A large deviation principle linking lineage statistics to fitness in microbial populations

Ethan Levien,[1, *] Trevor GrandPre,[2, *] and Ariel Amir[1]

[1] *School of Engineering and Applied Sciences, Harvard University, Cambridge, MA USA*
[2] *Department of Physics, University of California, Berkeley, CA USA*

(Dated: February 4, 2020)

In exponentially proliferating populations of microbes, the population typically doubles at a rate less than the average doubling time of a single-cell due to variability at the single-cell level. It is known that the distribution of generation times obtained from a single lineage is, in general, insufficient to determine a population's growth rate. Is there an *explicit* relationship between observables obtained from a single lineage and the population growth rate? We show that a population's growth rate can be represented in terms of averages over isolated lineages. This lineage representation is related to a large deviation principle that is a generic feature of exponentially proliferating populations. Due to the large deviation structure of growing populations, the number of lineages needed to obtain an accurate estimate of the growth rate depends exponentially on the duration of the lineages, leading to a non-monotonic convergence of the estimate, which we verify in both synthetic and experimental data sets.

A key determinant of fitness in microbial populations is the population growth rate [1–3]. For organisms such as *Escherichia coli* which undergo binary fission, the exponential growth rate of the population is determined by single-cell properties such as generation time, defined as the time from cell birth to division. In the simplest case where each cell in the population has a generation time of exactly τ_d, the number of cells in the population, denoted $N(t)$, will grow as $N(t) \sim e^{\Lambda t}$, where the exponential growth rate, Λ is given by $\Lambda = \ln(2)/\tau_d$. In reality, any clonal population of bacteria will exhibit a distribution of generation times due to a combination of factors, including intrinsic stochasticity of gene expression [4–7], asymmetric segregation of growth limiting resources at cell division [8–11] and environmental fluctuations [12]. Together these factors will result in a distribution of generation times, $\psi(\tau_d)$, throughout the history of the population. The relationship between this distribution and the population growth rate, Λ, has been the subject of numerous studies. A key result is the *Euler-Lotka equation* [1–3, 13–15],

$$\frac{1}{2} = \int_0^\infty \psi(\tau) e^{-\Lambda\tau} d\tau, \qquad (1)$$

which establishes a link between $\psi(\tau)$ and Λ. Equation (1) is a generalization of a relation originally obtained by Euler and later rediscovered by Lotka [16].

Despite the elegant simplicity of the Euler-Lotka equation, it obscures the underlying relationship between the stochastic dynamics along single lineages and the population growth rate. The reason is that $\psi(\tau_d)$, like Λ, is a property of the population rather than an intrinsic property of individual cells and it is therefore unclear how differences in single cell dynamics are reflected in $\psi(\tau_d)$. Only in the special case where the generation time of a newborn cell is completely uncorrelated with its immediate ancestor, or *mother*, does $\psi(\tau)$ correspond to the distribution of generation times along a single-lineage [1, 14]. In this limit one can, in principle, infer Λ without access to the entire population. In contrast, when generation times are correlated between mother and daughter cells, the distribution of generation times, $f(\tau)$, along a single lineage no longer contains enough information to deduce the growth exponent Λ using equation (1). Such correlations emerge naturally through feedback mechanisms and are required to maintain homeostasis of cell sizes [17, 18]: if intrinsic stochasticity within a cell causes it to grow abnormally large, the daughter cell will tend to have a shorter generation time in order to account for the additional size inherited from its mother. This leads to negative correlations between mother and daughter cells. There are numerous other mechanisms which can generate correlations between mother and daughter cells. For example, environmental fluctuations can induce positive correlations between mother and daughter cells [3, 13].

The discrepancy between the statistics of a lineage and those from the entire population, which determine the population's fitness, raises the question of how to quantify fitness from data obtained from a single lineage, or a collection of independent lineages; see Figure 1. Such data is typically obtained from mother machine experiments [19]. In these experiments, independent lineages are tracked for long periods of times in highly controlled conditions. Mother machine experiments enable detailed measurements of single-cell dynamics that would be impossible in bulk conditions. In contrast, bulk experiments can be used to probe population-level dynamics and measure fitness, but are blind to the physiological details at the microscopic level [19]. Here, we present a lineage representation of the population growth rate that connects the population dynamics to the statistics along a single lineage, or a collection of independent lineages. Our lineage representation reveals that a large deviation principle underlies population growth. This generalizes existing variational principles for the population growth rate (e.g. the results of [20]) to the context where generation times are correlated.

Lineage representation. A lineage-based representation of the population growth rate that is independent of the model specifics can be derived using the *division distribution*, denoted $p_T(n)$, which can be obtained from an exponentially growing population as follows. Suppose a population of cells is grown for a time T and assume that we have access to the generation times of individual cells and the genealogical relationships between cells, as shown in Figure 1. We emphasize that our only assumption is that the population grows exponentially as $T \to \infty$. We can randomly sample a lin-

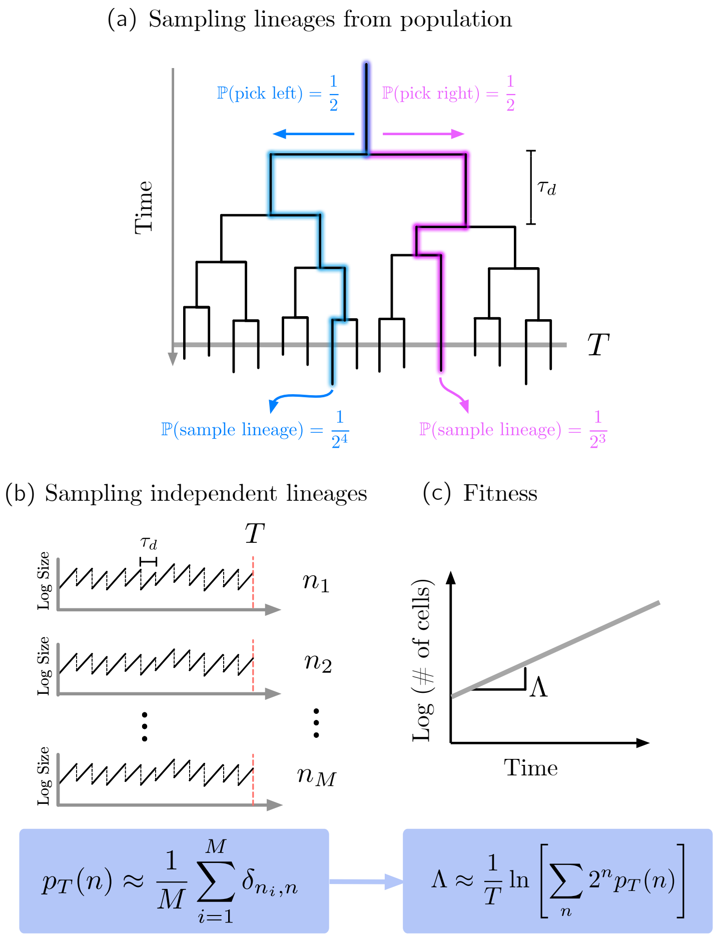
(a) Sampling lineages from population
Time
$\mathbb{P}$(pick left) $= \frac{1}{2}$
$\mathbb{P}$(pick right) $= \frac{1}{2}$
τ_d
T
$\mathbb{P}$(sample lineage) $= \frac{1}{2^4}$
$\mathbb{P}$(sample lineage) $= \frac{1}{2^3}$
(b) Sampling independent lineages
(c) Fitness
τ_d
T
Log Size
n_1
n_2
n_M
Log (# of cells)
Λ
Time
$p_T(n) \approx \frac{1}{M}\sum_{i=1}^{M}\delta_{n_i,n}$
$\Lambda \approx \frac{1}{T}\ln\left[\sum_n 2^n p_T(n)\right]$

