Selective Sweeps in Multilocus Models of Quantitative Traits

Pavlos Pavlidis,*,†,[1] Dirk Metzler,* and Wolfgang Stephan*
*Department of Biology II, Ludwig-Maximilians-University Munich, 82152 Planegg, Germany and †Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany

ABSTRACT We study the trajectory of an allele that affects a polygenic trait selected toward a phenotypic optimum. Furthermore, conditioning on this trajectory we analyze the effect of the selected mutation on linked neutral variation. We examine the well-characterized two-locus two-allele model but we also provide results for diallelic models with up to eight loci. First, when the optimum phenotype is that of the double heterozygote in a two-locus model, and there is no dominance or epistasis of effects on the trait, the trajectories of selected mutations rarely reach fixation; instead, a polymorphic equilibrium at both loci is approached. Whether a polymorphic equilibrium is reached (rather than fixation at both loci) depends on the intensity of selection and the relative distances to the optimum of the homozygotes at each locus. Furthermore, if both loci have similar effects on the trait, fixation of an allele at a given locus is less likely when it starts at low frequency and the other locus is polymorphic (with alleles at intermediate frequencies). Weaker selection increases the probability of fixation of the studied allele, as the polymorphic equilibrium is less stable in this case. When we do not require the double heterozygote to be at the optimum we find that the polymorphic equilibrium is more difficult to reach, and fixation becomes more likely. Second, increasing the number of loci decreases the probability of fixation, because adaptation to the optimum is possible by various combinations of alleles. Summaries of the genealogy (height, total length, and imbalance) and of sequence polymorphism (number of polymorphisms, frequency spectrum, and haplotype structure) next to a selected locus depend on the frequency that the selected mutation approaches at equilibrium. We conclude that multilocus response to selection may in some cases prevent selective sweeps from being completed, as described in previous studies, but that conditions causing this to happen strongly depend on the genetic architecture of the trait, and that fixation of selected mutations is likely in many instances.

To improve our understanding of the genetics of adaptation, recent approaches of molecular population genetics and genomics have attempted to detect signatures of positive selection in the genome (selective sweeps) (Stephan 2010). Typically, these studies reveal many genes or gene regions that may have been under positive selection. However, the relationship between the genes under selection and associated traits remains usually unknown. Here we follow the opposite direction by starting with phenotypes and working toward the genotypes. A phenotype may be determined by a multitude of genes as well as the environment. Multilocus population genetics has been developed in the last decades to describe the evolution of multilocus systems and phenotypes (Bürger 2000). Different types of selection, such as directional, stabilizing or disruptive selection, modify the genetic constitution of the population and favor either extreme or intermediate genotypic values of the trait. In this study we focus on stabilizing selection, which drives a trait toward a phenotypic optimum. We investigate the trajectory of an allele affecting the phenotypic trait (from low frequency up to an equilibrium value). We are particularly interested in exploring the parameter range of trajectories that fix and therefore might generate selective sweeps.

Historically, there has been a great interest in the maintenance of genetic variability under stabilizing selection, because stabilizing selection is assumed to operate on traits in various organisms, for example, the coat color in mice (Vignieri *et al.* 2010), human facial features (Perrett *et al.* 1994), plant defense mechanisms (Mauricio and Rausher 1997), enhancer elements in *Drosophila* (Ludwig *et al.* 2000), and vocalization in frogs and toads (Gerhardt 1994);

doi: 10.1534/genetics.112.142547
Manuscript received February 14, 2012; accepted for publication June 10, 2012
Supporting information is available online at http://www.genetics.org/content/suppl/2012/06/19/genetics.112.142547.DC1.
[1]Corresponding author: HITS gGmbH, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany. E-mail: pavlidisp@gmail.com

see also Endler (1986, Chap. V) for examples and discussion. Furthermore, it has been suggested that this type of selection exhausts genetic variation (Fisher 1930; Robertson 1956).

By contrast, many quantitative traits exhibit high levels of genetic variability. This contradiction motivated researchers to study the role of mutation (Lande 1975; Turelli 1984; Gavrilets and Hastings 1994; Bürger 1998), overdominance (Bulmer 1973; Gillespie 1984), migration (Tufto 2000), frequency-dependent selection through intraspecific competition for some resource (Bürger 2002; Bürger and Gimelfarb 2004), genotype–environment interaction (Gillespie and Turelli 1989), pleiotropy (Hill and Keightley 1988; Barton 1990; Zhang and Hill 2002), and epistasis (Zhivotovsky and Gavrilets 1992). Additionally, a lot of work has been devoted to exploring the ability of stabilizing selection in maintaining genetic variability of quantitative traits that are controlled by multiple loci in the absence of mutation. Theoretical focus was mainly on two-locus models, but also models of more than two loci have been analyzed.

Surprisingly, predictions about genetic variability depend profoundly on the number of loci. The two-locus model predicts that genetic variability may remain in the population due to stabilizing selection *per se*. On the other hand, in models with more than two loci the amount of genetic variability maintained by stabilizing selection is smaller. The reason is that the optimum can be approached very closely by various homozygous genotypes (Bürger 2000, Chap. VI) when there are more than two loci that control the trait. For the two-locus model, and assuming a symmetric viability model, such that the double heterozygous genotype is optimal and the fitness values of the remaining eight genotypes are symmetric about the optimum (*e.g.*, Bodmer and Felsenstein 1967; Karlin and Feldman 1970), it has been shown that there are nine equilibria (Bürger 2000), seven of which can be stable but not simultaneously. Those seven equilibria split into four classes (Bürger and Gimelfarb 1999): they can be polymorphic for both loci or one of them or totally monomorphic. Because analytical solutions of the two-locus model are available, analysis of this model plays an important role in our study.

To our knowledge, the first effort that bridges quantitative trait evolution and selective sweeps was made by Chevin and Hospital (2008). Their work was based on a seminal article by Lande (1983). Lande's model focuses on one locus of major effect on the trait and treats the remaining loci of minor effects as genetic background for this locus. It is assumed that heritable background variation is maintained at a constant amount by polygenic mutation and recombination (Lande 1975, 1983); also, the various loci that affect the trait are unlinked and there are no epistatic interactions. Chevin and Hospital (2008) used Lande's model to infer the deterministic trajectory of a beneficial mutation that affects a quantitative trait in the presence of background genetic variability. They studied both directional and stabilizing selection and showed that fixation needs longer time than in the classical one-locus model (*i.e.*, when genetic variability in the background is absent). In the case of stabilizing selection their approach (based on Lande's model) suggests that the occurrence of selective sweeps at quantitative trait loci (QTL) is expected to be very rare. In contrast to Chevin and Hospital (2008) the present study assumes an explicit number of loci that determine the trait, as this was done by Bodmer and Felsenstein (1967), Karlin and Feldman (1970), and Bürger (2000, Chap. VI). Therefore, the assumption of constant variability in the genetic background is relaxed since the genetic background is modeled explicitly.

We analyze the evolution of the deterministic multilocus model and also its stochastic analog assuming a finite constant effective population size. The focus is on the properties of the trajectory of a new mutation at a certain locus (called focal locus thereafter) that affects the trait under selection. We also examine the parameters (such as the recombination rate and the contribution of the alleles to the phenotype) that affect the fixation probability of the new mutation. Finally, conditioning on the trajectory we generate coalescent simulations and examine the properties of the genealogy and the associated polymorphism patterns. Results are presented for the classical two-locus two-allele model, but we also extend the analysis up to an eight-locus two-allele model.

Methods

The general model

We consider a diploid population of size N and a quantitative trait under selection. The quantitative trait is controlled by l diallelic loci with no epistatic interactions on the phenotype. There is no dominance of allelic effects on the trait (but there may be for their effects on fitness). The alleles at the locus i are labeled as L_i^1 and L_i^2. Allele L_i^1 contributes z_i^1 to the trait, and the contribution of L_i^2 is z_i^2 for each i. The optimum for the trait is set to 0 without loss of generality. The recombination fraction between loci i and $i+1$ is $r_i < 0.5$. At time $t = t_0$, the frequency of L_i^1 is $p_0(L_i^1)$; the loci are in linkage equilibrium ($D = 0$). The trait is assumed to be under a Gaussian fitness function; *i.e.*, if the phenotypic value of an individual is P, then its fitness is given by $W(P) = \exp(-P^2/\omega^2)$, where ω^2 determines how fast the fitness decreases away from the optimum (the smaller ω^2, the faster the fitness is decreasing). Note that due to the fitness function the effects of the alleles *on fitness* are not additive. Here, the phenotypic value P is determined solely by the genotype. However, it is straightforward to include environmental noise assuming that the environmental component is normally distributed with mean 0 and variance σ_E^2 by replacing ω^2 in the Gaussian fitness function by $\omega^2 + \sigma_E^2$ (Lande 1976). Sexes are equivalent and mating is random.

The population evolves forward in time from $t = t_0$ to the present $t = 0$, and generations do not overlap. The mutation rate is 0 (see Supporting Information, File S1, Implementation, for extensions of the model). In each generation, the

life cycle consists of (i) the zygote phase, (ii) viability selection, where individuals are selected as parents for the next generation according to their fitness value, (iii) recombination for each of the parents where gametes are formed, and (iv) random mating to form the zygotes of the next generation. In step iv, N matings take place among N individuals. Each mating produces one diploid offspring and each individual can participate in multiple matings as a male or female. In each generation, at the zygote phase, the frequencies of the alleles of the locus of interest are recorded and the trajectories are stored. Note that in this model selection and drift act simultaneously in step ii, where a finite number of individuals is chosen as parents for the next generation. Also, random genetic drift acts in steps iii and iv: from a pair of gametes only one recombinant is chosen to pass to the next generation.

Recursion equations of the multilocus model

As mentioned above, the two-locus two-allele model has been widely used and will serve here as a reference point. We are particularly interested in the work of Willensdorfer and Bürger (2003) who explore the equilibrium properties of the two-locus two-allele model for Gaussian selection under the assumption of a symmetric fitness function for which the double heterozygous genotype is optimal and the fitness values of the remaining eight genotypes are symmetric about the optimum (see also Table S1). The analysis of Willensdorfer and Bürger (2003) provides the existence and stability criteria for the equilibrium points of the model. The fitnesses of the nine possible genotypes are shown in Table S1B. Let x_1, x_2, x_3, and x_4 represent the frequencies of the gametes $L_1^1L_2^1$, $L_1^1L_2^2$, $L_1^2L_2^1$, and $L_1^2L_2^2$, respectively. A classical result (*e.g.*, Karlin and Feldman 1970; Willensdorfer and Bürger 2003) gives the recursion relations for the frequencies in the next generation as

$$\bar{W}x'_i = x_iW_i - \eta_i rD \quad \text{for } i = 1,2,3,4, \tag{1}$$

where $\eta_1 = \eta_4 = 1$ and $\eta_2 = \eta_3 = -1$. $W_i, i = 1,\ 2,\ 3,\ 4$, is the marginal fitness of gamete i and $D = x_1x_4 - x_2x_3$ measures the linkage disequilibrium. The average fitness is $\bar{W} = \sum_{i=1}^{4} x_iW_i$. No explicit solution for the system in Equation 1 is known (Reinhard Bürger, personal communication). Note that the model in Equation 1 is deterministic; *i.e.*, random genetic drift is neglected.

