Selective sweeps under dominance and inbreeding

Matthew Hartfield[1,2,3,*], Thomas Bataillon[2]

1 Department of Ecology and Evolutionary Biology, University of Toronto, Ontario M5S 3B2, Canada.

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

3 Institute of Evolutionary Biology, The University of Edinburgh, Edinburgh EH9 3FL, United Kingdom.

* m.hartfield@ed.ac.uk

Running Head: Sweeps under dominance and inbreeding

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

Website for simulation code:

`https://github.com/MattHartfield/SweepDomSelf`

Abstract

A major research goal in evolutionary genetics is to uncover loci experiencing positive selection. One approach involves finding 'selective sweeps', either formed by *de novo* mutation, or 'soft sweeps' arising from recurrent mutation or existing standing variation. Existing theory generally assumes outcrossing populations, and it is unclear how dominance affects soft sweeps. We consider how arbitrary dominance and inbreeding via self-fertilisation affect hard and soft sweep signatures. With increased self-fertilisation, they are maintained over longer map distances due to reduced effective recombination and faster beneficial allele fixation times. Dominance can affect sweep patterns in outcrossers if the derived variant originates from either a single novel allele, or from recurrent mutation. We apply models to infer properties of the *SLC24A5* sweep in European humans, and a sweep in the selfing plant *Arabidopsis thaliana*. These analyses highlight the challenges in distinguishing hard and soft sweeps, and propose methods to differentiate between scenarios.

Introduction

Inferring adaptive mutations from nucleotide polymorphism data is a major research goal in evolutionary genetics, and has been subject to extensive modelling work to determine the footprints they leave in genome data (Stephan 2019). The earliest models focussed on a scenario where a beneficial mutation arose as a single copy before rapidly fixing. Linked neutral mutations then 'hitchhike' to fixation with the adaptive variant, reducing diversity around the selected locus (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989). Hitchhiking also increases linkage disequilibrium at regions flanking the selected site, although it is minimal when measured at sites either side of the selected mutation (Thomson 1977; Innan and Nordborg 2003; McVean 2007). These theoretical expectations have spurred the creation of summary statistics for detecting sweeps, usually based on finding genetic regions exhibiting extended runs of homozygosity (Sabeti *et al.* 2002; Kim and Nielsen 2004; Voight *et al.* 2006; Ferrer-Admetlla *et al.* 2014; Vatsiou *et al.* 2016), or carrying an elevated number of derived variants (Fay and Wu 2000; Kim and Stephan 2002; Nielsen 2005; Boitard *et al.* 2009; Yang *et al.* 2018; Fujito *et al.* 2018).

Classic hitchhiking models consider 'hard' sweeps, where the common ancestor of an adaptive allele occurs after the onset of selection (Hermisson and Pennings 2017). Recent years have seen a focus on 'soft' sweeps, where the most recent common ancestor of a beneficial allele appeared before it became selected for (reviewed by Barrett and Schluter (2008); Messer and Petrov (2013); Hermisson and Pennings (2017)). Soft sweeps can originate from beneficial mutations being introduced by recurrent mutation at the target locus (Pennings and Hermisson 2006a,b),

or originating from existing standing variation that was either neutral or deleterious (Orr and Betancourt 2001; Innan and Kim 2004; Przeworski *et al.* 2005; Hermisson and Pennings 2005; Wilson *et al.* 2014; Berg and Coop 2015; Wilson *et al.* 2017). 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 may be present around the derived allele. This property is often used to detect soft sweeps in genetic data (Peter *et al.* 2012; Vitti *et al.* 2013; Garud *et al.* 2015; Garud and Petrov 2016; Schrider and Kern 2016; Sheehan and Song 2016; Harris *et al.* 2018a; Kern and Schrider 2018). Soft sweeps have been inferred in *Drosophila* (Karasov *et al.* 2010; Garud *et al.* 2015; Garud and Petrov 2016; Vy *et al.* 2017), humans (Peter *et al.* 2012; Schrider and Kern 2017), maize (Fustier *et al.* 2017), and pathogens including *Plasmodium falciparum* (Anderson *et al.* 2016) and HIV (Pennings *et al.* 2014). Yet determining how extensive soft sweeps are in nature remains a contentious issue (Jensen 2014; Harris *et al.* 2018b).

Up to now, there have only been a few investigations into how dominance affects sweep signatures. In a simulation study, Teshima and Przeworski (2006) explored how recessive mutations spend long periods 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 significantly high frequency to be detectable by genome scans (Teshima *et al.* 2006). Ewing *et al.* (2011) have carried out a general mathematical analysis of how dominance affects hard sweeps. Yet the impact of dominance on soft sweeps has yet to be explored in depth.

In addition, existing models have so far focussed on randomly mating populations, with haplotypes freely mixing between individuals over generations. Dif-

ferent reproductive modes alter how alleles are inherited, affecting the hitchhiking effect. Self-fertilisation, where male and female gametes produced from the same individual can fertilise one another, can alter adaptation rates and selection signatures (Hartfield *et al.* 2017). This mating system is prevalent amongst angiosperms (Igic and Kohn 2006), some animals (Jarne and Auld 2006) and fungi (Billiard *et al.* 2011). As the effects of dominance and self-fertilisation become strongly intertwined, it is important to consider both together. Dominant mutations are more likely to fix then recessive ones in outcrossers, as they have a higher initial selection advantage (Haldane 1927). Yet recessive alleles can fix more easily in selfers than in outcrossers as homozygote mutations are created more rapidly (Charlesworth 1992; Glémin 2012). Furthermore, a decrease in effective recombination rates in selfers (Nordborg *et al.* 1996; Nordborg 2000; Charlesworth and Charlesworth 2010) can interfere with selection acting at linked sites, e.g., making it likelier that deleterious mutations hitchhike to fixation with adaptive alleles (Hartfield and Glémin 2014), or competition between adaptive mutations at closely-linked loci increases the probability that rare mutations are lost by drift (Hartfield and Glémin 2016).

In a constant-sized population, beneficial mutations can be less likely to fix from standing variation (either neutral or deleterious) in selfers as they maintain lower diversity levels (Glémin and Ronfort 2013). Yet adaptation from standing variation becomes likelier in selfers compared to outcrossers under 'evolutionary rescue' scenarios, where swift adaptation is needed to prevent population extinction following environmental change. Here, rescue mutations are only present in standing variation as the population size otherwise becomes too small (Glémin and Ronfort 2013). Self-fertilisation further aids this process by creating beneficial

homozygotes more rapidly than in outcrossing populations (Uecker 2017).

Little data currently exists on the extent of soft sweeps in self-fertilisers. Many selfing organisms exhibit sweep-like patterns, including *Arabidopsis thaliana* (Long *et al.* 2013; Huber *et al.* 2014; Fulgione *et al.* 2018); *Caenorhabditis elegans* (Andersen *et al.* 2012); *Medicago truncatula* (Bonhomme *et al.* 2015); and *Microbotryum* fungi (Badouin *et al.* 2017). Detailed analyses of these cases has been hampered by a lack of theory on how hard and soft sweep signatures should manifest themselves under different self-fertilisation and dominance levels. Previous studies have only focussed on special cases; Hedrick (1980) analysed linkage disequilibrium caused by a hard sweep under self-fertilisation, while Schoen *et al.* (1996) modelled sweep patterns caused by modifiers that altered the mating system in different ways.

To this end, we develop a general selective sweep model. We determine the genetic diversity present following a sweep from either a *de novo* mutation, or from standing variation. We also determine the number of segregating sites and the site frequency spectrum, while comparing results to an alternative soft-sweep model where adaptive alleles arise via recurrent mutation. We subsequently use these results to determine characteristics of two selective sweeps, in humans and the selfing plant *Arabidopsis thaliana*. Note that we focus here on single sweep events, rather than characterising how sweeps affect genome-wide diversity (Elyashiv *et al.* 2016; Campos *et al.* 2017; Booker and Keightley 2018; Rettelbach *et al.* 2019).

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 $1 - \sigma$. 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 (Kimura 1971). Locus B determines fitness differences, carrying an allele that initially segregates neutrally for a period of time. After a period of time the allele becomes advantageous with selective advantage $1 + hs$ when heterozygous and $1 + s$ when homozygous, with $0 < h < 1$ and $s > 0$. We further assume that selection is strong (i.e., $N_e hs \gg 1$) so the sweep trajectory can be modelled deterministically. p_0 is the frequency at which the allele becomes selected for. Table 1 lists the notation used in the analysis.

Our overall goal is to determine how the spread of an adaptive allele at locus B affects genealogies underlying polymorphism at locus A, by considering whether neutral alleles at A are associated with the selected derived allele or ancestral neutral allele at locus B. A schematic is shown in Figure 1. We follow the approach of Berg and Coop (2015) and, looking backwards in time, break down the allele history of B into two phases. The first phase (the 'sweep phase') considers the derived allele at B being selectively favoured from an initial frequency p_0 and spreading through the population. The second phase (the 'standing phase') assumes that the derived allele is present at an frequency with mean p_0. During both phases, a pair of haplotypes can either coalesce, or one of them recombines onto the ancestral background.

Symbol	Usage
N	Population size (with $2N$ haplotypes)
σ	Proportion of matings that are self-fertilising
F	Wright's inbreeding coefficient, probability of identity-by-descent at a single gene, equal to $\sigma/(2-\sigma)$ at steady-state
Φ	Joint probability of identity-by-descent at two loci (Equation 1)
N_e	Effective population size, equal to $N/(1+F)$ with selfing
A, B	Loci carrying neutral, selected alleles
r	Recombination rate between loci A and B
r_{eff}	'Effective' recombination rate, approximately equal to $r(1-2F+\Phi)$ with selfing
R	$2Nr$, the population-level recombination rate
p_0	Frequency at which the derived allele at B becomes advantageous
$p_{0,A}$	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_{p_0}	Time in the past when derived locus became beneficial
$p(t)$	Frequency of beneficial allele at time t
$P_c(t)$	Probability of coalescence at time t
$P_r(t)$	Probability of recombination at time t
P_{NE}	Probability that neutral marker does not coalesce or recombine during sweep phase
$P_{NE,SL}$	P_{NE} using 'star-like' approximation (no coalescence during sweep phase)
$P_{R,Sw}$	Probability that neutral marker recombines during sweep phase
$P_{R,Sd}$	Probability that neutral marker recombines during standing phase
$P_{M,Sw}$	Probability that a lineage mutates during sweep phase
$P_{M,Sd}$	Probability that a lineage mutates during standing phase
H_l, H_h	'Effective' dominance coefficient for allele at low, high frequency
π	Pairwise diversity at site (π_0 is expected value without a sweep)
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
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.

