Article

Fixation Time for Competing Beneficial Mutations and Their Genomic Footprint

Wolfgang Stephan *

Posted Date: 30 April 2025

doi: 10.20944/preprints202504.2540.v1

Keywords: mathematical population genetics; nonnormalized allele frequencies; recurrent selective sweeps

Article

Fixation Time for Competing Beneficial Mutations and Their Genomic Footprint

Wolfgang Stephan

Faculty of Biology, Ludwig-Maximilian University of Munich; D-82152 Planegg-Martinsried, Germany

* Correspondence: w.stephan@lmu.de or stephan@bio.lmu.de

Abstract: For a highly beneficial mutation A at locus 1 spreading in a very large population, we have analyzed the scenario that at a closely linked locus 2 a second beneficial mutant B arises in repulsion; i.e. on a chromosome carrying the wildtype a at locus 1. Under the assumptions that the fitness of B is greater than that of A and that A- and B-carrying chromosomes can recombine at some rate r, recombinants AB may form and eventually fix. We present explicit formulas for the fixation time of AB under additive fitness of the mutants as a function of the frequency $X_2(0)$ of A at the time when B is introduced. Our analysis suggests that the effect of interference between the beneficial mutations is most pronounced for small values of $X_2(0) < 0.1$. Furthermore, we identify a threshold value for r, above which recombination speeds up fixation. Using published simulation data we also describe the genomic footprint of competing beneficial mutations. At neutral sites between the two linked selected loci an excess of intermediate-frequency variants may occur when interference is strong; i.e. $X_2(0)$ small. Finally, we discuss under which circumstances this scenario may be encountered in real data.

Keywords: mathematical population genetics; nonnormalized allele frequencies; recurrent selective sweeps

1. Introduction

The speed of beneficial mutations on their way to fixation in natural populations is a fundamental topic in population genetics. Knowing how fast selection can act to change allele frequencies is essential for understanding evolution. The time for a beneficial allele to spread through a natural population was investigated early in the history of population genetics, using deterministic models [1]. Later, propertiies of the fixation time of individual advantageous mutations under the influence of positive directional selection and genetic drift in populations of finite size have been derived by several authors [2 − 4].

However, the spread of two or more beneficial mutations that arise and interact during their fixation process is much less investigated. For two interacting mutations Otto and Barton [5] have studied the case that the second mutation is less beneficial than the first one. In contrast, Cuthbertson et al. [6] and Bossert and Pfaffelhuber [7] considered the scenario that the second mutation is fitter than the first one. Furthermore, they assumed that there is a chance that the first and second mutant recombine such that the recombinant type has the highest fitness and eventually fixes. Here we will focus on this latter case.

We consider the mathematical analysis of the fixation time in conjunction with the theory of selective sweeps. Although beneficial mutations are a comparatively small fraction of all new mutations, some of them may reach fixation and are thus important in evolution. If the fitness effects of these beneficial mutations are sufficiently strong, they may cause selective sweeps, i.e. localized reductions of genetic variation along genomes [8]. Such localized patterns of reduced genetic variation have been convincingly described in a variety of organisms.

 doi:10.20944/preprints202504.2540.v1

Detecting signatures of selective sweeps in genomes is a major goal of current population genetics, as it allows estimating the rate of beneficial mutations going to fixation and finding the genes involved in selection. The inference methods for detecting sweeps depend critically on assumptions on whether the benefical mutations occur sequentially (such that there is at most one beneficial allele on a chromosome on the way to fixation at a time) or whether beneficial alleles overlap with each other. Models of recurrent selective sweeps traditionally assume that in chromosomal regions of normal recombination rates at most one beneficial allele is on the way to fixation [9 – 11].

Here we follow the scenario proposed by Bossert and Pfaffelhuber [7]. Thus we assume that, while a highly beneficial mutation spreads in a natural population, a second beneficial mutation arises before the first one has fixed. Furthermore, we envision that the first mutation is less fit than the second one and that recombination may occur between the two chromosomes. Under these conditions a haplotype may be formed that is fitter than the two individual mutations and may therefore eventually fix. To model this process for a population of finite size, Bossert and Pfaffelhuber [7] used stochastic differential equations and calculated the fixation time of a recombinant haplotype under the assumption that it fixes. In contrast, we use a deterministic approach based on ordinary differential equations (ODEs). Thus, in our analysis an explicit assumption about the fixation of a recombinant is not necessary.

