# Effects of cell cycle variability on lineage and population measurements of mRNA abundance

Ruben Perez-Carrasco[1], Casper Beentjes[2], and Ramon Grima[3]

[1]Department of Mathematics, University College London, United Kingdom
[2]Mathematical Institute, University of Oxford, Oxford, United Kingdom
[3]School of Biological Sciences, University of Edinburgh, United Kingdom

March 24, 2020

## Abstract

Many models of gene expression do not explicitly incorporate a cell cycle description. Here we derive a theory describing how mRNA fluctuations for constitutive and bursty gene expression are influenced by the stochasticity in the duration of the cell cycle and DNA replication. By means of the analytical expressions for the moments, we show that the error introduced in the predicted mean number of mRNAs, when noise in the cell cycle duration is omitted, is a monotonic decreasing function of $\eta$ which is proportional to the ratio of the mean cell cycle duration and the mRNA lifetime; contrastingly, the error in the variance of the mRNA distribution peaks for intermediate values of $\eta$ consistent with genome-wide measurements in many organisms. Using eukaryotic cell data, we estimate the errors in the mean and variance to be at most 3% and 25%. Furthermore we derive an accurate negative binomial mixture approximation to the mRNA distribution and show that under certain conditions, bimodality can result from the doubling of the transcription rate when DNA duplicates. Finally, we show that for real genomic data, disregarding cell cycle stochasticity can introduce errors in the inference of transcription rates larger than 10%, supporting the relevance of the analysis presented.

## Introduction

Intrinsic noise in gene expression induces variability in the transcript number across a population of cells. Current microscopy techniques are able to capture this variability, which can be used to infer the kinetic parameters of transcription, thereby letting us quantify mechanisms in charge of the regulation of gene expression [1–3]. In order to make this inference possible, it is necessary to have an accurate stochastic dynamical model that is able to relate the details of the mRNA number distribution to the different transcriptional and post-transcriptional molecular mechanisms involved in mRNA processing. This has been extensively done by describing the dynamics of the system by means of the Master Equation, a Markovian description whose solution gives the probability of observing a certain number of mRNAs in a cell at a certain time [4]. Since the exact analytical solution of the Master Equation is only available for a few scenarios (e.g. [5–7]), the study of the probability distribution of mRNA transcript number is usually limited to calculating the moments of the distribution.

One particular mechanism that has been difficult to study analytically is the influence of the cell cycle on the distribution of mRNAs in a population of cells. The duration of the different phases of the cell cycle is stochastic, introducing noise not only in the time of mitosis when the molecular content is diluted, but also in the time at which DNA is replicated, which in turn increases the mRNA production rate [3]. In addition, during mitosis, the cellular content is divided, leading to a stochastic transcript bipartition [8].

Due to these different challenges, mathematical effort has been focused on limit cases, such as when the cell cycle duration is considered constant [6, 7, 9], or when DNA replication is omitted [10, 11]. Other studies have considered the effect of the cell cycle on protein fluctuations [5, 12–14]; the analysis in this case is simplified because unlike mRNA, protein lifetimes are very long and hence degradation is mostly due to dilution at cell division.

In addition, there are other factors beyond details of the cell cycle progression that can have a profound influence on transcript fluctuations. The symmetry of cellular division affects the number of transcripts in a cellular population. For instance, in a growing proliferating tissue, the continuous exponential appearance of young cells in a population introduces an asymmetry in the population cell age, favouring the proportion of cells at early stages of their cell cycle. This contrasts with the age structure of a homeostatic population where it is expected to find the cells equally distributed along their cell cycle [15, 16]. Since the average number of mRNAs in a cell increases with the time position in the cell cycle, we expect to observe larger mRNA content for the same type of cell in a homeostatic population compared to a growing population. Similar discrepancies arise when mRNA distributions measured from snapshots of a growing cell population are compared with the temporal tracking of the expression levels of a single cell over time, apparently contradicting ergodicity between single cells and the population. While this effect has been formalised mathematically [11], its relevance to the distributions of mRNA, or to the inference of different kinetic parameters, remains a conundrum.

In this paper we study the distribution of mRNA transcripts in single cells where expression can be bursty or non-bursty (both commonly observed, see for example [2]), with a cell cycle progression described as a number of stages having a stochastic duration. Our model also includes DNA replication, and differentiates between population and lineage (single cell trajectory) measurements of the mRNA distribution. Keeping the framework relevant to the experimental inference of kinetic parameters, we aim to answer the following question: how important is the inclusion of cell cycle variability for predicting the statistics of stochastic mRNA expression? With this objective in mind, we derive and analyze expressions for the error made in different observables of transcript abundance when a deterministic cell cycle (one of fixed length) is considered instead of a stochastic one. Furthermore, we apply our results to a genome-wide expression dataset to address the magnitude of the error made in the inference of the transcription rate when mathematical models with different cell cycle details are employed.

## Model Description

We consider a general model of stochastic gene expression that takes into account cell cycle variability (for an illustration see Fig. 1a and b) with the following properties:

1. The cell cycle is divided into $N$ stages. The duration of each stage $i$ is exponentially distributed with a rate $k_i$. This implies that the total cell cycle duration follows a hypoexponential distribution. Note that the number of stages in general will not coincide with the cell cycle phases. The number and duration of the different stages can be chosen by fitting the experimental cell cycle duration distribution.

2. The length of the mitotic phase is negligible and hence it is assumed to occur instantaneously after the end of the $N$-th stage. This leads to binomial partitioning of the mRNA between mother and daughter cells.

3. There is bursty or constitutive transcription of mRNA with rate $r_i$ of producing mRNAs per unit of time, and a decay rate $d_i$. When the transcription is bursty, the burst size follows a geometric distribution with mean $\beta_i$. All the parameters $r_i, d_i, \beta_i$ can vary depending on the stage $i$ along the cell cycle.

This constitutes the general model studied in this manuscript. Detailing cell stage specific rates of transcription and degradation is particularly relevant since it will bestow our model with the ability of accurately describing the dynamic nature of mRNA transcription [3, 17]. In addition, for the sake of clarity of our analysis we will also consider a particular case of the general model:

1. All the cell stage rates are identical along the cell cycle $k_i = k$. This implies that the total cell cycle duration follows an Erlang distribution. The number of cell cycle stages in this case can be easily determined from a best fit of an Erlang distribution to the experimental cell cycle duration [18]. In particular, the coefficient of variation (CV) of the Erlang distribution is equal to $\sqrt{1/N}$.

2. The degradation rate of the mRNA is independent of the cell cycle stage $d_i = d$.

3. There are $W$ stages prior to DNA replication and $N - W$ stages postreplication. The production rate of mRNA is proportional to the DNA content of the cell at each stage without dosage compensation, being $r_i = r$ for $i \leq W$ and $r_i = 2r$ for $i > W$. If transcription is considered to be bursty, the average burst size is constant along the cycle $\beta_i = \beta$.

Since in this particular scenario the cell cycle duration follows an Erlang distribution, it will be referred hereon as the Erlang model to distinguish it from the general model.

Stochastic simulations of the model can be used to study the effect of changing parameter values on the mRNA transcript number (Fig. 1c). Alternatively, we can study analytically the evolution of the probability $P_i(n,t)$ of finding a cell in stage $i$ with $n$ mRNAs at time $t$ by using a Master Equation description, that for

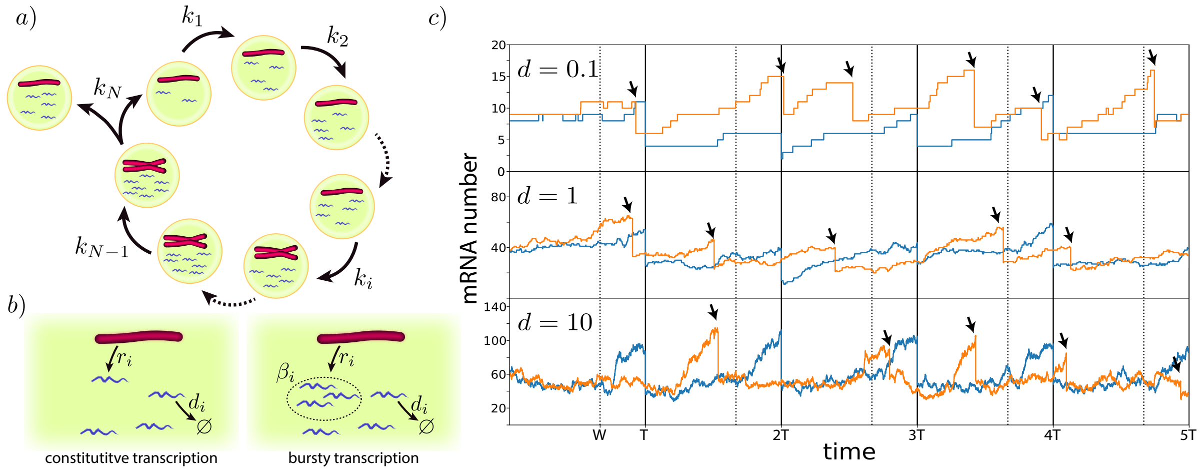

Figure 1: a,b) Schematic of the general model where mRNA dynamics take into account details of the cell cycle (a) including DNA replication, phase duration variability and bipartition at mitosis. During each cell cycle stage (b) mRNA dynamics is described as a production term (constitutive or bursty), and a linear degradation. c) Comparison of stochastic mRNA trajectories between a case where cell cycle duration is constant (blue) or stochastic (orange), for different degradation rates $d$. Arrows indicate stochastic division times. Stochastic cell cycle simulations use the Erlang model with a production rate per chromosome equal to is $r = 50d$ for a cell cycle with $N = 4$ stages, from which $W = 3$ stages occur prior to DNA replication ($w = W/N = 3/4$) indicated by dashed lines for the deterministic simulations.

