doi:10.1111/j.1558-5646.2011.01311.x

RECOMBINATION AND HITCHHIKING OF DELETERIOUS ALLELES

Matthew Hartfield[1] and Sarah P. Otto[2,3]

[1]*Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom*

[2]*Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada*

[3]*E-mail: otto@zoology.ubc.ca*

Received December 17, 2010
Accepted March 18, 2011

When new advantageous alleles arise and spread within a population, deleterious alleles at neighboring loci can hitchhike alongside them and spread to fixation in areas of low recombination, introducing a fixed mutation load. We use branching processes and diffusion equations to calculate the probability that a deleterious allele hitchhikes and fixes alongside an advantageous mutant. As expected, the probability of fixation of a deleterious hitchhiker rises with the selective advantage of the sweeping allele and declines with the selective disadvantage of the deleterious hitchhiker. We then use computer simulations of a genome with an infinite number of loci to investigate the increase in load after an advantageous mutant is introduced. We show that the appearance of advantageous alleles on genetic backgrounds loaded with deleterious alleles has two potential effects: it can fix deleterious alleles, and it can facilitate the persistence of recombinant lineages that happen to occur. The latter is expected to reduce the signals of selection in the surrounding region. We consider these results in light of human genetic data to infer how likely it is that such deleterious hitchhikers have occurred in our recent evolutionary past.

KEY WORDS: **Deleterious mutations, diffusion equations, genetic hitchhiking, Hill–Robertson effects, selective sweep.**

The first generation of evolutionary models of advantageous alleles focused on the dynamics of single selected loci in isolation from surrounding sites (Haldane, 1927; Fisher, 1930). Hill and Robertson (1966) demonstrated, however, that selection acting at one site in finite populations interferes with the efficacy of selection at surrounding sites, hampering the spread of neighboring beneficial alleles, even in the absence of fitness interactions among the sites. As pointed out by Hill and Robertson (1966), Charlesworth et al. (1993), and generalized by Rice (1999), selection on linked sites reduces the effective number of lineages contributing to future generations to those lineages with the highest fitness. Such genetic bottlenecks increase the power of drift relative to selection, such that advantageous alleles are less likely to spread and they spread more slowly than predicted by their direct effects on fitness. In a general analysis by Barton (1995), Hill–Robertson interference was shown to reduce the fixation probability of beneficial alleles linked to other selected sites. Breaking down interference among selected loci has also been shown to favor increased rates of sex and recombination (Otto and Barton, 1997; Barton and Otto, 2005; Roze and Barton, 2006).

In addition to affecting neighboring loci under selection, Maynard Smith and Haigh (1974) showed that the dynamics of a single selected locus impacts surrounding neutral loci. In particular, an advantageous allele sweeping through a population reduces, on average, the genetic variance around the site of a sweep (see also Thomson 1977). This phenomenon provides a mechanism for detecting regions experiencing selection, forming the basis for the Hudson-Kreitman-Aguade (HKA) test (Hudson et al., 1987), for example.

Relatively little attention has been paid, however, to the effect that selection on neighboring sites might have on the net fitness change associated with the fixation of a focal beneficial allele and

on the patterns of variation at surrounding selected sites (see, e.g., Yu and Etheridge (2010) regarding beneficial alleles segregating in the background, and Hadany and Feldman (2005) regarding deleterious alleles in the background). In this article, we consider a focal site carrying a new beneficial allele in the presence of neighboring sites subject to deleterious mutations. We calculate the chance that a linked deleterious allele hitchhikes to fixation along with the beneficial allele, as a function of the rate of recombination between them, and describe the implications for patterns of variation expected within the region of a selective sweep. This work builds upon a recent simulation study by Hadany and Feldman (2005), as well as complementary analytical work for asexual organisms (Johnson and Barton, 2002; Bachtrog and Gordo, 2004; Yu and Etheridge, 2008; Yu et al., 2010). Specifically, Hadany and Feldman (2005) demonstrated that beneficial alleles sweeping to fixation in a purely asexual population often carry along linked deleterious alleles. The fixation of deleterious alleles by hitchhiking generates a fixed mutation load that must await a future adaptive sweep by a back or compensatory mutation in order for it to be erased. Our work provides an analytical prediction of the probability of such undesirable hitchhikers, allowing for arbitrary rates of recombination between the sites under selection.

Empirical Background—Recent studies of amino-acid substitution data suggest that advantageous mutants are present at higher rates then previously assumed. Although precise values remain a matter of debate (Eyre-Walker, 2006), Bierne and Eyre-Walker (2004) estimated that approximately 45% of amino acid substitutions are adaptive in *Drosophila melanogaster*, equating to one substitution, on average, every 450 generations. Later studies have found that between 30 and 60% of substitutions in *D. melanogaster* coding and noncoding regions are adaptive (Andolfatto, 2005; Obbard et al., 2009; Andolfatto, 2007; Shapiro et al., 2007), highlighting the prevalence of beneficial mutation. Similar values have been observed in the wild mouse *Mus musculus castaneus* (Halligan et al., 2010). In hominids this rate tends to be lower; Boyko et al. (2008) and Eyre-Walker and Keightley (2009) found that on average 5% of amino-acid substitutions were adaptive if recent population bottlenecks were taken into account.

Another method to detect the presence of advantageous mutations is through investigating the underlying distribution of fitness effects among mutations. Using such a method Shaw et al. (2002) suggested that half of all mutations in *Arabidopsis thaliana* increased fitness (although see Keightley and Lynch (2003)). Even in fairly laboratory-adapted strains of *Saccharomyces cerevisiae*, Joseph and Hall (2004) estimated that around 6% of spontaneous mutations were beneficial (see also Hall and Joseph, 2010).

The strength of selection acting on beneficial alleles is also subject to much debate and is expected to depend on the nature of past environmental changes, both biotic and abiotic (Elena and Lenski, 2003). On the lower end, Jensen et al. (2008) estimated that advantageous mutants have had a mean selection coefficient of $s_a \approx 10^{-4}$ in *Drosophila*. On the upper end, very large selection coefficients have been detected in experimental evolution studies with bacteria and viruses, with an average $s_a \approx 2$ found in *Pseudomonas fluorescens* exposed to a novel carbon source (Barrett et al., 2006) and s_a ranging between 6 and 14 in the bacteriophage ϕX174 subjected to heat stress (Bull et al., 2000).

Although there is increasing evidence for the frequent spread of advantageous alleles, it is an inescapable fact that most spontaneous mutations that affect fitness are deleterious (Crow, 1970) and are maintained in populations at a low frequency by recurrent mutation (Wright, 1931). These mutation rates can be substantial; for example, the per-generation genomic deleterious mutation rate U_d in *Drosophila* has been estimated at 1.2 (Haag-Liautard et al., 2007; Keightley et al., 2009), with estimated rates of U_d of around 4.2 in hominids (Eöry et al., 2010). Deleterious mutation rates are lower in microbes, however. In nonmutator strains of yeast, Hill and Otto (2007) estimated $U_d = 0.013$ for mutations acting on sporulation ability and $U_d = 0.0003$ for those affecting growth rate.

If selection acts against deleterious mutations with a coefficient of s_d, then we would expect a total of $\sim U_d/s_d$ mutations to segregate within a population at mutation–selection balance (ignoring genetic associations among them). Even when U_d is less than one, the expected number of deleterious mutations carried by an individual may be much greater than one. Consequently, newly arisen advantageous alleles may occur within chromosomes also bearing deleterious alleles nearby. In the next section, we develop a model that describes the fate of a deleterious mutation that occurs in the genetic background of a novel beneficial allele. We later return to estimates of mutation rates and selection coefficients to assess how likely it is that deleterious alleles hitchhike to fixation, and how this depends on the mode of reproduction and the effective rate of recombination within a species.

Semi-Deterministic Model

We first present a semi-deterministic calculation of the fixation probability of a haplotype carrying both an advantageous and a deleterious allele using classic population genetics. In the next section, we build a stochastic diffusion model of the appearance and spread of this haplotype, but the calculations presented in this section help to develop an understanding of the key forces at work and so are a natural first step in investigating this problem.

We consider a finite population of N haploid chromosomes with discrete generations, using a standard Wright–Fisher model (Fisher, 1930; Wright, 1931). We are interested in the dynamics of a newly arisen beneficial allele at a locus A. The genome in which A first arises may carry one or more deleterious alleles. Deleterious alleles that are only loosely linked to locus A are unlikely to rise

Table 1. Table of haplotypes.

Haplotype	Fitness, w	Locus 1	Locus 2
$A_0\,B_0$	1	Wild type	Wild type
$A_1\,B_0$	$1 + s_a$	Beneficial	Wild type
$A_0\,B_1$	$1 - s_d$	Wild type	Deleterious
$A_1\,B_1$	$1 + s_a - s_d$	Beneficial	Deleterious

substantially in frequency and are ignored. We focus only on the single most closely linked of these deleterious mutations and call this second locus B, with recombination between A and B occurring at rate r. At locus A, the advantageous allele A_1 has a selective advantage s_a over the wild-type allele A_0. At locus B, the deleterious allele B_1 is selected against with selection coefficient s_d, relative to the wild-type allele B_0. We assume $s_a > s_d$, so that the advantageous-deleterious haplotype has a net beneficial effect, $s_{net} = s_a - s_d$. For clarity of presentation, we assume additive selection, but all of our analytical results continue to apply if s_d is replaced by $s_a - s_{net}$, wherever it occurs.

For each haplotype, we write a 0 subscript if the wild-type allele is present at the locus and a 1 subscript otherwise, in the order AB. All possible haplotypes, along with their fitness, are given in Table 1. In particular, the advantageous-deleterious haplotype is denoted A_1B_1, and when this haplotype first appears, the remainder of the population is either A_0B_0 (wild type) or A_0B_1 (bearing the deleterious allele). The latter haplotype (A_0B_1) is assumed to be rare and is ignored in the following analysis to simplify the calculations; simulations described in a later section indicate that this assumption introduces little bias. We also assume that no further mutation occurs at either of the loci during the course of the sweep, although the model can be modified to take this into account.

Let $p(t)$ denote the frequency of the A_1B_1 haplotype, where t is the number of generations since the beneficial allele arose and p_0 is its initial frequency (generally $1/N$). When the A_1B_1 haplotype first arises, it becomes established within the population with a probability u that is approximately twice the net selection coefficient, $2s_{net}$ (Haldane, 1927). It is further assumed that $s_{net} \ll 1$ and that the population size is large (see next section for results that apply in smaller populations).

