

GENETICS

Supporting Information

<http://www.genetics.org/content/suppl/2011/06/06/genetics.110.124297.DC1>

On the Fixation Process of a Beneficial Mutation in a Variable Environment

Hildegard Uecker and Joachim Hermisson

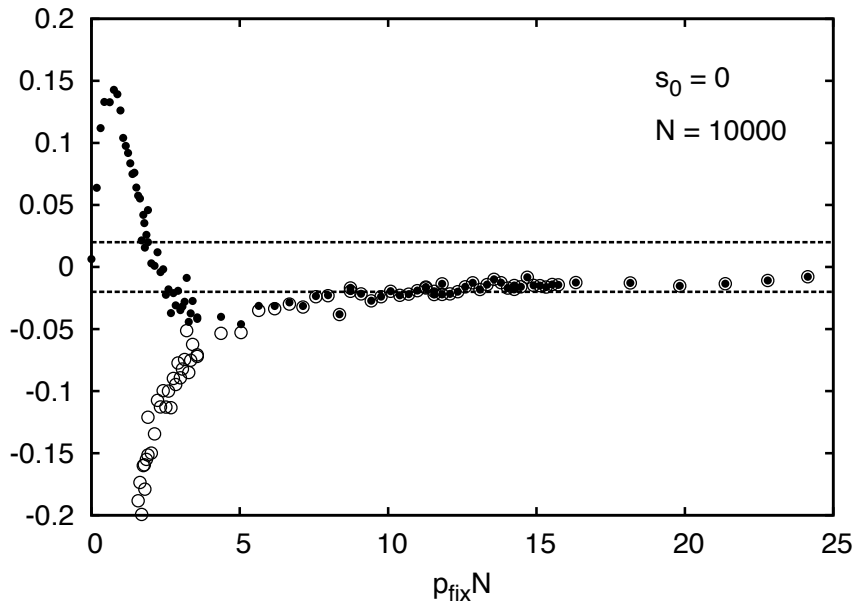


Figure S1: Relative deviation between analytical and simulation results for an allele with selective advantage $s(t) = s_1 t$. The empty dots are obtained from branching process approximation, the small filled dots from the corrected version Eq. (S1). One sees that the corrected version provides a much better approximation for small values of $p_{fix}N$; for large values of $p_{fix}N$, the results coincide. Simulations were performed for a population of $N = 10\,000$ individuals, and each simulation point is the average over $5 \cdot 10^7$ runs.

S1 Accuracy of the approximation

Weak selection. As pointed out in the main text, deviations from the exact solution are expected for weak selection.

For an allele with constant selective advantage and $N = const.$, a comparison to the exact solution is immediately possible (see main text).

As a further example, we consider the fixation probability of an allele with $s(t) = s_1 t$ and examine the relative error between theoretical prediction and simulation results in dependence of $p_{fix}N$ (see Figure S1). A relative error of less than 2% is found for $p_{fix}N < 10$. The results for small values of $p_{fix}N$ can be improved if we correct the branching approximation by a factor $1/(1 - \exp(-p_{fix}N))$:

$$p_{fix}^* = \frac{p_{fix}}{1 - \exp(-p_{fix}N)}. \quad (S1)$$

This heuristic approximation is inspired by Eq. (18). The relative deviation from the simulation results is added to Figure S1. It is seen that the approximation is considerably improved.

In Figure S2, we compare the distribution function of $T_{1/2}$ to simulation results for various values of s and N where s and N are constant. While agreement is excellent for high values of α , deviations increase for decreasing values of α . A maximum absolute deviation of ≥ 0.05 between theory and simulations is found for $\alpha = Ns \lesssim 10$.

