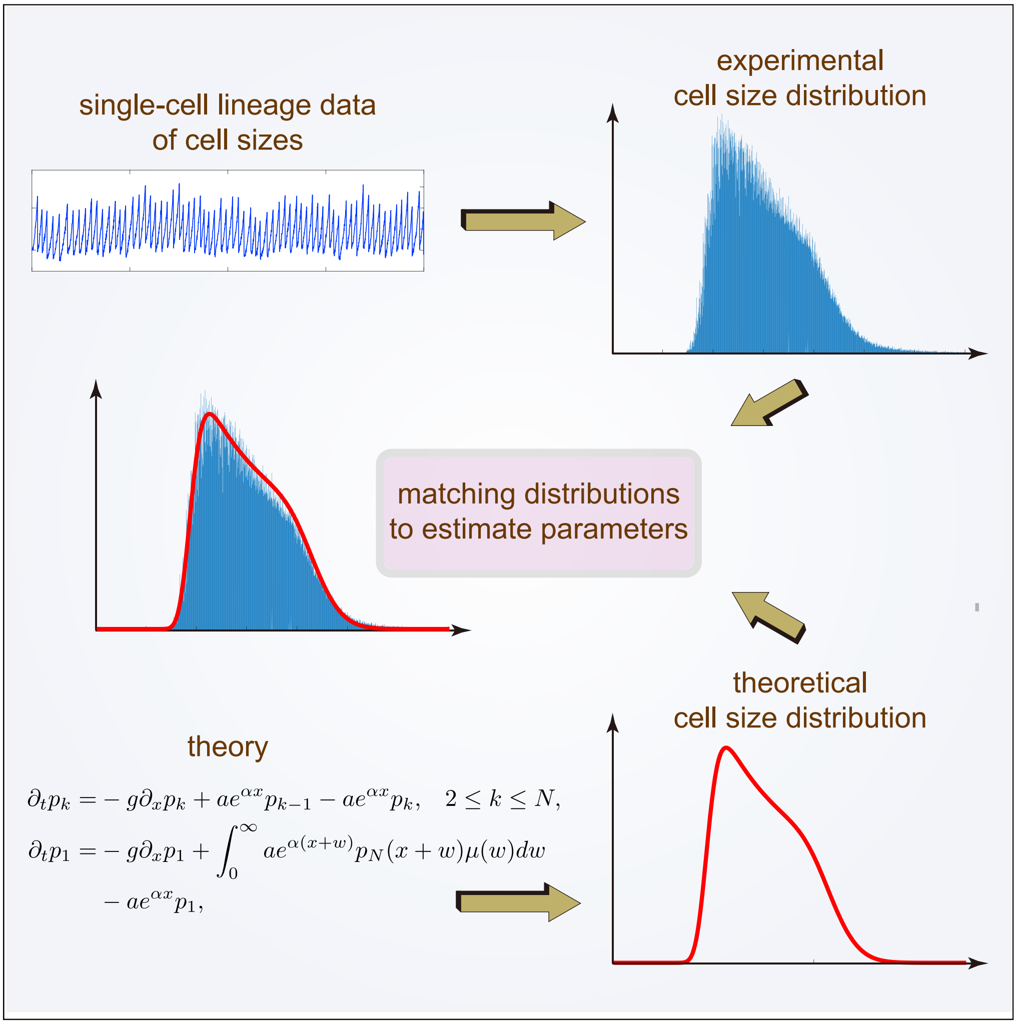
iScience

CellPress
OPEN ACCESS

Article

Cell size distribution of lineage data: analytic results and parameter inference



$$\partial_t p_k = -g\partial_x p_k + ae^{\alpha x}p_{k-1} - ae^{\alpha x}p_k, \quad 2 \le k \le N,$$

$$\partial_t p_1 = -g\partial_x p_1 + \int_0^\infty ae^{\alpha(x+w)}p_N(x+w)\mu(w)dw$$
$$- ae^{\alpha x}p_1,$$

Chen Jia,
Abhyudai Singh,
Ramon Grima

ramon.grima@ed.ac.uk

HIGHLIGHTS

Analytical expression is derived for the cell size distribution of lineage measurements

Theory explains the uncommon shape of the cell size distribution in *E. coli*

Multimodal size distribution is predicted for asymmetric division with random tracking

Size distribution matching gives accurate inference of size control strategy

Jia et al., iScience *24*, 102220
March 19, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102220

iScience

CellPress
OPEN ACCESS

Article

Cell size distribution of lineage data: analytic results and parameter inference

Chen Jia,[1] Abhyudai Singh,[2] and Ramon Grima[3,4,*]

SUMMARY

Recent advances in single-cell technologies have enabled time-resolved measurements of the cell size over several cell cycles. These data encode information on how cells correct size aberrations so that they do not grow abnormally large or small. Here, we formulate a piecewise deterministic Markov model describing the evolution of the cell size over many generations, for all three cell size homeostasis strategies (timer, sizer, and adder). The model is solved to obtain an analytical expression for the non-Gaussian cell size distribution in a cell lineage; the theory is used to understand how the shape of the distribution is influenced by the parameters controlling the dynamics of the cell cycle and by the choice of cell tracking protocol. The theoretical cell size distribution is found to provide an excellent match to the experimental cell size distribution of *E. coli* lineage data collected under various growth conditions.

INTRODUCTION

Cell size plays an important role in cellular processes; e.g. changes in cell volume or surface area have profound effects on metabolic flux and nutrient exchange Marshall et al. (2012), and therefore, it stands to reason that cell size should be actively maintained. In order for cells to achieve and maintain some characteristic size (size homeostasis), the amount of growth produced during the cell cycle must be controlled such that, on average, large cells at birth grow less than small ones.

There are three popular phenomenological models of cell size control leading to size homeostasis Vargas-Garcia et al. (2018): (i) the timer strategy which implies a constant time between successive divisions, (ii) the sizer strategy which implies cell division upon attainment of a critical size, and (iii) the adder strategy which implies a constant size addition between consecutive generations. The timer strategy is not viable for exponentially growing cells; in this case, size fluctuations diverge as the square root of the number of consecutive cell divisions implying that the timer strategy cannot maintain stable size distributions Jun and Taheri-Araghi (2015). In contrast, if cells grow linearly, a timer strategy is viable as a means to maintain size homeostasis Conlon and Raff (2003). Several studies have proposed that the sizer and adder strategies can explain experimental data in bacteria, yeast, and mammalian cells Campos et al. (2014); Taheri-Araghi et al. (2015); Tanouchi et al. (2015); Soifer et al. (2016); Chandler-Brown et al. (2017); Cadart et al. (2018). Cell size control mechanisms likely vary depending on growth conditions, strains, and species; for instance, in *Escherichia coli* (*E. coli*), evidence suggests a sizer mechanism in slow growth conditions and an adder in fast growth conditions Wallden et al. (2016).

Cell size statistics can be computed using data from cell lineages or population snapshots. To observe a single cell lineage, at each cell division event, one keeps track of only one of the newborn cells (daughter cells); thus, at an arbitrary time point, only a single cell is observed. To observe population snapshots, one tracks both daughters of each mother cell in the population and thus the evolution of the whole population over time. Recently, mathematical models have shown that cell size statistics calculated using lineage data, e.g. collected using mother machines, can vary considerably from those obtained from population snapshot data, e.g. collected using flow cytometry Thomas (2018); García-García et al. (2019); Totis et al. (2020a). In fact, differences between these two types of measurements are also observable in protein and mRNA count statistics Thomas (2017); Beentjes et al. (2020).

Modeling has elucidated various other interesting insights into cell size statistics; however, to our knowledge, no study thus far has attempted to explain the complex shapes of cell size distributions computed

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

[2]Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA

[3]School of Biological Sciences, University of Edinburgh, EH9 3JH, UK

[4]Lead contact

*Correspondence: ramon.grima@ed.ac.uk

https://doi.org/10.1016/j.isci.2021.102220

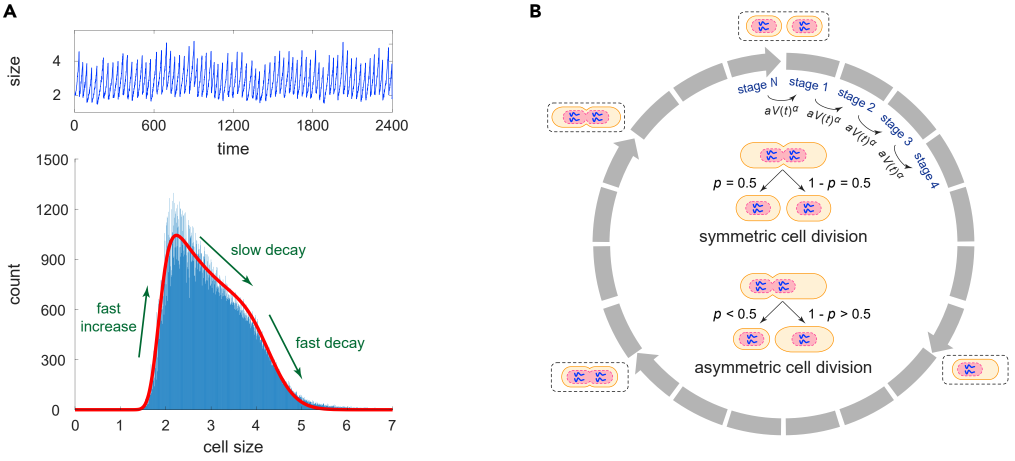


Figure 1. Cell size dynamics and a stochastic model describing it

(A) Single-cell time course data of cell length along a typical cell lineage measured in *E. coli* at 37°C (upper) and the histogram of cell sizes along all cell lineages (lower). The data shown are published in Tanouchi et al. (2015). When plotting the histogram, we use the data of all cell lineages at that temperature (160 lineages), each of which is recorded every minute over 70 generations. The cell size distribution computed from cell lineage measurements has an uncommon shape that is characterized by three features: a fast increase in the size count for small cells, followed by a slow decay for moderately large cells and a fast decay for large cells.

(B) Schematic illustrating a detailed model of cell size dynamics describing cell growth, multiple effective cell cycle stages, cell size control, and symmetric or asymmetric partitioning at cell division (see inset graph). Each cell can exist in N effective cell cycle stages. The transition rate from one stage to the next at a particular time t is proportional to the αth power of the cell size $V(t)$ with $\alpha > 0$ being the strength of cell size control and $a>0$ being the proportionality constant. This guarantees that larger cells at birth divide faster than smaller ones to achieve size homeostasis. At stage N, a mother cell divides into two daughters that are typically different in size via asymmetric cell division. Symmetric division is the special case where daughters are equisized.

from many generations of cell lineage measurements. This is because such high throughput data have become available only recently Tanouchi et al. (2017) and also since the majority of modeling approaches have analytically derived expressions for the first few moments of cell size statistics—these are not enough to characterize the highly non-Gaussian distribution of cell size computed from a cell lineage (see Figure 1A for a typical distribution for an *E. coli* lineage). These histograms are characterized by three features: a fast increase in the size count for small cells, a slow decay in the size count for moderately large cells, and a fast decay in the size count for large cells. We note that these distributions contain much more information than birth size distributions previously derived Amir (2014) since they reflect the full cell cycle dynamics.

Here, we develop a complete analytical theory of the cell size distribution in cell lineages. We formulate and solve a piecewise deterministic Markov model describing the evolution of the cell size over many generations, for all three size homeostasis strategies (timer, sizer, and adder). The model takes into account the major features responsible for the underlying dynamics: cell birth following division (including the asymmetric case and partitioning noise), exponential cell growth (including the case of noisy growth rates), variability in the duration of the cell cycle, and the user-defined choice of single-cell tracking protocols when division occurs, e.g., tracking always the smaller daughters, tracking always the larger daughters, or randomly picking one of the two daughters. The analytical solutions for the cell size distribution enable us to understand how the highly non-Gaussian shape of the distribution emerges from the underlying biophysical processes. Finally, by matching the analytical to the experimental cell size and doubling time distributions, we infer the values of various model parameters in *E. coli* for three different growth conditions.

RESULTS

Model specification

Here, we consider a detailed model of cell size dynamics across the cell cycle which is similar to the model proposed in Nieto et al. (2020a) but has more complicated cell division mechanisms such as asymmetric and stochastic partitioning (see Figure 1B for an illustration). The model is based on a number of

Table 1. Model parameters and their meaning

Parameters	Meaning
g	Exponential growth rate of cell size
N	Number of effective cell cycle stages
a	Proportionality constant for the transition rate between stages
α	Strength of cell size control
A	Mean generalized added size in each generation
p	Mean partition ratio of cell size at division

assumptions that are closely tied to experimental data. The assumptions are as follows, and the specific meaning of all model parameters is listed in Table 1.

1) The size of each cell grows exponentially in each generation with growth rate g. This assumption is supported by experiments in many cell types Godin et al. (2010).
2) Each cell can exist in N effective cell cycle stages, denoted by $1,2,\ldots,N$. The transition rate from one stage to the next at a particular time is proportional to the αth power of the cell size at that time, with $a>0$ being the proportionality constant Nieto et al. (2020a). In other words, the transition rate between stages at time t is equal to $aV(t)^{\alpha}$, where $\alpha > 0$ is the strength of cell size control and $V(t)$ is the cell size at that time. Under this assumption, larger cells at birth have larger transition rates between stages and thus, on average, have shorter cell cycle duration and lesser volume change than smaller ones; in the way size homeostasis is achieved.

Examples of possible biophysical mechanisms that can explain the power law form of the transition rate have been discussed in Nieto et al. (2020a). For instance, recent studies have suggested the accumulation of certain proteins up to a critical threshold as a possible mechanism in bacterial and fission yeast cell division Ghusinga et al. (2016); Sekar et al. (2018); Si et al. (2019); Patterson et al. (2019). Suppose that the concentration c of this "division protein" remains constant as the cell grows Weart and Levin (2003). Then, the number of molecules n of this protein is proportional to the cell volume V, i.e., $n = cV$. If this division protein makes polymers of α subunits, then the production rate of polymers will be proportional to n^{α}, which is proportional to V^{α}. Biologically, the multiple cell cycle stages considered here may be interpreted as different levels of the division protein polymer and cell division occurs when a certain number of polymers is reached. Under this mechanism, the rate of moving from one stage to the next is proportional to V^{α}. We emphasize that while this power law is compatible with certain biophysical mechanisms, it could also simply be understood as a phenomenological means to model size homeostasis.

Let V_b and V_d denote the cell sizes at birth and at division in a particular generation, respectively. Then, the increment in the αth power of the cell size across the cell cycle, $\Delta = V_d^{\alpha} - V_b^{\alpha}$, has an Erlang distribution with shape parameter N and mean $A = N\alpha g/a$ (see Section 1 in transparent methods for the proof). The quantity Δ will be referred to as the generalized added size in what follows. In our model, the noise in the generalized added size, characterized by the coefficient of variation squared, is equal to $1/N$. As N increases, the generalized added size, as well as V_b and V_d themselves, has smaller fluctuations. Since the cell cycle duration is given by $T=(1/g)\log(V_d/V_b)$, an increasing N also results in lesser fluctuations in the doubling time. Hence, our model allows the investigation of the influence of cell cycle duration variability on cell size dynamics.

We next focus on three crucial special cases. When $\alpha\to 0$, the transition rate between stages is a constant and thus the doubling time has an Erlang distribution that is independent of the birth size; this corresponds to the timer strategy. When $\alpha = 1$, the added size $V_d - V_b$ has an Erlang distribution that is independent of the birth size; this corresponds to the adder strategy. When $\alpha\to\infty$, the αth power of the cell size at division, V_d^{α}, has an Erlang distribution that is independent of the birth size; this corresponds to the sizer strategy. Intermediate strategies are naturally obtained for intermediate values of α; timer-like control is obtained when $0 < \alpha < 1$ and sizer-like control is obtained when $1<\alpha<\infty$ Nieto et al. (2020a).

