Selective sweeps under dominance and self-fertilisation

Matthew Hartfield[1,2,*], Thomas Bataillon[2]

1 Department of Ecology and Evolutionary Biology, University of Toronto, Ontario, Canada.

2 Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark.

* matthew.hartfield@birc.au.dk

Running Head: Sweeps under dominance and selfing

Key words: Adaptation; Dominance; Self-fertilisation; Selective Sweeps; *SLC24A5*

Abstract

A major research goal in evolutionary genetics is to uncover loci experiencing adaptation from genomic sequence data. One approach relies on finding 'selective sweep' patterns, where segregating adaptive alleles reduce diversity at linked neutral loci. Recent years have seen an expansion in modelling cases of 'soft' sweeps, where the common ancestor of derived variants predates the onset of selection. Yet existing theory assumes that populations are entirely outcrossing, and dominance does not affect sweeps. Here, we develop a model of selective sweeps that considers arbitrary dominance and non-random mating via self-fertilisation. We investigate how these factors, as well as the starting frequency of the derived allele, affect average pairwise diversity, the number of segregating sites, and the site frequency spectrum. With increased self-fertilisation, signatures of both hard and soft sweeps are maintained over a longer map distance, due to a reduced effective recombination rate and faster fixation times of adaptive variants. We also demonstrate that sweeps from standing variation can produce diversity patterns equivalent to hard sweeps. Dominance can affect sweep patterns in outcrossing populations arising from either a single novel mutation, or from recurrent mutation. It has little effect where there is either increased selfing or the derived variant arises from standing variation, since dominance only weakly affects the underlying adaptive allele trajectory. Different dominance values also alters the distribution of singletons (derived alleles present in one sample). We apply models to a sweep signature at the *SLC24A5* gene in European humans, demonstrating that it is most consistent with an additive hard sweep. These analyses highlight similarities between certain hard and soft sweep cases, and suggest ways of how to best differentiate between related scenarios. In addition, self-fertilising species can provide clearer signals of soft sweeps than outcrossers, as they are spread out over longer regions of the genome.

Author Summary

Populations adapt by fixing beneficial mutations. As a mutation spreads, it drags linked neutral variation to fixation, reducing diversity around adaptive genes. This footprint is known as a 'selective sweep'. Adaptive variants can appear either from a new mutation onto a single genotype; from recurrent mutation onto different genotypes; or from existing genetic variation. Each of these sources leaves subtly different selective sweep patterns in genetic data, which have been explored under simple biological cases. We present a general model of selective sweeps that includes self-fertilisation (where individuals produce both male and female gametes to fertilise one another), and dominance (where fitness differences exist between one and two gene copies within an individual). Soft sweep patterns are spread out over longer genetic regions in self-fertilising individuals, while dominance mainly affects sweeps in outcrossers from either a single or recurrent mutation. Applying models to a sweep signal associated with human skin pigmentation shows that this mutation was likely introduced into Eurasia from Africa in very few numbers. These models demonstrate to what extent soft sweeps can be detected in genome data, and how self-fertilising organisms can be good study systems for determining the extent of different adaptive modes.

Introduction

Inferring adaptation from nucleotide sequence data is a major research goal in evolutionary genetics. The earliest models focussed on the scenario where a beneficial mutation appeared in the population in a single copy before rapidly spreading to fixation. Linked neutral mutation would then 'hitchhike' to fixation with the adaptive variant, reducing diversity around the selected locus [1, 2]. Hitchhiking also causes a rapid increase in linkage disequilibrium at flanking regions to the selected site, although it is minimal when measured either side of the beneficial mutation [3–5]. These theoretical expectations have spurred the creation of summary statistics for detecting sweeps, based on finding regions of the genome exhibiting extended runs of homozygosity [6–10].

Classic hitchhiking models consider 'hard' sweeps, where the common ancestor of adaptive alleles occurs after its appearance [11]. Yet the last fifteen years have seen a focus on quantifying 'soft' sweeps, where the most recent common ancestor of the beneficial allele arose before the variant became selected for (reviewed in [11–13]). Soft sweeps can originate from beneficial mutations being introduced by recurrent mutation [14, 15], or from existing standing variation that was either neutral or deleterious [16–22]. A key property of soft sweeps is that the beneficial variant is present on multiple genetic backgrounds as it sweeps to fixation, so different haplotypes are present around the derived allele. This property is often used to detect soft sweeps in genetic data [23–28]. Soft sweeps have been inferred in several organisms, including *Drosophila* [25, 26], humans [23, 29], maize [30] and the malaria pathogen *Plasmodium falciparum* [31], although determining how extensive soft sweeps are in nature remains a contentious issue [32].

Up to now, almost all models of selective sweeps have made the same simplifying assumptions. In particular, there have been few analyses considering how dominance affects sweep signatures. In a simulation study, Teshima and Przeworski [33] determined how recessive mutations spend a long period of time at low frequencies, increasing the amount of recombination that acts on derived haplotypes, weakening signatures of 'hard' sweeps. Fully recessive mutations may need a long time to reach a high enough frequency so that they can be picked up by genome scans for adaptive loci [34]. Ewing *et al.* [35] have carried out a general mathematical analysis of dominance on 'hard' sweeps on genetic diversity. Yet the impact that dominance has on 'soft' sweeps has yet to be explored in depth.

In addition, existing models have overwhelmingly assumed that populations are sexual, with individuals haplotypes freely mixing between individuals. Different reproductive modes alters how alleles are inherited over subsequent generations and spread over time, therefore altering the hitchhiking effect. In particular, there is a renewed interest in studying the mechanisms of adaptation in self-fertilising species [36]. Self-fertilisation, where male and female gametes produced from the same individual can fertilise each other, is prevalent amongst angiosperms [37], some animals [38] and fungi [39]. Different levels of self-fertilisation is known to affect overall adaptation rates. Dominant mutations are likelier to fix then recessive ones in outcrossers, as they have a higher initial selection advantage [40]. Yet recessive alleles can fix more easily in selfers than in outcrossers as they rapidly create homozygote mutations [41, 42]. Hence the effects of dominance and self-fertilisation become strongly intertwined, so it is important to consider both together. Furthermore, a decrease in effective recombination rates in selfers [43] can amplify the effects of linked selection, making it likelier that deleterious mutations hitchhike to fixation with adaptive alleles [44], or nearby beneficial alleles are lost if one is already spreading through the population [45].

Self-fertilisation is also known to affect the degree to which adaptation proceeds from *de novo* mutation, or from standing variation. In a constant-sized population, fixation of beneficial mutations from standing variation (either neutral or deleterious) is generally less likely in selfers as lower levels of diversity are maintained [46]. Yet if adaptation from standing variation does occur, then the beneficial variant fixes more quickly in selfers than outcrossers, hence signatures of soft sweeps could become more marked [42, 46].

Furthermore, adaptation from standing variation becomes likelier in selfers under 'evolutionary rescue' scenarios, where swift adaptation needed to prevent population extinction. This is because the population size is greatly reduced, so the waiting time for the appearance of *de novo* rescue mutations becomes excessively long. Hence only adaptive mutations present in standing variation can contribute to preventing population extinction [46]. High selfing rates can further aid this process by creating beneficial homozygotes more rapidly than in outcrossing populations [47]. Therefore there is potential for soft sweeps to act in selfing organisms.

However, little data currently exists on the extent of soft sweeps in self-fertilisers. Many selfing organisms exhibit sweep-like regions, including *Arabidopsis thaliana* [48–50]; *Caenorhabditis elegans* [51]; *Medicago truncatula* [52]; and *Microbotryum* fungi [53]. Detailed analyses of these regions has been hampered by a lack of theory on how hard and soft sweep signatures should manifest themselves under different levels of self-fertilisation and dominance. Previous studies have only focussed on special cases; Hedrick [54] analysed linkage disequilibrium caused by a hard sweep under self-fertilisation, while Schoen *et al.* [55] modelled sweep patterns caused by modifiers that altered the mating system in different ways. A knowledge of expected diversity patterns following different types of sweeps can also be used to create more realistic statistical models for finding and quantifying novel adaptive candidate loci, while accounting for the mating system.

We present here a general model of selective sweeps. We determine the genetic diversity present following a sweep from either a *de novo* mutation, or from standing variation. The model assumes an arbitrary level of dominance and self-fertilisation. We first present general results for the probability of how genetic samples, carrying a recently-fixed beneficial mutation, are affected by recombination, dominance and selfing. We next determine how key summary statistics (pairwise diversity; number of segregating sites; and the site frequency spectrum) are affected by this general sweep model from standing variation. These results are compared to an alternative soft-sweep case where adaptive alleles arise via recurrent mutation. We also include a simulation study of how the distribution of singletons are affected under different sweep scenarios, complementing a recent study that used singleton densities to detect recent human adaptation [56]. We end by applying models to determine the history of a selective sweep at the *SLC24A5* gene in humans, to evaluate the evolutionary history of this adaptation, and determine if evidence exists for either non-additive dominance or a soft sweep signature.

Results

Model Outline

We consider a diploid population of size N (carrying $2N$ haplotypes in total). Individuals reproduce by self-fertilisation with probability σ, and outcross with probability $1 - \sigma$. The level of self-fertilisation can also be captured by the inbreeding coefficient $F = \sigma/(2 - \sigma)$ [57, 58]. There are two biallelic loci A, B with a recombination rate r between them. Locus A represents a region where neutral polymorphism accumulates under an infinite-sites model. Locus B determines fitness differences, carrying an allele that initially segregates at low frequency for a sizeable period of time. We are agnostic as to whether this allele is neutral or subject to weak selection, but note that an allele subject to strong purifying selection would have only recently appeared in the population, which we do not consider. Once the allele reaches a frequency f_0 it becomes advantageous, with selective advantage $1 + hs$ in heterozygote form and $1 + s$ as a homozygote, with $0 \leq h \leq 1$ and $s > 0$. We further assume that selection is strong (i.e., $N_e hs \gg 1$) so that the sweep trajectory can be modelled deterministically. Table 1 lists notation used in the model analysis.

Our overall goal is to determine how the emergence of an adaptive allele from standing variation at locus B affects genealogies underlying polymorphism at locus A. We model the genetic histories at A while considering the genetic background of neutral alleles (i.e., whether they are linked to the selected derived allele or ancestral neutral allele at locus B). A schematic of the process is shown in Fig 1. We follow the approach of Berg and Coop [21] and, looking backwards in time, break down the allele history into two phases. The first phase (the 'sweep phase')

Symbol	Usage
N	Population size (with $2N$ haplotypes)
σ	Proportion of matings that are self-fertilising
F	Wright's inbreeding coefficient, $\sigma/(2-\sigma)$ [57, 58]
N_e	Effective population size, equal to $N/(1+F)$ with selfing [59]
A, B	Loci carrying neutral, selected alleles
r	Recombination rate between loci A, B
r_{eff}	'Effective' recombination rate, approximately equal to $r(1-F)$ with selfing [43]
R	$2Nr$, the population-level recombination rate
f_0	Frequency at which the derived allele at B becomes advantageous
$f_{0,A}$	'Accelerated' effective starting frequency of B appearing as a single copy, conditional on fixation
s	Selective advantage of derived allele at B
h	Dominance coefficient of derived allele at B
t	Number of generations in the past from the present day
τ_{f_0}	Time in the past when derived locus became beneficial
$p(t)$	Frequency of beneficial allele at time t
P_{NR}	Probability that neutral marker does not recombine onto ancestral background during sweep phase
$P_{NR}(i\|n)$	Probability that i of n neutral markers do not recombine during sweep phase
H_l, H_h	'Effective' dominance coefficient for allele at low, high frequency
P_{coal}	Probability that two samples coalesce in the standing phase
$P_{coal,M}$	Probability that two samples coalesce instead of arising by different mutations
π	Pairwise diversity at site (π_0 is expected value without selection)
π_{SV}	Pairwise diversity following sweep from standing variation
π_M	Pairwise diversity following sweep from recurrent mutation
$\tilde{s}$	'Effective' selection coefficient to map hard sweep onto standing variation cases
$P_{ESF}(k\|i)$	Ewens' Samping Formula for the probability of k ancestral backgrounds formed from i non-recombined lineages
$\mathbb{E}(T_{tot})$	Expected time covered by entire genealogy
$\mathbb{E}(S)$	Expected number of segregating sites
μ	Probability of neutral mutation occurring per site per generation
μ_b	Probability of beneficial mutation occurring at target locus per generation
$\theta = 4N_e\mu$	Population level neutral mutation rate
$\Theta_b = 2N_e\mu_b$	Population level beneficial mutation rate

Table 1. Glossary of Notation.

considers the derived allele at B being selectively favoured and spreading through the population. The length of this phase is assumed to be sufficiently short ($t \sim 1/s$) so that no samples coalesce during this time, but they can recombine onto the ancestral background. The second phase (the 'standing phase') assumes that the derived allele is present at a fixed frequency f_0. Here, the two samples can either coalesce, or one of them recombines onto the ancestral background. Berg and Coop [21] showed that this assumption allows traditional coalescent results to be used to infer genetic properties of the sweep, after appropriate rescaling of the coalescent rate by f_0.

For tightly linked loci ($r \to 0$), the relatively rapid fixation time of the derived variant makes it unlikely for unique polymorphisms to arise on different haplotypes, reducing neutral diversity. Further from the target locus, recombination can transfer allele copies at A away from the selected background to the ancestral background, so diversity reaches neutral levels.

Self-fertilisation creates two key differences compared to traditional outcrossing models. First, the effective population size and recombination rate are scaled by factors $1/(1+F)$ and $1-F$ respectively [43,58]. Second, the trajectory of adaptive alleles, which determines expected diversity patterns following adaptation, depends on the levels of self-fertilisation (σ) and dominance (h). A goal of this analyses will be to determine how these processes interact to affect neutral variation following a sweep, and therefore the ability to detect different types of recent adaptation.

Throughout, analytical solutions are compared to results obtained from Wright-Fisher forward-in-time stochastic simulations. The simulation procedure itself is described in the 'Methods' section.

Probability of no recombination during sweep phase

Looking back in time following a sweep, sites linked to the beneficial allele can recombine onto the ancestral genetic background, so they exhibit the same diversity as putatively neutral regions. Let $p(t)$ be the frequency of the adaptive mutation at time t, defined as the number of generations prior to the present day. Further define $p(0) = 1$ (i.e., the allele is fixed at the present day), and τ_{f_0} the time in the past when the derived variant became beneficial (i.e., $p(\tau_{f_0}) = f_0$). If the neutral locus lies at a recombination distance r from the derived variant, then the probability that it will not recombine onto a neutral background is $1 - r(1 - p(t))$ [21]. We also define $r = r_{eff} = r(1 - F)$, which is the 'effective' recombination rate after accounting for the increased homozygosity created due to self-fertilisation [43]. More exact r_{eff} terms exist [60,61], but they are approximately equal to $r(1 - F)$ over short map distances. Using these more exact terms do not improve the accuracy of the analytical model relative to simulations for the parameters used (data not shown).

Over τ_{f_0} generations, the total probability that a single lineage does not recombine onto a neutral background, P_{NR}, equals:

$$\begin{aligned}
P_{NR} &= \prod_{t=0}^{\tau_{f_0}} (1 - r_{eff}(1 - p(t))) \\
&\approx \exp\left(-r_{eff} \int_{t=0}^{\tau_{f_0}} (1 - p(t))\mathrm{d}t\right) && \text{since } r_{eff} \ll 1 \\
&\approx \exp\left(-r_{eff} \int_{p=1}^{f_0} \frac{(1 - p(t))}{\mathrm{d}p/\mathrm{d}t}\mathrm{d}p\right) && \text{integrating over } p
\end{aligned} \tag{1}$$

We can calculate P_{NR} for general levels of self-fertilisation if the selection coefficient is not too weak (i.e., $1/N_e \ll s \ll 1$). Here the rate of change of the allele frequency is given by [42]:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -sp(1-p)(F+h-Fh+(1-F)(1-2h)p) \tag{2}$$

Note the negative factor in Eq 2 since we are looking back in time. By substituting Eq 2 into Eq 1, we obtain the following analytical solution for P_{NR}:

$$\begin{aligned} P_{NR} &= \exp\left(-\frac{r_{eff}}{H_l s}\log\left(1+\frac{H_l}{H_h}\left(\frac{1}{f_0}-1\right)\right)\right) \\ &= \left(1+\frac{H_l}{H_h}\left(\frac{1}{f_0}-1\right)\right)^{-r_{eff}/(H_l s)} \end{aligned} \tag{3}$$

Here, $H_l = F+h-Fh$, $H_h = 1-h+Fh$ are the 'effective' dominance coefficients when the beneficial variant is at a low or high frequency [42]. We can understand Eq 3 as follows. The beneficial mutation takes $(1/H_l s)\log(1+(H_l/H_h)(1/f_0-1))$ generations to go to fixation from initial frequency f_0. The rate at which the allele spreads depends on the ratio of the effective dominance coefficients H_l, H_h. These terms mediate the relative amount of time a beneficial allele spends at low and high frequencies, affecting the probability that a neutral marker recombines away from the selected background. Looking back in time, a proportion r_{eff} of neutral markers become unlinked from the beneficial allele each generation, so when the allele reaches its starting frequency f_0 a proportion P_{NR} of neutral markers remain linked to it [62].