We begin by formulating the differential equations for the basic allele frequency changes. Then we introduce nonnormalized variables (that are proportional to the allele frequencies) to find approximate solutions of this system of ODEs. Subsequently, we provide explicit formulas for the fixation time of the recombinant type. Finally, we apply our results to simulation data by Chevin et al. [12] to describe patterns of selective sweeps in the genome caused by the joint fixation of two mutations due to selection and recombination.

2. Model

We consider a two-locus model, with alleles A and a at locus 1 and B and b at locus 2, respectively. The upper-case letters denote beneficial alleles with selection coefficient s_2 at locus 1 and s_3 at locus 2, whereas the alleles with lower-case letters are assumed to be neutral (wildtype). This model has four haplotypes AB, Ab, aB, and ab, with frequencies given by the variables X_1, X_2, X_3 and X_4 (which add up to 1). Assuming additive selection, their relative fitnesses are $1 + s_1$, $1 + s_2$, $1 + s_3$ and 1, respectively, where $s_1 = s_2 + s_3$. Recombination between locus 1 and locus 2 occurs at rate r. Since we are interested in closely linked loci, we assume that $r \ll s_i \ll 1$ for $i = 2, 3$. In our deterministic setting (without genetic drift) the ODEs for the time change of the variables X_i are obtained by adding the change due to selection and the change due to recombination ([13], chapt. 2):

$$\begin{aligned}
\frac{dX_1}{dt} &= X_1\left(s_1 - \sum_{j=1}^{3} s_j X_j\right) - r(X_1X_4 - X_2X_3)\,, \\
\frac{dX_2}{dt} &= X_2\left(s_2 - \sum_{j=1}^{3} s_j X_j\right) + r(X_1X_4 - X_2X_3)\,, \qquad (1) \\
\frac{dX_3}{dt} &= X_3\left(s_3 - \sum_{j=1}^{3} s_j X_j\right) + r\left(X_1X_4 - X_2X_3\right)\,, \\
\frac{dX_4}{dt} &= X_4\left(-\sum_{j=1}^{3} s_j X_j\right) - r(X_1X_4 - X_2X_3)\,,
\end{aligned}$$

 doi:10.20944/preprints202504.2540.v1

where t measures time in generations.

For $r > 0$, which we assume throughout this paper, an exact solution of this system of ODEs is not known. An approximate solution can be obtained using nonnormalized variables Y_i ($i = 1, \ldots, 4$) (see [14] for mutation-selection and [15] for recombination-selection equations). These are related to the original variables of equations (1) as:

$$X_i = \frac{Y_i}{\sum_{j=1}^{4} Y_j}. \qquad (2)$$

Using (2) it can be shown that the nonnormalized variables satisfy the following ODEs:

$$\frac{dY_1}{dt} = s_1 Y_1 - \frac{r}{\sum_{i=1}^{4} Y_i}(Y_1 Y_4 - Y_2 Y_3),$$

$$\frac{dY_2}{dt} = s_2 Y_2 + \frac{r}{\sum_{i=1}^{4} Y_i}(Y_1 Y_4 - Y_2 Y_3), \qquad (3)$$

$$\frac{dY_3}{dt} = s_3 Y_3 + \frac{r}{\sum_{i=1}^{4} Y_i}(Y_1 Y_4 - Y_2 Y_3),$$

$$\frac{dY_4}{dt} = -\frac{r}{\sum_{i=1}^{4} Y_i}(Y_1 Y_4 - Y_2 Y_3).$$

Note that rescaling all Y_i by a constant leaves X_i invariant. To fix this scaling, we use $X_i(0) = Y_i(0)$ for all $i = 1, \ldots, 4$. As outlined in Appendix (A), the following approximate solutions of the ODEs (3) can be found, assuming that both mutations A and B arise on background ab (so that $X_1(0) = 0$):

$$Y_1(t) \approx rX_2(0)X_3(0)e^{s_1 t}\int_0^t \frac{1}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}}\,d\tau,$$

$$Y_2(t) \approx X_2(0)e^{s_2 t}\left[1 - rX_3(0)\int_0^t \frac{e^{s_3\tau}}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}}\,d\tau\right], \qquad (4)$$

$$Y_3(t) \approx X_3(0)e^{s_3 t}\left[1 - rX_2(0)\int_0^t \frac{e^{s_2\tau}}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}}\,d\tau\right],$$

$$Y_4(t) \approx X_4(0) + rX_2(0)X_3(0)\int_0^t \frac{e^{s_1\tau}}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}}\,d\tau.$$