3) 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 daughters that are exactly

the same in size via symmetric partitioning Amir (2014); Vargas-García and Singh (2018); Nieto-Acuna et al. (2019); Totis et al. (2020b); Nieto et al. (2020b); however, asymmetric cell division is common in biology. 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 Shi et al. (2020). Here, we follow the methodology that we devised in Jia and Grima (2020) and extend previous models by considering asymmetric partitioning at cell division: the mother cell divides into two daughters with different sizes.

If the partitioning of the cell size is symmetric, we track one of the two daughters randomly after division Brenner et al. (2015); Robert et al. (2018); if the partitioning is asymmetric, we either track the smaller daughter or track the larger daughter after division Zopf et al. (2013); Crane et al. (2014). Hence, our model corresponds to cell lineage measurements performed using a mother machine. Let V_d and $V_b{}'$ denote the cell sizes at division and just after division, respectively. If the partitioning is deterministic, then we have $V_b{}' = pV_d$, where $0 < p < 1$ is a constant with $p = 1/2$ corresponding to the case of symmetric division, $p < 1/2$ corresponding to smaller daughter tracking, and $p > 1/2$ corresponding to larger daughter tracking. However, in naturally occurring systems, the partitioning is appreciably stochastic. In this case, we assume that the partition ratio $V_b{}'/V_d$ has a beta distribution with mean p Nieto-Acuña et al. (2020), whose probability density function is given as follows:

$$f(z) = \frac{1}{B(p\nu, q\nu)} z^{p\nu-1}(1-z)^{q\nu-1},\ 0<z<1, \qquad \text{(Equation 1)}$$

where B is the beta function, $q = 1-p$, and $\nu > 0$ is referred to as the sample size parameter. The reason behind this assumption is that the partition ratio $V_b{}'/V_d$ should be a random number between 0 and 1, which is an important property of the beta distribution Nieto-Acuña et al. (2020). Then, the change in the logarithm of the cell size at division, $\log V_d - \log V_b{}' = \log(V_d/V_b{}')$, has the probability density function $\mu(w) = e^{-w}f(e^{-w})$, which can be written more explicitly as follows:

$$\mu(w) = \frac{1}{B(p\nu, q\nu)} e^{-p\nu w}(1-e^{-w})^{q\nu-1},\ w>0. \qquad \text{(Equation 2)}$$

When $\nu \to \infty$, the variance of the beta distribution tends to zero and thus stochastic partitioning reduces to deterministic partitioning, i.e., $f(z) = \delta(z-p)$ and $\mu(w) = \delta(w + \log p)$.

We next describe our stochastic model of cell size dynamics across the cell cycle. The microstate of the cell can be represented by an ordered pair (k,y), where k is the cell cycle stage which is a discrete variable and y is the cell size which is a continuous variable. Let $\tilde{p}_k(y)$ denote the probability density function of the cell size when the cell is in stage k. Note that the cell undergoes deterministic exponential growth in each stage and the system can hop between successive stages stochastically. Hence, the evolution of the cell size dynamics can be described by a *piecewise deterministic Markov process* whose Kolmogorov forward equation is given as follows:

$$\begin{aligned} \partial_t \tilde{p}_k &= -\partial_y\left(gy\tilde{p}_k\right) + ay^\alpha \tilde{p}_{k-1} - ay^\alpha \tilde{p}_k,\ 2 \le k \le N, \\ \partial_t \tilde{p}_1 &= -\partial_y\left(gy\tilde{p}_1\right) + \int_0^1 \frac{a}{z}\left(\frac{y}{z}\right)^\alpha \tilde{p}_N\left(\frac{y}{z}\right) f(z)dz - ay^\alpha \tilde{p}_1, \end{aligned} \qquad \text{(Equation 3)}$$

where $f(z)$ is the function defined in Equation (1). Similar hybrid models have, for example, been used to describe demographic noise in ecosystems Realpe-Gomez et al. (2012) and single-cell stochastic gene expression Lin and Buchler (2018); Jia et al. (2019). In the first equation above, the first term on the right-hand side represents the exponential growth of the cell size with growth rate g, and the second and third terms represent the transition between stages whose transition rate is proportional to the α th power of the cell size y. In the second equation, the second term corresponds to the partitioning of the cell size at division.

To solve Equation (3), the key step is to consider the dynamics of *the logarithmic cell size*, $x = \log y$, rather than the original cell size y. This is because the dynamic equation for the former is easier to solve. Let $p_k(x)$ denote the probability density function of the logarithmic cell size when the cell is in stage k. Since the probability density functions of the original and logarithmic cell sizes are related by $p_k(x) = y\tilde{p}_k(y)$, it follows from Equation (3) that the evolution of the logarithmic cell size is governed by the following master equation:

$$\begin{aligned}\partial_t p_k &= -g\partial_x p_k + ae^{\alpha x}p_{k-1} - ae^{\alpha x}p_k,\ 2\le k\le N,\\ \partial_t p_1 &= -g\partial_x p_1 + \int_0^\infty ae^{\alpha(x+w)}p_N(x+w)\mu(w)dw - ae^{\alpha x}p_1,\end{aligned} \qquad \text{(Equation 4)}$$

where $\mu(w)$ is the function defined in Equation (2).

Analytical distribution of the cell size along a cell lineage under deterministic partitioning

Recall that any probability distribution is fully determined by its characteristic function. Let $p(x)=\sum_{k=1}^{N}p_k(x)$ denote the probability density function of the logarithmic cell size. To obtain the analytical distribution of the cell size along a cell lineage, we introduce the characteristic function $G(\lambda) = \int_{-\infty}^{\infty} p(x)e^{i\lambda x}dx$, which is nothing but the inverse Fourier transform of $p(x)$. For simplicity, we first focus on deterministic partitioning at cell division, i.e., $\nu\to\infty$. Despite the biological complexity described by our model, the characteristic function can still be solved exactly in steady-state conditions (see Section 2 in transparent methods for the proof):

$$G(\lambda) = K\sum_{k=0}^{N-1}\sum_{l=0}^{k}C_{k,l}\left(\frac{A}{N}\right)^{l+1}\Gamma\left(1-\frac{i\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{l-\frac{i\lambda}{\alpha}}\prod_{n=0}^{\infty}a_N\left(p^{\alpha n}u\right)du, \qquad \text{(Equation 5)}$$

where $C_{k,l} = k!/l!(k\text{-}l)!$ is the combinatorial number, $a_N(u) = (1 + Au/N)^{-N}$ is a function of u, and

$$K=\left[\int_0^\infty \frac{1}{u}\left(a_N(u)^{-1}-1\right)\prod_{n=0}^{\infty}a_N\left(p^{\alpha n}u\right)du\right]^{-1}$$

is a normalization constant. Since the Fourier transform and the inverse Fourier transform are inverses of each other, taking the Fourier transform of the characteristic function gives the steady-state probability density function $p(x)$ of the logarithmic cell size. Finally, the probability density function of the original cell size $y = e^x$ along a cell lineage is given as follows:

$$\tilde{p}(y) = \frac{1}{y}p(\log y). \qquad \text{(Equation 6)}$$

The analytical solution is ideal since it allows a fast exploration of large swathes of parameter space without performing stochastic simulations.

To gain deeper insights into the cell size distribution, we next consider the limiting case of $N\to\infty$. In this case, the generalized added size $\boldsymbol{\Delta}$, as well as the cell cycle duration T, becomes deterministic, and thus, the system does not involve any stochasticity. As $N\to\infty$, we have $a_n(u) = e^{-Au}$, and thus, the characteristic function can be simplified to a large extent as follows (see Section 2 in transparent methods for the proof):

$$G(\lambda) = \frac{\overline{V}_d^{i\lambda}-\overline{V}_b^{i\lambda}}{\left(\log\overline{V}_d-\log\overline{V}_b\right)i\lambda}, \qquad \text{(Equation 7)}$$

where

$$\overline{V}_b = p\left(\frac{A}{1-p^{\alpha}}\right)^{\frac{1}{\alpha}},\ \overline{V}_d = \left(\frac{A}{1-p^{\alpha}}\right)^{\frac{1}{\alpha}},$$

are two constants. Taking the Fourier transform of $G(\lambda)$ shows that the logarithmic cell size has the uniform distribution

$$p(x) = \frac{1}{\log\overline{V}_d-\log\overline{V}_b}I_{\left[\log\overline{V}_b,\log\overline{V}_d\right]}(x), \qquad \text{(Equation 8)}$$

and thus the original cell size $y = e^x$ has the following distribution:

$$\tilde{p}(y) = \frac{1}{y}p(\log y) = \frac{1}{\left(\log\overline{V}_d-\log\overline{V}_b\right)y}I_{\left[\overline{V}_b,\overline{V}_d\right]}(y), \qquad \text{(Equation 9)}$$

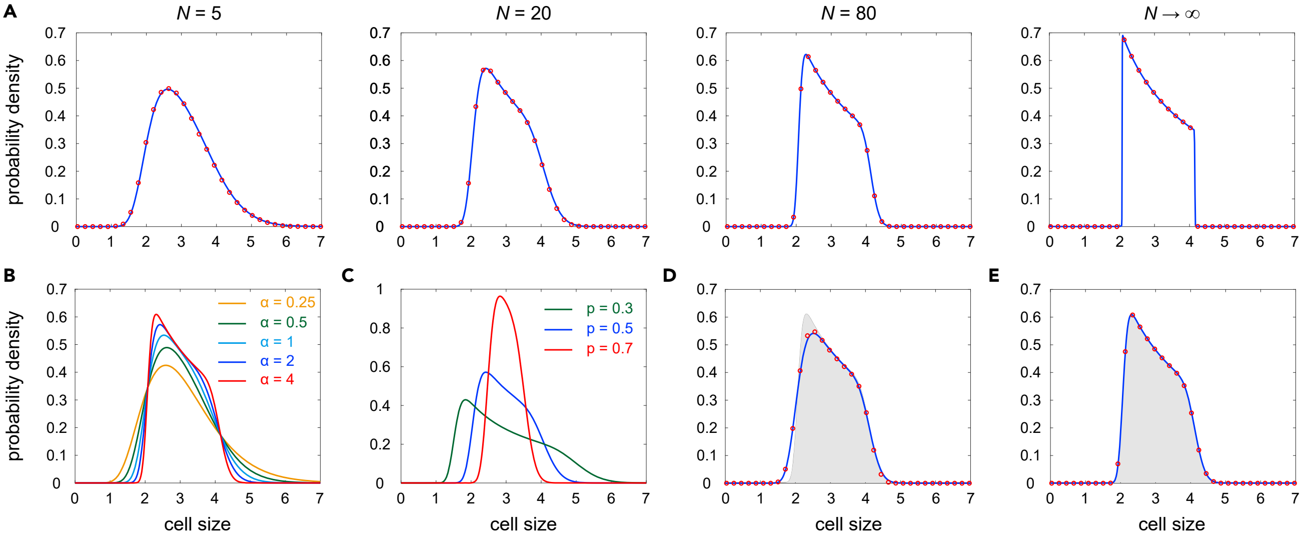


Figure 2. Influence of model parameters on the cell size distribution

(A) Cell size distribution as N increases. The red curve shows the analytical distribution given in Equation (6), and the red circles show the distribution obtained using the stochastic simulation algorithm proposed in Nieto et al. (2020b). The parameters are chosen as $\alpha = 2$, $p = 0.5$.

(B) Cell size distribution as α varies. The parameters are chosen as $N = 20$, $p = 0.5$.

(C) Cell size distribution as p varies. The parameters are chosen as $N = 20$, $\alpha = 2$.

(D) Comparison of the cell size distributions for the model with stochastic partitioning (blue curve and red circles) and the model with deterministic partitioning (solid gray region). The blue curve shows the approximate distribution given in Equation (20), and the red circles show the distribution obtained from simulations.

(E) Comparison of the cell size distributions for the model with stochastic growth rate (blue curve and red circles) and the model with deterministic growth rate (solid gray region). In (D) and (E), the parameters are chosen as $N = 30$, $\alpha = 3$, $p = 0.5$. In (A)-(E), the growth rate is chosen as $g = 0.02$ and the parameters A and a are chosen so that $\langle V\rangle = 3$ for the model with deterministic growth rate and deterministic partitioning. In (D), the standard deviation of the partition ratio is 10% of the mean; here, we assume that the partition ratios for different generations are i.i.d. normally distributed random variables. In (E), the standard deviation of the growth rate is 10% (red circles) or 50% (blue curve) of the mean; here, we assume that the growth rates for different generations are i.i.d. normally distributed random variables.

where $I_B(x)$ is the indicator function which takes the value of 1 when $x \in B$ and the value of 0 otherwise. This indicates that when cell cycle duration variability is small, the cell size has a distribution that is concentrated on the finite interval $[\overline{V}_b, \overline{V}_d]$, where $\overline{V}_b$ and $\overline{V}_d$ are the typical cell sizes at birth and at division, respectively.

Figures 2A–2C illustrate the distribution of the original cell size as a function of the parameters N, α, and p. It can be seen that as cell cycle duration variability become smaller (N increases), the analytical distribution given in Equation (6) converges to the limit distribution given in Equation (9). The cell size distribution has a regular shape for small N. As N increases, the shape of the distribution becomes more complicated. In particular, the distribution has three apparent sections: an exponential increase for small sizes, a power law decay for moderate sizes, and an exponential decay for large sizes. As $N \to \infty$, the dynamics becomes deterministic and the distribution has a compact support, characterized by infinite slopes of the two shoulders. In addition, we find that the influence of α on the cell size distribution is similar to the influence of N. Finally, increasing p gives rise to a distribution that is more symmetric and more concentrated.

Moments, noise, and skewness of the cell size distribution

Our analytic results can also be used to derive explicit expressions for several other quantities of interest. Recall that the probability density function $p(x)$ for the logarithmic cell size and the probability density function $\tilde{p}(y)$ for the original cell size are related by Equation (6). For any real number λ, the λth moment of the original cell size is given as follows:

$$\langle V^{\lambda}\rangle = \int_0^{\infty} y^{\lambda}\tilde{p}(y)dy = \int_{-\infty}^{\infty} e^{\lambda x}p(x)dx = F(\lambda),$$

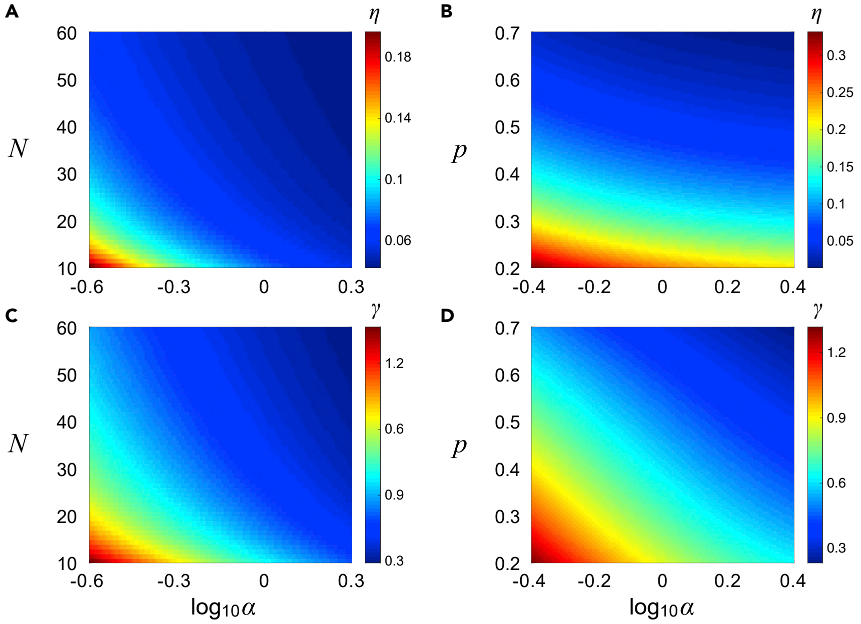


Figure 3. Noise and skewness of the cell size distribution

(A) Heatmap of the noise η versus α and N.