Note that for the special case $F = 0$ and $h = 1/2$, $H_l = H_h = 1/2$ and Eq 3 reduces to $(1/f_0)^{-(2r/s)}$. This is a standard result for the reduction of diversity following a sweep in outcrossing models with additive dominance [1, 21, 62, 63].

Probability of coalescence from standing variation

When the variant becomes advantageous at frequency f_0, we expect $\sim 2Nf_0$ haplotypes will carry it. We assume that f_0 remains fixed in time, so that different events occur with constant probabilities. Berg and Coop [21] have shown this assumption provides a good approximation to coalescent rates during the standing phase. The outcome during the standing phase can therefore be determined by considering two competing Poisson processes. The two haplotypes could coalesce; the waiting time for this event is exponentially distributed with rate $1/(2N_e f_0) = (1+F)/(2Nf_0)$, assuming N_e is reduced by a factor $1+F$ due to self-fertilisation [59]. Alternatively, one of the two samples could recombine onto the ancestral background; the exponential mean time for this event is $2r_{eff}(1-f_0)$ (note the factor of two as there are two samples under consideration). For two competing exponential distributions with rates λ_1 and λ_2, the probability of the first event occurring *given an event happens* equals $\lambda_1/(\lambda_1+\lambda_2)$ [64]. Hence the probability that two samples coalesce instead of recombine equals:

$$P_{coal} = \frac{\frac{1+F}{2Nf_0}}{\frac{1+F}{2Nf_0} + 2r_{eff}(1-f_0)} = \frac{1}{1+2R(1-F)f_0(1-f_0)/(1+F)} \tag{4}$$

where $R = 2Nr$ is the population-scaled recombination rate. Note the presence of the $(1-F)/(1+F) = 1-\sigma$ term, reflecting how selfing reduces the relative effect of recombination by this factor (by both increasing homozygosity, and reducing N_e so coalescence becomes more likely). Hence for a fixed recombination rate R, samples are more likely to coalesce with increased self-fertilisation, limiting the creation of different background haplotypes. Yet the same coalescent probability can be recovered by increasing the recombination distance by a factor $1/(1-\sigma)$; that is, if a longer genetic region is analysed.

Effective starting frequency from a de novo mutation

When a new beneficial mutation appears at a single copy, it is highly likely to go extinct by chance [40]. Beneficial mutations that increase in frequency faster than expected when rare are more able to overcome this stochastic loss and reach fixation. These beneficial mutations will hence display an apparent 'acceleration' in their logistic growth, equivalent to having a starting frequency that is greater than $1/(2N)$ [1, 65–67]. In Section A of the Supplementary *Mathematica* file (S1 File; S2 File for PDF copy), we outline how to determine the 'effective' starting frequency of hard sweeps that go to fixation, by comparing the sojourn time for the deterministic process to the stochastic diffusion process. We determine that 'hard' sweeps that go to fixation have the following elevated effective starting frequency:

$$f_{0,A} = \frac{1+F}{4NsH_l} \tag{5}$$

where $H_l = F + h - Fh$ is the effective dominance coefficient for mutations at a low frequency. This result is consistent with those obtained by Martin and Lambert [67], who obtained a distribution of effective starting frequencies using stochastic differential equations.

This acceleration effect can create substantial increases in the apparent f_0. The effect is strongest for recessive mutations; for example, for $N = 5,000$ and

$s = 0.05$ (as used in simulations below), $f_{0,A} = 0.01$ for recessive mutations with $h = 0.1$, an 100-fold increase above $f_0 = 1/(2N) = 0.0001$. $f_{0,A}$ is more modest for additive and dominant mutations; Eq 5 reduces to $1/2Ns$ with $h = 1/2$ or $F = 1$. Hence sweeps from standing variation whose actual f_0 lies below $f_{0,A}$ will produce sweep signatures that may appear similar to hard sweeps. As a consequence, in simulations we use a minimum $f_0 = 0.02$ for adaptation from standing variation cases, which lies above the highest possible value of $f_{0,A}$ for this parameter set.

Expected Pairwise Diversity

We can use P_{NR} and P_{coal} to calculate the expected pairwise diversity (denoted π) present on a genetic fragment flanking a beneficial allele following a sweep. Looking back in time, one of two possible outcomes can arise. Either two neutral sites linked to the adaptive mutant do not recombine during the sweep phase, and proceed to coalesce during the standing phase. This outcome occurs with probability $P_{NR} \cdot P_{coal}$, creating identical genotypes ($\pi = 0$) since this process occurs rapidly compared to the rate of neutral coalescence. Alternatively, one of the two samples will recombine onto the ancestral background with probability $1 - (P_{NR} \cdot P_{coal})$, so the samples will exhibit background neutral levels of diversity ($\pi = \pi_0$). Hence expected diversity following a sweep equals:

$$\begin{aligned}\mathbb{E}\left(\frac{\pi}{\pi_0}\right) &= 1 - (P_{NR} \cdot P_{coal}) \\ &= \left(\frac{1}{1 + 2R(1-F)f_0(1-f_0)/(1+F)}\right) \cdot \left(1 + \frac{H_l}{H_h}\left(\frac{1}{f_0} - 1\right)\right)^{-r(1-F)/(H_l s)}\end{aligned} \tag{6}$$

Eq 6 reflects similar formulas for diversity following soft sweeps in haploid outcrossing populations [15, 21]. Fig 2 plots Eq 6 with different dominance, self-fertilisation, and standing frequency values. The analytical solution fits well compared to simulations, although some inaccuracies appear when the mutation appears from a single initial copy. Under complete outcrossing, baseline levels of diversity are restored (i.e., $\pi/\pi_0 \to 1$) closer to the sweep origin for recessive mutations ($h = 0.1$), compared to co-dominant ($h = 0.5$) or dominant ($h = 0.9$) mutations. Hence recessive mutations produce weaker signatures of selective sweeps. Dominant and co-dominant mutations produce similar reductions in genetic diversity, so these cases may be hard to differentiate between from diversity data alone.

These patterns can be understood in terms of the underlying allele trajectories (Fig 3). For outcrossing populations, recessive mutations spend most of the sojourn time at low frequencies, maximising the number of recombination events over the sweep history, restoring neutral variation. These trajectories mimic those of sweeps from standing variation, which spend extended periods of time at low frequencies in the standing phase. Conversely, dominant mutations spend most of their time at a high frequency, so there is less chance for neutral markers to recombine onto the ancestral background. Similar results were found by Teshima and Przeworski [33].

As the degree of self-fertilisation increases, sweep signatures become similar to the co-dominant case as the derived allele is more likely to spread as a homozygote, reducing the influence that dominance exerts over beneficial allele trajectories (Fig 3(b)). In addition, sweep signatures stretch over longer physical regions due to the reduced effective recombination rate [43]. Increasing f_0 also causes sweeps with different dominance coefficients to produce comparable signatures. Here, beneficial mutation trajectories become alike after conditioning on starting at an elevated frequency. In particular, recessive mutations no longer spend the majority of their sojourn times at low frequencies, reducing the probability that neutral markers can recombine onto ancestral backgrounds (Fig 3(d)–(f)).

Overall, it appears that dominance only strongly affects diversity levels for hard sweeps in outcrossing populations. With increased levels of self-fertilisation, or if the mutation arises from standing variation, allele trajectories (and expected diversity patterns) become similar across different dominance values.

Different Sweep Scenarios can Yield Virtually Identical Signatures

Visual inspection of Fig 2 suggests that different sweep scenarios can produce equivalent reductions in genetic diversity. For example, reductions in diversity caused by a recessive mutation ($h < 0.1$) might be similar to those caused by a mutation with additive dominance ($h = 0.5$) but with a weaker selection coefficient. Similarly, a sweep from standing variation can be mistaken for a weaker hard sweep. Determining how different scenarios cause similar reductions in genetic diversity is useful when testing the most plausible sweep model underlying observed diversity patterns. Berg and Coop [21] argued that it was not possible to find an 'effective selection coefficient' $\tilde{s}$ that maps $\mathbb{E}(\pi/\pi_0)$ for a hard sweep onto results expected under a sweep from standing variation. However, we demonstrate in Section A of S3 File (with mathematical analyses in Section C of S1 File) how the argument of Berg and Coop [21] relies on an approximation that only holds when the population-level recombination rate is extremely low (specifically, when $4Nrf_0(1 - f_0) \ll 1$).

In fact, a sweep arising from standing variation with selective advantage s

can be mapped onto a hard sweep with intensity $\tilde{s}$, with general self-fertilisation and $h = 1/2$ (it does not appear possible to obtain a solution for any h). We equate Eq 6 for general f_0 to the special hard-sweep case $f_0 = 1/2N$ with selection coefficient $\tilde{s}$ (we do not use $f_{0,A}$ for the hard sweep to calculate tractable analytic solutions). After equating the two cases and solving for $\tilde{s}$, we obtain:

$$\begin{aligned}\tilde{s} &= -2r(1-\sigma)\log(2N)\left(\log\left(\left(\frac{1}{f_0}\right)^{-\frac{2r(1-\sigma)}{s}}\frac{1+r(1-\sigma)(2N-1)/(N)}{1+4Nr(1-\sigma)f_0(1-f_0)}\right)\right)^{-1}\\ &\approx -2r(1-\sigma)\log(2N)\left(\log\left(\left(\frac{1}{f_0}\right)^{-\frac{2r(1-\sigma)}{s}}\frac{1}{1+4Nr(1-\sigma)f_0(1-f_0)}\right)\right)^{-1} \qquad (7)\end{aligned}$$

The approximation in Eq 7 assumes $r_{eff}(2N-1)/(N) \ll 1$. To understand $\tilde{s}$, recall that the expected reduction in diversity following a a hard sweep with $f_0 = 1/2N$ is $(2N)^{-2r(1-\sigma)/\tilde{s}}$ (Eq 6, assuming $H_l = H_h = (1+F)/2$ due to additive dominance, and $P_{coal} \approx 1$). Inverting this term and solving for $\tilde{s}$ gives $\tilde{s} = -2r(1-\sigma)\log(2N)/\log(\mathbb{E}(\pi/\pi_0))$. Eq 7 is hence equivalent to the selective coefficient causing a hard sweep, given that the underlying diversity was actually shaped by a mutation arising from standing variation.

Fig 4(a) plots Eq 7 as a function of R, demonstrating that $\tilde{s}$ increases with the recombination rate. $\tilde{s}$ can be either less than or greater than s depending on f_0 and R. Increasing f_0 causes diversity to be restored closer to the beneficial allele as it is likelier that recombination occurs during the standing phase. Hence the $f_0 = 0.1$ case is equivalent to a hard sweep caused by a more weakly selected beneficial allele (Fig 4(b)).

In Section A of S3 File (with mathematical analyses in Section C of S1 File) we show that for an outcrossing population with any f_0, it is possible to find an effective selection coefficient $\tilde{s}_h$ so that a beneficial allele with $h = 1/2$ produces an equivalent sweep pattern to a mutation with arbitrary dominance. We also demonstrate that it is possible to find $\tilde{s}_F$ to map a co-dominant sweep in an outcrossing population onto an equivalent sweep under partial selfing, but this mapping only holds for hard sweeps.

Overall, these results caution that it will be necessary to compare a broad range of models when inferring the likeliest cause of selective sweep patterns, and that identifiability issues are to be expected when trying to determine which sweep model best fits diversity data. An example of these issues in relation to investigating sweep patterns in humans will be outlined in the section "Application to a selective sweep at the human *SLC24A5* gene".

Number of Segregating Sites

We can also calculate the total time underlying the genealogy, $\mathbb{E}(T_{tot})$, and therefore the expected number of segregating sites $\mathbb{E}(S)$. We consider n samples of the derived allele; looking back in time, i of these samples fail to recombine off the derived background during the sweep. The probability of this event can be drawn from a binomial distribution with probability P_{NR}. We denote this value $P_{NR}(i|n) \sim Bin(n, P_{NR})$. Out of these i samples, let k of them recombine during the sweep phase to create different ancestral backgrounds of the derived allele. Berg and Coop [21] demonstrated how the number of lineages that recombine away from the derived background can be determined using Ewens' Sampling Formula:

$$P_{ESF}(k|i) = S(i,k)\frac{R_{f_0}^k}{\prod_{l=1}^{i-1}(R_{f_0}+l)} \tag{8}$$

where $R_{f_0} = 4Nrf_0(1-f_0)$ is the scaled recombination rate acting on the ancestral background at frequency f_0, and $S(i,k)$ are Stirling numbers of the first kind [15, 21, 68]. Here, we use the rescaled version of R_{f_0} accounting for the reduced effective recombination rate and effective population size caused by self-fertilisation (see Eq 4):

$$P_{ESF}(k|i) = S(i,k)\frac{(2R(1-F)f_0(1-f_0)/(1+F))^k}{\prod_{l=1}^{i-1}((2R(1-F)f_0(1-f_0)/(1+F))+l)} \tag{9}$$

Finally, for the k neutral lineages created in the standing phase, along with the $n-i$ neutral lineages created in the sweep phase, the expected total time for the genealogy for all of them, in units of $2N_e$ generations, equals $\sum_{j=1}^{k+n-i-1} 1/j$ [69]. The total time covered by the genealogy is the product of these three terms, summed over all possible outcomes:

$$\mathbb{E}(T_{tot}) = \sum_{i=0}^{n} P_{NR}(i|n)\sum_{k=0}^{i} P_{ESF}(k|i)\sum_{j=1}^{k+n-i-1} 1/j \tag{10}$$

$\mathbb{E}(S)$ is $\theta\mathbb{E}(T_{tot})$ where $\theta = 4N_e\mu$ is the population level mutation rate [70]. Equivalent results for outcrossing populations are given by Pennings and Hermisson [15, Eq. 15] for adaptation from recurrent mutation, and Berg and Coop [21, Eq. 10] for adaptation from standing variation. Both these derivations assume $k > 1$ in the standing phase, as it was argued that $\mathbb{E}(T_{tot}) = 0$ so no segregating polymorphisms exist. Since simulation results show that this outcome is possible under low recombination rates, we do not include this conditioning in Eq 10.

Fig 5 plots $\mathbb{E}(S)$ alongside simulation results. The analytical solution provides a good fit but tends to overestimate simulations, as also observed by Berg and Coop [21]. Also note that fewer segregating sites are present with partial selfing, due to a reduction in the net mutation rate $\theta = 4N_e\mu$ caused by lower N_e.

Site Frequency Spectrum

The calculations for $\mathbb{E}(S)$ can be extended to determine the full site-frequency spectrum (SFS) following a sweep; that is, the probability that out of n samples, $l = 1$, $2 \ldots n - 1$ of them carry derived alleles. The full derivation is outlined in Section B of S3 File, and is similar to that used by Berg and Coop [21, Eq 15]. However we use a different approach when considering special cases where either all or none of the sampled lineages recombine away from the derived background during the sweep phase. In particular, if all lineages recombine away during the sweep phase, then the SFS reduces to the neutral case; if none do then only a singleton class is included to account for new mutations.

Fig 6 plots the expected SFS (Eq B14 in S3 File) alongside simulation results. Analytical results fit simulation data well, although there can be a tendency for it to underestimate the proportion of low- and high-frequency classes ($l = 1$ and 9 in Fig 6), and overestimate proportion of intermediate-frequency classes. Hard sweeps in either outcrossers or partial selfers are characterised by a large amount of singletons or highly-derived variants (Fig 6(a)), which is a typical selective sweep signature [71, 72]. As the initial selected frequency f_0 increases, so does the number of intermediate-frequency variants (Fig 6(b)). This signature is often seen as a characteristic of 'soft' sweeps [15, 21], reflecting the increased number of genetic backgrounds that the beneficial allele appears on. Yet recessive hard sweeps ($h = 0.1$ and $f_0 = 1/2N$) can produce SFS profiles that are similar to sweeps from standing variation, due to the increased number of recombination events occurring over the timespan of the sweep, especially at low frequencies for long periods of time. As with π/π_0, SFS patterns will not unambiguously discriminate between sweep scenarios.

With increased levels of self-fertilisation, both hard and soft sweep signatures are recovered if measuring the SFS further away from the beneficial allele (Fig 6(c), (d)). For example, a heightened number of intermediate-frequency alleles are observed in a sweep from standing variation (Fig 6(d)). Here too, one has to analyse a recombination distance that is $1/(1 - \sigma)$ times longer than in outcrossers to observe soft-sweep behaviour.