Willensdorfer and Bürger (2003) parametrize the model so that the effect of alleles L_1^1, L_1^2, L_2^1, and L_2^2 are $-\gamma_1/2$, $\gamma_1/2$, $-\gamma_2/2$, and $\gamma_2/2$, $\gamma_1 \geq \gamma_2 \geq 0$; then, the double heterozygote has the optimal phenotype. Let α_i denote the fitness for the γ_i phenotype, *i.e.*, $\alpha_i = \exp(-\gamma_i/\omega^2)$ under the Gaussian selection function. ω^2 quantifies the strength of selection ($1/\omega^2$ corresponds to s in Willensdorfer and Bürger 2003). They show that α_1 and α_2 and the recombination fraction r determine the equilibrium properties of the model (see also Willensdorfer and Bürger 2003, Equations 3.1, 3.2a, 3.2b, 3.8). We further demonstrate that the initial frequencies of the alleles determine to a large extent whether a new mutation will be fixed. For the symmetric fitness models in this study we use the model of Willensdorfer and Bürger (2003) as it was described above. Note, however, that by assuming that the effects of the alleles of the same locus are opposite (as has been assumed here) we study only a subcategory of symmetric fitness models. In general, when the optimum value is 0, any model with $z_1^1 + z_1^2 + z_2^1 + z_2^2 = 0$ will result in a symmetric fitness model.

While Equation 1 describes the evolution of two-locus two-allele models, for more loci the following classical equation (see Gimelfarb 1998 and references therein) describes the evolution of gametic frequencies,

$$p'(g_i) = W^{-1}\sum_j\sum_k p(g_j)p(g_k)w(g_j,g_k)H(g_i|g_j,g_k), \tag{2}$$

where $W = \sum_j\sum_k p(g_j)p(g_k)w(g_j,g_k)$, g_i, g_j, g_k are gametes, p, p' describe gametic frequencies at two successive generations, $w(g_j,g_k)$ denotes the fitness of the zygote formed by g_j and g_k, and $H(g_i|g_j,g_k)$ denotes the probability to obtain a g_i gamete from g_j and g_k parents (see File S1 for implementation details).

Lowess procedure

To capture the effect of the parameters on the equilibria in the system, we used the Lowess or Loess (locally weighted scatterplot smoothing) function. Lowess is a method that fits lines to scatterplots (Cleveland 1979). Lowess fits low-degree (usually 1 or 2) polynomial curves to localized subsets of the data. Thus, a global function (and therefore a model) is not required, and great flexibility can be gained. Here, we fit Lowess functions to data that assume two values (binary data): class 0 or class 1.

Coalescent simulations and SNP summary statistics

Assume a sample of k individuals from a present-day population ($t = 0$). Given the trajectory of the L_i^j allele, we implement coalescent simulations from $t = 0$ to the most recent common ancestor (MRCA) of the neutral genomic region around the locus L_i. The backward-in-time simulations are based on the structured coalescent model (Kaplan *et al.* 1989; Wakeley 2008; Teshima and Innan 2009; Ewing and Hermisson 2010). The population is subdivided into two genetic backgrounds: one class of lineages is linked to the L_i^1 allele and the other is linked to the L_i^2. Given the trajectory of the L_i^1 and L_i^2 allele, the genealogical history of linked neutral regions is considered separately for the two classes; recombination allows lineages to move between the two classes (as migration allows lineages to move in a structured population). We assume that the genealogies of the genomic region around the locus L_i are affected only by the locus L_i and not by the remaining loci. This simplification makes the backward simulations tractable and allowed us to use available simulation software (*e.g.*, Teshima and Innan 2009; Ewing and Hermisson 2010). However, as is mentioned in the following

sections, this is appropriate only when selection is relatively weak and the loci are weakly linked.

Four statistics of coalescent trees have been used. First, h (the height of the coalescent tree) measures the scaled time from the present to the MRCA of the sample; second, l is the total length of the coalescent; third, two quantities, b_L and b_N measure the balance of the coalescent when the root is placed at the node of the MRCA. b_L is based on the length of the subtrees on either side of the MRCA, whereas b_N uses the number of nodes (Equation 3):

$$b_L = 4\frac{l_L l_R}{l^2}, \quad b_N = 4\frac{n_L n_R}{n^2}. \tag{3}$$

l_L and l_R denote the total length of the left and right subtree of the MRCA, respectively. n_L and n_R are the numbers of nodes on the left and on the right side of the MRCA, respectively, and n is the total number of nodes (excluding the root), *i.e.*, $n = 2k-2$, where k is the sample size. b_L and b_N assume values in (0, 1]; when they equal 1 the coalescent trees are perfectly balanced ($l_L = l_R = l/2$ and similarly for the number of nodes), whereas smaller values denote some degree of imbalance. The summaries of the genealogies are related to the perturbations of the coalescent due to the action of selection. It is well known that in the neighborhood of a beneficial mutation directional selection reduces the height and the length of the coalescent and increases its imbalance (Li 2011). Several neutrality tests (*e.g.*, Fay and Wu 2000) are based on this perturbation of coalescent trees around the beneficial mutation.

Furthermore, we used SNP summary statistics to describe the polymorphism patterns in a present-day sample, as we move along the sequence alignment away from the L_1 locus. We measure the number of polymorphic sites and Tajima's (1989) D, which summarizes the site frequency spectrum. These summary statistics facilitate the comparison between polymorphism patterns that are created by the multilocus model and the one-locus selective sweep. Similarly to the summaries of the genealogies, they can describe perturbations of the polymorphism patterns that are created by the action of recent selection. It is well known that the level of polymorphism and the number of haplotypes are reduced around the target of selection, the site frequency spectrum is shifted toward low- and high-frequency derived variants, which cause negative values of Tajima's D, and the linkage disequilibrium increases on each side of the beneficial mutation (Kim and Stephan 2002; Kim and Nielsen 2004; Stephan *et al.* 2006).

Results

Trajectories of the L_1^1 allele at the focal locus L_1

Two-locus two-allele model with symmetric fitness function: The first goal of this analysis is to illustrate the effect of the parameters of Willensdorfer and Bürger's (2003) model on the fixation of the focal L_1^1 allele. Second, we introduce random genetic drift by simulating the evolution of a randomly mating population with effective population size N = 10,000. Then, we relax the assumption of the symmetric fitness matrix and analyze a more general fitness scheme.

Table 1 The parameter values that were used for the simulations of the two-locus two-allele model

Parameter	Value min.	Value max.
r	0	0.5
$p_0(L_1^1)$	0	0.2
$p_0(L_2^1)$	0	1
ω^2	1	10
z_1^1	−2	2
z_2^1	−2	2

r, recombination fraction; $p_0(L_1^1)$, initial frequency of the allele L_1^1; $p_0(L_2^1)$, initial frequency of the allele L_2^1; ω^2, strength of selection; z_1^1, contribution of L_1^1; and z_2^1, contribution of L_2^1.

Deterministic model: For the deterministic two-locus two-allele model with symmetric fitness matrix, we numerically solve the system of recursions described in Equation 1 and record the frequency of the L_1^1 allele for 10,000 generations. The fitness matrix is symmetric with respect to the double heterozygous genotype (Table S1). Recombination fractions between loci, initial allelic frequencies, ω^2, z_1^1, and z_2^1 are drawn from uniform distributions whose boundaries are defined in Table 1. The initial frequencies for the gametes $L_1^i L_2^j$, $i, j = 1, 2$, are given as the product $p_0(L_1^i)p_0(L_2^j)$, and therefore the initial value of D is 0.

Fixation of the allele is possible and this fixation may occur fast. These trajectories are similar to the trajectories obtained from the classical selective sweep theory. There is, however, a subset of trajectories that remain polymorphic for the focal locus. Furthermore, there is a class of trajectories that shows nonmonotonic behavior. The frequency initially increases and then may decrease to some equilibrium value.

To construct a trajectory we draw uniformly a value from the six-dimensional space described in Table 1. Since we draw random values for the parameters from the six-dimensional space, trajectories may become extinct, reach a polymorphic equilibrium point, or fix, depending on the parameter values. After 10,000 generations we record the final frequency of the trajectory. The proportion of parameter values that lead to a polymorphic equilibrium with frequency in the intervals (0, 0.5) and (0.5, 1) are 0.113 and 0.029, respectively. Similarly, the proportion of equilibrium points 0, 0.5, and 1 are 0.401, 0.415, and 0.041, respectively. Thus, the vast majority of trajectories lead to extinction or the polymorphic equilibrium value 0.5. Although it is not clear whether the frequencies reached after 10,000 generations represent equilibrium values, the fact that >40% of the trajectories remain polymorphic at frequency 0.5 is not surprising since the double heterozygote genotype is optimal for the symmetric fitness model. Thus, for the given parameter values the majority of trajectories either approach 0 or remain polymorphic at frequency 0.5 (Figure S1A).

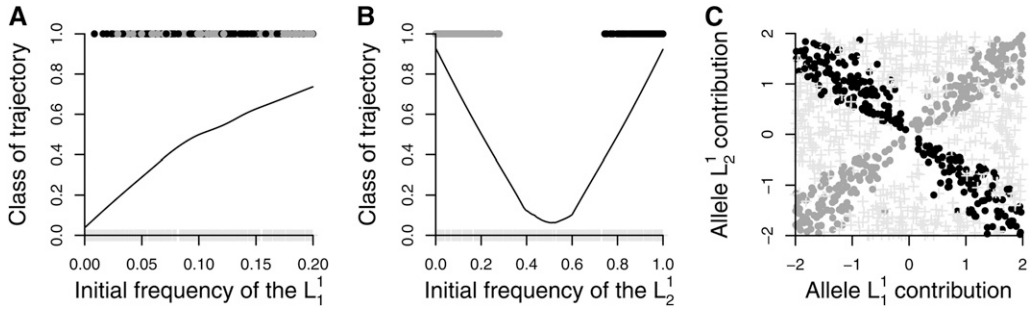

Figure 1 The impact of initial frequencies and allelic effects on determining the class of the trajectory for the comparison of the fixation class *vs.* the polymorphic class. Dark gray and black points depict trajectories that reach fixation, whereas light-colored points show trajectories that stay polymorphic. In dark gray $p_0(L_2^1) < 0.5$, whereas in black $p_0(L_2^1) > 0.5$. In A and B, light-colored points are on the line $y = 0$, while dark gray and black points are on the line $y = 1$. In C, light gray points are represented as pluses (+). The focal allele is L_1^1. The curves in A and B represent the Lowess smoothing function for the data. There are two classes of trajectories: class 0 that denotes polymorphic equilibria and class 1 that represents fixation. As shown in A, for very small values of $p_0(L_1^1)$the probability of a trajectory from the fixation class is small. In B, the initial frequency of L_2^1, $p_0(L_2^1)$, shows nonmonotonic behavior: small and large values of $p_0(L_2^1)$ make the fixation of L_1^1 possible. In C we can see how the contributions of the alleles interact. When $z_1^1 \approx z_2^1$ or $z_1^1 \approx -z_2^1$, then it is possible to obtain trajectories that reach fixation.

To identify the factors that determine the fixation of the L_1^1 allele, we compare different sets of trajectories pairwise. For example, comparison of the trajectories that reach fixation with the trajectories that remain at frequency 0.5 gives insight into the parameter values that affect these two sets. Thus, in the next sections the following two comparisons are made: (i) fixed trajectories (fixation class) *vs.* trajectories that stay at equilibrium frequency 0.5 (polymorphic class), and (ii) fixed trajectories *vs.* trajectories where the allele L_1^1 gets lost (extinction class). Throughout the text the fixation class is defined as the set of trajectories whose equilibrium frequency is in the range (0.999, 1] and the extinction class as the set of trajectories whose equilibrium frequency is in the range [0, 0.001). To simplify the analysis we use the set of trajectories whose equilibrium frequency is in the range (0.499, 0.501) to define the polymorphic class, unless mentioned differently.

