(2 - u)}{3N(6 - 4u + u^2) + 3(2 - u)}, \tag{C3}$$

$$\langle n \rangle_s^H = \left(\frac{6 - 4u + u^2}{2} + \frac{2 - u}{2N}\right)\alpha y. \tag{C4}$$

Note that it then follows that

$$\frac{d\mathcal{E}_{s,\text{int}}}{dN} = -\left\{\frac{8(2-u)^2}{3[N(u^2-4u+6)+(2-u)]^2}\right\}\alpha < 0, \quad (C5)$$

$$\frac{d\langle n\rangle_s^H}{dN} = -\left(\frac{2-u}{2N^2}\right)\alpha y < 0, \quad (C6)$$

$$\frac{d\left(\mathcal{E}_{s,\text{int}}/\langle n\rangle_s^H\right)}{dN} = \frac{2(2-u)\{2\alpha[N(3u^2-4u+2)+3(2-u)]+3N(u^2-4u+6)+3(2-u)\}}{3\alpha y[N(u^2-4u+6)+(2-u)]^3} > 0. \quad (C7)$$

Since the external component of the coefficient of variation squared has earlier been proved to decrease with N, it follows that for some positive functions a_1, a_2 of u, N, α we find

$$\frac{dCV_s^2}{dN} = \frac{a_1}{y} - a_2, \quad (C8)$$

which shows that CV_s^2 can be both increasing or decreasing as a function of N, depending on the exact values of α, y. In fact, by considering a_1 and a_2 and their extreme behavior (which is for $u = 1$) one can show that for

$$y > \frac{3+N}{5+7N} + \frac{1}{\alpha}\frac{3+9N}{2(5+7N)}, \quad (C9)$$

the protein noise for lineages is monotonically decreasing, whereas when this condition is not satisfied, we can find $u = M/N \in [0, 1]$ so that, perhaps counterintuitively, decreasing cell cycle variability, i.e., increasing N, leads to an increase in protein noise. On the other hand, this shows that if mRNA production is large enough, the protein noise does become monotonic as a function of cell cycle variability.

3. Effect of cell cycle variability N for population snapshots

For population snapshots we note from results in Appendix D, Eqs. (D23) and (D26) in particular, that the internal noise component is composed of

$$\mathcal{E}_{p,\text{int}}(u, N) = 1 + 2\alpha - \frac{\alpha(2-u)2^{1+u}}{3}\frac{N(2^{1/N}-1)}{2^{1/N}}, \quad (C10)$$

$$\langle n\rangle_p^H = \frac{2^{1-u}}{N(2^{1/N}-1)}\alpha y. \quad (C11)$$

From here it is easy to show that $\langle n\rangle_p^H$ and $\mathcal{E}_{p,\text{int}}(u, N)$ monotonically grow and decay, respectively, as N increases. As a result, we note that their ratio, the internal noise component $(\mathcal{E}_{p,\text{int}}/\langle n\rangle_p^H)$, decays monotonically when N increases, i.e., the opposite scenario to what happens for single-cell lineage measurements.

Finally, we will show that $\mathcal{E}_{p,\text{ext}}$ decays monotonically when N increases and as a result that, in contrast to the lineage framework, the total protein noise CV_p^2 is always decreasing as a function of increasing N, regardless of the values of α, y. We note that this proof is more cumbersome than for the lineage measurements and we only sketch the details here. We start by noting that, with $u \in [0, 1]$ as before,

$$\mathcal{E}_{p,\text{ext}}(u, N) = -\frac{2^u}{6}\frac{N(2^{1/N}-1)}{2^{2/N}}\{2^u[3(2^{\frac{1}{N}}-1)N(u-2)^2$$
$$+ 2^{\frac{1}{N}}(16-9u)+3u-4] - 12 \times 2^{1/N}\}$$
$$+ (-4^u + 3 \times 2^u - 1). \quad (C12)$$

which does not allow us to immediately make a conclusion about the behavior of the ratio of the two quantities. Therefore, we note that the internal component of the coefficient of variation squared in fact grows as N increases as shown by