In the Supplementary *Mathematica* file (Section E of S1 File) we plot SFS results for other recombination distances. In particular, these results demonstrate that with higher f_0, the SFS becomes similar to the neutral case over a shorter recombination distance than for hard sweeps, as reflected with results for expected pairwise diversity (Eq 6).

Soft sweeps from recurrent mutation

Until now, we have only focussed on a 'soft' sweep that arises from standing variation. An alternative type of 'soft' sweep is one where recurrent mutation at the selected locus introduces the beneficial allele onto different genetic backgrounds. We can examine this case by modifying existing results. Pennings and Hermisson

[15] demonstrated that the expected reduction in pairwise diversity $\mathbb{E}(\pi/\pi_0) = 1-[(P_{coal,M})(P_{NR})]$ where $P_{coal,M} = 1/(1+2\Theta_b)$ is the probability that two samples are identical by descent instead of arising on different genetic backgrounds by independent mutation events. Here, $\Theta_b = 2N_e\mu_b$ is the population level mutation rate at the beneficial locus. We can compare the signatures of these two different types of soft sweeps by using this solution, with P_{NR} as given by Eq 3 with $f_0 = 1/(2N)$, and $\Theta_b = 2N_e\mu_b = (2N\mu_b)/(1+F)$ in $P_{coal,M}$.

Fig 7(a), (b) compares $\mathbb{E}(\pi/\pi_0)$ in the standing variation case, and for the recurrent mutation case, under different levels of self-fertilisation. Several differences are apparent. First, while dominance only weakly affects sweep signatures arising from standing variation, it more strongly affects sweeps from recurrent mutation in outcrossing populations, as the underlying allele trajectories are affected by the level of dominance since each variant arises from an initial frequency $\sim 1/(2N)$ (Fig 3). Second, both models exhibit different behaviour close to the selected locus ($R \to 0$). The recurrent mutation model has diversity levels that are greater than zero, while the standing variation model exhibits no diversity. As R increases, diversity reaches higher levels in the standing variation case than for the recurrent mutation case. To determine the recombination rate when the recurrent mutation model exhibits higher diversity than the standing variation model, we assume that close to the adaptive mutant, it is very unlikely for samples to recombine during the sweep phase (i.e., $P_{NR} \approx 1$). It remains to determine when $P_{coal,M}$ is higher than P_{coal} under standing variation, which occurs when:

$$\begin{aligned} R \leq R_{Lim} &= \frac{\Theta_b}{f_0(1-f_0)(1-F)} \\ &\approx \frac{\Theta_b}{f_0(1-F)} \text{ for } f_0 \ll 1 \end{aligned} \tag{11}$$

Hence for a fixed Θ_b, the window where recurrent mutations creates higher diversity near the selected locus increases for lower f_0 or higher F, since both these factors reduces the potential for recombination to create new haplotypes during the standing phase. Eq 11 accurately reflects when standing variation sweeps exhibit higher diversity (Fig 7(a), (b)), but becomes inaccurate for $h = 0.1$ in outcrossing populations. Here, beneficial alleles have elevated fixation times, so some recombination is likely to occur during the sweep phase. We also observe that for higher selfing rates, the ratio of π_{SV} (diversity under sweep from standing variation) to π_M (diversity under sweep from recurrent mutation) becomes higher than in outcrossers (compare Fig 7(c) with 7(d)). This is because the effects of sweeps arising from recurrent mutation on diversity becomes diluted over a longer genetic distance under self-fertilisation, due to weakened effects of recombination.

We can also modify the expected SFS to account for recurrent mutation during the standing phase (see Section B in S3 File for details). These calculations verify that, close to the selected locus, sweeps from recurrent mutations show more intermediate-frequency variants than sweeps from standing variation. The situation is reversed once R exceeds R_{Lim}.

Distance between singletons

A selective sweep increases the mean distance between ‘singletons’, which are derived alleles that are only observed on a single haplotype. This phenomena was recently used to detect evidence for recent human adaptation [56]. We hence ran computer simulations to investigate the distribution of distances between the beneficial locus and the nearest singleton under different scenarios.

Singleton distances in fixed sweeps

We first measured the distance from the beneficial allele to the nearest singleton across 50 samples taken from a fixed sweep. These distances are compared to those obtained from the neutral background before a beneficial mutation was introduced (see Fig 11(a) in the Methods for a schematic). Due to the computational limitations of individual-based simulations, a large number of samples did not contain singletons (Fig 8(a)). Focussing on samples containing singletons in the neutral population, they are likelier to lie close to the target locus (Fig 8(b)). Sweeps reduce the overall frequency of observed singletons, and also increases the distance from the selected allele to the nearest singleton. These distributions are visibly different for sweeps of different dominance effects; recessive mutations ($h = 0.1$) cause a much stronger reduction in observed singleton densities than dominant adaptations ($h = 0.9$). This behaviour likely arises as recessive mutations increase in frequency closer to the present time, while dominant mutations reach a higher frequency earlier on (Fig 3). The rapid increase in frequency of recessive mutations in the recent past makes it even less likely for singletons to appear on selected backgrounds. This result is reflected in the SFS, where hard sweeps caused by recessive mutations also display a lower number of singletons (Fig 6(a)).

We showed that in outcrossing populations, a sweep arising from a recessive or dominant mutation can cause the same reduction in diversity as that caused by a co-dominant mutation, after rescaling the selection coefficient (Section A in S3 File). Hence we next measured the distribution of singleton distances for co-dominant sweeps but with different selection coefficients, to determine if similar patterns are produced to cases with different dominance. Weakly-selected mutations ($s = 0.01$) exhibit results that are similar to the neutral case, while strongly-selected mutations ($s = 0.09$) show a clear reduction in singleton densities (Fig 8(c), (d)). These patterns are opposite to what is observed for recessive and dominant mutations respectively, implying that singleton densities may provide clearer evidence regarding the dominance underlying a selective sweep.

Singleton distances in partial sweeps

We next investigated singleton distances from partial sweeps (i.e., those that have not completely fixed in the population). Specifically, we look at sweeps that have reached a frequency of 70% when they were sampled. The neutral expectation was calculated by measuring singleton distances around SNP that lie between a frequency of 65% – 75% (Fig 11(b) in the Methods). For the neutral case, there are always more singletons observed on the derived background, since it is present at a higher frequency (Fig 9(a)). Focussing on samples where a singleton was observed, we then see that the distributions are similar between ancestral and derived backgrounds (Fig 9(b)). On selected backgrounds, there are many more samples not carrying singletons (Fig 9(a)). For samples carrying singletons, fewer of them lie closer to the target locus on derived backgrounds, compared to ancestral backgrounds. Furthermore, singleton distances are uniformly distributed along the genetic tract on derived backgrounds, with visibly similar distances occurring irrespective of the dominance level (Fig 9(b)). Hence while singleton distances can provide evidence of ongoing adaptation, there appears to be very little power to infer the dominance level of the mutation.

In Section C of S3 File, we show that increasing either f_0 or F weakens the effect that h has on singleton distance distributions, in line with previous results showing how an increase in either of these values weakens the effect that dominance has on summary statistics. We also show that increasing the number of samples under investigation (from 50 to 1000) weakens the ability of singleton distributions to detect fixed sweeps as singleton distances will only be affected with very recent (i.e., very strong) selection [56]. However, evidence of an ongoing sweep (i.e., one observed at a frequency of 0.7) can still be seen if taking a large number of samples, as the distributions are markedly different between the ancestral and derived backgrounds.

Application to a selective sweep at the human *SLC24A5* gene

To demonstrate how these sweep models can be used to infer properties of genetic adaptation, we reanalyse a selective sweep at the *SLC24A5* gene in European human populations. The rs1426654 SNP harbours a G → A substitution that is strongly associated with skin pigmentation in Eurasian populations [73, 74]. It was long assumed that the derived A mutation was only present at a negligible frequency in Africa, yet recent data has shown it to be present at an elevated frequency in East Africa [74]. These East African populations harbour the same extended haplotype as in Eurasia, suggesting that the mutation was reintroduced into Africa following the out-of-Africa human expansion. Nevertheless, the recent discovery of these new haplotypes begs the question of whether the derived SNP was introduced into Eurasia at an elevated frequency or not. Hence we performed a maximum-likelihood fit of these analytical solutions to the sweep signature produced around the derived SNP in Europe, to determine whether it is consistent with a hard sweep, or instead one from either standing variation or recurrent mutation.

We downloaded diversity data from European populations in the 1000 Genomes phase 3 release, and fitted models to diversity data around the derived SNP (see Methods and Section G of S1 File for details). We implemented a nested model comparison, to test for the presence of either a sweep from standing variation, or from recurrent mutation. In both cases we also tested for the presence of non-additive dominance ($h \neq 1/2$). Results are outlined in Table 2. For the standing variation case, the best fitting model implicated that the sweep arose from a new mutation (a 'hard sweep') with additive dominance, with a selection coefficient $s = 0.065$ (see Fig 10(a) of a fit of this model to the sweep region). Models that included an elevated initial frequency also estimated unrealistically high selection coefficients, with s nearly equal to a thousand. These findings suggest that large sweep signatures, such as those observed in the *SLC24A5* gene, are extremely unlikely to be formed by adaptations arising from standing variation, in line with theoretical work (see also [21]). It was also not possible to discern a sweep assuming additive dominance from non-additive dominance; analysis of the likelihood surface shows that a ridge of maximum likelihood exists for constant hs, reinforcing the idea that it is not easy to discern non-additive dominance from diversity data alone (Section G of S1 File).

Model	Parameters	s	h (1/2)	x_0 ($1/2N_e$)	Θ (0)	LL	ΔAIC
HS, AD	*1*	*0.065*	–	–	–	*-4982.57*	*846*
HS, NAD	2	0.15	0.18	–	–	-4982.57	848
SV, AD	2	815	–	0.017	–	-4207.29	NA
SV, NAD	3	933	0.82	0.017	–	-4207.29	NA
RM, AD	**2**	**0.20**	–	–	**0.56**	**-4134.14**	**0**
RM, NAD	3	0.26	0.37	–	0.56	-4134.14	2

Table 2. Results of maximum-likelihood model fitting of *SLC24A5* sweep signature. Results are presented for a hard sweep model ('HS'); from standing variation ('SV'), or from recurrent mutation ('RM'). We also consider additive or non-additive dominance (denoted AD, NAD respectively). Numbers in brackets next to each parameter heading are the fixed values if they are not estimated for that particular model (as represented by a dash). ΔAIC is the difference in AIC between that model and the best fitting one (RM, AD, which is highlighted in bold). The italicised model HS, AD is the best fitting realistic model.

For the recurrent mutation model, the best-fitting model included a significant level of mutation at the target SNP ($\Theta = 0.56$). However, this high mutation rate leads to elevated diversity levels around the target SNP, which is not present in observed data (Fig 10(a)). We also tested for the presence of recurrent mutation by measuring H-statistics around the sweep region [25] (see Methods for formal definitions of these statistics), which measure the relative frequency of different haplotypes across samples. Specifically, high H_{12}, low H_2/H_1 values are consistent with a single haplotype fixing, in line with a hard sweep. Conversely, a reduced H_{12} and elevated H_2/H_1 values suggest multiple haplotypes fixing, which occurs following adaptation from standing variation or recurrent mutation. Fig 10(b) demonstrates that around the target SNP, H_1 is close to 1 while H_2/H_1 is near zero. Both results indicate that a single haplotype has fixed around the target SNP, which is not expected following a sweep from recurrent mutation [15]. It seems that the recurrent mutation model had the highest likelihood due to spikes of high diversity around the target SNP, which can be mistaken for a recurrent mutation effect if not checked against other analyses.

These models assume a fixed population size, but it is known that humans have a complex demographic history. European populations have likely undergone a contraction following migration from Africa, followed by extensive population growth [75]. To determine if this demography could have drastically affected our inference of different sweep signatures, we ran simulations using MSMS with inferred parameters to determine how sweep signatures are affected by this demographic history. Yet even under a growth-bottleneck model, a hard sweep model fits the observed sweep pattern better than either of the soft sweep models, after rescaling parameters by the different present-day N_e (Section D in S3 File, with plots also available in Section G of S1 File).

Furthermore, the derived A allele is present in African populations but at a low frequency (in 55 of 1063 African haplotypes in the 1000 Genomes dataset). This begs the question of whether the derived allele was introduced into Eurasia, but at too low a frequency to influence the maximum-likelihood model fit. Fig 10(c) shows phylogenetic trees of 20Kb regions either surrounding the target SNP, or upstream, downstream of the SNP. We observe that most European samples carrying the derived mutation cluster together, reflecting recent appearance and spread of the derived allele. However, within these clades we also observe some African haplotypes carrying the derived allele, suggesting that it was introduced into Eurasia due to out-of-Africa migration.

Overall, our model analyses determined that the derived SNP at the *SLC24A5* gene most likely followed 'hard' sweep dynamics. However, we also find evidence for ancestral African haplotypes forming the basis of the sweep. Hence the likeliest outcome is that the derived allele was introduced into Eurasia at a sufficiently low frequency so that its sweep dynamics were indistinguishable from a hard sweep. Given a selection coefficient of 0.065, co- dominance ($h = 0.5$) and $N_e = 10,000$, Eq 5 predicts an $f_{0,A}$ of 0.7%. It is likely that the derived haplotype was introduced at a lower frequency than this value.

Discussion

Summary of Theoretical Findings

While there has been many investigations into how different types of adaptation can be detected from next-generation sequence data [11, 13, 76, 77], these models assumed idealised sexually reproducing populations and beneficial mutations that have additive dominance ($h = 0.5$). Here we have created a general model of a selective sweep, with arbitrary levels of self-fertilisation and dominance. Our principal focus is on comparing a 'hard sweep' arising from a single allele copy to a 'soft sweep' arising from standing variation, but we have also considered the effect of adaptation from recurrent mutation (Fig 7).

We find that the qualitative patterns of different selective sweeps under selfing remain similar to expectations from classic outcrossing models. In particular, a sweep from standing variation still creates an elevated number of intermediate-frequency variants compared to a sweep from *de novo* mutation (Figs 6, 7). This pattern is a known signature of a 'soft sweep' [11,13,15,21], meaning that common statistical methods used for detecting them (e.g., observing an higher number of haplotypes than expected [24, 25]) can, in principle, still be applied to selfing organisms (but see the discussion below with regards to dominance). Under self-fertilisation, these signatures are stretched over longer physical regions than in outcrossers. These extensions arise as self-fertilisation affects gene genealogies during both the sweep and standing phases, but in different ways. During the sweep phase, beneficial alleles fix more rapidly under higher self-fertilisation as homozygote mutations are created more quickly [41,42]. In addition, the effective recombination rate is reduced by approximately $1 - F$ [43]. These two effects mean that neutral variants linked to an adaptive allele are less likely to recombine onto the neutral background during the sweep phase, as reflected in Eq 1 for P_{NR}. During the standing phase, two samples are more likely to coalesce with increased self-fertilisation since N_e is decreased by a factor $1/(1 + F)$ [59]. This effect, combined with an reduced effective recombination rate, means that the overall probability of recombination during the standing phase is reduced by a factor $1 - \sigma$ (Eqs 4, 9, B14 in S3 File). Hence intermediate-frequency variants, which could provide evidence of adaptation from standing variation, will be spread out over longer genomic regions. The elongation of sweep signatures means soft sweeps can be easier to detect in selfing organisms than in outcrossers, since the differences in diversity caused by sweeps are spread out over longer regions.

We have also investigated how dominance affects soft sweep signatures, since previous analyses have only focussed on how hard sweeps are affected with different dominance effects [33–35]. In outcrossing organisms, recessive mutations leave weaker sweep signatures than additive or dominant mutations as they spend more time at low frequencies, increasing the amount of recombination that restores neutral variation (Figs 2, 3). With increased self-fertilisation, dominance has less of an impact on sweep signatures as most mutations are homozygotes (Fig 3). However, dominance has different effects on separate types of 'soft' sweeps. Dominance only weakly affects sweeps from standing variation, as trajectories of beneficial alleles become similar once the variant's initial frequency greatly exceeds $1/2N$ (Figs 2, 3). Yet different dominance levels can affect sweep signatures if the beneficial allele is reintroduced from recurrent mutation (Fig 7). Hence if one wishes to understand how dominance affects selective sweep signatures, it is also important to consider the type of selective sweep underlying observed genetic diversity. We also showed how beneficial variants of different dominance values create distinct alterations in the distances to the nearest singleton (Fig 8). These results suggest that the distribution of low-frequency variants around a sweep can provide information on the dominance value underlying it. Investigating the utility of singletons to detect dominance effects seems a worthy future research direction, especially since in our example of estimating properties of the *SLC24A5* sweep, it is tricky to infer non-additive dominance from diversity data alone.