Willensdorfer and Bürger (2003) proved that two conditions are required for the fixation or extinction of L_1^1: $r \geq 1-\alpha_1\alpha_2\exp\sqrt{\ln\alpha_1 \ln\alpha_2}$ and $\gamma_1 \leq 2\gamma_2$ ($\alpha_1, \alpha_2, \gamma_1, \gamma_2$ are defined in *Methods*). These two conditions are derived from the stability analysis of Equation 1 at the monomorphic equilibria. An equivalent way to write the first condition is $r \geq 1-\exp(-(\gamma_1-\gamma_2)^2/\omega^2)$. When the effects of the alleles at the two loci are similar, then $1-\exp(-(\gamma_1-\gamma_2)^2/\omega^2) \approx (\gamma_1-\gamma_2)^2/\omega^2$, and the first relation holds for all but very small recombination fractions. The first condition is satisfied when selection on the double homozygotes $L_1^1L_2^2/L_1^1L_2^2$ and $L_1^2L_2^1/L_1^2L_2^1$ (their phenotypic values are $-\gamma_1+\gamma_2$ and $\gamma_1-\gamma_2$, respectively) is weak relative to recombination. An equivalent interpretation is that the first condition is satisfied when the squared difference of their effects (normalized by the selection intensity ω^2) is small relative to recombination. The second condition depends only on the phenotypes and is equivalent to $-\gamma_1+\gamma_2 \geq -\gamma_2$ or $\gamma_1-\gamma_2 \leq \gamma_2$, which means that the distance of the double homozygotes $L_1^1L_2^2/L_1^1L_2^2$ and $L_1^2L_2^1/L_1^2L_2^1$ from the optimum is smaller than the distance of the single homozygotes $L_1^1L_2^1/L_1^2L_2^1$ and $L_1^1L_2^2/L_1^2L_2^2$. Comparison of parameter values that result in fixation of the L_1^1 allele with the parameter values that result in a polymorphic equilibrium for the L_1^1 allele shows that the parameters $c_1 = r-1+\alpha_1\alpha_2\exp\sqrt{\ln\alpha_1 \ln\alpha_2}$ and $c_2 = 2\gamma_2-\gamma_1$ can separate the trajectories that fix from those that stay polymorphic. In case both c_1 and c_2 are positive (as required for the stability of the monomorphic equilibrium), 98.4% of the trajectories reach fixation (for our set of parameter values).

In addition to c_1 and c_2, there are other parameters that provide information about the equilibrium state of the trajectory, such as initial frequencies (Figure 1, A and B). An initial frequency of the L_2^1 allele close to the boundaries 0 or 1 yields fixation of the trajectory for the majority of the simulations, whereas intermediate initial frequencies lead to polymorphic equilibrium states (Figure 1B).

This finding may be explained as follows. Let the initial frequency of the L_2^1 allele be close to 0 for the trajectories that result in fixation of allele L_1^1 (*i.e.*, black points in Figure 1). Then, as illustrated in Figure 1C, these points are located in the proximity of the line $x = y$; *i.e.*, the contributions of the L_1^1 and L_2^1 alleles to the phenotype are approximately equal. Equivalently, the proportion of fixed trajectories is high when $z_1^1 \approx z_2^1$, given that the initial frequency of L_2^1 is low. Assuming that the initial frequency of L_2^1 is low, then the majority of the genotypes for the L_2 locus will be L_2^2/L_2^2, and a smaller proportion will be L_2^1/L_2^2. If the contribution of the L_2^1 allele is z_2^1, then the contribution of the L_2^2 allele is $z_2^2 = -z_2^1$, due to model assumptions. Thus, initially the L_2 locus brings an individual $2z_2^1$ units away from the optimum. Furthermore, since the initial frequency of the L_1^1 allele is small, the majority of the genotypes at the L_1 locus will be L_1^2/L_1^2, and a smaller proportion will be L_1^1/L_1^2. Thus, the majority of individuals have initially the $L_1^2L_2^2/L_1^2L_2^2$ genotype.

From this initial state, which is suboptimal, the population will move toward the optimal genotypes. Given that allele L_1^1 will not disappear and $z_1^1 \approx z_2^1$, then the optimal genotypes are the double heterozygote $L_1^1L_2^1/L_1^2L_2^2$ and the double homozygote for opposing alleles $L_1^1L_2^2/L_1^2L_2^2$.

Thus in this case there is a competition between alleles similar to the situation in other sweep models with multiple

loci (*e.g.*, Kirby and Stephan 1996). Given that the initial frequency of the L_1^1 allele is small, then fixation of the L_1^1 allele occurs when the frequency of the L_2^1 is also very small; otherwise the L_2^1 allele outcompetes the L_1^1 allele and the final state is a polymorphic equilibrium. When the initial frequency of L_1^1 is very small (as it is required for a classical selective sweep), then the percentage of trajectories that result in fixation of L_1^1 is reduced.

Furthermore, comparing the lowest frequencies for the L_1^1 allele in class 0 and class 1, we observe that the lowest initial frequency of L_1^1 in class 1 is at least one order of magnitude greater than in class 0 (6.1×10^{-3} *vs.* 1.1×10^{-5}). This means that classical selective sweeps (as described by Maynard Smith and Haigh 1974) may be rare under the symmetric fitness model (where $z_1^1 = -z_1^2$ and $z_2^1 = -z_2^2$) compared to sweeps from standing genetic variation and possible only if the effects of the alleles at different loci are similar in absolute value (Figure 1). If the difference between the absolute values of z_1^1 and z_2^1 is large, then the only optimum genotype is the double heterozygote.

The parameters z_i^j and $p_0(L_i^j)$ determine the initial distance of the population from the optimum and the variance of the genetic background. Previous studies (Lande 1983; Chevin and Hospital 2008) analyzed the initial distance from the optimum and the variance of the genetic background and emphasized their impact on the trajectory (Chevin and Hospital 2008, Equation 26b). We study the effects of the population's initial distance from the optimum and initial genetic background variance as well. Large initial distance from the optimum favors the fixation of the L_1^1 allele, whereas fixation is very rare for small initial distances from the optimum (Figure 2A). This is expected because L_1^1 is rarely advantageous when the population is already close to the optimum (it becomes advantageous only when its effect on the phenotype is very small).

Background genetic variance at time $t = 0$ has the effect opposite of that of the initial distance from the optimum (Figure 2B): larger values of background genetic variance disfavor the fixation of the L_1^1 allele. This can be explained because large values of background genetic variance imply intermediate frequencies for the alleles of the second locus, and consequently (see also Figure 1B) the vast majority of trajectories do not reach fixation. Since background genetic variance does not remain constant in the simulations, we assessed the effect of average background genetic variance (averaged over generations). It is similar to the initial background genetic variance (results not shown).

Stochastic model: Next we study the behavior of the stochastic model when the fitness matrix is symmetric. The population size $N = 10{,}000$. The simulation parameters are similar to the deterministic two-locus two-allele model with symmetric fitness matrix. We use the average frequency of the last 500 generations, $\hat{f}_{500}$, to define the equilibrium frequency. This is because the frequency of the L_1^1 allele does not remain constant but fluctuates due to random genetic drift. In Figure S1B, we plot the empirical cumulative distribution of $\hat{f}_{500}$. The proportion of trajectories with the equilibrium frequency 0.5 is largely reduced in the stochastic model. This is expected as a consequence of random genetic drift, which drives the frequency of the trajectory toward its absorbing states (compare Figure S1A with Figure S1B). To determine the importance of various parameters we plot them against the class of the trajectory. As above, for the case of deterministic trajectories three classes of trajectories are used in two comparison sets: (i) trajectories that result in fixation *vs.* trajectories that stay in a polymorphic equilibrium, and (ii) trajectories that result in fixation *vs.* trajectories that result in extinction. The definitions of the three trajectory classes (fixed, polymorphic, extinct) are given above. For the first comparison, the relation of six parameters with the class of the trajectory is depicted in Figure 3. We observe that the initial frequency $p_0(L_1^1)$, the strength of selection ω^2, and the c_1 parameter show a monotonic relationship with the probability of obtaining a trajectory of class 1. On the other hand, the initial frequency $p_0(L_2^1)$ and the contribution of the alleles L_1^1 and L_2^1 are nonmonotonic. In particular we observe that large absolute values for the contribution of L_2^1 and small absolute values for the contribution of L_1^1 favor the fixation of the allele against a polymorphic equilibrium state (Figure 3).

The role of ω^2 is crucial for the fixation of L_1^1 for the stochastic simulations. Large values of ω^2 (weak selection) result in a higher frequency of fixed trajectories. For small values of ω^2 $(1 < \omega^2 < 2)$ (one ω^2 unit represents a squared phenotypic unit), the frequency of polymorphic trajectories is about 88% in the stochastic set and about 66% in the deterministic set. For large values of ω^2 $(9 < \omega^2 < 10)$ the frequencies are 4.6 and 38%, respectively. Thus, when selection is not strong enough, stochastic trajectories are governed mostly by recombination and genetic drift and move toward the absorbing states. In the absence of mutations, these absorbing states correspond to the fixation or loss of the allele (Kimura 1983). Regarding the contributions of each locus to the genotypic value, the results are similar

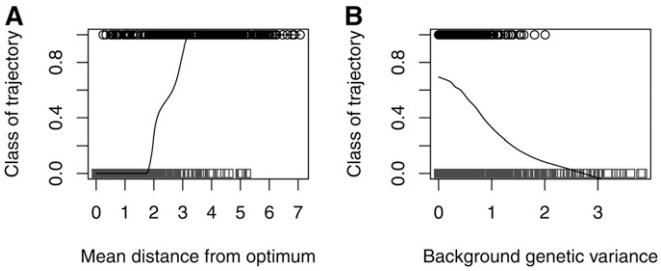

Figure 2 Initial mean distance from the optimum and background genetic variance affect the probability of fixation for the L_1^1 allele. Class 0 represents extinction of L_1^1 and class 1 represents fixation of L_1^1. Results are similar when class 0 represents the polymorphic equilibrium at frequency 0.5. (A) Small values of initial mean distance from the optimum disfavor fixation of theL_1^1 allele because L_1^1 is rarely beneficial. (B) Large values of initial background genetic variance disfavor fixation of L_1^1 because it implies intermediate frequencies of L_2^1 (and L_2^2), which has been shown (Figure 1B) to reduce the probability of fixation for L_1^1.