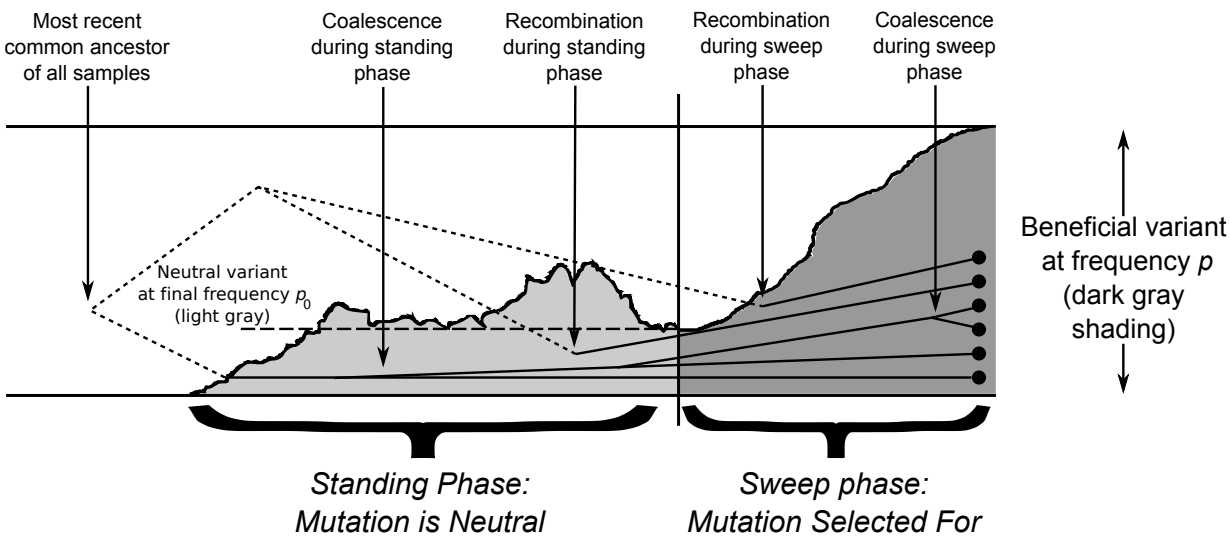

Figure 1. A schematic of the model. The history of the derived variant is separated into two phases. The 'standing phase' (shown in light gray), is when the derived variant is segregating at a frequency close to p_0 for a long period of time. The 'sweep phase' (shown in dark gray) is when the variant becomes selected for and starts increasing in frequency. Dots on the right-hand side represent a sample of haplotypes taken at the present day, with lines representing their genetic histories. Samples can either coalesce or recombine onto the ancestral background during either phase. Solid lines represent coalescent histories for the derived genetic background; dotted lines represent coalescent histories for the ancestral background.

For tightly linked loci (r close to 0), the relatively rapid fixation time of the derived variant makes it unlikely for a given neutral variant to be present on different backgrounds (with respect to the selected locus), 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 the initial level. Self-fertilisation creates two key differences compared to outcrossing. First, the adaptive allele trajectory, which underlies expected diversity patterns, is affected by the levels of self-fertilisation (σ) and dominance (h). Second, the effective population size (which determines the coalescence rate) and recombina-

tion frequency are scaled by factors $1/(1+F)$ and $1-2F+\Phi$ respectively, for $F=\sigma/(2-\sigma)$ the inbreeding coefficient (Wright 1951; Pollak 1987; Charlesworth 1992; Caballero and Hill 1992; Nordborg *et al.* 1996; Nordborg and Donnelly 1997; Nordborg 2000) and Φ the joint probability of identity-by-descent at the two loci (Roze 2009, 2016; Hartfield and Glémin 2016). At equilibrium, it equals:

$$\Phi = \frac{\sigma(2-\sigma-2(1-r)r(2-3\sigma))}{(2-\sigma)(2-(1-2(1-r)r)\sigma)} \tag{1}$$

Note that $1-2F+\Phi$ approximates to $1-F$ (as $\Phi \approx F$) for most combinations of recombination and selfing fractions, unless σ is close to one and r becomes high. The N_e scaling factor $1/(1+F)$ can also be a good approximation if there is non-Poisson variation in offspring, unless female fitness strongly affects reproduction number (Laporte and Charlesworth 2002). Although we focus on inbreeding via self-fertilisation, the scalings $N_e = N/(1+F)$ and $r_e \approx r(1-F)$ should also hold under other systems of regular inbreeding (Caballero and Hill 1992; Charlesworth and Charlesworth 2010, Box 8.4).

We will outline how both coalescence and recombination act during both of these phases, and use these calculations to determine selective sweep properties. Throughout, analytical solutions are compared to results from Wright-Fisher forward-in-time stochastic simulations; the procedure is described in the 'Methods' section. Although previous models tended to exclude coalescence during the sweep phase, including it is important for producing accurate matches with simulation results (Charlesworth, in prep.).

Probability of events during sweep phase

We first look at the probability of events (coalescence or recombination) acting throughout the sweep phase. Looking back in time following a sweep, sites linked to the beneficial allele can either coalesce or recombine onto the ancestral genetic background. Let $p(t)$ be the adaptive mutation frequency 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_{p_0} the time in the past when the derived variant became beneficial (i.e., $p(\tau_{p_0}) = p_0$).

For a pair of haplotype samples carrying the derived allele, if it is at frequency $p(t)$ at time t, this lineage pair can either coalesce or one of the haplotypes recombine onto the ancestral background. Each event occurs with probability:

$$\begin{aligned} P_c(t) &= \frac{1}{2N_e p(t)} = \frac{(1+F)}{2Np(t)} \\ P_r(t) &= 2r_{eff}(1-p(t)) = 2r(1-2F+\Phi)(1-p(t)) \end{aligned} \tag{2}$$

Equation 2 is based on those obtained by Kaplan *et al.* (1989), assuming that N_e is reduced by a factor $1+F$ due to self-fertilisation (Pollak 1987; Charlesworth 1992; Caballero and Hill 1992; Nordborg and Donnelly 1997), and $r_{eff} = r(1-2F+\Phi)$ is the 'effective' recombination rate after correcting for increased homozygosity due to self-fertilisation (Nordborg *et al.* 1996; Nordborg 2000; Charlesworth and Charlesworth 2010; Roze 2009, 2016; Hartfield and Glémin 2016).

We are interested in calculating (i) the probability P_{NE} that no coalescence or recombination occurs; (ii) the probability $P_{R,Sw}$ that recombination acts on a lineage to transfer it to the neutral background, assuming that no more than one

recombination event occurs per generation (see Campos and Charlesworth (2019) for derivations assuming multiple recombination events). We will go through these probabilities in turn to determine expected pairwise diversity. For P_{NE}, the total probability that the two lineages does not coalesce or recombine over τ_{p_0} generations equals:

$$
\begin{aligned}
P_{NE} &= \prod_{t=0}^{\tau_{p_0}} [1 - P_c(t) - P_r(t)] \\
&\approx \exp\left(-\int_{t=0}^{\tau_{p_0}} [P_c(t) + P_r(t)]\,\mathrm{d}t\right) && \text{assuming } P_c,\ P_r \ll 1 \\
&\approx \exp\left(-\int_{t=0}^{\tau_{p_0}} \left[\frac{1+F}{2Np(t)} + 2r(1-2F+\Phi)p(t)\right]\mathrm{d}t\right) \\
&\approx \exp\left(-\int_{p=1-\epsilon}^{p_0} \left[\frac{\frac{1+F}{2Np} + 2r(1-2F+\Phi)p}{\mathrm{d}p/\mathrm{d}t}\right]\mathrm{d}p\right) && \text{taking the integral over } p
\end{aligned}
\tag{3}
$$

Here ϵ is a small term and $1-\epsilon$ is the upper limit of the deterministic spread of the beneficial allele. We will discuss in the section 'Effective starting frequency from a de novo mutation' what a reasonable value for ϵ should be. We can calculate P_{NE} for a general self-fertilisation level 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 (Glémin 2012):

$$
\frac{\mathrm{d}p}{\mathrm{d}t} = -sp(1-p)(F + h - Fh + (1-F)(1-2h)p) \tag{4}
$$

Note the negative factor in Equation 4 since we are looking back in time. By substituting Equation 4 into Equation 3, we obtain an analytical solution for P_{NE},

although the resulting expression is complicated (Section A of Supplementary File S1).

To calculate $P_{R,Sw}$, the probability that recombination acts during the sweep, we first calculate the probability that recombination occurs when the beneficial allele is at frequency p'. Here, no events occur in the time leading up to p', then a recombination event occurs with probability $2r(1-2F+\Phi)(1-p')$. $P_{R,Sw}$ is obtained by summing this probability over the entire sweep from time 0 to τ_{p_0}, which can be approximated in continuous time by integration:

$$P_{R,Sw} \approx \int_{p'=1-\epsilon}^{p_0} \frac{P_{R,p'}}{\mathrm{d}p'/\mathrm{d}t}\mathrm{d}p' \tag{5}$$

where:

$$P_{R,p'} = \exp\left[-\int_{p=1-\epsilon}^{p'} \frac{\frac{1+F}{2Np} + 2r(1-2F+\Phi)p}{\mathrm{d}p/\mathrm{d}t}\mathrm{d}p\right] \cdot [2r(1-2F+\Phi)(1-p')]$$

$\mathrm{d}p/\mathrm{d}t$ is Equation 4 but with p' instead of p. We evaluate Equation 5 numerically.

'Star-like' approximation (no coalescence during the sweep phase)

It is possible to obtain more tractable analytical solutions by using a 'separation-of-timescales' approximation, and assume that no coalescence occurs during the sweep phase (Pennings and Hermisson 2006b; Berg and Coop 2015). Here we only have to calculate the probability that no recombination occurs during the sweep phase, which for a single lineage is equal to:

$$
\begin{aligned}
P_{NE,SL} &= \exp\left(-\int_{p=1-\epsilon}^{p_0} \frac{r_{eff}(1-p)}{\mathrm{d}p/\mathrm{d}t}\mathrm{d}p\right) \\
&= \exp\left(-\frac{r_{eff}}{H_l s}\log\left[1+\frac{H_l}{H_h}\left(\frac{1}{p_0}-1\right)\right]\right) \\
&= \left[1+\frac{H_l}{H_h}\left(\frac{1}{p_0}-1\right)\right]^{-r_{eff}/(H_l s)}
\end{aligned} \tag{6}
$$

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 (Glémin 2012). Note that for the special case $\sigma = 0$ and $h = 1/2$, $H_l = H_h = 1/2$ and Equation 6 reduces to $(1/p_0)^{-(2r/s)}$, which is equivalent to Equation 2 of Berg and Coop (2015) after scaling the selection coefficient by 1/2 to include semidominance.

Probability of coalescence from standing variation

When the variant becomes advantageous at frequency p_0, we expect $\sim 2Np_0$ haplotypes to carry it. We assume that p_0, and hence event probabilities, both remain invariant over time. Berg and Coop (2015) have shown this assumption provides a good approximation to coalescent rates during the standing phase. The outcome during the standing phase is thus determined by competing Poisson processes. The two haplotypes could coalesce, with a waiting time being exponentially distributed with rate $(1+F)/(2Np_0)$. Alternatively, one of the two haplotypes could recombine onto the ancestral background with mean waiting time $2r_{eff}(1-p_0)$. For two competing exponential distributions with rates λ_1 and λ_2, the probability of the first event occurring given an event happens equals $\lambda_1/(\lambda_1+\lambda_2)$ (Wakeley 2009).

Hence the probability that recombination occurs instead of coalescence equals:

$$P_{R,Sd} = \frac{2r_{eff}(1-p_0)}{\frac{1+F}{2Np_0} + 2r_{eff}(1-p_0)} = \frac{2R(1-2F+\Phi)p_0(1-p_0)/(1+F)}{1+2R(1-2F+\Phi)p_0(1-p_0)/(1+F)} \tag{7}$$

The probability of coalescence rather than recombination is one minus $P_{R,Sd}$. Here $R = 2Nr$ is the population-scaled recombination rate. The $(1-2F+\Phi)/(1+F)$ term in Equation 7 approximates to $1-\sigma$ if $\Phi \approx F$; this approximation holds unless both self-fertilisation and recombination are high. This term reflects how increased homozygosity reduces both effective recombination and N_e; with the latter making coalescence more likely. Sweeps from standing variation are characterised by recombination during this standing phase, so different background haplotypes go to fixation alongside the selected mutation. Equation 7 shows that for a fixed recombination rate R, haplotypes are more likely to coalesce with increased self-fertilisation, limiting the creation of different background haplotypes. However, the same coalescent probability can be recovered by increasing the recombination distance by a factor $\sim 1/(1-\sigma)$, so the various background haplotypes can be captured by a genetic scan if a longer genetic region is analysed.

