Modeling the growth of organisms validates a general relation between metabolic costs and natural selection

Efe Ilker[1,2] and Michael Hinczewski[2]

[1] *Physico-Chimie Curie UMR 168, Institut Curie, PSL Research University, 26 rue d'Ulm, 75248 Paris Cedex 05, France*

[2] *Department of Physics, Case Western Reserve University, Cleveland OH 44106*

Metabolism and evolution are closely connected: if a mutation incurs extra energetic costs for an organism, there is a baseline selective disadvantage that may or may not be compensated for by other adaptive effects. A long-standing, but to date unproven, hypothesis is that this disadvantage is equal to the fractional cost relative to the total resting metabolic expenditure. This hypothesis has found a recent resurgence as a powerful tool for quantitatively understanding the strength of selection among different classes of organisms. Our work explores the validity of the hypothesis from first principles through a generalized metabolic growth model, versions of which have been successful in describing organismal growth from single cells to higher animals. We build a mathematical framework to calculate how perturbations in maintenance and synthesis costs translate into contributions to the selection coefficient, a measure of relative fitness. This allows us to show that the hypothesis is an approximation to the actual baseline selection coefficient. Moreover we can directly derive the correct prefactor in its functional form, as well as analytical bounds on the accuracy of the hypothesis for any given realization of the model. We illustrate our general framework using a special case of the growth model, which we show provides a quantitative description of overall metabolic synthesis and maintenance expenditures in data collected from a wide array of unicellular organisms (both prokaryotes and eukaryotes). In all these cases we demonstrate that the hypothesis is an excellent approximation, allowing estimates of baseline selection coefficients to within 15% of their actual values. Even in a broader biological parameter range, covering growth data from multicellular organisms, the hypothesis continues to work well, always within an order of magnitude of the correct result. Our work thus justifies its use as a versatile tool, setting the stage for its wider deployment.

Discovering optimality principles in biological function has been a major goal of biophysics [1–6], but the competition between genetic drift and natural selection means that evolution is not purely an optimization process [7–9]. A necessary complement to elucidating optimality is clarifying under what circumstances selection is actually strong enough relative to drift in order to drive systems toward local optima in the fitness landscape. In this work we focus on one key component of this

problem: quantifying the selective pressure on the extra metabolic costs associated with a genetic variant. We validate a long hypothesized relation [10–12] between this pressure and the fractional change in the total resting metabolic expenditure of the organism.

The effectiveness of selection versus drift hinges on two non-dimensional parameters [13]: i) the *selection coefficient* s, a measure of the fitness of the mutant versus the wild-type. Mutants will have on average $1 + s$ offspring relative to the wild-type per wild-type generation time; ii) the *effective population* N_e of the organism, the size of an idealized, randomly mating population that exhibits the same decrease in genetic diversity per generation due to drift as the actual population (with size N). For a deleterious mutant ($s < 0$) where $|s| \gg N_e^{-1}$, natural selection is dominant, with the probability of the mutant fixing in the population exponentially suppressed. In contrast if $|s| \ll N_e^{-1}$, drift is dominant, with the fixation probability being approximately the same as for a neutral mutation [7]. Thus the magnitude of N_e^{-1} determines the "drift barrier" [14], the critical minimum scale of the selection coefficient for natural selection to play a non-negligible role.

The long-term effective population size N_e of an organism is typically smaller than the instantaneous actual N, and can be estimated empirically across a broad spectrum of life: it varies from as high as $10^9 - 10^{10}$ in many bacteria, to $10^6 - 10^8$ in unicellular eukaryotes, down to $\sim 10^6$ in invertebrates and $\sim 10^4$ in vertebrates [12, 15]. The corresponding six orders of magnitude variation in the drift barrier N_e^{-1} has immense ramifications for how we understand selection in prokaryotes versus eukaryotic organisms, particularly in the context of genome complexity [16–18]. For example, consider a mutant with an extra genetic sequence relative to the wild-type. We can separate s into two contributions [12], $s = s_c + s_a$, where s_c is the baseline selection coefficient associated with the metabolic costs of having this sequence, i.e. the costs of replicating it during cell division, synthesizing any associated mRNA / proteins, as well as the maintenance costs associated with turnover of those components. The difference $s_a = s - s_c$ is whatever adaptive advantage or disadvantage accrues due to the consequences of the sequence beyond its baseline metabolic costs. For a prokaryote with a low drift barrier N_e^{-1}, even the relatively low costs associated with replication and transcription are often under selective pressure [11, 12], unless $s_c < 0$ is compensated for an $s_a > 0$ of comparable or larger magnitude [19]. For the much greater costs of translation, the impact on growth rates of unnecessary protein production is large enough to be directly seen in experiments on bacteria [1, 20]. In contrast, for a eukaryote with sufficiently high N_e^{-1}, the same s_c might be effectively invisible to selection, even if $s_a = 0$. Thus even initially useless genetic material can be readily fixed in a population, making eukaryotes susceptible to non-coding "bloat" in the genome. But this also provides a rich palette of genetic materials from which the complex

variety of eukaryotic regulatory mechanisms can subsequently evolve [12, 21].

Part of the explanatory power of this idea is the fact that the s_c of a particular genetic variant should in principle be predictable from underlying physical principles. In fact, a very plausible hypothesis is that $s_c \approx -\delta C_T/C_T$, where C_T is the total resting metabolic expenditure of an organism per generation time, and δC_T is the extra expenditure of the mutant versus the wild-type. This relation can be traced at least as far back as the famous "selfish DNA" paper of Orgel and Crick [10], where it was mentioned in passing. But its true usefulness was only shown more recently, in the notable works of Wagner [11] on yeast and Lynch & Marinov [12] on a variety of prokaryotes and unicellular eukaryotes. By doing a detailed biochemical accounting of energy expenditures, they used the relation to derive values of s_c that provided intuitive explanations of the different selective pressures faced by different classes of organisms. The relation provides a Rosetta stone, translating biological thermodynamics into evolutionary terms. And its full potential is still being explored, most recently in describing the energetics of viral infection [22].

Our study poses a basic question: is this relation between s_c and metabolic expenditures true? For despite its plausibility and long pedigree, to our knowledge it has never been justified in complete generality from first principles. We do so through a general bioenergetic growth model, versions of which have been applied across the spectrum of life [23–25], from unicellular organisms to complex vertebrates. Even though the growth details, including the relative contributions of maintenance and synthesis costs, vary widely between different classes of organisms, we show that the relation is universal to an excellent approximation across the entire biological parameter range.

