Short-term adaptation on tunable fitness landscapes from standing genetic variation in recombining populations

Juan Li[1,2,*,a], André Amado[1,2,*,a], and Claudia Bank[1,2]

[1]Institute of Ecology and Evolution, University of Bern, Bern, Switzerland
[2]Swiss Institute for Bioinformatics, Lausanne, Switzerland
[*]These authors contributed equally to this work.
[a]To whom correspondence should be addressed: juan.li@unibe.ch, andre.amado@unibe.ch

Abstract

How does standing genetic variation affect polygenic adaptation in recombining populations? Despite a large body of work in quantitative genetics, epistatic and weak additive fitness effects among simultaneously segregating genetic variants are difficult to capture experimentally or to predict theoretically. In this study, we simulated adaptation on fitness landscapes with tunable ruggedness driven by standing genetic variation in recombining populations. We confirmed that recombination hinders the movement of a population through a rugged fitness landscape. When surveying the effect of epistasis on the fixation of alleles, we found that the combined effects of high ruggedness and high recombination rates lead to preferential fixation of alleles that had a high initial frequency. This indicates that positively epistatic alleles escape from being broken down by recombination only if they start at high frequency. We further infer single selection coefficients and pairwise epistasis along the adaptive path. When taking the final fixed genotype as the genetic background, we observe beneficial direct selection is overestimated and pairwise epistasis is underestimated. Quantitatively, the inferred ratio of epistasis and direct selection (≈ 1) is smaller than expected if the population was traversing the landscape randomly. Thus, adaptation on a rugged fitness landscape may lead to spurious signals of direct selection generated through epistasis. Our study highlights how the interplay of epistasis and recombination constrains the adaptation of a diverse population to a new environment.

Introduction

The role of epistasis, which is the genetic-background dependent effect of mutations on fitness, is a conundrum in evolutionary biology. Because of the complexities it causes, epistasis is very difficult to infer and characterize, although it undoubtedly exists in natural traits (Fenster et al. 1997; Crow 2010; Campbell et al. 2018). For instance, strong negative epistasis, manifested as hybrid incompatibility, maintains reproductive isolation between species (Fishman and Sweigart 2018). At the same time, the significance of epistasis during polygenic adaptation is a topic of active debate.

In theoretical biology, models of fitness landscapes, which map genotypes or phenotypes to fitness, have been a recurrent resource for probing the effect of epistasis on adaptation (Bank 2022). Theoretical work on the role of epistasis during adaptation on fitness landscapes has often assumed a strong selection weak mutation regime (SSWM, Gillespie 1984). SSWM implies that only one segregating site is under selection at any time. Under this assumption, only currently beneficial mutations (based on the current underlying genetic background) in an adaptive walk can fix. Thus, interference between selected variants and recombination are neglected. SSWM adaptation on fitness landscapes is well studied and allows for sophisticated analytical characterization of the adaptive process (*e.g.*, Orr 2002; Bank et al. 2016; Vaishnav et al. 2022). Despite the importance and beauty of the SSWM approximation, natural populations carry standing genetic variation that has evolutionary consequences, the study of which requires a relaxation of the SSWM assumption.

Standing genetic variation (SGV) plays a predominant role in local adaptation (Barrett and Schluter 2008; Marques et al. 2019; Bomblies and Peichel 2022), especially in the face of rapid environmental change. In such a situation, the typically low mutation rates are unlikely to be

sufficient to facilitate rapid adaptation by new mutations (Orr and Unckless 2014). In contrast, if polymorphism is already segregating in a recombining population, adaptation to a new niche can be achieved quickly. An empirical example comes from a songbird species (*Sinosuthora webbiana*, Lai et al. 2019). Here, shared standing variants were shown to facilitate high-altitude adaptation in two songbird populations. Interestingly, even the private SNP candidates of adaptation in the two populations are likely to have preexisted in the ancestral population, as was inferred under consideration of the demography of the populations.

Epistasis causes linkage disequilibrium (LD) among standing genetic variants; thus, excess LD could empirically indicate epistasis. For instance, LD among non-synonymous variants within a gene was shown to be significantly larger than LD among non-synonymous variants between genes, especially among interacting amino acids, but this pattern was not observed for synonymous variants (Stolyarova et al. 2022). However, LD for epistatically interacting gene clusters is altered by recombination. On the one hand, most new mutations rely on recombination to bring them together. On the other hand, recombination may break up LD between positively epistatic variants. This is because recombination shuffles genotypes, favoring the direct selection of individual alleles rather than their combined epistatic effects (Neher and Shraiman 2009). Due to its second-order effect, epistasis can only amplify individual genotypes to a high frequency when its effect is large enough to compensate for the removal due to recombination breakdown (Neher et al. 2013).

How does SGV affect adaptation on a fitness landscape in the presence of recombination? SGV can be shuffled by recombination creating new genotypes and thus opening up new adaptive opportunities. Only a small proportion of the fitness landscape, namely those genotypes which are composed of the segregating alleles, can be traversed by standing variation in the absence of new mutations. However, on a short time scale, this is likely to be a much larger number of genotypes than *de novo* mutation would create, especially when mutation rates are low, like in most organisms. Empirically, it is challenging to infer how a population has used its SGV to respond to environmental change and how epistasis has contributed to shaping the evolutionary trajectory.

Here, we approach this question computationally using a fitness landscape model with tunable epistasis. We study the short-term evolutionary dynamics of a haploid recombining population that evolves on fitness landscapes with varying degrees of epistasis by means of standing genetic variation. We quantify how recombination rates and epistasis affect the time until the population becomes monomorphic and the final population fitness. Moreover, we characterize the fixed alleles and infer pairwise epistasis along the adaptive path. We discuss our results in light of the contribution of epistasis to evolutionary trajectories in nature.

Methods

Modeling short-term adaptation in haploid recombining populations

We model the short-term evolution of a haploid recombining population in the presence of epistasis (Figure 1). We initialize a population in the genotype space that has evolved under neutrality; thus, we draw the segregating sites in the population from a neutral site frequency spectrum distribution (explained further below). When a sudden environmental change occurs, these previously neutral sites become selected according to a fitness landscape model. Since

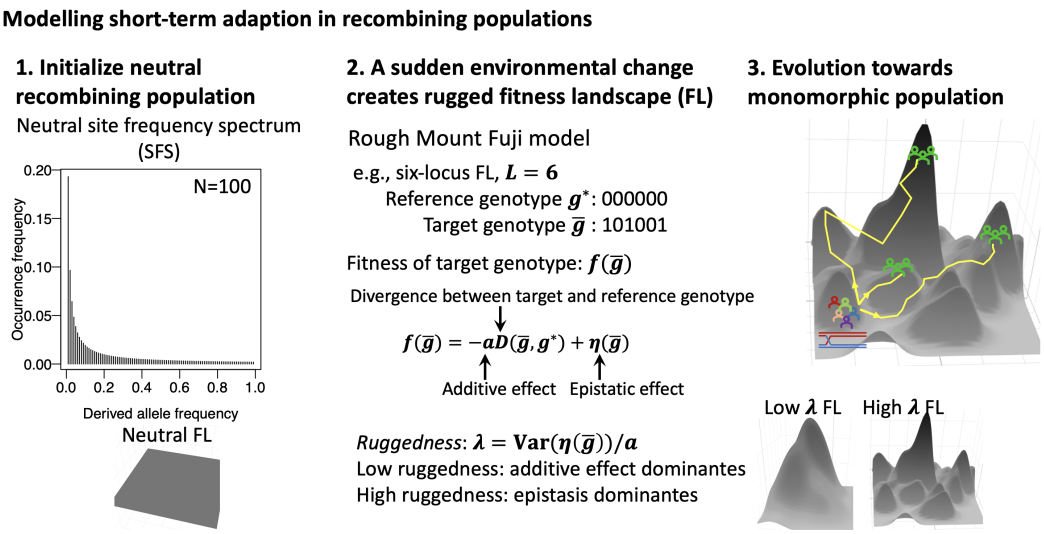


Figure 1. Illustration of the modeling approach. 1. We assume that SGV arose from neutral evolution. 2. Upon environmental change, the population experiences a rugged fitness landscape of the Rough-Mount-Fuji type. 3. Evolution occurs until the population is monomorphic.

we are interested in the short-term adaptation process that follows the change in environment, we neglect *de novo* mutations and focus only on the evolution of the preexisting segregating diversity. We use a Rough Mount Fuji fitness landscape model (RMF model Aita et al. 2000) to define the fitness landscape in which the population is evolving. As the population evolves, alleles become successively fixed due to selection and drift. The adaptation process ends when the population is monomorphic.

In our simulations, each population contains N haploid individuals and follows Wright-Fisher dynamics, with discrete non-overlapping generations. Every generation, individuals are grouped in pairs and allowed to recombine to form the gamete pool. We assume equidistant loci, such that the recombination probability between two neighboring loci is uniform, and the maximum recombination probability between two loci is 0.5 (here, all loci are independent). After recombination, N individuals are selected from the gamete pool with a probability proportional to their fitness (obtained from a preset fitness landscape). In the main text, we study $N = 5000$ to represent large populations. The supplementary material presents results for smaller populations of $N = 500$ and $N = 100$, representing strong(er) genetic drift scenarios. For each parameter set, we simulated 100 fitness landscapes. For each fitness landscape, we simulated adaptation for 100 initial genotype distributions.

Generating fitness landscapes according to a Rough Mont Fuji model

We generate fitness landscapes with a tunable fraction of epistasis using the RMF model (Aita et al. 2000). We assume genotypes with L diallelic loci (alleles 0 and 1). Allele 1 at each locus is assigned a constant additive fitness effect a. In addition, each genotype receives an epistatic contribution ($e_{\bar{g}}$) to its fitness. The fitness of genotype $\bar{g}$ is

$$f_{\bar{g}} = a \sum_{j=1}^{L} g_j + e_{\bar{g}} \,.$$

Here, g_j represents the allele at locus j ($g_j \in \{0, 1\}$). The epistatic term $e_{\bar{g}}$ is a random variable drawn from a normal distribution, $N(0, \sigma^2)$, with zero mean a standard deviation σ.

