Artificial selection for microbial collective composition can succeed or fail depending on the initial and target values

Juhee Lee[a,b], **Wenying Shou**[c,1], **and Hye Jin Park**[a,b,1]

[a]Department of Physics, Inha University, Incheon 22212, Republic of Korea; [b]Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea; [c]Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom

This manuscript was compiled on December 28, 2023

Microbial collectives, capable of functions beyond the reach of individual populations, can be enhanced through artificial selection. However, this process presents unique challenges. Here, we explore the 'waterfall' phenomenon, a metaphor describing how the success in achieving a desired genotype or species composition in microbial collectives can depend on both the target characteristics and initial conditions. We focus on collectives comprising fast-growing (F) and slow-growing (S) types, aiming to achieve specific S frequencies. Through simulations and analytical calculations, we show that intermediate target S frequencies might be elusive, akin to maintaining a raft's position within a waterfall, rather than above or below it. This challenge arises because intra-collective selection, favoring F during growth, is the strongest at intermediate S frequencies, which can overpower counteracting inter-collective selection effects. Achieving low target S frequencies is consistently possible as expected, but high target S frequencies require an initially high S frequency — similar to a raft that can descend but not ascend a waterfall. The range of attainable target frequencies is significantly influenced by the initial population size of the collectives, while the number of collectives under selection plays a less critical role. In scenarios involving more than two types, 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.

Artificial selection on microbial collectives | Variation and heredity of collective's composition | Stochastic process

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 (20–30), often with unimpressive outcomes.

One of the major challenges in selecting collectives is to ensure the inheritance of a collective function (31, 32). Inheritance can be compromised when genotype and species compositions change from a parent collective to offspring collectives. During maturation of a collective, genotype compositions within each species can be altered by intra-collective evolution, while species compositions can be shifted by 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 of two types with different growth rates to achieve a taget composition in the Adult collective. 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, 33), so long as the shared resource is depleted during collective maturation (24). In this case, both species initially 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).

Here, we mathematically examine the selection of target type compositions in two-type collectives when nutrient is always in excess. This allows us to analytically examine the evolutionary tipping point between intra-collective and inter-collective selection. We show that this tipping point creates a

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

The authors declare no competing interest.

[1]To whom correspondence should be addressed. E-mail:wenying.shou@gmail.com and E-mail: hyejin.park@inha.ac.kr

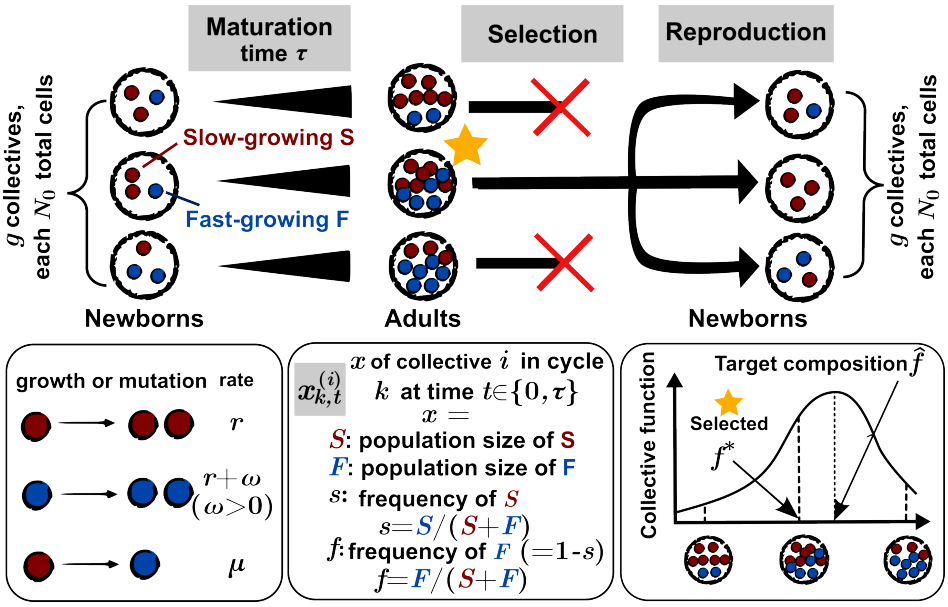

Fig. 1. Schematic for artificial selection on collectives. Each selection cycle begins with a total of g Newborn collectives (black open circles), each with N_0 total cells of slow-growing type (S, red-colored dots) and fast-growing type (F, blue-colored dots). During maturation (over time τ), S and F cells divide at rates r and $r+\omega$ ($\omega > 0$), respectively. 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 each 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 quantity 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.

"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 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 types.

Results and Discussions

We start with selecting for a target composition in collectives of fast-growing F and slow-growing S types. 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 types. We mainly describe in terms of F frequency $f = F/(S+F)$, which is related to S frequency $s = 1 - f$.

A selection cycle. A selection cycle (Fig. 1) starts with a total of g Newborn collectives. We consider two genotypes - slow-growing type (S) and fast-growing type (F), with S mutating to F at a rate μ. At the beginning of cycle k ($t = 0$), a 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 ($1 \le i \le g$) at time t ($0 \le t \le \tau$) 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 and $r+\omega$ ($\omega > 0$), respectively. If maturation time τ is too small, matured collectives ("Adults") do 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 $\tau = \ln(g+1)/r$, which guarantees sufficient cells to produce g Newborn collectives from a single Adult collective.

At the end of a cycle, the Adult with the highest function is chosen to reproduce g Newborn collectives each with N_0 cells ('Selection' and 'Reproduction' in Fig. 1). This allows us to employ the simplest version of the extreme value theory, and in our case, choosing top 1 outperforms a less stringent "choosing top 5%" selection regime. Collective function is dictated by the F frequency f, and is maximized at the target value $\hat{f}$. Thus among all Adult collectives, the chosen Adult has F frequency f^* closest to the target value $\hat{f}$. For the next cycle, the number of F cells in Newborns follow a binomial distribution with the mean value $N_0 f^*$. By repeating the selection cycle, we aim to achieve the target composition $\hat{f}$.

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. Note

that once a target frequency is achieved, inter-collective selection is always required to maintain that frequency due to the fitness difference between the two types.

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. 1 in Supplementary Information). If the target value is high (e.g. $\hat{f} = 0.9$, Figure 2**a** magenta), selection is successful 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. $d = |\langle f^* \rangle - \hat{f}| \leq 0.05$; computed absolute errors are shown in Supplementary Information Fig. S3). F frequency of the chosen collective eventually converges to the target value and stays around it, regardless of the initial frequency. In contrast, without collective-level selection (e.g. choosing a random collective to reproduce), F frequency increases until F reaches fixation (Supplementary information Fig. S4**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 an "higher-threshold" ($f^H \sim 0.7$, solid 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** red line), artificial selection succeeds when the initial frequency is below a "lower-threshold" (f^L, dashed 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.

In summary, two regions of target frequencies are "accessible" in the sense that a target value in the region can be maintained once reached; gold in Fig. 2**d**, **e**): (1) target frequencies above f^H ($\hat{f} > f^H$) and (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 how the distribution of F frequency f might change 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 favours 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,in F frequency over one cycle, the minimum f is selected ($f_{k+1}^* = f_{k+1}^{\min}$). 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.

Mathematically, 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 is

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

Equation 1 can be broken down into two parts: (i) $gP_{f_{k+1,\tau}}(f|f_k^*)$ describes that any one of the g collectives must achieve f, where $P_{f_{k+1,\tau}}(f|f_k^*)$ denotes the probability of an Adult ($t = \tau$) in cycle $k+1$ achieving f given f_k^*. (ii) $\left[1 - \int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*)\right]^{g-1}$ describes that the remaining $g-1$ collectives must all achieve a frequency above f (and thus not selected). $P_{f_{k+1,\tau}}(f'|f_k^*)$ is calculated by integrating across ζ the term $P_{f_{k+1,0}}(\zeta|f_k^*)$ (the probability density of $f_{k+1,0}$ of a Newborn in cycle $k+1$ achieving ζ given f_k^*) multiplied by the term $P_{f_{k+1,\tau}}(f'|f_{k+1,0} = \zeta)$ (the probability density of an Adult achieving f' given Newborn's frequency ζ, see Methods).

When we approximate the Adults' f in cycle $k+1$ as Guassian with mean $\bar{f}(\tau)$ and variance $\sigma_f^2(\tau)$, we can obtain an analytical solution of the change in f over one cycle - 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. 2 in Supplementary Information):

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

where

$$\overline{f}(\tau) = \frac{f_k^* e^{s\tau} + \frac{\mu}{s}(1 - f_k^*)(e^{s\tau} - 1)}{1 - f_k^* + f_k^* e^{s\tau} + \frac{\mu}{s}(1 - f_k^*)(e^{s\tau} - 1)}, \quad [3]$$

and

$$\begin{aligned}
\sigma_f^2(\tau) \quad = \quad & \frac{1}{N_0 e^{r\tau}\left((1 - f_k^*) + f_k^* e^{s\tau} + \frac{\mu(1-f_k^*)}{s}(e^{s\tau} - 1)\right)^4} \times [4] \\
& \Big[(1 - f_k^*)^2 \Big(f_k^*(1 - f_k^*) e^{(r+2s)\tau} + f_k^* e^{s\tau}(e^{(r+s)\tau} - 1) \\
& + \frac{\mu(1 - f_k^*)}{s}\left(\frac{2s}{r+2s} e^{2s\tau} - e^{s\tau} + \frac{r}{r+2s}\right)\Big) \\
& + \left(f_k^* e^{s\tau} + \frac{\mu}{s}(1 - f_k^*)(e^{s\tau} - 1)\right)^2 \\
& \times (f_k^*(1 - f_k^*) e^{r\tau} + (1 - f_k^*)(e^{r\tau} - 1))\Big].
\end{aligned}$$

e is Euler's number (2.71828...), and $\Phi^{-1}(\ldots)$ is an inverse cumulative function of standard normal distribution. For those unfamiliar with $\Phi^{-1}(\ldots)$, $y = \Phi^{-1}(x)$ (where $0 \leq x \leq 1$) is the number such that the probability of a standard normal random

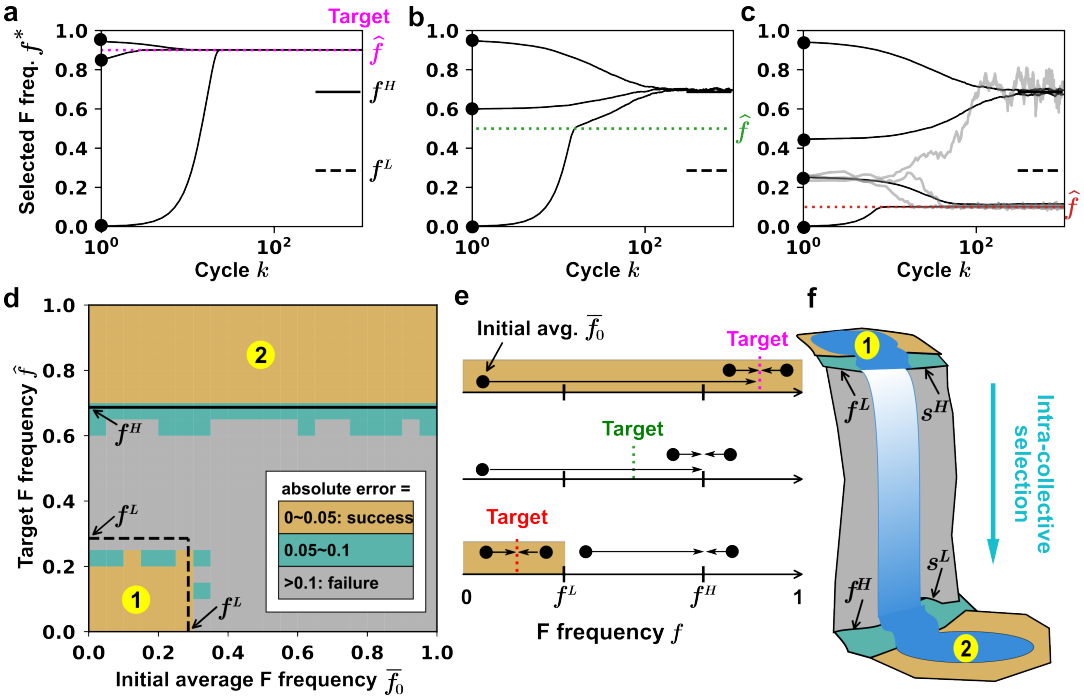

Fig. 2. Initial and target compositions determine the success of artificial selection on collectives. (a-c) Mutant frequency of the selected Adult collective (f^*) over cycles. The target frequency $\hat{f}$ is marked as a dotted line, while f^H (black solid line) and f^L (black dashed line) delineate boundary values that define the region of successful selection. **a** A high target F frequency ($\hat{f} = 0.9 > f^H$; magenta dotted line) can be achieved from any initial frequency (black dots). **b** An intermediate target frequency ($f^L < \hat{f} = 0.5 < f^H$; green dotted line) is never achievable, as all initial conditions converge to near f^H. **c** A low target freqeuncy ($\hat{f} = 0.1 < f^L$) is acheiveable, 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, other reached f^H. For parameters, we used wildtype growth rate $r = 0.5$, F growth advantage $\omega = 0.03$, mutation rate $\mu = 0.0001$, maturation time $\tau \approx 4.8$, and $N_0 = 1000$. The number of collective $g = 10$. Each black line is averaged from independent 300 realizations. **d** Two accessible regions (gold). Either high $\hat{f}$ (region 2) or low $\hat{f}$ starting from low initial f (region 1) can be achieved. We theoretically predict (by numerically integrating Eq. 1) the boundaries of success regions, f^H (black solid line) and f^L (black dashed lines), which agree with simulation results (gold regions). **e** 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. **f** The tension between intra-collective selection and inter-collective selection creates a "waterfall" phenomenon. See the main text for details.

variable being less than or equal to y is x (e.g. $0 = \Phi^{-1}(0.5)$). Eq. [2] contains all experimental parameters, including g (the total number of collectives under selection), τ (maturation time), μ (mutation rate), r (growth rate of the slow-growing S type), and ω (growth advantage of F over S). Equation [2] has the same shape as results from numerically integrating Eq. [1] and from stochastic simulations (Fig. 3**a**).