Growth model: Of an organism's net energy intake per unit time, some part of it is spent in locomotion and activities, some part of it stored (in certain organisms), and the remainder is consumed in the resting metabolism [24]. Let $\Pi(m(t))$ [unit: W] be the average power input into the resting metabolism, which can be an arbitrary function of the organism's current mass $m(t)$ [unit: g] at time t. Common choices for the functional form of $\Pi(m(t))$ will be discussed below. This power is partitioned into maintenance of existing biological mass (i.e. the turnover energy costs associated with the constant replacement of cellular components lost to degradation), and growth of new mass (i.e. synthesis of additional components during cellular replication) [26]. Energy conservation implies

$$\Pi(m(t)) = B(m(t))m(t) + E(m(t))\frac{dm}{dt}, \tag{1}$$

Here $B(m(t))$ [unit: W/g] is the maintenance cost per unit mass, and $E(m(t))$ [unit: J/g] is the synthesis cost per unit mass. We allow both these quantities to be arbitrary functions of $m(t)$.

Though we will derive our main result for the fully general model of Eq. (1), we will also explore a special case: $\Pi(m(t)) = \Pi_0 m^\alpha(t)$, $B(m(t)) = B_m$, $E(m(t)) = E_m$, with scaling exponent α and constants Π_0, B_m, and E_m [25]. Allometric scaling of $\Pi(m(t))$ with $\alpha = 3/4$ across many different species was first noted in the work of Max Kleiber in the 1930s [27], and with the assumption of time-independent $B(m(t))$ and $E(m(t))$ leads to a successful description of the growth curves of many higher animals [23, 24]. However, recently there has been evidence that $\alpha = 3/4$ may not be universal [28]. Higher animals still exhibit $\alpha < 1$ (with debate whether 3/4 or 2/3 is more appropriate [29]), but unicellular organisms have a broader range $\alpha \lesssim 2$. Thus we will use the model of Ref. [25] with an arbitrary species-dependent exponent α. Though time-independent $B(m(t))$ and $E(m(t))$ are reasonable as a first approximation, particularly for unicellular organisms, it is easy to imagine scenarios where for example the maintenance cost $B(m(t))$ might vary with $m(t)$ as part of the organism's developmental plan: as the organism approaches maturity, more energy might be allocated to reproductive functions [23] or heat production in endothermic animals [30], effectively increasing the cost of maintenance. Thus we initially consider the model in complete generality.

Baseline selection coefficient for metabolic costs: To derive an expression for s_c for the growth model of Eq. (1), we first focus on the time t_r to reproductive maturity, the typical time for the organism to grow from a mass $m(0) = m_0$ at birth to some mature mass $m_r = \epsilon m_0$. This time is related to the population growth rate r through $r = \ln(R_0)/t_r$, where R_0 is the basic reproductive ratio, the average number of offspring per individual surviving to reproductive maturity [25, 31]. For example in the case of binary fission of a unicellular organism, $\epsilon = 2$, and if one neglects cell deaths, $R_0 = 2$ as well. The value of ϵ can range much higher for more complex organisms, where m_r is typically the same order of magnitude as the asymptotic adult mass achieved in the long-time limit [23, 32]. Since $m(t)$ is a monotonically increasing function of t for any physically realistic growth model, we can invert Eq. (1) to write the infinitesimal time interval dt associated with an infinitesimal increase of mass dm as $dt = dm\, E(m)/G(m)$ where $G(m) \equiv \Pi(m) - B(m)m$ is the amount of power channeled to growth, and we have switched variables from t to m. Note that $G(m)$ must be positive over the m range to ensure that $dm/dt > 0$. Integrating dt gives us an expression for t_r,

$$t_r = \int_{m_0}^{\epsilon m_0} dm \frac{E(m)}{G(m)}. \tag{2}$$

Now consider a genetic variation in the organism that creates additional metabolic costs, but in keeping with our baseline assumption, does not alter biological function in any other respect. The

products of the genetic variation (i.e. extra mRNA transcripts or translated proteins) may alter the mass of the mutant, which we denote by $\tilde{m}(t)$. However the baseline assumption means that these changes do not affect the ability of the organism to assimilate energy for its resting metabolism, so that the left-hand side of Eq. (1) remains $\Pi(m(t))$, where $m(t)$ is now the *unperturbed* mass of the organism (the mass of all the pre-variation biological materials). The power input $\Pi(m(t))$ depends on $m(t)$ rather than $\tilde{m}(t)$ since only $m(t)$ contributes to the processes that allow the organism to process nutrients. It is also convenient to express our dynamics in terms of $m(t)$ rather than $\tilde{m}(t)$, since the condition defining reproductive time t_r remains unchanged, $m(t_r) = \epsilon m_0$, or in other words when the unperturbed mass reaches ϵ times the initial unperturbed mass m_0. Thus Eq. (1) for the mutant takes the form $\Pi(m(t)) = \tilde{B}(m(t)) + \tilde{E}(m(t))dm(t)/dt$, where $\tilde{B}(m(t)) = B(m(t)) + \delta B$ and $\tilde{E}(m(t)) = E(m(t)) + \delta E$ are the mutant maintenance and synthesis costs. We assume the perturbations δB and δE are independent of $m(t)$, though the theory can be generalized to models where the extra metabolic costs vary throughout the organism's development. In the Supplementary Information (SI), we show a sample calculation of δB and δE for mutations in *E.coli* and fission yeast involving short extra genetic sequences transcribed into non-coding RNA. This provides a concrete illustration of the framework we now develop.

