Obstruction of adaptation in diploids by recessive, strongly deleterious alleles

Zoe June Assaf[a,b,1], Dmitri A. Petrov[b,2], and Jamie R. Blundell[b,c,2]

Departments of [a]Genetics, [b]Biology, and [c]Applied Physics, Stanford University, Stanford, CA 94305

Edited by Montgomery W. Slatkin, University of California, Berkeley, CA, and approved April 3, 2015 (received for review December 30, 2014)

Recessive deleterious mutations are common, causing many genetic disorders in humans and producing inbreeding depression in the majority of sexually reproducing diploids. The abundance of recessive deleterious mutations in natural populations suggests they are likely to be present on a chromosome when a new adaptive mutation occurs, yet the dynamics of recessive deleterious hitchhikers and their impact on adaptation remains poorly understood. Here we model how a recessive deleterious mutation impacts the fate of a genetically linked dominant beneficial mutation. The frequency trajectory of the adaptive mutation in this case is dramatically altered and results in what we have termed a "staggered sweep." It is named for its three-phased trajectory: (*i*) Initially, the two linked mutations have a selective advantage while rare and will increase in frequency together, then (*ii*), at higher frequencies, the recessive hitchhiker is exposed to selection and can cause a balanced state via heterozygote advantage (the staggered phase), and (*iii*) finally, if recombination unlinks the two mutations, then the beneficial mutation can complete the sweep to fixation. Using both analytics and simulations, we show that strongly deleterious recessive mutations can substantially decrease the probability of fixation for nearby beneficial mutations, thus creating zones in the genome where adaptation is suppressed. These mutations can also significantly prolong the number of generations a beneficial mutation takes to sweep to fixation, and cause the genomic signature of selection to resemble that of soft or partial sweeps. We show that recessive deleterious variation could impact adaptation in humans and *Drosophila*.

recessive | hitchhiking | inbreeding depression | adaptation | selective sweep

In diploids, the fitness effect of having a single copy of a mutation depends not only on the mutation's selective effect, s, but also on its heterozygous effect, h. A fully recessive mutation ($h = 0$) is hidden in the heterozygote ($hs = 0$), and a fully dominant mutation ($h = 1$) is completely exposed ($hs = s$). Although it is generally agreed that beneficial mutations that reach fixation tend to be dominant (i.e., $h \geq 0.5$) (1, 2), both empirical data and theoretical models (3–6) suggest that many strongly and even moderately deleterious mutations are likely to be recessive (i.e., $h < 0.5$). For example, studies of de novo mutations from both mutation accumulation and mutagenesis experiments in *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans* have repeatedly found that strongly deleterious mutations tend to be completely recessive ($h \approx 0$) and more weakly deleterious mutations tend to be partially recessive ($h \approx 0.1$) (7–13). Furthermore, studies of natural populations have found that inbreeding depression is pervasive across sexually reproducing diploids and is mainly caused by recessive deleterious variation (reviewed in ref. 14). For example, in natural populations of *Drosophila*, approximately 30% of chromosomes carry a recessive lethal, and chromosomes that do not carry a recessive lethal suffer from at least 30% depression in homozygous fitness (15–24). These data suggest that many, if not most, deleterious mutations are likely to be fully or partially recessive, and such mutations can have a moderate to strong fitness effect in the homozygote.

It is thus possible that when a new adaptive mutation occurs, it will land on a chromosomal background containing one or more recessive deleterious mutations. It is well established that in finite populations, the fate of a new adaptive mutation should be affected by its genetically linked neighbors. For example, the rate of fixation of beneficial mutations at a single site will be lower if there are additional sites subject to positive or negative selection along the chromosome (25–30). This "Hill–Robertson" or "linkage" interference can be alleviated by recombination, allowing adaptive mutations to combine onto the same background or escape deleterious neighbors (31–35). In addition to affecting the rate of fixation of beneficial mutations, a deleterious mutation genetically linked to a beneficial mutation can "hitchhike" to high frequency or even fixation (36, 37). These theoretical predictions are beginning to be borne out in empirical data (38–47).

Despite the large body of work on hitchhiking and observations of recessive deleterious mutations in real organisms, there is a gap in our understanding of how recessive deleterious mutations affect adaptation. Models of hitchhiking have primarily focused on mutations with codominant effects, thus necessarily emphasizing the hitchhiking of weak deleterious mutations with stronger advantageous ones. Although recessive strongly deleterious mutations are expected to have a lower probability of hitchhiking to fixation than weaker codominant ones (48), recessive deleterious mutations can still have a profound impact on the dynamics of adaptation.

Here we develop a model for the dynamics of a dominant beneficial mutation that is initially genetically linked to a recessive deleterious mutation of larger effect ($|s_d| > |s_b|$), with varying rates of recombination. Provided the recessive deleterious mutation is sufficiently hidden in the heterozygote ($|h_d s_d| < |h_b s_b|$), hitchhiking occurs even when the beneficial mutation has a smaller fitness effect than its deleterious hitchhiker. We show that the

Significance

A deleterious mutation that is recessive is hidden in individuals containing only one copy (i.e., heterozygotes); however, individuals containing two copies (i.e., homozygotes) suffer negative effects. This class of mutation is responsible for a number of human genetic disorders, including cystic fibrosis and Tay-Sachs, in addition to causing the widespread phenomenon of inbreeding depression. Evidence suggests that recessive deleterious mutations may be abundant in nature, likely due to their ability to persist for long timescales at moderate frequencies. It is thus possible that when an adaptive mutation occurs, it lands on a chromosome containing a recessive deleterious mutation. We propose a model for this and find that recessive deleterious mutations can significantly slow adaptation and alter signatures of adaptation.

Author contributions: Z.J.A., D.A.P., and J.R.B. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: zjassaf@gmail.com.

[2]D.A.P. and J.R.B. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1424949112/-/DCSupplemental.

frequency trajectory of a beneficial mutation in this case is dramatically altered, causing what we have termed a "staggered sweep," whereby the linked mutations are balanced for a period before recombination unlinks them. This balancing selection is a type of associative overdominance. However, instead of the classic case of two recessive deleterious mutations in repulsion or a neutral mutation linked to a single overdominant one (49, 50), our case results from a recessive deleterious mutation linked to a dominant beneficial mutation of weaker selective effect.

Using our model, we show that recessive deleterious mutations can substantially lower the fixation probability of nearby beneficial mutations, thus creating zones around recessive deleterious mutations where adaptation is suppressed. Although it has been observed before in models of balancing selection that alleles balanced at low frequencies in finite populations can drift out of a population with higher probability than a neutral mutation (51–53), we have derived, for the first time (to our knowledge), an analytic solution for the distribution of extinction times. In addition to affecting the fixation probability of a beneficial mutation, we demonstrate that a recessive deleterious hitchhiker can lengthen the duration of a selective sweep and alter the genomic sweep signature, compared with a classic hard sweep. We show that these effects are strongest in small populations, at low recombination rates, and when the recessive deleterious hitchhiker is much stronger than the beneficial mutation in homozygotes. We end by estimating how common this effect could be in *D. melanogaster* and humans, showing that it may play a potentially important role in adaptation for both natural and laboratory populations.

Results

We first develop a heuristic understanding of staggered sweeps as an interaction between three effects: balancing selection, recombination, and genetic drift. We then derive analytic predictions for the dynamics of staggered sweeps and compare our results with forward simulations. Throughout our analysis, we emphasize scaling and parameter dependence over the details of constant factors. This is intentional. Scaling properties hold generally across different models studied, whereas the constant factors typically do not. For example, the fixation probability of a beneficial mutation is frequently quoted as $2s$; however, the constant 2 depends on the details of the stochastic model (it is inversely proportional to the variance in offspring number). The scaling with the selective effect, s, however, is general. For this reason, we do not distinguish between constant factors and instead emphasize the scaling (e.g., fixation probability of $\sim s$).

Model. Consider a population of N diploid individuals in which two sites are genetically linked along a chromosome such that an ancestral haplotype with no mutations is denoted OO. The first site can harbor a beneficial mutation with selective effect s_b and heterozygous effect h_b, the second site can harbor a deleterious mutation with effects s_d and h_d, and there is a recombination rate $l \times r$ between them (base pair distance × recombination rate per base pair per generation). When a new adaptive mutation lands on a chromosome harboring an existing recessive deleterious mutation, it generates a BD haplotype, and in the absence of recombination (for the moment) a BD/OO diploid will have heterozygote advantage. This is apparent from the fitnesses of the diploids,

$$\text{fitness of } OO/OO \text{ diploid} \approx 0 \qquad [1]$$

$$\text{fitness of } BD/OO \text{ diploid} \approx h_b s_b - h_d s_d \approx s_b \qquad [2]$$

$$\text{fitness of } BD/BD \text{ diploid} \approx s_b - s_d \approx -s_d \qquad [3]$$

where we assume selection coefficients to be small ($s_b \ll 1$, $s_d \ll 1$) and the selection coefficient of the deleterious mutation to be much larger than that of the beneficial mutation ($s_b \ll s_d$), and, for the final approximation, we further assume the beneficial mutation to be completely dominant and the deleterious mutation to be completely recessive ($h_b \approx 1$, $h_d \approx 0$). We emphasize that our model and analytic predictions can be extended to cases of partial dominance (i.e., $h_d \neq 0$, $h_b \neq 1$), as long as the effect of the beneficial mutation in the heterozygote is stronger than that of the deleterious mutation ($h_b s_b > h_d s_d$) but its effect in the homozygote is weaker ($s_b < s_d$) (*SI Text*, sections 1 and 2).

The new beneficial mutation will thus be subject to balancing selection on the BD haplotype and held at an equilibrium frequency. However, this balanced state is temporary (Fig. 1). If recombination generates a BO haplotype, unlinking the beneficial mutation from the recessive deleterious hitchhiker, then the beneficial mutation can complete the sweep to fixation, sometimes with a substantial delay (Fig. 1*A*). We call this a staggered sweep. Alternatively, the BD haplotype may drift to extinction before the beneficial mutation is able to escape via recombination (Fig. 1 *B* and *C*, red trajectories). Extinction is particularly probable if the number of copies of the BD haplotype at equilibrium is small and therefore drift is strong. As these simulations suggest, the probability of fixation and the sweep duration for a new beneficial mutation with a recessive deleterious hitchhiker can be significantly altered depending on the interaction of balancing selection, recombination, and drift, which we now quantify.

Balancing selection. At low frequencies, the BD haplotype will primarily occur in heterozygotes where the recessive deleterious allele is hidden. As a result, the BD haplotype will be subject to positive selection. In contrast, at high frequencies, the BD haplotype will primarily occur in homozygotes where the recessive strongly deleterious allele is exposed, resulting in selection that drives the BD haplotype down in frequency. These opposing forces cause balancing selection. The change in frequency per generation due to selection, $S(p)$, is thus frequency dependent,

$$\text{strength of selection} = S(p) \approx s_b p - s_d p^2 + O(s_d p^3) \qquad [4]$$

(valid for $p \ll 1$; *SI Text*, section 1). The first term reflects selection on heterozygotes with fitness $\approx s_b$ that occur with probability $\approx p$, and the second term reflects selection on homozygotes with fitness $\sim -s_d$ that occur with probability $\sim p^2$ (note that the term $O(s_d p^3)$ will be ignored from this point forward as it is very small for $p \ll 1$). The stable fixed point or equilibrium frequency can be found by setting $S(p) = 0$,

$$\text{equilibrium frequency} = p^* \approx \frac{s_b}{s_d} \qquad [5]$$

(derivations of p^* and $S(p)$ for arbitrary selection and dominance coefficients can be found in *SI Text*, section 1). In the absence of drift and recombination, a BD haplotype that reaches equilibrium frequency will remain there indefinitely.

Escape via recombination. The beneficial mutation can reach fixation if two conditions are met. First, there must be a recombination event in a heterozygote BD/OO individual that creates a BO haplotype. Second, this newly created BO haplotype must survive the effects of drift (to "establish" in the population) and proceed toward fixation. We call this "escape" of the beneficial mutation. In each generation, the probability of escape is proportional to the number of copies of the BD haplotype [$\sim N\, p(t)$], multiplied by the probability of a recombination event creating a BO haplotype ($\sim rl$), and the probability of establishment of this new BO haplotype ($\sim s_b$). Thus, the total probability of an escape event by time t is

$$\text{probability of escape} \approx rls_b N \int_0^t p(t')\,dt' \qquad [6]$$

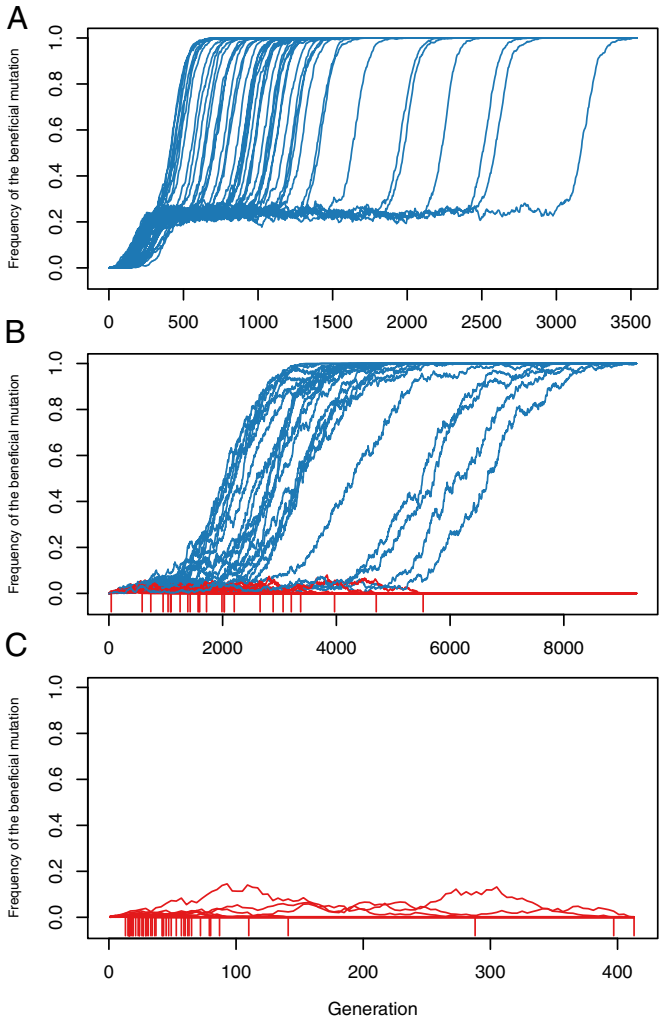

Fig. 1. Frequency trajectories of a beneficial mutation genetically linked to a recessive deleterious hitchhiker. Plotted trajectories are from 50 simulations that reached an equilibrium frequency of $p^* = s_b/s_d$, where blue indicates simulations that fixed the beneficial mutation and red indicates simulations in which it goes extinct (red tick marks below frequency zero mark the generation of extinction). (*A*) For $\alpha \gg 1$, selection dominates, and the *BD* haplotype staggers stably at the equilibrium frequency waiting for a recombination event that allows the beneficial mutation to escape on a *BO* haplotype to fixation. In this regime, loss of the beneficial mutation is very rare, and the staggered sweep can last for a substantial time. (*B*) For $\alpha \approx 1$, selection and drift are both important, and the *BD* haplotype staggers at the equilibrium frequency but with strong frequency fluctuations due to drift that can drive the *BD* haplotype to extinction (but not fixation) before an escape event occurs. (*C*) For $\alpha \ll 1$, drift dominates, and the *BD* haplotype never stably staggers; thus frequency changes are dominated by drift. Note that it cannot drift to fixation because of the recessive deleterious mutation (Fig. 2). Unless a recombination event occurs very early, the beneficial mutation will fluctuate to extinction on the *BD* haplotype.

(which is valid for a probability of escape $\ll 1$). Provided $p(t')$ is known, we can calculate the probability that the beneficial mutation escapes its deleterious background. We note that if the *BD* haplotype is held stably at equilibrium frequency, then $p(t') \approx p^*$. However, as discussed in the *Drift* section below, this is not always the case.

Drift. Despite the balancing selection driving the *BD* haplotype toward equilibrium frequency, sometimes the *BD* haplotype can go extinct before a recombination event occurs. This extinction event, which we call "loss," is driven by random fluctuations in frequency due to genetic drift. The variance in frequency per generation due to drift, $D(p)$, is

$$\text{strength of drift} = D(p) \approx \frac{p(1-p)}{N}. \quad [7]$$

To compare the relative effects of drift and selection, one can consider two characteristic timescales. The first, $\tau_S = \delta p/S$, is the time for selection alone to change the frequency of the balanced haplotype by δp. The second, $\tau_D = \delta p^2/D$, is the time for drift alone to change the frequency of the balanced haplotype by δp. The ratio of these two timescales tells one which of the two processes is faster, and hence which dominates the dynamics,

$$\frac{\text{speed of selection}}{\text{speed of drift}} = \frac{\tau_D}{\tau_S} = \delta p \frac{S}{D} = \delta p \frac{s_b p - s_d p^2}{p(1-p)/N}. \quad [8]$$

Because we are interested in changes in frequency that lead to extinction, the natural choice for δp is approximately p (*SI Text*, section 2). With this simplification, $\tau_D/\tau_S \approx N(s_b p - s_d p^2) = NS(p)$, which we can now use in Fig. 2 to understand the transition points between frequency ranges that are drift dominated or selection dominated. Selection will be more important whenever $S(p) = s_b p - s_d p^2$ (solid curve) is larger than $1/N$ (dashed lines), such that drift dominates for frequencies near 0,1, and p^* (light shading) and selection dominates in the alternate intervals (dark shading). Positive selection is strongest (causing the largest positive changes in frequency) at a frequency approximately midway to equilibrium, such that at frequency $p \approx p^*/2$ the strength of selection is $S_{max} \approx s_b^2/s_d$. As the relative strength of drift to selection becomes larger, the size of the drift intervals widens, even to the point where drift dominates for all frequencies below equilibrium (Fig. 2 *A* vs. *B*).