In the following derivation, we only consider those A_1 alleles that survive stochastic loss while rare. Once established, the frequency of A_1B_1 can be modeled by the standard deterministic equation for haploid selection (Haldane, 1924):

$$p(t) = \frac{p_0(1 + s_{net})^t}{p_0(1 + s_{net})^t + 1 - p_0}. \tag{1}$$

Among those alleles that succeed in fixing, the trajectory of the A_1B_1 haplotype is slightly faster, on average, than given by equation (1) (Maynard Smith and Haigh, 1974; Barton, 1994). This initial acceleration is taken into account in the diffusion model developed below; it turns out to have little effect, however, because rare recombination events that break apart the A_1B_1 haplotype are most likely to occur when the A_1B_1 haplotype is intermediate in frequency and not when it initially occurs.

Our goal is to calculate the probability, P, that the A_1B_1 haplotype is not broken apart by recombination before the advantageous A_1 allele fixes within the population. If such a recombination event has not yet occurred, there are approximately $p(t)$ of the A_1B_1 haplotypes and $1 - p(t)$ of the A_0B_0 haplotypes (ignoring the rare A_0B_1 individuals), so that matings between these two haplotypes occur at frequency $2p(t)(1 - p(t))$. Among the offspring of these matings, r will be recombinant, half of which will carry the most fit A_1B_0 haplotype and half of which will carry the least-fit A_0B_1 haplotype. Even once produced, the most fit recombinant may fail to establish itself within the population due to chance loss while rare. In Appendix A, we use branching processes to show that the probability that a single new A_1B_0 haplotype establishes within the population if it appears at time t equals

$$\Pi(t) = \frac{2s_a s_d}{s_a p(t) + s_d(1 - p(t))} + O(s^2). \tag{2}$$

The derivation of equation (2) accounts for the fact that the A_1B_0 haplotype has fitness $1 + s_a$ relative to the population mean fitness $1 + p(t)(s_a - s_d)$, which is changing over time according to equation (1). As expected, if the A_1B_0 recombinant haplotype arises while $p(t) \approx 0$, the recombinant lineage will establish with probability nearly equal to $2s_a$, the fixation probability of an advantageous A_1 allele in an otherwise wild-type population. Also as expected, if the A_1B_0 recombinant haplotype arises while $p(t) \approx 1$, the recombinant lineage will establish with probability nearly equal to $2s_d$, the fixation probability of a haplotype that has shed the deleterious allele B_1 in a population that otherwise carries both A_1 and B_1. We call A_1B_0 haplotypes that succeed in establishing while rare "successful recombinants."

Altogether, $\kappa(t) = rp(t)(1 - p(t))\Pi(t)$ is the probability that an A_1B_0 recombinant haplotype appears at time t and goes on to establish within the population. Note however that this calculation does not specify whether the A_1 or B_0 allele will fix first; in many cases, if a recombinant appears and fixes with probability $\Pi(t)$, the actual fixation of the A_1B_0 haplotype would occur after A_1 has reached fixation.

To calculate the overall probability, P, that the A_1B_1 haplotype is never broken apart by recombination, we must calculate the probability that in every generation, t, none of the N offspring are successful recombinants. Assuming weak selection such that both $\Pi(t)$ and $\kappa(t)$ are small, the probability that a deleterious hitchhiker will be carried to fixation by the spread of a linked

beneficial allele is given by

$$\begin{aligned} P &= \prod_{t=0}^{\infty}(1-\kappa(t))^N \\ &\approx \prod_{t=0}^{\infty} exp[-N\kappa(t)] \\ &= exp\left[\sum_{t=0}^{\infty} -N\kappa(t)\right] \\ &\approx exp\left[\int_{t=0}^{\infty} -N\kappa(t)dt\right]. \end{aligned} \tag{3}$$

Overall, P gives the probability that a fitter recombinant never establishes, assuming that the A_1B_1 haplotype is not lost stochastically when it first appears. The probability that the A_1B_1 haplotype succeeds in establishing initially and fixing within the population is thus $u\ (=2s_{net})$ times P. This equation is analogous to equation (16) in Yu and Etheridge (2010), who used a Moran model to estimate the fixation probability of two competing beneficial mutations, with recombination between the two loci.

Equation (3) can be solved by integrating over the allele frequency dynamics rather than over time and replacing the integral with

$$\int_{p=p_0}^{1} -\frac{N\ \kappa(p)}{dp/dt}\,dp. \tag{4}$$

In this haploid model with weak selection, $dp/dt = (s_a - s_d)\ p(1-p)$. Carrying out the integration, the probability that a fitter recombinant never establishes is given by

$$P \approx exp\left[-\frac{2\ N\ r\ s_a\ s_d\ \ln(s_a/s_d)}{(s_a-s_d)^2}\right],$$

where p_0 was assumed negligible relative to terms on the order of one. At this point, we can eliminate the population size from the result by measuring the net selection and recombination rates within the population, defined as $S_d = Ns_d$, $S_a = Ns_a$, $S_{net} = N(s_a - s_d)$, and $\rho = Nr$, yielding

$$P \approx \left(\frac{S_a}{S_d}\right)^{-\omega}, \tag{5}$$

where ω is the compound parameter defined by

$$\omega = 2\ \rho\ \frac{S_a\ S_d}{S_{net}^2}. \tag{6}$$

The hitchhiking process thus depends primarily on these scaled parameters and not separately on the population size and selection or recombination parameters. The above equations show that the probability of hitchhiking to fixation declines exponentially with the recombination rate between the loci and with the number of individuals within the population. The probability of hitchhiking is especially small when the strength of selection for the beneficial allele and against the deleterious allele is similar (S_{net} small), as this will cause the sweep of the A_1B_1 haplotype to take longer and allow for more recombination events.

To determine how small the recombination rate must be in order for hitchhiking to occur with a particular probability of interest, c, we set $P = c$ and solve for ρ:

$$\rho_{crit} = \frac{S_{net}}{S_d}\left[\frac{\ln\left(\frac{1}{c}\right)}{2\left(1+\frac{S_d}{S_{net}}\right)\ln\left(1+\frac{S_{net}}{S_d}\right)}\right]. \tag{7}$$

This gives us the recombination rate below which hitchhiking to fixation will occur with frequency greater than c, as a function only of the scaled selection coefficients S_d and S_{net}. At this point, we hold off discussing these results further until the next section, where we derive a stochastic solution.

Stochastic Model

The above analysis assumes that the population is very large, allowing us to combine stochastic results for the establishment of particular haplotypes while rare, with deterministic equations for the spread of these haplotypes. The above does not, however, take into account chance fluctuations in haplotype frequencies or the initial acceleration caused by considering only those trajectories where the beneficial allele becomes established (Maynard Smith and Haigh, 1974; Barton, 1994; Otto and Barton, 1997; Desai and Fisher, 2007). To account for these effects, we now derive a stochastic solution for this problem.

Again ignoring the rare deleterious-only lineage, we model the change in frequency, $p(t)$, of the A_1B_1 haplotype using a diffusion approximation. If a successful recombinant appears, however, the diffusion process is killed. As described by Karlin and Taylor (1981), the probability that the process is not ultimately killed, $P(p)$, given that A_1B_1 is currently at frequency p, satisfies

$$\frac{1}{2}V(p)\frac{d^2P(p)}{dp^2} + M(p)\frac{dP(p)}{dp} - K(p)P(p) = 0, \tag{8}$$

where $M(p)$ is the mean change in p over a time step measured in N generations; $V(p)$ is the variance in change of p; and $K(p)$ is the killing function, which denotes the probability of the process being "killed" while the A_1B_1 haplotype is at frequency p. In this model, killing occurs if recombination forms a fitter haplotype (i.e., A_1B_0) that succeeds in establishing within the population. To solve equation (8), we use the boundary conditions $P(0) = P(1) = 1$; that is, the system cannot be killed if the A_1B_1 or A_0B_0 haplotype is fixed. Further descriptions of similar diffusion models with killing are available in Karlin et al. (1967) and section 15.10 of Karlin and Taylor (1981); in particular, a related model is described where the diffusion process is killed whenever any recombinant is formed (A_1B_0 or A_0B_1), regardless of whether the recombinant succeeds in establishing.

As with standard diffusion models investigating an allele under weak directional selection in a haploid population (Kimura, 1970; Ewens, 2004), we obtain the values $M(p) = S_{net}\,p(1-p)$ and $V(p) = p(1-p)$, where $S_{net} = N(s_a - s_d)$ (see section 2 of Appendix S3). The killing term is obtained by taking the probability that the process is killed in a particular generation, $1-(1-\kappa)^N \approx N\kappa = N\,r\,p(1-p)\,\Pi$, and scaling in such a way that the killing term remains finite over the time step of N generations, as $N \to \infty$ (Karlin and Taylor, 1981). By doing so, we obtain the killing function $K(p) = \rho\,p(1-p)\,\pi(p)$, where $\rho = Nr$ and $\pi(p)$ is the scaled version of the establishment probability of the A_1B_0 recombinant, Π (eq. 2)

$$\pi(p) = \frac{2\,S_d\,(S_{net}+S_d)}{p\,S_{net}+S_d}. \tag{9}$$

The diffusion approximation assumes that S_{net}, S_d, and ρ remain finite as $N \to \infty$.

Plugging these diffusion coefficients into equation (8) and dividing by $p(1-p)$, the probability that the process is not killed, $P(p)$, given the current frequency p satisfies

$$\frac{1}{2}\frac{d^2P(p)}{dp^2} + S_{net}\frac{dP(p)}{dp} - \rho\,\pi(p)\,P(p) = 0. \tag{10}$$

If the process is not killed, there are two potential outcomes: fixation of A_0B_0 or fixation of A_1B_1. If we wish to know the probability that a particular advantageous allele that succeeds in fixing carries along with it a deleterious allele, we must rederive the diffusion model conditional on A_1 establishing within the population. In Appendix B, we show that the conditional probability $P^*(p)$ that the process is not killed (i.e., the deleterious allele B_1 fixes) among those cases where A_1 sweeps to fixation satisfies:

$$\frac{1}{2}\frac{d^2P^*(p)}{dp^2} + S_{net}\,\frac{1+e^{-2pS_{net}}}{1-e^{-2pS_{net}}}\frac{dP^*(p)}{dp} - \rho\,\pi(p)\,P^*(p) = 0. \tag{11}$$

The differential equations (10) and (11) were solved in *Mathematica* 6.0 (Supporting information), yielding the somewhat cumbersome equations (B5) and equation (B6), respectively. These can be solved numerically for the probability that the process is not ultimately killed (i.e., the probability that a successful recombinant never appears).