To make computations slightly more tractable, we then make the transformation to $q = 2^{1/N}$ (which yields $q > 1$ for $N > 0$ and $q \to 1$ when $N \to \infty$) and note that $dq/dN < 0$ for all N. This leaves us to show that $d\mathcal{E}_{p,\text{ext}}/dq > 0$ for all $q > 1$ in order to show that the external noise component satisfies $d\mathcal{E}_{p,\text{ext}}/dN < 0$. We start with

$$h_1(q) = \frac{d\mathcal{E}_{p,\text{ext}}}{dq} = \underbrace{\frac{2^u \log 2}{6(q \log q)^3}}_{>0} h_2(q). \quad (C13)$$

From here we note that $\lim_{q\downarrow 1} h_2(q) = \lim_{q\downarrow 1} h_2'(q) = \lim_{q\downarrow 1} h_2''(q) = 0$ and

$$\frac{d^3 h_2}{dq^3} = \underbrace{\frac{1}{q^3}}_{>0} h_3(q). \quad (C14)$$

The new function h_3 satisfies, for $u \in [0, 1]$,

$$\lim_{q\downarrow 1} h_3(q) = 6\{2^u[u^2\log(8) - 12u\log(2)$$
$$+ 2 + \log(4096)] - 6\} > 0,$$
$$\frac{dh_3}{dq} = \underbrace{\frac{1}{q}}_{>0} h_4(q).$$

In a similar fashion we proceed to show that for h_4 and $u \in [0, 1]$ we can write

$$\lim_{q\downarrow 1} h_4(q) = 2(2^u\{u[(u-4)\log(8) - 48]$$
$$+ 78 + \log(4096)\} - 42) > 0,$$
$$\lim_{q\downarrow 1} h_4'(q) = 2(3 \times 2^u\{u^2\log(2) - 2u[11 + \log(4)]$$
$$+ 38 + \log(16)\} - 78) > 0,$$
$$\frac{d^2 h_4}{dq^2} = \underbrace{\frac{8}{q}}_{>0} h_5(q),$$

where $h_5(q) = q[2^u(16 - 9u) - 12] + (-3 \times 2^u u + 5 \times 2^u - 3)$ is a positive linear function in q when $u \in [0, 1]$. By arriving at $h_5 > 0$ for $q > 1$ we find cascading back that $h_4'', h_4', h_4 > 0$ for $q > 1$. This then proves $h_3', h_3 > 0$ for $q > 1$ which in turn shows that $h_2''', h_2'', h_2', h_2 > 0$ for $q > 1$. Finally, having proven that $h_2 > 0$ we find $d\mathcal{E}_{p,\text{ext}}/dq = h_1 > 0$ for $q > 1$, which was what we needed to prove to show that $d\mathcal{E}_{p,\text{ext}}/dN < 0$.

APPENDIX D: POPULATION SNAPSHOTS

In this Appendix, we correct for the fact that in a population observation, the cell phases and observation times differ from those in single-cell measurements. In order to construct the PGF for the population observations we start from the expression

$$G_p(z) = \sum_{i=1}^{N} \Pi_{i,p} G_{i,p}(z), \tag{D1}$$

where $G_{i,p}(z)$ denotes the PGF for phase i in the population measurement case and $\Pi_{i,p}$ is the probability to observe a cell in phase i. Note that we take the distributions for each phase to be normalized to unity to reflect that the $G_{i,p}$ are in fact the marginal distributions for protein in each phase i. We will next show how to derive expressions for $G_i(z)$ and $\Pi_{i,p}$.

1. Phase distribution

While the probability of observing a cell in cell cycle phase i, denoted as Π_i, was obvious when performing single-cell measurements, this is not the case for population measurements. This is since each time a cell divides, two cells start in phase 1 and hence we generally expect that the probability of being in phase i decreases with i. We next derive an expression for the probability $\Pi_{i,p}$ for population measurements.

