Rafting a waterfall: Artificial selection for collective composition can succeed or fail depending on the initial and target values

Juhee Lee, Wenying Shou*, and Hye Jin Park*

Collectives, such as microbial communities, can perform functions beyond the capability of individual members. Enhancing these collective functions through artificial selection, however, presents significant challenges. Here, we explore the 'rafting-a-waterfall' phenomenon, a metaphor illustrating how the success in achieving a target population composition in microbial collectives depends on both the target characteristics and initial conditions. Specifically, collectives comprising fast-growing (F) and slow-growing (S) individuals were grown for a period of "maturation" time, and the collective with S-frequency closest to the target value is chosen to "reproduce" (inoculate) offspring collectives. Such collective selection is dictated by two opposing forces: during collective maturation, intra-collective selection acts like a waterfall, relentlessly driving the S-frequency to lower values, while during collective reproduction, inter-collective selection resembles a rafter striving to reach the target frequency. Due to this model structure, maintaining a target frequency requires the continued action of inter-collective selection. Using simulations and analytical calculations, we show that intermediate target S frequencies are the most challenging, akin to a target within the vertical drop of a waterfall, rather than above or below it. This arises because intra-collective selection is the strongest at intermediate S-frequencies, which can overpower inter-collective selection. While achieving low target S frequencies is consistently feasible, attaining high target S-frequencies requires an initially high S-frequency — much like a raft that can descend but not ascend a waterfall. The range of attainable target frequencies depends on the initial population size of the collectives: as the population size in Newborn collectives increases, the region of achievable target frequency is reduced until no frequency is achievable. In contrast, the number of collectives under selection plays a less critical role. In scenarios involving more than two populations, the evolutionary trajectory must navigate entirely away from the metaphorical 'waterfall drop.' Our findings illustrate that the strength of intra-collective evolution is frequency-dependent, with implications in experimental planning.

INTRODUCTION

Microbial collectives can carry out functions that arise from interactions among member species. These functions, such as waste degradation [1, 2], probiotics [3], and vitamin production [4], can be useful for human health and biotechnology. To improve collective functions, one can perform artificial selection (directed evolution) on collectives: Low-density "Newborn" collectives are allowed to "mature" during which cells proliferate and possibly mutate, and community function develops. "Adult" collectives with high functions are then chosen to reproduce, each seeding multiple offspring Newborns. Artificial selection of collectives have been attempted both in experiments [5–19] and in simulations [20–31], often with unimpressive outcomes.

One of the major challenges in selecting collectives is to ensure the inheritance of a collective function [32, 33]. Inheritance from a parent collective to offspring collectives can be compromised by changes in genotype and species compositions. During maturation of a collective, genotype compositions within each species can change due to intra-collective selection favoring fast-growing individuals, while species compositions can change due to ecological interactions. Furthermore, during reproduction of a collective, genotype and species compositions of offspring can vary stochastically from those of the parent.

Here, we consider the selection of collectives comprising two or three populations with different growth rates, and our goal is to achieve a target composition in the Adult collective. This is a common quest: whenever a collective function depends on both populations, collective function is maximised, by definition, at an intermediate frequency (e.g. too little of either population will hamper function [23]). Earlier work has demonstrated that nearly any target species composition can be achieved when selecting communities of two competing species with unequal growth rates [24, 34], so long as the shared resource is depleted during collective maturation [24]. In this case initially, both species evolved to grow faster, and the slower-growing species was preserved due to stochastic fluctuations in species composition during collective reproduction. Eventually, both species evolved to grow sufficiently fast to deplete the shared resource during collective maturation, and evolution in competition coefficients then acted to stabilize species ratio to the target value [24]. Regardless, earlier studies are often limited to numerical explorations, with prohibitive costs for a full characterization of the parameter space for such nested populations (population of collectives, and populations of mutants within a collective).

We mathematically examine the selection of composition in collectives consisting of populations growing at different rates. A selection cycle consists of three stages (Fig. 1). During collective maturation, intra-collective selection favors fast-growing individuals within a collective. At the end of maturation, inter-collective selection acts on collectives and favors those achieving the target composition. Finally during collective reproduction, offspring collectives sample stochastically from the parents, a process dominated by genetic drift. We made simplifying assumptions so that we can analytically examine the evolutionary tipping point between intra-collective and inter-collective selection. We show that this tipping point creates a "waterfall" effect which restricts not only which target compositions are achievable, but also the initial composition required to achieve the target. We also investigate how the range of achievable target composition is affected by the total population size in Newborns and the total number of collectives under selection. Finally, we show that the waterfall phenomenon extends to systems with more than 2 populations.

RESULTS AND DISCUSSIONS

To enable the derivation of an analytical expression, we have made the following simplifications. First, growth is always exponential, without complications such as resource limitation, ecological

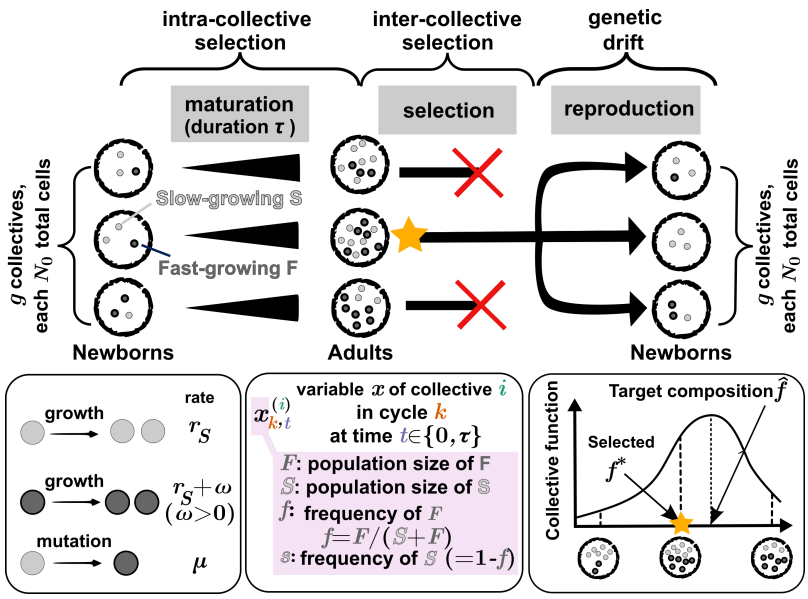

Figure 1. **Schematic for artificial selection on collectives.** Each selection cycle begins with a total of g Newborn collectives, each with N_0 total cells of slow-growing S population (light gray dots) and fast-growing F population (dark gray dots). During maturation (over time τ), S and F cells divide at rates r_S and $r_S + \omega$ ($\omega > 0$), respectively, and S mutates to F at rate μ. In the selection stage, the Adult collective with F frequency f closest to the target composition $\hat{f}$ is chosen to reproduce g Newborns for the next cycle. Newborns are sampled from the chosen Adult (yellow star) with N_0 cells per Newborn. The selection cycle is then repeated until the F frequency reaches a steady state, which may or may not be the target composition. To denote a variable x of i-th collective in cycle k at time t ($0 \le t \le \tau$), we use notation $x_{k,t}^{(i)}$ where $x \in \{S,\ F,\ s,\ f\}$. Note that time $t = 0$ is for Newborns and $t = \tau$ is for Adults.

interactions between the two populations, or density-dependent growth. Thus, the exponential growth equation can be used. Second, we consider only two populations (genotypes or species): the fast-growing F population with size F and the slow-growing S population with size S. We do not consider a spectrum of mutants or species, since with more than two populations, an analytical solution becomes very difficult. Finally, the single top-functioning community is chosen to reproduce, which allows us to employ the simplest version of the extreme value theory (see section below for further justification.)

Our goal is to select for collective composition in terms of F frequency $f = F/(S+F)$, or equivalently, S frequency $s = 1 - f$. More precisely, we want collectives such that after maturation time τ, $f(\tau)$ is as close to the target value $\hat{f}$ as possible (Fig. 1). Note that even if the target frequency has been achieved, since F frequency will always increase during maturation, inter-collective selection is required in each cycle to maintain the target frequency.

We will start with a complete model where S mutates to F at a nonzero mutation rate μ. We made this choice because it is more challenging to attain or maintain the target frequency when

the abundance of fast-growing F is further increased via mutations. This scenario is encountered in biotechnology: an engineered pathway will slow down growth, and breaking the pathway (and thus faster growth) is much easier than the other way around. When the mutation rate is set to zero, the same model can be used to capture collectives of two species with different growth rates. We show that intermediate F frequencies or equivalently, intermediate S frequencies, are the hardest targets to achieve. We then show that similar conclusions hold when selecting for a target composition in collectives of more than two populations.

Model structure

A selection cycle (Fig. 1) starts with a total of g Newborn collectives. At the beginning of cycle k ($t = 0$), each Newborn collective has a fixed total cell number $N_0 = S_{k,0}^{(i)} + F_{k,0}^{(i)}$ where $S_{k,t}^{(i)}$ and $F_{k,t}^{(i)}$ denote the numbers of S and F cells in collective i ($1 \leq i \leq g$) at time t ($0 \leq t \leq \tau$) of cycle k. The average F frequency among the g Newborn collectives is $\overline{f}_{k,0}$, such that the initial F cell number in each Newborn is drawn from the binomial distribution $\mathrm{Binom}(N_0, \overline{f}_{k,0})$.

Collectives are allowed to grow for time τ ('Maturation' in Fig. 1). During maturation, S and F grow at rates r_S and $r_S + \omega$ ($\omega > 0$), respectively. If maturation time τ is too small, a matured collective ("Adult") does not have enough cells to reproduce g Newborn collectives with N_0 cells. On the other hand, if maturation time τ is too long, fast-growing F will take over. Hence we set the maturation time $\tau = \ln(g+1)/r_S$, which guarantees sufficient cells to produce g Newborn collectives from a single Adult collective. At the end of a cycle, a single Adult with the highest function (with F frequency f closest to the target frequency $\hat{f}$) is chosen to reproduce g Newborn collectives each with N_0 cells ('Selection' and 'Reproduction' in Fig. 1). Note that even though S and F do not compete for nutrients, they compete for space: because the total number of cells transferred to the next cycle is fixed, an overabundance of one population will reduce the likelihood of the other being propagated.

Collective function is dictated by the Adult's F frequency f. Among all Adult collectives, the selected Adult is the one whose F frequency is closest to the target value, $\hat{f}$. In contrast with findings from an earlier study [23], choosing top 1 is more effective than the less stringent "choosing top 5%". In the earlier study, variation in the collective trait is partly due to nonheritable factors such as random fluctuations in Newborn biomass. In that context, a less stringent selection criterion proved more effective, as it helped retain collectives with favorable genotypes that might have exhibited suboptimal collective traits due to unfavorable nonheritable factors. However, since this study excludes nonheritable variations in collective traits, selecting the top 1 collective is more effective than selecting the top 5% (see Fig. 11 in Supplementary Information).

The selected Adult, with F frequency denoted as f^*, is then used to reproduce g offspring collectives, each with N_0 total cells. The number of F cells in a Newborn follows a binomial distribution $B(N_0, f^*)$. By repeating the selection cycle, we aim to achieve and maintain the target composition

Table I. Nomenclature

Variable	Representing
S	Number of slower-growing (S) cells
F	Number of faster-growing (F) cells
s	Frequency of S cells, $s = S/(S+F)$
f	Frequency of F cells, $f = F/(S+F) = 1-s$
r_S	Growth rate of S
$\omega > 0$	Growth rate advantage of F over S
μ	Mutation rate from S to F
g	Total number of collectives
τ	Maturation time
N	Total cell numbers in a collective, $N = S+F$
N_0	Total number of cells in Newborn, or Newborn size
$\hat{s}, \hat{f}$	Target frequency in s or f
f^*	F frequency of the selected collective in a cycle
f^L,f^H	*L*ow and *H*igh thresholds of inaccessible $\hat{f}$ ($f^L <$ inaccessible $\hat{f} < f^H$)
R_τ	Fold-growth of S cells over time τ, $R_\tau = e^{r_S \tau}$
W_τ	Fold ratio change of F cells over S cells over time τ, $W_\tau = e^{\omega\tau}$

$\hat{f}$.

Overall our model considers mutational stochasticity, as well as demographic stochasticity in terms of stochastic birth and stochastic sampling of a parent collective by offspring collectives. Other types of stochasticity, such as environmental stochasticity and measurement noise, are not considered and require future research.

The success of collective selection is constrained by the target composition, and sometimes also by the initial composition

Since intra-collective selection favors F, we expect that a higher target $\hat{f}$ (a lower target $\hat{s}$) is easier to achieve in the sense that the absolute error d between the target frequency $\hat{f}$ and the selected frequency averaged among independent simulations $\langle f^* \rangle$ is smaller than 0.05 (i.e. $d = |\langle f^* \rangle - \hat{f}| \leq 0.05$).

We fixed N_0, the total population size of a Newborn, to 1000, and obtained selection dynamics for various initial and target F frequencies by implementing stochastic simulations (Sec. I in Supplementary Information). If the target $\hat{f}$ is high (e.g. 0.9, Fig. 2a magenta), selection is successful (computed absolute errors in Supplementary Information Fig. 4): regardless of the initial frequency, f^* of the chosen collective eventually converges to the target $\hat{f}$ and stays around it. In contrast, without collective-level selection (e.g. choosing a random collective to reproduce), F

frequency increases until F reaches fixation (Supplementary information Fig. 3**b**).

In contrast, an intermediate target frequency (e.g. $\hat{f} = 0.5$; Fig. 2**b** green) is never achievable. High initial F frequencies (e.g. 0.6 and 0.95) decline toward the target, but stabilize at the "high-threshold" $f^H (\sim 0.7$, solid orange line segment in Fig. 2**a-c**) above the target. Low initial F frequencies (e.g. 0) increase toward the target, but then overshoot and stabilize at the f^H value.

If the target frequency is low (e.g. $\hat{f} = 0.1$; Fig. 2**c** dark blue), artificial selection succeeds when the initial frequency is below the "lower-threshold" f^L (dotted orange line segment in Fig. 2**a-c**). Initial F frequencies above f^L (e.g. 0.45 and 0.95) converge to f^H instead. Initial F frequencies near f^L display stochastic trajectories, converging to either f^H or the target.

To achieve target $\hat{f}$, inter-collective selection must overcome intra-collective selection. We can visualize the distributions of f over two consecutive cycles (bottom to top, Fig. 2**d**) where f started above target $\hat{f}$. When Newborns matured into Adults, the distribution of f up-shifted due to intra-collective selection. The distribution of f was then down-shifted toward the target due to inter-collective selection. If the magnitude of down-shift exceeded that of up-shift, progress toward the target was made. During reproduction, the distribution of f retained the same mean but became broader due to stochastic sampling by the Newborns from their parent.

In summary, two regions of target frequencies are "accessible" in the sense that a target value in the region can be reached (gold in Fig. 2**e**, **f**): (1) target frequencies above f^H ($\hat{f} > f^H$) or (2) target frequencies below f^L ($\hat{f} < f^L$) and starting at an average frequency below f^L ($\overline{f}_{1,0} < f^L$).

Intra-collective evolution is the fastest at intermediate F frequencies, creating the "waterfall" phenomenon

To understand what gives rise to the two accessible regions, we calculated $\triangle f$, the selection progress in F frequency over two consecutive cycles (Box 1, Eq. (2)). The solution (Fig. 3**a**, green) has the same shape as results from numerically integrating Eq. (1) (Fig. 3**a**, orange) and from stochastic simulations (Fig. 3**a,** blue).

If $\triangle f$ is negative, then inter-collective selection will succeed in countering intra-collective selection and reducing f toward the target. $\triangle f$ is negative if the selected f_k^* is low or high, but not if it is intermediate between f^L and f^H (Fig. 3**a**). This is because f increase during maturation is the most drastic when Newborn f is intermediate (Fig. 3**b**), for intuitive reasons: when Newborn f is low, the increase in f will be minor; when Newborn f is high, the fitness advantage of F over the population average is small and hence the increase is also minor. Thus, when Newborn F frequency is intermediate, inter-collective selection may not be able to counter intra-collective selection (Fig. 3**b** and Fig. 6**a** in Supplementary Information). Not surprisingly, similar conclusions are derived where S and F are slow-growing and fast-growing species which cannot be converted through mutations (Sec. IV and Fig. 8 in Supplementary Information).

Thus, inter-collective selection is akin to a raftman rowing the raft to a target, while intra-

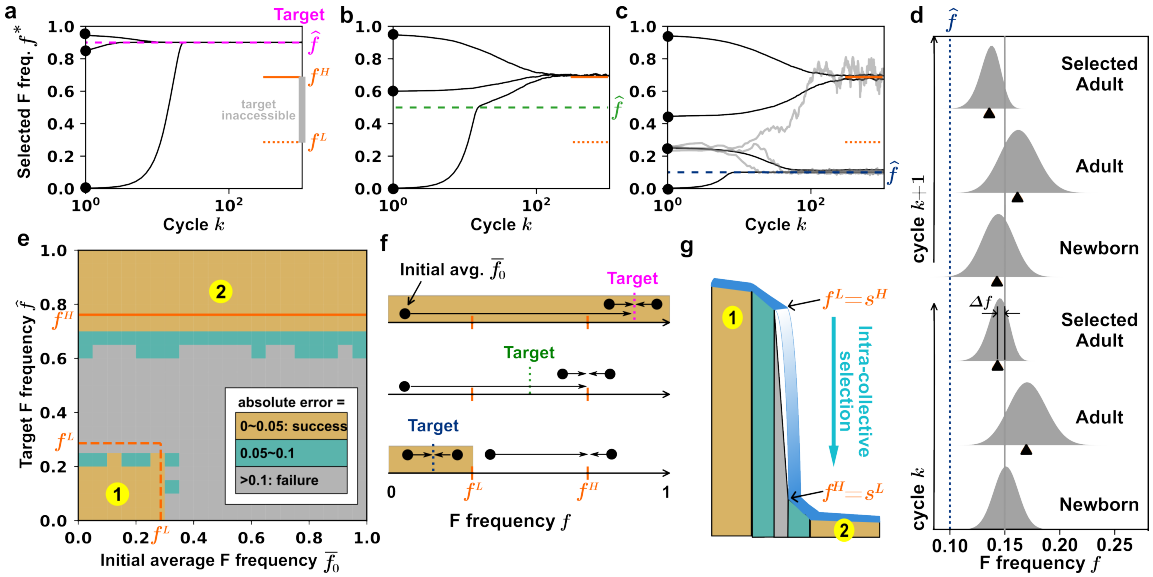

Figure 2. **Initial and target compositions determine the success of artificial selection on collectives. (a-c)** F frequency of the selected Adult collective (f^*) over cycles at different target $\hat{f}$ values (dashed lines). $\hat{f}$ between f^L and f^H (orange dotted and solid line segments) is inaccessible where selection will fail. **a** A high target F frequency (e.g. $\hat{f} = 0.9 > f^H$; magenta dotted line) can be achieved from any initial frequency (black dots). **b** An intermediate target frequency (e.g. $f^L < \hat{f} = 0.5 < f^H$; green dotted line) is never achievable, as all initial conditions converge to near f^H. **c** A low target frequency (e.g. $\hat{f} = 0.1 < f^L$) is achievable, but only from initial frequencies below f^L. For initial frequencies at f^L, stochastic outcomes (grey curves) are observed: while some replicates reached the target frequency, others reached f^H. For parameters, we used S growth rate $r_S = 0.5$, F growth advantage $\omega = 0.03$, mutation rate $\mu = 0.0001$, maturation time $\tau \approx 4.8$, and $N_0 = 1000$. The number of collectives $g = 10$. Each black line is averaged from independent 300 realizations. **d** Inter-collective selection opposes intra-collective selection. We plot probability density distributions of F frequency f during two consecutive cycles when selection is successful. Data correspond to cycles 31 and 32 from the second lowest initial point in **c**. Δf is a selection progress (see Box.). Black triangle: median. **e** Two accessible regions (gold). Either high $\hat{f}$ ($\hat{f} > f^H$; region 2) or low $\hat{f}$ starting from low initial f ($\hat{f} < f^L$ and $\overline{f}_{1,0} < f^L$; region 1) can be achieved. We theoretically predict (by numerically integrating Eq. (1)) f^H (orange solid line) and f^L (orange dotted line), which agree with simulation results (gold regions). **f** Example trajectories from initial compositions (black dots) to the target compositions (dashed lines). The gold areas indicate the region of initial frequencies where the target frequency can be achieved. **g** The tension between intra-collective selection and inter-collective selection creates a "waterfall" phenomenon. See the main text for details.

collective selection is akin to a waterfall. This metaphor is best understood in terms of S frequency $s = 1 - f$. The lower-threshold f^L corresponds to higher-threshold in $s^H = 1 - f^L$, while higher-threshold f^H corresponds to lower-threshold in $s^L = 1 - f^H$. Intra-collective selection is akin to a waterfall, driving the S frequency s from high to low (Fig. 2**g**). Intra-collective selection acts the strongest when s is intermediate ($s^L < s < s^H$), similar to the vertical drop of the fall. Intra-collective selection acts weakly at high ($> s^H$) or low ($< s^L$) s , similar to the gentle sloped upper

and lower pools of the fall (regions 1 and 2 of Fig. 2**e**, **g**). Thus, an intermediate target frequency can be impossible to achieve − a raft starting from the upper pool will be flushed down to s^L (f^H), while a raft starting from the lower pool cannot go beyond s^L (f^H). In contrast, a low target S frequency (in the lower pool) is always achievable. Finally, a high target S frequency (in the upper pool) can only be achieved if starting from the upper pool (as the raft can not jump to the upper pool if starting from below).