If drift is sufficiently strong, the *BD* haplotype can fluctuate to extinction. The probability that this occurs depends on the maximum strength of positive selection (S_{max}) relative to the strength of drift ($1/N$) (*SI Text*, section 2). We call this ratio α,

$$\alpha = \frac{\text{max speed of positive selection}}{\text{speed of drift}} = NS_{max} = \frac{Ns_b^2}{s_d}. \quad [9]$$

The dynamics of the *BD* haplotype are qualitatively different for $\alpha \gg 1$ and $\alpha \ll 1$. In the $\alpha \gg 1$ regime, there is a region below p^* where selection dominates (Fig. 2*A*), and thus the *BD* haplotype can establish in the population and stagger at the equilibrium frequency with relatively small fluctuations due to drift. In the $\alpha \ll 1$ regime, there is no region below p^* where selection dominates over drift (Fig. 2*B*), in which case there is no true establishment or balanced phase for the *BD* haplotype and it can easily go extinct. There is a crossover where $\alpha \approx 1$ in which selection and drift are of similar magnitudes. Below, we derive analytic expressions to predict the beneficial mutation's probability of fixation and total sweep duration in the two distinct regimes ($\alpha \gg 1$ and $\alpha \ll 1$) under the forces of selection, drift, and recombination. We confirm our results with simulations.

Predictions for the Regime of Strong Selection and Weak Drift ($\alpha \gg 1$).

Probability of fixation. In order for the beneficial mutation to reach fixation from equilibrium, it must avoid fluctuating to extinction before a recombination event can unlink it from its deleterious hitchhiker. The probability of fixation is therefore determined by which event occurs first: loss or escape.

Loss of the *BD* haplotype occurs only if it can fluctuate over the relatively high selective barrier (S_{max}). The process is similar to chemical reactions, where there is an "activation energy" and the reaction rate depends exponentially on the ratio of the barrier height to the strength of noise (e.g., Arrhenius' equation for

chemical reaction rates). Thus, it can be shown (*SI Text*, section 2.1) that the time to loss, independent of escape via recombination, is an exponentially distributed random variable with rate λ_l and mean τ_l

$$\text{rate of loss} = \lambda_l \approx \frac{s_b}{\ln[\alpha]} e^{-\alpha} \quad \text{mean loss time} = \tau_l \approx 1/\lambda_l, \qquad [10]$$

which we have verified using simulations (*SI Text*, section 4.1).

Escape of the beneficial mutation from equilibrium to fixation can also be modeled as an exponential process. In this regime of strong selection and weak drift, the *BD* haplotype is held closely to equilibrium; thus the probability of escape from Eq. **6** simplifies to $(rlNs_b)(s_b/s_d)t$. In this case, escape, independent of loss via drift, occurs at rate λ_e with mean τ_e,

$$\text{rate of escape} = \lambda_e \approx \frac{rlNs_b^2}{s_d} = rl\alpha \quad \text{mean escape time} = \tau_e \approx 1/\lambda_e, \qquad [11]$$

which we have verified using simulations (*SI Text*, section 4.2).

The overall probability of fixation for a new beneficial mutation that lands on a recessive deleterious background is then the product of the probability that a new *BD* haplotype establishes and reaches equilibrium ($\sim s_b$) and the probability that the first escape event occurs before the first loss event [$\approx \tau_l/(\tau_l + \tau_e)$],

$$\text{probability of fixation} \approx s_b \left(\frac{\tau_l}{\tau_l + \tau_e}\right), \qquad [12]$$

which we have verified using simulations [Fig. 3 *D–F* ($\alpha \gg 1$) and *SI Text*, section 4.4]. The probability of fixation of the adaptive mutation will be reduced relative to an identical adaptive mutation with no recessive deleterious hitchhiker whenever $\tau_l \lesssim \tau_e$, which is when loss is faster than escape.

Sweep time. If the beneficial mutation is not driven to extinction, the duration of a beneficial mutation's sweep to fixation can be substantially extended if it's genetically linked to a recessive deleterious hitchhiker (Fig. 1*A*). To understand when this occurs, consider that the total time of a successful sweep will, on average, be prolonged by the time it takes for an escape event to occur,

$$\text{average sweep time} \approx \frac{\ln[Ns_b]}{s_b} + \left(\frac{1}{\tau_e} + \frac{1}{\tau_l}\right)^{-1}, \qquad [13]$$

which we verified with simulations [Fig. 4 *C–F* ($\alpha \gg 1$) and *SI Text*, section 4.5]. The first term corresponds to the sweep time for a single adaptive mutation with no hitchhiker, and the second term is the extension in sweep time due to the staggered phase (for more discussion of the leveling off of sweep times at low recombination rates, as in Fig. 4 *C* and *D*, see *SI Text*, section 4.5).

Zones of altered adaptation around a recessive deleterious allele. We can understand the effect of recessive deleterious variation on the genome in terms of a base pair distance around every recessive deleterious mutation within which the dynamics of adaptation are altered (*SI Text*, section 3). A recessive deleterious hitchhiker will drastically suppress the probability of fixation of a linked beneficial mutation whenever $\tau_l < \tau_e$, translating to a distance, l_l, around the deleterious mutation within which new beneficial mutations of effect size s_b (or smaller) have a reduced chance of fixation ($l_l \approx \lambda_l/r\alpha$, red dashed lines in Fig. 3 *D–F*). This distance can be substantial for values of α that approach unity (Fig. 3*D*). Similarly, using Eq. **13**, we can derive a distance, l_e, around a recessive deleterious mutation within which new beneficial mutations of effect size s_b have an extended sweep time ($l_e \approx s_b/r\alpha \ln[Ns_b]$, red dashed lines in Fig. 4 *C–F*). Again, this distance is largest for smaller values of α (Fig. 4 *C* and *D*); however, the sweep duration is longest in large α regimes at small recombination rates (Fig. 4 *E* and *F*). Note that the overall impact of recessive deleterious mutations on adaptation will depend on the density and strength of these deleterious mutations and whether their zones of altered adaptation might overlap (further explored in *Discussion*).

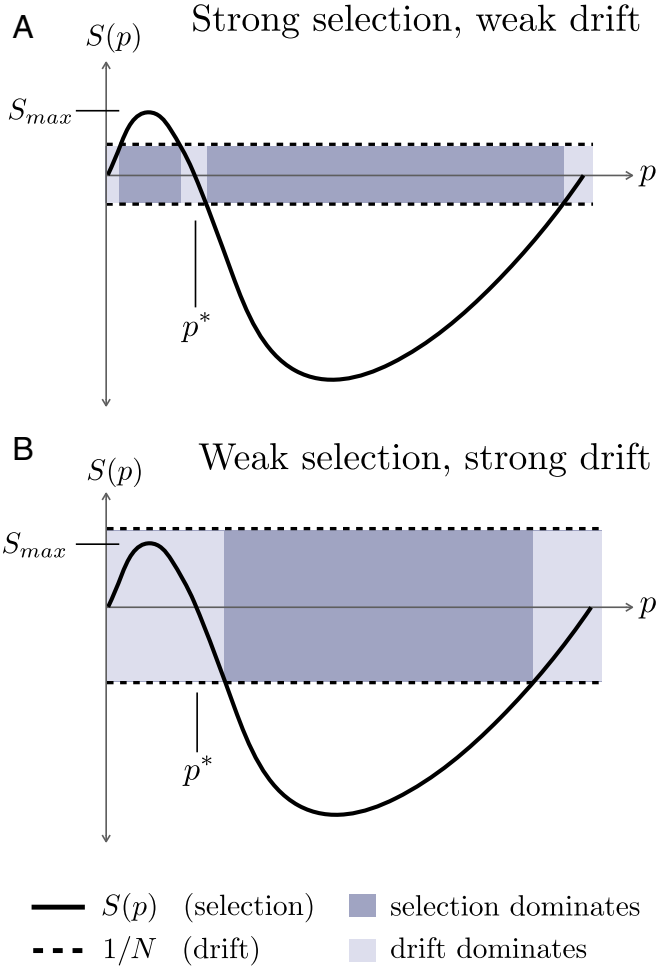

Fig. 2. A model of drift and balancing selection (without recombination). (*A*) Schematic of $\alpha = NS_{max} > 1$ regime, where the maximum positive effect of balancing selection ($S_{max} \approx s_b^2/s_d$) is stronger than the effect of drift ($1/N$). In this regime, the *BD* haplotype can balance at an equilibrium frequency and then either escape to fixation via recombination (as in Fig. 1 *A* and *B*, blue trajectories) or drift to extinction (as in Fig. 1*B*, red trajectories). The rate of drifting to extinction from equilibrium depends primarily on $\alpha \approx Ns_b^2/s_d$. (*B*) Schematic of $\alpha = NS_{max} < 1$ regime, where the maximum positive effect of balancing selection (S_{max}) is weaker than the effect of drift ($1/N$), and thus frequency dynamics of a balanced haplotype are always dominated by drift near (or below) equilibrium frequency. In this regime, the *BD* haplotype neither establishes nor stably balances, and thus primarily drifts to extinction before an escape event occurs (as in Fig. 1*C*).

Predictions for the Regime of Weak Selection and Strong Drift ($\alpha \ll 1$).

Probability of fixation. In this regime, selection for the beneficial mutation is not strong enough to hold the *BD* haplotype close to equilibrium frequency (as in Fig. 1*C*). Instead, the dynamics of the staggered sweep are dominated by a combination of drift (due to $1/N > S_{max}$) and selection against the recessive deleterious mutation. In this case, the beneficial mutation on the *BD* haplotype can easily drift to extinction, but it is very unlikely to drift to fixation. It can be shown (*SI Text*, section 2.3) that the upper limit to which the *BD* haplotype will typically drift is

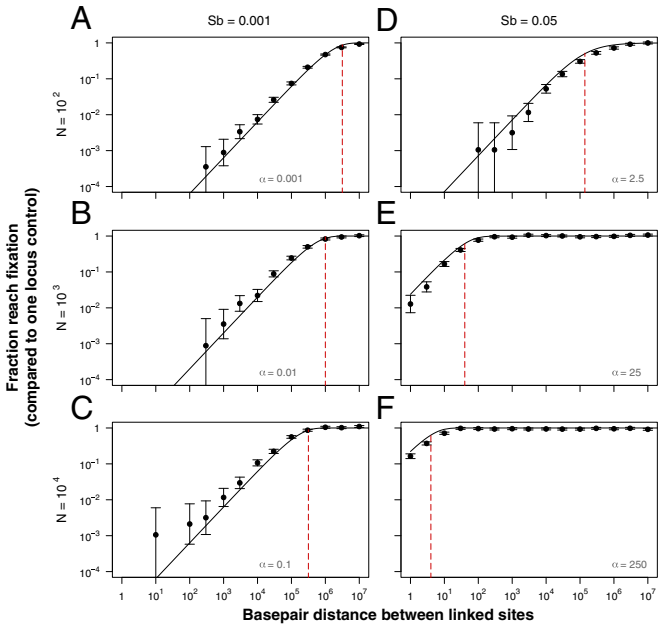

Fig. 3. Analytics predict simulation results for the fraction of staggered sweeps that reach fixation. The *y* axis is the fraction of simulations in which the beneficial mutation reaches fixation relative to a one-locus control with no hitchhiker, and the *x* axis is base pair distance between the linked sites. The population size (N) used in each panel is indicated by the figure row headings, and the beneficial mutation effect size (s_b) used in each panel is indicated by the figure column headings, such that *A*–*C* fall into the $\alpha < 1$ regime, and *D*–*F* fall into the $\alpha > 1$ regime. Points represent results of 1,000/s_b simulations (where bars indicate 95% binomial proportion confidence interval), solid lines indicate our analytic predictions (*A*–*C* use Eq. **S31** due to $\alpha < 1$, and *D*–*F* use Eq. **12** due to $\alpha > 1$). Red dashed lines are analytic predictions for the distance l_l below which the probability of fixation becomes suppressed (*A*–*C* use Eq. **S37** and *D*–*F* use Eq. **S33**). All simulations used $s_d = -0.1$. We have translated recombination rate $r \times l$ between the sites into base pair distance using a human recombination rate per base pair per generation $r = 10^{-8}$.

$\sim \sqrt{N/s_d}$ copies, at which point selection against homozygotes becomes stronger than drift and pushes the *BD* haplotype back down to lower frequencies. Thus, the distribution of extinction times of the *BD* haplotype will be approximately neutral (power-law distributed as $\sim 1/t^2$), but it cannot be much longer than $\sim \sqrt{N/s_d}$ generations (verified using simulations, *SI Text*, section 4.3).

We are interested in the probability that escape occurs before loss of the *BD* haplotype. Consider an interval dt in which there is a fraction of *BD* haplotypes that go extinct (which scales as dt/t^2 from neutrality), and a fraction of *BO* haplotypes that escape via recombination (which scales as $rls_b t^2$ from the total number of *BD* haplotypes that have existed by time t). The product of these two gives a constant probability of escape ($\sim rls_b\, dt$). Recalling that the *BD* haplotype will typically drift for only $\sim \sqrt{N/s_d}$ generations, we can use Eq. **6** to calculate the probability of fixation for the beneficial mutation to be $\sim rls_b \sqrt{N/s_d}$ (*SI Text*, section 2.3). We have verified with simulations that this predicts the probability of fixation of the beneficial mutation [Fig. 3 *A*–*C* ($\alpha \ll 1$) and *SI Text*, section 4.4].

Sweep time. Although the probability of fixation for the beneficial mutation can be significantly decreased in the $\alpha \ll 1$ regime, the total sweep time of the beneficial mutation is not expected to substantially change. This is because if it is to escape at all, it must do so in the first $\sim \sqrt{N/s_d}$ generations, which is generally a small fraction of the classic sweep time $\sim \ln(Ns_b)/s_b$. This prediction of no alteration in sweep time in the $\alpha \ll 1$ regime was confirmed with simulations (Fig. 4 *A* and *B* and *SI Text*, section 4.5).

Zones of altered adaptation around a recessive deleterious allele. We can again understand the effect of recessive deleterious variation on the genome in terms of a base pair distance around every recessive deleterious mutation within which the dynamics of adaptation are altered. Because the sweep time is not significantly altered in this regime of weak selection and no drift, there is no relevant distance within which sweep time is extended. However, a recessive deleterious hitchhiker will drastically suppress the probability of fixation of a linked beneficial mutation whenever $rl\sqrt{N/s_d} < 1$ (*SI Text*, sections 3.2 and 4.4). The distance in this case ($l_l \approx r^{-1}\sqrt{s_d/N}$) can be substantial for realistic N, s_d, and r. For example, all panels in the $\alpha \ll 1$ regime in Fig. 3 (Fig. 3 *A*–*C*) show that a beneficial mutation will have a reduction in its fixation probability even when it is ~1 Mb away from the recessive deleterious hitchhiker.

Simulations. To test our predictions, we conducted two-locus Wright–Fisher forward simulations (Figs. 3 and 4 and *SI Text*, sections 2 and 4). Simulations start by seeding the *BD* haplotype at 1 copy and the *OO* haplotype at $N-1$ copies, and recording the frequency of the four possible haplotypes over time. Simulations were performed over a wide range of selection coefficients ($s_b = 0.001 - 0.1$, $s_d = 0.01 - 1$), recombination rates ($lr = 10^{-8} - 10^{-1}$), and population sizes ($N = 10^2 - 10^5$), corresponding to ranges of $\alpha = Ns_b^2/s_d = 10^{-3} - 10^5$. All simulations used $s_b \le s_d$ and heterozygous effects of $h_d = 0$ and $h_b = 0.5$ such that the equilibrium frequency $p^* \approx s_b/2s_d \le 1/2$ (analytics were changed appropriately to allow for $h_b = 0.5$). We also included a one-locus control of an adaptive mutation with no deleterious hitchhiker for comparison. Note that figures are plotted using $r = 10^{-8}$ (corresponding to 1 cM/Mbp).

Figs. 3 and 4 show that our analytic expressions (black lines) predict simulation results (data points) and thus accurately capture the parameter dependence. Furthermore, the simulations confirm that an adaptive mutation that lands within a distance l_l of the recessive deleterious mutation (red dashed line in Fig. 3) has a reduced probability of fixation, and an adaptive mutation that lands within a distance l_e of a recessive deleterious mutation (red dashed line in Fig. 4) has an increased sweep time.

The effect of the recessive deleterious mutation on the probability of fixation of linked beneficial mutations can extend for very large genomic distances, especially for weakly adaptive sites (Fig. 3 and *SI Text*, sections 3.1, 3.2, and 4.4). For instance, when $s_d = 0.1$ and $N = 10^4$, any adaptive mutation with $s_b \le 0.001$ will satisfy $\alpha < 1$ and thus have a fixation probability that is substantially reduced within $l_l \approx 3 \times 10^5$ base pairs of the recessive deleterious mutation (Fig. 3*C*). However, if the effect size of the adaptive mutation is increased to $s_b = 0.05$, the probability of fixation recovers to about the same level as an adaptive mutation with no deleterious hitchhiker (Fig. 3*F*). In general, any recessive deleterious mutation will suppress fixation of nearby adaptive mutations if $s_b \le \sqrt{s_d/N}$ is satisfied.

The sweep time of adaptive mutations that reach fixation, on the other hand, is impacted for small α values that still satisfy $\alpha > 1$ (Fig. 4 and *SI Text*, sections 3.1, 3.2, and 4.5). In this regime, the adaptive mutation can reach a stable equilibrium frequency (due to $\alpha > 1$), where it has a slow rate of loss ($\tau_l > \ln[Ns_b]/s_b$) but the rate of new recombinants being generated in the population is not so large that the staggered sweep is resolved quickly ($\tau_e > \ln[Ns_b]/s_b$). For example, when a new adaptive mutation of effect $s_b = 0.05$ lands 10^3 base pairs away from a recessive deleterious mutation of effect $s_d = 0.1$ in a population of $N = 10^3$, the mean sweep time is ~15 times as long compared with a new adaptive mutation with no hitchhiker (Fig. 4*E*). In this scenario, where $\alpha = 25$, an extension in sweep time will occur even if the adaptive mutation lands within a distance $l_e \approx 10^4$ base pairs of the recessive deleterious mutation. As α increases, this distance

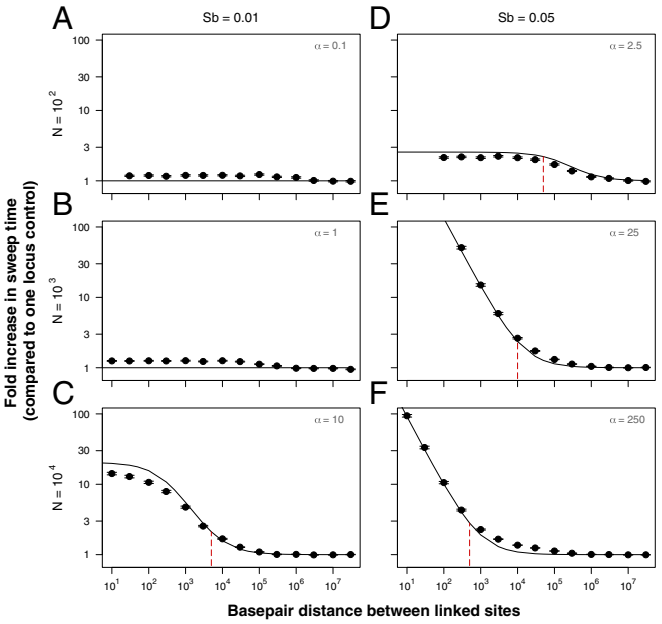

Fig. 4. Analytics predict simulation results for the mean sweep time of staggered sweeps. The *y* axis is the fold increase in the beneficial mutation's sweep time relative to a one-locus control with no hitchhiker, and the *x* axis is the base pair distance between the linked sites. The population size (N) used in each panel is indicated by the figure row headings, and the beneficial mutation effect size (s_b) used in each panel is indicated by the figure column headings, such that *A* and *B* have $\alpha \leq 1$, and *C*–*F* have $\alpha > 1$. Points represent results of 500 simulations in which fixation of the beneficial mutation occurred (required to calculate its sweep time), where bars indicate ±SE and solid lines indicate our analytic predictions (*A* and *B* have $\alpha < 1$ and thus no increase in sweep time; *C*–*F* have $\alpha > 1$ and thus use Eq. **13**). Red dashed lines are analytic predictions for the distance l_e at which the mean sweep time becomes extended (Eq. **S35** and *SI Text*, section 3.2). For more discussion of the leveling off of sweep times at low recombination rates in *C* and *D*, see *SI Text*, section 4.5. All simulations used $s_d = -0.1$. We have translated recombination rate $r \times l$ between the sites into base pair distance using a human recombination rate per base pair per generation $r = 10^{-8}$.

decreases such that the duration of staggered sweeps are most extended for intermediate values of α (Fig. 4*E*).