(B) Heatmap of the noise η versus α and p.

(C) Heatmap of the skewness γ versus α and N.

(D) Heatmap of the skewness γ versus α and p. The parameters are chosen as $p = 0.5$ in (A) and (C) and $N = 20$ in (B) and (D). In (A)-(D), the parameter A is chosen so that the mean cell size $\langle V \rangle = 3$.

where $F(\lambda)$ is the moment generating function of $p(x)$. This shows that the λth moment of the original cell size is exactly the moment generating function of the logarithmic cell size taken value at λ. Since the moment generating function $F(\lambda)$ and the characteristic function $G(\lambda)$ are related by $G(\lambda) = F(i\lambda)$, replacing the variable $i\lambda$ in Equation (5) by λ yields the moment generating function. Hence, the λth moment of the original cell size is given as follows:

$$\langle V^{\lambda} \rangle = F(\lambda) = K \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(\frac{A}{N}\right)^{l+1} \Gamma\left(1 - \frac{\lambda}{\alpha}\right)^{-1} \int_0^{\infty} u^{l-\frac{\lambda}{\alpha}} \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du. \quad \text{(Equation 10)}$$

In single-cell experiments, the noise in the cell size, characterized by the coefficient of variation squared, is given as follows:

$$\eta = \frac{\sigma^2}{\mu^2} = \frac{F(2)}{F(1)^2} - 1, \quad \text{(Equation 11)}$$

where μ is the mean and σ^2 is the variance. Figures 3A and 3B illustrate the noise η as a function of N, α, and p. Clearly, the fluctuations in the cell size become smaller with the increase of all the three parameters (see also Figure 2). This implies that small cell cycle duration variability and sizer-like strategy can lead to a more accurate control of the cell size.

A special case occurs when the cell cycle duration variability is very small, i.e., $N \gg 1$. In this case, replacing the variable $i\lambda$ in the characteristic function Equation (7) by λ yields the following equation:

$$\langle V^{\lambda} \rangle = F(\lambda) = \frac{1 - p^{\lambda}}{-\lambda \log p} \left(\frac{A}{1 - p^{\alpha}}\right)^{\frac{\lambda}{\alpha}}. \quad \text{(Equation 12)}$$

Thus, the noise in the cell size is given as follows:

$$\eta = -\frac{(1+p)\log p}{2(1-p)} - 1,$$

which is a decreasing function of p. Note that when N is small, the noise η is a function of both α and p (Figure 3B). However, when N is large, the noise only depends on p. It is easy to see that the noise in the cell size tends to infinity as $p \to 0$ and tends to zero as $p \to 1$. For the case of symmetric division ($p = 0.5$), the noise in the cell size is given by $\eta \approx 0.04$, which shows that the standard deviation of the cell size is roughly 20% of the mean.

Recall that the skewness of the cell size distribution is defined as follows:

$$\gamma = \langle \left(\frac{V-\mu}{\sigma} \right)^3 \rangle = \frac{F(3) - 3F(1)F(2) + 2F(1)^3}{\left[F(2) - F(1)^2\right]^{3/2}}, \qquad \text{(Equation 13)}$$

Figures 3C and 3D illustrate the skewness γ as a function of N, α, and p, from which we can see that the skewness increases with the decrease of all the three parameters. This implies that large cell cycle duration variability, timer-like division strategy, and tracking the smaller daughter at division lead to larger skewness of the cell size distribution. Moreover, we find that the skewness is always positive, which means that the cell size distribution is always right skewed. When $N \gg 1$, it follows from Equation (12) that the skewness only depends on p and is given as follows:

$$\gamma = \frac{2(1-p^3)(\log p)^2 + 9(1-p)(1-p^2)\log p + 12(1-p)^3}{6\left[-(1-p^2)\log p - (1-p)^2\right]^{3/2}},$$

which is also a decreasing function of p.

Analytical distribution of the cell cycle duration

In our model, the distribution of the doubling time can also be derived analytically in steady-state conditions. Actually, given that the birth size V_b is known, the conditional probability density of the cell cycle duration T has been obtained in Nieto et al. (2020a) as follows:

$$\mathbb{P}\left(T = t \middle| V_b^\alpha = x\right) = \frac{\alpha g N^N}{A^N (N-1)!} x^N \left(e^{\alpha g t} - 1\right)^{N-1} e^{\alpha g t - \frac{N}{A} x (e^{\alpha g t} - 1)}.$$

Here, we compute the unconditional distribution of the cell cycle duration. To this end, we find that the Laplace transform of V_b^α is given by the following equation (see Section 3 in transparent methods for the proof):

$$\langle e^{-\lambda V_b^\alpha} \rangle = \prod_{n=1}^{\infty} \left(1 + \frac{A p^{\alpha n} \lambda}{N} \right)^{-N} = \prod_{n=1}^{\infty} a_N\left(p^{\alpha n} u\right). \qquad \text{(Equation 14)}$$

Taking the inverse Laplace transform gives the probability density function of V_b^α. Finally, the distribution of the cell cycle duration T is given as follows:

$$\mathbb{P}(T = t) = \int_0^\infty \mathbb{P}\left(T = t \middle| V_b^\alpha = x\right) \mathbb{P}\left(V_b^\alpha = x\right) dx. \qquad \text{(Equation 15)}$$

A special case occurs when α is large (strong cell size control) or when p is small (smaller daughter tracking). Under the large α or small p approximation, the term $p^{\alpha n}$ is negligible for $n \geq 2$ and it suffices to keep only the first term in the infinite product given in Equation (14). In this case, the inverse Laplace transform has an explicit expression and the birth size distribution is given as follows:

$$\mathbb{P}(V_b = x) = \frac{N^N x^{\alpha N - 1} e^{-\frac{N}{A p^\alpha} x^\alpha}}{(N-1)! A^N p^{\alpha N}}. \qquad \text{(Equation 16)}$$

Inserting this equation into Equation (15) yields the doubling time distribution:

$$\mathbb{P}(T = t) = \frac{\alpha g (2N-1)!}{p^{\alpha N} [(N-1)!]^2} \cdot \frac{e^{\alpha g t} \left(e^{\alpha g t} - 1\right)^{N-1}}{\left(p^{-\alpha} + e^{\alpha g t} - 1\right)^{2N}}. \qquad \text{(Equation 17)}$$

We emphasize that in the special case of $N = 1$ and $p = 0.5$, our model reduces to the model in Osella et al. (2014) and the above two equations coincide with the results in that paper.

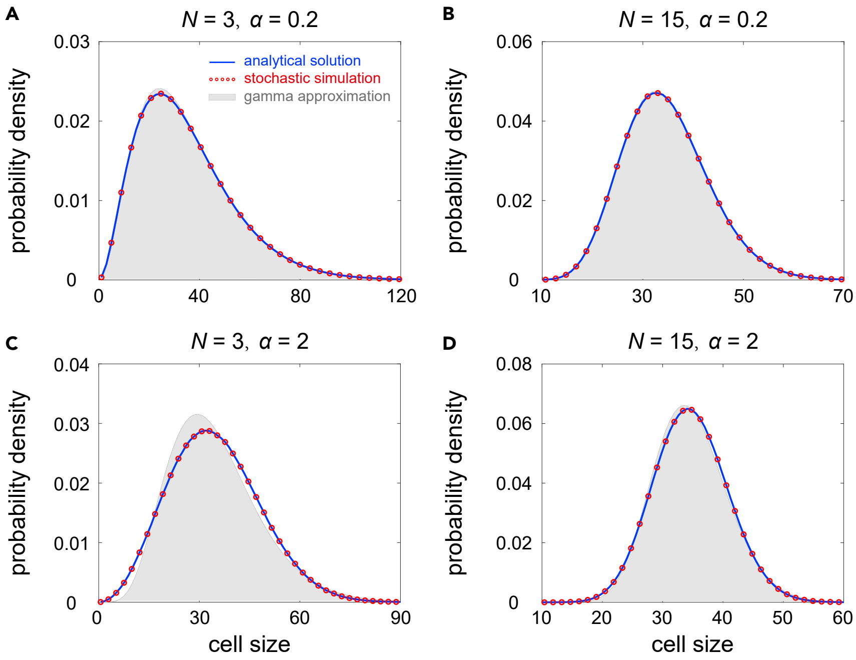


Figure 4. Distribution of the cell cycle duration and its approximation by the gamma distribution

We use the information of the sample mean and sample variance of the true distribution to determine the two parameters involved in the gamma approximation.

(A) Large cell cycle duration variability and small size control strength.

(B) Small cell cycle duration variability and small size control strength.

(C) Large cell cycle duration variability and large size control strength.

(D) Small cell cycle duration variability and large size control strength. In (A)-(D), the blue curve represents the analytical distribution given in Equation (15), the red circles represent the distribution obtained from simulations, and the gray region represents the gamma approximation. The parameters are chosen as $p = 0.5$, $g = 0.02$ and A and a are determined so that $\langle V \rangle = 3$.

Recent experiments Golubev (2016); Yates et al. (2017); Chao et al. (2019); Perez-Carrasco et al. (2020); Gavagnin et al. (2020) have shown that the cell cycle durations in various cell types are all well fitted by a gamma distribution. Therefore, it is natural to ask whether the doubling time in our model shares the same property. To see this, we illustrate the doubling time distribution and its approximation by the gamma distribution as N and α vary (Figure 4). It can be seen that the true distribution is in good agreement with its gamma approximation when α is small (Figures 4A and 4B). This is because a small α implies a timer-like size control, which leads to an approximately Erlang distributed doubling time due to the effect of multiple cell cycle stages and constant transition rates between them. When α is large, there are some slight differences between them for small N (Figure 4C); compared with the gamma approximation, the true distribution is more symmetric around its mean. However, when N is large, they are very close to each other and both well fitted by a normal distribution (any gamma distribution converges to the normal distribution as the shape parameter tends to infinity, see Figure 4D).

Correlation between sizes at birth and sizes at division

The birth size distribution derived above can be applied to study the correlation between birth and division sizes. Let V_b and V_d be the birth and division sizes in a particular generation, respectively, and let $V_b{}'$ and $V_d{}'$ be the birth and division sizes in the next generation, respectively. Since the generalized added size $\Delta = V_d^\alpha - V_b^\alpha$ is Erlang distributed and $V_b{}' = pV_d$, it is easy to obtain from Equation (14) that (see Section 4 in transparent methods for the proof)

$$\rho\left(V_b^\alpha, V_d^\alpha\right) = \rho\left(V_b^\alpha, V_b{}'^\alpha\right) = \rho\left(V_d^\alpha, V_d{}'^\alpha\right) = p^\alpha, \qquad \text{(Equation 18)}$$

where $\rho(X,Y)$ denotes Pearson's correlation coefficient between random variables X and Y. This characterizes the correlation between sizes at birth and sizes at division, as well as the correlation between the birth/

division sizes for mother and daughter cells. In particular, for the adder strategy ($\alpha = 1$), we have the following:

$$\rho(V_b, V_d) = \rho(V_b, V_b{}') = \rho(V_d, V_d{}') = p.$$

This implies that the size correlation for the adder only depends on p and is independent of other parameters. For the case of symmetric division ($p = 0.5$), this is consistent with the result obtained in Amir (2014), where the correlation coefficient between birth and division sizes is found to be approximately 0.5. In the presence of noise in partitioning, the formula for the correlation coefficient should be modified as follows (see Section 4 in transparent methods for the proof):

$$\rho\left(V_b^\alpha, V_d^\alpha\right) = \sqrt{\frac{N\left[(2K_1+1)K_2 - K_1^2\right] + K_2}{N\left[(2K_1+1)K_2 - K_1^2\right] + K_2 + 1}}, \qquad \text{(Equation 19)}$$

where

$$K_1 = \frac{B(\alpha + p\nu, q\nu)}{B(p\nu, q\nu) - B(\alpha + p\nu, q\nu)}, \quad K_2 = \frac{B(2\alpha + p\nu, q\nu)}{B(p\nu, q\nu) - B(2\alpha + p\nu, q\nu)}.$$

In this case, $\rho(V_b^\alpha, V_b{}'^\alpha)$ and $\rho(V_d^\alpha, V_d{}'^\alpha)$ are generally lower than $\rho(V_b^\alpha, V_d^\alpha)$ due to fluctuations in partitioning.

Distribution of the cell size along a cell lineage under stochastic partitioning and stochastic growth rate

Thus far, the analytical distribution of the cell size is obtained when the partitioning at division is deterministic. In the presence of noise in partitioning, it is very difficult to obtain the explicit expression of the cell size distribution. Fortunately, in naturally occurring systems, the stochasticity in partitioning is often very small. For example, recent cell lineage data Tanouchi et al. (2015) suggested that the coefficient of variation of the partition ratio $z = V_b{}'/V_d$ in *E. coli* is about 7%–9%. When noise in partitioning is small, we obtain an approximate expression for the cell size distribution, whose moment generating function is given by the following equation (see Section 5 in transparent methods for the proof):

$$F(\lambda) = K\sum_{k=0}^{N-1}\sum_{l=0}^{k} C_{k,l}\left(\frac{A}{N}\right)^{l+1}\Gamma\left(1 - \frac{\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{l-\frac{\lambda}{\alpha}}\prod_{n=0}^{\infty} a_N\left(p(\lambda)^{\alpha n}u\right)du, \qquad \text{(Equation 20)}$$

where K is a normalization constant and

$$p(\lambda) = \left(\int_0^1 f(x)x^{\lambda-\alpha}dx\right)^{\frac{1}{\lambda-\alpha}}.$$

To see the effect of stochastic partitioning, we illustrate the cell size distributions under deterministic and stochastic partitioning in Figure 2D with the standard deviation of the partition ratio z being 10% of the mean for the latter. Clearly, the approximate solution given in Equation (20) matches the simulation results very well. In addition, it can be seen that noise in partitioning gives rise to larger fluctuations in the cell size, characterized by the smaller slope of the left shoulder of the cell size distribution.

In addition to noise in partitioning, there is another important source of stochasticity, i.e., noise in the growth rate g. In many biological systems, such noise is also very small. For example, recent cell lineage data Tanouchi et al. (2015) suggested that the coefficient of variation of the growth rate g in *E. coli* is about 7%–8%. To see the influence of noise in the growth rate, we illustrate the cell size distributions under deterministic and stochastic growth rates in Figure 2E with the standard deviation of g being 10% or 50% of the mean for the latter (here we assume that the growth rates for different generations are i.i.d. normally distributed random variables). Interestingly, we find that noise in the growth rate has almost no effect on the cell size distribution, even when the noise is very large; this is in sharp contrast to noise in partitioning which has an apparent effect on the cell size distribution.