After drawing the genotype frequencies of initial population according to a neutral SFS, as described above, the fitness landscape is parameterized such that minor alleles receive the additive effect $a = 0.01$ (i.e., for each locus, the less frequent allele is assigned the additive beneficial effect). Due to this assignment, the mean fitness of the initial population is by definition lower than the maximal additive fitness. In addition, a random epistatic fitness contribution is assigned to each genotype, drawn from the normal distribution described above. The ratio of the standard deviation of the epistatic effect σ to the additive effect a determines the ruggedness of the fitness landscape. Low ruggedness implies that the additive effect dominates; high ruggedness indicates that the epistatic effect dominates. We probe fitness landscapes of different ruggedness by studying three values of σ, 0.001, 0.01, and 0.1. Hence, the ruggedness of our fitness landscape is 0.1, 1, or 10, which we refer to as smooth, intermediate and rugged fitness landscapes.

Defining the starting occupancy of the population on the genotype space

Choosing a starting occupancy of the genotype space is nontrivial since the starting frequencies greatly affect the outcome of the dynamics. To create a realistic initial frequency distribution of alleles and genotypes in the genotype space, we assume that previous neutral evolution has generated a neutral site frequency spectrum from which we draw allele frequencies for each locus, which are combined into genotype frequencies under the assumption of linkage equilibrium. Moreover, all alleles are defined to have an initial frequency above a genetic drift threshold, as detailed in the following subsection. This assumption reduces the homogenizing effect of random genetic drift in the first generations of our simulations. In the following, we describe these assumptions in detail.

Site frequency spectrum in the initial population

As stated before, we consider a haploid population with N individuals, originally at neutral mutation-drift equilibrium, suddenly exposed to a fitness landscape. Each genotype contains L diallelic loci. The allele frequency follows a discrete neutral site frequency spectrum (SFS, Hudson 2015). We define a minor and major allele for each locus, corresponding to the allele with the smaller and larger frequency in the population. The mean and variance of the minor allele frequencies (MAF) of a neutral SFS decrease with increasing population size (Figure S1). In the infinite loci approximation, the mean of minor allele frequencies is $\bar{p} = \int_0^{1/2} p\varphi(p)dp + \int_{1/2}^1 (1-p)\varphi(p)dp$, and its variance is $\sigma^2(p) = \int_0^{1/2} (p-\bar{p})^2\varphi(p)dp + \int_{1/2}^1 (1-p-\bar{p})^2\varphi(p)dp$, where $\varphi(p) = 2N\mu/p$ (Wright 1938). Lower-frequency alleles ($< 1/N$) shift the SFS towards lower frequencies when N increases. Hence, a larger population size results in a smaller mean MAF and a wider variance of the MAF.

Genetic drift threshold

Alleles present at a very low frequency in the initial population are easily lost by genetic drift, independently of their selective advantage. Since we are interested in studying the

population's response to selection, we impose a genetic drift threshold to the frequency of the alleles, which ensures a lower probability that selected alleles are eliminated by drift at the beginning of the evolutionary process. Assuming a haploid population of size N, an i-copy minor allele has a frequency i/N. The probability of losing this minor allele in the following generation is $p_{lost} = (1 - i/N)^N$. When N is large enough, this is $p \approx e^{-i}$, which is always true in our study since the smallest considered population size is 100.

Assuming we want to ensure that $p_{lost} < 0.05$, then $i \geq 3$, i.e., we should consider minor alleles with at least three copies in the initial population. We define this as the genetic drift threshold in the population. With or without genetic drift, the mean of the minor allele frequency is correlated with the population size (N), which spans 100 to 5000 in our simulations. The mean is always below 0.2; it is largest for $N = 100$ and drops to below 0.1 as the population size increases (Figure S1A and S1B). Using the drift threshold slightly elevates the mean and variance of minor allele frequencies (Figure S1C and S1D), resulting in a more diverse initial population.

Fraction of the fitness landscape covered by the initial population

When studying adaptation from SGV only, the initial state of the population plays a prominent role. Unless the fitness landscape is very small and the population is large, only a fraction of all genotypes in the genotype space will be represented in the initial population. To initialize the population, we adopt the neutral folded SFS in a discrete form (see Hudson 2015 for neutral SFS description). Minor allele frequencies are $1/N, 2/N, \ldots, N/2N$ (when N is even, or $(N-1)/2N$ if N is odd). Below, we assume N is an even number. Thus, the probability that a given allele appears i times in the population is given by $p_i = \frac{(1/i + 1/(N-i))}{\sum_{i=1}^{N-1} 1/i}$, summing the terms corresponding to i and $N - i$ (because the SFS is folded) and normalizing the distribution. Therefore, the mean of minor allele frequency i is $\bar{p}_{minor} = \sum_{i=1}^{N/2} i/N * p_i$, and $\bar{p}_{major} = 1 - \bar{p}_{minor}$. Both p_{major} and p_{minor} are arithmetic means. Suppose we sample the initial population from this distribution. In that case, we produce a set of genotypes akin to those sampled from a population in mutation-drift balance before changing the environment (the starting point of our simulations).

We compute the expected initial occupation of the fitness landscape with L loci. The probability of an individual carrying a given genotype with k minor alleles is approximately $h_k = \bar{p}_{minor}^k \, \bar{p}_{major}^{L-k}$. Thus, the probability that a specific genotype is not present in the population is $(1 - h_k)^N$. We can then sum the probability of a genotype being present in the population over all genotypes and divide this sum by the total number of possible genotypes. This corresponds to the expected fraction of the genotype space occupied by the initial population,

$$P_{occupied} = \frac{1}{2^L} \sum_{k=0}^{L} \frac{L!}{k!\,(L-k)!} \left[1 - (1 - h_k)^N\right].$$

The factor after the sum accounts for the number of genotypes that contain k minor alleles. Given a genetic drift threshold, we can re-estimate $P_{occupied}$ by setting the lower boundary as $i_t = 3$ (Figure 2).

Inferring selection along the adaptive path

In order to represent empirical cases in which the identity and order of fixed alleles may be known but a characterization of the combinatorially complete fitness landscape is rarely possible, we extract the direct selection coefficients and pairwise epistasis inferred along the adaptive trajectory. Note that there is no “adaptive walk” in the true sense since the population is polymorphic until the end of the simulation; we thus use the term “adaptive path” throughout the manuscript to denote the successive fixation of alleles. First, we define the final genotype or the initial major genotype as a reference genotype. Thus, we infer the single-locus effect as the effect of a single substitution between the fixed and the lost alleles along the adaptive trajectory, with the rest of the genetic background being that of the reference. We infer pairwise epistatic effects as the epistasis that is measured when considering the two-locus fitness landscape that spans the alleles at two consecutively fixed loci.

Data Availability

Data sharing not applicable to this article as no datasets were generated or analysed during the current study. All simulation code will be archived on Zenodo upon publication of the manuscript.

Results

Initial occupation of the fitness landscape

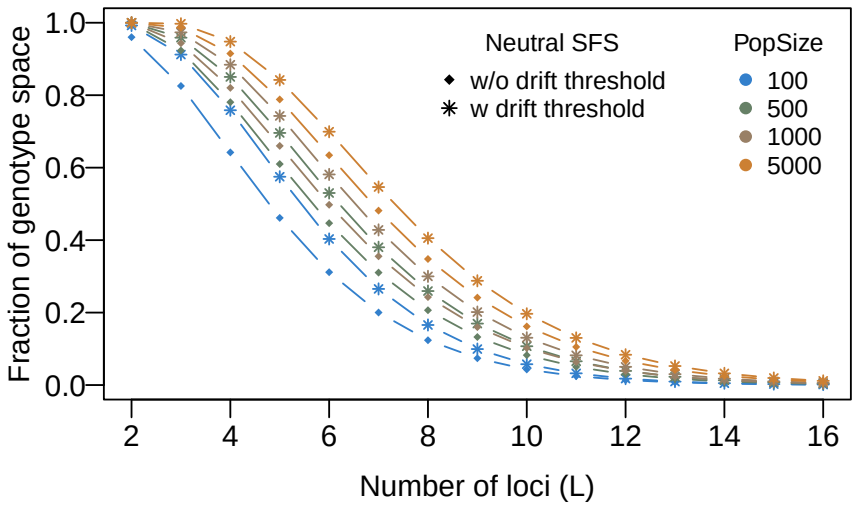


Figure 2. Approximate fraction of the genotype space initially covered by populations of different sizes. The filled diamond corresponds to populations generated without (w/o) the drift threshold, and the star corresponds to populations generated with (w) the drift threshold. Four population sizes (PopSize) are shown. As the number of loci increases, a smaller fraction of genotypes is covered by the initial population. For analytical results, see Methods.

Diallelic fitness landscapes are hypercubes of size $2^{\text{number of loci}}$. The number of genotypes in these landscapes increases rapidly as the number of loci increases. Consequently, even if

every locus is originally polymorphic, only a fraction of possible genotypes can be covered unless the fitness landscape is very small. Thus, without *de novo* mutation, recombination is necessary to explore any additional genotypes.

Based on the mean of minor allele frequencies, we computed the number of genotypes in the initial population given an L-locus genotype (Figure 2; see Methods). Here, the smallest landscape contains $2^2 = 4$ genotypes, whereas the largest reaches $2^{16} = 65536$ genotypes. Hence, already for a moderate number of loci, the size of the genotype space grows much larger than the population size used in our simulations ($N \in [100, 500, 5000]$, Figure 2). An initial population with SGV only covers the whole genotype space when the fitness landscape is very small. When the fitness landscape has more loci, a population can only cover a small fraction of genotypes in the genotype space. In these larger landscapes, nearly every individual is unique, and each genotype is mainly composed of major alleles.

Time to fixation is affected greatly by ruggedness and subtly by the recombination rate

In our simulations, adaptation occurs until the standing genetic variation in the population is exhausted and the population becomes monomorphic. The required time to fixation of a single genotype depends on a range of factors, namely the ruggedness of the landscape, the population size, and the recombination rate. The ruggedness of the fitness landscape has the strongest effect on the fixation time. Figures 3A and 3B show that populations achieve fixation one order of magnitude faster in the case of highly rugged landscapes relative to fitness landscapes with low or intermediate ruggedness. Two factors contribute to this result. Higher ruggedness entails stronger selection between genotypes since the fitness differences between neighboring genotypes are larger. On the other hand, highly rugged fitness landscapes have many local peaks where the population can get trapped and thus shorten the adaptation path until fixation. Therefore, higher ruggedness also leads to a wider variance in fixation times. The variance in the fixation time is always maximal when there is no recombination, a trend that is especially visible for fitness landscapes with low ruggedness.