F frequency can be reduced if the selected f_k^* is low or high, but not if it is intermediate between f^L and f^H (Fig. 3a). Increase in f during maturation is the most drastic when Newborn f is around 0.5 (Fig. 3**b** blue). This is intuitive: when Newborn f is low, the increase 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**).

Not surprisingly, similar conclusions are derived where S and F are slow-growing and fast-growing species which cannot be converted through mutations (Sec. 3 and Fig. S7 in Supplementary Infomation).

Overall, inter-collective selection is akin to a raftman, rowing the raft to a target, while intra-collective selection is akin to a waterfall. This metaphore is best understood in terms of S frequency $s = 1 - f$. Then, 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**f**). 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**d, f** and Fig. 3**a**). 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).

Manipulating experimental setups to expand the achievable target region. Variation among collectives is a key element for successful selection. Here, variation among collectives originates from the stochastic nature of growth and mutation during collective maturation, and of sampling effects during collective reproduction. From Eq. [4], the variance of f at time τ ($\sigma_f^2(\tau)$) scales with N_0^{-1}. Indeed, as Newborn size N_0 declines, the region of acheivable of target frequency expands

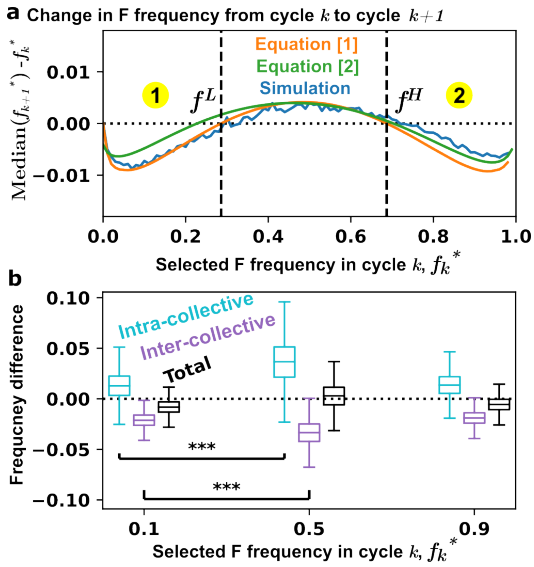

Fig. 3. Intra-collective selection and inter-collective selection jointly set the boundaries for selection success. a The change in F frequency over one cycle. When f_k^* is sufficiently low or high, inter-collective selection can lower the F frequency to below f_k^*. The points where $\mathrm{Median}(f_{k+1}^*) = f_k^*$ 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 indicate the changes in F frequency after intra-collective selection during maturation, after inter-collective selection, and over one selection cycle, respectively. 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.

(larger gold area in Fig. 4**a**). If the Newborn size N_0 is smaller than ~700, any target frequency can be reached when other parameters are fixed.

The number of collectives g also affects the selection outcomes. With more collectives, the chance to find a f closer to the target is more likely. Thus, a larger number of collectives broadens the region of success (Fig. 4b). However, the effect of g is not dramatic. To see why, we note that the only place that g appears is Eq. [2] in the form of $\Phi^{-1}(\frac{1}{g}) = \varphi(g)$. When g becomes large, $\Phi^{-1}(1/g)$ is asymptotically expressed as $\Phi^{-1}(1/g) \approx -\sqrt{2\ln g - \ln[\ln g] + \cdots}$ (Sec. 2 in Supplementary Information)(34). Thus, $\varphi(g)$ does not change dramatically as g varies.

The waterfall phenomenon in a higher dimension. In contrast to the above two-type case where the evolutionary trajectory travels along a one-dimensional line from the initial frequency to the final frequency, in a three-type case the trajectory travels in a two-dimensional plane. To examine the waterfall effect in a higher dimension, we investigate a three-type system where a faster-growing type (FF) grows faster than the fast-growing type (F) which grows faster than the slow-growing type (S) (Fig. 5**a** and Sec. 6 in Supplementary Information). In this system, a target type composition can be achieved if inter-collective selection reduces the frequencies of F as well

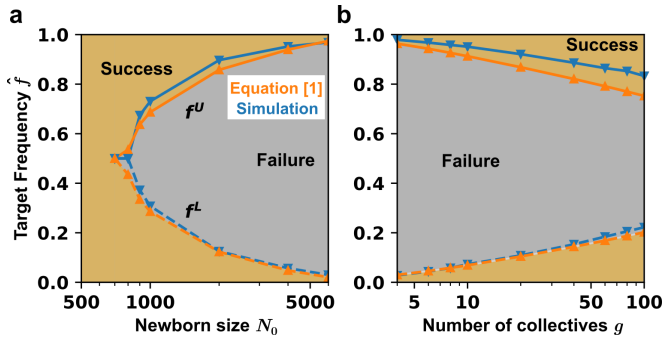

Fig. 4. Expanding the success region for artificial collective selection. **a** Reducing the population size in Newborn N_0 expands the region of success. In gold area, the probability that f^*_{k+1} becomes smaller than f^*_k in a cycle is more than 50%. We used $g = 10$ and $\tau \approx 4.8$. Figures 2-3 correspond to $N_0 = 1000$. **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 ($\tau = \log(100)/r$) is set to be long enough so that an Adult can generate at least 100 Newborns.

as FF (accessible region, gold in Fig. 5**b**). 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).

In conclusion, we have investigated the evolutionary trajectories of type composition in collectives under selection, which are governed by intra-collective selection (which favors fast-growing types) and inter-collective selection (which, in our case, strives to counter fast-growing types). Intra-collective selection has the strongest effect at intermediate frequencies of faster-growing types, 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 (low $\hat{s}$)is achievable from any initial position. In contrast, a more challenging target (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.

Materials and 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 $S^{(i)}_{k,0}$ slow-growing cells and $F^{(i)}_{k,0}$ fast-growing cells. At the first cycle, mean F frequency of collectives are set to be $\bar{f}_{1,0}$. $F^{(i)}_{1,0}$ is sampled from the binomial distribution with mean $N_0\overline{f}_{1,0}$. Then, $S^{(i)}_{1,0}(= N_0 - F^{(i)}_{1,0})$ S cells are in the collective i. In the maturation step, we calculate $S^{(i)}_{k,\tau}$ and $F^{(i)}_{k,\tau}$ by using stochastic simulation. We can simulate the division and mutation of each individual cell stochastically by using tau-leaping algorithm (36, 37)(see Fig. S4**a** and **b** in Supplementray Information). However, individual-based simulation requires a long computing time. Instead, we randomly sample $S^{(i)}_{k,\tau}$ and $F^{(i)}_{k,\tau}$ from the probability distribution $P(S^{(i)}_{k,\tau}, F^{(i)}_{k,\tau})$. To obtain $P(S^{(i)}_{k,\tau}, F^{(i)}_{k,\tau})$, we solve the master equation which describes a time evolution of probability distribution $P(S^{(i)}_{k,t}, F^{(i)}_{k,t})$ under the random processes (see Sec. 1 in Supplementary Information). We assumed that $P(S^{(i)}_{k,\tau}, F^{(i)}_{k,\tau})$ follows the Gaussian distribution. The mean and variance of the distribution are numerically obtained from ordinary differential equations which is derived from the master equation (see Sec. 1 in Supplementary Information). We choose the collective with the closest frequency to the target $\hat{f}$. The selected collective 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 $S^{(i)}_{k+1,0} = N_0 - F^{(i)}_{k+1,0}$.

Analytical approach to the conditional probability. The conditional probability distribution $\Psi(f^*_{k+1}|f^*_k)$ that we observe 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 $F^{(1)}_{k+1,0}, \cdots, F^{(g)}_{k+1,0}$ at the beginning of the cycle $k+1$, which can be mapped into $f^{(1)}_{k+1,0}, \cdots, f^{(g)}_{k+1,0}$ with the constraint of N_0 cells. If the number of cells in the selected collective is large enough, the joint conditional distribution function $P\left(f^{(1)}_{k+1,0}, \cdots, f^{(g)}_{k+1,0}|f^*_k\right)$ 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 $\mu = f^*_k$ 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

$$P_{f_{k+1,0}}(\zeta|f^*_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\zeta-\mu)^2}{2\sigma^2}\right). \quad [5]$$

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 $\mathcal{N}(\overline{f}(\tau), \sigma^2_f(\tau))$. The mean $\overline{f}(\tau)$ and variance $\sigma^2_f(\tau)$ are calculated by using means and variances of S and F (see Sec. 2 in Supplementary Information.) Since $\overline{f}(\tau)$ and $\sigma^2_f(\tau)$ also depend on ζ, the conditional probability distribution function of $f_{k+1,\tau}$ being f is given by

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

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 producting two Gaussian distribution functions and integrating

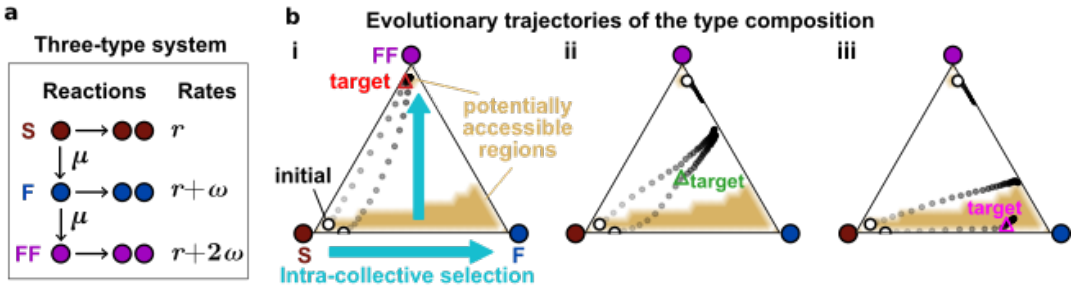

Fig. 5. In higher dimensions, the success of artificial selection requires the entire evolutionary trajectory remaining in the accessible region. a During collective maturation, a slow-growing type (S) (with growth rate r; dark red) can mutate to a fast-growing type (F) (with growth rate $r+\omega$; blue), which can mutate further into a faster-growing type (FF) (with growth rate $r+2\omega$; purple). 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. 1 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 (35).

overall ζ values, which is given by

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

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 replace $f_{k+1}^{\min}$ to 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. [3] and Eq. [4]. Then, the extreme value theory (38) estimates the median of the selected Adult by

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

The F frequency difference in Eq. [2] is obtained by substracting 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

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 (H.J.P.) 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, and Alex Yuan for constructive comments and discussions.

1. S Woo, I Song, HJ Cha, Fast and facile biodegradation of polystyrene by the gut microbial flora of *Plesiophthalmus davidis* larvae. *Appl. Environ. Microbiol.* **86**, e01361–20 (2020).
2. J Sun, A Prabhu, ST Aroney, C Rinke, Corrigendum: Insights into plastic biodegradation: community composition and functional capabilities of the superworm (*Zophobas morio*) microbiome in styrofoam feeding trials. *Microb. Genomics* **8** (2022).
3. JR Bober, CL Beisel, NU Nair, Synthetic biology approaches to engineer probiotics and members of the human microbiota for biomedical applications. *Annu. review biomedical engineering* **20**, 277–300 (2018).
4. EX Wang, MZ Ding, Q Ma, XT Dong, YJ Yuan, Reorganization of a synthetic microbial consortium for one-step vitamin c fermentation. *Microb. Cell Factories* **15**, 21 (2016).
5. CJ Goodnight, Experimental studies of community evolution i: The response to selection at the community level. *Evolution* **44**, 1614–1624 (1990).
6. CJ Goodnight, Experimental studies of community evolution ii: The ecological basis of the response to community selection. *Evolution* **44**, 1625–1636 (1990).
7. W Swenson, DS Wilson, R Elias, Artificial ecosystem selection. *Proc. Natl. Acad. Sci.* **97**, 9110–9114 (2000).
8. W Swenson, J Arendt, DS Wilson, Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environ. Microbiol.* **2**, 564–571 (2000).
9. K Panke-Buisse, AC Poole, JK Goodrich, RE Ley, J Kao-Kniffin, Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME journal* **9**, 980–989 (2015).
10. M Blouin, B Karimi, J Mathieu, TZ Lerch, Levels and limits in artificial selection of communities. *Ecol. Lett.* **18**, 1040–1048 (2015).
11. K Panke-Buisse, S Lee, J Kao-Kniffin, Cultivated sub-populations of soil microbiomes retain early flowering plant trait. *Microb. Ecol.* **73**, 394–403 (2017).
12. RJ Wright, MI Gibson, JA Christie-Oleza, Understanding microbial community dynamics to improve optimal microbiome selection. *Microbiome* **7**, 1–14 (2019).
13. MD Jochum, KL McWilliams, EA Pierson, YK Jo, Host-mediated microbiome engineering (hmme) of drought tolerance in the wheat rhizosphere. *PLoS One* **14**, e0225933 (2019).
14. AS Tiffany Raynaud, Marion Devers, M Blouin, Effect of the reproduction method in an artificial selection experiment at the community level. *Front. Ecol. Evol.* **7**, 416 (2019).
15. CY Chang, ML Osborne, D Bajic, A Sanchez, Artificially selecting bacterial communities using propagule strategies. *Evolution* **74**, 2392–2403 (2020).
16. ASM Jigyasa Arora, Margaret A. Mars Brisbin, Effects of microbial evolution dominate those of experimental host-mediated indirect selection. *PeerJ* **8**, e9350 (2020).
17. UG Mueller, et al., Artificial selection on microbiomes to breed microbiomes that confer salt tolerance to plants. *mSystems* **6**, e01125–21 (2021).
18. T Raynaud, M Devers-Lamrani, A Spor, M Blouin, Community diversity determines the evolution of synthetic bacterial communities under artificial selection. *Evolution* **76**, 1883–1895 (2022).
19. S Jacquiod, et al., Artificial selection of stable rhizosphere microbiota leads to heritable plant phenotype changes. *Ecol. Lett.* **25**, 189–201 (2022).
20. A Penn, "Modelling artificial ecosystem selection: A preliminary investigation" in *Advances in Artificial Life*, eds. W Banzhaf, J Ziegler, T Christaller, P Dittrich, JT Kim. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 659–666 (2003).
21. A Penn, I Harvey, "The role of non-genetic change in the heritability, variation, and response to selection of artificially selected ecosystems" in *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life*, eds. J Pollack, MA