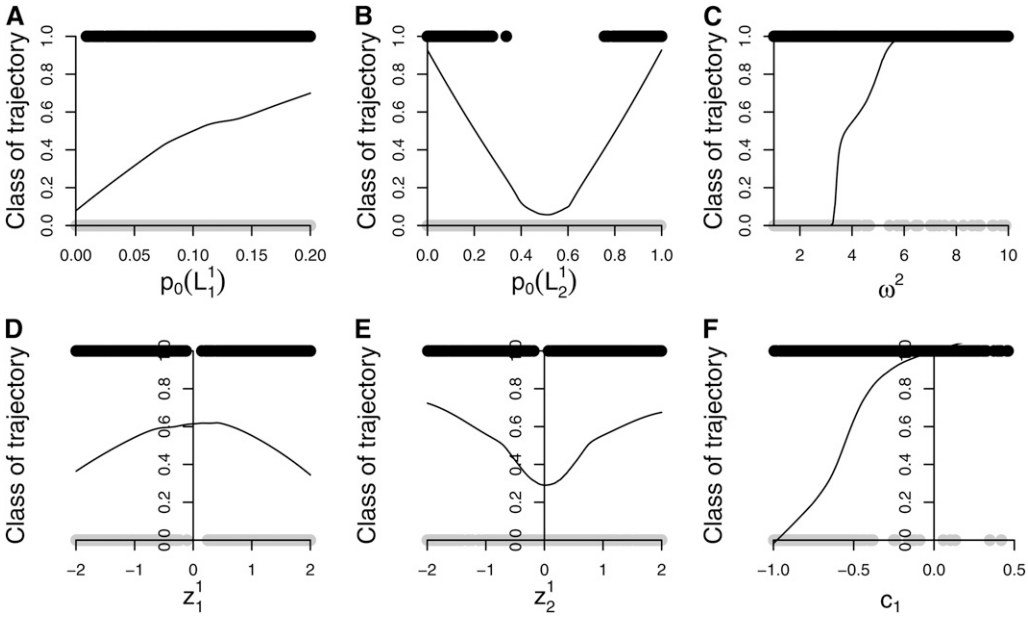

Figure 3 The relation of six parameters to the class of the trajectory (fixed *vs.* polymorphic). The curve in each subfigure represents the Lowess smoothing function of the data. (A) The initial frequency of the L_1^1 allele. (B) The initial frequency of the L_2^1 allele. (C) The parameter ω^2, which defines the strength of selection. (D) The contribution of the allele L_1^1 to the genotypic value. (E) The contribution of L_2^1 to the genotypic value. (F) The parameter c_1. Black points represent class 1 (trajectories that result in the fixation of L_1^1), whereas gray points represent class 0 (trajectories that result in a polymorphic state for L_1^1).

to the deterministic model; *i.e.*, the parameter values for the trajectories that fix are located in the proximity of the two diagonals.

Comparing the fixation class with the extinction class the following results have been obtained under the stochastic model. The roles of ω^2 and c_1 are not critical (results not shown). This means that small and large values of ω^2 have similar effects on the class of the trajectory. In this comparison both sets are associated with monomorphic (absorbing) states of the alleles. The strength of selection (at least for the values tested here) is not crucial, because maintaining either of the classes does not require strong selection (since both of the classes are absorbing states). The importance of c_1 has been explained above. In brief, c_1 is not informative for disentangling the monomorphic equilibria.

Finally, it should be mentioned that symmetry itself is not solely responsible for the frequency of the trajectory classes we observed in our simulations. Obviously, if the model is symmetric but the fitness of the heterozygote is sub-optimal (this may occur when the genotypic values are $d-\gamma$, d, $d+\gamma$, with $d$and $\gamma > 0$, and d is the distance of the heterozygote from the optimum), this model would still be symmetric but the fitness of the heterozygote would be less than the fitness of one homozygote. On the other hand, if we relax the symmetry assumption but require fitness advantage of the double heterozygote (the effects of the alleles are drawn uniformly as in previous simulations, but we require that the phenotypic value of the double heterozygote is the closest to the optimum), we do not observe the same patterns as in the symmetric model. In the symmetric model ~40% of the trajectories reach their equilibrium at frequency 0.5. For the nonsymmetric model, but with heterozygote advantage, the respective proportion of trajectories is ~10%. For the general model (see below), the respective proportion is ~0.02 (Figure S2). Thus, our simulations indicate that both symmetry and heterozygote advantage are responsible for the observed final frequencies of trajectories.

Two-locus two-allele model with general fitness function: We relax the assumption of symmetry of the fitness matrix in the deterministic two-locus two-allele model using a general fitness scheme. The parameter space is given in Table 1. Essentially, the difference between this model and the symmetric fitness model is that there is no restriction on the relations between the contributions of the alleles. Thus, the effects z_1^1, z_1^2, z_2^1, and z_2^2 are drawn uniformly from $[-2, 2]$. The optimum phenotypic value is again 0.

The shape of the trajectories in this model is similar to the symmetric fitness matrix model. The number of trajectories where the L_1^1 allele declines is similar for both models. However, in the case of the general fitness matrix model, the trajectories that result in the fixation of the L_1^1 allele occur more often than in the symmetric fitness model. This is shown by the comparison of the frequency distribution of the simulated trajectories after 10,000 generations (compare Figure S1A with Figure S1C). On the other hand, fewer trajectories stay at equilibrium frequency 0.5 (Figure S1C). This is expected because in the general fitness model the double heterozygous genotype is not necessarily associated with the highest fitness. In summary, the results indicate that in the general fitness model classical selective sweeps from rare variants may occur more often than in the symmetric fitness model.

An informative quantity for disentangling trajectories in which L_1^1 is getting fixed from those in which it stays polymorphic or disappears is the mean trait value at the beginning of the evolutionary trajectory. For mean initial trait values close to the optimum value 0, trajectories result in either extinction or polymorphic equilibrium for the L_1^1 allele

(results not shown). On the other hand, when the initial mean trait is far from the optimum, then fixing the L_1^1 allele becomes more probable. When the mean value for the trait under selection is far from the optimum, then the allele can be beneficial. On the other hand, when the population is already at the optimum or close to it, then the L_1^1 allele will not be favored in general.

Initial background genetic variance does not appear to have the same effect as in the symmetric model. In the symmetric model, we observed that the proportion of trajectories that reach fixation decreases as the initial genetic variance increases and that for large values of initial background genetic variance the proportion of trajectories that reach fixation diminishes (Figure 2B). In the general model, we observe only a slight decrease of the proportion of fixed trajectories as the initial background genetic variance increases.

The comparison between the symmetric and the general fitness model may be open to discussion because of the different sampling spaces for the allelic effects. In particular, for the two-locus two-allele model, we uniformly sample random values from $[-2,\ 2]^2$ in the symmetric model, whereas we sample from $[-2,\ 2]^4$ for the general model. Thus, initial conditions are not comparable. We used here the symmetric model to create results comparable to the Willensdorfer and Bürger (2003) model for which analytical results are available. However, the general model is a reasonable extension of the symmetric model with less restrictive assumptions, and thus we present in the next section results about the proportion of fixed trajectories for the general model only (avoiding comparisons with the symmetric fitness scheme).

Biallelic models of up to eight loci with general fitness function: We study the effect of the number of loci on the trajectory of a focal mutation at locus L_1. Both deterministic and stochastic trajectories have been implemented either by using Equation 2 to describe the evolution of more than two loci (deterministic trajectories) or with stochastic simulations. Assuming that l loci are present, we study the general model where the effects of loci are sampled in $[-2,\ 2]^{2l}$. We could not extend our simulations beyond the eight-locus model due to computational resource limitations.

We first analyzed the effect of the number of loci on the percentage of fixed trajectories in the deterministic model. For nonsymmetric fitness matrices, the maximum number of fixed trajectories occurs for two loci and then decreases as the number of loci increases (Figure 4). When the initial frequency of the L_1^1 allele is up to 0.2, the percentage of trajectories that reach fixation tends to reflect sweeps from standing variation. We also decreased the initial frequency of the L_1^1 allele to 1×10^{-3}, 1×10^{-4}, and 1×10^{-5} to approximate sweeps from new mutations more precisely. Under the general fitness model the effect of the initial frequency of the L_1^1 allele is negligible for the two-locus model: about one-third of trajectories reach fixation for either initial frequency for the two-locus two-allele model (Figure 4),

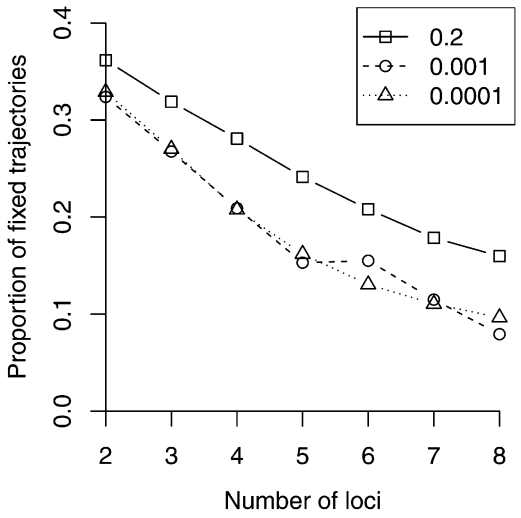

Figure 4 Effect of the number of loci on the proportion of trajectories that reach fixation for the deterministic model. The general fitness model has been studied. Parameter values used in the simulations are given in Table 1. In the legend box the number denotes the initial frequency of L_1^1. For example, 0.001 illustrates the proportion of fixed trajectories for the general fitness matrix when the initial frequency is less than 0.001.

slighly less than the percentage of trajectories that reach fixation when the initial frequency of L_1^1 is up to 0.2. For more than two loci the effect of the initial frequency of L_1^1 is weak but not negligible.

The effect of the number of loci on the percentage of fixed trajectories in the stochastic model is as follows. Due to drift the probability of a polymorphic equilibrium is reduced, and the population evolves mostly toward the absorbing states. For small initial frequencies of the L_1^1 allele ($<$0.001) the highest frequency of fixed trajectories occurs for the two-locus two-allele model ($\sim$15.4%) and for large initial frequencies ($<$0.2) the highest proportion of fixed trajectories is 21%. The proportion of trajectories that reach fixation is lower in the stochastic model than the deterministic model. Since the initial frequency of L_1^1 is relatively low, the decrease of the proportion of trajectories that reach fixation should be attributed to random genetic drift (compare Figures 4 and 5). The frequency of fixed trajectories drops approximately linearly as the number of loci increases (Figure 5).

To examine how much the frequency of fixed trajectories is reduced by the presence of other loci under selection (beyond the effect of drift), we compare the proportion of fixed trajectories in multilocus models to a single-locus model with genotypes *AA*, *Aa*, *aa*, and constant selection coefficient s_0, similar to that found initially in the multilocus model. Thus, relative fitnesses are $1 + s_0$, $1 + 0.5s_0$, and 1 for the three genotypes, respectively. Under the multilocus model, we estimated the selection coefficient s_0 of L_1^1 at time $t = 0$, using Equation A4 of Chevin and Hospital (2008). This assumes that selection is relatively weak. Thus, we simulated additional data sets for which selection is relatively weak ($\omega^2 = 100$). The initial frequency $p_0(L_1^1)$ of the L_1^1 allele is

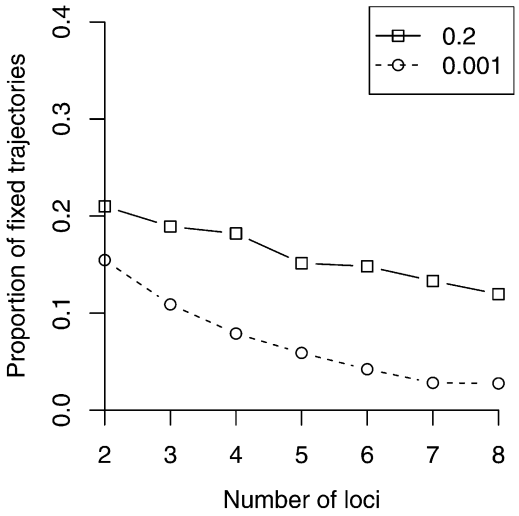

Figure 5 Effect of the number of loci on the frequency of trajectories that reach fixation for the stochastic model. Fixation is possible even for small initial frequencies of L_1^1 and the proportion of fixed trajectories decreases as the number of loci increases.