For large fitness landscapes ($L = 15$), the least rugged landscapes yield the largest fixation times for low recombination rates, whereas the maximum fixation times for high recombination rates are reached for fitness landscapes of intermediate ruggedness. All these effects are systematic but weak as compared to the variation among replicates. For large and highly rugged landscapes, we observed a minimum of the fixation time and its variance at intermediate recombination rates. The population size also plays a crucial role in determining the fixation time, with larger populations reaching fixation relatively faster than smaller populations, measured in units of the population size (Figure S3). In general, the effects of population size and landscape ruggedness on the fixation time are much larger than the effect of landscape size and recombination rate.

Fitness gain decreases with increasing recombination at high ruggedness

Recombination breaks the linkage among positively epistatic alleles (*e.g.*, Neher and Shraiman 2009). Thus, high recombination should reduce the potential gain of mean

population fitness in epistatic fitness landscapes. On the other hand, recombination brings together beneficial alleles in additive landscapes, thereby reducing Hill-Robertson interference (HRI; Hill and Robertson 1966). We calculated the fitness change from the initial mean population fitness to the final genotype, normalized by the distance of the mean fitness of the initial population to the global fitness peak. As expected, in fitness landscapes with low ruggedness, the fitness gain increases with the recombination rate ($\lambda = 0.1$ or 1, Figure 3 and S4). Here, increasingly many beneficial mutations (i.e, those with an additive beneficial effect) are fixed because recombination reduces HRI. We observe the reverse pattern in fitness landscapes of higher ruggedness. Here, the gain in population fitness decreases with increasing recombination rates (Figure 3, S4C and S4F).

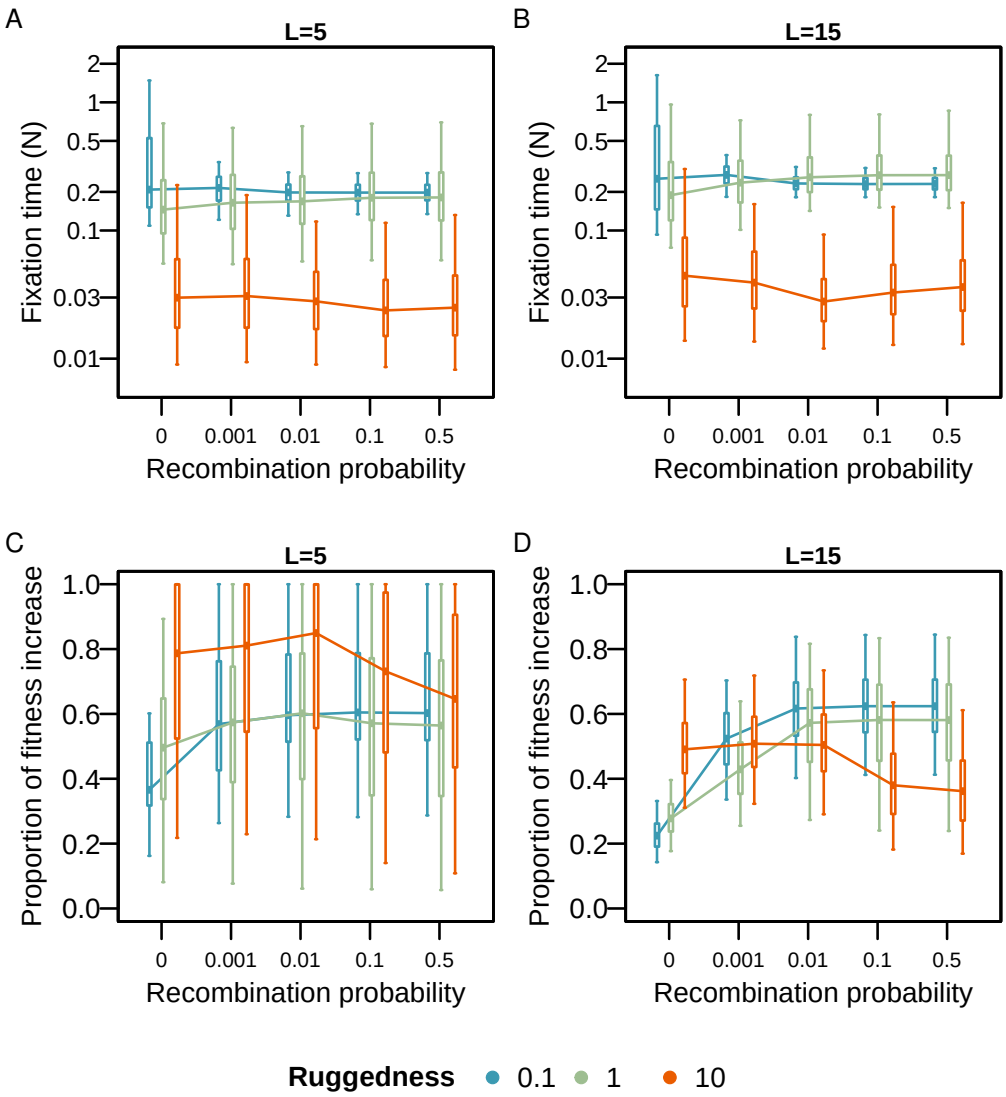


Figure 3. Fixation time and relative fitness change. A & B. The distribution of times until the population becomes monomorphic, scaled by the population size. C & D. Fitness change between the initial mean fitness and the fitness of the final genotype, normalized to the distance of the mean initial fitness to the global peak. For both small (A; $L = 5$) and large (B; $L = 15$) fitness landscapes, fixation occurs most rapidly in rugged fitness landscapes. Although the fixation time is similar across recombination rates, the fitness gain through adaptation decreases with the recombination rate for rugged fitness landscapes (C & D), whereas it increases with the recombination rate for low and intermediate ruggedness in large fitness landscapes (D). For each parameter set, we simulated 100 fitness landscapes and allowed 100 populations with different initial genotype distributions to adapt on each fitness landscape. Each population contains 5000 individuals ($N = 5000$). We study fitness landscapes of varying ruggedness (0.1, 1, 10) by studying three epistasis parameters, 0.001, 0.01, and 0.1, while maintaining a constant additive effect 0.01. See also Methods.

Initial frequencies determine which alleles become fixed on rugged fitness landscapes

The ruggedness of the fitness landscape greatly affects whether minor alleles become fixed. We assigned the positive additive effect to minor alleles; therefore, if the direct selection is effective, minor alleles should predominantly become fixed resulting in increased divergence. As expected, divergence increases with recombination probabilities for fitness landscapes with low and intermediate ruggedness ($\lambda = 0.1$ or 1, Figure 4A, 4B and S6). Here, high recombination rates favor the fixation of genotypes with more beneficial (minor) alleles (Figure 4C-E, S7 and S8).

Divergence decreases with increasing ruggedness of the fitness landscape, and this effect is further amplified by recombination ($\lambda = 10$, Figure 4A and 4B). Indeed, we observed that the initial frequency of fixed alleles increases with recombination (Figure 4C, S7 and S8). Thus, the proportion of fixed minor alleles declines. As the epistatic coefficient is independent of allele frequencies, we suspect that a high recombination rate favors the fixation of combinations of positively epistatic high-frequency alleles. Moreover, with increasing ruggedness, high-fitness genotypes are more evenly distributed across the whole fitness landscape. Thus, the chance that the initial population contains high-fitness genotypes increases. The relationship between divergence and ruggedness levels is less pronounced at lower population sizes because genetic drift overwhelms selection ($N = 100$, Figure S6).

Across all epistasis parameters, mostly major alleles become fixed during the first few steps of the adaptive path (Figure 3C&D). This can likely be attributed to genetic drift and competition among many selected genotypes. During these initial fixation steps, individually beneficial (minor) alleles are lost even in the presence of recombination. At low ruggedness, the mean initial allele frequencies of fixed alleles over the fixation trajectory for different recombination probabilities separate very early, indicating that recombination begins to save beneficial alleles from being lost (step 2, Figure 3C&D). Therefore, at high recombination, more minor (beneficial) alleles become fixed over time (Figure 4D). At high ruggedness, the mean initial allele frequencies of fixed alleles over the fixation trajectory remain indistinguishable until fixation step 7 (Figure 3C&D). Here, many high-fitness genotypes are distributed randomly across the fitness landscape, and hence there is strong competition among many selected genotypes. Because several of these genotypes may exist initially, selection can act on them early during the adaptive path (Figure S5). Hence, high-frequency alleles involved in positively epistatic genotypes become fixed rapidly.

Populations evolving on fitness landscapes of intermediate ruggedness follow the pattern of low-ruggedness fitness landscapes during the first steps of the adaptive path. As alleles become fixed, the total additive component in the subset of the fitness landscape in which alleles are still segregating becomes smaller, but the epistatic component remains at its original size. Therefore, epistasis becomes more dominant as the adaptive walk progresses, and fewer minor alleles become fixed towards the end of the adaptive walk. At the last fixation step, the mean initial frequency of the fixed allele for intermediate ruggedness converges to the value for high ruggedness, where fixation is random.