Box 1: Changes in the distribution of F frequency f after one cycle

We consider the case where f_k^*, the F frequency of the selected Adult at cycle k, is above the target value ($f_k^* > \hat{f}$). This case is particularly challenging because intra-collective evolution favors fast-growing F and thus will further increase f away from the target. From f_k^*, Newborns of cycle $k+1$ will have f fluctuating around f_k^*, and after they mature, the minimum f is selected ($f_{k+1}^* = \min\left[f_{k+1,\tau}^{(1)}, f_{k+1,\tau}^{(2)}, \cdots, f_{k+1,\tau}^{(g)}\right]$). If the selected composition at cycle $k+1$ can be reduced compared to that of cycle k (i.e. $f_{k+1}^* < f_k^*$), the system can evolve to the lower target value.

To find f_k^* values such that $f_{k+1}^* < f_k^*$, we used the median value of the conditional probability distribution Ψ of f_{k+1}^* (the F frequency of the selected Adult at cycle $k+1$ being f) given the selected f_k^* at cycle k (mathematical details in Sec. II in Supplementary Information). If the median value (Median$[\Psi(f_{k+1}^*|f_k^*)]$) is smaller than f_k^*, then selection will likely be successful since the selected Adult in cycle $k+1$ has more than 50% chance to have a reduced F frequency compared to cycle k.

There are two points where the median values are the same as f_k^* (Fig. 3**a**), which are assigned as lower-threshold (f^L) and higher-threshold (f^H).

Following the extreme value theory, the conditional probability density function $\Psi(f_{k+1}^* = f|f_k^*)$ of f_{k+1}^* (the F frequency of the selected Adult at cycle $k+1$ being f) given the selected f_k^* at cycle k is

$$\Psi(f_{k+1}^* = f|f_k^*) = gP_{f_{k+1,\tau}}(f|f_k^*)\left[1 - \int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*)\right]^{g-1}. \quad (1)$$

Equation (1) can be described as the product between two terms related to probability: (i) $gP_{f_{k+1,\tau}}(f|f_k^*)$ describes the probability density that any one of the g Adult collectives achieves f given f_k^*, and (ii) $\left[1 - \int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*)\right]^{g-1}$ describes the probability that all other $g-1$ collectives achieve frequencies above f and thus not selected.

Since computing the exact formula of Adults' f distribution in cycle $k+1$ is hard, we approximate it as Gaussian with mean $\bar{f}(\tau)$ and variance $\sigma_f^2(\tau)$. The Gaussian approximation on Eq. (1) requires sharp Gaussian distributions of $S(\tau)$ and $F(\tau)$ (i.e. $\bar{S}(\tau) \gg \sigma_S(\tau)$ and $\bar{F}(\tau) \gg \sigma_F(\tau)$). Compared to Gaussian, the exact $S(\tau)$ (negative binomial) distribution and

$F(\tau)$ (Luria-Delbrück) distribution are right-skewed and heavy-tailed. However these problems are alleviated when the initial numbers of S and F cells are not small (on the order of 100). Additionally, the sharpness of distributions could be achieved (see Fig. 1 in Supplementary Information).

To obtain an analytical solution of the change in f over one cycle, we first assume that in a Newborn collective, the number of S cells is distributed as Gaussian with mean $\overline{S}_0 = N_0(1-f_k^*)$ and variance $\sigma_{S,0}^2 = N_0 f_k^*(1-f_k^*)$. Then, the number of F cells, $F_0 = N_0 - S_0$, is distributed as Gaussian with mean $\overline{F}_0 = N_0 f_k^*$ and variance $\sigma_{f,0}^2 = N_0 f_k^*(1-f_k^*)$. From these, we can calculate for Adult collectives the mean and variance of population sizes $F(\tau)$ (i.e. $\bar{F}(\tau)$, $\sigma_F^2(\tau)$) and $S(\tau)$ (i.e. $\bar{S}(\tau)$, $\sigma_S^2(\tau)$) (mathematical details in Sec. I in Supplementary Information). This task is simplified by the exponential growth of S and F: $R_\tau = e^{r_S \tau}$ describes the fold-growth of S over maturation time τ, and since ω is the fitness advantage of F over S, $W_\tau = e^{\omega\tau}$ describes the fold change of F/S over time τ. From R_τ, W_τ, $\frac{\mu}{\omega}$ (mutation rate scaled with the fitness difference), f_k^* (F frequency in the selected collective at cycle k), N_0 (Newborn size), $\frac{\omega}{r_S}$ (relative fitness advantage), we can calculate the mean and variance of F frequency among the Adults of $k+1$ cycle ($\overline{f}(\tau)$; $\sigma_f^2(\tau)$, detailed formula in Eqs. (40) and (41) in Supplementary Information).

Selection progress - the difference between the median value of the conditional probability distribution $\Psi(f_{k+1}^*|f_k^*)$ and the selected frequency of f_k^* (Sec. II in Supplementary Information) - can be expressed as:

$$\triangle f = \text{Median}[\Psi(f_{k+1}^*|f_k^*)] - f_k^* = \overline{f}(\tau) + \left[\Phi^{-1}\left(\frac{\ln 2}{g}\right)\right]\sigma_f(\tau) - f_k^*, \tag{2}$$

where $\Phi^{-1}(\ldots)$ is the inverse cumulative function of standard normal distribution (see main text for an example). We chose median because compared to mean, it is easier to get analytical expression since $\Phi^{-1}(\ldots)$ is known in a closed form. Regardless, using median generated results similar to simulations (Fig. 7 in Supplementary Information). As expected, selection progress $\triangle f$ is governed by both the mean ($\overline{f}(\tau)$) and the variation ($\sigma_f(\tau)$) in f among Adults.

When $\mu = 0$, $\overline{f}(\tau)$ and $\sigma_f(\tau)$ can be simplified to:

$$\overline{f}(\tau) = \frac{f_k^*}{\frac{1-f_k^*}{W_\tau} + f_k^*}, \tag{3}$$

and

$$\sigma_f^2(\tau) = \frac{1}{N_0 W_\tau^2} \frac{f_k^*(1-f_k^*)\left(2 - 2f_k^* + 2f_k^{*2} - \frac{1-f_k^*}{R_\tau W_\tau} - \frac{f_k^*}{R_\tau}\right)}{\left(\frac{1-f_k^*}{W_\tau} + f_k^*\right)^4}. \quad (4)$$

In the limit of small $f_k^* \ll 1$, Equation (3) becomes $\overline{f}(\tau)|_{f_k^* \ll 1} \approx f_k^* W_\tau$ while Equation (4) becomes $\sigma_f^2(\tau)|_{f_k^* \ll 1} = (2 - \frac{1}{R_\tau W_\tau}) f_k^* W_\tau^2 / N_0$. Thus, both Newborn size (N_0) and fold-change in F/S during maturation (W_τ) are important determinants of selection progress.

Manipulating experimental setups to expand the achievable target region

In Eq. (2) (Box 1), selection progress $\triangle f$ depends on the total number of collectives under selection (g). $\triangle f$ also depends on the mean and the standard deviation of Adult F frequency f — $\overline{f}(\tau)$ and $\sigma_f(\tau)$. Eqs. (3) and (4) (Box 1) provide simplified expressions of $\overline{f}(\tau)$ and $\sigma_f(\tau)$ when mutation rate μ has been set to 0. When mutation rate μ is not zero (Eqs. (40) and (41) in Section II of Supplementary Information), selection progress is additionally influenced by $\frac{\mu}{\omega}$ (mutation rate μ scaled with fitness difference ω).

Our goal is to make $\triangle f$ as negative as possible so that any increase in f during collective maturation may be reduced. From Eq. (2) (Box 1), a small $\overline{f}(\tau)$ will facilitate collective-level selection. Additionally, a large $\sigma_f(\tau)$ will also facilitate collective-level selection due to negative $\Phi^{-1}\left(\frac{\ln 2}{g}\right)$. Note that since $\frac{\ln 2}{g}$<0.5 for $g \geq 2$, $\Phi^{-1}\left(\frac{\ln 2}{g}\right)$ — corresponding to the number y such that the probability of a standard normal random variable being less than or equal to y is $\frac{\ln 2}{g}$— is negative.

From Eq. (4) (Box 1), $\sigma_f(\tau)$ will be large if Newborn size N_0 is small. Indeed, as Newborn size N_0 declines, the region of achievable target frequency expands (larger gold area in Fig. 4**a**). If the Newborn size N_0 is sufficiently small (e.g. ≤ 700 in our parameter regime), any target frequency can be reached. An analytical approximation of the maximal Newborn size permissible for all target frequencies is given in section III in Supplementary Information.

From Eqs. (3) and (4) (Box 1), maturation time τ affects $\overline{f}(\tau)$ through $W_\tau = e^{\omega\tau}$ (the fold change in F/S over τ), and affects $\sigma_f^2(\tau)$ additionally through $R_\tau = e^{r_S \tau}$ (fold-growth of S over τ). Longer τ increases $\overline{f}(\tau)$ and is thus detrimental to selection progress. The relationship between $\sigma_f(\tau)$ and τ is not monotonic (Fig. 6**c** in Supplementary Information), meaning that an intermediate value of τ is the best for achieving large $\sigma_f(\tau)$. However, the effect of $\overline{f}(\tau)$ dominates that of $\sigma_f(\tau)$ and therefore, the region of success monotonically reduces with longer maturation time (Fig. 4**c**). Similarly, $\overline{f}(\tau)$ will be small if ω (fitness advantage of F over S) is small. Indeed, as ω become larger, the region of success becomes smaller (Fig. 9 in Supplementary Information).

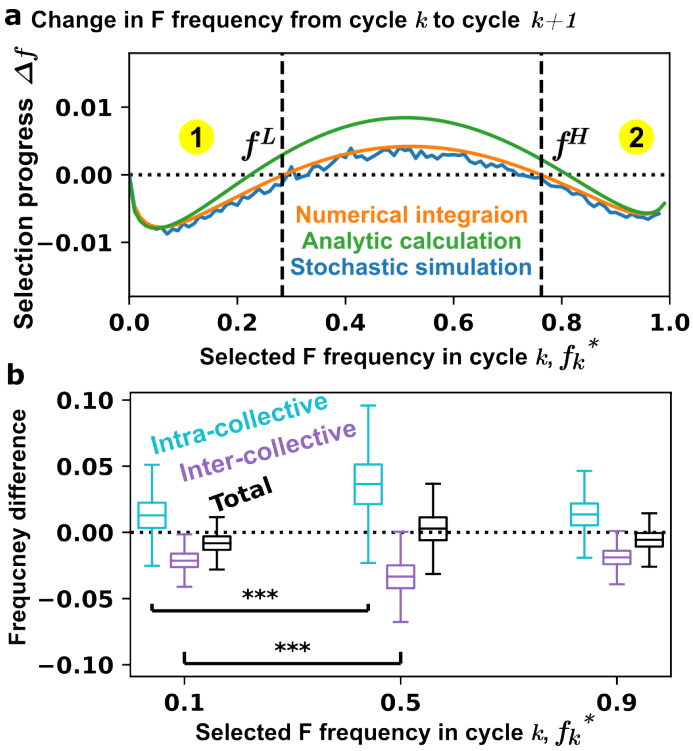

Figure 3. **Intra-collective selection and inter-collective selection jointly set the boundaries for selection success.** **a** The change in F frequency over one cycle. When f_k^* is sufficiently low or high, inter-collective selection can lower the F frequency to below f_k^* ($\Delta f < 0$). The points where $\Delta f = 0$ (in the orange line) are denoted as f^L and f^H, corresponding to the boundaries in Fig. 2. **b** The distributions of frequency differences obtained by 1000 numerical simulations. The cyan, purple, and black box plots respectively indicate the changes in F frequency after intra-collective selection (the mean frequency among the 100 Adults minus the mean frequency among the 100 Newborns during maturation), after inter-collective selection (the frequency of the 1 selected Adult minus the mean frequency among the 100 Adults), and over one selection cycle (the frequency of the selected Adult of one cycle minus that of the previous cycle). The box ranges from 25% to 75% of the distribution, and the median is indicated by a line across the box. The upper and lower whiskers indicate maximum and minimum values of the distribution. ***: $P < 0.001$ in an unpaired t-test.

g, the number of collectives under selection, also affects selection outcomes. As g increases, the value of $\Phi^{-1}\left(\frac{\ln 2}{g}\right)$ becomes more negative, and so does $\triangle f$ - meaning collective-level selection will be more effective. Intuitively, with more collectives, the chance of finding a f closer to the target is more likely. Thus, a larger number of collectives broadens the region of success (Fig. 4**b**). However, the effect of g is not dramatic. To see why, we note that the only place that g appears is Eq. (2) in $\Phi^{-1}\left(\frac{1}{g}\right)$. When g becomes large, $\Phi^{-1}\left(\frac{1}{g}\right)$ is asymptotically expressed as $\Phi^{-1}\left(\frac{1}{g}\right) \approx -\sqrt{2\ln g - \ln[\ln g] + \cdots}$ (Section II in Supplementary Information)[35], and thus does not change dramatically as g varies.

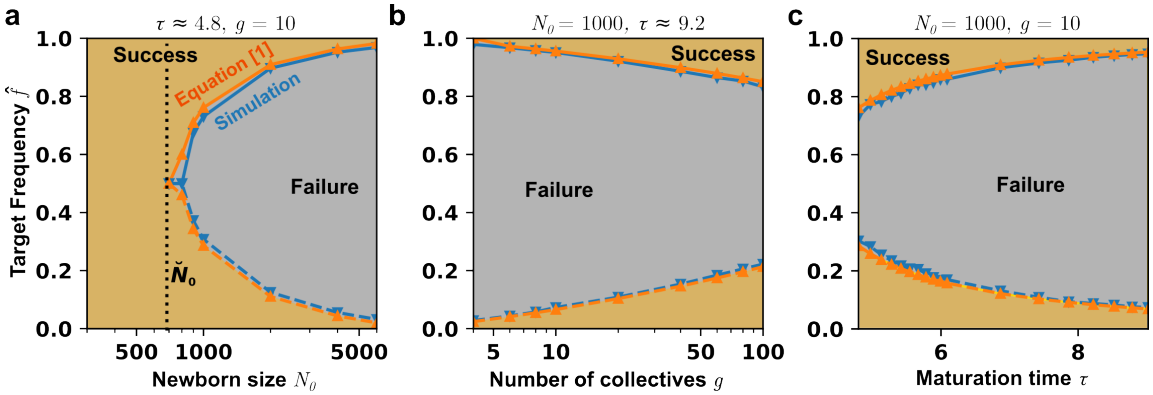

Figure 4. **Expanding the success region for artificial collective selection.** **a** Reducing the population size in Newborn N_0 expands the region of success. In gold area, the probability that f^*_{k+1} becomes smaller than f^*_k in a cycle is more than 50%. We used $g = 10$ and $\tau \approx 4.8$. Figures 2-3 correspond to $N_0 = 1000$**.** Black dotted line indicates the critical Newborn size $\check{N}_0$, below which all target frequencies can be achieved. **b** Increasing the total number of collectives g also expands the region of success, although only slightly. We used a fixed Newborn size $N_0 = 1000$. The maturation time τ ($\tau = \log(100)/r_S \approx 9.2$) is set to be long enough so that an Adult can generate at least 100 Newborns. **c** Increasing the maturation time τ shrinks the region of success. We used a fixed Newborn size $N_0 = 1000$ and number of collectives $g = 10$.

The waterfall phenomenon in a higher dimension

To examine the waterfall effect in a higher dimension, we investigate a three-population system where a faster-growing population (FF) grows faster than the fast-growing population (F) which grows faster than the slow-growing population (S) (Fig. 5**a** and Sec. VIII in Supplementary Information). In the three-population case, the evolutionary trajectory travels in a two-dimensional plane. A target population composition can be achieved if inter-collective selection can sufficiently reduce the frequencies of F as well as FF (accessible region, gold in Fig. 5**b**).

From numerical simulations, we identified two accessible regions: a small region near FF and a band region spanning from S to F (gold in Fig. 5**b** i). Intuitively, the rate at which FF grows faster than S+F is greater than the rate at which F grows faster than S (see section VIII in Supplementary Information). Thus, the problem can initially be reduced to a two-population problem (i.e. FF versus F+S; Fig. 5**c** left), and then expanded to a three-population problem (Fig. 5**c** right).

Similar to the two-population case, targets in the inaccessible region are never achievable (Fig. 5**b** ii), while those in the FF region are always achievable (Fig. 5**b** i). Strikingly, a target composition in an accessible region may not be achievable even when the initial composition is within the same region: once the composition escapes the accessible region, the trajectory cannot return back to the accessible region (Fig. 5**b**iii, the leftmost initial condition). However, if the initial position is closer to the target in the accessible region, the target becomes achievable (Fig. 5**b** iii, initial condition near the bottom). Note that here, selection outcome is path-dependent in the sense of being

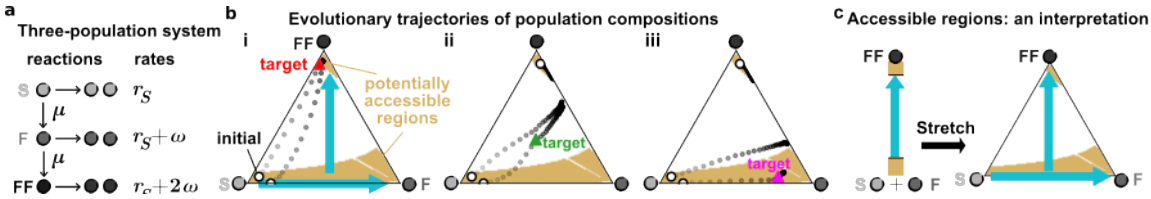

Figure 5. **In higher dimensions, the success of artificial selection requires the entire evolutionary trajectory remaining in the accessible region.** **a** During collective maturation, a slow-growing population (S) (with growth rate r_S; light gray) can mutate to a fast-growing population (F) (with growth rate $r_S + \omega$; medium gray), which can mutate further into a faster-growing population (FF) (with growth rate $r_S + 2\omega$; dark gray). Here, the rates of both mutational steps are μ, and $\omega > 0$. **b** Evolutionary trajectories from various initial compositions (open circles) to various targets. Intra-collective evolution favors FF over F (vertical blue arrow) over S (horizontal blue arrow). The accessible regions are marked gold (see Sec. I in Supplementary Information). We obtain final compositions starting from several initial compositions while aiming for different target compositions in i, ii, and iii. The evolutionary trajectories are shown in dots with color gradients from the initial to final time. (i) A target composition with a high FF frequency is always achievable. (ii) A target composition with intermediate FF frequency is never achievable. (iii) A target composition with low FF frequency is achievable only if starting from an appropriate initial composition such that the entire trajectory never meanders away from the accessible region. The figures are drawn using mpltern package [36]. **c** The accessible region in the three-population problem is interpreted as an extension of the two-population problem. First the accessible region between FF and S+F is given, and then the S+F region is stretched into S and F.

sensitive to initial conditions. This phenomenon is distinct from hysteresis where path-dependence results from whether a tuning parameter is increased or decreased.

In conclusion, we have investigated the evolutionary trajectories of population compositions in collectives under selection, which are governed by intra-collective selection (which favors fast-growing populations) and inter-collective selection (which, in our case, strives to counter fast-growing populations). Intra-collective selection has the strongest effect at intermediate frequencies of faster-growing populations, potentially creating an inaccessible region analogous to the vertical drop of a waterfall. High and low target frequencies are both accessible, analogous to the lower- and the upper-pools of a waterfall, respectively. A less challenging target (high $\hat{f}$; low $\hat{s}$) is achievable from any initial position. In contrast, a more challenging target (low $\hat{f}$; high $\hat{s}$) is only achievable if the entire trajectory is contained within the region, similar to a raft striving to reach a point in the upper-pool must start at and remain in the upper pool. Our work suggests that the strength of intra-collective selection is not constant, and that strategically choosing an appropriate starting point can be essential for successful collective selection.