the general model with constitutive mRNA transcription (bursty case is detailed in the Appendix A) reads,

$$\begin{aligned}\frac{\partial P_1(n,t)}{\partial t} =& -k_1 P_1(n,t) + k_N P_N'(n,t) + \\ & r_1\left(P_1(n-1,t) - P_1(n,t)\right) + \\ & d_1\left((n+1)P_1(n+1,t) - nP_1(n,t)\right), \qquad (1)\\ \frac{\partial P_i(n,t)}{\partial t} =& -k_i P_i(n,t) + k_{i-1}P_{i-1}(n,t) + \\ & r_i\left(P_i(n-1,t) - P_i(n,t)\right) + \\ & d_i\left((n+1)P_i(n+1,t) - nP_i(n,t)\right),\ i \in [2,N]. \\ & \qquad (2)\end{aligned}$$

The first and second terms in these equations describe the exit from, and entrance to, the present cell cycle stage. The third term models transcription and the fourth term mRNA decay. Note that binomial partitioning during mitosis is explicitly taken into account by the second term of Eq. (1). This process implies:

$$P_N'(n,t) = \sum_{m=0}^{\infty} \binom{m}{n} 2^{-m} P_N(m,t), \qquad (3)$$

where we take the convention $m$ choose $n$ equals zero when $n > m$.

## Factorial Moments in Cyclo-stationary Conditions

Defining the generating function $G_i = \sum_n z^n P_i(n)$, the Master Equations Eqs. (1)-(2) can be written as

$$\begin{aligned}\frac{\partial G_1(z,t)}{\partial t} =& -k_1 G_1(z,t) + k_N G_N\left(\frac{1+z}{2}, t\right) + \\ & r_1(z-1)G_1(z,t) + \\ & d_1(1-z)\frac{d}{dz}G_1(z,t), \qquad (4)\\ \frac{\partial G_i(z,t)}{\partial t} =& -k_i G_i(z,t) + k_{i-1}G_{i-1}(z,t) + \\ & r_i(z-1)G_i(z,t) + \\ & d_i(1-z)\frac{d}{dz}G_i(z,t),\ i \in [2,N]. \qquad (5)\end{aligned}$$

From the definition of the generating function it follows that the unnormalised $\ell$-th factorial moment of the mRNA distribution in stage $j$ is given by:

$$(n_j)_\ell \equiv \sum_n n(n-1)...(n-\ell+1)P_j(n) = G_j^{(\ell)}(1), \quad (6)$$

where the superscript $(\ell)$ means differentiating $\ell$ times. Enforcing cyclo-stationary conditions (steady-state for the mRNA distribution of each individual cell stage) by setting the time derivatives in Eqs. (4)-(5) to zero, differentiating $p$ times the resulting equations and using the definition of the factorial moments above, we obtain:

$$\begin{aligned} 0 &= -k_1(n_1)_p + k_N\left(\frac{1}{2}\right)^p (n_N)_p \\ &\quad + r_1 p(n_1)_{p-1} - d_1 p(n_1)_p, \qquad (7) \\ 0 &= -k_i(n_i)_p + k_{i-1}(n_{i-1})_p \\ &\quad + r_i p(n_i)_{p-1} - d_i p(n_i)_p, \quad i \in [2, N]. \qquad (8) \end{aligned}$$

Eq. (8) can be brought into the form:

$$(n_{i+1})_p = f_i(n_i)_p + g_i,\ i \in [2, N], \tag{9}$$

where we have used the definitions:

$$f_i = \frac{k_i}{k_{i+1} + pd_{i+1}},\ g_i = \frac{r_{i+1}p(n_{i+1})_{p-1}}{k_{i+1} + pd_{i+1}}. \tag{10}$$

Since these are first-order non-homogeneous recurrence relations with variable coefficients, their solution can be written as:

$$(n_j)_p = \delta_j(n_1)_p + \theta_j,\ j \in [2, N], \tag{11}$$

where we have used the definitions:

$$\theta_j = \delta_j \sum_{m=1}^{j-1} \frac{g_m}{\delta_{m+1}},\ \delta_j = \prod_{k=1}^{j-1} f_k. \tag{12}$$

Solving Eq. (11) for $(n_N)_p$ and substituting in Eq. (7), after some simplification we obtain:

$$(n_1)_p = \frac{2r_1 p(n_1)_{p-1} + k_N\left(\frac{1}{2}\right)^{p-1}\theta_N}{2(d_1 p + k_1) - k_N\left(\frac{1}{2}\right)^{p-1}\delta_N}. \tag{13}$$

Note that the solution of the unnormalised $p$-th factorial moment depends on knowledge of the unnormalised $p-1$-th factorial moment. Hence, because of this dependency, all factorial moments need knowledge of the zeroth order factorial moment $(n_j)_0$, which corresponds with the probability of finding the cell at stage $j$. By the definition of Eq. (6) we see that $(n_j)_0 = G_j(1)$. Setting $p = 0$ in Eqs (7) and (8) one obtains:

$$(n_i)_0 = \left(k_i \sum_{j=1}^{N} k_j^{-1}\right)^{-1}. \tag{14}$$

Hence summarising, Eqs. (11), (13), and (14) together give the solution to the unnormalised $p$-th factorial moment of the mRNA numbers in cell stage $j$. Note that to obtain the normalised $p$-th factorial moment one divides the unnormalised $p$-th factorial moment by $G_j(1) = \sum_n P_j(n) = (n_j)_0$.

The factorial moments for the general model with bursty transcription can be derived following the same steps. This procedure shows that the first factorial moment is equal to the constitutive case, whereas the factorial moments for higher orders in the bursty case are larger than in the constitutive case (see Appendix A).

## Lineage Measurements

The moments of the distribution can be used to compute the mRNA distribution statistics for different tissues. For instance, the mean number of mRNAs can be calculated as the average along the cell cycle stages of the expected number of mRNAs at each stage ($(n_i)_1/(n_i)_0$) weighted by the probability $\pi_i$ of finding a cell in a tissue at a certain stage $i$. Following this methodology, the expressions for the mean and the variance are,

$$\langle n \rangle = \sum_{i=1}^{N} \pi_i \frac{(n_i)_1}{(n_i)_0}, \qquad \sigma^2 = \sum_{i=1}^{N} \pi_i \frac{(n_i)_2}{(n_i)_0} + \langle n \rangle(1 - \langle n \rangle). \tag{15}$$

We will start our analysis studying the scenario in which the mRNA content of a single cell is tracked in time at regular intervals and, after division, the tracking keeps following only one of the daughter cells. This scenario is equivalent to the mRNA distribution of the cells forming a homeostatic tissue, where after each division one of the cells leaves the population, keeping constant the number of cells in the tissue [15, 16]. This scenario will be referenced in the text as the "lineage" case, to differentiate it from the mRNA distribution across a growing proliferating population of cells, that will be referred to as the "population" case. In the lineage case, the probability $\pi_i$ of finding a cell at the $i$-th cell cycle stage corresponds with $(n_i)_0$ (Eq. 14) being inversely proportional to the cell stage advance rate $k_i$,

$$\pi_i = (n_i)_0. \tag{16}$$

For the Erlang model this is $\pi_i = 1/N$, and the explicit expression for the mean transcript can be obtained by introducing Eqs. (13), (14) and (16) in (15), obtaining,

$$\langle n \rangle = \hat{n} + (1 - w)\hat{n} - \frac{\hat{n}}{\eta}\left(1 - \frac{\left(\frac{1}{1+\eta\Delta}\right)^{(1-w)/\Delta}}{2 - \left(\frac{1}{1+\eta\Delta}\right)^{1/\Delta}}\right), \tag{17}$$

where for the sake of clarity we have written the expression in terms of the coefficient of variation of the cell cycle $\sqrt{\Delta} = \sqrt{1/N}$. In addition, we have introduced the mean mRNA number in the absence of cell cycle $\hat{n} = r/d$, the fraction of the cell cycle before DNA replication $w = W/N$, and the nondimensional parameter $\eta = dT$ that compares the degradation timescale with the dilution timescale given by the average cycle duration $T = N/k$ (see Table 1). Note that $\eta$ is proportional to the ratio between the mRNA half-life $t_{1/2}$ and the cell cycle duration $T$ following $\eta = T\ln(2)/t_{1/2}$.

| | Meaning | Erlang model |
|---|---|---|
| $N$ | Number of cell stages | |
| $W$ | Cell stages prior to replication | |
| $k_i$ | Rate of advance of cell cycle stage $i$ | $k$ |
| $r_i$ | Transcription rate during stage $i$ | $r$ if $i \leq W$<br>$2r$ if i>W |
| $d_i$ | mRNA degradation during stage $i$ | $d$ |
| $\beta_i$ | mean burst size during stage $i$ | $\beta$ |
| $w$ | Proportion of cell cycle before DNA replication | $W/N$ |
| $\hat{n}$ | Stationary average mRNA number in absence of cell cycle | $r/d$ |
| $T$ | Average cell cycle duration | $N/k$ |
| $\Delta$ | Squared coefficient of variation of cell cycle duration | $1/N$ |
| $\eta$ | mRNA degradation relative to cell division rate | $dT$ |

Table 1: Description of the different parameters used to describe the cell cycle, mRNA dynamics, and their relationship in the Erlang model. Parameters in shadowed rows can be derived from the rest of the parameters.