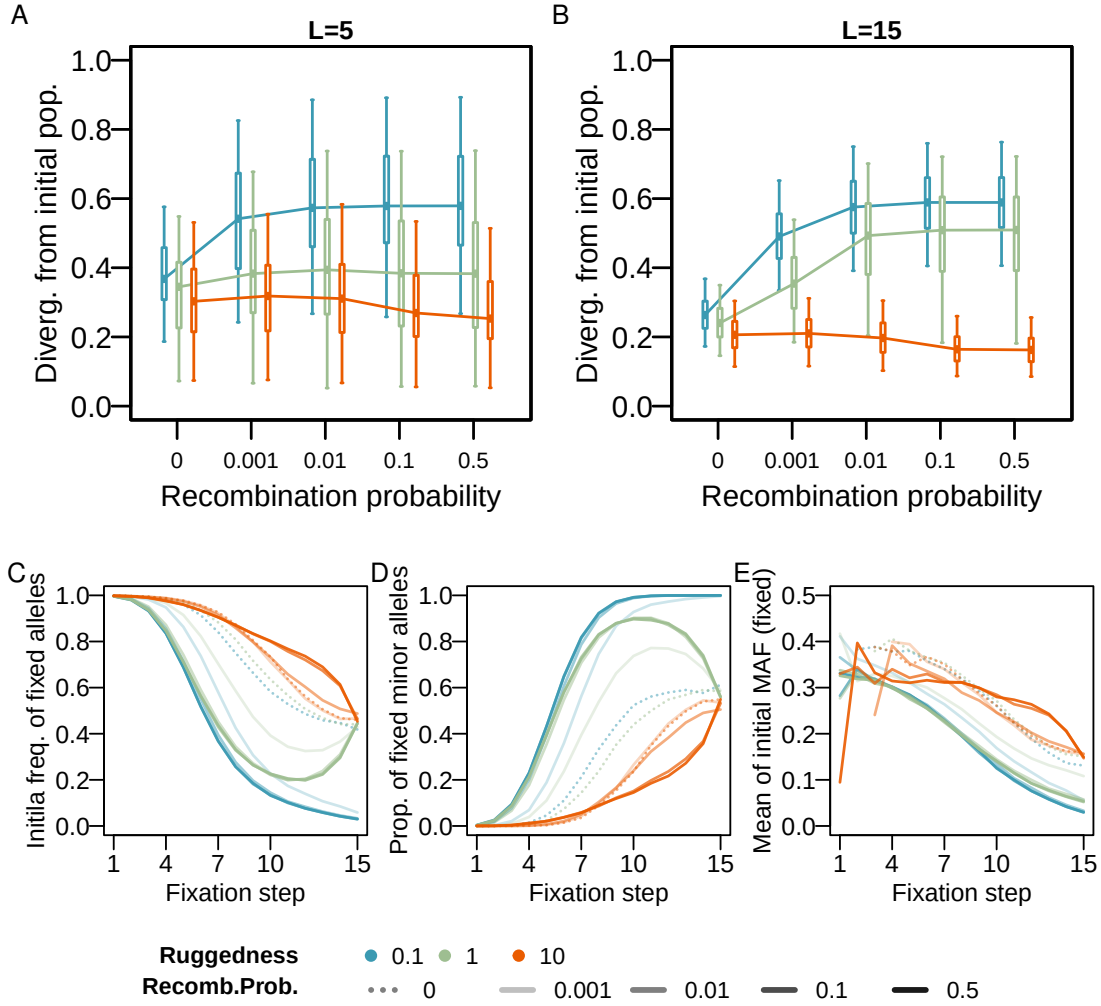


Figure 4. Genetic change between the initial population and the final genotype. To measure the divergence between the initial and the final populations, we calculated the average Hamming distance between the final genotype and the initial population. A. Divergence for small fitness landscape ($L = 5$). B. Divergence for large fitness landscape ($L = 15$). The final genotype is closer to the initial population when both ruggedness and recombination rates are high. C & D show the initial frequencies of the fixed alleles at each fixation step. C. The mean initial frequency of fixed alleles. D. Proportion of fixed alleles which are minor alleles in the initial population. E. The mean initial minor allele frequencies of fixed alleles. More high-frequency alleles are fixed by the stronger interplay of recombination and epistasis in rugged fitness landscapes. All other simulation details are as in Methods and Figure 3.

Misinference of epistasis and selection coefficients along the adaptive path

When adaptation from SGV is observed in nature, rebuilding the complete underlying fitness landscape experimentally is not feasible. Even extracting two-locus fitness landscapes is a large experimental endeavor, but it is more realistic and has been done occasionally. Here, we use our simulation data to infer single-locus selection effects and pairwise epistasis along the observed adaptive paths, thereby showing which estimates would be obtained in experimental tests of pairwise subsets of a fitness landscape. We define as reference genetic background either the final fixed genotype, or the major genotype, i.e. the genotype composed of the major alleles at each locus. We then measure the single-locus effects of the fixed alleles and the pairwise epistasis between two consecutively fixed alleles.

In general, our inference yields poor estimates of both the extent of direct selection and pairwise epistasis. When the landscape is smooth, the inferred one-locus selection coefficient increases from -0.01 to 0.01 with the sequential fixation of alleles (Figure 5A, left panel), independent of which genotype is chosen as the reference. This pattern reflects that initially, genetic drift fixes the major, deleterious allele. Towards the end of the adaptive path, the direct selection coefficient is inferred correctly, except when the recombination rate is low or zero. For smooth fitness landscapes, the low pairwise epistasis in the larger landscape is consistently underestimated (Figure 5B, right panel).

As the epistasis in the underlying larger fitness landscape increases, there appears a large difference between the two potential reference genotypes (Figure 5, S9 and S10A). When the reference genotype is the final genotype, both single-locus effects and pairwise epistasis tend to be overestimated. In contrast, when the major genotype is chosen as the reference, both one-locus effects and pairwise epistasis are underestimated. In general, the size of single-locus effects that is inferred is of the order of the epistasis in the landscape. The amount of pairwise epistasis that is inferred on the basis of the final genotype is in the correct order for intermediate and rugged fitness landscapes, whereas almost no pairwise epistasis is estimated when the major genotype is chosen as a reference.

The recombination rate also influences the obtained estimates. Low recombination rates lower the inferred single-locus effects for smooth and intermediately rugged fitness landscapes. In contrast, they exaggerate the single-locus estimates for very rugged fitness landscapes, when the final genotype is chosen as a reference. Moreover, the estimates vary more between the beginning and the end of the adaptive path at high recombination rates. For example, the estimates of pairwise epistasis, obtained with the final genotype as a reference, are approximately correct for the last pair of fixed mutations for intermediately and strongly epistatic fitness landscapes when the recombination rate is high ($r = 0.1$ or $r = 0.5$), whereas the (incorrect) estimate remains almost constant for lower recombination rates.

We observed similar patterns of misinference for small population sizes and small fitness landscapes ($N = 100$, Figure S9 and S10). Notably, strong genetic drift in small populations leads to generally lower estimates of single-locus selection coefficients and pairwise epistasis, except in very rugged fitness landscapes. Moreover, for fitness landscapes of intermediate and large ruggedness, the variation in estimates from the beginning and the end of the adaptive path, and between reference genotypes, is lower when the fitness landscape is small ($N = 100$, Figure S9 and S10).

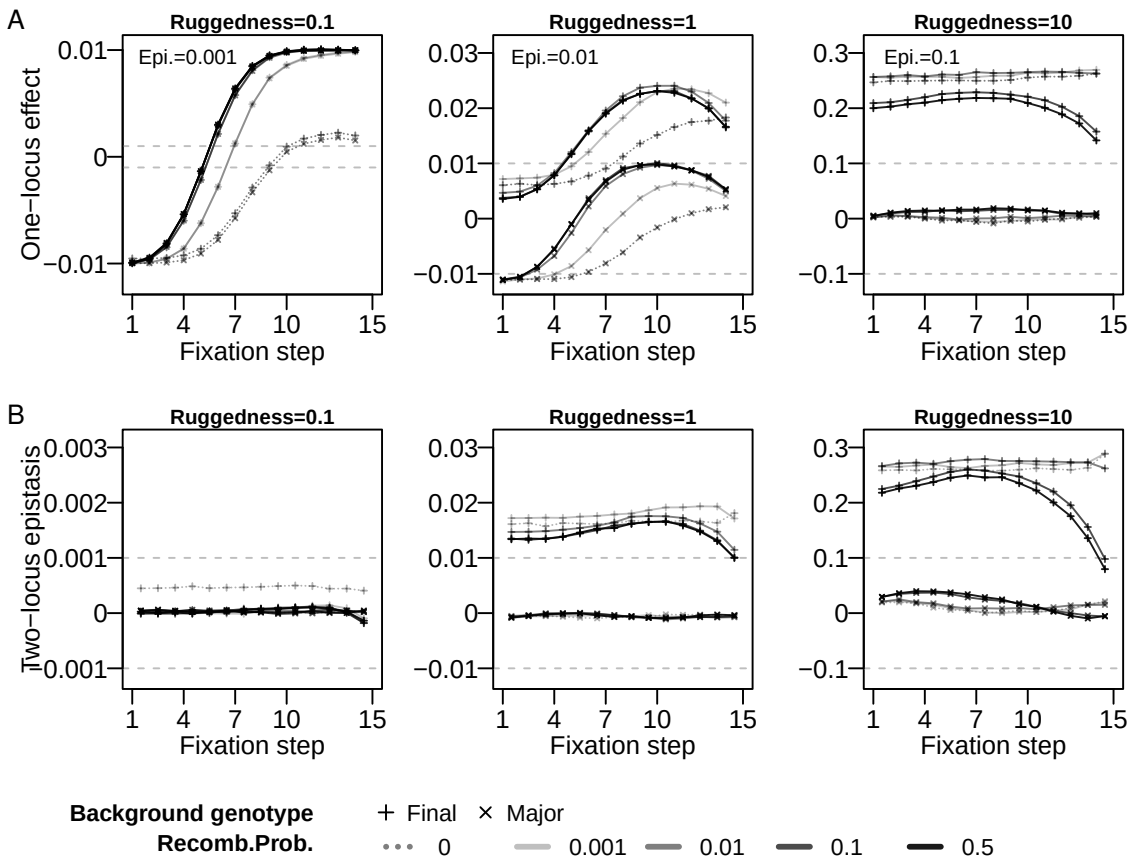


Figure 5. Selection coefficients inferred from the adaptive path. We inferred the one-locus effect (A) and the pairwise epistasis (B) based on two monomorphic genetic backgrounds: the initial major genotype (crosses) and the final fixed genotype (pluses). We inferred the one-locus effect as the selection coefficient of the fixed allele compared to the lost allele at each fixation step, where all other loci are fixed for the focal genetic background. We inferred pairwise epistasis as the epistasis observed from two consecutive fixed alleles on the focal background. Adaptation of polymorphic populations effectively smooths the fitness landscape: one-locus effects are overestimated, and epistasis is underestimated. The ruggedness parameter is indicated above each panel, and its epistatic magnitude (top left in A) is highlighted by dashed lines in each panel. Note the different y-axis ranges between panels. Simulation details are as in Methods and Figure 3.

Adaptation proceeds along a smoother-than-average trajectory on the fitness landscape

We found above that epistasis is misinferred as direct selection when taking the final genotype as a reference. To quantify this effect for an adaptive path, we can compute the ratio of pairwise epistasis and one-locus effects $(\text{pairwise epistasis})/(\text{one-locus effect})$ as a proxy of the roughness of the experienced fitness landscape. Pairwise epistasis is calculated from four genotypes (*e.g.*, $f_{11} - f_{01} - f_{10} + f_{00}$, Figure S2) and the one-locus effect is from two genotypes (*e.g.*, $f_{01} - f_{00}$, Figure S2). Thus, their variances are four times and two times of $Var(e_{\bar{g}})$ in the model, Therefore, we would expect this ratio to be on average $\sqrt{2} \approx 1.4$ for highly rugged fitness landscapes in our study. Contrary to this expectation, the observed ratio for highly rugged fitness landscapes is around 1 for most of the adaptive path, implying that the trajectory is smoother than expected.

