Condition for successful artificial selection on microbial collectives and optimal experimental setup

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

[a]Departement of Physics, Inha University, Incheon 22212, Republic of Korea ; [b]Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea ; [c]Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle WA 98109-1024, Washington, United States of America; [d]Division of Biosciences, University College London, London WC1E 6BT, United Kingdom

This manuscript was compiled on March 7, 2023

Microbial collectives – communities of microbes of the same or different species – can carry out functions that are often not achievable by individual microbes in isolation. Examples of collective functions include waste degradation, food fermentation, and vitamin production that are beneficial to humans. A human-desired collective function might be achieved by artificial selection on collectives. However, such selection on collectives is challenging due to the lack of theoretical understanding. Here, we examine a simple microbial system to develop a theoretical approach for understanding the conditions for successful artificial selection on collectives. In each selection cycle, multiple low-density "Newborn" collectives with wild and mutant cells are allowed to "mature" during which cells grow and possibly mutate. At the end of a cycle, "Adult" collective with the highest function is selected and is split into multiple Newborns for the next cycle. Repeating selection cycles, we aim to obtain the target composition that gives an optimal collective function. Investigating the possibility to get closer the target composition through artificial selection, we predict whether the selection is successful at given experimental setups. This approach allows us to find which experimental parameters are important for successful selection. Our results show that small bottleneck sizes play an important role while the effect of increasing collective numbers is not always significant.

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

Microbial collectives can carry out functions that single species cannot. For example, marine microorganisms work together to degrade chitin (1, 2). Collective functions include human-desired traits such as waste degradation (3) and vitamin production (4). Consequently, the evolution of collective functions has recently drawn great interests. As a way of modifying collective functions, artificial selection or directed evolution on collectives (5–8) has been suggested. The main idea is to allow low-density "Newborn" collectives to mature so that mutations can arise and collective functions can be evolved. One then chooses high-function "Adult" collectives to reproduce by splitting each into multiple offspring Newborns. Experimental and numerical studies have tested artificial selection of collectives (9–26). While some of the selection efforts seem successful to improve collective functions, others highlight challenges in selecting collectives (9, 10, 12–16, 18, 20). Even under identical experimental conditions, replicating experiments show the opposite results in (10): some improve collective functions and others are insensitive to the selection.

Recent theoretical studies have striven to identify essential ingredients for effective selection of collectives (21–26). Doulcier *et. al.* have shown that any target species composition can be achieved through directed evolution on collectives of two competing species except for extreme cases (21). This success is based on the evolution of species' traits such as growth rates and interaction strengths. If the trait space is unlimited, there will be trait sets that give the target composition at equilibrium. Thus the target composition can be achieved when the trait sets can be selected. However, if no such sets exist, the protocol used in (21) does not work to obtain the target composition. Hence short-term evolution experiments (possible traits are fixed) or the cases where the range of trait space is limited demand another mechanism or theory for the successful selection.

Here, we consider a situation where the trait space is fixed and any possible equilibrium composition cannot give a target composition. For simplicity, we use single-species collectives with two genotypes, wild and mutant genotypes with different growth rates. We show that the target genotype composition can be selected when the artificial selection is strong enough to offset composition changes during maturation. By obtaining a mathematical expression of the composition distribution after a selection cycle, we suggest a condition for successful artificial selection. Furthermore, we investigate the optimal bottleneck size and collective numbers. Although we only focus on faster-growing mutants in the main text, our approach is applicable to any situations where the target composition is different from the composition at equilibrium.

Collectives are composed of wildtype and mutant type cells (see Fig. 1). We consider the case where mutation is beneficial on growth. The case of mutant growing slower than wildtype is shown in Sec. 3 in Supplementary information, and not surprisingly, results are robust.

Each collective grows from N_0 cells. In the first cycle, the average mutant frequency is $\overline{f}_0$ (e.g. the initial mutant number in each collective is drawn from the binomial distribution $\mathrm{Binom}(N_0, \overline{f}_0)$). We denote the numbers of wildtype and mutant cells in collective i as $w_{k,t}^{(i)}$ and $m_{k,t}^{(i)}$, respectively. At the initial of each cycle k ($t = 0$), $w_{k,0}^{(i)} + m_{k,0}^{(i)} = N_0$. Collectives are allowed to grow for time τ ('Maturation' in Fig. 1). During maturation, wildtype and mutant cells grow at rates r and $r + s$, respectively where $s > 0$. Wildtype cells mutate at rate μ. If maturation time τ is too small, matured collectives

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: w.shou@ucl.ac.uk and E-mail: hyejin.park@inha.ac.kr

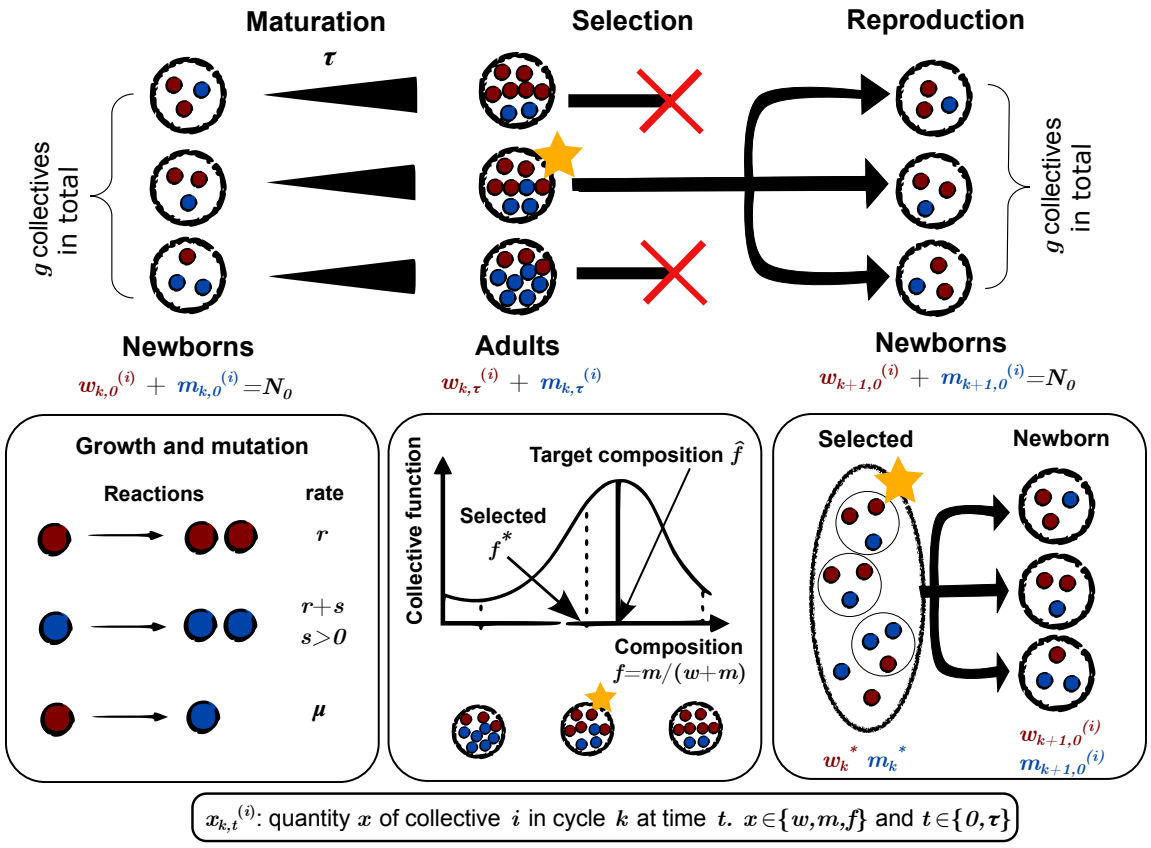

Fig. 1. Schematic figure for artificial selection on collectives. One cycle of the protocol consists of three stages: maturation, selection, and reproduction. Each cycle begins with a total of g Newborn collectives (large black circles) with N_0 total cells per Newborn. For simplicity, we consider two types of cells in collectives called wildtype (red-colored dots) and mutant type (blue-colored dots). During maturation (over time τ), wildtype and mutant cells divide at rates r and $r+s$ ($s>0$), respectively. Wildtype mutates to a mutant genotype at rate μ. In the selection stage, the Adult collective with the highest function is chosen (marked by yellow stars) to reproduce Newborns for the next cycle. Here, we assume that the collective function only depends on the frequency of mutants $f=m/(w+m)$ (w and m are the numbers of wildtype and mutant type cells) and has the maximum value at a target composition $\hat{f}$. The collective with the closest value to $\hat{f}$ is chosen to reproduce Newborns for the next cycle. Newborns are sampled from the chosen Adult with N_0 cells per each. By repeating cycles, we could obtain the target composition, which optimizes the collective function. To denote a quantity x in i-th collective at cycle k with time t ($0 \le t \le \tau$), we use notation $x_{k,t}^{(i)}$ where $x \in \{w,\ m,\ f\}$. Note that time $t=0$ is for Newborns and $t=\tau$ is for Adults.

("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 mutants will take over. Hence we set the maturation time $\tau = \log(g+1)/r$, which guarantees to produce g Newborn collectives from a single Adult. At the end of the cycle, the Adult with the highest function is chosen to reproduce g Newborn collectives with N_0 cells ('Selection' and 'Reproduction' in Fig. 1). We assume that collective function is maximized at the target composition $\hat{f}$ and the more similar the composition to $\hat{f}$, the higher the function of a collective. The chosen Adult has genotype composition f^* closest to the target value $\hat{f}$ among all Adults. For the next cycle, g Newborn collectives with N_0 cells are sampled from the selected Adult collective ('Reproduction' in Fig. 1). The sampled mutant cell numbers in Newborns follow a binomial distribution with the mean value $N_0 f^*$. By repeating the cycles, we aim to achieve the target composition $\hat{f}$ which gives to the highest collective function.

Trajectories of collective composition. First, we obtain composition trajectories at a given target and initial compositions by implementing stochastic simulations (see Supplementary Information Sec. 1 for the details). Figure 2**a** shows the success of artificial selection. The selected collective's composition gets closer to the target composition and stays around it while mutants fixate in collectives without collective-level selection. Compared to the average compositions without artificial selection, the selected collective composition reaches the target value faster (see Fig. 2**b**). More importantly, artificial selection stabilizes the collectives' composition around the target composition for a long time.

However, not all target values can be achievable or maintainable. When artificial selection cannot cancel out the mutant frequency increase during maturation, a target composition cannot be obtained or maintained. We examine various target frequencies whether artificial selection is effective (see Fig. 2**c**). With a low initial composition ($\overline{f}_0 = 0.05$), artificial selection works for the low and high target values ($\hat{f}=0.1$ and $\hat{f}=0.9$). In contrast, the intermediate target value ($\hat{f}=0.5$) is initially achieved but the composition cannot be maintained, and the variance of $\langle f^* \rangle$ over cycles increases. Such an achievable target range also depends on the initial composition. For a higher initial mutant frequency $\overline{f}_0 = 0.5$, the low target value $\hat{f}=0.1$ now becomes unachievable (see Fig. 2**d**).

Altogether, artificial selection is successful in two regions as in Fig. 3: (1) high target frequencies (above around 0.7) and (2) low target frequencies with low initial frequencies (below around 0.3). We denote "success" as the relative error d

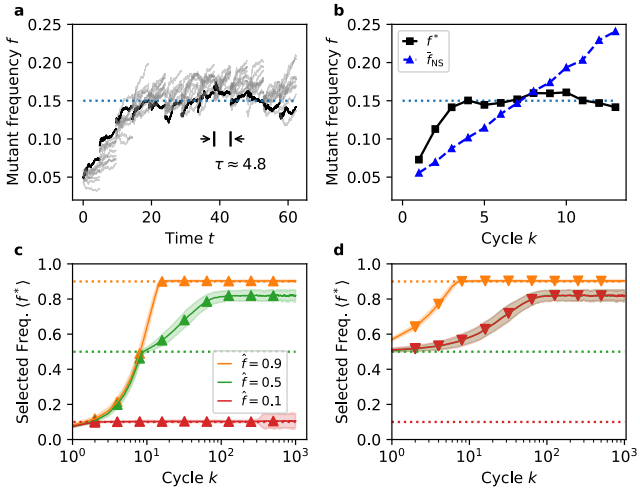

Fig. 2. Initial and target compositions determine the success of artificial selection on collectives. a Trajectories of Mutant frequency for 10 collectives ($g = 10$) over time. The collective whose frequency is closest to the target value is selected in every cycle (black lines). The gray lines denote the other collectives. For parameters, we used $\overline{f}_0 = 0.05$, $r = 0.5$, $s = 0.03$, $\mu = 0.0001$ and $\tau \approx 4.8$. With these parameters, the target collective composition $\hat{f} = 0.15$ can be achieved. **b** Comparison between frequency trajectories with and without selection clearly shows the effect of artificial selection. The black line indicates mutant frequency of the selected collective f_k^* at each cycle in **a**. The blue line indicates the average trajectory without selection $\overline{f}_{\rm NS}$ (the average of $g = 10$ individual lineages without collective-level selection at the end of each cycle that the Newborns are diluted from g Adults with N_0 cells.) The f^* approaches the target value faster than $\overline{f}_{\rm NS}$ and stabilizes its value around the target. Hereafter, we use cycle time k to present the results. **c** and **d** show the composition of selected collectives with various $\hat{f}$ for $\overline{f}_0 = 0.1$ and $\overline{f}_0 = 0.5$, respectively. Each lines are averaged for independent 300 realizations. The shaded area indicates the standard deviation. Even though artificial selection alters the frequency trajectory, it does not guarantee to select the collectives with the target composition.

between the target frequency $\hat{f}$ and the average of selected one $\langle f^* \rangle$ is smaller than 5% (i.e. $d = |\langle f^* \rangle - \hat{f}|/\hat{f} \leq 5\%$, computed relative errors are shown in Supplementary Information Fig. S3.) Otherwise, artificial selection cannot compensate for the mutant frequency increase during maturation. We analytically calculate the boundaries of regions, f^U and f^L, that match well with the simulation results in Fig. 3. Below, we explain how to get the condition under which a target frequency can be successfully achieved by artificial selection in detail.