Random tracking protocol can lead to complex multimodal cell size distributions

If cell division is asymmetric, the two daughters are different in size and thus far we have assumed that the smaller/larger daughter (such as the bud/mother cell in budding yeast) is tracked after division Zopf et al. (2013); Crane et al. (2014). We have seen that whether the smaller or the larger daughter is tracked, the cell

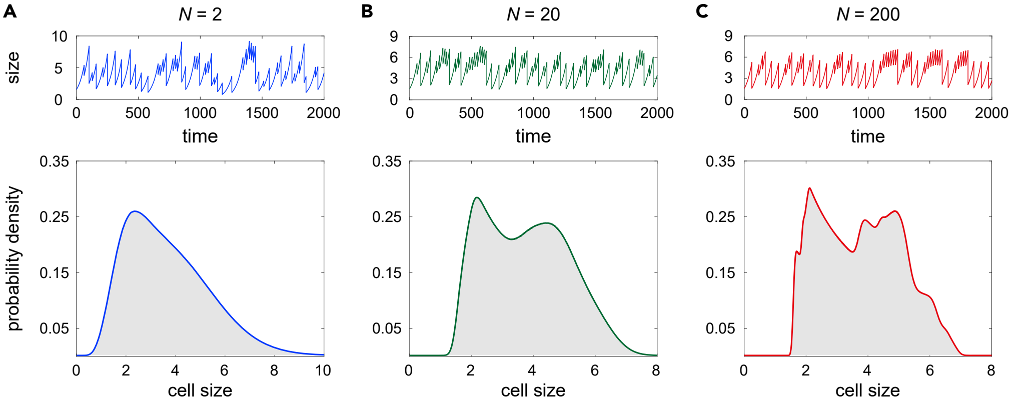


Figure 5. Cell size distribution for asymmetric cell division under the random tracking protocol

After division, one of the two daughters is randomly tracked with probability 1/2.

(A) Typical stochastic trajectory of the cell size (upper) and the cell size distribution (lower) in the case of large cell cycle duration variability (N = 2).

(B) Same as (A) but for moderate cell cycle duration variability (N = 20).

(C) Same as (A) but for small cell cycle duration variability (N = 200). In (A)-(C), the colored curve and the gray region show the cell size distributions obtained from two independently repeated stochastic simulations. The parameters are chosen as p = 0.3, α = 2, A = 25.

size distribution along a cell lineage is always unimodal and right skewed and larger daughter tracking yields lesser fluctuations in size than smaller daughter tracking. Next, we consider another tracking protocol, namely where we track one of the two daughters randomly with probability 1/2 after division Tanouchi et al. (2015); Brenner et al. (2015); Robert et al. (2018). Clearly, the three types of tracking protocols (tracking a random daughter, the smaller daughter, or the larger daughter) are exactly the same for symmetric cell division; however, they are remarkably different for asymmetric cell division.

For the random tracking protocol, the probability density function of the partition ratio $z = V_b'/V_d$ is given by the following equation (here the noise in partitioning is ignored):

$$f(z) = \frac{1}{2}\delta(z-p) + \frac{1}{2}\delta(z-q), \qquad \text{(Equation 21)}$$

where $0 < p \leq 1/2$ is the ratio of the size of the smaller daughter to the size of the mother cell and $q = 1-p$. Figure 5 illustrates the simulated cell size distribution under the random tracking protocol. Interestingly, we find that the shape of the distribution undergoes two stochastic bifurcations as cell cycle duration variability becomes smaller (N increases). When N is small, the cell size distribution is in general unimodal (Figure 5A), as in the case of smaller/larger daughter tracking. When N is moderate, random tracking is capable of producing a bimodal cell size distribution (Figure 5B), where the two peaks can be attributed intuitively to the subpopulations of smaller and larger daughters, respectively. Surprisingly, when N is large, we find that random tracking can give rise to a complex cell size distribution that displays multiple peaks (Figure 5B), two major peaks and some minor peaks. Increasing cell cycle duration variability (decreasing N) smoothens the cell size distribution, by first removing the smaller peaks and then merging the two major peaks into one.

Parameter inference using synthetic data

Recent breakthroughs in microfluidic devices have made it possible to monitor the single-cell volume dynamics along a cell lineage over many generations. Given such cell lineage data, an important question is whether all the parameters involved in our model can be inferred accurately. Parameter inference is crucial since it provides insights on the strength of cell size control, as well as cell cycle duration variability in various cell types.

The steps of our parameter estimation method are described as follows. First, the data of cell sizes at birth and at division in each generation, V_b and V_d, can be easily extracted from the cell lineage data. Since $\Delta = V_d^\alpha - V_b^\alpha$ is Erlang distributed with shape parameter N and mean A, once the parameter α is determined, both the parameters N and A can be determined by fitting the data of $V_d^\alpha - V_b^\alpha$ to an Erlang distribution. In addition, since $\rho(V_b^\alpha, V_d^\alpha) = p^\alpha$ (assuming deterministic partitioning), once the parameter α is determined, the parameter p can also be inferred from the correlation coefficient between V_b and V_d. For clarity, let $N(\alpha)$, $A(\alpha)$, and $p(\alpha)$ denote the optimal estimates of N, A, and p given the value of α, respectively. They can be inferred from the data of birth and division sizes as follows:

$$A(\alpha) = \langle \Delta \rangle, \; N(\alpha) = \frac{A(\alpha)^2}{\langle \Delta^2 \rangle - \langle \Delta \rangle^2}, \qquad \text{(Equation 22)}$$

$$p(\alpha) = \rho\left(V_b^\alpha, V_d^\alpha\right)^{1/\alpha}. \qquad \text{(Equation 23)}$$

If the partitioning at division is stochastic, then Equation (23) should be replaced by Equation (19). Next, the parameter α is determined by an optimal fit of the experimental to the theoretical cell size distribution using the least square criterion. Specifically, we determine α by solving the following optimization problem:

$$\min_\alpha \sum_{i=1}^{M} \left| p\left(x_i; \alpha, p(\alpha), N(\alpha), A(\alpha)\right) - \widehat{p}(x_i) \right|^2, \qquad \text{(Equation 24)}$$

where $p(x;\alpha,p,N,A)$ is the theoretical cell size distribution given the parameters α,p,N,A, $\widehat{p}(x)$ is the sample cell size distribution obtained from experiments, x_i are some reference points, and M is the number of bins chosen. Once α is estimated, the values of p, N, and A, are automatically determined. The reason why we do not estimate p directly as the mean of the partition ratio $V_b{}'/V_d$ is that the cell size distribution is sensitive to the value of p. A comparatively small error in p will result in a comparatively large change in the cell size distribution.

Since the cell size distribution is a function of $A = N\alpha g/a$, which depends on the ratio of g and a, it is impossible to infer the growth rate g from the cell size distribution. Finally, the growth rate g is determined by an optimal fit of the experimental to the theoretical/simulated doubling time distribution using the least square criterion. Once g is inferred, the last parameter a can be determined from the estimated α, N, and A as $a = N\alpha g/A$.

To verify the effectiveness of our method, we use our model to generating synthetic data of cell size dynamics. To make the time course data better mimic real biological processes, we add some noise to both the growth rate g and the partition ratio z. We then perform parameter inference by fitting the noisy data to two models: the model with deterministic partitioning (model I) and the model with stochastic partitioning (model II). The parameters input to the synthetic data and the parameters estimated using the above method are given in Table 2, where three sets of input parameters are chosen to cover large swathes of parameter space and to include three types of control strategies (timer-like, adder, and sizer-like). It can be seen that fitting the noisy data to both models leads to an accurate estimation of p and g and a relatively accurate estimation of N. However, fitting the data to model I gives rise to a systematic underestimation of α and A and an overestimation of a due to stochasticity in partitioning. Fitting the data to model II can remarkably improve the accuracy of estimation of these three parameters.

Experimental validation of the theory

To test our theory, we apply it to study the single-cell time course data of the cell size collected for *E. coli* in Tanouchi et al. (2017). In this data set, the time course data of the cell length were recorded every minute for 279 cell lineages over 70 generations using a mother machine microfluidic device 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 at each temperature by fitting the data to both model I and model II. The estimated parameters and the estimation errors are listed in Table 3 and are depicted by the box plots in Figure 6D.

From the estimated parameters, it can be seen that both models lead to similar estimation of p, N, and g. However, the introduction of partitioning noise into the model leads to higher estimation of α and A and lower estimation of a; this is consistent with our observation for synthesis data. For model I, the strength of cell size control, α, is estimated to be 0.7–0.9 for the three growth conditions, which are all lower than 1. Incorporating partitioning noise into the model gives rise to a higher estimate of α; for model II, α is estimated to be 0.8–1.2 for the three temperatures, implying that the size control strategy in *E. coli* is close to the adder. Moreover, higher temperature leads to a higher strength α than lower temperature.

Table 2. Parameter inference using synthesis data

	α	p	N	A	g	a
Input parameters	0.5	0.4	30	0.79	0.01	0.191
Estimated parameters using model I	0.44 ± 0.02	0.40 ± 0.0002	28.70 ± 0.82	0.65 ± 0.06	0.0100 ± 0.0001	0.195 ± 0.015
Estimated parameters using model II	0.50 ± 0.04	0.40 ± 0.0003	28.10 ± 1.52	0.79 ± 0.11	0.0100 ± 0.0002	0.179 ± 0.016
	α	p	N	A	g	α
Input parameters	1	0.5	20	2.08	0.02	0.192
Estimated parameters using model I	0.77 ± 0.06	0.50 ± 0.0003	21.15 ± 0.88	1.26 ± 0.18	0.0200 ± 0.0002	0.263 ± 0.026
Estimated parameters using model II	0.99 ± 0.08	0.50 ± 0.0003	19.20 ± 1.14	2.04 ± 0.35	0.0200 ± 0.0003	0.187 ± 0.027
	α	p	N	A	g	α
Input parameters	2	0.6	10	9.39	0.03	0.064
Estimated parameters using model I	1.29 ± 0.04	0.60 ± 0.0005	12.90 ± 0.32	2.73 ± 0.20	0.0301 ± 0.0003	0.183 ± 0.010
Estimated parameters using model II	1.92 ± 0.04	0.60 ± 0.0004	10.40 ± 0.52	8.29 ± 0.57	0.0300 ± 0.0004	0.072 ± 0.004

The cell lineage data are generated from the model with stochastic partitioning, where some noise is added to the growth rate g and the partition ratio z with the coefficients of variation of both parameters being chosen as 7% (here we assume that the growth rate is constant in each generation and the growth rates/partition ratios across different generations are i.i.d. normally distributed random variables). For each set of model parameters, we generate synthetic data simulating 50 cell lineages. For each cell lineage, the model parameters are estimated by fitting the synthetic data to both the model with deterministic partitioning at cell division (model I) and the model with stochastic partitioning (model II). The value in each cell shows the mean and standard deviation of the estimated parameter computed over 50 cell lineages.

Previous papers Tanouchi et al. (2015); Modi et al. (2017); Thomas (2018) have proposed an alternative approach to determining the regulation strength. This approach assumes that the birth size V_b and the division size V_d in each generation are related by the following:

$$V_d = \beta V_b + \gamma + \varepsilon, \qquad \text{(Equation 25)}$$

where $0 \leq \beta \leq 2$ and $\gamma \geq 0$ are two constants and ε is Gaussian white noise. Under this assumption, β characterizes the strength of cell size control with $\beta = 0$, $\beta = 1$, and $\beta = 2$ corresponding to the sizer, adder, and timer strategies, respectively. Using the data of birth and division sizes for different generations, the parameter β can be easily determined as the slope of the regression line of V_d on V_b. Based on the lineage data, β is estimated to be 1.08 for cells at 25°C, 1.02 for cells at 27°C, and 0.93 for cells at 37°C (Figure S1), all of which are close to 1. This implies a size control strategy close to the adder, which is consistent with the predictions of our model. Compared with this method, our distribution matching method seems to be more reliable since the linear relationships between V_b and V_d are actually very weak with numerous outliers and a low R^2 around 0.3 (Figure S1).

Figures 6A and 6B illustrate the experimental cell size and cell cycle duration distributions using the measurements of all cell lineages versus the theoretical distributions using the estimated parameters. Here, the theoretical distributions are plotted based on model II, but both models give rise to similar distribution shapes. Interestingly, both experimental distributions at all the three temperatures coincide perfectly with our model, which implies that our model can indeed reproduce the cell size dynamics in *E. coli* very well. In addition, it supports the main assumption of choosing the rate of moving from one effective cell cycle stage to the next to be a power law of the cell size.

Based on the lineage data, the correlation coefficients between the birth and division sizes, V_b and V_d, for the three temperatures are 0.533 ± 0.099, 0.503 ± 0.121, and 0.431 ± 0.139, respectively, and the correlation coefficients between mother and daughter birth sizes, V_b and V_b', are 0.475 ± 0.107, 0.445 ± 0.134, and 0.399 ± 0.146, respectively. Here, the errors represent the standard deviations computed over all cell

Table 3. Model parameters estimated using *E. coli* cell lineage data at three different temperatures

Model I	25°C	27°C	37°C
α	0.690 ± 0.0332	0.751 ± 0.0374	0.868 ± 0.1037
p	0.442 ± 0.0020	0.451 ± 0.0016	0.448 ± 0.0039
N	13.761 ± 0.6831	33.147 ± 1.3734	36.768 ± 2.3979
A	1.1357 ± 0.0971	1.1508 ± 0.0995	1.7673 ± 0.4553
g	0.0123 ± 0.0001	0.0153 ± 0.0001	0.0249 ± 0.0002
a	0.1120 ± 0.0176	0.3418 ± 0.0223	0.4713 ± 0.0449
Model II	25°C	27°C	37°C
α	0.833 ± 0.0315	1.010 ± 0.0402	1.151 ± 0.0730
p	0.444 ± 0.0022	0.450 ± 0.0023	0.446 ± 0.0034
N	15.222 ± 0.6791	35.317 ± 1.4840	38.222 ± 1.6418
A	1.5820 ± 0.1152	1.9551 ± 0.1511	3.2246 ± 0.4863
g	0.0124 ± 0.0002	0.0153 ± 0.0002	0.0251 ± 0.0003
a	0.1074 ± 0.0074	0.2883 ± 0.0196	0.3561 ± 0.0430
ν	156.75 ± 5.0945	206.97 ± 7.4727	229.25 ± 9.8141

The parameters p, α, N, A are determined by fitting the experimental to the theoretical cell size distribution. The parameters g and a are determined by fitting the experimental to the theoretical doubling time distribution. Two theoretical models are used: the model with deterministic partitioning (model I) and the model with stochastic partitioning (model II). For model II, once the parameter p is estimated, the sample size parameter ν in Equation (1) can be inferred by fitting the partition ratio data to a beta distribution. The estimation error for each parameter was computed using bootstrap. Specifically, we performed parameter inference 50 times; for each estimation, the theoretical model was fitted to the data of 30 randomly selected cell lineages. The estimation error was then calculated as the standard deviation over 50 repeated samplings. Here, the bootstrap technique is used because each cell lineage was only measured for 70 generations, and the data of a single lineage are insufficient to estimate all the parameters accurately.

lineages at that temperature. Stochastic simulations based on model II with the parameters estimated in Table 3 show that the correlation coefficients between V_b and V_d for the three growth conditions are 0.564, 0.546, and 0.495, respectively, and the correlation coefficients between V_b and $V_b{}'$ are 0.506, 0.447, and 0.396, respectively. Clearly, our estimated parameters capture the size correlations between mother and daughter cells very well.

To further evaluate the performance of our model, we also examine the doubling time correlations between mother and daughter cells. Experimentally, the correlation coefficients between two successive cell cycle durations for the three growth conditions are −0.109 ± 0.109, −0.117 ± 0.114, and −0.152 ± 0.111, respectively, whereas simulations based on model II with the estimated parameters predict the correlation coefficients to be −0.185, −0.160, and −0.173, respectively. Both theory and experiments reveal a negative correlation between successive doubling times, and their values do not differ too much. This again verifies the effectiveness of our model and our parameter inference approach.

Typically, a mother cell divides into two daughters that are different in size due to stochasticity in partitioning and possible asymmetric cell division Nieto-Acuña et al. (2020). Note that the data of cell sizes just before division and just after division, V_d and $V_b{}'$, can be easily extracted from the cell lineage data and thus the parameter p can be estimated as the mean partition ratio $\langle V_b{}'/V_d \rangle$. An interesting characteristic implied by the *E. coli* data is that at cell division, the smaller daughter is always tracked with the mean partition ratio p being estimated to be about 0.46 for all the three growth conditions (0.459 ± 0.040 for cells at 25°C, 0.461 ± 0.039 for cells at 27°C, and 0.464 ± 0.034 for cells at 37°C).