Discussion

Using a two-locus model, we have shown that recessive deleterious mutations can (*i*) decrease the rate of fixation and (*ii*) increase the sweep time of linked adaptive mutations with similar or weaker fitness effects. Here we discuss to what extent these two effects are likely to occur in real populations and how they can alter signatures of selection.

Impact of Recessive Deleterious Mutations in Real Populations. Every recessive deleterious mutation has a zone around it within which adaptive events are suppressed (*SI Text*, section 3), and thus their overall effect on adaptation will critically depend on the density and strength of these deleterious mutations. Although these mutations are likely to appear in functional regions of the genome, their actual densities are not yet well understood. Thus, focusing on coding genes, we use what data can be found for the densities of deleterious mutations for both *D. melanogaster* and humans to derive order-of-magnitude estimates for the impact of recessive deleterious variation on adaptation in real populations (Table 1 and *SI Text*, section 3.5).

Data on the abundance of recessive lethals ($s_d = 100\%$) are perhaps the most unambiguous. They are thought to occur at a rate of $\sim 0.3 - 1.0$ per genome in both humans and *Drosophila* (24, 54), which translates to 1 per ~20,000 coding genes in humans and 1 per ~12,000 coding genes in *Drosophila*. The density of mildly deleterious recessive mutations is less well characterized; however, they are thought to be much more abundant (14). One estimate in *Drosophila* places the number of mildly deleterious recessive mutations ($s_d \sim 1\%$) at about 200 per autosome (20), which is 1 per ~30 coding genes. There are no data like these for humans yet; thus we use the *Drosophila* density as a starting point for a range of possible densities. The size of the zones around these deleterious mutations depends on multiple parameters (N, s_d, s_b, r), where, for example, a recessive lethal in a natural *Drosophila* population impacts a region of ~100 kb, but in humans, the size of this zone is ~1 Mb (Table 1, column 4).

The overall impact of recessive deleterious mutations on adaptation will depend not only on their densities but also on the density of functional regions (in this case, coding genes) around each deleterious mutation. Estimates of gene density are currently quoted at about 1 gene per ~100 kb in humans and 1 gene per ~10 kb in *Drosophila* (55, 56), yet a significant proportion of genes are clustered in both organisms, especially in humans (57, 58). Thus, we frame our estimates in terms of coding gene densities within a given zone size, where we used the *Drosophila* and human reference genomes to estimate the gene density around every coding gene (column 4 brackets in Table 1; method in *SI Text*, section 3). For example, a recessive lethal in a human population is likely to suppress adaptation for ~1 Mb around itself, which contains on average ~20 other coding genes (*SI Text*, section 3.3). Given the density of recessive lethals, this translates to only about 0.1% of human coding genes within which adaptation is suppressed (Table 1).

Although the impact of recessive lethals is unlikely to be very dramatic, mildly deleterious mutations paint a very different picture (Table 1). For example, a mildly deleterious mutation in a wild *Drosophila* population affects about 10 kb around itself, within which there are typically three other coding genes (*SI Text*, section 3.3). This translates to ~10% of coding genes within *Drosophila* that are impacted. These effects become exacerbated in the context of smaller population sizes, as might be seen in experimental populations. For example, in a *Drosophila* population of $N = 1{,}000$ flies, beneficial mutations with effect size ≤0.3% will have suppressed fixation probabilities in the entirety of the *Drosophila* genome (Table 1). If we consider the beneficial mutations that do fix (*SI Text*, section 3.5), we find that beneficial mutations with comparable selective effect to the deleterious mutation ($s_b \approx 1\%$) will be subject to staggered phases within the majority of the genome (~65% of coding genes). This is interesting in light of a number of experimental evolutions in small populations of *Drosophila* ($N \approx 10^2$–10^3), where studies have often shown a number of alleles that initially increase in frequency with a rate and direction suggestive of selection but then do not finish the sweep to fixation (59, 60) [older experiments also exhibited similar patterns (61–65)]. Possible explanations [including individually overdominant loci or selection on polygenic traits (66)] are still under investigation; however, it is plausible that this behavior could in part be due to linked recessive deleterious alleles in small populations causing staggered sweeps.

For humans, there are no current estimates for the densities of weakly deleterious recessive mutations; however, if we consider a range of densities from 1 mildly deleterious mutation every 10 coding genes to every 100 coding genes, we find that anywhere from 3% to 30% of coding genes may be subject to a reduced rate of adaptation (Table 1). This result strongly emphasizes the need for more information regarding the actual numbers of weakly deleterious recessive mutations segregating in human populations, as their combined effect could potentially result in a significantly suppressed rate of fixation of weakly adaptive mutations, particularly in small populations. We note that in classical population genetics, the term $Ns_b > 1$ is generally required

Table 1. Estimates of the the proportion of the genome in which the probability of fixation of the beneficial mutation is decreased for *Drosophila melanogaster* and humans

Organism	Number of coding genes	Recessive deleterious effect (s_d), %	Zone of reduced adaptation [genes in zone]	Beneficial effect (s_b) impacted, %	Density of recessive deleterious	Proportion of adaptive mutations impacted, %
Drosophila	~12,000	100	100 kb [20 genes]	≤0.10	1/(genome)	~0.1
(wild, $N = 10^6$)		1	10 kb [3 genes]	≤0.01	1/(30 genes)	~10
Drosophila	~12,000	100	3 Mb [400 genes]	≤3.00	1/(genome)	~3
(laboratory, $N = 10^3$)		1	300 kb [40 genes]	≤0.30	1/(30 genes)	~100
Human ($N = 10^4$)	~20,000	100	1 Mb [20 genes]	≤1.00	1/(genome)	~0.1
		1	100 kb [3 genes]	≤0.10	1/(100 genes)	~3
					1/(30 genes)	~10
					1/(10 genes)	~30

Due to variation in functional density across genomes and organisms, we frame our estimates in terms of coding genes and their densities, using the *Drosophila* and human reference genomes (*SI Text*, section 3). Column information is as follows: column 1 indicates the species and population size considered, column 2 indicates the number of coding genes in a haploid set of autosomes, column 3 indicates the recessive deleterious mutation effect size of interest, column 4 indicates the zone around this recessive deleterious mutation within which adaptation is suppressed (*SI Text*, sections 3.1 and 3.2), where the bracketed information is the number of coding genes that typically appear in a region of this size (centered on a coding gene, *SI Text*, section 3.3), column 5 indicates the beneficial mutation effect size of interest (where we consider all beneficial mutations which fall within the $\alpha < 1$ regime because they behave similarly and are greatly impacted), column 6 indicates the densities of recessive deleterious mutations as obtained from refs. 20, 24, and 54, where we use a range of possible densities for mildly deleterious mutations in humans due to a lack of information, and column 7 indicates the proportion of new adaptive mutations of the given effect size impacted (i.e., the proportion of coding genes in a genome within which adaptation is suppressed), which can be substantial.

to be satisfied for adaptation to proceed. However, in a population with densely distributed recessive mutations, the requirement becomes $Ns_b^2/s_d > 1$, a factor s_b/s_d smaller. If genomes are indeed rich in recessive deleterious variation, it may be the case that beneficial mutations must have substantially larger fitness effects ($s_b > \sqrt{s_d/N}$ instead of $s_b > 1/N$) to spread through a population.

Altered Signatures of Selection. A beneficial mutation that begins its selective sweep with a linked recessive deleterious hitchhiker may be expected to leave an altered genomic signature of selection upon reaching fixation compared with a hard sweep. Under the classic hard sweep model of adaptation, a single de novo adaptive mutation occurs on a single haplotype and drives it to fixation. As a result, diversity in the immediate vicinity of the adaptive site is expected to be greatly reduced at the completion of the sweep and should recover to background levels farther away in the genome (67). In contrast, during a successful staggered sweep in which recombination must unlink a recessive deleterious hitchhiker, the haplotype that begins the sweep will always be distinct from the haplotype that finishes the sweep. It is thus possible in this scenario for both haplotypes to persist in the population after fixation occurs, producing higher levels of haplotype diversity around the beneficial mutation.

To illustrate this point, we simulated staggered sweeps, hard sweeps, and soft sweeps with linked neutral diversity using SLiM, a program for forward population genetic simulations of linked loci (68) (soft sweep simulations used a high beneficial mutation rate, such that beneficial mutations occurred on multiple haplotypes and swept concurrently). Upon fixation of the beneficial mutation, we plotted patterns of heterozygosity around the adaptive site, and found that staggered sweeps leave distinct signatures (Fig. 5; additional statistics in *SI Text*, section 5). At the conclusion of a staggered sweep, there can be multiple haplotypes present in the population at substantial frequencies (Fig. 5 *A* vs. *B*), a signature that is qualitatively similar to that generated by a soft or partial sweep. Additionally, compared with both hard and soft sweeps, staggered sweeps consistently leave higher levels of heterozygosity above the region where the recessive deleterious hitchhiker was unlinked (Fig. 5*C*). This asymmetry is caused by the requirement that, in order for fixation to occur, at least two haplotypes must participate in the selective sweep on the side of the beneficial mutation that contained the deleterious hitchhiker. These staggered sweep signatures are interesting in light of the many studies that have found partial sweeps (69–71), a signature characterized by a single beneficial mutation (or haplotype, if the mutation is not yet identified) that has spread through a population but not yet reached fixation. Our model suggests that linked recessive deleterious variation could potentially be another factor contributing to such signatures observed in natural populations.

Possible Extensions to Our Model. Our model is unique in that our assumption of recessivity allows for the consideration of deleterious mutations with appreciable selective effects, a class of deleterious mutations that are observed in natural populations and yet are generally precluded from hitchhiking models that assume codominance (due to $s_d > s_b$). Thus, for a deleterious mutation with a given effect size s_d, a linked beneficial mutation should have a higher probability of fixation if that deleterious hitchhiker is recessive. This is due to the fact that the *BD* haplotype can potentially establish (as in the $\alpha > 1$ regime) or drift at low frequencies (as in the $\alpha < 1$ regime), which, in either case, allows the beneficial mutation to persist for longer times in the population (and thus have more chances at escape) than if its hitchhiker had codominant effects. One possible extension would be to require the deleterious mutation to be weaker ($s_d \ll s_b$), which would not cause temporary balancing selection, as in our model, but may allow the beneficial mutation on the *BD* haplotype to initially establish with a higher probability than if the hitchhiker had codominant effects (i.e., $\sim s_b$ instead of $\sim s_b - s_d$). Another interesting extension would be to build a model with an arbitrary number, n, of deleterious sites, in which the deleterious effect is correlated with the dominance coefficient. This would better reflect what is observed in nature, and additionally would allow for the interesting case where many neighboring weakly deleterious recessive mutations can cumulatively cause a balanced state (if $ns_d > s_b$), and which would also require longer timescales to be unlinked. In general, the ability of recessive deleterious mutations to hitchhike to appreciable frequencies suggests they could play a role in the dynamics of rapid adaptation and potentially in the maintenance of genetic variation. Our model of a single beneficial and deleterious mutation is appropriate if mutation rates are small (i.e., $N\mu \ll 1$). An interesting extension would thus be to incorporate arbitrary rates of beneficial and deleterious mutations (similar to ref. 30),

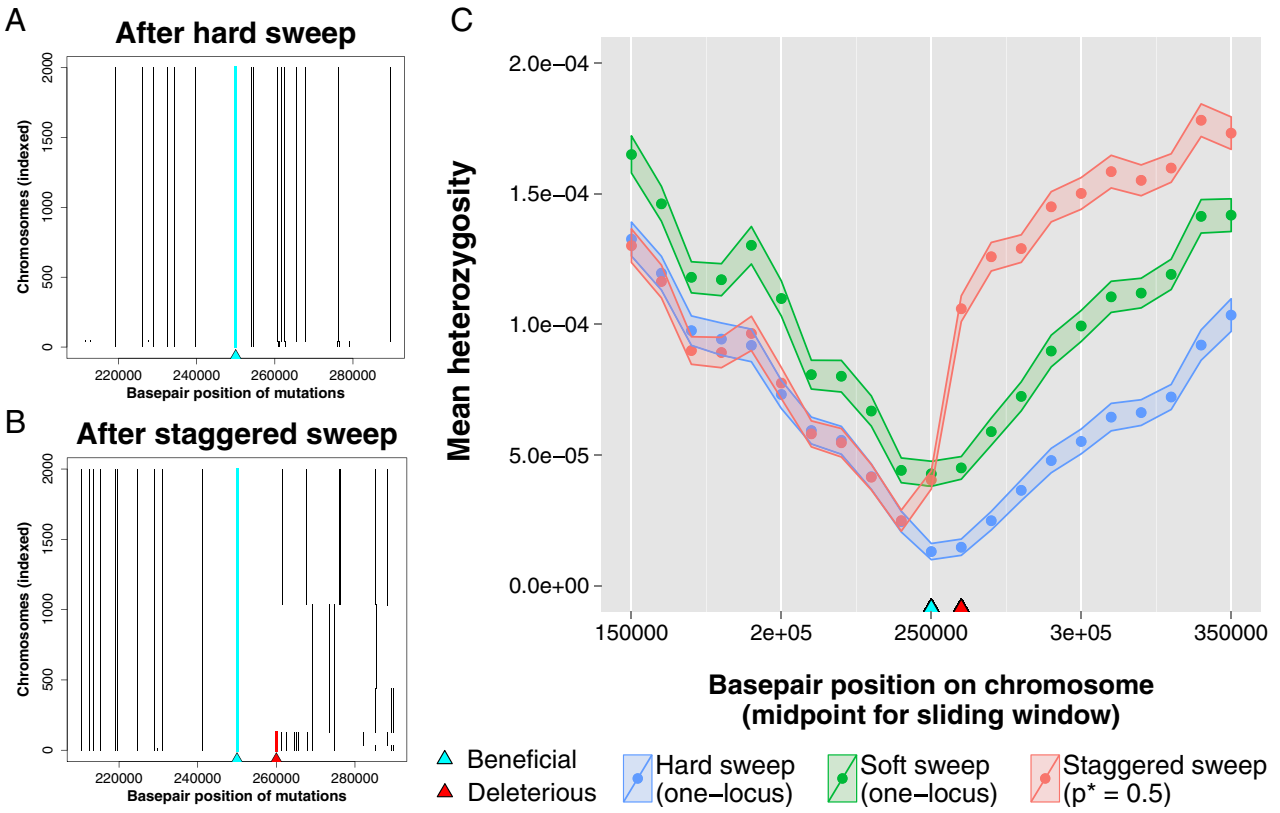

Fig. 5. Altered signatures of selection in the genome after a staggered sweep. Simulations were performed using SliM (68) to generate and track neutral diversity around the adaptive site, where simulations used $s_b = 0.05$ and $N = 1{,}000$ diploids. (*A*) Haplotypes present in a population at the conclusion of single simulation of a hard sweep in which a beneficial mutation on a haplotype containing only neutral mutations was seeded at establishment frequency; note that a single haplotype dominates the population. (*B*) Haplotypes present in a population at the conclusion of single simulation of a staggered sweep in which a beneficial mutation on a haplotype containing both neutral mutations and a single recessive deleterious mutation ($s_d = 0.05$) 10 kb away (where $r = 10^{-8}$) was seeded at establishment frequency. Note that recombination has unlinked the beneficial and recessive deleterious mutations, such that multiple haplotypes are at high frequency in the population after fixation of the beneficial mutation. (*C*) Mean heterozygosity across 200 simulations calculated in sliding windows of length 30 kb with step size 10 kb, where the ribbon around data points indicates the SEM. Results are plotted for hard sweep simulations in which a new adaptive mutation occurs on a single haplotype, soft sweep simulations in which a new adaptive mutation occurs on multiple haplotypes ($Nu_b \approx 1$), and staggered sweep simulations in which an adaptive mutation occurs on a single haplotype background containing a recessive deleterious mutation ($s_d = 0.05$).

allowing for beneficial mutations to be so frequent that when a new adaptive event occurs the effects of the previous one may not have yet been "forgotten."

Conclusions. Studies of linkage interference due to deleterious hitchhikers have largely been concerned with mutations of weak effects, as these are able to hitchhike to fixation with a beneficial mutation of larger effect. However, we have shown here that recessive mutations, which behave like weakly deleterious mutations at low frequencies but like strongly deleterious mutations at high frequencies, can significantly interfere with the rate and dynamics of adaptation. We find that a single recessive deleterious mutation will inhibit adaptation for large genomic distances around itself, and linked adaptive mutations sweeping through a population may stagger at an intermediate frequency for an extended time before reaching fixation. The consequences of recessive deleterious variation for adaptation are amplified in small populations, for closely linked sites, for weakly adaptive events, and for populations that harbor substantial recessive deleterious load. Although definitive experimental data for staggered sweeps are yet to be acquired, the evidence for abundant recessive deleterious variation from both natural populations (inbreeding depression data) and de novo mutations (mutation accumulation and mutagenesis experiments) suggests that staggered sweeps may be important during adaptation.