Effective starting frequency for a de novo mutation

When a new beneficial mutation appears at a single copy, it is highly likely to go extinct by chance (Fisher 1922; Haldane 1927). 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 $1/(2N)$ (Maynard Smith 1976; Barton 1998; Desai and Fisher 2007; Martin and Lambert 2015). Correcting for this acceleration is important to accurately model hard sweep signatures, and inform on the minimum level of standing variation needed to differentiate a hard sweep from one originating from standing variation.

In Section B of Supplementary File S1, we determine that hard sweeps that go to fixation have the following effective starting frequency:

$$p_{0,A} = \frac{1+F}{4NsH_l} \tag{8}$$

where $H_l = F + h - Fh$ is the effective dominance coefficient for mutations at a low frequency. This result is consistent with those of Martin and Lambert (2015), who obtained a distribution of effective starting frequencies using stochastic differential equations. This acceleration effect can create substantial increases in the apparent p_0, especially for recessive mutations. For example, with $N = 5,000$, $s = 0.05$ and $h = 0.1$, $p_{0,A} = 0.01$, an 100-fold increase above $p_0 = 1/(2N) = 0.0001$.

Effective final frequency: The effective final frequency of the derived allele, at which its spread is no longer deterministic, can be obtained by changing H_l to $H_h = 1 - h + Fh$ in Equation 8. Van Herwaarden and Van der Wal (2002) determined that the sojourn time for an allele with dominance coefficient h that is increasing in frequency, is the same for an allele decreasing in frequency with dominance $1 - h$. Glémin (2012) showed that this result also holds under any inbreeding value F (see also Charlesworth, in prep.).

Expected Pairwise Diversity

We can use P_{NE}, $P_{R,sw}$ and $P_{R,sd}$ to calculate the expected pairwise diversity (denoted π) present around a sweep. During the sweep phase, if an event were to occur then the two neutral sites can coalesce; since this event were to have occurred in the recent past then little diversity exist between them. Alternatively, one of the two samples can recombine onto the neutral background, so they will have the same pairwise diversity as the background population (π_0). If the two samples trace back to the standing phase (with probability P_{NE}) then the same logic applies. Hence the expected diversity following a sweep equals:

$$\mathbb{E}\left(\frac{\pi}{\pi_0}\right) = P_{R,sw} + (P_{NE} \cdot P_{R,sd}) \tag{9}$$

The full solution to Equation 9 can be obtained by plugging in the relevant parts from Equations 3, 5 and 7, which we evaluate numerically. An analytical approximation can be obtained by using the 'star-like' result for P_{NE} (Equation 6). In this case we are interested in calculating the probability of coalescence during the standing phase $P_{C,sd} = 1 - P_{R,sd}$, and the expected pairwise diversity approximates to:

$$\begin{aligned}\mathbb{E}_{SL}\left(\frac{\pi}{\pi_0}\right) &= 1 - (P_{NE} \cdot P_{C,sd}) \\ &= 1 - \left[\frac{1}{1 + 2R(1 - 2F + \Phi)p_0(1 - p_0)/(1 + F)}\right] \cdot \left[1 + \frac{H_l}{H_h}\left(\frac{1}{p_0} - 1\right)\right]^{-2r(1-2F+\Phi)/(H_l s)}\end{aligned} \tag{10}$$

Equation 10 reflects similar formulas for diversity following soft sweeps in hap-

loid outcrossing populations (Pennings and Hermisson 2006b; Berg and Coop 2015). There is a factor of two in the power term to account for two lineages. Note that both Equations 9 and 10 are undefined for $h = 0$ or 1 with $\sigma = 0$; these cases can be derived separately.

Figure 2 plots Equation 9 with different dominance, self-fertilisation, and standing frequency values. The analytical solution fits well compared to forward-in-time simulations, yet slightly overestimates them for high self-fertilisation frequencies. Under complete outcrossing, baseline diversity is restored (i.e., $\mathbb{E}(\pi/\pi_0)$ goes to 1) closer to the sweep origin for recessive mutations ($h = 0.1$), compared to semidominant ($h = 0.5$) or dominant ($h = 0.9$) mutations. Dominant and semidominant mutations produce similar reductions in genetic diversity, so these cases may be hard to differentiate from diversity data alone.

These patterns can be understood by examining the underlying allele trajectories, using logic described by Teshima and Przeworski (2006) (Figure 3). For outcrossing populations, recessive mutations spend most of the sojourn time at low frequencies, maximising recombination events and restoring neutral variation. These trajectories mimic 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 high frequencies, reducing the chance for neutral markers to recombine onto the ancestral background.

As self-fertilisation increases, sweep signatures become similar to the co-dominant case as the derived allele is more likely to spread as a homozygote, weakening the influence that dominance exerts over beneficial allele trajectories. Increasing p_0 also causes sweeps with different dominance coefficients to produce comparable signatures, as beneficial mutation trajectories become similar after conditioning

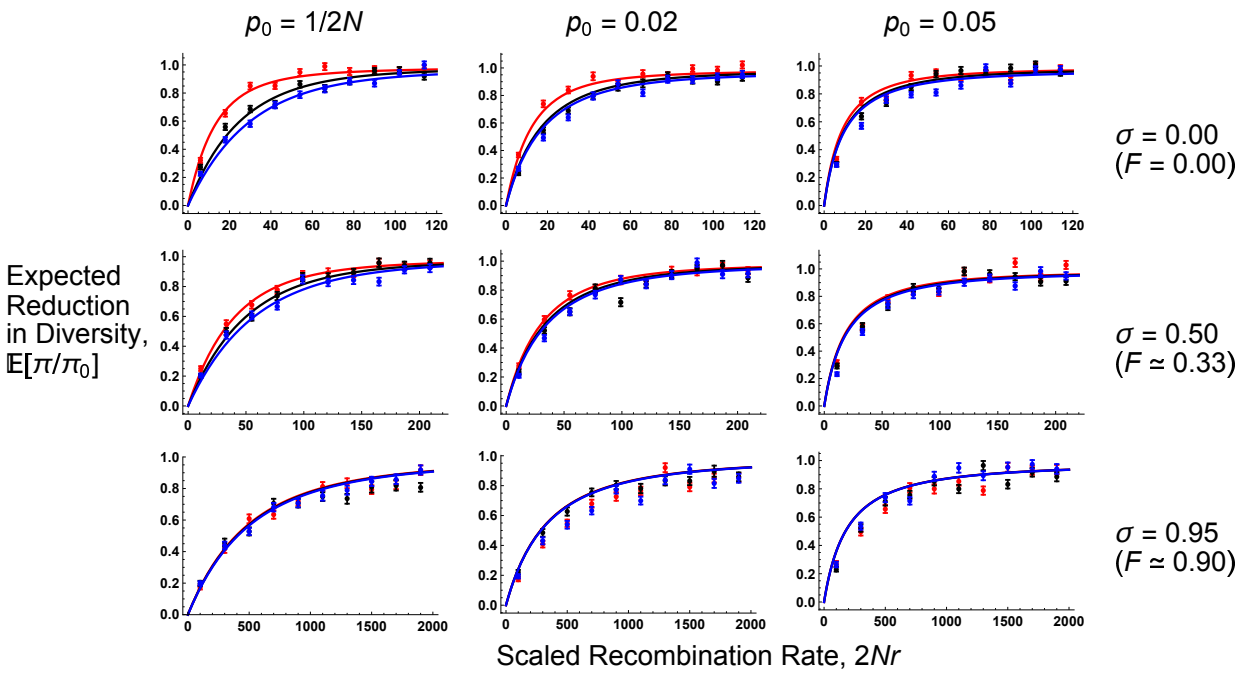

Figure 2. Expected pairwise diversity following a selective sweep. Plots of $\mathbb{E}(\pi/\pi_0)$ as a function of the recombination rate scaled to population size $2Nr$. Lines are analytical solutions (Equation 9), points are forward-in-time simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$ (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). Values of p_0 and self-fertilisation rates σ used are shown for the relevant row and column; note the $x-$axis range changes with the self-fertilisation rate. For $p_0 = 1/2N$ we use $p_{0,A}$ in our model, as given by Equation 8. Further results are plotted in Section C of Supplementary File S1.

on starting at an elevated frequency.

In Supplementary File S2, we show that the star-like approximation (Equation 10) systematically predicts higher diversity levels compared to the full solution (Equation 9). This is because Equation 10 assumes that no coalescence occurs during the sweep phase, which is only accurate for very strongly selected mutations (Charlesworth, in prep.). We also compare forward-in-time simulations to coalescent simulations for the outcrossing case. Both methods yield similar results,

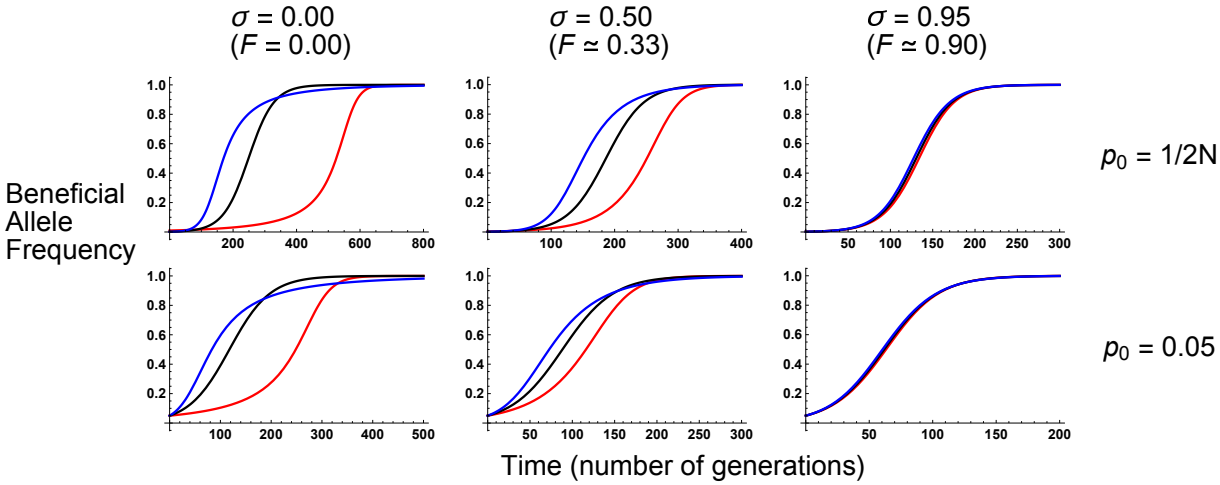

Figure 3. Beneficial allele trajectories. These were obtained by numerically evaluating the negative of Equation 4 forward in time. $N = 5,000$, $s = 0.05$, and h equals either 0.1 (red lines), 0.5 (black lines), or 0.9 (blue lines). Values of p_0 and self-fertilisation rates σ used are shown for the relevant row and column. Note the different $x-$axis scales used in each panel. Further results are plotted in Section C of Supplementary File S1.

although forward-in-time simulations produce slightly higher diversity estimates compared to coalescent simulations.

Site Frequency Spectrum

The star-like approximation can be used to obtain analytical solutions for the number of segregating sites and the site frequency spectrum (i.e., the probability that $l = 1, 2 \ldots n-1$ of n alleles carry derived variants). The full derivation for these statistics are outlined in Supplementary File S2. Figure 4 plots the SFS (Equation A11 in Supplementary File S2) alongside simulation results. Analytical results fit the simulation data well after including an inflated singleton class to account for new mutations that occur during the sweep phase (Berg and Coop

2015). There is a tendency for it to underestimate the proportion of low- and high-frequency classes ($l = 1$ and 9 in Figure 4), and overestimate the proportion of intermediate-frequency classes. Hard sweeps in either outcrossers or partial selfers are characterised by a large number of singletons or highly-derived variants (Figure 4), which is a typical selective sweep signature (Braverman *et al.* 1995; Kim and Stephan 2002). As the initial frequency p_0 increases, so does the number of intermediate-frequency variants (Figure 4). This signature is often seen as a characteristic of soft sweeps (Pennings and Hermisson 2006b; Berg and Coop 2015). Recessive hard sweeps ($h = 0.1$ and $p_0 = 1/2N$) can produce SFS profiles that are similar to sweeps from standing variation, as there are an increased number of recombination events occurring since the allele is at a low frequency for long time periods (Figure 3). With increased self-fertilisation, both hard and soft sweep signatures (e.g., increased number of intermediate-frequency alleles) are recovered when measuring the SFS at a longer recombination distance than in outcrossers (Figure 4, bottom row).

