

# Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation

Chen Jia<sup>1</sup>, Ramon Grima<sup>2,\*</sup>

<sup>1</sup> Applied and Computational Mathematics Division, Beijing Computational Science Research Center, Beijing 100193, China

<sup>2</sup> School of Biological Sciences, University of Edinburgh, U.K.

\* Correspondence: Ramon.Grima@ed.ac.uk

## Abstract

The stochasticity of gene expression is manifested in the fluctuations of mRNA and protein copy numbers within a cell lineage over time. While data of this type can be obtained for many generations, most mathematical models are unsuitable to interpret such data since they assume non-growing cells. Here we develop a theoretical approach that quantitatively links the frequency content of lineage data to subcellular dynamics. We elucidate how the position, height, and width of the peaks in the power spectrum provide a distinctive fingerprint that encodes a wealth of information about mechanisms controlling transcription, translation, replication, degradation, bursting, promoter switching, cell cycle duration, cell division, and gene dosage compensation. Predictions are confirmed by analysis of single-cell *Escherichia coli* data obtained using fluorescence microscopy. Furthermore, by matching the experimental and theoretical power spectra, we infer the temperature-dependent gene expression parameters, without the need of measurements relating fluorescence intensities to molecule numbers.

## Introduction

In recent years, measurements of the size, division events, and the content of single cells over extended time (many generations) has been made possible due to advances in microfluidic devices and live-cell imaging [1–3]. While the existing data are typically for proteins, such measurements are also in principle possible for mRNAs particularly with the advent of new methods to visualize RNA dynamics in live cells using bright and stable fluorescent RNAs [4]. The data in these experiments are sampled at a rate that is much higher than the frequency of cell division thus providing us with a means to understand the temporal variation of gene expression as a cell progresses through its cell cycle.

A common feature of these time traces is a noisy oscillatory variation of the fluorescence (from fluorescently labelled proteins) with time with a period that is roughly coincident with the interval between two successive cell division events. A sawtooth type of temporal pattern is expected due to a sharp dip in the protein numbers at cell division stemming from the partitioning of the contents of the mother cell amongst two daughter cells. While these oscillations are regular in some cases, very often they display a significant degree of noisiness. This

reflects the inherent stochasticity of transcription, translation, and replication [5–7], noise introduced or modified by homeostatic mechanisms such as those that compensate for the doubling of gene copies at replication [8, 9], and non-genetic sources of noise such as variability in the duration of the cell cycle from one generation to the next [1, 10, 11] and variability in the number of molecules allocated to a new born cell at cell division [12, 13]. Hence it follows that a measure of the regularity of an oscillation, such as the power spectrum of fluorescence fluctuations calculated over a lineage, encapsulates within it a large amount of information about the inherent chemical and physical processes, both deterministic and stochastic, that shape cellular dynamics.

An essential first step to link the properties of the power spectrum to the underlying dynamic intracellular processes is the derivation of a mathematical formula for the spectrum as a function of gene expression rate parameters. For this purpose, the standard stochastic models in the literature that are based on the two-stage or three stage-models of gene expression [14] are not useful because they do not provide a mathematical description of processes along a cell lineage. These model transcription and translation, and implicitly model dilution due to cell division via an effective decay reaction; however simulated time traces of the protein numbers based on these models using the stochastic simulation algorithm will not display any noisy oscillatory behaviour since the partitioning of molecules at cell division is not taken into account. This also implies a lack of description of important cell cycle features such as the high variability in the interdivision time that is a characteristic of experimental time traces. Recently models that surmount the aforementioned limitations of the standard models have been studied leading to mathematical formulae for the mean and variance of mRNA/protein numbers [15–17] and also for the steady-state distributions of these numbers [18, 19] calculated across a cell lineage. These statistical measures provide different information than the power spectrum; notably the former, unlike the latter, neither provide an understanding of the correlations between molecule numbers at two time points nor of the frequency composition of fluctuations in molecule numbers [20].

In this article, for the first time, we calculate in closed-form the power spectrum of fluctuations across a lineage in a stochastic gene expression model with a high level of biological realism, including a description of transcription, translation, degradation, bursting, promoter switching, DNA replication, gene dosage

compensation, and symmetric/asymmetric partitioning at cell division. The analytical expressions give insight into how the regularity and noisiness of the oscillations in the mRNA/protein abundance across generations are related to the rate parameters associated with the various subcellular processes at play; the theory also makes various predictions that are then verified by analysis of a publicly available single-cell data set of *Escherichia coli* followed over 70 generations for three different growth conditions; finally we show how matching the experimental and theoretical power spectra enables an accurate inference of all the rate parameters.

## Results

### Model specification

We consider a detailed model of stochastic gene expression in an individual cell which includes promoter switching, transcriptional or translational bursting, cell cycle duration variability, gene replication, gene dosage compensation, and symmetric or asymmetric cell division (see Fig. 1(a) for an illustration). The specific meaning of all model parameters is listed in Table 1 and the biological values of some key parameters in different cell types are listed in table S1. The model is based on a number of assumptions that are closely tied to experimental data. The assumptions are as follows.

Meaning of model parameters	
$u$	gene switching rate from OFF to ON before replication
$gu$	gene switching rate from OFF to ON after replication
$v$	gene switching rate from ON to OFF
$\rho_0$	burst production rate when the gene is OFF
$\rho_1$	burst production rate when the gene is ON
$B$	mean burst size of the gene product
$d$	degradation rate of the gene product
$N$	number of cell cycle stages
$N_0$	number of cell cycle stages before replication
$w$	proportion of the cell cycle before replication
$a$	transition rate from one cell cycle stage to the next
$f$	cell cycle frequency
$\eta$	ratio of the degradation rate and the cell cycle frequency
$p$	allocation probability in asymmetric binomial partitioning
$\rho_{\text{eff}}$	effective burst production rate before replication
$\kappa\rho_{\text{eff}}$	effective burst production rate after replication
$d_{\text{eff}}$	effective decay rate of the gene product

Table 1. **Model parameters and their meaning.**

1) The promoter of the gene of interest can switch between an inactive state  $G_0$  and an active state  $G_1$  with switching rates  $u$  and  $v$  before gene replication [14]. Dosage compensation is modeled as a change in the switching rate of the promoter from the inactive to the active state upon replication with its value being  $u$  before replication and  $gu$  after replication. This assumption is supported by experiments [8].

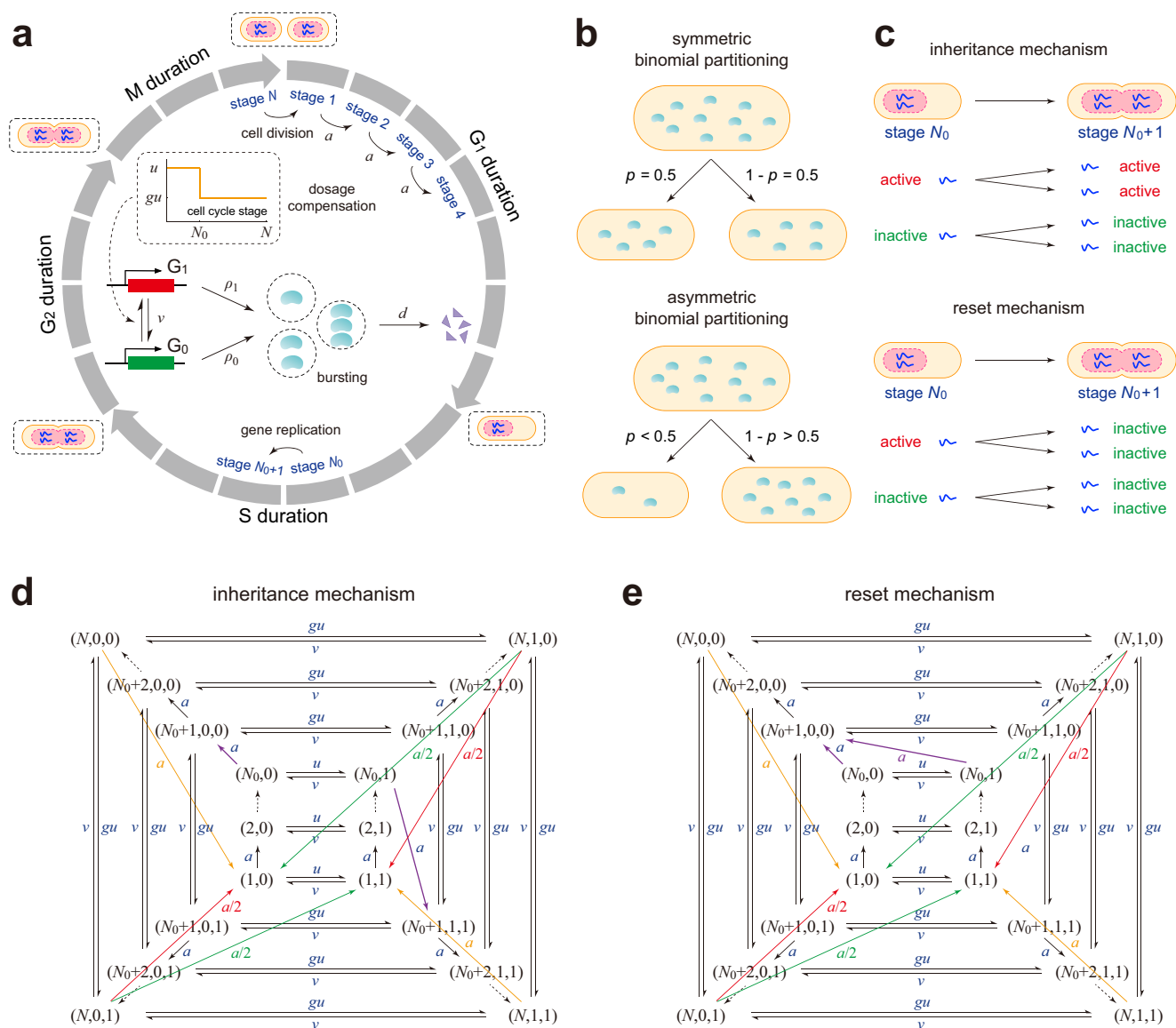
2) In each gene state  $G_i$  ( $i = 0, 1$ ), the synthesis of the gene product of interest, mRNA or protein, occurs at a rate  $\rho_i$  in bursts of a random size sampled from an arbitrary probability distribution  $\mu = (\mu_n)$ . This means that in each burst, there is a probability  $\mu_n$  of producing  $n$  copies of the gene product. In previous papers, the synthesis of mRNA in each gene state is assumed to be non-bursty [21], i.e. the mRNA molecules are produced one at a time. In this case,  $\mu_n = \delta_{1,n}$  is the Kronecker delta which takes the value of 1 when  $n = 1$  and the value of 0 otherwise. On the other hand, in models where mRNA is not explicitly described, the effective synthesis of protein in each gene state is usually assumed to be bursty with the burst size sampled from a geometric distribution  $\mu_n = p_B^n(1 - p_B)$ , where  $p_B = B/(1 + B)$  with  $B$  being the mean burst size; this is due to rapid synthesis of protein molecules from a single short-lived mRNA molecule [22, 23]. Therefore, the arbitrariness of the burst size distribution allows us to analyze the dynamics of both mRNA and protein in a unified model.

3) The gene product is degraded via first-order kinetics with rate constant  $d$ , which is a common assumption supported by experiments [24].

4) Each cell can exist in  $N$  effective cell cycle stages, denoted by  $1, 2, \dots, N$ , with  $a$  being the transition rate from one stage to the next, which is assumed to be the same for all stages. Since the transition time between stages is exponentially distributed, the duration of the cell cycle is Erlang distributed with mean  $N/a$  and thus the cell cycle frequency is  $f = a/N$ . In our model, the noise in the doubling time, characterized by the coefficient of variation squared, is equal to  $1/N$ . As  $N \rightarrow \infty$ , the noise vanishes and thus the doubling time becomes fixed. Hence, our model allows the investigation of the influence of cell cycle duration variability on stochastic gene expression.

We emphasize that the effective cell cycle stages introduced here do not directly correspond to the four biological cell cycle phases of eukaryotic cells ( $G_1$ , S,  $G_2$ , and M) since the durations of the latter are typically not exponentially distributed. In our model, a cell cycle phase corresponds to multiple effective cell cycle stages (Fig. 1(a)). By introducing a number of effective cell cycle stages, our model has the property that the total duration of the cell cycle and the durations of individual cell cycle phases are all Erlang distributed. This is in agreement with experiments in various cell types [17, 25–28].