These approximations were first established and tested by Yun Song (personal communication). Numerical analysis suggests that they are generally excellent for $r < min(s_2, s_3)/10$.

3. Fixation time

Using equations (4) we can find the fixation time of two interfering mutations. We call an allele or haplotype fixed when it reaches frequency $1 - \delta$ and denote this time T. Thus T measures the time from some initial frequency at $t = 0$ to $1 - \delta$. The initial frequency may be the frequency of a newly arising mutation. In a haploid population of size N, which we consider here, this initial frequency is given by $1/N$. In our case, however, we are interested in the fixation of the double mutant AB whose initial frequency is $X_1(0) = 0$, as AB arises during the fixation process due to recombination. Thus, the fixation time of AB is found by solving the equation

$$X_1(T) = 1 - \delta, \qquad (5)$$

 doi:10.20944/preprints202504.2540.v1

where δ is a small number.

Next we express equation (5) in terms of nonnormalized variables and obtain

$$X_1(T) = \left(1 + \sum_{i=2}^{4} \frac{Y_i(T)}{Y_1(T)}\right)^{-1} = 1 - \delta\,.$$

Using (4) with $X_2(0) > X_3(0)$ and $s_3 > s_2$, this equation can be approximated by

$$\delta \approx \frac{Y_2(T)}{Y_1(T)} + \frac{Y_3(T)}{Y_1(T)}\,. \qquad (6)$$

Evaluating the terms on the right-hand side of equation (6) requires that we find useful approxiations of the integrals in equations (4) because closed formulae for these integrals are not known (see Appendix (B)). Besides the two aforementioned assumptions, we assume that T is sufficiently large such that the second mutation is eventually dominating the first one; i.e. $X_2(0)e^{s_2 T} \ll X_3(0)e^{s_3 T}$, and population size is large $(N > 10^5)$ such that selection is strong $(Ns_i > 100)$. Under these assumptions we find for the integrals $I_i(T)$ defined in Appendix (B):

$$I_1(T) \approx -\frac{ln\big(X_2(0)\big)}{s_2\big(1-X_2(0)\big)}\,,$$

$$I_2(T) < \frac{T}{X_3(0)}\,, \qquad (7)$$

$$I_3(T) \approx \frac{1}{X_2(0)}\left[\frac{s_3}{s_2(s_3-s_2)}\,ln\big(X_2(0)\big) - \frac{1}{s_3-s_2}\,ln\big(X_3(0)\big)\right]\,.$$

Inserting these formulas into equation (6), this equation can be written in the following form:

$$\delta I_1(T) \approx \frac{1}{X_3(0)e^{s_3 T}}\left(\frac{1}{r} - X_3(0)I_2(T)\right) + \frac{1}{X_2(0)e^{s_2 T}}\left(\frac{1}{r} - X_2(0)I_3(T)\right). \quad (8)$$

We can neglect the first term on the right-hand side of equation (8) for the following reasons: first, because we assumed that $X_2(0)e^{s_2 T} \ll X_3(0)e^{s_3 T}$, and second since $X_3(0)I_2(T)$ is bounded by $\frac{1}{r}$, the term $\frac{X_3(0)I_2(T)}{X_2(0)e^{s_2 T}}$ can be neglected compared to $\frac{1}{X_2(0)e^{s_2 T}}\frac{1}{r}$. This leads to equation (9):

$$-\frac{\delta X_2(0) ln\big(X_2(0)\big)}{s_2\big(1-X_2(0)\big)} \approx e^{-s_2 T}\left(\frac{1}{r} - \frac{s_3 ln\big(X_2(0)\big)}{s_2(s_3-s_2)} + \frac{ln\big(X_3(0)\big)}{s_3-s_2}\right). \qquad (9)$$