FIG. 1. (a) A population tree starting from a single ancestor. The distinction is made between single lineages (highlighted) and a population (black). Lineages can be sampled by traveling down the tree and randomly selecting a daughter cell at the end of each branch. The probability of selecting any specific lineage with n divisions is 2^{-n}. (b) M independent lineages of length T. For the ith lineage, n_i is the number of cell divisions along that lineage. For each lineage we have shown the cell size, which typically increases exponentially between divisions, as a function of time. The lineage division distribution can be approximated from these independent lineages by recording the division events and using the highlighted formula. (c) A growing population of cells from which one can compute the fitness directly by counting the number of cells as a function of time (or utilizing equation (1)). Using the lineage distribution of divisions we can obtain the fitness from independent lineages.

eage from the tree by starting from the ancestral cell in the population and randomly selecting one of its daughter cells with equal probability to obtain the next cell in the lineage. Repeating this procedure yields a single lineage, as shown by the highlighted paths in Figure 1.

If $N(T)$ is the number of cells in the population at time T, then there are exactly $N(T)$ lineages, as each cell in the final population corresponds to a distinct lineage. However, by randomly selecting a lineage in the *forward* manner described above, lineages with more divisions are less likely to be selected, since each division decreases the chance that we will travel down that specific path through the tree. In particular, the probability of drawing any specific lineage from the tree is 2^{-n}. It follows that the empirical division distribution, denoted $\hat{p}_T(n)$, of observing exactly n divisions in a lineage sampled using this procedure is given by [20, 21]

$$\hat{p}_T(n) = 2^{-n}N(n,T), \tag{2}$$

where $N(n,T)$ is a random variable representing the number of lineages with n divisions in a specific realization of a growing population. Note that $\hat{p}_T(n)$, is also a random variable, and will therefore differ between different realizations of the population tree. By averaging over many realization of the tree, we obtain the division distribution:

$$p_T(n) \equiv \langle \hat{p}_T(n)\rangle_{\text{trees}}. \tag{3}$$

It is important to remember that $p_T(n)$ is distinct from what has been called the *retrospective distribution*, defined as the probability of observing n divisions in a lineage obtained by uniformly sampling a cell from the population at time T and following its ancestors back in time [21].

Given a specific realization of the population, the total number of cells can be represented in terms of the empirical division distribution as

$$N(T) = \sum_n N(n,T) = \sum_n 2^n \hat{p}_T(n). \tag{4}$$

Averaging over many realization of the population now yields the average population size in terms of the division distribution

$$\langle N(T)\rangle_{\text{trees}} = \sum_n 2^n p_T(n). \tag{5}$$

Importantly, $p_T(n)$ is not a random variable that depends on a specific realization of the tree, rather it is an intrinsic property of lineages and has no information about the correlations between sister cells. It can therefore be obtained by running many *independent* lineages: if we have a collection of M independent lineages of length T and observe n_i divisions in the ith lineage, then

$$p_T(n) = \lim_{M\to\infty}\frac{1}{M}\sum_i^M \delta_{n_i,n}; \tag{6}$$

see Figure 1 (b). The long-time population growth rate, defined as

$$\Lambda \equiv \lim_{T\to\infty}\frac{1}{T}\ln N(T), \tag{7}$$

can now be related to an average over independent lineages according to the *lineage representation*,

$$\Lambda = \lim_{T\to\infty}\frac{1}{T}\ln\langle 2^n\rangle_p . \tag{8}$$

Here, the angular brackets denote an average over $p_T(n)$. The definition of Λ in equation (7) is justified in SM section 1, where we have shown that this limit is self-averaging (i.e. in the long-time limit, only a single realization of the population is needed to obtain Λ). The lineage representation in equation (8) establishes a relationship between the lineage dynamics and the population fitness. A similar formulation was used in [21] to quantify how selection acts on an observable in a growing population; however, our formulation differs in that the average is taken over *independent* lineages rather than lineages from a single growing population.

In order to apply the lineage representation of the population growth rate to real data, we must develop an understanding of how quickly it converges in the number of lineages, M, and the duration of each lineage, T. As we will show below, some care needs to be taken when selecting M and T, as the lineage representation has a non-monotonic convergence in T. However, before presenting our convergence analysis, we establish a relationship between the lineage representation and the large deviation principle underlying the growth process. This structure is best introduced with the simple example below.

Explicit calculation of $p_T(n)$ for discrete Langevin model. We now perform an explicit calculation of $p_T(n)$ for a specific model in which generation times undergo a discrete Langevin process along a lineage, this model is referred to as the *random generation time* model in the literature [1, 2]. In this model, the generation time τ of a cell is related to its mother's generation time, τ', according to

$$\tau = \tau_0(1-c) + \tau' c + \xi \tag{9}$$

where ξ is a Gaussian with mean zero and variance σ_ξ^2. The parameter c controls the strength of correlations between mother and daughter cells, and for $c = 0$ we retrieve the classical Bellman-Harris branching process [22]. It can be seen that the average generation time along a lineage is $\langle\tau\rangle = \tau_0$. The population growth rate for this model has previously been obtained using a recursive approach in ref [2]. Here, we are able to derive the same result by implementing the lineage representation. Additionally, we obtain the complete distribution of divisions which elucidates the underlying large deviation structure of the population growth process.

We proceed by writing the probability that the nth division will occur along a lineage at time T:

$$\begin{aligned} q_T(n) = &\int_0^\infty \int_0^\infty \cdots \int_0^\infty \delta\left(\sum_i t_i - T\right) \\ &\times \prod_i \frac{1}{\sqrt{2\pi\sigma_\xi^2}} e^{-(t_i - ct_{i-1} - \tau_0(1-c))^2/2\sigma_\xi^2} dt_i. \end{aligned} \tag{10}$$

Note that this is distinct from $p_T(n)$, the probability of observing exactly n divisions *before* reaching time T. However, since we are interested primarily in the large deviations, the two distributions are interchangeable. By making a change of variables and performing a Gaussian integral (see SM section 2), equation (10) leads to the simple formula

$$p_T(n) = K e^{-n\frac{(1-c)^2}{2\sigma_\tau^2}\left(\tau_0 - \frac{T}{n}\right)^2}, \tag{11}$$

where K is a normalization constant independent of n and $\sigma_\tau^2 = \sigma_\xi^2/(1-c^2)$ is the variance in τ taken over a single lineage. Notice that the distribution of divisions is not Gaussian but due to the quadratic dependence of the exponent on T/n, that of the inverse of the number of divisions is approximately Gaussian. We elaborate on this fact in SM section 4.