ACKNOWLEDGMENTS. The authors would like to thank Daniel Fisher for helpful discussions. We would also like to thank Gavin Sherlock, Hua Tang, Carlos Bustamante, Philipp Messer, Mike McLaren, and all members of the D.A.P. laboratory, as well as Daniel Weissman, Michael Desai, Matthew Hartfield, Deborah Charlesworth, Brian Charlesworth, and Nick Barton for useful feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-114747 (to Z.J.A.) and by National Institutes of Health Grants RO1GM100366 and RO1GM097415 (to D.A.P.).

1. Haldane JBS (1927) A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Math Proc Cambridge Philos Soc* 23(7):838–844.
2. Charlesworth B (1998) Adaptive evolution: The struggle for dominance. *Curr Biol* 8(14):R502–R504.
3. Fisher RA (1928) The possible modification of the response of the wild type to recurrent mutations. *Am Nat* 62(1):115–116.
4. Wright S (1934) Physiological and evolutionary theories of dominance. *Am Nat* 68(714):24–53.
5. Kacser H, Burns JA (1981) The molecular basis of dominance. *Genetics* 97(3-4):639–666.
6. Manna F, Martin G, Lenormand T (2011) Fitness landscapes: An alternative theory for the dominance of mutation. *Genetics* 189(3):923–937.
7. Mukai T, Chigusa S, Yoshikawa I (1965) The genetic structure of natural populations of *Drosophila melanogaster*. 3. Dominance effect of spontaneous mutant polygenes controlling viability in heterozygous genetic backgrounds. *Genetics* 52(3):493–501.

8. Houle D, Hughes KA, Assimacopoulos S, Charlesworth B (1997) The effects of spontaneous mutation on quantitative traits. II. Dominance of mutations with effects on life-history traits. *Genet Res* 70(1):27–34.
9. García-Dorado A, Caballero A (2000) On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* 155(4):1991–2001.
10. Chavarrías D, López-Fanjul C, García-Dorado A (2001) The rate of mutation and the homozygous and heterozygous mutational effects for competitive viability: A long-term experiment with *Drosophila melanogaster*. *Genetics* 158(2):681–693.
11. Fry JD, Nuzhdin SV (2003) Dominance of mutations affecting viability in *Drosophila melanogaster*. *Genetics* 163(4):1357–1364.
12. Peters AD, Halligan DL, Whitlock MC, Keightley PD (2003) Dominance and overdominance of mildly deleterious induced mutations for fitness traits in *Caenorhabditis elegans*. *Genetics* 165(2):589–599.
13. Szafraniec K, Wloch DM, Sliwa P, Borts RH, Korona R (2003) Small fitness effects and weak genetic interactions between deleterious mutations in heterozygous loci of the yeast *Saccharomyces cerevisiae*. *Genet Res* 82(1):19–31.
14. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nat Rev Genet* 10(11):783–796.
15. Ives PT (1945) The genetic structure of American populations of *Drosophila melanogaster*. *Genetics* 30(2):167–196.
16. Band HT, Ives PT (1963) Comparison of lethal + semilethal frequencies in second and third chromosomes from a natural population of *Drosophila melanogaster*. *Can J Genet Cytol* 14:351–357.
17. Mukai T, Cardellino RA, Watanabe TK, Crow JF (1974) The genetic variance for viability and its components in a local population of *Drosophila melanogaster*. *Genetics* 78(4):1195–1208.
18. Mukai T, Yamaguchi O (1974) The genetic structure of natural populations of *Drosophila melanogaster*. XI. Genetic variability in a local population. *Genetics* 76(2):339–366.
19. Simmons MJ, Crow JF (1977) Mutations affecting fitness in *Drosophila* populations. *Annu Rev Genet* 11:49–78.
20. Latter BD (1998) Mutant alleles of small effect are primarily responsible for the loss of fitness with slow inbreeding in *Drosophila melanogaster*. *Genetics* 148(3):1143–1158.
21. Sved JA (1971) An estimate of heterosis in *Drosophila melanogaster*. *Genet Res* 18(1): 97–105.
22. Sved JA (1975) Fitness of third chromosome homozygotes in *Drosophila melanogaster*. *Genet Res* 25(2):197–200.
23. Wilton AN, Sved JA (1979) X-chromosomal heterosis in *Drosophila melanogaster*. *Genet Res* 34(3):303–315.
24. Kusakabe S, Yamaguchi Y, Baba H, Mukai T (2000) The genetic structure of the Raleigh natural population of *Drosophila melanogaster* revisited. *Genetics* 154(2):679–685.
25. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269–294.
26. Birky CW, Jr, Walsh JB (1988) Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci USA* 85(17):6414–6418.
27. Otto SP, Barton NH (1997) The evolution of recombination: removing the limits to natural selection. *Genetics* 147(2):879–906.
28. Barton NH (2010) Genetic linkage and natural selection. *Philos Trans R Soc Lond B Biol Sci* 365(1552):2559–2569.
29. Neher RA, Shraiman BI, Fisher DS (2010) Rate of adaptation in large sexual populations. *Genetics* 184(2):467–481.
30. Good BH, Desai MM (2014) Deleterious passengers in adapting populations. *Genetics* 198(3):1183–1208.
31. Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78(2): 737–756.
32. Peck JR (1994) A ruby in the rubbish: Beneficial mutations, deleterious mutations and the evolution of sex. *Genetics* 137(2):597–606.
33. Barton NH (1995) A general model for the evolution of recombination. *Genet Res* 65(2):123–145.
34. Roze D, Barton NH (2006) The Hill-Robertson effect and the evolution of recombination. *Genetics* 173(3):1793–1811.
35. Hartfield M, Keightley PD (2012) Current hypotheses for the evolution of sex and recombination. *Integr Zool* 7(2):192–209.
36. Hadany L, Feldman MW (2005) Evolutionary traction: The cost of adaptation and the evolution of sex. *J Evol Biol* 18(2):309–314.
37. Hartfield M, Otto SP (2011) Recombination and hitchhiking of deleterious alleles. *Evolution* 65(9):2421–2434.
38. Luksza M, Lässig M (2014) A predictive fitness model for influenza. *Nature* 507(7490): 57–61.
39. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA (2013) Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci USA* 110(8): 2910–2915.
40. Lang GI, et al. (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500(7464):571–574.
41. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nat Rev Genet* 4(6):457–469.
42. Chun S, Fay JC (2011) Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet* 7(8):e1002240.
43. Williamson SH, et al. (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3(6):e90.
44. Huff CD, et al. (2012) Crohn's disease and genetic hitchhiking at IBD5. *Mol Biol Evol* 29(1):101–111.
45. Fay JC (2013) Disease consequences of human adaptation. *Appl Transl Genom* 2: 42–47.
46. Cruz F, Vilà C, Webster MT (2008) The legacy of domestication: Accumulation of deleterious mutations in the dog genome. *Mol Biol Evol* 25(11):2331–2336.
47. Lu J, et al. (2006) The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends Genet* 22(3):126–131.
48. Hartfield M, Glémin S (2014) Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. *Genetics* 196(1):281–293.
49. Clegg MT (1978) Dynamics of correlated genetic systems. II. Simulation studies of chromosomal segments under selection. *Theor Popul Biol* 13(1):1–23.
50. Ohta T, Kimura M (1970) Development of associative overdominance through linkage disequilibrium in finite populations. *Genet Res* 16(2):165–177.
51. Robertson A (1962) Selection for heterozygotes in small populations. *Genetics* 47:1291–1300.
52. Ewens W, Thomson G (1970) Heterozygote selective advantage. *Ann Hum Genet* 33(4):365–376.
53. Nei M, Roychoudhury AK (1973) Probability of fixation and mean fixation time of an overdominant mutation. *Genetics* 74(2):371–380.
54. Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M (2015) An estimate of the average number of recessive lethal mutations carried by humans. *Genetics* 199(4): 1243–1254.
55. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
56. Adams MD, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287(5461):2185–2195.
57. Adachi N, Lieber MR (2002) Bidirectional gene organization: A common architectural feature of the human genome. *Cell* 109(7):807–809.
58. Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol* 1(1):5.
59. Orozco-terWengel P, et al. (2012) Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol Ecol* 21(20):4931–4941.
60. Burke MK, et al. (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467(7315):587–590.
61. Clegg MT, Kidwell JF, Kidwell MG, Daniel NJ (1976) Dynamics of correlated genetic systems. I. Selection in the region of the Glued locus of *Drosophila melanogaster*. *Genetics* 83(4):793–810.
62. Yoo BH (1980) Long-term selection for a quantitative character in large replicate populations of *Drosophila melanogaster*: Part 1. Response to selection. *Genet Res* 35(1):1–17.
63. Yoo BH (1980) Long-term selection for a quantitative character in large replicate populations of *Drosophila melanogaster*: Part 2. Lethals and visible mutants with large effects. *Genet Res* 35(1):19–31.
64. Robertson FW, Reeve E (1952) Studies in quantitative inheritance I. The effects of selection of wing and thorax length in *Drosophila melanogaster*. *J Genet* 50(3):414–448.
65. Frankham R, Jones LP, Barker JSF (1968) The effects of population size and selection intensity in selection for a quantitative character in *Drosophila*. 3. Analyses of the lines. *Genet Res* 12(3):267–283.
66. Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180(3):1645–1660.
67. Smith JM, Haigh J (2007) The hitch-hiking effect of a favourable gene. *Genet Res* 89(5-6):391–403.
68. Messer PW (2013) SLiM: Simulating evolution with selection and linkage. *Genetics* 194(4):1037–1039.
69. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci USA* 103(1):135–140.
70. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3):e72.
71. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* 31(5):1275–1291.

Supporting Information

Assaf et al. 10.1073/pnas.1424949112

SI Text

1. Balancing Selection for Arbitrary Dominance and Selection Coefficients

For arbitrary dominance the fitnesses of the diploids in the population are

$$w_{OO/OO} = 1 \quad \text{[S1]}$$

$$w_{BD/OO} = (1+h_b s_b)(1-h_d s_d) = 1 + h_b s_b - h_d s_d - h_b s_b h_d s_d \quad \text{[S2]}$$

$$w_{BD/BD} = (1+s_b)(1-s_d) = 1 + s_b - s_d - s_b s_d. \quad \text{[S3]}$$

The change in frequency of a BD haplotype at frequency p_t at time t due to selection can be derived,

$$p_{t+1} = p_t^2 \log\left(\frac{w_{BD/BD}}{\overline{w}_{pop}}\right) + p_t q_t \log\left(\frac{w_{BD/OO}}{\overline{w}_{pop}}\right). \quad \text{[S4]}$$

Using a series expansion in p and taking the continuous time limit, this becomes a differential equation describing the dynamics of the frequency of the *BD* haplotype, p, due to selection

$$S(p) = (h_b s_b - h_d s_d)p + (s_b - 3h_b s_b - s_d + 3h_d s_d)p^2 + (-s_b + 2h_b s_b + s_d - 2h_d s_d)p^3, \quad \text{[S5]}$$

which is valid for small p, although also qualitatively correct for $p \sim 1$.

We can approximate Eq. **S5** with the assumption $s_b \ll s_d \ll 1$, giving

$$S(p) \approx s_b p - s_d p^2 + s_d p^3. \quad \text{[S6]}$$

The first term reflects selection on heterozygotes with fitness $\sim s_b$ that occur with probability $\sim p$, the second term reflects selection on homozygotes with fitness $\sim -s_b$ that occur with probability $\sim p^2$, and the third term comes from a decrease in population mean fitness at higher frequencies of *BD*. This cubic equation is the origin of the shape of the selection curve in Fig. 2. Because $p^* \ll 1$ when $s_b \ll s_d$, the cubic term can be ignored as it is smaller than the other terms by a factor of $\sim p^*$. The cubic term, however, is important if one enters a regime in which $Ns_d \lesssim 1$; however, this regime is not considered in the paper as it is biologically implausible for most populations.

If we wish to use Eq. **S5** to derive an equilibrium frequency for arbitrary dominance and selection coefficients, we find

$$p^* = \frac{h_d s_d - h_b s_b + h_b s_b h_d s_d}{s_b - 2h_b s_b - s_d + 2h_d s_d - s_b s_d + 2h_b s_b h_d s_d}. \quad \text{[S7]}$$

2. Drift to Extinction from Equilibrium No Recombination

2.1. Strong Selection, Weak Drift Regime ($\alpha > 1$). In the absence of recombination, the rate of extinction depends on two things: the strength of selection [$S(p)$] pushing the *BD* haplotype toward equilibrium p^*, and the variance in frequency due to drift [$D(p)$] that enables the *BD* haplotype to fluctuate to extinction against selection. Under these two effects, the frequency of the *BD* haplotype, p, is governed by the stochastic equation

$$\delta p \approx S(p)\delta t + \eta\sqrt{D\delta t} \quad \text{[S8]}$$

with $S(p) \approx s_b p - s_d p^2 + s_d p^3$ and where $D(p) = p(1-p)/N$ is the variance from drift and η is Gaussian distributed noise with zero mean and unit variance. In the long time limit, the probability density $\rho(p|t, p_0)$ generated by Eq. **S8**, which describes the likelihood of observing a *BD* haplotype at frequency p after time t, given that it started at p_0, has weights at $p = 0$ and $p = 1$ only (which must sum to unity). What this means for the *BD* haplotype is that after a long enough time, it must eventually fluctuate to extinction or fixation. In our particular case, because $s_b \ll s_d$, the probability of fluctuating to extinction is far larger than that of fluctuating to fixation.

Before this time, however, the probability density must move from being concentrated at $p = p_0$ (where it all started) to being concentrated at $p = 0$ (extinction). We want to estimate how quickly this process happens. Consider starting all of the probability density at $p = p^*$. Initially, this probability density spreads out a small amount around p^*, due to drift. However, because of strong selection, it remains sharply peaked around $p = p^*$ and reaches a selection–drift steady state: Its shape does not change, but it begins to decay (i.e., its amplitude decays) at some characteristic rate λ such that $\rho(p|t, p_0) \approx \exp(-\lambda t)$. The rate, λ, is small because we are in the limit of strong selection, and fluctuating to extinction is improbable. Our goal here is to estimate λ, which determines the rate at which *BD* haplotypes go extinct.

To do this, we consider the stochastic equation for δp from Eq. **S8**. For changes in frequency δp such that $\delta p/p \ll 1$, both $S(p)$ (selection) and $D(p)$ (drift) can be assumed to be constant, and Eq. **S8** reduces exactly to standard diffusion. The solution for probability density in this case is therefore a Gaussian over δp,

$$\rho(\delta p|t)d(\delta p) = \frac{d(\delta p)}{\sqrt{2\pi D t}} \exp\left[-\frac{(\delta p - St)^2}{2Dt}\right]. \quad \text{[S9]}$$

This makes sense: The mean change in frequency is St, which is the change that would have been expected without drift, whereas the variance around this grows linearly in time Dt, with a coefficient that is exactly the variance introduced by the drift term. Multiplying out the quadratic and exponentiating the constant out front, this can be written as

$$\rho(\delta p|t) = \frac{1}{\sqrt{2\pi}} \exp\left[\beta - \frac{\beta}{2}\left(\frac{t}{\tau} + \frac{\tau}{t}\right) - \frac{\ln(t/\beta\tau)}{2}\right] \quad \text{[S10]}$$

where

$$\tau = \frac{\delta p}{S} \quad \text{and} \quad \beta = \frac{S\delta p}{D}. \quad \text{[S11]}$$

Writing the density in this way is particularly useful because one can straightforwardly see when selection dominates or when drift dominates. The parameter β determines which of the two is most important: when $|\beta| \gg 1$, selection dominates, whereas for $|\beta| \ll 1$, drift dominates. One way of interpreting β is as a ratio of two timescales,

$$\beta = \frac{\text{time for drift to change frequency by } \delta p}{\text{time or selection to change frequency by } \delta p} = \frac{(\delta p^2/D)}{(\delta p/S)} = \frac{S\delta p}{D}. \quad \text{[S12]}$$

If the timescale for selection to change the frequency by δp is much faster than that for drift (large β), then selection drives the dynamics. The reverse is true for small β in which drift drives the dynamics. Focusing on the case of strong selection and considering Eq. **S10**, one can see that when $\beta \gg 1$, the $\ln(t/\beta\tau)$ term is small compared with the other two; thus it can be ignored. The log density then becomes

$$\ln \rho(\delta p|t) \approx \beta - \frac{\beta}{2}\left(\frac{t}{\tau}+\frac{\tau}{t}\right). \quad \textbf{[S13]}$$

In our specific case, we are interested in the *BD* haplotype fluctuating from $p=p^*$ to $p=0$ (i.e., negative δp) in the presence of a positive selection (S) pushing the *BD* haplotype upward toward $p=p^*$. This means that we are interested in the case of $\delta p<0$ and $S>0$. Considering the expressions for τ and β, one can see that such a case corresponds to both $\beta<0$ and $\tau<0$. One can see from Eq. **S13** that the probability density is sharply peaked at $t=|\tau|$, regardless of the sign of β. This somewhat counterintuitive result means the following: When selection is strong ($|\beta| \gg 1$), if the *BD* haplotype moves against the flow of selection, driven purely by fluctuations, then the path it takes will be similar to the path it would have taken if it had moved with the flow of selection [a result similar to that found by Maruyama (1)]. The key thing to note, however, is that when selection is strong, then the *BD* haplotype is very unlikely to fluctuate against the flow of selection.

To quantify just how unlikely this is, we remember that in the rare cases where movement occurs against the flow, then the most probable paths are close to a time $t=|\tau|$. Therefore, the probability of changing in frequency δp in the direction opposite of v (i.e., β negative) is obtained by substituting this time into Eq. **S13**, giving

$$Pr(\text{moving by } \delta p \text{ against } v) \approx \exp(-2|\beta|). \quad \textbf{[S14]}$$

The above probability result assumed that S and D were constant over a small interval δp. To calculate the probability of fluctuating over the entire selective barrier $S(p)$ to extinction, one must therefore take the product of the probabilities of many such jumps, each with its own $\beta(p)$. This product gives an integral in the exponent

$$P(p^* \to 0) \approx \exp\left[2\int_{p^*}^{0} dp \frac{S(p)}{D(p)}\right]. \quad \textbf{[S15]}$$

In our specific case,

$$S(p) \approx s_b p - s_d p^2 \quad \textbf{[S16]}$$

$$D(p) \approx p/N \quad \textbf{[S17]}$$

$$p^* \approx s_b/s_d. \quad \textbf{[S18]}$$

The integral can be solved exactly giving

$$P(p^* \to 0) \approx \exp\left(-\frac{Ns_b^2}{s_d}\right). \quad \textbf{[S19]}$$

This result has the rather simple interpretation that what primarily determines the probability is the barrier height. Evaluating the strength of selection and noise at this highest point and multiplying by the distance it has to move, p^*, also gives exactly the above result. The parameter

$$\alpha \equiv Ns_b^2/s_d \quad \textbf{[S20]}$$

is therefore particularly important in the dynamics because it alone determines the probability of fluctuating to extinction

$$P(p^* \to 0) \approx A\exp[-B\alpha] \quad \textbf{[S21]}$$

with A and B constants $O(1)$.