$P^*(p_0)$ as given by (B6) is the main quantity of interest in this article. It describes the probability that an A_1 allele that fixes within a population carries along with it a linked deleterious allele B_1, given that the initial frequency of the A_1B_1 haplotype is p_0. Although equations (B5) and (B6) should be used in any numerical analysis, further insight is provided by approximating $P*(p_0)$ as an exponentially decreasing function of the recombination rate (as inferred in the semi-deterministic analysis). Assuming that selection is strong relative to drift ($S_d, S_{net} \gg 1$), that the frequency of the A_1B_1 haplotype when the A_1 allele first appears is negligibly small ($p_0 \ll 1$), and that recombination is not too frequent ($\rho \ll S_d, S_{net}$), we obtain:

$$P^*(p) \approx \left(e^{-1/S_d}\,\frac{S_a}{S_d}\right)^{-\omega} \tag{12}$$

(see details in section 3 of Appendix S3). Again, this can be used to calculate a critical value of recombination above which hitchhiking is unlikely to occur. Specifically, we solve equation (12) for the rate of recombination necessary for the deleterious B_1 allele to fix with probability c, given that the beneficial allele A_1 initially appears with B_1 and ultimately fixes

$$\rho_{crit} = \frac{S_{net}}{S_d}\left[\frac{\ln\left(\frac{1}{c}\right)}{2\left(1+\frac{S_d}{S_{net}}\right)\left(\ln\left(1+\frac{S_{net}}{S_d}\right)-1/S_d\right)}\right]. \tag{13}$$

For example, when $c = 1/2$, the term in square brackets is approximately 1/4 as long as neither S_d nor S_{net} is too small (see the figure in section 3 of the Appendix S3). Thus, as a rough rule of thumb (using unscaled parameters), the recombination rate r must be less that 1/4 of $s_{net}/(Ns_d)$ for there to be at least a 50% chance that the deleterious allele hitchhikes to fixation.

Hitchhiking events are thus likely to occur over larger regions of the genome if the net selection coefficient acting on the A_1B_1 haplotype, s_{net}, is stronger because sweeps occur faster. Conversely, the stronger the disadvantage of the deleterious allele, s_d, the less likely a hitchhiker will fix because recombinant A_1B_0 haplotypes are so much more fit. Finally, the larger the population size, the less likely that a hitchhiker will fix, simply because there are more individual chances for recombination to occur while the population remains polymorphic.

These patterns are illustrated in Figure 1, which gives the probability that the deleterious B_1 allele hitchhikes to fixation given that the beneficial A_1 allele fixes, with darker shading corresponding to higher probabilities. These contour plots are based on the exact solution (B6) to the diffusion equation for $P^*(p)$. The thick dashed curves show the approximate equation (13) for the critical value of the recombination rate, ρ, below which we expect deleterious alleles to hitchhike to fixation more than c of the time ($c = 10\%$, 50%, or 90%) when they occur on the haplotype bearing a new beneficial allele; these curves accurately follow the appropriate contour lines as long as selection is not too weak (roughly, $S_{net}, S_d \geq 2$).

COMPARISON TO THE CASE OF A LINKED NEUTRAL ALLELE

The dynamics of neutral loci are likely to be affected by the spread nearby of a beneficial allele whenever r is approximately less than s_a (Maynard Smith and Haigh, 1974). This rule cannot be used to compare to equation (13) directly, however, because our criteria for being "affected" is now quite strict: the linked B_1 allele must fix due to the sweep. We thus briefly describe a corresponding

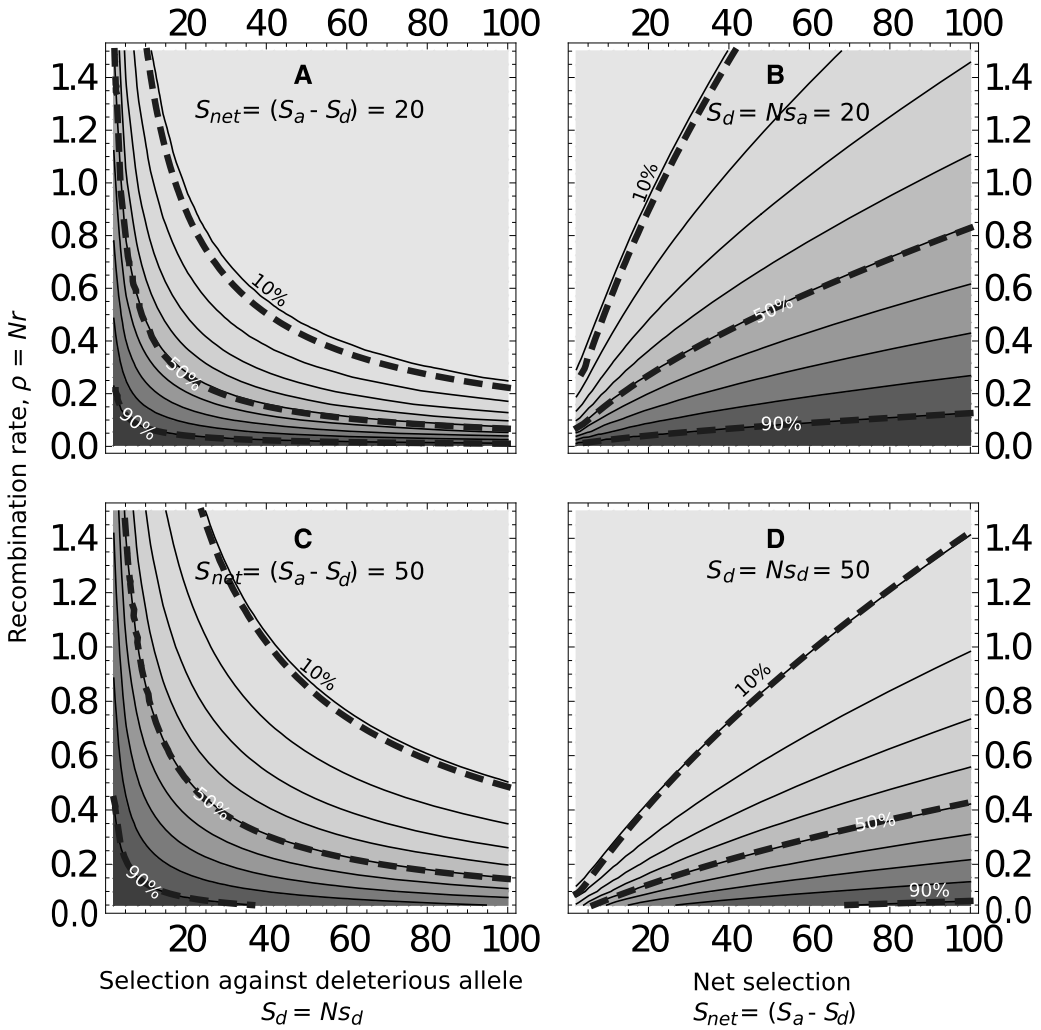


Figure 1. Contour plots of the fixation probability of the deleterious B_1 allele, given that the A_1B_1 haplotype appears initially at frequency of $1/N$ and that the A_1 allele is not lost stochastically (10% contour intervals based on equation B6). The graphs are shown for N = 10,000, although the results are not very sensitive to N, as long as the scaled parameters are held constant. In each case, ρ is plotted along the y-axis versus S_d along the x-axis (left panels) with S_{net} = 20 (top) or 50 (bottom) or versus S_{net} along the x-axis (right panels) with S_d = 20 (top) or 50 (bottom). The dashed curves show the predicted thresholds below which there is a greater than c = 10%, 50%, and 90% probability of hitchhiking, based on equation 13; in each case this threshold coincides closely with the appropriate contours.

model for the case when B is neutral (full details are provided in the section 4 of Appendix S3).

The diffusion equations remain essentially the same, except that the killing term must be revised now that the recombinant A_1B_0 haplotype is no more fit than the A_1B_1 haplotype that is spreading through the population. We assume that, whenever a recombinant A_1B_0 haplotype appears, the probability that this haplotype becomes the ancestor of the population at some distant future point in time is very nearly $1/(Np)$. This assumes that any individual carrying the A_1 allele alive at that time is equally likely to be the lucky one to ultimately fix and give rise to the entire descendant population. Using $1/(Np)$ instead of Π for the fixation probability of the recombinant A_1B_0 haplotype, we obtain the revised killing function, $K(p) = \rho\, p(1 - p)\, 1/p$, for use in the diffusion equation (8), assuming that allele A_1 fixes. The conditional probability of the process not being killed was then obtained using *Mathematica* 6.0.

Focusing on the conditional probability that the process reaches fixation on A_1 before being killed by the appearance of a successful recombinant, we again obtained an approximation assuming that selection is strong relative to drift

$$P^*(p_0) = (2\, e^{\gamma}\, S_{net})^{-\rho/S_{net}}, \tag{14}$$

where $\gamma = 0.577$ is Euler's constant. We have persisted in referring to the net selection on the A_1B_1 haplotype as S_{net} despite the fact that now $S_{net} = S_a$ for ease of comparison with the previous case.

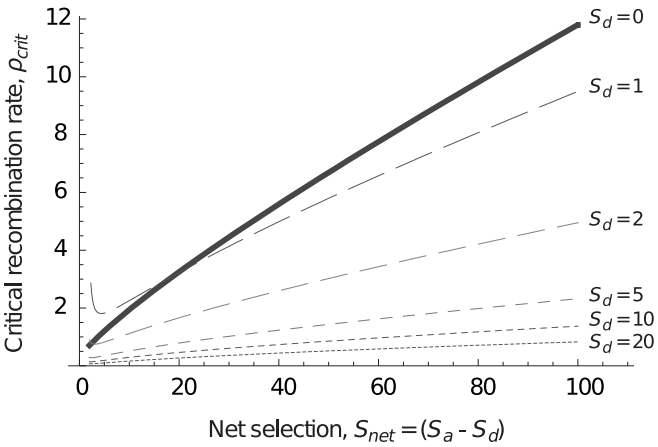


Figure 2. The critical value of the recombination rate, ρ_{crit}, below which there is a greater than $c = 50\%$ probability that the deleterious B_1 allele will hitchhike to fixation along with the advantageous allele, as a result of their initial association based on the approximate equations (13) and (15). In each case, ρ_{crit} is plotted along the *y*-axis versus S_{net} along the *x*-axis, for varying values of S_d. The case of a neutral linked allele at the *B* locus is given by the thick top curve. (The upturns in some of the curves near the origin as well as crossing of some of the curves are caused by inaccuracies in these approximations when selection is weak relative to drift.)

Again, solving this equation for the critical value of ρ below which hitchhiking to fixation occurs more than a proportion c of the time, we get

$$\rho_{crit}^{neutral} = S_{net}\left[\frac{\ln\left(\frac{1}{c}\right)}{\gamma + \ln(2\ S_{net})}\right]. \qquad (15)$$

For $c = 1/2$, the term in square brackets is approximately 1/4 when $S_{net} = 5$, and it continues to decline (but slowly) as S_{net} increases. Thus, as a rough rule of thumb, r must be less than $\approx$1/4 of s_{net} for there to be a 50% chance that a neutral allele hitchhikes to fixation. Again, such hitchhiking events are likely to occur over larger regions of the genome when the sweeps are faster (s_{net} large). The key difference, however, from the case with a deleterious hitchhiker is the absence of Ns_d in the denominator of this rule, which makes it easier to satisfy than the case of a deleterious hitchhiker (assuming selection is strong relative to drift). Figure 2 shows just how much more likely it is for alleles at locus B to hitchhike to fixation along with allele A when the B locus is neutral (thick top curve) than when it is subject to selection against deleterious mutations (dashed curves).