5) Cell division occurs when the cell transitions from effective stage  $N$  to the next stage 1. At division, most previous papers assume that the mother cell divides into two via symmetric binomial partitioning: each molecule has an equal chance to be allocated to one of the two daughters [19, 29]. However, asymmetric cell division is common in biology [13, 30]. For instance, *Saccharomyces cerevisiae* divides asymmetrically into two daughters with different sizes. *Escherichia coli* may also undergo asymmetric division with old daughters receiving fewer gene products than new daughters [31]. Here we extend previous models by considering asymmetric binomial partitioning at cell division: the probability for a molecule being allocated to one daughter is  $p \leq 1/2$  and the probability of being allocated to the other is  $q = 1 - p$  (Fig. 1(b)). After cell division, we randomly track one of the two daughters with probability  $1/2$ ; hence our



**Fig. 1. The model.** (a) Schematic illustrating the model describing  $N$  effective cell cycle stages, gene replication at stage  $N_0$ , promoter switching between active (red) and inactive (green) states, bursty production of the gene product in the two gene states, degradation, and gene dosage compensation induced by a decrease in the activation rate of the gene after replication (see inset graph). (b) At stage  $N$ , a mother cell divides into two daughters that are typically different in size (asymmetric division) with the larger daughter inheriting more molecules. Symmetric division is the special case where the daughters are equisized. (c) At replication, the gene states of the two daughter copies can be the same as that of the mother copy (inheritance mechanism) or else they are both reset to the inactive state (reset mechanism). (d),(e) Transition diagram of cellular states under the two mechanisms. Before replication, a cellular state can be represented by an ordered pair  $(r, i)$ , where  $r$  is the cell cycle stage and  $i$  is the gene state of the mother copy; after replication, a cellular state can be represented by an ordered triple  $(r, i, j)$ , where  $i$  and  $j$  are the gene states of the two daughter copies. See main text for explanation of the colored arrows.

model corresponds to cell lineage measurements performed using a mother machine such as in [32].

6) The replication of the locus containing the gene of interest occurs over a period that is much shorter than the rest of the cell cycle. Note that the replication of the whole genome within a cell cannot be assumed to be instantaneous. However, since the replication time of a particular gene is much shorter than the total duration of the S phase, it is reasonable to consider it to be instantaneous. Specifically, we assume that gene replication occurs instantaneously when the cell transitions from a fixed effective stage  $N_0 \in [1, N - 1]$  to the next stage  $N_0 + 1$ . We

shall refer to the gene copy that is replicated as the mother copy and to the duplicated gene copies as the daughter copies. Under this assumption, for haploid cells, there is only one mother copy during the first  $N_0$  stages and two daughter copies during the last  $N - N_0$  stages; for diploid cells, the number of gene copies varies from two to four upon replication. For diploid cells, we assume that the two alleles act independently of each other [33, 34].

7) After replication, the two daughter copies can either inherit the gene state from the mother copy or be both reset to the inactive state [18]. To distinguish between them, we refer to the former as the inheritance mechanism and the latter as the reset

mechanism (Fig. 1(c)). The consideration behind the former is the copying of the landscape of histone modifications (implicated in gene activation) during DNA replication [35]. One plausible explanation for the latter is that to avoid the potential risk of conflict between replication and transcription [36], it is likely that in the region where replication is ongoing or just completed, there is no transcription, indicating an inactive state.

We next describe our stochastic model for haploid cells. The results for diploid cells can be easily deduced from the haploid case using allelic expression independence (see Methods). Based on the stage of cell cycle progression and the states of the gene copies, there are many possible cellular states for a cell. Before replication, the cell can exist in  $2N_0$  cellular states that can be represented by an ordered pair  $\alpha = (r, i)$ , where  $r \in [1, N_0]$  is the cell cycle stage and  $i = 0, 1$  is the gene state of the mother copy; after replication, the cell can exist in  $4(N - N_0)$  cellular states that can be represented by an ordered triple  $\alpha = (r, i, j)$ , where  $r \in [N_0 + 1, N]$  is the cell cycle stage and  $i, j = 0, 1$  are the gene states of the two daughter copies. In sum, there are a total of  $K = 4N - 2N_0$  possible cellular states.

The evolution of cellular state transitions is governed by a Markovian model whose master equation is given in section S1. The transition diagrams of the Markovian model under the inheritance and reset mechanisms are illustrated in Fig. 1(d),(e), respectively. The purple arrows show that upon replication, the cellular state will transition from  $(N_0, i)$  to  $(N_0 + 1, i, i)$  for the inheritance mechanism and from  $(N_0, i)$  to  $(N_0 + 1, 0, 0)$  for the reset mechanism. After division, we randomly track one of the two daughters and thus the cell will transition from  $(N, i, j)$  to  $(1, i)$  or  $(1, j)$ . The orange arrows illustrate those transitions from  $(N, i, i)$  to  $(1, i)$  with rate  $a$ , the red arrows illustrate those from  $(N, i, j), i \neq j$  to  $(1, i)$  with rate  $a/2$ , and the green arrows illustrate those from  $(N, i, j), i \neq j$  to  $(1, j)$  with rate  $a/2$ .

It then follows that the microstate of the gene of interest can be represented by  $(\alpha, n)$ , where  $\alpha$  is the cellular state and  $n$  is the copy number of the gene product. The evolution of the complete stochastic gene expression dynamics is governed by a master equation which is given in Methods.

### General expressions for the power spectrum of single-cell measurements across generations

Experiments suggest that the periodicity of the cell cycle can induce oscillatory behavior in gene expression [1]. In fact, if we use a deterministic model to describe the synthesis and degradation of the gene product and assume that the molecule numbers halve after a deterministic doubling time, then the solution of the deterministic rate equation (the time series of gene product abundances) will be periodic under all choices of rate parameters. However, numerous time lapse experiments [6, 37] have shown that the time course data of expression levels in single cells do not always appear oscillatory, due to various sources of noise such as cell cycle duration variability, gene copy number variability, cell division asymmetry, gene expression bursting, and promoter switching. Here we examine how these sources of noise influence the robustness of sustained oscillations.

Let  $n(t)$  denote the copy number of the gene product in a cell at time  $t$ . Stochastic gene expression oscillations are often

characterized by two functions: the autocorrelation function and the power spectrum. The former  $R(t) = \text{Cov}_{ss}(n(0), n(t))$  is defined as the steady-state covariance of  $n(0)$  and  $n(t)$ , while the latter  $G(\xi) = \int_{-\infty}^{\infty} R(|t|)e^{-2\pi i \xi t} dt$  is defined as the Fourier transform of the former, where  $\xi \geq 0$  denotes the frequency. In general, sustained oscillations cannot be observed if the power spectrum  $G(\xi)$  is a monotonically decreasing function of  $\xi$ . In contrast, a non-monotonic power spectrum with one or more off-zero peaks implies the presence of sustained oscillations; the dominant frequencies of these oscillations are the values of  $\xi$  at which the modes of the power spectrum occur.

While our model is complex due to the large variety of biological processes that it captures, its autocorrelation function can still be computed exactly as

$$R(t) = (m_1 + m_2)e^{Wt}\mathbb{1} - (m_1\mathbb{1})^2 + B \int_0^t m_1 e^{Qs} S e^{W(t-s)} \mathbb{1} ds, \quad (1)$$

where  $m_1 = (m_{1\alpha})$  and  $m_2 = (m_{2\alpha})$  are two row vectors whose components are the first and second factorial moments of the gene product abundance in all cellular states,  $\mathbb{1}$  is the column vector whose components are all 1,  $S = \text{diag}(\rho_\alpha)$  is the diagonal matrix whose diagonal elements are the burst production rates in all cellular states,  $Q = (q_{\alpha\beta})$  is generator matrix of cellular state transitions (see Fig. 1(d),(e) for the transition diagram which has some orange, red, and green arrows), and  $W = (w_{\alpha\beta})$  is another matrix obtained from  $Q$  by replacing the rates of orange arrows by  $a/2$ , replacing the rates of red arrows by  $pa/2$ , replacing the rates of green arrows by  $qa/2$ , and subtracting  $d$  from the diagonal entries (see Methods for the proof and the detailed expressions of each term).

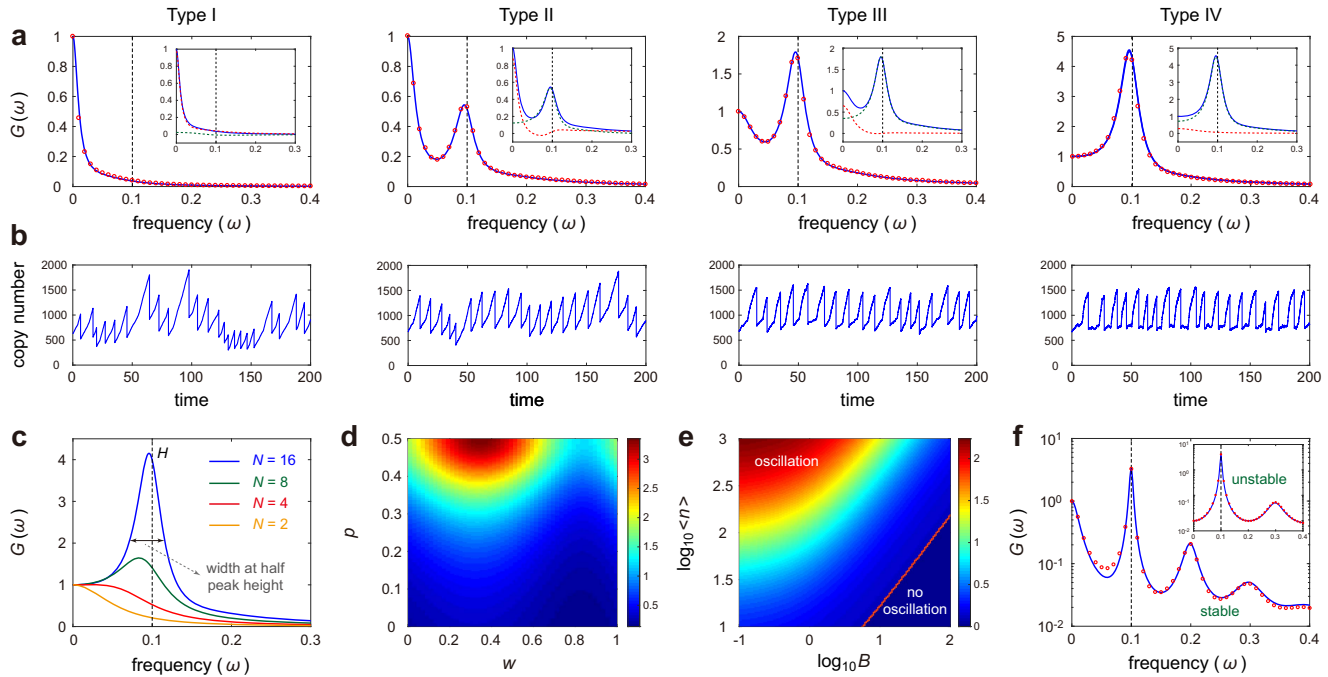
Note that the autocorrelation function  $R(t)$  in Eq. (1) is expressed in matrix form. A more explicit expression can be obtained by expanding the matrix exponentials  $e^{Wt}$  and  $e^{Qs}$  in terms of their eigenvalues and eigenvectors. We find that the autocorrelation function (power spectrum) can be rewritten as the linear combination of  $2K - 1$  exponential (Lorentzian) functions:

$$R(t) = \sum_{k=1}^{2K-1} u_k e^{\lambda_k t}, \quad G(\xi) = \sum_{k=1}^{2K-1} \frac{-2u_k \lambda_k}{4\pi^2 \xi^2 + \lambda_k^2}, \quad (2)$$

where  $K$  is the number of cellular states,  $\lambda_k$  ( $0 \leq k \leq 2K - 1$ ) are all the eigenvalues of the two matrices  $Q$  and  $W$ , and  $u_k$  are suitable constants (see Methods for the proof and the specific expressions of  $u_k$ ). Since  $Q$  and  $W$  are both  $K \times K$  matrices, they have a total of  $2K$  eigenvalues.