Recall that in our estimation procedure, we do not use the information of V_d and $V_b{}'$ to determine the parameter p; rather, we infer p by an optimal fit of the experimental to the theoretical cell size distribution. The estimate of p in Table 3 is 0.44–0.45 for the three temperatures, which is slightly lower than the value of 0.46 estimated using V_d and $V_b{}'$. The reason of this discrepancy is that after cell division, over 60% generations have a small but sharp decline in the cell size (see Figure 6C for the cell size dynamics of five typical

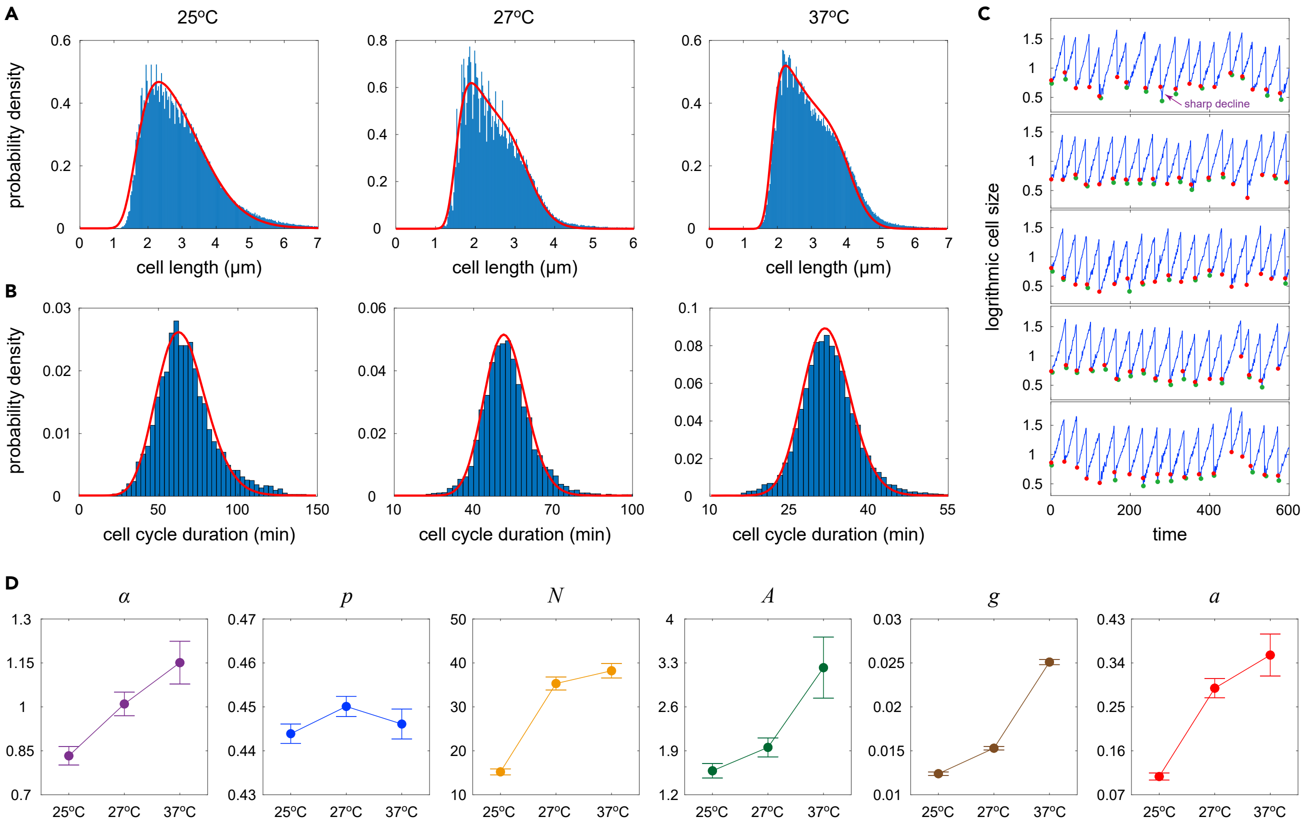


Figure 6. Fitting the experimental cell size and doubling time distributions to theory

(A) Experimental cell size distributions (blue bars) at the three temperatures and their optimal fitting to model II (red curve) where partitioning is stochastic. Here, the theoretical distribution is computed using Equation (20).

(B) Experimental doubling time distributions (blue bars) at the three temperatures and their optimal fitting to model II (red curve), where the theoretical distribution is computed using stochastic simulations.

(C) Five typical trajectories of cell size dynamics for cells at 37°C. The red dots show the cell sizes just after division, and the green dots show the minimal cell sizes in each generation. For over 60% generations, there is a small abrupt decline in the cell size after division, shown as the sharp drop from a red dot to a green dot.

(D) Means and standard deviations of all model parameters estimated by fitting the data to model II. The parameter values can be found in Table 3.

cell lineages with the red dots being the cell sizes just after division and the green dots being the minimal cell sizes in each generation; the small sharp drop in the cell size after division is shown as the transition from a red dot to a green dot). Therefore, the realistic effective partition ratio should be computed using the green dots rather than the red dots. This explains why the parameter p estimated in Table 3 is lower than the mean partition ratio $\langle V_b'/V_d \rangle$.

A possible explanation of these abrupt declines is that the green dots are actually the true beginning of a new cell cycle and the red dots correspond to an intermediate time point during cell division. In our model, division is modeled as an instantaneous transition from stage N to stage 1. However, cell division is never instantaneous in real systems. This finite division time effect may cause the abrupt drops observed in the lineage data. To verify the estimate of p, we compute the ratio of the size associated with a green dot to the division size in the previous generation, which can be viewed as the realistic partition ratio. The mean of the realistic partition ratio is estimated to be 0.44–0.45 for all the three temperatures (0.445 ± 0.038 for cells at 25°C, 0.442 ± 0.034 for cells at 27°C, and 0.450 ± 0.032 for cells at 37°C), which are very close to the estimates of p in Table 3.

A natural question is whether the lineage data used here can be described by simpler models with fewer parameters. In fact, the models in many previous papers can be viewed as special cases of our model. In Osella et al. (2014); Nieto-Acuna et al. (2019); Totis et al. (2020a), the authors focus on the special case of only one cell cycle stage (N = 1) and the majority of previous papers focus on the special case of symmetric division ($p = 0.5$) Kohram et al. (2021). Note that symmetric division mentioned here means that the mean partition ratio equals 0.5 and we allow the partitioning to be stochastic. To see whether the data studied here can be reproduced by simpler models, we fit the cell size distribution to the model with $N = 1$ as well as the model with $p = 0.5$ (Figure S2). Both models fail to capture the unusual shape of the cell cycle distribution. This suggests that our model seems to be minimal in order to describe real lineage data.

DISCUSSION

In this work, we have analytically derived the cell size distribution of measurements obtained from a cell lineage. We have solved two models. The first model assumes that (i) the birth size is a fixed (generation independent) fraction of the division size in the last generation; (ii) the cell grows exponentially between birth and division events where the growth rate is a generation independent constant; (iii) the length of the cell cycle is stochastic; (iv) size homeostasis is enforced by timer-like, sizer-like, or adder strategies. A second model was also solved which relaxes the assumption (i) above, namely it allows for a stochastic ratio of the birth to division size.

The main features of the experimental cell size distribution in *E. coli*, namely a fast increase in the size count for small cells, a slow decay for moderately large cells, and a fast decay for large cells, are reproduced by the analytical solution of both models when the parameters N and α are large enough; this implies that these features emerge when the variability in the cell cycle duration is not too large and when adder or sizer-like mechanisms enforce size homeostasis. We also find that noise in partitioning at cell division (noise in the ratio of the birth to division size) has a considerable influence on the shape of the cell size distribution, whereas noise in the growth rate hardly exerts any influence; this is in agreement with an earlier moment-based study Modi et al. (2017).

Our theory predicts that large cell cycle duration variability, timer-like division strategy, and tracking the smaller daughter at division lead to larger skewness and coefficient of variation of the cell size distribution. We have furthermore shown that (i) the distribution of cell cycle duration that emerges from our model is well approximated by a gamma distribution that has been measured experimentally for many cell types Golubev (2016); (ii) if cells divide asymmetrically, they are tracked randomly after division, and if cell cycle duration variability is intermediate or low, then the cell size distribution is multimodal.

Lastly, we have shown that the theoretical distributions provide an excellent fit to the experimental *E. coli* cell size and doubling time distributions reported in Tanouchi et al. (2017) for three different growth conditions. This match provides support for the implicit assumption of our model that the speed of the cell cycle (the transition rate between effective stages) monotonically increases with the cell volume and specifically has a power law dependence on the cell volume. Note that while this law is compatible with certain biophysical mechanisms (as discussed earlier in the model specification section), it can also be seen as phenomenological means to model cell size homeostasis; in fact more generally and beyond the context of our model, the usage of kinetic rates with power laws has found widespread applications in the effective modeling of complex biochemical kinetics in cells Savageau and Voit (1987). Finally, based on the matching of the experimental to the theoretical cell size and doubling time distributions, we have estimated all the model parameters directly from *E. coli* cell lineage data and found that the regulation strength α exhibits a weak increase with temperature. The estimated values of α (using model II, the most accurate model in this paper) ranging between 0.8 and 1.2 confirm the previous results that some *E. coli* strains use the adder strategy to achieve size homeostasis Tanouchi et al. (2015); Wallden et al. (2016). Simulations with the inferred parameters using distribution matching also captured the size and doubling time correlations between mother and daughter cells—this provides further evidence of the accuracy of the deduced model.

In previous studies, simpler models of cell size regulation with fewer parameters have been proposed Amir (2014); Osella et al. (2014); Vargas-García and Singh (2018); Nieto-Acuna et al. (2019); Totis et al. (2020a); Nieto et al. (2020a); Lin and Amir (2020). Here, we have shown that inference using the constraints of previous models such as $N = 1$ or $p = 0.5$ do not reproduce the correct shape of the lineage cell size distribution (Figure S2). To gain more insights, we compare our model with the classical model proposed in Amir (2014). In that paper, the doubling time in a particular generation is assumed to be a function of the birth size. Under this assumption, the model is not Markovian; this is because given an arbitrary time t, the evolution of the system after time t not only depends on the cell size at that time but also depends on the cell sizes in the past (since the birth time is generally smaller than time t). In our work, by assuming multiple cell cycle stages and the coupling between the size and the rate of cell cycle progression, we are able to model cell size dynamics as a Markov process. By using the powerful tool of Markov processes, we manage to compute the exact solution of the cell size distribution of lineage measurements, which is difficult for a non-Markovian model like in Amir (2014). In addition, in Amir (2014), (i) the partitioning at cell division is assumed to be symmetric (the mean partition ratio p is assumed to be 0.5) and (ii) in one of the two models considered there, the doubling time distribution is assumed to be Gaussian. In naturally occurring systems, the partitioning at cell division is often asymmetric Shi et al. (2020). Here, we have shown that asymmetry and stochasticity in partitioning will greatly affect the distribution shape and assuming symmetric division will lead to large errors in data fitting. Note also that generally the doubling time distribution is not Gaussian (while it looks like being normal in Figure 6B for *E. coli*, it is not for most organisms Golubev (2016); Yates et al. (2017); Chao et al. (2019); Perez-Carrasco et al. (2020); Gavagnin et al. (2020)), rather it is right skewed and well approximated by a gamma or Erlang distribution which our model can capture but the method in Amir (2014) cannot.

Concluding, the major advance in our work is the analytic derivation of the cell size distribution of lineage measurements, while previous studies focused more on population measurements Xia et al. (2020); Xia and Chou (2021) or moments of lineage measurements. The advantages of the analytical distribution are (i) the ease and speed with which one can explore the dependence of cell size statistics on parameters across large swathes of parameter space (compared to stochastic simulations) and (ii) the reliable estimation of parameters from data based on distribution matching which is generally much more robust than moment-based estimation Munsky et al. (2018); Öcal et al. (2019). The present model presents a framework onto which one can build further biological realism; current research work aims to extend the model to include gene expression and its correlation to cell size resulting in concentration homeostasis of mRNAs and proteins Padovan-Merhar et al. (2015); Shahrezaei and Marguerat (2015); Vargas-Garcia et al. (2018); Bertaux et al. (2018); Lin and Amir (2018); Cao and Grima (2020).

Limitations of the study

A major assumption of our model is that cell growth is exponential. While this is a common assumption and holds for various cell types, it is not universal. Hence, the present model cannot, for example, predict the cell size distributions in *Schizosaccharomyces pombe* (fission yeast) where the increase of cell size with time after birth is non-exponential Nobs and Maerkl (2014); Nakaoka and Wakamoto (2017).

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ramon Grima (ramon.grima@ed.ac.uk).

Material availability

This study did not generate new unique reagents.

Data and code availability

All data needed to evaluate the conclusions in the paper are present in the paper and in Tanouchi et al. (2017).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102220.

ACKNOWLEDGMENTS

C.J. acknowledges support from the NSAF grant in National Natural Science Foundation of China with grant No. U1930402. A.S. is supported by the National Institute of Health Grant 1R01GM126557. 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, A.S, and R.G interpreted the theoretical results. C.J and R.G jointly wrote the manuscript with input from A.S.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 28, 2020
Revised: January 29, 2021
Accepted: February 17, 2021
Published: March 19, 2021

REFERENCES

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

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

Bertaux, F., Marguerat, S., and Shahrezaei, V. (2018). Division rate, cell size and proteome allocation: impact on gene expression noise and implications for the dynamics of genetic circuits. R. Soc. Open Sci. *5*, 172234.

Brenner, N., Braun, E., Yoney, A., Susman, L., Rotella, J., and Salman, H. (2015). Single-cell protein dynamics reproduce universal fluctuations in cell populations. Eur. Phys. J. E *38*, 102.

Cadart, C., Monnier, S., Grilli, J., Sáez, P.J., Srivastava, N., Attia, R., Terriac, E., Baum, B., Cosentino-Lagomarsino, M., and Piel, M. (2018). Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. Nat. Commun. *9*, 1–15.

Campos, M., Surovtsev, I.V., Kato, S., Paintdakhi, A., Beltran, B., Ebmeier, S.E., and Jacobs-Wagner, C. (2014). A constant size extension drives bacterial cell size homeostasis. Cell *159*, 1433–1446.

Cao, Z., and Grima, R. (2020). Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. Proc. Natl. Acad. Sci. U S A *117*, 4682–4692.

Chandler-Brown, D., Schmoller, K.M., Winetraub, Y., and Skotheim, J.M. (2017). The adder phenomenon emerges from independent control of pre-and post-start phases of the budding yeast cell cycle. Curr. Biol. *27*, 2774–2783.

Chao, H.X., Fakhreddin, R.I., Shimerov, H.K., Kedziora, K.M., Kumar, R.J., Perez, J., Limas, J.C., Grant, G.D., Cook, J.G., Gupta, G.P., et al. (2019). Evidence that the human cell cycle is a series of uncoupled, memoryless phases. Mol. Syst. Biol. *15*, e8604.

Conlon, I., and Raff, M. (2003). Differences in the way a mammalian cell and yeast cells coordinate cell growth and cell-cycle progression. J. Biol. *2*, 7.

Crane, M.M., Clark, I.B., Bakker, E., Smith, S., and Swain, P.S. (2014). A microfluidic system for studying ageing and dynamic single-cell responses in budding yeast. PLoS One *9*, e100042.

García-García, R., Genthon, A., and Lacoste, D. (2019). Linking lineage and population observables in biological branching processes. Phys. Rev. E *99*, 042413.

Gavagnin, E., Vittadello, S.T., Guanasingh, G., Haass, N.K., Simpson, M.J., Rogers, T., and Yates, C.A. (2020). Synchronised Oscillations in Growing Cell Populations Are Explained by Demographic Noise. bioRxiv. https://doi.org/10.1101/2020.03.13.987032.