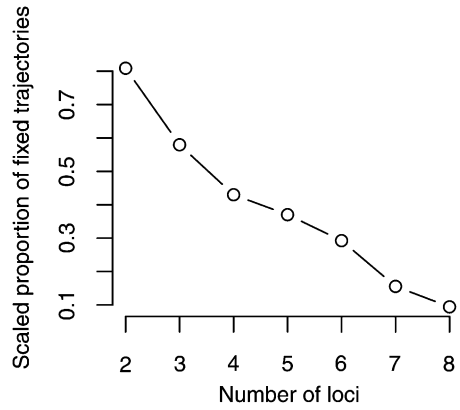

Figure 6 The ratio of the proportion of fixed trajectories/$\bar{P}(p, s_0, N)$*vs.* the number of loci for the general model. The probability $P(p, s_0, N)$ is calculated as described in the main text. The initial frequency of L_1^1 is 0.0001, and the ω^2 value is 100.

0.0001. Even though $p_0(L_1^1)$ could take any value in (0, 1), we chose $p_0(L_1^1) = 0.0001$ to study the fixation probability of a new mutation appearing in a single copy. Then we calculated the fixation probability for the one-locus model using the equation (Kimura 1962)

$$P(p, s_0, N) = \frac{1 - e^{-2Ns_0 p}}{1 - e^{-2Ns_0}}, \tag{4}$$

where N= 10,000 is the effective population size, and p is the initial frequency of the A allele. The probability of fixation is calculated by averaging $P(p, s_0, N)$ over all estimated $s_0$values. Figure 6 shows the proportion of fixed trajectories/$\bar{P}(p, s_0, N)$, where $\bar{P}(p, s_0, N)$ is the average $P(p, s_0, N)$ over all estimated s_0, for the general fitness matrix for two- to eight-locus models.

Under the general model, for more than two loci the effect of the initial background genetic variance ($t = 0$) and average background genetic variance (averaged over generations) is similar to the two-locus model: there is a slight decrease in the proportion of trajectories that reach fixation as the initial genetic variance increases. Very small initial background genetic variance favors the fixation of L_1^1. A plausible explanation might be that when initial background genetic variance is very small then heterozygosity of the background loci is small, and thus the system behaves as possessing fewer loci. In the general fitness model, the proportion of fixed trajectories increases as the number of loci decreases (Figure 4).

The ratio of the effect of the focal allele on the phenotype to the mean initial trait value: Another quantity of interest for multilocus models is the ratio of the effect of the focal allele to the mean initial trait value (henceforth denoted as φ). The results shown here refer to the deterministic model. φ is interesting because the mean initial phenotypic value may increase with the number of loci (results not shown). For the symmetric fitness model we observe the following patterns for φ: for the two-locus two-allele model the variance of φ is smaller for trajectories that reach fixation than trajectories that reach an equilibrium at frequency 0.5 or vanish (Figure S3A). The variance of φ is smaller for trajectories that reach equilibrium at a frequency in (0.5, 1) than trajectories that reach equilibrium at a frequency in (0, 0.5). Furthermore, φ is strictly negative for trajectories that fix and takes small absolute values. Negative φ implies that the signs of the initial mean trait value and the effect of L_1^1 are opposite. Therefore, L_1^1 is beneficial initially. Small absolute values indicate that z_1^1 is small relative to the mean phenotypic value. This is because fixing alleles of large effect, even if they are beneficial initially, will result in overshooting the phenotypic optimum; thus they will become deleterious during the course of evolution. This is the case especially for low initial frequencies of L_1^1, since alleles at other loci that start in higher frequencies will be more likely to fix first, thus reducing the distance to the phenotypic optimum (Chevin and Hospital 2008). For symmetric models with more than two loci, the pattern is different than for the two-locus two-allele case: trajectories that reach fixation may also have positive φ; that is, initially deleterious alleles may later become beneficial and eventually fix. This may occur if fixation of alleles from the genetic background will result in overshooting the phenotypic optimum during the course of evolution. Furthermore, the variance of φ appears more uniform for various final frequencies (Figure S3B).

For the general fitness model, under the two-locus two-allele case the variance of φ is greater than for the symmetric two-locus two-allele model, and φ may be either positive or negative (Figure S4A). Thus, even if the effect of the focal locus on the phenotype is large compared to the initial mean phenotypic value, fixation of the focal allele might still occur. Furthermore, fixation of the focal allele is possible even

when its effect is initially deleterious (*i.e.*, φ is positive). Finally, trajectories that go to fixation or extinction have greater variance of φ than trajectories that stay polymorphic. For models with more than two loci, results are similar to the two-locus two-allele model (Figure S4B).

Coalescent simulations conditioning on the trajectory of the L_1^1 allele

In this section we describe our coalescent simulations that were performed to obtain (i) the genealogies and (ii) the neutral polymorphism patterns in the neighborhood of the focal locus. The results are approximate because of two reasons. First, conditioning on the frequency of one allele implies that the coalescent rates of all genotypes that carry this allele are equal. However, in the case of multilocus models this is not true. For example, the coalescent rate of the L_1^1 allele depends on whether it is located on gamete $L_1^1L_2^1$ or $L_1^1L_2^2$ since the dynamics of these two gametes are different (see Equation 2). This is also shown in Figure S5, where a random pair of trajectories for the $L_1^1L_2^1$ and $L_1^1L_2^2$ gametes is drawn for the deterministic two-allele two-locus model with symmetric fitness matrix. The growth rates of these trajectories and their equilibrium frequencies are different. Therefore, the coalescent rate of the L_1^1 allele depends on the gamete that carries it. Note that the reason mentioned above is not equivalent to multiple selective sweeps where the trajectory at one locus is affected by the other loci because here we calculate the trajectory of the focal allele by summing up the frequencies of all gametes that carry the focal allele. The second reason is that all involved loci affect the dynamics of a neutral locus, when the recombination fraction is <0.5. In other words, a neutral allele may be under the hitchhiking effects of two (or more) selected loci. Thus, simulating the genealogy of a neutral site would require tracking the frequencies of all gametes backward in time instead of tracking the frequency of a single allele. Such an analysis is beyond the goals of this article. Combining the two previous arguments indicates that this approach yields reliable results in a close neighborhood of the sweep for cases of relatively weak selection and weakly linked loci.

Given a trajectory, coalescent simulations require specifying the time point from which the backward process is considered. That means, the genealogies will be strikingly different if the backward process initiates 100 or 5000 generations after the onset of the L_1^1 allele. Thus, an arbitrary time point is required, which represents the beginning of the backward simulation process. Here, we have used 100 generations after the trajectory has reached its equilibrium frequency. This time point is temporally close to the onset of the L_1^1allele. Therefore, the signature of the trajectory on the neutral polymorphisms is still present in the data.

Backward simulations have been performed using either a modified version of the software mbs (Teshima and Innan 2009) or the msms software (Ewing and Hermisson 2010). Our mbs algorithm implements the infinite site model, in contrast to the original software, and it calculates and outputs statistics related to the coalescent trees, such as the height, the total length, and the balance of the coalescent. msms is more efficient. Both algorithms produce equivalent results. For the coalescent simulations we have used parameters related to human data. Assuming that the mutation rate $\mu = 10^{-9}$/bp/year (*e.g.*, Zhao *et al.* 2006), then $\theta = 4N\mu = 0.001$/bp/generation. The ratio $\rho/\theta = 1$. The effective population size $N = 10{,}000$ and remains constant. Simulations are performed for a 0.5-Mb genomic segment. The locus is located in the middle of the simulated segment. The sample size is 50. For a given equilibrium frequency bin (see below), we have chosen randomly one trajectory whose initial frequency is <0.001. This is done to resemble closely a selective event of a new variant. For a given trajectory, 1000 coalescent simulations are performed. Finally, the summary statistics for the coalescent trees are computed at the recombination breakpoints for each simulation, and the results are binned. The following example demonstrates the binning process: if the (arbitrary) positions (in base pairs) $u_1 = 103989,\ u_2 = 103995,\ u_3 = 105000$ are breakpoints (*i.e.*, the genealogy may change), and the bin size is set to 100, then u_1 and u_2 will be in the same bin (1039), whereas u_3 will be in a different bin (1050). The results from the same bin are averaged over the whole set of simulations. We repeat this process for four sets of trajectories in which the equilibrium frequency is (i) 1 (fixation), (ii) 0.9 to 1, (iii) 0.3 to 0.4, and (iv) trajectories that show a nonmonotonic behavior. Trajectories in [0.9, 1] represent trajectories near fixation and trajectories in [0.3, 0.4] represent trajectories that stay in low frequency. Results for trajectories that eventually stay at a polymorphic equilibrium at low frequency are largely similar to results for the trajectories in [0.3, 0.4]. The results presented here are obtained from the analysis of the two-locus two-allele stochastic model with symmetric fitness matrix. The results for the remaining models are similar because the shape of the trajectories is similar for the various equilibrium frequencies, and we condition only on the trajectory of the focal allele (Figure 7). For trajectories that result in fixation (solid black line in Figure 7), the signatures of selective sweeps emerge in the proximity of the locus under selection: coalescent trees are shorter in length and height, and they are imbalanced in the proximity of the L_1 locus. For trajectories that result in a polymorphic equilibrium the signatures are weaker. For example, when the equilibrium frequency is between 0.9 and 1, the total length of the coalescent is smaller, and the tree imbalance is larger. However, the height of the coalescent tree is similar to the neutral expectation (thick gray solid line in Figure 7). Interestingly, the imbalance of the coalescent is higher for trajectories in the bin [0.9, 1) than the trajectories that result in fixation of the allele in the proximity of the L_1 locus. The explanation is as follows. When fixation of the L_1^1 allele has occurred, all genealogical lines coalesce in the recent past in the proximity of the locus L_1. Thus a short tree that is not imbalanced is generated because no