The first term of Eq. (17) corresponds to the classical scenario without cell cycle. The second term of Eq. (17) introduces the effect of DNA replication for the case in which the mRNA degradation timescale is much shorter than the cell cycle length ($\eta \to \infty$). Finally, the third term in Eq. (17) describes the contribution when mRNA degradation occurs at comparable timescale to the cell cycle duration. This latter contribution increases monotonically with the cell cycle variability $\Delta$ (see Appendix B), and is minimal in the limit of $\Delta \to 0$ (deterministic cell cycle duration). In this deterministic limit Eq. (17) reduces to the simpler form

$$\langle n \rangle^* = \lim_{\Delta \to 0} \langle n \rangle = \hat{n} + (1-w)\hat{n} - \frac{\hat{n}}{\eta}\left(1 - \frac{e^{-\eta(1-w)}}{2-e^{-\eta}}\right), \tag{18}$$

which agrees with a different calculation using deterministic rate equations (see Appendix C). Comparison of Eqs. (17) and (18) allows us to quantify the relative error $R$ made in the expected number of mRNA when the cell cycle variability is not considered in the description of the mRNA dynamics,

$$R \equiv \frac{\langle n \rangle - \langle n \rangle^*}{\langle n \rangle} = 1 - \frac{\eta(2-w) - \left(1 - \frac{\mathrm{e}^{-\eta(1-w)}}{2-\mathrm{e}^{-\eta}}\right)}{\eta(2-w) - \left(1 - \frac{(1+\eta\Delta)^{-(1-w)/\Delta}}{2-(1+\eta\Delta)^{-1/\Delta}}\right)}. \tag{19}$$

Note that $R$ is only a function of $\eta$, $w$ and $\Delta$, and therefore independent of the mRNA production rate (see Fig. 2a). The error is always positive (see Appendix B) and increases with the cell cycle time variability $\Delta$, reaching its maximum for $\Delta = 1$ (which is the maximum $\Delta$ attainable for an Erlang process since $N \geq 1$). Similarly, since the expression for the first moment is identical in the bursty and constitutive cases (see Appendix A), the mean transcript number and its error are also independent of how bursty the transcription is.

For a given cell type, the average time at which replication of a given gene occurs and the cell cycle duration variability can be considered constant (provided external conditions are not changed), and hence the value of the error $R$ for different genes will be determined exclusively by $\eta$, which compares the mean cell cycle duration and mRNA lifetime, and can vary significantly from gene to gene [19]. The error decreases with $\eta$ (see Fig 2a,b), vanishing for $\eta \gg 1$ corresponding with the scenario where mRNA lifetime is much shorter than the cell cycle duration. On the other hand, the relative error $R$ is maximum for low values of $\eta$, describing the case of stable mRNAs for which degradation rates are much lower than the proliferation rate of the cell (analytical expressions for the mRNA distribution for this case can be found following the method described in [5]),

$$\lim_{\eta \to 0} R = \frac{\Delta}{\Delta + (2-w) + \frac{2}{2-w}}.$$

Interestingly, this maximal error depends on the properties of the cell cycle through $w$ and $\Delta$ and it is maximised for intermediate levels of the DNA replication time $w = 2 - \sqrt{2} \simeq 0.6$, which is comparable to biological values of the relative duration of the $G_1$ phase which typically varies between $w = 0.25$ and $w = 0.75$ [3, 24] (excluding cells which have arrested $G_1$ phases), achieving a maximal relative error of $R \simeq 15\%$, corresponding to $\Delta = 1/2$ and $w = 2 - \sqrt{2}$.

The relative error can be more precisely estimated given data for specific types of cells. For example the cell cycle duration distribution in NIH 3T3 mouse embryonic fibroblasts has been described by an Erlang distribution with $CV^2 \simeq 1/12$ (which implies $N = 12$ effective cell cycle stages) [18] and the $G_1$ phase occupies roughly a fraction $w = 0.4$ of the cell cycle [24]. The maximum error $R$ for these parameters shows a relative error around 3% (Fig. 2b), while for most of the transcriptome ($\eta \sim 1$, see Fig. 2c) is $R \simeq 1\%$, indicating that in these cases the cell cycle duration variability can be ignored if the mean mRNA is all that we are interested in.

Making use of the second order moments of the distribution, we can extend the analysis to other statistic observables, allowing us to quantify the error in the variance, $R_\sigma$, of mRNA fluctuations made when ne-

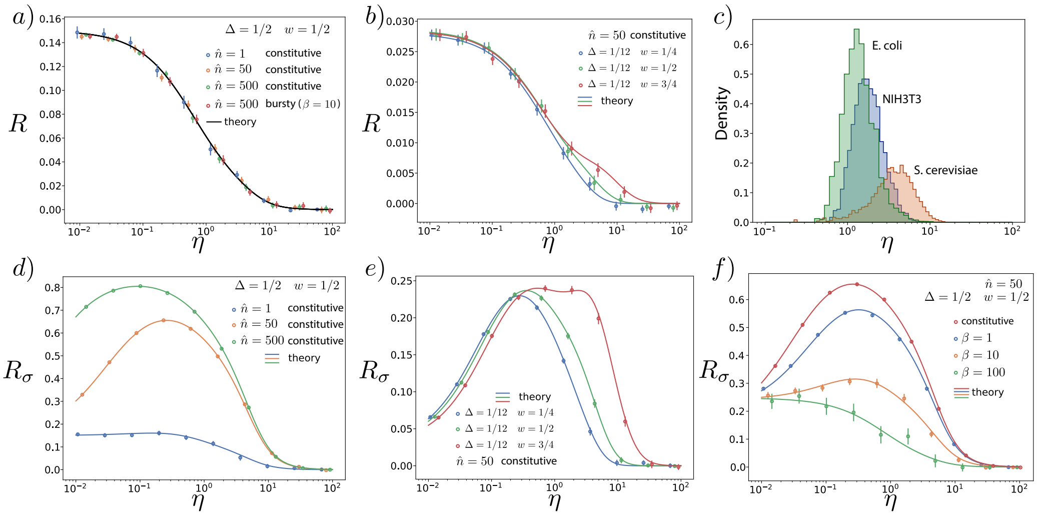

Figure 2: Relative error made in the average number of mRNAs ($R$) and its variance ($R_\sigma$) when considering the cell cycle to be deterministic instead of Erlang distributed in a non-proliferating population or a lineage. Panels compare the theoretical results (lines) with stochastic simulations (circles). a,b) Relative error $R$ of the mean number of mRNA. c) Genome-wide values of $\eta$ for three different cell types. NIH3T3 mouse fibroblast data was obtained from [19]. Degradation rates for *S. cerevisiae* cultured in yeast extract peptone dextrose were obtained from [20] and its cell cycle duration from [21]. Stability data for the transcripts of *E. coli* cultured in Lysogeny broth were obtained from [22], while its cell cycle duration from [23]. Averages are done over trajectories of duration $t = 600T \max(1/dT, 1/\rho T, 1)$. Panels a), b) and f) show averages over 50 trajectories for all conditions except for $\hat{n} = 1$ that shows an average of 500 trajectories. Panels c) and d) show averages over 200 trajectories. Error bars indicate the standard error of the mean.

glecting cell-cycle variability,

$$R_\sigma = \frac{\sigma^2 - \sigma^{2*}}{\sigma^2}, \tag{20}$$

where $\sigma^{2*}$ is the variance of the mRNA distribution in the deterministic cell cycle limit ($\Delta \to 0$). For the Erlang model, combining the expression for the variance (Eq. 15) with the factorial moments (Eqs. 13-14) we obtain an error for the variance $R_\sigma$ that is much larger than the one observed in the mean. Additionally, $R_\sigma$ does not have a monotonic dependence on the degradation rate, but is maximal for intermediate values of the degradation rate ($\eta \sim 1$, see Fig. 2d,e,f). Interestingly, this region of values of $\eta$ corresponds with most of the transcripts genome-wide for different species (see Fig. 2c). In particular, for the NIH 3T3 cells the error reaches $R_\sigma \simeq 25\%$ (Fig. 2e), and can reach values as high as 80% for $\Delta = 1/2$ (Fig.2d). In contrast to the error in the mean, the error in the variance will depend on the transcription rate and the transcriptional burstiness. Analysis of $R_\sigma$ for the bursty model shows that $R_\sigma$ decreases with the burst size, reflecting that increase in the variance due to the bursty gene expression reduces the relative impact of the contribution from cell cycle variability (Fig. 2f). Nevertheless, despite this reduction, the error $R_\sigma$ is still above 10% for many scenarios including both bursty and constitutive expression (Fig. 2d,e,f). Furthermore, in contrast to the error in the mean, $R_\sigma$ depends on the DNA replication position $w$ in such a way that genes replicating later in the cell cycle (larger $w$) not only show larger errors, but also for a broader range of degradation rates (see Fig. 2 e).