METHODS

Stochastic simulations

A selection cycle is composed of three steps: maturation, selection, and reproduction. At the beginning of the cycle k, a collective i has $S_{k,0}^{(i)}$ slow-growing cells and $F_{k,0}^{(i)}$ fast-growing cells. At the first cycle, mean F frequency of collectives are set to be $\bar{f}_{1,0}$. $F_{1,0}^{(i)}$ is sampled from the binomial distribution with mean $N_0\overline{f}_{1,0}$. Then, $S_{1,0}^{(i)}(= N_0 - F_{1,0}^{(i)})$ S cells are in the collective i. In the maturation step, we calculate $S_{k,\tau}^{(i)}$ and $F_{k,\tau}^{(i)}$ by using stochastic simulation. We can simulate the division and mutation of each individual cell stochastically by using tau-leaping algorithm [37, 38](see Fig. 3**a** and **b** in Supplementary Information). However, individual-based simulations require a long computing time. Instead, we randomly sample $S_{k,\tau}^{(i)}$ and $F_{k,\tau}^{(i)}$ from the joint probability density distribution $P(S_{k,\tau}^{(i)}, F_{k,\tau}^{(i)})$. To obtain $P(S_{k,\tau}^{(i)}, F_{k,\tau}^{(i)})$, we solve the master equation which describes a time evolution of probability distribution $P(S_{k,t}^{(i)}, F_{k,t}^{(i)})$ under the random processes (see Sec. I in Supplementary Information). We assumed that $S_{k,\tau}^{(i)}$ and $F_{k,\tau}^{(i)}$ are independent (as S and F populations grow independently without ecological interactions), and thus $P(S_{k,\tau}^{(i)}, F_{k,\tau}^{(i)})$ is product of two probability density functions $P(S_{k,\tau}^{(i)})$ and $P(F_{k,\tau}^{(i)})$. Each distribution follows a Gaussian distribution,with the mean and variance numerically obtained from ordinary differential equations derived from the master equation (see Sec. I in Supplementary Information). We choose the collective with the closest frequency to the target $\hat{f}$ to generates g Newborns. The number of F cells is sampled from the binomial distribution with the mean of $N_0 f_k^*$. We start a new cycle with those Newborn collectives. Then, the number of S cells in a collective i is $S_{k+1,0}^{(i)} = N_0 - F_{k+1,0}^{(i)}$.

Analytical approach to the conditional probability

The conditional probability distribution $\Psi(f_{k+1}^*|f_k^*)$of observing f_{k+1}^* at a given f_k^* is calculated by the following procedure. Given the selected collective in cycle k with f_k^*, the collective-level reproduction proceeds by sampling g Newborn collectives with N_0 cells in cycle $k+1$. Each Newborn collective contains certain F numbers $F_{k+1,0}^{(1)}, \cdots, F_{k+1,0}^{(g)}$ at the beginning of the cycle $k+1$, which can be mapped into $f_{k+1,0}^{(1)}, \cdots, f_{k+1,0}^{(g)}$ with the constraint of N_0 cells. If the number of cells in the selected collective is large enough, the joint conditional distribution function $P\left(f_{k+1,0}^{(1)}, \cdots, f_{k+1,0}^{(g)}|f_k^*\right)$ is well described by the product of g independent and identical Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. So we consider the frequencies of g Newborn collectives as g identical copies of the Gaussian random variable $f_{k+1,0}$. The mean and variance of $f_{k+1,0}$ are given by $m = f_k^*$ and $\sigma^2 = f_k^*(1 - f_k^*)/N_0$. Then, the conditional probability distribution function of $f_{k+1,0}$ being ζ is

given by

$$P_{f_{k+1,0}}(\zeta|f_k^*) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\zeta - m)^2}{2\sigma^2}\right). \tag{5}$$

After the reproduction step, the Newborn collectives grow for time τ. The frequency is changed from the given frequency ζ to f by division and mutation processes. We assume that the frequency f of an Adult is also approximated by a Gaussian random variable $\mathcal{N}(\overline{f}(\tau), \sigma_f^2(\tau))$. The mean $\overline{f}(\tau)$ and variance $\sigma_f^2(\tau)$ are calculated by using means and variances of S and F (see Sec. II in Supplementary Information.) Since $\overline{f}(\tau)$ and $\sigma_f^2(\tau)$ also depend on ζ, the conditional probability distribution function of $f_{k+1,\tau}$ being f is given by

$$P_{f_{k+1,\tau}}(f|\zeta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(f - \overline{f}(\tau))^2}{2\sigma_f^2(\tau)}\right). \tag{6}$$

The conditional probability distribution of an Adult collective in cycle $k+1$ ($f_{k+1,\tau}$) to have frequency f at a given f_k^* is calculated by multiplying two Gaussian distribution functions and integrating overall ζ values, which is given by

$$P_{f_{k+1,\tau}}(f|f_k^*) = \int_0^1 d\zeta \; P_{f_{k+1,\tau}}(f|\zeta) \; P_{f_{k+1,0}}(\zeta|f_k^*). \tag{7}$$

Since we select the minimum frequency $f_{k+1}^{\min}$ among g identical copies of $f_{k+1,\tau}$, the conditional probability distribution function of $f_{k+1}^{\min}$ follows a minimum value distribution, which is given in Eq. (1). Here, for the case of $\hat{f} < f_k^*$, the selected frequency f_{k+1}^* is the minimum frequency $f_{k+1}^{\min}$. So we have $\Psi(f_{k+1}^*|f_k^*)$ by replacing $f_{k+1}^{\min}$ with f_{k+1}^*.

We assume that the conditional probability distribution in Eq. (7) follows a normal distribution, whose mean and variances are describe by Eq. (40) and Eq. (41). Then, the extreme value theory [39] estimates the median of the selected Adult by

$$\text{Median}(f_{k+1}^*) = \overline{f}(\tau) + \left[\Phi^{-1}(\frac{\ln 2}{g})\right] \sigma_f(\tau). \tag{8}$$

The selection progress Δf in Eq. (2) is obtained by subtracting f_k^* from Eq. (8).

DATA AND SOURCE CODE AVAILABILITY

Data and source code of stochastic simulations are available in https://github.com/schwarzg/artificial_selection_collective_composition

---

[1] Woo, S., Song, I. & Cha, H. J. Fast and facile biodegradation of polystyrene by the gut microbial flora of *Plesiophthalmus davidis* larvae. *Applied and Environmental Microbiology* **86**, e01361–20 (2020).

[2] Sun, J., Prabhu, A., Aroney, S. T. & Rinke, C. Corrigendum: Insights into plastic biodegradation: community composition and functional capabilities of the superworm (*Zophobas morio*) microbiome in styrofoam feeding trials. *Microbial Genomics* **8** (2022).

[3] Bober, J. R., Beisel, C. L. & Nair, N. U. Synthetic biology approaches to engineer probiotics and members of the human microbiota for biomedical applications. *Annual Review of Biomedical Engineering* **20**, 277–300 (2018).

[4] Wang, E.-X., Ding, M.-Z., Ma, Q., Dong, X.-T. & Yuan, Y.-J. Reorganization of a synthetic microbial consortium for one-step vitamin c fermentation. *Microbial Cell Factories* **15**, 21 (2016).

[5] Goodnight, C. J. Experimental studies of community evolution i: The response to selection at the community level. *Evolution* **44**, 1614–1624 (1990).

[6] Goodnight, C. J. Experimental studies of community evolution ii: The ecological basis of the response to community selection. *Evolution* **44**, 1625–1636 (1990).

[7] Swenson, W., Wilson, D. S. & Elias, R. Artificial ecosystem selection. *Proceedings of the National Academy of Sciences* **97**, 9110–9114 (2000).

[8] Swenson, W., Arendt, J. & Wilson, D. S. Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environmental Microbiology* **2**, 564–571 (2000).

[9] Blouin, M., Karimi, B., Mathieu, J. & Lerch, T. Z. Levels and limits in artificial selection of communities. *Ecology Letters* **18**, 1040–1048 (2015).

[10] Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E. & Kao-Kniffin, J. Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal* **9**, 980–989 (2015).

[11] Panke-Buisse, K., Lee, S. & Kao-Kniffin, J. Cultivated sub-populations of soil microbiomes retain early flowering plant trait. *Microbial Ecology* **73**, 394–403 (2017).

[12] Jochum, M. D., McWilliams, K. L., Pierson, E. A. & Jo, Y.-K. Host-mediated microbiome engineering (hmme) of drought tolerance in the wheat rhizosphere. *PLoS One* **14**, e0225933 (2019).

[13] Wright, R. J., Gibson, M. I. & Christie-Oleza, J. A. Understanding microbial community dynamics to improve optimal microbiome selection. *Microbiome* **7**, 1–14 (2019).

[14] Raynaud, T., Devers, M., Spor, A. & Blouin, M. Effect of the reproduction method in an artificial selection experiment at the community level. *Frontiers in Ecology and Evolution* **7**, 416 (2019).

[15] Jigyasa Arora, A. S. M., Margaret A. Mars Brisbin. Effects of microbial evolution dominate those of experimental host-mediated indirect selection. *PeerJ* **8**, e9350 (2020).

[16] Chang, C.-Y., Osborne, M. L., Bajic, D. & Sanchez, A. Artificially selecting bacterial communities using propagule strategies. *Evolution* **74**, 2392–2403 (2020).

[17] Mueller, U. G. *et al.* Artificial selection on microbiomes to breed microbiomes that confer salt tolerance to plants. *mSystems* **6**, e01125–21 (2021).

[18] Jacquiod, S. *et al.* Artificial selection of stable rhizosphere microbiota leads to heritable plant phenotype changes. *Ecology Letters* **25**, 189–201 (2022).

[19] Raynaud, T., Devers-Lamrani, M., Spor, A. & Blouin, M. Community diversity determines the evolution of synthetic bacterial communities under artificial selection. *Evolution* **76**, 1883–1895 (2022).

[20] Penn, A. "Modelling artificial ecosystem selection: A preliminary investigation". In Banzhaf, W., Ziegler, J., Christaller, T., Dittrich, P. & Kim, J. T. (eds.) *Advances in Artificial Life*, 659–666 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003).

[21] Penn, A. & Harvey, I. "The role of non-genetic change in the heritability, variation, and response to selection of artificially selected ecosystems". In Pollack, J., Bedau, M. A., Husbands, P., Watson, R. A. & Ikegami, T. (eds.) *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life*, vol. 9, 352 (MIT Press, 2004).

[22] Williams, H. T. P. & Lenton, T. M. Artificial selection of simulated microbial ecosystems. *Proceedings of the National Academy of Sciences* **104**, 8918 (2007).

[23] Xie, L., Yuan, A. E. & Shou, W. Simulations reveal challenges to artificial community selection and possible strategies for success. *PLoS Biology* **17**, e3000295 (2019).

[24] Doulcier, G., Lambert, A., De Monte, S. & Rainey, P. B. Eco-evolutionary dynamics of nested darwinian populations and the emergence of community-level heredity. *eLife* **9**, e53433 (2020).

[25] Xie, L. & Shou, W. Steering ecological-evolutionary dynamics to improve artificial selection of microbial communities. *Nature Communications* **12**, 6799 (2021).

[26] Chang, C.-Y. *et al.* Engineering complex communities by directed evolution. *Nature Ecology & Evolution* **5**, 1011 (2021).

[27] Fraboul, J., Biroli, G. & De Monte, S. Artificial selection of communities drives the emergence of structured interactions. *Journal of Theoretical Biology* **571**, 111557 (2023).

[28] Lalejini, A., Dolson, E., Vostinar, A. E. & Zaman, L. Artificial selection methods from evolutionary computing show promise for directed evolution of microbes. *eLife* **11**, e79665 (2022).

[29] Zaccaria, M., Sandlin, N., Soen, Y. & Momeni, B. Partner-assisted artificial selection of a secondary function for efficient bioremediation. *iScience* **26**, 107632 (2023).

[30] Vessman, B., Guridi-Fernández, P., Arias-Sánchez, F. I. & Mitri, S. Novel artificial selection method improves function of simulated microbial communities. *bioRxiv [preprint]* (2023). Https://doi.org/10.1101/2023.01.08.523165 (accessed 18 December 2023).

[31] Arias-Sánchez, F. I., Vessman, B., Haym, A., Alberti, G. & Mitri, S. Artificial selection improves pollutant degradation by bacterial communities. *Nature Communications* **15**, 7836 (2024).

[32] Xie, L., Yuan, A. E. & Shou, W. A quantitative genetics framework for understanding the selection response of microbial communities. *bioRxiv [preprint]* (2023). Https://doi.org/10.1101/2023.10.24.563725 (accessed 28 December 2023).

[33] Thomas, J. L., Rowland-Chandler, J. & Shou, W. Artificial selection of microbial communities: what have we learnt and how can we improve? *Current Opinion in Microbiology* **77**, 102400 (2024).

[34] Rainey, P. B. Major evolutionary transitions in individuality between humans and ai. *Philosophical Transactions of the Royal Society B* **378**, 20210408 (2023).

[35] Philip, J. The function inverfc theta. *Australian Journal of Physics* **13**, 13 (1960).

[36] Ikeda, Y., Grabowski, B. & Körmann, F. mpltern 0.3.0: ternary plots as projections of Matplotlib (2019). URL https://doi.org/10.5281/zenodo.3528355.

[37] Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics* **115**, 1716–1733 (2001).

[38] Cao, Y., Gillespie, D. T. & Petzold, L. R. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics* **124**, 044109 (2006).

[39] Gumbel, E. J. *Statistics of extremes* (Columbia university press, 1958).

ACKNOWLEDGMENTS

J. Lee and H.J. Park were supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT), Grant No. RS-2023-00214071 and RS-2024-00460958 and by an appointment to the JRG Program at the APCTP through the Science and Technology Promotion Fund and Lottery Fund of the Korean Government. W. Shou was supported by an Academy of Medical Sciences Professorship and a Royal Society Wolfson Fellowship. This was also supported by the Korean Local Governments–Gyeongsangbuk-do Province and Pohang City and INHA UNIVERSITY Research Grant. We thank Su-Chan Park, Li Xie, Alex Yuan, and Botond Major for constructive comments and discussions.

AUTHOR INFORMATION

Juhee Lee

Present address: Integrated Science Lab, Umea university, Sweden

Authors and affiliations

Department of Physics, Inha University, Incheon 22212, Republic of Korea
Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea

Juhee Lee & Hye Jin Park

Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom

Wenying Shou

Author contributions

J. Lee., W. Shou, H. J. Park designed research; J. Lee performed research; J. Lee, W. Shou, H. J. Park wrote the paper.

Corresponding author

Wenying shou (wenying.shou@gmail.com) and Hye Jin park (hyejin.park@inha.ac.kr).

COMPETING INTERESTS

No competing intersests.

Supplementary Information

CONTENTS

I. Stochastic simulation of the selection cycle 2

Maturation 2

Selection 6

Reproduction 7

Simulation Result 8

II. Conditional probability distribution of the selected collective frequency f^* and selection progress Δf 8

III. Critical newborn size $\breve{N}_0$ to allow all target frequencies 13

IV. Selection without mutation $\mu = 0$ 14

V. Stronger or weaker advantages ω 15

VI. Deleterious mutation $\omega < 0$ 15

VII. Selecting more than one collective 16

VIII. Extension to three-population system 17

IX. Derivation of equations 23

Equations (10) and (11) 24

Equations (16) and (17) 24

Equation (18) and (20) 26

Equation (22) and (23) 27

Equation (24) 28

Equation (25) 31

Equations (32)-(34) 31

References 32

In this supplementary information (SI), we present the mathematical details of all calculations in the main text, as well as additional figures. The SI is organized as follows: In section I, we provide the mathematical details of stochastic simulation of the selection cycle. In section II, we show the mathematical process to obtain the analytical expression of the selection progress in Eq. (2) in the main text. In section III, we provide the analytical approximation of the critical Newborn size. In section IV, we discuss a limiting case when mutation rate is zero. Section V provides the boundary of the region of success with different selective advantages. In section VI, we discuss specific scenarios under which mutation converts F to S. In section VII, we present an example in which more than one collective is selected for reproduction. In section VIII, we provide the mathematical details of the selection cycle for a three-population system. Finally, in section IX, we provide detailed derivations of equations from section I to section II.

I. STOCHASTIC SIMULATION OF THE SELECTION CYCLE

In the main text, we design a simple model of artificial selection on collectives. The selection cycle starts with g "Newborn" collectives which consist of two populations - slow-growing population (S) and fast-growing population (F). S mutate to F at a rate μ. The Newborns mature for a fixed time τ. The matured collective ("Adult") with the highest function (with F frequency f closest to the target $\hat{f}$) is chosen to reproduce g Newborn collectives each with N_0 cells.

In our selection cycle, variation among collectives mainly resulted from demographic noises during cell birth, cell mutation, and collective reproduction. In this section, we provide details of the simulation.

Maturation

Here, we calculate the cell numbers during maturation. Each collective i ($i = 1, ..., g$) has $S^{(i)}_{k,\mathrm{t}}$ S cells and $F^{(i)}_{k,t}$ F cells where k is the cycle number and t indicates time ($0 \leq t \leq \tau$). At the beginning of cycle k ($t = 0$), each Newborn collective has a total of $N_0 = S^{(i)}_{k,0} + F^{(i)}_{k,0}$ cells. The collectives are allowed to "mature" for $t = \tau$ during which S and F grow at rates r_S and $r_S + \omega$ ($\omega > 0$), respectively. In this subsection, we ignore the cycle number index k and the collective index (i) for convenience. That is, we denote $S^{(i)}_{k,t}$ and $F^{(i)}_{k,t}$ as $S(t)$ and $F(t)$, respectively.

We describe cell divisions of S and F cells and mutation from S to F with following chemical reaction rules:

$$\mathrm{S} \xrightarrow{r_S} \mathrm{S} + \mathrm{S}, \tag{1}$$

$$\mathrm{F} \xrightarrow[\omega>0]{r_S+\omega} \mathrm{F} + \mathrm{F}, \tag{2}$$

$$\mathrm{S} \xrightarrow{\mu} \mathrm{F}, \tag{3}$$

One can run an individual-based simulation by counting the number of events occurring during collective maturation via the tau-leaping algorithm [1, 2] to generate a sample trajectory of $S(t)$ and $F(t)$ for each collective. However, the individual-based simulation requires long computing times due to a large number of random events to be counted. Hence we used a 'sampling method' by sampling the numbers of S and F cells in collectives from a joint probability density distribution

(jpdf) $P(S,F,t)$ which denotes the probability density to have S number of S cells and F number of F cells at time t in the cycle. To do so, we require an analytical expression of $P(S,F,t)$.

First, we assume that the chemical reactions in Eqs. (1)-(3) occur independently, and never occur simultaneously within a short time interval $[t, t+dt)$. Then, the differential of $P(S,F,t)$ with respect to time is given by

$$\begin{aligned}\frac{dP(S,F,t)}{dt} &= r_S(S-1)P(S-1,F,t) \\ &+(r_S+\omega)(F-1)P(S,F-1,t) \\ &+\mu(S+1)P(S+1,F-1,t) \\ &-(r_S S+(r_S+\omega)F+\mu S)P(S,F,t).\end{aligned} \tag{4}$$

This master equation describes a probability density 'flux' at the state (S,F). The first term describes the scenario where a single birth event of a S cell happens during time interval $[t, t+dt)$, which changes the collective's composition from $(S-1,F)$ to (S,F). Similarly, the second term comes from a birth event of a F cell. The third term indicates the mutation event from $(S+1,F-1)$ to (S,F). The last term corresponds to the outflow of probability density by birth and mutation processes, which describes the changes from (S,F) to any other states.

Calculating the exact form of $P(S,F,t)$ is not simple. Instead, we assume that the mutation rate is much smaller than the growth rates, and hence the correlation between S and F is sufficiently small. Additionally, S and F do not interact ecologically. Then, we can express $P(S,F,t)$ as a product of two probability density functions (pdf) of $S(t)$ and $F(t)$, $P(S,F,t) = P(S,t)P(F,t)$. We assume that each pdf of S and F can be approximated as Gaussian ($\mathcal{N}$), which is supported by the Central Limit Theorem and Fig. 1. In more detail, the cell numbers S and F are mainly determined by growth (Eqs. (1) and (2)), and also mutations (Eq. (3)). Even though the number of events would be different among different realizations, the mean numbers of events will follow Gaussian distributions. So, we can simply assume that the distributions of cell numbers also follow Gaussian distributions. This assumption requires that the distributions have insignificant skewness and no heavy tails, which we will numerically check afterwards. The pdfs of $S(t)$ and $F(t)$ are given by

$$P(S,t) = \frac{1}{\sqrt{2\pi\sigma_S^2(t)}} e^{-\frac{(S(t)-\overline{S}(t))^2}{2\sigma_S^2(t)}} \tag{5}$$

and

$$P(F,t) = \frac{1}{\sqrt{2\pi\sigma_F^2(t)}} e^{-\frac{(F(t)-\overline{F}(t))^2}{2\sigma_F^2(t)}}. \tag{6}$$

That is, $P(S,F,t)$ is written as

$$P(S,F,t) = \frac{1}{\sqrt{2\pi\sigma_S^2(t)}} \frac{1}{\sqrt{2\pi\sigma_F^2(t)}} e^{-\frac{(S(t)-\overline{S}(t))^2}{2\sigma_S^2(t)} - \frac{(F(t)-\overline{F}(t))^2}{2\sigma_F^2(t)}} \tag{7}$$

Now we need means ($\overline{S}(t)$ and $\overline{F}(t)$) and variances ($\sigma_S^2(t)$ and $\sigma_F^2(t)$) of S and F cell numbers to express the distribution analytically.