The exponential form of equation (11) along with equation (8) suggests that the population growth rate is dominated by a particular value of n which maximizes the exponent of $2^n p_T(n)$. Treating n as a continuous variable and solving for n in $\frac{\partial}{\partial n}\left[n \ln 2 + \ln p_T(n)\right] = 0$ yields the dominant number of divisions

$$n_c = T \Big/ \sqrt{\langle\tau\rangle^2 - \frac{2\ln(2)\sigma_\tau^2}{(1-c)^2}}. \tag{12}$$

Note that in the limit where $\sigma_\tau^2 \to 0$, we find $n_c = T/\langle\tau\rangle$, which is the number of divisions corresponding to the average generation time. Substituting only the dominant value of n from equation (12) into equation (8) gives us the formula for the bulk population growth rate: $\Lambda = n_c \ln(2)/T + \ln p(n_c, T)/T$. After some simplification, we obtain

$$\Lambda = \frac{2\ln(2)/\langle\tau\rangle}{1 + \sqrt{1 - 2\ln(2)\frac{\sigma_\tau^2}{\langle\tau\rangle^2}\frac{1+c}{1-c}}}, \tag{13}$$

which is in agreement with previous computations using an alternative approach [2]. From equation (13), we can see how the three model parameters, namely $\langle\tau\rangle$, σ_τ^2 and c, affect population growth. In particular, growth is increased when σ_τ^2 and c are increased, while increasing $\langle\tau\rangle$ decreases growth.

Large deviation principle. In order to connect the observation of the previous section to large deviation theory, we introduce the *time averaged division rate* $\gamma = n/T$, so that the distribution of division rates given by equation (11) can be expressed as

$$p_T(\gamma) \propto e^{-TI(\gamma)} \tag{14}$$

with

$$I(\gamma) = \frac{\gamma(1-c)^2}{2\sigma_\tau^2}\left(\tau_0 - 1/\gamma\right)^2. \tag{15}$$

The exponential dependence of $p_T(\gamma)$ on T is known as a *large deviation principle* and suggests that for large T averages over $p_T(\gamma)$ are dominated by a single value of γ [23]. For the remainder of this paper we will assume this large deviation principle is satisfied.

In SM section 3, we show that when $I''(\langle\gamma\rangle_p) \gg 1$ we can make a Gaussian approximation of $p_T(\gamma)$ to obtain

$$\Lambda \approx \frac{\ln(2)}{\langle\tau\rangle} + \frac{T\ln(2)^2\sigma_\gamma^2}{2} \tag{16}$$

with $\sigma_\gamma^2 = 1/TI''(\langle\gamma\rangle_p)$. This illustrates a central result of the large deviation formulation; namely, regardless of the model specifics, corrections to the population growth rate due to variation in generation times scale inversely with the curvature of the large deviation rate function.

Convergence of lineage representation. We now address the question: How accurately can we estimate Λ given M lineages with durations T? To quantify the accuracy of an estimate, denoted $\hat{\Lambda}_{\text{lin}}$, we use the averaged squared deviation

$$\text{err}(\hat{\Lambda}_{\text{lin}})^2 = \left\langle \left((\hat{\Lambda}/\Lambda - 1\right)^2 \right\rangle_{\mathcal{E}}. \tag{17}$$

Here, the average $\langle\cdot\rangle_{\mathcal{E}}$ represents the average over many realization of the ensemble of M lineages of duration T, not

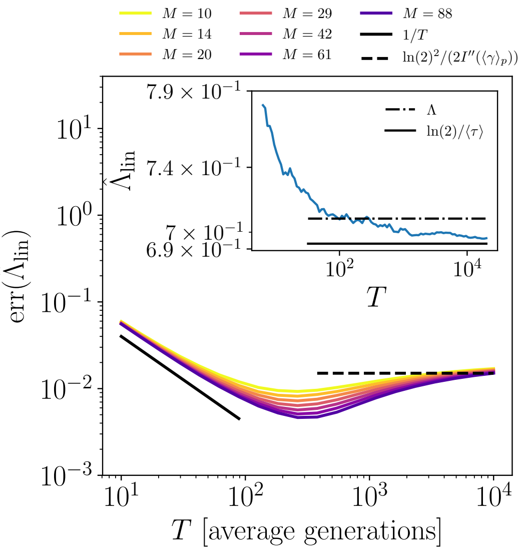

FIG. 2. Convergence of the error from the lineage representation as a function of the lineage durations, T, for different numbers of lineages, M. Data was generated from lineage simulations of the Langevin model with $\langle\tau\rangle = 1$, $c = 0.2$ and $\sigma_\tau = 0.2$. Here it can clearly be seen that the error initially scales as $1/T$, eventually increasing to approach the limit imposed on the sampling error by the large deviation rate function; see equation (19). The inset shows the lineage representation $\hat{\Lambda}_{\mathrm{lin}}$ as a function of T using $M = 80$ lineages. This plot is noisy because only a single ensemble of lineages is used, in contrast err in the main plot is computed by averaging over many ensembles of lineages.

to be confused with the averages elsewhere that are taken over lineages within an ensemble. Two distinct factors contribute to the error: First, the estimate of Λ obtained from the lineage representation will be subject to a systematic error resulting from the fact that given an infinite number of lineages each with a finite duration T, the lineage representation approximates the *arithmetic mean* fitness at time T: $\Lambda_{T,a} = 1/T \ln\langle N\rangle$. This is distinct from Λ. (In fact, it is not even the correct measure of fitness for a population grown over a finite period of time, which is given by the *geometric mean* $1/T\langle \ln N\rangle$; see ref. [24] for an explanation of why this is.) We refer to this error as *finite duration error*, and as we have shown in the SM section 4, it will scale inversely with T.

The second factor contributing to $\mathrm{err}(\hat{\Lambda})$ is sampling error in the approximation of the average $\langle 2^n\rangle_p$ from a finite number of lineages. As shown in the SM section 4, when

$$\frac{1}{\langle 2^n\rangle_p}\sqrt{\frac{\mathrm{var}(2^n)}{M}} = \sqrt{\frac{2^{T\ln(2)/I''(\langle\gamma\rangle_p)} - 1}{M}} \ll 1 \qquad (18)$$

the contribution of the sampling error to $\mathrm{err}(\hat{\Lambda}_{\mathrm{lin}})$ will grow exponentially in T for any fixed M, eventually dominating the error resulting from finite lineage durations. As T becomes large, the distribution of γ becomes much more narrow, so an ever-increasing number of lineages are needed to sample the variation in γ; however, as we have seen in equation (16), knowledge of the variation is needed to resolve the effects of generation time variability on population growth. In the long-time limit, all information about the variation is lost for finite M and the lineage representation simply retrieves the zeroth order term in equation (16):

$$\lim_{T\to\infty} \hat{\Lambda}_{\mathrm{lin}} = \ln(2)\langle\gamma\rangle_p. \qquad (19)$$

This demonstrates that the T and M limits do not commute, because as we have already seen, if we first take $M \to \infty$ the lineage representation converges to the exact population growth rate. As a result, there is a “goldilocks effect”: if T is too small the estimate will be inaccurate due to the finite duration error, while if T is too large we encounter the limit given by equation (19). The best estimate is in fact obtained by using an intermediate T where both effects are minimized. This prediction is validated numerically for the Langevin model in Fig. 2. We have also generated the same data for a more biophysically realistic model of cell growth (the cell-size regulation mode [17]), and found the results are qualitatively similar; see SM section 5.