## Population Measurements

When considering a proliferating population of cells, the continuous appearance of synchronised cells at an initial cell cycle stage establishes a different age distribution than the one derived in the lineage scenario (Fig. 3a). Specifically, after mitosis, one cell at stage $N$ leaves the population to give rise to two cells at stage 1, enhancing the probability of finding cells in the population at initial stages of their cell cycle. The population values for the probability of observing a cell in the $i$-th cell stage $\pi_i$, can be obtained by considering the evolution of the average number of cells in cell cycle stage $i$ at time $t$, denoted by $C_i(t)$.

$$\frac{\mathrm{d}C_1(t)}{\mathrm{d}t} = -k_1 C_1(t) + 2k_N C_N(t), \tag{21}$$

$$\frac{\mathrm{d}C_i(t)}{\mathrm{d}t} = -k_i C_i(t) + k_{i-1} C_{i-1}(t) \quad i = 2, \ldots, N, \tag{22}$$

where the factor 2 in the first equation stands for cellular division: every time a cell divides (leaving stage $N$), two cells start at stage 1. In the lineage case this factor becomes 1. More generally, for cases with asym-

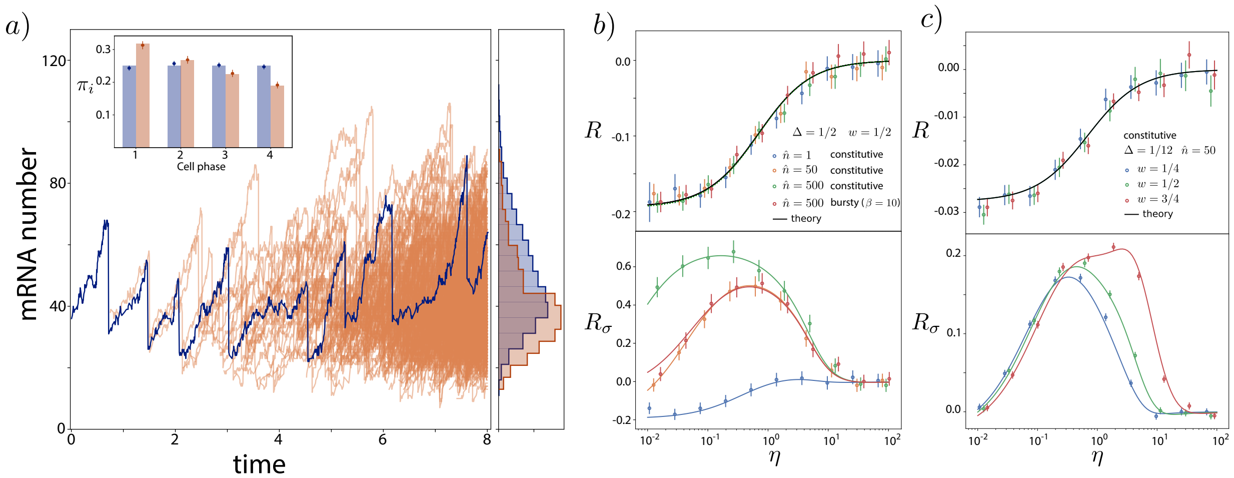

Figure 3: a) Comparison between the mRNA content over a single cell trajectory in time (blue) with the mRNA distribution of a proliferating population (orange). Histograms for both mRNA distributions (right) compare the average of 100 trajectory realizations with the snapshot of a single population at $t = 7T$. Parameters used are $T = 1$ , $N = 4$, $W = 2$ , $\hat{n} = 50$, $\eta = 1$. Inset) Probability distribution $\pi_i$ of finding a cell at different cell cycle stages for single trajectory (blue) and a proliferating population (orange). Stochastic simulations for $\pi_i$ (circles) are compared with theoretical results (bars) obtained from Eq. (16) (that for the Erlang case is constant $\pi_i = 1/N$) and from Eq. (24). b,c) Relative error made in the average number of mRNAs ($R$) and its variance ($R_\sigma$) when considering the cell cycle to be deterministic instead of Erlang distributed in a growing proliferative population. Comparison includes theoretical results (lines) and stochastic simulations (circles). Simulations in b) show the average of 5000 snapshots at a time $10T$ and in c) the average of 25000 snapshots at a time $10T$. Error bars indicate the standard error of the mean.

metric division (after mitosis some cells leave the population with a certain probability) this factor 2 can be replaced by a factor $\alpha \in [0, 2]$ [16]. While for eqs. (21,22) the number of cells $C_i(t)$ will grow in time, the relative cell stage distribution in the population will eventually reach a steady-state for which we can write the ansatz $C_i(t)/C_1(t) \equiv \lambda_i$. Specifically, for the Erlang case, introducing the definition of $\lambda_i$ in Eq. (22) yields the relationship

$$\lambda_i = 2^{(1-i)/N}, \tag{23}$$

that gives the explicit values for the probability $\pi_i$ of observing a cell in the population at stage $i$,

$$\pi_i(t) = \frac{C_i(t)}{\sum_{i=1}^{N} C_i(t)} = \frac{2^{1/N} - 1}{2^{(i/N)-1}}, \quad i = 1, \ldots, N, \tag{24}$$

differing from the lineage stage distribution (Eq. (16), which for the Erlang case is constant ($\pi_i = 1/N$). This discrepancy was confirmed by simulations (see inset of Fig 3a).

In addition to differences in $\pi_i$, cells in the population case are also found more likely at earlier times inside each stage than in the lineage case. For the Erlang model, the distribution of times that each cell has been in its current cell stage follows an exponential distribution $\sim \text{Exp}(k2^{1/N})$ (see Appendix D and [5]). Given the Markovian nature of the process, this effect is equivalent to reducing $k$ and having a faster cell advance through the cell cycle. Therefore, using the expressions for $\pi_i$ from Eq. (24), and the new effective rates of cell stage advance $k \to k2^{1/N}$ in Eqs. (11) and (13-14), allows us to obtain the factorial moments for population measurements. The mean number of mRNA in this scenario is

$$\langle n \rangle = \hat{n} \frac{2^{1-w} \eta \Delta}{2^{\Delta} + \eta \Delta - 1}. \tag{25}$$

It is straightforward to show that $\langle n \rangle$ increases monotonically with $\Delta$ (similar to the lineage case). The exactness of Eq. (25) is confirmed by stochastic simulations in Figure 3. In the limit of a deterministic cell cycle, Eq. (25) reduces to the simpler form:

$$\langle n \rangle^* = \lim_{\Delta \to 0} \langle n \rangle = \frac{2^{1-w} \eta \hat{n}}{\eta + \ln(2)}. \tag{26}$$

This agrees with a different calculation using deterministic rate equations (see Appendix C). Similar to the lineage case, this allows us to write explicitly an expression for the relative error $R$ in the average number of mRNAs made when omitting the stochasticity of the cell cycle,

$$R \equiv \frac{\langle n \rangle - \langle n \rangle^*}{\langle n \rangle} = 1 - \frac{2^{\Delta} + \eta \Delta - 1}{\Delta(\eta + \ln(2))}. \tag{27}$$

As in the lineage case, the error is a monotonic decreasing function of $\eta$, and increases with $\Delta$ reaching an error that is similar to the single cell case ($R \simeq 20\%$) (see Fig. 3b,c). Nevertheless, in contrast to the lineage case, the error is negative, indicating that the expected