We also derived an 'effective selection coefficient' $\tilde{s}$ so that sweeps from standing variation will produce a pattern of diversity reduction equivalent to a hard sweep (Eq 7; Fig 4), and an $\tilde{s}_h$ so that a non-additive sweep in an outcrossing population can be mapped onto a co-dominant sweep (Section A in S3 File). These derivations imply that different types of sweep models can lead to similar outcomes, which may prove problematic when making inferences from genomic data [78, Supplementary Material]. Yet it may be apparent if some sweep signatures arise from standing variation or not, if unrealistic parameters are needed to produce expected patterns of diversity. In particular, for the *SLC24A5* sweep to appear from standing variation, the underlying selection coefficient must be unrealistically large (Table 2). Hence adaptation from elevated standing variation (greater than 0.7%) is unlikely for this case.

Soft sweeps from recurrent mutation or standing variation?

Our theoretical results shed light onto how to distinguish between soft sweeps that arise from either standing variation, or from recurrent mutation. Both models are characterised by an elevated number of intermediate-frequency haplotypes, in comparison to a hard sweep. Yet sweeps arising from recurrent mutation produces intermediate-frequency haplotypes closer to the beneficial locus, while sweeps from standing variation produce intermediate-frequency haplotypes further away from the adaptive locus (Fig 7 and Section B in S3 File). Eq 11 provides a simple condition for the recombination distance needed so a sweep from standing variation exhibits higher diversity than one from recurrent mutation. The size of this region increases under higher self-fertilisation.

This result has implications for inferring different types of sweeps. If multiple swept haplotypes are present over long genetic distances, this observation implies that the beneficial allele underlying the sweep likely originated from standing variation as opposed to recurrent mutation. This phenomenon could explain the elevated H_2/H_1 statistics, and reduced H_{12} values upstream of the *SLC24A5* SNP (Fig 10(b)), especially given that we know the derived SNP to be present at a low frequency in Africa. However, if this was truly a selective sweep arising from an elevated starting frequency, we also expect elevated H_2/H_1 values downstream of the SNP, which we do not observe. A simpler explanation for the elevated haplotype diversity is that the recombination rate is higher upstream of the SNP than downstream, which has broken down the sweep signature to a greater extent in this region (see Fig 12 in the Methods for the actual recombination map).

Different haplotype structure between sweeps from either standing variation or recurrent mutation should be more pronounced in self-fertilising organisms, due to the reduction in effective recombination rates. However, if investigating sweep patterns over longer genetic regions, it becomes likelier that genetic diversity will be affected by multiple beneficial mutations spreading throughout the genome. Competing selective sweeps can lead to elevated diversity near a target locus for two reasons. First, selection interference increases the fixation time of individual mutations, allowing more recombination that can restore neutral diversity [79]. In addition, competing selective sweeps can drag different sets of neutral variation to fixation, creating asymmetric reductions in diversity [80]. Further investigations of selective sweep patters across long genetic distances will prove to be a rich area of future research.

Using models to determine properties of selective sweeps

Analysis of the *SLC24A5* sweep signature

An emerging approach to quantifying properties of genetic adaptation involve fitting sweep models to regions displaying high substitution rates compared to an outgroup [78, 81, 82]. Inspired by these works, we demonstrated how the general sweep models can be used to determine adaptation properties by applying them to the *SLC24A5* gene in European humans. Overall, the sweep pattern best matches a classic 'hard' sweep signature (Table 2; Fig 10). However, since the derived allele is known to be present at a low frequency in Africa, it also appears that the derived allele was introduced into Eurasia at a sufficiently low frequency so that the resulting signature is equivalent to a 'hard' sweep, even if the mutation did not appear after out-of-Africa migration (Fig 10(c)). This analyses demonstrates how adaptive mutations arising from standing variation have to be present at a sufficiently high frequency (above the 'accelerated' $f_{0,A}$ given by Eq 5) to be reliably distinguished from classic hard sweeps. In addition, analysis of this specific sweep region also demonstrates the utility of combining model fitting of genetic diversity with other statistics (e.g., haplotype structure, phylogenetic relationships) to fully work out the evolutionary history of individual selective sweeps.

One potential difficulty arising out of model analysis is that of estimating dominance coefficients. Sweep models where h was non-additive did not explain the data better than a co-dominant sweep. Nevertheless, there are several ad-hoc reasons why the underlying mutation is likely to be approximately co-dominant. Recessive hard sweeps appear similar to sweeps from standing variation (with a weaker reduction in diversity at linked regions) and are heterozygous for long periods of time (Fig 3(a)). Hence the strong sweep signature, and high frequency of the derived allele in European populations, makes it unlikely for this mutation to be recessive. Similarly, strongly dominant mutations take a long period of time to fully fix, in contrast to the observed near-fixation of the derived *SLC24A5* SNP. It will be important to extent inference methods to more accurately quantify dominance of adaptive mutations. One promising approach could be to analyse singleton densities, which appear to differ under recessive and dominant sweeps (Fig 8).

Potential model applications to self-fertilising organisms

Existing software for finding sweep signatures in nucleotide data are commonly based on finding regions with a site-frequency spectrum matching what is expected under a selective sweep [83, 84]. The more general models developed here can therefore be used to create more specific sweep-detection methods while accounting for self-fertilisation. However, a recent analysis found that signatures of soft sweeps can be incorrectly inferred if analysing genetic regions that flank hard sweeps, which was named the 'soft shoulder' effect [85]. Due to the reduction in recombination in selfers, these model results indicate that 'soft-shoulder' footprints could be present over long genetic distances, and should be accounted for. One remedy to this problem is to not just classify genetic regions as being subject to either a 'hard' or 'soft' sweep, but also as being linked to a region subject to one of these sweeps [27].

Further investigations of selective sweeps under self-fertilisation will also be aided by the creation of new simulation methods that account for this mating system. It is common to test sweep models by comparing results to coalescent simulations of adaptation [86,87], but existing simulations do not account for self-fertilisation. Creating new simulation programs will prove important to further explore other key properties of selective sweeps (e.g., haplotype structure, singleton densities, power calculations) under selfing across larger sample and population sizes. We therefore hope that these results will stimulate the creation of new simulation and inference software to further explore how adaptation is affected by different reproductive modes.

Methods

Exact simulations, including dominance and self-fertilisation

Simulations were coded in C and are based on Wright-Fisher population dynamics. These are available in S4 File or online at `https://github.com/MattHartfield/DomSelfAdapt`. There exists N diploid individuals, each containing two haplotypes consisting of a stretch of genetic material at which neutral mutations can accumulate via an infinite-sites model. The far left hand side of the tract contains the locus at which the beneficial allele can arise.

Each generation the entire population is replaced. First, the number of self-fertilisation reproductions is drawn from a Binomial distribution with probability σ. It is then decided which specific reproduction events will occur by selfing. To create offspring, a first parent is chosen with probability proportional to its fitness, then one of its two haplotypes is selected with equal probability. If selfing arises, then the offspring's second haplotype is chosen from the same parent, which could be the same as the first. Otherwise a second parent is selected, with probability proportional to its fitness, then one of its haplotypes is chosen. The number of recombination events per haplotype is drawn from a Poisson distribution with mean r. Crossover locations are uniformly distributed over the fragment length. Offspring haplotypes are subsequently created by initially copying over the first sampled parental haplotype, then switching over to copying the second parental haplotype after passing a recombination breakpoint. Selection and recombination is repeated in this manner for all N individuals.

New neutral polymorphisms are then added. The number of mutations to be added to the entire population is chosen from a Poisson distribution with mean

$2N\mu$. For each new mutation, it can appear in one of the $2N$ haplotypes with equal probability, with its location selected from a uniform distribution. A 'garbage-collection' routine is then executed to remove non-polymorphic loci. Fig 11(a) outlines how polymorphisms are distributed in the simulation.

The simulation is split into two parts. A 'burn-in' phase is run first to generate background neutral diversity, where the population evolves without any beneficial alleles present for $20N$ generations. 100 different populations are created for each neutral parameter set. In the second part, the adaptive mutation is introduced into a single haplotype chosen at random; it is initially neutral until its frequency matches or exceeds f_0, at which point it has selective advantage s and dominance coefficient h acting upon it. We can set $f_0 = 1/2N$ so that the mutation is beneficial from the outset (a 'hard' sweep). The beneficial allele is then tracked until it is either lost, or reaches the 'census' frequency at which the selective sweep is analysed, after which we randomly sample haplotypes from the population to create final outputs.

Measuring mean pairwise diversity; number of segregating sites; site frequency spectrum

After the beneficial allele has gone to fixation, we sampled 10 haplotypes 10 times from each burn-in population to create 1000 simulation estimates. For each of these statistics, mutations are placed in one of 10 bins depending of the distance from the sweep, with the relevant statistic calculated per bin. Mean values, along with 95% confidence intervals, are calculated over all 1000 outputs.

Measuring distances between singleton mutations

We sampled 50 or 1000 haplotypes once from each base population, creating 100 total datasets. We also sample the same number of haplotypes from the burn-in population to determine the neutral distribution of distances.

We investigated cases where the sweep has either gone to fixation, or where the population is sampled after the beneficial allele exceeds a frequency of 0.7. When the beneficial allele is sampled at fixation, the distance from the adaptive locus to the nearest singleton is measured over all samples. The distance is normalised to between 0 and 1, where 0 is the location of the selected locus and 1 the furthest right-hand edge. We also note how many samples did not contain singletons. When the sweep is sampled at a frequency of 0.7, we measure singleton distances separately for samples carrying either the ancestral or derived allele. For the neutral burn-in population, we first found derived alleles that were present at a frequency between 0.65 and 0.75. For each of these, we measured the upstream distance to the nearest singleton, if present. If not, we check if a singleton existed downstream of the reference variant, and the singleton distance is calculated as the distance of the nearest singleton from the left-hand edge of the genome, plus the upstream distance from the reference variant to the right-hand edge (Fig 11(b)). Summing distances in this manner is valid as we assume polymorphisms are uniformly distributed throughout the genome. Otherwise we noted if no singleton existed on the haplotype.

Human sweep data analyses

Data processing

Data was retrieved from the 1000 Genomes phase 3 version 3 integrated call set (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) [88]. The five European populations (CEU, FIN, GBR, IBN, TSI) were investigated; related individuals were removed (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/20140625_related_individuals.txt`) giving 503 total individuals. SNP data was obtained using *VCFtools* [89] over a 1Mb region, between locations 47,930,001 and 48,930,000 on Chromosome 15 (the rs1426654 target SNP is at location 48,426,484). Only biallelic SNPs in Hardy-Weinburg equilibrium (with $P-$value greater than 10^{-6}) were retained; indels were removed. Pairwise diversity was calculated in 20Kb bins over this region. Baseline diversity (i.e., that expected in the absence of a selective sweep) was determined by calculating mean diversity values at flanking regions both up- and downstream of the sweep. Specifically, we measure the mean diversity between locations 47,930,001 and 48,220,000 upstream of the target SNP, and between locations 48,640,000 and 48,930,000 downstream of the target SNP (Fig 12(a)). Diversity estimates up- and downstream were divided by the mean values between these regions (Fig 12(b)). Sex-averaged recombination maps for each bin were obtained from Bhérer *et al.* [90] (Fig 12(c)).

Model fitting

Sweep models were fitted to this diversity data using the maximum likelihood procedure of Sattath *et al.* [81]. Two nested models were considered; one where a sweep arose from standing variation (Equation 6), or where the sweep arose from recurrent mutation (as described in the 'Soft sweeps from recurrent mutation' section). Since we are analysing human data assuming a fixed population size, we set $F = 0$ and $N_e = 10,000$ [91]. Due to the large number of polymorphisms per bin, we assume that observed pairwise diversity at recombination distance r is normally distributed with mean values equal to the expected values given by the models (denoted $m(r)$), and variance $v(r) = m(r)(1-m(r))/n$ for n the number of segregating sites in that bin. The log-likelihood for the data under these models, as measured over all b bins, equals $-\sum_b(\log(2\pi v(r))/2 + (\hat{K}(r) - m(r))^2)/(2v(r)),$ where $\hat{K}(r)$ is the relative diversity in each bin.

Maximum likelihood for each model was found using the 'FindMaximum' function in *Mathematica* version 11.2 [92]. In all models we estimated the selection coefficient s. We then used a nested model structure to determine if evidence existed for non-additive dominance ($h \neq 1/2$); standing variation of the selective sweep ($f_0 > 1/2N_e$); or recurrent mutation at the target SNP location ($\Theta \neq 0$). We set options in 'FindMaximum' so that $s > 0$, and $0 < h < 1$, $f_0 > 1/2N_e$ and $\Theta > 0$ if these parameters were not fixed. We compared six models: (i) fixed $h = 1/2$, $f_0 = 1/2N_e$ (hard sweep with additive dominance); (ii) variable h, fixed $f_0 = 1/2N_e$ (hard sweep with non-additive dominance); (iii) fixed $h = 1/2$, variable f_0 (standing variation sweep with additive dominance); (iv) variable h, f_0 (standing variation sweep with non-additive dominance); (v) fixed $h = 1/2$, variable Θ (recurrent mutation with additive dominance) (iv) variable h, Θ (recurrent mutation with non-additive dominance). Note that for hard sweep models, we do not use $f_{0,A}$ to ensure a tractable model fit. Using $f_0 = 1/2N_e$ should not prove problematic for inferring different types of sweeps, as long as estimated f_0 for the standing variation cases lie above $f_{0,A}$, so the two cases can be differentiated. Since estimated $f_0 \sim 1.7\%$ and $f_{0,A} \sim 0.7\%$, this condition is fulfulled.

To calculate the H statistics of Garud *et al.* [25], haplotype counts in each of the 20Kb bins were obtained using the '--hapcount' function in *VCFtools*. From these the relevant haplotype statistics were calculated per bin. Let there be K unique haplotypes in a bin, ordered so that p_1 is the frequency of the most common haplotype, p_2 the frequency of the second common haplotype, and so on. Then $H_1 = \sum_i^K (p_i^2)$, $H_{12} = (p_1 + p_2)^2 + \sum_{i=3}^K (p_i)^2$, and $H_2 = H_1 - p_1^2$. We also calculated the ratio H_2/H_1.

Human Sweep Simulations

We ran simulations of the selective sweep using MSMS [87] to determine expected diversity patterns under different demographic scenarios. To ensure tractable simulations, we simulated 100 haplotypes using a genetic region of length 200Kb, with the selected site located in the middle of the region. The scaled neutral mutation rate $4N_e\mu$ equalled 188.8 (assuming $N_e = 10,000$), reflecting a per-basepair rate of 2.36×10^{-8} as recently used by Field *et al.* [56]; the scaled recombination rate $2N_er$ was set to 55.4 reflecting the sex-averaged recombination rate over the region as determined by Bhérer *et al.* [90]. Three sweep scenarios were simulated: (i) a hard sweep (ii) a sweep from standing variation with initial selected frequency 1.7% (iii) a sweep from recurrent mutation with $\Theta = 2N_e\mu_b = 0.56$. Input values reflect those obtained from the maximum likelihood model fitting. Simulations were run assuming two demographic scenarios; either a constant population of size $N_e = 10,000$, or a growth-bottleneck demographic mimicking human migration out of Africa (parameters used are outlined in Fig 1(d) of Schrider *et al.* [93]). For the latter model, other parameters were scaled by the present-day $N_e = 35,900$. In both the growth-bottleneck models and constant-sized models assuming a 1.7% starting frequency, MSMS requires the user to set a time in the past when selection started acting on the beneficial mutation. In these cases, starting times were set so that the sweep reached fixation in the present day. We also simulated pairwise diversity from a neutral growth-bottleneck demographic scenario, to determine expected baseline diversity in the absence of a selective sweep. All results are averages over 1,000 simulation runs. A complete list of command lines and parameters are outlined in S5 Table.

Phylogenetic analyses

Biallelic SNPs in Hardy-Weinburg equilibrium ($P > 10^{-6}$) were extracted from the five European populations and the five African populations (ESN, GWD, LWK, MSL, YRI) in the 1000 Genomes dataset, in bins of size 20Kb, from between basepair locations 48,320,000–48,340,000, 48,420,000–48,440,000, and 48,500,000–48,520,000 on chromosome 15. Distance matrices were then created for all pairwise comparisons of individuals, where the distance between two individuals is defined as the sum of all differences over all segregating sites (e.g., a heterozygote-homozygote difference at a SNP adds 1 to the distance; a derived homozygote-ancestral homozygote difference adds 2). Phylogenetic trees were created from these matrices by neighbour-joining, using the 'nj' function in the 'ape' package for R [94, 95].