How much data do we need to be confident we are not encountering the limit given by equation (19)? The sampling error will have a negligible effect on the estimate when equation (18) is satisfied. This condition can be rewritten as

$$M \gg 2^{T\ln(2)/I''(\langle\gamma\rangle_p)}. \qquad (20)$$

This implies that the number of lineages needed to avoid encountering the sampling error grows exponentially with the duration of the lineages and the generation time variance. In order to be confident that the finite duration error is small enough to resolve the second term in equation (16), we must select $T \gg I''(\gamma_c)$. This means that $T\ln(2)/I''(\langle\gamma\rangle_p)$ is a large quantity. For example, if we want to ensure that the finite duration error is an order of magnitude smaller than the generation time variance, a safe value of M will generally be larger than $2^{10\times\ln(2)} \approx 120$. Existing data sets obtained from mother machines contain on the order of one hundred lineages, making the application of the lineage representation plausible. In SM section 6, we have explored the applications of the lineage algorithm to mother machine data, where we have found that the dependence of $\hat{\Lambda}_{\mathrm{lin}}$ on T is qualitatively consistent with the theory presented above and Fig. 2.

Discussion. Experimental advances over the last few decades have made it possible to observe the stochastic dynamics of growth and division in bacteria with increasing levels of precision [19, 25, 26]. These observations have revealed universal principles underlying microbial growth, such as the adder mechanisms for maintaining homeostasis of cell sizes [17, 18, 27–29]. Bulk experiments in which bacteria are grown exponentially or in competition assays can be used to compare the fitness of different strains, and in principle elucidate how these physiological differences map to fitness. However, the equivalence between growth in bulk experiments and those used to observe single-cell traits remains unclear, because of the different environments which cells are subjected to in these experiments. In this paper, we have presented a lineage representation that links single-lineages to the population growth.

We have also found that a large deviation principle underlies the population dynamics which applies to the distribution of division rates among lineages in a growing population.

This implies that the population becomes dominated by lineages with a certain optimal division rate. This idea generalizes an *optimal lineage* principle introduced by Wakamoto et al. which was used to calculate the population growth rate within the context of a model with uncorrelated generation times [20]. Using the large deviation framework, we have quantified exactly how much data is needed to resolve the effects of cell-to-cell variability on population growth from single lineages. We expect that this work will serve as a guide for future experimental studies seeking to link single-cell observations to fitness.

We thank Jie Lin and Somenath Bakshi for helpful discussions related to this work. We thank Jane Kondev for providing feedback on the manuscript. We acknowledge funding support from National Science Foundation under Award DMS-1902895 (EL), Californian Alliance Research Exchange with NSF Grants 1647273/1742065/1306595/1306683/1306747/1306760 (TG) and NSF CAREER 1752024 (AA).

[*] These two authors contributed equally

[1] J. Lin and A. Amir, Cell Systems **5**, 358 (2017).
[2] J. Lin and A. Amir, Physical Review E **101** (2020), 10.1103/PhysRevE.101.012401.
[3] E. Levien, J. Kondev, and A. Amir, bioRxiv (2019).
[4] Z. Wang and J. Zhang, PNAS; Proceedings of the National Academy of Sciences **108**, E67 (2011).
[5] F. Duveau, A. Hodgins-Davis, B. P. Metzger, B. Yang, S. Tryban, E. A. Walker, T. Lybrook, and P. J. Wittkopp, eLife **7** (2018), 10.7554/elife.37272.
[6] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297** (2002).
[7] Y. Sughiyama and T. J. Kobayashi, Physical Review E **95** (2017), 10.1103/physreve.95.012131.
[8] L. Chao, C. U. Rang, A. M. Proenca, and J. U. Chao, PLOS Computational Biology **12**, e1004700 (2016).
[9] S. Vedel, H. Nunns, A. Košmrlj, S. Semsey, and A. Trusina, Cell Systems **3**, 187 (2016).
[10] A. Marantan and A. Amir, Physical Review E **94** (2016), 10.1103/PhysRevE.94.012405.
[11] J. Min, J. Lin, and A. Amir, Physical Review Letters **122** (2019), 10.1103/PhysRevLett.122.068101.
[12] B. Claudi, P. Spröte, A. Chirkova, N. Personnic, J. Zankl, N. Schürmann, A. Schmidt, and D. Bumann, Cell **158**, 722 (2014).
[13] E. O. Powell, Microbiology **15**, 492 (1956).
[14] J. L. Lebowitz and S. I. Rubinow, Journal of Mathematical Biology **1**, 17 (1974).
[15] R. Garcia-Garcia, A. Genthon, and D. Lacoste, Physical Review E **99** (2019), 10.1103/PhysRevE.99.042413.
[16] A. J. Lotka, Publications of the American Statistical Association **16**, 121 (1907).
[17] A. Amir, Physical Review Letters **112** (2014), 10.1103/PhysRevLett.112.208102.
[18] P.-Y. Ho, J. Lin, and A. Amir, Annual Review of Biophysics **47**, 251 (2018).
[19] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun, Current Biology **20**, 1099 (2010).
[20] Y. Wakamoto, A. Y. Grosberg, and E. Kussell, Evolution **66**, 115 (2012).
[21] T. Nozoe, E. Kussell, and Y. Wakamoto, PLoS genetics **13**, e1006653 (2017).
[22] R. Bellman and T. Harris, Annals of Mathematics **55**, 280 (1952).
[23] H. Touchette, Physics Reports **478**, 1 (2009).
[24] R. C. Lewontin and D. Cohen, Proceedings of the National Academy of sciences **62**, 1056 (1969).
[25] S. Luro, L. Potvin-Trottier, B. Okumus, and J. Paulsson, Nature Methods **17**, 93 (2019).
[26] D. Camsund, M. J. Lawson, J. Larsson, D. Jones, S. Zikrin, D. Fange, and J. Elf, Nature Methods **17**, 86 (2019).
[27] C. Cadart, S. Monnier, J. Grilli, P. J. Sáez, N. Srivastava, R. Attia, E. Terriac, B. Baum, M. Cosentino-Lagomarsino, and M. Piel, Nature Communications **9** (2018), 10.1038/s41467-018-05393-0.
[28] Y.-J. Eun, P.-Y. Ho, M. Kim, S. LaRussa, L. Robert, L. D. Renner, A. Schmid, E. Garner, and A. Amir, Nature Microbiology **3**, 148 (2017).
[29] M. M. Logsdon, P.-Y. Ho, K. Papavinasasundaram, K. Richardson, M. Cokol, C. M. Sassetti, A. Amir, and B. B. Aldridge, Current Biology **27**, 3367 (2017).
[30] F. Jafarpour, Physical Review Letters **122** (2019), 10.1103/PhysRevLett.122.118101.
[31] F. Jafarpour, Physical Review X **8** (2018), 10.1103/PhysRevX.8.021007.
[32] Y. Tanouchi, A. Pai, H. Park, S. Huang, N. E. Buchler, and L. You, Scientific Data **4** (2017), 10.1038/sdata.2017.36.