An important special case occurs when promoter switching is much faster than cell cycle progression and gene product degradation, i.e.  $f, d \ll u, v$  [34]. In this case, gene switching dynamics will reach rapid equilibrium. As a result, the effective burst production rate is given by  $\rho_{\text{eff}} = (\rho_1 u + \rho_0 v)/(u + v)$  before replication and given by  $\kappa \rho_{\text{eff}} = 2(\rho_1 g u + \rho_0 v)/(g u + v)$  after replication, where  $\kappa \geq 1$  is a factor characterizing the change in the burst production rate due to gene replication and dosage compensation. In the absence of dosage compensation, we have  $g u = u$  and thus  $\kappa = 2$ . In the fast switching regime, the two gene states can be combined into a single one and hence the cellular state is only determined by the  $N$  effective cell cycle





**Fig. 2. Properties of the power spectrum.** (a) For significant cell cycle duration variability, the spectra can be of four types. Note that the spectra are normalized such that  $G(0) = 1$ . Theory (blue) matches stochastic simulations using Gillespie’s algorithm (red circles). Vertical dashed lines show the cell cycle frequency. Insets show the decomposition of the spectrum into two parts controlling the zero and off-zero peaks (red and green dashed curves, respectively). (b) Typical trajectories for each type of spectra. (c) The off-zero peak becomes higher as cell cycle duration variability decreases ( $N$  increases). (d),(e) Heat maps showing the dependence of the height of the off-zero peak (relative to the zero peak) on the parameters controlling the asymmetry of partitioning ( $p$ ), the replication point ( $w$ ), the burstiness of gene expression ( $B$ ), and the mean number of gene product molecules ( $\langle n \rangle$ ). Type I spectra are observed in the region marked “no oscillation”. (f) For low cell cycle duration variability, higher-order harmonics of the cell cycle frequency appear; for midway replication, an unstable gene product, e.g. most mRNAs, has a spectrum with no peaks at even harmonics, while a stable gene product, e.g. most proteins, has peaks at all harmonics. Note that the blue curves are computed using the analytical expression given in Eq. (2) and the red circles are computed using the approximate expressions given in Eq. (5) for stable products and Eq. (15) for unstable products. See section S10 for the technical details of this figure.

stages. In this case, we do not need to distinguish between the inheritance and the reset mechanisms because they lead to the same oscillatory behavior.

We next examine the properties of stochastic oscillations under fast promoter switching. In this regime,  $Q$  and  $W$  reduce to  $N \times N$  matrices and the power spectrum in Eq. (2) simplifies because all the eigenvalues can be computed explicitly. The eigenvalues of  $Q$  are given by  $\lambda_k = -a + a\omega_k$  ( $0 \leq k \leq N-1$ ) and the eigenvalues of  $W$  are given by  $\lambda_{N+k} = -d - a + 2^{-1/N}a\omega_k$  ( $0 \leq k \leq N-1$ ), where  $\omega_k = e^{2\pi ki/N}$  are all the  $N$ th roots of unity. In addition, the steady-state mean of gene product abundances can be also calculated explicitly as (see section S3 for the proof)

$$\langle n \rangle = \frac{\rho_{\text{eff}} B}{\eta f} \left[ w + \kappa(1-w) + \frac{(\kappa-1)a_{\eta,N}^w + 1 - \kappa a_{\eta,N}}{\eta(2a_{\eta,N} - 1)} \right], \quad (3)$$

where  $w = N_0/N$  is the proportion of the cell cycle before gene replication,  $\eta = d/f$  is the ratio of the degradation rate and the cell cycle frequency which serves as a measure for the stability of the gene product, and  $a_{\eta,N} = (1 + \eta/N)^N \approx e^\eta$  when  $N \gg 1$ .

To validate the analytical expression of the power spectrum, we compare it with the numerical spectrum calculated by means of the Wiener-Khinchin theorem, which states that  $G(\xi) = \lim_{T \rightarrow \infty} \langle |\hat{n}_T(\xi)|^2 \rangle / T$ , where  $\hat{n}_T(\xi) = \int_0^T n(t) e^{-2\pi i \xi t} dt$  is the

truncated Fourier transform of a single stochastic trajectory over the interval  $[0, T]$  and the angled brackets denote the ensemble average over trajectories (Fig. 2(a)). To guarantee the accuracy of the numerical spectrum, we calculate the time-ensemble average over 5000 stochastic trajectories simulated using Gillespie’s algorithm with the maximum simulation time being chosen as  $30N/a$  (about 30 cell cycles). Here we normalize the power spectrum such that  $G(0) = 1$ . Clearly, the analytical (blue curve) and numerical (red circles) spectra coincide perfectly with each other. In general, the numerical simulations of the power spectrum turn out to be very slow. The analytical solution is hence crucial because it allows a fast exploration of large swathes of parameter space.

### Single-cell time traces can be classified into four different types according to their power spectra shape

Our gene expression model displays four different types of power spectra (Fig. 2(a)): (i) the spectrum is unimodal and monotonically decreasing with a peak at zero, (ii) the spectrum is bimodal with the height of the off-zero peak less than 1, (iii) the spectrum is bimodal with the height of the off-zero peak greater than 1, and (iv) the spectrum is unimodal and bell-shaped with the height of the off-zero peak greater than 1. For convenience, we

refer to (i)-(iv) as type I, II, III, and IV spectra, respectively. The robustness of oscillations increases as the spectrum changes from type I to type IV (Fig. 2(b)); this is since the increasing height of the off-zero peak relative to the zero peak implies increasing power in a narrow range of frequencies centered about the cell cycle frequency. To better understand the analytical solution of the power spectrum, we decompose it into two parts:

$$G(\xi) = \sum_{k=1}^{N-1} \frac{-2u_k \lambda_k}{4\pi^2 \xi^2 + \lambda_k^2} + \sum_{k=N}^{2N-1} \frac{-2u_k \lambda_k}{4\pi^2 \xi^2 + \lambda_k^2}, \quad (4)$$

where the first part is the contribution of the eigenvalues of  $Q$  and the second part is the contribution of the eigenvalues of  $W$  (insets of Fig. 2(a)). Clearly, the first part (green curves) mainly controls the off-zero peak and the second part (red curves) mainly controls the zero peak as well as the decay of the spectrum. When cell cycle duration variability is small ( $N \gg 1$ ), the first eigenvalue of  $W$  is given by  $\lambda_N = -d - a(1 - 2^{-1/N}) = -d - a(1 - e^{-(\ln 2)/N}) \approx -d - (\ln 2)a/N = -d_{\text{eff}}$ , where  $d_{\text{eff}} = d + (\ln 2)f$  is the effective decay rate of the gene product, which is the sum of the decay rate due to active degradation and the decay rate due to dilution at cell division [38]. This explains why the second part characterizes the decay of the spectrum.

### General scaling properties of the height and width of the off-zero spectral peak

To see the effect of cell cycle duration variability on sustained oscillations, we illustrate how the power spectrum given by Eq. (2) varies with  $N$  (Fig. 2(c)). When  $N$  is very small, the spectrum only has a peak at zero, implying that no regular oscillations can be observed. However, as  $N$  increases, the spectrum becomes non-monotonic with the off-zero peak becoming higher and closer to (but still less than) the cell cycle frequency  $f$  (shown as a vertical dashed line). This indicates that there is a threshold cell cycle duration variability below which the periodicity of the cell cycle leads to sustained oscillations in gene expression.

Moreover, as  $f$  increases while keeping  $N$  and  $\langle n \rangle$  fixed, the power spectrum becomes much wider but the height of the off-zero peak remains exactly the same (fig. S1(a)). In fact, if the cell cycle frequency is increased from  $f$  to  $\alpha f$  with some  $\alpha > 1$ , then the coefficients  $u_k$  in Eq. (2) will remain the same, but the eigenvalues  $\lambda_k$  will be replaced by  $\alpha \lambda_k$  (see section S2 for the proof). Hence, the power spectrum (autocorrelation function) will be stretched (compressed) along the horizontal axis by a factor of  $\alpha$ . This explains why the height of the off-zero spectral peak is independent of the cell cycle frequency.

### Symmetric division, midway replication, and non-bursty expression enhance the regularity of oscillations

Oscillations are also affected by the gene replication time, asymmetric cell division, gene expression bursting, and gene expression mean. Fig. 2(d),(e) illustrate the height of the off-zero spectral peak as a function of  $w$ ,  $p$ ,  $B$ , and  $\langle n \rangle$ . It is clear that the off-zero peak becomes lower as  $B$  increases and as  $p$  and  $\langle n \rangle$  decrease. The decline of the peak height with increasing  $B$ , decreasing  $p$ , and decreasing  $\langle n \rangle$  is expected

since all of them correspond to an increase in the fluctuations of gene product abundances which counteracts the regularity of oscillations; indeed noise above a certain threshold can even completely destroy sustained oscillations (Fig. 2(e)). This is in sharp contrast to a negative feedback genetic loop, where random bursting can promote the regularity of oscillations [39, 40]. In addition, we also find that sustained oscillations are the most regular when  $w$  is neither too small nor too large (Fig. 2(d)).

### Contrasting the properties of the off-zero peak for stable and unstable gene products

To further understand these observations, we consider two important special cases. In bacteria and yeast, most proteins have very long half-lives, i.e.  $\eta \ll 1$  (table S1). For such stable gene products, when cell cycle duration variability is not too large, the power spectrum can be simplified as

$$G(\xi) \approx \frac{2A\langle n \rangle f}{3(4\pi^2 \xi^2 + \lambda_N^2)} + \frac{\rho_{\text{eff}}^2 B^2 N}{2f} \sum_{k=1}^{[N/2]} H_k G_k(\xi), \quad (5)$$

where  $A$  and  $H_k, k \geq 1$  are suitable constants and

$$G_k(\xi) = \frac{2k^2 \pi^2 f^2 - (\pi^2 \xi^2 - a^2 \cos \theta_k \sin^2 \theta_k / 2)}{4k^6 \pi^6 f^4 + N^2 (\pi^2 \xi^2 - a^2 \cos \theta_k \sin^2 \theta_k / 2)^2}, \quad (6)$$

with  $\theta_k = 2k\pi/N$  (see Methods for the expressions of  $A$  and  $H_k$  and section S4 for the proof). In Eq. (5), the decay of the power spectrum is mainly controlled by the first term, while the off-zero peak is controlled by the function  $G_1(\xi)$  in the second term. The influence of the remaining functions  $G_k(\xi), k \geq 2$  in the second term will be discussed later.

From Eq. (6) with  $k = 1$ , the position of the off-zero peak is given by  $\xi = (a/\pi) \cos \theta_1 \sin(\theta_1/2) < a\theta_1/2\pi = a/N$ , which is smaller than the cell cycle frequency  $f = a/N$ . When  $N \gg 1$ , the peak position is approximately equal to  $f$  since  $\sin \theta \approx \theta$  and  $\cos \theta \approx 1$  when  $\theta$  is small (Fig. 2(c)). Moreover, the width of the off-zero peak, characterized by the difference of the frequencies at which the spectrum attains half of its peak value (see Fig. 2(c) for an illustration), is given by  $D = 2\pi f/N$ . In other words, the width is proportional to both the cell cycle frequency and cell cycle duration variability (in agreement with fig. S1(a)). In the case of bursty gene expression, the height of the off-zero peak is given by (see section S4 for the proof)

$$H = \frac{J_1 \langle n \rangle N / 6\pi^4}{\langle n \rangle [2J(\gamma + \frac{2}{3N}) + \frac{C}{\pi^2 N}] + \frac{3}{\ln 2} + 2B(\gamma - 2 + \frac{3}{\ln 2})}, \quad (7)$$

where  $\gamma = 2(1 - 4pq)/(1 + 2pq)$  is a function of  $p$ , and

$$J_k = \frac{[w + \kappa(1 - w) + \frac{(\kappa-1)\sin(2k\pi w)}{2k\pi}]^2 + \frac{(\kappa-1)^2 \sin^4(k\pi w)}{k^2 \pi^2}}{[\kappa - \frac{1}{3}(\kappa-1)w(4-w)]^2},$$

$$J = \frac{[w + \kappa(1 - w)]^2}{[\kappa - \frac{1}{3}(\kappa-1)w(4-w)]^2}, \quad (8)$$

and  $C = \sum_{k=1}^{\infty} J_k / k^2$  are all functions of  $\kappa$  and  $w$ . In the non-bursty case, the term involving  $B$  in Eq. (7) vanishes. Note that since  $p \leq 1/2$  is the probability that a molecule is allocated

to one daughter and  $q = 1 - p$  is the probability of being allocated to the other daughter, it follows that  $0 \leq \gamma \leq 2$ ; hence the parameter  $\gamma$  is a dimensionless measure of the asymmetry of cell division. From Eq. (7), it is easy to see that  $H$  decreases with increasing  $B$  and decreasing  $p$  and  $\langle n \rangle$ . This is in full agreement with the numerical results shown in Fig. 2(d),(e) which are computed using Eq. (2).