Boundary of artificial-selection-dominant region. To achieve a target mutant frequency, inter-collective selection must counter the rise in frequency of fast-growing mutants during collective maturation. Specifically, we consider the challenging case where the mutant frequency of the selected Adult f_k^* is above the target value at cycle k ($f_k^* > \hat{f}$). Then, Newborns of cycle $k+1$ will have mutant frequencies fluctuating around f_k^*. If all Adult collectives have larger frequencies than the target value, then the minimum mutant frequency among the adults ($f_{k+1}^{\rm min}$) is selected ($f_{k+1}^* = f_{k+1}^{\rm min}$). 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 the mutant frequency in the selected Adult at cycle $k+1$ at a given f_k^* is calculated by taking the minimum value distribution among g Adult collectives' frequencies after the

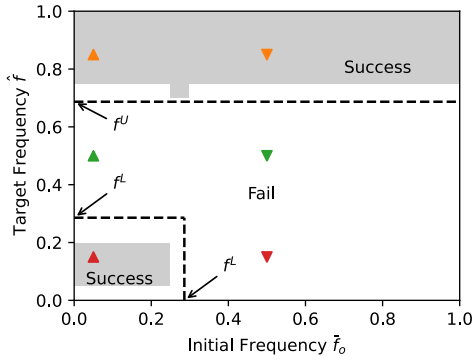

Fig. 3. There are two regions in the initial and target compositions where artificial selection can be successful. The shaded area indicates the region where collective selection is successful (see the main text). The parameters used in Fig. 2**c** and 2**d** are marked with color- and shape-matched triangle symbols. Either high target frequencies or low target frequencies starting from low initial frequencies can be achieved. We theoretically predict the boundaries of regions, f^U and f^L (black dashed lines), which agree with simulation results (shaded area) well.

maturation

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

Here, $P_{f_{k+1,\tau}}(f|f_k^*)$ is the conditional probability distribution of $f_{k+1,\tau}$ being f given f_k^*. $P_{f_{k+1,\tau}}(f|f_k^*)$ is calculated by integrating across all Newborn mutant frequency ζ (of cycle $k+1$) the probability density of achieving ζ given f_k^* multiplied by the probability density of achieving Adult mutant frequency f in cycle $k+1$. $C_{f_{k+1,\tau}}(f|f_k^*) = \int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*)$ is the conditional cumulative distribution function of $P_{f_{k+1,\tau}}(f|f_k^*)$ (see Methods). In essence, equation [1] states that the probability density of having f_{k+1}^* being f is: the probability $P_{f_{k+1,\tau}}(f|f_k^*)$ that one of the collectives has frequency f multiplied by the rest $g-1$ collectives all having frequency higher than f $\left(\left[1 - C_{f_{k+1,\tau}}(f|f_k^*)\right]^{g-1}\right)$, finally multiplied by the number of cases that one collective among g has a frequency f.

Figure 4**a** shows the median value of the conditional probability distribution $\Psi(f_{k+1}^* = f|f_k^*)$ with a shift $-f_k^*$ so that the frequency change between the selected collectives in successive selection steps is clearly shown. Results obtained by stochastic simulations (blue) are similar to those from analytic calculations (orange) (see Fig. S4 in Supplementary information for full distributions).

An increase in mutant frequency during maturation is most drastic when Newborn mutant frequency is around 0.5. This is intuitive: when Newborn mutant frequency is low, the increase will be minor; when Newborn mutant frequency is high, the fitness advantage of mutant over the population average is small and hence the increase is also minor. When Newborn mutant frequency is intermediate, the increase is the greatest. Indeed, this intuition is supported by simulations (see Fig. 4**a**). Thus, when Newborn mutant frequency is near 0.5, artificial selection may not overcome this drastic increase in mutant frequency. We draw the boundary values f^L and f^U such that the frequency can be lowered for $f < f^L$ and $f > f^U$ (see Fig. 4**a**). Otherwise, the frequency of the selected collective increases on average in the next cycle.

Calculated f^L and f^U delineate regions where artificial

selection is strong enough to lower the mutant frequency to the target value as in Fig. 3. For high target values ($\hat{f} > f^U$), slowly increasing mutant frequency can be offset by artificial selection. Consequently, the target value can be maintained. For low target values ($\hat{f} < f^L$), if the starting frequency is low ($\overline{f}_0 < f^L$), slowly increasing mutant frequency can also be offset by artificial selection, leading to maintenance of the target frequency. However, if the initial frequency $\overline{f}_0$ is larger than f^L, rapidly increasing mutant frequency can no longer be offset by artificial selection. Thus, successful regions for artificial selection should cover the range where 1) $\hat{f} > f^U$ and 2) $\hat{f}, \overline{f}_0 < f^L$. The boundaries result from this calculation are consistent with simulations (Fig. 3**a**).

Experimental setups for successful selection. According to the conditional probability distribution $\Psi(f^*_{k+1}|f^*_k)$ in Eq. [1], the change of selected frequencies Δf^*_k depends on (1) the distribution of mutant frequency in Adult collectives which dictates the minimal value to be selected; and (2) the number of collectives (g) we breed. Thus optimizing the experimental setups such as maturation time and Newborn size (which affect the distribution of mutant frequency) and the number of collectives can increase the chance for successful artificial selection.

Because a larger variation in Adult collectives provides a higher chance to choose the target composition, a large variation is a key element for successful selection. Variation is originated from the stochastic nature in the reproduction and maturation steps. Because smaller Newborn collective size N_0 induces larger variations in mutant frequency in Newborn, smaller N_0 is essential to successful selection (see Eq. [27] in Supplementary Information.) The simulation results indeed demonstrate this (see Fig. 5**a**.) We observed that by reducing N_0, the region that allows successful collective selection is broadened (larger grey area in Fig. 5**a**). If the Newborn size N_0 is smaller than $N_0 = 900$, we can always achieve our target composition.

The number of collectives g also affects the range of success. Because only one collective is selected, the selected mutant frequency is the closest to the target value among g collectives. With more collectives, the chance to find a closer value to the target is more likely. Thus, a larger number of collectives broadens the region of success (Fig. 5**b**).

We have investigated whether we could achieve the desired collective composition through artificial collective selection. The desired collective composition can be achieved only when artificial selection on collectives can cancel out the change induced by cell-level dynamics within collectives, and successful artificial selection depends on the initial and the target compositions. We found that the collective-to-collective variation in the reproduction step plays an important role in achieving and maintaining the desired collective composition, and that the size of Newborns and the number of collectives expand the success region.

Our calculations consider two genotypes. The maturation time is not too long to prevent the mutant from fixing within a collective but is long enough to generate the g Newborns. Those assumptions on timescales are different from the previous work done on (21). Their timescale on the maturation is long enough, so the collectives reach the fixed point within the maturation step. By accumulating novel mutations on trait set, the target composition can be achieved in (21). While

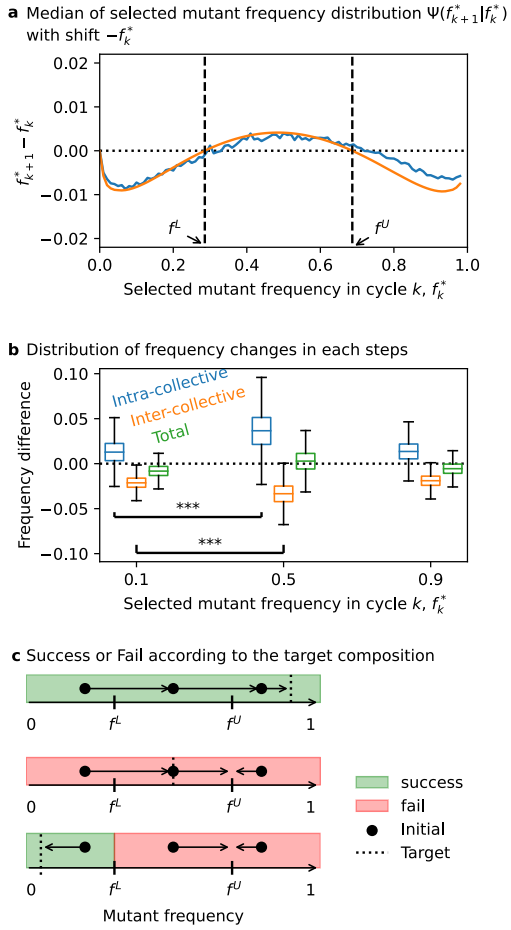

Fig. 4. Competition between natural selection and artificial selection determines the boundaries. a The median value of the conditional distribution of the frequency changes in a cycle, $f^*_{k+1} - f^*_k$. Positive values of $f^*_{k+1} - f^*_k$ mean that cycles started from f^*_k end at the frequency larger than f^*_k with more than 50% of probability, which is artificial selection cannot cancel out the mutant frequency increase. On the other hand, when f^*_k is sufficiently low or high, as artificial selection can compensate for the frequency increase, median values become negative. The points where the median of the distribution becomes zero are denoted as f^L and f^U, respectively. The values correspond to the boundaries in Fig. 3. **b** The distributions of frequency differences obtained by 1000 numerical simulations. The blue box plots indicate the frequency distributions after the maturation step, and the orange boxes are differences between the selected frequency and the mean of g collectives. Increase of frequencies during maturation for $f^*_k = 0.5$ shows maximal differences which cannot be canceled out by selection. The asterisk means the result of unpaired t-test; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. NS, not significant; **c** The role of boundaries in determining success or failure. The arrows indicate examples of trajectory from the initial compositions (black dots) to the target compositions (dashed lines). The green(red) area indicates the region of initial compositions that can(cannot) select to the target frequency $\hat{f}$ with $\hat{f} > f^U$(upper panel), $f^L \leq \hat{f} < f^U$(middle panel), and $\hat{f} \leq f^L$(lower panel). Once the composition f enters within $f^L \leq f < f^U$, the frequency must reach at least f^U because artificial selection is not effective.

their work has investigated such long-term selection in the open evolutionary dynamics with multiple traits, our work focuses on the selection within the shorter timescale, which is available in experiments and engineering.

Evolutionary trajectories in lower versus higher dimensions. We have predicted the conditions for successful artificial selection. During collective maturation, mutant cells outgrow wildtype cells. To obtain a target composition, inter-collective

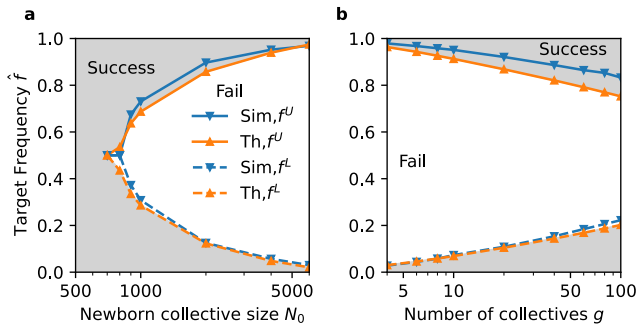

Fig. 5. Optimal experimental setups to expand the success region for artificial collective selection. **a** Change of boundary values f^U and f^L with varying the bottleneck size N_0. In gray area, the probability that f^*_{k+1} becomes smaller than f^*_k in a cycle is more than 50%. We used $g = 10$ and $\tau \approx 4.8$. **b** Effect of collective numbers g with the fixed bottleneck 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. 'Sim' indicates the result from stochastic simulations. 'Th' is calculated from the analytical approach.

selection should effectively reduce the mutant frequency. By using this fact, we made a theoretical prediction of whether a given target composition can be achieved. This approach performs well for a simple model with two types. While we have presented the results for the simple system, the approach can be generically applied to more realistic and complex systems. If frequency changes induced by natural selection can be fully compensated in the selection step, artificial selection can stabilize the frequency around the given value. Marking the compositions that can be stabilized by artificial selection, we obtain the accessible region in frequency space.

The accessible region provides the information that the target composition can stay or not, but the region cannot directly tell whether the target composition is reachable from a prepared initial composition. For example in two-strain or two-type populations, transitions from the high-frequency part of the accessible region to the low-frequency part are not allowed (see Fig. 4**c**). It is because artificial selection cannot reduce the mutant frequency against its fast change due to natural selection. When we reverse the direction from low frequency to high frequency, the transition is successful because the increasing mutant frequency is favorable in both natural and artificial selection. Likewise, it is required to evaluate possible trajectories and whether it is able to follow by artificial selection or not.

While the two-type population has only one trajectory from the initial to the target due to its low dimensionality, the evolution of composition can choose various trajectories in higher dimensional composition space. So one can find trajectories that are not passing outside of the accessible region. However, the trajectory is not always chosen in the actual test because the collective selection is directed to the target composition nevertheless it heads outside of the accessible region.