Changes in the metabolic terms will result in a perturbation to the reproduction time, $\tilde{t}_r = t_r + \delta t_r$, and consequently the growth rate $\tilde{r} = r + \delta r$. The corresponding baseline selection coefficient s_c can be exactly related to $\tilde{s}_c \equiv -\delta t_r/t_r$, the fractional change in t_r, through $s_c = R_0^{\tilde{s}_c/(1-\tilde{s}_c)} - 1$ (see SI). This relation can be approximated as $s_c \approx \ln(R_0)\tilde{s}_c$ when $|\tilde{s}_c| \ll 1$, the regime of interest when making comparisons to drift barriers $N_e^{-1} \ll 1$. In this regime $\tilde{s}_c \approx \delta r/r$, the fractional change in growth rate. $\tilde{s}_c$ can be written in a way that directly highlights the contributions of δE and δB to $\tilde{s}_c$. This formulation involves the mathematical trick of introducing a function $p(m) \equiv t_r^{-1} E(m)/G(m)$. Eq. (2) then implies that $\int_{m_0}^{\epsilon m_0} dm\, p(m) = 1$, so we can treat $p(m)$ as a normalized "probability" over the mass range m_0 to ϵm_0. Note that since $p(m) = t_r^{-1} dt/dm$, the average of any function $F(m(t))$ over the reproductive time scale t_r can be expressed as as $\langle F \rangle \equiv \int_{m_0}^{\epsilon m_0} dm\, F(m) p(m)$. Expanding Eq. (2) for t_r to first order in the perturbations δE and δB, the coefficient $\tilde{s}_c = -\delta t_r/t_r = -\sigma_E \delta E/\langle E \rangle - \sigma_B \delta B/\langle B \rangle$, with positive dimensionless prefactors

$$\sigma_E \equiv \langle E \rangle \langle E^{-1} \rangle, \quad \sigma_B \equiv \langle B \rangle \langle \Theta^{-1} \rangle. \tag{3}$$

Here $\Theta(m) \equiv G(m)/m$. The magnitude of σ_B versus σ_E describes how much fractional increases in maintenance costs matter for selection relative to fractional increases in synthesis costs.

Relating the baseline selection coefficient to the fractional change in total resting metabolic

costs: The final step in our theoretical framework is to connect the above considerations to the total resting metabolic expenditure C_T of the organism per generation time t_r, given by $C_T = \zeta \int_0^{t_r} dt\, \Pi(m(t)) = \zeta t_r \langle \Pi \rangle$. In order to facilitate comparison with the experimental data of Ref. [12], compiled in terms of phosphate bonds hydrolyzed [P], we add the prefactor ζ which converts from units of J to P. Assuming an ATP hydrolysis energy of 50 kJ/mol under typical cellular conditions, we set $\zeta = 1.2 \times 10^{19}$ P/J. The genetic variation discussed above leads to a modified total cost $\tilde{C}_T = C_T + \delta C_T$, and the fractional change $\delta C_T/C_T$ can be expressed in a form analogous to $\tilde{s}_c$, namely $\delta C_T/C_T = \sigma'_E \delta E/\langle E \rangle + \sigma'_B \delta B/\langle B \rangle$, with

$$\sigma'_E \equiv \langle E \rangle \langle \Pi \rangle^{-1} \langle \Pi E^{-1} \rangle, \quad \sigma'_B \equiv \langle B \rangle \langle \Pi \rangle^{-1} \langle \Pi \Theta^{-1} \rangle. \tag{4}$$

The connection between s_c and $\delta C_T/C_T$ can be constructed by comparing Eq. (3) with Eq. (4). We see that $\tilde{s}_c = -\delta C_T/C_T$ for all possible perturbations δE and δB only when $\sigma_E = \sigma'_E$ and $\sigma_B = \sigma'_B$. Thus the accuracy of the relation hinges on the degree to which these coefficients agree with one another. The relative differences can be written as:

$$\begin{aligned} \left| 1 - \frac{\sigma'_E}{\sigma_E} \right| &= \left| 1 - \frac{\langle \Pi E^{-1} \rangle}{\langle \Pi \rangle \langle E^{-1} \rangle} \right| \leq \kappa(\Pi)\kappa(E^{-1}), \\ \left| 1 - \frac{\sigma'_B}{\sigma_B} \right| &= \left| 1 - \frac{\langle \Pi \Theta^{-1} \rangle}{\langle \Pi \rangle \langle \Theta^{-1} \rangle} \right| \leq \kappa(\Pi)\kappa(\Theta^{-1}). \end{aligned} \tag{5}$$

where $\kappa(F) \equiv \sqrt{\langle F^2 \rangle - \langle F \rangle^2}/\langle F \rangle$ and we have used the Cauchy-Schwarz inequality. These bounds imply two cases when the relation is exact: i) $\kappa(\Pi) = 0$, which means $\Pi(m)$ is a constant independent of m; ii) $\kappa(\Pi) > 0$ and $\kappa(E^{-1}) = \kappa(\Theta^{-1}) = 0$, which means $E(m)$ and $\Theta(m)$ are independent of m. Outside these cases, the relation $\tilde{s}_c \approx -\delta C_T/C_T$ is an approximation. To see how well it holds, it is instructive to investigate the allometric growth model described earlier, where $\Pi(m(t)) = \Pi_0 m^\alpha(t)$, $E(m(t)) = E_m$ and $B(m(t)) = B_m$.

Testing the relation in an allometric growth model. Though the values of the model parameters will vary between species, we know that the exponent $\alpha \lesssim 2$ [25, 28], and there is a rough scale for E_m and B_m that can be established through comparison to metabolic data compiled in Ref. [12] covering a variety of prokaryotes and unicellular eukaryotes. This data consisted of two quantities, C_G and C_M, which reflect the growth and maintenance contributions to C_T. Using Eq. (1) to decompose $\Pi(m(t))$, we can write $C_T = C_G + t_r C_M$, where $C_G = \zeta \int_{m_0}^{\epsilon m_0} dm\, E(m) = \zeta(\epsilon - 1) m_0 E_m$ is the expenditure for growing the organism, and $C_M = \zeta \langle Bm \rangle = \zeta B_m \langle m \rangle$ is the mean metabolic expenditure for maintenance per unit time. As shown in the SI, the simplest version of the allometric growth model predicts linear scaling of both C_G and C_M with cell volume. Best fits of the model to the data, shown in Fig. 1, yield global interspecies averages: $E_m = 1,300$ J/g and