Finally, we introduce population size N into this equation by writing $\delta = \frac{1}{N}$, $X_3(0) = \frac{1}{N}$ and $X_2(0) = \frac{x_{20}}{N}$, where x_{20} is the number of A alleles at $t = 0$. Then solving the equation for T yields

$$T \approx \frac{1}{s_2}\left[2ln(N) - ln\left(\frac{x_{20}}{s_2}\right) - ln\left(\frac{ln\left(\frac{N}{x_{20}}\right)}{1-\frac{x_{20}}{N}}\right) + ln\left(\frac{1}{r} + \frac{s_3}{s_2}\frac{ln\left(\frac{N}{x_{20}}\right)}{s_3-s_2} - \frac{ln(N)}{s_3-s_2}\right)\right]. \qquad (10)$$

Table 1 shows that this result agrees very well with the numerical solution of equation (8).

As expected, the formula for T is complex. However, in the interesting parameter range of small r values such that $\frac{1}{r} \gg \left|\frac{s_3}{s_2}\frac{ln\left(\frac{N}{x_{20}}\right)}{s_3-s_2} - \frac{ln(N)}{s_3-s_2}\right|$, we may approximate equation (10) as

 doi:10.20944/preprints202504.2540.v1

$$T \approx \frac{1}{s_2}\left[2ln(N) - ln\left(\frac{r}{s_2}\frac{x_{20}}{1-\frac{x_{20}}{N}}ln\left(\frac{N}{x_{20}}\right)\right)\right]. \qquad (11)$$

The first term on the right-hand side of equation (11) equals the fixation time of a new allele starting at frequency $\frac{1}{N}$ and ending at $1-\frac{1}{N}$, driven by positive directional selection with selection coeffcient s_2. This term also appears in the result of Bossert and Pfaffelhuber [7]. The denominator $s_2$can be explained as follows. To go to fixation, the successful recombinant AB with fitness $1+s_2+s_3$ has to compete against the – at the time – dominant aB type with fitness $1+s_3$, having a fitness advantage s_2.

Furthermore, unless r is very small, the second term in equation (11) is negative such that T is smaller than $\frac{2ln(N)}{s_2}$. This is not surprising as we are dealing here with an equation describing continuous input of new AB alleles due to recombination, similar to the case of fixation under continuous mutation pressure and positive directional selection ([16], equation (8)). Thus, based on equation (11) we obtain a relatively simple formula for the threshold of r

$$r_c = \frac{s_2}{x_{20}}\frac{1-\frac{x_{20}}{N}}{ln\left(\frac{N}{x_{20}}\right)}, \qquad (12)$$

above which recombination speeds up fixation time, whereas for $r < r_c$ the second term in equation (11) turns positive, such that the fixation time T becomes larger than $\frac{2l\ (N)}{s_2}$, meaning that the input of recombinants ceases.

4. Genomic footprint of competing mutations

As an application we analyze simulation data from a study of genetic variation at neutral sites located between two selected loci [12]. The data were obtained using Monte Carlo simulations of a Wright-Fisher model ([13], chapt. 3) with two selected loci and three neutral loci. The three neutral sites are located between the selected loci as described in Table 1. Table 1 also contains the parameter values used in the simulations. They meet the assumptions of our analysis, except for the populaton size. In the simulations $N = 20000$ was used, while in our derivation $N > 10^5$ was suggested. To check whether this causes problems, we compared the analytical results for T from equation (10) with the numerical solutions of equation (8) for $N = 20000$. However, as Table 1 (columns 6 and 7) shows, no discrepancies could be found.

The first observation concerns T as a function of $X_2(0)$. Since $r = 0.005$ was used in all simulations and in all cases r is larger than the threshold r_c (equation 12), we expect that T is smaller than $\frac{2ln(N)}{s_2} = 198.1$ and decreases with increasing $X_2(0) = \frac{x_{20}}{N}$. This is indeed the case. The most pronounced effect of $X_2(0)$ on fixation time is observed for small values of $X_2(0) < 0.1$. For larger values of $X_2(0)$, however, fixation time is relatively constant. This observation is consistent with the formulas for T, especially equation (11), which shows that, for given $\frac{r}{s_2}$, fixation time depends about logarithmally on x_{20}. This formula also says that recombination is most important in speeding up fixation when the second mutation is introduced at low $X_2(0)$ values. Here the interference between the two mutations is largest.