At present, we have calculated the probability of fluctuating to extinction. To translate this into an estimate for the rate of decay, λ, we must remember that in deriving the above result for the probability, we used the fact that each of the small jumps in frequency δp that the *BD* haplotype takes to reach $p=0$ lasts a time $\tau=\delta p/S$. We can therefore calculate how long it takes for that probability to decay by adding up all of the τ on its journey from $p=p^*$ to $p=0$. However, our calculation must only include the region where $\beta>1$ (i.e., where selection dominates). Therefore, the frequency limits in the integral are not $p=p^*$ and $p=0$ but rather $p=p^*-1/Ns_b$ to $p=1/Ns_b$, because these define the region of the barrier where $\beta>1$ (see Fig. 2*A*). Therefore

$$\text{time to extinction} = \int_{p^*-1/Ns_b}^{1/Ns_b} \frac{dp}{S(p)} = \int_{p^*-1/Ns_b}^{1/Ns_b} \frac{dp}{s_b p - s_d p^2} \quad \textbf{[S22]}$$

This integral will be dominated by where the flow of selection is slowest, which is the start and the finish. Because of the symmetry of the quadratic around the point $p=p^*/2$, we can integrate from $1/Ns_b$ to $p^*/2$ and then double the result. We can also realize that because $p \ll 1$ the quadratic term can be ignored, giving

$$\text{time to extinction} \approx 2\int_{1/Ns_b}^{p^*/2} \frac{dp}{s_b p} = \frac{2}{s_b}\ln\left(\frac{Ns_b p^*}{2}\right) \approx \frac{2}{s_b}\ln(\alpha). \quad \textbf{[S23]}$$

This result again makes sense: This time is exactly the characteristic time it would have taken the *BD* haplotype to rise from $1/Ns_b$ to p^*, were it to have done so deterministically under the influence of selection alone. Approximating the rate is now straightforward: Given that an amount of probability $\exp(-Ns_b^2/s_d)$ is lost in an amount of time $\sim 1/s_b \ln(\alpha)$, it means the rate of decay, λ, must be approximately given by

$$\lambda_l \approx \frac{e^{-\alpha}}{(1/s_b)\ln(\alpha)}, \quad \textbf{[S24]}$$

which is the result quoted Eq. **10** of the main text, such that mean time to extinction $\tau_l \approx 1/\lambda_l$, and which agrees very well with the simulations described in *SI Text*, section 4. We verify the slightly counterintuitive dependence of λ on s_b by performing 10^5 simulations at various s_b but at fixed α in Fig. S3*J*, as well as for various N, s_b, and s_d without fixing α (Fig. S3 *A*–*I*).

2.2. A Note on α for Arbitrary Dominance Coefficients. For the case of arbitrary dominance coefficients ($h_b \neq 1$, $h_d \neq 0$), a relatively straightforward extension can be made to our model by changing α to reflect the fitness of heterozygotes (i.e., replace s_b with $h_b s_b - h_d s_d$). In this case, our approximations will hold as long as there is still heterozygote advantage, or, in other words, $h_b s_b > h_d s_d$ and $h_b s_b - h_d s_d < s_d$. Thus, α becomes

$$\alpha \approx N\frac{(h_b s_b - h_d s_d)^2}{s_d}, \quad \textbf{[S25]}$$

which will still predict the rate of loss. However, note that in the rate of escape quoted in the main paper,

$$\lambda_e \approx rlNs_b \frac{s_b}{s_d} \approx rl\alpha, \quad \text{[S26]}$$

the first s_b value comes from the probability of establishment for a new *BO* recombinant haplotype, and the second s_b value comes from the equilibrium frequency p^*. Thus, for arbitrary dominance, the probability of escape becomes

$$\lambda_e \approx rlNs_b \frac{h_b s_b - h_d s_d}{s_d}. \quad \text{[S27]}$$

2.3. Strong Drift, Weak Selection Regime ($\alpha < 1$). When $\alpha < 1$, there is no region below p^* in which selection dominates over drift. However, there is a region above p^* where selection will dominate over drift. This will occur when the strength of selection $[S(p) = s_b p - s_d p^2]$ approximately matches the strength of drift ($\sim \pm 1/N$). Because this occurs above p^*, the dominant term in $S(p)$ is the quadratic term, so the balance between selection and drift is a balance involving selection against the homozygotes,

$$-s_d p^2 \approx \pm 1/N \quad p \approx \sqrt{N/s_d}\,. \quad \text{[S28]}$$

This means that when the *BD* haplotype is below the frequency of $\sqrt{N/s_d}$, its dynamics are governed by drift, i.e., are largely neutral. Under neutrality, the cumulative probability that a lineage which started at one copy ($n_o = 1$) has gone extinct ($n = 0$) by time t is

$$\Pr[n = 0|t] \approx e^{-n_o/t} \approx e^{-1/t} \approx 1 - 1/t. \quad \text{[S29]}$$

Thus, the probability that a neutral lineage goes extinct in an interval dt near t is $\sim dt/t^2$, giving a distribution of extinction times that are power law distributed as $\sim 1/t^2$.

The distribution of extinction times of the *BD* haplotype in the $\alpha < 1$ regime is expected to follow this neutral power-law distribution, but with a cutoff at $t \approx \sqrt{N/s_d}$ (due to the *BD* haplotype being unable to drift to high frequencies where the recessive deleterious mutation is exposed). The fact that the maximum time a *BD* haplotype can remain in the population in this regime is $t \approx \sqrt{N/s_d}$, which is the same as the maximum frequency it can reach before selection against the homozygotes pushes it down in frequency, is no accident. They are the same because, under neutrality, the number of generations it takes to change by order n is $t \approx n$. We verify this in the simulations section (section 4.3).

Recalling that the total probability of an escape event by time t is

$$\text{probability of escape} \approx rls_b N \int_0^t p(t')dt' \quad \text{[S30]}$$

(which is valid for a probability of escape $\ll 1$). Also recall that, as we just showed and as quoted in *Predictions for the Regime of Weak Selection and Strong Drift ($\alpha \ll 1$)*, in an interval dt, there is a fraction $\sim dt/t^2$ of *BD* haplotypes that go extinct and a fraction $\sim rls_b t^2$ of *BO* haplotypes that escape via recombination, giving a probability of escape $\sim rls_b\, dt$ that is constant in time as long as the BD haplotypes persist. Thus, the probability of fixation in this $\alpha \ll 1$ regime is

$$\text{probability of fixation} \approx rls_b \int_0^{\sqrt{N/s_d}} dt' = rls_b \sqrt{N/s_d} \quad \text{[S31]}$$

where the upper limit in the integral comes from the fact that it is unlikely for the *BD* haplotype to drift for longer than $\sim \sqrt{N/s_d}$ generations.

3. Zones of Altered Adaptation

3.1. Strong Selection, Weak Drift ($\alpha > 1$).

Zone of suppressed probability of fixation. In this regime, the probability that a *BD* haplotype fixes can be significantly reduced relative to the equivalent beneficial mutation with no deleterious hitchhiker, but only if it is likely to fluctuate to extinction before escape can occur. This means that $\tau_l < \tau_e$. Using the expressions for these times from Eqs. **10** and **11** in the main text, the condition for the fixation probability to be significantly reduces becomes

$$\frac{\ln[\alpha]}{s_b} e^{\alpha} \ll 1/rl\alpha. \quad \text{[S32]}$$

Rearranging this equation defines a base pair distance, l_l, around a recessive deleterious mutation within which the probability of fixation of beneficial mutations is reduced relative to the case of no hitchhiker,

$$l_l = \frac{s_b}{r\alpha \ln[\alpha]} e^{-\alpha}. \quad \text{[S33]}$$

Zone of increased sweep time. Even if a beneficial mutation is not driven to extinction, the duration of a beneficial mutation's sweep to fixation can be substantially extended if it's genetically linked to a recessive deleterious hitchhiker. To understand when this occurs, consider that the total time of a successful sweep will be prolonged by the time it takes for an escape event to occur,

$$\text{average sweep time} \approx \frac{\ln[Ns_b]}{s_b} + \left(\frac{1}{\tau_e} + \frac{1}{\tau_l}\right)^{-1}. \quad \text{[S34]}$$

The first term corresponds to the sweep time for a single adaptive mutation with no hitchhiker, and the second term comes from conditioning on fixation such that the extension in sweep time is determined by the minimum of τ_e or τ_l. Typically, a sweep will be significantly extended provided the additional time $\min(\tau_e, \tau_l) \gg \ln[Ns_b]/s_b$. Using the expressions for τ_e and τ_l from Eqs. **10** and **11** in the main text and rearranging, we can again cast this in terms of a base pair distance,

$$l_e = \frac{s_b}{r\alpha \ln[Ns_b]}, \quad \text{[S35]}$$

which is the distance to a recessive deleterious mutation within which a new beneficial mutation must land to have its sweep time significantly extended.

3.2. Strong Drift, Weak Selection ($\alpha < 1$).

Zone of suppressed probability of fixation. The probability of fixation in this regime, from Eq. **6** and *Predictions for the Regime of Weak Selection and Strong Drift ($\alpha \ll 1$)*, is

$$\Pr(\text{fixation}) \approx rls_b \int_0^{\sqrt{N/s_d}} dt = rls_b \sqrt{N/s_d}, \quad \text{[S36]}$$

where the upper limit in the integral comes from the fact that it is highly unlikely for a *BD* haplotype to drift for longer than $\sim \sqrt{N/s_d}$ generations. Compared with a beneficial mutation with no hitchhikers (where probability of fixation is $\sim s_b$), a recessive deleterious hitchhiker will significantly suppress the probability of fixation whenever $rl\sqrt{N/s_d} < 1$. This can be used to define a base pair distance, l_l, around any recessive deleterious mutation within which a new beneficial mutation of effect size $\sim s_b$ (or smaller) has a reduced chance of fixation,

$$l_l = \frac{1}{r}\sqrt{\frac{s_d}{N}}. \quad \text{[S37]}$$

Zone of increased sweep time. In this regime, there can be no significant increase in sweep time, because the loss time is always small relative to the sweep time.

3.3. Quantifying Coding Gene Density for *Drosophila melanogaster* and Humans. Given that deleterious regions are likely to occur in functional regions, and that functional regions are likely to be clustered in the genome (particularly for humans), we framed our estimates of functional density in terms of the density of coding genes around every coding gene. For this analysis in humans, we downloaded the University of California, Santa Cruz (UCSC) knownCanonical table of genes and their corresponding positions, where each gene is only represented by one isoform, and the set of gene predictions are "based on data from RefSeq, GenBank, Rfam, and the tRNA Genes track. . .This is a moderately conservative set of predictions" (UCSC website, genome.ucsc.edu/cgi-bin/hgTables). We then excluded all genes that do not appear in the knownGenePep list (i.e., excluded noncoding) and excluded all genes that are not located on autosomes. Then, using the annotated locations of all genes in this set, we quantified the number of coding genes that fall within a window centered around the midpoint of every coding gene, where the window sizes were defined by our zone sizes. Note that this means, for example, we had 19,353 data points for coding gene densities because there are currently 19,353 coding genes annotated on the human autosomes. Results can be seen in Fig. S1 *A* and *B*.

We did a similar analysis for *Drosophila*, where we downloaded from flybase.org the dmel-all-gene- list, excluded all genes that do not appear on chromosomes 2 or 3, excluded all genes that do not appear in the dmel-all-translation- list (i.e., excluded noncoding), and only used the first isoform listed for every gene (i.e., if multiple isoforms were present, we used the -PA isoform). Then, using the annotated locations of all genes in this set, we quantified the number of coding genes that fall within a window centered around the midpoint of every coding gene. In this case, for *Drosophila*, there are currently 11,631 annotated coding genes on chromosomes 2 and 3. Results can be seen in Fig. S1 *C–F*.

3.4. Estimates for *Drosophila melanogaster* and Humans for the Proportion of the Genome in Which the Fixation Probability of Beneficial Mutations Is Reduced. Table 1 gives predictions for the proportion of coding genes in the *Drosophila* and human genomes that are subject to reduced fixation probabilities for new beneficial mutations. These predictions are based on coding gene densities (described in section 3.3) and the predicted zone size around the deleterious mutations (described in sections 3.1 and 3.2). Because this zone is maximized in the $\alpha = Ns_b^2/s_d \leq 1$ regime, we used a zone size of $l_l = \sqrt{s_d/N}/r$. Note that this is independent of the beneficial mutation effect size, and thus all beneficial mutations that satisfy $s_b \leq \sqrt{s_d/N}$ behave similarly, emphasizing that weakly adaptive mutations may be particularly susceptible to the effects of hidden recessive deleterious variation in the genome. Details of how we constructed Table 1 are found below. Method for Table 1:

Column 1 = Organism and assumed population size

Column 2 = Number of coding genes on autosomes, see section 3.3

Column 3 = Recessive deleterious effect size of interest

Column 4 = l_l for $\alpha < 1 = \sqrt{s_d/N}/r$

where brackets indicate number of genes which appear in this zone, given by section 3.3

Column 5 = the zone size is the same for all $s_b \leq \sqrt{s_d/N}$

Column 6 = n_d/g_c where

n_d = number of recessive deleterious mutations in a haploid set of autosomes (2–4)
g_c = number of coding genes in reference genome

Column 7 = column 4 brackets × column 6.

3.5. Estimates for *Drosophila melanogaster* and Humans for the Proportion of the Genome in Which the Sweep Time of Beneficial Mutations Is Extended. If we consider the beneficial mutations that do reach fixation, we find that a substantial proportion of the human and *Drosophila* genomes will cause there to be a staggered phase during the beneficial mutation's selective sweep, particularly due to more mildly deleterious recessive mutations (Table S1). Although the impact of recessive deleterious variation on the probability of fixation of beneficial mutations can be simplified in terms of which mutations fall within the $\alpha = Ns_b^2/s_d < 1$ regime, the potential for extended sweep times is not as easily simplified. Extensions in sweep times do not occur in the $\alpha < 1$ regime; instead, genomes will be most affected by staggered sweeps for intermediate values of α. In this regime, the adaptive mutation can reach a stable equilibrium frequency ($\alpha \not\ll 1$) where it has a slow rate of loss ($\tau_l > \ln[Ns_b]/s_b$), but the rate of new recombinants being generated in the population is not so large that the staggered sweep is resolved quickly ($\tau_e > \ln[Ns_b]/s_b$). One example where this regime likely applies is experimental evolutions (in obligately sexual diploids, like fruit flies), where the beneficial effect sizes may be strong but the population size is small. For example, in a population of $N = 1{,}000$ flies, a beneficial mutation with effect size $s_b = 1\%$ will be subject to staggered sweep phases within ~65% of coding genes.

To give a better sense of the parameter range affected, see Fig. S1 *G–I*, where *G* and *H* are for more mildly deleterious mutations (which can affect substantial portions of the genome), and *I* is for recessive lethals (which do not have a substantial effect). Experimental evolutions in *Drosophila* correspond to blue and purple lines in Fig. S1 *G–I*. Details of how these plots were constructed can be found below. Note that the plots are stepwise due to the measurements of the clustering of genes (zone size is rounded to the nearest order of magnitude and translated to the mean number of genes within that distance; see section 3.3). Method for Table S1:

Column 1 = Organism and assumed population size

Column 2 = Number of coding genes on autosomes, see section 3.3

Column 3 = Recessive deleterious effect size of interest

Column 4 = $\dfrac{s_b}{r\alpha \ln[Ns_b]}$

where brackets indicate number of genes which appear in this zone, given by section 3.3

Column 5 = Beneficial effect size of interest

Column 6 = n_d/g_c where

n_d = number of recessive deleterious mutations in a haploid set of autosomes (2–4)
g_c = number of coding genes in reference genome

Column 7 = column 4 brackets × column 6.

Our method for plots in Fig. S1 *G–I* was as follows: For Fig. S1 *G–I*, the *y* axis is count × density. Count denotes the number of genes that appear in region l_e (see section 3.3); $l_e = s_b/r\alpha\ln[Ns_b]$, where r = 10–8 and $\alpha = Ns_b^2/s_d$. Density is 1/30 for $s_d = 0.01$, 1/12,000 for dmel $s_d = 1$, and 1/20,000 for human $s_d = 1$. The *x* axis is s_b, where minimum s_b is set by $\alpha = Ns_b^2/s_d$, due to moving into the $\alpha < 1$ regime where no substantial extension in sweep time occurs, and maximum s_b is set by s_d, due to the beneficial mutation being stronger than the deleterious mutation.

4. Simulations

Two-locus Wright–Fisher forward simulations of an adaptive mutation genetically linked to a recessive deleterious mutation were performed to test our predictions. A wide range of parameters were varied, including selection coefficients ($s_b = 0.001 - 0.1$, $s_d = 0.01 - 1$), recombination rates ($c = r/l = 10^{-8} - 10^{-1}$), and diploid population sizes ($N = 10^2 - 10^6$), corresponding to ranges of $\alpha = Ns_b^2/s_d = 10^{-3} - 10^5$. All simulations used $s_b \leq s_d$ and heterozygous effects of $h_d = 0$ and $h_b = 0.5$ such that the equilibrium frequency $p^* \approx s_b/2s_d \leq 1/2$. We also included a one-locus control beneficial mutation (with no hitchhiker) for comparison. Diploid fitness was calculated as multiplicative across loci (as in Eqs. **S1**–**S3**). Simulations tracked frequencies of haplotypes (i.e., *BD*, *OO*, *BO*, *DO*), where each generation consisted of recombination and selection in diploids, and then drift of haplotypes (i.e., multinomial sampling). Simulations concluded when the beneficial mutation approached fixation or extinction (i.e., $p \gg 1 - 1/2N$ or $p \ll 1/2N$).