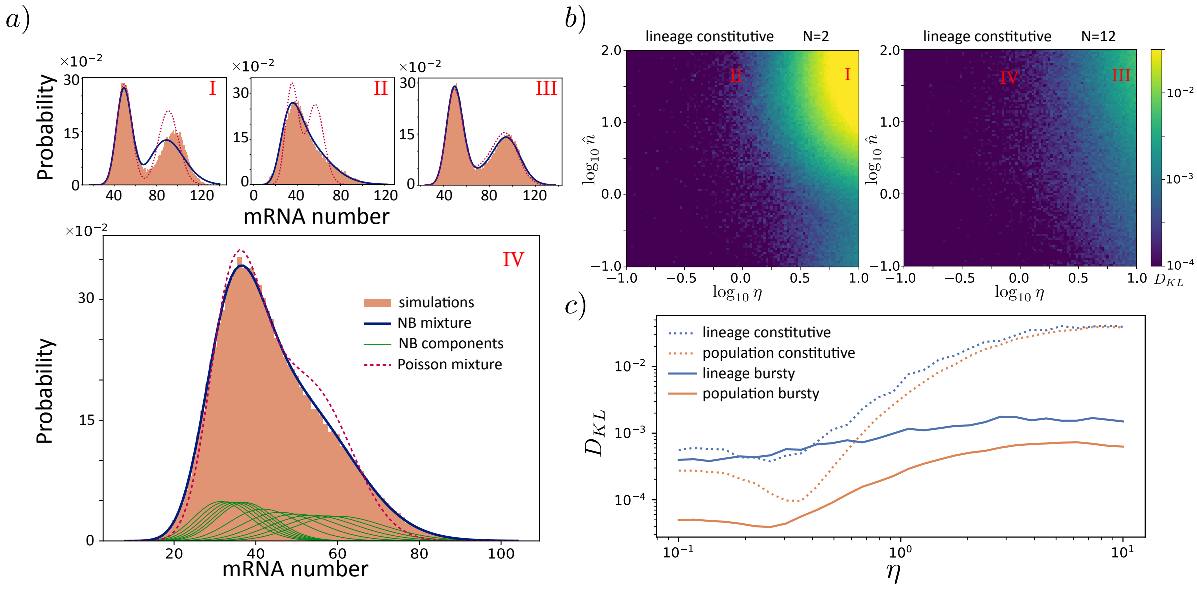

Figure 4: Approximation of the mRNA distribution as a mixture of negative binomials. a) Comparison between the approximations for the mRNA distribution (Poisson mixture and Negative Binomial mixture) with simulations for four different cases I ($N = 2$, $\eta = 10$, $\hat{n} = 50$ ), II ($N = 2$, $\eta = 1$, $\hat{n} = 50$ ), III ($N = 12$, $\eta = 10$, $\hat{n} = 50$ ), IV ($N = 12$, $\eta = 1$, $\hat{n} = 50$ ). Case IV also includes the individual components of the NB mixture. b) Kullback-Leibler divergence of the NB binomial mixture from stochastic simulations for lineage distributions of constitutive mRNA expression. c) Comparison of the Kullback-Leibler divergence from simulations of the NB mixture for combinations of lineage/population measures and constitutive/bursty ($\beta = 10$) expression. Distributions for a) and b) result from trajectories over a time $t = 6 \cdot 10^4 T \max(1/dT, 1/\rho T, 1)$, while for panel c) we used $t = 6 \cdot 10^6 T/d \cdot \max(1/dT, 1/\rho T, 1)$ for lineage measurements, and $t = 20T$ for population measurements.

number of mRNA decreases with the variability of the cell cycle duration. Strikingly, the error is independent of $w$, hence independent of the relative duration of G1 and G2 phases. Analysis of the error in the variance, $R_\sigma$, results in similar observations to those of the lineage measurements, where $R_\sigma$ depends on the transcription rate and the transcription burstiness, resulting in errors much larger than $R$ ($R_\sigma > 50\%$) that peaks at intermediate values of the degradation rate corresponding to the most frequent values of $\eta$ measured genome-wide for different species (see Fig. 2c). As in the lineage case, the error $R_\sigma$ depends on the replication position during the cell cycle $w$, so genes replicating later in the cell cycle show larger errors for broader ranges of mRNA stability.

## mRNA Distribution Approximation

The exact mRNA distribution of our model is known only for some limit cases such as $\eta \to 0$ [5]. Nevertheless, for more general realistic cases, we can use the moment derivation to reconstruct an approximate distribution. In particular, our analysis provides analytical expressions for the moments of the distribution at each cell stage $i$. Exclusively using the first moments, we can approximate the total mRNA population as a mixture of $N$ Poisson distributions $\tilde{P}(N,t) = \sum_i \pi_i \text{Pois}(\langle n \rangle_i)$, where the weights $\pi_i$ correspond to the probability of finding a cell at cell stage $i$ obtained in Eqs. (16) and (24). Similarly, including the second moments, we can describe the probability as a mixture of negative binomial distributions $\tilde{P}(n,t) = \sum_i \pi_i \text{NB}(\langle n \rangle_i, \sigma_i)$, where each component $\text{NB}(\langle n \rangle_i, \sigma_i^2)$ is a negative binomial distribution with mean $\langle n \rangle_i$ and variance $\sigma_i^2$ (see Eq. 15 ). Results for the lineage case, show that while the Poisson mixture failed to recover the distribution obtained from stochastic simulations in most scenarios (see Fig. 4a), the negative binomial mixture resulted in a very good prediction, able to recover the broad tails and bimodality of the mRNA distribution. In order to accurately assess the goodness of the reconstructed distribution, we computed the Kullback-Leibler divergence of the negative binomial mixture $\tilde{P}(n,t)$ from the simulated exact distribution (see Fig. 4b). We observed that the approximation only fails for regimes with very unstable mRNAs that are highly expressed. On the other hand, the approximation improves for larger values of $N$, closer to experimental values for the cell cycle duration variability ($CV^2 = 1/N = 1/12$) [18] (compare left and right panels of Fig. 4b). Comparison of the distributions for bursty expression and population measurements, using their corresponding moments and stages distributions, $\pi_i$, yielded an even better approximation with values of the Kullback-Leibler divergence orders of magnitude lower than the lineage case (see Fig. 4c).

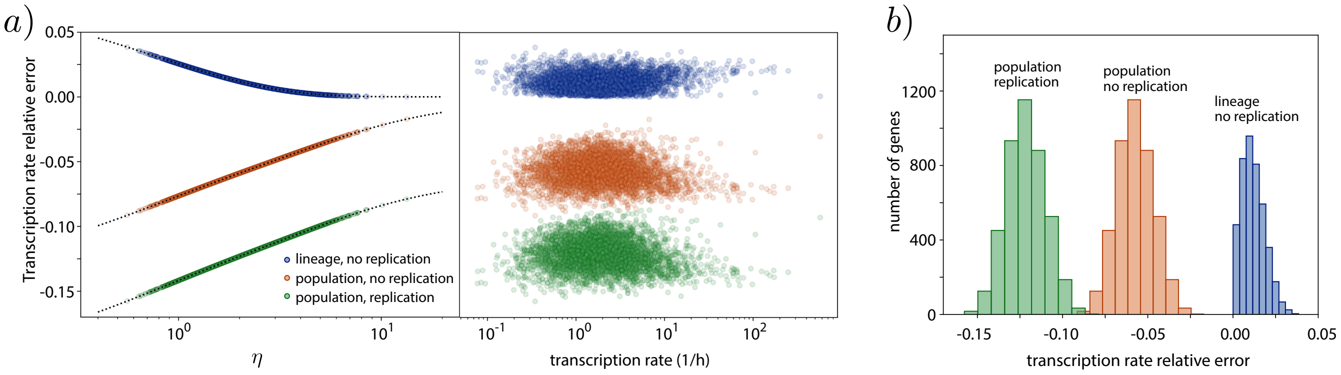

Figure 5: Relative error of inferred mean transcription rates with deterministic models at a genomic scale. a) Relative error in the inferred transcription rate for the 5028 genes reported in [19], as a function of the relative degradation rate $\eta$ and the reported transcription rate. The relative errors are calculated between different models with stochastic cell cycle and the deterministic cell cycle model without replication used in [19] (see Appendix F). Discrepancies between models are a function of $\eta$ (dotted line). b) Histogram showing the amount of genes for different levels of transcription rate error for three different stochastic cell cycle models. The stochastic cell cycle used is an Erlang model with $k = 0.145h^{-1}$ and $N = 4$ using the reported cell cycle duration and variability in [19] (see Appendix F), while replication is considered to occur at the middle of the cell cycle ($w = 1/2$).

## Genome-wide Transcription Rate Inference Error

Our results so far have been focused on analyzing the errors that different models introduce on the mRNA statistics. Likewise, it is relevant to assess the error that different models introduce in the inference of biochemical parameters from experimental data. For this purpose we analyzed the genome-wide data from [19] and compared their transcription rate inference based on a lineage model with constant cell cycle and no replication (obtained by solving an expression equivalent to $\langle n^* \rangle$ with $w = 0$ in Eq. (18), see Appendix F) against different models incorporating stochastic cell cycle duration (see Fig. 5 and Appendix F). Interestingly, since the average mRNA number is proportional to the transcription rate (see eqs. 17 and 25), for given values of the cell cycle duration and gene replication, the relative error made when omitting cell cycle variation is a function depending only on the degradation rate through the parameter $\eta$ (Fig. 5a). Since [19] reported no correlation between mRNA stability and transcription rate, this resulted in an absence of correlation between the error and the speed at which genes are transcribed (Fig. 5a). Additionally, in agreement with the error of the average mRNA number $R$, the error in the transcriptional rate estimate increases with the stability of the mRNA. When comparing the error expected for different models, small errors were observed for the lineage case with no DNA replication (See Fig. 5b). Nevertheless, for more realistic scenarios where the error is evaluated for a growing population case with DNA replication [19], more than 90% of the genes detected understimate the transcription rate with an error bigger than 10% (see Fig. 5b).

## Discussion

Most of the models employed to study gene regulation ignore the effect that a detailed stochastic cell cycle description has on gene expression. The model and methodology developed in this paper not only allows one to analytically evaluate the role of features such as cell duration stochasticity or DNA replication in the transcript population, but also provides a straightforward way of discriminating the scenarios for which such details are relevant for the description of the system. This is of paramount importance when mathematical models are used to infer parameters from experimental data, where the precision of the information demands the use of the right level of abstraction [25].

Specifically, this approach contrasts with alternative strategies that either ignore cell cycle effects or fit mRNA populations to arbitrary population mixtures, impeding the inference of mechanistic information of the transcriptional parameters. This is of particular relevance for current data analysis where mRNA labelling techniques give access to mRNA abundance distributions in populations of cells. In order to extract mechanistic information of the transcriptional process from these distributions, it is paramount to link the details of the distribution to the properties of the different biomolecular mechanisms [3, 26]. While in this paper we analysed the error in the transcription rate estimation due to neglecting cell cycle variability and replication, future work will address how taking into account such details may also affect the inference of other biochemical parameters such as gene activation and deactivation rates, or the mean burst size. The necessity of such study becomes apparent from the mRNA distributions obtained, which can be approximated accurately by negative binomials in scenarios with constitutive gene expression, challenging the common practice to use negative binomial distributions as a signature of bursty transcription [1,26,27].