The fact that neutral alleles are much more likely to hitchhike to fixation than linked deleterious alleles has another important implication. Namely, the presence of a linked deleterious allele increases the chance that surrounding genetic variation will be rescued by recombination. Had there been no linked sites under selection, we would expect a region surrounding a sweep to be entirely fixed when $\rho < \rho_{crit}^{neutral}$ in the majority of cases (eq. 15). If a beneficial allele first occurs on a chromosome containing a deleterious allele, however, this region is greatly reduced to $\rho < \rho_{crit}$ (eq. 13), as illustrated in Figure 2. Consequently, linkage to sites carrying deleterious alleles reduces the impact of selective sweeps, making it less likely that surrounding genetic variation will be lost.

Turning this argument around, a recently fixed beneficial allele might have been strongly selected but appear to have been weakly selected based on the amount of genetic variation remaining in the region. This is because recombinants were favored that untied the beneficial allele from the deleterious genetic baggage with which it arose. Furthermore, we would expect that genetic variation should more often be rescued by the appearance of more fit recombinants on the side of a selective sweep that bears a higher density of other sites under selection. In Supporting information, we simulate a three-locus model with one locus subject to advantageous mutation, one locus being a neutral marker, and one locus subject to recurrent deleterious mutation, with the beneficial mutant placed on a randomly selected genetic background. As confirmed in Figure S1 the sweep of neutral diversity is less severe in cases where selection acts on the locus subject to deleterious mutations.

TWO COMPETING BENEFICIAL MUTATIONS

The above analyses can also be used to solve a related problem of beneficial mutations competing for fixation in the presence of recombination, as considered by Yu and Etheridge (2010). If a beneficial allele is rising in frequency when a second beneficial allele appears at a linked site, then it is possible for the first beneficial allele to be lost if the second allele is more strongly favored if it appears with the wildtype allele at the first locus, and if a recombinant that brings together both alleles onto the same haplotype fails to establish in time.

Although technically there are three chromosome types to be considered before the recombinant appears (00, 10, and the new 01, where the "1" now indicates a beneficial mutation at the first and second sites), we can approximate this scenario as did Yu and Etheridge (2010) by assuming that the 00 wild type is rapidly eliminated, so that the frequencies of 01 and 10 sum roughly to one. This approximation performs surprisingly well for this problem because rare recombination events do not occur until the 10 and 01 haplotypes are both common.

Equation (1) then describes the spread of the more fit 01 haplotype, whose frequency is $\approx p(t)$ (frequency of 10 $\approx 1 - p(t)$), with s_{net} equal to the difference in fitness between 10 and 01 individuals. Equation (2) describes the fixation probability of a recombinant double mutant, with s_a and s_d giving the selective advantage of the double mutant when it appears in a population

predominantly composed of 10 and 01 individuals, respectively. All of the subsequent results described above then follow. Figure S4 shows that equations (5) and (12) provide an excellent estimate of the probability that recombination successfully rescues both beneficial mutations. Although similar in spirit to the work of Yu and Etheridge (2010), our analyses have the advantage of providing closed-form solutions that appear to accurately capture the stochastic nature of recombination rescuing combinations of beneficial alleles at two selected loci.

Two-locus simulations

To investigate the accuracy of the above results, we compare both the semi-determinstic and stochastic models to Monte Carlo simulations. Simulations start with a population of N haploid chromosomes, each consisting of two linked loci. Fitness is assumed to be additive.

An initial proportion $p_0 = 1/N$ of the population is assigned the advantageous-deleterious A_1B_1 haplotype. The rest of the population bears the A_0B_0 haplotype. It is assumed that the A_0B_1 haplotype is present at a negligibly small frequency, and while it is not considered in the initial population it is tracked if it appears by recombination.

A new generation is formed by selecting two parents with probability proportional to their fitness. Recombination between the two parental loci then occurs with Poisson probability r. This is repeated until N new offspring are created. A new generation is created in this way until the A_1B_1 genotype is either fixed or is lost from the population. This entire process is repeated 20,000 times to build up an overall probability of fixation along with 95% confidence intervals. We focus attention on the processes where the advantageous allele fixes.

Results are plotted in Figure 3. Simulation data match up very well to all three solutions for the probability of hitchhiking $P^*(p)$: semi-deterministic equation (5), diffusion equation (B6), and the approximation to the diffusion equation (12). All three solutions offer similar results when we changed the population size, as long as ρ, S_d, S_a are held constant. Differences between the solutions only become apparent when selection becomes weak. Stochastic effects then play more of a role, especially where the A_1B_1 haplotype is oversampled and rises to fixation faster than expected, so that the diffusion with killing (B6) provides a slightly more accurate solution. Additional figures presented in section 3 of the Appendix S3 show that the analytical solutions perform less well as selection strengthens in very small populations (e.g., $s_d = 0.1$ with N = 100 or 1000); in these cases, the diffusion approximation assuming weak selection breaks down and the fixation probability of the deleterious allele is underestimated.

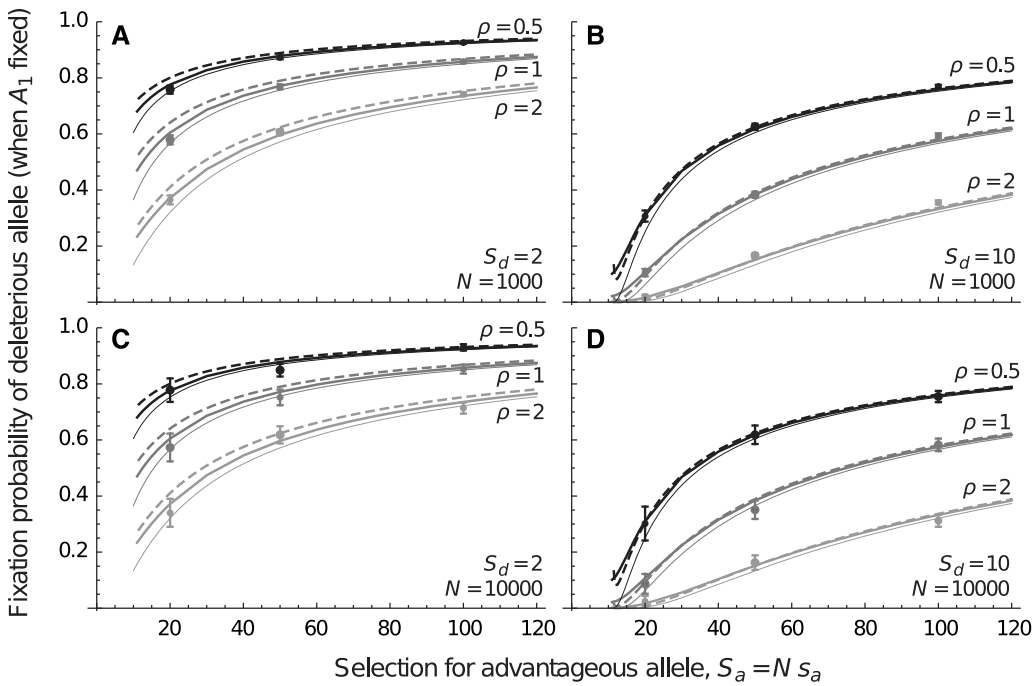


Figure 3. **Fixation probability of the deleterious B_1 allele, given that the A_1B_1 haplotype appears initially at frequency of $1/N$ and that the A_1 allele is not lost when rare, for different recombination rates $\rho = Nr$. Plots compare the solution to the semi-deterministic model (5) (thin solid), the full solution to the diffusion (B6) (thick solid), the approximation to the diffusion (12) (thick dashed), and simulation results based on the Wright–Fisher model (points). Bars indicate 95% confidence intervals here and throughout. Parameters are $N = 1000$ (A and B) and $N = 10,000$ (C and D), with $S_d = 2$ (A and C) and $S_d = 10$ (B and D).**

Multilocus simulations

Although the above two-locus models offer tractable results, novel advantageous alleles may arise in genomes with multiple mutant alleles. Therefore, we switch to using multilocus computer simulations to investigate the mutation load generated by the rise to fixation of an advantageous allele, given that such mutations arise at rate U in a genome with total map length R, where each new deleterious mutations is assigned a random position between 0 and R. The methods used for these simulations are based on Hartfield et al. (2010) and detailed in Supporting information.

We then determined the mean number of deleterious alleles that fix along with each beneficial mutation, assuming multiplicative selection. Simulations with different S_a values are compared to the control case, $S_a = 0$, in Figures 4 and S2. These results corroborate the two-locus model; the mean number of deleterious mutants that fix declines with the rate of recombination and rises with the strength of selection on the advantageous mutant, S_a. The mean number of fixed deleterious alleles also stays approximately the same as N increases, if the compound parameters S_a, S_d, NR, and NU are held constant.

Increasing the recombination rate also raises the fixation probability of the advantageous mutant (Fig. S3), which is a well-known result (Peck, 1994; Barton, 1995). Thus recombination is doubly advantageous, as it reduces the number of deleterious alleles that fix in a population following a selective sweep and it increases the likelihood that such an advantageous mutant can establish when rare. This is the likeliest cause of strong selection acting on a modifier for increased recombination in the presence of advantageous and deleterious mutations (Hartfield et al., 2010).

APPLYING RESULTS TO HUMAN GENETIC DATA

How likely is deleterious hitchhiking to occur in nature? To answer this, we use human data as an example. Deleterious mutants are maintained at a mutation–selection balance frequency of $q = \mu/s_d$ (Wright, 1931), where s_d measures selection against the deleterious allele in heterozygotes. Thus an estimate for the number of deleterious mutants segregating throughout a genome is U/s_d, for U the diploid per-genome deleterious mutation rate, which has been recently estimated as $U = 4.2$ (Eöry et al., 2010).

U measures deleterious mutations arising across the entire genome, with the majority appearing in noncoding regions (Eöry et al., 2010). Thus we assume all deleterious mutations have a fixed, weak value of s_d. This will slightly overestimate the number of deleterious mutants segregating, as we do not consider stronger deleterious mutations that can arise in coding regions (Eyre-Walker et al., 2006; Boyko et al., 2008).