Next we discuss the simulation results of Chevin et al. [12] in the light of our analysis. Variation at the neutral loci (close to the selected ones) shows typical hitchhiking effects [8]; i.e. variation is reduced relative to the neutral standard level such that stronger selection acting at locus 2 ($s_3 > s_2$) leads to a greater reduction than at the neutral site near locus 1. Furthermore, variation at the neutral locus in the middle between the two selected loci is greater than that at the loci near the

 doi:10.20944/preprints202504.2540.v1

selected sites, which is also a typical hitchhiking effect, because the distance to the selected loci (and hence recombination rate) is larger.

Increasing $X_2(0)$ generally leads to stronger hitchhiking effects such that levels of neutral variation decrease with $X_2(0)$. This can be clearly observed at the neutral locus close to locus 1, whereas at the neutral locus close to the stronger selected site this is hardly visible. At the neutral locus in the middle there is also a strong decay of variation with increasing levels of $X_2(0)$. The effect of $X_2(0)$ on hitchhiking is likely due to the interference of the two mutations. The longer they compete with each other on their way to fixation, the weaker their hitchhiking effect. This has already been observed in other studies (e.g. [17]).

Finally, we discuss *D*, a statistic introduced by Tajima [18]. In Table 1 (column 5) only the *D* values for the neutral locus in the middle between locus 1 and 2 are shown. All *D* values at the other two loci are negative as expected from the theory of genetic hitchhiking. A negative *D* is observed when an allele has either a lower or higher frequency than expected by the neutral theory. Interestingly, however, Chevin et al. [12] observed strongly positive *D* values for $X_2(0) = 0.007, 0.024$, and 0.077, whereas *D* is around zero or negative for larger $X_2(0)$. Positive values of *D* indicate that alleles are at intermediate frequencies, such as predicted for balancing selection. In our case, however, this is probably not a valid hypothesis, at least concerning the standard models of balancing selection. A plausible hypothesis proposed by Bossert and Pfaffelhuber [7] is that positive *D* may be observed when a haplotype structure arises in the genome through recombination between different haplotypes consisting of multiple polymorphic loci. Haplotype structures exist in populations only if polymorphisms at individual loci tend to be in intermediate frequency (such that the less frequent variants are not too rare). This may be the case for $X_2(0) = 0.007, 0.024$, and 0.077, but not for the larger $X_2(0)$ values, for which diversity is more heavily reduced (Table 1). An alternative, though related hypothesis postulates that the dynamics of the two selected mutations (while in repulsion) reaches nonnegligible frequencies at similar times such that recombination may produce haplotypes with the two favorable alleles in coupling [12].

If these hypotheses are correct, a genomic footprint of competing beneficial mutations may be detected by measuring Tajima's *D* and/or linkage disequilibrium. In general, footprints associated with selective sweeps caused by the fixation of beneficial mutations can be found in genetic data if their characteristic pattern of variation, such as a dip of nucleotide diversity around a selected site or a haplotype structure revealed by linkage disequilibrium, persist for some time. For Wright-Fisher populations such signatures may be detected for up to $0.1N$ generations after fixation of the driving mutations [19, 20].

5. Discussion

For a highly beneficial mutation *A* at locus 1 spreading in a very large population, we have analyzed the scenario when a second beneficial mutant *B* arises in repulsion; i.e. on a chromosome carrying the wildtype *a* at locus 1. Under the assumptions that the fitness of *B* is greater than that of *A* and that *A*- and *B*-carrying chromosomes can recombine, recombinants *AB* may form and eventually fix. We present approximate formulas for the fixation time of *AB* under additive fitness of the mutations as a function of $X_2(0)$, the frequency of *A* at the introduction of *B*. The latter parameter turns out to be useful for describing the interference between competing beneficial mutations.