Discussion

Probabilistic fitness landscape models are a playground to study the effect of epistasis on adaptation. Here, we study short-term adaptation from standing genetic variation of recombining haploid populations on fitness landscapes of varying ruggedness. We determine realistic initial population compositions and discuss the complex interaction of epistasis and recombination during adaptation. In addition, we show that rugged fitness landscapes could leave a signal of strong direct selection in empirical data.

The role of epistasis in adaptation and its interaction with recombination is a challenging topic, and our study confirms the complex effects of recombination on adaptation in the presence of epistasis.

On additive fitness landscapes, recombination is necessary to bring together beneficial alleles that segregate on different genetic backgrounds. When a locus is under strong direct selection, fixation probability and time are correlated negatively (Kimura and Ohta 1969). For multiple loci, high recombination rates increase the fixation probability and accelerate fixation by combining beneficial alleles in a single genotype (Figure 3A&B, Roze and Barton 2006). At first sight, surprisingly, we find almost no effect of the recombination rate on the time until fixation of a monomorphic genotype (Figure 3A&B). However, this result is explained when taking into consideration the divergence from the initial genotype and the fitness gain during adaption at different recombination rates (Figure 3C&D and 4A&B). When there is no recombination, no new genotypes can be created, and fixation occurs on average at lower fitness and with less average divergence from the initial genotype composition. Therefore, although the adaptive path is “shorter”, the fixation time is similar across recombination rates. Thus, in agreement with the abovementioned classical results, adaptation on smooth fitness landscapes occurs more slowly and less effectively at low recombination rates.

Epistasis is a type of second-order selection that is affected by recombination in a more complex fashion than direct selection. On the one hand, epistasis is only expressed when different allelic combinations segregate simultaneously (i.e., recombination or high mutation rates are required to create the different haplotypes); on the other hand, recombination breaks apart positively epistatic combinations of alleles, thereby reducing the selection effect of epistasis. The positive frequency-dependent nature of epistasis drives alleles at epistatically interacting loci towards fixation (Ayala and Campbell 1974). In our study, we initialize populations at linkage equilibrium. Due to the shape of the SFS, many alleles are initially at very high or low frequencies and many high-fitness genotypes are present initially in highly rugged fitness landscapes. This explains our observation that here, high recombination rates ($r = 0.1$ or $r = 0.5$) are detrimental to fitness gain during adaptation.

Consistent with previous reports of an “optimal” recombination rate for crossing a two-locus fitness valley (Weissman et al. 2010), we observed small-effect but systematic evidence for a minimum time to fixation while maintaining an optimal fitness gain for an intermediate recombination rate in large and highly rugged fitness landscapes ($r = 10$ and $L = 15$; see Figure 3B&D). Similarly, for small and highly rugged fitness landscapes, the same intermediate recombination rate seems to result in a maximum fitness gain while maintaining an almost constant fixation time across all recombination rates ($r = 10$ and $L = 5$; see Figure 3A&C). This pattern is also visible for small intermediately rugged fitness landscapes, in which direct selection and epistasis are of the same order of magnitude. However, in large intermediately rugged fitness landscape, increasing recombination is always beneficial for fitness gains. We

suspect that the added value of recombination, by bringing multiple alleles of additive beneficial effect into the same genotype overrides the intermediate optimum for large fitness landscapes. Here, a smaller proportion of genotypes in the fitness landscape exist in the initial population (and at lower frequencies, which reduces the effectiveness of epistasis).

Notably, the divergence from the initial population always decreases with increasing ruggedness. That is because the number of fitness peaks increases with the amount of ruggedness; with the unpredictability of the fitness landscape, also the (high-fitness) peaks are more evenly distributed across the fitness landscape (Bank 2022). Thus, the probability of high-fitness genotypes being present in the initial population increases with the ruggedness of the fitness landscape. Therefore, a significant fitness increase can often be achieved even without recombination (Figure 3C&D). On the other hand, the fitness increase achieved by populations evolving in large and highly rugged fitness landscapes is on average much lower than when evolution occurred in less rugged fitness landscapes. Thus, populations that evolve in large and highly rugged fitness landscapes are in danger of getting stranded at suboptimal genotypes, especially if there is strong recombination. Previous work studying the evolution of recombining populations in fitness landscapes by means of *de novo* mutations has described a similar pattern: when recombination is strong, populations can get stalled at a suboptimal peak for much longer than with intermediate recombination (de Visser et al. 2009; Nowak et al. 2014).

In highly rugged fitness landscapes, the larger the product of genotype frequency and strength of epistasis the more likely a genotype can become fixed. This is referred to as genotype selection, also known as ”clonal condensation” (Neher et al. 2013). In their work, populations were initialized as a diverse sample (such that each allele was at equal frequency) and subsequently allowed to evolve via selection and recombination. As a consequence, a clonal structure emerged at high recombination rates. In our study, ”clonal condensation” should be more pronounced since a population initiated from a neutral SFS contains genotypes with varying allele frequencies.

Both theoretical and empirical studies have highlighted the crucial role of standing genetic variation for adaptation to environmental challenges (e.g., Barrett and Schluter 2008; Orr and Unckless 2014; Lai et al. 2019; Marques et al. 2019; Bomblies and Peichel 2022). Once adaptation has occurred, it is a challenge to infer how much epistasis and direct selection, respectively, contributed to the adaptive path. This is an important question because it indicates whether adaptation will be repeatable on different genetic backgrounds in other populations exposed to similar environmental challenge. If direct selection is primarily responsible for adaptation, the same alleles will contribute to adaptation repeatedly. If epistasis plays a dominant role, adaptation is contingent on the genetic background and different genotypes will be spreading in different populations. Which scenario is at play is especially important for guided interventions in conservation biology (so-called genetic rescue), because epistasis could render the introduction of genetic material ineffective or even detrimental.

We showed here that inferring the one locus effects of fixed alleles and the pairwise epistasis between successively fixed alleles is not sufficient to quantify the shape of the underlying fitness landscapes. Throughout the parameter range, both selection coefficients and epistasis are misinferred. Specifically, in rugged fitness landscapes and when the final genotype is used as the reference genetic background, we consistently overestimate single-locus effects to be on the order of magnitude of the epistasis parameter, even if the true underlying additive effects are orders of magnitude smaller. Our selection inference is also greatly dependent on

the choice of reference genetic background. When the major genotype (i.e., the genotype with the initially most common alleles at each locus) is used as a reference, epistasis is inferred to be almost zero, and single locus selection is consistently underestimated.

One potential reason for this finding is that in a rugged fitness landscape, adaptive walks take smoother-than-average routes. Previous work has discussed this in various different contexts. For example, subsets of larger fitness landscapes that are comprised of alleles that were fixed along an adaptive walk are on average smoother than subsets comprised of random alleles (Blanquart et al. 2014). Using a rather different approach, regression models are often poor descriptors of the local fitness landscape when samples were chosen among adapted genotypes (Du Plessis et al. 2016).

From an empirical point of view, an important take-home message is that inferring strong direct selection of recently fixed individual alleles could also be a signature of epistasis. Future work could take a genomic approach and test whether genomic signatures of epistasis are distinguishable from those of direct selection. We hypothesize that signatures of soft sweeps are virtually indistinguishable from signatures of epistasis.

In this study, we modeled a haploid recombining population with SGV. We chose this simplification because dominance in diploid populations results in additional complexities, and existing fitness landscape models do not provide a definition of the dominance of additive effects and or the dominance of epistasis (Turelli and Orr 2000). Future work should address this problem by building sensible diploid models of fitness landscapes that include dominance in a parameter-light way. Moreover, the RMF model is neat for its simplicity but results in regular fitness landscapes. Although some experimental fitness landscapes were consistent with an underlying RMF landscape (Szendro et al. 2013; Bank et al. 2016), there is increasing evidence that real fitness landscapes are not as regular as the RMF model suggests (e.g.,Pokusaeva et al. 2019). There are efforts to build fitness landscape models which account for irregularities of interactions (e.Reddy and Desai 2021) but these are not as easily studied and implemented as the RMF model, which requires only a single parameter to vary epistasis.

Acknowledgements

We thank David McLeod, Stephan Peischl and the members from THEE division for suggestions that improved this work. This work was supported by funding from ERC Starting Grant 804569 (FIT2GO), SNSF Project Grant 315230/204838/1 (MiCo4Sys), and HFSP Young Investigator Grant RGY0081/2020 to CB.

References

T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo, and Y. Husimi. Analysis of a local fitness landscape with a model of the rough mt. fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopolymers: Original Research on Biomolecules*, 54(1): 64–79, 2000.

F. J. Ayala and C. A. Campbell. Frequency-dependent selection. *Annual review of Ecology and systematics*, pages 115–138, 1974.

C. Bank. Epistasis and adaptation on fitness landscapes. *Annual Review of Ecology, Evolution, and Systematics*, 53, 2022.

C. Bank, S. Matuszewski, R. T. Hietpas, and J. D. Jensen. On the (un) predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49): 14085–14090, 2016.

R. D. Barrett and D. Schluter. Adaptation from standing genetic variation. *Trends in ecology & evolution*, 23(1):38–44, 2008.

F. Blanquart, G. Achaz, T. Bataillon, and O. Tenaillon. Properties of selected mutations and genotypic landscapes under fisher's geometric model. *Evolution*, 68(12):3537–3554, 2014.

K. Bomblies and C. L. Peichel. Genetics of adaptation. *Proceedings of the National Academy of Sciences*, 119(30):e2122152119, 2022.

R. F. Campbell, P. T. McGrath, and A. B. Paaby. Analysis of epistasis in natural traits using model organisms. *Trends in Genetics*, 34(11):883–898, 2018.

J. F. Crow. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1241–1244, 2010.

J. A. G. de Visser, S.-C. Park, and J. Krug. Exploring the effect of sex on empirical fitness landscapes. *the american naturalist*, 174(S1):S15–S30, 2009.

L. Du Plessis, G. E. Leventhal, and S. Bonhoeffer. How good are statistical models at approximating complex fitness landscapes? *Molecular biology and evolution*, 33(9): 2454–2468, 2016.

C. B. Fenster, L. F. Galloway, and L. Chao. Epistasis and its consequences for the evolution of natural populations. *Trends in Ecology & Evolution*, 12(7):282–286, 1997.