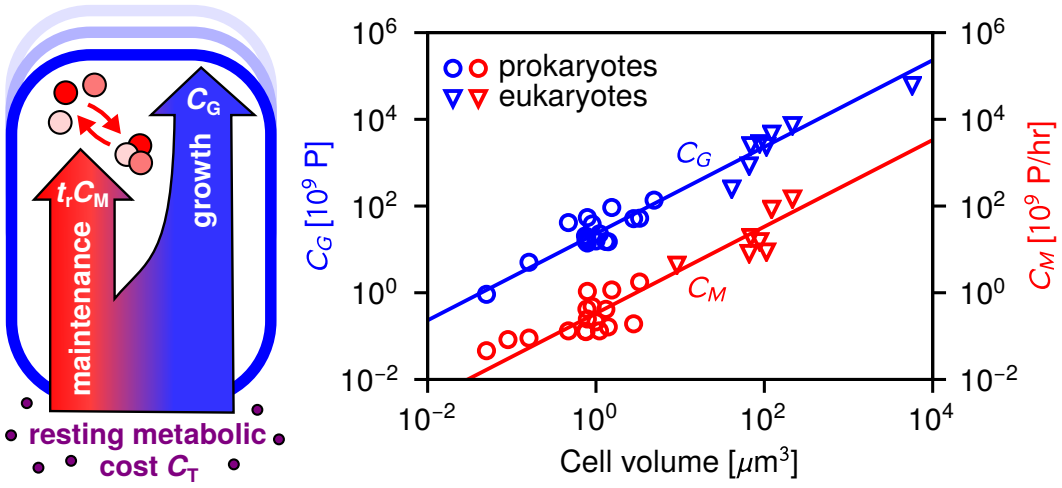

FIG. 1. The growth C_G (blue) and maintenance C_M (red) contributions to an organism's total resting metabolic cost $C_T = C_G + t_r C_M$ per generation time t_r. The symbols (circles = prokaryotes, triangles = unicellular eukaryotes) represent data tabulated in Ref. [12], as a function of cell volume. C_G and C_M are measured in units of 10^9 P (phosphate bonds hydrolyzed), and 10^9 P/hr respectively. As shown in SI Eq. (S6), the allometric growth model predicts linear scaling of C_G and C_M with cell volume. The solid lines represent best fits to Eq. (S6) with parameters $E_m = 1,300$ J/g and $B_m = 7 \times 10^{-3}$ W/g.

$B_m = 7 \times 10^{-3}$ W/g. As discussed in the SI, these values are remarkably consistent with earlier, independent estimates, for unicellular and higher organisms [24, 25, 33, 34].

Since $E(m(t)) = E_m$ is a constant in the allometric growth model, $\sigma_E = 1$ from Eq. (3). Additionally, because $\kappa(E^{-1}) = 0$ we know that $\sigma_E = \sigma'_E$ holds exactly. So the only aspect of the approximation that needs to be tested is the similarity between σ_B and σ'_B. Fig. 2A shows σ_B versus σ'_B for the range $\alpha = 0 - 3$, which includes the whole spectrum of biological scaling [28] up to $\alpha = 2$, plus some larger α for illustration. The E_m and B_m parameters have been set to the unicellular best-fit values quoted above, and $\epsilon = 2$. For a given α, the coefficient Π_0 has been set to yield a certain division time t_r, ranging from $t_r = 1$ hr (purple pair of curves at the bottom) to $t_r = 40$ hr (red pair of curves at the top). These roughly encompass both the fast and slow extremes of typical unicellular reproductive times. In all cases σ'_B is in excellent agreement with σ_B. For the range $\alpha \leq 2$ the discrepancy is less than 15%, and it is in fact zero at the special points $\alpha = 0$ (where $\kappa(\Pi) = 0$) and $\alpha = 1$ (where $\kappa(\Theta^{-1}) = 0$). Clearly the approximation begins to break down at $\alpha \gg 1$, but it remains sound in the biologically relevant regimes. Note that σ_B values for $t_r = 1$ hr are ~ 0.03, reflecting the minimal contribution of maintenance relative to synthesis costs in determining the selection coefficient for fast-dividing organisms. Indeed, an often used approximation in this case is to ignore maintenance costs altogether, in which case $\tilde{s}_c = -\delta C_T / C_T$

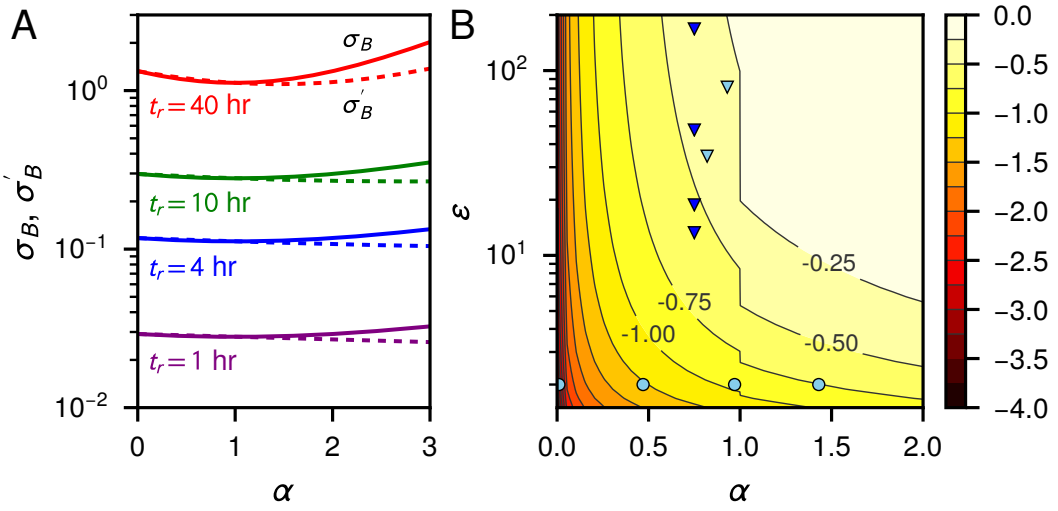

FIG. 2. A: σ_B (solid curves) from Eq. (3) and σ'_B (dashed curves) from Eq. (4) versus α, for the allometric growth model with $E_m = 1,300$ J/g, $B_m = 7 \times 10^{-3}$ W/g, and $\epsilon = 2$. At any given α, the parameter Π_0 for each pair of curves (different colors) is chosen to correspond to particular reproductive times t_r, indicated in the labels. B: Contour diagram showing the logarithm of the maximum possible discrepancy $\log_{10} |1 - \sigma'_B/\sigma_B|$ for any allometric growth model parameters, as a function of α and ϵ. To illustrate biological ranges α and ϵ, the symbols correspond to data for various species drawn from the growth trajectories analyzed in Ref. [25] (light blue) and Ref. [23] (dark blue). Circles are unicellular organisms, and triangles multicellular organisms (a detailed list is provided in the SI).

exactly, since only $\sigma_E = \sigma_{E'} = 1$ matters. The result for $\tilde{s}_c$ in this limit is consistent with microbial metabolic flux theory [35], where the maintenance costs are typically neglected. As t_r increases, so does σ_B. At the other extreme, when $t_r = 40$ hr, $\sigma_B \sim 1.1 - 1.3 > \sigma_E$, and the influence of maintenance becomes comparable to synthesis.