Bedau, P Husbands, RA Watson, T Ikegami. (MIT Press), Vol. 9, p. 352 (2004).
22. HTP Williams, TM Lenton, Artificial selection of simulated microbial ecosystems. *Proc. Natl. Acad. Sci.* **104**, 8918 (2007).
23. L Xie, AE Yuan, W Shou, Simulations reveal challenges to artificial community selection and possible strategies for success. *PLoS Biol.* **17**, e3000295 (2019).
24. G Doulcier, A Lambert, S De Monte, PB Rainey, Eco-evolutionary dynamics of nested darwinian populations and the emergence of community-level heredity. *Elife* **9**, e53433 (2020).
25. L Xie, W Shou, Steering ecological-evolutionary dynamics to improve artificial selection of microbial communities. *Nat. Commun.* **12**, 6799 (2021).
26. CY Chang, et al., Engineering complex communities by directed evolution. *Nat. Ecol. & Evol.* **5**, 1011 (2021).
27. J Fraboul, G Biroli, S De Monte, Artificial selection of communities drives the emergence of structured interactions. *J. Theor. Biol.* **571**, 111557 (2023).
28. A Lalejini, E Dolson, AE Vostinar, L Zaman, Artificial selection methods from evolutionary computing show promise for directed evolution of microbes. *Elife* **11**, e79665 (2022).
29. M Zaccaria, N Sandlin, Y Soen, B Momeni, Partner-assisted artificial selection of a secondary function for efficient bioremediation. *Iscience* **26**, 107632 (2023).
30. B Vessman, P Guridi-Fernández, FI Arias-Sánchez, S Mitri, 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. L Xie, AE Yuan, W Shou, 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).
32. JL Thomas, J Rowland-Chandler, W Shou, Artificial selection of microbial communities: what have we learnt and how can we improve? *Curr. Opin. Microbiol.* **77**, 102400 (2024).
33. PB Rainey, Major evolutionary transitions in individuality between humans and ai. *Philos. Transactions Royal Soc. B* **378**, 20210408 (2023).
34. J Philip, The function inverfc theta. *Aust. J. Phys.* **13**, 13 (1960).
35. Y Ikeda, B Grabowski, F Körmann, mpltern 0.3.0: ternary plots as projections of Matplotlib (2019).
36. DT Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. *The J. Chem. Phys.* **115**, 1716–1733 (2001).
37. Y Cao, DT Gillespie, LR Petzold, Efficient step size selection for the tau-leaping simulation method. *The J. Chem. Phys.* **124**, 044109 (2006).
38. EJ Gumbel, *Statistics of extremes.* (Columbia university press), (1958).

Supplementary Information for

Artificial selection for microbial collective composition can succeed or fail depending on the initial and target values

Juhee Lee, Wenying Shou, and Hye Jin Park

Wenying Shou.
E-mail:wenying.shou@gmail.com
Hye Jin Park.
E-mail: hyejin.park@inha.ac.kr

This PDF file includes:

Figs. S1 to S10 (not allowed for Brief Reports)
SI References

1. Stochastic simulation for the evolution of collective function

In the main text, we design a simple model of artificial selection on collectives. The collectives are composed of two genotypes, slow-growing (S) and fast-growing (F) genotypes. We start with g 'Newborn' collectives and each collective i ($i = 1, ..., g$) has $S^{(i)}_{k,t}$ S cells and $F^{(i)}_{k,t}$ F cells. The index k is a cycle number and t is time indices where $0 \leq t \leq \tau$. At the beginning of cycle k ($t = 0$), any collective has a fixed total cell number $N_0 = S^{(i)}_{k,0} + F^{(i)}_{k,0}$. The collectives are allowed to "mature" for $t = \tau$, where S and F populations grow at rates of r and $r + \omega$, respectively. S cells also mutate to F cells at a rate of μ per cell, and the back mutation from F to S is ignored. After maturation, we choose the collective with F frequency the closest to the target value $\hat{f}$, and the chosen collective "reproduce" g newborn collectives each with N_0 cells for the next cycle.

Maturation. In this subsection, we ignore the cycle number index k and collective index (i) in convenience. So we denote $S(t)$ and $F(t)$ as representative notations of $S^{(i)}_{k,t}$ and $F^{(i)}_{k,t}$, respectively. The initial S and F cell numbers in a Newborn collective are $S^{(i)}_{k,0}$ and $F^{(i)}_{k,0}$. We consider mutation that is beneficial to grow ($\omega > 0$).

We consider the population growth of $S(t)$ and $F(t)$ as a stochastic process. We use $P(S, F, t)$ to denote the probability to have S S cells and F F at time t from the beginning of a cycle. Birth and mutation events stochastically happen which are represented by the three chemical reaction rules:

$$S \xrightarrow{r} S + S, \quad [1]$$

$$F \xrightarrow[s>0]{r+s} F + F, \quad [2]$$

$$S \xrightarrow{\mu} F, \quad [3]$$

where S and F denote individual cells of a S type and a F type, respectively. The reaction rules can be simulated via tau-leaping algorithm (1, 2), but the individual-based simulation requires long simulation times to get enough results. Hence we go over probability distribution $P(S, F, t)$ to sample compositions instead of tracking all events that occurred in individual collectives. First, we calculate the distribution $P(S, F, t)$ analytically, and then sample numbers of S and F cells from this distribution at the end of cycle $t = \tau$. Below, we describe the derivation of $P(S, F, t)$.

The master equation of $P(S, F, t)$, which follows reaction rules Eq. [1-3], is given by

$$\begin{aligned} \frac{dP(S,F,t)}{dt} &= r(S-1)P(S-1,F,t) \\ &\quad +(r+s)(F-1)P(S,F-1,t) \\ &\quad +\mu(S+1)P(S+1,F-1,t) \\ &\quad -(rS+(r+s)F+\mu S)P(S,F,t). \end{aligned} \quad [4]$$

Here, dt is small such that at most one cell birth or mutation event could occur. 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. The last term corresponds to the outflow of probability density by birth and mutation processes, which induces the changes in composition from (S, F) to others.

Calculating the exact form of $P(S, F, t)$ is not simple. Instead, we consider the probability distributions of $S(t)$ and $F(t)$ at time t. We assume that the mutation rate is much smaller than the growth rates, and hence the correlation between S and F is sufficiently small. Then, $P(S, F, t) \sim P(S, t)P(F, t)$. The distributions of S and F can be approximated as Gaussian ($\mathcal{N}$), which is supported by the Central Limit Theorem and Figure S1 (note the concordance between the blue and green distributions). That is, the probability distribution $P(S, F, t) \sim \mathcal{N}(S(t)|\overline{S}(t), \sigma^2_S(t))\mathcal{N}(F(t)|\overline{F}(t), \sigma^2_F(t))$.

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

$$\frac{d\overline{S}(t)}{dt} = r\overline{S}(t) - \mu\overline{S}(t), \quad [5]$$

$$\frac{d\overline{F}(t)}{dt} = (r+s)\overline{F}(t) + \mu\overline{S}(t). \quad [6]$$

We assume that the mutation rate μ is very smaller than r and ω. By solving Eq. [5] and Eq. [6], the means $\overline{S}(t)$ and $\overline{F}(t)$ are given by

$$\overline{S}(t) = \overline{S}_o e^{(r-\mu)t} \approx \overline{S}_o e^{rt}, \quad [7]$$

$$\overline{F}(t) = \overline{F}_o e^{(r+s)t} + \frac{\mu\overline{S}_o}{s+\mu}\left(e^{(r+s)t} - e^{(r-\mu)t}\right) \approx \overline{F}_o e^{(r+s)t} + \frac{\mu\overline{S}_o}{s}(e^{(r+s)t} - e^{rt}), \quad [8]$$

where $\overline{S}_o \equiv \overline{S}(0)$ and $\overline{F}_o \equiv \overline{F}(0)$ are the mean number of S and F cells at the beginning of cycle $t = 0$. Note that the second term of Eq. [8] is consistent with the studies on the distribution of the number of F cells starting from S_0 S cells (3).

We define the second momenta of S and F as

$$\overline{S^2} = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S^2 P(S,F,t), \tag{9}$$

$$\overline{F^2} = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} F^2 P(S,F,t). \tag{10}$$

Then, corresponding differential equations are given by

$$\frac{d\overline{S^2}}{dt} = 2(r-\mu)\overline{S^2} + (r+\mu)\overline{S}, \tag{11}$$

$$\frac{d\overline{F^2}}{dt} = 2(r+s)\overline{F^2} + (r+s)\overline{F} + 2\mu\overline{SF} + \mu\overline{S}. \tag{12}$$

The solution of Eq. [11] is

$$\overline{S^2}(t) = \frac{r+\mu}{\mu-r}\overline{S}_o\left[e^{(\mu-r)t}-1\right]e^{2(r-\mu)t} + \overline{S^2}_o e^{2(r-\mu)t} \tag{13}$$

$$\approx \overline{S_o^2}e^{2rt} + \overline{S}_o e^{rt}(e^{rt}-1), \tag{14}$$

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

$$\sigma_S^2(t) = \sigma_{S,0}^2 e^{2rt} + \overline{S}_o e^{rt}(e^{rt}-1), \tag{15}$$

where $\sigma_{S,0}^2 \equiv \sigma_S^2(0)$ is a variance of S cell numbers at the beginning of cycle $t=0$. In Eq. [12], we require $\overline{SF}(t) = \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} SFP(S,F,t)$ to calculate $\overline{F^2}(t)$. The master Eq. [4] provides a differential equation for $\overline{SF}(t)$ as

$$\frac{d\overline{SF}}{dt} = (2r+s-\mu)\overline{SF} + \mu(\overline{S^2}-\overline{S}). \tag{16}$$

The solution of Eq. [16] is given by

$$\begin{aligned}\overline{SF}(t) = {} & \overline{SF}_o e^{(2r+s)t} \\ & + \frac{\mu}{s}(\overline{S^2}_0 + \overline{S}_0)(e^{(2r+s)t} - e^{2rt}) \\ & - \frac{2\mu\overline{S}_0}{r+s}(e^{(2r+s)t} - e^{rt}).\end{aligned} \tag{17}$$

By using Eq. [17], the solution of Eq. [12] is given by

$$\begin{aligned}\overline{F^2}(t) = {} & \overline{F^2}_0 e^{2(r+s)t} + \overline{F}_0 e^{(r+s)t}(e^{(r+s)t}-1) \\ & + \frac{\mu\overline{S}_o}{s}\left(\frac{2s}{r+2s}e^{2(r+s)t} - e^{(r+s)t} + \frac{r}{r+2s}e^{rt}\right) \\ & + \frac{2\mu\overline{SF}_0}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) \\ & + O(\mu^2),\end{aligned} \tag{18}$$

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

$$\begin{aligned}\sigma_F^2(t) = {} & \sigma_{F,0}^2 e^{2(r+s)t} + \overline{F}_0 e^{(r+s)t}(e^{(r+s)t}-1) \\ & + \frac{\mu\overline{S}_o}{s}\left(\frac{2s}{r+2s}e^{2(r+s)t} - e^{(r+s)t} + \frac{r}{r+2s}e^{rt}\right) \\ & + \frac{2\mu(\overline{SF}_0 - \overline{S}_0\overline{F}_0)}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) \\ & + O(\mu^2).\end{aligned} \tag{19}$$

Using Eq. [7,8,15,19], we construct Gaussian distribution functions for $S(t)$ and $F(t)$, $\mathcal{N}(S|\overline{S}(\tau), \sigma_S^2(\tau))$ and $\mathcal{N}(F|\overline{F}(\tau), \sigma_F^2(\tau))$. Then, numbers of S cells and F cells are randomly sampled from each distribution: $S(\tau)$ from $\mathcal{N}(S|\overline{S}(\tau), \sigma_S^2(\tau))$ and $F(\tau)$ from $\mathcal{N}(F|\overline{F}(\tau), \sigma_F^2(\tau))$. This process requires only two sampling steps for each collective.

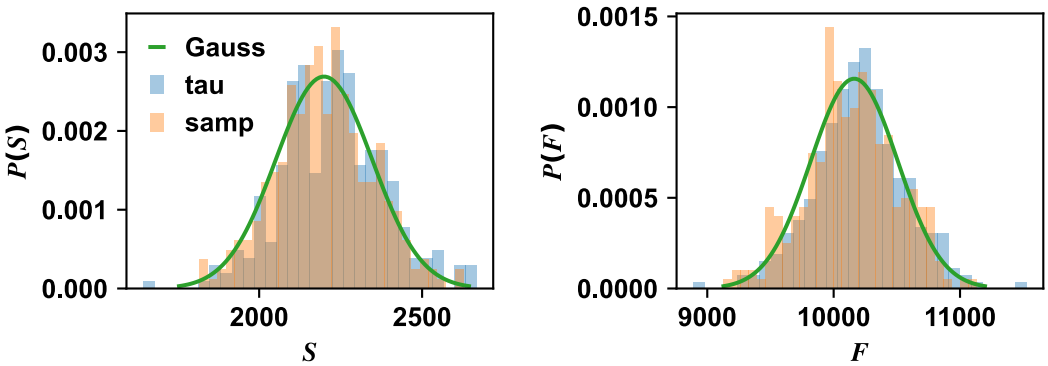

Fig. S1. Comparison between the calculated Gaussian distribution ("Gauss", with the mean and variances computed from Eq. [7,8,15,19]) and simulations using tau-leaping ("tau") and sampling ("samp") methods. The simulations run 500 times. The initial number of cells are $S_0 = 200$ and $F_0 = 800$. The parameters are $r = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, and $\tau = 4.8$.