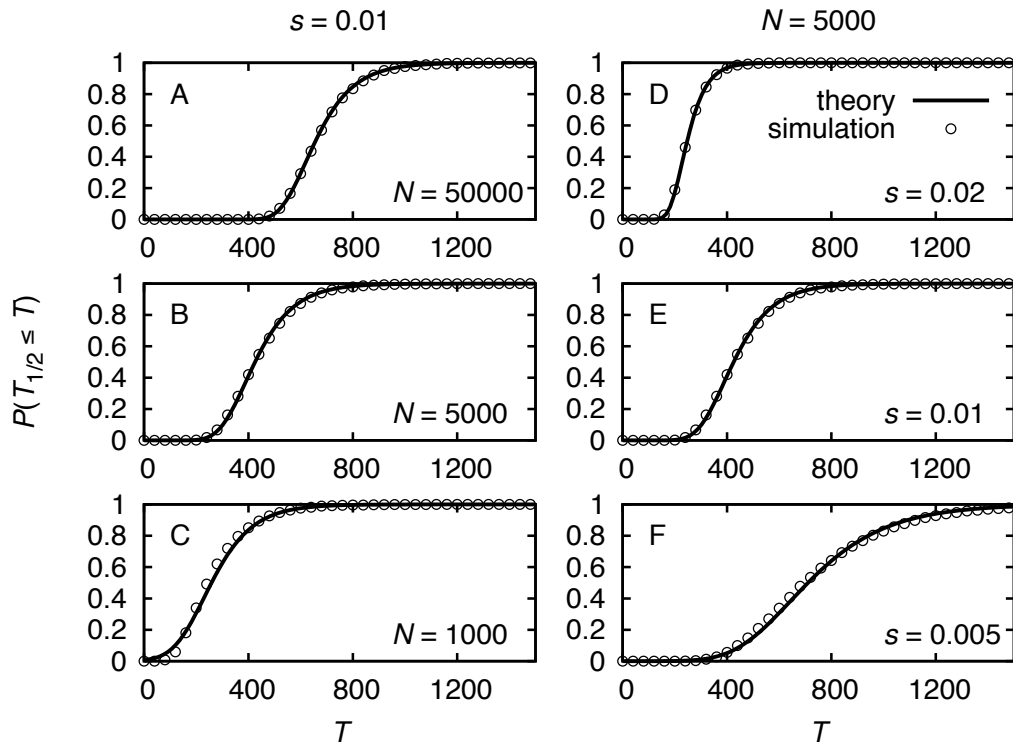


Figure S2: Distribution of the time to reach frequency $x_c = 0.5$ for various values of s and N , where s and N are constant. Analytical and simulation results are compared. Simulation results are averaged over 5000 runs. A maximum absolute deviation of ≥ 0.05 between theory and simulations is found for $\alpha = Ns \lesssim 10$ (see panel C).

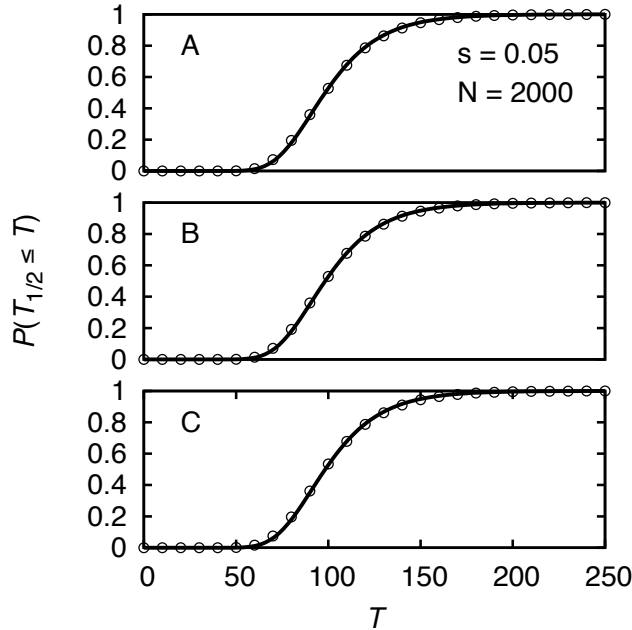


Figure S3: Distribution function of $T_{1/2}$ for three different modes of density regulation. The selection strength is $s = 0.05$ and the initial population size $N_0 = 2000$. A: Inherently constant population size, i.e. $b(t, N_t) = d(t, N_t) = 0$, with $\xi(t, N) = 1$. B: Lotka-Volterra dynamics, i.e. $b(t, N_t) = b + \rho(1 - N_t/K)$ and $d(t, N_t) = b$, with $b = 1.0$, $r = 1.0$ and $K = N_0 = 2000$. C: Freely fluctuating population size, i.e. $b(t, N_t) = d(t, N_t) = b$, with $b = 1.0$. Simulations are averaged over 50 000 runs. The maximum absolute deviation is of the same order of magnitude for all three scenarios.