Soft sweeps from recurrent mutation

So far, 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. During the sweep phase, markers linked to the derived background can not only change state by recombination, but also via mutation. If the derived allele is at frequency p then the probability of a mutation event is $2\mu_b(1-p)/p$, for μ_b the mutation probability (Pennings and

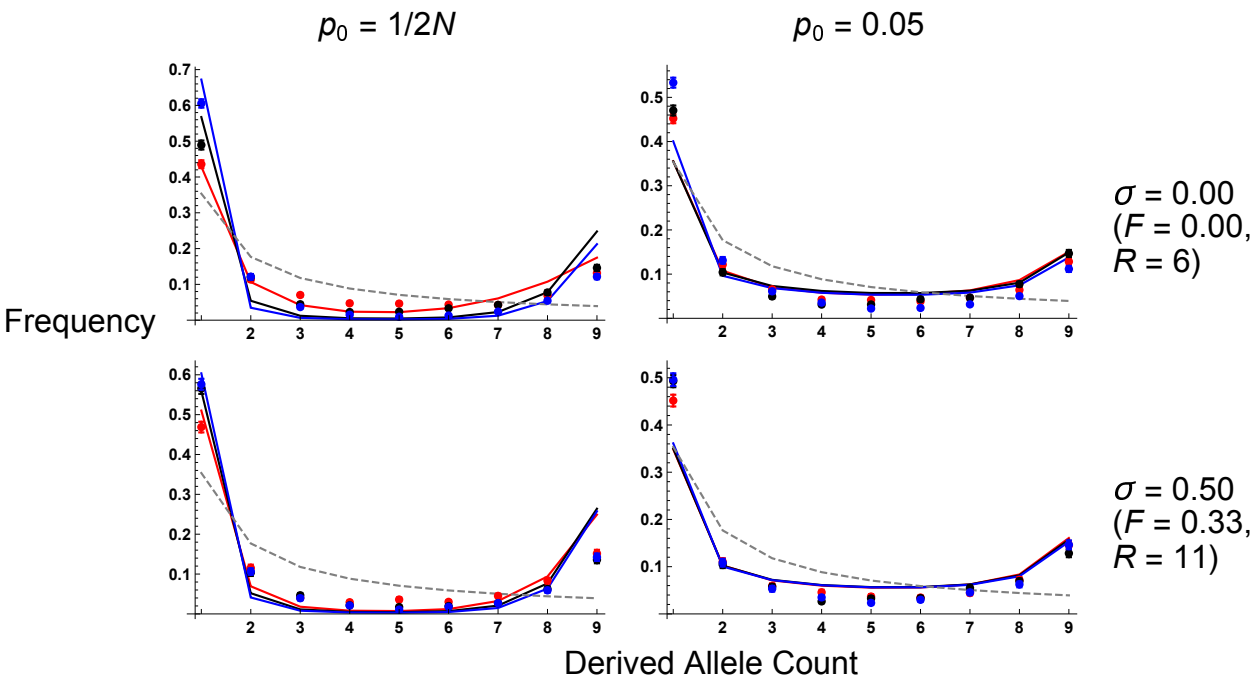

Figure 4. Expected site frequency spectrum, in flanking regions to the adaptive mutation, following a selective sweep. Lines are analytical solutions (Equation A11 in Supplementary File S2), points are simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$, and dominance coefficient $h = 0.1$ (red lines, points), 0.5 (black lines, points), or 0.9 (blue lines, points). The neutral SFS is also included for comparisons (grey dashed line). Values of p_0, self-fertilisation rates σ and recombination distances R are shown for the relevant row and column. Results for other recombination distances are in Section E of Supplementary File S1.

Hermisson 2006b). In this case the expected reduction in diversity now equals:

$$\mathbb{E}\left(\frac{\pi_M}{\pi_0}\right) = P_{R,sw} + P_{M,sw} + (P_{NE} \cdot P_{M,sd}) \tag{11}$$

where $P_{R,sw}$, P_{NE} are modified to include mutation arising during the sweep phase:

$$P_{R,Sw} \approx \int_{p'=1-\epsilon}^{p_0} \frac{P_{R,p'}}{\mathrm{d}p'/\mathrm{d}t} \mathrm{d}p' \tag{12}$$

where:

$$P_{R,p'} = \exp\left[-\int_{p=1-\epsilon}^{p} \frac{\frac{1+F}{2Np} + 2r(1-2F+\Phi)p + \frac{2\mu_b(1-p)}{p}}{\mathrm{d}p/\mathrm{d}t} \mathrm{d}p \right] \cdot [2r(1-2F+\Phi)(1-p')] \tag{13}$$

and:

$$P_{NE} \approx \exp\left(-\int_{p=1-\epsilon}^{p_{0,A}} \left[\frac{\frac{1+F}{2Np} + 2r(1-2F+\Phi)p + \frac{2\mu_b(1-p)}{p}}{\mathrm{d}p/\mathrm{d}t} \right] \mathrm{d}p \right) \tag{14}$$

$P_{M,sw}$ is the mutation probability during the sweep phase, and is similar to Equation 12 except that $2r(1-2F+\Phi)(1-p')$ is replaced by $2\mu_b(1-p')/p'$, for p' is the derived allele frequency when the event occurs. $P_{M,sd}$ is the probability that, at the sweep origin, the derived allele appears by mutation instead of coalescing, and is defined in a similar manner to $P_{R,sd}$ (Equation 7):

$$P_{M,Sd} = \frac{\frac{2\mu_b(1-p_{0,A})}{p_{0,A}}}{\frac{1+F}{2Np_{0,A}} + \frac{2\mu_b(1-p_{0,A})}{p_{0,A}}} = \frac{2\Theta_b(1-p_{0,A})}{1+F+2\Theta_b(1-p_{0,A})} \tag{15}$$

where $\Theta_b = 2N\mu_b$. The coalescence probability is one minus $P_{M,Sd}$. We can also obtain an analytical solution using a 'star-like' approximation, which assumes that

no coalescence or mutation events occur during the sweep phase:

$$\mathbb{E}_{SL}\left(\frac{\pi_M}{\pi_0}\right) = 1 - \left[\frac{1}{1+2(1-p_{0,A})\Theta_b/(1+F)}\right] \cdot \left[1 + \frac{H_l}{H_h}\left(\frac{1}{p_{0,A}} - 1\right)\right]^{-2r(1-2F+\Phi)/(H_l s)} \tag{16}$$

Figure 5 compares $\mathbb{E}(\pi/\pi_0)$ in the standing variation case, and for the recurrent mutation case, under different levels of self-fertilisation. While dominance only weakly affects sweep signatures arising from standing variation under outcrossing, it more strongly affects sweeps from recurrent mutation in outcrossing populations, as each variant arises from an initial frequency close to $1/(2N)$ (Figure 3). Second, the two models exhibit different behaviour close to the selected locus (R close to zero). The recurrent mutation model has diversity levels that are greater than zero, while the standing variation model exhibits little diversity. As R increases, diversity reaches higher levels in the standing variation case than for the recurrent mutation case. Assuming weak recombination (so that $P_{NE} \approx 1$), the recombination rate at which a sweep from recurrent mutation yields higher diversity than one from standing variation is given when the coalescence probability is higher for the mutation model than that for the standing variation case. This change occurs at:

$$R \leq R_{Lim} = \frac{\Theta_b}{p_0(1-2F+\Phi)} \tag{17}$$

$$\approx \frac{\Theta_b}{p_0(1-F)} \tag{18}$$

The last approximation arises as $\Phi \approx F$ unless F is close to one, and recombi-

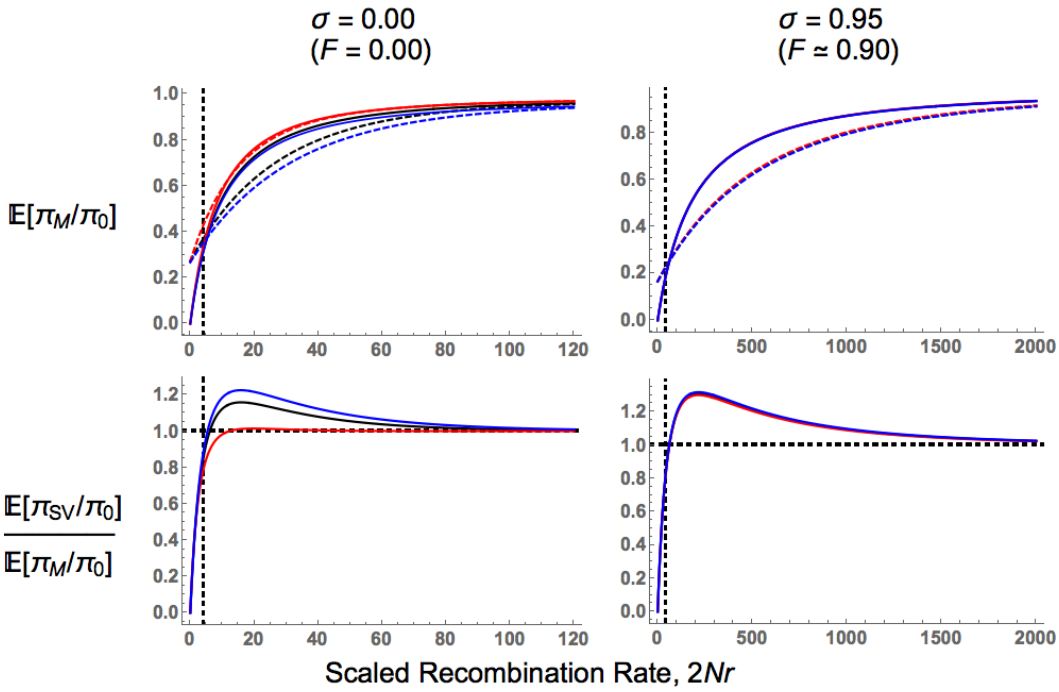

Figure 5. Comparing sweeps from recurrent mutation to those from standing variation. Top row: comparing the reduction in diversity following a soft sweep (Equation 11), from either standing variation ($p_0 = 0.05$, solid lines) or recurrent mutation (using $P_{coal,M}$ with $\Theta_b = 0.2$, dashed lines). $N = 5,000$, $s = 0.05$, and dominance coefficient $h = 0.1$ (red lines), 0.5 (black lines), or 0.9 (blue lines). Bottom row: 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 the respective plot for the top row. Vertical dashed black line indicates R_{Lim} (Equation 18), the predicted recombination rate where $\pi_{SV}/\pi_M = 1$ (horizontal dashed line in the bottom-row plots). Note the different $x-$axis lengths between panels. Results are also plotted in Section F of Supplementary File S1.

nation rates are high (r approximately greater than 0.1). Hence for a fixed Θ_b, the window where recurrent mutations create higher diversity near the selected locus increases for lower p_0 or higher F, since both these factors reduces the potential for recombination to create new haplotypes during the standing phase. Equation 18 is generally accurate when sweeps from standing variation have higher diversity than

sweeps with recurrent mutations (Figure 5, bottom row), but becomes inaccurate for $h = 0.1$ in outcrossing populations, as some events are likely to occur during the sweep phase. In Section B in Supplementary File S2 we show how similar results apply to the SFS.