4.1. Extinction Times in the $\alpha > 1$ Regime. These simulations seeded the *BD* haplotype at frequency p^* and the *OO* haplotype at frequency $1 - p^*$, used zero recombination rate, and then recorded the generation of extinction of the beneficial mutation (i.e., the *BD* haplotype). A range of parameters were used, and each parameter set was performed for 1,000 simulations. Parameters were chosen such that $\alpha = Ns_b^2/s_d > 1$, $Ns_d > 1$ and mean time to extinction $\tau_l = 1/\lambda_l \gg \ln[Ns_b]/s_b$ (to ensure extinction times are not confounded with the time it typically takes to traverse from p^* to 0).

The extinction times are predicted to be exponentially distributed with rate

$$\lambda_l \approx \frac{e^{-\alpha}}{(1/s_b)\ln(\alpha)} \quad \textbf{[S38]}$$

called the "rate of loss," where

$$\alpha = Ns_b^2/s_d. \quad \textbf{[S39]}$$

Results of a few sample parameter sets can be seen in Fig. S2, in which both histograms of extinction times (Fig. S2 *A*–*C*) and QQ plots of exponential quantiles (Fig. S2 *D*–*F*) highlight that the extinction times indeed look to be exponentially distributed (with our predicted rate parameter indicated in red).

To test whether the rate of loss $\lambda_l \approx \exp(-\alpha) * (s_b/\ln(\alpha))$ from Eq. **S24** accurately captures the correct scaling, or more specifically is truly exponential in the parameter $\alpha = Ns_b^2/s_d$, we performed a series of simulations in which a rate of loss was inferred from the distribution of extinction times using the R function fitdistr (from the downloadable R statistics package "MASS"). If λ_l is indeed exponential in α, then we would expect that if we hold all parameters constant except one (thus changing α via a single parameter), this observed rate of loss would behave such that

$$\text{rate} = \exp[-\alpha] * (s_b/\ln[\alpha]) \quad \textbf{[S40]}$$

$$\text{rate} * (\ln[\alpha]/s_b) = \exp[-\alpha] \quad \textbf{[S41]}$$

$$\ln[\text{rate} * (\ln[\alpha]/s_b)] = -\alpha. \quad \textbf{[S42]}$$

This is shown in Fig. S3, where changing a single variable (either N, s_b, or s_d) affects the observed rate of loss only via the parameter α, indicating that λ_l accurately captures the scaling in the exponent. The term λ_l will predict the scaling for the rate of loss but not its exact form, so for comparison with simulations, we fit constants to a subset of simulations [finding $\lambda_l = 0.1\exp(-0.3\alpha) * (s_b/\ln(\alpha))$] and used these coefficients throughout the rest of our predictions for the rate of loss, the probability of fixation, and the sweep time in simulations.

4.2. Escape Times in the $\alpha > 1$ Regime. These simulations seeded the *BD* haplotype at frequency p^* and the *OO* haplotype at frequency $1 - p^*$, used nonzero recombination rates, and then recorded the generation of fixation or extinction of the beneficial mutation. A range of parameters were used, and each parameter set was performed for 1,000 simulations. We tried to choose parameters that satisfied $\alpha = Ns_b^2/s_d \gg 1$ such that drift to extinction should not confound measurements of escape time. Furthermore, we chose parameters to satisfy $Ns_d > 1$ (such that *BD* does not drift to fixation) and to satisfy $\tau_e = 1/\lambda_e \gg \ln[Ns_b]/s_b$ (such that escape times are not confounded with the time it typically takes to traverse from p^* to 1).

The escape times are predicted to be exponentially distributed with rate

$$\lambda_e \approx c\alpha \quad \textbf{[S43]}$$

called the "rate of escape," where

$$\alpha = Ns_b^2/s_d. \quad \textbf{[S44]}$$

Results of a few sample parameter sets can be seen in Fig. S4, in which both histograms of escape times (Fig. S4 *A*–*C*) and QQ plots of exponential quantiles (Fig. S4 *D*–*F*) highlight that the escape times indeed look to be exponentially distributed (with our predicted rate parameter indicated in red).

To test whether $c\alpha = cNs_b^2/s_d$ captures the correct scaling for the rate of escape, we performed a series of simulations in which a rate of escape was inferred from the distribution of fixation times using the R function fitdistr (from the library MASS). This observed rate is expected to scale linearly in response to c, N, s_b^2 and $1/s_d$. This is shown in Fig. S5, where each plot has the results from a series of simulations in which three parameters have fixed values and one parameter varies (either N, s_b, s_d, or c). The term λ_e will predict the scaling for the rate of escape but not its exact form, so, for comparison with simulations, we fit a constant using a subset of simulations (finding $\lambda_e = 0.8c\alpha$), and used this coefficient throughout the rest of our predictions for the rate of escape, the probability of fixation, and the sweep time in simulations.

4.3. Extinction Times in the $\alpha < 1$ Regime. To observe the extinction times in the $\alpha < 1$ regime, we performed 100,000 simulations in the $\alpha < 1$ regime (also requiring $Ns_d > 1$ so there is selection against the recessive deleterious mutation), seeding the *BD* haplotype at frequency $1/2N$ and the *OO* haplotype at frequency $1 - 1/2N$, and then recording the generation of extinction. Results are shown in Fig. S6. One way to view a power law is to plot the complementary cumulative distribution on a log–log plot, such that $\ln(\Pr[T > t]) = \ln(t)$. The neutral expectation is indicated by the black line, and the distribution of extinction times for neutral simulations are shown in the light blue histogram (note that it follows the neutral expectation, although finite sampling leads to poorer resolution at the tip of the tail). The distribution of extinction times from simulations in the $\alpha < 1$ regime are shown in the dark blue, pink, and yellow histograms, where they follow the neutral expectation up until a cutoff time that scales with $\sim\sqrt{N/s_d}$. The term $\sqrt{N/s_d}$ will predict the scaling for the extinction time but not its exact form, so, for comparison with simulations, we fit a constant to a subset of simulations (finding $4\sqrt{N/s_d}$) and used this coefficient throughout the rest of our predictions for the loss time when $\alpha < 1$, the probability of fixation, and the sweep time in simulations.

4.4. Probability of Fixation of the Beneficial Mutation (Both α Regimes). These simulations seeded the haplotype at frequency $p=1/2N$ and the *OO* haplotype at frequency $1-1/2N$, used nonzero recombination rates, and then recorded the generation of fixation or extinction of the beneficial mutation. A range of parameters were used, and each parameter set was performed for $1{,}000/s_b$ simulations such that a one-locus control beneficial mutation (with no deleterious hitchhiker) is expected to fix in ∼1,000 simulations.

For comparison with simulation results, we used the following analytic predictions for the probability of fixation of a beneficial mutation that enters a population on a *BD* haplotype:

$$\Pr(\text{fixation}) \approx \begin{cases} \left(\frac{1-e^{-s_b}}{1-e^{-2Ns_b}}\right)\left(\frac{\lambda_e}{\lambda_e+\lambda_l}\right)\left(1-e^{-(\lambda_e+\lambda_l)t_{max}}\right) & (\text{for } \alpha>1) \\ \left(\frac{1-e^{-s_b}}{1-e^{-2Ns_b}}\right)\left(1-e^{-c4\sqrt{N/s_d}}\right) & (\text{for } \alpha<1) \end{cases} \quad \textbf{[S45]}$$

where

$$\lambda_l \approx \frac{s_b}{\ln[\alpha]} e^{-\alpha} \quad (\text{for } \alpha>1) \quad \textbf{[S46]}$$

$$\lambda_e \approx c\alpha \quad (\text{for } \alpha>1) \quad \textbf{[S47]}$$

$$\alpha = \frac{Ns_b^2}{s_d} \text{ and } c = rl. \quad \textbf{[S48]}$$

A full panel of simulation results with analytic predictions for the probability of fixation can be seen in Fig. S7.

The first term in the probability of fixation in both regimes is the probability of establishment, where we have replaced the approximation s_b with the more exact form $1-e^{-s_b}/1-e^{-2Ns_b}$ for the sake of comparison with simulations (particularly important for small population sizes). Note that in the $\alpha>1$ regime, the third term is the probability that escape occurs before the simulations end at time t_{max} (due to not being able to run simulations for infinite generations). Simulations ran for, at most, $t_{max}=10^9$ generations.

A beneficial mutation of effect s_b will have a decreased probability of fixation if it falls within a recombination distance c_l of the recessive deleterious mutation with effect s_d in a population of N. In the $\alpha>1$ regime, this is found by setting $\lambda_l=\lambda_e$, and in the $\alpha<1$ regime, this is found by setting $c=1/\sqrt{N/s_d}$. For comparison with simulations, we thus use the following analytic predictions:

$$c_l \approx \begin{cases} \max\left(\frac{s_b}{r\alpha\ln[\alpha]e^{\alpha}}, \frac{1}{\alpha\, t_{max}}\right) & (\text{for } \alpha>1) \\ \sqrt{s_d/N} & (\text{for } \alpha<1) \end{cases}. \quad \textbf{[S49]}$$

These predictions for c_l can be seen as the vertical dashed gray lines in Fig. S7.

4.5. Sweep Time of the Beneficial Mutation (Both α Regimes). These simulations seeded the *BD* haplotype at frequency $p=1/2N$ and the *OO* haplotype at frequency $1-1/2N$, used nonzero recombination rates, and then recorded the generation of fixation or extinction of the beneficial mutation. A range of parameters were used, and each parameter set was performed until 500 fixation events occurred.

For comparison with simulation results, we used the following analytic predictions for the sweep time of a beneficial mutation that enters a population on a *BD* haplotype:

$$\text{total sweep time} \approx \begin{cases} \frac{4\ln[2Ns_b]}{s_b} + \left(\frac{1}{\tau_e}+\frac{1}{\tau_l}\right)^{-1} & (\text{for } \alpha>1) \\ \frac{4\ln[2Ns_b]}{s_b} + \left(\frac{1}{\tau_e}+\frac{1}{4\sqrt{N/s_d}}\right)^{-1} & (\text{for } \alpha<1) \end{cases}. \quad \textbf{[S50]}$$

A full panel of simulation results can be seen in Fig. S8. Note that the first term (in both regimes) has added factors of 2 that come from the sweep time predictions for diploids. The second term (in both regimes) is derived from the faster of two processes—either the escape time (τ_e) or the loss time (τ_l for $\alpha>1$ and $\sqrt{N/s_d}$ for $\alpha<1$).

The sweep time will not be changed if $\ln[Ns_b]/s_b$ is larger than τ_e, τ_l, or $\sqrt{N/s_d}$. If $\ln[Ns_b]/s_b < \tau_e < (\tau_l \parallel \sqrt{N/s_d})$, then escape will generally occur before extinction of the *BD* haplotype, and the sweep will be extended by τ_e. If $\ln[Ns_b]/s_b < (\tau_l \parallel \sqrt{N/s_d}) < \tau_e$, then loss occurs before escape; however, we are conditioning on fixation, and thus, in the rare instances where the beneficial mutation does reach fixation, it must do so in a time τ_l or $\sqrt{N/s_d}$ (depending on the regime). For example, in Fig. 4 *C* and *D*, we have a case of $\ln[Ns_b]/s_b < \tau_l$, and thus there is still an extension in the sweep time (seen as the leveling off of sweep times at low recombination rates).

A beneficial mutation of effect s_b will have an increased sweep time if it falls within a recombination distance c_e of the recessive deleterious mutation with effect s_d in a population of N. This can be found by setting $1/\lambda_e = \ln[Ns_b]/s_b$; thus, for comparison with simulations, we use the following analytic predictions:

$$c_e = \frac{s_b}{\alpha\ln[Ns_b]} \quad (\text{for } \alpha>1 \text{ and } \alpha<1). \quad \textbf{[S51]}$$

These predictions for c_e can be seen as the vertical dashed red lines in Fig. S8. Note that in the $\alpha<1$ regime, the sweep time is not expected to be substantially altered, because, in this regime, where $\alpha=Ns_b^2/s_d<1$, the beneficial mutation must escape in a time $\sqrt{N/s_d}$, which is generally less than $\ln[Ns_b]/s_b$.

4.6. Testing Robustness of Model if Deleterious Allele Segregating at High Frequency. Our model predictions should be robust to cases where the recessive deleterious allele is segregating on an *OD* haplotype in many individuals in the population, because, in our case of $N\mu_d<1$, the deleterious mutation will typically reach, at most, $\sqrt{N/s_d}$ copies, or a frequency of $1/\sqrt{Ns_d}$. To understand this, consider that the deleterious variation will reduce the mean (log) fitness advantage of a new *BD* haplotype, due to causing the *BD* haplotype to now appear in *BD*/*OD* diploids in addition to appearing in *BD*/*OO* diploids,

$$\overline{w}_{BD} \approx s_b \times (\text{frequency of } OO) + (s_b - s_d) \times (\text{frequency of } OD) \quad \textbf{[S52]}$$

$$\overline{w}_{BD} \approx s_b\left(1-\frac{1}{\sqrt{Ns_d}}\right) + (s_b-s_d)\left(\frac{1}{\sqrt{Ns_d}}\right)$$

$$\overline{w}_{BD} \approx s_b - \frac{s_b}{\sqrt{Ns_d}} + \frac{s_b}{\sqrt{Ns_d}} - s_d\frac{1}{\sqrt{Ns_d}} \quad \textbf{[S53]}$$

$$\overline{w}_{BD} \approx s_b - \sqrt{\frac{s_d}{N}}. \quad \textbf{[S54]}$$

Thus, the mean selective advantage of the *BD* haplotype can be substantially reduced by segregating deleterious variation when $s_b \lesssim \sqrt{s_d/N}$, which, interestingly, is when $\alpha=Ns_b^2/s_d \lesssim 1$. However, as we have shown, whenever $\alpha<1$, the selective advantage

of the *BD* is irrelevant because the dynamics are determined by drift and selection against the recessive deleterious allele. Therefore, having multiple copies of an *OD* haplotype will not change our results. We confirmed that our analytics still hold in such a case, using simulations that seeded an *OD* haplotype at $\sqrt{N/s_d}$ copies when the *BD* haplotype appears.

5. Additional Signatures of Staggered Sweeps

Some additional statistics were calculated for the diversity around a beneficial mutation that had a recessive deleterious hitchhiker during its sweep to fixation (Fig. S9). Hard sweeps were simulated by seeding a beneficial mutation with effect s_b on a single haplotype at establishment frequency, staggered sweeps by doing the same but where the single haplotype also contained a recessive deleterious mutation with effect s_d a distance l away, and soft sweeps by seeding the beneficial mutation on a new haplotype at establishment frequency every t generations, where t was drawn from an exponential distribution with rate s_b (thus $\theta \approx NU_b = 1$ beneficial mutation entering the population every generation with an establishment probability of $\sim s_b$).

1. Maruyama T (1974) The age of an allele in finite population. *Gen Res* 23(2):137–143.
2. Kusakabe S, Yamaguchi Y, Baba H, Mukai T (2000) The genetic structure of the Raleigh natural population of *Drosophila melanogaster* revisited. *Genetics* 154(2): 679–685.
3. Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M (2015) An estimate of the average number of recessive lethal mutations carried by humans. *Genetics* 199(4):1243–1254.
4. Latter BD (1998) Mutant alleles of small effect are primarily responsible for the loss of fitness with slow inbreeding in *Drosophila melanogaster*. *Genetics* 148(3):1143–1158.

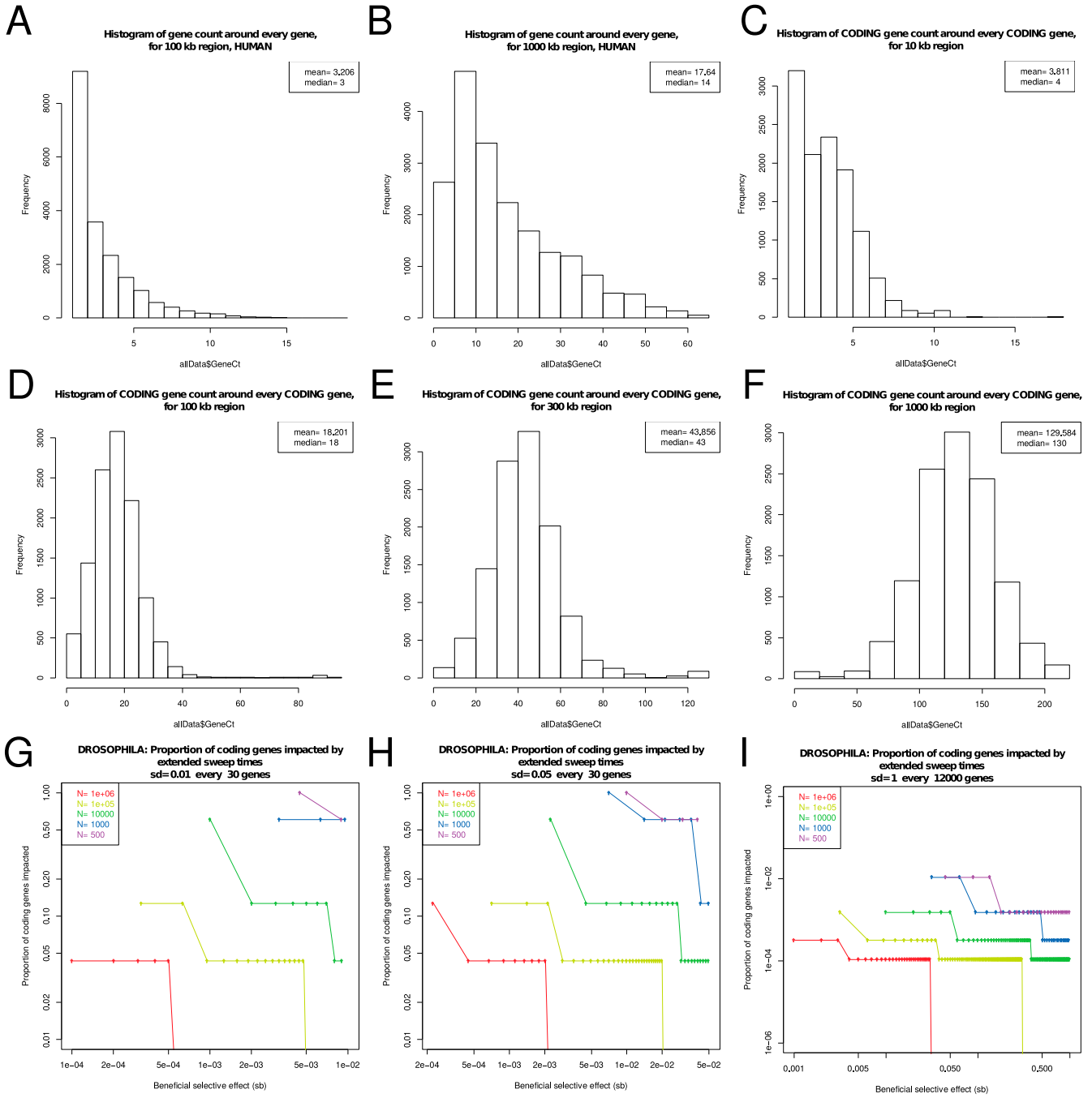

Fig. S1. Histograms of the density of coding genes around every coding gene in the reference genomes of human and *Drosophila* (*A*–*F*) and predictions for the proportion of coding genes affected in the *Drosophila* genome by staggered phases as a function of s_b (*G*–*I*). (*A* and *B*) For humans, the number of coding genes which appear in a (*A*) 100-kb and (*B*) 1-Mb window (respectively), where 19,353 windows, each centered on the midpoint of a coding gene, were used (per window size). (*C*–*F*) For *Drosophila*, the number of coding genes that appear in a (*C*) 10-kb, (*D*) 100-kb, (*E*) 300-kb, and (*F*) 1-Mb window (respectively), where 11,631 windows, each centered on the midpoint of a coding gene, were used (per window size). (*G*–*I*) Proportion of coding genes in the *Drosophila* genome impacted by extended staggered phases due to deleterious mutations with effect sizes (*G*) 1%, (*H*) 5%, and (*I*) 100%, as a function of s_b where the minimum is set by when $\alpha = 1$ and the maximum is set by $s_b = s_d$, and where each line indicates calculations for a different population size.