Supporting information

S1 File. Supplementary *Mathematica* File. *Mathematica* notebook of algebraic derivations and simulation comparisons (.nb format).

S2 File. Supplementary *Mathematica* File (PDF). *Mathematica* notebook of algebraic derivations and simulation comparisons (.pdf format).

S3 File. Supplementary Text File. Additional results and figures pertaining to effective reduction in diversity under different scenarios; deriving the site-frequency spectrum; further results on singleton distributions; and simulation results of *SLC24A5* sweep region.

S4 File. Simulation Code. Forward-in-time simulation code written in C. Also available from `https://github.com/MattHartfield/DomSelfAdapt`.

S5 Table. Simulation Command Lines. List of MSMS command lines used to simulate a sweep at the *SLC24A5* region under different scenarios.

Acknowledgments

We would like to thank Sally Otto for providing information on the elevated effective starting frequency of beneficial mutations; Dan Schrider for providing information on how to simulate demographic scenarios; and Mikkel Schierup for feedback on the human sweep analyses.

References

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

2. Kaplan NL, Hudson RR, Langley CH. The "Hitchhiking Effect" Revisited. Genetics. 1989;123(4):887–899.

3. Thomson G. The effect of a selected locus on linked neutral loci. Genetics. 1977;85(4):753–788.

4. Innan H, Nordborg M. The Extent of Linkage Disequilibrium and Haplotype Sharing Around a Polymorphic Site. Genetics. 2003;165(1):437.

5. McVean GAT. The Structure of Linkage Disequilibrium Around a Selective Sweep. Genetics. 2007;175(3):1395–1406.

6. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832–837.

7. Kim Y, Nielsen R. Linkage Disequilibrium as a Signature of Selective Sweeps. Genetics. 2004;167(3):1513–1524.

8. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biol. 2006;4(3):e72.

9. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Mol Biol Evol. 2014;31(5):1275–1291.

10. Vatsiou AI, Bazin E, Gaggiotti OE. Detection of selective sweeps in structured populations: a comparison of recent methods. Mol Ecol. 2016;25(1):89–103.

11. Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017;8(6):700–716.

12. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008;23(1):38–44.

13. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28(11):659–669.

14. Pennings PS, Hermisson J. Soft Sweeps II – Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol Biol Evol. 2006;23(5):1076–1084.

15. Pennings PS, Hermisson J. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genet. 2006;2(12):e186.

16. Orr HA, Betancourt AJ. Haldane's Sieve and Adaptation From the Standing Genetic Variation. Genetics. 2001;157(2):875–884.

17. Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA. 2004;101(29):10667–10672.

18. Przeworski M, Coop G, Wall JD. The Signature of Positive Selection on Standing Genetic Variation. Evolution. 2005;59(11):2312–2323.

19. Hermisson J, Pennings PS. Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics. 2005;169(4):2335–2352.

20. Wilson BA, Petrov DA, Messer PW. Soft Selective Sweeps in Complex Demographic Scenarios. Genetics. 2014;198(2):669–684.

21. Berg JJ, Coop G. A Coalescent Model for a Sweep of a Unique Standing Variant. Genetics. 2015;201(2):707–725.

22. Wilson BA, Pennings PS, Petrov DA. Soft Selective Sweeps in Evolutionary Rescue. Genetics. 2017;205(4):1573–1586.

23. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between Selective Sweeps from Standing Variation and from a *De Novo* Mutation. PLoS Genet. 2012;8(10):e1003011.

24. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. Annu Rev Genet. 2013;47(1):97–120.

25. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. PLoS Genet. 2015;11(2):e1005004.

26. Garud NR, Petrov DA. Elevated Linkage Disequilibrium and Signatures of Soft Sweeps Are Common in *Drosophila melanogaster*. Genetics. 2016;203(2):863–880.

27. Schrider DR, Kern AD. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. 2016;12(3):e1005928.

28. Sheehan S, Song YS. Deep Learning for Population Genetic Inference. PLoS Comput Biol. 2016;12(3):e1004845.

29. Schrider DR, Kern AD. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol Biol Evol. 2017;34(8):1863–1877.

30. Fustier MA, Brandenburg JT, Boitard S, Lapeyronnie J, Eguiarte LE, Vigouroux Y, et al. Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. Mol Ecol. 2017;26(10):2738–2756.

31. Anderson TJC, Nair S, McDew-White M, Cheeseman IH, Nkhoma S, Bilgic F, et al. Population Parameters Underlying an Ongoing Soft Sweep in Southeast Asian Malaria Parasites. Mol Biol Evol. 2016;34(1):131–144.

32. Jensen JD. On the unfounded enthusiasm for soft selective sweeps. Nat Commun. 2014;5.

33. Teshima KM, Przeworski M. Directional Positive Selection on an Allele of Arbitrary Dominance. Genetics. 2006;172(1):713–718.

34. Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome Res. 2006;16(6):702–712.

35. Ewing G, Hermisson J, Pfaffelhuber P, Rudolf J. Selective sweeps for recessive alleles and for other modes of dominance. J Math Bio. 2011;63(3):399–431.

36. Hartfield M, Bataillon T, Glémin S. The Evolutionary Interplay between Adaptation and Self-Fertilization. Trends Genet. 2017;33(6):420–431.

37. Igic B, Kohn JR. The distribution of plant mating systems: study bias against obligately outcrossing species. Evolution. 2006;60(5):1098–1103.

38. Jarne P, Auld JR. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution. 2006;60(9):1816–1824.

39. Billiard S, López-Villavicencio M, Devier B, Hood ME, Fairhead C, Giraud T. Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. Biol Rev Camb Philos Soc. 2011;86(2):421–442.

40. Haldane JBS. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. Math Proc Cambridge Philos Soc. 1927;23(7):838–844.

41. Charlesworth B. Evolutionary Rates in Partially Self-Fertilizing Species. Am Nat. 1992;140(1):126–148.

42. Glémin S. Extinction and fixation times with dominance and inbreeding. Theor Popul Biol. 2012;81(4):310–316.

43. Nordborg M. Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. Genetics. 2000;154(2):923–929.

44. Hartfield M, Glémin S. Hitchhiking of Deleterious Alleles and the Cost of Adaptation in Partially Selfing Species. Genetics. 2014;196(1):281–293.

45. Hartfield M, Glémin S. Limits to Adaptation in Partially Selfing Species. Genetics. 2016;203(2):959–974.

46. Glémin S, Ronfort J. Adaptation and Maladaptation in Selfing and Outcrossing Species: New Mutations Versus Standing Variation. Evolution. 2013;67(1):225–240.

47. Uecker H. Evolutionary rescue in randomly mating, selfing, and clonal populations. Evolution. 2017;71(4):845–858.

48. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 2013;45(8):884–890.

49. Huber CD, Nordborg M, Hermisson J, Hellmann I. Keeping It Local: Evidence for Positive Selection in Swedish *Arabidopsis thaliana*. Mol Biol Evol. 2014;31(11):3026–3039.

50. Fulgione A, Koornneef M, Roux F, Hermisson J, Hancock AM. Madeiran *Arabidopsis thaliana* Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. Mol Biol Evol. 2018;35(3):564–574.

51. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, et al. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat Genet. 2012;44(3):285–290.

52. Bonhomme M, Boitard S, San Clemente H, Dumas B, Young N, Jacquet C. Genomic Signature of Selective Sweeps Illuminates Adaptation of Medicago truncatula to Root-Associated Microorganisms. Mol Biol Evol. 2015;32(8):2097–2110.

53. Badouin H, Gladieux P, Gouzy J, Siguenza S, Aguileta G, Snirc A, et al. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. Mol Ecol. 2017;26(7):2041–2062.

54. Hedrick PW. Hitchhiking: A Comparison of Linkage and Partial Selection. Genetics. 1980;94(3):791–808.

55. Schoen DJ, Morgan MT, Bataillon T. How Does Self-Pollination Evolve? Inferences from Floral Ecology and Molecular Genetic Variation. Philos Trans R Soc Lond B Biol Sci. 1996;351(1345):1281–1290.

56. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760–764.

57. Wright S. The genetical structure of populations. Ann Eugen. 1951;15:323–354.

58. Caballero A, Hill WG. Effects of Partial Inbreeding on Fixation Rates and Variation of Mutant Genes. Genetics. 1992;131(2):493–507.

59. Nordborg M, Donnelly P. The Coalescent Process With Selfing. Genetics. 1997;146(3):1185–1195.

60. Roze D. Diploidy, Population Structure, and the Evolution of Recombination. Am Nat. 2009;174(S1):S79–S94.

61. Roze D. Background Selection in Partially Selfing Populations. Genetics. 2016;203(2):937–957.

62. Barton NH. Genetic Hitchhiking. Philos Trans R Soc Lond B Biol Sci. 2000;355:1553–1562.

63. Stephan W, Wiehe THE, Lenz MW. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol. 1992;41:237–254.

64. Wakeley J. Coalescent theory: an introduction. vol. 1. Greenwood Village, Colorado: Roberts & Company Publishers; 2009.

65. Barton NH. The effect of hitch-hiking on neutral genealogies. Genet Res. 1998;72:123–133.

66. Desai MM, Fisher DS. Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection. Genetics. 2007;176(3):1759–1798.

67. Martin G, Lambert A. A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations. Theor Popul Biol. 2015;101:40–46.

68. Abramowitz M, Stegun IA. Handbook of Mathematical Functions. New York: Dover Publications, Inc.; 1970.

69. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7(2):256–276.

70. Hudson RR. Gene Genealogies and the Coalescent Process. In: Futuyma DJ, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. vol. 7. Oxford Univ. Press, Oxford; 1990. p. 1–42.

71. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995;140(2):783–796.

72. Kim Y, Stephan W. Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. Genetics. 2002;160(2):765–777.

73. Lamason RL, Mohideen MAPK, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. Science. 2005;310(5755):1782–1786.

74. Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. Science. 2017;358(6365).

75. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet. 2009;5(10):e1000695.

76. Pritchard JK, Di Rienzo A. Adaptation - not by sweeps alone. Nat Rev Genet. 2010;11(10):665–667.

77. Stephan W. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol Ecol. 2016;25(1):79–88.

78. Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A Genomic Map of the Effects of Linked Selection in Drosophila. PLoS Genet. 2016;12(8):e1006130.

79. Kim Y, Stephan W. Selective Sweeps in the Presence of Interference Among Partially Linked Loci. Genetics. 2003;164(1):389–398.

80. Chevin LM, Billiard S, Hospital F. Hitchhiking Both Ways: Effect of Two Interfering Selective Sweeps on Linked Neutral Variation. Genetics. 2008;180(1):301–316.

81. Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in *Drosophila simulans*. PLoS Genet. 2011;7(2):e1001302.

82. Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, et al. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. PLoS Genet. 2013;9(12):e1003995.

83. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15(11):1566–1575.

84. Boitard S, Schlötterer C, Futschik A. Detecting Selective Sweeps: A New Approach Based on Hidden Markov Models. Genetics. 2009;181(4):1567–1578.

85. Schrider DR, Mendes FK, Hahn MW, Kern AD. Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. Genetics. 2015;200(1):267–284.

86. Spencer CCA, Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics. 2004;20(18):3673–3675.

87. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010;26(16):2064–2065.

88. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

89. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–2158.

90. Bhérer C, Campbell CL, Auton A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat Commun. 2017;8:14994.

91. Jorde LB, Bamshad M, Rogers AR. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays. 1998;20(2):126–136.

92. Wolfram Research, Inc . Mathematica Edition: Version 11.2. Champaign, Illinois: Wolfram Research, Inc.; 2017.

93. Schrider DR, Shanku AG, Kern AD. Effects of Linked Selective Sweeps on Demographic Inference and Model Selection. Genetics. 2016;204(3):1207–1223.

94. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–290.

95. R Development Core Team. R: A Language and Environment for Statistical Computing; 2014. Available from: http://www.R-project.org.

Figures

Fig 1. A schematic of the model. The history of the derived variant is separated into two phases. The 'standing phase' (shown in light gray), is when the derived variant is segregating at a frequency f_0 for a long period of time. The 'sweep phase' (shown in dark gray) is when the variant becomes selected for and starts increasing in frequency. Dots on the right-hand side represent samples of the derived haplotype taken in the present day, with lines representing their genetic histories. Samples can recombine onto the ancestral background either during the sweep phase or the standing phase. Solid lines represent coalescent histories on the derived genetic background; dotted lines represent coalescent histories on the ancestral background.

Fig 2. Expected pairwise diversity following a selective sweep. Plots of $\mathbb{E}(\pi/\pi_0)$ as a function of the recombination rate scaled to population size $2Nr$. Lines are analytical solutions (Eq 6), points are simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$ (note μ is scaled by N in simulations, not N_e), and dominance coefficient $h = 0.1$ (red lines, points), 0.5 (black lines, points), or 0.9 (blue lines, points). Rate of self-fertilisation equals 0; 0.5; or 0.95 (note the $x-$axis range changes with the self-fertilisation rate). The sweep arose from either a single *de novo* mutation (actual $f_0 = 1/2N$; note we use $f_{0,A}$ in our model, as given by Eq 5), standing variation with $f_0 = 0.02$; or $f_0 = 0.05$. Further results are plotted in Section B of S1 File.

Fig 3. Beneficial allele trajectories. These were obtained by numerically evaluating the negative of Eq 2 forward in time. $N = 5,000$, $s = 0.05$, and h equals either 0.1 (red lines), 0.5 (black lines), or 0.9 (blue lines). Values of f_0 and self-fertilisation rates used are shown at the end of the relevant row and column. Note the different $x-$axis scales used in each panel. Further results are plotted in Section B of S1 File.

Fig 4. Effective reductions in diversity under different scenarios. (a) $\tilde{s}$ (Eq 7) as scaled to s, as a function of $R = 2Nr$. $f_0 = 1/2N$ (black line), 0.02 (red line) or 0.1 (blue line). (b) Plot of π/π_0 (Eq 6) using $\tilde{s}$. Solid lines represent $f_0 = 1/2N$ (black line), 0.02 (red line) or 0.1 (blue line). Points are Eq 6 assuming $f_0 = 1/2N$, but using $\tilde{s}$ (Eq 7) evaluated for $f_0 = 0.02$ (circles) or 0.1 (squares). The population is outcrossing; similar results exist for partial selfing ($\sigma = 0.5$) if measuring over a longer recombination distance (Section A of S2 File). Other parameters are $N = 5,000$, $s = 0.05$, $h = 0.5$.

Fig 5. Expected number of segregating sites following a selective sweep. A plot of $\mathbb{E}(S)$, as a function of the recombination rate scaled to population size $2Nr$. Lines are analytical solutions (Eq 10 multiplied by θ), points are simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$ (so $\theta = 4N_e\mu$ per bin is 4 for $\sigma = 0$, 3 for $\sigma = 0.5$, and 2.1 for $\sigma = 0.95$), and dominance coefficient $h = 0.1$ (red lines, points), 0.5 (black lines, points), or 0.9 (blue lines, points). Rate of self-fertilisation σ equals 0, 0.5, or 0.95 as denoted on the right-hand side; note the different x-axes ranges. The sweep arose from either a single *de novo* mutation or standing variation with $f_0 = 0.05$, as denoted at the top of the figure. Further results are plotted in Section D of S1 File.

Fig 6. Expected site frequency spectrum, in flanking regions to the adaptive mutation, following a selective sweep. Lines are analytical solutions (Eq B14 in the Supplementary Material), points are simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$ (so the effective mutation rate per bin is 4 for $\sigma = 0$ and 3 for $\sigma = 0.5$), and dominance coefficient $h = 0.1$ (red lines, points), 0.5 (black lines, points), or 0.9 (blue lines, points). The neutral SFS is also included for comparisons (grey dashed line). Rate of self-fertilisation $\sigma = 0$ or 1/2, as denoted on the right-hand side. The SFS is measured at a recombination distance of $R = 6$ for $\sigma = 0$, or $R = 11$ for $\sigma = 0.5$. The sweep arose from either a single *de novo* mutation or standing variation with $f_0 = 0.05$, as denoted above the panels. Results for other recombination distances are in Section E of S1 File.

Fig 7. Comparing sweeps from recurrent mutation to those from standing variation. (a), (b): Comparing the reduction in diversity following a 'soft' sweep (Eq 6), from either standing variation ($f_0 = 0.05$, solid lines) or recurrent mutation (using $P_{coal,M}$ with $\Theta_b = 0.2$, dashed lines). $N = 5,000$, $s = 0.05$, and dominance coefficient $h = 0.1$ (red lines), 0.5 (black lines), or 0.9 (blue lines). Populations are either outcrossing (a) or highly selfing ($\sigma = 0.95$; (b)). (c), (d): Plotting the ratio of the diversity following a sweep from standing variation (π_{SV}) to one from recurrent mutation (π_M). Parameters for each panel are as in (a) and (b) respectively. Vertical dashed black line indicates R_{Lim} (Eq 11), the predicted recombination rate where $\pi_{SV}/\pi_M = 1$ (horizontal dashed line in (c), (d)). Note the different $x-$axis lengths between panels (a), (c) and (b), (d). Results are also plotted in Section F of S1 File.