A deleterious allele must have $N_e s_d \geq 1$ in order for selection to overcome the effects of genetic drift (Kimura, 1983). Therefore, assuming deleterious alleles are very weakly selected

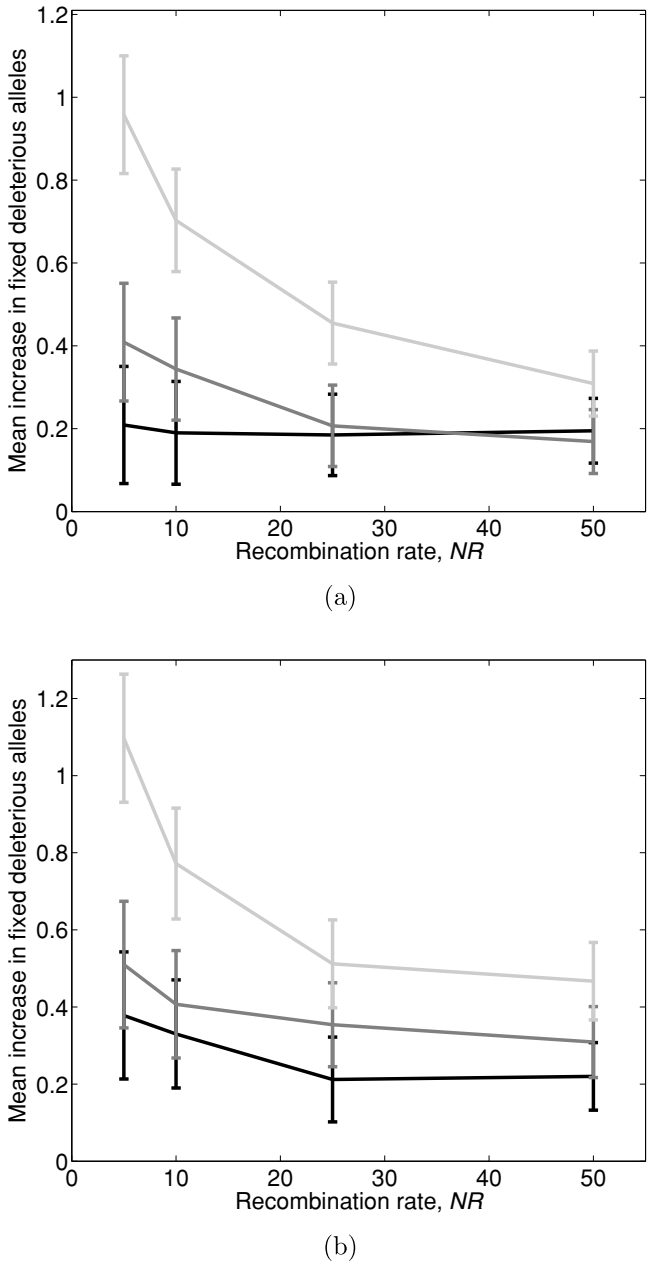


Figure 4. The increase in the number of deleterious alleles that fix genome-wide for a given S_a, subtracting off the number that fix in the $S_a = 0$ case, as a function of the total map length NR (see Fig. S2 for the raw data). Only cases where the advantageous allele has fixed are considered. $S_a = 20$ (black line), 40 (dark gray line), or 80 (light gray line). $S_d = 10$, $NU = 50$, and (a) $N = 500$ or (b) $N = 1000$.

($N_e s_d = 1$, with human $N_e = 10{,}000$; Jorde et al. 1998), we expect $U/s_d = 4.2/0.0001 = 42{,}000$ such deleterious alleles segregating at any time, roughly half of which lie in each haploid set of 3 Gb in the human genome. Including the site of the beneficial mutation, the average distance between two selected sites is thus 142.9 kb. Assuming that selected sites are randomly distributed across

the genome (i.e., ignoring clustering), this distance would be approximately exponentially distributed. In this case, the closest of the deleterious alleles lying to either side of the beneficial allele would also be exponentially distributed with mean 71.4 kb. As a rough guide, the average recombination rate is 1 cM/Mb in a human genome (Broman et al., 1998), thus the closest deleterious allele lies, on average, at a distance of $N_e r = 7.14$. The fixation probability of the deleterious allele with the advantageous mutant would then be 18.8% for $N_e s_a = 5$, 37.1% for $N_e s_a = 25$, and 62.1% for $N_e s_a = 100$, obtained by integrating the hitchhiking probability (B6) over an exponentially distributed distance with mean $N_e r = 7.14$. These calculations are explained in more detail in section 5 of the Appendix S3. If we assumed $N_e s_d = 10$, then by following a similar logic we calculate that the mean distance to the nearest deleterious allele is $N_e r = 71.4$, and the estimated fixation probability of a deleterious allele is 0.8% for $N_e s_a = 25$ and 2.5% for $N_e s_a = 100$.

Overall, these calculations suggest that in humans, deleterious mutants will hitchhike at appreciable frequencies only if they are very weakly selected ($N_e s_d < 10$). However, this is only an initial calculation that deserves to be revised to take into account fine-scale recombination rates (McVean et al., 2004) and clustering of mutations around coding regions. For now we note that if clustering causes the average recombination distance to a deleterious allele to drop tenfold, then the hitchhiking probabilities calculated above increase substantially, rising for $N_e s_d = 1$ to 68%, 85%, 94% with $N_e s_a = 5$, 25, 100, respectively, and for $N_e s_d = 10$ to 7%, 20% with $N_e s_a = 25$ and 100.

Discussion

As long as genetic variance in fitness is present within a population, new beneficial alleles can arise in genomes that, by chance, carry deleterious alleles at linked sites. Consequently, if they remain associated, deleterious alleles can hitchhike to fixation as an advantageous allele sweeps through the population. Even if recombination occurs between the two loci, there can still be a good chance of both alleles fixing, if either the recombinant fails to appear in time or is lost by chance when it does appear. Williamson et al. (2007) found possible evidence of such hitchhiking causing the high prevalence of the hereditary hemochromatosis mutation C282Y, due to a selective sweep occurring 150 kb away from the HFE gene where the deleterious C282Y allele is located.

To our knowledge, this article represents the first theoretical study on how recombination affects the hitchhiking to fixation of deleterious alleles. Using both a semi-deterministic and a diffusion approach, we show that in regions of low recombination there is a high probability that a deleterious mutant would be swept to fixation if linked to an advantageous mutant (Fig. 1). This probability approaches one as the deleterious effect s_d tends towards zero and the overall advantage of the A_1B_1 haplotype s_{net} is larger. Outside this parameter range, we find that hitchhiking is likely (greater then 50% chance) if $r \lesssim s_{net}/(4Ns_d)$ (more precisely, equation 13). A promising empirical approach would be to investigate areas around the genome that show high d_N/d_S values. Such regions are assumed to be subject to recurrent sweeps (Nielsen, 2005). If deleterious alleles do hitchhike, then around these sites there should be signs of increased load, such as increased indel frequency, or lower frequency of optimal codon usage. Such a negative relationship between *dN* and optimal codon usage was found in *Drosophila* by Betancourt and Presgraves (2002).

Furthermore, we determined that the hitchhiking of tightly linked deleterious alleles reduces the region in which the sweep is likely to fix surrounding sites (compare eq. 15 to eq. 13). This is important as it implies that deleterious hitchhiking can alter experimental estimates of the strength of such sweeps. A potential example of these effects was reported by Clegg et al. (1980), who found that linkage disequilibrium in *D. melanogaster* broke down more quickly than expected (geometric decay at a ratio $1 - r$), based on the surrounding markers being neutral and on measured recombination rates between the selected and neutral markers. This observation could be explained by recombination untangling advantageous alleles from deleterious backgrounds (see also Fig. S1). Further work is warranted to explore the impact of neighboring selected sites on patterns of neutral sequence variability in a fully multilocus framework. In particular, a full treatment requires an exploration not only of the primary effects of a selective sweep at a focal site, but also of how hitchhiking of deleterious alleles can cause secondary sweeps as wild-type alleles reestablish themselves at surrounding sites.

Our work also sheds light on the results found by Hartfield et al. (2010), who showed that a modifier gene for increased recombination is more likely to fix in a population that is subject to both deleterious and advantageous mutation, compared to the deleterious-only mutation case (Keightley and Otto, 2006). The increased selection acting on a recombination modifier when both deleterious and advantageous mutants are present together, compared to when just deleterious or just advantageous mutations are present, suggests that uncoupling advantageous mutants from deleterious backgrounds provides a substantial amount of selection on a recombination modifier (Peck, 1994; Hartfield et al., 2010).

Our preliminary calculations suggest that in obligately sexual species with long genetic map lengths (such as the human genome), recombination is frequent enough to prevent all but weakly deleterious mutants from hitchhiking with advantageous mutants. Our calculations assumed, however, that mutations affecting fitness arise at equal rates throughout the genome, which ignores the clustering of fitness-impacting sites near genic regions. If recombination rates between selected sites are low, either

because of this clustering or because of cold spots in recombination, the probability that deleterious alleles hitchhike to fixation rises substantially. Similarly, in species that frequently inbreed (e.g., selfing) or reproduce asexually, the effective amount of recombination may be much lower, substantially increasing the probability of deleterious alleles hitchhiking to fixation. In asexuals with no recombination, the subsequent mutation accumulation can be extremely detrimental (Hadany and Feldman, 2005).

In conclusion, sex and recombination both enhance the probability of beneficial alleles establishing and hinder the fixation of deleterious alleles within a lineage. If this can be shown empirically to be a potent selective force on recombination rates, then this would provide key insight into why sex and recombination are prevalent, which remains an open question in evolutionary genetics (Otto, 2009).

ACKNOWLEDGMENTS

We would like to thank P. Keightley for his support and advice on using human genetic data. We also thank P. Keightley, N. Barton, J. Hermisson, and two anonymous referees for comments on the manuscript. MH is funded by a Biotechnology and Biological Sciences Research Council studentship; SO is funded by the Natural Sciences and Engineering Research Council of Canada.

LITERATURE CITED

Abramowitz, M., and I. Stegun. 1970. Handbook of mathematical functions. Dover Publications, Inc., New York.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149–1152.

———. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 17:1755–1762.

Bachtrog, D., and I. Gordo. 2004. Adaptive evolution of asexual populations under Muller's ratchet. Evolution 58:1403–1413.

Barrett, R. D. H., R. Craig MacLean, and G. Bell. 2006. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. Biol. Lett. 2:236–238.

Barton, N. H. 1994. The reduction in fixation probability caused by substitutions at linked loci. Genet. Res. 64:199–208.

———. 1995. Linkage and the limits to natural selection. Genetics 140:821–841.

Barton, N. H., and S. P. Otto. 2005. Evolution of recombination due to random drift. Genetics 169:2353–2370.

Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in *Drosophila*. Proc. Natl. Acad. Sci. USA 99:13616–13620.

Bierne, N., and A. Eyre-Walker. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol. Biol. Evol. 21:1350–1360.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. 63:861–869.