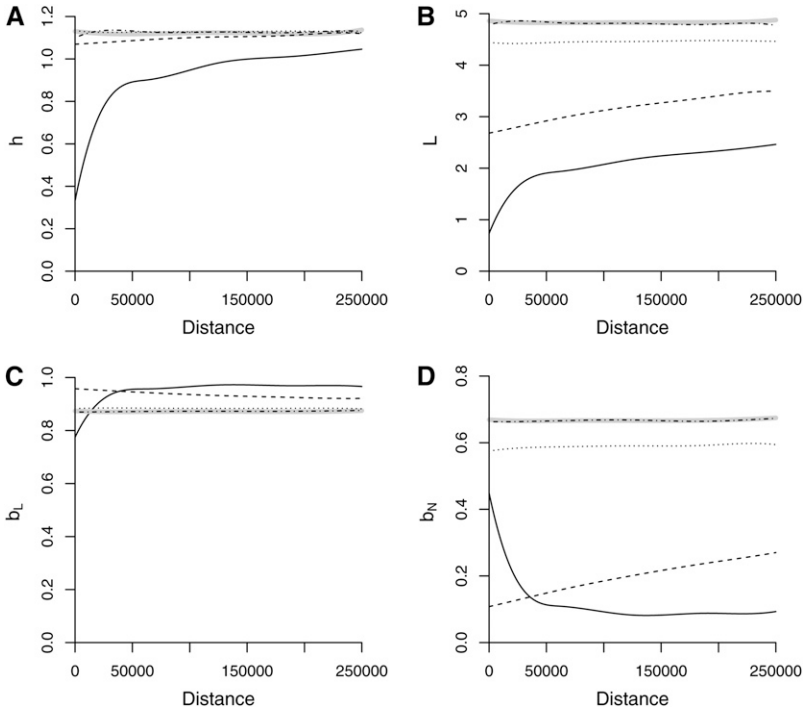

Figure 7 Summary statistics for the coalescent trees as a function of the distance from the locus. The solid line refers to the equilibrium frequency 1 (fixation), the dashed line refers to the equilibrium frequency in [0.9, 1), the dotted line refers to the frequency in [0.3, 0.4), and the thick gray solid line to neutral simulations with the same parameter values. The dash-dotted line presents the results for the nonmonotonic trajectories. Note that the results for the nonmonotonic trajectories overlap completely with the neutral curves. Each point is based on 2500 simulated trajectories. A coalescent tree was generated by conditioning on each trajectory. Trajectories are generated from the two-locus two-allele stochastic model assuming the multidimensional parameter space in Table S1. (A) The height of the tree is shown. (B) The total length of the coalescent. (C and D) The statistics b_L and b_N, respectively.

genealogical line has escaped the coalescence. Imbalance is generated further from the locus L_1, because recombination breaks the linkage between a neutral site and L_1 (Fay and Wu 2000; Kim and Stephan 2002). On the other hand, trajectories in the bin [0.9, 1) generate imbalanced genealogies very close to the locus because a large fraction of the present-day lines carry the allele L_1^1 (and coalesce in the recent past), whereas a small fraction of the present-day lines carry the allele L_1^2 and coalesce further in the past.

Furthermore, we have used molecular population genetics summary statistics to describe the properties of the polymorphisms in the proximity of the L_1^1 allele. A sliding-window approach with window size 5 kb and offset 1 kb has been implemented. The length of the genomic fragment and the position of the L_1 locus are provided in the previous section. The summary statistics are described in *Methods*. For each window the mean value of each summary statistic is calculated over the simulated data sets.

Tajima's D (Tajima 1989) is negative over the whole region for frequencies >0.9. For fixed trajectories, Tajima's D becomes less negative closely to the L_1. For trajectories close to fixation, Tajima's D obtains its most negative value exactly at the location of the locus L_1. We can associate Tajima's D with the b_N statistic. The number of polymorphic sites follows, as expected, the total length of the coalescent. For trajectories that result in L_1^1 frequencies >0.9, the expected number of polymorphic sites is minimum at the proximity of the beneficial mutation and increases gradually as we move away from it, similarly to the classical results for the one-locus model (*e.g.*, Kim and Stephan 2002).

Discussion

Overview

In this study, we explore selective sweeps in multilocus two-allele models of a quantitative trait. Selection works on the phenotype based on a Gaussian fitness function. The Gaussian function seems an appropriate choice for many quantitative traits (Endler 1986; Willensdorfer and Bürger 2003), because it naturally formalizes the concept of the evolution toward an optimum value. Furthermore, it is sufficiently flexible to allow for modeling both stabilizing and directional selection. Stabilizing selection is modeled by assuming that the optimal genotypic value is located between the extreme genotypic values that an individual may obtain. Directional selection can be modeled by assuming that the optimum is more extreme than the genotypic values that the individual may have. Therefore, the allele frequencies shift toward the direction of fixation of the most extreme genotype favored by selection.

Previous studies (Bodmer and Felsenstein 1967; Karlin and Feldman 1970) suggest that multiple equilibrium points exist in two-locus two-allele models with a Gaussian fitness function. Furthermore, conditions are provided for their existence and stability. However, the trajectories of the alleles toward the equilibrium points have not been explored. This study focuses on the trajectory of an allele, which initially is in low frequency and moves toward its equilibrium points.

An important result of our analysis shows that selective sweeps that initiate from a very low frequency of the L_1^1 allele (*i.e.*, the beneficial allele at the focal locus) are very rare in the two-locus two-allele model with symmetric fitness matrix. Multiple conditions need to be satisfied to

achieve fixation. First, the contribution of one of the alleles at the second locus (*e.g.*, L_2^1) should be approximately equal to the contribution of the L_1^1 allele. Second, the initial frequency of the L_2^1 allele needs to be very low. Under this regime the population is initially dominated by the L_1^2 and L_2^2 alleles, and thus the population is initially far away from its optimum value (since they have similar contributions as well). Thus, the L_1^1 competes with the L_2^1 allele; since their contribution is similar their initial frequencies may determine the fate of the trajectory of the allele. In fact this result suggests that in the two-allele two-locus model a selective sweep becomes possible when the second locus is nearly monomorphic, *i.e.*, when the model resembles the one-locus two-allele model. Since fixation of L_1^1 becomes more probable as its initial frequency increases, a model of sweeps from standing genetic variation may be more likely.

Relaxing the assumption of symmetric z_i^j values, we show that the fixation of the L_1^1 allele becomes more likely. This is because the optimum genotype does not correspond necessarily to the heterozygous state. For example, if the contributions of the L_1^1, L_1^2, L_2^1, and L_2^2 alleles are -1, 1, 1, 1, respectively, and the optimum genotypic value is at 0, then the L_1^1 allele is clearly beneficial: the second locus contributes $+2$ to the genotypic value, and only the $L_1^1L_1^1$ genotype may bring the population to the optimum by contributing -2. Figures 4 and 5 show that in the general fitness model a larger proportion of trajectories results in fixation compared to the symmetric fitness model.

Assuming an effective population size $N = 10000$ we also explore the effects of random genetic drift. Genetic drift increases the proportion of the trajectories that reach monomorphic states (Figure S1C). This is expected because genetic drift pushes the model toward its absorbing states. Therefore, selection needs to be sufficiently strong to maintain the polymorphic state of the trajectory. This is illustrated clearly in Figure 3, where the ω^2 value is small for the vast majority of trajectories that are polymorphic at equilibrium.

When more than two loci are modeled, the proportion of trajectories that reach fixation decreases as the number of loci increases (Figures 4, 5, and 6). The proportion of trajectories that become extinct increases, whereas that of trajectories that remain polymorphic decreases. This is in agreement with the results of Bürger (2000) who shows that when the trait is determined by more than four loci the monomorphic equilibrium points become more likely.

Model limitations

We presented a study of selective sweeps in multilocus models based on numerical calculations and computer simulations. We studied the effects of recombination rates between loci, locus contributions to the phenotypic value, selection intensity ω^2, and the initial allelic frequencies on the fixation of a focal allele. Our approach allows for background genetic variance to change over time in response to genetic drift (in the stochastic model), selection and recombination. Therefore, to some extent our approach is more realistic than previous studies where constant background genetic variance was assumed (Lande 1983; Chevin and Hospital 2008). On the other hand, we do not allow for mutations, the optimum phenotypic value remains constant, generations are discrete, mating is random, and there is no interaction between genotype and environment, while all these factors should maintain higher genetic variance in the trait than pure stabilizing selection as modeled here. With stochastic simulations several of these assumptions can be relaxed (for example, assortative mating, changing of the optimum phenotypic value, and allowing mutations). Relaxing these assumptions, however, would complicate the analysis and the interpretations of the results.

Furthermore, by using Lowess smoothing functions we study the effects of parameters independently of each other. Clearly, several parameters interact (*e.g.*, initial allelic frequencies and ω^2 values), affect the probability of fixation jointly, or confound each other. Due to the large number of parameters, however, a joint study of all parameters or parameter combinations would be prohibitive.

Additionally, we demonstrated that a significant percentage of trajectories may reach fixation under our simulation parameter values. This result might be invalid for multilocus models in general, since we omit mutations during simulations and we study only a single trait.

A further limitation of our analysis might be that different simulated models assume different numbers of free parameters. For example, as mentioned above, we sample allelic effects in $[-2, 2]^2$ for the symmetric two-locus two-allele model but in $[-2, 2]^4$ for the general two-locus two-allele model. Thus, comparing the proportions of fixed trajectories between models with different sampling space might be problematic. Similarly, when the number of loci increases, the whole range of possible genotypic values increases given that allelic effects are sampled from the same range (here $[-2, 2]$). This has the important side effect that stabilizing selection becomes stronger on average as the number of loci increases, since the deviation of initial phenotypic values from the optimum increases as the number of loci increases. It is possible that the proportion of fixed trajectories is affected by the initial selection intensity and results illustrated in Figures 4, 5, and 6 are to some extent confounded by the differences in the initial phenotypic values in models that include more than two loci.

Coalescent trees for trajectories that approach fixation are similar to coalescent trees of classical selective sweeps

Conditioning on the trajectory of the L_1^1 allele, coalescent simulations have been implemented. As mentioned previously, this is correct only for weak selection and weak linkage. However, such a first approximation is useful to study the genealogical properties and the patterns of neutral polymorphism around the L_1 locus. When the L_1^1 allele fixes in the population, then the genealogies around the L_1 locus are

similar to the classical selective sweep (given that the initial frequency of L_1^1 is small). The coalescent trees are on average imbalanced and short in the proximity of L_1, as expected in a classical selective sweep model. The imbalance increases as we move away from the focal locus, before it reverts to neutral levels.

Coalescent trees for trajectories that stay polymorphic at intermediate or low frequencies resemble neutrality

When the trajectories do not reach fixation, then a part or all signatures of a selective sweep become invisible, depending on the equilibrium frequency of the trajectory. For example, when the equilibrium frequency is between 0.9 and 1, then the height of the coalescent tree equals the neutral expectations, because ancestral alleles (L_1^2) exist in the present-day sample. For smaller equilibrium frequencies (*e.g.*, 0.3–0.4) both the coalescent and polymorphism summaries resemble the neutral expectations.

Depending on the parameter values, a large fraction of trajectories is maintained at some equilibrium value and does not reach fixation. For these trajectories analysis of incomplete sweeps (Sabeti *et al.* 2002; Voight *et al.* 2006; Tang *et al.* 2007) may be useful. There is, however, an essential difference between incomplete sweeps and sweeps in multilocus models that were studied in this article. Incomplete sweeps are on the way to fixation, whereas the sweeps studied here remain at equilibrium frequency. Therefore, the signatures of selection will be visible only in the cases in which the equilibrium frequency has been reached recently. If the trajectory remained at the equilibrium level (either polymorphic or monomorphic for the focal allele) for too long, then the signatures of selection will fade away due to recombination.

Our results indicate that detection of selection from polymorphism patterns in multilocus models may be hard. When the focal allele fixes in the population, then the statistical tools that are used to detect sweeps in one-locus two-allele models may be useful (*e.g.*, Kim and Stephan 2002; Nielsen *et al.* 2005; Pavlidis *et al.* 2010). This is also true for equilibrium points close to fixation. Even if the patterns appear to be different than those of fixed trajectories, the direction of perturbations is similar to the classical sweep models, and therefore the same statistical tools may be used. However, for smaller equilibrium frequencies some or all signatures of selection studied in this article disappear.