Selection. After the maturation step is finished in cycle k, we compute the frequencies of the F cells in each collectives $\{f_{k,\tau}^{(1)}, \cdots, f_{k,\tau}^{(g)}\}$ where we denote the frequency of collective i at time $t = \tau$ in cycle k as $f_{k,\tau}^{(i)}$. We select one frequency which is the closest value to the target frequency $\hat{f}$. The frequency selected at the cycle k is denoted by f_k^*. In mathematical expression, the selected frequency is

$$f_k^* = \text{argmin}_{f_{k,\tau}^{(i)}, i \in \{1, \cdots, g\}} |f_{k,\tau}^{(i)} - \hat{f}|. \quad [20]$$

Reproduction. We divide the selected collective into g Newborn collectives. We sample $F_{k+1,0}^{(i)}$ F cells from the selected collective for the Newborns in the next cycle $k+1$. The exact description of reproduction step is described by consecutive sampling without replacement from the selected collective. Then the joint probability distribution $P(F_{k+1,0}^{(1)}, \cdots, F_{k+1,0}^{(g)})$ follows a multivariate hypergeometric distribution. The selected collective is assumed to have S_k^* S cells and F_k^* F cells. If the size of selected collective $N_k^* = S_k^* + F_k^*$ is large enough, the consecutive sampling is approximated to the independent binomial sampling (see Fig. S2). 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 $F_{k+1,0}^{(i)}$ is given by

$$P_{F_{k+1,0}^{(i)}}(F) = \frac{N_0!}{(N_0 - F)!F!} (f_k^*)^F (1 - f_k^*)^{N_0 - F}. \quad [21]$$

After the 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. These Newborn collectives start the maturation in the next cycle.

Simulation Result. In the main text, we simulate artificial selection by using stochastic simulation. In Fig. S3, we draw 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). In the colormap, errors higher than 0.15 are marked with gray, which indicates the failure of selection. The dashed lines indicate the same boundary in Fig 2**d** in the main text.

Figure S4 presents the composition trajectories of all collectives using tau-leaping algorithm. The selected collectives have the closest compotisition to the target composition $\hat{f}$ at the end of each cycle. The selected collective can have smaller frequency than its parent collective, so the composition can be lowered after cycles.

2. Conditional probability distribution of the selected collective frequency f^*

Since maturation and reproduction are stochastic processes, the frequency f_{k+1}^* of the selected collective at cycle $k+1$ starting from f_k^* is also described by the probability distribution. In this case, f_{k+1}^* depends on f_k^*, and the distribution of f_{k+1}^* follows the conditional probability distribution $\Psi(f_{k+1}^* | f_k^*)$. Below, we describe calculation of the conditional probability distribution $\Psi(f_{k+1}^* | f_k^*)$.

We focus on the situation where the change of frequency in the selection step has the opposite direction of the maturation step. That means the target frequency $\hat{f}$ is smaller than f_k^*, and the collective with the smallest frequency will be selected among g Newborn collectives. We can calculate the probability distribution of the Newborn collectives' frequencies at the end of cycle $k+1$ by using Eq. [7,8,15,19] and Eq. [21]. Then, we obtain a distribution of the minimum value among g copies of the collectives. The minimum value distribution is the conditional probability distribution $\Psi(f_{k+1}^* | f_k^*)$.

Let us start the calculation from the reproduction step at the end of cycle k. The probability distribution of the F cell numbers in Newborn collectives is given in Eq. [21]. If the total number of cells in a Newborn collective N_0 is large enough, Eq. [21] is approximated by the Gaussian distribution $F_{k+1,0}^{(i)} \sim \mathcal{N}(N_0 f_k^*, f_k^*(1 - f_k^*)N_0)$. Then, the probability density function that $f_{k+1,0}^{(i)}$ to be ζ in Newborn collective i is

$$P_{f_{k+1,0}^{(i)}}(\zeta | f_k^*) = \frac{\sqrt{N_0}}{\sqrt{2\pi f_k^*(1 - f_k^*)}} e^{-\frac{N_0(\zeta - f_k^*)^2}{2 f_k^*(1 - f_k^*)}}. \quad [22]$$

The Newborn collective i maturates in cycle $k+1$ with the initial cell numbers $S_{k+1,0}^{(i)} = N_0(1 - \zeta)$ and $F_{k+1,0}^{(i)} = N_0 \zeta$. Since all Newborn collectives are independent and identically distributed, we can drop the superscript (i) for simplicity. We can express $S(t)$ and $F(t)$ based on Eq. [7,8,15,19] as

$$S(t) \approx \overline{S}(t) + \sigma_S(t) G_S(t), \quad [23]$$

$$F(t) \approx \overline{F}(t) + \sigma_F(t) G_F(t), \quad [24]$$

where $G_S(t)$ and $G_F(t)$ are random variable following the standard distribution $\mathcal{N}(0, 1)$. Here we ignore cycle index $k+1$ in subscription for convenience. Then, we can approximately write $f(t)$ as

$$\begin{aligned} f(t) &\approx \frac{\overline{F}(t) + \sigma_F(t) G_F(t)}{\overline{S}(t) + \sigma_S(t) G_S(t) + \overline{F}(t) + \sigma_F(t) G_F(t)} \\ &\approx \overline{f}(t) + \sigma_f(t) G_f(t). \end{aligned} \quad [25]$$

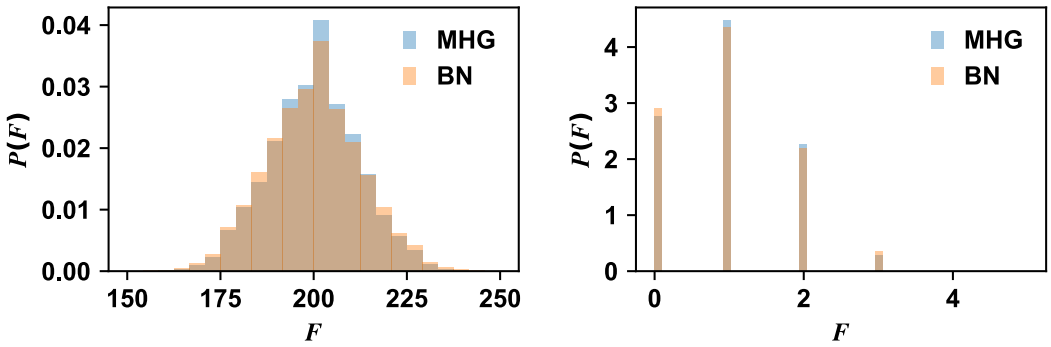

Fig. S2. Comparison between consecutive sampling and independent binomial sampling. A parent collective is divided into 10 collectives. The histogram labeled with 'MHG' is the probability mass function of F of the fifth collective sampled via multivariate hypergeometric distribution. The independent binomial sampling is labeled with 'BN'. 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.

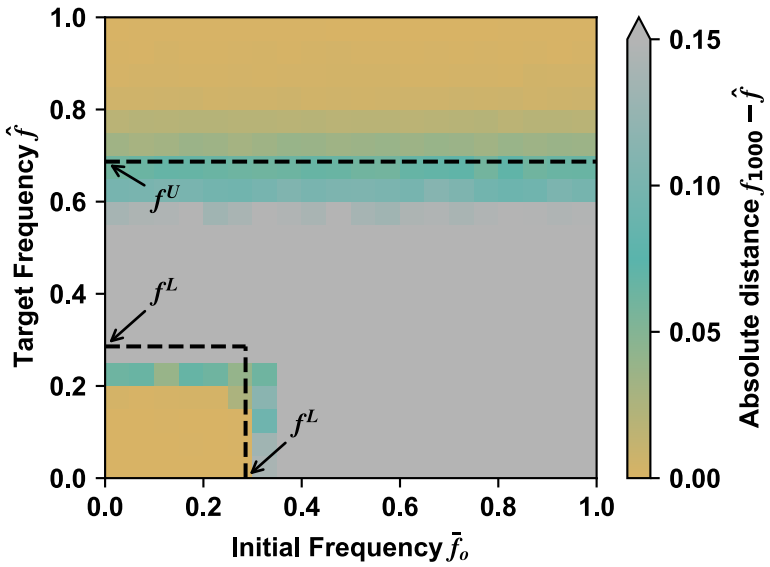

Fig. S3. Color map of the absolute error $d = |\langle f^* \rangle - \hat{f}|$ between frequency $\langle f^* \rangle$ of the averaged selected collectives at the end of simulations ($k = 1000$) and the target frequency $\hat{f}$. The dashed lines are drawn by the arguments in the main text. For parameters, we used $r = 0.5$, $\omega = 0.03$, $\mu = 0.0001,\ N_0 = 1000,\ g = 10$ and $\tau \approx 4.8$.

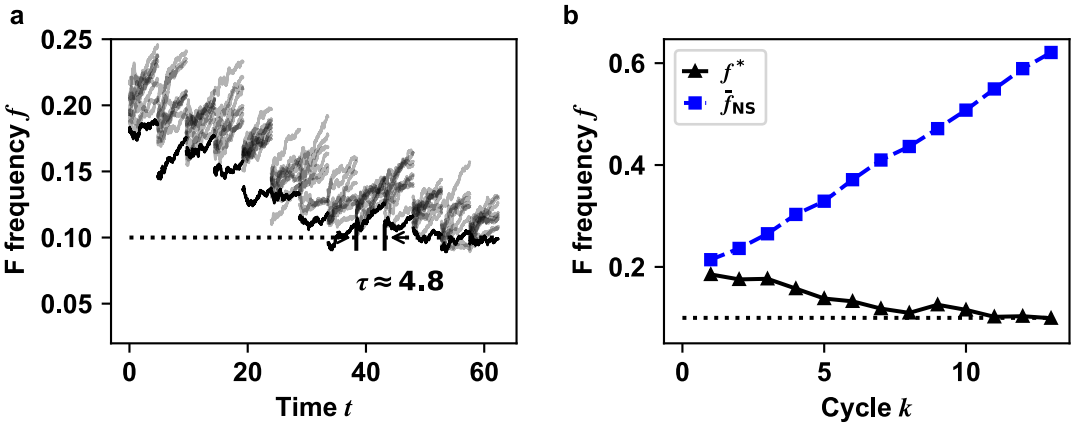

Fig. S4. a Trajectories of F frequency for 10 collectives ($g = 10$) over time. The collective whose frequency is closest to the target value is selected in every cycle (black lines). The gray lines denote the other collectives. For parameters, we used S growth rate $r = 0.5$, F growth advantage $\omega = 0.03$, mutation rate $\mu = 0.0001$, maturation time $\tau \approx 4.8$, and $N_0 = 1000$. **b** Comparison between frequency trajectories with (black) and without (blue) selection 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 $\overline{f}_{\mathrm{NS}}$ (the average of $g = 10$ individual lineages without collective-level selection at the end of each cycle).

The mean of f is given by

$$\overline{f}(t) = \frac{\overline{F}(t)}{\overline{S}(t)+\overline{F}(t)} = \frac{\zeta e^{st} + \frac{\mu}{s}(1-\zeta)(e^{st}-1)}{1-\zeta+\zeta e^{st}+\frac{\mu}{s}(1-\zeta)(e^{st}-1)}, \quad [26]$$

and the variance is

$$\begin{aligned}\sigma_f^2(t) &= \frac{(1-\overline{f}(t))^2\sigma_F^2(t)+\overline{f}(t)^2\sigma_S^2(t)}{\overline{N}(t)^2} \quad [27]\\ &= \frac{1}{N_0 e^{rt}\left((1-\zeta)+\zeta e^{st}+\frac{\mu(1-\zeta)}{s}(e^{st}-1)\right)^4}\times \\ &\left[(1-\zeta)^2\left(\zeta e^{st}(e^{(r+s)t}-1)+\frac{\mu(1-\zeta)}{s}\left(\frac{2s}{r+2s}e^{(r+2s)t}-e^{st}+\frac{r}{r+2s}\right)\right)\right. \\ &\left.+\left(\zeta e^{st}+\frac{\mu}{s}(1-\zeta)(e^{st}-1)\right)^2(1-\zeta)(e^{rt}-1)\right].\end{aligned}$$

The average size of collectives is $\overline{N}(t) = \overline{S}(t)+\overline{F}(t)$. In terms of a probability distribution, a random variable that a Adult's F frequency $f(\tau) = f_\tau$ follows the Gaussian distribution $\mathcal{N}(\overline{f}_\tau, \sigma_{f_\tau}^2)$ whose the probability density function is given by

$$P_{f_\tau}(f|\zeta) = \frac{1}{\sqrt{2\pi}\sigma_{f_\tau}} e^{-\frac{(f-\overline{f_\tau})^2}{2\sigma_{f_\tau}^2}}. \quad [28]$$

With indices of cycle k and $k+1$, the conditional probability that a offspring collective has frequency f after the maturation in cycle $k+1$ from the parent frequency f_k^* is

$$P_{f_{k+1,\tau}}(f|f_k^*) = \int_0^1 d\zeta\ P_{f_{k+1,\tau}}(f|\zeta)\ P_{f_{k+1,0}}(\zeta|f_k^*). \quad [29]$$

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

$$\begin{aligned}C_{f_{k+1}^{\min}}(f|f_k^*) &= \text{Prob}[f_{k+1}^{\min} \le f|f_k^*] \\ &= 1-\text{Prob}[f_{k+1,\tau}^{\min} \ge f|f_k^*] \\ &= 1-\text{Prob}[(f_{k+1,\tau}^{(1)} \ge f)\wedge(f_{k+1,\tau}^{(2)} \ge f)\wedge\cdots\wedge(f_{k+1,\tau}^{(g)} \ge f)|f_k^*]. \quad [30]\end{aligned}$$

The above Eq. [30] simply states that $f_{k+1}^{\min} \le f$ if every single of g collectives at the end of maturation (time τ) in cycle $k+1$ have frequencies greater than or equal to f. Frequencies are independent and identically distributed, and thus $C_{f_{\mathrm{k+1}}^{\min}}(f|f_k^*) = 1-[\text{Prob}(f_{k+1,\tau} \ge f)|f_k^*]^g$. Note that $\text{Prob}[f_{k+1,\tau} \ge f|f_k^*] = \int_f^1 df' P_{f_{k+1,\tau}}(f'|f_k^*) = 1-\int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*) = 1-C_{f_{k+1,\tau}}(f|f_k^*)$, and Eq. [30] becomes

$$C_{f_{k+1}^{\min}}(f|f_k^*) = 1-(1-C_{f_{k+1,\tau}}(f|f_k^*))^g. \quad [31]$$

Then, the probability density function $\Psi(f_{k+1}^*|f_k^*)$ is obtained by differentiating Eq. [31] with respect to f and replacing $f \to f_{k+1}^*$,

$$\Psi(f_{k+1}^*|f_k^*) = g(1-C_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*))^{g-1}P_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*). \quad [32]$$

We compute the probability density function [32] by using numerical integration and compare it with the stochastic simulation results in Fig. S5. Both distributions agree with each other.