Our analysis suggests that the effect of interference between beneficial mutations is most pronounced for small values of $X_2(0) < 0.1$. In this parameter range fixation time decreases substantially with $X_2(0)$. However, for larger values fixation time is relatively constant (Table 1). This agrees with the formulas for *T*, especially equation (11), which shows that *T* depends about

 doi:10.20944/preprints202504.2540.v1

logarithmally on $X_2(0)$. Similarly, the effect of interference on the genomic footprint of competing mutations can be clearly discerned. For small values of $X_2(0) = 0.007, 0.024,$ and 0.077, a strongly positive D was observed, whereas D is around zero or negative for larger $X_2(0)$ (Table 1). Positive values of D indicate that alleles are in intermediate frequencies. This may be observed when a haplotype structure arises in the genome through recombination between different allelic types consisting of multiple polymorphic loci [7, 12].

Finally, we address the question whether we can expect observing patterns of overlapping selective sweeps due to competing mutations in genomes. Estimates of the average selection coefficient s and the rate ν at which beneficial mutations arise and go to fixation (i.e. selective substitutions) are known for some species including *Drosophila melanogaster*. For instance, Jensen et al. [21] analyzed a dataset of genetic variation from the euchromatic part of the genome of a *D. melanogaster* population from Africa, which is – roughly speaking – the recombining portion of chromosomes. They obtained the following estimates: $s = 0.002$, $N = 5 \times 10^6$ and $\nu = 4.2 \times 10^{-11}$ per generation per nucleotide site. Since under selection and genetic drift the mean fixation time (conditional on fixation) for a diploid species such as *D. melanogaster* is $T = \frac{4}{s} ln(2Ns)$ [3], we find that the probability of a second substitution arising on a chromosome during the sojourn of the first one to fixation is $T\nu = 8.3 \times 10^{-7}$ per nucleotide site. Multiplying $T\nu$ with the size of the euchromatic part of a chromosome (in *D. melanogaster* approximately 24 Mb $= 2.4 \times 10^7$ base pairs), we find that on average at about 20 sites of a chromosome strongly selected substitutions could arise and compete with the first mutation during its sojourn to fixation.

An example that some of these selected substitutions occur in close proximity in the genome is found at the *polyhomeotic* locus of a European population of *D. melanogaster*. Voigt et al. [22] report a case in which five selected substitutions (i.e. nearly fixed variants between Europe and Africa) are located in the 5-kb intergenic region between *polyhomeotic proximal* and the gene *CG3835* within a segment of 2.28 kb. They showed that these five selected variants are involved in adaptation of *D. melanogaster* to the higher temperature in Europe compared to that of the ancestral species range in Africa. Variation is generally low in the whole *polyhomeotic* region and Tajima's *D* is strongly negative, as expected after a sweep. However, using a larger dataset than in her previous study, Susanne Voigt (personal communication) found evidence that the five beneficial substitutions that likely caused the sweep did not act independently in a sequential manner but were selected as haplotype block. As a consequence, an elevated level of Tajima's *D* in the fragment containing the five selected substitutions was not detected.

A more promising example in the context of interference between beneficial mutations may be the *Agouti* locus in deer mice. Here the precise mutations required for adaptation to light-colored soil of the Nebraska Sand Hills have been identified [23]. The authors claim that – contrary to the aforementioned *Drosophila* case – the light Sand Hills phenotype is the result of independent selection on many mutations within the *Agouti* locus spanning about 120 kb. Thus, in this case a genomic footprint of interference between beneficial mutations may be encountered in sequence data.

Appendices

(A) Derivation of the approximate solutions of ODEs (3)

We write the solutions in the form

$$Y_1 = rX_2(0)X_3(0)e^{s_1 t} I_1(t)\,,$$

$$Y_2 = X_2(0)e^{s_2 t}\big(1 - rX_3(0)I_2(t)\big)\,, \qquad \text{(A1)}$$

 doi:10.20944/preprints202504.2540.v1

$$Y_3 = X_3(0)e^{s_3 t}\left(1 - rX_2(0)I_3(t)\right),$$

$$Y_4 = X_4(0) + rX_2(0)X_3(0)I_4(t),$$

where $I_i(0) = 0$ for $i = 1, \ldots, 4$.