The means are defined by $\overline{S}(t) = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S\, P(S,F,t)$ and $\overline{F}(t) = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} F\, P(S,F,t)$. The differential equations for means are obtained by applying the definition to the master equation

in Eq. (4), as

$$\frac{d\overline{S}(t)}{dt} = r_S\overline{S}(t) - \mu\overline{S}(t), \tag{8}$$

$$\frac{d\overline{F}(t)}{dt} = (r_S + \omega)\overline{F}(t) + \mu\overline{S}(t). \tag{9}$$

We assume that the mutation rate μ is much smaller than r_S and ω. By solving Eq. (8) and Eq. (9), the means $\overline{S}(t)$ and $\overline{F}(t)$ are given by

$$\overline{S}(t) = \overline{S}_o e^{(r_S-\mu)t} \approx \overline{S}_o e^{r_S t}, \tag{10}$$

$$\overline{F}(t) = \overline{F}_o e^{(r_S+\omega)t} + \frac{\mu\overline{S}_o}{\omega+\mu}\left(e^{(r_S+\omega)t} - e^{(r_S-\mu)t}\right) \approx \overline{F}_o e^{(r_S+\omega)t} + \frac{\mu\overline{S}_o}{\omega}\left(e^{(r_S+\omega)t} - e^{r_S t}\right), \tag{11}$$

where $\overline{S}_o \equiv \overline{S}(0)$ and $\overline{F}_o \equiv \overline{F}(0)$ are the mean numbers of S and F cells at the beginning of cycle, $t = 0$. Note that the second term of Eq. (11) is consistent with previous studies [3]. Now we introduce factors $R_t = e^{r_S t}$ and $W_t = e^{\omega t}$ in Eqs. (10) and (11) in order to simplify the formula. R_t is the multiplying factor by which the S cell number increases after time t. W_t is the fold change in F/S. Then, we can rewrite

$$\overline{S}(t) = \overline{S}_o R_t, \tag{12}$$

$$\overline{F}(t) = \overline{F}_o R_t W_t + \frac{\mu\overline{S}_o}{\omega} R_t(W_t - 1) \tag{13}$$

We define the second momenta of S and F as

$$\overline{S^2} = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S^2 P(S,F,t), \tag{14}$$

$$\overline{F^2} = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} F^2 P(S,F,t). \tag{15}$$

Then, the corresponding differential equations are given by

$$\frac{d\overline{S^2}}{dt} = 2(r_S - \mu)\overline{S^2} + (r_S + \mu)\overline{S}, \tag{16}$$

$$\frac{d\overline{F^2}}{dt} = 2(r_S + \omega)\overline{F^2} + (r_S + \omega)\overline{F} + 2\mu\overline{SF} + \mu\overline{S}. \tag{17}$$

The solution of Eq. (16) is

$$\overline{S^2}(t) = \frac{r_S + \mu}{\mu - r_S}\overline{S}_o\left[e^{(\mu - r_S)t} - 1\right]e^{2(r_S-\mu)t} + \overline{S^2}_o e^{2(r_S-\mu)t} \tag{18}$$

$$\approx \overline{S_o^2} e^{2r_S t} + \overline{S}_o e^{r_S t}(e^{r_S t} - 1), \tag{19}$$

where $\overline{S_o^2} \equiv \overline{S^2}(0)$ is the second moment of initial values. Thus, the variance $\sigma_S^2(t) = \overline{S^2}(t) - [\overline{S}(t)]^2$ is

$$\sigma_S^2(t) = \sigma_{S,0}^2 e^{2r_S t} + \overline{S}_o e^{r_S t}(e^{r_S t} - 1) \tag{20}$$

$$= \sigma_{S,0}^2 R_t^2 + \overline{S}_o R_t(R_t - 1) \tag{21}$$

where $\sigma_{S,0}^2 \equiv \sigma_S^2(0)$ is a variance of S cell numbers at $t = 0$. In Eq. (17), we require $\overline{SF}(t) = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} SFP(S,F,t)$ to calculate $\overline{F^2}(t)$. Eq. (4) provides a differential equation for $\overline{SF}(t)$ as

$$\frac{d\overline{SF}}{dt} = (2r_S + \omega - \mu)\overline{SF} + \mu\left(\overline{S^2} - \overline{S}\right). \tag{22}$$

The solution of Eq. (22) is given by

$$\begin{aligned}\overline{SF}(t) = \overline{SF}_o e^{(2r_S+\omega)t} \\ &+\frac{\mu}{\omega}\left(\overline{S^2}_0 + \overline{S}_0\right)\left(e^{(2r_S+\omega)t} - e^{2r_S t}\right) \\ &-\frac{2\mu\overline{S}_0}{r_S+\omega}\left(e^{(2r_S+\omega)t} - e^{r_S t}\right).\end{aligned} \tag{23}$$

By using Eq. (23), the solution of Eq. (17) is given by

$$\begin{aligned}\overline{F^2}(t) = \overline{F^2}_0 e^{2(r_S+\omega)t} + \overline{F}_0 e^{(r_S+\omega)t}\left(e^{(r_S+\omega)t} - 1\right) \\ &+\frac{\mu\overline{S}_o}{\omega}\left(\frac{2\omega}{r+2\omega}e^{2(r_S+\omega)t} - e^{(r_S+\omega)t} + \frac{r}{r+2\omega}e^{r_S t}\right) \\ &+\frac{2\mu\overline{SF}_0}{\omega}e^{(r_S+\omega)t}\left(e^{(r_S+\omega)t} - e^{r_S t}\right) \\ &+O(\mu^2),\end{aligned} \tag{24}$$

where $\overline{F_o^2} \equiv \overline{F^2}(0)$ is the second moment of initial values . Thus, the variance $\sigma_F^2(t) = \overline{F^2}(t) - [\overline{F}(t)]^2$ is given, up to the order of μ, by

$$\begin{aligned}\sigma_F^2(t) = \sigma_{F,0}^2 e^{2(r_S+\omega)t} + \overline{F}_0 e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\ &+\frac{\mu\overline{S}_o}{\omega}\left(\frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} - e^{(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right) \\ &+\frac{2\mu(\overline{SF}_0 - \overline{S}_0\overline{F}_0)}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) \\ &+O(\mu^2).\end{aligned} \tag{25}$$

Using $R_\tau = e^{r_S\tau}$ and $W_\tau = e^{\omega\tau}$, we rewrite

$$\begin{aligned}\sigma_F^2(t) = \sigma_{F,0}^2 R_t^2 W_t^2 + \overline{F}_0 R_t W_t(R_t W_t - 1) \\ &+\frac{\mu\overline{S}_o R_t}{\omega}\left(\frac{2\omega}{r_S+2\omega}R_t W_t^2 - W_t + \frac{r_S}{r_S+2\omega}\right) \\ &+\frac{2\mu(\overline{SF}_0 - \overline{S}_0\overline{F}_0)}{\omega}R_t^2 W_t(W_t - 1) \\ &+O(\mu^2).\end{aligned} \tag{26}$$

Using Eqs. (10),(11),(20),(25), we construct pdfs for $S(t)$ and $F(t)$ at the end of cycle $t = \tau$. Then, we randomly sample a number from $P(S,\tau)$ for $S(\tau)$ and another number from $P(F,\tau)$ for $F(\tau)$. Those two numbers are cell numbers in a single Adult. We repeat this process for each Newborn to get cell numbers of all Adults. Note that the initial values for the Newborn i are

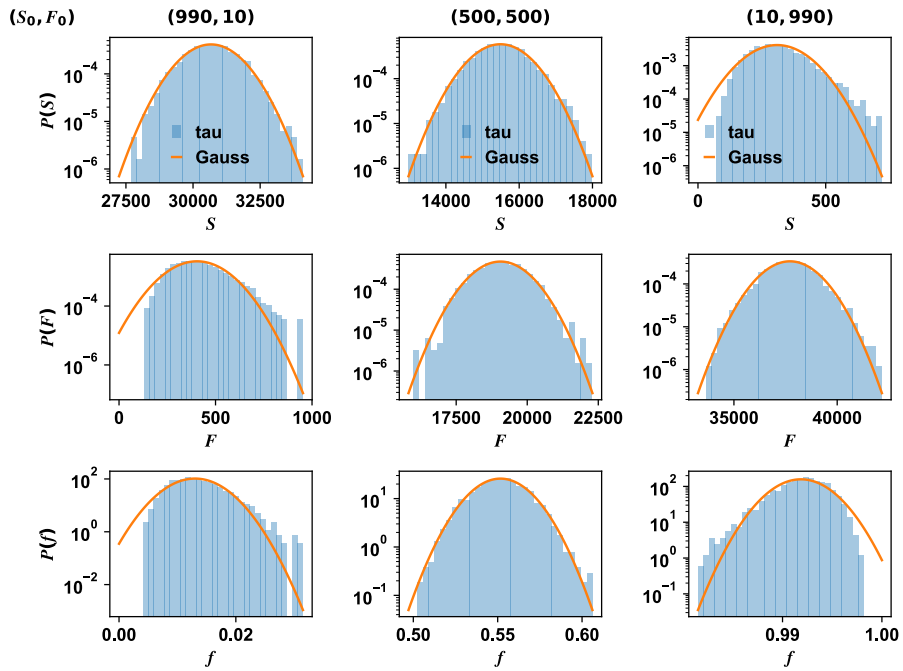

FIG. 1. Comparison between the calculated Gaussian distribution ("Gauss", with the mean and variances computed from Eqs. (10),(11),(20),(25)) and simulations using tau-leaping ("tau"). The simulations run 3000 times. The initial number of cells are $(S_0, F_0) = (990, 10), (500, 500),$ and $(10, 990)$ for each column. The parameters $r = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, and $\tau = 4.8$ are used.

$\overline{S}_0 = S^{(i)}_{k,0}$, $\overline{F}_0 = F^{(i)}_{k,0}$, $\sigma^2_{S,0} = 0$, and $\sigma^2_{F,0} = 0$. This process only requires two random numbers per collective while the result is consistent with the individual-based simulation.

Now, we check validity of Gaussian approximation for probability density functions of S and F populations. If we consider mutation from S to F as death in S population, then the process in S corresponds to a branching process with death. Also, the birth process in F including mutation results in a Luria-Delbrück distribution[3]. Thus the distributions of Adults' S and F numbers are more skewed and heavy-tailed than Gaussian. But this problem is alleviated by larger initial S and F numbers and when the maturation time τ is not very long (see Fig. 1). Since we usually consider larger initial cell numbers, we use the Gaussian approximation on S and F populations in further calculations.

Selection

After sampling cell numbers of each Adult in the maturation step, we compute the F frequencies in each collectives $\{f^{(1)}_{k,\tau}, \cdots, f^{(g)}_{k,\tau}\}$. We denote the F frequency of collective i at time $t = \tau$ in cycle k as $f^{(i)}_{k,\tau}$. Among the g Adults, we select one collective with the F frequency which is the closest value to the target frequency $\hat{f}$. The selected Adult's F frequency value is denoted by f^*_k.

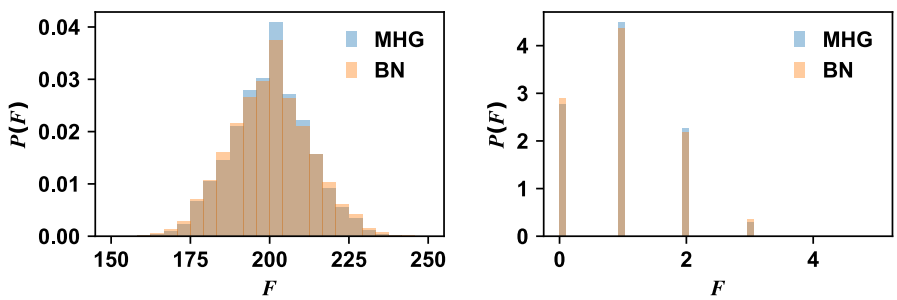

FIG. 2. Congruence between consecutive sampling (MHG for multivariate hypergeometric distribution) and independent binomial (BN) sampling. The initial numbers of cells are $S = 8000$ and $F = 2000$ for the left panel, and $S = 20$ and $F = 5$ for the right panel. 10000 samples are drawn for each distribution. Here, a parent collective is divided into 10 collectives.

In mathematical expression, the selected frequency is defined by

$$f_k^* = \text{argmin}_{f_{k,\tau}^{(i)}, i \in \{1,\cdots,g\}} |f_{k,\tau}^{(i)} - \hat{f}|. \tag{27}$$

Reproduction

Using the chosen Adult, we generate g Newborn collectives for the next cycle $k+1$. The most natural way is consecutive random sampling N_0 cells from the selected Adult without replacement. In mathematical expression, we first randomly sample $F_{k+1,0}^{(1)}$ F cells and draw $S_{k+1,0}^{(1)} = N_0 - F_{k+1,0}^{(1)}$ S cells from the selected Adult. Next, we sample $F_{k+1,0}^{(2)}$ F cells and $S_{k+1,0}^{(2)} = N_0 - F_{k+1,0}^{(2)}$ S cells from the remaining cells in the Adult. We repeat the process g times. Then the jpdf to choose $F_{k+1,0}^{(1)}, \cdots, F_{k+1,0}^{(g)}$ F cells, $P(F_{k+1,0}^{(1)}, \cdots, F_{k+1,0}^{(g)})$, follows a multivariate hypergeometric distribution.

If we assume that the selected Adult size $N_k^* = S_k^* + F_k^*$ is large enough compared to Newborn size N_0, the consecutive sampling is well approximated to the independent binomial sampling (see Fig. 2). Thus, we independently sample g numbers of F cells, $\{F_{k+1,0}^{(1)}, \cdots, F_{k+1,0}^{(g)}\}$, from the binomial distribution. The probability mass function of each $F_{k+1,0}^{(i)}$ is given by

$$P_{F_{k+1,0}^{(i)}}(F) = \frac{N_0!}{(N_0 - F)!F!} (f_k^*)^F (1 - f_k^*)^{N_0 - F}. \tag{28}$$

After sampling, the numbers of S cells are set to be $S_{k+1}^{(i)}(0) = N_0 - F_{k+1,0}^{(i)}$ for each collective. We can now start cycle $k+1$ with these Newborn collectives. By repeating the above three steps (maturation, selection, and reproduction), we run the simulation until F frequency reaches a stationary state.

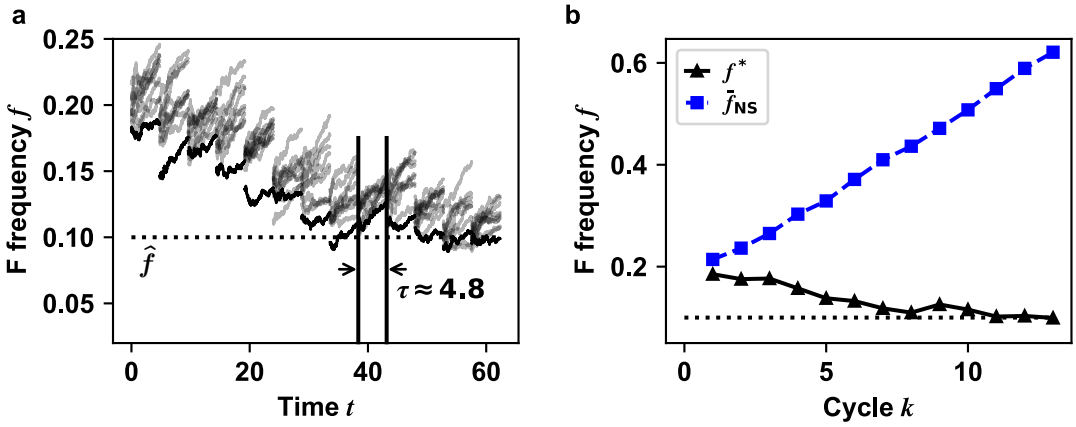

FIG. 3. **a** Trajectories of F frequency for 10 collectives ($g = 10$) over time. The collective whose frequency is closest to the target value is selected in every cycle (black lines). The gray lines denote the other collectives. For parameters, we used S growth rate $r_S = 0.5$, F growth advantage $\omega = 0.03$, mutation rate $\mu = 0.0001$, maturation time $\tau \approx 4.8$, and $N_0 = 1000$. **b** Comparison between frequency trajectories with selection (the chosen one Adult producing all offspring; black) and without selection (each Adult producing one offspring; blue) clearly shows the effect of artificial selection. The black line indicates F frequency of the selected collective f_k^* at each cycle in **a**. The blue line indicates the average trajectory without selection $\overline{f}_{\rm NS}$ (the average of $g = 10$ individual lineages without inter-collective selection at the end of each cycle).

Simulation Result

Figure 3 presents the composition trajectories of all collectives using tau-leaping algorithm in the maturation step. The selected Adults have the closest composition to the target composition $\hat{f}$. The selected Adult can have smaller F frequency than its parent Adult, so F frequency can be lowered after cycles.

In Fig. 4, we plot the absolute error d between the target frequency $\hat{f}$ and $\langle f^* \rangle$ (*i.e.* $d = |\langle f^* \rangle - \hat{f}|$) at the end of simulations (1000 cycles). Since the computing time for Tau-leaping algorithm (individual-based simulation) to reach 1000 cycles is very long, we used the sampling scheme in the above subsection. In the colormap, errors higher than 0.15 are marked with gray, which indicates selection failure. The dashed lines indicate the same boundary in Fig 2e in the main text.

II. CONDITIONAL PROBABILITY DISTRIBUTION OF THE SELECTED COLLECTIVE FREQUENCY f^* AND SELECTION PROGRESS Δf

In the main text, we identify the region of success by using selection progress $\Delta f = \text{Median}[\Psi(f_{k+1}^*|f_k^*)] - f_k^*$, which is obtained from conditional pdf of f_{k+1}^* (the F frequency of the selected Adult at cycle $k+1$) given the selected f_k^* at cycle k, written as $\Psi(f_{k+1}^* = f|f_k^*)$. We consider the challenging case where f_k^* is above the target value ($f_k^* > \hat{f}$), and therefore the Adult with minimum F frequency will be selected. To get an analytical expression of $\Psi(f_{k+1}^* = f|f_k^*)$, we first find the conditional

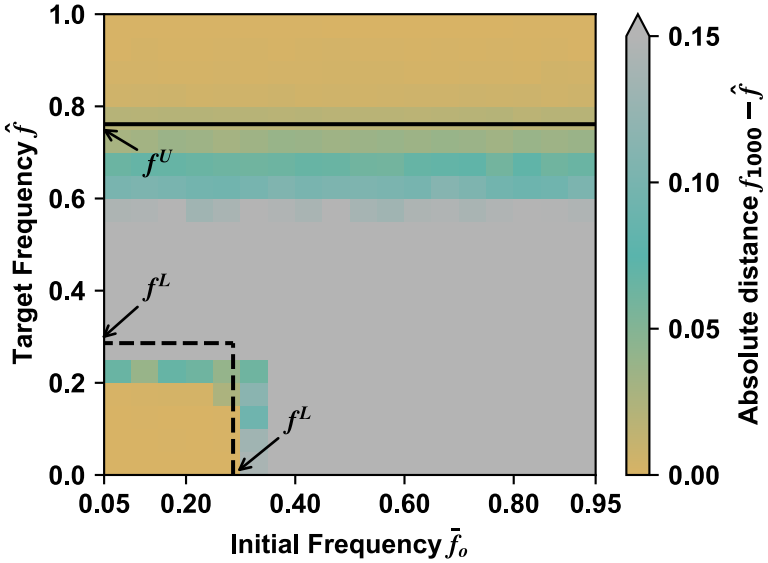

FIG. 4. Color map of the absolute error $d = |\langle f^* \rangle - \hat{f}|$ averaged selected collectives at the end of simulations ($k = 1000$) and the target frequency $\hat{f}$. The solid and dashed lines are drawn by the arguments in the main text. For parameters, we used $r_S = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, $N_0 = 1000$, $g = 10$ and $\tau \approx 4.8$. The result is the average of 300 independent simulations. Compared to Fig. 2e, this figure has a higher resolution.

pdf of f of Adults in cycle $k+1$ given f_k^* at cycle k. Then, we find $\Psi(f_{k+1}^*|f_k^*)$ from the minimum value distribution of F frequencies among g Adults. Below, we describe mathematical details of this process.