To make a more comprehensive analysis of the validity of the $\tilde{s}_c \approx -\delta C_T/C_T$ relation, we do a computational search for the worst case scenarios: for each value of α and ϵ, we can numerically determine the set of other growth model parameters that gives the largest discrepancy $|1 - \sigma'_B/\sigma_B|$. Fig. 2B shows a contour diagram of the results on a logarithmic scale, $\log_{10} |1 - \sigma'_B/\sigma_B|$, as a function of α and ϵ. Estimated values for α and ϵ from the growth trajectories of various species are plotted as symbols to show the typical biological regimes. While the maximum discrepancies are smaller for the parameter ranges of unicellular organisms (circles) compared to multicellular ones (triangles), in all cases the discrepancy is less than 50%. To observe a serious error, where σ'_B is no longer the same order of magnitude as σ_B, one must go to the large α, large ϵ limit (top right of the diagram) which no longer corresponds to biologically relevant growth trajectories.

We thus reach the conclusion that the baseline selection coefficient for metabolic costs can be reliably approximated as $s_c \approx -\ln(R_0)\delta C_T/C_T$. As in the original hypothesis [10–12], $-\delta C_T/C_T$ is the dominant contribution to the scale of s_c, with corrections provided by the logarithmic factor $\ln(R_0)$. Our derivation puts the relation for s_c on a solid footing, setting the stage for its wider deployment. It deserves a far greater scope of applications beyond the pioneering studies of Refs. [11, 12, 22]. Knowledge of s_c can also be used to deduce the adaptive contribution $s_a = s - s_c$ of a mutation, which has its own complex connection to metabolism [36]. The latter requires measurement of the overall selection coefficient s, for example from competition/growth assays, and the calculation of s_c from the relation, assuming the underlying energy expenditures are well characterized. The s_c relation underscores the key role of thermodynamic costs in shaping the interplay between natural selection and genetic drift. Indeed, it gives impetus to a major goal for future research: a comprehensive account of those costs for every aspect of biological function, and how they vary between species, what one might call the "thermodynome". Relative to its more mature omics brethren—the genome, proteome, transcriptome, and so on—the thermodynome is still in its infancy, but fully understanding the course of evolutionary history will be impossible without it.

ACKNOWLEDGMENTS

The authors thank useful correspondence with M. Lynch, and feedback from B. Kuznets-Speck and C. Weisenberger. E.I. acknowledges support from Institut Curie.

[1] E. Dekel and U. Alon, Nature **436**, 588 (2005).

[2] K. A. Dill, K. Ghosh, and J. D. Schmit, Proc. Natl. Acad. Sci. **108**, 17876 (2011).

[3] P. R. ten Wolde, N. B. Becker, T. E. Ouldridge, and A. Mugler, J. Stat. Phys. **162**, 1395 (2016).

[4] D. Hathcock, J. Sheehy, C. Weisenberger, E. Ilker, and M. Hinczewski, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 16 (2016).

[5] C. Zechner, G. Seelig, M. Rullan, and M. Khammash, Proc. Natl. Acad. Sci. **113**, 4729 (2016).

[6] S. Fancher and A. Mugler, Phys. Rev. Lett. **118**, 078101 (2017).

[7] M. Kimura, Genetics **47**, 713 (1962).

[8] T. Ohta, Nature **246**, 96 (1973).

[9] T. Ohta and J. H. Gillespie, Theor. Popul. Biol. **49**, 128 (1996).

[10] L. E. Orgel and F. H. Crick, Nature **284**, 604 (1980).

[11] A. Wagner, Mol. Biol. Evol. **22**, 1365 (2005).

[12] M. Lynch and G. K. Marinov, Proc. Natl. Acad. Sci. **112**, 15690 (2015).

[13] J. H. Gillespie, *Population genetics: a concise guide* (JHU Press, 2010).

[14] W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch, Proc. Natl. Acad. Sci. **109**, 18488 (2012).

[15] B. Charlesworth, Nature Rev. Genet. **10**, 195 (2009).

[16] M. Lynch and J. S. Conery, Science **302**, 1401 (2003).

[17] M. Lynch, Mol. Biol. Evol. **23**, 450 (2005).

[18] E. V. Koonin, BMC Biol. **14**, 114 (2016).

[19] I. Sela, Y. I. Wolf, and E. V. Koonin, Proc. Natl. Acad. Sci. **113**, 11399 (2016).

[20] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, Science **330**, 1099 (2010).

[21] R. J. Taft, M. Pheasant, and J. S. Mattick, Bioessays **29**, 288 (2007).

[22] G. Mahmoudabadi, R. Milo, and R. Phillips, Proc. Natl. Acad. Sci. **114**, E4324 (2017).

[23] G. B. West, J. H. Brown, and B. J. Enquist, Nature **413**, 628 (2001).

[24] C. Hou, W. Zuo, M. E. Moses, W. H. Woodruff, J. H. Brown, and G. B. West, Science **322**, 736 (2008).

[25] C. P. Kempes, S. Dutkiewicz, and M. J. Follows, Proc. Natl. Acad. Sci. **109**, 495 (2012).

[26] S. Pirt, Proc. R. Soc. Lond. B **163**, 224 (1965).

[27] M. Kleiber, Hilgardia **6**, 315 (1932).

[28] J. P. DeLong, J. G. Okie, M. E. Moses, R. M. Sibly, and J. H. Brown, Proc. Natl. Acad. Sci. **107**, 12941 (2010).

[29] C. R. White and R. S. Seymour, Proc. Natl. Acad. Sci. **100**, 4046 (2003).

[30] J. Werner, N. Sfakianakis, A. D. Rendall, and E. M. Griebeler, J. Theor. Biol. **444**, 83 (2018).

[31] V. M. Savage, J. F. Gillooly, J. H. Brown, G. B. West, and E. L. Charnov, Am. Nat. **163**, 429 (2004).

[32] R. E. Ricklefs, Proc. Natl. Acad. Sci. **107**, 10314 (2010).

[33] M. E. Moses, C. Hou, W. H. Woodruff, G. B. West, J. C. Nekola, W. Zuo, and J. H. Brown, Am. Nat. **171**, 632 (2008).

[34] A. Maitra and K. A. Dill, Proc. Natl. Acad. Sci. **112**, 406 (2015).