In order to verify the central idea of our approach, we also investigated the three-strain system (see Fig. 6**a** and Sec. 4 in Supplementary Information). We add double-mutated cells that grow faster than the single mutant which grows faster than the wild type. In this system, we can mark the accessible region where the artificial selection reduces double mutants as well as reduce single mutants (gray regions in Fig. 6**b**). We checked that the success of artificial selection depends on the trajectory followed, even if the target composition is located in the accessible region. For example, a target composition in the low-frequency part of the accessible region is not achievable even from the same part (see right panel of Fig. 6**b**). Once the composition escapes the accessible region, the selected composition cannot return back to the accessible region that contains the target point. This highlights the importance of predicting the trajectory for successful artificial selection and suggests that knowledge of trajectory is necessary to predict the success of artificial selection.

Materials and Methods

Stochastic simulations. Our stochastic simulation is composed of three steps: maturation, selection, and reproduction. Initially, the number of mutant cells in a collective i, $m^{(i)}_{1,0}$, is sampled from the binomial distribution with mean $N_0\overline{f}_0$. Then, $w^{(i)}_{1,0}(= N_0 - m^{(i)}_{1,0})$ wildtype cells are in the collective i. In the maturation step, we simulate the division and mutation of each cell stochastically. For Fig. 2**a** and **b**, the tau-leaping algorithm is used while for Fig. 2**c** and **d**, we directly sampled the wildtype and mutant cell numbers from the approximated frequency distribution to accelerate simulations (see Sec. 1 in Supplementary Information.) After the maturation step, each collective's function is evaluated by the difference of frequency to the target $\hat{f}$. We choose the collective with the closest frequency to the target. The selected collective generates g Newborns. The number of mutant cells is sampled from the binomial distribution with the mean of $N_0 f^*_k$, where f^*_k is the frequency of the selected collective. Then, the number of wildtype cells in the collective i is $w^{(i)}_{k+1,0} = N_0 - m^{(i)}_{k+1,0}$. We start a new cycle with those Newborn collectives.

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 inferred by the following process. Given the selected collective at 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 mutant numbers $m^{(1)}_{k+1,0}, \cdots, m^{(g)}_{k+1,0}$ at the beginning of 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. The first and the second subscripts indicate a cycle index and time from the beginning of the cycle, respectively. 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

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

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 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 w and m (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

$$P_{f_{k+1,\tau}}(f|\zeta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(f-\overline{f}(\tau))^2}{2\sigma^2_f(\tau)}\right). \quad [3]$$

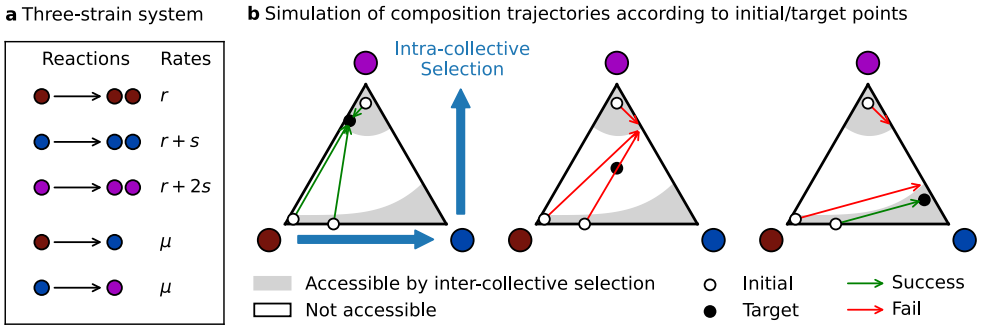

Fig. 6. Success of artificial selection depends on the trajectory in higher dimension. a The reaction rules of microbiome in the maturation step. In the maturation step, the (single) mutant can mutate further into a double-mutant with mutation rate μ. The double-mutant has a growth rate of $r + 2s$. Other processes are the same in Fig. 1. **b** The collectives attempt to evolve from various initial mutant frequencies (empty circles) to the target frequency (black dots). The composition changes due to intra-collective selection are shown in blue bold arrows. The accessible region where matured compositions can return back to their initial compositions by collective selection is marked (see Sec. 1 in Supplementary Information). If the region is inaccessible, the composition changes due to the intra-collective selection are too fast to be canceled out by inter-collective selection. We obtain final compositions starting from several initial compositions aiming to get target compositions, drawing straight arrows from the initial to final compositions. The arrow is colored by green when the simulation ends around the target composition, otherwise, the color is red. If a target composition has a high(middle) double-mutant frequency, the target is always achievable(not achievable) in the left(middle) panel. But, even though we target a low double-mutant frequency, the target composition is achievable only when the trajectory from the initial point is always located in an accessible region (right panel). Each arrow indicates the composition change from initial to final compositions in each simulation.

The conditional probability distribution of Adult collective to have frequency $f_{k+1,\tau}$ at a given f_k^* can be calculated by producing two Gaussian distribution functions and integrating overall ζ values, which is given by

$$P_{f_{k+1,\tau}}(f|f_k^*) = \int_0^1 d\zeta \; P_{f_{k+1,\tau}}(f|\zeta) \; P_{f_{k+1,0}}(\zeta|f_k^*). \qquad [4]$$

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}^*.

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 NRF grant funded by the Korea government (MSIT) Grant No.2020R1A2C1101894 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. This was also supported by the Korean Local Governments–Gyeongsangbuk-do Province and Pohang City. We thank Su-Chan Park and Li Xie for constructive comments and discussions. This work was supported by INHA UNIVERSITY Research Grant.

1. S Beier, S Bertilsson, Uncoupling of chitinase activity and uptake of hydrolysis products in freshwater bacterioplankton. *Limnol. Oceanogr.* **56**, 1179–1188 (2011).
2. S Beier, S Bertilsson, Bacterial chitin degradation—mechanisms and ecophysiological strategies. *Front. microbiology* **4**, 149 (2013).
3. S Woo, I Song, HJ Cha, Fast and facile biodegradation of polystyrene by the gut microbial flora of plesiophthalmus davidis larvae. *Appl. environmental microbiology* **86**, e01361–20 (2020).
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. Á Sánchez, et al., Directed evolution of microbial communities. *Annu. review biophysics* **50**, 323–341 (2021).
6. FI Arias-Sánchez, B Vessman, S Mitri, Artificially selecting microbial communities: If we can breed dogs, why not microbiomes? *PLoS biology* **17**, e3000356 (2019).
7. DS Wilson, *The natural selection of populations and communities.* (Benjamin/Cummings Pub. Co.), (1980).
8. CJ Goodnight, Heritability at the ecosystem level. *Proc. Natl. Acad. Sci.* **97**, 9365–9366 (2000).
9. W Swenson, DS Wilson, R Elias, Artificial ecosystem selection. *Proc. Natl. Acad. Sci.* **97**, 9110–9114 (2000).
10. W Swenson, J Arendt, DS Wilson, Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environ. Microbiol.* **2**, 564–571 (2000).
11. HTP Williams*, TM Lenton, Artificial selection of simulated microbial ecosystems. *Proc. Natl. Acad. Sci.* **104**, 8918 (2007).
12. 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).
13. K Panke-Buisse, S Lee, J Kao-Kniffin, Cultivated sub-populations of soil microbiomes retain early flowering plant trait. *Microb. ecology* **73**, 394–403 (2017).
14. M Blouin, B Karimi, J Mathieu, TZ Lerch, Levels and limits in artificial selection of communities. *Ecol. Lett.* **18**, 1040–1048 (2015).
15. RJ Wright, MI Gibson, JA Christie-Oleza, Understanding microbial community dynamics to improve optimal microbiome selection. *Microbiome* **7**, 1–14 (2019).
16. CY Chang, ML Osborne, D Bajic, A Sanchez, Artificially selecting bacterial communities using propagule strategies. *Evolution* **74**, 2392–2403 (2020).
17. 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).
18. UG Mueller, et al., Artificial selection on microbiomes to breed microbiomes that confer salt tolerance to plants. *mSystems* **6**, e01125–21 (2021).
19. 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).
20. ASM Jigyasa Arora, Margaret A. Mars Brisbin, Effects of microbial evolution dominate those of experimental host-mediated indirect selection. *PeerJ* **8**, e9350 (2020).
21. 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).
22. L Xie, AE Yuan, W Shou, Simulations reveal challenges to artificial community selection and possible strategies for success. *PLoS biology* **17**, e3000295 (2019).
23. AJ Black, P Bourrat, PB Rainey, Ecological scaffolding and the evolution of individuality. *Nat. ecology & evolution* **4**, 426–436 (2020).
24. L Xie, W Shou, Steering ecological-evolutionary dynamics to improve artificial selection of microbial communities. *Nat. Commun.* **12**, 6799 (2021).
25. CY Chang, et al., Engineering complex communities by directed evolution. *Nat. ecology & evolution* **5**, 1011 (2021).
26. B Vessman, P Guridi-Fernández, FI Arias-Sánchez, S Mitri, Novel artificial selection method improves function of simulated microbial communities. *bioRxiv* (2023).

Supplementary Information for

Condition for successful artificial selection on microbial collectives and optimal experimental setup

Juhee Lee, Wenying Shou, and Hye Jin Park

Wenying Shou.
E-mail:w.shou@ucl.ac.uk
Hye Jin Park.
E-mail: hyejin.park@inha.ac.kr

This PDF file includes:

Figs. S1 to S6 (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, wildtype and mutant genotypes. We start with g 'Newborn' collectives with N_0 cells, and each collective i ($i = 1, ..., g$) has $w_{k,0}^{(i)}$ wild-type cells and $m_{k,0}^{(i)}$ mutant cells. The index k is a cycle number. The collectives are allowed to "mature" for $t = \tau$, where wild-type and mutant populations grow at a rate of r and $r + s$, respectively. wildtype cells also mutate to mutant cells at a rate of μ per cell, and the back mutation from mutant to wildtype is ignored. After maturation, we choose the collective with mutant 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 $w(t)$ and $m(t)$ as representative notations of $w_k^{(i)}(t)$ and $m_k^{(i)}(t)$, respectively. The initial number of cells are the Newborn collectives $w_{k,0}^{(i)}$ and $m_{k,0}^{(i)}$. We consider mutation that is beneficial to grow ($s > 0$).

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

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

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

$$W \xrightarrow{\mu} M, \quad [3]$$

where W and M denote individual cells of a wildtype and a mutant, 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 $P(w, m, t)$ to sample compositions instead of tracking all events that occurred in individual collectives. First, we calculate the distribution $P(w, m, t)$ analytically, and then sample numbers of wildtype and mutant cells from this distribution at the end of cycle $t = \tau$. Below, we describe the derivation of $P(w, m, t)$.

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

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

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 in wildtype cells happens during time interval $[t, t + dt)$, which changes the collective's composition from $(w - 1, m)$ to (w, m). Similarly, the second term comes from a birth event in mutant cells. 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 (w, m) to others.

Calculating the exact form of $P(w, m, t)$ is not simple. Instead, we consider the probability distributions of $w(t)$ and $m(t)$ at time t. We assume that the mutation rate is much smaller than the growth rates, and hence the correlation between w and m is sufficiently small. Then, $P(w, m, t) \sim P(w, t)P(m, t)$. The distributions of w and m 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 $P(w, m, t) \sim \mathcal{N}(w(t)|\overline{w}(t), \sigma_w^2(t))\mathcal{N}(m(t)|\overline{m}(t), \sigma_m^2(t))$.

The means are defined by $\overline{w}(t) = \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} wP(w, m, t)$ and $\overline{m}(t) = \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} mP(w, m, t)$. The differential equations for means are obtained by applying the definition to the master equation in Eq. [4], as

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

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

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

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

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

where $\overline{w}_o \equiv \overline{w}(0)$ and $\overline{m}_o \equiv \overline{m}(0)$ are the mean number of wildtype and mutant 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 mutant cells starting from w_0 wild-type cells (3).