To get the analytic approximation of the median of Eq. [32], we assume that the F frequency distribution of Adult collective is Gaussian form. Thus, instead of calculating marginal probability with ζ variable as in Eq. [29], we only calculate the mean $\overline{f}(\tau)$ and variance $\sigma_f^2(\tau)$, from Newborn's frequency distribution of $\mathcal{N}(f_k^*, f_k^*(1-f_k^*)/N_0)$:

$$\overline{f}(\tau) = \frac{\overline{F}(\tau)}{\overline{S}(\tau)+\overline{F}(\tau)} = \frac{f_k^* e^{s\tau}+\frac{\mu}{s}(1-f_k^*)(e^{s\tau}-1)}{1-f_k^*+f_k^* e^{s\tau}+\frac{\mu}{s}(1-f_k^*)(e^{s\tau}-1)}, \quad [33]$$

$$\begin{aligned}\sigma_f^2(\tau) &= \frac{1}{N_0 e^{r\tau}\left((1-f_k^*)+f_k^* e^{s\tau}+\frac{\mu(1-f_k^*)}{s}(e^{s\tau}-1)\right)^4}\times \quad [34]\\ &\left[(1-f_k^*)^2\left(f_k^*(1-f_k^*)e^{(r+2s)\tau}+f_k^* e^{s\tau}(e^{(r+s)\tau}-1)+\frac{\mu(1-f_k^*)}{s}\left(\frac{2s}{r+2s}e^{(r+2s)\tau}-e^{s\tau}+\frac{r}{r+2s}\right)\right),\right. \\ &\left.+\left(f_k^* e^{s\tau}+\frac{\mu}{s}(1-f_k^*)(e^{s\tau}-1)\right)^2(f_k^*(1-f_k^*)e^{r\tau}+(1-f_k^*)(e^{r\tau}-1))\right]\end{aligned}$$

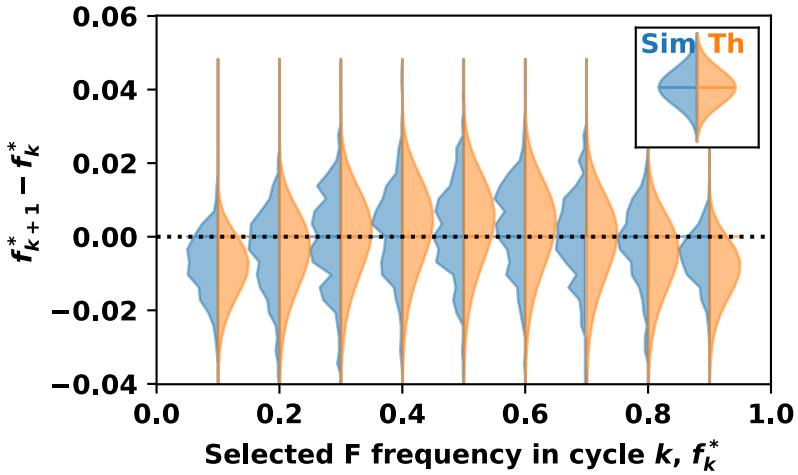

Fig. S5. The probability density functions of the selected collective's frequency f^*_{k+1} with the offset of f^*. The blue distribution is obtained using 300 stochastic simulation results. The orange distribution represents Eq. [32] computed by numerical integration. The median values of the distributions are shown in Fig. S3**a** in the main text.

which are Eq. [3] and Eq.[4] in the main text, respectively.

We note that σ_f^2 is proportional to N_0^{-1}. The variances in Eq. [15] and [19] are linearly scale with N_0 by using $\overline{S}_0 = N_0\zeta$ and $\overline{F}_0 = N_0(1-\zeta)$ from Eq. [22]. The mean collective size $\overline{N}(t) = \overline{S}(t) + \overline{F}(t)$ is also proportional to N_0 because the average cell numbers in Eq. [7] and [8] are linear to N_0. Thus, the scaling relation of Eq. [27] is given by$\sigma_f^2(t) = [(1-\overline{f}(t))^2\sigma_F^2(t) + \overline{f}(t)^2\sigma_S^2(t)]/\overline{N}(t)^2 \sim N_0/N_0^2 \sim 1/N_0$.

The median (Median$[\Psi(f_{k+1}^*|f_k^*)] \equiv \tilde{f}$) of Eq. [32] satisfies $C_{f_{k+1}^{\min}}(\tilde{f}|f_k^*) = 1/2$, which means $\tilde{f} = C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right)$. If we assume the distribution Eq. [32] is the Gaussian form, the inverse function $C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right)$ is writen as

$$C^{-1}_{f_{k+1,\tau}}\left(\frac{\ln 2}{g}\right) = \overline{f}(\tau) + \left[\Phi^{-1}(\frac{\ln 2}{g})\right]\sigma_f(\tau), \qquad [35]$$

where $\Phi^{-1}(y)$ is an inverse cumulative density function (CDF) of the normal distribution with the mean $\overline{f}(\tau)$ in Eq.[34] and the standard deviation $\sigma_f(\tau)$, a square root of Eq.[34]. Substracting f_k^* from Eq. [35] gives Eq. [2] in the main text.

Further, we get an asymptotic expression of $\Phi^{-1}(\ln 2/g)$ when g is large (or $\Phi^{-1}(y)$ with small y). 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 an 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 $\Phi^{-1}(y) = -\sqrt{2}\mathrm{erfc}^{-1}(2y)$. Here, we introduce a method in (5). 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 in both side, we have

$$x^2 \approx -\ln y - \frac{1}{2}\ln \pi x^2. \qquad [36]$$

Replacing x^2 in the righthand side in Eq. [36] into the expression itself, we get continued logarithmic form of

$$x^2 \approx -\ln y - \frac{1}{2}\ln \pi(-\ln y - \frac{1}{2}\ln \pi(-\ln y - \frac{1}{2}\ln \cdots). \qquad [37]$$

Inserting $x = \mathrm{erfc}^{-1}(y)$ (square root of Eq. [37]) 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

$$\Phi^{-1}\left(\frac{\ln 2}{g}\right) \approx -\sqrt{2\ln g - \ln[\ln g] + \cdots}. \qquad [38]$$

3. Selection without mutation

When the mutation rate is zero, two genotypes behave as two distinct species. So the S corresponds to the slower-growing species and the F corresponds to the faster-growing species. Then, our problem is generalized to the selection of two-species collective. The compositional change is provided by Eq. [35] with setting $\mu = 0$. Corresponding $\bar{f}$ and σ_f^2 become

$$\overline{f}(\tau) = \frac{f_k^* e^{s\tau}}{1 - f_k^* + f_k^* e^{s\tau}}, \qquad [39]$$

$$\begin{aligned}\sigma_f^2(\tau) \quad = \quad & \frac{1}{N_0 e^{r\tau}\left((1-f_k^*) + f_k^* e^{s\tau}\right)^4} \times \\ & \left[(1-f_k^*)^2\left(f_k^*(1-f_k^*)e^{(r+2s)\tau} + f_k^* e^{s\tau}(e^{(r+s)\tau} - 1)\right)\right. \\ & \left. + (f_k^* e^{s\tau})^2\left(f_k^*(1-f_k^*)e^{r\tau} + (1-f_k^*)(e^{r\tau} - 1)\right)\right],\end{aligned} \qquad [40]$$

Equations [39] and [40] 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. S7.

4. Selecting more than one collective

In the main text, we choose one collective which have the closest frequency value to the target among g collectives. Such 'Top 1' stratedgy has an advantage on applying extreme value theory while it has potentially lose the 'unlucky' Adults (6). Such 'unlucky' Adults are further to the target value than the selected one, but have possiblity to be selected in next cycle. The 'Top tier' stratedgy may secure the 'unlucky' Adult by choosing more than one collective (6). So we test the 'top-tier' stratedgy by choosing 5 among 100 Adults with the distance from the target value in Fig. S8. The top-tier stratedgy is shown to be inefficient in the two-genotype system. This is because the system is too simple, so lower ranked Adult is lower again in next cycle.

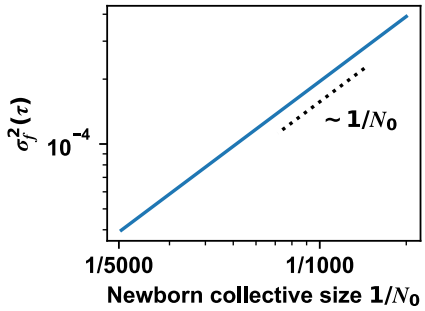

Fig. S6. Scaling relation of F frequency variance (Eq. [27]) with Newborn collective size N_0. The initial F frequency is 0.3. The parameters are $r = 0.5$, $\omega = 0.03$, $\mu = 0.0001$, and $\tau \approx 4.8$.

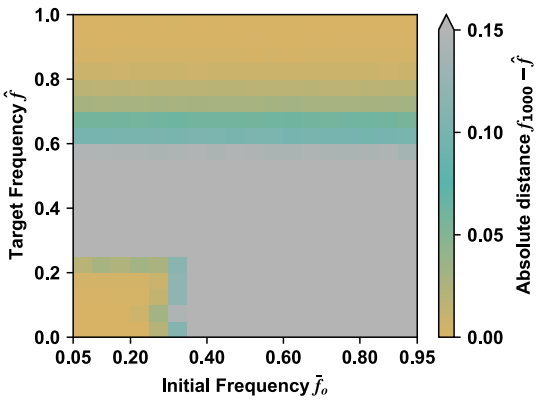

Fig. S7. Color map of the absolute error $d = |\langle f^* \rangle - \hat{f}|$ between frequency $\langle f^* \rangle$ of the averaged selected collectives at the end of simulations ($k = 1000$) and the target frequency $\hat{f}$. For parameters, we used $r = 0.5$, $\omega = 0.03$, $\mu = 0,\ N_0 = 1000,\ g = 10$ and $\tau \approx 4.8$.

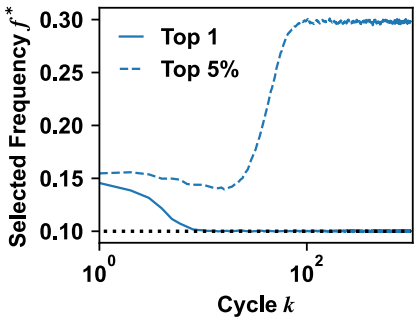

Fig. S8. Comparison of selecting Top-tier 5 with Top 1. We breed 100 collectives and choose 5 collectives with the closest to the target value.

5. Deleterious mutation

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 even 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 conditional probability distribution $\Psi(f^*_{k+1}|f^*_k)$ is a maximum value distribution of Eq. [29]. Thus, instead of Eq. [30], we look for the cumulative distribution function of the maximum value $f^{\max}_{k+1}$ such that

$$\begin{aligned} C_{f^{\max}_{k+1}}(f|f^*_k) &= \text{Prob}[f^{\max}_{k+1} \geq f|f^*_k] \\ &= \text{Prob}[(f^{(1)}_{k+1}(\tau) \geq f) \wedge (f^{(2)}_{k+1}(\tau) \geq f) \wedge \cdots \wedge (f^{(g)}_{k+1}(\tau) \geq f)|f^*_k]. \end{aligned} \quad [41]$$

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

$$C_{f^{\max}_{k+1}}(f|f^*_k) = (C_{f_{k+1,\tau}}(f|f^*_k))^g. \quad [42]$$

Likewise in the previous section, we get the conditional probability density function by differentiating Eq. [42] with respect to f and replacing $f \to f^*_{k+1}$ as

$$\Psi(f^*_{k+1}|f^*_k) = gC_{f_{k+1,\tau}}(f^*_{k+1}|f^*_k)^{g-1}P_{f_{k+1,\tau}}(f^*_{k+1}|f^*_k). \quad [43]$$

The distribution in Eq. [43] is evaluated for various f^*_k in Fig. S9**a** with numerical simulations and the median values of distributions are presented in Fig. S9**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. 2**d** in the main text.

6. Extension to the system including double-mutation

In the main text, we categorize the target frequencies based on whether the frequency change in the maturation step can be compensated in the selection step. Here, we extend the idea into more complex systems in order to suggest a possible generalization. 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. Each mutation adds the growth rate with ω. So the F and FF type have growth rates with $r + \omega$ and $r + 2\omega$, respectively. The mutation rate is μ. So, the birth and mutation events are written by the chemical reactions:

$$S \xrightarrow{r} S + S, \quad [44]$$

$$F \xrightarrow[s>0]{r+s} F + F, \quad [45]$$

$$FF \xrightarrow[s>0]{r+2s} FF + FF \quad [46]$$

$$S \xrightarrow{\mu} F, \quad [47]$$

$$F \xrightarrow{\mu} FF \quad [48]$$

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

$$\begin{aligned} \frac{dP(S,F,FF,t)}{dt} &= r(S-1)P(S-1,F,FF,t) - rSP(S,F,FF,t) \\ &+(r+s)(F-1)P(S,F-1,FF,t) - (r+s)FP(S,F,FF,t) \\ &+(r+2s)(FF-1)P(S,F,FF-1,t) - (r+2s)FFP(S,F,FF,t) \\ &+\mu(S+1)P(S+1,F-1,FF,t) - \mu S\,P(S,F,FF,t) \\ &+\mu(F+1)P(S,F+1,FF-1,t) - \mu F\,P(S,F,FF,t). \end{aligned} \quad [49]$$

The composition of collective i in cycle k is now represented with two frequencies $(f^{(i)}_k(t), h^{(i)}_k(t)) \equiv \mathbf{p}^{(i)}_k(t)$ where the F frequency $f^{(i)}_k(t) = F^{(i)}_k(t)/(S^{(i)}_k(t) + F^{(i)}_k(t) + FF^{(i)}_k(t))$ and the FF frequency $h^{(i)}_k(t) = FF^{(i)}_k(t)/(S^{(i)}_k(t) + F^{(i)}_k(t) + FF^{(i)}_k(t))$. 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 2 to obtain the conditional probability $\Psi(\mathbf{p}^*_{k+1}|\mathbf{p}^*_k)$ by using the master Eq. [49].