Application to selective sweep signatures

To demonstrate how these sweep models can be used to infer properties of genetic adaptation, we use the analytical solutions to examine selective sweep signatures in two different organisms. We examine a selective sweep at the *SLC24A5* gene in human populations, and one in the self-fertilising plant *Arabidopsis thaliana*. In both cases we infer the selection strength acting at the target site by maximum likelihood. For the human sweep we also test for a sweep from standing variation, and if there's evidence for semi-dominance ($h = 1/2$) against general dominance. In both cases we fit models to diversity data, using estimates at the flanking regions to account for baseline diversity in the absence of the sweep (thus considering the effects of e.g., local background selection, demographic history). We use published recombination maps to model inferred meiotic recombination events (both crossovers and gene conversion) over the sweep region.

The *SLC24A5* gene in humans

The rs1426654 SNP is a G $\rightarrow$ A substitution that is strongly associated with skin pigmentation in Eurasian populations (Lamason *et al.* 2005; Crawford *et al.* 2017). 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 (Crawford *et al.* 2017). 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 (see also Lin *et al.* (2018)). These data raise the question of whether the derived SNP was introduced into Eurasia at an elevated frequency or not. Hence we investigated whether the sweep signature is consistent with a hard sweep or one from standing variation, using data from the 1000 Genomes project. Preliminary calculations led to unrealistic parameters for s, p_0 being chosen, that would have meant the beneficial mutation appeared and fixed very recently. To ensure that unrealistic parameters were not given strong statistical support, we also included a likelihood term to consider the allele age, so parameters that implied unrealistically young or old sweeps were not given statistical support (see Methods and Supplementary Files S3, S4 for details).

Results are outlined in Table 2, and plotted in Figure 6(a), (b). We fit models over two different lengths; one looking at a 400Kb region up- and downstream from the substituted sites, and at 600Kb up- and downstream from the substitution. The model with the highest statistical support, as determined by $AICc$, depends on the length analysed. For the 400Kb case, the best-fitting model was a hard semi-dominant sweep, with a selection coefficient $s = 0.059$. For the 600Kb case, the best-fitting model was instead a strong dominant sweep from standing variation ($s = 0.141$, $h = 0.95$, $p_0 = 0.0108$). Both selection estimates are similar to those obtained by Beleza *et al.* (2012).

This example shows how different sweep models can yield similar signatures in diversity data. The similarity between models is reflected in the small $\Delta AICc$ between models, and also in the Akaike weights $\mathcal{W}_i$, which can be interpreted as

Analysis 400Kb up, downstream of *SLC24A5*				
Parameter	**HS, SD**	HS, GD	SV, SD	SV, GD
s	**0.0592**	0.0584	0.0590	0.138
h (1/2)	–	0.653	–	0.950
$p_0(p_{0,A})$	–	–	0.000844	0.0110
$\Delta AICc$	**0**	1.57	2.28	1.94
$\mathcal{W}_i$	**0.464**	0.212	0.148	0.176

Analysis 600Kb up, downstream of *SLC24A5*				
s	0.0647	0.0615	0.0643	**0.141**
h (1/2)	–	0.702	–	**0.950**
$p_0(p_{0,A})$	–	–	0.000772	**0.0108**
$\Delta AICc$	1.59	1.87	3.77	**0**
$\mathcal{W}_i$	0.226	0.196	0.0761	**0.501**

Table 2. Maximum-likelihood model-fitting results for the *SLC24A5* sweep signature. Results are presented for a hard sweep model ('HS') or from standing variation ('SV'). We also consider semi-dominance ($h = 1/2$) or general dominance (denoted 'SD', 'GD' 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). $\Delta AICc$ is the difference in $AICc$ between that model and the best fitting one. $\mathcal{W}_i$ are Akaike weights for each model.

an approximate posterior probability for a given model (Burnham and Anderson 2002). The highest $\mathcal{W}_i$ equals 0.464 or 0.501, indicating that there is a sizeable probability that an alternative model underpins the *SLC24A5* signature.

A sweep in *Arabidopsis thaliana*

The model plant *Arabidopsis thaliana* is highly self-fertilising, with estimated outcrossing rates on the order of a few percent (Platt *et al.* 2010; Bomblies *et al.* 2010). Huber *et al.* (2014) found candidate selective sweep regions in samples taken from southern Sweden. One of these, at location 12.855Mb on chromosome 1, shows a

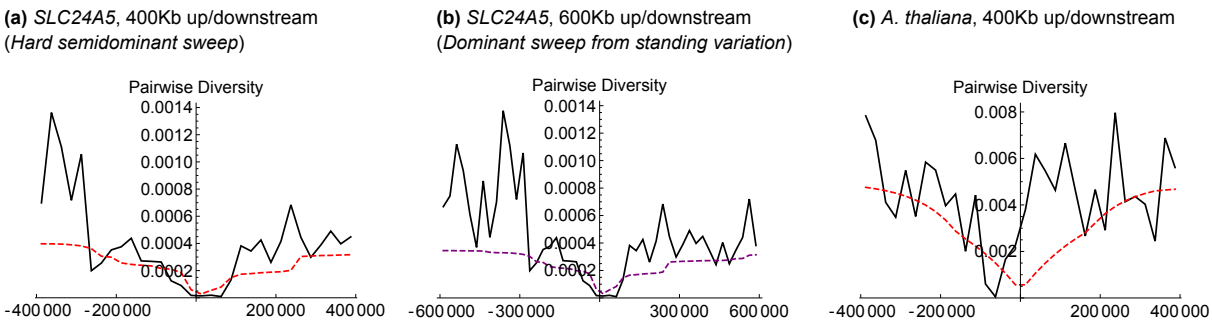

Figure 6. Selective sweep signature analyses. (a) Pairwise diversity around the *SLC24A5* sweep (black line), along with the best-fitting model (red dashed line), as a function of distance from the target substitution. Diversity was measured 400Kb up- and downstream of the target site; negative values indicate distances upstream from the sweep. (b) As (a) but measuring diversity 600Kb up- and downstream of the target site. The best-fitting model was different, as indicated by the purple dashed line. (c) Pairwise diversity around the *Arabidopsis thaliana* sweep, and the best fitting model (red dashed line). Diversity was measured 400Kb up- and downstream of the target site; results for diversity measured 600Kb up- and downstream is shown in Figure G in Supplementary File S2.

characteristic reduction in diversity. We apply our models to determine the selection coefficient underlying this sweep pattern, while correcting N_e to correct for self-fertilisation. We found during the human sweep analysis how it was necessary to correct for the expected allele age to infer realistic values of both the starting allele frequency and selective advantage. To our knowledge no estimates of allele age exist for this sweep, so we restrict our analysis to determining the selection coefficient while assuming a hard sweep.

Table 3 shows that the estimated selection coefficients are smaller than for the human sweep, either 0.00315 or 0.00462 depending on the genetic length analysed. If we infer the sweep using the N_e estimates without rescaling to include self-fertilisation, then the selection coefficients are higher (0.176 and 0.251 for 400Kb

Distance Around Sweep	Selection Coefficient	
	With Selfing	Ignoring Selfing
400Kb around sweep	0.00315	0.176
600Kb around sweep	0.00462	0.251

Table 3. Selection coefficients as inferred by maximum-likelihood model fitting to the *A. thaliana* sweep. Results are presented over two different lengths, either including selfing effects or not.

and 600Kb cases respectively). These results demonstrate how not including the reproductive mode can lead to overestimates of the selection coefficients underlying sweep signatures. The best-fitting sweep is plotted in Figure 6(c) for the 400Kb case (the 600Kb case is shown in Figure G of the Supplementary File S2).

Discussion

Summary of Theoretical Findings

While there has been many investigations into how different sweep processes can be detected from next-generation sequence data (Pritchard and Di Rienzo 2010; Messer and Petrov 2013; Stephan 2016; Hermisson and Pennings 2017), these models generally assumed idealised randomly mating populations and beneficial mutations that are semidominant ($h = 0.5$). Here we have created a more general selective sweep model, with arbitrary self-fertilisation and dominance levels. 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 also consider the case of recurrent mutation (Figure 5).

We find that the qualitative patterns of different selective sweeps under selfing

remain similar to expectations from 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 (Figures 4, 5). This pattern is standard for soft sweeps (Pennings and Hermisson 2006b; Messer and Petrov 2013; Berg and Coop 2015; Hermisson and Pennings 2017) so existing statistical methods for detecting them (e.g., observing an higher than expected number of haplotypes; Vitti *et al.* (2013); Garud *et al.* (2015)) can, in principle, also be applied to selfing organisms. 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 in different ways. During the sweep phase, beneficial alleles fix more rapidly under higher self-fertilisation as homozygous mutations are created more rapidly (Charlesworth 1992; Glémin 2012). In addition, the effective recombination rate is reduced by approximately $1 - F$ (Nordborg *et al.* 1996; Nordborg 2000; Charlesworth and Charlesworth 2010), and slightly further for highly inbred populations (Roze 2009, 2016). 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 Equation 3 for P_{NE}. During the standing phase, two haplotypes are more likely to coalesce under high levels of self-fertilisation since N_e is decreased by a factor $1/(1+F)$ (Pollak 1987; Charlesworth 1992; Caballero and Hill 1992; Nordborg and Donnelly 1997). This effect, combined with a reduced effective recombination rate, means that the overall recombination probability during the standing phase is reduced by a factor $\sim(1-\sigma)$ (Equations 7). 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.

We have also investigated how dominance affects soft sweep signatures, since previous analyses have only focussed on how dominance affects hard sweeps (Teshima and Przeworski 2006; Teshima *et al.* 2006; Ewing *et al.* 2011). 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 (Figures 2, 3). With increased self-fertilisation, dominance has a weaker impact on sweep signatures as most mutations are homozygous (Figure 3). We also show that the SFS for recessive alleles can resemble a soft sweep, with a higher number of intermediate-frequency variants than for other hard sweeps (Figure 4). Dominance only weakly affects sweeps from standing variation, as trajectories of beneficial alleles become similar once the variant's initial frequency exceeds $1/(2N)$ (Figures 2, 3). Yet different dominance levels can affect sweep signatures if the beneficial allele is reintroduced by recurrent mutation (Figure 5). Hence if one wishes to understand how dominance affects sweep signatures, it is also important to consider which processes underlies observed patterns of genetic diversity.

Soft sweeps from recurrent mutation or standing variation?

These theoretical results shed light onto how to distinguish between soft sweeps that arise either from standing variation, or from recurrent mutation. Both models are characterised by an elevated number of intermediate-frequency variants, in comparison to a hard sweep. Yet sweeps arising from recurrent mutation produce intermediate-frequency variants closer to the beneficial locus, compared to sweeps from standing variation (Figure 5). Equation 18 provides a simple condition for

R_{Lim}, the recombination distance needed for a sweep from standing variation to exhibit higher diversity than one from recurrent mutation; the size of this region increases under higher self-fertilisation.

Differences in 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, when investigating sweep patterns over broad 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 (Kim and Stephan 2003). In addition, competing selective sweeps can drag different sets of neutral variation to fixation, creating asymmetric reductions in diversity around a substitution (Chevin *et al.* 2008). Further investigations of selective sweep patterns across long genetic distances will prove to be a rich area of future research.

Using models to analyse empirical selective sweep signatures

We demonstrated how our more general sweep models can be used to determine adaptation properties in two cases. First we reanalysed diversity around the *SLC24A5* gene in European humans. We found that the sweep pattern either best matches a hard semidominant sweep, or a strongly-selected dominant sweep that had a starting selected frequency of $\sim$1%. However, it is unclear which scenario is the most realistic, since the most strongly supported model depends on

the length of the analysed region. Given that the derived SNP is also present in Africa at a low frequency, then this ambiguity could have arisen if the derived allele was introduced into Eurasia at an elevated frequency that was close to the effective starting frequency of a hard sweep (Equation 8). These analyses demonstrate how different sweep scenarios can yield similar sweep outcomes, creating additional hypotheses for the evolutionary history of an adaptive mutation. We also determined that knowing the demographic history is useful if wishing to estimate the starting frequency of a sweep from standing variation. Higher p_0 values mean that a larger selection coefficient is needed to produce a sweep with the same reduction in diversity, which can lead to unrealistically high p_0 and s values being estimated. We correct for this problem by calculating the allele age associated with these parameter estimates, penalising estimates that lead to overly old or young sweeps based on demographic history.