We define the second momenta of w and m as

$$\overline{w^2} = \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} w^2 P(w,m,t), \quad [9]$$

$$\overline{m^2} = \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} m^2 P(w,m,t). \quad [10]$$

Then, corresponding differential equations are given by

$$\frac{d\overline{w^2}}{dt} = 2(r-\mu)\overline{w^2} + (r+\mu)\overline{w}, \quad [11]$$

$$\frac{d\overline{m^2}}{dt} = 2(r+s)\overline{m^2} + (r+s)\overline{m} + 2\mu\overline{wm} + \mu\overline{w}. \quad [12]$$

The solution of Eq. [11] is

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

$$\approx \overline{w_o^2}e^{2rt} + \overline{w}_o e^{rt}(e^{rt}-1), \quad [14]$$

where $\overline{w_o^2} \equiv \overline{w^2}(0)$ is an initial value of the second moment. Thus, the variance $\sigma_w^2(t) = \overline{w^2}(t) - [\overline{w}(t)]^2$ is

$$\sigma_w^2(t) = \sigma_{w,0}^2 e^{2rt} + \overline{w}_o e^{rt}(e^{rt}-1), \quad [15]$$

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

$$\frac{d\overline{wm}}{dt} = (2r+s-\mu)\overline{wm} + \mu(\overline{w^2}-\overline{w}). \quad [16]$$

The solution of Eq. [16] is given by

$$\begin{aligned}\overline{wm}(t) &= \overline{wm}_o e^{(2r+s)t} \\ &+ \frac{\mu}{s}(\overline{w^2}_0 + \overline{w}_0)(e^{(2r+s)t} - e^{2rt}) \\ &- \frac{2\mu\overline{w}_0}{r+s}(e^{(2r+s)t} - e^{rt}).\end{aligned} \quad [17]$$

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

$$\begin{aligned}\overline{m^2}(t) &= \overline{m^2}_0 e^{2(r+s)t} + \overline{m}_0 e^{(r+s)t}(e^{(r+s)t}-1) \\ &+ \frac{\mu\overline{w}_o}{s}\left(\frac{2s}{r+2s}e^{2(r+s)t} - e^{(r+s)t} + \frac{r}{r+2s}e^{rt}\right) \\ &+ \frac{2\mu\overline{wm}_0}{s}e^{(r+s)t}(e^{(r+s)t}-e^{rt}) \\ &+ O(\mu^2),\end{aligned} \quad [18]$$

where $\overline{m_o^2} \equiv \overline{m^2}(0)$ is an initial value of the second moment. Thus, the variance $\sigma_m^2(t) = \overline{m^2}(t) - [\overline{m}(t)]^2$ is given, up to the order of μ, by

$$\begin{aligned}\sigma_m^2(t) &= \sigma_{m,0}^2 e^{2(r+s)t} + \overline{m}_0 e^{(r+s)t}(e^{(r+s)t}-1) \\ &+ \frac{\mu\overline{w}_o}{s}\left(\frac{2s}{r+2s}e^{2(r+s)t} - e^{(r+s)t} + \frac{r}{r+2s}e^{rt}\right) \\ &+ \frac{2\mu(\overline{wm}_0 - \overline{w}_0\overline{m}_0)}{s}e^{(r+s)t}(e^{(r+s)t}-e^{rt}) \\ &+ O(\mu^2).\end{aligned} \quad [19]$$

Using Eq. [7,8,15,19], we construct Gaussian distribution functions for $w(t)$ and $m(t)$, $\mathcal{N}(w|\overline{w}(\tau),\sigma_w^2(\tau))$ and $\mathcal{N}(m|\overline{m}(\tau),\sigma_m^2(\tau))$, at the end of cycle. Then, numbers of wildtype cells and mutant cells are randomly sampled from each distribution: $w(\tau)$ from $\mathcal{N}(w|\overline{w}(\tau),\sigma_w^2(\tau))$ and $m(\tau)$ from $\mathcal{N}(m|\overline{m}(\tau),\sigma_m^2(\tau))$. This process requires only two sampling steps for each collective.

Selection. After the maturation step is finished in cycle k, we compute the frequencies of the mutant 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 $f_{k,\tau}^{(i)} \equiv f_k^{(i)}(\tau)$. 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

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

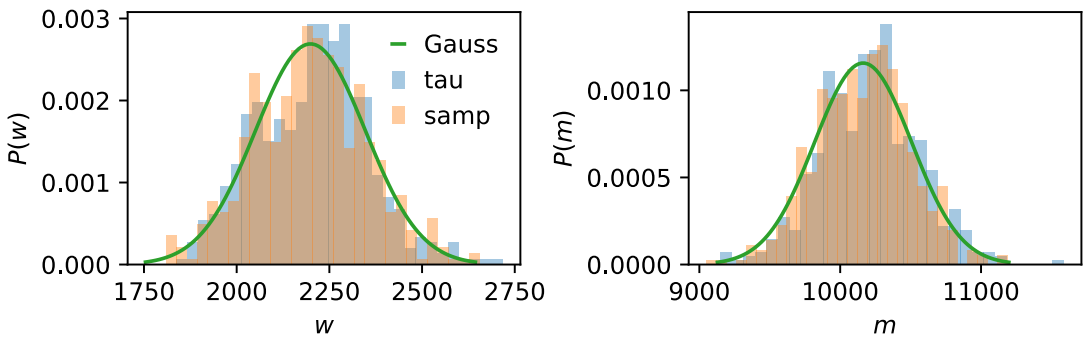

Fig. S1. Comparison between the calculated Gaussian distribution ("Gauss", with the mean and variances computed from Eq. [7,8,15,19]) and simulations using tau-leaping ("tau") and sampling ("samp") methods. The simulations run 500 times. The line labeled with 'Gauss' is the Gaussian distribution. The initial number of cells are $w_0 = 200$ and $m_0 = 800$. The parameters are $r = 0.5$, $s = 0.03$, $\mu = 0.0001$, and $\tau = 4.8$.

Reproduction. We divide the selected collective into g offspring collectives. We sample $m^{(i)}_{k+1,0}$ mutant cells and $w^{(i)}_{k+1,0}$ wildtype 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 $P(m^{(1)}_{k+1,0}, \cdots, m^{(g)}_{k+1,0})$ follows a multivariate hypergeometric distribution. The selected collective is assumed to have w^*_k wildtype cells and m^*_k mutant cells. If the size of selected collective $N^*_k = w^*_k + m^*_k$ is large enough, the consecutive sampling is approximated to the independent binomial sampling (see Fig. S2). Thus, we independently sample g numbers of mutants, $\{m^{(1)}_{k+1,0}, \cdots, m^{(g)}_{k+1,0}\}$, from the binomial distribution. The probability mass function of $m^{(i)}_{k+1,0}$ is given by

$$P_{m^{(i)}_{k+1,0}}(m) = \frac{N_0!}{(N_0-m)!m!}(f^*_k)^m(1-f^*_k)^{N_0-m}. \quad [21]$$

After the sampling, the numbers of wildtype cells are set to be $w^{(i)}_{k+1}(0) = N_0 - m^{(i)}_{k+1,0}$ 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 relative error d between the target frequency $\hat{f}$ and $\langle f^* \rangle$ (*i.e.* $d = |\langle f^* \rangle - \hat{f}|/\hat{f} \leq 5\%$) at the end of simulations (1000 cycles). In the colormap, errors higher than 5% are marked with white, which indicates the failure of selection. The dashed lines indicate the same boundary in Fig 3 in the main text.

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 first 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 $\Psi(f^*_{k+1}|f^*_k)$.

We start from the reproduction step at the end of cycle k. The probability distribution of the mutant 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, equation [21] is approximated by the Gaussian distribution $m^{(i)}_{k+1,0} \sim \mathcal{N}(N_0 f^*_k, f^*_k(1-f^*_k)/N_0)$. Then, the probability density function that $f^{(i)}_{k+1,0}$ to be ζ in Newborn collective i is

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

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

$$w(t) \approx \overline{w}(t) + \sigma_w(t)G_w(t) \quad [23]$$

$$m(t) \approx \overline{m}(t) + \sigma_m(t)G_m(t), \quad [24]$$

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

$$\begin{aligned} f(t) &\approx \frac{\overline{m}(t) + \sigma_m(t)G_m(t)}{\overline{w}(t) + \sigma_w(t)G_w(t) + \overline{m}(t) + \sigma_m(t)G_m(t)} \\ &\approx \overline{f}(t) + \sigma_f(t)G_f(t). \end{aligned} \quad [25]$$

The mean of f is given by

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

and the variance is

$$\sigma^2_f(t) = \frac{(1-\overline{f}(t))^2\sigma^2_m(t) + \overline{f}(t)^2\sigma^2_w(t)}{\overline{N}(t)^2}. \quad [27]$$

The average size of collectives is $\overline{N}(t) = \overline{w}(t) + \overline{m}(t)$. In terms of a probability distribution, the random variable that the frequency at the end of cycle $f(\tau) = f_\tau$ follows the Gaussian distribution $\mathcal{N}(\overline{f}(\tau), \sigma^2_f(\tau))$ whose the probability density function is given by

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

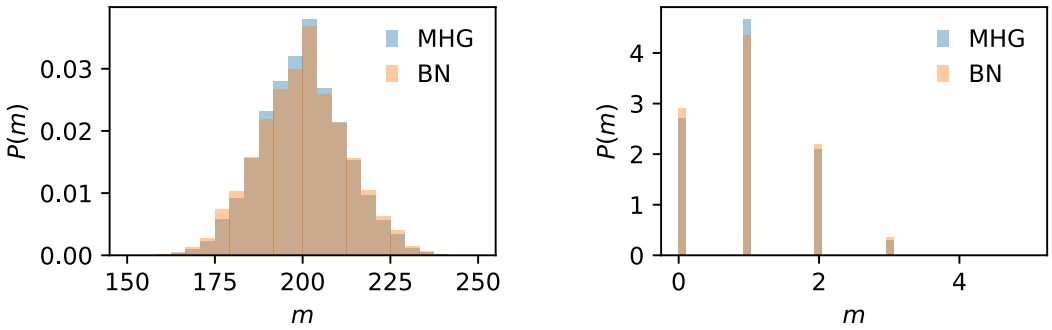

Fig. S2. Comparison between consecutive sampling and independent binomial sampling. A parent collective is divided into 10 collectives. The histogram labeled with 'MHG' is the probability mass function of m of the fifth collective sampled via multivariate hypergeometric distribution. The independent binomial sampling is labeled with 'BN'. The initial numbers of cells are $w = 8000$ and $m = 2000$ for the left panel, and $w = 20$ and $m = 5$ for the right panel. 10000 samples are drawn for each distribution.

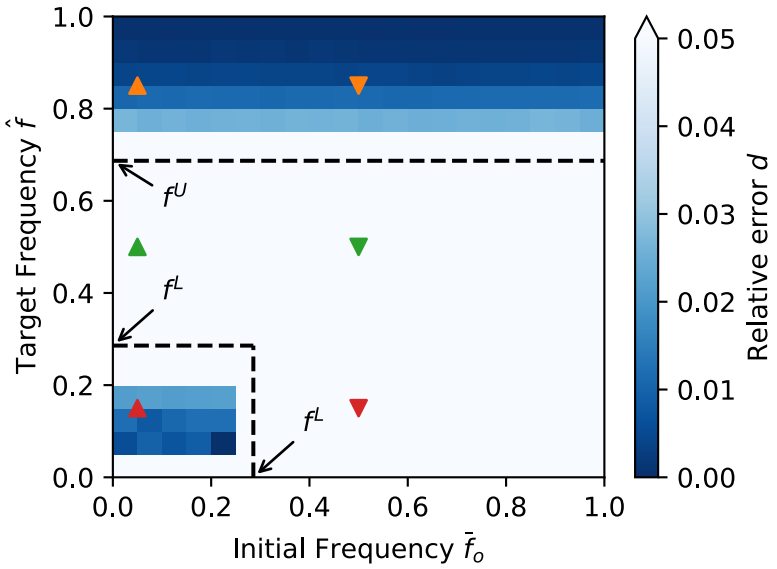

Fig. S3. Color map of the relative error $d = |\langle f^* \rangle - \hat{f}|/\hat{f} \leq 5\%$ between frequency $\langle f^* \rangle$ of the averaged selected collectives at the end of simulations and the target frequency $\hat{f}$. The dashed lines are drawn by the arguments in the main text. For parameters, we used $\overline{f}_0 = 0.05$, $r = 0.5$, $s = 0.03$, $\mu = 0.0001$ and $\tau \approx 4.8$.

With indices of cycle numbers 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

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

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

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

The above Eq. [30] simply states that $f^{\min} \leq f$ 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 $C_{f^{\min}}(f|f_k^*) = 1 - [\text{Prob}(f_{k+1,\tau} \geq f)|f_k^*]^g$. Note that $\text{Prob}[f_{k+1,\tau} \geq f|f_k^*] = \int_f^1 df' P_{f_{k+1,\tau}}(f'|f_k^*) = 1 - \int_0^f df' P_{f_{k+1,\tau}}(f'|f_k^*) = 1 - C_{f_{k+1,\tau}}(f|f_k^*)$, and Eq. [30] becomes

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

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

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

We compute the probability density function [32] by using numerical integration and compare it with the stochastic simulation results in Fig. S4. Both distributions are agreed to each other.

3. 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 the growth. The same analogy can be applied even when the mutation is deleterious. In the case of $s = -0.03$, the target frequency lower than around 0.3 and larger than around 0.7 can be selected (see Fig. S5**a**). Since the sign of s is opposite to the result in the main text, the diagram is reversed from Fig. 3.

Since the mutant population grows slower than the wildtype population ($s < 0$), the mutant 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 $f^{\max}$ such that

$$\begin{aligned} C_{f^{\max}}(f|f_k^*) &= \text{Prob}[f^{\max} \geq f|f_k^*] \\ &= \text{Prob}[(f_{k+1}^{(1)}(\tau) \geq f) \wedge (f_{k+1}^{(2)}(\tau) \geq f) \wedge \cdots \wedge (f_{k+1}^{(g)}(\tau) \geq f)|f_k^*]. \end{aligned} \qquad [33]$$

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

$$C_{f^{\max}}(f|f_k^*) = (C_{f_{k+1,\tau}}(f|f_k^*))^g. \qquad [34]$$

Likewise in the previous section, we get the conditional probability density function by differentiating Eq. [34] with respect to f and replacing $f \to f_{k+1}^*$ as

$$\Psi(f_{k+1}^*|f_k^*) = gC_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*)^{g-1} P_{f_{k+1,\tau}}(f_{k+1}^*|f_k^*). \qquad [35]$$

The distribution in Eq. [35] is evaluated for various f_k^* in Fig. S5**b** with numerical simulations.