Mode of density-regulation. To test whether the mode of density-regulation influences the outcome, some exemplary simulations for three different scenarios of density regulation were performed ($b(t, N_t) = d(t, N_t) = \text{const.}$, Lotka-Volterra dynamics, i.e. $b(t, N_t) = b + \rho(1 - N_t/K)$ and $d(t, N_t) = b$, and $b(t, N_t) = d(t, N_t) = 0$ with accordingly chosen ξ). Figure S3 shows a comparison of the simulation results to the theoretical curve. We see that demographic stochasticity has only a slight effect (note that we used a relatively low (initial) population size of $N_0 = 2000$ for the examples to make demographic stochasticity strong).

S2 Diffusion approximation

Pioneered by Kimura (1957), the diffusion approximation has been established as the second main approach for the analytical analysis of the fixation process. In this appendix, we show how an approximation for the fixation probability with variable selection and population size can be derived within the diffusion framework. We also use diffusion results for the mean fixation time for constant s and N to estimate the error of the branching process approach in this case.

As in the main text, we assume that the population inhabits a variable environment. All key model parameters may explicitly depend on time. In particular, let $s = s(t)$ be the selection coefficient of the A allele and $N = N(t)$ the

total population size. Finally, the variance in offspring number is $\sigma^2 = \sigma^2(t)$. Following standard practice, we can combine N and σ^2 in the variance effective population size, $N_e(t) = N(t)/\sigma^2(t)$.

We can formulate the model as a diffusion process as follows. Let $g_t(x \rightarrow x + \epsilon)$ be the transition probability that the allele frequency changes from x at time t to $x + \epsilon$ at time $t + dt$. The diffusion is characterized by the mean and the variance of the transition probability in the limit of small dt and ϵ :

$$E_{x,t}[\epsilon]dt = \int_{-\infty}^{\infty} \epsilon g_t(x \rightarrow x + \epsilon)d\epsilon, \quad (\text{S2a})$$

$$\text{Var}_{x,t}[\epsilon]dt = \int_{-\infty}^{\infty} \epsilon^2 g_t(x \rightarrow x + \epsilon)d\epsilon + \mathcal{O}[dt^2]. \quad (\text{S2b})$$

To leading order in N_e^{-1} ,

$$E_{x,t}[\epsilon] = s(t)x(1-x), \quad (\text{S3a})$$

$$\text{Var}_{x,t}[\epsilon] = \frac{\sigma^2(t)}{N(t)}x(1-x) = \frac{x(1-x)}{N_e(t)}. \quad (\text{S3b})$$

Define $f(x, \tau|p, t)$ as the probability density for an allele frequency of x at time $t + \tau$, given that the frequency at time t was p . Then

$$\begin{aligned} f(x, \tau + dt|p, t) &= \int_{-\infty}^{\infty} g_t(p \rightarrow p + \epsilon)f(x, \tau|p + \epsilon, t + dt)d\epsilon \\ &\approx f(x, \tau|p, t + dt) \\ &\quad + \left(\frac{\partial f(x, \tau|p, t + dt)}{\partial p} E_{p,t}[\epsilon] + \frac{1}{2} \frac{\partial^2 f(x, \tau|p, t + dt)}{\partial p^2} \text{Var}_{p,t}[\epsilon] \right) dt. \end{aligned} \quad (\text{S4})$$

In the limit $dt \rightarrow 0$, we obtain the Kolmogorov backward equation for the conditional density $f(x, \tau|p, t)$ of a time-inhomogeneous diffusion:

$$\frac{\partial f}{\partial \tau} - \frac{\partial f}{\partial t} = s(t)p(1-p) \frac{\partial f}{\partial p} + \frac{p(1-p)}{2N_e(t)} \frac{\partial^2 f}{\partial p^2}. \quad (\text{S5})$$

If $s(t)$ and $N_e(t)$ are both constant, $\partial f/\partial t = 0$ and (S5) reduces to the backward equation of the classical homogeneous case. With time dependence, the state of the population can no longer be described by the allele frequency x alone; f depends on both, x and t . Note that we measure time on a generations scale, not in $2N$ generations, as it is often done in diffusion theory.