In addition, incorporating the methodology developed in this paper to gene regulatory networks will provide a route to better understanding the stochastic details of gene expression in growing tissues. This is of special relevance in embryo development, where the details of intrinsic noise are known to play a major role in the formation of spatial domains of gene expression in the patterning of embryonic tissues [28–30].

## Acknowledgments

R.P.-C. acknowledges support from the UCL Mathematics Clifford Fellowship. R.G. acknowledges support from the Leverhulme Trust (Grant No.RPG-2018-423). C.B. acknowledges the Clarendon Fund and New College, Oxford, for funding.

## References

[1] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA synthesis in mammalian cells.," PLoS Biol., vol. 4, p. e309, oct 2006.

[2] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast," Nat. Struct. Mol. Biol., vol. 15, no. 12, pp. 1263–1271, 2008.

[3] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, "Single-cell analysis of transcription kinetics across the cell cycle," Elife, vol. 5, pp. 1–24, jan 2016.

[4] N. Van Kampen, Stochastic Processes in Physics and Chemistry. North-Holland Personal Library, Elsevier Science, 2011.

[5] C. H. L. Beentjes, R. Perez-Carrasco, and R. Grima, "Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics," Phys. Rev. E, vol. 101, p. 032403, mar 2020.

[6] Z. Cao and R. Grima, "Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells.," Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 9, pp. 4682–4692, 2020.

[7] 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 Math. Phys. Eng. Sci., vol. 471, no. 2180, 2015.

[8] D. Huh and J. Paulsson, "Random partitioning of molecules at cell division," Proc. Natl. Acad. Sci. U. S. A., vol. 108, no. 36, pp. 15004–15009, 2011.

[9] R. Dessalles, V. Fromion, and P. Robert, "Models of protein production along the cell cycle: An investigation of possible sources of noise," PLoS One, vol. 15, no. 1, pp. 1–25, 2020.

[10] A. Schwabe and F. J. Bruggeman, "Contributions of cell growth and biochemical reactions to nongenetic variability of cells," Biophys. J., vol. 107, no. 2, pp. 301–313, 2014.

[11] P. Thomas, "Making sense of snapshot data: Ergodic principle for clonal cell populations," J. R. Soc. Interface, vol. 14, no. 136, 2017.

[12] M. Soltani and A. Singh, "Effects of cell-cycle-dependent expression on random fluctuations in protein levels," R. Soc. Open Sci., vol. 3, no. 12, 2016.

[13] 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., vol. 12, p. e1004972, aug 2016.

[14] 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, vol. 99, p. 042416, apr 2019.

[15] R. S. Nowakowski, S. B. Lewin, and M. W. Miller, "Bromodeoxyuridine immunohistochemical determination of the lengths of the cell cycle and the DNA-synthetic phase for an anatomically defined population," J. Neurocytol., vol. 18, no. 3, pp. 311–318, 1989.

[16] E. Hannezo, J. Prost, and J.-F. Joanny, "Growth, homeostatic regulation and stem cell dynamics in tissues," J. R. Soc. Interface, vol. 11, p. 20130895, apr 2014.

[17] N. Battich, J. Beumer, B. de Barbanson, L. Krenning, C. S. Baron, M. E. Tanenbaum, H. Clevers, and A. van Oudenaarden, "Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies," Science (80-. )., vol. 367, pp. 1151–1156, mar 2020.

[18] C. A. Yates, M. J. Ford, and R. L. Mort, "A Multi-stage Representation of Cell Proliferation as a Markov Process," Bull. Math. Biol., vol. 79, pp. 2905–2928, dec 2017.

[19] 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, vol. 473, pp. 337–342, may 2011.

[20] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown, "Precision and functional specificity in mRNA decay," Proc. Natl. Acad. Sci., vol. 99, pp. 5860–5865, apr 2002.

[21] S. Fred, "Getting started with yeast," Methods Enzymol., vol. Volume 350, pp. 3–41, 2002.

[22] D. W. Selinger, "Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation," Genome Res., vol. 13, pp. 216–223, feb 2003.

[23] S. T. Liang, M. Ehrenberg, P. Dennis, and H. Bremer, "Decay of rpIN and lacZ mRNA in Escherichia coli," J. Mol. Biol., vol. 288, no. 4, pp. 521–538, 1999.

[24] A. T. Hahn, J. T. Jones, and T. Meyer, "Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors," Cell Cycle, vol. 8, no. 7, pp. 1044–1052, 2009.

[25] D. J. Warne, R. E. Baker, and M. J. Simpson, "Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art," J. R. Soc. Interface, vol. 16, p. 20180943, feb 2019.

[26] L. Ham, R. D. Brackston, and M. P. H. Stumpf, "Extrinsic Noise and Heavy-Tailed Laws in Gene Expression," Phys. Rev. Lett., vol. 124, p. 108101, mar 2020.

[27] L. H. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding, "General properties of transcriptional time series in Escherichia coli," Nat. Genet., vol. 43, no. 6, pp. 554–560, 2011.

[28] B. Zoller, S. C. Little, and T. Gregor, "Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting," Cell, vol. 175, pp. 835–847.e25, oct 2018.

[29] A. C. Oates, "What's all the noise about developmental stochasticity?," Development, vol. 138, pp. 601–7, feb 2011.

[30] N. C. Lammers, V. Galstyan, A. Reimer, S. A. Medin, C. H. Wiggins, and H. G. Garcia, "Multimodal transcriptional control of pattern formation in embryonic development," Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 2, pp. 836–847, 2020.

[31] 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, vol. 311, pp. 1600–3, mar 2006.

[32] O. G. Berg, "A model for the statistical fluctuations of protein numbers in a microbial population," J. Theor. Biol., vol. 71, pp. 587–603, apr 1978.

# Appendix

## A Bursty mRNA transcription model

Considering the general model, we can introduce bursty mRNA as a reaction with a burst rate $\nu_i$ at cell stage $i$. The number of mRNAs $\ell$ produced in a burst at cell stage $i$ follows a geometric distribution $\xi_i(\ell)$ with average number $\beta_i$ of transcripts produced per burst i.e. an average rate $r_i = \nu_i \beta_i$ of mRNAs produced per unit of time [31]. The explicit geometric probability distribution follows

$$\xi_i(\ell) = \frac{1}{1+\beta_i}\left(\frac{\beta_i}{1+\beta_i}\right)^{\ell}. \tag{28}$$

The resulting Master Equation reads,

$$\begin{aligned}\frac{\partial P_1(n,t)}{\partial t} &= -k_1 P_1(n,t) + k_N P_N'(n,t) + \\ &\frac{r_1}{\beta_1}\left(\sum_{\ell=1}^{n}\left[P_1(n-\ell,t)\xi_1(\ell)\right] - P_1(n)\left(1-\xi_1(0)\right)\right) + \\ &\qquad d_1((n+1)P_1(n+1,t) - nP_1(n,t)), \quad (29)\\ \frac{\partial P_i(n,t)}{\partial t} &= -k_i P_i(n,t) + k_{i-1}P_{i-1}(n,t) \\ &+ \frac{r_i}{\beta_i}\left(\sum_{\ell=1}^{n}\left[P_i(n-\ell,t)\xi_i(\ell)\right] - P_i(n)\left(1-\xi_i(0)\right)\right) + \\ &\qquad d_i((n+1)P_i(n+1,t) - nP_i(n,t)),\ i \in [2,N]. \quad (30)\end{aligned}$$

As in the constitutive case, we can use these system of differential equations to obtain the steady state factorial moments of the distribution by introducing the generating function $G_i = \sum_n z^n P_i(n)$. In particular the terms corresponding to bursty transcription follow the sum,

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{\ell=1}^{n} z^n P(n-\ell,t)\xi_i(\ell) &= \sum_{\ell=1}^{\infty}\sum_{n=0}^{\infty} z^n P(n-\ell,t)\xi_i(\ell) = \\ \sum_{\ell=1}^{\infty}\xi_i(\ell)z^{\ell}G(z) &= G(z)\left(\frac{1}{1+\beta_i(1-z)} - \frac{1}{1+\beta_i}\right).\end{aligned} \tag{31}$$

This results in the system of differential equations,

$$\begin{aligned}\frac{\partial G_1(z,t)}{\partial t} &= -k_1 G_1(z,t) + k_N G_N\left(\frac{1+z}{2},t\right) + \\ &\left(\frac{1}{1+\beta_1(1-z)} - 1\right)\frac{r_1}{\beta_1}G_1(z,t) + \\ &d_1(1-z)\frac{d}{dz}G_1(z,t), \quad (32)\\ \frac{\partial G_i(z,t)}{\partial t} &= -k_i G_i(z,t) + k_{i-1}G_{i-1}(z,t) + \quad (33)\\ &\left(\frac{1}{1+\beta_i(1-z)} - 1\right)\frac{r_i}{\beta_i}G_i(z,t) + \\ &d_i(1-z)\frac{d}{dz}G_i(z,t),\ i \in [2,N]. \quad (34)\end{aligned}$$