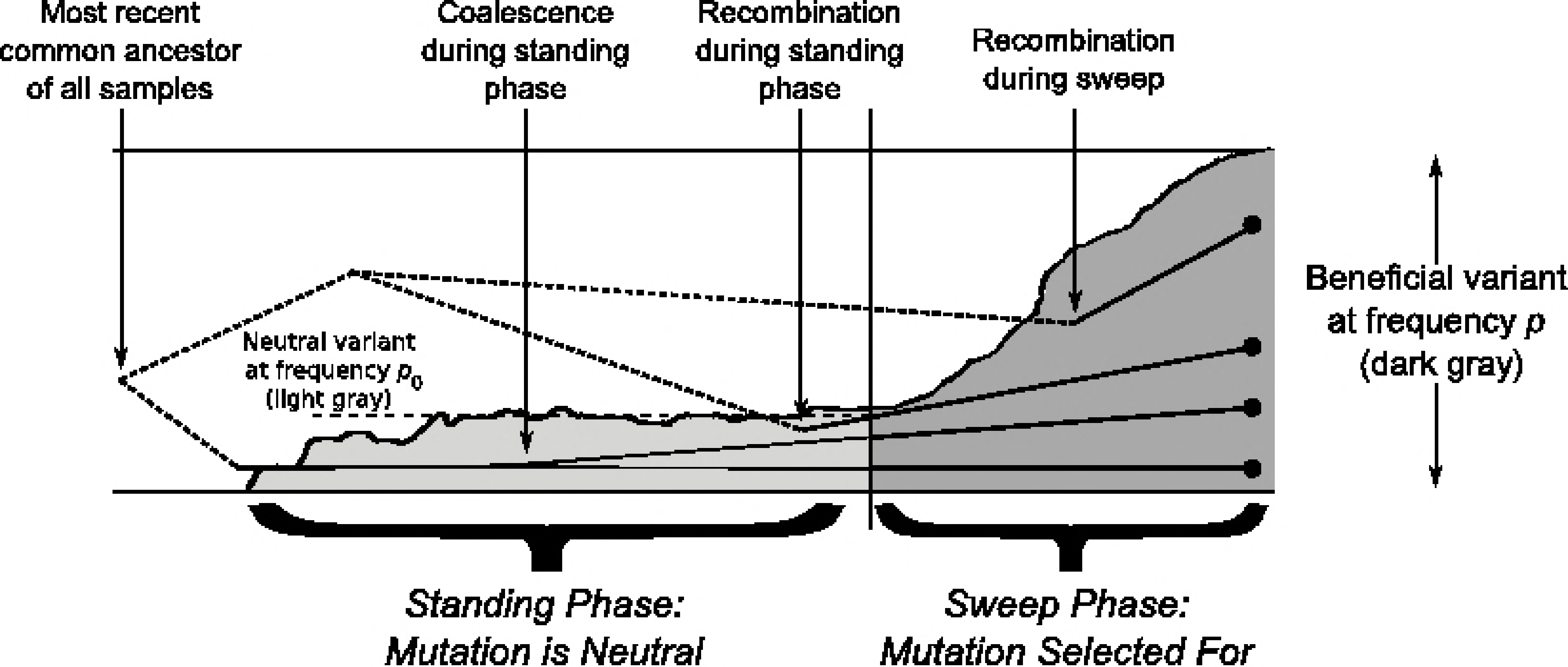
Fig 8. Histograms of distances from the selected locus to the nearest singleton. The distance is scaled to the maximum length of the sampled genome (e.g., a distance of 0.5 means that a singleton lies halfway along the sampled haplotype). A distance ">1" indicates that no singleton was observed, and therefore lies beyond the sampled haplotype. x-axis annotations denote the mid-point of each bin (e.g. '0.05' indicates distance of 0 to less than 0.1). Distances are measured from either the neutral burn-in population, or one where a 'hard' sweep ($f_0 = 1/2N$) has fixed. $N = 5,000$, $F = 0$, $4N\mu = 40$, $R = 2Nr = 4$ across the whole genetic sample. (a), (c) are log-counts of the distances for all samples over all 100 simulations; (b), (d) are the frequency of distances over samples where a singleton was observed. In (a), (b) $s = 0.05$ and the dominance coefficient h varies, with values as given in the plot legend. For (c), (d), $h = 0.5$ and s varies, with values as given in the plot legend.

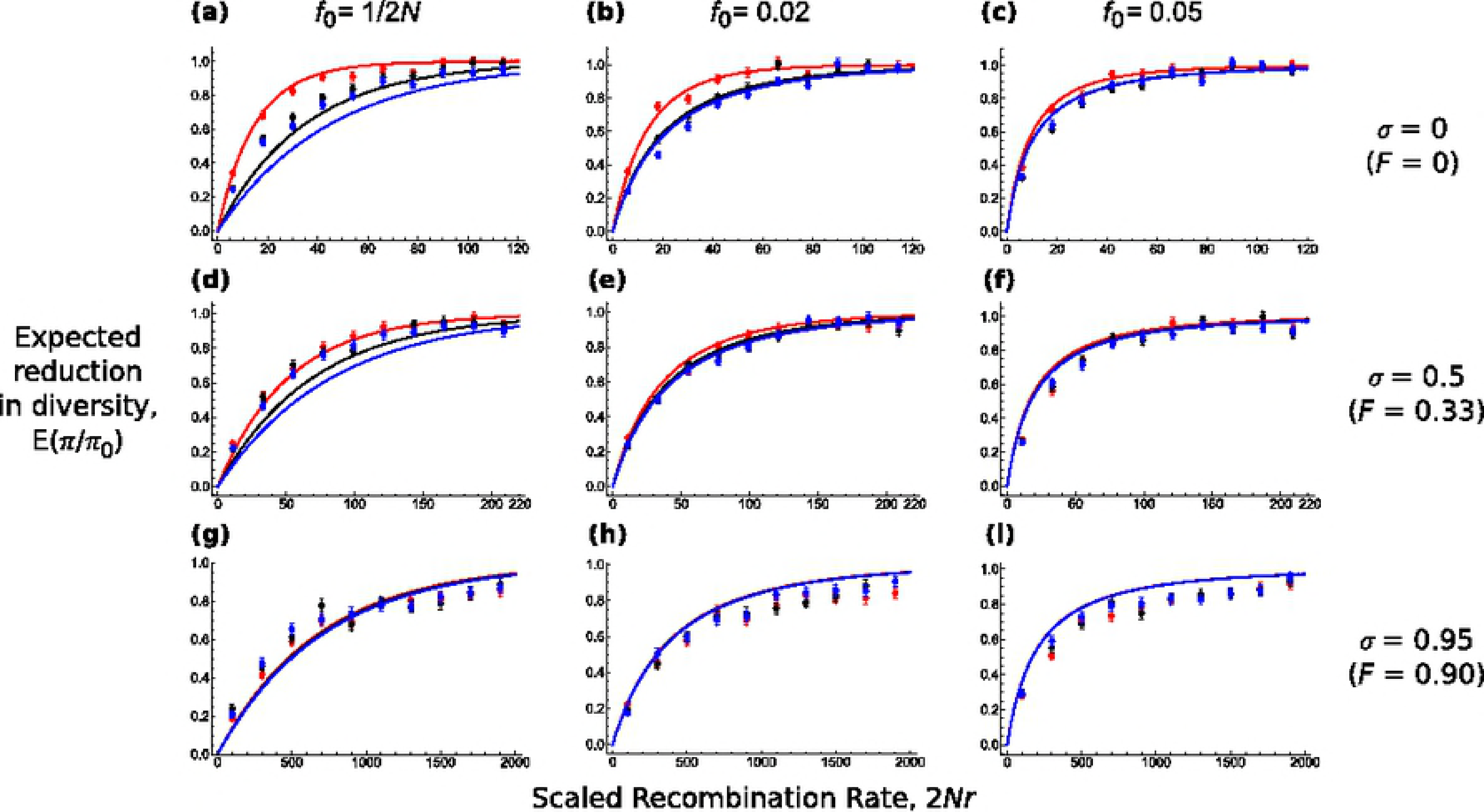
Fig 9. Plots of distances from the selected locus to the nearest singleton, for a partial sweep. Distances are measured from either the neutral burn-in population (grey dashed lines), or one where a 'hard' sweep ($f_0 = 1/2N$) has reached a frequency of 70% (coloured lines). (a) Ratio of the log-counts of the distances for derived and ancestral alleles. (b) Ratio of the frequency of singleton distances for derived and ancestral alleles for each bin (e.g. '0.05' indicates distance of 0 to less than 0.1). Measurements are taken over all samples in all simulations. All plots are log-counts of the distances for all samples over all simulations. $N = 5,000$, $F = 0$, $4N\mu = 40$, $R = 2Nr = 4$ across the whole genome. In sweep cases, $s = 0.05$ with the dominance coefficient $h = 0.1$ (red lines), 0.5 (black lines) or 0.9 (blue lines). Black dashed line indicates the 1-to-1 ratio, where the derived and ancestral classes have the same frequency.

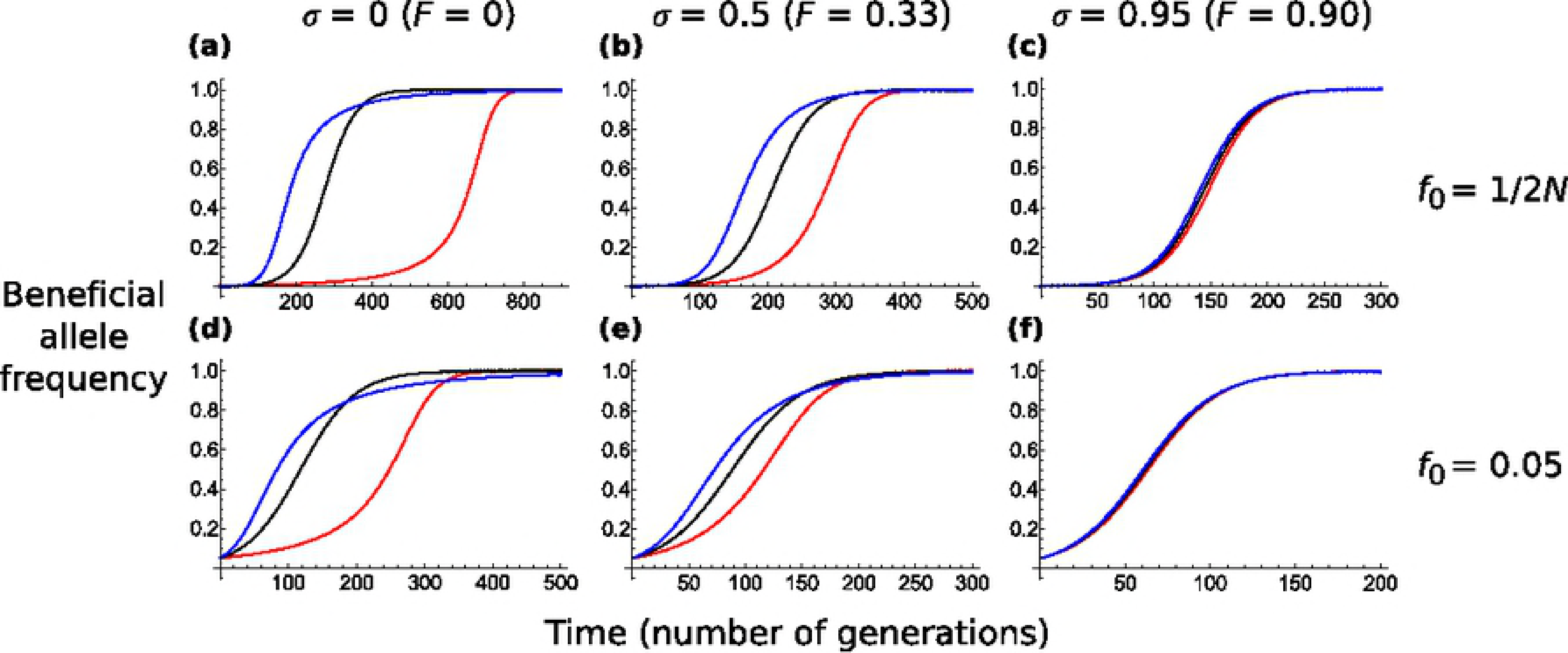
Fig 10. Analysis of the *SLC24A5* sweep signature in humans. (a) Plot of diversity around the derived SNP in the *SLC24A5* gene, scaled to baseline values (see Methods for details), as a function of the distance from the target SNP as measured in basepairs. Negative values denote distance upstream of the target SNP; positive values denote downstream distances. Red dashed line denotes the 'hard sweep' model; blue dashed line is the recurrent mutation model. (b) Plot of two haplotype statistics, H_{12} (black line) and H_2/H_1 (red line) over the sweep region. (c) Unrooted phylogenetic trees of European and African samples from the 1000 Genomes dataset at different distances from the target SNP; covered distances are denoted in the headings. Arrows indicate instances where African haplotypes carrying the derived SNP (blue triangles) are present in the clade of European samples that cluster due to the sweep.

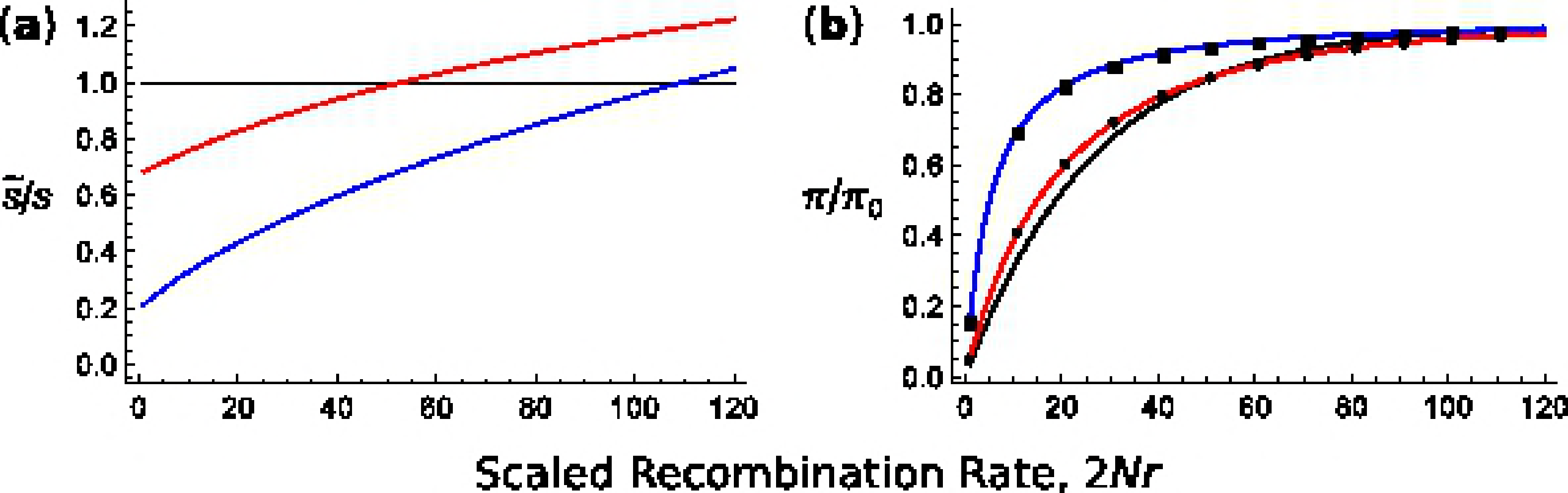
Fig 11. Schematic of how neutral polymorphisms accumulate in simulations. (a) The selected locus is located at the far left-hand side, with a neutral tract stretching out to its right. Polymorphisms accumulate along this tract, with locations standardised to be between 0 and 1. The recombination rate per reproduction is drawn from a Poisson distribution with mean r. 'Singletons' are polymorphisms where the derived allele is present in only one sample, with one present at location 0.65. (b) Measuring singleton distances using segregating target SNPs at a reference point. In the top sample the nearest singleton is located upstream of the target SNP, with distance 0.3 between them. In the bottom sample the singleton is located downstream of the SNP. Hence the total distance is given as the upstream distance to the right-hand edge (0.5), plus the distance of the singleton from the left-hand edge (0.1), giving a total distance of 0.6.