Denote now as $P(p, t)$ the fixation probability of an allele with frequency p in the population at time t . Since fixation is the probability of eventual absorption of the diffusion at the boundary $x = 1$, we can write $P(p, t)$ in terms of the transition probability as

$$P(p, t) = \lim_{\substack{\tau \rightarrow \infty \\ y \rightarrow 1}} \int_y^1 f(x, \tau|p, t)dx. \quad (\text{S6})$$

Integrating (S5) over any frequency interval $[x_1, x_2]$, we see that the Kolmogorov backward equation also holds for the corresponding probabilities. For the probability of eventual fixation, in particular, the τ dependence vanishes and we obtain (Kimura and Ohta, 1974)

$$-\dot{P}(p, t) = s(t)p(1-p)P'(p, t) + \frac{p(1-p)}{2N_e(t)}P''(p, t), \quad (\text{S7})$$

where \dot{P} and P' denote derivatives with respect to t and p , respectively. In contrast to the time-homogeneous case, there is no exact solution to the time-inhomogeneous equation. However, it is possible to derive an approximate solution for small p by setting $1 - p \approx 1$ in (S7). This approximation was first used by Kimura and Ohta (1974) to derive the fixation probability for a logistically growing population. As we will see, a full solution of the general time-dependent model is possible under the same assumption. We use the following Ansatz,

$$P(p, t) = 1 - \exp(-Q(t)p). \quad (\text{S8})$$

Substituting this relation into the approximated PDE (ignoring $1 - p$ terms in Eq. (S7)) leads to the following ODE for $Q(t)$,

$$\frac{dQ(t)}{dt} = -s(t)Q(t) + \frac{Q^2(t)}{2N_e(t)}. \quad (\text{S9})$$

A general solution to this differential equation can be found,

$$Q(t) = \frac{2}{\int_t^\infty \frac{1}{N_e(\tau)} \exp\left(-\int_t^\tau s(y)dy\right) d\tau}, \quad (\text{S10})$$

and we obtain an approximation for the fixation probability from Eq. (S8). If the fixation process starts from a single copy of the derived allele that enters the population at time $t = 0$, in particular, we find

$$p_{fix,diff} := P(1/N(0), 0) = 1 - \exp(-Q(0)/N(0)) \approx Q(0)/N(0). \quad (\text{S11})$$

This expression may be directly compared to the result for the fixation probability p_{fix} (16b) from the branching process derivation. We find that

$$p_{fix} = \frac{p_{fix,diff}}{1 + p_{fix,diff}/2}. \quad (\text{S12})$$

Both expressions are thus equal to leading order in $p_{fix,diff}$ (i.e. to leading order in the selection strength).

The approximate solution from the diffusion implies the absence of competition among mutant alleles (which appears in (S7) through the p^2 terms). This is equivalent to the independence assumption of the branching process. Note also that the approximation is constructed in a way that it fulfills the boundary condition $P(0, t) = 0$, but not the condition $P(1, t) = 1$. It is therefore only valid if the initial frequency of the allele, p , is sufficiently small.

For constant selection, the diffusion equation (with correct boundary conditions at $p = 0$ and $p = 1$) can be solved exactly. In particular, the derivation of the mean fixation time is possible (Kimura and Ohta, 1969; Ewens, 2004). This allows for a comparison of our approximate solution with the exact diffusion result in this case. From Eq. (47), we obtain the following expression for the mean fixation time in the diffusion limit ($s \rightarrow 0$, $\alpha = sN/\xi = \text{const.}$):

$$\langle \tau_{fix} \rangle^{(1)} = \frac{2}{\alpha} \left(\int_0^{\alpha/2} \ln\left(\frac{\alpha - Y}{Y}\right) \exp(-Y) dY \right). \quad (\text{S13})$$

Within the diffusion framework, the mean fixation time is given by (see Ewens, 2004, p.140ff and p.167ff),

$$\begin{aligned} \langle \tau_{fix} \rangle^{(2)} &= \frac{2}{\alpha[\exp(\alpha) - 1]} \int_0^1 \frac{[\exp(\alpha x) - 1][\exp(\alpha(1-x)) - 1]}{x} dx \\ &= \frac{2}{\alpha[\exp(\alpha) - 1]} \int_0^\alpha \frac{\exp(\alpha) - \exp(Y) - \exp(\alpha - Y) + 1}{Y} dY \\ &\approx \frac{2}{\alpha} \left(\int_0^\alpha \frac{1 - \exp(-Y)}{Y} dY + \exp(-\alpha) \int_0^\alpha \frac{1 - \exp(Y)}{Y} dY \right). \end{aligned} \quad (\text{S14})$$

In the last step, we approximated $(\exp(\alpha) - 1)^{-1} \approx \exp(-\alpha)$. This approximation is of order $o(\exp(\alpha))$. By integration by parts,

$$\int_0^\alpha \ln\left(\frac{\alpha - Y}{Y}\right) \exp(-Y) dY = \int_0^\alpha \frac{1 - \exp(-Y)}{Y} dY + \exp(-\alpha) \int_0^\alpha \frac{1 - \exp(Y)}{Y} dY. \quad (\text{S15})$$