On the other hand, in bacteria and yeast, most mRNAs have very short half-lives, i.e.  $\eta \gg 1$  (table S1). For such unstable gene products, we also derive a simplified expression of the power spectrum which is given in Methods. The width of the off-zero peak is still given by  $D = 2\pi f/N$  but the height is given by

$$H = \frac{\sin^2(w\pi)N^2}{2C(w)\pi^2}, \quad (9)$$

where  $C(w) = \sum_{k=1}^{\infty} \sin^2(kw\pi)/k^2$  is a function of  $w$  (see section S5 for the proof). We emphasize that this formula is derived in the limit of  $\eta \rightarrow \infty$ . In naturally occurring systems, the value of  $\eta$  for an unstable gene product usually does not exceed 100 (table S1) and hence applying this formula may lead to some errors, especially when  $\kappa$  is very close to 1. It is interesting to note that for unstable gene products, the height of the off-zero peak depends less on asymmetric division, gene expression mean, random bursting, and dosage compensation, whereas from Eq. (7), it is clear that the opposite is true for stable gene products. For both stable and unstable cases, as cell cycle duration variability becomes smaller ( $N$  increases), the off-zero peak becomes narrower and higher.

Recall that for quasi-symmetric cell division, sustained oscillations are the most regular when  $w$  is neither too small nor too large (Fig. 2(d)). For stable gene products with  $\eta \ll 1$ , it follows from Eq. (7) that the maximal regularity is obtained when  $w \approx 0.29$  in the case of symmetric division, large gene expression mean, and no dosage compensation. For unstable gene products with  $\eta \gg 1$ , it follows from Eq. (9) that the maximal regularity is obtained when  $w = 0.5$ .

Note that our conclusions for the differences between the power spectra of (unstable) mRNA and (stable) protein are also confirmed by numerical simulations of a more complex model with both mRNA and protein descriptions (rather than an effective protein description with bursting dynamics as described in point 2 in the Model Specification section).

### Increasing the asymmetry of cell division induces a sharp change in the power spectrum for stable gene products

Our theory further reveals an important difference between symmetric ( $p = 1/2$ ) and asymmetric ( $p < 1/2$ ) cell division for stable gene products. When the gene expression mean is large, it follows from Eq. (7) that the height of the off-zero peak reduces to  $H \approx J_1 N / 6\pi^4 [2J(\gamma + 2/3N) + C/\pi^2 N]$ . For symmetric division, we have  $\gamma = 0$  and thus the height depends on  $N$  quadratically as  $H \approx J_1 N^2 / 2\pi^2 (4\pi^2 J + 3C)$  (fig. S2(a)). For asymmetric division, however, we have  $0 < \gamma \leq 2$  and thus when cycle cycle duration variability is very small ( $N \gg 1$ ), the height depends on  $N$  linearly as  $H \approx J_1 N / 12\pi^4 J \gamma$  (fig. S2(b)). As cell division asymmetry becomes stronger, the height transitions sharply from the  $N^2$  law to the  $N$  law (fig. S2(c)). This

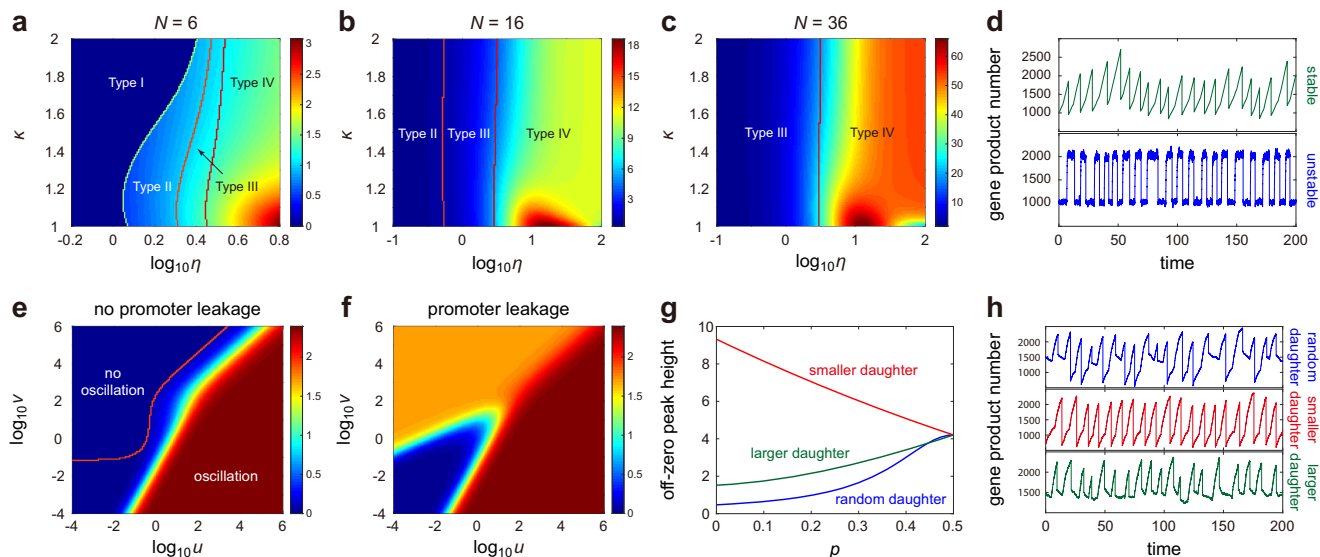
implies that for a certain cell cycle duration variability (fixed  $N$ ), symmetric division leads to a much higher degree of regularity in oscillations for stable products than asymmetric division. In contrast, we find that for unstable gene products, there is no analogous transition because the height of the off-zero peak is always proportional to  $N^2$  and is independent of  $p$  which controls the asymmetry of cell division (see Eq. (9)).

### Single-cell time traces can display higher-order harmonics of the cell cycle frequency

Interestingly, when cell cycle duration variability is very small, besides the peak at the cell cycle frequency  $f$ , the power spectrum also has peaks at integer multiples of  $f$  (Fig. 2(f) and fig. S2(a),(b)). This shows that besides the fundamental period of the mean doubling time  $T = 1/f$ , the system also has the hidden periods of  $T/2$  and even  $T/3$ . Note that in Fig. 2(a),(c), we have not observed higher-order harmonics because the cell cycle duration variability is not sufficiently small ( $N$  is not sufficiently large). Similar peaks at higher-order harmonics have been previously reported for biochemical systems with feedback loops, due to the combination of intrinsic noise and nonlinearity in the law of mass action [41]. In the present model, the propensities of the reactions are all linear in molecule numbers and hence the hidden periods cannot be attributed to the same mechanism as in [41].

These hidden frequencies can be better understood using our analytical results. Recall that the spectral peak at  $f$  is controlled by the function  $G_1(\xi)$  in the second term of Eq. (5). In fact, the remaining functions  $G_k(\xi)$ ,  $k \geq 2$  in the second term control the peaks at higher-order harmonics  $kf$ . From Eq. (6), for stable gene products, the height and width of the spectral peak at the  $k$ th harmonic frequency are given by  $J_k H / J_1 k^4$  and  $k^2 D$ , respectively, where  $J_k$  is given in Eq. (8). In particular, when  $w$  is very small or very large, we have  $J_k \approx 1$  for all  $k$  and hence the height of the spectral peak at  $f$  is  $2^4 = 16$  times greater than that at  $2f$  and  $3^4 = 81$  times greater than that at  $3f$ . Moreover, the width of the spectral peak at  $f$  is  $2^2 = 4$  times lesser than that at  $2f$  and  $3^2 = 9$  times lesser than that at  $3f$ . This is in full agreement with our simulation results which show that the peaks at higher-order harmonic frequencies are much lower and wider than the peak at the fundamental frequency (Fig. 2(f) and fig. S2(a),(b)).

For unstable gene products, similar phenomena are also observed but the characteristics of the higher-order spectral peaks are slightly different. Actually, the height and width of the spectral peak at the  $k$ th harmonic frequency are given by  $\sin^2(kw\pi)H / \sin^2(w\pi)k^4$  and  $k^2 D$ , respectively (see Methods for the proof). When  $w$  is very small or very large, we have  $\sin^2(kw\pi) / \sin^2(w\pi) \approx k^2$  and hence the height of the spectral peak at  $f$  is  $2^2 = 4$  times greater than that at  $2f$  and  $3^2 = 9$  times greater than that at  $3f$ . Another interesting prediction is that for midway replication ( $w = 0.5$ ), there are no peaks at even harmonics since the height at  $2kf$  is zero (inset of Fig. 2(f)). As a result, we find that for midway replication, stable gene products yield higher peaks at higher-order harmonics, while for early or late replication, unstable gene products yield higher peaks at higher-order harmonics.



**Fig. 3. Further properties of the power spectrum.** (a)-(c) Heat maps showing the dependence of the height of the off-zero peak (relative to the zero peak) on the parameters controlling gene product stability ( $\eta$ ) and dosage compensation ( $\kappa$ ). For stable gene products with  $\eta \ll 1$ , it is only possible to observe type I-III spectra; type IV can be produced by unstable gene products with  $\eta \gg 1$ . Strong dosage compensation ( $\kappa = 1$ ) in (b),(c) leads to maximal oscillation regularity at an intermediate gene product stability. (d) Typical trajectories for stable and unstable gene products. With the same gene expression mean, unstable products lead to more robust sustained oscillations than stable ones. (e) Heat map showing the dependence of the height of the off-zero peak (relative to the zero peak) on promoter switching rates  $u$  and  $v$  in the absence of promoter leakage. (f) Same as (e) but in the presence of promoter leakage. Fast promoter switching, leakage, and active gene state dominance enhance oscillation regularity. (g) Height of the off-zero peak (relative to the zero peak) versus the asymmetric allocation probability  $p$  for three types of tracking protocols at cell division: tracking one of the two daughters randomly (blue), tracking the smaller daughter (red), and tracking the larger daughter (green). (h) Typical trajectories for the three protocols generated using Gillespie's algorithm. Note that (a)-(c) and (e)-(g) are obtained from numeric evaluation of Eq. (2). See section S11 for the technical details of this figure.

### Bifurcations between the four types of power spectra are observed as the gene product stability is varied

The dependence of oscillations on gene product stability  $\eta$  is much more complicated. Fig. 3(a)-(c) illustrate the height of the off-zero spectral peak as a function of  $\eta$ ,  $\kappa$ , and  $N$  when the gene expression mean is large. Since type I spectra are monotonically decreasing, the height of the off-zero peak is set to be zero for convenience. For stable gene products with  $\eta \ll 1$ , it is only possible to observe type I, II, and III spectra. It can be analytically shown that for the case of symmetric division, large gene expression mean, midway replication ( $w = 0.5$ ), and no dosage compensation, stable gene products yield type I spectra when  $N \leq 6$ , type II spectra when  $7 \leq N \leq 28$ , and type III spectra when  $N \geq 29$  (see section S4 for the proof).

When  $N$  is small and  $\eta$  is not restricted to small values, all the four types of power spectra can be observed; as  $\eta$  increases, the system undergoes three stochastic bifurcations from type I to type II, then to type III, and finally to type IV (Fig. 3(a)); when  $N$  is moderate, type I spectra cannot occur; as  $\eta$  increases, the system undergoes two stochastic bifurcations from type II to type III, and then to type IV (Fig. 3(b) and fig. S1(b)); when  $N$  is large, both type I and II spectra fail to occur; as  $\eta$  increases, the system undergoes only one stochastic bifurcation from type III to type IV (Fig. 3(c)). These results show that sustained oscillations tend to occur when (i)  $N$  is small but  $\eta$  is large (unstable gene products with large cell cycle duration variability) or (ii)  $N$  is moderate or large. In the latter case of moderate or low cell cycle duration variability, oscillations can be observed for all values

of  $\eta$  and hence are expected for both stable and unstable gene products.

The reason why unstable gene products yield more robust sustained oscillations can be understood as follows. For stable gene products, since active degradation is weak, molecule numbers increase approximately linearly with time and hence the time traces of gene product abundances appear like a sawtooth wave (upper panel of Fig. 3(d)). Due to cell cycle duration variability and noise due to partitioning at division, the expression levels at the end (or beginning) of each generation have large fluctuations. Such noise gives rise to the zero peak of the power spectrum, which explains why for small  $\eta$ , the spectra are of types I-III. In contrast, for unstable gene products, since the degradation rate is large, molecule numbers quickly reach a steady state and hence the time traces of gene product abundances appear like a square wave (lower panel of Fig. 3(d)). The two levels of the square wave correspond to the steady-state levels before and after replication. Once the expression level deviates from the steady-state value, the large degradation rate will help it relax to the steady state rapidly and hence the expression levels at the end (or beginning) of each generation have relatively small fluctuations. This explains why for unstable gene products with large  $\eta$ , the spectra are of type IV, i.e. do not have a peak at zero frequency, but rather the power is concentrated in a narrow bandwidth of frequencies close to the cell cycle frequency.