[35] J. Berkhout, E. Bosdriesz, E. Nikerel, D. Molenaar, D. de Ridder, B. Teusink, and F. J. Bruggeman, Genetics **194**, 505 (2013).

[36] M. N. Price and A. P. Arkin, Genome Biol. Evol. **8**, 1917 (2016).

Supplementary Information: Modeling the growth of organisms validates a general relation between metabolic costs and natural selection

Efe Ilker[1,2] and Michael Hinczewski[2]

[1] *Physico-Chimie Curie UMR 168, Institut Curie, PSL Research University, 26 rue d'Ulm, 75248 Paris Cedex 05, France*

[2] *Department of Physics, Case Western Reserve University, Cleveland OH 44106*

1. Derivation of the relation between s_c and $\tilde{s}_c$

In the main text we posited an approximate relation $s_c \approx \ln(R_0)\tilde{s}_c$ between the baseline selection coefficient s_c and the fractional change in growth rate $\tilde{s}_c$ due to a genetic variation. Here we derive an exact relation between the two quantities, generalizing the approach used in Ref. [1] for the specific case of binary fission. We then show how the approximation used in the main text arises in the limit $|\tilde{s}_c| \ll 1$.

Consider a group of wild-type organisms with population $N_w(t)$ as a function of time, and a group of mutant organisms with population $N_m(t)$. Under our baseline assumption (neglecting adaptive contributions), the selection coefficient associated with the mutation is s_c, and both types of organisms have the same average number of surviving offspring per generation R_0. For example $R_0 = 2$ for binary fission neglecting cell deaths. In general, $R_0 = p_{\rm r} f$, where $p_{\rm r}$ is the fraction of the population to survive until the age of reproduction, and f is the average fecundity [2, 3].

We assume both populations are in a regime of exponential (Malthusian) growth, so that $N_w(t) = N_w(0)\exp(rt)$ and $N_m(t) = N_w(0)\exp((r+\delta r)t)$ with respective population growth rates r and $r+\delta r$. If t_r is the mean generation time of the wild-type, and $t_r + \delta t_r$ that of the mutant, the growth rates are given by $r = \ln(R_0)/t_r$ and $r + \delta r = \ln(R_0)/(t_r + \delta t_r)$. If the populations were initially equal, $N_m(0) = N_w(0)$, we can thus write the ratio of populations at any subsequent time as

$$\frac{N_m(t)}{N_w(t)} = e^{t\delta r} = R_0^{t\left(\frac{1}{t_r+\delta t_r} - \frac{1}{t_r}\right)}. \tag{S1}$$

On the other hand, after n wild-type generations ($t = nt_r$) the ratio of the two populations is related to the selection coefficient (as conventionally defined in population genetics) through

$$\frac{N_m(nt_r)}{N_w(nt_r)} = (1+s_c)^n. \tag{S2}$$

Plugging $t = nt_r$ into Eq. (S1) and comparing to Eq. (S2), we see that

$$s_c = R_0^{-\frac{\delta t_r}{t_r+\delta t_r}} - 1. \tag{S3}$$

If we define $\tilde{s}_c \equiv -\delta t_r/t_r$, then Eq. (S3) can be rewritten as

$$s_c = R_0^{\frac{\tilde{s}_c}{1-\tilde{s}_c}} - 1. \tag{S4}$$

Note that in the case where $|\tilde{s}_c| \ll 1$, or $|\delta t_r| \ll t_r$, we can also write $\tilde{s}_c \approx \delta r/r$, and so interpret $\tilde{s}_c$ as the fractional change in growth rate. In this same limit we can expand Eq. (S4) for small $\tilde{s}_c$,

$$s_c = \ln(R_0)\tilde{s}_c + \frac{1}{2}\ln(R_0)\left(2 + \ln(R_0)\right)\tilde{s}_c^2 + \cdots . \tag{S5}$$

Keeping only the leading order term, linear in $\tilde{s}_c$, yields the approximation $s_c \approx \ln(R_0)\tilde{s}_c$.

2. Fitting of allometric growth model to experimental data

As discussed in the main text, we can decompose C_T into two components, $C_T = C_G + t_r C_M$, where $C_G = \zeta \int_{m_0}^{\epsilon m_0} dm\, E(m)$ is the expenditure for growing the organism, and $C_M = \zeta\langle Bm\rangle$ is the mean metabolic expenditure for maintenance per unit time. For the allometric growth model, these contributions to C_T simplify to $C_G = \zeta(\epsilon - 1)m_0 E_m$ and $C_M = \zeta B_m \langle m\rangle$. Ref. [4] noted that C_G and C_M collected from experimental data scaled nearly linearly with cell volume, with allometric exponents of 0.97 ± 0.04 and 0.88 ± 0.07 respectively. In fact, the simplest version of the allometric model predicts exactly linear scaling, using the following assumptions. Since the data tabulated in Ref. [4] covers prokaryotes and unicellular eukaryotes, we take $\epsilon = 2$. Since the mass of the organism varies between m_0 and $2m_0$ over time t_r, we approximate $\langle m\rangle \approx (3/2)m_0$. Any errors in this approximation, or variance in ϵ, will not change the order of magnitude of the estimated model parameters. We relate the experimentally observed cell volume V to the mean cell mass $\langle m\rangle$ by assuming a typical cell is 2/3 water (density $\rho_{\rm wat} = 10^{-12}\,\mathrm{g}/\mu\mathrm{m}^3$) and 1/3 dry biomass (density $\rho_{\rm dry} \approx 1.3 \times 10^{-12}\,\mathrm{g}/\mu\mathrm{m}^3$) [5]. Hence $\langle m\rangle = (2\rho_{\rm wat} + \rho_{\rm dry})V/3 \equiv \rho_{\rm cell} V$. We thus find:

$$C_G = (4/3)\zeta E_m \rho_{\rm cell} V, \qquad C_M = \zeta B_m \rho_{\rm cell} V. \tag{S6}$$

For each expression we have only one unknown parameter, E_m and B_m respectively. Best fits to the Ref. [4] data, shown in main text Fig. 1B, yield global interspecies averages of the parameters, $E_m = 1,300$ J/g and $B_m = 7 \times 10^{-3}$ W/g.