Next we compare the time derivatives of equations (A1) with those of the corresponding equations (3). Assuming $r \ll s_i$ then leads to

$$\frac{dI_4}{dt} \approx \frac{e^{s_1 t}}{X_4(0)+X_2(0)e^{s_2 t}+X_3(0)e^{s_3 t}} . \qquad \text{(A2)}$$

This immediately yields the last equation in (4) (up to first order in r). In a similar way, we obtain $\frac{dI_i}{dt}$ for $i = 1, 2, 3$ by comparing the time derivatives of equations (A1) with those of the corresponding equations (3).

(B) Approximations of the integrals in equations (4)
We begin by approximating the integral

$$I_1(t) = \int_0^t \frac{1}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}} d\tau . \qquad \text{(B1)}$$

The function to integrate has a maximum at $\tau = 0$ and decays quickly to zero (within about $\hat{t} \approx -\frac{ln(X_2(0))}{s_2}$ generations). Because $X_3(0) = \frac{1}{N}$, we may assume for large populations that within this short decay time $X_3(0)e^{s_3\tau} \ll X_2(0)e^{s_2\tau}$ and neglect the last term in the denominator. For $t > \hat{t}$ the integral can then be approximated by a standard formula. This leads to the approximation for $I_1(T)$ given in equations (7). This approximation is excellent for large population sizes $N > 10^5$.

Next we consider the integral

$$I_2(t) = \int_0^t \frac{e^{s_3\tau}}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}} d\tau . \qquad \text{(B2)}$$

Here the function to integrate increases monotonically up to a level of $\frac{1}{X_3(0)}$ for large t. This leads to the upper bound of this integral given in equations (7). A more precise approximation is not required in this case (see main text below equation (8)).

Finally, we approximate the integral

 doi:10.20944/preprints202504.2540.v1

$$I_3(t) = \int_0^t \frac{e^{s_2\tau}}{X_4(0)+X_2(0)e^{s_2\tau}+X_3(0)e^{s_3\tau}}d\tau \ . \qquad \text{(B3)}$$

The function to integrate has a maximum at $\hat{t} = \frac{1}{s_3} ln\left(\frac{s_2}{s_3-s_2}\frac{X_4(0)}{X_3(0)}\right) \approx -\frac{1}{s_3} ln\big(X_3(0)\big)$. For $t \le \hat{t}$ we integrate the function $\frac{e^{s_2t}}{X_4(0)+X_2(0)e^{s_2t}}$ and for larger times $\frac{1}{X_2(0)+X_3(0)e^{(s_3-s_2)t}}$. This leads to the third formula in equations (7), which was obtained assuming large populations.

In general, the analytical results in this paper were derived for large populations ($N > 10^5$). In the applying these results to the simulation data of Chevin et al. [12], we had to check numerically whether they are still valid for a smaller population size of 20000 used in these simulations. We found that the fixation times analytically calculated agree very well with the numerical results (see Table 1).

Table 1. Simulation data of Chevin et al. [12] and fixation times.

$X_2(0)$	π_l	π_m	π_r	D	T_{ana}	T_{num}
0.007	0.333	0.491	0.191	0.503	162.7	163.4
0.024	0.311	0.473	0.201	0.535	157.0	152.1
0.077	0.276	0.420	0.198	0.468	141.8	141.8
0.224	0.244	0.323	0.192	0.019	133.1	133.2
0.5	0.184	0.184	0.168	−0.621	127.1	127.3

The first five columns show the simulation data: $X_2(0)$, the frequency of the first mutation when the second mutation is introduced; π_l, relative genetic diversity at a neutral locus between loci 1 and 2 near locus 1 (here 'relative' refers to expected diversity under neutrality); π_m, relative genetic diversity at neutral locus in the middle between loci 1 and 2; π_r, relative genetic diversity at a neutral locus between loci 1 and 2 near locus 2; D, Tajima's [18] measure of the deviation of the level of variation from neutrality at the locus in the middle. T_{ana} is the fixation time from equation (10)

 doi:10.20944/preprints202504.2540.v1

measured in generations, and T_{num} is obtained by solving equation (8) numerically. The parameter values are: $N = 20000, X_3(0) = \frac{1}{N}, \delta = \frac{1}{N}, s_2 = 0.1, s_3 = 0.2, r = 0.005$.