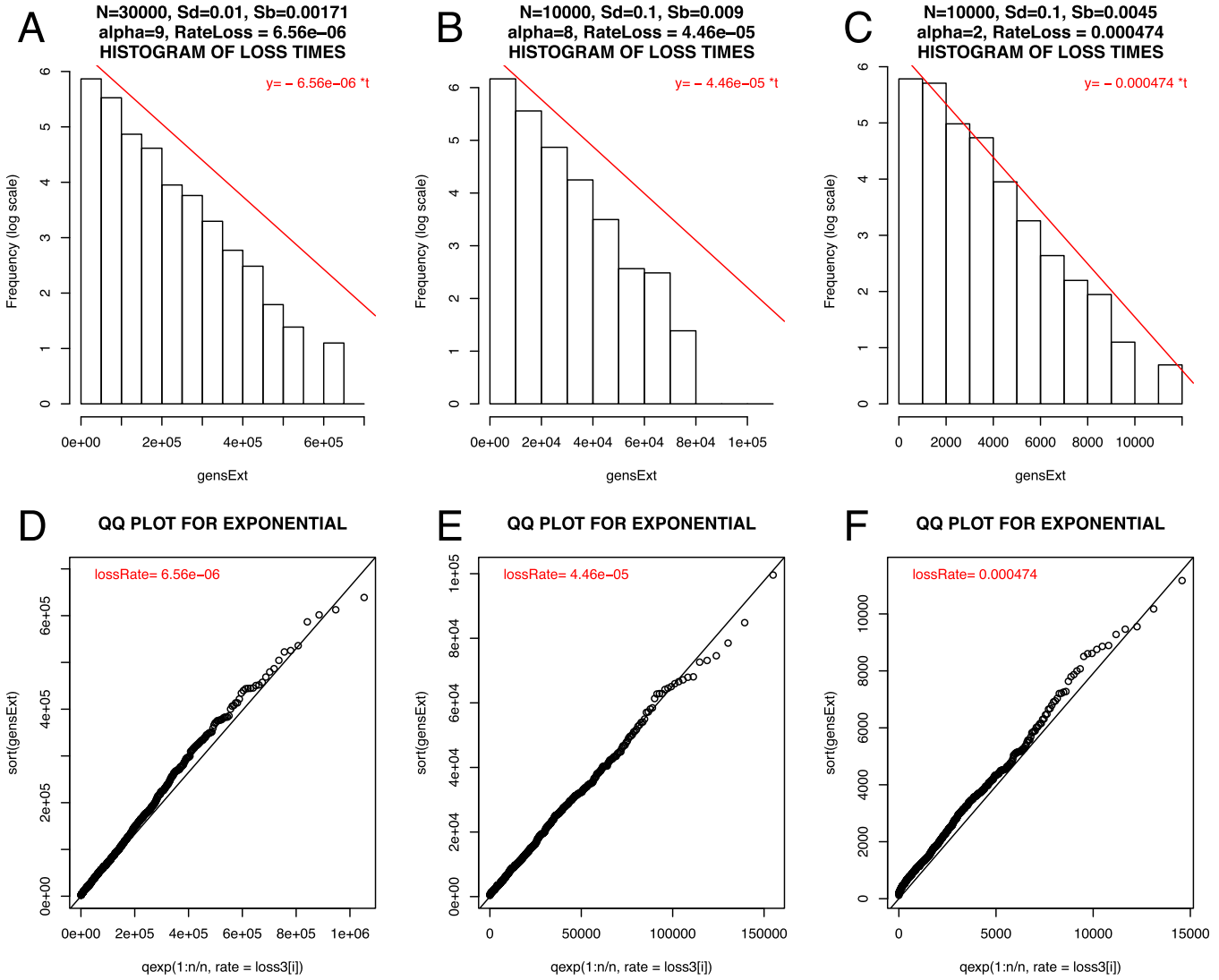

Fig. S2. Histograms and QQ plots demonstrating that the extinction times in the $\alpha > 1$ regime are exponentially distributed (a fuller set of simulation results across parameter regimes can be seen in Fig. S3). *A–C* consist of histograms of extinction times from 1,000 simulations (the parameter set used is indicated in the histogram title), such that the *x* axis is the extinction time *t* and the *y* axis is the frequency of events on a log scale. Note that if extinction times are indeed exponentially distributed with rate λ_l, then we expect these histograms to be described by the line $log(y) \approx -\lambda_l x$, which is indeed the case and can be seen by the red line, which is an exponential density curve with the analytically predicted rate parameter λ_l. Note that short extinction times ($1/\lambda_l = \tau_l \ll \ln[Ns_b]/s_b$) will not be exponentially distributed, sometimes causing an uptick in the distribution for extinction times near zero. *D–F* consist of the corresponding QQ plot for each histogram above it, where the *x* axis is the theoretical quantiles for an exponential distribution with rate λ_l and the *y* axis is the sample quantiles of extinction times obtained from simulations. The fact that the points lie along the straight line indicate that the extinction times are likely exponentially distributed.

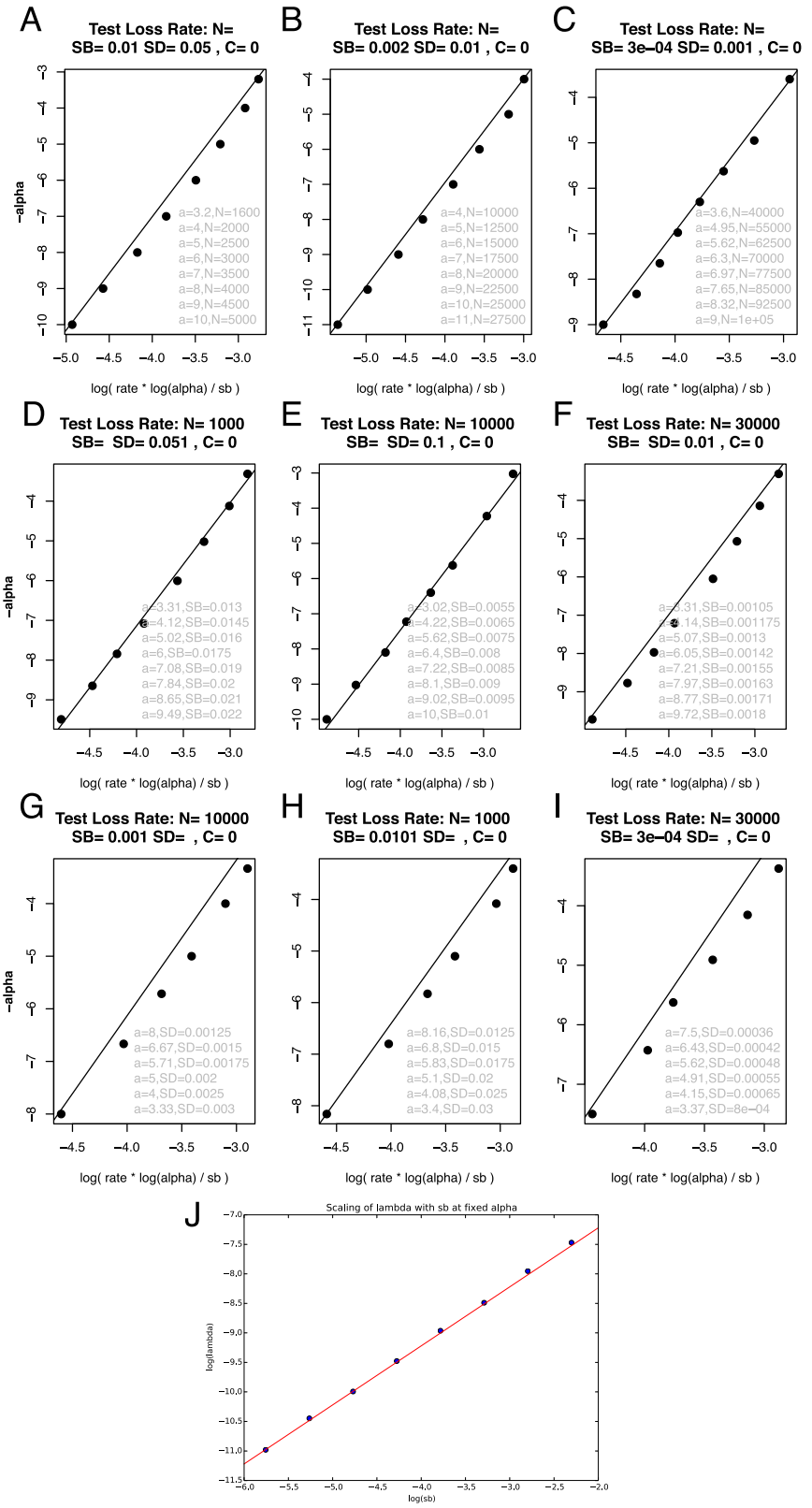

Fig. S3. Testing whether the rate of loss is indeed exponential in the parameter $\alpha = Ns_b^2/s_d$, as described by Eq. **S43**. Each subplot contains results from a set of simulations in which two parameters were held constant and one parameter varied; however, the observed rate of loss will be predicted through the parameter α (see Eq. **S43**). The rate of loss was inferred from 1,000 simulations, where an exponential distribution was fit to the extinction times using the R function fitdistr from the package MASS. Parameter values used (for N, s_b, s_d, and always $c = 0$) are indicated in the heading of each subplot (with blank values for the parameter varied, details for which are found in the gray text). For example, in *A*, each black data point indicates the results from 1,000 simulations in which $s_b = 0.01$ and $s_d = 0.05$ and N is varied, such that the value of N used is indicated by the gray text with the corresponding α (written "a") next to it. *A*–*C*

Legend continued on following page

use fixed values of s_b and s_d (values indicated in headings) and varies N (values indicated in gray text), *D–F* use fixed N and s_d and varied s_b, and *G–I* use fixed N and s_b and varied s_d. *J* verifies the s_b dependence of λ by plotting the rate of decay λ at fixed α over a range of s_b. These plots show that the rate of loss observed in simulations is indeed exponential in α.

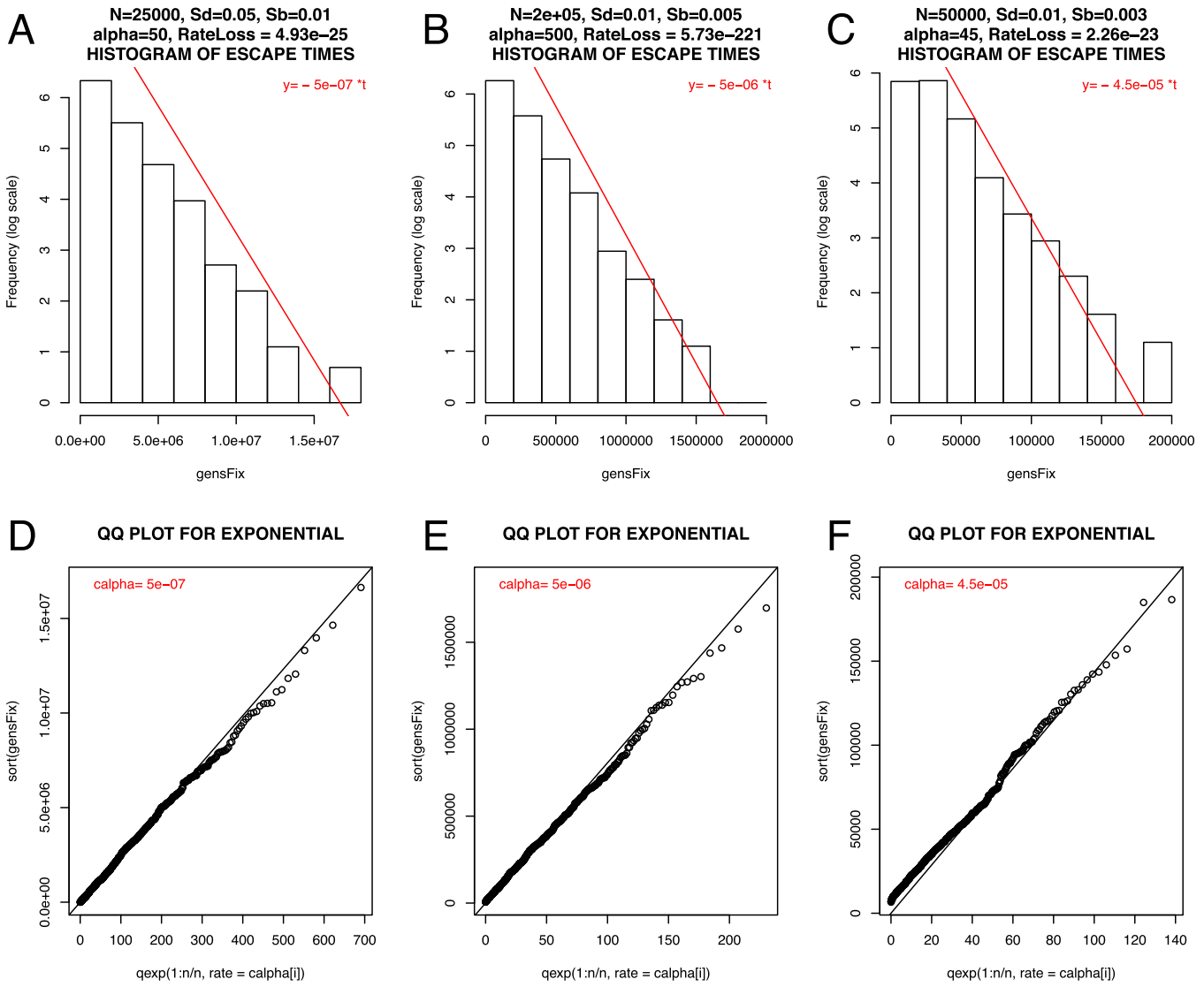

Fig. S4. Histograms and QQ plots demonstrating that the escape times in the $\alpha > 1$ regime are exponentially distributed (a fuller set of simulation results across parameter regimes can be seen in Fig. S5). *A–C* consist of histograms of escape times from 1,000 simulations, such that the *x* axis is the escape time *t* and the *y* axis is the frequency of events on a log scale, and where the parameter set used is indicated in the histogram title such that *A* uses $N = 25{,}000$, $s_b = 0.01$, $s_d = 0.05$; *B* uses $N = 200{,}000$, $s_b = 0.005$, $s_d = 0.01$; and *C* uses $N = 50{,}000$, $s_b = 0.003$, $s_d = 0.01$. Note that if escape times are indeed exponentially distributed with rate λ_e, then we expect these histograms to be described by the line $log(y) \approx -\lambda_e x$, which is indeed the case and can be seen by the red line, which is an exponential density curve with the analytically predicted rate parameter λ_e. Note that short escape times ($1/\lambda_e = \tau_e \ll \ln[Ns_b]/s_b$) will not be exponentially distributed, sometimes causing an uptick in the distribution for escape times near zero. *D–F* consist of the corresponding QQ plot for each histogram above it (such that *D* uses the same variables as *A*, etc.), where the *x* axis is the theoretical quantiles for an exponential distribution with rate λ_e and the *y* axis is the sample quantiles of escape times obtained from simulations. The fact that the points lie along the straight line indicate that the escape times are likely exponentially distributed.