A hallmark of multilocus two-allele models are the nonmonotonic trajectories. This class of trajectories is absent from one-locus two-allele models. These trajectories quickly approach a certain frequency, but eventually they decline either to extinction or to some other equilibrium frequency. The difference between their maximum frequency and the equilibrium frequency may be quite large. In the simulated data sets, we observed differences even larger than 0.5. However, the polymorphism and coalescent patterns seem to be very similar to the neutral expectations. Thus, these trajectories may be completely invisible using the summary statistics studied in this article. Summarizing the results, it may be claimed that the statistical tools that have been developed to detect selective sweeps may detect only a small proportion of the multilocus selection cases, namely only those cases that result in fixed trajectories or equilibrium trajectories close to fixation. Tools that are used for detecting incomplete sweeps may be useful when the trajectory has reached its equilibrium frequency very recently. For trajectories that have reached their equilibrium frequency further in the past, we expect that recombination will destroy the signatures of selection. In fact, the results imply that positive or stabilizing selection may occur at a much higher rate than previous studies that analyze selective sweeps report (*e.g.*, Li and Stephan 2006). However, the majority of the cases remains undetectable since both the coalescent trees (as summarized here) and the polymorphism summary statistics do not deviate from neutrality.

Comparison of the present study to Chevin and Hospital (2008)

To our knowledge the only study of selective sweeps at QTL was done by Chevin and Hospital (2008). They assume an infinite number of unlinked and independent loci that control a trait. Moreover they assume that the variability in the genetic background remains constant during the selective phase and that the effect of the focal locus on the trait value is small compared to the effect of the genetic background. These assumptions enable them to solve the trajectory of a new allele analytically for linear, exponential, and Gaussian fitness functions. Chevin and Hospital (2008) focus mainly on the trajectories that reach fixation, but they also study trajectories that initially increase and vanish eventually. Under their model polymorphic equilibria are also possible for certain initial conditions (Luis-Miguel Chevin, personal communication). The results of Chevin and Hospital (2008) indicate that trajectories of new alleles evolve slightly slower than classical selective sweeps, and given that the allele will be fixed, selective sweeps look slightly older than the classical one-locus selective sweep.

In our study, we model the genetic basis of a quantitative trait explicitly, taking into account that a finite number of loci makes the model mathematically intractable. Therefore, computer simulations were employed to study the trajectory of a new beneficial allele. The contribution of alleles may be arbitrary as well as the recombination fraction between the loci. We provide information about the role of various parameters on the fixation of the trajectories, but we also study extensively the trajectories that remain polymorphic. The present study may thus be considered complementary to the study of Chevin and Hospital (2008) for finite multilocus models, providing information about the trajectories of new alleles and the polymorphism patterns generated by selective sweeps in multilocus models. Furthermore, our study complements the deterministic analyses of Willensdorfer and Bürger (2003) and Gimelfarb

(1998) by including random genetic drift. We expect that the results of this study as well as those of the previous theoretical investigations will be essential for the development of software for the detection of selective sweeps in multilocus models.

Acknowledgments

We are very grateful to two anonymous reviewers and Luis-Miguel Chevin for their valuable suggestions. This work has been supported by grants from the Deutsche Forschungsgemeinschaft Research Unit 1078 to W.S. (Ste 325/12) and D.M. (Me 3134/3) and a grant from the Volkswagen-Foundation to P.P. (I/824234).

Literature Cited

Barton, N. H., 1990 Pleiotropic models of quantitative variation. Genetics 124: 773–782.

Bodmer, W. F., and J. Felsenstein, 1967 Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics 57: 237–265.

Bulmer, M. G., 1973 The maintenance of the genetic variability of polygenic characters by heterozygous advantage. Genet. Res. 22: 9–12.

Bürger, R., 1998 Mathematical properties of mutation-selection models. Genetica 102: 279–298.

Bürger, R., 2000 *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Hoboken, NJ.

Bürger, R., 2002 On a genetic model of intraspecific competition and stabilizing selection. Am. Nat. 160: 661–682.

Bürger, R., and A. Gimelfarb, 1999 Genetic variation maintained in multilocus models of additive quantitative traits under stabilizing selection. Genetics 152: 807–820.

Bürger, R., and A. Gimelfarb, 2004 The effects of intraspecific competition and stabilizing selection on a polygenic trait. Genetics 167: 1425–1443.

Chevin, L.-M., and F. Hospital, 2008 Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645–1660.

Cleveland, W., 1979 Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74: 829–836.

Endler, J., 1986 *Natural Selection in the Wild*. Princeton University Press, Princeton.

Ewing, G., and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Fisher, R., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Gavrilets, S., and A. Hastings, 1994 Maintenance of multilocus variability under strong stabilizing selection. J. Math. Biol. 32: 287–302.

Gerhardt, H., 1994 The evolution of vocalization in frogs and toads. Annu. Rev. Ecol. Syst. 25: 293–324.

Gillespie, J. H., 1984 Pleiotropic overdominance and the maintenance of genetic variation in polygenic characters. Genetics 107: 321–330.

Gillespie, J. H., and M. Turelli, 1989 Genotype-environment interactions and the maintenance of polygenic variation. Genetics 121: 129–138.

Gimelfarb, A., 1998 Stable equilibria in multilocus genetic systems: a statistical investigation. Theor. Popul. Biol. 54: 133–145.

Hill, W. G., and P. D. Keightley, 1988 Interrelations of mutation, population size, artificial and natural selection, pp. 57–70 in *Proceedings of the Second International Conference on Quantitative Genetics*, edited by B. S. Weir, E. J. Eisen, M. M. Goodman and G. Namkoong. Sinauer, Sunderland, MA.

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "Hitchhiking Effect" revisited. Genetics 123: 887–899.

Karlin, S., and M. W. Feldman, 1970 Linkage and selection: two locus symmetric viability model. Theor. Popul. Biol. 1: 39–71.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713–719.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. Genetics 144: 635–645.

Lande, R., 1976 The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet. Res. 26: 221–235.

Lande, R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity 50: 47–65.

Li, H., 2011 A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol. Biol. Evol. 28: 365–375.

Li, H., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2: e166.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.

Mauricio, R., and M. Rausher, 1997 Experimental manipulation of putative selective agents provides evidence for the role of natural enemies in the evolution of plant defense. Evolution 51: 1435–1444.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.

Pavlidis, P., J. D. Jensen, and W. Stephan, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907–922.

Perrett, D. I., K. A. May, and S. Yoshikawa, 1994 Facial shape and judgements of female attractiveness. Nature 368: 239–242.

Robertson, A., 1956 The effect of selection against extreme deviants based on deviation or on homozygosis. J. Genet. 54: 236–248.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Stephan, W., 2010 Genetic hitchhiking *vs.* background selection: the controversy and its implications. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1245–1253.

Stephan, W., Y. S. Song, and C. H. Langley, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172: 2647–2663.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e171.
Teshima, K. M., and H. Innan, 2009 mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. BMC Bioinformatics 10: 166.
Tufto, J., 2000 Quantitative genetic models for the balance between migration and stabilizing selection. Genet. Res. 76: 285–293.
Turelli, M., 1984 Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Popul. Biol. 25: 138–193.
Vignieri, S. N., J. G. Larson, and H. E. Hoekstra, 2010 The selective advantage of crypsis in mice. Evolution 64: 2153–2158.
Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72.
Wakeley, J., 2008 *Coalescent Theory*. Roberts, Greenwood Village, CO.
Willensdorfer, M., and R. Bürger, 2003 The two-locus model of Gaussian stabilizing selection. Theor. Popul. Biol. 64: 101–117.
Zhang, X.-S., and W. G. Hill, 2002 Joint effects of pleiotropic selection and stabilizing selection on the maintenance of quantitative genetic variation at mutation–selection balance. Genetics 162: 459–471.
Zhao, Z., N. Yu, Y.-X. Fu, and W.-H. Li, 2006 Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. Genetics 174: 399–409.
Zhivotovsky, L. A., and S. Gavrilets, 1992 Quantitative variability and multilocus polymorphism under epistatic selection. Theor. Popul. Biol. 42: 254–283.

Communicating editor: L. M. Wahl

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2012/06/19/genetics.112.142547.DC1

Selective Sweeps in Multilocus Models of Quantitative Traits

Pavlos Pavlidis, Dirk Metzler, and Wolfgang Stephan

DOI: 10.1534/genetics.112.142547

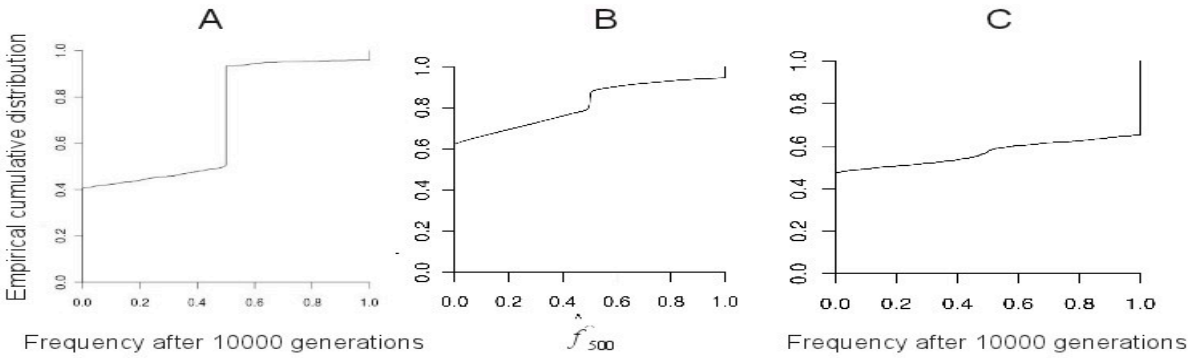

Figure S1 Empirical cumulative distribution for final trajectory frequencies. A) Frequencies obtained after 10000 simulated generations for the two-locus two-allele deterministic model with symmetric fitness. A distribution of frequencies is obtained in [0,1]. The vast majority of frequencies for the allele are either 0, 0.5, or 1. B) The empirical cumulative distribution for the equilibrium frequency of the stochastic two-locus two-allele model (where the equilibirum frequency is obtained by averaging over the last 500 generations; see main text). C) Empirical cumulative distribution of the frequencies of the trajectories after 10000 generations for the deterministic two-locus two-allele model assuming a general fitness matrix. In contrast to the symmetric fitness model, where 4.15% of the trajectories fix, in the general fitness model 34.65% of the trajectories reach fixation. The percentage of trajectories that stay at equilibrium frequency 0.5 is smaller.

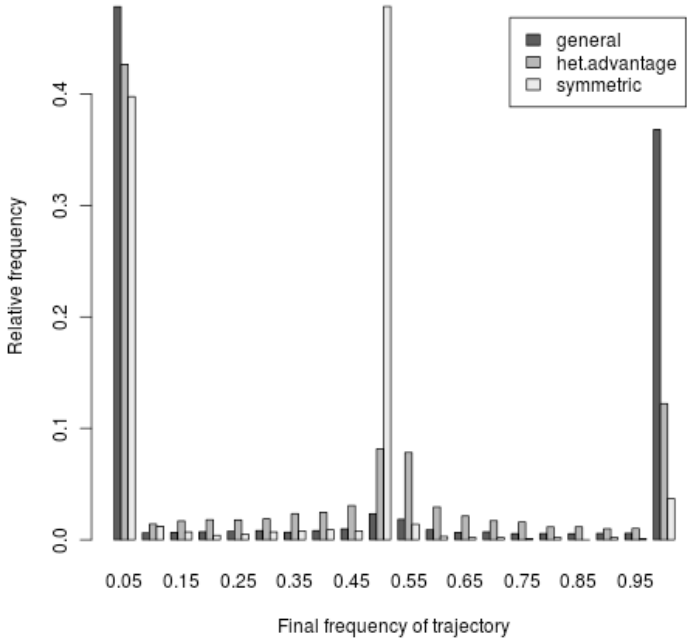

Figure S2 Histogram of final frequency of trajectories under the symmetric (light colored), general (black) and heterozygote advantage (but no symmetry) model (gray).