We thus obtain for the error term $\Delta(\alpha)$ of the branching result:

$$\Delta(\alpha) = \langle \tau_{fix} \rangle^{(2)} - \langle \tau_{fix} \rangle^{(1)} = \frac{2}{\alpha} \int_{\alpha/2}^\alpha \ln\left(\frac{\alpha - Y}{Y}\right) \exp(-Y) dY + o(\exp(-\alpha)). \quad (\text{S16})$$

Since

$$\begin{aligned} \exp\left(\frac{\alpha}{2}\right) \Delta(\alpha) &= \frac{2}{\alpha} \int_{\alpha/2}^\alpha \ln\left(\frac{\alpha - Y}{Y}\right) \exp\left[-\left(Y - \frac{\alpha}{2}\right)\right] dY \\ &= \int_0^1 \ln\left(\frac{1 - Y}{1 + Y}\right) \exp\left(-\frac{\alpha}{2} Y\right) dY \xrightarrow{\alpha \rightarrow \infty} 0, \end{aligned} \quad (\text{S17})$$

we conclude that $\Delta(\alpha)$ is of order $o(\exp(-\alpha/2))$.

S3 Additional explanations

Derivation of Eq. (4a) and (4b). To calculate the expected frequency change $E[\Delta x | x_t, N_t]$ in an infinitesimal time interval dt , we need to consider all events that change the frequency $x_t = n/N_t$ of mutants in the population. These events are summarized in Table 1. The second and third column give the probability of the event and the change Δx induced by it. Taking all events, their respective probabilities and effects in account, we obtain:

$$\begin{aligned} E[\Delta x | x_t, N_t] &= \left[(\xi(t, N_t) + s(t, N_t)) \frac{n(N_t - n)}{N_t} \left(\frac{n+1}{N_t} - \frac{n}{N_t} \right) + \xi(t, N_t) \frac{n(N_t - n)}{N_t} \left(\frac{n-1}{N_t} - \frac{n}{N_t} \right) \right. \\ &\quad + b(t, N_t) n \left(\frac{n+1}{N_t+1} - \frac{n}{N_t} \right) + b(t, N_t) (N_t - n) \left(\frac{n}{N_t+1} - \frac{n}{N_t} \right) + d(t, N_t) n \left(\frac{n-1}{N_t-1} - \frac{n}{N_t} \right) \\ &\quad \left. + d(t, N_t) (N_t - n) \left(\frac{n}{N_t-1} - \frac{n}{N_t} \right) \right] dt \\ &= s(t, N_t) x_t (1 - x_t) dt \end{aligned} \quad (\text{S18})$$

Since $E[\Delta x | x_t, N_t]^2$ is of order $\mathcal{O}((dt)^2)$, it holds for the variance of Δx :

$$\text{Var}[\Delta x | x_t, N_t] = E[(\Delta x)^2 | x_t, N_t] - E[\Delta x | x_t, N_t]^2 \approx E[(\Delta x)^2 | x_t, N_t]. \quad (\text{S19})$$

event	probability	Δx
a mutant replaces a resident	$(\xi(t, N_t) + s(t, N_t)) \frac{n(N_t - n)}{N_t} dt$	$\left(\frac{n+1}{N_t} - \frac{n}{N_t}\right)$
a resident replaces a mutant	$\xi(t, N_t) \frac{n(N_t - n)}{N_t} dt$	$\left(\frac{n-1}{N_t} - \frac{n}{N_t}\right)$
a mutant is born	$b(t, N_t) n dt$	$\left(\frac{n+1}{N_t+1} - \frac{n}{N_t}\right)$
a resident is born	$b(t, N_t)(N_t - n) dt$	$\left(\frac{n}{N_t+1} - \frac{n}{N_t}\right)$
a mutant dies	$d(t, N_t) n dt$	$\left(\frac{n-1}{N_t-1} - \frac{n}{N_t}\right)$
a resident dies	$d(t, N_t)(N_t - n) dt$	$\left(\frac{n}{N_t-1} - \frac{n}{N_t}\right)$

Table 1: Summary of the events that change the frequency x of mutants in the population.