Likewise in the main text, small enough changes in the maturation step are required to make artificial selection effective. According to Eq. [8], the maximal speed of frequency changes in the maturation step appears around $f = 0.5$ also in $s < 0$ case. So, the boundary values f^L and f^U have the same definition as in the main text, and we look for the median of $\Psi(f_{k+1}^*|f_k^*)$ with the shift of $-f_k^*$ (see Fig. S5**c**). The boundaries are well agreed with Fig. S5**a**.

4. 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. Using the idea, we obtain the conditional probability to select a collective with f_k^* at the cycle k from the selected collective in the previous cycle $k-1$. Here, we extend the idea into more complex collectives in order to suggest a possible generalization. We assume that collectives consist of three genotypes with wildtype, single-mutant, and double-mutant. The growth rate of wildtype is r. Each mutation adds the growth rate with s. So the single and double-mutant

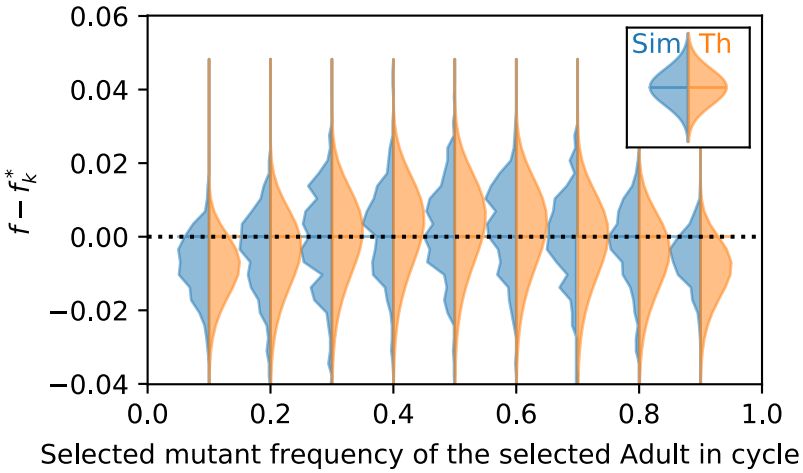

Fig. S4. The probability density functions of the selected collective's frequency f_{k+1}^* with the offset of f^*. The blue distribution is obtained using 300 stochastic simulation results. The orange distribution represents Eq. [32] computed by numerical integration. The median values of the distributions are shown in fig. 4**a** in the main text.

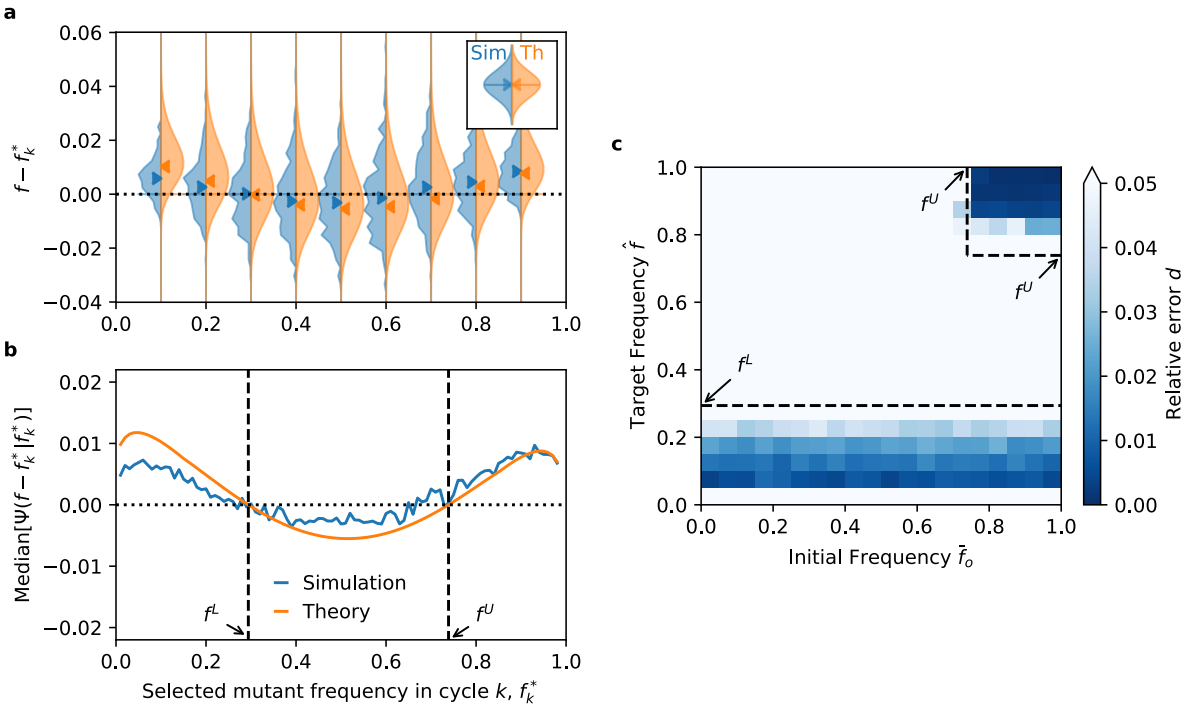

Fig. S5. Artificial selection also works for deleterious mutation. a Conditional Probability density functions of mutant frequency change $f - f_k^*$ in a cycle for various f_k^* values. The left-hand side distribution is obtained from simulations and the right-hand side distribution is numerically obtained by evaluating Eq. [35]. 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, $\mathrm{Median}[\Psi(f - f_k^*)] = 0$ 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 $\bar{f}_0$. 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**.

have growth rates with $r+s$ and $r+2s$, respectively. The mutation rate is μ. So, the birth and mutation events are written by the chemical reactions:

$$W \xrightarrow{r} W+W, \quad [36]$$
$$M \xrightarrow[s>0]{r+s} M+M, \quad [37]$$
$$V \xrightarrow[s>0]{r+2s} V+V, \quad [38]$$
$$W \xrightarrow{\mu} M, \quad [39]$$
$$M \xrightarrow{\mu} V. \quad [40]$$

Here, V is an individual of a double mutant. We write a master equation of the processes for $P(w,m,v,t)$ which is the probability to have w wildtype cells, m single-mutant cells, and v double-mutant cells at time t.

$$\begin{aligned}\frac{dP(w,m,v,t)}{dt} &= r(w-1)P(w-1,m,v,t) - rwP(w,m,v,t) \\ &+(r+s)(m-1)P(w,m-1,v,t) - (r+s)mP(w,m,v,t) \\ &+(r+2s)(v-1)P(w,m,v-1,t) - (r+2s)vP(w,m,v,t) \\ &+\mu(w+1)P(w+1,m-1,v,t) - \mu wP(w,m,v,t) \\ &+\mu(m+1)P(w,m+1,v-1,t) - \mu mP(w,m,v,t).\end{aligned} \quad [41]$$

The composition of collective i in cycle k is now represented with $(f_k^{(i)}(t), h_k^{(i)}(t)) \equiv \mathbf{p}_k^{(i)}(t)$ where $f_k^{(i)}(t) = m_k^{(i)}(t)/(w_k^{(i)}(t) + m_k^{(i)}(t) + v_k^{(i)}(t))$ and $h_k^{(i)}(t) = v_k^{(i)}(t)/(w_k^{(i)}(t) + m_k^{(i)}(t) + v_k^{(i)}(t))$. Then, the target composition is set to be $(\hat{f}, \hat{h})$. The composition of the selected collective 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 $\Psi(\mathbf{p}_k^*|\mathbf{p}_{k-1}^*)$ by using the master Eq. [41].

At the reproduction step in cycle $k-1$, we choose N_0 cells from $(f_{k-1}^*, h_{k-1}^*) \equiv \mathbf{p}_{k-1}^*$. 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 $m_{k,0}, v_{k,0}$ cells are represented by

$$P(m_{k,0}, v_{k,0}|f_{k-1}^*, h_{k-1}^*) = \frac{N_0!}{(N_0 - m_{k,0} - v_{k,0})! m_{k,0}! v_{k,0}!}(1 - f_{k-1}^* - h_{k-1}^*)^{N_0 - m_{k,0} - v_{k,0}}(f_{k-1}^*)^{m_{k,0}}(h_{k-1}^*)^{v_{k,0}}, \quad [42]$$

where the number of wildtype $w_{k,0}$ is automatically set to be $w_{k,0} = N_0 - m_{k,0} - v_{k,0}$. Then, the approximated multivariate normal distribution is $\mathcal{N}(N_0\mathbf{p}_{k-1}^*, N_0\mathbf{M}_k)$ where the mean distribution is $\boldsymbol{p}_{k-1}^* = (f_{k-1}^*, h_{k-1}^*)$ and covariance matrix is $\mathbf{M}_k$. The diagonal terms of $\mathbf{M}_k$ 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

$$\mathbf{M}_k = \begin{bmatrix} \frac{f_{k-1}^*(1-f_{k-1}^*)}{N_0^2} & -\frac{f_{k-1}^* h_{k-1}^*}{N_0^2} \\ -\frac{f_{k-1}^* h_{k-1}^*}{N_0^2} & \frac{h_{k-1}^*(1-h_{k-1}^*)}{N_0^2} \end{bmatrix} \equiv \begin{bmatrix} \sigma_{f_{k,0}}^2 & \sigma_{f_{k,0}h_{k,0}} \\ \sigma_{f_{k,0}h_{k,0}} & \sigma_{h_{k,0}}^2 \end{bmatrix}. \quad [43]$$

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

$$P_{\boldsymbol{\rho}_{k,0}}(\boldsymbol{\rho}|\mathbf{p}_{k-1}^*) = \frac{1}{\sqrt{2\pi}^2\sqrt{\det \boldsymbol{M}_k}} e^{-\frac{1}{2}(\boldsymbol{\rho}-\boldsymbol{p}_{k-1}^*)\boldsymbol{M}_k^{-1}(\boldsymbol{\rho}-\boldsymbol{p}_{k-1}^*)^T}. \quad [44]$$

At the beginning of cycle k, a newborn collective starts to maturate from (w_0, m_0, v_0) cells (for convenience, cycle index k is dropped.) Their initial covariance matrix is $N_0^2\mathbf{M}_k$. In terms of (ζ, η), each initial numbers are $w_0 = N_0(1-\zeta-\eta)$, $m_0 = N_0\zeta$,

and $v_0 = N_0\eta$. By using Eq. [41], we can write ordinary differential equations up to the second moment.

$$\frac{d\overline{w}(t)}{dt} = r\overline{w}(t) - \mu\overline{w}(t) \quad [45]$$

$$\frac{d\overline{m}(t)}{dt} = (r+s)\overline{m}(t) + \mu(\overline{w}(t) - \overline{m}(t)) \quad [46]$$

$$\frac{d\overline{v}(t)}{dt} = (r+2s)\overline{v}(t) + \mu\overline{m}(t) \quad [47]$$

$$\frac{d\overline{w^2}}{dt} = 2(r-\mu)\overline{w^2} + (r+\mu)\overline{w} \quad [48]$$

$$\frac{d\overline{m^2}}{dt} = 2(r+s-\mu)\overline{m^2} + (r+s+\mu)\overline{m} + \mu(2\overline{wm} + \overline{w}) \quad [49]$$

$$\frac{d\overline{v^2}}{dt} = 2(r+2s)\overline{v^2} + (r+2s)\overline{v} + \mu(2\overline{mv} + \overline{m}) \quad [50]$$

$$\frac{d\overline{wm}}{dt} = (2r+s-2\mu)\overline{wm} + \mu(\overline{w^2} - \overline{w}) \quad [51]$$

$$\frac{d\overline{mv}}{dt} = (2r+3s-\mu)\overline{mv} + \mu(\overline{wv} + \overline{m^2} - \overline{m}) \quad [52]$$

$$\frac{d\overline{wv}}{dt} = (2r+2s-\mu)\overline{wv} + \mu\overline{wm}. \quad [53]$$

The initial conditions of the system in coupled equations Eq. [45-53] are obtained by the mean and (co)variances of Eq. [42]. By solving equations numerically, we obtain a set of mean cell numbers $(\overline{w}, \overline{m}, \overline{v})$ and a set of variances $(\sigma_w^2, \sigma_m^2, \sigma_v^2)$ as well as covariances $(\sigma_{wm}, \sigma_{mv}, \sigma_{wv})$. We assume that the covariances are smaller than the variances. We consider w, m, and v as Gaussian random variables

$$w(t) \approx \overline{w}(t) + \sigma_w(t)G_w(t) \quad [54]$$

$$m(t) \approx \overline{m}(t) + \sigma_m(t)G_m(t) \quad [55]$$

$$v(t) \approx \overline{v}(t) + \sigma_v(t)G_v(t). \quad [56]$$

Then, the frequency of single-mutant becomes

$$\begin{aligned} f(t) &\approx \frac{\overline{m}(t) + \sigma_m(t)G_m(t)}{\overline{w}(t) + \sigma_w(t)G_w(t) + \overline{m}(t) + \sigma_m(t)G_m(t) + \overline{v}(t) + \sigma_v(t)G_v(t)} \\ &\approx \overline{f}(t) + \sigma_f(t)G_f(t), \end{aligned} \quad [57]$$

where $\overline{f} = \overline{m}/(\overline{w} + \overline{m} + \overline{v})$ and $\sigma_f^2 = \left(\overline{f}^2(\sigma_w^2 + \sigma_v^2) + (1-\overline{f}^2)\sigma_m^2\right)/(\overline{w} + \overline{m} + \overline{v})$. Similarly, the frequency of double-mutant is

$$\begin{aligned} h(t) &\approx \frac{\overline{v}(t) + \sigma_v(t)G_v(t)}{\overline{w}(t) + \sigma_w(t)G_w(t) + \overline{m}(t) + \sigma_m(t)G_m(t) + \overline{v}(t) + \sigma_v(t)G_v(t)} \\ &\approx \overline{h}(t) + \sigma_h(t)G_h(t), \end{aligned} \quad [58]$$

where $\overline{h} = \overline{v}/(\overline{w} + \overline{m} + \overline{v})$ and $\sigma_h^2 = \left(\overline{h}^2(\sigma_w^2 + \sigma_m^2) + (1-\overline{h}^2)\sigma_v^2\right)/(\overline{w} + \overline{m} + \overline{v})$. If the covariances are small enough, we can approximate the joint probability distribution of $(f(\tau), h(\tau)) \equiv (f_\tau, h_\tau) = \mathbf{p}_\tau$ at the end of maturation step as

$$P_{\mathbf{p}_\tau}(\mathbf{p}|\boldsymbol{\rho}) = \frac{1}{\sqrt{2\pi}\sigma_{f_\tau}} e^{-\frac{(f-\overline{f}_\tau)^2}{2\sigma_{f_\tau}^2}} \frac{1}{\sqrt{2\pi}\sigma_{h_\tau}} e^{-\frac{(h-\overline{h}_\tau)^2}{2\sigma_{h_\tau}^2}}. \quad [59]$$