Ghusinga, K.R., Vargas-Garcia, C.A., and Singh, A. (2016). A mechanistic stochastic framework for regulating bacterial cell division. Sci. Rep. *6*, 1–9.

Godin, M., Delgado, F.F., Son, S., Grover, W.H., Bryan, A.K., Tzur, A., Jorgensen, P., Payer, K., Grossman, A.D., Kirschner, M.W., et al. (2010). Using buoyant mass to measure the growth of single cells. Nat. Methods *7*, 387–390.

Golubev, A. (2016). 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.

Jia, C., and Grima, R. (2020). Frequency Domain Analysis of Fluctuations of mRNA and Protein Copy Numbers within a Cell Lineage: Theory and Experimental Validation. bioRxiv. https://doi.org/10.1101/2020.09.23.309724.

Jia, C., Zhang, M.Q., and Qian, H. (2019). Analytic Theory of Stochastic Oscillations in Single-Cell Gene Expression. arXiv, arXiv:1909.09769.

Jun, S., and Taheri-Araghi, S. (2015). Cell-size maintenance: universal strategy revealed. Trends Microbiol. *23*, 4–6.

Kohram, M., Vashistha, H., Leibler, S., Xue, B., and Salman, H. (2021). Bacterial growth control mechanisms inferred from multivariate statistical analysis of single-cell measurements. Curr. Biol. *31*, 1–10.

Lin, J., and Amir, A. (2018). Homeostasis of protein and mRNA concentrations in growing cells. Nat. Commun. *9*, 1–11.

Lin, J., and Amir, A. (2020). From single-cell variability to population growth. Phys. Rev. E *101*, 012401.

Lin, Y.T., and Buchler, N.E. (2018). Efficient analysis of stochastic gene dynamics in the non-adiabatic regime using piecewise deterministic Markov processes. J. R. Soc. Interface *15*, 20170804.

Marshall, W.F., Young, K.D., Swaffer, M., Wood, E., Nurse, P., Kimura, A., Frankel, J., Wallingford, J., Walbot, V., Qu, X., et al. (2012). What determines cell size? BMC Biol. *10*, 1–22.

Modi, S., Vargas-Garcia, C.A., Ghusinga, K.R., and Singh, A. (2017). Analysis of noise mechanisms in cell-size control. Biophys. J. *112*, 2408–2418.

Munsky, B., Li, G., Fox, Z.R., Shepherd, D.P., and Neuert, G. (2018). Distribution shapes govern the discovery of predictive models for gene regulation. Proc. Natl. Acad. Sci. U S A *115*, 7533–7538.

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

Nieto, C., Arias-Castro, J., Sánchez, C., Vargas-García, C., and Pedraza, J.M. (2020a). Unification of cell division control strategies through continuous rate models. Phys. Rev. E *101*, 022401.

Nieto, C., Vargas-Garcia, C., and Pedraza, J.M. (2020b). Continuous Rate Modelling of Bacterial Stochastic Size Dynamics. bioRxiv. https://doi.org/10.1101/2020.09.29.319251.

Nieto-Acuña, C., Arias-Castro, J.C., Vargas-García, C., Sánchez, C., and Pedraza, J.M. (2020). Correlation between protein concentration and bacterial cell size can reveal mechanisms of gene expression. Phys. Biol. *17*, 045002.

Nieto-Acuna, C.A., Vargas-Garcia, C.A., Singh, A., and Pedraza, J.M. (2019). Efficient computation of stochastic cell-size transient dynamics. BMC Bioinformatics *20*, 1–6.

Nobs, J.-B., and Maerkl, S.J. (2014). Long-term single cell analysis of S. pombe on a microfluidic microchemostat array. PLoS One *9*, e93466.

Öcal, K., Grima, R., and Sanguinetti, G. (2019). Parameter estimation for biochemical reaction networks using Wasserstein distances. J. Phys. A Math. Theor. *53*, 034002.

Osella, M., Nugent, E., and Lagomarsino, M.C. (2014). Concerted control of Escherichia coli cell division. Proc. Natl. Acad. Sci. U S A *111*, 3431–3435.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352.

Patterson, J.O., Rees, P., and Nurse, P. (2019). Noisy cell-size-correlated expression of cyclin b drives probabilistic cell-size homeostasis in fission yeast. Curr. Biol. *29*, 1379–1386.

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

Realpe-Gomez, J., Galla, T., and McKane, A.J. (2012). Demographic noise and piecewise deterministic Markov processes. Phys. Rev. E *86*, 011137.

Robert, L., Ollion, J., Robert, J., Song, X., Matic, I., and Elez, M. (2018). Mutation dynamics and fitness effects followed in single cells. Science *359*, 1283–1286.

Savageau, M.A., and Voit, E.O. (1987). Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. Math. Biosci. *87*, 83–115.

Sekar, K., Rusconi, R., Sauls, J.T., Fuhrer, T., Noor, E., Nguyen, J., Fernandez, V.I., Buffing, M.F., Berney, M., Jun, S., et al. (2018). Synthesis and degradation of FtsZ quantitatively predict the first cell division in starved bacteria. Mol. Syst. Biol. *14*, e8623.

Shahrezaei, V., and Marguerat, S. (2015). Connecting growth with gene expression: of noise and numbers. Curr. Opin. Microbiol. *25*, 127–135.

Shi, C., Chao, L., Proenca, A.M., Qiu, A., Chao, J., and Rang, C.U. (2020). Allocation of gene products to daughter cells is determined by the age of the mother in single Escherichia coli cells. Proc. R. Soc. B *287*, 20200569.

Si, F., Le Treut, G., Sauls, J.T., Vadia, S., Levin, P.A., and Jun, S. (2019). Mechanistic origin of cell-size control and homeostasis in bacteria. Curr. Biol. *29*, 1760–1770.

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

Taheri-Araghi, S., Bradde, S., Sauls, J.T., Hill, N.S., Levin, P.A., Paulsson, J., Vergassola, M., and Jun, S. (2015). Cell-size control and homeostasis in bacteria. Curr. Biol. *25*, 385–391.

Tanouchi, Y., Pai, A., Park, H., Huang, S., Buchler, N.E., and You, L. (2017). Long-term growth data of Escherichia coli at a single-cell level. Sci. Data *4*, 1–5.

Tanouchi, Y., Pai, A., Park, H., Huang, S., Stamatov, R., Buchler, N.E., and You, L. (2015). A noisy linear map underlies oscillations in cell size and gene expression in bacteria. Nature *523*, 357–360.

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

Thomas, P. (2018). Analysis of cell size homeostasis at the single-cell and population level. Front. Phys. (Lausanne) *6*, 64.

Totis, N., Acuna, C.A.N., Vargas-Garcia, C.A., Kuper, A., Singh, A., and Waldherr, S. (2020a). Cell Size Statistics in Cell Lineages and Population Snapshots with Different Growth Regimes and Division Strategies. bioRxiv. https://doi.org/10.1101/2020.05.15.094698.

Totis, N., Nieto, C., Küper, A., Vargas-García, C., Singh, A., and Waldherr, S. (2020b). A population-based approach to study the effects of growth and division rates on the dynamics of cell size statistics. IEEE Control Syst. Lett. *5*, 725–730.

Vargas-Garcia, C.A., Ghusinga, K.R., and Singh, A. (2018). Cell size control and gene expression homeostasis in single-cells. Curr. Opin. Syst. Biol. *8*, 109–116.

Vargas-García, C.A., and Singh, A. (2018). Elucidating cell size control mechanisms with stochastic hybrid systems. In 2018 IEEE Conference on Decision and Control (CDC) (IEEE), pp. 4366–4371.

Wallden, M., Fange, D., Lundius, E.G., Baltekin, Ö., and Elf, J. (2016). The synchronization of replication and division cycles in individual E. coli cells. Cell *166*, 729–739.

Weart, R.B., and Levin, P.A. (2003). Growth rate-dependent regulation of medial FtsZ ring formation. J. Bacteriol. *185*, 2826–2834.

Xia, M., and Chou, T. (2021). Kinetic Theory for Structured Populations: Application to Stochastic Sizer-Timer Models of Cell Proliferation. arXiv, arXiv:2101.03470.

Xia, M., Greenman, C.D., and Chou, T. (2020). PDE models of adder mechanisms in cellular proliferation. SIAM J. Appl. Math. *80*, 1307–1335.

Yates, C.A., Ford, M.J., and Mort, R.L. (2017). A multi-stage representation of cell proliferation as a Markov process. Bull. Math. Biol. *79*, 2905–2928.

Zopf, C., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-cycle dependence of transcription dominates noise in gene expression. PLoS Comput. Biol. *9*, e1003161.

iScience, Volume *24*

Supplemental information

Cell size distribution of lineage data: analytic results and parameter inference

Chen Jia, Abhyudai Singh, and Ramon Grima

Supplementary figures

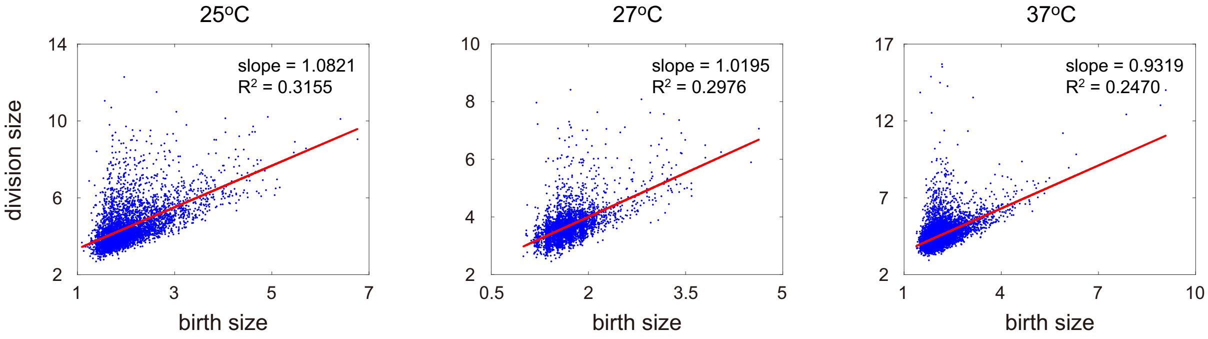


Figure S1. **Scatter plot of birth sizes versus division sizes for the three growth conditions, related to Table 3 and Figure 6.** The red line represents the regression line with its slope and the R^2 value shown in the upper-right corner.

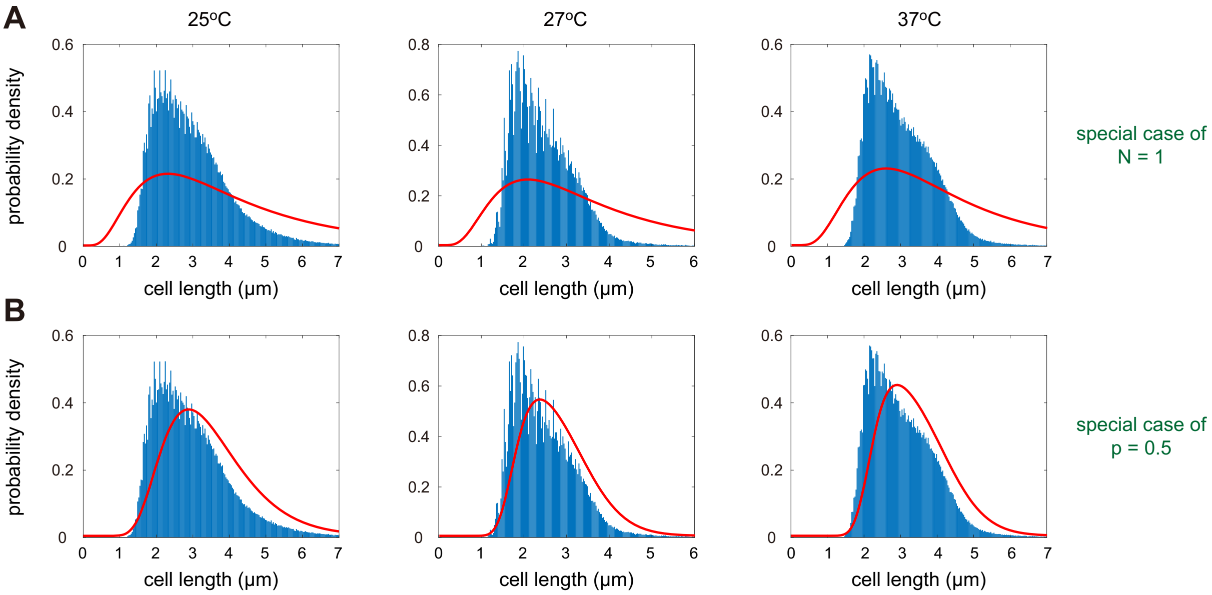


Figure S2. **Fitting the experimental cell size distribution to simpler models, related to Figure 6.** (**A**) Simpler model with only one cell cycle stage ($N = 1$). (**B**) Simpler model with symmetric division ($p = 0.5$). In (A), we use the correlation coefficient between the birth and division sizes (see Eq. (23) in the main text) to infer the parameter p. In (A),(B), we use the mean of the generalized added size to infer the parameter A (see Eq. (22) in the main text).

Transparent methods

1 Distribution of the generalized added size

Let V_b and V_d denote the cell sizes at birth and at division in a particular generation, respectively. In the main text, we have stated that the generalized added size $V_d^\alpha - V_b^\alpha$ is Erlang distributed with shape parameter N and mean A, where $A = N\alpha g/a$. To see this, note that when V_b is fixed, the cell size in this generation is given by $V(t) = V_b e^{gt}$. Since the transition rate from one stage to the next at time t is equal to $aV(t)^\alpha$, the distribution of the transition time T is given by

$$\mathbb{P}(T > t) = e^{-\int_0^t aV(s)^\alpha ds} = e^{-\int_0^t aV_b^\alpha e^{\alpha gs} ds} = e^{-\frac{aV_b^\alpha}{\alpha g}(e^{\alpha gt}-1)}. \tag{1}$$

This shows that

$$\mathbb{P}(V_b^\alpha(e^{\alpha gT} - 1) > t) = e^{-\frac{at}{\alpha g}}.$$

Hence $V_b^\alpha(e^{\alpha gT} - 1)$ is exponentially distributed with mean $\alpha g/a$. Note that $V_b^\alpha(e^{\alpha gT} - 1)$ is the increment of the αth power of the cell size in a particular cell cycle stage. Therefore, the total increment of the αth power of the cell size in each generation is the independent sum of N exponentially distributed random variables with mean $\alpha g/a$. This shows that $V_d^\alpha - V_b^\alpha$ has an Erlang distribution with shape parameter N and mean $N\alpha g/a = A$.