L. Fishman and A. L. Sweigart. When two rights make a wrong: the evolutionary genetics of plant hybrid incompatibilities. *Annual review of plant biology*, 69:707–731, 2018.

J. H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, pages 1116–1129, 1984.

W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3):269–294, 1966.

R. R. Hudson. A new proof of the expected frequency spectrum under the standard neutral model. *Plos one*, 10(7):e0118087, 2015.

M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763, 1969.

Y.-T. Lai, C. K. Yeung, K. E. Omland, E.-L. Pang, Y. Hao, B.-Y. Liao, H.-F. Cao, B.-W. Zhang, C.-F. Yeh, C.-M. Hung, et al. Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences*, 116(6): 2152–2157, 2019.

D. A. Marques, J. I. Meier, and O. Seehausen. A combinatorial view on speciation and adaptive radiation. *Trends in ecology & evolution*, 34(6):531–544, 2019.

R. A. Neher and B. I. Shraiman. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16):6866–6871, 2009.

R. A. Neher, M. Vucelja, M. Mezard, and B. I. Shraiman. Emergence of clones in sexual populations. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01008, 2013.

S. Nowak, J. Neidhart, I. G. Szendro, and J. Krug. Multidimensional epistasis and the transitory advantage of sex. *PLoS computational biology*, 10(9):e1003836, 2014.

H. A. Orr. The population genetics of adaptation: the adaptation of dna sequences. *Evolution*, 56(7):1317–1330, 2002.

H. A. Orr and R. L. Unckless. The population genetics of evolutionary rescue. *PLoS genetics*, 10(8):e1004551, 2014.

V. O. Pokusaeva, D. R. Usmanova, E. V. Putintseva, L. Espinar, K. S. Sarkisyan, A. S. Mishin, N. S. Bogatyreva, D. N. Ivankov, A. V. Akopyan, S. Y. Avvakumov, et al. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS genetics*, 15(4):e1008079, 2019.

G. Reddy and M. M. Desai. Global epistasis emerges from a generic model of a complex trait. *Elife*, 10:e64740, 2021.

D. Roze and N. H. Barton. The hill–robertson effect and the evolution of recombination. *Genetics*, 173(3):1793–1811, 2006.

A. V. Stolyarova, T. V. Neretina, E. A. Zvyagina, A. V. Fedotova, A. S. Kondrashov, and G. A. Bazykin. Complex fitness landscape shapes variation in a hyperpolymorphic species. *Elife*, 11:e76073, 2022.

I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. De Visser. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01005, 2013.

M. Turelli and H. A. Orr. Dominance, epistasis and the genetics of postzygotic isolation. *Genetics*, 154(4):1663–1679, 2000.

E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, and A. Regev. The evolution, evolvability and engineering of gene regulatory dna. *Nature*, 603(7901):455–463, 2022.

D. B. Weissman, M. W. Feldman, and D. S. Fisher. The rate of fitness-valley crossing in sexual populations. *Genetics*, 186(4):1389–1410, 2010.

S. Wright. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences*, 24(7):253–259, 1938.

Supplementary Information

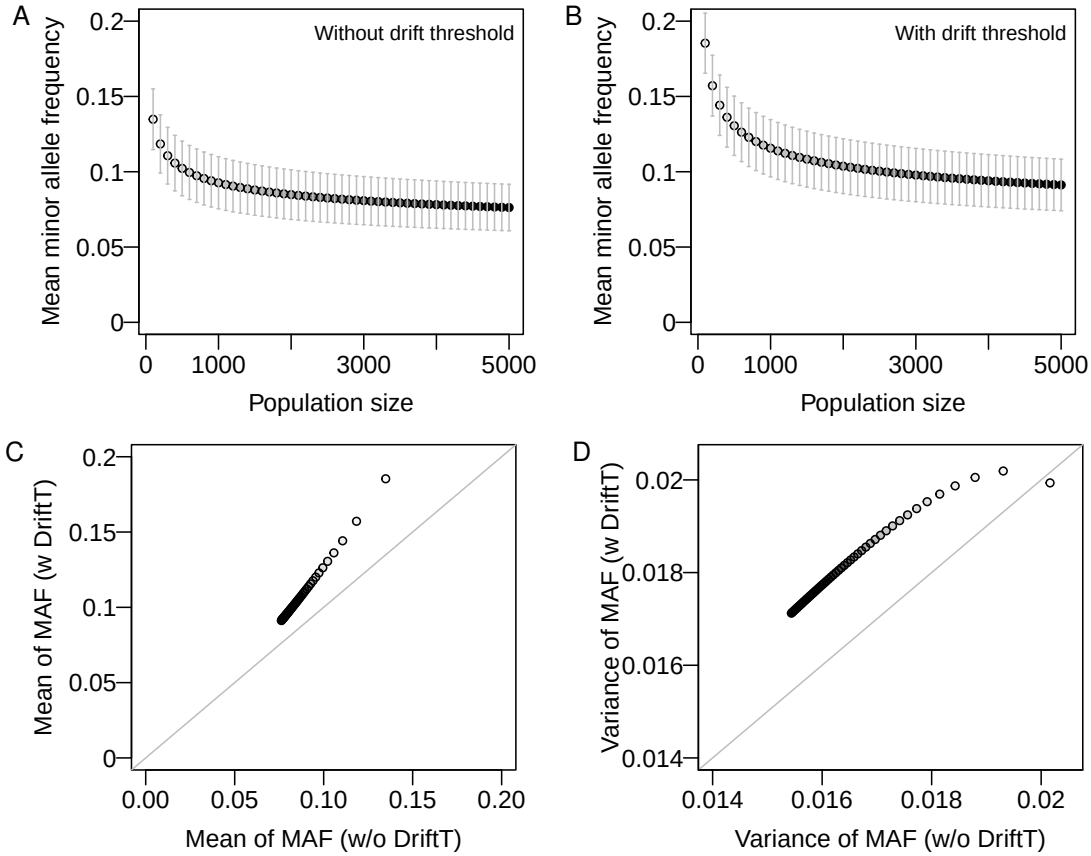
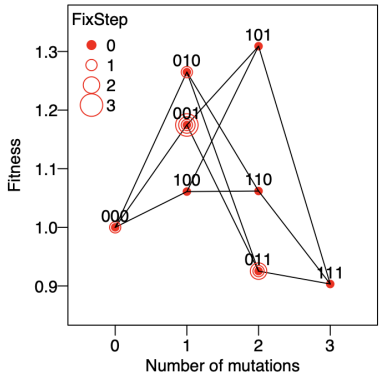


Figure S1. Mean and standard deviation (SD) of minor allele frequencies (MAF) in the initial population. Mean and one SD of minor allele frequencies are plotted by population size (A&B). The site frequency spectrum spans all possible allele frequencies (without genetic drift, A) or is truncated using the genetic drift threshold ($MAF \geq 3$, DriftT). Smaller population sizes show larger means and wider SD. The mean (C) and SD (D) are slightly elevated under the genetic drift threshold, where the population size ranges from 100 to 5000 with a step size of 100, indicated by the transparency of the black color.



Fixation step	1	2	3
Fixation locus	1	3	2
Fixed allele	0	1	0

Genetic backgroud (final genotype): 001

One-locus effect	Consective two-locus effect
$f_{step1} = f_{001} - f_{101}$	$epi_{step\ 1\&2} = f_{001} - (f_{101} + f_{000} - f_{100})$
$f_{step2} = f_{001} - f_{000}$	$epi_{step\ 2\&3} = f_{001} - (f_{011} + f_{000} - f_{010})$
$f_{step3} = f_{001} - f_{011}$	

Figure S2. Example of selection coefficients on an adaptive path. We used a three-locus fitness landscape to demonstrate how we calculate the additive effect (one-locus effect) and the pair-wise epistasis (two-locus effect) on the adaptive path for a given reference genetic background. In this figure, the reference genetic background is the final genotype. The one-locus effect is the selection coefficient of the fixed allele compared to the lost allele (*e.g.*, $f_{step1}, f_{step2}, f_{step3}$); the two-locus effect is the epistatic coefficient of the fixed genotype, composed of two consecutive fixed alleles (*e.g.*, $epi_{step1\&2}, epi_{step2\&3}$). In this figure, each locus contains two alleles, 0 and 1. Genotype (000) is the reference genotype. The final genotype is 001. The alleles in red denote the fixed alleles at each step.

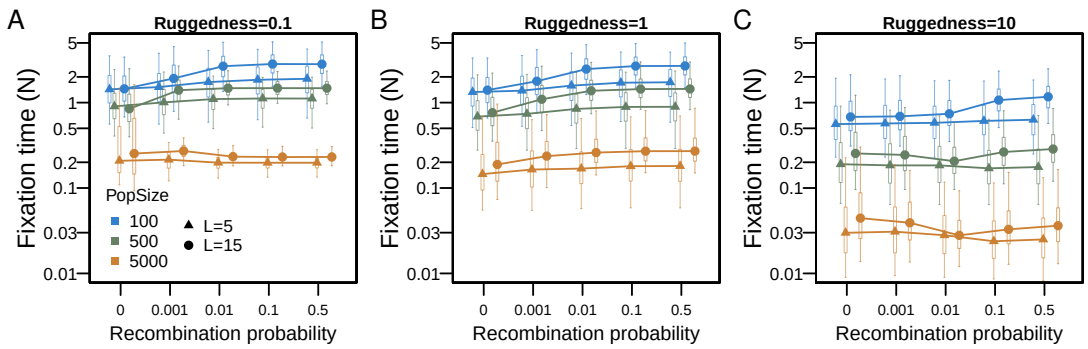


Figure S3. Distribution of fixation times scaled by the population size (N). We consider two sizes of the fitness landscapes, 5 loci ($L = 5$, triangles) and 15 loci ($L = 15$, solid circles). The three population sizes (PopSize) are 100, 500 and 5000, which occupy a small or large fraction of genotype space (see Figure 2). Whereas there is no effect of recombination on fixation time on small fitness landscapes, the fixation time is variable across recombination rates on large fitness landscapes. The additive effect in the underlying RMF fitness landscapes is 0.01. We study fitness landscapes of different ruggedness (0.1, 1, 10) by choosing three standard deviations (σ) of epistasis, 0.001, 0.01, and 0.1.