SUPPLEMENTAL MATERIALS

1. Argument that Λ is self-averaging

We show that, provided fluctuations in generation times have a finite correlation time,

$$\Lambda_T = \frac{1}{T}\ln N(T) \tag{S1}$$

converges to a deterministic value in the large T limit. This will establish that equation (7) is an appropriate definition for the long-term fitness. We begin by defining averages of Λ_T in two ways: First, we consider the geometric mean

$$\Lambda_{T,g} = \frac{1}{T}\langle \ln N(T)\rangle \tag{S2}$$

where the average is taken over many realizations of the population. Second, we consider the arithmetic mean fitness:

$$\Lambda_{T,a} = \frac{1}{T}\ln\langle N(T)\rangle. \tag{S3}$$

We will say that Λ_T is self-averaging if the geometric and mean fitness converge to the same value in the long-time limit. The distinction between mean and geometric fitness has been shown to be important when considering populations growing in the presence of environmental stochasticity. In this context, it is often the case that the two are not equal and that the geometric fitness is the more appropriate measure of the long-term viability of a population; see ref. [24] for a detailed discussion of this point. In the context of branching processes, the fact that Λ_T is self-averaging has been established for the Bellman-Harris branching process model; see ref. [22]. For completeness, we provide an argument that is slightly more general, and allows for mother and daughter cells to have correlated generation times.

Setting $N(T) = \langle N(T)\rangle + dN$, we have

$$\begin{aligned}\Lambda_{T,g} &= \frac{1}{T}\left\langle \ln\left[\langle N(T)\rangle\left(1+\frac{dN}{\langle N(T)\rangle}\right)\right]\right\rangle \\ &\approx \Lambda_{T,a} - \frac{1}{2T}\frac{\langle dN^2\rangle}{\langle N(T)^2\rangle} \\ &= \Lambda_{T,a} - \frac{1}{2T}\mathrm{CV}_N^2\end{aligned} \tag{S4}$$

where CV_N^2 is the coefficient of variation of $N(T)$. We now argue that CV_N^2 converges to a constant as T becomes large. To do this, we write down the differential equation for the probability distribution of $N(t)$, denoted $P(N,t)$. Let $\alpha(t)$ be the per unit time, per capita probability of a cell dividing. The time dependence in $\alpha(t)$ comes from the fact that the distribution of ages will take some time to converge to its steady-state [30, 31]. Since the division rates of individual cells are age-dependent, the per capita division rates throughout the population will depend on time. It has previously been shown that if there is any variability in generation times and the fluctuations in generation times have a finite correlation time, the distribution of ages will eventually become time-invariant [3, 14]. Hence $\alpha(t)$ will converge to a constant, $\bar{\alpha}$, in the long-time limit. Thus, after a long period of time, the dynamics of $P(N,t)$ are approximated by

$$\frac{d}{dt}P(N,t) \approx \bar{\alpha}(N-1)P(N-1,t) - \bar{\alpha}NP(N,t). \tag{S5}$$

To proceed we analyze the moments of N taking the initial condition to be $\langle N(t_0)^k\rangle = n_k$. It is important that we are only considering the dynamics after an initial transient, otherwise equation (S5) will not be a valid approximation. For the first moment of N, we find

$$\frac{d}{dt}\langle N\rangle = \bar{\alpha}\sum\left[(N+1)NP(N,t) - N^2P(N,t)\right] = \bar{\alpha}\langle N\rangle \tag{S6}$$

implying $\langle N\rangle = n_1 e^{(t-t_0)\bar{\alpha}}$ and $\bar{\alpha} = \Lambda_{T,a}$. Similarly,

$$\begin{aligned}\frac{d}{dt}\langle N^2\rangle &\approx \bar{\alpha}\sum N^2\left[(N-1)P(N-1,t) - NP(N,t)\right] \\ &= \bar{\alpha}\left[n_1 e^{t\bar{\alpha}} + 2\langle N^2\rangle\right].\end{aligned} \tag{S7}$$

Solving this ODE with $\langle N^2\rangle$ gives

$$\langle N^2\rangle \approx e^{(t-t_0)\bar{\alpha}}\left[n_2 e^{(t-t_0)\bar{\alpha}} + n_1\left(e^{(t-t_0)\bar{\alpha}} - 1\right)\right] \tag{S8}$$

Since both $\langle N^2\rangle$ and $\langle N\rangle^2$ grow asymptotically as $e^{2\bar{\alpha}t}$, the coefficient of variation of $N(t)$ will converge to a constant. In particular,

$$\mathrm{CV}_N^2(t) = \frac{n_2}{n_1^2} + \frac{1-e^{-\bar{\alpha}(t-t_0)}}{n_1} \to \frac{n_2+n_1}{n_1^2}. \tag{S9}$$

It follows that

$$\lim_{T\to\infty} \Lambda_{T,a} = \lim_{T\to\infty} \Lambda_{T,g}, \tag{S10}$$

which implies

$$\Lambda = \lim_{T\to\infty} \Lambda_T \tag{S11}$$

is non-random. The predictions of this section are verified in Figure S1, where we have shown that the exponential growth rate of $N(t)$ is deterministic and CV_N converges to a constant, although our calculation does not give an explicit formula for this constant.

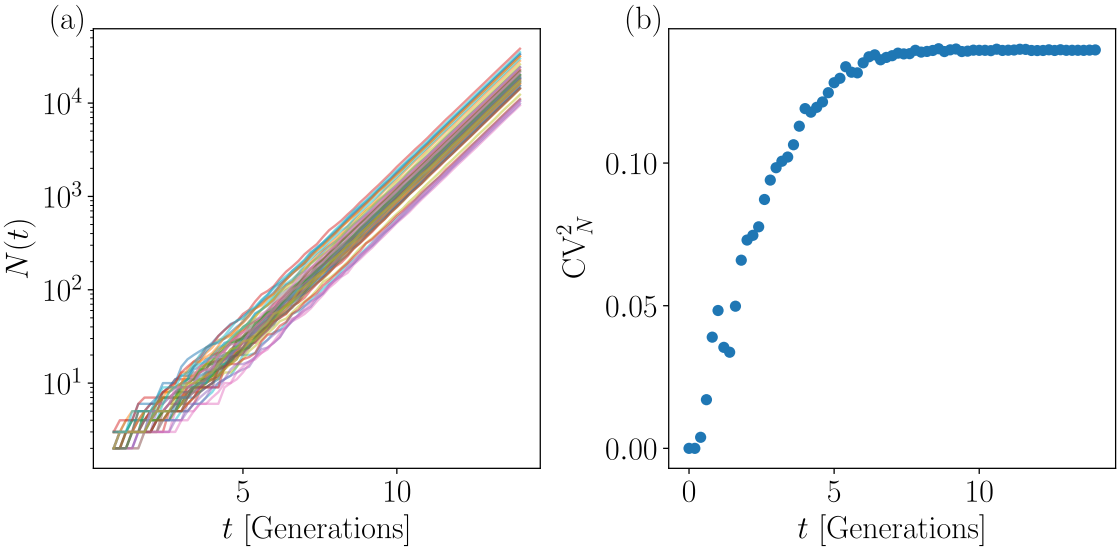

FIG. S1. (a) Many realizations of $N(t)$ on a log plot. For all realizations, the slope of $\ln N(t)$ is eventually constant and independent of the realization. Thus Λ is well-defined. (b) The coefficient of variation of $N(t)$ as a function of time. The CV is computed from 500 simulations of $N(t)$. Data was generated from simulations of the full population for the Langevin model with $\langle\tau\rangle = 1$, $c = 0.2$ and $\sigma_\tau = 0.2$.