Enforcing the steady-state by setting the time derivatives in Eqs. (32) and (33) to zero, differentiating $p$ times the resulting equations and using the definition of the factorial moments $(n_i)_k$ we obtain:

$$\begin{aligned}0 =& -k_1(n_1)_p + k_N\left(\frac{1}{2}\right)^p (n_N)_p + \\ & r_1\sum_{j=0}^{p-1}\frac{p!}{j!}\beta_1^{p-j-1}(n_1)_j - d_1 p (n_1)_p, \quad (35)\\ 0 =& -k_i(n_i)_p + k_{i-1}(n_{i-1})_p + \\ & r_i\sum_{j=0}^{p-1}\frac{p!}{j!}\beta_i^{p-j-1}(n_i)_j - d_i p (n_i)_p,\ i \in [2,N]. \quad (36)\end{aligned}$$

The normalization of the factorial moments $(n_j)_0$, obtained for $p = 0$ with the normalization condition $\sum_j G_j(1) = 1$, is the same as in the constitutive case, and only depends on the cell stage advance rates

$$(n_1)_0 = \left(1 + \sum_{i=2}^{N}\prod_{j=2}^{i}\frac{k_{j-1}}{k_j}\right)^{-1}, \tag{37}$$

$$(n_i)_0 = (n_1)_0 \prod_{j=2}^{i}\frac{k_{j-1}}{k_j},\ i \in [2,N]. \tag{38}$$

Similarly to the constitutive case, equation (36) can be written in the form,

$$(n_i)_p = f_{i-1}(n_{i-1})_p + \tilde{g}_{i-1},\ i \in [2,N], \tag{39}$$

where we have used same definition for $f_i$ as in the constittutive production case, but $\tilde{g}_i$ replaces $g_i$, which instead of depending on the immediately lower order factorial moment $p-1$, depends on all the moments lower than $p$:

$$f_i = \frac{k_i}{k_{i+1}+pd_{i+1}}, \quad \tilde{g}_i = \frac{r_{i+1}\sum_{j=0}^{p-1}\frac{p!}{j!}\beta_{i+1}^{p-j-1}(n_{i+1})_j}{k_{i+1}+pd_{i+1}}. \tag{40}$$

These first-order non-homogeneous recurrence relations have the solution

$$(n_j)_p = \delta_j (n_1)_p + \tilde{\theta}_j,\ j \in [2, N], \tag{41}$$

where we have used the definitions:

$$\tilde{\theta}_j = \delta_j \sum_{m=1}^{j-1} \frac{\tilde{g}_m}{\delta_{m+1}},\ \delta_j = \prod_{k=1}^{j-1} f_k. \tag{42}$$

Substituting this solution in Eq. (35) we obtain

$$(n_1)_p = \frac{2r_1 \sum_{j=0}^{p-1} \frac{p!}{j!} \beta_1^{p-j-1} (n_1)_j + k_N \left(\frac{1}{2}\right)^{p-1} \tilde{\theta}_N}{2(d_1 p + k_1) - k_N \left(\frac{1}{2}\right)^{p-1} \delta_N}. \tag{43}$$

Comparing these results we can immediately see that the expected value of mRNAs is the same in the bursty case and the constitutive case considering the same average rate of mRNA production at cell stage $i$: $r_i = \nu_i \beta_i$. Differences arise for higher moments. In particular all the factorial moments of the bursty scenario with $p > 1$ are larger than the factorial moments of the constitutive case since,

$$r_i \sum_{j=0}^{p-1} \frac{p!}{j!} \beta_i^{p-j-1} (n_i)_j = \tag{44}$$

$$r_i p (n_i)_{p-1} + r_i \sum_{j=0}^{p-2} \frac{p!}{j!} \beta_i^{p-j-1} (n_i)_j > p r_i (n_i)_{p-1}.$$

## B Monotonic dependence of mean mRNA on the coefficient of variation of the cell cycle duration for lineage observations

By Eq. (17), we have for $\eta > 0$

$$\begin{aligned} \langle n \rangle &= w\hat{n} + (1-w)2\hat{n} - \frac{\hat{n}}{\eta} \left( 1 - \frac{\left(\frac{1}{1+\eta\Delta}\right)^{(1-w)/\Delta}}{2 - \left(\frac{1}{1+\eta\Delta}\right)^{1/\Delta}} \right) \\ &= C + Df(\Delta), \end{aligned} \tag{45}$$

where $w$ is a fraction, $C, D$ are constants ($D$ is positive) and

$$f(\Delta) = \frac{\left(\frac{1}{1+\eta\Delta}\right)^{(1-w)/\Delta}}{2 - \left(\frac{1}{1+\eta\Delta}\right)^{1/\Delta}}. \tag{46}$$

If we define $x = (1+\eta\Delta)^{1/\Delta}$ we note that since $\Delta > 0$ we have $x \in (1, e^{\eta})$ and also $x(\Delta)$ is monotonically decreasing, which follows from

$$\frac{dx}{d\Delta} = \underbrace{\frac{(1+\eta\Delta)^{-1+1/\Delta}}{\Delta^2}}_{>0} \underbrace{(\eta\Delta - (1+\eta\Delta)\log(1+\eta\Delta))}_{<0} < 0. \tag{47}$$

We then note that using this transformation we get:

$$f(\Delta) = g(x) = \frac{x^w}{2x-1}, \tag{48}$$

which satisfies

$$\frac{dg(x)}{dx} = \underbrace{\frac{x^{w-1}}{(1-2x)^2}}_{>0} \underbrace{(2x(w-1)-w)}_{<0} < 0. \tag{49}$$

Using this we find

$$\frac{df(\Delta)}{d\Delta} = \underbrace{\frac{dx}{d\Delta}}_{<0} \underbrace{\frac{dg(x)}{dx}}_{<0} > 0, \tag{50}$$

which proves strict monotonicity of $\langle n \rangle$ as a function of $\Delta > 0$.

## C Alternative derivation of Eqs. (18) and (26) from deterministic rate equations

Consider a cell cycle of fixed duration $T$ with replication (and consequent doubling of transcription) occurring at time $\tau = wT$ (where $w$ is a fraction). If the transcription rate before replication is $r$, the mRNA decay rate is $d$ and $n(t)$ is the deterministic estimate for the mean number of mRNA molecules at time $t$ then a deterministic model for this process is:

$$\frac{dn(t)}{dt} = \begin{cases} r - dn(t), & \text{if } 0 \le t < \tau \\ 2r - dn(t), & \text{if } \tau \le t \le T. \end{cases} \tag{51}$$

In the cyclo-stationary limit, binomial partitioning (when cell division occurs at the end of the cell cycle) leads to the boundary condition $2n(0) = n(T)$. Note that $t$ in this context means the cell age and not absolute time and hence it can only vary between 0 and $T$. Solving these differential equations we obtain the solution:

$$n(t) = \begin{cases} \hat{n}\left(1 + \frac{e^{d(\tau-t)}}{1-2e^{\eta}}\right), & \text{if } 0 \le t < \tau \\ 2\hat{n}\left(1 + \frac{e^{d(\tau-t+T)}}{1-2e^{\eta}}\right), & \text{if } \tau \le t \le T, \end{cases} \tag{52}$$

where $\hat{n} = r/d$. Let $f(t)dt$ be the probability of observing a cell of age between $t$ and $t + \mathrm{d}t$ where $\mathrm{d}t$ is an infinitesimal time interval. It then follows that:

$$f(t)\mathrm{d}t = \lim_{N\to\infty} \pi_j (j = Nt/T), \tag{53}$$

where $\pi_j$ is the probability of observing a cell in cell cycle stage $j$. Note that since $\Delta = 1/N$, the limit of $N \to \infty$ at constant $T$ is the same as the limit of $\Delta \to 0$. Since $T = N/k$, in this limit we have infinite cell stages $N$ advancing with an infinite rate $k$ i.e. the cell spends an infinitesimal small time $\mathrm{d}t = 1/k$ at each stage. Knowing that $\pi_i = 1/N$ for lineage measurements we have

$$f(t) = \lim_{N\to\infty} \frac{\pi_j}{\mathrm{d}t} = \frac{1}{T}. \qquad (54)$$

For the population case we substitute $i/N = t/T$ in Eq. (24) take limit of large $N$ and finally use $N = kT$ to obtain,

$$f(t) = \frac{\ln(2)}{T} 2^{(1-t/T)}. \qquad (55)$$

Note that both Eqs. (54) and (55) are well known and have been in common use for more than 40 years [32]. Finally we obtain the mean number of mRNA averaged over the cell cycle $\bar{n} = \int_0^T f(t)n(t)dt$. For the lineage measurements this yields

$$\bar{n} = w\hat{n} + (1-w)2\hat{n} - \frac{\hat{n}}{\eta}\left(1 - \frac{e^{-\eta(1-w)}}{2-e^{-\eta}}\right), \qquad (56)$$

whilst for the population measurements we obtain

$$\bar{n} = \frac{2^{1-w}\eta\hat{n}}{\eta + \ln(2)}. \qquad (57)$$

These expressions agree exactly with Eq. (18) and Eq. (26) which were derived from a Master Equation approach in the limit of zero variability in the cell cycle duration for the case of lineage and population measurements, respectively.