2 Cell size distribution under deterministic partitioning

2.1 General case

Here we compute the analytical distribution of the cell size. For simplicity, we first focus on the case of deterministic partitioning at cell division, i.e. $\mu(y) = \delta(y + \log p)$. In this case, Eq. (4) in the main text reduces to

$$\begin{aligned}\partial_t p_k &= -\, g\partial_x p_k + ae^{\alpha x}p_{k-1} - ae^{\alpha x}p_k, \quad 2 \le k \le N,\\ \partial_t p_1 &= -\, g\partial_x p_1 + ap^{-\alpha}e^{\alpha x}p_N(x - \log p) - ae^{\alpha x}p_1.\end{aligned} \tag{2}$$

To proceed, for each cell cycle stage k, we introduce the moment generating function

$$F_k(\lambda) = \int_{-\infty}^{\infty} p_k(x)e^{\lambda x}dx, \quad F(\lambda) = \int_{-\infty}^{\infty} p(x)e^{\lambda x}dx,$$

where $p(x) = \sum_{k=1}^N p_k(x)$ is the probability density function of the logarithmic cell size. Then Eq. (2) can be converted to the following differential equations:

$$\begin{aligned}\partial_t F_k(\lambda) &= g\lambda F_k(\lambda) - aF_k(\lambda+\alpha) + aF_{k-1}(\lambda+\alpha), \quad 2 \le k \le N,\\ \partial_t F_1(\lambda) &= g\lambda F_1(\lambda) - aF_1(\lambda+\alpha) + ap^\lambda F_N(\lambda+\alpha).\end{aligned}$$

At the steady state, we have

$$\begin{aligned}g\lambda F_k(\lambda) - aF_k(\lambda+\alpha) + aF_{k-1}(\lambda+\alpha) = 0, \quad 2 \le k \le N,\\ g\lambda F_1(\lambda) - aF_1(\lambda+\alpha) + ap^\lambda F_N(\lambda+\alpha) = 0.\end{aligned} \tag{3}$$

Note that the first row of Eq. (3) is recursive with respect to k, which indicates that F_{k-1} can be represented by F_k for each $2 \leq k \leq N$. Hence F_k can be represented by F_N as

$$F_k(\lambda) = \sum_{l=0}^{N-k} C_{N-k,l} \left(-\frac{A}{\alpha N}\right)^l (\lambda - \alpha) \cdots (\lambda - l\alpha) F_N(\lambda - l\alpha), \quad 1 \leq k \leq N. \tag{4}$$

Summing over k in the above equation yields

$$F(\lambda) = \sum_{k=1}^{N} F_k(\lambda) = \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(-\frac{A}{\alpha N}\right)^l (\lambda - \alpha) \cdots (\lambda - l\alpha) F_N(\lambda - l\alpha). \tag{5}$$

Inserting Eq. (4) into the second row of Eq. (3) shows that F_N is the solution to the following functional equation:

$$p^{\lambda-\alpha} F_N(\lambda) = \sum_{l=0}^{N} C_{N,l} \left(-\frac{A}{\alpha N}\right)^l (\lambda - \alpha) \cdots (\lambda - l\alpha) F_N(\lambda - l\alpha). \tag{6}$$

Complex computations show that the solution to this functional equation can be computed explicitly as follows.

Theorem 1. The solution to Eq. (6) is given by

$$F_N(\lambda) = \frac{KA}{N} \Gamma\left(1 - \frac{\lambda}{\alpha}\right)^{-1} \int_0^\infty u^{-\frac{\lambda}{\alpha}} \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du, \tag{7}$$

where

$$a_N(u) = \left(1 + \frac{Au}{N}\right)^{-N}$$

is a function of u and

$$K = \left[\int_0^\infty \frac{1}{u}(a_N(u)^{-1} - 1) \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du\right]^{-1}$$

is a normalization constant.

Proof. The proof of this theorem is highly nontrivial and will be given in Section 6. □

Inserting Eq. (7) into Eq. (5) gives the following explicit expression for the moment generating function:

$$F(\lambda) = K \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(\frac{A}{N}\right)^{l+1} \Gamma\left(1 - \frac{\lambda}{\alpha}\right)^{-1} \int_0^\infty u^{l-\frac{\lambda}{\alpha}} \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du. \tag{8}$$

For convenience, we also introduction the characteristic function

$$G(\lambda) = \int_{-\infty}^{\infty} p(x) e^{i\lambda x} dx,$$

which is nothing but the inverse Fourier transform of $p(x)$. Clearly, the moment generating function and the characteristic function are related by $G(\lambda) = F(i\lambda)$. Thus we finally obtain the following explicit expression of the characteristic function:

$$G(\lambda) = K \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(\frac{A}{N}\right)^{l+1} \Gamma\left(1 - \frac{i\lambda}{\alpha}\right)^{-1} \int_0^\infty u^{l-\frac{i\lambda}{\alpha}} \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du. \tag{9}$$

Since the Fourier transform and inverse Fourier transform are inverses of each other, taking the Fourier transform of the characteristic function $G(\lambda)$ yields the probability density $p(x)$ of the logarithmic cell size. Finally, the probability density of the original cell size $y = e^x$ is given by

$$\tilde{p}(y) = \frac{1}{y} p(\log y).$$

2.2 Case of small cell cycle duration variability

We next focus on the special case where cell cycle duration variability is very small, i.e. $N \gg 1$. In this limit, we have $a_N(u) = e^{-Au}$ and thus we have

$$\prod_{n=0}^{\infty} a_N(p^{\alpha n} u) = \prod_{n=0}^{\infty} e^{-Ap^{\alpha n} u} = e^{-\frac{Au}{1-p^\alpha}}.$$

This implies that

$$\int_0^\infty u^{l-\frac{\lambda}{\alpha}} \prod_{n=0}^{\infty} a_N(p^{\alpha n} u) du = \int_0^\infty u^{l-\frac{\lambda}{\alpha}} e^{-\frac{Au}{1-p^\alpha}} du = \Gamma\left(l+1-\frac{\lambda}{\alpha}\right)\left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}-l-1}.$$

Inserting this equation into Eq. (8) yields

$$F(\lambda) = K \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(\frac{A}{N}\right)^{l+1} \Gamma\left(1-\frac{\lambda}{\alpha}\right)^{-1} \Gamma\left(l+1-\frac{\lambda}{\alpha}\right)\left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}-l-1}.$$

By virtue of the fact that

$$\frac{\Gamma(l+x)}{\Gamma(x)} = x(x+1)\cdots(x+l-1) = (x)_l,$$

we obtain

$$F(\lambda) = K \left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}} \sum_{k=0}^{N-1} \sum_{l=0}^{k} C_{k,l} \left(\frac{1-p^\alpha}{N}\right)^{l+1} \left(1-\frac{\lambda}{\alpha}\right)_l,$$

where $(x)_l = x(x+1)\cdots(x+l-1)$ is the Pochhammer symbol. Moreover, using the hockey-stick identity

$$\sum_{k=1}^{N-1} C_{k,l} = C_{N,l+1},$$

we obtain

$$\begin{aligned} F(\lambda) &= K \left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}} \sum_{l=0}^{N-1} \sum_{k=l}^{N-1} C_{k,l} \left(\frac{1-p^\alpha}{N}\right)^{l+1} \left(1-\frac{\lambda}{\alpha}\right)_l \\ &= K \left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}} \sum_{l=0}^{N-1} C_{N,l+1} \left(\frac{1-p^\alpha}{N}\right)^{l+1} \left(1-\frac{\lambda}{\alpha}\right)_l \\ &= \frac{K}{N} \left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}} \sum_{l=0}^{N-1} \frac{C_{N-1,l}}{l+1} \left(\frac{1-p^\alpha}{N}\right)^{l+1} \left(1-\frac{\lambda}{\alpha}\right)_l. \end{aligned}$$

In the limit of $N \to \infty$, we have

$$\sum_{l=0}^{N-1} \frac{C_{N-1,l}}{l+1} \left(\frac{1-p^\alpha}{N}\right)^{l} \left(1-\frac{\lambda}{\alpha}\right)_l = \frac{\alpha(1-p^\lambda)}{\lambda(1-p^\alpha)},$$

This identity, together with the fact that $F(0) = 1$, finally shows that

$$F(\lambda) = \frac{1-p^\lambda}{-\lambda\log p}\left(\frac{A}{1-p^\alpha}\right)^{\frac{\lambda}{\alpha}}. \tag{10}$$

Direct computation shows that $F(\lambda)$ can be rewritten as

$$F(\lambda) = \frac{\bar{V}_d^\lambda - \bar{V}_b^\lambda}{(\log\bar{V}_d - \log\bar{V}_b)\lambda},$$

where

$$\bar{V}_b = p\left(\frac{A}{1-p^\alpha}\right)^{\frac{1}{\alpha}}, \quad \bar{V}_d = \left(\frac{A}{1-p^\alpha}\right)^{\frac{1}{\alpha}}.$$

Replacing λ by $i\lambda$ in the above equation gives the characteristic function

$$G(\lambda) = \frac{\bar{V}_d^{i\lambda} - \bar{V}_b^{i\lambda}}{(\log\bar{V}_d - \log\bar{V}_b)i\lambda}.$$

Taking the Fourier transform of the characteristic function shows that logarithmic cell size has the uniform distribution

$$p(x) = \frac{1}{\log\bar{V}_d - \log\bar{V}_b}I_{[\log\bar{V}_b,\log\bar{V}_d]}(x),$$

and thus the original cell size $y = e^x$ has the following distribution:

$$\tilde{p}(y) = \frac{1}{y}p(\log y) = \frac{1}{(\log\bar{V}_d - \log\bar{V}_b)y}I_{[\bar{V}_b,\bar{V}_d]}(y), \tag{11}$$

where $I_A(x)$ is the indicator function which takes the value of 1 when $x \in A$ and takes the value of 0 otherwise.

3 Cell cycle duration distribution

Let V_b and V_d denote the cell sizes at birth and at division in a particular generation, respectively, and let T denote the corresponding cell cycle duration. Since the cell size growth exponentially, we have

$$V_d = V_b e^{gT}.$$

This shows that

$$T = \frac{1}{g}\log\left(\frac{V_d}{V_b}\right). \tag{12}$$

Recall that $V_d^\alpha - V_b^\alpha$ has an Erlang distribution with shape parameter N and rate parameter N/A. Thus given that $V_b^\alpha = x$, the probability density function of V_d^α is given by

$$\mathbb{P}(V_d^\alpha = y|V_b^\alpha = x) = \frac{N^N}{A^N(N-1)!}(y-x)^{N-1}e^{-\frac{N}{A}(y-x)}, \quad y \geq x.$$

Thus given that $V_b^\alpha = x$, it follows from Eq. (12) that the probability density function of the cell cycle duration T is given by

$$\mathbb{P}(T = t|V_b^\alpha = x) = \frac{\alpha g N^N}{A^N(N-1)!}x^N(e^{\alpha gt}-1)^{N-1}e^{\alpha gt - \frac{N}{A}x(e^{\alpha gt}-1)}.$$

We next compute the distribution of V_b. To this end, let $V_b(k)$ and $V_d(k)$ denote the cell sizes at birth and at division in the kth generation, respectively. Under the assumption of deterministic partitioning, we have $V_b(k+1) = pV_d(k)$ and thus we obtain the recursive equation

$$V_b^\alpha(k+1) = p^\alpha[V_b^\alpha(k) + \Delta_k], \quad k \geq 0,$$

where $\Delta_k = V_d^\alpha(k) - V_b^\alpha(k)$ is the generalized added size in the kth generation, which has an Erlang distribution with shape parameter N and rate parameter N/A. Using the recursive equation repeatedly, we obtain

$$V_b^\alpha(k) = p^{k\alpha}V_b(0)^\alpha + p^{k\alpha}\Delta_0 + p^{(k-1)\alpha}\Delta_1 + \cdots + p^\alpha\Delta_{k-1}. \tag{13}$$

Since $\Delta_0, \Delta_1, \Delta_2, \cdots$ are i.i.d. Erlang distributed random variables with shape parameter N and rate parameter N/A, the Laplace transform of Δ_n is given by

$$\mathbb{E}e^{-\lambda\Delta_n} = \left(1 + \frac{A\lambda}{N}\right)^{-N} = a_N(\lambda).$$

It thus follows from (13) and the independence of $V_b(0), \Delta_0, \Delta_1, \Delta_2, \cdots$ that

$$\mathbb{E}e^{-\lambda V_b^\alpha(k)} = \mathbb{E}e^{-\lambda p^{k\alpha}V_b(0)^\alpha} \prod_{n=1}^{k} a_N(p^{n\alpha}\lambda). \tag{14}$$

Since the distribution of $V_b(k)$ converges to the steady-state distribution of the birth size as $k \to \infty$, taking $k \to \infty$ in Eq. (14) shows that the Laplace transform of V_b^α is given by

$$\mathbb{E}e^{-\lambda V_b^\alpha} = \prod_{n=1}^{\infty} a_N(p^{\alpha n}u) = \prod_{n=1}^{\infty}\left(1 + \frac{Ap^{\alpha n}\lambda}{N}\right)^{-N}. \tag{15}$$

Taking the inverse Laplace transform gives the probability density function of V_b^α, from which is the probability density function of V_b can be obtained. Finally, the distribution of the cell cycle duration T is given by

$$\mathbb{P}(T = t) = \int_0^\infty \mathbb{P}(T = t|V_b^\alpha = x)\mathbb{P}(V_b^\alpha = x)dx. \tag{16}$$

A special case occurs when α is large (strong cell-size control) or when p is small (smaller daughter tracking). Under the large α or small p approximation, the term $p^{\alpha n}$ is negligible for $n \geq 2$ and it suffices to keep only the first term in the infinite product given in Eq. (15). In this case, the laplace transform of V_b^α reduces to

$$\mathbb{E}e^{-\lambda V_b^\alpha} \approx \left(1 + \frac{Ap^\alpha\lambda}{N}\right)^{-N}.$$

Taking the inverse Laplace transform gives the birth size distribution

$$\mathbb{P}(V_b = x) = \frac{N^N x^{\alpha N-1}e^{-\frac{N}{Ap^\alpha}x^\alpha}}{(N-1)!A^N p^{\alpha N}}.$$

Inserting this equation into Eq. (16) gives the doubling time distribution

$$\mathbb{P}(T = t) = \frac{\alpha g(2N-1)!}{p^{\alpha N}[(N-1)!]^2} \cdot \frac{e^{\alpha gt}(e^{\alpha gt} - 1)^{N-1}}{(p^{-\alpha} + e^{\alpha gt} - 1)^{2N}}.$$

4 Correlation between birth and division sizes

Let V_b and V_d denote the cell sizes at birth and at division in a particular generation, respectively, and let V_b' and V_d' denote the the birth and division sizes in the next generation, respectively. We first focus on the correlation between the birth size V_d and the division size V_d. Since the generalized added size $\Delta = V_d^\alpha - V_b^\alpha$ is independent of V_b, we have

$$\mathrm{Cov}(V_b^\alpha, V_d^\alpha) = \mathrm{Cov}(V_b^\alpha, V_b^\alpha + \Delta) = \mathrm{Var}(V_b^\alpha),$$

as well as

$$\mathrm{Var}(V_d^\alpha) = \mathrm{Var}(V_b^\alpha + \Delta) = \mathrm{Var}(V_b^\alpha) + \mathrm{Var}(V_b^\Delta),$$

where $\mathrm{Cov}(X, Y)$ denotes the covariance between random variables X and Y and $\mathrm{Var}(X)$ denotes the variance of X. This shows that

$$\rho(V_b^\alpha, V_d^\alpha) = \frac{\mathrm{Cov}(V_b^\alpha, V_d^\alpha)}{\sqrt{\mathrm{Var}(V_b^\alpha)\mathrm{Var}(V_d^\alpha)}} = \sqrt{\frac{\mathrm{Var}(V_b^\alpha)}{\mathrm{Var}(V_b^\alpha) + \mathrm{Var}(\Delta)}}, \tag{17}$$

where $\rho(X, Y)$ denotes the covariance between X and Y. Since Δ is Erlang distributed with shape parameter N and mean A, we have

$$\mathrm{Var}(\Delta) = \frac{A^2}{N}. \tag{18}$$

Moreover, since we have obtained the Laplace transform of V_b^α, it is easy to compute its variance. In particular, it follows from Eq. (15) that V_b^α has the same law as the independent sum of an infinite number of random variables $X_1, X_2, \cdots$, where X_n is Erlang distributed with shape parameter N and mean $N/Ap^{\alpha n}$. This shows that

$$\mathrm{Var}(V_b^\alpha) = \sum_{n=1}^\infty \frac{A^2 p^{2\alpha n}}{N} = \frac{A^2 p^{2\alpha}}{N(1 - p^{2\alpha})}.$$