## Strong dosage compensation causes resonance-like behavior at intermediate gene product stability

The pattern of sustained oscillations is also influenced by dosage compensation. When dosage compensation is weak ( $\kappa$  is close to 2), the height of the off-zero spectral peak (relative to the zero peak) is an increasing function of  $\eta$  (Fig. 3(a)-(c)). In this case, the more unstable are gene products, the more capable they are of exhibiting regular stochastic oscillations. However, when dosage compensation is strong ( $\kappa$  is close to 1) and cell cycle duration variability is not too large, there is an optimal  $\eta$  such that the height is maximized (Fig. 3(b),(c)); oscillations are the most regular when  $\eta$  is around 15, which falls within the biological range of a typical mRNA in bacteria (table S1). The reason of this phenomenon can be understood as follows. Oscillations cannot be very regular for small  $\eta$  due to noise in cell cycle duration and partitioning at division, as discussed earlier. For large  $\eta$ , oscillations also cannot be very regular because when  $\kappa$  is close to 1, there is little change in the effective burst production rate across the cell cycle and hence the steady-state expression levels before and after replication (shown as the two levels of the square wave in Fig. 3(d)) will merge into one. As a result, when dosage compensation is strong, oscillations are the most regular at an intermediate  $\eta$  value.

## Fast promoter switching, leakage, and active gene state dominance enhance oscillation regularity

Finally, we investigate the effect of promoter switching on gene expression oscillations. Fig. 3(e),(f) illustrate the height of the off-zero spectral peak as a function of the promoter switching rates,  $u$  and  $v$ , in the absence and presence of promoter leakage, where promoter leakage means that there is a nonzero burst production rate when the gene is in the inactive state. For the case of no leakage, oscillations are manifest when the gene is mostly active, i.e.  $u \gg v$ , and fail to occur when the gene is mostly inactive, i.e.  $u \ll v$ , likely due to an exceptionally small gene expression mean. In the presence of leakage, however, we observe a strong oscillation when  $u \gg v$  and a weaker oscillation when  $u \ll v$ . Due to promoter leakage, even when the gene is mostly inactive, there is still a smaller but non-vanishing gene expression mean, which leads to the weaker oscillation observed. In addition, we find that the height of the off-zero peak is exceptionally small when promoter switching is very slow [42], i.e.  $u, v \ll 1$ , regardless of whether there is promoter leakage or not. We stress here that while Fig. 3(e),(f) display the results for the inheritance mechanism, similar results also hold for the reset mechanism.

## The power spectrum in asymmetrically dividing cells is strongly influenced by the single-cell tracking protocol

In some previous papers [1, 43], to track a cell lineage, one of the two daughters was randomly selected at cell division with probability  $1/2$ . However, for asymmetric cell division, the two daughters are of different sizes and another possible protocol is to track the smaller/larger daughter (such as the bud/mother cell in budding yeast) at cell division [3, 44]. Assuming well-mixing, the probability of a newborn cell receiving a gene product

molecule is equal to the ratio of the volume of the newborn to the volume of the mother cell and hence on average the smaller daughter receives fewer gene products than the larger one. We remind the reader that in our model the probability for a molecule to be allocated to one daughter is  $p \leq 1/2$  and the probability of being allocated to the other is  $q = 1 - p$ . Hence it follows that the daughter with allocation probability  $p < 1/2$  at division is the smaller daughter (Fig. 1(b)). Thus far we assumed that one of the two daughter cells is randomly tracked with probability  $1/2$  after cell division. Now we study two other tracking protocols, namely where we always follow the smaller or the larger daughter after cell division. For a cellular state  $(N, i, j)$  in effective cell cycle stage  $N$ , suppose that  $i$  and  $j$  record the gene states of the daughter copies that will be allocated to the smaller and larger daughter, respectively. If the smaller daughter is tracked after division, the cell will transition from a cellular state  $(N, i, j)$  in stage  $N$  to the cellular state  $(1, i)$  in stage 1 with rate  $a$ . If the larger daughter is tracked after division, the cell will transition from  $(N, i, j)$  to  $(1, j)$  with rate  $a$ . These considerations show that the transition diagram of cellular states should be modified as follows: if the smaller (larger) daughter is tracked at division, then the green (red) arrows in Fig. 1(d),(e) should be deleted and the transition rates of the red (green) arrows should be changed from  $a/2$  to  $a$  (fig. S3).

In this case, the autocorrelation function has the same form as in Eq. (1), where  $Q = (q_{\alpha\beta})$  is the generator matrix of cellular state transitions (see fig. S3 for the transition diagram which has some colored arrows) and  $W = (w_{\alpha\beta})$  is another matrix obtained from  $Q$  by replacing the rates of orange and red (green) arrows by  $pa$  ( $qa$ ) and subtracting  $d$  from the diagonal entries if the smaller (larger) daughter is tracked after cell division (see section S6 for the detailed expressions of each term).

To compare the three types of tracking protocols at cell division (tracking a random daughter, the smaller daughter, or the larger daughter), we illustrate the height of the off-zero spectral peak as a function of  $p$  for moderately unstable gene products with  $\eta = 2$  (Fig. 3(g)). Clearly, the three tracking protocols give rise to the same oscillatory behavior for symmetric division. However, for asymmetric division, the smaller daughter tracking protocol yields a much higher off-zero peak than the other two protocols, implying the most regular oscillations. The reason for this observation is as follows. For unstable gene products, the molecule number relaxes to the steady-state value rapidly and hence the expression levels just before division are roughly the same for the three protocols. However just after division, the smaller daughter receives less molecules than the larger one. Hence it follows that smaller daughter tracking yields larger amplitudes of oscillations in the time traces than larger daughter tracking (Fig. 3(h)). Compared to random tracking, smaller/larger daughter tracking leads to less noisy oscillations, presumably since the latter introduces a deterministic element in the tracking process. This explains why random tracking leads generally to less robust oscillations than the other two tracking protocols (Fig. 3(g)). Note that while increasing cell division asymmetry (decreasing  $p$ ) reduces the robustness of oscillations for random and larger daughter tracking, it leads to the opposite effect for small daughter tracking.

In fig. S4, we also investigate the differences between the

power spectra obtained from the three tracking protocols as a function of the parameter  $\eta$  which is a measure of gene product stability. We find that the differences increase with gene product stability. This can be explained as follows. The abrupt change in the number of molecules at division as we switch from a mother cell to a daughter cell is sensitive to the choice of protocol; these low frequency fluctuations contribute to the height of the zero peak which is sizeable for very stable products and fairly small for very unstable products (Fig. 3(a)-(c)). Hence the choice of protocol has most effect on the spectra of stable products.

### Experimental validation of the theory and its application to parameter inference

To test our theory, we apply it to study oscillations in single-cell gene expression data collected for *E. coli* in [2]. In this data set, the time course data of fluorescence intensity of a constitutively expressed yellow fluorescent protein was recorded every minute for 279 cell lineages over 70 generations using a mother machine under three different growth conditions (25°C, 27°C, and 37°C). At the three temperatures, there are a total of 65, 54, and 160 cell lineages measured, respectively. Based on such data, it is possible to estimate all the parameters involved in our model for each cell lineage. The medians of the estimated parameters for all cell lineages are listed in Table 2 and the distributions of the estimated parameters are given in figs. S5 and S6. In the following, we briefly describe the estimation procedures.

	25°C	27°C	37°C
$f$ (min <sup>-1</sup> )	0.0148	0.0187	0.0307
$T = 1/f$ (min)	67.5676	53.4759	32.5733
$N$	14	31	31
$w$	0.2000	0.1450	0.1224
$\kappa$	2.0621	2.0000	2.9412
$\beta$	42.0719	15.4767	13.0344
$\langle n \rangle$	99.2042	199.5930	287.8168
$\rho_{\text{eff}}$ (min <sup>-1</sup> )	0.1299	0.2303	0.6259
$\kappa\rho_{\text{eff}}$ (min <sup>-1</sup> )	0.2232	0.5319	1.3809
$B$	4.7718	5.2099	4.8606

Table 2. **The medians of the estimated parameters for all cell lineages at three different temperatures.** The distributions of the estimated parameters can be found in figs. S5 and S6. The medians rather than the means are reported here since the estimation of the former is more robust than that of the latter, with respect to outliers.

Since the protein is constitutively expressed [2], there is no promoter switching, i.e.  $u = v = 0$ . Moreover, since the protein is very stable, it is reasonable to assume that it has negligible degradation, i.e.  $d = \eta = 0$  [2, 43]. Based on the time course data, we estimated the average power spectrum over all cell lineages at each temperature by means of the Wiener-Khinchin theorem (Fig. 4(a)), where we have normalized the spectrum so that  $G(0) = 1$ . Clearly, the average power spectra are of type II for all the three growth conditions. As the temperature increases, the position and height of the off-zero spectral peak both increase,

implying more robust oscillations. The position of the off-zero peak is very close to the cell cycle frequency  $f$ , which can be easily estimated from the data of doubling times (Table 2).

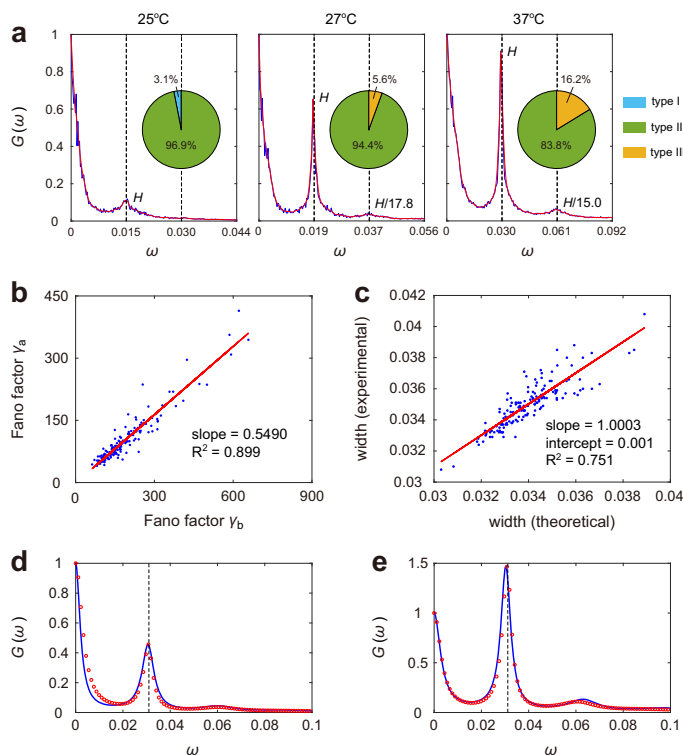
We find that the doubling time data for all cell lineages are well fitted by an Erlang distribution; the parameter  $N$  can then be estimated as the inverse of the coefficient of variation squared of this distribution. The medians of the estimated  $N$  for the three growth conditions are 14, 31, and 31, respectively. It can be seen that cells cultured at 27°C and 37°C have much smaller cell cycle duration variability than those cultured at 25°C. Our theory demonstrates that smaller cell cycle duration variability gives rise to a higher off-zero spectral peak (Fig. 2(c)). This explains why the off-zero peak is significantly higher at higher temperatures (Fig. 4(a)). Our theoretical result also predicts that when  $N$  is large, the power spectrum has peaks at higher-order harmonics of the cell cycle frequency (Fig. 2(f)). This is in excellent agreement with our data analysis, which shows that the average power spectra at 27°C and 37°C have an apparent peak at the second harmonic frequency. Interestingly, the ratio of the heights of the spectral peaks at  $f$  and  $2f$  is estimated to be 17.8 and 15.0 for the two temperatures, both of which are very close to the theoretical value of  $2^4 = 16$  predicted by our theory for stable gene products when  $w$  is small.

To perform a more detailed analysis, we also estimated the power spectrum for each cell lineage by fitting the time course data with an autoregressive (AR) model, which is a standard model in time series analysis [45], with the order of the AR model being determined by minimizing the Akaike information criterion (see section S7 for details). Our theoretical results show that for stable gene products, only type I, II, and III spectra can be observed (Fig. 3(a)-(c)). This is in full agreement with our data analysis with the percentages of the three types of power spectra being illustrated by the pie charts in Fig. 4(a). Clearly, type II spectra are dominant for all the three growth conditions. For cells at 25°C, only types I and II are observed; for cells at higher temperatures, only types II and III are observed. The percentage of type III spectra is higher for cells at 37°C than cells at 27°C.