We use again Table 1 for the calculation:

$$\begin{aligned}
E[(\Delta x)^2 | x_t, N_t] &= \left[(\xi(t, N_t) + s(t, N_t)) \frac{n(N_t - n)}{N_t} \left(\frac{n+1}{N_t} - \frac{n}{N_t}\right)^2 + \xi(t, N_t) \frac{n(N_t - n)}{N_t} \left(\frac{n-1}{N_t} - \frac{n}{N_t}\right)^2 \right. \\
&\quad + b(t, N_t) n \left(\frac{n+1}{N_t+1} - \frac{n}{N_t}\right)^2 + b(t, N_t)(N_t - n) \left(\frac{n}{N_t+1} - \frac{n}{N_t}\right)^2 \\
&\quad \left. + d(t, N_t) n \left(\frac{n-1}{N_t-1} - \frac{n}{N_t}\right)^2 + d(t, N_t)(N_t - n) \left(\frac{n}{N_t-1} - \frac{n}{N_t}\right)^2 \right] dt \\
&= \left[\frac{1}{N_t} (2\xi(t, N_t) + s(t, N_t)) x_t (1 - x_t) + \frac{1}{N_t} \frac{n(N_t - n)}{(N_t + 1)^2} b(t, N_t) + \frac{1}{N_t} \frac{n(N_t - n)}{(N_t - 1)^2} d(t, N_t) \right] dt \\
&\approx \left[\frac{1}{N_t} (2\xi(t, N_t) + s(t, N_t)) x_t (1 - x_t) + \frac{1}{N_t} x_t (1 - x_t) b(t, N_t) + \frac{1}{N_t} x_t (1 - x_t) d(t, N_t) \right] dt \\
&= \frac{2\xi(t, N_t) + b(t, N_t) + d(t, N_t) + s(t, N_t)}{N_t} x_t (1 - x_t) dt = \frac{x_t (1 - x_t)}{N_{e,t}} dt,
\end{aligned} \tag{S20}$$

where we approximated $N_t + 1 \approx N_t$ and $N_t - 1 \approx N_t$ in the course of the calculation.

Derivation of Eq. (34). In order to proof Eq. (34), we proof the following relation from which Eq. (34) follows immediately by choosing $z = 0$:

$$\frac{d^n P(z, t)}{dz^n} = n! \left[\frac{B^{n-1}(t)}{(A(t) - B(t)(z-1))^n} + \frac{B^n(t)(z-1)}{(A(t) - B(t)(z-1))^{n+1}} \right], \quad n \geq 1 \tag{S21}$$

with the probability generating function (cf. Eq. (10))

$$P(z, t) = 1 + \frac{z-1}{A(t) - B(t)(z-1)}. \tag{S22}$$

We carry out a proof by induction. Eq. (S21) builds our induction hypothesis.

- Base case $n = 1$:

$$\frac{dP(z, t)}{dz} = \frac{1}{A(t) - B(t)(z-1)} + \frac{B(t)(z-1)}{(A(t) - B(t)(z-1))^2}.$$

- Inductive step $n \rightarrow n + 1$:

$$\begin{aligned}
\frac{d^{n+1}P(z, t)}{dz^{n+1}} &= \frac{d}{dz} \frac{d^n P(z, t)}{dz^n} \\
&= \frac{d}{dz} \left[\frac{B^{n-1}(t)}{(A(t) - B(t)(z-1))^n} + \frac{B^n(t)(z-1)}{(A(t) - B(t)(z-1))^{n+1}} \right] \\
&= n! \left[n \frac{B^{n+1}(t)B(t)}{(A(t) - B(t)(z-1))^{n+1}} + \frac{B^n(t)}{(A(t) - B(t)(z-1))^{n+1}} + (n+1) \frac{B^n(t)B(t)(z-1)}{(A(t) - B(t)(z-1))^{n+2}} \right] \\
&= (n+1)! \left[\frac{B^n(t)}{(A(t) - B(t)(z-1))^{n+1}} + \frac{B^{n+1}(t)(z-1)}{(A(t) - B(t)(z-1))^{n+2}} \right].
\end{aligned}$$

It follows:

$$p_n(t) = \frac{1}{n!} \frac{d^n P(z, t)}{dz^n} \Big|_{z=0} = \frac{B^{n-1}(t)}{(A+B)^n(t)} - \frac{B^n(t)}{(A+B)^{n+1}(t)} = \frac{B^{n-1}(t)A(t)}{(A+B)^{n+1}(t)}, \quad n \geq 1. \quad (\text{S23})$$