Bull, J. J., M. R. Badgett, and H. A. Wichman. 2000. Big-benefit mutations in a bacteriophage inhibited with heat. Mol. Biol. Evol. 17:942–950.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Clegg, M. T., J. F. Kidwell, and C. R. Horch. 1980. Dynamics of correlated genetic systems. *V. Rates of decay of linkage disequilibria in experimental populations of Drosophila melanogaster*. *Genetics* 94:217–234.

Crow, J. F. 1970. Genetic loads and the cost of natural selection. Pp. 128–177, *in* K. I. Kojima, ed. Mathematical topics in population genetics, *Biomathematics*, vol. 1. Springer-Verlag, Berlin.

Desai, M. M., and D. S. Fisher. 2007. Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics 176:1759–1798.

Elena, S. F., and R. E. Lenski. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat. Rev. Genet. 4:457–469.

Eöry, L., D. L. Halligan, and P. D. Keightley. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol. Biol. Evol. 27:177–192.

Ewens, W. J. 2004. Mathematical population genetics: 1. Theoretical introduction, *Interdisciplinary applied mathematics*, Vol. 27. 2nd ed. Springer, New York.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. Trends Ecol. Evol. 21:569–575.

Eyre-Walker, A., and P. D. Keightley. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26:2097–2108.

Eyre-Walker, A., M. Woolfit, and T. Phelps. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891–900.

Fisher, R. A. 1930. The genetical theory of natural selection. The Clarendon Press, Oxford.

Haag-Liautard, C., M. Dorris, X. Maside, S. Macaskill, D. L. Halligan, B. Charlesworth, and P. D. Keightley. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature 445:82–85.

Hadany, L., and M. W. Feldman. 2005. Evolutionary traction: the cost of adaptation and the evolution of sex. J. Evol. Biol. 18:309–314.

Haldane, J. 1924. A mathematical theory of natural and artificial selection, part I. Trans. Cambridge Philos. Soc. 23:19–41.

Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection, part V: Selection and mutation. Math. Proc. Cambridge Philos. Soc. 23:838–844.

Hall, D. W., and S. B. Joseph. 2010. A high frequency of beneficial mutations across multiple fitness components in *Saccharomyces cerevisiae*. Genetics 185:1397–1409.

Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6:e1000825.

Hartfield, M., S. P. Otto, and P. D. Keightley. 2010. The role of advantageous mutations in enhancing the evolution of a recombination modifier. Genetics 184:1153–1164.

Hill, J. A., and S. P. Otto. 2007. The role of pleiotropy in the maintenance of sex in yeast. Genetics 175:1419–1427.

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Jensen, J. D., K. R. Thornton, and P. Andolfatto. 2008. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet. 4:e1000198.

Johnson, T., and N. H. Barton. 2002. The effect of deleterious alleles on adaptation in asexual populations. Genetics 162:395–411.

Jorde, L. B., M. Bamshad, and A. R. Rogers. 1998. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays 20:126–136.

Joseph, S. B., and D. W. Hall. 2004. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. Genetics 168:1817–1825.

Karlin, S., J. L. McGregor, and W. Bodmer. 1967. The rate of production of recombinants between linked genes in finite populations. Proc. Fifth Berkeley Symp. on Math. Stat. Prob. 4:403–414.

Karlin, S., and H. M. Taylor. 1981. A second course in stochastic processes. Academic Press, New York.

Keightley, P. D., and M. Lynch. 2003. Towards a realistic model of mutations affecting fitness. Evolution 57:683–685.

Keightley, P. D., and S. P. Otto. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443:89–92.

Keightley, P. D., U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195–1201.

Kimura, M. 1970. Stochastic processes in population genetics. Pp. 178–209, *in* K. -I. Kojima, ed. Mathematical Topics in Population Genetics, *Biomathematics*, vol. 1. Springer-Verlag; Heidelberg; New York, Berlin.

———. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

Kimura, M., and T. Ohta. 1970. Probability of fixation of a mutant gene in a finite population when selective advantage decreases with time. Genetics 65:525–534.

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

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Nielsen, R. 2005. Molecular signals of natural selection. Annu. Rev. Genet. 39:197–218.

Obbard, D. J., J. J. Welch, K.-W. Kim, and F. M. Jiggins. 2009. Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet. 5:e1000698.

Otto, S. P. 2009. The evolutionary enigma of sex. Am. Nat. 174:S1–S14.

Otto, S. P., and N. H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. Genetics 147:879–906.

Peck, J. R. 1994. A ruby in the rubbish: Beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137:597–606.

Rice, W. R. 1999. Free content genetic polarization: unifying theories for the adaptive significance of recombination. J. Evol. Biol. 12:1047–1049.

Roze, D., and N. H. Barton. 2006. The Hill-Robertson effect and the evolution of recombination. Genetics 173:1793–1811.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu, D. A. Turissini, S. Fang, H.-Y. Wang, R. R. Hudson, R. Nielsen, Z. et al. 2007. Adaptive genic evolution in the *Drosophila* genomes. Proc. Natl. Acad. Sci. USA 104:2271–2276.

Shaw, F. H., C. J. Geyer, and R. G. Shaw. 2002. A comprehensive model of mutations affecting fitness and interferences for *Arabidopsis thaliana*. Evolution 56:453–463.

Thomson, G. 1977. The effect of a selected locus on linked neutral loci. Genetics 85:753–788.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3:e90.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Yu, F., and A. M. Etheridge. 2008. Rate of adaptation of large populations. Pp. 3–27, *in* P. Pontarotti, ed. Evolutionary Biology from Concept to Application. Springer-Verlag, Berlin.

———. 2010. The fixation probability of two competing beneficial mutations. Theor. Popul. Biol. 78:36–45.

Yu, F., A. M. Etheridge, and C. Cuthbertson. 2010. Asymptotic behavior of the rate of adaptation. Ann. Appl. Probab. 20:978–1004.

Associate Editor: J. Hermisson

Appendix A

DERIVATION OF $\Pi(t)$, THE PROBABILITY OF ESTABLISHMENT OF A RECOMBINANT HAPLOTYPE

When the recombinant A_1B_0 haplotype is produced, it appears within a population that is already changing due to the spread of the A_1B_1 haplotype. Thus, we cannot calculate the probability of fixation of the recombinant A_1B_0 haplotype based solely on its fitness $1 + s_a$ relative to the current population mean $1 + p(t)(s_a - s_d)$. Rather, we must also account for future changes in the population mean fitness as the A_1B_1 haplotype rises in frequency. To do so, we develop a time-inhomogeneous branching process that explicitly follows the dynamics of $p(t)$ (given by eq. 1) that occur after the appearance of the recombinant A_1B_0 haplotype. A previous diffusion analysis by Kimura and Ohta (1970) also calculated the fixation probability for a favorable allele whose benefit declined over time, but the focus of their analysis was on a case where selection declines linearly over time, whereas here the selection coefficient favoring A_1B_0 declines according to a logistic function of time, given by $s(t) = s_a - p(t)(s_a - s_d)$.

Let $\Pi(t)$ be the fixation probability of the recombinant A_1B_0 haplotype at generation t, given that the current frequency of the A_1B_1 haplotype is $p(t)$. In a population of constant size, the average parent has one surviving offspring, but we assume that the A_1B_0 haplotype is more fit and so has an average of $1 + s(t)$ offspring. Using branching process logic (Haldane, 1927), the recombinant A_1B_0 haplotype will ultimately be lost (with probability $1 - \Pi(t)$) if and only if all j offspring inheriting the haplotype also fail to leave any descendants over the long run (with probability $(1 - \Pi(t + 1))^j$). Assuming a Poisson distribution for the number of offspring j and summing over this distribution, we obtain a recursion for $\Pi(t)$:

$$1 - \Pi(t) = \sum_{j=0}^{\infty} e^{-(1+s(t))} \frac{(1 + s(t))^j}{j!} (1 - \Pi(t + 1))^j = exp[-(1 + s(t))\ \Pi(t + 1)]. \quad \text{(A1)}$$

Solving for $\Pi(t + 1)$ and subtracting $\Pi(t)$, we obtain the change in fixation probability over time, which we assume is slow enough that it can be well approximated by the differential

equation

$$\frac{d\Pi}{dt} = -\frac{ln[1-\Pi(t)]}{1+s(t)} - \Pi(t). \tag{A2}$$

With weak selection ($s(t) \ll 1$), $\Pi(t)$ is of the same order as $s(t)$ and the above simplifies to

$$\frac{d\Pi}{dt} = -\frac{1}{2}\Pi(t)^2 + s(t)\ \Pi(t) + O(s^2) \tag{A3}$$

(Barton, 1995). This differential equation can be solved when selection on the recombinant haplotype varies according to $s(t) = s_a - p(t)(s_a - s_d)$ by first replacing the variable t with the variable p using the chain rule and $dp/dt = s_{net}p(1-p)$ (section 1 of Appendix S3). To leading order in the selection coefficients, the resulting solution for the fixation probability of the recombinant A_1B_0 haplotype is given by equation (2).

Appendix B

DERIVING THE DIFFUSION PROCESS WITH KILLING CONDITIONAL ON FIXATION OF THE A_1 ALLELE

Conditioning on the fixation of A_1 implies that either the A_1B_1 haplotype fixes (if the process is not killed) or the recombinant successfully establishes and leads to the fixation of the A_1B_0 haplotype (if the process is killed). Either way, the A_1B_1 haplotype cannot be lost while it is rare. We must thus adjust the drift term in the diffusion, $M(p)$, to account for the fact that the A_1B_1 haplotype will, on average, rise more rapidly when rare among those processes where the A_1B_1 haplotype is not lost. The variance term $V(p)$ and the killing term $K(p)$ are unchanged in the conditioned model, as these terms depend only on the current frequency of the A_1B_1 haplotype and not on its ultimate fate. From equation (9.5) in chapter 15 of Karlin and Taylor (1981), the conditional drift term $M^*(p)$ is given by.

$$M^*(p) = S_{net}\ p(1-p) + \frac{s(p)}{S(p)}\ p(1-p), \tag{B1}$$

where

$$s(p) = exp\left[-\int_0^p \frac{2M(\eta)}{V(\eta)}d\eta\right] \tag{B2}$$

$$S(p) = \int_0^p s(\xi)\, d\xi. \tag{B3}$$

Here, the values of $M(p)$ and $V(p)$ are for the unconditional diffusion process as outlined in the main part of the article. Plugging these terms into equations (B2) and (B3) and evaluating the integrals, we obtain the conditional drift term:

$$M^*(p) = S_{net}\ p(1-p)\ \frac{1+e^{-2pS_{net}}}{1-e^{-2pS_{net}}}. \tag{B4}$$

This revised drift term is then placed in equation (8), along with the variance and killing terms, which remain unchanged. Dividing the result by $p(1-p)$ yields equation (11) in the main text.