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 probability mass function to choosing $F_{k,0}, FF_{k,0}$ cells are represented by

$$P(F_{k+1,0}, FF_{k+1,0}|f^*_k, h^*_k) = \frac{N_0!}{(N_0 - F_{k+1,0} - FF_{k+1,0})!F_{k+1,0}!FF_{k+1,0}!}(1 - f^*_k - h^*_k)^{N_0 - F_{k+1,0} - FF_{k+1,0}}(f^*_k)^{F_{k+1,0}}(h^*_k)^{FF_{k+1,0}}, \quad [50]$$

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

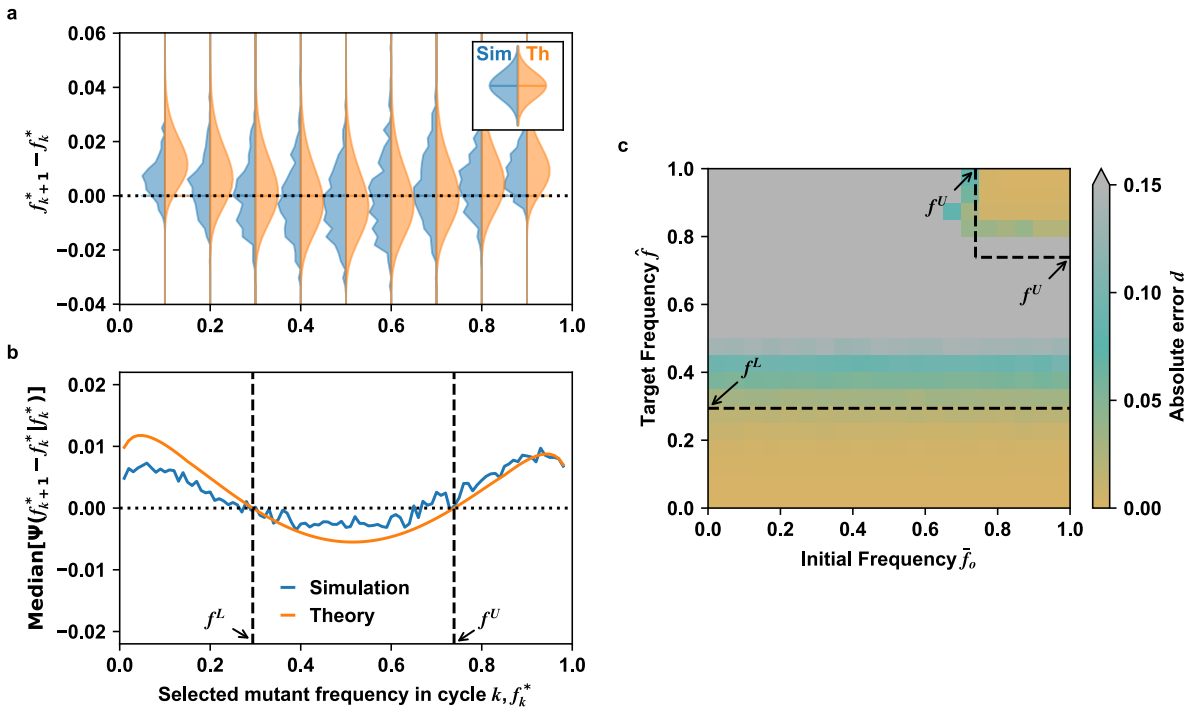

Fig. S9. Artificial selection also works for deleterious mutation. **a** Conditional Probability density functions of $f^*_{k+1} - f^*_k$ for various f^*_k values. The left-hand side distribution is obtained from simulations and the right-hand side distribution is numerically obtained by evaluating Eq. [43]. 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, $\mathrm{Median}[\Psi(f^*_{k+1} - f^*_k|f^*_k)] = 0$ 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 $\bar{f}_{1,0}$. 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**.

diagonal terms of $\mathbf{M}_{k+1}$ are variances $\sigma_X^2 = \overline{X^2} - [\overline{X}]^2$ and the offdiagonal terms are covariances $\sigma_{XY} = \overline{XY} - \overline{XY}$. The matrix is given by

$$\mathbf{M}_{k+1} = \begin{bmatrix} \frac{f_k^*(1-f_k^*)}{N_0^2} & -\frac{f_k^* h_k^*}{N_0^2} \\ -\frac{f_k^* h_k^*}{N_0^2} & \frac{h_k^*(1-h_k^*)}{N_0^2} \end{bmatrix} \equiv \begin{bmatrix} \sigma^2_{f_{k+1,0}} & \sigma_{f_{k+1,0}h_{k+1,0}} \\ \sigma_{f_{k+1,0}h_{k+1,0}} & \sigma^2_{h_{k+1,0}} \end{bmatrix}. \quad [51]$$

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 the joint probability distribution function is given by

$$P_{\boldsymbol{\rho}_{k+1,0}}(\boldsymbol{\rho}|\mathbf{p}_k^*) = \frac{1}{\sqrt{2\pi}^2\sqrt{\det \boldsymbol{M}_{k+1}}} e^{-\frac{1}{2}(\boldsymbol{\rho}-\mathbf{p}_k^*)\boldsymbol{M}_{k+1}^{-1}(\boldsymbol{\rho}-\mathbf{p}_k^*)^T}. \quad [52]$$

At the beginning of cycle k, a newborn collective starts to maturate 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. [49], we can write ordinary differential equations up to the second moment.

$$\frac{d\overline{S}(t)}{dt} = r\overline{S}(t) - \mu\overline{S}(t), \quad [53]$$

$$\frac{d\overline{F}(t)}{dt} = (r+s)\overline{F}(t) + \mu(\overline{S}(t) - \overline{F}(t)), \quad [54]$$

$$\frac{d\overline{FF}(t)}{dt} = (r+2s)\overline{FF}(t) + \mu\overline{F}(t), \quad [55]$$

$$\frac{d\overline{S^2}}{dt} = 2(r-\mu)\overline{S^2} + (r+\mu)\overline{S}, \quad [56]$$

$$\frac{d\overline{F^2}}{dt} = 2(r+s-\mu)\overline{F^2} + (r+s+\mu)\overline{F} + \mu(2\overline{SF} + \overline{F}), \quad [57]$$

$$\frac{d\overline{FF^2}}{dt} = 2(r+2s)\overline{FF^2} + (r+2s)\overline{FF} + \mu(2\overline{FFF} + \overline{F}), \quad [58]$$

$$\frac{d\overline{SF}}{dt} = (2r+s-2\mu)\overline{SF} + \mu(\overline{S^2} - \overline{S}), \quad [59]$$

$$\frac{d\overline{FFF}}{dt} = (2r+3s-\mu)\overline{FFF} + \mu(\overline{SFF} + \overline{F^2} - \overline{F}), \quad [60]$$

$$\frac{d\overline{SFF}}{dt} = (2r+2s-\mu)\overline{SFF} + \mu\overline{SF}. \quad [61]$$

The initial conditions of the system in coupled Eqs. [53-61] are obtained by the mean and (co)variances of Eq. [50]. 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

$$S(t) \approx \overline{S}(t) + \sigma_S(t)G_S(t), \quad [62]$$

$$F(t) \approx \overline{F}(t) + \sigma_F(t)G_F(t), \quad [63]$$

$$FF(t) \approx \overline{FF}(t) + \sigma_{FF}(t)G_{FF}(t). \quad [64]$$

Then, the F frequency becomes

$$\begin{aligned} f(t) &\approx \frac{\overline{F}(t) + \sigma_F(t)G_F(t)}{\overline{S}(t) + \sigma_S(t)G_S(t) + \overline{F}(t) + \sigma_F(t)G_F(t) + \overline{v}(t) + \sigma_{FF}(t)G_{FF}(t)} \\ &\approx \overline{f}(t) + \sigma_f(t)G_f(t), \end{aligned} \quad [65]$$

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

$$\begin{aligned} h(t) &\approx \frac{\overline{FF}(t) + \sigma_{FF}(t)G_{FF}(t)}{\overline{S}(t) + \sigma_S(t)G_S(t) + \overline{F}(t) + \sigma_F(t)G_F(t) + \overline{FF}(t) + \sigma_{FF}(t)G_{FF}(t)} \\ &\approx \overline{h}(t) + \sigma_h(t)G_h(t), \end{aligned} \quad [66]$$

where $\overline{h} = \overline{FF}/(\overline{S} + \overline{F} + \overline{FF})$ and $\sigma_h^2 = \left(\overline{h}^2(\sigma_S^2 + \sigma_F^2) + (1-\overline{h}^2)\sigma_{FF}^2\right)/(\overline{S} + \overline{F} + \overline{FF})$. 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

$$P_{\mathbf{p}_\tau}(\mathbf{p}|\boldsymbol{\rho}) = \frac{1}{\sqrt{2\pi}\sigma_{f_\tau}} e^{-\frac{(f-\overline{f}_\tau)^2}{2\sigma^2_{f_\tau}}} \frac{1}{\sqrt{2\pi}\sigma_{h_\tau}} e^{-\frac{(h-\overline{h}_\tau)^2}{2\sigma^2_{h_\tau}}}. \quad [67]$$

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

$$P_{\mathbf{p}_{k+1,\tau}}(\mathbf{p}|\mathbf{p}^*_k) = \int_0^1 d\zeta \int_0^1 d\eta P_{\mathbf{p}_{k+1,\tau}}(\mathbf{p}|\zeta,\eta) P_{\mathbf{p}_{k+1,0}}(\zeta,\eta|\mathbf{p}^*_k). \quad [68]$$

We select the Adult collective among g Adult collectives such that the change in frequencies during maturation could be compensate. During maturation, a freuquency distribution moves different direction 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. [55]. Thus, we choose the minimum value $h^{\min}_{k+1}$ 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 temporally 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

$$\begin{aligned}
C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&= (\int_0^1 - \int_f^1) df' (\int_0^1 - \int_h^1) dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&= \int_0^1 df' \int_0^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) - \int_h^1 dh' \int_0^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&\quad - \int_0^1 df' \int_h^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) + \int_f^1 df' \int_h^1 dh' P_{\mathbf{p}^*_{k+1}}(f',h'|\boldsymbol{p}^*_k) \\
&= 1 - \Pr[f^*_{k+1} \geq f|\boldsymbol{p}^*_k] - \Pr[h^*_{k+1} \geq h|\boldsymbol{p}^*_k] + \Pr[f^*_{k+1} \geq f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k]. \quad [69]
\end{aligned}$$

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

$$\begin{aligned}
\Pr[f^*_{k+1} \geq f \wedge h^*_{k+1} \geq h|\boldsymbol{p}^*_k] &= \Pr[f^{(1)}_{k+1} \geq f \wedge h^{(1)}_{k+1} \geq h \wedge f^{(2)}_{k+1} \geq f \wedge h^{(2)}_{k+1} \geq h \wedge \cdots \wedge f^{(g)}_{k+1} \geq f \wedge h^{(g)}_{k+1} \geq h|\boldsymbol{p}^*_k] \\
&= [\Pr[f_{k+1} \geq f \wedge h_{k+1} \geq h|\,\boldsymbol{p}^*_k]]^g \\
&= [1 - \Pr[f_{k+1} < f|\boldsymbol{p}^*_k] - \Pr[h_{k+1} < h|\boldsymbol{p}^*_k] + \Pr[f_{k+1} < f \wedge h_{k+1} < h]|\,\boldsymbol{p}^*_k]^g \\
&= \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_k) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^g, \quad [70]
\end{aligned}$$

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

$$C_{f_{k+1}}(f|\boldsymbol{p}^*_k) = \int_0^f df' \int_0^1 dh' P_{\mathbf{p}_{k+1}}(f',h'|\boldsymbol{p}^*_k), \quad [71]$$

$$C_{h_{k+1}}(h|\boldsymbol{p}^*_k) = \int_0^1 df' \int_0^h dh' P_{\mathbf{p}_{k+1}}(f',h'|\boldsymbol{p}^*_k). \quad [72]$$

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

$$C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) = 1 - \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k)\right]^g - \left[1 - C_{h_{k+1}}(h|\boldsymbol{p}^*_k)\right]^g + \left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_k) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^g. \quad [73]$$

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

$$\begin{aligned}
P_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}^*_{k+1}}(f,h|\boldsymbol{p}^*_k) \\
&= g(g-1)\left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-2} \quad [74] \\
&\quad \times (P_{f_k}(f|\boldsymbol{p}^*_{k-1}) - \partial_f C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k))(C_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) - \partial_h C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)) \\
&\quad + g\left[1 - C_{f_{k+1}}(f|\boldsymbol{p}^*_k) - C_{h_{k+1}}(h|\boldsymbol{p}^*_{k-1}) + C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k)\right]^{g-1} P_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}^*_k),
\end{aligned}$$

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 $f_{k+1}^* = \max(f_{k+1}^{(1)}, f_{k+1}^{(2)}, \cdots, f_{k+1}^{(g)})$. We rewrite the joint cumulative distribution function $C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*)$ to be little different from Eq. [69] because now we have to utilize the condition $f^* < f$ instead of $f^* > f$,

$$\begin{aligned} C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}_{k+1}^*}(f',h'|\boldsymbol{p}_k^*) \\ &= \int_0^f df' (\int_0^1 - \int_h^1) dh' P_{\mathbf{p}_{k+1}^*}(f',h'|\boldsymbol{p}_k^*) \\ &= \int_0^f df' \int_0^1 dh' P_{\mathbf{p}_{k+1}^*}(f',h'|\boldsymbol{p}_k^*) - \int_0^f df' \int_h^1 dh' P_{\mathbf{p}_{k+1}^*}(f',h'|\boldsymbol{p}_k^*). \\ &= \Pr[f_{k+1}^* < f|\boldsymbol{p}_k^*] - \Pr[f_{k+1}^* < f \wedge h_{k+1}^* \geq h|\boldsymbol{p}_k^*]. \end{aligned} \quad [75]$$