□

Symmetry of the frequency path in a constant environment. Let Ω_1 be the set of paths which lead from $n_0 = 1$ to $n = n_c - 1$, and Ω_2 be the set of paths which lead from $n = N - n_c$ to $n = N - 1$. Then there exists a bijection between these two sets $f : \Omega_1 \rightarrow \Omega_2$ which can basically be constructed by mirroring the path at $x_c = 0.5$. Concretely: Be $\omega_1 = (n_1, \dots, n_l) \in \Omega_1$ where $n_1 = 1$ and $n_l = n_c - 1$. Define then $f(\omega_1) = \omega_2 = (m_1, \dots, m_l)$ by $m_i = N - n_{1+l-i}$. Consider now the paths $\omega'_1 = (n_1, \dots, n_l, n_c)$ and $\omega'_2 = (m_1, \dots, m_l, N)$. It now holds: (1) The probability that path ω'_1 is realized equals the probability that ω'_2 is realized as both make the same number of steps in both directions. (2) The distribution of the run-time along both paths ω'_1 and ω'_2 is identical because the transition rates of the Moran model defined in Eq. (3) are symmetric. From (1) and (2) it follows that the distribution of the time to reach fixation starting from $(N - n_c + 1)/N \approx (1 - x_c)$ is the same as the distribution of the time to reach frequency $x_c = n_c/N$ starting from $1/N \approx 0$. (This is true whether one considers the first, the last or the mean passage time; the latter is approximated by our theory.)

References

- Ewens, W. J., 2004 *Mathematical Population Genetics*. Springer, New York et al., second edition.
- Kimura, M., 1957 Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics* **28**: 882–901.
- Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**: 763–771.
- Kimura, M., and T. Ohta, 1974 Probability of gene fixation in an expanding finite population. *Proceedings of the National Academy of Sciences USA* **71**: 3377 – 3379.

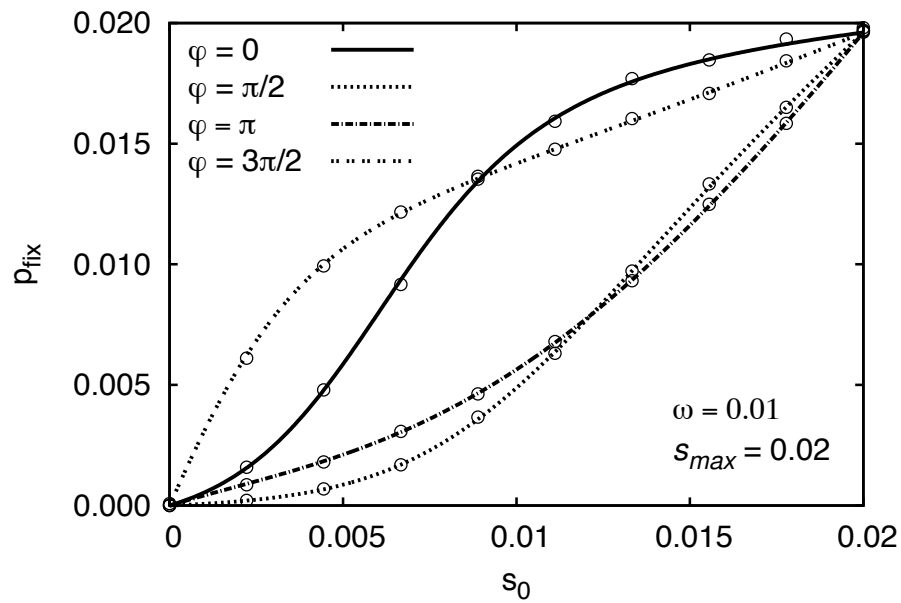


Figure S4: Fixation probability for a mutation with periodically changing selection strength $s(t) = s_0 + (s_{\text{max}} - s_0) \cos(\omega t + \varphi)$ in dependence of s_0 for various values of φ ; $s_{\text{max}} = 0.02$, $\omega = 0.01$. Small values of s_0 correspond to relatively long and pronounced periods of selective disadvantage. As s_0 approaches s_{max} the fixation probability tends to $\approx 2s_{\text{max}}$ as expected. Comparison to simulated data shows that the theory provides an accurate prediction of the fixation probability also for scenarios in which the mutation temporarily gets disadvantageous. For $s_0 \leq 0$, however, it predicts a fixation probability of zero and therefore underestimates the true value. Simulations were performed for a population of 100 000 individuals, and each simulation point is the average over 10^6 runs.