We also analyse a sweep in the highly self-fertilising plant *A. thaliana*, estimating that a sweep on chromosome 1 had more modest selection coefficients (0.32–0.46%). These estimates are slightly higher but of the same order as those estimated by Huber *et al.* (2014) for sweeps affecting Southern Swedish populations of *A. thaliana*. We also demonstrate that estimates of s would be greatly inflated if one does not account for how self-fertilisation affects N_e, showing the importance of correcting for the mating system when inferring sweep properties.

These models assumed a fixed population size, but many populations have sizes that have changed over time. European humans have likely undergone a bottleneck followed by an expansion (Gutenkunst *et al.* 2009). To test if this demography could have drastically affected our results, we simulate a hard sweep under a growth-bottleneck population but with the same scaled $2N_es$ value as inferred for

a hard sweep in the 400Kb case (Table 2). Both the absolute and relative diversity closely matches that from the *SLC24A5* sweep (Figure H in Supplementary File S2). This example suggests that estimates of hard sweep properties from diversity data are robust to population demography, if they are scaled by the present-day effective population size.

Potential applications to self-fertilising organisms

Existing methods for finding sweep signatures in nucleotide polymorphism data are commonly based on finding regions with a site-frequency spectrum matching what is expected under a selective sweep (Nielsen *et al.* 2005; Boitard *et al.* 2009; Huber *et al.* 2016). The more general models developed here can be used to create more specific sweep-detection methods that include self-fertilisation. However, a recent analysis found that soft-sweep signatures can be incorrectly inferred if analysing genetic regions that flank hard sweeps, which was named the 'soft shoulder' effect (Schrider *et al.* 2015). Due to the reduction in recombination in selfers, these model results indicate that 'soft-shoulder' footprints can arise over long genetic distances and should be taken into account. 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 (Schrider and Kern 2016). These more general calculations can also be extended to quantify to what extent background selection and sweeps jointly shape genome-wide diversity in self-fertilising organisms (Elyashiv *et al.* 2016; Campos *et al.* 2017; Booker and Keightley 2018; Rettelbach *et al.* 2019).

The model fitting exercises presented here also demonstrate caveats with re-

spect to inferring adaptation properties from sweep signatures. Inference requires using other parameters (e.g., population size, self-fertilisation rate, generation time, pairwise diversity) that have a broad range of estimates; a more thorough analysis should take this uncertainty into account. We also found that estimates of the *SLC24A5* sweep from standing variation gave unrealistic values, unless we also corrected for the expected allele age. Since very recent demographic history was unavailable for *A. thaliana*, we did not try and determine whether that sweep arose from standing variation. A greater knowledge of expected allele ages for sweeps in southern Sweden would be useful in determining whether they arose from standing variation. Alternatively, future analysis of sweep properties would benefit from considering multiple summary statistics alongside population diversity (e.g., the site-frequency spectrum, haplotype structure). Machine-learning methods are currently proving popular to classify whether sweeps are hard or soft by using information from multiple summary statistics (Schrider and Kern 2018; Sheehan and Song 2016). It would be useful to see if these methods can also quantify properties of selective sweeps.

Methods

Simulation Procedures

Forward-in-time simulations

Simulations were ran using SLiM version 3.2 (Haller and Messer 2019), with simulation scripts available from GitHub (`https://github.com/MattHartfield/SweepDomSelf`) and Supplementary File S5. There exists N diploid individuals of

length 100,000 nucleotides. The target locus carrying the derived allele is present at the left-hand end of the haplotype, while the remaining loci can carry neutral mutations only. Selected alleles have a homozygous selective advantage s and dominance coefficient h. Mutation and recombination parameters are input as population-level rates, which are subsequently scaled down to obtain per-locus mutation rates, or a per-inter-base-pair recombination probability.

A 'burn-in' phase is first ran to generate background neutral diversity, where the population evolves without any beneficial alleles present for $20N$ generations; this population was subsequently saved. The second phase acts differently depending on whether the beneficial allele is instantly selected for (a hard sweep), or whether it went through a neutral phase. If a hard sweep was simulated, then the beneficial allele was introduced into a single individual as a heterozygote, and tracked until it is fixed or lost. If the latter, the burn-in population is reloaded, the random seed changed and the beneficial mutation reintroduced. The procedure is repeated until the mutation has fixed.

If the derived mutation was initially neutral, then following the burn-in a neutral allele is introduced into a random individual as a heterozygote, and tracked until it is lost or it reaches a frequency p_0. If it is lost then the burn-in population is reloaded, the random seed changed and the derived allele is reintroduced. If the mutation reaches the target frequency p_0 then it is then converted into a selected mutation, and is tracked until fixation or loss. If the beneficial mutation is subsequently lost then the simulation is stopped and restarted from scratch (i.e., the burn-in population is also regenerated).

100 burn-in populations were generated for each parameter set. After the beneficial allele has gone to fixation, we sampled 10 haplotypes 10 times from each

burn-in population to create 1,000 replicate simulations. Mutations are placed in one of 10 bins depending on the distance from the sweep. Relevant statistics (pairwise diversity, relative to neutral expectations; number of segregating sites; site frequency spectrum) were calculated per bin. Mean values are calculated over all 1,000 outputs. 95% confidence intervals were calculated by bootstrapping the data 1,000 times.

Data analyses

Model fitting approach

Sweep models were fitted to diversity data using a variant of the the maximum likelihood procedure of Sattath *et al.* (2011). The expected reduction in diversity (Equation 9) is equal to the probability that recombination acts on a lineage during the sweep. Hence the probability that recombination acts to restore background diversity is drawn from a Bernoulli process with probability $\mathbb{E}(\pi/\pi_0)$, meaning that expected diversity equals $\pi_0\mathbb{E}(\pi/\pi_0)$ and the variance is $(\pi_0)^2\mathbb{E}(\pi/\pi_0)(1-\mathbb{E}(\pi/\pi_0))$. Thus for diversity estimates π_i in bin i, we calculate the normalised values $Z_i = (\pi_i - \pi_0\mathbb{E}(\pi/\pi_0)) \,/\, \left(\sqrt{(\pi_0)^2\mathbb{E}(\pi/\pi_0)(1-\mathbb{E}(\pi/\pi_0))}\right)$ for parameter estimates underlying $\mathbb{E}(\pi/\pi_0)$ (Equation 9). The likelihood of individual Z_i values $\mathcal{L}_i$ were obtained from a normal distribution; the log-likelihood of the model fit is $\ell_M = \sum_i \log(\mathcal{L}_i)$.

For the human sweep analyses, preliminary model fitting for the standing-variation case inferred high initial frequencies and selection coefficients, implying that the beneficial allele was unrealistically young (arising during the last few thousand years). Hence we also factored in the allele age to penalise unrealis-

tic values. For given parameter estimates, we calculated the fixation time using Glémin (2012, Equation 21) assuming 28 years per generation (Beleza *et al.* 2012). Beleza *et al.* (2012) argued that *SLC24A5* and other pigmentation genes emerged around 11,000 to 19,000 years ago, coinciding with population spread after the last glacial maximum. Hence after calculating the allele age for a certain parameter set, it was normalised by assuming a mean time of 15,000 years and standard deviation of $4000/(\Phi^{-1}(0.975))$ for Φ the inverse of the normal quantile function (so 95% of the normal distribution lied between 11,000 and 19,000 years). The likelihood of the allele age ℓ_τ is equal to the natural log of the likelihood as obtained from a standard normal distribution. The total likelihood is the sum of this likelihood with the model-fitting likelihood, i.e., $\ell_T = \ell_\tau + \ell_M$. We visually checked the normality of each model fit by inspecting QQ plots of the residuals (Supplementary Files S3, S4).

Human sweep data processing and parametrisation

Data was retrieved from the 1000 Genomes phase 3 version 3 integrated call set (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) (The 1000 Genomes Project Consortium 2015). The five European populations (CEU, FIN, GBR, IBN, TSI) were investigated, and related individuals were removed (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/20140625_related_individuals.txt`) giving 503 total individuals. *VCFtools* (Danecek *et al.* 2011) was used to create an initial file of variants around the target SNP, by retaining biallelic sites in Hardy-Weinberg equilibrium (with $P-$value greater than 10^{-6}) and removing indels between locations 46,926,485 and 49,926,484 on chromosome 15 (the rs1426654 target SNP is at location 48,426,484). Pairwise

diversity was then calculated in 25Kb windows around the SNP using the 'scikit-allel' package (available from `https://zenodo.org/record/822784`). Diversity was measured either between positions 48,026,485 – 48,826,484 (i.e., 400Kb up- and downstream of the target SNP), or between positions 47,826,485 – 49,026,484 (i.e., 600Kb up- and downstream of the target SNP). Linkage disequilibrium (as measured by r^2) decays to $\sim$0.1 for European populations over this bin size (The 1000 Genomes Project Consortium 2015). Baseline diversity (i.e., that expected in the absence of a selective sweep) was determined by calculating mean diversity values in the first and last 7 windows of each diversity dataset. Specifically, for the 400Kb dataset, the upstream diversity was measured between locations 48,026,485 – 48,176,484 and the downstream diversity was measured between locations 48,651,485 – 48,826,484. For the 600Kb dataset, the upstream diversity was measured between locations 47,826,485 – 48,001,484 and the upstream diversity was measured between locations 48,851,485 – 49,026,484. Sex-averaged recombination maps for each bin were obtained from Bhérer *et al.* (2017). The raw data is shown in Figure I of Supplementary File S2.

A nested model approach was implemented to determine if there was evidence for dominance that differed from $h = 1/2$, or an elevated starting frequency. Since we analysed human data assuming a fixed population size, we set $F = 0$ and $N_e = 10,000$ (Jorde *et al.* 1998). We compared four models: (i) fixed $h = 1/2$, $p_0 = p_{0,A}$ (hard sweep with semidominance); (ii) variable h, fixed $p_0 = p_{0,A}$ (hard sweep with general dominance); (iii) fixed $h = 1/2$, variable p_0 (standing variation sweep with semidominance); (iv) variable h, p_0 (standing variation sweep with general dominance).

The maximum likelihood parameter set for each model was found using the

'FindMaximum' function in *Mathematica* version 11.2 (Wolfram Research, Inc. 2017). We bounded s between 0.01 and 0.2, with a starting value of 0.02. If h was estimated then the starting value was 0.5 and search values were bounded between 0.05 and 0.95 as analytical solutions break down for extreme h values. If p_0 was estimated then the starting value was set to 0.001, with the search bounded between 0.000844 – 0.5 for the 400Kb dataset, and 0.000772 – 0.5 for the 600Kb dataset. The lower bound in both cases was set to $p_{0,A}$ (Equation 8) with $s = 0.0592$ and $s = 0.0647$ respectively, i.e., the selection coefficients as determined for the hard sweep. The upper bound of 0.5 was set as very high p_0 values would not cause sweep-like patterns.