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

$$P_{\mathbf{p}_{k,\tau}}(\mathbf{p}|\mathbf{p}_{k-1}^*) = \int_0^1 d\zeta \int_0^1 d\eta P_{\mathbf{p}_{k,\tau}}(\mathbf{p}|\zeta,\eta) P_{\mathbf{p}_{k,0}}(\zeta,\eta|\mathbf{p}_{k-1}^*). \quad [60]$$

At the end of cycle k, we will choose an appropriate collective among g collectives that follows the conditional probability distribution in Eq. [71]. In (f,h) space, the direction of change in frequencies during maturation depends on the initial composition (f_{k-1}^*, h_{k-1}^*). So, we take different extreme value distributions according to the direction. We note that the mean change in h is always positive in whole (f,h) space since $d\overline{v}/dt$ is always positive in Eq. [47]. Thus, we choose the minimum value $h^{\min}$ in every selection step.

If the mean $\overline{f_{k,\tau}^{(i)}} = \int_0^1 df' \int_0^1 dh'\ f P_{\mathbf{p}_{k,\tau}}(f',h'|\mathbf{p}_{k-1}^*)$ is larger than f_{k-1}^*, the minimum value among $f_{k,\tau}^{(1)}, f_{k,\tau}^{(2)}, \cdots, f_{k,\tau}^{(g)}$ 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$f_k^* = \min(f_{k,\tau}^{(1)}, f_{k,\tau}^{(2)}, \cdots, f_{k,\tau}^{(g)})$ and $h_k^* = \min(h_{k,\tau}^{(1)}, h_{k,\tau}^{(2)}, \cdots, h_{k,\tau}^{(g)})$. We temporally drop cycle index k and time index τ for simplicity. Then, the joint cumulative distribution function $C_{\mathbf{p}^*}(f, h|\boldsymbol{p}_{k-1}^*) = \Pr[f^* < f \wedge h^* < h|\boldsymbol{p}_{k-1}^*]$ is

$$\begin{aligned}
C_{\mathbf{p}^*}(f, h|\boldsymbol{p}_{k-1}^*) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&= (\int_0^1 - \int_f^1) df' (\int_0^1 - \int_h^1) dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&= \int_0^1 df' \int_0^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) - \int_h^1 dh' \int_0^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&\quad - \int_0^1 df' \int_h^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) + \int_f^1 df' \int_h^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&= 1 - \Pr[f^* \geq f|\boldsymbol{p}_{k-1}^*] - \Pr[h^* \geq h|\boldsymbol{p}_{k-1}^*] + \Pr[f^* \geq f \wedge h^* \geq h|\boldsymbol{p}_{k-1}^*]. \qquad [61]
\end{aligned}$$

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

$$\begin{aligned}
\Pr[f^* \geq f \wedge h^* \geq h|\boldsymbol{p}_{k-1}^*] &= \Pr[f^{(1)} \geq f \wedge h^{(1)} \geq h \wedge f^{(2)} \geq f \wedge h^{(2)} \geq h \wedge \cdots \wedge f^{(g)} \geq f \wedge h^{(g)} \geq h|\boldsymbol{p}_{k-1}^*] \\
&= \left[\Pr[f^{(i)} \geq f \wedge h^{(i)} \geq h \middle| \boldsymbol{p}_{k-1}^*]\right]^g \\
&= \left[1 - \Pr[f^{(i)} < f|\boldsymbol{p}_{k-1}^*] - \Pr[h^{(i)} < h|\boldsymbol{p}_{k-1}^*] + \Pr[f^{(i)} < f \wedge h^{(i)} < h] \middle| \boldsymbol{p}_{k-1}^*\right]^g \\
&= \left[1 - C_{f^{(i)}}(f|\boldsymbol{p}_{k-1}^*) - C_{h^{(i)}}(h|\boldsymbol{p}_{k-1}^*) + C_{\mathbf{p}^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*)\right]^g, \qquad [62]
\end{aligned}$$

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

$$C_{f^{(i)}}(f|\boldsymbol{p}_{k-1}^*) = \int_0^f df' \int_0^1 dh' P_{\mathbf{p}^{(i)}}(f', h'|\boldsymbol{p}_{k-1}^*), \qquad [63]$$

$$C_{h^{(i)}}(h|\boldsymbol{p}_{k-1}^*) = \int_0^1 df' \int_0^h dh' P_{\mathbf{p}^{(i)}}(f', h'|\boldsymbol{p}_{k-1}^*). \qquad [64]$$

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

$$C_{\mathbf{p}^*}(f, h|\boldsymbol{p}_{k-1}^*) = 1 - \left[1 - C_{f^{(i)}}(f|\boldsymbol{p}_{k-1}^*)\right]^g - \left[1 - C_{h^{(i)}}(h|\boldsymbol{p}_{k-1}^*)\right]^g + \left[1 - C_{f^{(i)}}(f|\boldsymbol{p}_{k-1}^*) - C_{h^{(i)}}(h|\boldsymbol{p}_{k-1}^*) + C_{\mathbf{p}^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*)\right]^g. \qquad [65]$$

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

$$\begin{aligned}
P_{\mathbf{p}_k^*}(f, h|\boldsymbol{p}_{k-1}^*) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}_k^*}(f, h|\boldsymbol{p}_{k-1}^*) \\
&= g(g-1)\left[1 - C_{f_k^{(i)}}(f|\boldsymbol{p}_{k-1}^*) - C_{h_k^{(i)}}(h|\boldsymbol{p}_{k-1}^*) + C_{\mathbf{p}_k^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*)\right]^{g-2} \qquad [66] \\
&\quad \times (P_{f_k^{(i)}}(f|\boldsymbol{p}_{k-1}^*) - \partial_f C_{\mathbf{p}^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*))(P_{h_k^{(i)}}(h|\boldsymbol{p}_{k-1}^*) - \partial_h C_{\mathbf{p}_k^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*)) \\
&\quad + g\left[1 - C_{f_k^{(i)}}(f|\boldsymbol{p}_{k-1}^*) - C_{h_k^{(i)}}(h|\boldsymbol{p}_{k-1}^*) + C_{\mathbf{p}_k^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*)\right]^{g-1} P_{\mathbf{p}_k^{(i)}}(f, h|\boldsymbol{p}_{k-1}^*),
\end{aligned}$$

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

If the mean $\overline{f_{k,\tau}^{(i)}}$ is smaller than f_{k-1}^* at the end of the maturation step, the chosen collective is likely to have maximum f values among g matured collectives. Then, the definition of f^* is written by $f^* = \max(f^{(1)}, f^{(2)}, \cdots, f^{(g)})$. We rewrite the joint cumulative distribution function $C_{\mathbf{p}^*}(f, h|\boldsymbol{p}_{k-1}^*)$ to be little different with Eq. [61] because now we have to utilize the condition $f^* < f$ instead of $f^* > f$,

$$\begin{aligned}
C_{\mathbf{p}^*}(f, h|\boldsymbol{p}_{k-1}^*) &= \int_0^f df' \int_0^h dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&= \int_0^f df' (\int_0^1 - \int_h^1) dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) \\
&= \int_0^f df' \int_0^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*) - \int_0^f df' \int_h^1 dh' P_{\mathbf{p}^*}(f', h'|\boldsymbol{p}_{k-1}^*). \\
&= \Pr[f^* < f|\boldsymbol{p}_{k-1}^*] - \Pr[f^* < f \wedge h^* \geq h|\boldsymbol{p}_{k-1}^*]. \qquad [67]
\end{aligned}$$

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

$$\begin{aligned}
\Pr[f^* < f \wedge h^* \geq h|\boldsymbol{p}^*_{k-1}] &= \Pr[f^{(1)} < f \wedge h^{(1)} \geq h \wedge f^{(2)} < f \wedge h^{(2)} \geq h \wedge \cdots \wedge f^{(2)} < f \wedge h^{(2)} \geq h|\boldsymbol{p}^*_{k-1}] \\
&= \left[\Pr[f^{(i)} < f \wedge h^{(i)} \geq h|\boldsymbol{p}^*_{k-1}]\right]^g \\
&= \left[\Pr[f^{(i)} < f|\boldsymbol{p}^*_{k-1}] - \Pr[f^{(i)} < f \wedge h^{(i)} < h|\boldsymbol{p}^*_{k-1}]\right]^g \\
&= \left[C_{f^{(i)}}(f|\boldsymbol{p}^*_{k-1}) - C_{\mathbf{p}^{(i)}}(f,h|\boldsymbol{p}^*_{k-1})\right]^g . \qquad [68]
\end{aligned}$$

Thus, the joint cumulative distribution function is given by

$$C_{\mathbf{p}^*}(f,h|\boldsymbol{p}^*_{k-1}) = \left[C_{f^{(i)}}(f|\boldsymbol{p}^*_{k-1})\right]^g - \left[C_{f^{(i)}}(f|\boldsymbol{p}^*_{k-1}) - C_{\mathbf{p}^{(i)}}(f,h|\boldsymbol{p}^*_{k-1})\right]^g \qquad [69]$$

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

$$\begin{aligned}
P_{\mathbf{p}^*_k}(f,h|\boldsymbol{p}^*_{k-1}) &= \frac{\partial^2}{\partial f \partial h} C_{\mathbf{p}^*_k}(f,h|\boldsymbol{p}^*_{k-1}) \\
&= g(g-1)\left[C_{f^{(i)}_k}(f|\boldsymbol{p}^*_{k-1}) - C_{\mathbf{p}^{(i)}_k}(f,h|\boldsymbol{p}^*_{k-1})\right]^{g-2} \\
&\quad \times (P_{f^{(i)}_k}(f|\boldsymbol{p}^*_{k-1}) - \partial_f C_{\mathbf{p}^{(i)}_k}(f,h|\boldsymbol{p}^*_{k-1}))\partial_h C_{\mathbf{p}^{(i)}_k}(f,h|\boldsymbol{p}^*_{k-1}) \\
&\quad + g\left[C_{f^{(i)}_k}(f) - C_{\mathbf{p}^{(i)}_k}(f,h|\boldsymbol{p}^*_{k-1})\right]^{g-1} P_{\mathbf{p}^{(k)}_k}(f,h|\boldsymbol{p}^*_{k-1}). \qquad [70]
\end{aligned}$$

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

$$\Psi(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}) = \begin{cases}
g(g-1)\left[1 - C_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - C_{h^{(i)}_k}(h|\mathbf{p}^*_{k-1}) + C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1})\right]^{g-2} \\
\times (P_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - \partial_f C_{\mathbf{p}^{(i)}}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}))(P_{h^{(i)}_k}(h) - \partial_h C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1})) \\
+ g\left[1 - C_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - C_{h^{(i)}_k}(h|\mathbf{p}^*_{k-1}) + C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1})\right]^{g-1} P_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}) \\
\qquad\qquad \text{for } \overline{f_{k,\tau}} - f^*_{k-1} \geq 0 \text{ and } \overline{h_{k,\tau}} - h^*_{k-1} \geq 0 \\
g(g-1)\left[C_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - C_{\mathbf{p}^{(i)}_k}(f,h|\boldsymbol{p}^*_{k-1})\right]^{g-2} \\
(P_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - \partial_f C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}))\partial_h C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}) \\
+ g\left[C_{f^{(i)}_k}(f|\mathbf{p}^*_{k-1}) - C_{\mathbf{p}^{(i)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1})\right]^{g-1} P_{\mathbf{p}^{(k)}_k}(\mathbf{p}^*_k|\mathbf{p}^*_{k-1}) \\
\qquad\qquad \text{for } \overline{f_{k,\tau}} - f^*_{k-1} < 0 \text{ and } \overline{h_{k,\tau}} - h^*_{k-1} \geq 0
\end{cases} . \qquad [71]$$

Using Eq. [71], we get the mean values of f^* and h^* as

$$\begin{aligned}
\overline{f^*_k} &= \int_0^1 df' \int_0^1 dh' \; f' \times P_{f^*_k,h^*_k}(f',h'|f^*_{k-1},h^*_{k-1}), \qquad [72] \\
\overline{h^*_k} &= \int_0^1 df' \int_0^1 dh' \; h' \times P_{f^*_k,h^*_k}(f',h'|f^*_{k-1},h^*_{k-1}). \qquad [73]
\end{aligned}$$

We define the accessible region in frequency space where the signs of the changes in both double-mutant frequency and single-mutant frequency after a cycle are opposite to that of maturation (see Fig. S6),

$$\text{sign}\left(\overline{f^*_k} - f^*_{k-1}\right) \times \text{sign}\left(\overline{f_{k,\tau}} - f^*_{k-1}\right) \leq 0 \text{ and } \text{sign}\left(\overline{h^*_k} - h^*_{k-1}\right) \times \text{sign}\left(\overline{h_{k,\tau}} - h^*_{k-1}\right) \leq 0, \qquad [74]$$

where $\overline{f_{k,\tau}}$ and $\overline{h_{k,\tau}}$ are the mean values after the maturation step in cycle k. Figure S6**a** shows a concise explanation of the classification of regions where the initial and target compositions satisfy the above conditions in Eq. [74]. 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. S6**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. S6**b**. Both mutant frequencies tend to increase, and the inter-collective selection can compensate for these changes if the composition changes slowly when the mutant frequencies are small. However, if the changes occur too rapidly when the double-mutant frequency is intermediate, the frequency cannot be stabilized. So the accessible region is limited to the regions where the composition changes slowly.