2. Details in derivation of $p_T(n)$ for Langevin model

Here we provide details on the calculation of $p_T(n)$ for the Langevin model. Starting with equation (10) of the main text, we replace the δ function with $\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega x} d\omega$ and introduce the constant $b = \tau_0(1-c)$. This leads to an equivalent form

$$q_T(n) \propto \int_{-\infty}^{\infty} e^{-i\omega T}\left[\int_0^\infty \cdots \int_0^\infty \prod_{j=0}^{n} e^{i\omega t_j} e^{(t_j-ct_{j-1}-b)^2/2\sigma_\xi^2} dt_j\right] d\omega. \tag{S12}$$

We can complete the square in the integrand and define the constants $\bar{b} = b + i\omega 2\sigma_\xi^2/(2(1-c))$ and $d = i\omega c 2\sigma_\xi^2 2/(1-c)$. This leads to the integral

$$q_T(n) \propto \int_{-\infty}^{\infty} e^{-i\omega T}\left[\int_0^\infty \cdots \int_0^\infty \prod_{j=0}^{n} e^{(t_j-ct_{j-1}-\bar{b})^2/2\sigma_\xi^2} e^{(\bar{b}^2-b^2)n/2\sigma_\xi^2} e^{t_n d} dt_j\right] d\omega. \tag{S13}$$

In order to evaluate this integral, we make the transformation $x_i = t_i - ct_{i-1} - \bar{b}$. Note that the Jacobian from this change of variables is 1. When the noise in generation times is small enough for the contribution from the negative part of the integral to be negligible, we can make the approximation that integrals are evaluated over the real line. This leads to

$$q_T(n) \propto \int_{-\infty}^{\infty} e^{-i\omega T} e^{-(\bar{b}^2-b^2)n/2\sigma_\xi^2} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=0}^{n} e^{-x_j^2/2\sigma_\xi^2} e^{dt_n} dx_j \right] d\omega. \tag{S14}$$

Also, notice that t_n can be expressed as a sum over all the x_is as

$$t_n = x_n + cx_{n-1} + c^2 x_{n+2} + \cdots . \tag{S15}$$

For $|c| < 1$, which is always the case in our model, this only has an affect on a finite number of terms, thus the nth term can be neglected for large n. We find that

$$q_T(n) = \int_{-\infty}^{\infty} e^{-i\omega T} e^{n(\bar{b}^2-b^2)/2\sigma_\xi^2} d\omega. \tag{S16}$$

Using the identity

$$\frac{\bar{b}^2 - b^2}{2\sigma_\xi^2} = \frac{-\omega^2 2\sigma_\xi^2}{4(1-c)^2} + \frac{i\omega b}{1-c} \tag{S17}$$

and computing the integral, we see that the formula for q is given by equation (11) of the main text. To obtain $p_T(n)$ from $q_T(n)$, we note that $q_T(n)$ is the probability of observing n divisions weighted by the frequency of passing through age zero, which is time invariant (see [3, 14]), hence in the long time limit the large deviation rate functions for these two distributions will be equivalent.

3. Variational formulation of population growth rate

Using the large deviation formulation, we can obtain an alternative representation of Λ. First, we can take a Laplace Transform of equation (14) of the main text, yielding

$$\left\langle e^{T\gamma\beta} \right\rangle_p = \int e^{T(\gamma\beta - I(\gamma))} d\gamma. \tag{S18}$$

Setting $\beta = \ln 2$, we see that the population growth rate is given by a saddle point over γ written as

$$\Lambda = \max_\gamma \left[\gamma \ln 2 - I(\gamma)\right]. \tag{S19}$$

This formulation is equivalent to equation (8) of the main text, but makes an explicit connection between the growth rate and the Large deviation rate function. In addition, we can cast equation (S19) as a cumulant expansion of equation (S18) with $\beta = \ln(2)$. This gives

$$\Lambda = \langle\gamma\rangle_p \ln(2) + \frac{T\ln(2)^2}{2}\sigma_\gamma^2 + \cdots, \tag{S20}$$

where $\sigma_\gamma^2 = 1/TI''(\langle\gamma\rangle_p)$ is the variance of γ with respect to $p_T(\gamma)$. We see that the first term in equation (S20) is simply the growth rate calculated from the doubling time of a single lineage, $\ln(2)/\langle\tau\rangle$, while including the second order term gives us equation (S20) becomes equation (16) of the main text. In the context of the Langevin model, the curvature of the large deviation rate function is given by $I''(\langle\gamma\rangle_p) \approx (1-c)^2 \langle\tau\rangle^3/\sigma_\tau^2$. It is straightforward to check that equation (S20) is consistent with equation (13) to leading order in $\sigma_\tau^2/\langle\tau\rangle^2$ and c. If either one of these parameters is large, then the distribution $p_T(\gamma)$ will be too broad for the first term in the cumulant expansion to give a good approximation.

4. Analysis of convergence

We analyze the convergence of the lineage representation by first considering the limit in which we have an infinite number of lineages ($M \to \infty$), each with a large, but finite duration T. In this case, the lineage representation gives the arithmetic mean fitness, $\Lambda_{T,a}$. This can be obtained from the saddle point approximation,

$$\left\langle 2^{\gamma T} \right\rangle = \frac{K}{T} \int_0^\infty e^{-TI(\gamma) + \gamma T \ln(2)} d\gamma \tag{S21}$$

$$\approx \frac{K}{T} \sqrt{\frac{2\pi}{TI''(\gamma_c)}} e^{-TI(\gamma_c) + \gamma_c T \ln(2)}. \tag{S22}$$

Note that the factor $1/T$ outside the integral comes from the Jacobian when we change variables from n to γ. It follows that

$$\Lambda_T = \frac{1}{T}\ln\left\langle 2^{\gamma T}\right\rangle = \Lambda + \frac{1}{T}\ln\left[\frac{K}{T^{3/2}}\sqrt{\frac{2\pi}{I''(\gamma_c)}}\right]. \tag{S23}$$

In order for the normalization of $p_T(n)$ to hold, K must grow as $\sqrt{T}$. We therefore conclude that the convergence in T is (ignoring logarithmic terms) $O(1/T)$. The specific value of the prefactor will depend on the initial transient dynamics when the population is small, and therefore cannot be computed using the large deviation estimates, which are valid when T is large.