Funding: This research was supported by retirement benefits.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed are contained in Table 1.

Conflicts of Interest: The author declares no conflict of interest.

Acknowledgements: I am very grateful to Peter Pfaffelhuber (U Freiburg) whose paper [7] helped me to conceptualize this study, Yun Song (UC Berkeley) for supporting this project during a sabbatical in his lab and providing equations (3) and (4), and Luis-Miguel Chevin (CNRS Montpellier) for supplying his simulation data. I also thank Ellen Baake (U Bielefeld) for introducing me to references on nonnormalized allele frequencies and Susanne Voigt (TU Dresden) for her recent analyses of *polyhomeotic* sequences.

References

1. Fisher, R.A. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **1922**, *42*, 321-341.
2. Kimura, M.; Ohta, T. The average number of generations until the fixation of a mutant gene in a finite population. *Genetics* **1969**, *61*, 763-771.
3. van Herwaarden, O.A.; van der Wal, N.J. Extinction time and age of an allele in a large finite population. *Theoretical Population Biology* **2002**, *61*, 311-318.
4. Hermisson, J.; Pennings, P.S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **2005**, *169*, 2335-2355.
5. Otto, S.P.; Barton, N.H. The evolution of recombination: removing the limits to natural selection. *Genetics* **1997**, *147*, 879-906.
6. Cuthbertson, C.A.; Etheridge, A.; Yu, F. Fixation probability for competing selective sweeps. *Electronic Journal of Probability* **2012**, *17*, 1-36.
7. Bossert, S.; Pfaffelhuber, P. The fixation probability and time for a doubly beneficial mutant. *Stochastic Processes and their Applications* **2018**, *128*, 4018-4050.
8. Maynard Smith, J.; Haigh, J. The hitchhiking effect of a favourable gene. *Genetical Research* **1974**, *23*, 23-35.
9. Kaplan, N.L.; Hudson, R.R.; Langley, C.H. The 'hitchhiking effect' revisited. *Genetics* **1989**, *123*, 887-899.
10. Wiehe, T.H.E.; Stephan, W. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution* **1993**, *10*, 842-854.
11. Stephan, W. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Molecular Biology and Evolution* **1995**, *12*, 959-962.
12. Chevin, L.-M.; Billiard, S.; Hospital, F. Hitchhiking both ways: Effect of two interfering selective sweeps on linked neutral variation. *Genetics* **2008**, *180*, 301-316.
13. Ewens, W.J. (2004) *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd ed.; Springer: Berlin, Germany, 2004.
14. Thompson, C.J.; McBride, J.L. On Eigen's theory of self-organization of matter and the evolution of biological macromolecules. *Mathematical Biosciences* **1974**, *21*, 127-142.
15. Baake, M.; Baake, E. An exactly solved model for mutation, recombination and selection. *Canadian Journal of Mathematics* **2003**, *55*, 3-41.
16. Stephan, W. The classical hitchhiking model with continuous mutation pressure and purifying selection. *Ecology and Evolution* **2021**, *11*, 15896-15904.
17. Kim, Y.; Stephan, W. Selective sweeps in the presence of interference among partially linked loci. *Genetics* **2003**, *164*, 389-398.
18. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **1989**, *123*, 585-595.
19. Kim, Y.; Stephan, W. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **2000**, *155*, 1415-1427.
20. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **2002**, *160*, 1179-1189.

 doi:10.20944/preprints202504.2540.v1

21. Jensen, J.D.; Thornton, K.R.; Andolfatto, P. An approximate Bayesian estimator suggests strong recurrent sweeps in Drosophila. *PLoS Genetics* **2008**, *4*, e1000198.
22. Voigt, S.; Laurent, S.; Litovchenko, M.; Stephan, W. Positive selection at the *polyhomeotic* locus led to decreased thermosensitivity of gene expression in temperate *Drosophila melanogaster*. *Genetics* **2015**, *200*, 591-599.
23. Linnen, C.R.; Poh, Y.-P.; Peterson, B.K.; Barrett, R.D.H.; Larson, J.G.; Jensen, J.D.; Hoekstra, H.E. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* **2013**, *339*, 1312-1316.