For each model the goodness of fit was determined using the corrected Akaike Information Criterion ($AICc$), with lower values indicating stronger statistical support:

$$AICc = 2K - 2\ell_T + \frac{2K(K+1)}{n-K-1} \tag{19}$$

Here, K is the number of parameters to be estimated and n the number of datapoints (32 for the 400Kb dataset and 48 for the 600Kb dataset). We also calculated the difference between the $AICc$ for model i and the lowest $AICc$ value ($\Delta AICc_i$), and Akaike weights $\mathcal{W}_i$:

$$\mathcal{W}_i = \frac{\exp(-\frac{1}{2}\Delta AICc_i)}{\sum_i \exp(-\frac{1}{2}\Delta AICc_i)} \tag{20}$$

Arabidopsis thaliana sweep data processing and parametrisation

Data was obtained from the 1001 genomes database (`http://1001genomes.org/`). The '1001 Admixture Map' tool (`http://1001genomes.github.io/admixture-map/`) was used to see and obtain individuals from Southern Sweden that were less than ∼2km apart, as Huber *et al.* (2014) found that samples from within this distance were likely to be clones of each other. 48 individuals remained; the 'VCF subset' tool on the 1001 genomes website (`http://tools.1001genomes.org/vcfsubset/`) was used to download genome data from chromosome 1 between positions 1 and 30,415,258. The dataset was filtered to remove indels and non-biallelic sites. To filter out over- and under-assembled regions, we first obtained the mean depth per site from this reduced dataset using *VCFTools* (using the `--site-mean-depth` option). We then kept polymorphic sites whose mean depth lay within the 99.5% quantiles of a Poisson distribution, with the same mean as the distribution of depth results (mean = 27.1; 99.5% quantiles = 14–43).

Pairwise diversity was then calculated in 25Kb windows around the SNP, either between positions 12,455,001 – 13,255,000 (i.e., 400Kb up- and downstream of the sweep region), or between positions 12,255,001 – 13,455,000 (i.e., 600Kb up- and downstream of the target SNP), using the 'scikit-allel' package. Mean r^2 for Southern Swedish populations decays to ∼0.2 over this bin size (Long *et al.* 2013). Baseline diversity was determined by calculating mean diversity values in the first and last 7 windows of each diversity dataset. Hence for the 400Kb dataset, the upstream diversity was measured between locations 12,455,001 – 12,630,000, and the upstream diversity was measured between locations 13,080,001 – 13,255,000. For the 600Kb dataset, the upstream diversity was measured between locations

12,255,001 – 12,430,000 and the upstream diversity was measured between locations 13,280,001 – 13,455,000. Raw recombination maps for each bin were obtained from cross P10 of Salome *et al.* (2012), as used by Fulgione *et al.* (2018). The raw data is shown in Figure J of Supplementary File S2.

For the model analysis we assumed a self-fertilisation fraction of 0.97 (Platt *et al.* 2010), giving $F \approx 0.94$, although we note that estimates of the selfing rate can vary among geographic regions (Platt *et al.* 2010; Bomblies *et al.* 2010). We fixed $N_e = 194,800$, which was the population size estimated by Huber *et al.* (2014) for southern Sweden. This value is higher than estimates obtained by Gossmann *et al.* (2010) but lower than obtained by Cao *et al.* (2011). Our model requires the use of $N = N_e/(1+F)$ i.e. the population size prior to the effects of self-fertilisation, which is approximately equal to 485,437. We also increase the scaled recombination rate $R = 2Nr = 2N_e(1+F)r$. The estimates from Salome *et al.* (2012) are actual recombination rates as measured between outcrossed progeny. Both the rescaled N_e and raw R values are scaled by the respective inbreeding values during the model fit. We also performed model fitting without first rescaling the population size by $1+F$, to determine how it affects parameter estimation.

The selection coefficient was determined as the value that maximised ℓ_M only, assuming $h = 1/2$ and $p_0 = p_{0,A}$, with the search of s bounded between 0.0001 and 2 with an initial starting value of 0.1. We used wide bounds because of the higher estimated selection coefficients in cases where the selfing rate was not accounted for. We also used a higher starting condition since the model was seen to converge on very low estimates of s that were local maxima, even though they did not give the largest likelihood over the parameter range.

Acknowledgments. We would like to thank Sally Otto for providing information on the elevated effective starting frequency of beneficial mutations; Dan Schrider for information on how to simulate human demographic scenarios; Mikkel Schierup for feedback on the human sweep analyses; Brian Charlesworth on providing advice on modelling selective sweeps, sharing unpublished results, and providing comments on the manuscript; Christian Huber and Andrea Fulgione for answering questions on detecting sweeps in *A. thaliana*; Angela Hancock for useful discussions; and Ben Haller for answering questions about SLiM. MH was supported by a Marie Curie International Outgoing Fellowship (MC-IOF-622936) and a NERC Independent Research Fellowship (NE/R015686/1). MH and TB also acknowledge financial support from the European Research Council under the European Union's Seventh Framework Program (FP7/20072013, ERC Grant 311341).

References

Andersen, E. C., Gerke, J. P., Shapiro, J. A., Crissman, J. R., Ghosh, R., Bloom, J. S., Felix, M.-A., and Kruglyak, L. 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat. Genet.*, 44(3): 285–290, doi: 10.1038/ng.1050.

Anderson, T. J. C., Nair, S., McDew-White, M., Cheeseman, I. H., Nkhoma, S., Bilgic, F., McGready, R., Ashley, E., Pyae Phyo, A., White, N. J., and Nosten, F. 2016. Population parameters underlying an ongoing soft sweep in southeast asian malaria parasites. *Mol. Biol. Evol.*, 34(1): 131–144, doi: 10.1093/molbev/msw228.

Badouin, H., Gladieux, P., Gouzy, J., Siguenza, S., Aguileta, G., Snirc, A., Le Prieur, S., Jeziorski, C., Branca, A., and Giraud, T. 2017. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Mol. Ecol.*, 26(7): 2041–2062, doi: 10.1111/mec.13976.

Barrett, R. D. H. and Schluter, D. 2008. Adaptation from standing genetic variation. *Trends Ecol. Evol.*, 23(1): 38–44, doi: 10.1016/j.tree.2007.09.008.

Barton, N. H. 1998. The effect of hitch-hiking on neutral genealogies. *Genet. Res.*, 72: 123–133.

Beleza, S., Santos, A. M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M. D., Parra, E. J., and Rocha, J. 2012. The timing of pigmentation lightening in europeans. *Mol. Biol. Evol.*, 30(1): 24–35.

Berg, J. J. and Coop, G. 2015. A coalescent model for a sweep of a unique standing variant. *Genetics*, 201(2): 707–725, doi: 10.1534/genetics.115.178962.

Bhérer, C., Campbell, C. L., and Auton, A. 2017. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, 8: 14994, doi: 10.1038/ncomms14994.

Billiard, S., López-Villavicencio, M., Devier, B., Hood, M. E., Fairhead, C., and Giraud, T. 2011. Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. *Biol. Rev. Camb. Philos. Soc.*, 86(2): 421–442, doi: 10.1111/j.1469-185X.2010.00153.x.

Boitard, S., Schlötterer, C., and Futschik, A. 2009. Detecting selective sweeps: A

new approach based on hidden markov models. *Genetics*, 181(4): 1567–1578, doi: 10.1534/genetics.108.100032.

Bomblies, K., Yant, L., Laitinen, R. A., Kim, S.-T., Hollister, J. D., Warthmann, N., Fitz, J., and Weigel, D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet.*, 6(3): e1000890, doi: 10.1371/journal.pgen.1000890.

Bonhomme, M., Boitard, S., San Clemente, H., Dumas, B., Young, N., and Jacquet, C. 2015. Genomic signature of selective sweeps illuminates adaptation of *Medicago truncatula* to root-associated microorganisms. *Mol. Biol. Evol.*, 32(8): 2097–2110.

Booker, T. R. and Keightley, P. D. 2018. Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. *Mol. Biol. Evol.*, 35(12): 2971–2988, doi: 10.1093/molbev/msy188.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2): 783–796.

Burnham, K. P. and Anderson, D. R. 2002. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York; London, second edition.

Caballero, A. and Hill, W. G. 1992. Effects of partial inbreeding on fixation rates and variation of mutant genes. *Genetics*, 131(2): 493–507.

Campos, J. L. and Charlesworth, B. 2019. The effects on neutral variability of

recurrent selective sweeps and background selection. *Genetics*, 212(1): 287–303, doi: 10.1534/genetics.119.301951.

Campos, J. L., Zhao, L., and Charlesworth, B. 2017. Estimating the parameters of background selection and selective sweeps in drosophila in the presence of gene conversion. *Proc. Natl. Acad. Sci. USA*, 114(24): E4762–E4771, doi: 10.1073/pnas.1619434114.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.*, 43(10): 956–963, doi: 10.1038/ng.911.

Charlesworth, B. 1992. Evolutionary rates in partially self-fertilizing species. *Am. Nat.*, 140(1): 126–148.

Charlesworth, B. and Charlesworth, D. 2010. *Elements of Evolutionary Genetics*. Roberts & Company Publishers, Greenwood Village, Colo.

Chevin, L.-M., Billiard, S., and Hospital, F. 2008. Hitchhiking both ways: Effect of two interfering selective sweeps on linked neutral variation. *Genetics*, 180(1): 301–316, doi: 10.1534/genetics.108.089706.

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S. P., Jensen, J. D., Campbell, M. C., Beggs, W., Hormozdiari, F., Mpoloka, S. W., Mokone, G. G., Nyambo, T., Meskel, D. W., Belay, G., Haut, J., NISC Comparative Sequencing Program, Rothschild, H., Zon, L., Zhou, Y., Kovacs, M. A.,

Xu, M., Zhang, T., Bishop, K., Sinclair, J., Rivas, C., Elliot, E., Choi, J., Li, S. A., Hicks, B., Burgess, S., Abnet, C., Watkins-Chow, D. E., Oceana, E., Song, Y. S., Eskin, E., Brown, K. M., Marks, M. S., Loftus, S. K., Pavan, W. J., Yeager, M., Chanock, S., and Tishkoff, S. A. 2017. Loci associated with skin pigmentation identified in African populations. *Science*, 358(6365), doi: 10.1126/science.aan8433.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158, doi: 10.1093/bioinformatics/btr330.

Desai, M. M. and Fisher, D. S. 2007. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics*, 176(3): 1759–1798, doi: 10.1534/genetics.106.067678.

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., and Sella, G. 2016. A genomic map of the effects of linked selection in Drosophila. *PLoS Genet.*, 12(8): e1006130, doi: 10.1371/journal.pgen.1006130.

Ewing, G., Hermisson, J., Pfaffelhuber, P., and Rudolf, J. 2011. Selective sweeps for recessive alleles and for other modes of dominance. *J. Math. Bio.*, 63(3): 399–431, doi: 10.1007/s00285-010-0382-4.

Fay, J. C. and Wu, C.-I. 2000. Hitchhiking Under Positive Darwinian Selection. *Genetics*, 155(3): 1405–1413.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. 2014. On de-

tecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.*, 31(5): 1275–1291, doi: 10.1093/molbev/msu077.

Fisher, R. A. 1922. On the dominance ratio. *Proc. R. Soc. Edinburgh*, 42: 321–341.

Fujito, N. T., Satta, Y., Hayakawa, T., and Takahata, N. 2018. A new inference method for detecting an ongoing selective sweep. *Genes Genet. Syst.*, 93(4): 149–161, doi: 10.1266/ggs.18-00008.

Fulgione, A., Koornneef, M., Roux, F., Hermisson, J., and Hancock, A. M. 2018. Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. *Mol. Biol. Evol.*, 35(3): 564–574, doi: 10.1093/molbev/msx300.

Fustier, M. A., Brandenburg, J. T., Boitard, S., Lapeyronnie, J., Eguiarte, L. E., Vigouroux, Y., Manicacci, D., and Tenaillon, M. I. 2017. Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. *Mol. Ecol.*, 26(10): 2738–2756, doi: 10.1111/mec.14082.

Garud, N. R. and Petrov, D. A. 2016. Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. *Genetics*, 203(2): 863–880, doi: 10.1534/genetics.115.184002.

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. 2015. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLoS Genet.*, 11(2): e1005004, doi:10.1371/journal.pgen.1005004.

Glémin, S. 2012. Extinction and fixation times with dominance and inbreeding. *Theor. Popul. Biol.*, 81(4): 310–316, doi: 10.1016/j.tpb.2012.02.006.