Let us start from the reproduction step from the selected Adult in cycle k. We reproduce g Newborns in next cycle $k+1$. Then the probability distribution of the F cell numbers in Newborn collectives is given in Eq. (28). If the total number of cells in a Newborn collective N_0 is large enough, Eq. (28) is approximated by the Gaussian distribution $F_{k+1,0}^{(i)} \sim \mathcal{N}(N_0 f_k^*, N_0 f_k^*(1-f_k^*))$. Then, the probability density function that $f_{k+1,0}^{(i)}$ to be ζ in Newborn collective i is

$$P_{f_{k+1,0}^{(i)}}(\zeta|f_k^*) = \frac{\sqrt{N_0}}{\sqrt{2\pi f_k^*(1-f_k^*)}} e^{-\frac{N_0(\zeta - f_k^*)^2}{2f_k^*(1-f_k^*)}}. \tag{29}$$

The Newborn collective i has initial cell numbers $S_{k+1,0}^{(i)} = N_0(1-\zeta)$ and $F_{k+1,0}^{(i)} = N_0\zeta$. From here, we ignore cycle index $k+1$ in subscript and i superscript for convenience.

Next, we write conditional pdf of Adults' F frequency with given Newborn F frequency ζ. We assume that cell numbers in Adult $S(\tau)$ and $F(\tau)$ follow Gaussian distributions as in Eqs. (5) and (6). Based on Eqs. (10), (11), (20), and (25), we have

$$S(\tau) \approx \overline{S}(\tau) + \sigma_S(\tau) G_S(\tau), \tag{30}$$

$$F(\tau) \approx \overline{F}(\tau) + \sigma_F(\tau) G_F(\tau), \tag{31}$$

where $G_S(\tau)$ and $G_F(\tau)$ are random variable following the standard distribution $\mathcal{N}(0,1)$. Note that each Gaussian is sharp if Newborn size N_0 is sufficiently large ($\bar{S}(\tau) \gg \sigma_S(\tau)$ and $\bar{F}(\tau) \gg \sigma_F(\tau)$). Then, we can approximately write $f(\tau)$ as

$$\begin{aligned} f(\tau) &\approx \frac{\overline{F}(\tau) + \sigma_F(\tau) G_F(\tau)}{\overline{S}(\tau) + \sigma_S(t) G_S(\tau) + \overline{F}(\tau) + \sigma_F(\tau) G_F(\tau)} \\ &\approx \overline{f}(\tau) + \sigma_f(t) G_f(\tau). \end{aligned} \tag{32}$$

The mean of f is given by

$$\overline{f}(\tau) = \frac{\overline{F}(\tau)}{\overline{S}(\tau) + \overline{F}(\tau)} = \frac{\zeta W_\tau + \frac{\mu}{\omega}(1-\zeta)(W_\tau - 1)}{1 - \zeta + \zeta W_\tau + \frac{\mu}{\omega}(1-\zeta)(W_\tau - 1)}, \tag{33}$$

and the variance is

$$\begin{aligned} \sigma_f^2(\tau) &= \frac{(1-\overline{f}(\tau))^2 \sigma_F^2(\tau) + \overline{f}(\tau)^2 \sigma_S^2(\tau)}{\overline{N}(\tau)^2} \\ &= \frac{1}{N_0 R_\tau \left((1-\zeta) + \zeta W_\tau + \frac{\mu}{\omega}(1-\zeta)(W_\tau - 1)\right)^4} \times \\ &\quad \left[(1-\zeta)^2 \left(\zeta W_\tau (R_\tau W_\tau - 1) + \frac{\mu}{\omega}(1-\zeta) \left(\frac{2\omega/r_S}{1+2\omega/r_S} R_\tau W_\tau^2 - W_\tau + \frac{1}{1+2\omega/r_S} \right) \right) \right. \\ &\quad \left. + \left(\zeta W_\tau + \frac{\mu}{\omega}(1-\zeta)(W_\tau - 1) \right)^2 (1-\zeta)(R_\tau - 1) \right]. \end{aligned} \tag{34}$$

where $R_\tau = e^{r_S \tau}$ and $W_\tau = e^{\omega \tau}$. The average Adult size is $\overline{N}(\tau) = \overline{S}(\tau) + \overline{F}(\tau)$. Thus the Adult's F frequency $f(\tau) = f_\tau$ follows the Gaussian distribution $\mathcal{N}(\overline{f}_\tau, \sigma_{f_\tau}^2)$ whose pdf is given by

$$P_{f_\tau}(f|\zeta) = \frac{1}{\sqrt{2\pi}\sigma_{f_\tau}} e^{-\frac{(f-\overline{f}_\tau)^2}{2\sigma_{f_\tau}^2}}. \tag{35}$$

Next, we get the conditional pdf of $f_{k+1,\tau}$ (offspring Adult's F frequency in cycle $k+1$) given f_k^* . We multiply Eqs. (29) and (35) and take integral over ζ:

$$P_{f_{k+1,\tau}}(f|f_k^*) = \int_0^1 d\zeta \; P_{f_{k+1,\tau}}(f|\zeta) \; P_{f_{k+1,0}}(\zeta|f_k^*). \tag{36}$$

After maturation in cycle $k+1$, the Adult with the smallest frequency is selected among g Adult collectives, denoted as $f_{\mathrm{k+1}}^{\min}$. The pdf of $f_{k+1}^{\min}$ is obtained by the theory of extreme value statistics [4]. The cumulative distribution function (cdf) of the minimum value $f_{k+1}^{\min}$ is given by

$$\begin{aligned} C_{f_{k+1}^{\min}}(f|f_k^*) &= \mathrm{Prob}[f_{k+1}^{\min} \le f | f_k^*] \\ &= 1 - \mathrm{Prob}[f_{k+1,\tau}^{\min} \ge f | f_k^*] \\ &= 1 - \mathrm{Prob}[(f_{k+1,\tau}^{(1)} \ge f) \wedge (f_{k+1,\tau}^{(2)} \ge f) \wedge \cdots \wedge (f_{k+1,\tau}^{(g)} \ge f) | f_k^*]. \end{aligned} \tag{37}$$

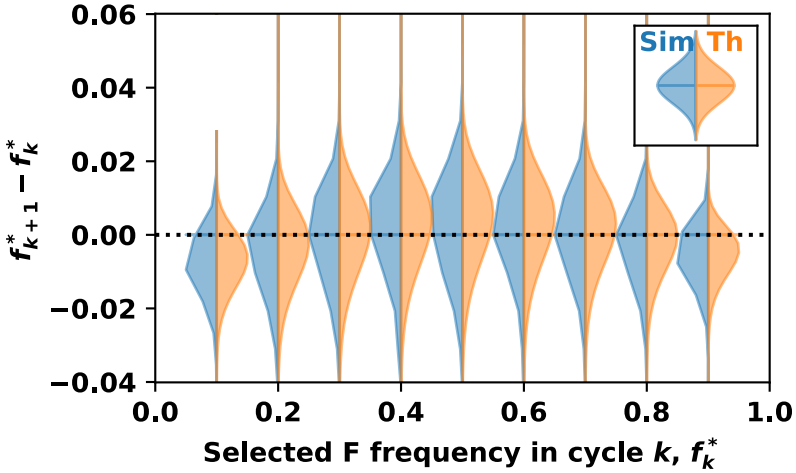

FIG. 5. The probability density functions of the selected Adult's F frequency f^*_{k+1} subtracted by f^*_k. For simulations (blue), at each f^*_k, we performed 1000 stochastic simulations. The orange distribution represents Eq. (39) computed by numerical integration. The median values of the distributions are shown in Fig. 4**a** in the main text.

Since frequencies are independent and identically distributed, $C_{f^{\min}_{k+1}}(f|f^*_k) = 1 - [\mathrm{Prob}(f_{k+1,\tau} \geq f)|f^*_k]^g$. Note that $\mathrm{Prob}[f_{k+1,\tau} \geq f|f^*_k] = \int_f^1 df' P_{f_{k+1,\tau}}(f'|f^*_k) = 1 - \int_0^f df' P_{f_{k+1,\tau}}(f'|f^*_k) = 1 - C_{f_{k+1,\tau}}(f|f^*_k)$, and Eq. (37) becomes

$$C_{f^{\min}_{k+1}}(f|f^*_k) = 1 - \left(1 - C_{f_{k+1,\tau}}(f|f^*_k)\right)^g . \tag{38}$$

Then, the probability density function $\Psi(f^*_{k+1}|f^*_k)$ is obtained by differentiating Eq. (38) with respect to f and replacing $f \to f^*_{k+1}$,

$$\Psi(f^*_{k+1}|f^*_k) = g\left(1 - C_{f_{k+1,\tau}}(f^*_{k+1}|f^*_k)\right)^{g-1} P_{f_{k+1,\tau}}(f^*_{k+1}|f^*_k). \tag{39}$$

We compute the probability density function (39) by using numerical integration and compare it with the stochastic simulation results in Fig. 5. The two distributions are similar.

To get the analytic approximation of the median of Eq. (39), we assume that the Adult's F frequency distribution is Gaussian. Then we only need to calculate the mean $\overline{f}(\tau)$ and variance $\sigma^2_f(\tau)$ of Adult's F frequency. Instead of calculating the integral with respect to ζ in Eq. (36), we put a set of initial value from Newborn's F frequency distribution $\mathcal{N}(f^*_k, f^*_k(1-f^*_k)/N_0)$ in Eqs. (12),(13),(20),(26): $\overline{S}_0 = N_0(1-f^*_k)$, $\overline{F}_0 = N_0 f^*_k$, $\sigma^2_{S,0} = N_0 f^*_k(1-f^*_k)$, $\sigma^2_{F,0} = N_0 f^*_k(1-f^*_k)$, and $\overline{SF}_0 - \overline{S}_0\overline{F}_0 = -N_0 f^*_k(1-f^*_k)$. Then we have

$$\overline{f}(\tau) = \frac{\overline{F}(\tau)}{\overline{S}(\tau) + \overline{F}(\tau)} = \frac{f^*_k W_\tau + \frac{\mu}{\omega}(1-f^*_k)(W_\tau - 1)}{1 - f^*_k + f^*_k W_\tau + \frac{\mu}{\omega}(1-f^*_k)(W_\tau - 1)}, \tag{40}$$

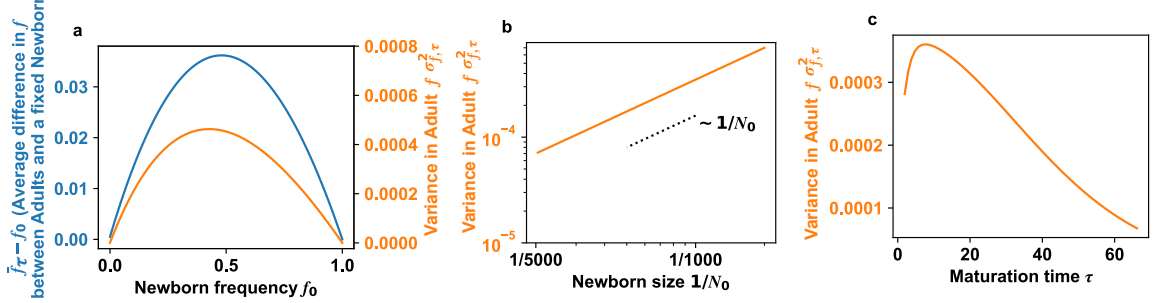

FIG. 6. **a** Mean (Eq. (33)) and variance (Eq. (34)) of f values of Adult collectives with respect to the Newborn frequency f_0. **b** Scaling relation of F frequency variance (Eq. (41)) with Newborn collective size N_0. The initial F frequency is 0.5. The parameters are $r_S = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, and $\tau \approx 4.8$. **c** Relation of F frequency variance (Eq. (41)) with maturation time τ. Other parameters are the same as **b**.

$$\sigma_f^2(\tau) = \frac{(1-f_k^*)}{N_0 R_\tau \left(1 - f_k^* + f_k^* W_\tau + \frac{\mu}{\omega}(1-f_k^*)(W_\tau - 1)\right)^4} \times \tag{41}$$
$$\Bigg[f_k^* W_\tau \left\{ (2 - 2f_k^* + 2f_k^{*2}) R_\tau W_\tau - (1-f_k^*) - f_k^* W_\tau \right\}$$
$$+ \frac{\mu}{\omega}(1-f_k^*) \Bigg\{ (1-f_k^*) \left(\left(\frac{2\omega}{r_S + 2\omega} - 2f_k^* \right) R_\tau W_\tau^2 - W_t + \frac{r_S}{r_S + 2\omega} + 2f_k^* R_\tau W_\tau \right)$$
$$+ 2f_k^* \big((1+f_k^*) R_\tau - 1 \big) W_\tau (W_\tau - 1) \Bigg\} + O(\mu^2) \Bigg]$$

which give rise to Eq. (3) and Eq. (4) in the main text, respectively. The functional form of Eqs. (40) and (41) are plotted in Fig. 6**a**.

The median ($\text{Median}[\Psi(f_{k+1}^*|f_k^*)] \equiv \tilde{f}$) of Eq. (39) satisfies $C_{f_{k+1}^{\min}}(\tilde{f}|f_k^*) = \frac{1}{2}$, which means $\tilde{f} = C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right)$. If we assume that the distribution Eq. (39) is Gaussian, then the inverse function $C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right)$ can be written as

$$\text{Median}[\Psi(f_{k+1}^*|f_k^*)] = C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right) = \overline{f}(\tau) + \left[\Phi^{-1}\left(\frac{\ln 2}{g}\right)\right] \sigma_f(\tau), \tag{42}$$

where $\Phi^{-1}(y)$ is an inverse cumulative density function (CDF) of the normal distribution with mean $\overline{f}(\tau)$ in Eq. (41) and standard deviation $\sigma_f(\tau)$, a square root of Eq. (41). Subtracting f_k^* from Eq. (42) gives the selection progress

$$\Delta f = \text{Median}[\Psi(f_{k+1}^*|f_k^*)] - f_k^* = \overline{f}(\tau) + \left[\Phi^{-1}\left(\frac{\ln 2}{g}\right)\right] \sigma_f(\tau) - f_k^* \tag{43}$$

which is Eq. (2) in the main text.

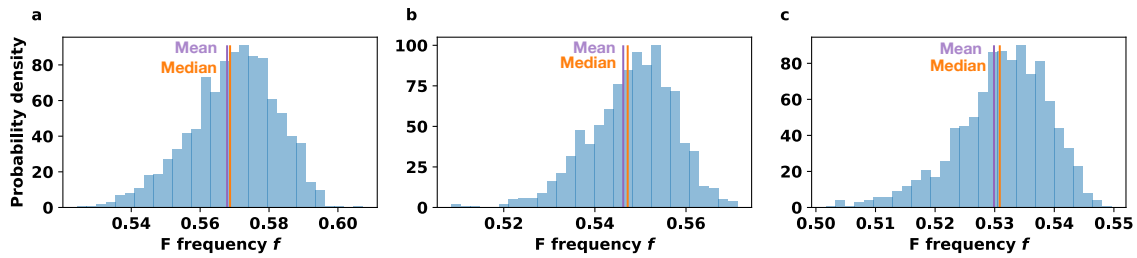

FIG. 7. Median (orange) and mean (violet) have similar distributions. We performed 1000 simulations to get probability density. **a** $g = 10$, **b** $g = 100$, and **c** $g = 1000$. Initial F frequency is $f_k^* = 0.5$. The parameters are $r_S = 0.5$, $\omega = 0.03$, $\mu = 0.0001$ and $\tau = \ln[1000]/r_S$.

Further, we get an asymptotic expression of $\Phi^{-1}(\ln 2/g)$ when g is large (or $\Phi^{-1}(y)$ with small y). Here, we introduce a method from [5]. We start from the CDF of the standard normal distribution, $\Phi(x) = (\mathrm{erfc}(-x/\sqrt{2})/2$ where the function $\mathrm{erfc}(x)$ is the complementary error function. To get the expression of Φ^{-1}, we need an asymptotic expression of inverse of $y = \mathrm{erfc(x)}$ function ($x = \mathrm{erfc}^{-1}(y)$) as the inverse CDF $\Phi^{-1}(y) = -\sqrt{2}\mathrm{erfc}^{-1}(2y)$. The known asymptotic expansion of $y = \mathrm{erfc}(x)$ for large x is $\mathrm{erfc}(x) \approx e^{-x^2}/x\sqrt{\pi}$. By taking logarithm of both sides, we have

$$x^2 \approx -\ln y - \frac{1}{2}\ln \pi x^2. \tag{44}$$

Replacing x^2 on the righthand side in Eq. (44) into the expression itself, we get continued logarithmic form of

$$x^2 \approx -\ln y - \frac{1}{2}\ln \pi(-\ln y - \frac{1}{2}\ln \pi(-\ln y - \frac{1}{2}\ln \cdots). \tag{45}$$

Inserting $x = \mathrm{erfc}^{-1}(y)$ (square root of Eq. (45)) into the inverse CDF $\Phi^{-1}(y) = -\sqrt{2}\mathrm{erfc}^{-1}(2y)$, we have $\Phi^{-1}(y) \approx -\sqrt{-2\ln 2y - \ln\pi(-\ln 2y - \cdots)}$. So, the asymptotic expression of $\Phi^{-1}(\ln 2/g)$ is given by

$$\Phi^{-1}\left(\frac{\ln 2}{g}\right) \approx -\sqrt{2\ln g - \ln[\ln g] + \cdots}. \tag{46}$$

III. CRITICAL NEWBORN SIZE $\check{N}_0$ TO ALLOW ALL TARGET FREQUENCIES

First we note that σ_f^2 in Eq. (41) is proportional to N_0^{-1} for the following reasons. Variance $\sigma_S^2(t)$ in Eq. (20) scales linearly with N_0 since both $\overline{S}_0$ and $\sigma_{S,0}^2$ scale linearly with N_0. Variance $\sigma_F^2(t)$ in Eq. (25) also scales linearly with N_0 because $\sigma_{F,0}^2$, $\overline{F}_0$, $\overline{S}_0$, and covariance $\overline{SF}_0 - \overline{S}_0\overline{F}_0$ all scale linearly with N_0. The mean Adult size $\overline{N}(\tau) = \overline{S}(\tau) + \overline{F}(\tau)$ is also proportional to N_0 because the average cell numbers in Eqs. (10) and (11) are linear with respect to N_0. Thus, the scaling relation of Eq. (41) is given by $\sigma_f^2(\tau) = [(1 - f(\tau))^2\sigma_F^2(\tau) + \overline{f}(\tau)^2\sigma_S^2(\tau)]/\overline{N}(t)^2 \sim N_0/N_0^2 \sim 1/N_0$.

Small N_0 makes all target frequencies achievable shown in Fig. 4**a** in the main text. That is because small N_0 induces large σ_f, and thus N_0 smaller than a certain critical value $\breve{N}_0$ makes the selection progress Δf always negative regardless of the value of f_k^* (*i.e.* $\Delta f = \overline{f}(\tau) + \left[\Phi^{-1}(\frac{\ln 2}{g})\right]\sigma_f(\tau) - f_k^* < 0$). That means the inter-collective selection overcomes intra-collective selection in any target frequencies. To get an analytical approximation of the critical newborn size $\breve{N}_0$, we simply assume that selection progress Δf is maximum at $f_k^* = 1 - f_k^* = \frac{1}{2}$ where the changes in $\bar{f}$ and σ_f^2 are fastest. If the maximum value of Δf is zero, all other values of Eq. (42) are negative, which naturally states all targets are achievable. Putting $f_k^* = \frac{1}{2}$, Eqs. (40) and (41) become

$$\overline{f}(\tau)|_{f_k^*=\frac{1}{2}} = \frac{\overline{F}(\tau)}{\overline{S}(\tau) + \overline{F}(\tau)} = \frac{W_\tau + \frac{\mu}{\omega}(W_\tau - 1)}{1 + W_\tau + \frac{\mu}{\omega}(W_\tau - 1)}, \tag{47}$$

and

$$\begin{aligned}\sigma_f^2(\tau)|_{f_k^*=\frac{1}{2}} \approx & \frac{2}{N_0 R_\tau \left(1 + W_\tau + \frac{\mu}{\omega}(W_\tau - 1)\right)^4} \times \\ & \Bigg[W_\tau \left(3R_t W_\tau - 1 - W_\tau\right) \\ & + \frac{\mu}{\omega}\left\{\left(\left(\frac{2\omega}{r_S + 2\omega} - 1\right) R_t W_\tau^2 - W_t + \frac{r_S}{r_S + 2\omega} + R_t W_\tau\right) + (3R_t - 2)W_\tau(W_\tau - 1)\right\}\Bigg]\end{aligned} \tag{48}$$

So, by setting $\Delta f|_{f_k^*=\frac{1}{2}} = \overline{f}(\tau)|_{f_k^*=\frac{1}{2}} + \left[\Phi^{-1}(\frac{\ln 2}{g})\right]\sigma_f(\tau)|_{f_k^*=\frac{1}{2}} - \frac{1}{2} = 0$ with $N_0 = \breve{N}_0$, we get a solution of

$$\begin{aligned}\breve{N}_0 = & \left[\Phi^{-1}\left(\frac{\ln 2}{g}\right)\right]^2 \frac{8}{R_\tau \left[1 + W_\tau + \frac{\mu}{\omega}(W_\tau - 1)\right]^2 \left[1 - W_\tau - \frac{\mu}{\omega}(W_\tau - 1)\right]^2} \times \\ & \Bigg[W_\tau \left(3R_\tau W_\tau - 1 - W_\tau\right) \\ & + \frac{\mu}{\omega}\left\{\left(\left(\frac{2\omega}{r_S + 2\omega} - 1\right) R_\tau W_\tau^2 - W_\tau + \frac{r_S}{r_S + 2\omega} + R_\tau W_\tau\right) + (3R_\tau - 2)W_\tau(W_\tau - 1)\right\}\Bigg]\end{aligned} \tag{49}$$

Thus, all target frequencies are successfully selected with Newborn size N_0 smaller than $\breve{N}_0$. If mutation rate is zero, the critical value becomes

$$\breve{N}_0 = \left[\Phi^{-1}\left(\frac{\ln 2}{g}\right)\right]^2 \frac{8W_\tau \left(3R_\tau W_\tau - W_\tau - 1\right)}{R_\tau \left(W_\tau^2 - 1\right)^2}. \tag{50}$$

IV. SELECTION WITHOUT MUTATION $\mu = 0$

When the mutation rate is zero, two genotypes behave as two distinct species. The compositional change is provided by Eq. (42) with setting $\mu = 0$. Corresponding $\bar{f}$ in Eq. (40) and σ_f^2 in Eq.