The conditional diffusion process requires some care, however, with the boundary conditions. The probability that the process is not killed given that the A_1B_1 haplotype is fixed remains one, $P^*(1) = 1$, as before. Conditioning assumes, however, that the $p = 0$ boundary is never reached. Rather than assigning $P^*(0)$, we instead assume that $P^*(p)$ varies little over very small values of p, given that the process will ultimately reach $p = 1$ if it is not killed. Thus, we use $dP^*(0)/dp = 0$ as a second boundary condition.

Solving equation (10), we find that the probability that the process is never killed, regardless of whether A_0 or A_1 ultimately fixes is

$$\begin{aligned} P(p) = &\left(U^0_{-\omega}[-2(pS_{net}+S_d)]\left(L^{-1}_{\omega}[-2S_d]\right.\right. \\ &\left. - L^{-1}_{\omega}[-2(S_{net}+S_d)]\right) - L^{-1}_{\omega}[-2(pS_{net}+S_d)] \\ &\times \left(U^0_{-\omega}[-2S_d] - U^0_{-\omega}[-2(S_{net}+S_d)]\right)\Big) \\ &/\left(U^0_{-\omega}[-2(S_{net}+S_d)]\, L^{-1}_{\omega}[-2S_d]\right. \\ &\left. - U^0_{-\omega}[-2S_d]\, L^{-1}_{\omega}[-2(S_{net}+S_d)]\right), \end{aligned} \tag{B5}$$

whereas the solution to equation (11), conditioned on the fixation of the beneficial A_1 allele, given by B6 (below). Here, $U^b_a[z] = U[a, b, z]$ is the Tricomi confluent hypergeometric function, $L^{\alpha}_n[x]$ the generalized Laguerre polynomial (Abramowitz and Stegun, 1970), and ω is the compound parameter given by equation (6) in the main text. Additional details regarding the derivation and solutions for these equations are provided in a *Mathematica* 6.0 file (Supporting information, section 2).

$$P^*(p) = \left(\frac{1-e^{-2S_{net}}}{1-e^{-2pS_{net}}}\right) \times \frac{U^0_{-\omega}[-2(pS_{net}+S_d)]\, L^{-1}_{\omega}[-2S_d] - U^0_{-\omega}[-2S_d]\, L^{-1}_{\omega}[-2(pS_{net}+S_d)]}{U^0_{-\omega}[-2(S_{net}+S_d)]\, L^{-1}_{\omega}[-2S_d] - U^0_{-\omega}[-2S_d]\, L^{-1}_{\omega}[-2(S_{net}+S_d)]} \tag{B6}$$

Supporting Information

The following supporting information is available for this article:

Appendix S1. Testing the effect of recombination on the fixation of a linked, neutral allele.
Appendix S2. Methods used for multilocus simulations.
Appendix S3. Derivations in mathematica (HartfieldOttoSM.nb file available for download).
Figure S1. The mean frequency of a linked neutral allele following a successful selective sweep, given as a function of the recombination rate Nr between different sites.
Figure S2. The mean number of deleterious alleles that fix genome-wide following the completion of a successful selection sweep, given as a function of the recombination rate NR (see Fig. 4 for data presented relative to $S_a = 0$). $S_a = 0$ (black dashed line), 20 (black solid line), 40 (dark gray), or 80 (light gray).
Figure S3. Fixation probability of the advantageous mutant in multilocus simulations, as a function of the recombination rate NR. $S_a = 0$ (black dashed line), 20 (black solid line), 40 (dark gray), or 80 (light gray).
Figure S4. Fixation probability of a recombinant carrying two beneficial mutations.

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

S.1 Testing the effect of recombination on the fixation of a linked, neutral allele

We investigate the effect of deleterious hitchhiking on neutral sequence, and how recombination affects this, by extending the two-locus simulations. In these new simulations, there are three linked loci; the deleterious site, a neutral locus and a third locus where the sweep is present, all separated by a recombination distance r. Additionally, there is now recurrent mutation occurring at rate μ at the deleterious locus, with no back-mutation.

The initial set-up is different as well: initially the deleterious allele is present at a fixed mutation-selection balance frequency of μ/s_d (Wright 1931), or 50% if $s_d = 0$. When the advantageous allele is introduced in a single copy, it is placed within a random individual that does not necessarily carry a deleterious allele. This is because we want to measure the difference in diversity due to background selection, averaged over all possible initial backgrounds. A neutral allele is also introduced in the individual in which the sweep first arises. This neutral marker allows us to measure the extent to which the initial neutral diversity is reduced to the singly allele that happens to be adjacent to the new mutation.

The population then undergoes the same cycle of selection, recombination, then mutation at the deleterious locus. The sweep is tracked until it is fixed or lost. If fixed, the frequency of the neutral allele is noted. The sweep is then reintroduced 1,000,000 times, and the mean final frequency of the neutral allele is measured.

If the mean frequency is near one, this implies that little recombination has taken place between the neutral and selectively favoured allele over the course of the sweep. A lower value implies that a higher level of recombination has taken

place, resulting in the advantageous allele becoming separated from the neutral allele that it was originally linked to. The results of the main text predict that recombination must be even tighter (lower ρ_{crit}, see Figure 2) for hitchhiking to fix nearby deleterious alleles, compared to the case of nearby neutral alleles, suggesting that those recombination events that do occur in the presence of surrounding selected sites are more likely to establish themselves, on average, within the population during a selective sweep. The increased establishment of recombinant chromosomes should result in reduced effects on linked neutral diversity as well. In particular, in these simulations, we predict that a beneficial allele will not drag a neighboring neutral allele to as high a frequency in the presence of another selected site in the surrounding region.

Figure S.1 plots the results of these simulations. In all cases tested we verify the prediction that if $S_d > 0$, the initially linked neutral allele does not sweep to as high a frequency on average, compared to the $S_d = 0$ case. This indicates that the diversity present at linked sites is more likely to be preserved by recombination when beneficial alleles arise at sites surrounded by others subject to selection. The reductions observed in our simulations are modest; this is due to there being only one linked deleterious site, with a low mutation rate ($\mu = 0.0005$). A larger effect would be observed if the mutation rate was higher or more linked deleterious loci were present.

In summary, we argue that linked selected sites should reduce the impact that a selective sweep has on surrounding neutral diversity. Our reasoning focuses on the increased probability that recombinant chromosomes will establish within the population during the sweep, because they uncouple a beneficial allele from any deleterious alleles within its genetic background (main text, Barton (1995)). Veri-

fying that this is indeed the case in a multi-locus framework deserves future work. In particular, our simulations have not accounted for the variation in fitness among recombinant chromosomes due to additional selected sites throughout the genome, beyond the one neighboring selected site. Furthermore, we have not accounted for the cascade of secondary sweeps that occur whenever fitter recombinants arise and drag along with them their own suite of alleles.

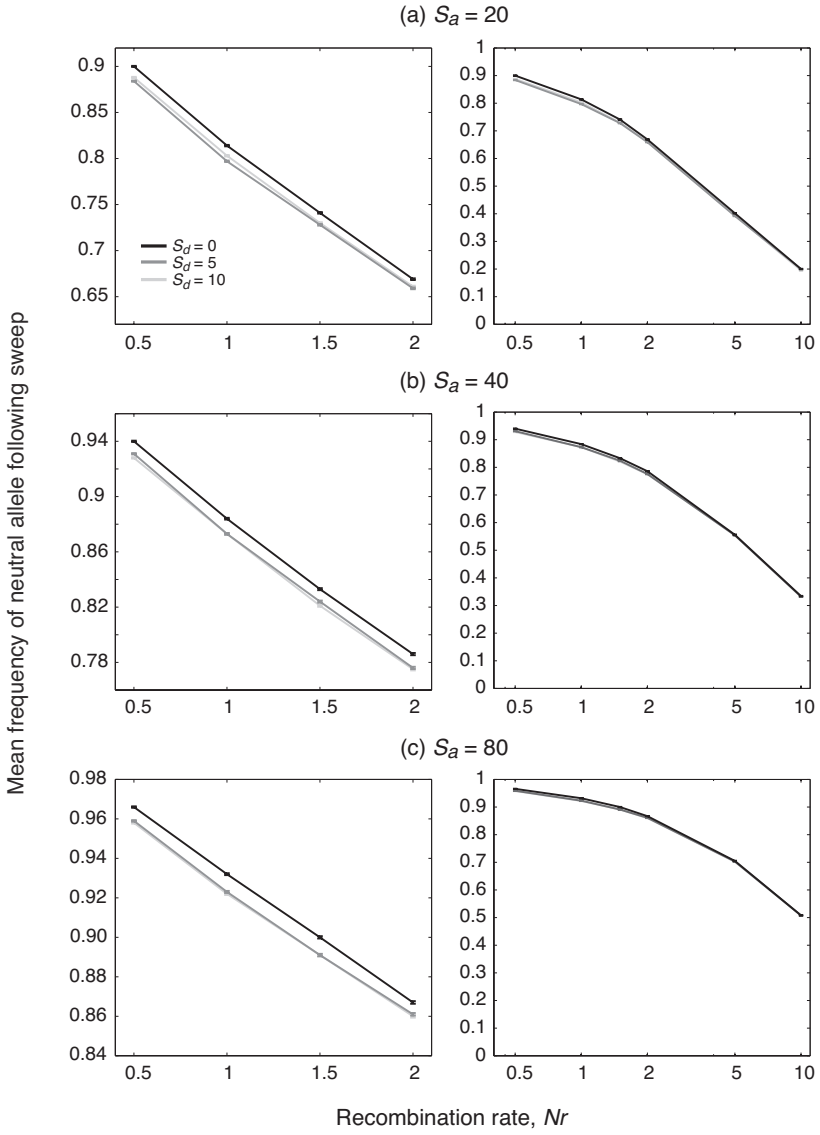


Figure S.1: The mean frequency of a linked neutral allele following a successful selective sweep, given as a function of the recombination rate Nr between different sites. The left- and right-hand panels report the same simulations, with the left-hand panels zoomed into the region (tighter linkage) where the impact of a neighboring selected site on the patterns of neutral diversity was greatest. $S_d = 0$ (black line), 5 (dark gray line) or 10 (light gray line). $N = 1000$, $\mu = 0.0005$, with (a) $S_a = 20$, (b) $S_a = 40$, or (c) $S_a = 80$.

S.2 Methods used for Multilocus Simulations

Initially there is a haploid population of N chromosomes with an infinite number of loci per chromosome. Each locus has a wildtype allele or a deleterious allele with selection s_d acting against it. Fitness is multiplicative, so initially in the absence of the advantageous allele the fitness of an individual is $(1 - s_d)^k$, where k is the number of deleterious mutants present in an individual chromosome.