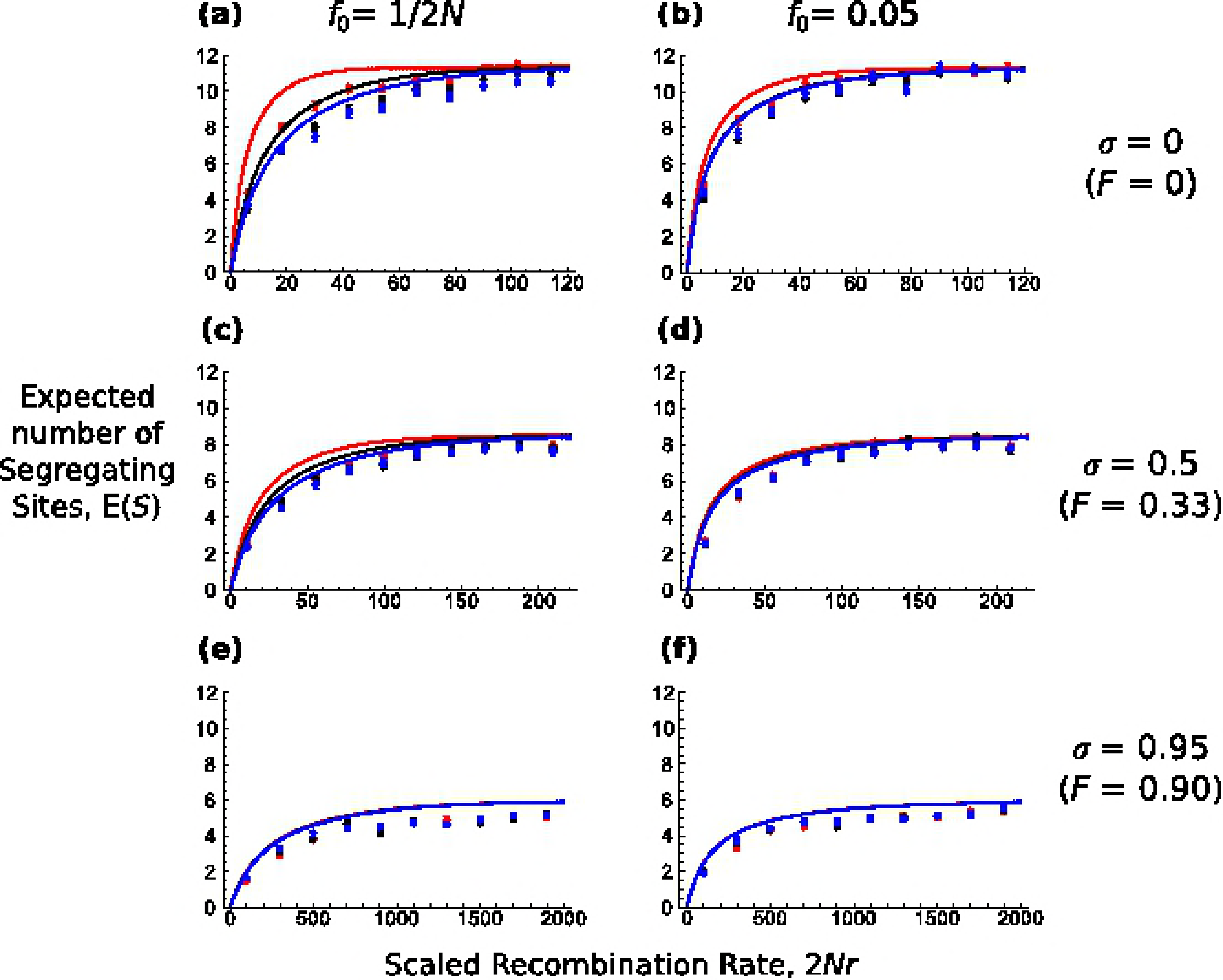
Fig 12. Diversity and recombination data around the *SLC24A5* sweep region. (a) Plot of raw pairwise diversity in 20Kb bins, as a function of distance from the target SNP. Dashed grey lines show mean diversity values when measured either upstream or downstream of the target SNP. (b) Relative diversity measurements, after dividing raw diversity measurements by the mean values from either up- or downstream of the target SNP. (c) Cumulative recombination distance from the target SNP, as obtained from Bhérer *et al.* [90], scaled by $2N = 20,000$.

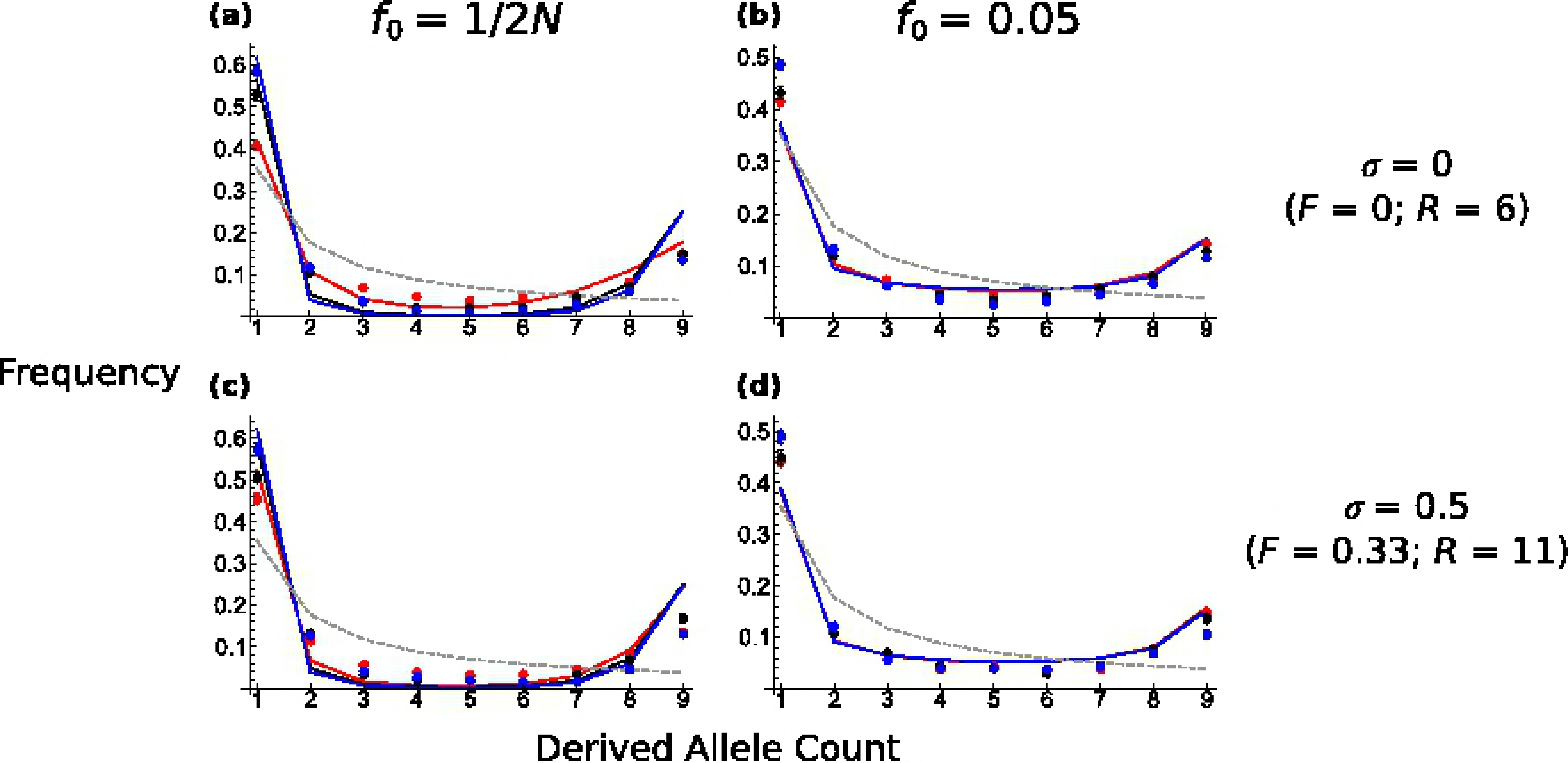

Most recent common ancestor of all samples
Coalescence during standing phase
Recombination during standing phase
Recombination during sweep
Neutral variant at frequency p_0 (light gray)
Beneficial variant at frequency p (dark gray)
Standing Phase: Mutation is Neutral
Sweep Phase: Mutation Selected For

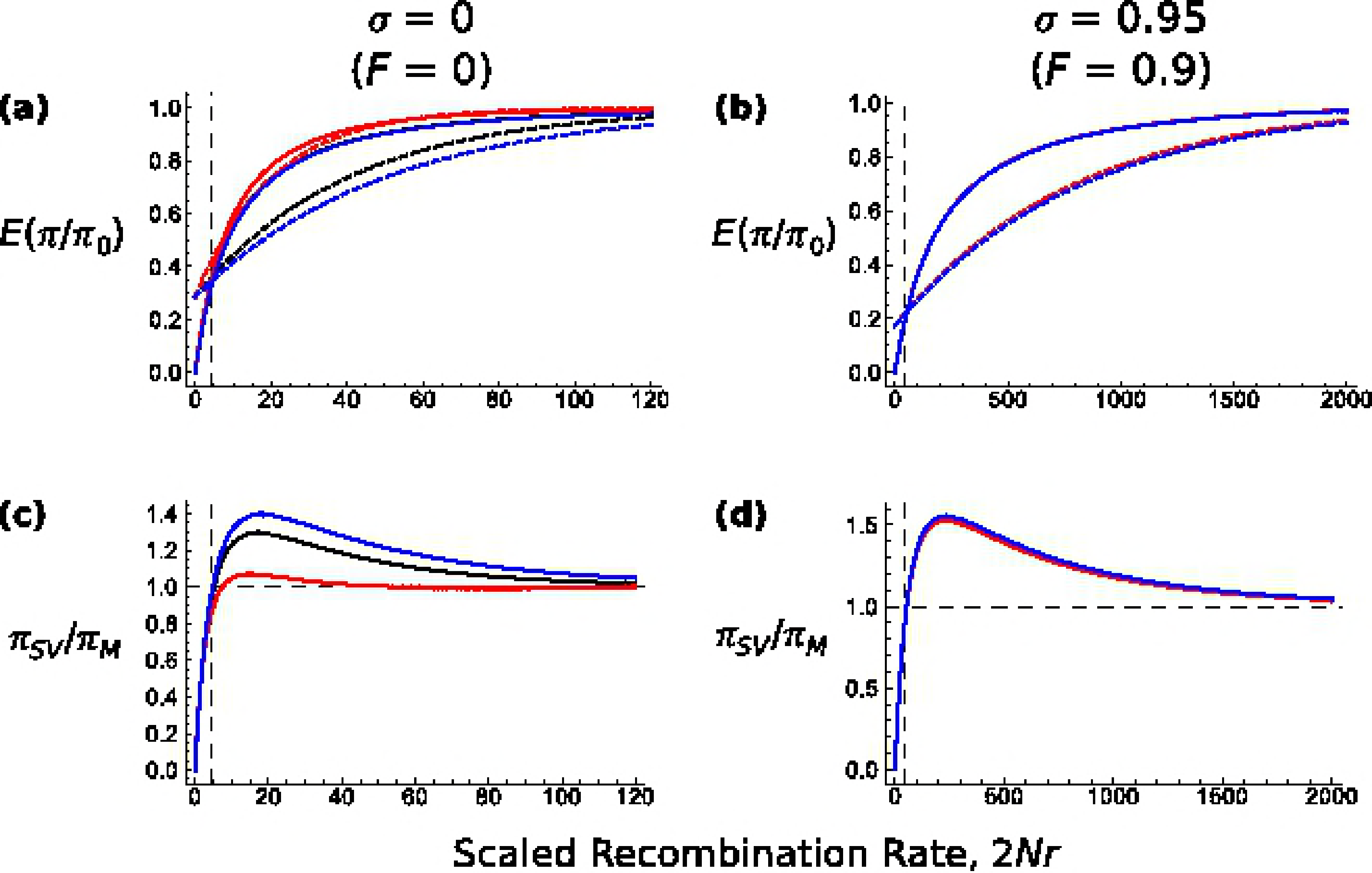

(a)
f_0= 1/2N
(b)
f_0= 0.02
(c)
f_0= 0.05
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
σ = 0
(F = 0)
(d)
(e)
(f)
0
50
100
150
200
220
σ = 0.5
(F = 0.33)
(g)
(h)
(i)
0
500
1000
1500
2000
σ = 0.95
(F = 0.90)
Expected reduction in diversity, E(π/π_0)
Scaled Recombination Rate, 2Nr

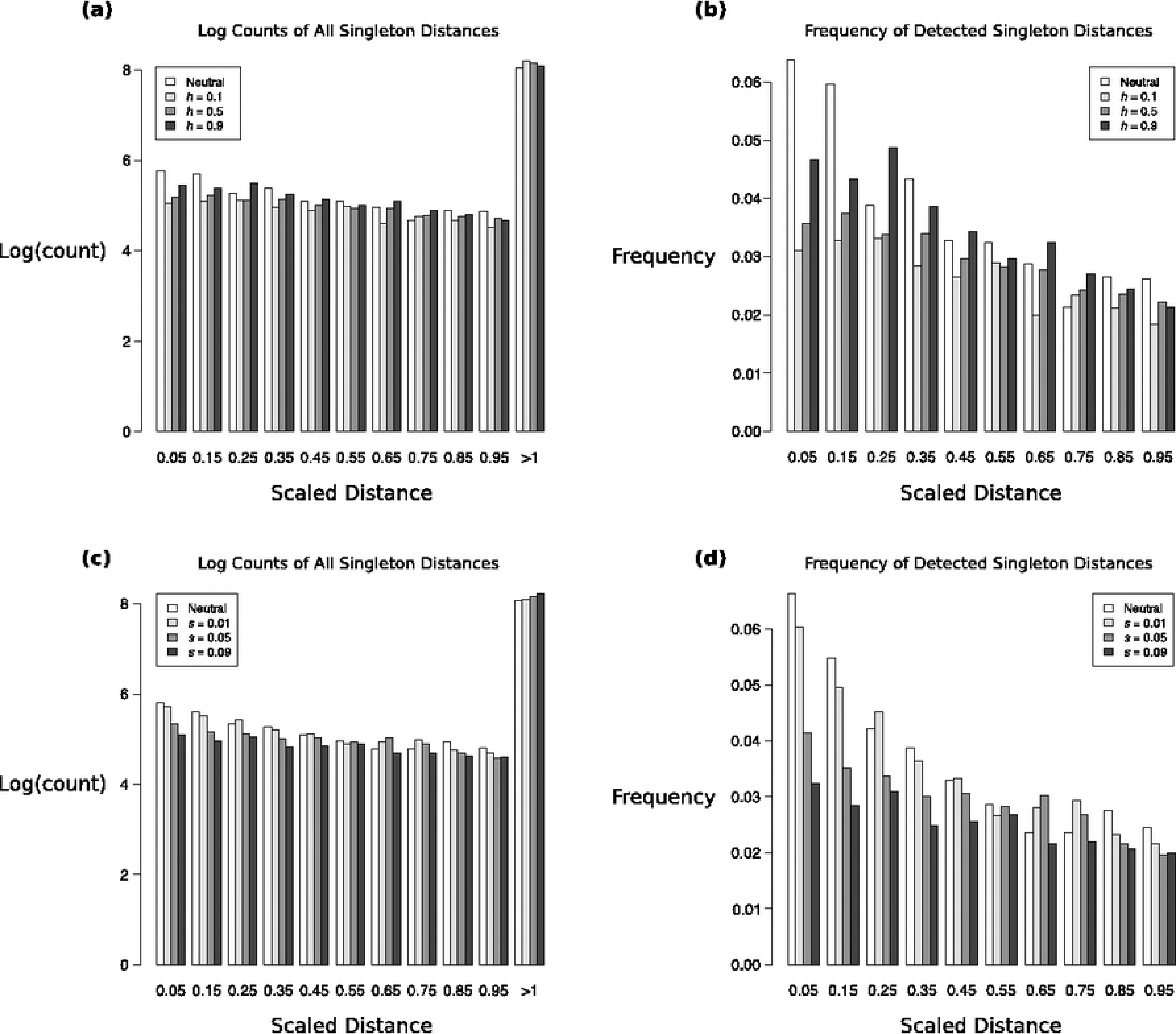

$\sigma = 0$ ($F = 0$)
$\sigma = 0.5$ ($F = 0.33$)
$\sigma = 0.95$ ($F = 0.90$)
(a)
(b)
(c)
(d)
(e)
(f)
$f_0 = 1/2N$
$f_0 = 0.05$
Beneficial allele frequency
Time (number of generations)
0.2
0.4
0.6
0.8
1.0
200
400
600
800
100
300
500
50
150
250