Since the burst production rate increases from  $\rho_{\text{eff}}$  to  $\kappa\rho_{\text{eff}}$  upon replication, we can fit the data (recorded per minute) between two cell division times by the following mean-field approximation:

$$\hat{n}(t+1) - \hat{n}(t) = \begin{cases} \rho_{\text{eff}}, & \text{if } t - T_k \leq w(T_{k+1} - T_k), \\ \kappa\rho_{\text{eff}}, & \text{if } t - T_k > w(T_{k+1} - T_k), \end{cases} \quad (10)$$

where  $T_k$  is the  $k$ th cell division time and  $t$  is an arbitrary time point (in minute) between two consecutive division times  $T_k$  and  $T_{k+1}$ . Here we use a piecewise linear function to approximate the time series of expression levels between two division events. The first (second) part of the piecewise linear function approximates the time series before (after) replication and thus should have an average slope of  $\rho_{\text{eff}}$  ( $\kappa\rho_{\text{eff}}$ ). By fitting the time course data  $n(t)$  with the approximation  $\hat{n}(t)$  given in Eq. (10), we obtained the least squares estimates of  $w$  and  $\kappa$  for each cell lineage by minimizing the distance  $\sum_{t=0}^{M-1} [n(t) - \hat{n}(t)]^2$  between the two, where  $M$  is the number of time points for each cell lineage (see section S8 for details). The medians of the estimated  $w$  for the three growth conditions are 0.20, 0.15, and 0.12, respectively, which decrease with temperature. This is possibly because at



**Fig. 4. Analysis of single-cell time course data of fluorescence intensities in *E. coli* published in [2].** (a) Average power spectra over all cell lineages at three different temperatures (blue curves) and their smoothed approximations (red curves) using the Gaussian filter. The average spectra are estimated by means of the Wiener-Khinchin theorem, where 65, 54, and 160 cell lineages are averaged over, respectively, for the three growth conditions. The power spectrum for each cell lineage is also estimated by fitting the time course data with an AR model. The pie charts display the percentages of various types of power spectra for all cell lineages. (b) Comparison between the Fano factor  $\gamma_b$  of the fluorescence intensities just before division and the Fano factor  $\gamma_a$  of the fluorescence intensities just after division for all cell lineages at 37°C. (c) Comparison between the widths of the experimental power spectra obtained using the AR model technique and the theoretical power spectra determined using the estimated parameters for all cell lineages at 37°C. (d),(e) Comparison between the experimental (blue curve) and theoretical (red circles) power spectra for two typical cell lineages at 37°C.

higher temperatures, the growth rate of *E. coli* cells is faster and thus replication would occur more continuously, leading to a shift in  $w$  towards zero [46]. In addition, the medians of the estimated  $\kappa$  are close to 2 for all the three growth conditions (Table 2), implying weak dosage compensation.

The remaining parameters to be estimated are  $\rho_{\text{eff}}$  and  $B$ , where  $B$  is the mean translational burst size, i.e. the average number of protein molecules produced per mRNA lifetime. To estimate them, recall that  $\rho_{\text{eff}}B$  represents the mean number of protein molecules produced per unit time. However, what was measured in the data set is the fluorescence intensity of protein molecules, instead of the real copy number. Hence, it is crucial to determine the proportionality constant between fluorescence intensities and copy numbers. To do this, we note that under the assumption of symmetric binomial partitioning at division,

the expression levels just before and just after a particular cell division time are coupled by (see section S9 for the proof)

$$\gamma_a = \frac{\gamma_b}{2} + \frac{\beta}{2}, \quad (11)$$

where  $\gamma_b$  ( $\gamma_a$ ) is the Fano factor (the variance divided by the mean) of the fluorescence intensities just before (after) division and  $\beta$  is the fluorescence intensity per protein molecule. Note that both  $\gamma_b$  and  $\gamma_a$  can be estimated for each cell lineage. To test this relationship, we show  $\gamma_a$  as a function of  $\gamma_b$  for all cell lineages at the three temperatures (see Fig. 4(b) for cells at 37°C and fig. S7(a) for cells at lower temperatures), from which we observed a strong linear relationship with a high  $R^2$  and a slope close to 0.5. Then the proportionality constant  $\beta$  for a given temperature can be estimated as  $\beta = 2\langle\gamma_a\rangle - \langle\gamma_b\rangle$ , where the angled brackets denote the sample means over all cell lineages at that temperature. The estimated  $\beta$  for the three growth conditions are 42.1, 15.5, and 13.0, respectively. In particular, our analysis shows that the fluorescence intensity per protein molecule decreases with temperature, which is likely because lower temperatures are more conducive to the correct folding of the fluorescent protein [47].

From the fluorescence intensity data and the inferred  $\beta$ , it is easy to estimate the gene expression mean  $\langle n \rangle$  for each cell lineage (Table 2). Note that cells at 27°C and 37°C have similar cell cycle duration variability with  $N = 31$  but the latter yield a higher off-zero spectral peak. This is because cells at 37°C have a larger gene expression mean, which augments the height of the off-zero spectral peak. Moreover, since  $\rho_{\text{eff}}$ ,  $B$ ,  $w$ ,  $\kappa$ , and  $\langle n \rangle$  are related by Eq. (3) and we have estimated  $w$  and  $\kappa$ , we can obtain the estimate of  $\rho_{\text{eff}}B$ . Finally, we estimated  $\rho_{\text{eff}}$  and  $B$  separately by equating the heights of the off-zero peak of the experimental power spectrum obtained using the AR model technique and the theoretical power spectrum determined by Eq. (2) using all the estimated parameters. From Table 2, we can see that the mean burst frequency (average number of bursts produced per unit time)  $\rho_{\text{eff}}$  increases with temperature approximately linearly but the mean burst size  $B$  is not temperature dependent.

Thus far we have shown how, by means of the theoretical expressions of the power spectrum and the gene expression mean, we can estimate all the model parameters for each cell lineage from the time course data. The accuracy of these estimates is verified in two different ways. First, in Fig. 4(d),(e), we show a good agreement between the experimental power spectrum and the theoretical power spectrum evaluated using the inferred model parameters for two typical cell lineages. As a second test of the accuracy of parameter inference, we compare the widths of the experimental and theoretical power spectra for all cell lineages (see Fig. 4(c) for cells at 37°C and fig. S7(b) for cells at lower temperatures). It can be seen from Fig. 4(c) that the widths for the two spectra show a strong linear relationship with a slope close to 1, a negligible intercept, and an  $R^2$  of 0.75. We stress that while the model parameters were not estimated from the full spectrum curve but only from the height of the off-zero peak, the full theoretical spectrum matches the experimental spectrum reasonably well.



## Discussion

In this work, we have investigated the frequency decomposition of the copy number fluctuations of a gene product (mRNA or protein) within a cell lineage, by deriving expressions for the power spectrum of fluctuations in a detailed stochastic model of gene expression. This model takes into account the salient experimental observations of intracellular dynamics including promoter switching, transcriptional and translational bursting, variability in the duration of the cell cycle, variability in the gene copy number due to replication, the copying or resetting of the gene state during replication, gene dosage compensation, and partitioning of molecules due to symmetric or asymmetric cell division.

Our study differs from previous ones in four main respects: (i) our model is more grounded in biological reality than other models in the literature due to the large number of subcellular and cellular processes that it accounts for, as mentioned above; (ii) a number of studies have derived the distribution of molecule numbers or more commonly the moments for models that have some similarity to ours [15–19, 48]; however in contrast, here we derive expressions for the power spectrum which provide information about the frequency content of lineage data. This type of analysis has previously been only reported for models of non-growing cells [20, 49]; (iii) rather than a parameter inference based on the matching of the moments or the distribution of molecule numbers calculated from the data to those of a stochastic model [50], we showcase a power spectrum based parameter inference method; (iv) our model takes into account the details of the experimental protocol used for tracking cells across a lineage.

Our novel theory provides expressions for the height and width of the off-zero spectral peak (with its position close to the cell cycle frequency) as a function of all rate parameters in the model. The ratio of the height of this peak to the power at zero frequency provides a means to classify the power spectra into two main types: (i) the spectra with the ratio less than 1 (types I and II) and (ii) the spectra with the ratio greater than 1 (types III and IV). The periodicity in molecule number variation induced by cell division dominates over subcellular noise for (ii) while the reverse is the case for (i). Type I is further differentiated from type II by specifying that in the former, there is only one peak at zero frequency whereas in the latter there is a dominant peak at zero frequency and a lesser one at approximately the cell cycle frequency. Similarly type III is further differentiated from type IV by specifying that in the former, there is a dominant peak at approximately the cell cycle frequency and a lesser one at zero frequency while in the latter there is only one peak at approximately the cell cycle frequency. The theory also predicts that while the spectra for fast decaying (unstable) gene products can be of all four types, the spectra for slow decaying (stable) gene products can only be of types I-III.

Our analysis of experimental data for *E. coli* shows that the type of spectra of single-cell trajectories depends on the temperature: lower temperatures favour type I and II spectra while higher temperatures favour type II and III spectra. Overall the most common spectrum was Type II implying that for many cells, the “forces” inducing periodicity of molecule numbers

are typically slightly less strong than the “forces” inducing subcellular noise; the strength of the latter increases with decreasing temperature. None of the 279 cell lineages had a type IV spectrum for protein fluctuations, in accordance with the theoretical result that stable gene products cannot display such a spectrum.

Our theory made a number of other testable predictions: (i) the height of the off-zero spectral peak (relative to that of the zero peak) increases with decreasing cell cycle duration variability and increasing mean expression levels, while the width of the off-zero peak is proportional to the cell cycle frequency and to the cell cycle duration variability; (ii) if the cell cycle duration variability is small enough then the spectra display peaks at higher-order harmonics of the cell cycle frequency. For stable proteins, the height of the spectral peak at the second harmonic (twice the cell cycle frequency) is 16 times larger than that at the cell cycle frequency for early or late replication. Both predictions were confirmed by analysis of lineage data for *E. coli*.

Furthermore our theory made numerous other predictions which require the design of new experiments and which cannot be tested using current data. A brief summary of these predictions is as follows: (i) the dependence of the height of the off-zero spectral peak on parameters is very different for gene products that are stable, e.g. most proteins, compared to gene products that are unstable, e.g. most mRNAs. For stable gene products, the height of the off-zero peak decreases with increasing mean burst size, decreasing mean molecule number, and increasing asymmetry of cell division. Furthermore the height is maximal for replication occurring (almost) a third of the way through the cycle. In contrast for unstable gene products, the height of the off-zero peak is maximal for replication occurring midway through the cycle but is (almost) independent of asymmetric division, gene expression mean, random bursting, and dosage compensation. Independent of stability, fast promoter switching enhances the height of the off-zero peak (relative to that of the zero peak) whereas slow switching has the opposite effect; (ii) the strength of gene dosage compensation is reflected in how the spectrum varies with gene product stability. For weak dosage compensation, molecules which have a larger decay rate also exhibit a spectrum with a higher off-zero peak. For strong dosage compensation, molecules which have a decay rate that is neither too large nor too small are the ones which exhibit the most pronounced off-zero peaks; (iii) the power spectrum in asymmetrically dividing cells is strongly influenced by the choice of single-cell tracking protocol. In particular the spectra obtained from following the smaller daughter after division have a significantly higher off-zero peak than those obtained from following the larger daughter or from random tracking. This last prediction is likely the easiest of the three to check by redoing the cell tracking analysis in budding yeast which displays asymmetric cell division. Predictions (i) and (ii) ideally require the ability to obtain live-cell fluorescence data for mRNA over tens of generations and for different types of mRNA with widely varying decay rates, which are presently not easy to obtain. The development of such methods, particularly those with minimal perturbation of transcription and translation, is still in an active area of research [4].

We have also showcased the use of the power spectrum to



determine the values of all the rate parameters in our model from cell lineage data. Two strengths of our inference procedure are (i) we do not need an experimental determination of the relationship between total fluorescence intensity and molecule number; this is determined automatically from the relationship between fluctuations in the fluorescence intensity just before and just after division; (ii) our analysis takes into account noise due to partitioning and variability in the cell cycle duration which has recently been shown to be crucial to obtain an accurate estimation of the burst frequency and the burst size [19]. Our analysis confirms there is no or very weak dosage compensation in *E. coli* in agreement with previous studies [46], i.e. at replication, the expression increases roughly two-fold, in agreement with an expected doubling of the number of gene copies. Furthermore we find that for constitutive expression, while the translational burst frequency increases approximately linearly with temperature, the translational burst size is temperature invariant. Previous studies were conducted at one temperature [5] and hence could not quantify the thermal dependence of gene expression parameters.

Concluding we have performed an exact frequency analysis of mRNA and protein fluctuations in a detailed model of gene circuit elements central to gene expression control. As we have shown, there is a wealth of information about subcellular processes hidden in the frequency content of mRNA and protein fluctuations within a cell lineage, and we hope our results will further stimulate work in this area.