Now consider the estimate obtained from a finite number of lineages. For large M, by applying the central limit theorem, we get

$$\left\langle 2^{\gamma T}\right\rangle_M \approx \left\langle 2^{\gamma T}\right\rangle + \sqrt{\frac{\mathrm{Var}(2^{\gamma T})}{M}}\eta \tag{S24}$$

where η is a standard normal variable. The rate function can be approximated as a Gaussian by writing

$$I(\gamma) \approx \frac{1}{2}I''(\langle\gamma\rangle_p)(\gamma - \langle\gamma\rangle_p)^2 = \frac{(\gamma-\langle\gamma\rangle_p)^2}{2T\sigma_\gamma^2}. \tag{S25}$$

Recall that $\langle\gamma\rangle_p$ is the average of γ with respect to $p_T(\gamma)$. In this approximation, γ is a normally distributed random variable with variance

$$\sigma_\gamma^2 = \frac{1}{TI''(\langle\gamma\rangle_p)}. \tag{S26}$$

For a random variable X with mean μ and variance σ^2, $\langle e^X\rangle = e^{\mu+\sigma^2/2}$. Approximating γ with a Gaussian and applying this identity to $2^{\gamma T}$, we have

$$\langle 2^{\gamma T}\rangle \approx e^{T\ln(2)\langle\gamma\rangle + T^2\ln(2)^2\sigma_\gamma^2/2}, \tag{S27}$$

$$\left\langle \left(2^{\gamma T}\right)^2\right\rangle = e^{2T\ln(2)\langle\gamma\rangle + T^2\ln(2)^2 2\sigma_\gamma^2} \tag{S28}$$

$$\approx \langle 2^{\gamma T}\rangle^2 2^{T^2\ln(2)\sigma_\gamma^2} = \langle 2^{\gamma T}\rangle^2 2^{T\ln(2)/I''(\langle\gamma\rangle_p)}. \tag{S29}$$

Hence for the variance, we have

$$\mathrm{Var}(2^{\gamma T}) = \left\langle \left(2^{\gamma T}\right)^2\right\rangle - \langle 2^{\gamma T}\rangle^2 \tag{S30}$$

$$= \left\langle 2^{\gamma T}\right\rangle^2\left(2^{T\ln(2)/I''(\langle\gamma\rangle_p)} - 1\right). \tag{S31}$$

The predictions for the statistics of γ are validated numerically in Figure S2.

When $\sqrt{\mathrm{Var}(2^{\gamma T})/M}/\langle 2^{\gamma T}\rangle \ll 1$, we have

$$\hat{\Lambda}_{\mathrm{lin}} = \frac{1}{T}\ln\left(\left\langle 2^{\gamma T}\right\rangle + \sqrt{\frac{\mathrm{Var}(2^{\gamma T})}{M}}\eta\right) \tag{S32}$$

$$= \frac{1}{T}\ln\left\langle 2^{\gamma T}\right\rangle + \frac{1}{T}\ln\left(1 + \frac{1}{\langle 2^{\gamma T}\rangle}\sqrt{\frac{\mathrm{Var}(2^{\gamma T})}{M}}\eta\right) \tag{S33}$$

$$\approx \Lambda_T + \frac{1}{T\left\langle 2^{\gamma T}\right\rangle}\sqrt{\frac{\mathrm{Var}(2^{\gamma T})}{M}}\eta \tag{S34}$$

$$\approx \Lambda + \underbrace{\frac{1}{T}\ln\left[\frac{K}{T^{3/2}}\sqrt{\frac{2\pi}{I''(\gamma_c)}}\right]}_{\text{finite duration error}} + \underbrace{\frac{1}{T}\sqrt{\frac{2^{T\ln(2)/I''(\langle\gamma\rangle_p)}-1}{M}}\eta}_{\text{sampling error}}. \tag{S35}$$

From this calculation, we see that when equation (20) of the main text is satisfied the sampling error is $O(M^{-1/2})$. However, for any fixed M the the second term in equation (S35) grows exponentially in T. Hence for large enough T the expansion breaks down and we need to obtain the error in a different way. To this end, we consider the limit of a finite number of lineages, each with an infinite duration. Let γ_i be the empirical division rate along the ith of M lineages. In the limit $T\to\infty$,

$$\lim_{T\to\infty}\gamma_i = \langle\gamma\rangle_p. \tag{S36}$$

Therefore, if we first take the T limit, rather than having a distribution of γ each lineages gives the same value of $\gamma_i = \langle\gamma\rangle$. As a result, the lineage algorithm simply gives

$$\Lambda_{\text{lin}} = \ln(2)\langle\gamma\rangle_p. \tag{S37}$$

We conclude that for any finite M the lineage representation is not effective for large T. In particular, the representation only gives reasonable results if equation (20) is satisfied. This also establishes that the T and M limits of the lineage representation do not commute, and it is only if we first take the M limit that the estimate converges.

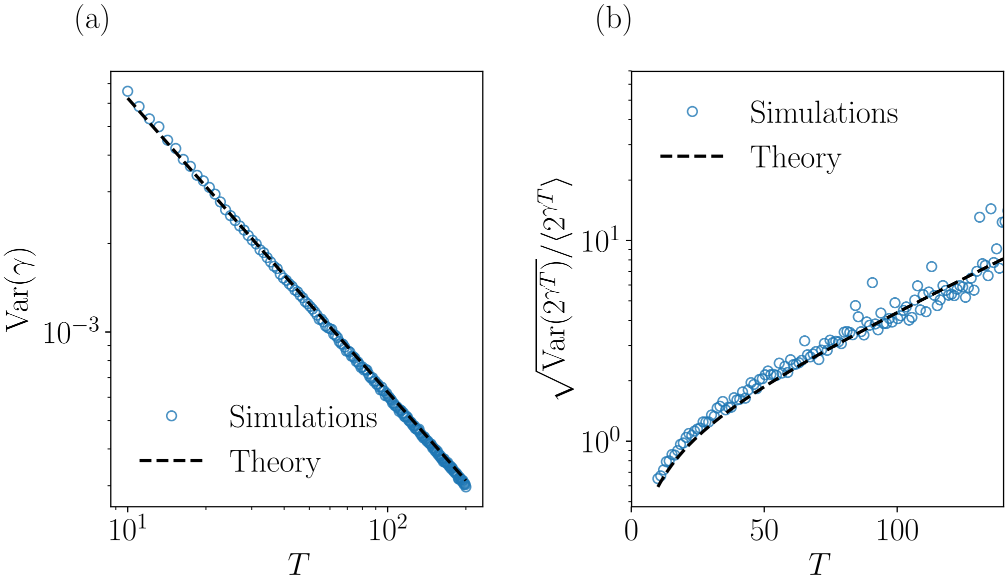

FIG. S2. Validation of the predictions for the statistics of γ using the Langevin model. (a) A demonstration of the linear decay in the variance of γ with time as predicted by equation (S26). (b) The coefficient of variation of $2^{\gamma T}$ compared to the theoretical prediction of equation (S30) obtained from a Gaussian approximation. Data was generated from lineage simulations of the Langevin model with $\langle\tau\rangle = 1$, $c = 0.2$ and $\sigma_\tau = 0.2$.