The fitted values are consistent with earlier approaches, once water content is accounted for (i.e. to get E_m per dry biomass, multiply the value by ≈ 3, so $E_m^{\rm dry} = 3,900$ J/g). The synthesis cost

E_m has a very narrow range across many species, with $E_m = 1,100 - 1,800$ J/g in bird and fish embryos, and $4,000 - 7,500$ J/g for mammal embryos and juvenile fish, birds, and mammals [6]. This energy scale seems to persist down to the prokaryotic level, with $E_m^{dry} = 3,345$ J/g estimated for *E. coli* [3]. $E_m^{\rm dry}$ also appears in a different guise as the inverse of the "energy efficiency" ε of *E. coli* growth in the model of Ref. [7]; converting the optimal observed $\varepsilon \approx 15$ dry g/(mol ATP) yields $E_m^{dry} = \zeta/\varepsilon = 3,333$ J/g, consistent with the other estimates cited above, as well as our fitted value. The ratio B_m/E_m was estimated for various species in Ref. [3], and found to vary in the range $10^{-6} - 10^{-5}$ s^{-1} from prokaryotes to unicellular eukaryotes, entirely consistent with our fitted value of $B_m/E_m = 5 \times 10^{-6}$ s^{-1}. The scale shifts for larger, multicellular species, but not dramatically. For example for a subset of mammals with scaling $\alpha = 3/4$, adult mass sizes $m_a = 10 - 6.5 \times 10^5$ g, and typical values of $B_0 \approx 0.022$ W/g$^{3/4}$, $E_m \approx 7000$ J/g [8], we get a range of $B_m/E_m = 10^{-7} - 10^{-6}$ s^{-1}. We thus have confidence that the growth model provides a description of the metabolic expenditures (in terms of growth and maintenance contributions) that is consistent both with the empirical data of Ref. [4] and parameter expectations based on a variety of earlier approaches.

For the symbols in the contour diagram of main text figure Fig. 2B, we used parameters extracted from growth trajectories analyzed in Ref. [3] (light blue) and Ref. [9] (dark blue). Circles (left to right) are unicellular organisms ($\epsilon = 2$): *T. weissflogii*, *L. borealis*, *B. subtilis*, *E. coli*. Triangles (top to bottom) are multicellular organisms: guinea pig, *C. pacificus*, hen, *Pseudocalanus sp.*, guppy, cow. For the multicellular case the plotted values of ϵ correspond to asymptotic adult mass in units of m_0. This is an upper bound on ϵ, though the actual ϵ should typically be comparable [9, 10].

3. Sample calculation of the baseline selection coefficient: short, non-coding RNA in *E. coli* and fission yeast

To illustrate a calculation of baseline selection coefficients in the framework developed in the main text, let us consider a specific biological example: a mutant with a short (< 200 bp) sequence in the genome that is transcribed into non-coding RNA, and which is not present in the wild-type. We will focus on two organisms, the prokaryote *E. coli* and the unicellular eukaryote *S. pombe* (fission yeast). To date we know that at least some subset of non-coding RNA transcripts have functional roles in these organisms [11, 12]. The evolution of such regulatory sequences will be shaped both by the selective advantage s_a of having the sequence in the genome, and the baseline disadvantage s_c from the extra energetic costs of copying and transcription.

Before calculating s_c, we first establish the validity of the growth model for these organisms.

The model parameters fitted for the data from prokaryotes and unicellular eukaryotes in Fig. 1B of the main text are $E_m = 1,300$ J/g and $B_m = 7 \times 10^{-3}$ W/g. The corresponding growth and maintenance contributions to the total resting metabolic cost per generation, C_G and C_M, are given by Eq. (S6). Using $\zeta = 1.2 \times 10^{19}$ P/J (recall that P corresponds to ATP or ATP equivalents hydrolyzed), $\rho_{\rm cell} = 1.1 \times 10^{-12}$ g/μm^3, and typical cell volumes $V_{E.coli} = 1$ μm^3 [5], $V_{S.pombe} = 106$ μm^3 [13], we find: $C_G^{E.coli} = 2.30 \times 10^{10}$ P, $C_M^{E.coli} = 3.34 \times 10^8$ P/hr, $C_G^{S.pombe} = 2.43 \times 10^{12}$ P, $C_M^{S.pombe} = 3.54 \times 10^{10}$ P/hr. These agree well in magnitude with the literature estimates compiled in the SI of Ref. [4] (all normalized to 20°C): $C_G^{E.coli} = 1.57 \times 10^{10}$ P, $C_M^{E.coli} = 2.13 \times 10^8$ P/hr, $C_G^{S.pombe} = 2.35 \times 10^{12}$ P, $C_M^{S.pombe} = 8.7 \times 10^9$ P/hr. Thus the globally fitted E_m and B_m values are physically reasonable for both organisms.

The extra sequence in the mutant leads to perturbations in both synthesis cost per unit mass, δE, and maintenance cost per unit mass, δB. To calculate the first, we use the following estimates based on the analysis in Ref. [4]: for a sequence of length L, the total DNA-related synthesis cost is $d_\xi L$, where the label $\xi =$ *E. coli* or *S. pombe*. Here the prefactor $d_{E.coli} \approx 101$ P and $d_{S.pombe} \approx 263$ P. If the steady-state average number of corresponding mRNA trascripts in the cell is N_r, the additional ribonucleotide synthesis costs are $\approx 46 N_r L$ in units of P. Hence we have, per unit mass,

$$\delta E \approx \frac{\zeta^{-1} L}{\rho_{\rm cell} V_\xi} \left(d_\xi + 46 N_r \right), \tag{S7}$$

with the ζ^{-1} prefactor converting from P to J, so that δE has units of J/g. The same analysis [4] yields the maintenance cost per unit time for replacing transcripts after degradation, $\approx 2 N_r L \gamma_\xi$ in units of P/s, where $\gamma_{E.coli} = 0.003$ s^{-1} and $\gamma_{S.pombe} = 0.001$ s^{-1} are the RNA degradation rates for the two organisms. Per unit mass, the maintenance perturbation δB is given by

$$\delta B \approx \frac{2 \zeta^{-1} L N_r \gamma_\xi}{\rho_{\rm cell} V_\xi}, \tag{S8}$$

in units of W/g.

The final step is to calculate the prefactors σ_E and σ_B from Eq. (3) in the main text. For this we need to choose a particular growth model exponent α, and we set $\alpha = 1$, corresponding to the assumption of exponential cell mass growth. In this case $\sigma_E = 1$ for both organisms, while $\sigma_B^{E.coli} = 0.0140$, $\sigma_B^{S.pombe} = 0.121$. The choice of α has a minimal influence on the prefactors: $\sigma_E = 1$ exactly for any model with a constant function $E(m) = E_m$. Moreover, any α value in the biologically relevant range of $0 \leq \alpha \leq 2$ yields a σ_B value within 5% of the $\alpha = 1$ result for each organism.