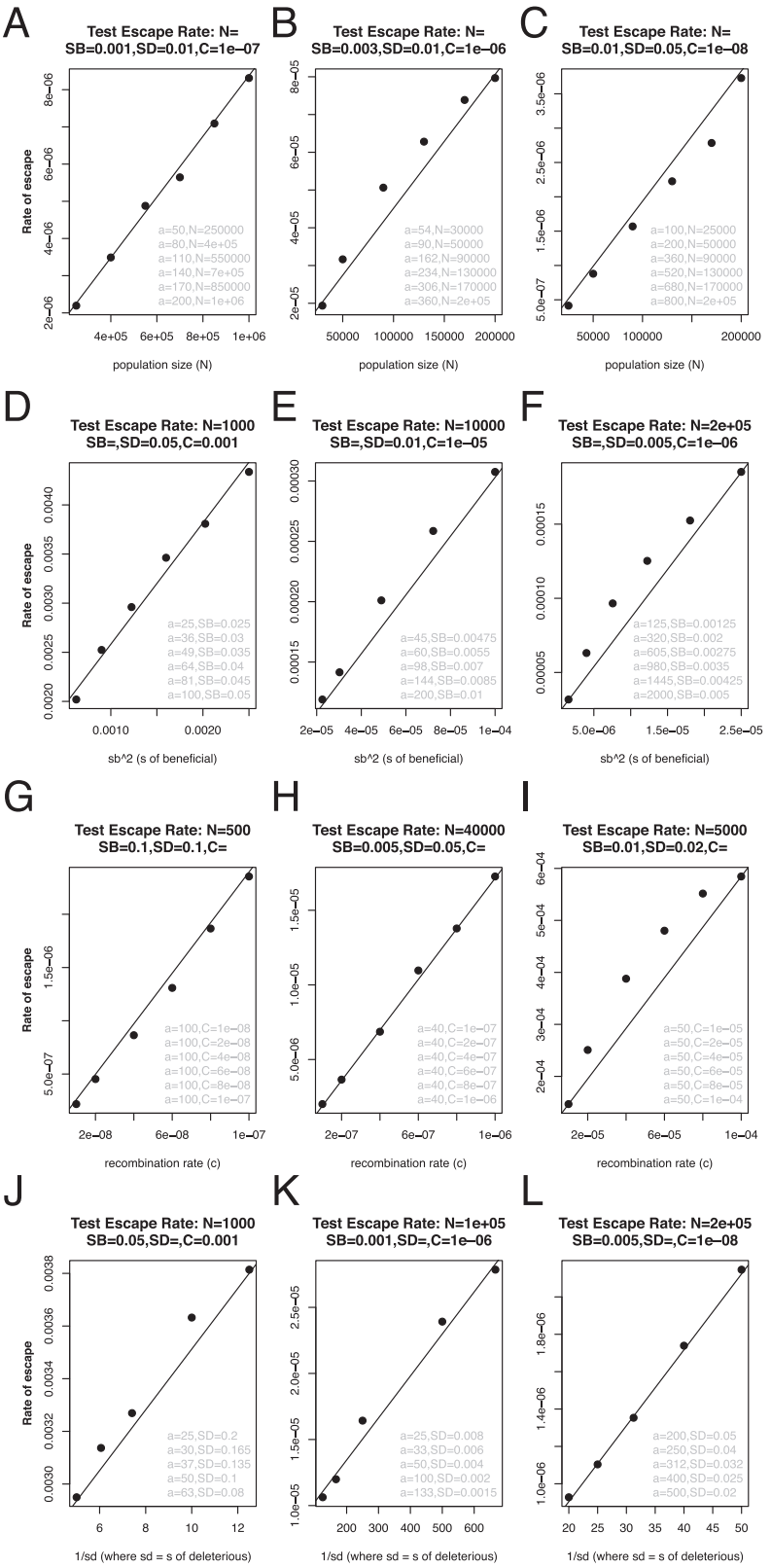

Fig. S5. Testing whether the rate of escape indeed scales with $c\alpha = Ns_b^2/s_d$. Each subplot contains results from a set of simulations in which three parameters were held constant and one parameter varied, such that the observed rate of escape should scale with the varied parameter (as written on the *x* axis). Each data point indicates the results from 1,000 simulations, where an exponential distribution was fit to the escape times using the R function fitdistr from the package MASS and this inferred rate of escape recorded (*y* axis). Parameter values used (for $N, s_b, s_d, c = 0$) are indicated in the heading of each subplot (with blank values for the parameter varied, details for which are found in the gray text). For example, in *A*, each black data point indicates the results from 1,000 simulations in which $s_b = 0.001$, $s_d = 0.01$, $c = 10^{-7}$, and N is varied across data points where the value of N used is indicated by the gray text with the

Legend continued on following page

corresponding α (written "a") next to it. *A*–*D* use fixed values of s_b, s_d, and *c* (values indicated in headings) and varied *N* (values indicated in gray text), *D*–*F* use fixed values of *N*, s_d, and *c* and varied s_b, *G*–*I* use fixed values of *N*, s_b, and s_d and varied *c*, and *J*–*L* use fixed values of *N*, s_b, and *c* and varied s_d. These plots show that the rate of escape scales with $c\alpha$.

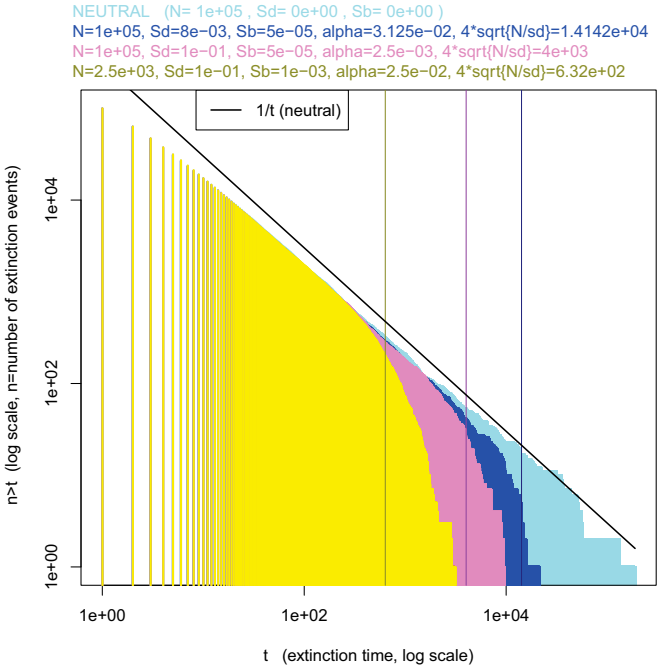

Fig. S6. Histogram of extinction times in the $\alpha < 1$ regime, where each color indicates the results from 100,000 simulations using the parameter set indicated at the top of the plot in the corresponding color. The light blue histogram indicates the results from a neutral simulation, where the black $y = 1/t$ line is the corresponding expectation for the distribution under neutrality. The dark blue, pink, and yellow histograms show that under the $\alpha < 1$ regime, the distribution of extinction times looks neutral up until a cutoff at $t \approx 4\sqrt{N/s_d}$.

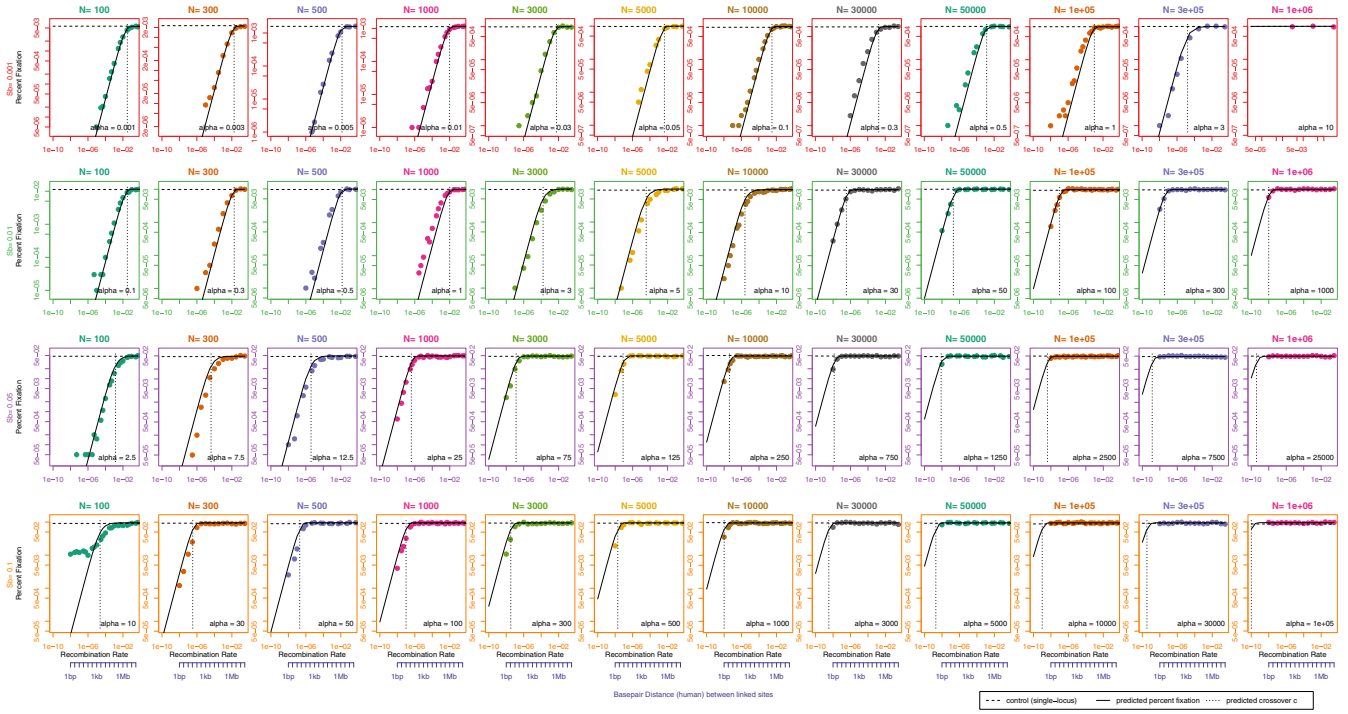

Fig. S7. Probability of fixation (*y* axis) vs. the recombination rate *c* (*x* axis). These plots show the comparison of simulations to analytics, where each data point is the result from 1,000/s_b simulations, and the horizontal dashed line is the expectation for a single beneficial mutation with no deleterious hitchhiker. Note that the population size *N* is indicated at the top of each column in a font color that corresponds to the color of the data points, the recessive deleterious effect is $s_d = 0.1$, and the beneficial mutation effect s_b is indicated at the left side of each row in a font color that corresponds to the color of the plot axes. Gray vertical dashed lines indicate the predicted crossover point c_l, such that recombination distances below this have a substantially decreased probability of fixation compared with a one-locus control beneficial mutation with no hitchhiker.

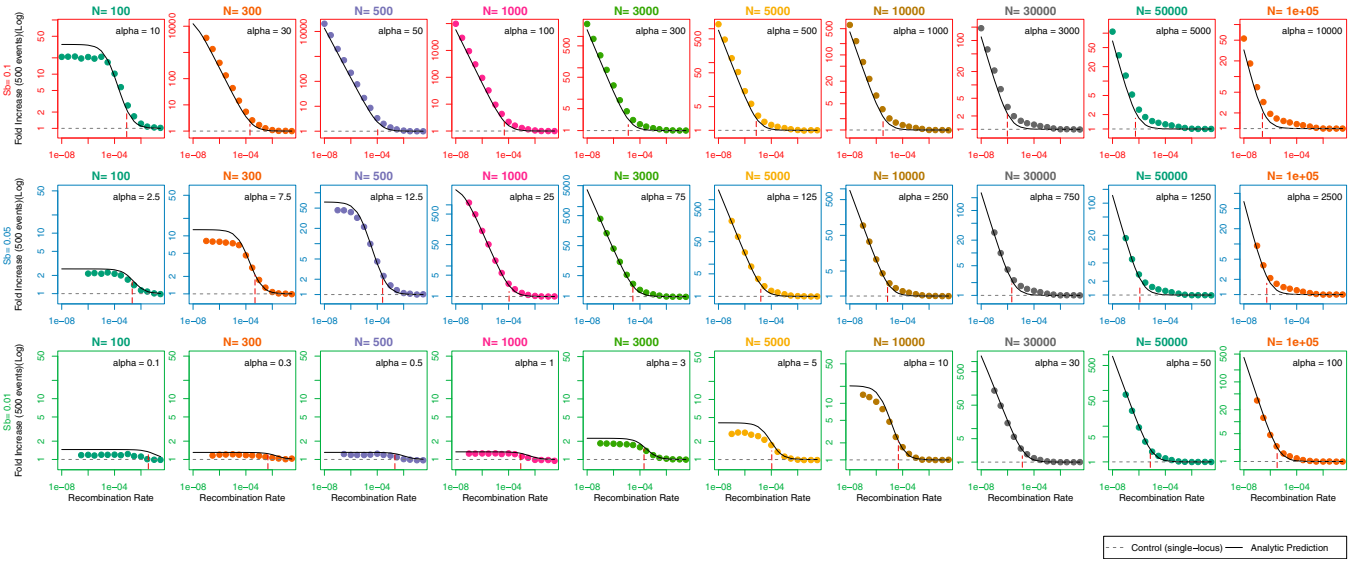

Fig. S8. Sweep time (*y* axis) vs. the recombination rate *c* (*x* axis). These plots show the comparison of simulations to analytics, where each data point is the result from 500 simulations in which fixation occurred, and the horizontal dashed line is the expectation for the sweep time of a single beneficial mutation with no deleterious hitchhiker. Note that the population size *N* is indicated at the top of each column in a font color that corresponds to the color of the data points, the recessive deleterious effect is $s_d = 0.1$, and the beneficial mutation effect s_b is indicated at the left side of each row in a font color that corresponds to the color of the plot axes. Red vertical dashed lines indicate the predicted crossover point c_e, such that recombination distances below this have an increased sweep time compared with a one-locus control beneficial mutation with no hitchhiker.

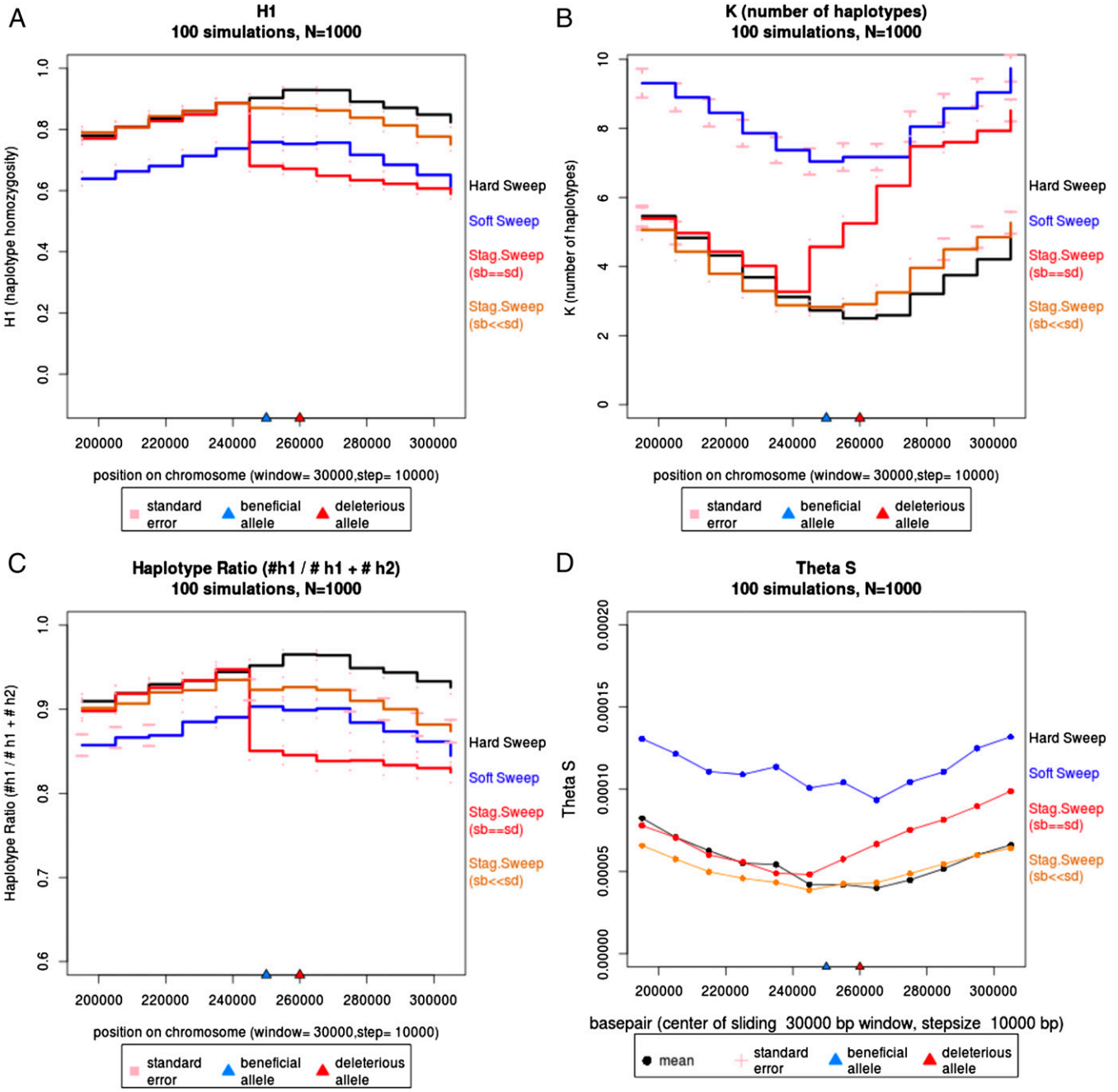

Fig. S9. Additional sweep signatures, averaged across 100 simulations using a diploid population size of N = 1,000 and a beneficial mutation effect size $s_b = 0.05$, where black lines indicate a hard sweep, blue lines indicate a soft sweep ($N\mu_b = 1$), red lines indicate a staggered sweep where $s_d = s_b = 0.05$, and orange lines indicate a staggered sweep where $s_d = 0.5$ such that $s_b \ll s_d$. For calculating statistics a window size of 30,000 base pairs with step size of 10,000 base pairs was used. Pink bars are SEMs. (*A*) The homozygosity of the most common haplotype ($H1 = p_1^2$), (*B*) the number of haplotypes in the population ("K"), (*C*) the ratio of the most common to the second-most common haplotype, and (*D*) θ_s diversity.

Table S1. Estimates for *Drosophila melanogaster* and humans for the proportion of the genome in which the sweep time of the beneficial mutation is extended

Organism and population size	Number of coding genes	Recessive deleterious effect (s_d), %	Zone of extended sweep time [genes in zone]	Beneficial effect (s_b) impacted, %	Density of recessive deleterious	Proportion of adaptive mutations impacted, %
Drosophila	~12,000					
$N = 10^6$		100	100 b [0.01 gene]	~5	1/(genome)	~0
		5	10 b [0.001 gene]	~5	1/(30 genes)	~0.003
		1	10 b [0.001 gene]	~1	1/(30 genes)	~0.003
$N = 10^3$		100	1 Mb [130 genes]	~5	1/(genome)	~1
		5	10 kb [3 genes]	~5	1/(30 genes)	~10
		1	100 kb [20 genes]	~1	1/(30 genes)	~65
$N = 10^2$		100	—	~5	1/(genome)	—
		5	1 Mb [130 genes]	~5	1/(30 genes)	~100
		1	10 Mb [10^3 genes]	~1	1/(30 genes)	~100
Human ($n = 10^4$)	~20,000	100	100 kb [3 genes]	~1	1/(genome)	~0.01
		1	100 kb [3 genes]	~0.10	1/(100 genes)	~3
					1/(30 genes)	~10
					1/(10 genes)	~30

Dashes indicate no effect on sweep time due to $\alpha \leq 1$, in which case the beneficial mutation effect size of interest is unlikely to sweep due to the recessive deleterious variation in the genome.