Let the average number of cells under cyclostationary assumptions in cell cycle phase i at time t be denoted by $\langle C_i(t) \rangle$. Then, it immediately follows from the model specification that when each phase of the cell cycle is exponentially distributed with parameter k, the time-evolution equations are given by

$$\frac{d\langle C_1(t) \rangle}{dt} = -k\langle C_1(t) \rangle + 2k\langle C_N(t) \rangle, \tag{D2}$$

$$\frac{d\langle C_i(t) \rangle}{dt} = -k\langle C_i(t) \rangle + k\langle C_{i-1}(t) \rangle, \quad i = 2, \dots, N. \tag{D3}$$

The factor 2 in the first equation stands for cellular division: every time a cell divides (leaving phase N), two cells start in phase 1. In the case that we are tracking single-cell lineages, then the factor 2 is replaced by 1. More generally, for cases with asymmetric division (after division some cells differentiate) this factor 2 can be replaced by a factor $\nu \in [0, 2]$. We are interested in finding the probability that a cell is in phase i, which is given by

$$\Pi_i(t) = \frac{\langle C_i(t) \rangle}{\sum_{i=1}^{N} \langle C_i(t) \rangle}. \tag{D4}$$

Note that this expression is valid for both population snapshot and lineage observations. While this quantity will change with time, it will eventually approach a steady-state value and this is what we are here interested in. We make the ansatz that given long enough time, $\langle C_i(t) \rangle = \lambda_i \langle C_1(t) \rangle$ where λ_i are some time-independent constants which need to be determined (except for $\lambda_1 = 1$ which follows immediately). Substituting this assumption in Eq. (D4) we find

$$\Pi_{i,p} = \frac{\lambda_i}{1 + \sum_{j=2}^{N} \lambda_j}. \tag{D5}$$

Hence, next we determine the values of λ_i. Substituting the ansatz in Eqs. (D2) and (D3), we find the relationship

$$\frac{\lambda_{i-1}}{\lambda_i} = 2\lambda_N, \quad i = 2, \dots, N \tag{D6}$$

with $\lambda_1 = 1$. Solving this recurrence relation one finds

$$\lambda_i = 2^{(1-i)/N}. \tag{D7}$$

Substituting the latter expression in Eq. (D5) and simplifying one finally obtains

$$\Pi_{i,p} = \frac{2^{1/N} - 1}{2^{(i/N)-1}}, \quad i = 1, \dots, N \tag{D8}$$

which in the limit of $N \to \infty$ yields the familiar age structure in a population which doubles at mitosis, i.e., $f(t) = 2^{1-t/T} \log 2 / T$. Note that this result is in contrast to the phase distribution in the single-cell measurement case, which was given by a uniform distribution, i.e., $\Pi_{i,s} = 1/N$ for all i. In a population we are more likely to observe cells in an early cell phase compared to lineage data. Note that incidentally we can derive the population growth rate (see next section) from this formalism, which is necessary to derive the age distribution in a population for each phase.

2. Population growth rate

To derive the population growth rate, we consider $\langle C(t) \rangle = \sum_i \langle C_i(t) \rangle$, the expected total number of cells, i.e., the size of the population. From Eqs. (D2) and (D3) we then find in the long-time limit when the proportion of cells in each phase relative to the total number of cells, i.e., $\Pi_{i,p}$, becomes stable that the growth of the cell population is given by

$$\frac{d\langle C(t) \rangle}{dt} = k\langle C_N(t) \rangle = k\Pi_{N,p}\langle C(t) \rangle = k(2^{1/N} - 1)\langle C(t) \rangle. \tag{D9}$$

This shows that the average number of cells in the population grows like $\langle C(t) \rangle \propto e^{\lambda t}$ where $\lambda = k(2^{1/N} - 1)$. Note that in the limit of $N \to \infty$ and $k = N/T$ this becomes $\lambda = \log 2 / T$, showing that the population doubles in size after every cell cycle when we consider the case of a deterministic cell cycle of length T.