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

$$\begin{aligned} \Pr[f_{k+1}^* < f \wedge h_{k+1}^* \geq h|\boldsymbol{p}_k^*] &= \Pr[f_{k+1}^{(1)} < f \wedge h_{k+1}^{(1)} \geq h \wedge f_{k+1}^{(2)} < f \wedge h_{k+1}^{(2)} \geq h \wedge \cdots \wedge f_{k+1}^{(2)} < f \wedge h_{k+1}^{(2)} \geq h|\boldsymbol{p}_k^*] \\ &= [\Pr[f_{k+1} < f \wedge h_{k+1} \geq h|\boldsymbol{p}_k^*]]^g \\ &= [\Pr[f_{k+1} < f|\boldsymbol{p}_k^*] - \Pr[f_{k+1} < f \wedge h_{k+1} < h|\boldsymbol{p}_k^*]]^g \\ &= \left[C_{f_{k+1}}(f|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*)\right]^g. \end{aligned} \quad [76]$$

Thus, the joint cumulative distribution function is given by

$$C_{\mathbf{p}_{k+1}^*}(f,h|\boldsymbol{p}_k^*) = \left[C_{f_{k+1}}(f|\boldsymbol{p}_k^*)\right]^g - \left[C_{f_{k+1}}(f|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*)\right]^g. \quad [77]$$

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

$$\begin{aligned} P_{\mathbf{p}_{k+1}^*}(f,h|\boldsymbol{p}_k^*) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}_{k+1}^*}(f,h|\boldsymbol{p}_k^*) \\ &= g(g-1)\left[C_{f_{k+1}}(f|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*)\right]^{g-2} \\ &\quad \times (C_{f_{k+1}}(f|\boldsymbol{p}_k^*) - \partial_f C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*))\partial_h C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*) \\ &\quad + g\left[C_{f_{k+1}}(f|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*)\right]^{g-1} P_{\mathbf{p}_{k+1}}(f,h|\boldsymbol{p}_k^*). \end{aligned} \quad [78]$$

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

$$\Psi(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*) = \begin{cases} g(g-1)\left[1 - C_{f_{k+1}}(f_{k+1}^*|\boldsymbol{p}_k^*) - C_{h_{k+1}}(h_{k+1}^*|\mathbf{p}_k^*) + C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*)\right]^{g-2} \\ \times (P_{f_{k+1}}(f_{k+1}^*|\mathbf{p}_k^*) - \partial_f C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*))(P_{h_{k+1}}(h) - \partial_h C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*)) \\ + g\left[1 - C_{f_{k+1}}(f_{k+1}^*|\boldsymbol{p}_k^*) - C_{h_{k+1}}(h_{k+1}^*|\boldsymbol{p}_k^*) + C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*)\right]^{g-1} P_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*) \\ \qquad \text{for } \overline{f_{k+1,\tau}} - f_k^* \geq 0 \text{ and } \overline{h_{k+1,\tau}} - h_k^* \geq 0 \\ g(g-1)\left[C_{f_{k+1}}(f_{k+1}^*|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\boldsymbol{p}_k^*)\right]^{g-2} \\ (P_{f_{k+1}}(f_{k+1}^*|\mathbf{p}_k^*) - \partial_f C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*))\partial_h C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*) \\ + g\left[C_{f_{k+1}}(f_{k+1}^*|\boldsymbol{p}_k^*) - C_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*)\right]^{g-1} P_{\mathbf{p}_{k+1}}(\mathbf{p}_{k+1}^*|\mathbf{p}_k^*) \\ \qquad \text{for } \overline{f_{k+1,\tau}} - f_k^* < 0 \text{ and } \overline{h_{k+1,\tau}} - h_k^* \geq 0 \end{cases}. \quad [79]$$

Using Eq. [79], we get the mean values of f^* and h^* as

$$\overline{f_{k+1}^*} = \int_0^1 df' \int_0^1 dh' \; f' \times P_{f_{k+1}^*, h_{k+1}^*}(f',h'|f_k^*,h_k^*), \quad [80]$$

$$\overline{h_{k+1}^*} = \int_0^1 df' \int_0^1 dh' \; h' \times P_{f_{k+1}^*, h_{k+1}^*}(f',h'|f_k^*,h_k^*). \quad [81]$$

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. S10),

$$\text{sign}\left(\overline{f_{k+1}^*} - f_k^*\right) \times \text{sign}\left(\overline{f_{k+1,\tau}} - f_k^*\right) \leq 0 \text{ and } \text{sign}\left(\overline{h_{k+1}^*} - h_k^*\right) \times \text{sign}\left(\overline{h_{k+1,\tau}} - h_k^*\right) \leq 0, \quad [82]$$

where $\overline{f_{k+1,\tau}}$ and $\overline{h_{k+1,\tau}}$ are the mean values after the maturation step in cycle $k+1$. Figure S10**a** shows a concise explanation of the classification of regions where the initial and target compositions satisfy the above conditions in Eq. [82]. 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 gray-colored area in Fig. S10**b**. Similar to the case of the two-genotype population, the accessible region is shaped by the flow velocity of the composition in the maturation step, as depicted in the flow diagram in Fig. S10**b**. 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.

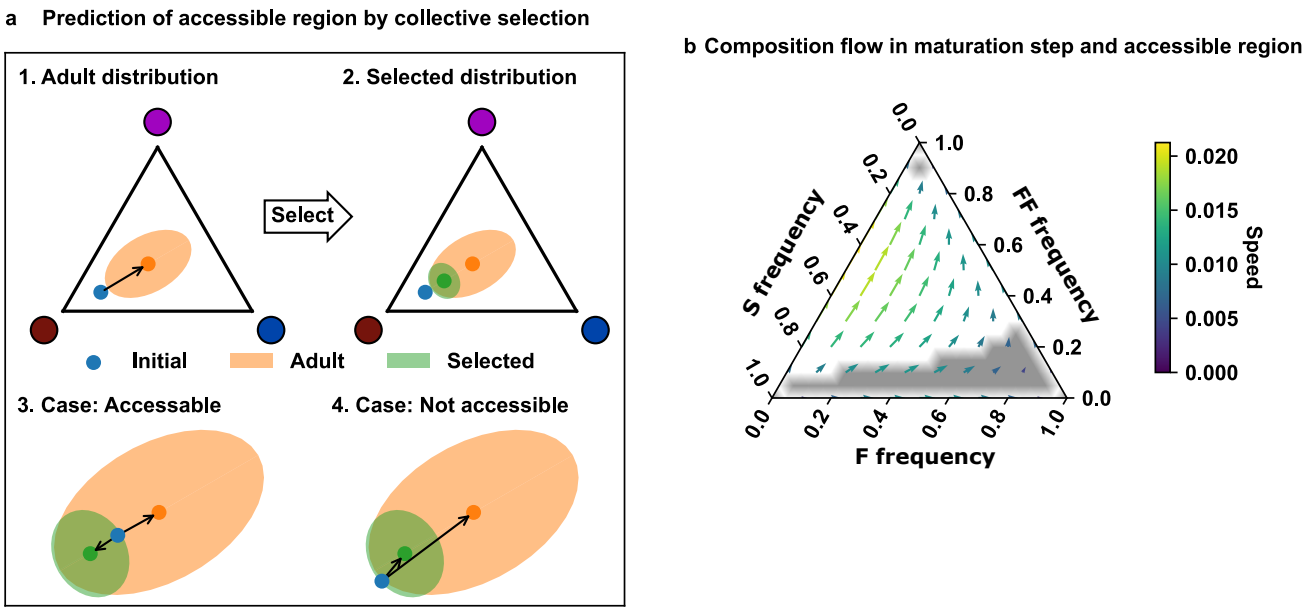

Fig. S10. **a** The schematic procedure to determine accessible regions by collective selection. From the given parent composition (blue dot), the probability distribution of offspring Adults (Eq. [68]) is computed (marked in the orange-colored area). From Adult compositions, the probability distribution of the selected collective (Eq.[79]) is computed (marked in the green area). If the signs of the changes in both F frequency and FF frequency after the selection (from blue dot to green dot) are opposite to that of maturation (from blue dot to orange dot), the given composition is accessible. Otherwise, the composition is not accessible and will change after cycles. **b** The accessible regions are marked by the gray area. The vector field is the flow of compositions during maturation. The length and color of the arrows indicate the speed of composition changes. The figures are drawn using *mpltern* package (7).

7. Derivation of equations

In this section, we go over the derivation of Eq. [7-27] for readers not equipped with advanced mathematics training.

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

Equations [7] and [8]. Equation [7] is straightforwardly solved by integrating Eq. [5]. Equation [8] is obtained from Eq. [6] using integration factor $e^{-(r+\omega)t}$:

$$\left[\frac{d\overline{F}(t)}{dt} - (r+s)\overline{F}(t)\right] e^{-(r+s)t} = \mu\overline{S}_o e^{(r-\mu)t} e^{-(r+s)t}, \tag{83}$$

$$\frac{d\left[\overline{F}(t)e^{-(r+s)t}\right]}{dt} = \mu\overline{S}_o e^{-(s+\mu)t}. \tag{84}$$

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

$$\overline{F}(t) = \overline{F}_o e^{(r+s)t} + \frac{\mu\overline{S}_o}{s+\mu}\left(e^{(r+s)t} - e^{(r-\mu)t}\right) \approx \overline{F}_o e^{(r+s)t} + \frac{\mu\overline{S}_o}{s}\left(e^{(r+s)t} - e^{rt}\right). \tag{85}$$

Equations [11] and [12]. Applying Eq. [4], we have

$$\begin{aligned}\frac{d\overline{S^2}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} S^2 \frac{dP(S,F,t)}{dt} \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[rS^2(S-1)P(S-1,F,t) + S^2(r+s)(F-1)P(S,F-1,t) \\ &\quad +\mu(\mathrm{S}+1)S^2P(S+1,F-1,t) - (rS + (r+s)F + \mu S)S^2P(S,F,t)].\end{aligned} \tag{86}$$

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 $\sum_{F=0}^{\infty}\sum_{S=1}^{\infty}\left[rS^2(S-1)P(S-1,F,t)\right]$. Let $\alpha = S-1$, and this becomes $\sum_{F=0}^{\infty}\sum_{\alpha=0}^{\infty}\left[r(\alpha+1)^2\alpha P(\alpha,F,t)\right]$. We reassign α as S, and obtain:

$$\begin{aligned}&\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[r(S+1)^2 S\, P(S,F,t)\right] - \sum_{F=0}^{\infty}\sum_{S=0}^{\infty} rS^3P(S,F,t) \\ &= r\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[(2S+1)S\right]P(S,F,t) \\ &= r(2\overline{S^2} + \overline{S}).\end{aligned} \tag{87}$$

We collect the two green terms, and similarly obtain:

$$\begin{aligned}&\sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[S^2(r+s)(F-1)P(S,F-1,t) - (r+s)FS^2P(S,F,t)\right] \\ &= (r+s)\sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[S^2F\,P(S,F,t) - FS^2P(S,F,t)\right] \\ &= 0.\end{aligned} \tag{88}$$

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

$$\begin{aligned}&\mu\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[S(S-1)^2P(S,F,t) - S^3P(S,F,t)\right] \\ &= \mu\sum_{F=0}^{\infty}\sum_{S=0}^{\infty}\left[(-2S^2+S)P(S,F,t)\right] \\ &= \mu\left(-2\overline{S^2} + \overline{S}\right).\end{aligned} \tag{89}$$

Now, add the three parts together, and we have

$$\begin{aligned}\frac{d\overline{S^2}}{dt} &= \mu\left(-2\overline{S^2}+\overline{S}\right)+r(2\overline{S^2}+\overline{S}) \\ &= 2\overline{S^2}(r-\mu)+\overline{S}(r+\mu), \end{aligned} \quad [90]$$

which is Eq. [11]. Likewise,

$$\begin{aligned}\frac{d\overline{F^2}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty} F^2\frac{dP(S,F,t)}{dt} \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[rF^2(S-1)P(S-1,F,t)+F^2(r+s)(F-1)P(S,F-1,t) \\ &\quad +\mu(S+1)F^2P(S+1,F-1,t)-(rS+(r+s)F+\mu S)F^2P(S,F,t)] \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[rF^2S+(F+1)^2(r+s)F+\mu S(F+1)^2-(rS+(r+s)F+\mu S)F^2\right]P(S,F,t) \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[(F+1)^2(r+s)F+\mu S(F+1)^2-((r+s)F+\mu S)F^2\right]P(S,F,t) \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[(r+s)F(2F+1)+\mu S(2F+1)\right]P(S,F,t) \quad [91] \\ &= 2(r+s)\overline{F^2}+(r+s)\overline{F}+2\mu\overline{SF}+\mu\overline{S} \quad [92]\end{aligned}$$

which is Eq. [12].