## Methods

### Oscillations for diploid cells

Suppose that we have computed the analytical solutions of the autocorrelation function  $R(t)$  and the power spectrum  $G(\xi)$  for haploid cells. For diploid cells, we assume that the two alleles act independently of each other. Under the assumption of allelic expression independence, let  $n(t)$  and  $n'(t)$  denote the expression levels of the two alleles at time  $t$ , which are independent and identically distributed. Then for diploid cells, the autocorrelation function is given by

$$\begin{aligned} R_{\text{diploid}}(t) &= \text{Cov}_{\text{ss}}(n(0) + n'(0), n(t) + n'(t)) \\ &= 2\text{Cov}_{\text{ss}}(n(0), n(t)) = 2R(t), \end{aligned}$$

and thus the power spectrum is given by

$$G_{\text{diploid}}(\xi) = 2G(\xi).$$

Since the two functions only differ by a constant for haploid and diploid cells, they lead to the same oscillatory behavior. Thus we only need to focus on haploid cells in what follows.

### Master equation describing cell-cycle dependent stochastic gene expression

Recall that before gene replication, each cellular state can be represented as  $\alpha = (r, i)$ , where  $r$  is the cell cycle stage and  $i$  is the gene state of the mother copy; after gene replication, each cellular state can be represented as  $\alpha = (r, i, j)$ , where  $i$  and  $j$  are the gene states of the two daughter copies. Let  $\pi_\alpha$  denote

the probability of observing cellular state  $\alpha$ . The evolution of the cellular state dynamics is then governed by the following master equation (see section S1 for the master equation written using the original model parameters):

$$\dot{\pi}_\alpha = \sum_{\beta \neq \alpha} (\pi_\beta q_{\beta\alpha} - \pi_\alpha q_{\alpha\beta}),$$

where  $q_{\alpha\beta}$  is the transition rate from cellular state  $\alpha$  to cellular state  $\beta$ . The transition diagrams for all cellular states under the inheritance and reset mechanisms are illustrated in Fig. 1(d),(e), respectively. The microstate of the gene of interest can be represented by the ordered pair  $(\alpha, n)$ , where  $\alpha$  is the cellular state and  $n$  is the copy number of the gene product. At division, the cell will transition from some microstate  $(N, i, j, n)$  in cell cycle stage  $N$  to another microstate  $(1, i, m)$  or  $(1, j, m)$  in cell cycle stage 1 with  $m \leq n$ . To model asymmetric cell division, we need to distinguish the following two types of transitions between cellular states:

$$\begin{aligned} D_1 &= \{(\alpha, \beta) : \alpha = (N, i, j), \beta = (1, i)\}, \\ D_2 &= \{(\alpha, \beta) : \alpha = (N, i, j), \beta = (1, j)\}, \end{aligned}$$

where  $\alpha$  is a cellular state in cell cycle stage  $N$  and  $\beta$  is a cellular state in cell cycle stage 1. For any pair of cellular states  $(\alpha, \beta)$  in  $D_1$  or  $D_2$ , the transition rate from  $\alpha$  to  $\beta$  is given as follows:

$$\begin{aligned} q_{\alpha\beta} &= \frac{a}{2}, & (\alpha, \beta) \in D_1 \setminus D_2, \\ q_{\alpha\beta} &= \frac{a}{2}, & (\alpha, \beta) \in D_2 \setminus D_1, \\ q_{\alpha\beta} &= a, & (\alpha, \beta) \in D_1 \cap D_2. \end{aligned}$$

The cellular state transitions in  $D_1 \setminus D_2$ ,  $D_2 \setminus D_1$ , and  $D_1 \cap D_2$  are marked by the red, green, and orange arrows in Fig. 1(d),(e), respectively. Due to asymmetric binomial partitioning at cell division, the transition rate from microstate  $(\alpha, n)$  to  $(\beta, m)$  is given as follows:

$$\begin{aligned} (\alpha, n) &\xrightarrow{\frac{a}{2} C_{n,m} p^m q^{n-m}} (\beta, m), & (\alpha, \beta) \in D_1 \setminus D_2, \\ (\alpha, n) &\xrightarrow{\frac{a}{2} C_{n,m} q^m p^{n-m}} (\beta, m), & (\alpha, \beta) \in D_2 \setminus D_1, \\ (\alpha, n) &\xrightarrow{\frac{a}{2} C_{n,m} (p^m q^{n-m} + q^m p^{n-m})} (\beta, m), & (\alpha, \beta) \in D_1 \cap D_2, \end{aligned}$$

where  $C_{n,m} = n!/m!(n-m)!$  is the combinatorial number and  $p$  is the probability of asymmetric binomial partitioning. Let  $p_{\alpha,n}$  denote the probability of being in microstate  $(\alpha, n)$ . Then the evolution of stochastic gene expression dynamics is governed by

the following master equation:

$$\begin{aligned}\dot{p}_{\alpha,n} = & \sum_{m=0}^{n-1} \rho_{\alpha} \mu_{n-m} p_{\alpha,m} + (n+1)d p_{\alpha,n+1} \\ & + \sum_{\substack{\beta \neq \alpha \\ (\beta, \alpha) \notin D_1 \cup D_2}} q_{\beta\alpha} p_{\beta,n} \\ & + \frac{a}{2} \sum_{\substack{\beta \neq \alpha \\ (\beta, \alpha) \in D_1}} \sum_{m=n}^{\infty} C_{m,n} p^n q^{m-n} p_{\beta,m} \\ & + \frac{a}{2} \sum_{\substack{\beta \neq \alpha \\ (\beta, \alpha) \in D_2}} \sum_{m=n}^{\infty} C_{m,n} q^n p^{m-n} p_{\beta,m} \\ & - \left( \rho_{\alpha} \sum_{m=1}^{\infty} \mu_m + nd + \sum_{\beta \neq \alpha} q_{\alpha\beta} \right) p_{\alpha,n},\end{aligned}$$

where  $\rho_{\alpha}$  is the total burst production rate (for the mother copy or for the two daughter copies) in cellular state  $\alpha$ ,  $\mu_n$  is the burst size distribution of the gene product,  $d$  is the degradation rate of the gene product, and  $q_{\alpha\beta}$  is the transition rate from cellular state  $\alpha$  to cellular state  $\beta$ . Since there are only two gene states, there are only five possible values for  $\rho_{\alpha}$ , which are given by

$$\rho_{\alpha} = \begin{cases} \rho_0, & \alpha = (r, 0), \\ \rho_1, & \alpha = (r, 1), \\ 2\rho_0, & \alpha = (r, 0, 0), \\ 2\rho_1, & \alpha = (r, 1, 1), \\ \rho_0 + \rho_1, & \alpha = (r, 0, 1), (r, 1, 0), \end{cases}$$

where  $\rho_0$  and  $\rho_1$  are the burst product rate (for a single gene copy) in the inactive and active gene states, respectively.

### Analytical solutions for the autocorrelation function and the power spectrum

Here we only consider the bursty case. The results in the non-bursty case can be derived in the same way (see section S2 for details). Let  $\alpha(t)$  denote the state of the gene and let  $n(t)$  denote the copy number of the gene product in an individual cell at time  $t$ , respectively. To proceed, let

$$\pi_{\alpha}(t) = \sum_{n=0}^{\infty} p_{\alpha,n} = \mathbb{P}(\alpha(t) = \alpha),$$

$$m_{1\alpha}(t) = \sum_{n=0}^{\infty} n p_{\alpha,n} = \mathbb{E}n(t) I_{\{\alpha(t)=\alpha\}},$$

$$m_{2\alpha}(t) = \sum_{n=0}^{\infty} n(n-1) p_{\alpha,n} = \mathbb{E}n(t)(n(t)-1) I_{\{\alpha(t)=\alpha\}},$$

be the first three unnormalized factorial moments of gene product abundances in cellular state  $\alpha$ , where  $I_A$  is the indicator function of the set  $A$ . Then the row vectors  $\pi = (\pi_{\alpha})$ ,  $m_1 = (m_{1\alpha})$ , and  $m_2 = (m_{2\alpha})$  satisfy the following differential equations (see section S2 for the proof):

$$\begin{aligned}\dot{\pi}(t) &= \pi(t)Q, \\ \dot{m}_1(t) &= m_1(t)W_{11} + B\pi(t)S, \\ \dot{m}_2(t) &= m_2(t)W_{22} + 2Bm_1(t)S + 2B^2\pi(t)S,\end{aligned}\tag{12}$$

where  $S = \text{diag}(\rho_{\alpha})$  is the diagonal matrix whose diagonal entries are the burst production rates in all cellular states and  $W_{11}$  and  $W_{22}$  are two matrices defined as

$$\begin{aligned}W_{11} &= Q^{(1)} + pQ^{(2)} + qQ^{(3)} - dI, \\ W_{22} &= Q^{(1)} + p^2Q^{(2)} + q^2Q^{(3)} - 2dI,\end{aligned}$$

where  $Q^{(i)} = (q_{\alpha\beta}^{(i)})$  are three matrices defined as

$$\begin{aligned}q_{\alpha\beta}^{(1)} &= q_{\alpha\beta} I_{(\alpha,\beta) \notin D_1 \cup D_2}, \\ q_{\alpha\beta}^{(2)} &= \frac{a}{2} I_{(\alpha,\beta) \in D_1}, \quad q_{\alpha\beta}^{(3)} = \frac{a}{2} I_{(\alpha,\beta) \in D_2}.\end{aligned}$$

From Eq. (12), at the steady state, the first and second factorial moments are given by

$$\begin{aligned}m_1 &= -B\pi S W_{11}^{-1}, \\ m_2 &= -2(Bm_1 + B^2m_0)S W_{22}^{-1}.\end{aligned}\tag{13}$$

Therefore, the steady-state gene expression mean is given by

$$\langle n \rangle = m_1 \mathbb{1} = -B\pi S W_{11}^{-1} \mathbb{1},$$

where  $\mathbb{1}$  denotes the column vector whose components are all 1.

Since Eq. (12) is a set of linear differential equations, its time-dependent solution is given by

$$m_1(t) = m_1(0)e^{W_{11}t} + B \int_0^t \pi(0)e^{Qs} S e^{W_{11}(t-s)} ds.$$

From now on, we assume that the system has reached the steady state. Given the initial cellular state  $\alpha(0) = \alpha$  and initial copy number  $n(0) = n$  of the gene product, it follows that

$$\mathbb{E}n(t) = m_1(t)\mathbb{1} = n e_{\alpha} e^{W_{11}t} \mathbb{1} + B \int_0^t e_{\alpha} e^{Qs} S e^{W_{11}(t-s)} \mathbb{1} ds,$$

where  $e_{\alpha}$  denotes the row vector whose  $\alpha$ th component is 1 and all other components are 0. This clearly shows that

$$\begin{aligned}\mathbb{E}[n(t)|\alpha(0), n(0)] \\ = n(0)e_{\alpha(0)} e^{W_{11}t} \mathbb{1} + B \int_0^t e_{\alpha(0)} e^{Qs} S e^{W_{11}(t-s)} \mathbb{1} ds.\end{aligned}$$

Therefore, at the steady state, we have (see section S2 for the proof)

$$\begin{aligned}\mathbb{E}n(0)n(t) &= \sum_{\alpha} \mathbb{E}n(0) I_{\{\alpha(0)=\alpha\}} \mathbb{E}[n(t)|\alpha(0), n(0)] \\ &= (m_1 + m_2) e^{W_{11}t} \mathbb{1} + B \int_0^t m_1 e^{Qs} S e^{W_{11}(t-s)} \mathbb{1} ds,\end{aligned}$$

where  $m_1$  and  $m_2$  are the steady-state first and second factorial moments given in Eq. (13). Since the autocorrelation function is defined as  $R(t) = \mathbb{E}n(0)n(t) - \mathbb{E}n(0)\mathbb{E}n(t)$ , we finally obtain the explicit expression of the autocorrelation function, which is given in Eq. (1).