3. Population age distribution

Let us consider $\langle C_i(t, \tau) \rangle$ the average number of cells in a population that are in cell phase i at time t that have been in that cell phase for a duration τ, i.e., are of an age τ. After a small time duration δ, all the cells will either advance to an age $\tau + \delta$ or advance to the next cell phase. Therefore, we can write the conservation equation

$$\langle C_i(t + \delta, \tau + \delta) \rangle = \langle C_i(t, \tau) \rangle - \langle C_i(t, \tau) \rangle k\delta, \tag{D10}$$

where k is the rate of advancing to the next cell phase. Assuming that there is a stationary distribution for the age of the cell population at a phase i, $\pi_i(\tau)$, we can write $\langle C_i(t, \tau) \rangle$ as $\langle C_i(t, \tau) \rangle = \langle C_i(t) \rangle \pi_i(\tau)$, where $\langle C_i(t) \rangle$ as before is the expected number of cells in cell phase i at time t. Introducing this factorization of $\langle C_i(t, \tau) \rangle$ in Eq. (D10), and taking the

limit $\delta \rightarrow 0$, we get the differential equation

$$\frac{d\langle C_i(t)\rangle}{dt}\pi_i(\tau) + \frac{d\pi_i(\tau)}{d\tau}\langle C_i(\tau)\rangle = -\langle C_i(t)\rangle\pi_i(\tau)k, \quad \text{(D11)}$$

where we have used the chain rule to compute the derivative of $d\langle C_i(x,x)\rangle/dx|_{t,\tau}$. Given that in the long-time limit, the population of cells is growing at an average rate λ, which we derived in the previous section, and since the probability of finding a cell in a certain phase i, Π_i, is constant in time, the number of cells in a given phase has to grow with the same rate as the population. This implies that therefore $d\langle C_i(t)\rangle/dt = \lambda\langle C_i(t)\rangle$. Introducing this equality in Eq. (D11), we get an equation solely for $\pi_i(\tau)$,

$$\lambda\pi_i(\tau) + \frac{d\pi_i(\tau)}{d\tau} = -\pi_i(\tau)k, \quad \text{(D12)}$$

which can be solved by

$$\pi_i(\tau) = (\lambda + k)e^{-(\lambda+k)\tau} = k2^{1/N}e^{-k2^{1/N}\tau}. \quad \text{(D13)}$$

Finally, we note that there is a direct link between the age distribution at measurement and the phase length distribution on a population level. In order to achieve a stable distribution for the age at measurement in the long-time limit we note that the population phase length distribution has to be consistent with the age distribution. Since the population age distribution at observation in each phase is still exponential, but now with rate $k2^{1/N}$, the distribution for the time it takes to progress to

the next cell phase in a population must also be exponentially distributed with the same rate. This means that in a population of cells we observe a quicker progression through the phases (in distribution on a population level) than when following a single cell. This can be understood from the observation that cells that quickly progress through their phases (and thus divide quickly) will be more abundant in populations than in single-cell lineages.

4. Population snapshot observation distribution

By the results of the previous section, we can derive the distribution for population snapshot measurements from the single-cell measurement statistics via a rescaling of the phase progression rates. The distribution for the length of the cell phases is still exponential, which means the Markovian nature of the problem remains intact in the population case. This implies that in all the equations for the PGF derived earlier (for single lineages) we replace $k \mapsto k2^{1/N}$ to find the corresponding population equations. The recipe for calculating population observations therefore is as follows:

(1) To obtain $G_{i,p}(z)$, replace $k \mapsto k2^{1/N}$ in the expressions for the lineage distributions $G_{i,s}(z)$, e.g., in Eq. (50).

(2) Average over the different phases via Eq. (D1) using $\Pi_{i,p}$ from Eq. (D8).