The success of artificial selection is dependent on the location of the target composition relative to the accessible region. As seen in Figs. S6**c**, **e**, **h**, and **k**, the target composition locates within the accessible region, resulting in a successful selection. Conversely, if the target composition is not within the accessible region, as shown in Figs. S6**d**, **g**, and **j**, artificial selection fails.

It is important to note that the accessible region does not always predict the success of artificial selection. This region only indicates locally stable compositions but does not guarantee that the target composition can be reached from any initial

composition. This is demonstrated in Figs. S6**f**, **i**, and **j**, where the selected collectives cannot approach the target values because the trajectories should pass through regions where natural selection is strong. Additionally, even when a path exists to reach the target composition without leaving the accessible region, as in Fig. S6**f**, the composition may not be able to return to the accessible region once it falls outside of it. Thus, the trajectory information is necessary to fully predict the success of artificial selection, which cannot be obtained solely from the accessible region.

5. Derivation of equations

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

Assumptions: $\mu \ll s, r$

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

$$\left[\frac{d\overline{m}(t)}{dt} - (r+s)\overline{m}(t)\right] e^{-(r+s)t} = \mu\overline{w}_o e^{(r-\mu)t} e^{-(r+s)t} \tag{75}$$

$$\frac{d\left[\overline{m}(t)e^{-(r+s)t}\right]}{dt} = \mu\overline{w}_o e^{-(s+\mu)t} \tag{76}$$

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

$$\overline{m}(t) = \overline{m}_o e^{(r+s)t} + \frac{\mu\overline{w}_o}{s+\mu}\left(e^{(r+s)t} - e^{(r-\mu)t}\right) \approx \overline{m}_o e^{(r+s)t} + \frac{\mu\overline{w}_o}{s}\left(e^{(r+s)t} - e^{rt}\right). \tag{77}$$

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

$$\begin{aligned}\frac{d\overline{w^2}}{dt} &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} w^2 \frac{dP(w,m,t)}{dt} \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}[rw^2(w-1)P(w-1,m,t) + w^2(r+s)(m-1)P(w,m-1,t) \\ &\quad +\mu(w+1)w^2P(w+1,m-1,t) - (rw + (r+s)m + \mu w)w^2P(w,m,t)]\end{aligned} \tag{78}$$

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 w starts from 0 or 1 because the term is zero for $w = 0$. Thus, the first violet-colored term is equivalent to $\sum_{m=0}^{\infty}\sum_{w=1}^{\infty}\left[rw^2(w-1)P(w-1,m,t)\right]$. Let $\alpha = w-1$, and this becomes $\sum_{m=0}^{\infty}\sum_{\alpha=0}^{\infty}\left[r(\alpha+1)^2\alpha P(\alpha,m,t)\right]$. We reassign α as w, and obtain:

$$\begin{aligned}&\sum_{m=0}^{\infty}\sum_{w=0}^{\infty}\left[r(w+1)^2wP(w,m,t)\right] - \sum_{m=0}^{\infty}\sum_{w=0}^{\infty} rw^3P(w,m,t) \\ &= r\sum_{m=0}^{\infty}\sum_{w=0}^{\infty}\left[(2w+1)w\right]P(w,m,t) \\ &= r(2\overline{w^2} + \overline{w}).\end{aligned} \tag{79}$$

We collect the two green terms, and similarly obtain:

$$\begin{aligned}&\sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[w^2(r+s)(m-1)P(w,m-1,t) - (r+s)mw^2P(w,m,t)\right] \\ &= (r+s)\sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[w^2mP(w,m,t) - mw^2P(w,m,t)\right] \\ &= 0.\end{aligned} \tag{80}$$

Finally, we collect the two orange terms. For the first orange term, the sum is the same regardless of whether we start from $w = 0$ or -1. Let w start from -1, and we have $\sum_{m=0}^{\infty}\sum_{w=-1}^{\infty}\left[\mu(w+1)w^2P(w+1,m-1,t)\right]$. Let $\alpha = w+1$,then the term becomes $\sum_{m=0}^{\infty}\sum_{\alpha=0}^{\infty}\left[\mu\alpha(\alpha-1)^2P(\alpha,m-1,t)\right]$. We reassign α as w, and additionally apply index change on $m-1$:

$$\begin{aligned}&\mu\sum_{m=0}^{\infty}\sum_{w=0}^{\infty}\left[w(w-1)^2P(w,m,t) - w^3P(w,m,t)\right] \\ &= \mu\sum_{m=0}^{\infty}\sum_{w=0}^{\infty}\left[(-2w^2+w)P(w,m,t)\right] \\ &= \mu\left(-2\overline{w^2} + \overline{w}\right).\end{aligned} \tag{81}$$

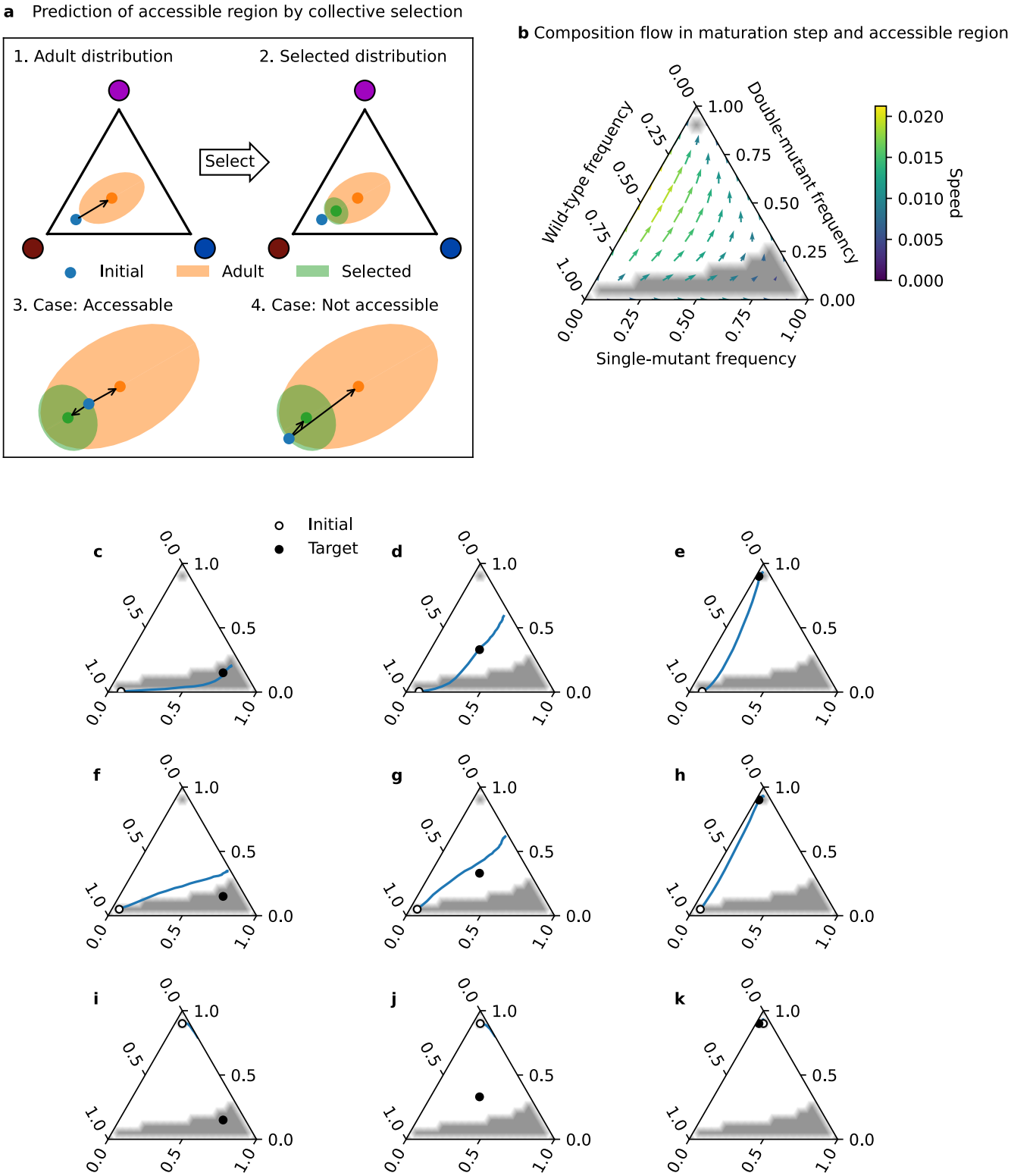

Fig. S6. **a** The schematic procedure to determine accessible regions by collective selection. From the given parent composition (blue dot), the probability distribution of offspring Adults (Eq. [60]) is computed (marked in the orange-colored area). From the Adult compositions, the probability distribution of the selected offspring (Eq.[71]) is computed (marked in the green area). If the signs of the changes in both double-mutant frequency and single-mutant 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 in maturation. The length and color of the arrows indicate the speed of composition changes. **c**-**k** Trajectories of mean compositions for 30 simulations starting from (m_0,v_0)=(85,5), (50,50), and (50,900) with the targets $(\hat{f},\hat{h})$=(0.7,0.15), (0.33,0.33), and (0.02,0.9). The axis labels are the same with **b**. The figures are drawn using *mpltern* package (5).

Now, add the three parts together, and we have

$$\begin{aligned}\frac{d\overline{w^2}}{dt} &= \mu\left(-2\overline{w^2}+\overline{w}\right)+r(2\overline{w^2}+\overline{w}) \\ &= 2\overline{w^2}(r-\mu)+\overline{w}(r+\mu), \end{aligned} \quad [82]$$

which is Eq. [11]. Likewise,

$$\begin{aligned}\frac{d\overline{m^2}}{dt} &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty} m^2\frac{dP(w,m,t)}{dt} \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}[rm^2(w-1)P(w-1,m,t)+m^2(r+s)(m-1)P(w,m-1,t) \\ &\quad +\mu(w+1)m^2P(w+1,m-1,t)-(rw+(r+s)m+\mu w)m^2P(w,m,t)] \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[rm^2w+(m+1)^2(r+s)m+\mu w(m+1)^2-(rw+(r+s)m+\mu w)m^2\right]P(w,m,t) \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[(m+1)^2(r+s)m+\mu w(m+1)^2-((r+s)m+\mu w)m^2\right]P(w,m,t) \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[(r+s)m(2m+1)+\mu w(2m+1)\right]P(w,m,t) \quad [83] \\ &= 2(r+s)\overline{m^2}+(r+s)\overline{m}+2\mu\overline{wm}+\mu\overline{w} \quad [84]\end{aligned}$$

which is Eq. [12].