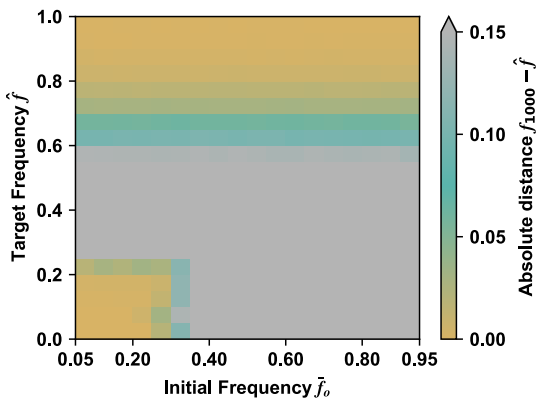

FIG. 8. Simulation with zero mutation rate. Color map of the absolute error $d = |\langle f^* \rangle - \hat{f}|$ between frequency $\langle f^* \rangle$ of the averaged selected collectives at the end of simulations ($k = 1000$) and the target frequency $\hat{f}$. For parameters, we used $r_S = 0.5$, $\omega = 0.03$, $\mu = 0$, $N_0 = 1000$, $g = 10$ and $\tau \approx 4.8$.

(41) become

$$\overline{f}(\tau) = \frac{f_k^* W_\tau}{1 - f_k^* + f_k^* W_\tau}, \tag{51}$$

$$\sigma_f^2(\tau) = \frac{f_k^*(1 - f_k^*) W_\tau \left[(2 - 2f_k^* + 2f_k^{*2}) R_\tau W_\tau - (1 - f_k^*) - f_k^* W_\tau \right]}{N_0 R_\tau \left(1 - f_k^* + f_k^* W_\tau\right)^4} \tag{52}$$

Equations (51) and (52) suggest that when a community consists of two competing species, we obtain similar conclusions on the accessible region for target composition. The stochastic simulation results are present in Fig. 8.

V. STRONGER OR WEAKER ADVANTAGES ω

The solution of equation (2) in main text provides the boundary values with varying the ω, the fitness advantage of F over S. We numerically calculate the solutions and plot in Fig. 9.

VI. DELETERIOUS MUTATION $\omega < 0$

In the main text, we show the target composition can be achieved in some ranges of initial and target values when the mutation is beneficial to growth. The same analogy can be applied when the mutation is deleterious. Since the F cells grow slower than the S cells ($\omega < 0$), the F frequency naturally decreases in the maturation step. Then, the challenging case is selecting larger F frequency against to the intra-collective selection. So the conditional probability distribution $\Psi(f_{k+1}^* | f_k^*)$ that

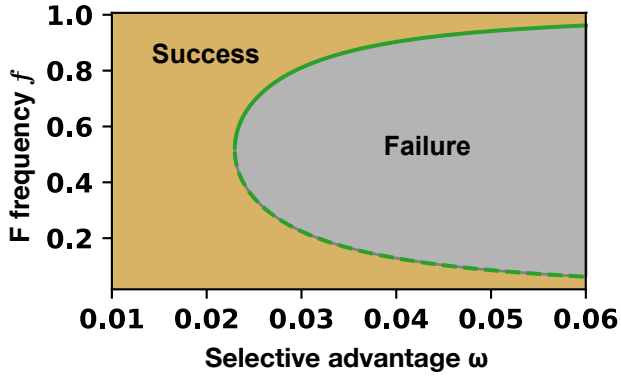

FIG. 9. Change of success region in varying selective advantage ω. $r_S = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, $N_0 = 1000$, $g = 10$ and $\tau \approx 4.8$.

we consider now is a maximum value distribution of Eq. (36). Thus, instead of Eq. (37), we look for the cumulative distribution function of the maximum value $f_{k+1}^{\max}$ such that

$$\begin{aligned} C_{f_{k+1}^{\max}}(f|f_k^*) &= \text{Prob}[f_{k+1}^{\max} \geq f|f_k^*] \\ &= \text{Prob}[(f_{k+1}^{(1)}(\tau) \geq f) \wedge (f_{k+1}^{(2)}(\tau) \geq f) \wedge \cdots \wedge (f_{k+1}^{(g)}(\tau) \geq f)|f_k^*]. \end{aligned} \tag{53}$$

If all frequencies are independent and identical distributed random variables, the cumulative distribution function becomes

$$C_{f_{k+1}^{\max}}(f|f_k^*) = \left(C_{f_{k+1,\tau}}(f|f_k^*)\right)^g. \tag{54}$$

Likewise in the previous section, we get the conditional probability density function by differentiating Eq. (54) with respect to f and replacing $f \to f_{k+1}^*$ as

$$\Psi(f_{k+1}^*|f_k^*) = g\left(C_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*)\right)^{g-1} P_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*). \tag{55}$$

The distribution in Eq. (55) is evaluated for various f_k^* in Fig. 10**a** with numerical simulations, and the median values of distributions are presented in Fig. 10**b**. In the case of $\omega = -0.03$, the target frequency lower than around 0.3 and larger than around 0.7 can be selected. Since the sign of ω is opposite to the result in the main text, the diagram is reversed from Fig. 2e in the main text.

VII. SELECTING MORE THAN ONE COLLECTIVE

In the main text, we choose one collective which has the closest frequency to the target among g collectives. Such 'top 1' strategy allows us to apply extreme value theory. However, 'top 1' may be too restrictive [6]. Thus, we test the 'top-tier' strategy by choosing top 5 among 100 Adults (Fig. 11). The top-tier strategy is shown to be inefficient in our system. This is because in [6], nonheritable variations−such as stochastic fluctuations in species composition introduced by pipetting − caused nonheritable variations in collective function. Nonheritable variations could potentially

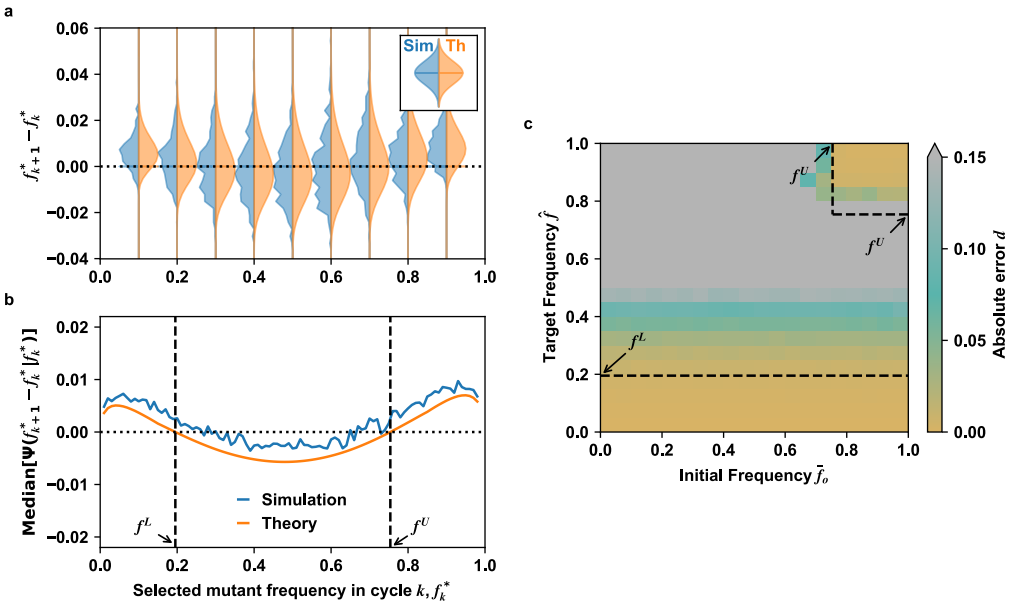

FIG. 10. Artificial selection also works for deleterious mutation. **a** Conditional probability density functions of $f_{k+1}^* - f_k^*$ for various f_k^* values. The left-hand side distribution is obtained from simulations and the right-hand side distribution is numerically obtained by evaluating Eq. (55). Small triangles inside indicate the median values of the distributions. **b** the median value of distributions at a given f_k^*. The points where the shifted median becomes zero, Median$[\Psi(f_{k+1}^* - f_k^* | f_k^*)] = 0$ are denoted as f^L and f^U, respectively. **c** The relative error between the target frequency $\hat{f}$ and the ensemble averaged selected frequency $\langle f_k^* \rangle$ is measured after 1000 cycles starting from the initial frequency $\overline{f}_{1,0}$. Either the lower target frequencies or the higher target frequencies starting from the high initial frequencies can be achieved. The black dashed lines indicate the predicted boundary values f^U and f^L in **a**.

mask desired mutations if these mutations happened to occur in an 'unlucky' environment that yielded lower collective functions. Hence, lenient selection would allow the preservation of these mutations. In contrast here, stochastic fluctuations in genotype composition is heritable: a parent Newborn with lower F frequency f will tend to have offspring Newborns with lower f values. Hence, top-1 is more effective in this study.

VIII. EXTENSION TO THREE-POPULATION SYSTEM

We assume that collectives consist of three genotypes with slow-growing(S), fast-growing(F), and faster-growing(FF) types. The growth rate of S is r_S. Each mutation adds ω to the growth rate. Thus, the F and FF types have growth rates $r_S + \omega$ and $r_S + 2\omega$, respectively. The mutation rate

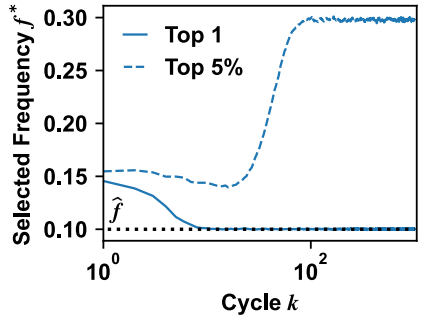

FIG. 11. Selecting Top-5% outperforms selecting Top 1. We bred 100 collectives and chose either top-1 collective (solid line) or top-5 collectives (dashed line) with f closest to the target value $\hat{f}$ (black dotted line).

is μ. So, the birth and mutation events are written by the chemical reactions:

$$\mathrm{S} \xrightarrow{r_S} \mathrm{S}+\mathrm{S}, \tag{56}$$

$$\mathrm{F} \xrightarrow[\omega>0]{r_S+\omega} \mathrm{F}+\mathrm{F}, \tag{57}$$

$$\mathrm{FF} \xrightarrow[\omega>0]{r_S+2\omega} \mathrm{FF}+\mathrm{FF} \tag{58}$$

$$\mathrm{S} \xrightarrow{\mu} \mathrm{F}, \tag{59}$$

$$\mathrm{F} \xrightarrow{\mu} \mathrm{FF} \tag{60}$$

We write a master equation of the processes for $P(S, F, FF, t)$ which is the probability to have S, F and FF numbers of S, F, and FF cells at time t, respectively.

$$\begin{aligned}\frac{dP(S,F,FF,t)}{dt} &= r_S(S-1)P(S-1,F,FF,t) - r_S S P(S,F,FF,t) \\ &+(r_S+\omega)(F-1)P(S,F-1,FF,t) - (r_S+\omega)FP(S,F,FF,t) \\ &+(r_S+2\omega)(FF-1)P(S,F,FF-1,t) - (r_S+2\omega)FFP(S,F,FF,t) \\ &+\mu(S+1)P(S+1,F-1,FF,t) - \mu S\,P(S,F,FF,t) \\ &+\mu(F+1)P(S,F+1,FF-1,t) - \mu F\,P(S,F,FF,t).\end{aligned} \tag{61}$$

The composition of collective i in cycle k is now represented with two frequencies $(f_k^{(i)}(t), h_k^{(i)}(t)) \equiv \mathbf{p}_k^{(i)}(t)$ where the F frequency is $f_k^{(i)}(t) = F_k^{(i)}(t)/(S_k^{(i)}(t)+F_k^{(i)}(t)+FF_k^{(i)}(t))$ and the FF frequency is $h_k^{(i)}(t) = FF_k^{(i)}(t)/(S_k^{(i)}(t)+F_k^{(i)}(t)+FF_k^{(i)}(t))$. Then, the target composition is set to be $(\hat{f}, \hat{h})$. The composition of the selected Adult in cycle k is $(f_k^*, h_k^*) \equiv \mathbf{p}_k^*$. We apply the processes used in the above section II to obtain the conditional probability $\Psi(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*)$ by using the master Eq. (61).

At the reproduction step in cycle k, we choose N_0 cells from the selected Adult whose composition is $(f_k^*, h_k^*) \equiv \mathbf{p}_k^*$. Then, Newborn collectives are independently sampled from a multinomial distribution. For convenience, we drop the collective index (i). Then, the conditional joint probability

mass function of $F_{k+1,0}, FF_{k+1,0}$ cells is represented by

$$P(F_{k+1,0}, FF_{k+1,0}|f_k^*, h_k^*) = \frac{N_0!}{(N_0 - F_{k+1,0} - FF_{k+1,0})! F_{k+1,0}! FF_{k+1,0}!} \times (1 - f_k^* - h_k^*)^{N_0 - F_{k+1,0} - FF_{k+1,0}} (f_k^*)^{F_{k+1,0}} (h_k^*)^{FF_{k+1,0}}, \tag{62}$$

where the number of S $S_{k+1,0}$ is automatically set to be $S_{k+1,0} = N_0 - F_{k+1,0} - FF_{k+1,0}$. Then, the approximated multivariate normal distribution is $\mathcal{N}(N_0 \mathbf{p}_k^*, N_0 \mathbf{M}_{k+1})$ where the mean distribution is $\boldsymbol{p}_k^* = (f_k^*, h_k^*)$ and covariance matrix is $\mathbf{M}_{k+1}$. The diagonal terms of $\mathbf{M}_{k+1}$ are variances $\sigma_X^2 = \overline{X^2} - [\overline{X}]^2$ and the off-diagonal terms are covariances $\sigma_{XY} = \overline{XY} - \overline{X}\,\overline{Y}$. The matrix is given by

$$\mathbf{M}_{k+1} = \begin{bmatrix} \frac{f_k^*(1-f_k^*)}{N_0^2} & -\frac{f_k^* h_k^*}{N_0^2} \\ -\frac{f_k^* h_k^*}{N_0^2} & \frac{h_k^*(1-h_k^*)}{N_0^2} \end{bmatrix} \equiv \begin{bmatrix} \sigma^2_{f_{k+1,0}} & \sigma_{f_{k+1,0} h_{k+1,0}} \\ \sigma_{f_{k+1,0} h_{k+1,0}} & \sigma^2_{h_{k+1,0}} \end{bmatrix}. \tag{63}$$

Then a Newborn's composition $\boldsymbol{\rho}(\zeta, \eta)$ follows the multivariate Gaussian distribution $(\zeta, \eta) \sim \mathcal{N}(\boldsymbol{p}_k^*, \mathbf{M}_{k+1})$ whose joint probability distribution is given by

$$P_{\boldsymbol{\rho}_{k+1,0}}(\boldsymbol{\rho}|\mathbf{p}_k^*) = \frac{1}{\sqrt{2\pi}^2 \sqrt{\det \boldsymbol{M}_{k+1}}} e^{-\frac{1}{2}(\boldsymbol{\rho} - \boldsymbol{p}_k^*)\boldsymbol{M}_{k+1}^{-1}(\boldsymbol{\rho} - \boldsymbol{p}_k^*)^T}. \tag{64}$$

At the beginning of cycle k, a newborn collective starts from (S_0, F_0, FF_0) cells (for convenience, cycle index k is dropped.) In terms of (ζ, η), each initial numbers are $S_0 = N_0(1-\zeta-\eta)$, $F_0 = N_0\zeta$, and $FF_0 = N_0\eta$. Their initial covariance matrix is $N_0^2 \mathbf{M}_{k+1}$. By using Eq. (61), we can write ordinary differential equations up to the second moment.

$$\frac{d\overline{S}(t)}{dt} = r_S \overline{S}(t) - \mu \overline{S}(t), \tag{65}$$

$$\frac{d\overline{F}(t)}{dt} = (r_S + \omega)\overline{F}(t) + \mu(\overline{S}(t) - \overline{F}(t)), \tag{66}$$

$$\frac{d\overline{FF}(t)}{dt} = (r_S + 2\omega)\overline{FF}(t) + \mu \overline{F}(t), \tag{67}$$

$$\frac{d\overline{S^2}}{dt} = 2(r_S - \mu)\overline{S^2} + (r + \mu)\overline{S}, \tag{68}$$

$$\frac{d\overline{F^2}}{dt} = 2(r_S + \omega - \mu)\overline{F^2} + (r_S + \omega + \mu)\overline{F} + \mu(2\overline{SF} + \overline{F}), \tag{69}$$

$$\frac{d\overline{FF^2}}{dt} = 2(r_S + 2\omega)\overline{FF^2} + (r_S + 2\omega)\overline{FF} + \mu(2\overline{FFF} + \overline{F}), \tag{70}$$

$$\frac{d\overline{SF}}{dt} = (2r_S + \omega - 2\mu)\overline{SF} + \mu(\overline{S^2} - \overline{S}), \tag{71}$$

$$\frac{d\overline{FFF}}{dt} = (2r_S + 3\omega - \mu)\overline{FFF} + \mu(\overline{SFF} + \overline{F^2} - \overline{F}), \tag{72}$$

$$\frac{d\overline{SFF}}{dt} = (2r_S + 2\omega - \mu)\overline{SFF} + \mu \overline{SF}. \tag{73}$$

The initial conditions of the system in coupled Eqs. (65)-(73) are obtained by the mean and (co)variances of Eq. (62). By solving equations numerically, we obtain a set of mean cell numbers $(\overline{S}, \overline{F}, \overline{FF})$ and a set of variances $(\sigma_S^2, \sigma_F^2, \sigma_{FF}^2)$ as well as covariances $(\sigma_{SF}, \sigma_{FFF}, \sigma_{SFF})$. We assume that the covariances are smaller than the variances. We consider S, F, and FF as Gaussian random variables

$$S(t) \approx \overline{S}(t) + \sigma_S(t) G_S(t), \tag{74}$$

$$F(t) \approx \overline{F}(t) + \sigma_F(t) G_F(t), \tag{75}$$

$$FF(t) \approx \overline{FF}(t) + \sigma_{FF}(t) G_{FF}(t). \tag{76}$$

Then, the F frequency becomes

$$\begin{aligned} f(t) &\approx \frac{\overline{F}(t) + \sigma_F(t) G_F(t)}{\overline{S}(t) + \sigma_S(t) G_S(t) + \overline{F}(t) + \sigma_F(t) G_F(t) + \overline{FF}(t) + \sigma_{FF}(t) G_{FF}(t)} \\ &\approx \overline{f}(t) + \sigma_f(t) G_f(t), \end{aligned} \tag{77}$$

where $\overline{f} = \overline{F}/(\overline{S} + \overline{F} + \overline{FF})$ and $\sigma_f^2 = \left(\overline{f}^2(\sigma_S^2 + \sigma_{FF}^2) + (1 - \overline{f}^2)\sigma_F^2\right)/(\overline{S} + \overline{F} + \overline{FF})$. Similarly, the FF frequency is

$$\begin{aligned} h(t) &\approx \frac{\overline{FF}(t) + \sigma_{FF}(t) G_{FF}(t)}{\overline{S}(t) + \sigma_S(t) G_S(t) + \overline{F}(t) + \sigma_F(t) G_F(t) + \overline{FF}(t) + \sigma_{FF}(t) G_{FF}(t)} \\ &\approx \overline{h}(t) + \sigma_h(t) G_h(t), \end{aligned} \tag{78}$$

where $\overline{h} = \overline{FF}/(\overline{S} + \overline{F} + \overline{FF})$ and $\sigma_h^2 = \left(\overline{h}^2(\sigma_S^2 + \sigma_F^2) + (1 - \overline{h}^2)\sigma_{FF}^2\right)/(\overline{S} + \overline{F} + \overline{FF})$. The dynamic flow of F and FF frequencies during maturation is shown in Fig. 12**a**. If the covariances are small enough, we can approximate the joint probability distribution of Adult's composition $(f_\tau, h_\tau) = \mathbf{p}_\tau$ as

$$P_{\mathbf{p}_\tau}(\mathbf{p}|\boldsymbol{\rho}) = \frac{1}{\sqrt{2\pi}\sigma_{f_\tau}} e^{-\frac{(f-\overline{f}_\tau)^2}{2\sigma_{f_\tau}^2}} \frac{1}{\sqrt{2\pi}\sigma_{h_\tau}} e^{-\frac{(h-\overline{h}_\tau)^2}{2\sigma_{h_\tau}^2}}. \tag{79}$$

With cycle index k, we get conditional probability of matured collectives $P_{\mathbf{p}_{k+1,\tau}}(\mathbf{p}|\mathbf{p}_k^*)$ by

$$P_{\mathbf{p}_{k+1,\tau}}(\mathbf{p}|\mathbf{p}_k^*) = \int_0^1 d\zeta \int_0^1 d\eta P_{\mathbf{p}_{k+1,\tau}}(\mathbf{p}|\zeta,\eta) P_{\mathbf{p}_{k+1,0}}(\zeta,\eta|\mathbf{p}_k^*). \tag{80}$$

We select the Adult collective among g Adult collectives such that the change in frequencies during maturation could be compensated. During maturation, a frequency distribution moves in different directions in (f, h) space depending on the initial composition (f_k^*, h_k^*) So, we take different directions to obtain the extreme value distributions. Considering only the sign of the frequency changes in f and h, we take either maximum or minimum. The mean change in h is always positive in the whole (f, h) space since $d\overline{FF}/dt$ is always positive in Eq. (67). Thus, we choose the minimum value $h_{k+1}^{\min}$ in every selection step.