Using these steps, we can then solve for the snapshot observation distribution for model IV with replication which we will write as $G_p(z) = G_{N,p}(z)W(z)$, where

$$G_{N,p}(z) = \prod_{s=0}^{\infty}\left[1 - \frac{2xy}{x(N2^{1/N} + 2y) - N2^{1/N}2^s}\right]^{N-M}\left[1 - \frac{xy}{x(N2^{1/N} + y) - N2^{1/N}2^s}\right]^M, \quad \text{(D14)}$$

$$W(z) = (2^{\frac{1}{N}} - 1)N(x - 1)\left(\frac{\left[\frac{2xy}{N(x-1)} + 2^{\frac{1}{N}}\right]^{N-M}\left\{\left[\frac{xy}{N(x-1)} + 2^{\frac{1}{N}}\right]^M - 1\right\}}{(2^{\frac{1}{N}} - 1)N(x - 1) + xy} + \frac{\left[\frac{2xy}{N(x-1)} + 2^{\frac{1}{N}}\right]^{N-M} - 1}{(2^{\frac{1}{N}} - 1)N(x - 1) + 2xy}\right). \quad \text{(D15)}$$

This result then yields the explicit snapshot distribution for all the other models considered in this paper. We recover the special case of model III by taking $M = N$ which in turn simplifies the observation distribution to

$$G_{N,p}(z) = \prod_{s=0}^{\infty}\left[1 - \frac{xy}{x(N2^{1/N} + y) - N2^{1/N}2^s}\right]^N, \quad \text{(D16)}$$

$$W(z) = (2^{\frac{1}{N}} - 1)N(x - 1)\frac{\left[\frac{xy}{N(x-1)} + 2^{\frac{1}{N}}\right]^N - 1}{(2^{\frac{1}{N}} - 1)N(x - 1) + xy}. \quad \text{(D17)}$$

Finally, we obtain a generalized version of model II by taking $N \rightarrow \infty$; we call this generalized because unlike the one in the main text, here for completeness we return to $u = M/N$ so that we can model replication in model II. This yields

$$G_{N,p}(z) = \prod_{s=0}^{\infty}\exp\left[\frac{(1-u)2xy}{2^s - x}\right]\exp\left(\frac{uxy}{2^s - x}\right) = \exp\left\{[2(1-u) + u]xy\sum_{s=0}^{\infty}\frac{1}{2^s - x}\right\}, \quad \text{(D18)}$$

$$W(z) = (\log 2)(x - 1)\left[\frac{\exp\left\{\frac{(1-u)[x(2y+\log 2)-\log 2]}{x-1}\right\}\left(\exp\left\{\frac{u[x(y+\log 2)-\log 2]}{x-1}\right\} - 1\right)}{x(y + \log 2) - \log 2} + \frac{\exp\left\{\frac{(1-u)[x(2y+\log 2)-\log 2]}{x-1}\right\} - 1}{x(2y + \log 2) - \log 2}\right]. \quad \text{(D19)}$$

Note that for $u = 1$ this further simplifies to

$$G_p(z) = \frac{(x - 1)\log 2}{xy + (x - 1)\log 2}\left\{\exp\left[\frac{xy + (x - 1)\log 2}{x - 1}\right] - 1\right\}\exp\left(xy\sum_{s=0}^{\infty}\frac{1}{2^s - x}\right), \quad \text{(D20)}$$

which is the population equivalent of model II presented in the main text. In particular, we can now write the mean of protein numbers for generalized models II, III, and IV, respectively,

$$\langle n \rangle_p = \frac{2^{1-u}}{\log 2} \alpha y, \tag{D21}$$

$$\langle n \rangle_p^E = \frac{1}{N(2^{1/N} - 1)} \alpha y, \tag{D22}$$

$$\langle n \rangle_p^H = \frac{2^{1-(M/N)}}{N(2^{1/N} - 1)} \alpha y, \tag{D23}$$