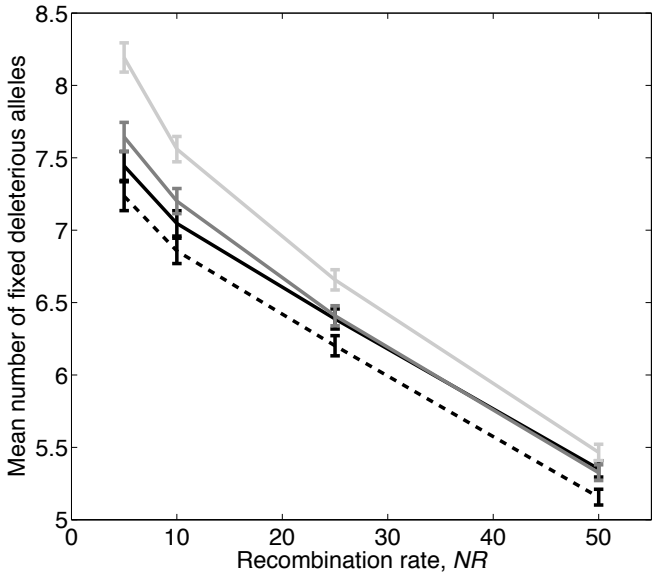
New generations are created by selection, recombination, then mutation. Two parents are chosen with replacement from the population, with probability proportional to their fitness. Recombination then occurs, with the number of crossovers across the chromosome selected from a Poisson distribution with mean R. One of these parents is selected to be the template for the offspring genome. Each mutant has a map position assigned to it as it appears, which is drawn from a uniform $[0, 1]$ distribution. For each crossover event, the position of the recombination event is also drawn from a $[0, 1]$ uniform distribution. The allelic states are then swapped at sites whose map distances exceed the recombination distance. If two crossovers are chosen, locus states are swapped at sites whose map distances lie between the two recombination distances. The number of crossovers is capped at two for ease of computing, which leads to little loss of accuracy if R is small (Hartfield et al. 2010).

For each offspring, the number of new deleterious mutants is chosen from a Poisson distribution with mean U. Each new mutant is assigned to a new locus. Back mutation also occurs at a deleterious allele with probability $\mu = 10^{-8}$. Overall, the whole cycle is repeated N times to repopulate the gene pool.

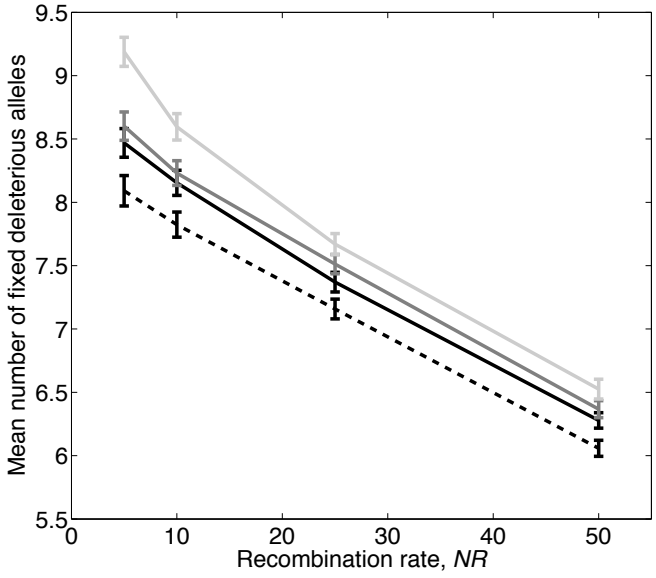
There is an initial burn-in of $2N$ generations, so that the population reaches

a mutational steady-state. A 'garbage collection' routine is executed every 50 generations during the burn-in; mutants that are lost from the population are cleared to free memory, as well as deleterious mutants that have fixed, so that we do not consider deleterious mutants that accumulate through Muller's Ratchet before the advantageous allele is introduced. Following the burn-in, the state of the population is saved and mutation is turned off, so deleterious mutants that do fix tend to be driven to fixation by hitchhiking, rather than an on-going ratchet mechanism. An advantageous allele is then added to a random chromosome at a random site. This allele increases the fitness of the host chromosome to $(1 + s_a)(1 - s_d)^k$.

The advantageous allele is then tracked until it is fixed or lost from the population. During this time a different garbage collection routine is run every 50 generations, which only clears mutants lost from the population in order to free memory. Fixed deleterious alleles are not cleared at this stage, so the mean number that fix with the sweep can be measured. If the advantageous mutant reaches fixation, then all remaining deleterious mutants are tracked until they are fixed or lost, to determine how many deleterious mutants fix. The advantageous mutant is reintroduced from the burn-in population 3,000 times, and its fixation probability is calculated, along with the average number of fixed deleterious mutants. This is repeated for 4,000 burn-ins to build a probability distribution for these statistics.

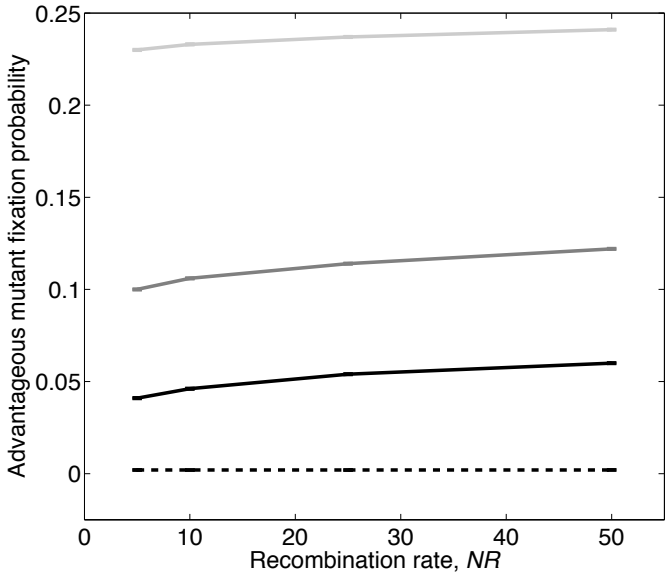


(a)

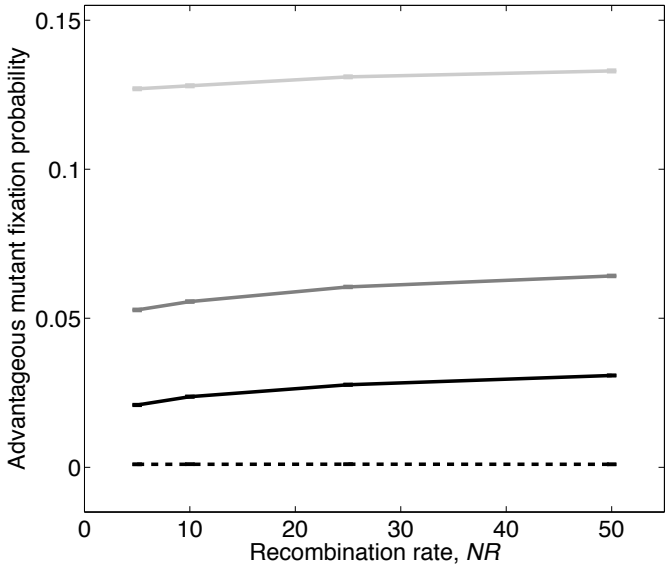


(b)

Figure S.2: The mean number of deleterious alleles that fix genome-wide following the completion of a successful selection sweep, given as a function of the recombination rate NR (see Figure 4 for data presented relative to $S_a = 0$). $S_a = 0$ (black dashed line), 20 (black solid line), 40 (dark gray) or 80 (light gray). $S_d = 10$, $NU = 50$, and (a) $N = 500$ or (b) $N = 1000$.



(a)



(b)

Figure S.3: Fixation probability of the advantageous mutant in multilocus simulations, as a function of the recombination rate NR. $S_a = 0$ (black dashed line) 20 (black solid line), 40 (dark gray) or 80 (light gray). $S_d = 10$, $NU = 50$, and (a) $N = 500$ or (b) $N = 1000$.

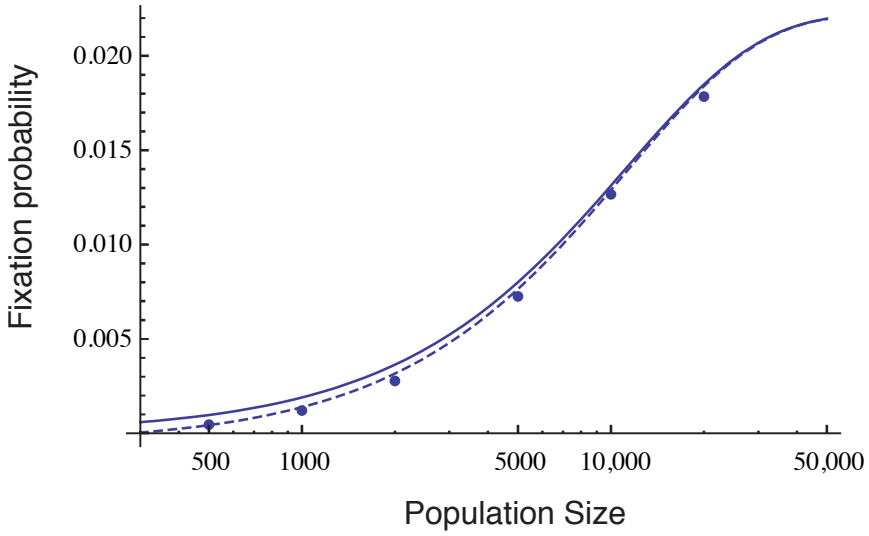


Figure S.4: Fixation probability of a recombinant carrying two beneficial mutations. Using the notation of Yu and Etheridge (2010), $s\gamma$ is the selective advantage of the first beneficial mutation to occur, s is the selective advantage of the second beneficial allele, and $s(1+\gamma)$ is the selective advantage of the recombinant double mutant. The semi-deterministic solution 5 (solid curve) and the diffusion solution 12 (dashed curve) are presented alongside simulation results (dots) for the parameters considered in Figure 4a of Yu and Etheridge (2010), where $s_{net} = s(1-\gamma)$ equals the difference in fitness between 10 and 01 individuals, and where $s_a = s$ and $s_d = s\gamma$ give the selective advantage of the double mutant when it appears in a population predominantly composed of 10 and 01 individuals, respectively. These curves are multiplied by the establishment probability of the second beneficial allele, given by equation 2 with the selection coefficients now reflecting the advantage of 10 spreading within a population of 00 individuals ($s_{net} = s\gamma$), within which a 01 mutant appears with advantage $s_a = s$ over the 00 wildtype. Parameters as in Figure 4a of Yu and Etheridge (2010): $s = 0.02$, $\gamma = 0.8$, $r = 0.00001$, with a starting frequency of haplotype 10 of 0.2 at the time that the second beneficial mutation appears in a 01 haplotype.