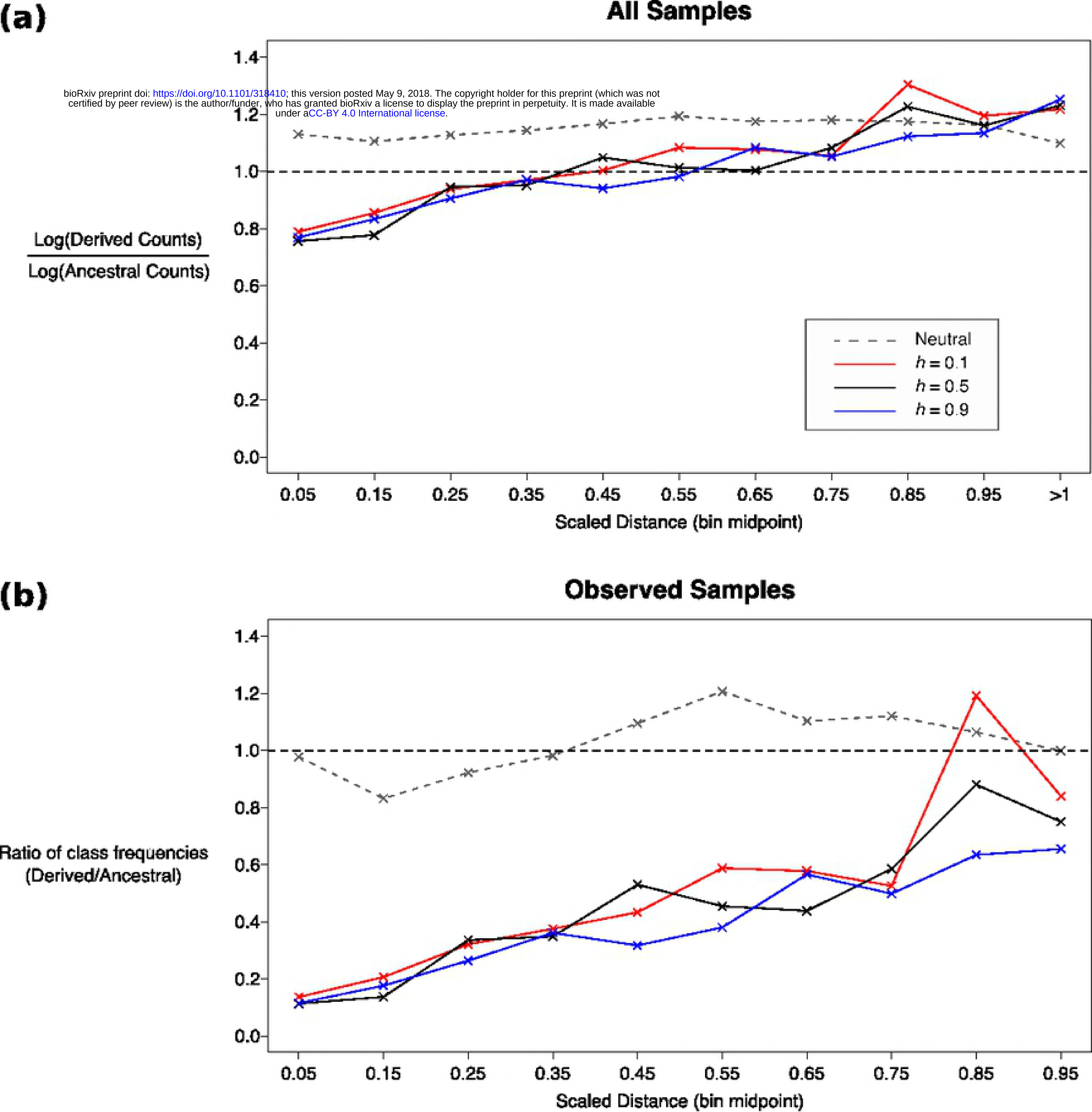

(a)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$\tilde{s}/s$
0
20
40
60
80
100
120
(b)
1.0
0.8
0.6
0.4
0.2
0.0
π/π_0
0
20
40
60
80
100
120
Scaled Recombination Rate, $2Nr$

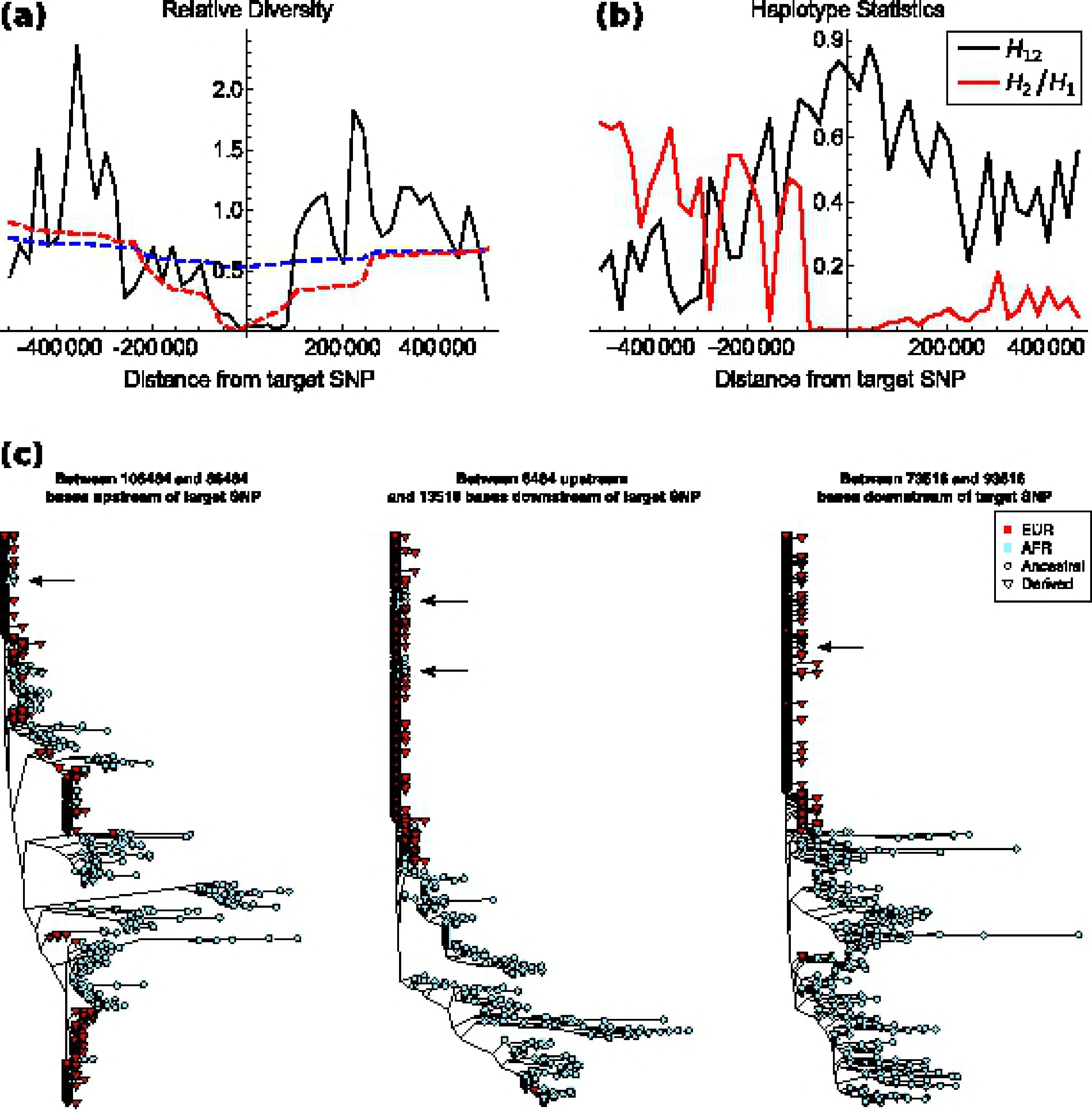

(a)
$f_0 = 1/2N$
(b)
$f_0 = 0.05$
$\sigma = 0$
$(F = 0)$
(c)
(d)
$\sigma = 0.5$
$(F = 0.33)$
(e)
(f)
$\sigma = 0.95$
$(F = 0.90)$
Expected number of Segregating Sites, E(S)
Scaled Recombination Rate, 2Nr
0
2
4
6
8
10
12
0
20
40
60
80
100
120
50
150
200
500
1000
1500
2000

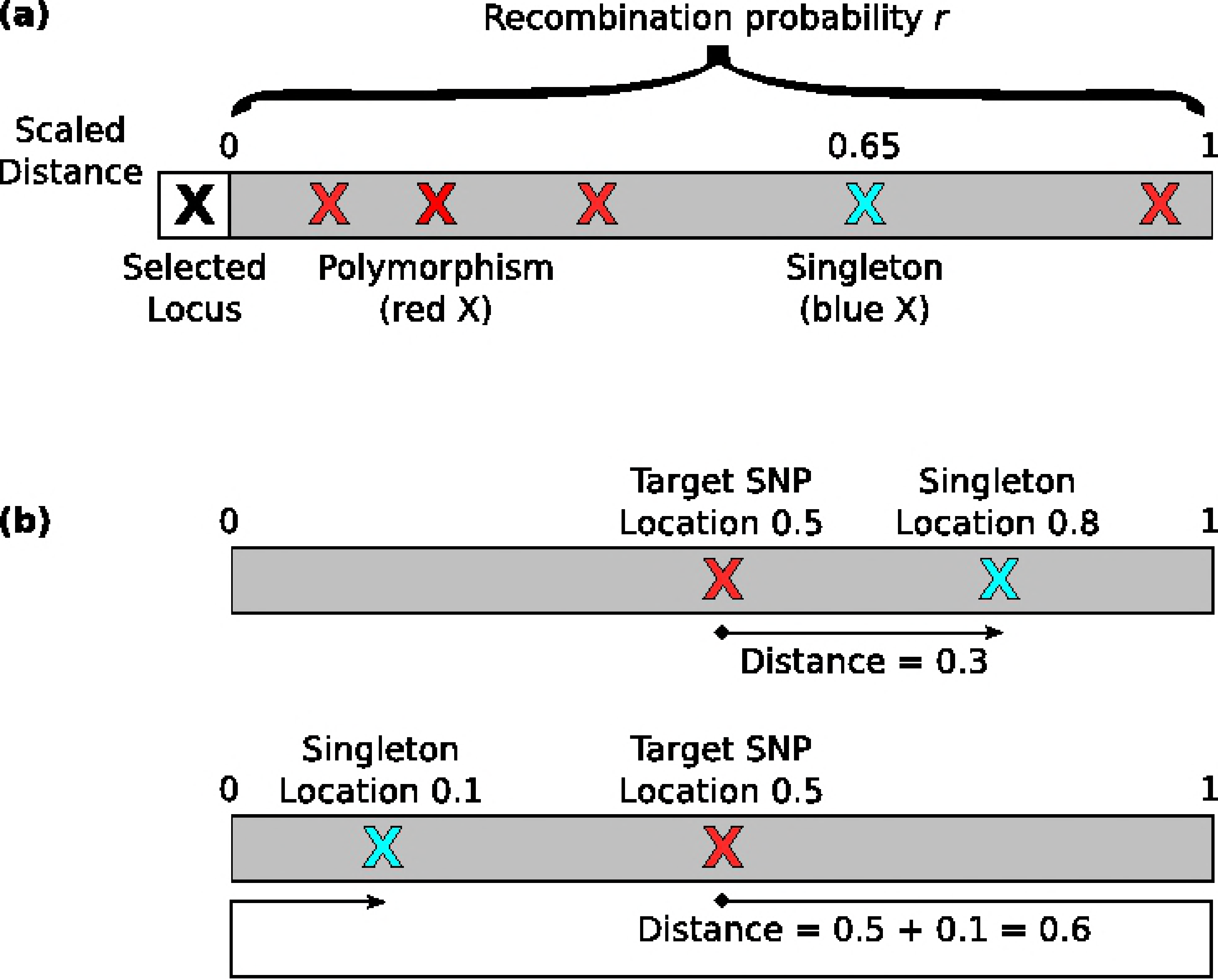

(a)
$f_0 = 1/2N$
(b)
$f_0 = 0.05$
$\sigma = 0$
$(F = 0; R = 6)$
Frequency
(c)
(d)
$\sigma = 0.5$
$(F = 0.33; R = 11)$
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
Derived Allele Count

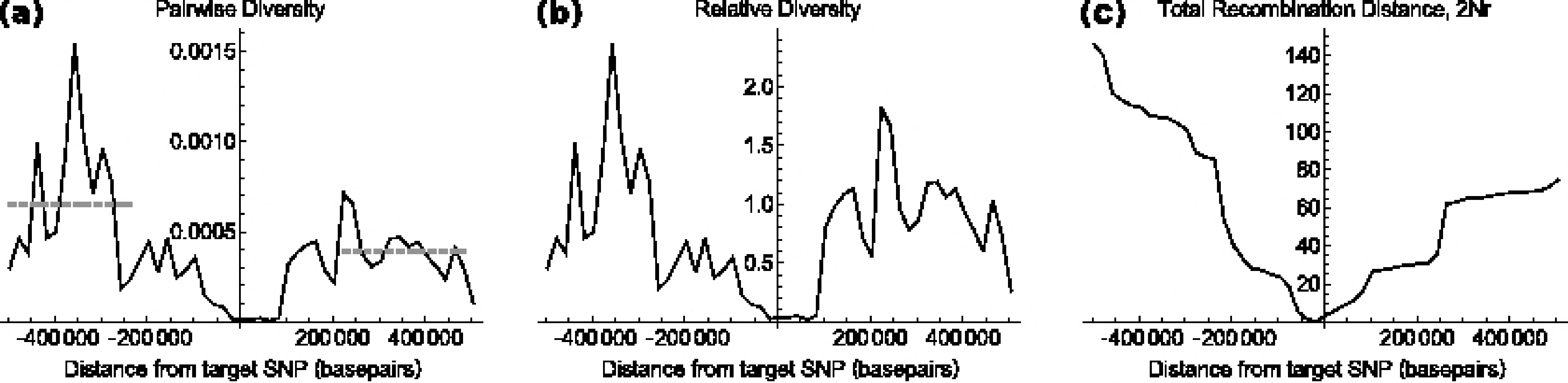

σ = 0
(F = 0)
σ = 0.95
(F = 0.9)
(a)
(b)
(c)
(d)
E(π/π0)
πSV/πM
1.0
0.8
0.6
0.4
0.2
0.0
1.4
1.2
1.5
0.5
0
20
40
60
80
100
120
500
1000
1500
2000
Scaled Recombination Rate, 2Nr


(a)
Log Counts of All Singleton Distances
Neutral
h = 0.1
h = 0.5
h = 0.9
Log(count)
0
2
4
6
8
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
>1
Scaled Distance
(b)
Frequency of Detected Singleton Distances
Neutral
h = 0.1
h = 0.5
h = 0.9
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
Scaled Distance
(c)
Log Counts of All Singleton Distances
Neutral
s = 0.01
s = 0.05
s = 0.09
Log(count)
0
2
4
6
8
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
>1
Scaled Distance
(d)
Frequency of Detected Singleton Distances
Neutral
s = 0.01
s = 0.05
s = 0.09
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
Scaled Distance


(a)
All Samples
Log(Derived Counts)
Log(Ancestral Counts)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Neutral
h = 0.1
h = 0.5
h = 0.9
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
>1
Scaled Distance (bin midpoint)
(b)
Observed Samples
Ratio of class frequencies
(Derived/Ancestral)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
Scaled Distance (bin midpoint)


(a)
Relative Diversity
2.0
1.5
1.0
0.5
-400000
-200000
200000
400000
Distance from target SNP
(b)
Haplotype Statistics
H_{12}
H_2/H_1
0.8
0.6
0.4
0.2
-400000
-200000
200000
400000
Distance from target SNP
(c)
Between 106484 and 86484 bases upstream of target SNP
Between 6484 upstream and 13516 bases downstream of target SNP
Between 73516 and 93516 bases downstream of target SNP
EUR
AFR
Ancestral
Derived


(a)
Recombination probability *r*
Scaled Distance
0
0.65
1
X
X X X X X
Selected Locus
Polymorphism (red X)
Singleton (blue X)
(b)
0
Target SNP Location 0.5
Singleton Location 0.8
1
X X
Distance = 0.3
0
Singleton Location 0.1
Target SNP Location 0.5
1
X X
Distance = 0.5 + 0.1 = 0.6


(a)
Pairwise Diversity
0.0015
0.0010
0.0005
-400000
-200000
200000
400000
Distance from target SNP (basepairs)
(b)
Relative Diversity
2.0
1.5
1.0
0.5
-400000
-200000
200000
400000
Distance from target SNP (basepairs)
(c)
Total Recombination Distance, 2Nr
140
120
100
80
60
40
20
-400000
-200000
200000
400000
Distance from target SNP (basepairs)