Equation [13] and [15]. Using integration factor $e^{-2(r-\mu)t}$ and Eq. [11], we have:

$$\left[\frac{d\overline{S^2}}{dt}-2(r-\mu)\overline{S^2}\right]e^{-2(r-\mu)t} = (r+\mu)\overline{S}_o e^{(r-\mu)t}e^{-2(r-\mu)t}, \quad [93]$$

$$\frac{d\left[\overline{S^2}e^{-2(r-\mu)t}\right]}{dt} = (r+\mu)\overline{S}_o e^{(\mu-r)t}, \quad [94]$$

$$\overline{S^2}e^{-2(r-\mu)t} = \frac{r+\mu}{\mu-r}\overline{S}_o\left[e^{(\mu-r)t}-1\right]+\overline{S^2}_o. \quad [95]$$

Since $\mu \ll r$,we have

$$\begin{aligned}\overline{S^2} &= \frac{r+\mu}{\mu-r}\overline{S}_o\left[e^{(\mu-r)t}-1\right]e^{2(r-\mu)t}+\overline{S^2}_o e^{2(r-\mu)t} \\ &\approx \overline{S}_o e^{2rt}\left[1-e^{-rt}\right]+\overline{S^2}_o e^{2rt}, \end{aligned} \quad [96]$$

where $\overline{S^2}_o$ is the expected S^2 at time 0. For Eq. [15],

$$\begin{aligned}\sigma_S^2(t) &= \overline{S^2}(t)-[\overline{S}(t)]^2 \\ &\approx \overline{S}_o e^{2rt}\left[1-e^{-rt}\right]+\overline{S^2}_o e^{2rt}-\left(\overline{S}_o e^{rt}\right)^2 \\ &= \overline{S}_o e^{2rt}\left[1-e^{-rt}\right]+e^{2rt}\left(\overline{S^2}_o-\left(\overline{S}_o\right)^2\right)=\overline{S}_o e^{2rt}\left[1-e^{-rt}\right]+e^{2rt}\sigma_S^2(0). \end{aligned} \quad [97]$$

Equation [16] and [17]. Since $\overline{S\,F}=\sum_{S=0}^{\infty}\sum_{F=0}^{\infty}S\,F\,P(w,m,t)$,

$$\begin{aligned}\frac{d\overline{S\,F}}{dt} &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[S\,Fr(S-1)P(S-1,F,t)+S\,F(r+s)(F-1)P(S,F-1,t) \\ &\quad +S\,F\mu(S+1)P(S+1,F-1,t) \\ &\quad -S\,F(rS+(r+s)F+\mu S)P(S,F,t)] \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}[(S+1)FrS+S(F+1)(r+s)F+(S-1)(F+1)\mu S \\ &\quad -S\,F(rS+(r+s)F+\mu S)]P(S,F,t) \\ &= \sum_{S=0}^{\infty}\sum_{F=0}^{\infty}\left[2rS\,F+sS\,F-\mu S\,F+\mu S^2-\mu S)\right]P(S,F,t) \\ &= (2r+s-\mu)\overline{S\,F}+\mu(\overline{S^2}-\overline{S}). \end{aligned} \quad [98]$$

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

$$\frac{d[\overline{S\,F}e^{-(2r+s-\mu)t}]}{dt} = \mu(\overline{S^2}-\overline{S})e^{-(2r+s-\mu)t}. \quad [99]$$

Thus, we have

$$\begin{aligned}
\overline{S\,F}e^{-(2r+s-\mu)t} - \overline{S\,F}_o &= \mu\int(\overline{S^2}-\overline{S})e^{-(2r+s-\mu)t}dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{S}_o\left[e^{(\mu-r)t}-1\right]e^{2(r-\mu)t}+\overline{S^2}_o e^{2(r-\mu)t}-\overline{S}_o e^{(r-\mu)t})e^{-(2r+s-\mu)t}dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{S}_o\left[e^{-(\mu-r)t}-e^{2(r-\mu)t}\right]e^{-(2r+s-\mu)t}+\overline{S^2}_o e^{2(r-\mu)t}e^{-(2r+s-\mu)t} \\
&\quad -\overline{S}_o e^{(r-\mu)t-(2r+s-\mu)t})dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{S}_o e^{-(r+s)t}-\frac{r+\mu}{\mu-r}\overline{S}_o e^{-(\mu+s)t}+\overline{S^2}_o e^{-(\mu+s)t}-\overline{S}_o e^{-(r+s)t})dt \\
&= \mu\left\{\frac{2r}{\mu-r}\overline{S}_o\frac{e^{-(r+s)t}-1}{-(r+s)}+(\overline{S^2}_o-\frac{r+\mu}{\mu-r}\overline{S}_o)\frac{e^{-(\mu+s)t}-1}{-(\mu+s)}\right\}, \quad [100]
\end{aligned}$$

which results in

$$\begin{aligned}
\overline{S\,F}(t) &= \overline{S\,F}_o e^{(2r+s-\mu)t}+\mu\left\{\frac{2r}{\mu-r}\overline{S}_o\frac{e^{-(r+s)t}-1}{-(r+s)}+(\overline{S^2}_o-\frac{r+\mu}{\mu-r}\overline{S}_o)\frac{e^{-(\mu+s)t}-1}{-(\mu+s)}\right\}e^{(2r+s-\mu)t} \\
&= \overline{S\,F}_o e^{(2r+s-\mu)t}+\frac{2r\mu}{\mu-r}\overline{S}_o\frac{e^{(r-\mu)t}-e^{(2r+s-\mu)t}}{-(r+s)}+\mu(\overline{S^2}_o-\frac{r+\mu}{\mu-r}\overline{S}_o)\frac{e^{2(r-\mu)t}-e^{(2r+s-\mu)t}}{-(\mu+s)} \\
&\approx \overline{S\,F}_o e^{(2r+s)t}-\frac{2\mu\overline{S}_o}{r+s}\left(e^{(2r+s)t}-e^{rt}\right)+\frac{\mu}{s}(\overline{S^2}_o+\overline{S}_o)\left(e^{(2r+s)t}-e^{2rt}\right). \quad [101]
\end{aligned}$$

Equation [18]. From Eq. [12], we have

$$\left(\frac{d\overline{F^2}}{dt}-2(r+s)\overline{F^2}\right)e^{-2(r+s)t} = \frac{d\left(\overline{F^2}e^{-2(r+s)t}\right)}{dt} = \left((r+s)\overline{F}+2\mu\overline{S\,F}+\mu\overline{S}\right)e^{-2(r+s)t}. \quad [102]$$

The right-hand side becomes

$$\begin{aligned}
&\approx(r+s)\left(\overline{F}_o e^{(r+s)t}+\frac{\mu\overline{S}_o}{s+\mu}\left(e^{(r+s)t}-e^{(r-\mu)t}\right)\right)e^{-2(r+s)t}+2\mu\overline{S\,F}_o e^{-(s+\mu)t}+\mu\overline{S}_o e^{-(r+2s+\mu)t}+o(\mu^2) \\
&=(r+s)\left(\overline{F}_o e^{-(r+s)t}+\frac{\mu\overline{w}_o}{s+\mu}\left(e^{-(r+s)t}-e^{-(r+2s+\mu)t}\right)\right)+2\mu\overline{S\,F}_o e^{-(s+\mu)t}+\mu\overline{S}_o e^{-(r+2s+\mu)t}+o(\mu^2) \\
&=(r+s)\left(\frac{\overline{F}_o s+N_0\mu}{s+\mu}e^{-(r+s)t}-\frac{\mu\overline{S}_o}{s+\mu}e^{-(r+2s+\mu)t}\right)+2\mu\overline{S\,F}_o e^{-(s+\mu)t}+\mu\overline{S}_o e^{-(r+2s+\mu)t}+o(\mu^2). \quad [103]
\end{aligned}$$

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

$$\begin{aligned}
\overline{F^2}e^{-2(r+s)t}-\overline{F^2}_o &\approx (r+s)\left(\frac{\overline{F}_o s+N_0\mu}{(s+\mu)(r+s)}(1-e^{-(r+s)t})+\frac{\mu\overline{S}_o}{(s+\mu)(r+2s+\mu)}(e^{-(r+2s+\mu)t}-1)\right) \\
&\quad +\frac{2\mu\overline{S\,F}_o}{s+\mu}(1-e^{-(s+\mu)t})+\frac{\mu\overline{S}_o}{r+2s+\mu}(1-e^{-(r+2s+\mu)t})+o(\mu^2). \quad [104]
\end{aligned}$$

Then, we have

$$
\begin{aligned}
\overline{F^2}(t) &\approx \overline{F^2}_o e^{2(r+s)t} + (r+s)\left(\frac{\overline{F}_o s + N_0\mu}{(s+\mu)(r+s)}(e^{2(r+s)t} - e^{(r+s)t}) + \frac{\mu\overline{S}_o}{(s+\mu)(r+2s+\mu)}(e^{(r-\mu)t} - e^{2(r+s)t})\right) \\
&\quad + \frac{2\mu\overline{S\,F}_o}{s+\mu}(e^{2(r+s)t} - e^{(2r+s-\mu)t}) + \frac{\mu\overline{S}_o}{r+2s+\mu}(e^{2(r+s)t} - e^{(r-\mu)t}) + o(\mu^2) \\
&= \overline{F^2}_o e^{2(r+s)t} + \left(\frac{\overline{F}_o s + N_0\mu}{s}(e^{2(r+s)t} - e^{(r+s)t}) + (\frac{1}{s} - \frac{1}{r+2s})\mu\overline{S}_o(e^{rt} - e^{2(r+s)t})\right) \\
&\quad + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \frac{\mu\overline{S}_o}{r+2s}(e^{2(r+s)t} - e^{rt}) + o(\mu^2) \\
&= \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{\overline{F}_o}{\overline{S}_o}(e^{2(r+s)t} - e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&= \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + \frac{\mu\overline{F}_o}{s}(e^{2(r+s)t} - e^{(r+s)t}) + o(\mu^2) \\
&= \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o(1 + \frac{\mu}{s})e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&= \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \qquad [105] \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2).
\end{aligned}
$$

Equation [19].

$$
\begin{aligned}
\sigma_F^2(t) &= \overline{F^2}(t) - [\overline{F}(t)]^2 \\
&\approx \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{S\,F}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&\quad - \left[\overline{F}_o e^{(r+s)t} + \frac{\mu\overline{S}_o}{s}(e^{(r+s)t} - e^{rt})\right]^2 \\
&= \overline{F^2}_o e^{2(r+s)t} + \frac{2\mu\overline{SF}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&\quad - \overline{F}_o^2 e^{2(r+s)t} - 2\overline{F}_o e^{(r+s)t}\frac{\mu\overline{S}_o}{s}(e^{(r+s)t} - e^{rt}) - \left[\frac{\mu\overline{S}_o}{s}(e^{(r+s)t} - e^{rt})\right]^2 \\
&= \sigma_{F,0}^2 e^{2(r+s)t} + \frac{2\mu(\overline{S\,F}_o - \overline{S}_o\overline{F}_o)}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{F}_o e^{(r+s)t}(e^{(r+s)t} - 1) \qquad [106] \\
&\quad + \frac{\mu\overline{S}_o}{s}\left[-e^{(r+s)t} + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2).
\end{aligned}
$$

Equations [25]-[27]. 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.

$$\begin{aligned} f &\approx \frac{\overline{F}+\sigma_F G_F}{\overline{S}+\sigma_S G_S+\overline{F}+\sigma_F G_F} \\ &= \frac{\overline{F}\left(1+\frac{\sigma_F}{\overline{F}}G_F\right)}{(\overline{S}+\overline{F})(1+\frac{\sigma_S G_S+\sigma_F G_F}{\overline{S}+\overline{F}})} \\ &\approx \frac{\overline{F}}{\overline{S}+\overline{F}}\left(1+\frac{\sigma_F}{\overline{F}}G_F\right)\left(1-\frac{\sigma_S G_S+\sigma_F G_F}{\overline{S}+\overline{F}}\right) \\ &\approx \overline{f}\left(1+\frac{\sigma_F}{\overline{F}}G_F-\frac{\sigma_S G_S+\sigma_F G_F}{\overline{S}+\overline{F}}\right) \\ &= \overline{f}\left(1+\left(\frac{\sigma_F}{\overline{F}}-\frac{\sigma_F}{\overline{S}+\overline{F}}\right)G_F-\frac{\sigma_S}{\overline{S}+\overline{F}}G_S\right). \end{aligned} \tag{107}$$

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

$$\begin{aligned} \overline{f}(t) &= \frac{\overline{F}(t)}{\overline{S}(t)+\overline{F}(t)} \approx \frac{\overline{F}_o e^{(r+s)t}+\frac{\mu\overline{S}_o}{s}(e^{(r+s)t}-e^{rt})}{\overline{S}_o e^{rt}+\overline{F}_o e^{(r+s)t}+\frac{\mu\overline{S}_o}{s}(e^{(r+s)t}-e^{rt})} \\ &= \frac{\overline{F}_o e^{st}+\frac{\mu\overline{S}_o}{s}(e^{st}-1)}{\overline{S}_o+\overline{F}_o e^{st}+\frac{\mu\overline{S}_o}{s}(e^{st}-1)} \\ &= \frac{\overline{f}_o e^{st}+\frac{\mu}{s}(1-\overline{f}_o)(e^{st}-1)}{(1-\overline{f}_o)+\overline{f}_o e^{st}+\frac{\mu}{s}(1-\overline{f}_o)(e^{st}-1)}. \end{aligned} \tag{108}$$

Note that the initial value of mean $\overline{f}(0)$ is equal to the mean of the binomial distribution Eq. [22], f_k^*. The variance is

$$\begin{aligned} \sigma_f^2(t) &= \overline{f}^2\left(\left(\frac{\overline{S}}{(\overline{S}+\overline{F})\overline{F}}\right)^2\sigma_F^2(t)+\left(\frac{1}{\overline{S}+\overline{F}}\right)^2\sigma_S^2(t)\right) \\ &= \overline{f}^2\left(\frac{(1-\overline{f}(t))^2}{\overline{N}(t)^2(\overline{f}(t))^2}\sigma_F(t)+\frac{1}{\overline{N}(t)^2}\sigma_S^2(t)\right) \\ &= \frac{(1-\overline{f}(t))^2\sigma_F^2(t)+\overline{f}(t)^2\sigma_S^2(t)}{\overline{N}(t)^2}. \end{aligned} \tag{109}$$

References

1. DT Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. *The J. Chem. Phys.* **115**, 1716–1733 (2001).
2. Y Cao, DT Gillespie, LR Petzold, Efficient step size selection for the tau-leaping simulation method. *The J. Chem. Phys.* **124**, 044109 (2006).
3. Q Zheng, Progress of a half century in the study of the luria–delbrück distribution. *Math. Biosci.* **162**, 1–32 (1999).
4. EJ Gumbel, *Statistics of extremes.* (Columbia university press), (1958).
5. J Philip, The function inverfc theta. *Aust. J. Phys.* **13**, 13 (1960).
6. L Xie, AE Yuan, W Shou, Simulations reveal challenges to artificial community selection and possible strategies for success. *PLoS Biol.* **17**, e3000295 (2019).
7. Y Ikeda, B Grabowski, F Körmann, mpltern 0.3.0: ternary plots as projections of Matplotlib (2019).