where we recall $u = M/N$ in the limit of $N \to \infty$. We can also write expressions for the variances for generalized models II, III, and IV:

$$\sigma_p^2 = \frac{2^{1-u}}{\log 2} \alpha y + \frac{4(u \log 2 - 2 \log 2 + 2^{-u}3)}{3 \log 2} \alpha^2 y$$
$$+ \frac{2^{1-2u}\{-4^u[u^2 \log^2 2 - u2 \log 2(1 + 2 \log 2) + 2 + 4 \log^2(2) + 4 \log 2] + 2^{u+1}(3 + 2 \log 2) - 2\}}{\log^2 2} \alpha^2 y^2, \tag{D24}$$

$$\left(\sigma_p^E\right)^2 = \frac{1}{N(2^{1/N} - 1)} \alpha y + \left[\frac{2}{(2^{\frac{1}{N}} - 1)N} - \frac{1}{3} 2^{2 - \frac{1}{N}}\right] \alpha^2 y + \left[\frac{1}{(2^{\frac{1}{N}} - 1)^2 N^2} - \frac{2^{1 - \frac{2}{N}}(3N + 1)}{3N}\right] \alpha^2 y^2, \tag{D25}$$

$$\left(\sigma_p^H\right)^2 = \frac{2^{1-(M/N)}}{N(2^{1/N} - 1)} \alpha y + \frac{4\left[3 \times 2^{-\frac{M}{N}} + 2^{-1/N}(2N - M) + M - 2N\right]}{3(2^{\frac{1}{N}} - 1)N} \alpha^2 y$$
$$+ \Bigg(-\frac{2^{1 - \frac{2}{N}}(3(2^{\frac{1}{N}} - 1)\{-M^2 + 2^{\frac{1}{N}}[-4MN + (M-3)M + 4N^2] + 4MN + M - 4N^2\})}{3(2^{\frac{1}{N}} - 1)^2 N^2}$$
$$+ \frac{\{4^{\frac{M}{N}}[-5\, 2^{\frac{1}{N}+2}N + 4N + 2^{\frac{N+2}{N}}(8N + 3)] + 3[2^{\frac{1}{N}}(2N + 3) - 2N]2^{\frac{M+N+1}{N}} - 3\, 2^{\frac{N+2}{N}}\}}{3(2^{\frac{1}{N}} - 1)^2 N^2}\Bigg) \alpha^2 y^2. \tag{D26}$$

Note that generalized model II for $u = 1$ simplifies to model II in the main text:

$$\langle n \rangle_p = \frac{\alpha y}{\log 2}, \quad \sigma_p^2 = \frac{1}{\log 2} \alpha y + \frac{6 - 4 \log 2}{3 \log 2} \alpha^2 y + \frac{1 - \log^2 2}{\log^2 2} \alpha^2 y^2 \simeq 1.44 \alpha y + 1.55 \alpha^2 y + 0.08 \alpha^2 y^2.$$

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

[2] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, Stochasticity in gene expression: From theories to phenotypes, Nat. Rev. Genet. **6**, 451 (2005).

[3] D. Schnoerr, G. Sanguinetti, and R. Grima, Approximation and inference methods for stochastic biochemical kinetics-a tutorial review, J. Phys. A: Math. Theor. **50**, 093001 (2017).

[4] V. Shahrezaei and P. S. Swain, Analytical distributions for stochastic gene expression, Proc. Natl. Acad. Sci. USA **105**, 17256 (2008).

[5] J. Dattani and M. Barahona, Stochastic models of gene transcription with upstream drives: Exact solution and sample path characterization, J. R. Soc., Interface **14**, 20160833 (2017).

[6] N. Popović, C. Marr, and P. S. Swain, A geometric analysis of fast-slow models for stochastic gene expression, J. Math. Biol. **72**, 87 (2016).