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

$$\left[\frac{d\overline{w^2}}{dt}-2(r-\mu)\overline{w^2}\right]e^{-2(r-\mu)t} = (r+\mu)\overline{w}_o e^{(r-\mu)t}e^{-2(r-\mu)t}, \quad [85]$$

$$\frac{d\left[\overline{w^2}e^{-2(r-\mu)t}\right]}{dt} = (r+\mu)\overline{w}_o e^{(\mu-r)t}, \quad [86]$$

$$\overline{w^2}e^{-2(r-\mu)t} = \frac{r+\mu}{\mu-r}\overline{w}_o\left[e^{(\mu-r)t}-1\right]+\overline{w^2}_o. \quad [87]$$

Since $\mu \ll r$,we have

$$\begin{aligned}\overline{w^2} &= \frac{r+\mu}{\mu-r}\overline{w}_o\left[e^{(\mu-r)t}-1\right]e^{2(r-\mu)t}+\overline{w^2}_o e^{2(r-\mu)t} \\ &\approx \overline{w}_o e^{2rt}\left[1-e^{-rt}\right]+\overline{w^2}_o e^{2rt}, \end{aligned} \quad [88]$$

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

$$\begin{aligned}\sigma_w^2(t) &= \overline{w^2}(t)-[\overline{w}(t)]^2 \\ &\approx \overline{w}_o e^{2rt}\left[1-e^{-rt}\right]+\overline{w^2}_o e^{2rt}-\left(\overline{w}_o e^{rt}\right)^2 \\ &= \overline{w}_o e^{2rt}\left[1-e^{-rt}\right]+e^{2rt}\left(\overline{w^2}_o-(\overline{w}_o)^2\right)=\overline{w}_o e^{2rt}\left[1-e^{-rt}\right]+e^{2rt}\sigma_w^2(0). \end{aligned} \quad [89]$$

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

$$\begin{aligned}\frac{d\overline{wm}}{dt} &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}[wmr(w-1)P(w-1,m,t)+wm(r+s)(m-1)P(w,m-1,t) \\ &\quad +wm\mu(w+1)P(w+1,m-1,t) \\ &\quad -wm(rw+(r+s)m+\mu w)P(w,m,t)] \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}[(w+1)mrw+w(m+1)(r+s)m+(w-1)(m+1)\mu w \\ &\quad -wm(rw+(r+s)m+\mu w)]P(w,m,t) \\ &= \sum_{w=0}^{\infty}\sum_{m=0}^{\infty}\left[2rwm+swm-\mu mw+\mu w^2-\mu w)\right]P(w,m,t) \\ &= (2r+s-\mu)\overline{wm}+\mu(\overline{w^2}-\overline{w}). \end{aligned} \quad [90]$$

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

$$\frac{d[\overline{wm}e^{-(2r+s-\mu)t}]}{dt} = \mu(\overline{w^2}-\overline{w})e^{-(2r+s-\mu)t}. \quad [91]$$

Thus, we have

$$\begin{aligned}
\overline{wm}e^{-(2r+s-\mu)t}-\overline{wm}_o &= \mu\int(\overline{w^2}-\overline{w})e^{-(2r+s-\mu)t}dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{w}_o\left[e^{(\mu-r)t}-1\right]e^{2(r-\mu)t}+\overline{w^2}_o e^{2(r-\mu)t}-\overline{w}_o e^{(r-\mu)t})e^{-(2r+s-\mu)t}dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{w}_o\left[e^{-(\mu-r)t}-e^{2(r-\mu)t}\right]e^{-(2r+s-\mu)t}+\overline{w^2}_o e^{2(r-\mu)t}e^{-(2r+s-\mu)t} \\
&\quad -\overline{w}_o e^{(r-\mu)t-(2r+s-\mu)t})dt \\
&= \mu\int(\frac{r+\mu}{\mu-r}\overline{w}_o e^{-(r+s)t}-\frac{r+\mu}{\mu-r}\overline{w}_o e^{-(\mu+s)t}+\overline{w^2}_o e^{-(\mu+s)t}-\overline{w}_o e^{-(r+s)t})dt \\
&= \mu\left\{\frac{2r}{\mu-r}\overline{w}_o\frac{e^{-(r+s)t}-1}{-(r+s)}+(\overline{w^2}_o-\frac{r+\mu}{\mu-r}\overline{w}_o)\frac{e^{-(\mu+s)t}-1}{-(\mu+s)}\right\}, \quad [92]
\end{aligned}$$

which results in

$$\begin{aligned}
\overline{wm}(t) &= \overline{wm}_o e^{(2r+s-\mu)t}+\mu\left\{\frac{2r}{\mu-r}\overline{w}_o\frac{e^{-(r+s)t}-1}{-(r+s)}+(\overline{w^2}_o-\frac{r+\mu}{\mu-r}\overline{w}_o)\frac{e^{-(\mu+s)t}-1}{-(\mu+s)}\right\}e^{(2r+s-\mu)t} \\
&= \overline{wm}_o e^{(2r+s-\mu)t}+\frac{2r\mu}{\mu-r}\overline{w}_o\frac{e^{(r-\mu)t}-e^{(2r+s-\mu)t}}{-(r+s)}+\mu(\overline{w^2}_o-\frac{r+\mu}{\mu-r}\overline{w}_o)\frac{e^{2(r-\mu)t}-e^{(2r+s-\mu)t}}{-(\mu+s)} \\
&\approx \overline{wm}_o e^{(2r+s)t}-\frac{2\mu\overline{w}_o}{r+s}\left(e^{(2r+s)t}-e^{rt}\right)+\frac{\mu}{s}(\overline{w^2}_o+\overline{w}_o)\left(e^{(2r+s)t}-e^{2rt}\right). \quad [93]
\end{aligned}$$

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

$$\left(\frac{d\overline{m^2}}{dt}-2(r+s)\overline{m^2}\right)e^{-2(r+s)t}=\frac{d\left(\overline{m^2}e^{-2(r+s)t}\right)}{dt}=((r+s)\overline{m}+2\mu\overline{wm}+\mu\overline{w})\,e^{-2(r+s)t}. \quad [94]$$

The right-hand side becomes

$$\begin{aligned}
&\approx(r+s)\left(\overline{m}_o e^{(r+s)t}+\frac{\mu\overline{w}_o}{s+\mu}\left(e^{(r+s)t}-e^{(r-\mu)t}\right)\right)e^{-2(r+s)t}+2\mu\overline{wm}_o e^{-(s+\mu)t}+\mu\overline{w}_o e^{-(r+2s+\mu)t}+o(\mu^2) \\
&=(r+s)\left(\overline{m}_o e^{-(r+s)t}+\frac{\mu\overline{w}_o}{s+\mu}\left(e^{-(r+s)t}-e^{-(r+2s+\mu)t}\right)\right)+2\mu\overline{wm}_o e^{-(s+\mu)t}+\mu\overline{w}_o e^{-(r+2s+\mu)t}+o(\mu^2) \\
&=(r+s)\left(\frac{\overline{m}_o s+N_0\mu}{s+\mu}e^{-(r+s)t}-\frac{\mu\overline{w}_o}{s+\mu}e^{-(r+2s+\mu)t}\right)+2\mu\overline{wm}_o e^{-(s+\mu)t}+\mu\overline{w}_o e^{-(r+2s+\mu)t}+o(\mu^2). \quad [95]
\end{aligned}$$

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:

$$\begin{aligned}
\overline{m^2}e^{-2(r+s)t}-\overline{m^2}_o &\approx (r+s)\left(\frac{\overline{m}_o s+N_0\mu}{(s+\mu)(r+s)}(1-e^{-(r+s)t})+\frac{\mu\overline{w}_o}{(s+\mu)(r+2s+\mu)}(e^{-(r+2s+\mu)t}-1)\right) \\
&\quad +\frac{2\mu\overline{wm}_o}{s+\mu}(1-e^{-(s+\mu)t})+\frac{\mu\overline{w}_o}{r+2s+\mu}(1-e^{-(r+2s+\mu)t})+o(\mu^2). \quad [96]
\end{aligned}$$

Then, we have

$$
\begin{aligned}
\overline{m^2}(t) &\approx \overline{m^2}_o e^{2(r+s)t} + (r+s)\left(\frac{\overline{m}_o s + N_0\mu}{(s+\mu)(r+s)}(e^{2(r+s)t} - e^{(r+s)t}) + \frac{\mu\overline{w}_o}{(s+\mu)(r+2s+\mu)}(e^{(r-\mu)t} - e^{2(r+s)t})\right) \\
&\quad + \frac{2\mu\overline{wm}_o}{s+\mu}(e^{2(r+s)t} - e^{(2r+s-\mu)t}) + \frac{\mu\overline{w}_o}{r+2s+\mu}(e^{2(r+s)t} - e^{(r-\mu)t}) + o(\mu^2) \\
&= \overline{m^2}_o e^{2(r+s)t} + \left(\frac{\overline{m}_o s + N_0\mu}{s}(e^{2(r+s)t} - e^{(r+s)t}) + (\frac{1}{s} - \frac{1}{r+2s})\mu\overline{w}_o(e^{rt} - e^{2(r+s)t})\right) \\
&\quad + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \frac{\mu\overline{w}_o}{r+2s}(e^{2(r+s)t} - e^{rt}) + o(\mu^2) \\
&= \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{\overline{m}_o}{\overline{w}_o}(e^{2(r+s)t} - e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&= \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + \frac{\mu\overline{m}_o}{s}(e^{2(r+s)t} - e^{(r+s)t}) + o(\mu^2) \\
&= \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o(1 + \frac{\mu}{s})e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&= \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2).
\end{aligned} \tag{97}
$$

Equation [19].

$$
\begin{aligned}
\sigma_m^2(t) &= \overline{m^2}(t) - [\overline{m}(t)]^2 \\
&\approx \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&\quad - \left[\overline{m}_o e^{(r+s)t} + \frac{\mu\overline{w}_o}{s}(e^{(r+s)t} - e^{rt})\right]^2 \\
&= \overline{m^2}_o e^{2(r+s)t} + \frac{2\mu\overline{wm}_o}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[(-e^{(r+s)t}) + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2) \\
&\quad - \overline{m}_o^2 e^{2(r+s)t} - 2\overline{m}_o e^{(r+s)t}\frac{\mu\overline{w}_o}{s}(e^{(r+s)t} - e^{rt}) - \left[\frac{\mu\overline{w}_o}{s}(e^{(r+s)t} - e^{rt})\right]^2 \\
&= \sigma_{m,0}^2 e^{2(r+s)t} + \frac{2\mu(\overline{wm}_o - \overline{m}_o\overline{w}_o)}{s}e^{(r+s)t}(e^{(r+s)t} - e^{rt}) + \overline{m}_o e^{(r+s)t}(e^{(r+s)t} - 1) \\
&\quad + \frac{\mu\overline{w}_o}{s}\left[-e^{(r+s)t} + \frac{2s}{r+2s}e^{2(r+s)t} + \frac{r}{r+2s}e^{rt}\right] + o(\mu^2).
\end{aligned} \tag{98}
$$

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.

$$
\begin{aligned}
f &\approx \frac{\overline{m} + \sigma_m G_m}{\overline{w} + \sigma_w G_w + \overline{m} + \sigma_m G_m} \\
&= \frac{\overline{m}\left(1 + \frac{\sigma_m}{\overline{m}}G_m\right)}{(\overline{w} + \overline{m})(1 + \frac{\sigma_w G_w + \sigma_m G_m}{\overline{w} + \overline{m}})} \\
&\approx \frac{\overline{m}}{\overline{w} + \overline{m}}\left(1 + \frac{\sigma_m}{\overline{m}}G_m\right)\left(1 - \frac{\sigma_w G_w + \sigma_m G_m}{\overline{w} + \overline{m}}\right) \\
&\approx \overline{f}\left(1 + \frac{\sigma_m}{\overline{m}}G_m - \frac{\sigma_w G_w + \sigma_m G_m}{\overline{w} + \overline{m}}\right) \\
&= \overline{f}\left(1 + \left(\frac{\sigma_m}{\overline{m}} - \frac{\sigma_m}{\overline{w} + \overline{m}}\right)G_m - \frac{\sigma_w}{\overline{w} + \overline{m}}G_w\right).
\end{aligned} \tag{99}
$$

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

$$\begin{aligned}
\overline{f}(t) &= \frac{\overline{m}(t)}{\overline{w}(t) + \overline{m}(t)} \approx \frac{\overline{m}_o e^{(r+s)t} + \frac{\mu \overline{w}_o}{s}(e^{(r+s)t} - e^{rt})}{\overline{w}_o e^{rt} + \overline{m}_o e^{(r+s)t} + \frac{\mu \overline{w}_o}{s}(e^{(r+s)t} - e^{rt})} \\
&= \frac{\overline{m}_o e^{st} + \frac{\mu \overline{w}_o}{s}(e^{st} - 1)}{\overline{w}_o + \overline{m}_o e^{st} + \frac{\mu \overline{w}_o}{s}(e^{st} - 1)} \\
&= \frac{\overline{f}_o e^{st} + \frac{\mu}{s}(1 - \overline{f}_o)(e^{st} - 1)}{(1 - \overline{f}_o) + \overline{f}_o e^{st} + \frac{\mu}{s}(1 - \overline{f}_o)(e^{st} - 1)}. \qquad [100]
\end{aligned}$$

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

$$\begin{aligned}
\sigma_f^2(t) &= \overline{f}^2 \left(\left(\frac{\overline{w}}{(\overline{w} + \overline{m})\overline{m}} \right)^2 \sigma_m^2(t) + \left(\frac{1}{\overline{w} + \overline{m}} \right)^2 \sigma_w^2(t) \right) \\
&= \overline{f}^2 \left(\frac{(1 - \overline{f}(t))^2}{\overline{N}(t)^2 (\overline{f}(t))^2} \sigma_m(t) + \frac{1}{\overline{N}(t)^2} \sigma_w^2(t) \right) \\
&= \frac{(1 - \overline{f}(t))^2 \sigma_m^2(t) + \overline{f}(t)^2 \sigma_w^2(t)}{\overline{N}(t)^2}. \qquad [101]
\end{aligned}$$

References

1. DT Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. *The J. chemical physics* **115**, 1716–1733 (2001).
2. Y Cao, DT Gillespie, LR Petzold, Efficient step size selection for the tau-leaping simulation method. *The J. chemical physics* **124**, 044109 (2006).
3. Q Zheng, Progress of a half century in the study of the luria–delbrück distribution. *Math. biosciences* **162**, 1–32 (1999).
4. EJ Gumbel, *Statistics of extremes.* (Columbia university press), (1958).
5. Y Ikeda, B Grabowski, F Körmann, mpltern 0.3.0: ternary plots as projections of Matplotlib (2019).