Putting everything together, we now can calulate all the components of main text Eq. (3) for $\tilde{s}_c$, namely σ_E, σ_B, δE, δB, $\langle E \rangle = E_m$, and $\langle B \rangle = B_m$. Had we chosen instead to use the $\delta C_T / C_T$

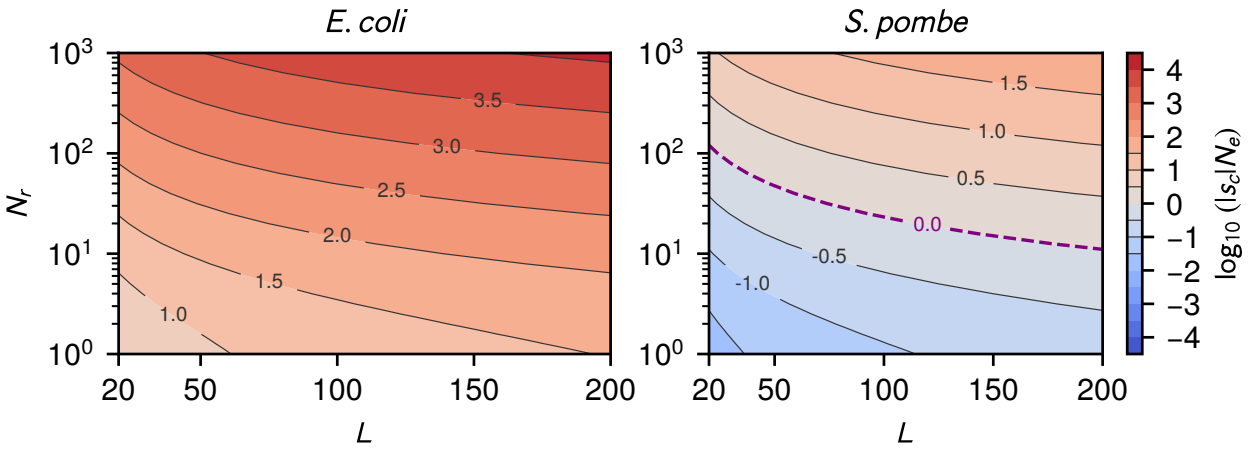

FIG. S1. Contour diagrams of $\log_{10}(|s_c|N_e)$ as a function of sequence length L and mean RNA transcript number N_r per cell for *E. coli* (left) and *S. pombe* (right). The dashed line in the diagram on the right corresponds to $|s_c| = N_e^{-1}$.

approximation of main text Eq. (4), the only discrepancy would have been in the fact that $\sigma'_B \neq \sigma_B$, since $\sigma'_E = \sigma_E = 1$. However the discrepancy is small, with $|1 - \sigma'_B/\sigma_B| < 0.09$ for both organisms in the range $0 \leq \alpha \leq 2$.

Fig. S1 shows contour diagrams of $\log_{10}(|s_c|N_e)$ as a function of L and N_r for *E. coli* and *S. pombe*. Here $s_c = \ln(2)\tilde{s}_c$, assuming $R_0 = 2$, and the effective population sizes are $N_e^{E.coli} = 2.5 \times 10^7$ [14], $N_e^{S.pombe} = 1.2 \times 10^7$ [15]. For *E.coli*, with its smaller metabolic expenditures per generation relative to fission yeast, the cost of the extra sequence is always significant: $|s_c| > N_e^{-1}$ for the entire range of L and N_r considered, even for the smallest length ($L = 20$ bp) and a single transcript per cell on average, $N_r = 1$. Thus there will always be strong selective pressure to remove the extra sequence, unless s_c is compensated for by a comparable or greater adaptive advantage s_a. In contrast, for *S. pombe* there is a regime of L and N_r where $|s_c| < N_e^{-1}$ (the region below the dashed line). Here the selective disadvantage of the extra sequence is weaker than genetic drift, and such a genetic variant could fix in the population at roughly the same rate as a neutral mutation even if it conferred no selective advantage, $s_a = 0$. While this makes fission yeast more tolerant of genomic "bloat" relative to *E. coli*, initially non-functional extra genetic material could subsequently facilitate the development of novel regulatory mechanisms.

[1] L.-M. Chevin, Biol. Lett. **7**, 210 (2011).

[2] V. M. Savage, J. F. Gillooly, J. H. Brown, G. B. West, and E. L. Charnov, Am. Nat. **163**, 429 (2004).

[3] C. P. Kempes, S. Dutkiewicz, and M. J. Follows, Proc. Natl. Acad. Sci. **109**, 495 (2012).

[4] M. Lynch and G. K. Marinov, Proc. Natl. Acad. Sci. **112**, 15690 (2015).

[5] R. Milo and R. Phillips, *Cell biology by the numbers*, `http://book.bionumbers.org/` (2017).

[6] M. E. Moses, C. Hou, W. H. Woodruff, G. B. West, J. C. Nekola, W. Zuo, and J. H. Brown, Am. Nat. **171**, 632 (2008).

[7] A. Maitra and K. A. Dill, Proc. Natl. Acad. Sci. **112**, 406 (2015).

[8] C. Hou, W. Zuo, M. E. Moses, W. H. Woodruff, J. H. Brown, and G. B. West, Science **322**, 736 (2008).

[9] G. B. West, J. H. Brown, and B. J. Enquist, Nature **413**, 628 (2001).

[10] R. E. Ricklefs, Proc. Natl. Acad. Sci. **107**, 10314 (2010).

[11] R. Raghavan, E. A. Groisman, and H. Ochman, Genome Res. **21**, 1487 (2011).

[12] H. S. Leong, K. Dawson, C. Wirth, Y. Li, Y. Connolly, D. L. Smith, C. R. Wilkinson, and C. J. Miller, Nat. Commun. **5**, 3947 (2014).

[13] F. Chang, Mol. Biol. Cell **28**, 1819 (2017).

[14] J. Charlesworth and A. Eyre-Walker, Mol. Biol. Evol. **23**, 1348 (2006).

[15] A. Farlow, H. Long, S. Arnoux, W. Sung, T. G. Doak, M. Nordborg, and M. Lynch, Genetics **201**, 737 (2015).