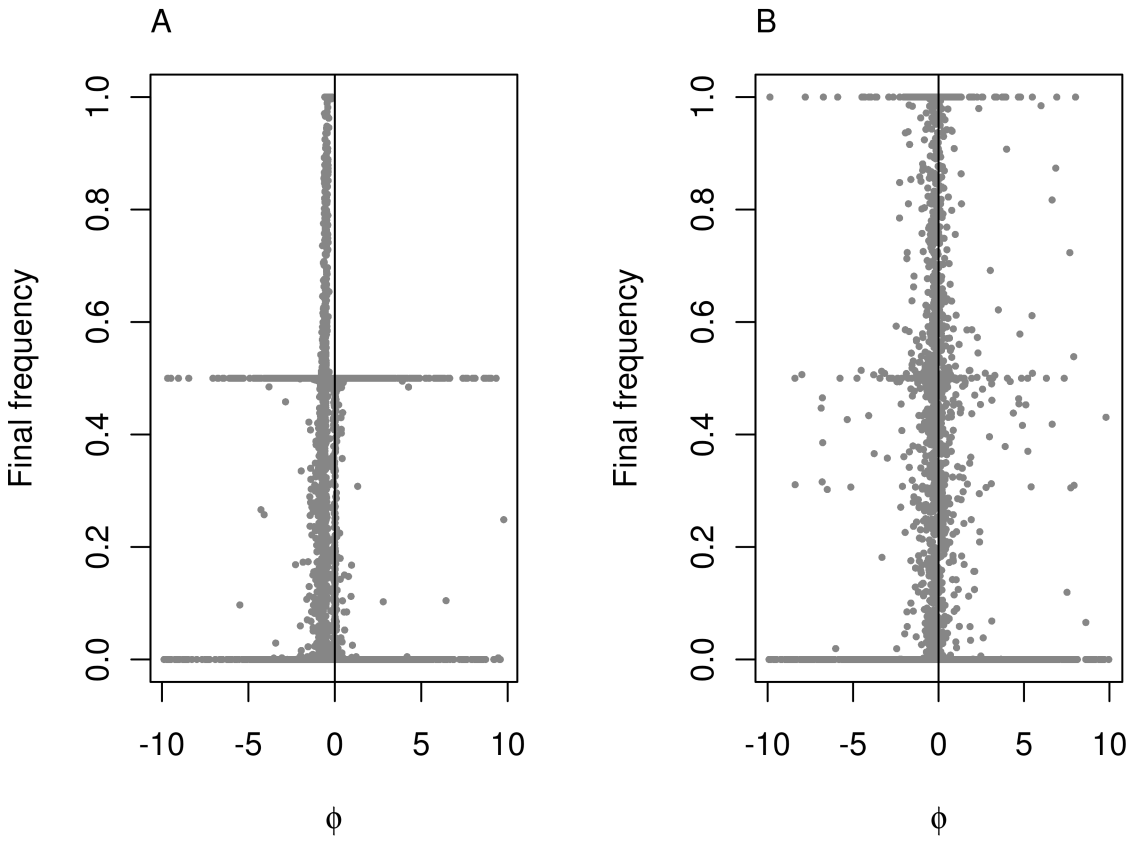

Figure S3 Effect of the ratio φ of z_1^1 to "initial mean phenotypic value" under symmetric fitness models. A) Two-locus two-allele model, B) four-locus two-allele model. For illustration only φ values between -10 and 10 are shown. There are, however, outliers whose absolute value is greater than 100 (this may happen when the initial mean phenotypic value is very small). The pattern for models with more than two loci are similar to B). As shown in A) for the two-locus two-allele model, the variance of φ is smaller for trajectories that reach fixation than trajectories that either stay polymorphic at 0.5 or vanish. Trajectories with final frequency in (0.5, 1) have a smaller variance than trajectories with final frequency in (0, 0.5). This pattern is less apparent for models with more than two loci.

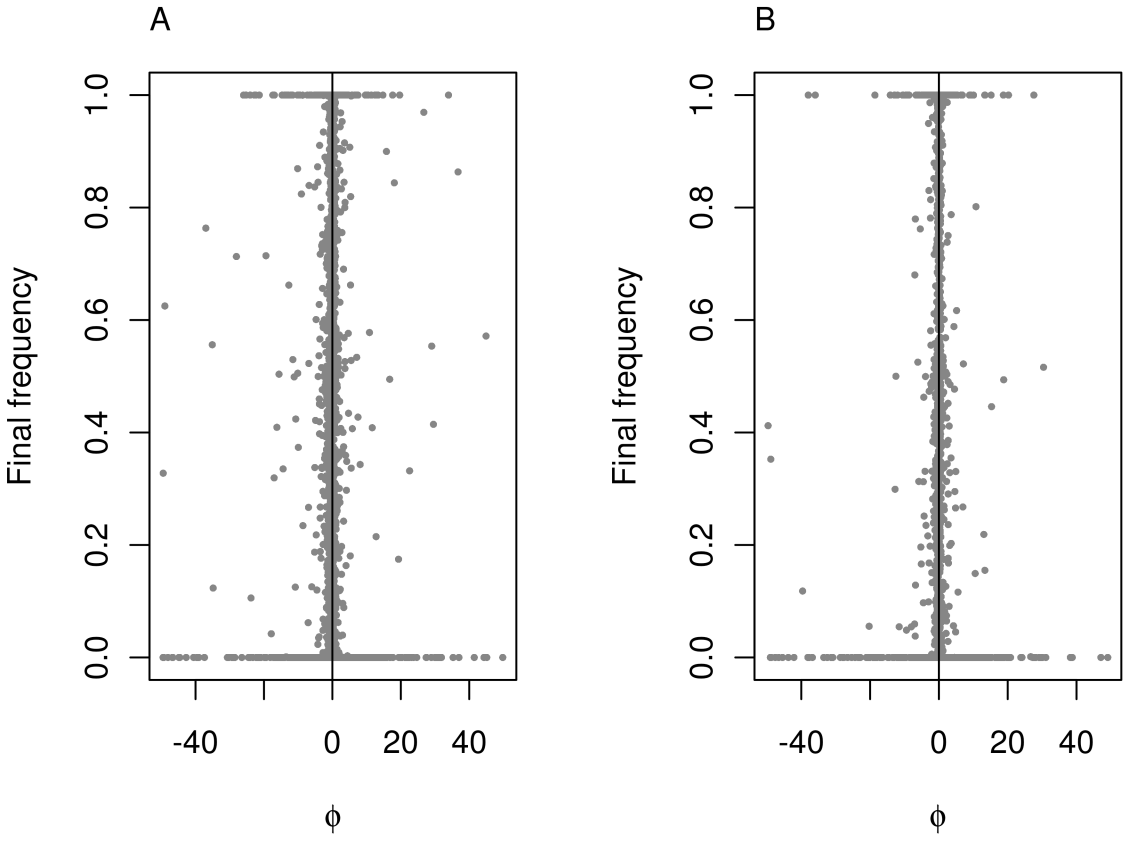

Figure S4 Effect of the ratio φ of z_1^1 to "initial mean phenotypic value" under general fitness models. A) two-locus two-allele model, B) four-locus two-allele model. For illustration only φ values between -50 and 50 are shown. Variance of φ for the absorbing states (frequency 1.0 or 0.0) is greater than for the intermediate polymorphic frequencies.

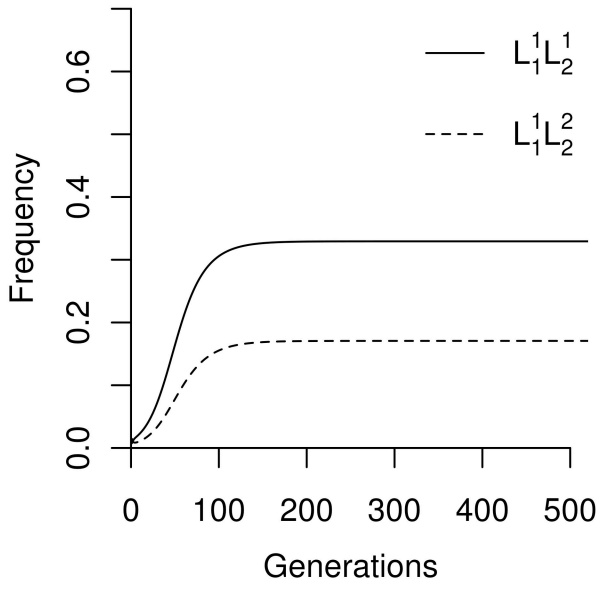

Figure S5 The growth rates for the $L_1^1L_2^1$ and $L_1^1L_2^2$ trajectories are different. The $L_1^1L_2^1$ trajectory grows faster than the $L_1^1L_2^2$ trajectory. Therefore, the coalescent rate for the L_1^1 allele depends on whether it is located on the gamete $L_1^1L_2^1$ or on $L_1^1L_2^2$.

Table S1 (A) Genotypic values and (B) fitnesses for the symmetric fitness model

A

	$L_2^1 L_2^1$	$L_2^1 L_2^2$	$L_2^2 L_2^2$
$L_1^1 L_1^1$	$-\gamma_1 - \gamma_2$	$-\gamma_1$	$-\gamma_1 + \gamma_2$
$L_1^1 L_1^2$	$-\gamma_2$	0	γ_2
$L_1^2 L_1^2$	$\gamma_1-\gamma_2$	γ_1	$\gamma_1+\gamma_2$

B

	$L_2^1 L_2^1$	$L_2^1 L_2^2$	$L_2^2 L_2^2$
$L_1^1 L_1^1$	$1-d$	$1-b$	$1-a$
$L_1^1 L_1^2$	$1-c$	1	$1-c$
$L_1^2 L_1^2$	$1-a$	$1-b$	$1-d$

File S1

Implementation

Forward simulations have been implemented in a C++ program available on the webpage http://bio.lmu.de/~pavlidis or upon request (pavlidisp@gmail.com). The stochastic model assumes N diploid individuals. The number of loci l may be arbitrary. However, large (>20) values of l may require extensive computational time and memory. For each generation, N individuals are chosen as fathers, and N as mothers according to their fitness value. For each sex this is done by multinomial sampling with parameters N and $(F_1, ..., F_N)$, where F_i is the fitness of the ith individual normalized by the average fitness of the population. The same individual is possible to be a mother and a father. Then, recombination occurs for each parent, and a recombinant chromosome is generated that will pass to the next generation. Mating is random. All measurements (frequencies of alleles, average fitness, average trait value etc.) are calculated in the zygote step.

The code provides further extensions to the classical two- and l-locus models as described in the main text. First, it allows for different optimum values for male and female individuals. Second, the optimum for the trait may change after time t (t follows either an exponential distribution, or it is predefined by the user) to a new value, which is either uniform or predefined by the user. Additionally, mutations can be assumed to occur for each locus. The environmental effect follows a Gaussian distribution, or is absent. The effective population size is constant, but an extension to a (stepwise) changing population size can be readily implemented.