[7] D. Bratsun, D. Volfson, L. S. Tsimring, and J. Hasty, Delay-induced stochastic oscillations in gene regulation, Proc. Natl. Acad. Sci. USA **102**, 14593 (2005).

[8] N. Kumar, A. Singh, and R. V. Kulkarni, Transcriptional bursting in gene expression: analytical results for general stochastic models, PLoS Comput. Biol. **11**, e1004292 (2015).

[9] T. Tian and K. Burrage, Stochastic models for regulatory networks of the genetic toggle switch, Proc. Natl. Acad. Sci. USA **103**, 8372 (2006).

[10] R. Grima, D. R. Schmidt, and T. J. Newman, Steady-state fluctuations of a genetic feedback loop: An exact solution, J. Chem. Phys. **137**, 035104 (2012).

[11] P. Thomas, N. Popović, and R. Grima, Phenotypic switching in gene regulatory networks, Proc. Natl. Acad. Sci. USA **111**, 6994 (2014).

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

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

[14] N. Friedman, L. Cai, and X. S. Xie, Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression, Phys. Rev. Lett. **97**, 168302 (2006).

[15] M. R. Maurizi, Proteases and protein degradation inescherichia coli, Experientia **48**, 178 (1992).

[16] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and Erin K O'Shea, Quantification of protein half-lives in the budding yeast proteome, Proc. Natl. Acad. Sci. USA **103**, 13004 (2006).

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

[18] I. G. Johnston and N. S. Jones, Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions, Proc. R. Soc. A **471**, 20150050 (2015).

[19] A. H. K. Roeder, V. Chickarmane, A. Cunha, B. Obara, B. S. Manjunath, and E. M. Meyerowitz, Variability in the control of cell division underlies sepal epidermal patterning in arabidopsis thaliana, PLoS Biol. **8**, e1000367 (2010).

[20] A. Golubev, Applications and implications of the exponentially modified gamma distribution as a model for time variabilities related to cell proliferation and gene expression, J. Theor. Biol. **393**, 203 (2016).

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

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

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

[24] L. Robert, J. Ollion, J. Robert, X. Song, I. Matic, and M. Elez, Mutation dynamics and fitness effects followed in single cells, Science **359**, 1283 (2018).

[25] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise, Nature (London) **441**, 840 (2006).

[26] J. Jędrak, M. Kwiatkowski, and A. Ochab-Marcinek, 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).

[27] P. Thomas, Making sense of snapshot data: Ergodic principle for clonal cell populations, J. R. Soc., Interface **14**, 20170467 (2017).

[28] C. A. Yates, M. J. Ford, and R. L. Mort, A multi-stage representation of cell proliferation as a markov process, Bull. Math. Biol. **79**, 2905 (2017).

[29] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, Probing gene expression in live cells, one protein molecule at a time, Science **311**, 1600 (2006).

[30] P. Wang, S. Hayden, and Y. Masui, Transition of the blastomere cell cycle from cell size-independent to size-dependent control at the midblastula stage in xenopus laevis, J. Exp. Zool. **287**, 128 (2000).

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

[32] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells, Science **329**, 533 (2010).

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

[34] C. Miller, B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken *et al.*, Dynamic transcriptome analysis measures rates of mrna synthesis and decay in yeast, Mol. Syst. Biol. **7**, 458 (2011).

[35] M. Hashimoto, T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto, Noise-driven growth rate gain in clonal cellular populations, Proc. Natl. Acad. Sci. USA **113**, 3251 (2016).

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

[37] D. Siegal-Gaskins and S. Crosson, Tightly regulated and heritable division control in single bacterial cells, Biophys. J. **95**, 2063 (2008).

[38] B. Cerulus, A. M. New, K. Pougach, and K. J. Verstrepen, Noise and epigenetic inheritance of single-cell division times influence population fitness, Curr. Biol. **26**, 1138 (2016).

[39] See https://bitbucket.org/CasperBeentjes/protein-cell-cycle-ssa.