5. Test of convergence on cell-size regulation model

Here, we test the lineage representation on a more biophysically realistic model of microbial growth. This model has been used previously in the literature to understand how cells maintain homeostasis of their sizes [17, 18]. The central assumption of the cell-size regulation model is that cells grow exponentially at the single-cell level and divide when they reach a size v_{div} depending on their size at birth, v_{birth}. Since cells grow exponentially, the generation time of a cell satisfies $v_{\text{div}} = v_{\text{birth}} e^{\lambda\tau}$, or

$$\tau = \frac{1}{\lambda}\ln\left(\frac{v_{\text{div}}}{v_{\text{birth}}}\right). \tag{S38}$$

Here, λ is the single-cell growth rate. For simplicity, we will assume that cells divide symmetrically; thus v_{birth} is obtained by dividing the cell's mother's size at division by 2. Phenotypic variability is introduced in the form of both variability in single-cell growth rates and noise in the division volumes. To implement this, we take the growth rate λ and division volume v_{div} of a cell to obey [2]

$$\begin{aligned} \ln\lambda &= \ln\langle\lambda\rangle + \eta_\lambda \\ v_{\text{div}} &= 2(1-\alpha)v_{\text{birth}} + 2\alpha v_0 + \eta_v \end{aligned} \tag{S39}$$

where η_λ and η_v are independent normally distributed random variables with variances σ_λ^2 and σ_v^2, respectively. The first equation captures the fact that growth rates approximately follow a log-normal distribution for small noise, while the second equation tells us how cells decide when to divide based on their volume. The parameter α determines the cell-size regulation strategy. When $\alpha = 1$ (known as a "sizer" strategy), cells divide at a critical size, while when $\alpha = 1/2$ (known as an "adder" strategy) cells add a constant size v_0 between birth and division. We refer to refs [1, 18] for an in-depth discussion of the cell-size control model and its implications for population growth.

The convergence of the lineage representation for the cell-size regulation model is shown in Figure S3, which should be compared to Figure 2. We see that the behavior of the error in T and M is qualitatively similar to the the Langevin model, with the convergence being non-monotonic in T.

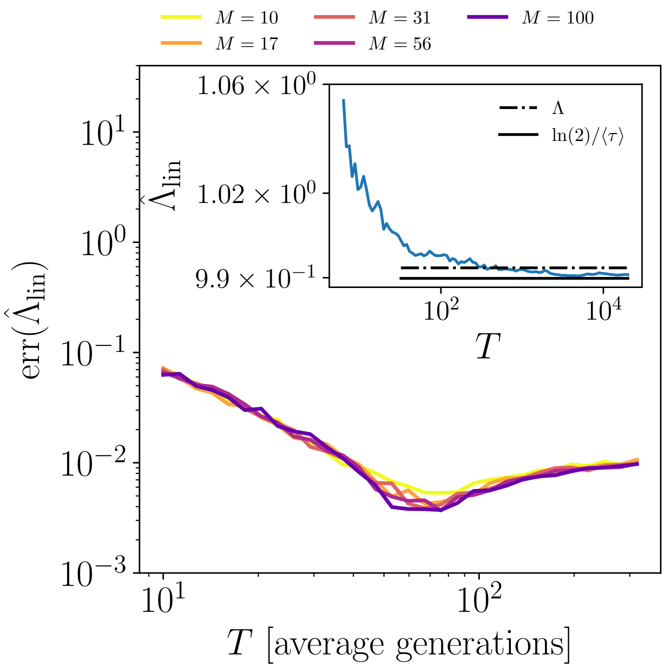

FIG. S3. Here we have run simulations in the same manner as Fig. 2 of the main text, but used the cell-size regulation model instead of the Langevin model to compute the generation times. Parameter values used are $\sigma_v = 0.2$, $\sigma_\lambda = 0.1$ and $\alpha = 1/2$.

6. Application to experimental data

Here, we apply the lineage representation to the mother machine data from ref [32]. In these experiments, *E. coli* were grown in three different temperatures. Each experiment resulted in a collection of roughly $M \approx 80$ independent lineages (recordings of division times) with durations on the order of 100 generations. In order to explore how the duration of the lineages affects our estimate, we have computed $\hat{\Lambda}_{\mathrm{lin}}$ for different values of T by truncating the lineages. Along with the lineage algorithm, we have computed two other measurements of growth: First, we have computed the first order term in equation (16), $\ln(2)/\langle\tau\rangle$. Second, we have computed the average single-cell elongation rate along the lineages. If v_{birth} and v_{div} are the initial and final volume of a cell over the course of a cell cycle, then the single-cell elongation rate is defined as $\lambda = 1/\tau \ln(v_{\mathrm{div}}/v_{\mathrm{birth}})$. By taking the average of λ over all cells in each experiment, we obtained the average single-cell elongation rates, $\langle\lambda\rangle$.

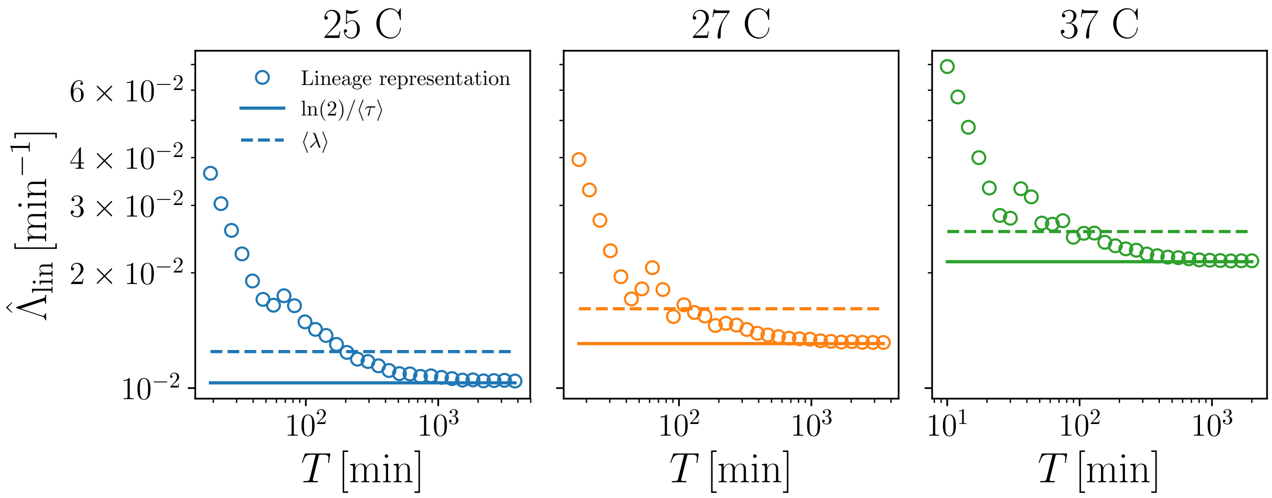

FIG. S4. The application of the lineage representation to experimental data. Each panel corresponds to an experiment done in a different temperature (indicated above the panel). Error bars for $\ln(2)/\langle\tau\rangle$ and $\langle\lambda\rangle$ can be computed from the standard error and are small enough to be invisible in these figures.

The results of our analysis are shown in Figure S4. Here, we see that for large T, $\hat{\Lambda}_{\mathrm{lin}} \approx \ln(2)/\langle\tau\rangle$; therefore, the lineage algorithm is not resolving the higher order effects on the population growth rate. Comparing the trajectories of $\hat{\Lambda}_{\mathrm{lin}}$ with those from the insets in Figure 2 of the main text and Figure S3, we see a qualitative agreement. In particular, they seem to decrease initially before fluctuating around a constant, before decreasing to $\ln(2)/\langle\tau\rangle$. This indicates that our theory indeed captures the performance of the lineage representation on real data, although without the corresponding population growth rate measurements, we cannot make any claims about the accuracy of the growth estimates.