## D Derivation of the distribution of cell stage durations in population measurements

We let $C_i(t,\tau)$ denote the number of cells in a population that are in cell stage $i$ at time $t$ that have been in that cell state for a duration $\tau$. After a small time duration $\delta$ all the cells will either advance to an age $\tau+\delta$ or advance to the next cell stage. Therefore we can write the conservation equation

$$C_i(t+\delta, \tau+\delta) = C_i(t,\tau) - C_i(t,\tau)k_i\delta. \qquad (58)$$

Assuming that there is a stationary distribution for the stage age of the cell population at a stage $i$, $p_i(\tau)$, we can write $C_i(t,\tau)$ as $C_i(t,\tau) = C_i(t)p_i(\tau)$. Where $C_i(t)$ is the number of cells at cell stage $i$. Introducing this factorization of $C_i(t,\tau)$ in (58), and taking the limit $\delta \to 0$, we get the relationship,

$$\frac{\mathrm{d}C_i(t)}{\mathrm{d}t} p_i(\tau) + \frac{\mathrm{d}p_i(\tau)}{\mathrm{d}\tau} C_i(\tau) = -C_i(t)p_i(\tau)k_i, \qquad (59)$$

where we have used the chain rule to compute the derivative of $\mathrm{d}N_i(x,x)/\mathrm{d}x|_{t,\tau}$. Since the probability of finding a cell in a certain stage $i$, $\pi_i$, is constant in time, the number of cells at a given stage has to grow with the same rate as the population, therefore $\mathrm{d}C_i(t)/\mathrm{d}t = KC_i(t)$, being $K$ the growth rate of the population. Introducing this equality in equation 59, we get an equation for $p_i(\tau)$.

$$Kp_i(\tau) + \frac{\mathrm{d}p_i(\tau)}{\mathrm{d}\tau} = -p_i(\tau)k_i. \qquad (60)$$

That gives,

$$p_i(\tau) = (K + k_i)\,\mathrm{e}^{-(K+k_i)\tau}. \qquad (61)$$

In the Erlang distributed model, $k_i = k$ is constant, and the rate of growth of the population can be calculated from the conservation equation for the total number of cells $C(t)$,

$$C(t+\delta) = C(t) + C_N(t)k_n\delta = C(t) + C(t)k_N\pi_N\delta, \qquad (62)$$

where $n$ is the number of stages of the cell cycle. From this equation, we obtain that the rate of exponential growth of the population is $K = k_N\pi_N$. Using the value of $\pi_N$ from Eq. (24), we obtain that for the Erlang model, the stage age distribution of cell cycle stage $i$ is

$$p_i(\tau) = k2^{1/N}\mathrm{e}^{-k2^{1/N}\tau}. \qquad (63)$$

## E Computational analysis

The simulations for the general cell cycle model (including Erlang distributed times), were made using a custom made Gillespie algorithm where cell cycle stages are treated as one extra reaction (Algorithm 1). After the last stage of the cell cycle is completed, the cell cycle time is reset and the number of mRNAs is reduced by sampling a binomial distribution $B(n, 1/2)$ where $n$ is the number of mRNAs before cell division.

On the other hand, to simulate a cell cycle where the different stages have deterministic duration, the Gillespie algorithm has been modified to take into account if a deterministic cell stage change would take place before the next stochastic reaction time (see Algorithm 2). The rest of the details of the algorithm are the same as in the general cell cycle model.

To obtain statistics from lineage measurements, each trajectory was sampled by choosing evenly distributed time points. For population measurements, several simulations are run in parallel, one for each cell. After

**Algorithm 1** General model with stochastic cell cycle with constitutive expression

```
$n = \hat{n}$                                          ▷ Initial number of mRNA
$j = 1$                                                ▷ Initial cell stage
$t = 0$                                                ▷ Initial time
while $t < t_{max}$ do
    Compute reaction channels propensities
    $p_+ = r_j$                                        ▷ Propensity of mRNA production
    $p_- = nd_j$                                       ▷ Propensity of mRNA degradation
    $p_c = k_i$                                        ▷ Propensity of cell cycle advance
    Select next reaction channel
    $u = \text{UniformRandom}(0, p_+ + p_m + p_c)$     ▷ Random number to select next reaction channel
    if $u < p_+$ then                                  ▷ mRNA production selected
        $n += 1$
        $\tau = \text{ExponentialRandom}(1/p_+)$
    else if $u < (p_+ + p_-)$ then                     ▷ mRNA degradation selected
        $n -= 1$
        $\tau = \text{ExponentialRandom}(1/p_-)$
    else                                               ▷ Cell cycle stage advance selected
        $j += 1$
        $\tau = \text{ExponentialRandom}(1/p_c)$
        if $j > N$ then                                ▷ Cell cycle has finished
            $n = \text{BinomialRandom}(n, 1/2)$        ▷ Bipartition of mRNA
            $j = 1$                                    ▷ Reset cell cycle
    $t += \tau$
```

**Algorithm 2** Deterministic cell cycle model

```
$n = \hat{n}$                                          ▷ Initial number of mRNA
$t = \text{Random}(f(t))$                              ▷ Initial time following cyclostationary distribution
$j = \lceil Nt/T \rceil$                               ▷ Initial cell stage
$t_{next} = s_1$                                       ▷ Time left on the current cell state
while $t < t_{max}$ do
    Compute reaction channels propensities
    $p_+ = r_j$                                        ▷ Propensity of mRNA production
    $p_- = nd_j$                                       ▷ Propensity of mRNA degradation
    Select next stochastic reaction channel
    $u = \text{UniformRandom}(0, p_+ + p_m)$           ▷ Random number to select next reaction channel
    if $u < p_+$ then                                  ▷ mRNA production proposed
        $D_n = 1$                                      ▷ Proposed change in number of mRNAs
        $\tau = \text{ExponentialRandom}(1/p_+)$
    else                                               ▷ mRNA degradation proposed
        $D_n = -1$                                     ▷ Proposed change in number of mRNAs
        $\tau = \text{ExponentialRandom}(1/p_-)$
    Comparison of proposed stochastic reaction channel with cell stage advance
    if $t_{next} > \tau$ then                          ▷ Stochastic reaction selected
        $n += D_n$                                     ▷ Update number of mRNA
        $t_{next} -= \tau$                             ▷ Update time of current cell stage
    else                                               ▷ Cell cycle stage advance
        $j += 1$
        $\tau = t_{next}$
        if $j > N$ then                                ▷ Cell cycle has finished
            $n = \text{BinomialRandom}(n, 1/2)$        ▷ Bipartition of mRNA
            $j = 1$                                    ▷ Reset cell cycle
            $t_{next} = s_1$                           ▷ Start timer of first cell cycle stage
        else
            $t_{next} = s_j$                           ▷ Start timer of next cell cycle stage
    $t += \tau$
```

each cell division event, a new cell is introduced in the simulation containing the remaining mRNA from the binomial partition of the mother cell. In order to achieve a steady state behaviour with deterministic cell cycle it was necessary to initiate each replicate following the corresponding age distribution (Eq. (54) or Eq. (55) ). Statistics from the population measurements are done across all the cells at a particular time snapshot.

## F Inference of transcription rates and error calculation

Using the expression for the average number of mRNAs in the lineage measurements given by Eq. (17), we can write the transcription rate parameter $r$ for the Erlang model as a function of the average number of mRNAs observed and the rest of the parameters of the model,

$$r = \frac{\langle n \rangle \eta}{T} \left( 2 - w - \frac{1}{\eta} \left( 1 - \frac{(\frac{1}{1+\eta\Delta})^{(1-w)/\Delta}}{2 - (\frac{1}{1+\eta\Delta})^{1/\Delta}} \right) \right)^{-1} . \tag{64}$$

Similarly, we can write an expression for $r$ for the population case using Eq. (25),

$$r = \frac{\langle n \rangle \eta}{T} \frac{2^{\Delta} + \eta\Delta - 1}{2^{1-w}\eta\Delta}. \tag{65}$$

We can use both Eqs. (64) and (65) to obtain the average transcription rate $\bar{r}$ along the cell cycle for lineage or population cases,

$$\bar{r} = rw + 2r(1-w). \tag{66}$$

In the limit with a deterministic cell cycle duration and no replication $r_{exp} = \bar{r}(\Delta \to 0, w = 1)$ we recover the expression used in [19], which returns a value of transcription rate for each gene given the measured decay rate and average number of mRNA transcripts. By contrast, in order to compute $\bar{r}$ in a general case we need to evaluate the cell cycle duration variability $\Delta$. For the reported average cell length 27.5h and its standard deviation of 13h, the number of effective states $N$ can be obtained from the coefficient of variation of the cell cycle length $N = 1/\Delta = 1/CV^2 \simeq 4$.

Introducing the calculated value of $\Delta$ in Eqs. (64-66), we can evaluate $\bar{r}$ for lineage and population cases for different DNA replication positions along the cell cycle. In the text we study a case without replication $w = 1$ and a case with replication at the middle of the cell cycle $w = 1/2$.

In order to evaluate how our predictions differ from the reported transcription rates $r_{exp}$, we compute the relative error $\varepsilon$

$$\varepsilon = \frac{r_{exp} - \bar{r}}{r_{exp}}. \tag{67}$$

Note that since $\bar{r}$ is linear in the average transcript number $\langle n \rangle$, the resulting error is independent of $\langle n \rangle$. Therefore, differences in the error $\varepsilon$ among the different genes reported in [19] will only depend on their degradation rate.