If the mean $\overline{f^{(i)}_{k+1,\tau}} = \int_0^1 df' \int_0^1 dh'\, f P_{\mathbf{p}_{k+1,\tau}}(f', h'|\mathbf{p}^*_k)$ is larger than f^*_k, the minimum value among $f^{(1)}_{k+1,\tau}, f^{(2)}_{k+1,\tau}, \cdots, f^{(g)}_{k+1,\tau}$ will be chosen in selection step to compensate the frequency change in maturation step. Let us denote the selected valued of f and h as$f^*_{k+1} = \min(f^{(1)}_{k+1,\tau}, f^{(2)}_{k+1,\tau}, \cdots, f^{(g)}_{k+1,\tau})$ and $h^*_{k+1} = \min(h^{(1)}_{k+1,\tau}, h^{(2)}_{k+1,\tau}, \cdots, h^{(g)}_{k+1,\tau})$. We temporarily drop time index τ for simplicity. Then, the joint cumulative distribution function $C_{\mathbf{p}^*_{k+1}}(f, h|\boldsymbol{p}^*_k) = \Pr[f^*_{k+1} < f \wedge h^*_{k+1} < h|\boldsymbol{p}^*_k]$ is

$$
\begin{aligned}
C_{\mathbf{p}^*_{k+1}}(f, h|\boldsymbol{p}^*_k) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) \\
&= \left(\int_0^1 - \int_f^1\right) df' \left(\int_0^1 - \int_h^1\right) dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) \\
&= \int_0^1 df' \int_0^1 dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) - \int_h^1 df' \int_0^1 dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) \\
&\quad - \int_0^1 df' \int_h^1 dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) + \int_f^1 df' \int_h^1 dh' P_{\mathbf{p}^*_{k+1}}(f', h'|\boldsymbol{p}^*_k) \\
&= 1 - \Pr[f^*_{k+1} \geq f|\boldsymbol{p}^*_k] - \Pr[h^*_{k+1} \geq h|\boldsymbol{p}^*_k] + \Pr[f^*_{k+1} \geq f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k]. \quad (81)
\end{aligned}
$$

The probability $\Pr[f^*_{k+1} \geq f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k]$ can be converted as

$$
\begin{aligned}
&\Pr[f^*_{k+1} \geq f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k] \\
&\quad = \Pr[f^{(1)}_{k+1} \geq f \wedge h^{(1)}_{k+1} \geq h \wedge f^{(2)}_{k+1} \geq f \wedge h^{(2)}_{k+1} \geq h \wedge \cdots \wedge f^{(g)}_{k+1} \geq f \wedge h^{(g)}_{k+1} \geq h|\boldsymbol{p}^*_k] \\
&\quad = [\Pr[f_{k+1} \geq f \wedge h_{k+1} \geq h|\, \boldsymbol{p}^*_k]]^g \\
&\quad = [1 - \Pr[f_{k+1} < f|\boldsymbol{p}^*_k] - \Pr[h_{k+1} < h|\boldsymbol{p}^*_k] + \Pr[f_{k+1} < f \wedge h_{k+1} < h|\, \boldsymbol{p}^*_k]]^g \\
&\quad = \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_k) + C_{\mathbf{p}_{k+1}}(f, h|\boldsymbol{p}^*_k)\right]^g, \quad (82)
\end{aligned}
$$

where $C_{\mathbf{p}_{k+1}}(f, h|\boldsymbol{p}^*_k) = \int_0^f df' \int_0^h dh' P_{\mathbf{p}_{k+1}}(f', h'|\boldsymbol{p}^*_k)$ is a conditional joint cumulative distribution function of $(f^{(i)}, h^{(i)})$. The marginal cumulative distribution functions are

$$
C_{f_{k+1}}(f|\boldsymbol{p}^*_k) = \int_0^f df' \int_0^1 dh' P_{\mathbf{p}_{k+1}}(f', h'|\boldsymbol{p}^*_k), \quad (83)
$$

$$
C_{h_{k+1}}(h|\boldsymbol{p}^*_k) = \int_0^1 df' \int_0^h dh' P_{\mathbf{p}_{k+1}}(f', h'|\boldsymbol{p}^*_k).ype \quad (84)
$$

Similarly, the probabilities $\Pr[f^*_{k+1} \geq f|\boldsymbol{p}^*_k]$ and $\Pr[h^*_{k+1} \geq h|\boldsymbol{p}^*_k]$ are converted into $\Pr[f^*_{k+1} \geq f|\boldsymbol{p}^*_k] = \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k)\right]^g$ and $\Pr[h^*_{k+1} \geq h|\boldsymbol{p}^*_k] = \left[1 - C_{h_{k+1}}(h|\boldsymbol{p}^*_k)\right]^g$. Thus, the joint cumulative distribution function is

$$
\begin{aligned}
C_{\mathbf{p}^*_{k+1}}(f, h|\boldsymbol{p}^*_k) = 1 &- \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k)\right]^g - \left[1 - C_{h_{k+1}}(h|\boldsymbol{p}^*_k)\right]^g \quad (85) \\
&+ \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_k) + C_{\mathbf{p}_{k+1}}(f, h|\boldsymbol{p}^*_k)\right]^g.
\end{aligned}
$$

Then, the conditional probability of the selected collective is given by

$$\begin{aligned}
P_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) \\
&= g(g-1)\left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-2} \\
&\quad \times (P_{f_{k+1}}(f|\boldsymbol{p}^*_{k-1}) - \partial_f C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k))(P_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) - \partial_h C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)) \\
&\quad + g\left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-1} P_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k),
\end{aligned} \tag{86}$$

where $\partial_f C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) = \frac{\partial}{\partial f} C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) = \int_0^h dh' P_{\mathbf{p}_{k+1}}(f,h'|\boldsymbol{p}^*_k)$ and $\partial_h C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) = \frac{\partial}{\partial h} C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) = \int_0^f df' P_{\mathbf{p}_{k+1}}(f',h|\boldsymbol{p}^*_k)$.

If the mean $\overline{f_{k+1,\tau}}$ is smaller than f^*_k , the chosen collective is likely to have maximum f values among g matured collectives. Then, the definition of f^* is written by $f^*_{k+1} = \max(f^{(1)}_{k+1}, f^{(2)}_{k+1}, \cdots, f^{(g)}_{k+1})$. We rewrite the joint cumulative distribution function $C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)$ to be little different from Eq. (81) because now we have to utilize the condition $f^* < f$ instead of $f^* > f$,

$$\begin{aligned}
C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&= \int_0^f df' (\int_0^1 - \int_h^1) dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&= \int_0^f df' \int_0^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) - \int_0^f df' \int_h^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k). \\
&= \Pr[f^*_{k+1} < f|\boldsymbol{p}^*_k] - \Pr[f^*_{k+1} < f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k].
\end{aligned} \tag{87}$$

The probability $\Pr[f^*_{k+1} < f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k]$ is converted as

$$\begin{aligned}
&\Pr[f^*_{k+1} < f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k] \\
&= \Pr[f^{(1)}_{k+1} < f \wedge h^{(1)}_{k+1} \geq h \wedge f^{(2)}_{k+1} < f \wedge h^{(2)}_{k+1} \geq h \wedge \cdots \wedge f^{(2)}_{k+1} < f \wedge h^{(2)}_{k+1} \geq h|\boldsymbol{p}^*_k] \qquad (88) \\
&= [\Pr[f_{k+1} < f \wedge h_{k+1} \geq h|\boldsymbol{p}^*_k]]^g \\
&= [\Pr[f_{k+1} < f|\boldsymbol{p}^*_k] - \Pr[f_{k+1} < f \wedge h_{k+1} < h|\boldsymbol{p}^*_k]]^g \\
&= \left[C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^g. \qquad (89)
\end{aligned}$$

Thus, the joint cumulative distribution function is given by

$$C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) = \left[C_{f_{k+1}}(f|\boldsymbol{p}^*_k)\right]^g - \left[C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^g. \tag{90}$$

In this case, the conditional probability distribution function is given by

$$\begin{aligned}
P_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) \\
&= g(g-1)\left[C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-2} \\
&\quad \times (P_{f_{k+1}}(f|\boldsymbol{p}^*_k) - \partial_f C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)) \partial_h C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k) \\
&\quad + g\left[C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-1} P_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k).
\end{aligned} \tag{91}$$

By replacing (f,h) to $\mathbf{p}^*_{k+1}$, we finally obtain the conditional probability distribution $\Psi(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)$,

$$\Psi(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k) = \begin{cases} g(g-1)\left[1 - C_{f_{k+1}}(f^*_{k+1}|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h^*_{k+1}|\mathbf{p}^*_k) + C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)\right]^{g-2} \\ \times(P_{f_{k+1}}(f^*_{k+1}|\mathbf{p}^*_k) - \partial_f C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k))(P_{h_{k+1}}(h) - \partial_h C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)) \\ +g\left[1 - C_{f_{k+1}}(f^*_{k+1}|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h^*_{k+1}|\mathbf{p}^*_k) + C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)\right]^{g-1} P_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k) \\ \qquad\qquad \text{for } \overline{f_{k+1,\tau}} - f^*_k \geq 0 \text{ and } \overline{h_{k+1,\tau}} - h^*_k \geq 0 \\ g(g-1)\left[C_{f_{k+1}}(f^*_{k+1}|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\boldsymbol{p}^*_k)\right]^{g-2} \\ (P_{f_{k+1}}(f^*_{k+1}|\mathbf{p}^*_k) - \partial_f C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k))\partial_h C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k) \\ +g\left[C_{f_{k+1}}(f^*_{k+1}|\boldsymbol{p}^*_k) - C_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)\right]^{g-1} P_{\mathbf{p}_{k+1}}(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k) \\ \qquad\qquad \text{for } \overline{f_{k+1,\tau}} - f^*_k < 0 \text{ and } \overline{h_{k+1,\tau}} - h^*_k \geq 0 \end{cases}. \tag{92}$$

Using Eq. (92), we get the mean values of f^* and h^* as

$$\overline{f^*_{k+1}} = \int_0^1 df' \int_0^1 dh' \; f' \times P_{f^*_{k+1},h^*_{k+1}}(f',h'|f^*_k,h^*_k), \tag{93}$$

$$\overline{h^*_{k+1}} = \int_0^1 df' \int_0^1 dh' \; h' \times P_{f^*_{k+1},h^*_{k+1}}(f',h'|f^*_k,h^*_k). \tag{94}$$

We define the accessible region in frequency space where the signs of the changes in both F frequency and FF frequency after a cycle are opposite to that of maturation (see Fig. 12),

$$\text{sign}\left(\overline{f^*_{k+1}} - f^*_k\right) \times \text{sign}\left(\overline{f_{k+1,\tau}} - f^*_k\right) \leq 0 \text{ and } \text{sign}\left(\overline{h^*_{k+1}} - h^*_k\right) \times \text{sign}\left(\overline{h_{k+1,\tau}} - h^*_k\right) \leq 0, \tag{95}$$

where $\overline{f_{k+1,\tau}}$ and $\overline{f^*_{k+1}}$ are the mean values of F frequencies after the maturation step in cycle $k+1$ before and after selection, respectively, and h values are defined similarly for FF. Or, if the condition is not met, the composition of the selected collective may diverge from the target composition after several cycles. The accessible regions are marked in the gold-colored area in Fig. 12**b**. Similar to the two-population case, the accessible region is shaped by the flow velocity of the composition during the maturation step, as depicted in the flow diagram in Fig. 12**a**. Both F and FF frequencies tend to increase, and the inter-collective selection can compensate for these changes if the composition changes slowly when the F and FF frequencies are small. However, if the changes occur too rapidly when the FF frequency is intermediate, the frequency cannot be stabilized. So the accessible region is limited to the regions where the composition changes slowly.

This is explainable by projecting the three-population problem into the two-population problem. The selective advantage of FF relative to the rest of collective mainly determines the accessible region. The growth rate of the rest varies from r_S to $r_S + w$ according to F frequency, so the mean growth rate of the rest is written by $\bar{r_S} = r_S + f'\omega$ where f' is F frequency in S+F. Then, the corresponding selective advantage of FF is $\bar{\omega} = (2-f')\omega$ which varies between ω to 2ω. Using $\bar{r_S}$ and $\bar{\omega}$ similar to Sec. II, we get bounds of accessible region (see dashed-line in Fig. 12**b**). The boundary from projected problem agreed well with the original three-population problem.

IX. DERIVATION OF EQUATIONS

In this section, we go over the derivation of Eq. (10)-(34) for readers not equipped with advanced mathematics training.

Assumptions: $\mu \ll \omega, r_S$

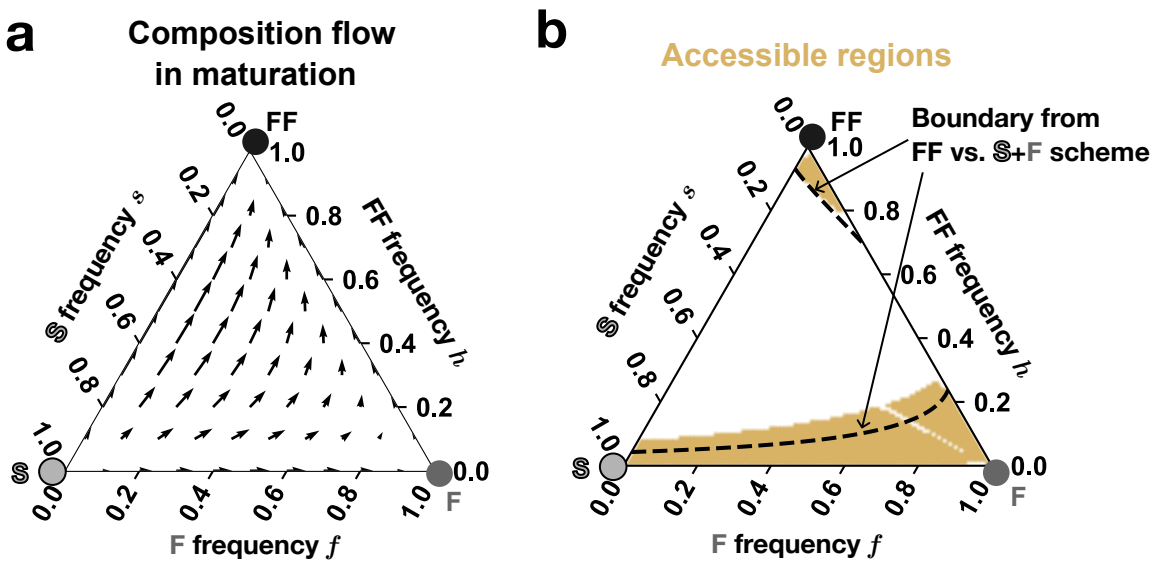

FIG. 12. **a** The flow of composition change in F and FF frequencies at each composition (f, h). Top corner indicates that FF cells fix in the collective. Right bottom corner means collectives with only F cells while collectives contain S cells only at left bottom corner. Arrow length means the speed of change. **b** The accessible regions are marked by the gold area. If the signs of changes in both F frequency and FF frequency after inter-collective selection are opposite to those during maturation, then the given composition is accessible. Otherwise, the composition is not accessible and will change after cycles. Dashed lines are the boundary of accessible region by projecting the collective into a two-population problem (FF vs. S+F). The figures are drawn using *mpltern* package [7].

Equations (10) and (11)

Equation (10) is straightforwardly solved by integrating Eq. (8). Equation (11) is obtained from Eq. (9) using integration factor $e^{-(r_S+\omega)t}$:

$$\left[\frac{d\overline{F}(t)}{dt} - (r_S+\omega)\overline{F}(t)\right] e^{-(r_S+\omega)t} = \mu\overline{S}_o e^{(r_S-\mu)t} e^{-(r_S+\omega)t}, \tag{96}$$

$$\frac{d\left[\overline{F}(t)e^{-(r_S+\omega)t}\right]}{dt} = \mu\overline{S}_o e^{-(\omega+\mu)t}. \tag{97}$$

Integrating both sides, we get $\overline{F}(t)e^{-(r_S+\omega)t} - \overline{F}(0) = \frac{\mu\overline{S}_o}{\omega+\mu}\left(1 - e^{-(\omega+\mu)t}\right)$. Thus,

$$\overline{F}(t) = \overline{F}_o e^{(r_S+\omega)t} + \frac{\mu\overline{S}_o}{\omega+\mu}\left(e^{(r_S+\omega)t} - e^{(r_S-\mu)t}\right) \approx \overline{F}_o e^{(r_S+\omega)t} + \frac{\mu\overline{S}_o}{\omega}\left(e^{(r_S+\omega)t} - e^{r_S t}\right). \tag{98}$$

Equations (16) and (17)

Applying Eq. (4), we have

$$\begin{aligned}\frac{d\overline{S^2}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S^2 \frac{dP(S,F,t)}{dt} \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[r_S S^2(S-1)P(S-1,F,t) + S^2(r_S+\omega)(F-1)P(S,F-1,t) \\ &\quad +\mu(\mathrm{S}+1)S^2P(S+1,F-1,t) - (rS + (r_S+\omega)F + \mu S)S^2P(S,F,t)].\end{aligned} \tag{99}$$

We collect the two violet-colored terms and change the order of summation. Note that the first violet-colored term does not change regardless of whether S starts from 0 or 1 because the term is zero for $S=0$. Thus, the first violet-colored term is equivalent to $\sum_{F=0}^{\infty}\sum_{S=1}^{\infty}\left[r_S S^2(S-1)P(S-1,F,t)\right]$. Let $\alpha = S-1$, and this becomes $\sum_{F=0}^{\infty}\sum_{\alpha=0}^{\infty}\left[r_S(\alpha+1)^2\alpha P(\alpha,F,t)\right]$. We reassign α as S, and obtain:

$$\begin{aligned}&\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[r_S(S+1)^2 S\, P(S,F,t)\right] - \sum_{F=0}^{\infty}\sum_{S=0}^{\infty} r_S S^3 P(S,F,t) \\ &= r_S \sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[(2S+1)S\right]P(S,F,t) \\ &= r_S(2\overline{S^2}+\overline{S}).\end{aligned} \tag{100}$$

We collect the two green terms, and similarly obtain:

$$\begin{aligned}&\sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[S^2(r_S+\omega)(F-1)P(S,F-1,t) - (r_S+\omega)FS^2P(S,F,t)\right] \\ &= (r_S+\omega)\sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[S^2F\,P(S,F,t) - FS^2P(S,F,t)\right] \\ &= 0.\end{aligned} \tag{101}$$

Finally, we collect the two orange terms. For the first orange term, the sum is the same regardless of whether we start from $S=0$ or -1. Let S start from -1, and we have

$$\sum_{F=0}^{\infty}\sum_{S=-1}^{\infty}\left[\mu(S+1)S^2P(S+1,F-1,t)\right]. \tag{102}$$

Let $\alpha = S+1$,then the term becomes $\sum_{F=0}^{\infty}\sum_{\alpha=0}^{\infty}\left[\mu\alpha(\alpha-1)^2P(\alpha,F-1,t)\right]$. We reassign α as S, and additionally apply index change on $F-1$:

$$\begin{aligned}&\mu\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[S(S-1)^2P(S,F,t) - S^3P(S,F,t)\right] \\ &= \mu\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[(-2S^2+S)P(S,F,t)\right] \\ &= \mu\left(-2\overline{S^2}+\overline{S}\right).\end{aligned} \tag{103}$$

Now, add the three parts together, and we have

$$\begin{aligned}\frac{d\overline{S^2}}{dt} &= \mu\left(-2\overline{S^2}+\overline{S}\right)+r_S(2\overline{S^2}+\overline{S})\\ &= 2\overline{S^2}(r_S-\mu)+\overline{S}(r_S+\mu),\end{aligned} \tag{104}$$

which is Eq. (16). Likewise,

$$\begin{aligned}\frac{d\overline{F^2}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}F^2\frac{dP(S,F,t)}{dt}\\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[r_SF^2(S-1)P(S-1,F,t)+F^2(r_S+\omega)(F-1)P(S,F-1,t)\\ &\quad +\mu(S+1)F^2P(S+1,F-1,t)-(r_SS+(r_S+\omega)F+\mu S)F^2P(S,F,t)]\\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[r_SF^2S+(F+1)^2(r_S+\omega)F+\mu S(F+1)^2-(r_SS+(r_S+\omega)F+\mu S)F^2\right]P(S,F,t)\\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[(F+1)^2(r_S+\omega)F+\mu S(F+1)^2-((r_S+\omega)F+\mu S)F^2\right]P(S,F,t)\\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[(r_S+\omega)F(2F+1)+\mu S(2F+1)\right]P(S,F,t) \qquad (105)\\ &= 2(r_S+\omega)\overline{F^2}+(r_S+\omega)\overline{F}+2\mu\overline{SF}+\mu\overline{S} \qquad (106)\end{aligned}$$

which is Eq. (17).