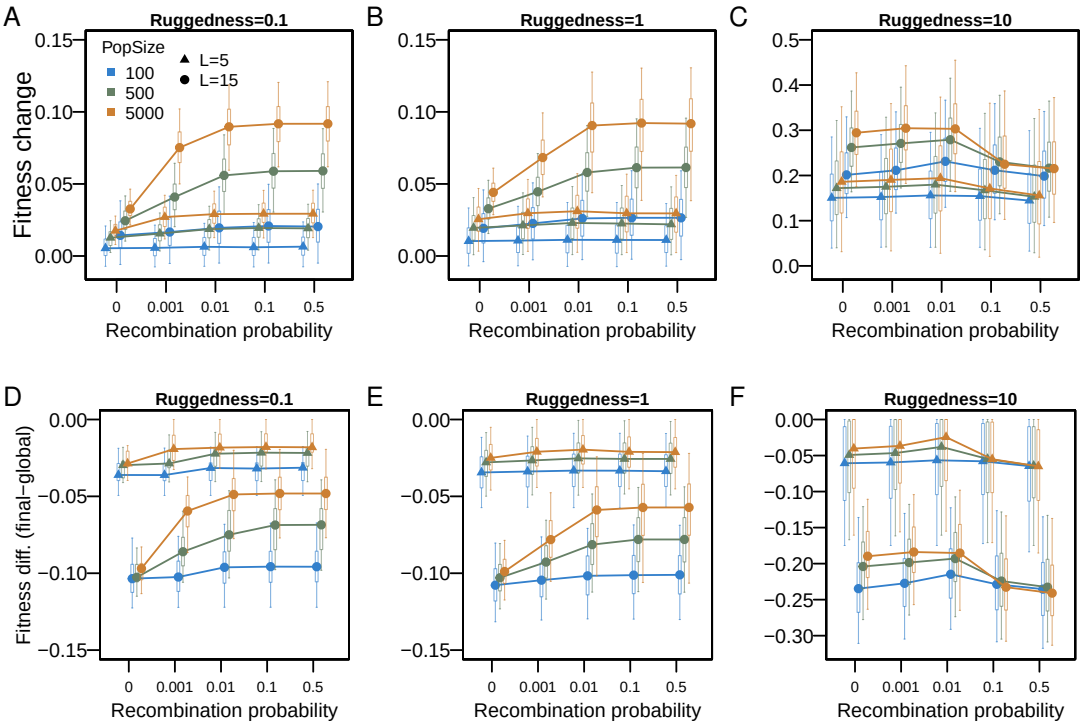


Figure S4. Fitness gain. Relative fitness change between the initial and final populations (fixed genotype) is defined as the absolute fitness change divided by the mean fitness of the initial population (A-C). D-E. Fitness difference between the final genotype and global peak (Fitness diff. (final-global)). Here, fitness difference is defined as the fitness difference between the fixed genotype and the global peak normalized by the fitness of the global peak. Simulation parameters are as in Methods and Figure S3.

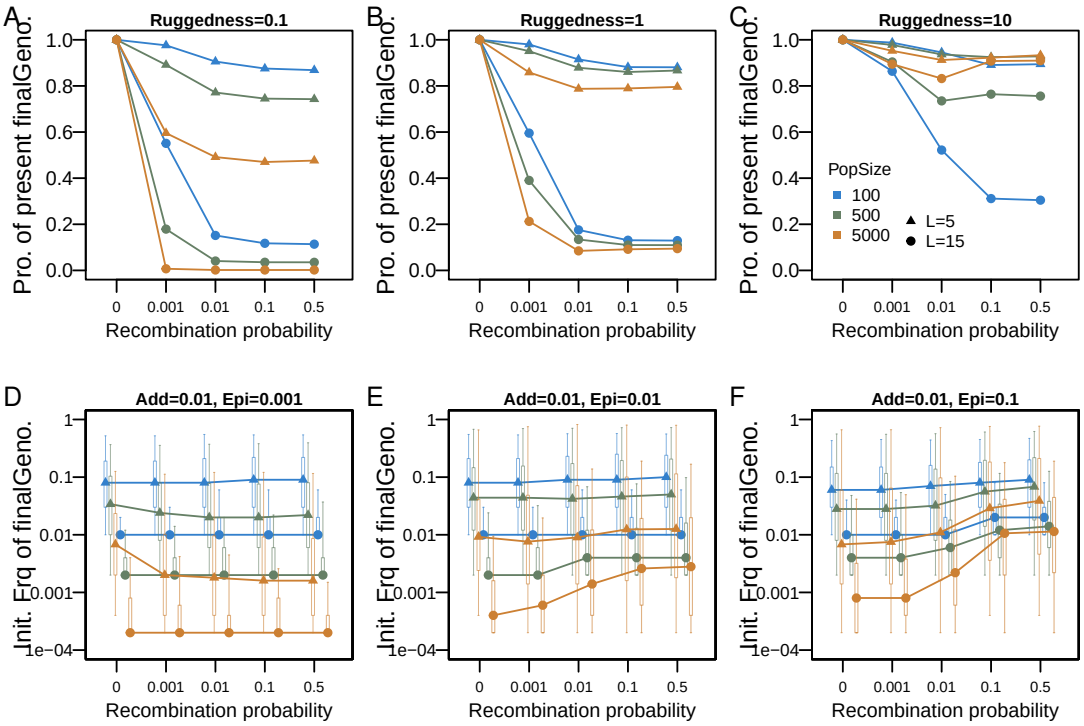


Figure S5. Presence of the final genotype in the initial population. A-C. Proportion of simulations in which the final genotype is present in the initial population. D-F. Distribution of initial frequencies of the final genotype. Here, only genotypes with at least one copy in the initial population are included. Simulation parameters are as in Methods and Figure S3.

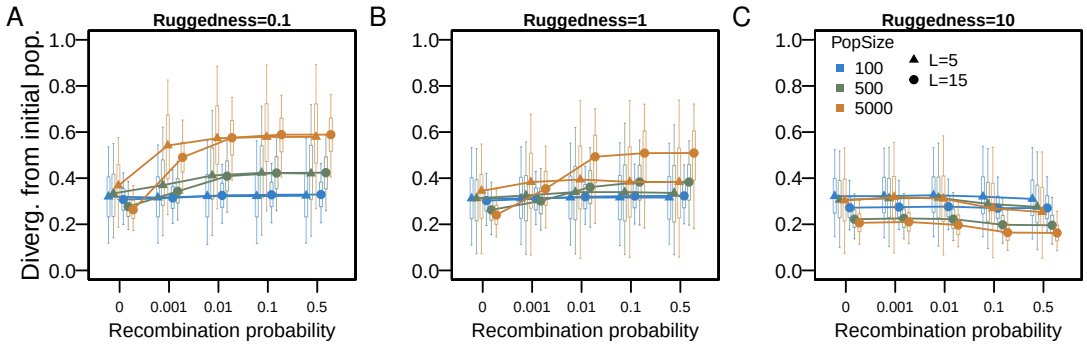


Figure S6. Distribution of the Hamming distance between the final genotype and the initial population. Simulation parameters are as in Methods and Figure S3.

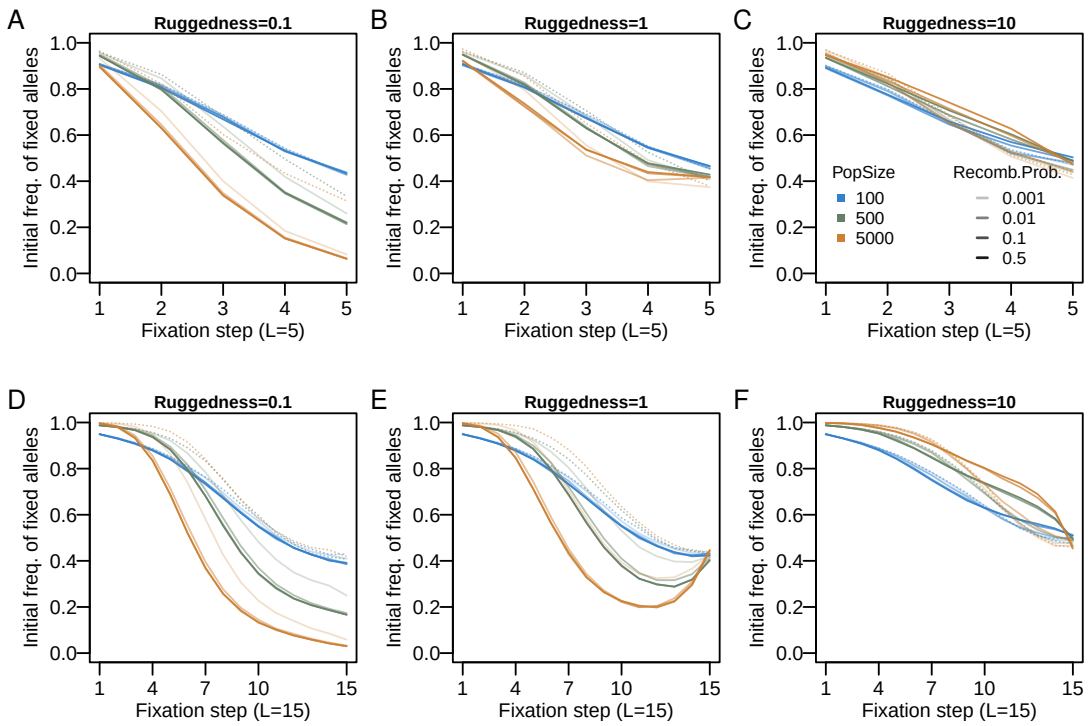


Figure S7. Mean of the initial frequency of fixed alleles on the adaptive path. The fitness landscape contains 5 loci (A-C) and 15 loci (D-F). The dotted line indicates no recombination. Simulation parameters are as in Methods and Figure S3.