We next express  $R(t)$  more explicitly in terms of simple functions. To this end, we assume that all eigenvalues of  $Q$ , as well as all eigenvalues of  $W_{11}$ , are mutually distinct (in fact, any matrix can be approximated by such matrices to any degree

of accuracy). Let  $\lambda_0, \lambda_1, \dots, \lambda_{K-1}$  be all eigenvalues of  $Q$  and let  $\lambda_K, \lambda_{K+1}, \dots, \lambda_{2K-1}$  be all eigenvalues of  $W_{11}$ . By the Perron-Frobenius theorem, one eigenvalue of the generator matrix  $Q$  must be zero and other eigenvalues must have negative real parts. Similarly, all eigenvalues of  $W_{11}$  must have negative real parts, which implies that

$$\lambda_0 = 0, \quad \text{Re}(\lambda_1), \dots, \text{Re}(\lambda_{2K-1}) < 0,$$

where  $\text{Re}(x)$  denotes the real part of  $x$ . From Eq. (1), it can be deduced that the correlation function is a linear combination of exponential functions:

$$R(t) = \sum_{k=1}^{2K-1} u_k e^{\lambda_k t}, \quad (14)$$

where  $u_k$  are suitable constants. Taking the derivatives on both sides of Eqs. (1) and (14) and evaluating at  $t = 0$  yield

$$\sum_{k=1}^{2K-1} u_k \lambda_k^n = v_n,$$

where

$$v_n = (m_1 + m_2) W_{11}^n \mathbb{1} + B \sum_{k=1}^n m_1 Q^{n-k} S W_{11}^{k-1} \mathbb{1}.$$

This equation can be rewritten in matrix form as

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{2K-1} \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_{2K-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{2K-1} & \lambda_2^{2K-1} & \dots & \lambda_{2K-1}^{2K-1} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{2K-1} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{2K-1} \end{pmatrix},$$

where the matrix  $V$  on the left-hand side is a Vandermonde matrix. This shows that

$$\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{2K-1} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{2K-1} \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_{2K-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{2K-1} & \lambda_2^{2K-1} & \dots & \lambda_{2K-1}^{2K-1} \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{2K-1} \end{pmatrix}.$$

To proceed, recall that the inverse of the Vandermonde matrix is given by  $V^{-1} = (b_{kl})$ , where

$$b_{kl} = (-1)^{l-1} \left[ \frac{\prod_{\substack{1 \leq m_1 < \dots < m_{2K-1-l} \leq 2K-1 \\ m_1, \dots, m_{2K-1-l} \neq k}} \lambda_{m_1} \dots \lambda_{m_{2K-1-l}}}{\lambda_k \prod_{\substack{1 \leq m \leq 2K-1 \\ m \neq k}} (\lambda_m - \lambda_k)} \right].$$

Therefore, the coefficients  $u_k$  are given by

$$u_k = \sum_{l=1}^{2K-1} b_{kl} v_l.$$

Finally, the autocorrelation function can be written as

$$R(t) = \sum_{k=1}^{2K-1} u_k e^{\lambda_k t} = \sum_{k,l=1}^{2K-1} b_{kl} v_l e^{\lambda_k t}.$$

Taking the Fourier transform of the autocorrelation function gives the power spectrum, which is given in Eq. (2). The simplification of the gene expression mean and the power spectrum in the regime of fast promoter switching can be found in sections S2-S5.

## Power spectrum for stable gene products

For stable gene products with  $\eta \ll 1$ , the power spectrum can be simplified as in Eq. (5), where

$$A = 2(\ln 2)^2 J \langle n \rangle \left( \gamma + \frac{2}{3N} \right) + \frac{1}{(\ln 2)^2} + \frac{2B}{3 \ln 2} \left( \gamma - 2 + \frac{3}{\ln 2} \right),$$

if gene expression is bursty,

$$A = 2(\ln 2)^2 J \langle n \rangle \left( \gamma + \frac{2}{3N} \right) + \frac{1}{(\ln 2)^2},$$

if gene expression is non-bursty, and

$$H_k = \left[ w + \kappa(1-w) + \frac{(\kappa-1) \sin(2k w \pi)}{2k \pi} \right]^2 + \frac{(\kappa-1)^2 \sin^4(k w \pi)}{k^2 \pi^2},$$

is a function of  $\kappa$  and  $w$ , where  $J$  is given in Eq. (8).

## Power spectrum for unstable gene products

For unstable gene products with  $\eta \ll 1$ , when cell cycle duration variability is not too large, the power spectrum can be written more explicitly as (see section S5 for the proof)

$$G(\xi) \approx \frac{2(\kappa-1)^2 \rho_{\text{eff}}^2 B^2 N \sin^2(k w \pi)}{\eta^2 f} \sum_{k=1}^{[N/2]} G_k(\xi), \quad (15)$$

where

$$G_k(\xi) = \frac{2k^2 \pi^2 f^2 + (\pi^2 \xi^2 - a^2 \cos \theta_k \sin^2 \theta_k / 2)}{4k^6 \pi^6 f^4 + N^2 (\pi^2 \xi^2 - a^2 \cos \theta_k \sin^2 \theta_k / 2)^2},$$

with  $\theta_k = 2k\pi/N$ . Here the function  $G_1(\xi)$  controls the first off-zero spectral peak and the functions  $G_k(\xi), k \geq 2$  control the peaks at higher-order harmonic frequencies. In analogy to the case of stable gene products, the position of the off-zero peak is given by  $\xi = (a/\pi) \cos \theta_1 \sin(\theta_1/2) < a\theta_1/2\pi = a/N$ , which is smaller than the cell cycle frequency  $f$ . When  $N \gg 1$ , the peak position is approximately equal to  $f$  since  $\sin \theta \approx \theta$  and  $\cos \theta \approx 1$  when  $\theta$  is small. Moreover, the width of the off-zero peak is given by  $D = 2\pi f/N$  and the height of the off-zero peak is given in Eq. (9). From Eq. (15), the height and width of the spectral peak at the  $k$ th harmonic frequency are given by  $\sin^2(k w \pi) / \sin^2(w \pi) k^4 H$  and  $k^2 D$ , respectively. In particular, when cell cycle duration variability is small, the height of the spectral peak at  $f$  is  $2^4 \sin^2(w \pi) / \sin^2(2w \pi)$  times greater than that at  $2f$  and  $3^4 \sin^2(w \pi) / \sin^2(3w \pi)$  times greater than that at  $3f$ . Moreover, the width of the spectral peak at  $f$  is  $2^2 = 4$  times lesser than that at  $2f$  and  $3^2 = 9$  times lesser than that at  $3f$ .

## References

- [1] Brenner, N. *et al.* Single-cell protein dynamics reproduce universal fluctuations in cell populations. *European Physical Journal E* **38**, 102 (2015).

- [2] Tanouchi, Y. *et al.* Long-term growth data of *Escherichia coli* at a single-cell level. *Scientific Data* **4**, 1–5 (2017).
- [3] Crane, M. M., Clark, I. B., Bakker, E., Smith, S. & Swain, P. S. A microfluidic system for studying ageing and dynamic single-cell responses in budding yeast. *PloS one* **9**, e100042 (2014).
- [4] Chen, X. *et al.* Visualizing RNA dynamics in live cells with bright and stable fluorescent RNAs. *Nature Biotechnology* **37**, 1287–1293 (2019).
- [5] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- [6] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
- [7] Peterson, J. R., Cole, J. A., Fei, J., Ha, T. & Luthey-Schulten, Z. A. Effects of DNA replication on mRNA noise. *Proceedings of the National Academy of Sciences* **112**, 15886–15891 (2015).
- [8] Skinner, S. O. *et al.* Single-cell analysis of transcription kinetics across the cell cycle. *Elife* **5**, e12175 (2016).
- [9] Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular cell* **58**, 339–352 (2015).
- [10] Tyson, D. R., Garbett, S. P., Frick, P. L. & Quaranta, V. Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. *Nature methods* **9**, 923–928 (2012).
- [11] Hölzel, M. *et al.* Myc/Max/Mad regulate the frequency but not the duration of productive cell cycles. *EMBO reports* **2**, 1125–1132 (2001).
- [12] Huh, D. & Paulsson, J. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences* **108**, 15004–15009 (2011).
- [13] Huh, D. & Paulsson, J. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* **43**, 95 (2011).
- [14] Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. USA* **105**, 17256–17261 (2008).
- [15] Soltani, M., Vargas-Garcia, C. A., Antunes, D. & Singh, A. Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLoS computational biology* **12**, e1004972 (2016).
- [16] Soltani, M. & Singh, A. Effects of cell-cycle-dependent expression on random fluctuations in protein levels. *Royal Society open science* **3**, 160578 (2016).
- [17] Perez-Carrasco, R., Beentjes, C. & Grima, R. Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. *J. R. Soc. Interface* **17**, 20200360 (2020).
- [18] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. USA* **117**, 4682–4692 (2020).
- [19] Beentjes, C. H., Perez-Carrasco, R. & Grima, R. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Phys. Rev. E* **101**, 032403 (2020).
- [20] Simpson, M. L., Cox, C. D. & Sayler, G. S. Frequency domain analysis of noise in autoregulated gene circuits. *Proceedings of the National Academy of Sciences* **100**, 4551–4556 (2003).
- [21] Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
- [22] Bokes, P., King, J. R., Wood, A. T. & Loose, M. Multiscale stochastic modelling of gene expression. *J. Math. Biol.* **65**, 493–520 (2012).
- [23] Jia, C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. *Phys. Rev. E* **96**, 032402 (2017).
- [24] Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* **99**, 5860–5865 (2002).
- [25] Golubev, A. 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**, 203–217 (2016).
- [26] Yates, C. A., Ford, M. J. & Mort, R. L. A multi-stage representation of cell proliferation as a Markov process. *Bull. Math. Biol.* **79**, 2905–2928 (2017).
- [27] Chao, H. X. *et al.* Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Mol. Syst. Biol.* **15** (2019).
- [28] Gavagnin, E. *et al.* Synchronised oscillations in growing cell populations are explained by demographic noise. *bioRxiv* (2020).
- [29] Johnston, I. G. & Jones, N. S. Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **471**, 20150050 (2015).
- [30] Summers, H. D. *et al.* Statistical analysis of nanoparticle dosing in a dynamic cellular system. *Nature nanotechnology* **6**, 170–174 (2011).
- [31] Shi, C. *et al.* Allocation of gene products to daughter cells is determined by the age of the mother in single *Escherichia coli* cells. *Proceedings of the Royal Society B* **287**, 20200569 (2020).
- [32] Robert, L. *et al.* Mutation dynamics and fitness effects followed in single cells. *Science* **359**, 1283–1286 (2018).
- [33] Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
- [34] Sepúlveda, L. A., Xu, H., Zhang, J., Wang, M. & Golding, I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science* **351**, 1218–1222 (2016).
- [35] Reverón-Gómez, N. *et al.* Accurate recycling of parental histones reproduces the histone modification landscape during DNA replication. *Mol. Cell* **72**, 239–249 (2018).
- [36] Hamperl, S. & Cimprich, K. A. Conflict resolution in the genome: how transcription and replication make it work. *Cell* **167**, 1455–1467 (2016).
- [37] Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455**, 485 (2008).
- [38] Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302 (2006).
- [39] Toner, D. & Grima, R. Effects of bursty protein production on the noisy oscillatory properties of downstream pathways. *Sci. Rep.* **3**, 2438 (2013).
- [40] Jia, C., Zhang, M. Q. & Qian, H. Analytic theory of stochastic oscillations in single-cell gene expression. *arXiv preprint arXiv:1909.09769* (2019).
- [41] Thomas, P., Straube, A. V., Timmer, J., Fleck, C. & Grima, R. Signatures of nonlinearity in single cell noise-induced oscillations. *J. Theor. Biol.* **335**, 222–234 (2013).
- [42] Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences* **111**, 6994–6999 (2014).



- [43] Tanouchi, Y. *et al.* A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature* **523**, 357–360 (2015).
- [44] Zopf, C., Quinn, K., Zeidman, J. & Maheshri, N. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput. Biol.* **9**, e1003161 (2013).
- [45] Xie, Z. *Case studies in time series analysis*, vol. 3 (World Scientific, 1993).
- [46] Walker, N., Nghe, P. & Tans, S. J. Generation and filtering of gene expression noise by the bacterial cell cycle. *BMC biology* **14**, 11 (2016).
- [47] Zhang, C., Liu, M.-S. & Xing, X.-H. Temperature influence on fluorescence intensity and enzyme activity of the fusion protein of GFP and hyperthermophilic xylanase. *Appl. Microbiol. Biotechnol.* **84**, 511–517 (2009).
- [48] Dessalles, R., Fromion, V. & Robert, P. Models of protein production along the cell cycle: An investigation of possible sources of noise. *Plos one* **15**, e0226016 (2020).
- [49] Cox, C. D. *et al.* Frequency domain analysis of noise in simple gene circuits. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **16**, 026102 (2006).
- [50] Munsky, B., Trinh, B. & Khammash, M. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology* **5**, 318 (2009).

## Acknowledgements

We thank Ruben Perez-Carrasco and Irina Kalita for stimulating discussions.

## Funding

C. J. acknowledges support from the NSAF grant in National Natural Science Foundation of China with grant No. U1930402. R. G. acknowledges support from the Leverhulme Trust (RPG-2018-423).

## Author contributions

R. G. conceived the original idea. C. J. performed the theoretical derivations and numerical simulations. C. J. and R. G. interpreted the theoretical results, analyzed the experimental data, and jointly wrote the manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Data and materials availability

All data needed to evaluate the conclusions in the paper are present in the paper and in [2].