Inserting the above two equations into Eq. (17) shows that

$$\rho(V_b^\alpha, V_d^\alpha) = \sqrt{p^{2\alpha}} = p^\alpha.$$

We next focus on the correlation between two successive birth sizes and the correlation between two successive division sizes. Since $V_b' = pV_d$, the correlation coefficient between V_b^α and $V_b'^\alpha$ is exactly the same as that between V_b^α and V_d^α, i.e.

$$\rho(V_b^\alpha, V_b'^\alpha) = \rho(V_b^\alpha, V_d^\alpha) = p^\alpha.$$

Finally, since $V_b' = pV_d$, the correlation coefficient between V_d^α and $V_d'^\alpha$ is exactly the same as that between $V_b'^\alpha$ and $V_d'^\alpha$, i.e.

$$\rho(V_d^\alpha, V_d'^\alpha) = \rho(V_b'^\alpha, V_d'^\alpha) = \rho(V_b^\alpha, V_d^\alpha) = p^\alpha.$$

We next focus on the model with stochastic partitioning. In this case, Eqs. (17) and (18) still hold and thus the key is to compute the variance of V_b^α. Let $R = V_b'/V_d$ be the partition ratio, which is assumed to be beta distributed. Since $V_b' = RV_d$, we have

$$V_b'^\alpha = R^\alpha(V_b^\alpha + \Delta).$$

This shows that $R^\alpha(V_b^\alpha+\Delta)$ and V_b^α have the same distribution. Thus we obtain

$$\mathbb{E}V_b^\alpha = \mathbb{E}R^\alpha(V_b^\alpha+\Delta) = \mathbb{E}R^\alpha\mathbb{E}(V_b^\alpha+\Delta), \tag{19}$$

$$\mathbb{E}V_b^{2\alpha} = \mathbb{E}R^{2\alpha}(V_b^\alpha+\Delta)^2 = \mathbb{E}R^{2\alpha}\mathbb{E}(V_b^\alpha+\Delta)^2, \tag{20}$$

where we have used the fact that the partition ratio R is independent of the birth size V_b and generalized added size Δ. Since R has a beta distribution with mean p and sample size parameter ν, we have

$$\mathbb{E}R^\alpha = \frac{1}{B(p\nu,q\nu)}\int_0^\infty z^{\alpha+p\nu-1}(1-z)^{q\nu-1}dz = \frac{B(\alpha+p\nu,q\nu)}{B(p\nu,q\nu)}.$$

It then follows from Eq. (19) that

$$\mathbb{E}V_b^\alpha = \frac{A\mathbb{E}R^\alpha}{1-\mathbb{E}R^\alpha} = AK_1, \tag{21}$$

where

$$K_1 = \frac{\mathbb{E}R^\alpha}{1-\mathbb{E}R^\alpha} = \frac{B(\alpha+p\nu,q\nu)}{B(p\nu,q\nu)-B(\alpha+p\nu,q\nu)}.$$

Similarly, it follows from Eq. (20) that

$$\mathbb{E}V_b^{2\alpha} = A^2K_2\left(2K_1+1+\frac{1}{N}\right), \tag{22}$$

where

$$K_2 = \frac{\mathbb{E}R^{2\alpha}}{1-\mathbb{E}R^{2\alpha}} = \frac{B(2\alpha+p\nu,q\nu)}{B(p\nu,q\nu)-B(2\alpha+p\nu,q\nu)}.$$

Combining Eqs. (21) and (22) shows that

$$\mathrm{Var}(V_b^\alpha) = \mathbb{E}V_b^{2\alpha} - (\mathbb{E}V_b^\alpha)^2 = A^2K_2\left(2K_1+1+\frac{1}{N}\right) - A^2K_1^2,$$

Inserting this equation into Eq. (17) finally shows that

$$\rho(V_b^\alpha,V_d^\alpha) = \sqrt{\frac{N\left[(2K_1+1)K_2-K_1^2\right]+K_2}{N\left[(2K_1+1)K_2-K_1^2\right]+K_2+1}}.$$

5 Cell size distribution under stochastic partitioning

We next focus on the case of stochastic partitioning at cell division. In this case, Eq. (4) in the main text can be converted to the following differential equations satisfied by the moment generating function:

$$\partial_t F_k(\lambda) = g\lambda F_k(\lambda) - aF_k(\lambda+\alpha) + aF_{k-1}(\lambda+\alpha), \quad 2\le k\le N,$$
$$\partial_t F_1(\lambda) = g\lambda F_1(\lambda) - aF_1(\lambda+\alpha) + a\hat{\mu}(\lambda)F_N(\lambda+\alpha),$$

where

$$\hat{\mu}(\lambda) = \int_0^\infty \mu(y)e^{-\lambda y}dy = \int_0^1 f(x)x^\lambda dx$$

is the Laplace transform of $\mu(y)$. At the steady state, in analogy to the derivation of Eq. (5), we obtain

$$F(\lambda) = \sum_{k=0}^{N-1}\sum_{l=0}^{k} C_{k,l}\left(-\frac{A}{\alpha N}\right)^l(\lambda-\alpha)\cdots(\lambda-l\alpha)F_N(\lambda-l\alpha), \tag{23}$$

where F_N is the solution to the following functional equation:

$$\hat{\mu}(\lambda-\alpha)F_N(\lambda)=\sum_{l=0}^{N}C_{N,l}\left(-\frac{A}{\alpha N}\right)^l(\lambda-\alpha)\cdots(\lambda-l\alpha)F_N(\lambda-l\alpha). \tag{24}$$

To proceed, we defined a new function

$$p(\lambda)=\left(\int_0^1 f(x)x^{\lambda-\alpha}dx\right)^{\frac{1}{\lambda-\alpha}}.$$

With this notation, Eq. (24) can be rewritten as

$$p(\lambda)^{\lambda-\alpha}F_N(\lambda)=\sum_{l=0}^{N}C_{N,l}\left(-\frac{A}{\alpha N}\right)^l(\lambda-\alpha)\cdots(\lambda-l\alpha)F_N(\lambda-l\alpha). \tag{25}$$

Comparing Eqs. (7) and (25), we can see that when the noise in partitioning is small, the solution to the above functional equation is approximately given by

$$F_N(\lambda)=\frac{KA}{N}\Gamma\left(1-\frac{\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{-\frac{\lambda}{\alpha}}\prod_{n=0}^{\infty}a_N(p(\lambda)^{\alpha n}u)du,$$

where the constant K is given by

$$K=\left[\int_0^\infty\frac{1}{u}(a_N(u)^{-1}-1)\prod_{n=0}^{\infty}a_N(p(0)^{\alpha n}u)du\right]^{-1},$$

with $p(0)$ being the limit of $p(\lambda)$ as $\lambda\to 0$. Inserting this equation into Eq. (23) shows that the moment generating function is given by

$$F(\lambda)=K\sum_{k=0}^{N-1}\sum_{l=0}^{k}C_{k,l}\left(\frac{A}{N}\right)^{l+1}\Gamma\left(1-\frac{\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{l-\frac{\lambda}{\alpha}}\prod_{n=0}^{\infty}a_N(p(\lambda)^{\alpha n}u)du.$$

Replacing the variable λ in the the moment generating function by $i\lambda$ yields the characteristic function

$$G(\lambda)=K\sum_{k=0}^{N-1}\sum_{l=0}^{k}C_{k,l}\left(\frac{A}{N}\right)^{l+1}\Gamma\left(1-\frac{i\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{l-\frac{i\lambda}{\alpha}}\prod_{n=0}^{\infty}a_N(p(i\lambda)^{\alpha n}u)du.$$

Taking the Fourier transform of the characteristic function gives the the probability density $p(x)$ of the logarithmic cell size. Then the probability density of the original cell size $y=e^x$ is given by

$$\tilde{p}(y)=\frac{1}{y}p(\log y).$$

6 Proof of Theorem 1

Here we shall give the detailed proof of Theorem 1 in Section 2. To this end, we must revisit the cell size distribution for the adder strategy.

6.1 Cell size distribution for the adder strategy

For the adder strategy ($\alpha = 1$), we can compute the cell size distribution in an alternative way. For simplicity, we consider the case of deterministic partitioning, i.e. $f(z) = \delta(z - p)$. In this case, Eq. (3) in the main text reduces to

$$\begin{aligned} \partial_t \tilde{p}_k(y) &= -\partial_y[gy\tilde{p}_k(y)] + ay\tilde{p}_{k-1}(y) - ay\tilde{p}_k(y), \quad 2 \le k \le N, \\ \partial_t \tilde{p}_1(y) &= -\partial_y[gy\tilde{p}_1(y)] + \frac{ay}{p^2}\tilde{p}_N\left(\frac{y}{p}\right) - ay\tilde{p}_1(y). \end{aligned} \tag{26}$$

To proceed, for each cell cycle stage k, we introduce the Laplace transform

$$H_k(\lambda) = \int_0^\infty \tilde{p}_k(y)e^{-\lambda y}dy, \quad H(\lambda) = \int_0^\infty \tilde{p}(y)e^{-\lambda y}dy,$$

where $\tilde{p}(y) = \sum_{k=1}^N \tilde{p}_k(y)$ is the probability density function of the cell size. Then Eq. (26) can be converted to the following differential equations:

$$\begin{aligned} \partial_t H_k(\lambda) &= (g\lambda + a)\partial_\lambda H_k(\lambda) - a\partial_\lambda H_{k-1}(\lambda), \quad 2 \le k \le N, \\ \partial_t H_1(\lambda) &= (g\lambda + a)\partial_\lambda H_1(\lambda) - a\partial_\lambda H_N(p\lambda). \end{aligned}$$

At the steady state, we have

$$\begin{aligned} &(g\lambda + a)H_k'(\lambda) - aH_{k-1}'(\lambda) = 0, \quad 2 \le k \le N, \\ &(g\lambda + a)H_1'(\lambda) - aH_N'(p\lambda) = 0. \end{aligned} \tag{27}$$

Note that the first row of Eq. (27) is recursive with respect to k, which indicates that H_{k-1}' can be represented by H_k' for each $2 \le k \le N$. Hence H_k' can be represented by H_N' as

$$H_k'(\lambda) = \left(1 + \frac{A'\lambda}{N}\right)^{N-k} H_N'(\lambda), \quad 1 \le k \le N, \tag{28}$$

where $A' = Ng/a$. Inserting this equation into the second row of Eq. (27) shows that H_N' is the solution to the following functional equation:

$$H_N'(\lambda) = \left(1 + \frac{A'\lambda}{N}\right)^{-N} H_N'(p\lambda) = a_N(\lambda)H_N'(p\lambda).$$

Using the above equation repeatedly yields

$$H_N'(\lambda) = H_N'(0)\prod_{n=1}^\infty a_N(p^n\lambda). \tag{29}$$

Inserting the above equation into Eq. (28) and summing over k, we obtain

$$H'(\lambda) = \sum_{k=1}^N H_k'(\lambda) = \frac{N}{A'}H_N'(0)(a_N(\lambda)^{-1} - 1)\prod_{n=1}^\infty a_N(p^n\lambda).$$

Since $H(0) = 1$, integrating the above equation yields

$$H(\lambda) = 1 + \frac{N}{A'}H_N'(0)\int_0^\lambda \frac{1}{u}(a_N(u)^{-1} - 1)\prod_{n=0}^\infty a_N(p^n u)du.$$

Finally, using the fact that $H(\infty) = 0$, we obtain

$$H(\lambda) = 1 - K \int_0^{\lambda} \frac{1}{u}(a_N(u)^{-1} - 1) \prod_{n=0}^{\infty} a_N(p^n u) du,$$

where

$$K = -\frac{N}{A'} H_N'(0) = \left[\int_0^{\infty} \frac{1}{u}(a_N(u)^{-1} - 1) \prod_{n=0}^{\infty} a_N(p^n u) du \right]^{-1}.$$

is a normalization constant. Taking the inverse Laplace transform gives the cell size distribution.

6.2 Detailed proof

We are now in a position to prove Theorem 1.

Proof. For simplicity, we first focus on the adder strategy ($\alpha = 1$). Recall that the probability density functions at a certain stage for the original cell size and the logarithmic cell size are related by

$$\tilde{p}_k(y) = \frac{1}{y} p_k(\log y).$$

Using the change of variables formula, we obtain

$$F_N(\lambda) = \int_{-\infty}^{\infty} p_k(x) e^{\lambda x} dx = \int_0^{\infty} \tilde{p}_k(y) y^{\lambda} dy.$$

Straightforward computations shows that

$$\int_0^{\infty} x^{\lambda-1} e^{-yx} dx = \Gamma(\lambda) y^{-\lambda}.$$

Combining the above two equations shows that

$$\begin{aligned} F_N(-\lambda) &= \int_0^{\infty} \tilde{p}_k(y) y^{-\lambda} dy \\ &= \Gamma(\lambda)^{-1} \int_0^{\infty} \tilde{p}_k(y) dy \int_0^{\infty} x^{\lambda-1} e^{-yx} dx \\ &= \Gamma(\lambda)^{-1} \int_0^{\infty} x^{\lambda-1} dx \int_0^{\infty} \tilde{p}_k(y) e^{-xy} dy \\ &= \Gamma(\lambda)^{-1} \int_0^{\infty} H_k(x) x^{\lambda-1} dx. \end{aligned}$$

It follows from Eq. (29) that

$$H_N(x) = \frac{KA'}{N} \int_x^{\infty} \prod_{n=0}^{\infty} a_N(p^n u) du.$$

Thus we obtain

$$\begin{aligned} F_N(-\lambda) &= \frac{KA'}{N} \Gamma(\lambda)^{-1} \int_0^{\infty} x^{\lambda-1} dx \int_x^{\infty} \prod_{n=0}^{\infty} a_N(p^n u) du \\ &= \frac{KA'}{N} \Gamma(\lambda)^{-1} \int_0^{\infty} \prod_{n=0}^{\infty} a_N(p^n u) du \int_0^{u} x^{\lambda-1} dx \\ &= \frac{KA'}{N} \Gamma(\lambda+1)^{-1} \int_0^{\infty} u^{\lambda} \prod_{n=0}^{\infty} a_N(p^n u) du. \end{aligned}$$

It follows from Eq. (6) that $F_N(\lambda)$ is the solution to the functional equation

$$p^{\lambda-1}F_N(\lambda) = \sum_{l=0}^{N} C_{N,l}\left(-\frac{A'}{N}\right)^l (\lambda-1)\cdots(\lambda-l)F_N(\lambda-l). \tag{30}$$

Thus the solution to the above functional equation can be computed explicitly as

$$F_N(\lambda) = \frac{KA'}{N}\Gamma(1-\lambda)^{-1}\int_0^\infty u^{-\lambda}\prod_{n=0}^{\infty} a_N(p^n u)du. \tag{31}$$

We next focus on the case of a general size control strategy (arbitrary α). In this case, it follows from Eq. (6) that $F_N(\lambda)$ is the solution to the functional equation

$$p^{\lambda-\alpha}F_N(\lambda) = \sum_{l=0}^{N} C_{N,l}\left(-\frac{A}{\alpha N}\right)^l (\lambda-\alpha)\cdots(\lambda-\alpha)F_N(\lambda-\alpha). \tag{32}$$

Comparing the functional forms of Eqs. (30) and (32), it is easy to see that the solution to Eq. (32) can be obtain from the function given in Eq. (31) by replacing λ by λ/α, replacing A' by $A = \alpha A'$, and replacing p by p^α. Thus we finally obtain

$$F_N(\lambda) = \frac{KA}{N}\Gamma\left(1-\frac{\lambda}{\alpha}\right)^{-1}\int_0^\infty u^{-\frac{\lambda}{\alpha}}\prod_{n=0}^{\infty} a_N(p^{\alpha n} u)du. \tag{33}$$

This gives the desired result. □