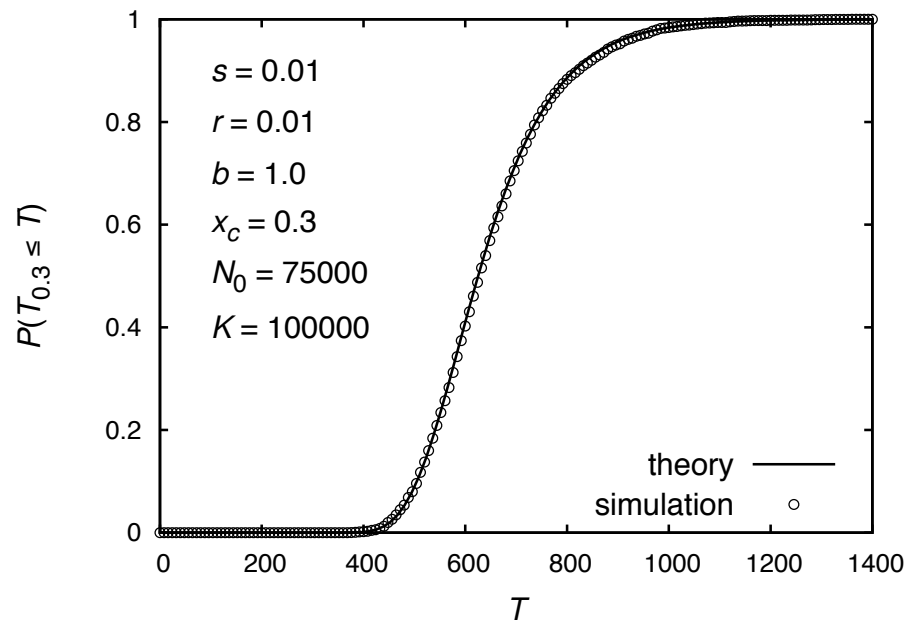


Figure S5: Distribution of the time to reach frequency $x_c = 0.3$ in a population which grows logistically according to the second scenario with $\xi = 0$ (see text). Analytical and simulation results are compared. The simulation curve is the average over 5000 runs.

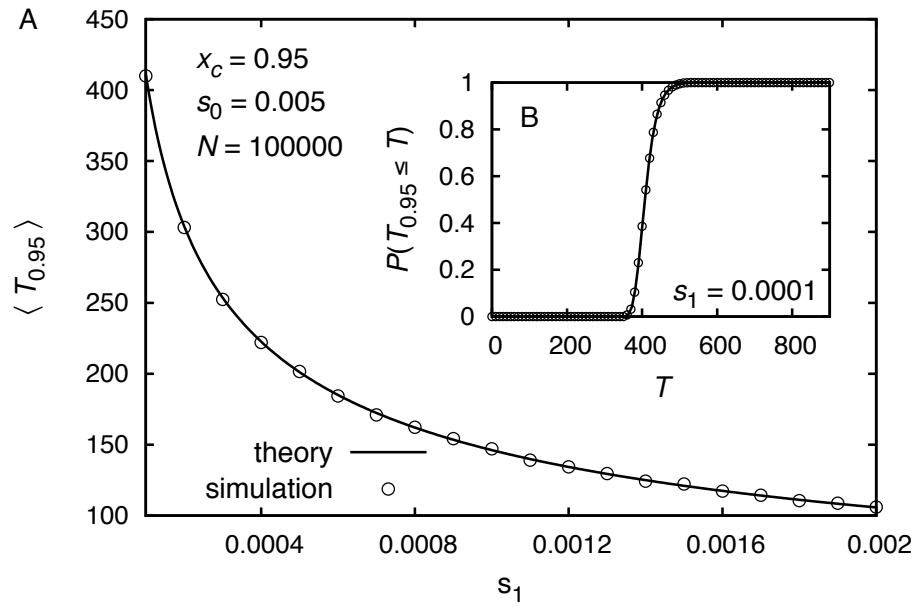


Figure S6: A: Mean time to reach frequency $x_c = 0.95$ for linearly increasing selection $s(t) = s_0 + s_1 t$ in dependence of s_1 . Every simulation point is the average over 100 runs. B: Distribution function for a fixed value of s_1 . For the simulation curve 5000 runs were performed.