Equation (18) and (20)

Using integration factor $e^{-2(r_S-\mu)t}$ and Eq. (16), we have:

$$\left[\frac{d\overline{S^2}}{dt}-2(r_S-\mu)\overline{S^2}\right]e^{-2(r_S-\mu)t}=(r_S+\mu)\overline{S}_oe^{(r_S-\mu)t}e^{-2(r_S-\mu)t}, \tag{107}$$

$$\frac{d\left[\overline{S^2}e^{-2(r_S-\mu)t}\right]}{dt}=(r_S+\mu)\overline{S}_oe^{(\mu-r_S)t}, \tag{108}$$

$$\overline{S^2}e^{-2(r_S-\mu)t}=\frac{r_S+\mu}{\mu-r_S}\overline{S}_o\left[e^{(\mu-r_S)t}-1\right]+\overline{S^2}_o. \tag{109}$$

Since $\mu \ll r_S$, we have

$$\begin{aligned}\overline{S^2} &= \frac{r_S+\mu}{\mu-r_S}\overline{S}_o\left[e^{(\mu-r_S)t}-1\right]e^{2(r_S-\mu)t}+\overline{S^2}_oe^{2(r_S-\mu)t}\\ &\approx \overline{S}_oe^{2r_St}\left[1-e^{-r_St}\right]+\overline{S^2}_oe^{2r_St},\end{aligned} \tag{110}$$

where $\overline{S^2}_o$ is the expected S^2 at time 0. For Eq. (20),

$$\begin{aligned}\sigma_S^2(t) &= \overline{S^2}(t) - [\overline{S}(t)]^2 \\ &\approx \overline{S}_o e^{2r_S t}\left[1 - e^{-r_S t}\right] + \overline{S^2}_o e^{2r_S t} - \left(\overline{S}_o e^{r_S t}\right)^2 \\ &= \overline{S}_o e^{2r_S t}\left[1 - e^{-r_S t}\right] + e^{2r_S t}\left(\overline{S^2}_o - \left(\overline{S}_o\right)^2\right) = \overline{S}_o e^{2r_S t}\left[1 - e^{-r_S t}\right] + e^{2r_S t}\sigma_S^2(0). \quad (111)\end{aligned}$$

Equation (22) and (23)

Since $\overline{S\,F} = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S\,F\,P(S,F,t)$,

$$\begin{aligned}\frac{d\overline{S\,F}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[S\,F r_S(S-1)P(S-1,F,t) + S\,F(r_S+\omega)(F-1)P(S,F-1,t) \\ &\quad + S\,F\mu(S+1)P(S+1,F-1,t) \\ &\quad - S\,F(r_S S + (r_S+\omega)F + \mu S)P(S,F,t)] \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[(S+1)F r_S S + S(F+1)(r_S+\omega)F + (S-1)(F+1)\mu S \\ &\quad - S\,F(r_S S + (r_S+\omega)F + \mu S)]P(S,F,t) \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[2r_S S\,F + \omega S\,F - \mu S\,F + \mu S^2 - \mu S)\right]P(S,F,t) \\ &= (2r_S + \omega - \mu)\overline{S\,F} + \mu(\overline{S^2} - \overline{S}). \quad (112)\end{aligned}$$

We can solve this, again using the integration factor technique above:

$$\frac{d[\overline{S\,F}e^{-(2r_S+\omega-\mu)t}]}{dt} = \mu(\overline{S^2} - \overline{S})e^{-(2r_S+\omega-\mu)t}. \quad (113)$$

Thus, we have

$$\begin{aligned}
\overline{S\,F}e^{-(2r_S+\omega-\mu)t} - \overline{S\,F}_o &= \mu\int(\overline{S^2}-\overline{S})e^{-(2r_S+\omega-\mu)t}dt \\
&= \mu\int\left(\frac{r_S+\mu}{\mu-r_S}\overline{S}_o\left[e^{(\mu-r_S)t}-1\right]e^{2(r_S-\mu)t}\right. \\
&\quad \left.+\overline{S^2}_o e^{2(r_S-\mu)t} - \overline{S}_o e^{(r_S-\mu)t}\right)e^{-(2r_S+\omega-\mu)t}dt \\
&= \mu\int\left(\frac{r_S+\mu}{\mu-r_S}\overline{S}_o\left[e^{-(\mu-r_S)t}-e^{2(r_S-\mu)t}\right]e^{-(2r_S+\omega-\mu)t}\right. \\
&\quad \left.+\overline{S^2}_o e^{2(r_S-\mu)t}e^{-(2r_S+\omega-\mu)t} - \overline{S}_o e^{(r_S-\mu)t-(2r_S+\omega-\mu)t}\right)dt \\
&= \mu\int\left(\frac{r_S+\mu}{\mu-r_S}\overline{S}_o e^{-(r_S+\omega)t} - \frac{r_S+\mu}{\mu-r_S}\overline{S}_o e^{-(\mu+\omega)t}\right. \\
&\quad \left.+\overline{S^2}_o e^{-(\mu+\omega)t} - \overline{S}_o e^{-(r_S+\omega)t}\right)dt \\
&= \mu\left\{\frac{2r_S}{\mu-r_S}\overline{S}_o\frac{e^{-(r_S+s)t}-1}{-(r_S+\omega)}\right. \\
&\quad \left.+\left(\overline{S^2}_o - \frac{r_S+\mu}{\mu-r_S}\overline{S}_o\right)\frac{e^{-(\mu+\omega)t}-1}{-(\mu+\omega)}\right\},
\end{aligned} \tag{114}$$

which results in

$$\begin{aligned}
\overline{S\,F}(t) &= \overline{S\,F}_o e^{(2r_S+\omega-\mu)t} + \mu\left\{\frac{2r_S}{\mu-r_S}\overline{S}_o\frac{e^{-(r_S+\omega)t}-1}{-(r_S+\omega)} + (\overline{S^2}_o - \frac{r_S+\mu}{\mu-r_S}\overline{S}_o)\frac{e^{-(\mu+\omega)t}-1}{-(\mu+\omega)}\right\}e^{(2r_S+\omega-\mu)t} \\
&= \overline{S\,F}_o e^{(2r_S+\omega-\mu)t} + \frac{2r_S\mu}{\mu-r_S}\overline{S}_o\frac{e^{(r_S-\mu)t}-e^{(2r_S+\omega-\mu)t}}{-(r_S+\omega)} \\
&\quad +\mu\left(\overline{S^2}_o - \frac{r_S+\mu}{\mu-r_S}\overline{S}_o\right)\frac{e^{2(r_S-\mu)t}-e^{(2r_S+\omega-\mu)t}}{-(\mu+\omega)} \\
&\approx \overline{S\,F}_o e^{(2r_S+\omega)t} - \frac{2\mu\overline{S}_o}{r_S+\omega}\left(e^{(2r_S+\omega)t}-e^{r_S t}\right) + \frac{\mu}{\omega}(\overline{S^2}_o+\overline{S}_o)\left(e^{(2r_S+\omega)t}-e^{2r_S t}\right).
\end{aligned} \tag{115}$$

Equation (24)

From Eq. (17), we have

$$\left(\frac{d\overline{F^2}}{dt} - 2(r_S+\omega)\overline{F^2}\right)e^{-2(r_S+\omega)t} = \frac{d\left(\overline{F^2}e^{-2(r_S+\omega)t}\right)}{dt} = \left((r_S+\omega)\overline{F} + 2\mu\overline{S\,F} + \mu\overline{S}\right)e^{-2(r_S+\omega)t}. \tag{116}$$

The right-hand side becomes

$$
\begin{aligned}
\approx&(r_S+\omega)\left(\overline{F}_o e^{(r_S+\omega)t}+\frac{\mu\overline{S}_o}{\omega+\mu}\left(e^{(r_S+\omega)t}-e^{(r_S-\mu)t}\right)\right)e^{-2(r_S+\omega)t}\\
&+2\mu\overline{S\,F}_o e^{-(\omega+\mu)t}+\mu\overline{S}_o e^{-(r_S+2\omega+\mu)t}+o(\mu^2)\\
=&(r_S+\omega)\left(\overline{F}_o e^{-(r_S+\omega)t}+\frac{\mu\overline{S}_o}{\omega+\mu}\left(e^{-(r_S+\omega)t}-e^{-(r_S+2\omega+\mu)t}\right)\right)\\
&+2\mu\overline{S\,F}_o e^{-(\omega+\mu)t}+\mu\overline{S}_o e^{-(r_S+2\omega+\mu)t}+o(\mu^2)\\
=&(r_S+\omega)\left(\frac{\overline{F}_o\omega+N_0\mu}{\omega+\mu}e^{-(r_S+\omega)t}-\frac{\mu\overline{S}_o}{\omega+\mu}e^{-(r_S+2\omega+\mu)t}\right)\\
&+2\mu\overline{S\,F}_o e^{-(\omega+\mu)t}+\mu\overline{S}_o e^{-(r_S+2\omega+\mu)t}+o(\mu^2).
\end{aligned}
\tag{117}
$$

Note that we have checked that the second and third terms of $\overline{SF}$ can be ignored after we compare the full calculation with this simpler version. Integrate both sides:

$$
\begin{aligned}
\overline{F^2}e^{-2(r_S+\omega)t}-\overline{F^2}_o\approx&(r_S+\omega)\left(\frac{\overline{F}_o\omega+N_0\mu}{(\omega+\mu)(r_S+\omega)}(1-e^{-(r_S+\omega)t})\right.\\
&\left.+\frac{\mu\overline{S}_o}{(\omega+\mu)(r_S+2\omega+\mu)}(e^{-(r_S+2\omega+\mu)t}-1)\right)\\
&+\frac{2\mu\overline{S\,F}_o}{\omega+\mu}(1-e^{-(\omega+\mu)t})+\frac{\mu\overline{S}_o}{r_S+2\omega+\mu}(1-e^{-(r_S+2\omega+\mu)t})+o(\mu^2)
\end{aligned}
\tag{118}
$$

Then, we have

$$
\begin{aligned}
\overline{F^2}(t) \approx\ & \overline{F^2}_o e^{2(r_S+\omega)t} + (r_S+\omega)\Bigg(\frac{\overline{F}_o\omega + N_0\mu}{(\omega+\mu)(r_S+\omega)}(e^{2(r_S+\omega)t} - e^{(r_S+\omega)t}) \\
& + \frac{\mu\overline{S}_o}{(\omega+\mu)(r_S+2\omega+\mu)}(e^{(r_S-\mu)t} - e^{2(r_S+\omega)t})\Bigg) \\
& + \frac{2\mu\overline{S\,F}_o}{\omega+\mu}(e^{2(r_S+\omega)t} - e^{(2r_S+\omega-\mu)t}) + \frac{\mu\overline{S}_o}{r_S+2\omega+\mu}(e^{2(r_S+\omega)t} - e^{(r_S-\mu)t}) + o(\mu^2) \\
=\ & \overline{F^2}_o e^{2(r_S+\omega)t} + \left(\frac{\overline{F}_o\omega + N_0\mu}{\omega}(e^{2(r_S+\omega)t} - e^{(r_S+\omega)t}) + \left(\frac{1}{\omega} - \frac{1}{r_S+2\omega}\right)\mu\overline{S}_o(e^{r_S t} - e^{2(r_S+\omega)t})\right) \\
& + \frac{2\mu\overline{S\,F}_o}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \frac{\mu\overline{S}_o}{r_S+2\omega}(e^{2(r_S+\omega)t} - e^{r_S t}) + o(\mu^2) \\
=\ & \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{S\,F}_o}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
& + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{\overline{F}_o}{\overline{S}_o}(e^{2(r_S+\omega)t} - e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2) \\
=\ & \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{S\,F}_o}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
& + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + \frac{\mu\overline{F}_o}{\omega}(e^{2(r_S+\omega)t} - e^{(r_S+\omega)t}) + o(\mu^2) \\
=\ & \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{S\,F}_o}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o(1+\frac{\mu}{\omega})e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
& + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2) \\
=\ & \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{S\,F}_o}{\omega}e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
& + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2).
\end{aligned}
\tag{119}
$$

Equation (25)

$$
\begin{aligned}
\sigma_F^2(t) &= \overline{F^2}(t) - [\overline{F}(t)]^2 \\
&\approx \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{S\,F}_o}{\omega} e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2) \\
&\quad - \left[\overline{F}_o e^{(r_S+\omega)t} + \frac{\mu\overline{S}_o}{\omega}(e^{(r_S+\omega)t} - e^{r_S t})\right]^2 \\
&= \overline{F^2}_o e^{2(r_S+\omega)t} + \frac{2\mu\overline{SF}_o}{\omega} e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{\omega}\left[(-e^{(r_S+\omega)t}) + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2) \\
&\quad - \overline{F}_o^2 e^{2(r_S+\omega)t} - 2\overline{F}_o e^{(r_S+\omega)t}\frac{\mu\overline{S}_o}{\omega}(e^{(r_S+\omega)t} - e^{r_S t}) - \left[\frac{\mu\overline{S}_o}{\omega}(e^{(r_S+\omega)t} - e^{r_S t})\right]^2 \\
&= \sigma_{F,0}^2 e^{2(r_S+\omega)t} + \frac{2\mu(\overline{S\,F}_o - \overline{S}_o\overline{F}_o)}{\omega} e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - e^{r_S t}) \qquad (120) \\
&\quad + \overline{F}_o e^{(r_S+\omega)t}(e^{(r_S+\omega)t} - 1) + \frac{\mu\overline{S}_o}{\omega}\left[-e^{(r_S+\omega)t} + \frac{2\omega}{r_S+2\omega}e^{2(r_S+\omega)t} + \frac{r_S}{r_S+2\omega}e^{r_S t}\right] + o(\mu^2).
\end{aligned}
$$

Equations (32)-(34)

To derive this equation, we use the fact that $1/(1+x) \sim 1 - x$ for small x. We will omit (t) for simplicity. Also note that we are considering relatively large populations so that the standard deviation is much smaller than the mean.

$$
\begin{aligned}
f &\approx \frac{\overline{F} + \sigma_F G_F}{\overline{S} + \sigma_S G_S + \overline{F} + \sigma_F G_F} \\
&= \frac{\overline{F}\left(1 + \frac{\sigma_F}{\overline{F}} G_F\right)}{(\overline{S} + \overline{F})(1 + \frac{\sigma_S G_S + \sigma_F G_F}{\overline{S} + \overline{F}})} \\
&\approx \frac{\overline{F}}{\overline{S} + \overline{F}}\left(1 + \frac{\sigma_F}{\overline{F}} G_F\right)\left(1 - \frac{\sigma_S G_S + \sigma_F G_F}{\overline{S} + \overline{F}}\right) \\
&\approx \overline{f}\left(1 + \frac{\sigma_F}{\overline{F}} G_F - \frac{\sigma_S G_S + \sigma_F G_F}{\overline{S} + \overline{F}}\right) \\
&= \overline{f}\left(1 + \left(\frac{\sigma_F}{\overline{F}} - \frac{\sigma_F}{\overline{S} + \overline{F}}\right) G_F - \frac{\sigma_S}{\overline{S} + \overline{F}} G_S\right). \qquad (121)
\end{aligned}
$$

Recall that if $A \sim \mathcal{N}(\mu_A, \sigma_A)$, $B \sim \mathcal{N}(\mu_B, \sigma_B)$, then $A - B \sim \mathcal{N}(\mu_A - \mu_B, \sqrt{\sigma_A^2 + \sigma_B^2})$. Thus, f is distributed as a Gaussian with the mean of

$$\begin{aligned}
\overline{f}(t) &= \frac{\overline{F}(t)}{\overline{S}(t) + \overline{F}(t)} \approx \frac{\overline{F}_o e^{(r_S+\omega)t} + \frac{\mu \overline{S}_o}{\omega}(e^{(r_S+\omega)t} - e^{r_S t})}{\overline{S}_o e^{r_S t} + \overline{F}_o e^{(r_S+\omega)t} + \frac{\mu \overline{S}_o}{\omega}(e^{(r_S+\omega)t} - e^{r_S t})} \\
&= \frac{\overline{F}_o e^{\omega t} + \frac{\mu \overline{S}_o}{\omega}(e^{\omega t} - 1)}{\overline{S}_o + \overline{F}_o e^{\omega t} + \frac{\mu \overline{S}_o}{\omega}(e^{\omega t} - 1)} \\
\omega &= \frac{\overline{f}_o e^{\omega t} + \frac{\mu}{\omega}(1 - \overline{f}_o)(e^{\omega t} - 1)}{(1 - \overline{f}_o) + \overline{f}_o e^{\omega t} + \frac{\mu}{\omega}(1 - \overline{f}_o)(e^{\omega t} - 1)}. \qquad (122)
\end{aligned}$$

Note that the initial value of mean $\overline{f}(0)$ is equal to the mean of the binomial distribution Eq. (29), f_k^*. The variance is

$$\begin{aligned}
\sigma_f^2(t) &= \overline{f}^2 \left(\left(\frac{\overline{S}}{(\overline{S} + \overline{F})\overline{F}} \right)^2 \sigma_F^2(t) + \left(\frac{1}{\overline{S} + \overline{F}} \right)^2 \sigma_S^2(t) \right) \\
&= \overline{f}^2 \left(\frac{(1 - \overline{f}(t))^2}{\overline{N}(t)^2 (\overline{f}(t))^2} \sigma_F(t) + \frac{1}{\overline{N}(t)^2} \sigma_S^2(t) \right) \\
&= \frac{(1 - \overline{f}(t))^2 \sigma_F^2(t) + \overline{f}(t)^2 \sigma_S^2(t)}{\overline{N}(t)^2}. \qquad (123)
\end{aligned}$$

---

[1] Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics* **115**, 1716–1733 (2001).

[2] Cao, Y., Gillespie, D. T. & Petzold, L. R. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics* **124**, 044109 (2006).

[3] Zheng, Q. Progress of a half century in the study of the luria–delbrück distribution. *Mathematical Biosciences* **162**, 1–32 (1999).

[4] Gumbel, E. J. *Statistics of extremes* (Columbia university press, 1958).

[5] Philip, J. The function inverfc theta. *Australian Journal of Physics* **13**, 13 (1960).

[6] Xie, L., Yuan, A. E. & Shou, W. Simulations reveal challenges to artificial community selection and possible strategies for success. *PLoS Biology* **17**, e3000295 (2019).

[7] Ikeda, Y., Grabowski, B. & Körmann, F. mpltern 0.3.0: ternary plots as projections of Matplotlib (2019). Available at https://doi.org/10.5281/zenodo.3528355.