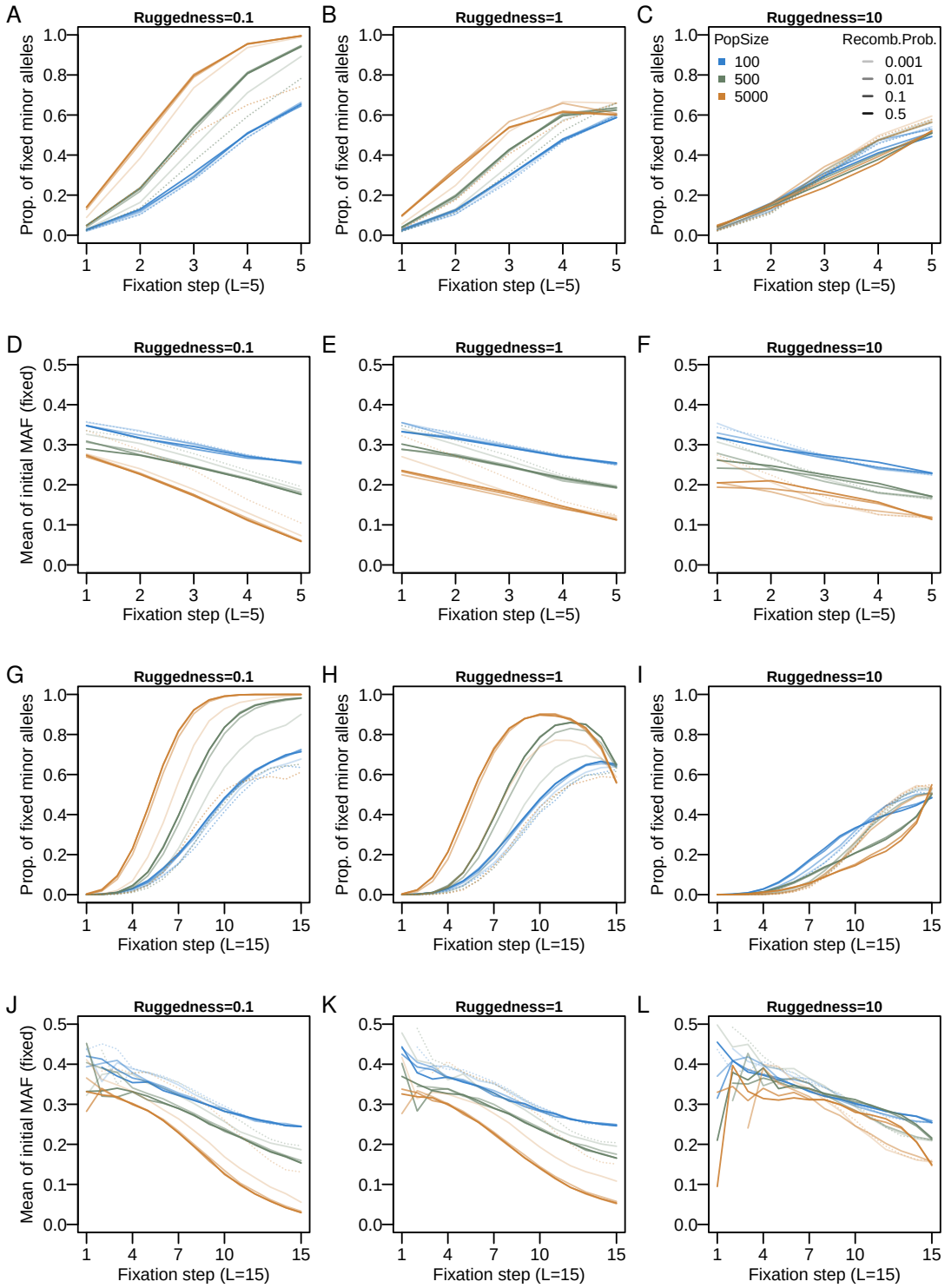


Figure S8. Characterization of fixed minor alleles. The fitness landscape contains 5 (A-F) or 15 loci (G-L). A-C & G-I Mean fraction of fixed minor alleles. D-F & J-L Mean of initial allele frequencies for fixed minor alleles. Fewer minor alleles are fixed at higher ruggedness, and those minor alleles fixed tend to have higher initial frequencies. The dotted line indicates no recombination. Simulation parameters are as in Methods and Figure S3.

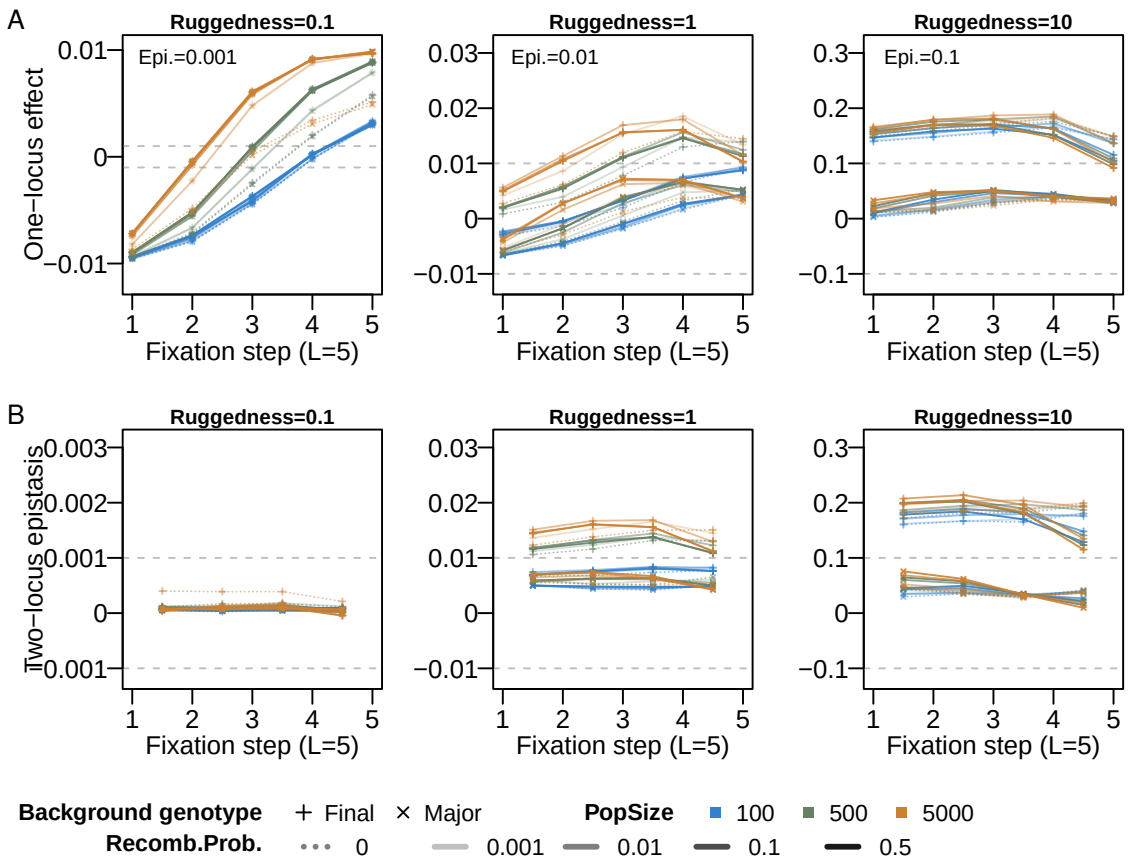


Figure S9. Selection coefficients inferred from the adaptive path. The fitness landscape contains 5 loci. The mean of direct selection coefficients (one-locus effect, A) and two-locus epistasis (B) at each step is obtained from the genetic background using the major (crosses) or final (pluses) genotype. Simulation parameters are as in Methods, Figure 5 and S3.

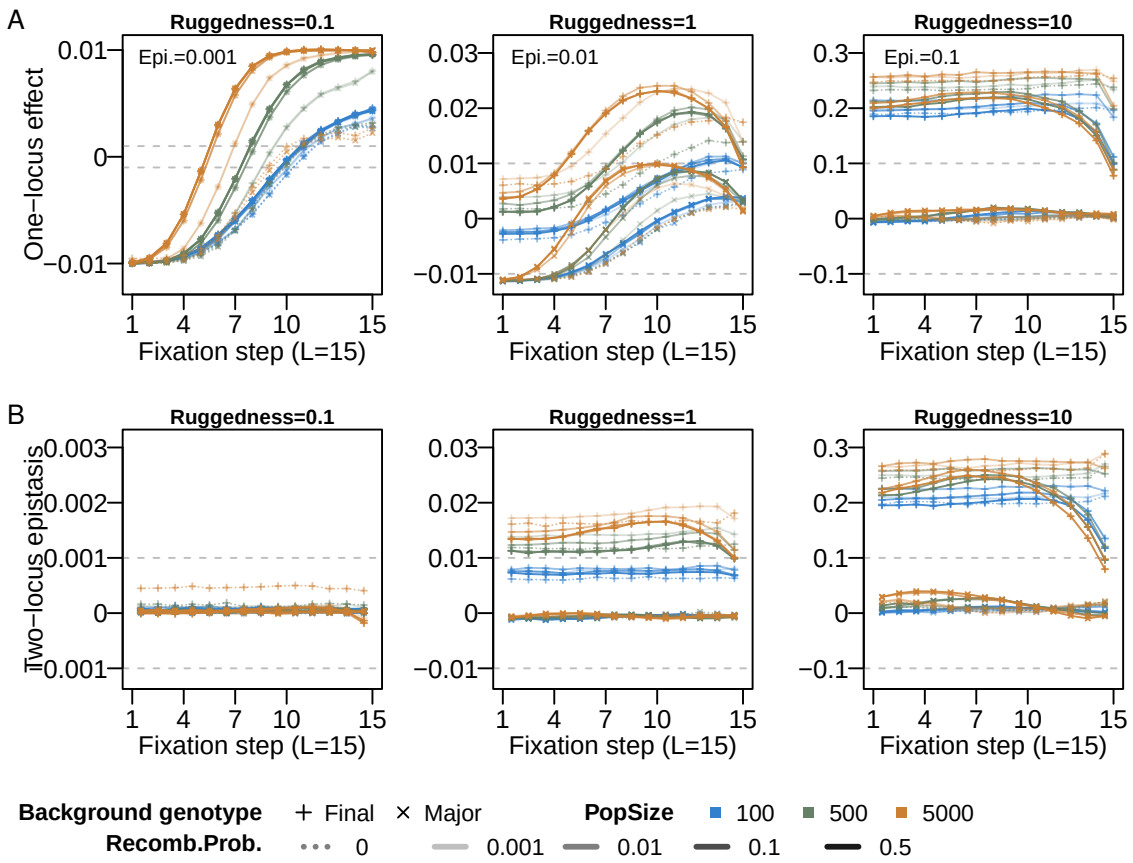


Figure S10. Selection coefficients inferred from the adaptive path. The fitness landscape contains 15 loci. The mean of direct selection coefficients (one-locus effect, A) and two-locus epistasis (B) at each step is obtained from the genetic background using the major (crosses) or final (pluses) genotype. Simulation parameters are as in Methods, Figure 5 and S3.