Glémin, S. and Ronfort, J. 2013. Adaptation and maladaptation in selfing and outcrossing species: New mutations versus standing variation. *Evolution*, 67(1): 225–240, doi: 10.1111/j.1558-5646.2012.01778.x.

Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., and Eyre-Walker, A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.*, 27(8): 1822–1832, doi: 10.1093/molbev/msq079.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet.*, 5(10): e1000695, doi: 10.1371/journal.pgen.1000695.

Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Math. Proc. Cambridge Philos. Soc.*, 23(7): 838–844, doi: 10.1017/S0305004100015644.

Haller, B. C. and Messer, P. W. 2019. Slim 3: Forward genetic simulations beyond the wright–fisher model. *Mol. Biol. Evol.*, 36(3): 632–637, doi: 10.1093/molbev/msy228.

Harris, A. M., Garud, N. R., and DeGiorgio, M. 2018a. Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. *Genetics*, 210(4): 1429–1452, doi: 10.1534/genetics.118.301502.

Harris, R. B., Sackman, A., and Jensen, J. D. 2018b. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLoS Genet.*, 14(12): e1007859, doi: 10.1371/journal.pgen.1007859.

Hartfield, M. and Glémin, S. 2014. Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. *Genetics*, 196(1): 281–293, doi: 10.1534/genetics.113.158196.

Hartfield, M. and Glémin, S. 2016. Limits to adaptation in partially selfing species. *Genetics*, 203(2): 959–974, doi: 10.1534/genetics.116.188821.

Hartfield, M., Bataillon, T., and Glémin, S. 2017. The evolutionary interplay between adaptation and self-fertilization. *Trends Genet.*, 33(6): 420–431, doi: 10.1016/j.tig.2017.04.002.

Hedrick, P. W. 1980. Hitchhiking: A comparison of linkage and partial selection. *Genetics*, 94(3): 791–808.

Hermisson, J. and Pennings, P. S. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4): 2335–2352, doi: 10.1534/genetics.104.036947.

Hermisson, J. and Pennings, P. S. 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.*, 8(6): 700–716, doi: 10.1111/2041-210X.12808.

Huber, C. D., Nordborg, M., Hermisson, J., and Hellmann, I. 2014. Keeping It Local: Evidence for Positive Selection in Swedish *Arabidopsis thaliana*. *Mol. Biol. Evol.*, 31(11): 3026–3039, doi: 10.1093/molbev/msu247.

Huber, C. D., DeGiorgio, M., Hellmann, I., and Nielsen, R. 2016. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol. Ecol.*, 25(1): 142–156, doi: 10.1111/mec.13351.

Igic, B. and Kohn, J. R. 2006. The distribution of plant mating systems: study bias against obligately outcrossing species. *Evolution*, 60(5): 1098–1103, doi: 10.1111/j.0014-3820.2006.tb01186.x.

Innan, H. and Kim, Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. USA*, 101(29): 10667–10672, doi: 10.1073/pnas.0401720101.

Innan, H. and Nordborg, M. 2003. The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. *Genetics*, 165(1): 437.

Jarne, P. and Auld, J. R. 2006. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. *Evolution*, 60(9): 1816–1824, doi: 10.1111/j.0014-3820.2006.tb00525.x.

Jensen, J. D. 2014. On the unfounded enthusiasm for soft selective sweeps. *Nat. Commun.*, 5, doi: 10.1038/ncomms6281.

Jorde, L. B., Bamshad, M., and Rogers, A. R. 1998. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays*, 20(2): 126–136.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. 1989. The "hitchhiking effect" revisited. *Genetics*, 123(4): 887–899.

Karasov, T., Messer, P. W., and Petrov, D. A. 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet.*, 6(6): e1000924, doi: 10.1371/journal.pgen.1000924.

Kern, A. D. and Schrider, D. R. 2018. diploS/HIC: An updated approach to classifying selective sweeps. *G3*, 8(6): 1959–1970, doi: 10.1534/g3.118.200262.

Kim, Y. and Nielsen, R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3): 1513–1524, doi: 10.1534/genetics.103.025387.

Kim, Y. and Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2): 765–777.

Kim, Y. and Stephan, W. 2003. Selective sweeps in the presence of interference among partially linked loci. *Genetics*, 164(1): 389–398.

Kimura, M. 1971. Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.*, 2(2): 174–208.

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., Parra, E. J., Mangini, N. J., Grunwald, D. J., Shriver, M. D., Canfield, V. A., and Cheng, K. C. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755): 1782–1786, doi: 10.1126/science.1116238.

Laporte, V. and Charlesworth, B. 2002. Effective population size and population subdivision in demographically structured populations. *Genetics*, 162(1): 501–519.

Lin, M., Siford, R. L., Martin, A. R., Nakagome, S., Möller, M., Hoal, E. G., Bustamante, C. D., Gignoux, C. R., and Henn, B. M. 2018. Rapid evolution

of a skin-lightening allele in southern African KhoeSan. *Proc. Natl. Acad. Sci. USA*, 115(52): 13324–13329, doi: 10.1073/pnas.1801948115.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjalmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandakova, T., Lysak, M. A., Seren, U., Hellmann, I., and Nordborg, M. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.*, 45(8): 884–890, doi: 10.1038/ng.2678.

Martin, G. and Lambert, A. 2015. A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations. *Theor. Popul. Biol.*, 101: 40–46, doi: 10.1016/j.tpb.2015.01.004.

Maynard Smith, J. 1976. What determines the rate of evolution? *Am. Nat.*, 110(973): 331–338, doi: 10.1086/283071.

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

McVean, G. A. T. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3): 1395–1406, doi: 10.1534/genetics.106.062828.

Messer, P. W. and Petrov, D. A. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.*, 28(11): 659–669, doi: 10.1016/j.tree.2013.08.003.

Nielsen, R. 2005. Molecular signals of natural selection. *Annu. Rev. Genet.*, 39(1): 197–218, doi: 10.1146/annurev.genet.39.073003.112420.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.*, 15(11): 1566–1575, doi: 10.1101/gr.4252305.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics*, 154(2): 923–929.

Nordborg, M. and Donnelly, P. 1997. The coalescent process with selfing. *Genetics*, 146(3): 1185–1195.

Nordborg, M., Charlesworth, B., and Charlesworth, D. 1996. Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. *Proc. R. Soc. B*, 263(1373): 1033–1039, doi: 10.1098/rspb.1996.0152.

Orr, H. A. and Betancourt, A. J. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, 157(2): 875–884.

Pennings, P. S. and Hermisson, J. 2006a. Soft Sweeps II – Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Mol. Biol. Evol.*, 23(5): 1076–1084, doi: 10.1093/molbev/msj117.

Pennings, P. S. and Hermisson, J. 2006b. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLoS Genet.*, 2(12): e186, doi: 10.1371/journal.pgen.0020186.

Pennings, P. S., Kryazhimskiy, S., and Wakeley, J. 2014. Loss and Recovery of Genetic Diversity in Adapting Populations of HIV. *PLoS Genet.*, 10(1): e1004000, doi: 10.1371/journal.pgen.1004000.

Peter, B. M., Huerta-Sanchez, E., and Nielsen, R. 2012. Distinguishing between selective sweeps from standing variation and from a *De Novo* mutation. *PLoS Genet.*, 8(10): e1003011, doi: 10.1371/journal.pgen.1003011.

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E. B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J., and Borevitz, J. O. 2010. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.*, 6(2): e1000843, doi: 10.1371/journal.pgen.1000843.

Pollak, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics*, 117(2): 353–360.

Pritchard, J. K. and Di Rienzo, A. 2010. Adaptation - not by sweeps alone. *Nat. Rev. Genet.*, 11(10): 665–667, doi: 10.1038/nrg2880.

Przeworski, M., Coop, G., and Wall, J. D. 2005. The signature of positive selection on standing genetic variation. *Evolution*, 59(11): 2312–2323, doi: 10.1111/j.0014-3820.2005.tb00941.x.

Rettelbach, A., Nater, A., and Ellegren, H. 2019. How linked selection shapes the diversity landscape in *Ficedula* flycatchers. *Genetics*, 212(1): 277–285, doi: 10.1534/genetics.119.301991.

Roze, D. 2009. Diploidy, population structure, and the evolution of recombination. *Am. Nat.*, 174(S1): S79–S94, doi: 10.1086/599083.

Roze, D. 2016. Background selection in partially selfing populations. *Genetics*, 203(2): 937–957, doi: 10.1534/genetics.116.187955.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909): 832–837, doi: 10.1038/nature01140.

Salome, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., and Weigel, D. 2012. The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity*, 108(4): 447–455, doi: 10.1038/hdy.2011.95.

Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., and Sella, G. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genet.*, 7(2): e1001302, doi: 10.1371/journal.pgen.1001302.

Schoen, D. J., Morgan, M. T., and Bataillon, T. 1996. How does self-pollination evolve? inferences from floral ecology and molecular genetic variation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 351(1345): 1281–1290, doi: 10.1098/rstb.1996.0111.

Schrider, D. R. and Kern, A. D. 2016. S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet.*, 12(3): e1005928, doi: 10.1371/journal.pgen.1005928.

Schrider, D. R. and Kern, A. D. 2017. Soft sweeps are the dominant mode of adaptation in the human genome. *Mol. Biol. Evol.*, 34(8): 1863–1877, doi: 10.1093/molbev/msx154.

Schrider, D. R. and Kern, A. D. 2018. Supervised machine learning for population genetics: A new paradigm. *Trends Genet.*, doi: 10.1016/j.tig.2017.12.005.

Schrider, D. R., Mendes, F. K., Hahn, M. W., and Kern, A. D. 2015. Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, 200(1): 267–284, doi: 10.1534/genetics.115.174912.

Sheehan, S. and Song, Y. S. 2016. Deep learning for population genetic inference. *PLoS Comput. Biol.*, 12(3): e1004845, doi: 10.1371/journal.pcbi.1004845.

Stephan, W. 2016. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.*, 25(1): 79–88, doi: 10.1111/mec.13288.

Stephan, W. 2019. Selective sweeps. *Genetics*, 211(1): 5–13, doi: 10.1534/genetics.118.301319.

Teshima, K. M. and Przeworski, M. 2006. Directional positive selection on an allele of arbitrary dominance. *Genetics*, 172(1): 713–718, doi: 10.1534/genetics.105.044065.

Teshima, K. M., Coop, G., and Przeworski, M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.*, 16(6): 702–712, doi: 10.1101/gr.5105206.

The 1000 Genomes Project Consortium 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, doi: 10.1038/nature15393.

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

Uecker, H. 2017. Evolutionary rescue in randomly mating, selfing, and clonal populations. *Evolution*, 71(4): 845–858, doi: 10.1111/evo.13191.

Van Herwaarden, O. A. and Van der Wal, N. J. 2002. Extinction time and age of an allele in a large finite population. *Theor. Popul. Biol.*, 61(3): 311–318, doi: 10.1006/tpbi.2002.1576.

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. 2016. Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.*, 25(1): 89–103, doi: 10.1111/mec.13360.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. 2013. Detecting natural selection in genomic data. *Annu. Rev. Genet.*, 47(1): 97–120, doi: 10.1146/annurev-genet-111212-133526.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.*, 4(3): e72, doi: 10.1371/journal.pbio.0040072.

Vy, H. M. T., Won, Y.-J., and Kim, Y. 2017. Multiple modes of positive selection shaping the patterns of incomplete selective sweeps over african populations of *Drosophila melanogaster*. *Mol. Biol. Evol.*, 34(11): 2792–2807, doi: 10.1093/molbev/msx207.

Wakeley, J. 2009. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers, Greenwood Village, Colorado.

Wilson, B. A., Petrov, D. A., and Messer, P. W. 2014. Soft Selective Sweeps in Complex Demographic Scenarios. *Genetics*, 198(2): 669–684, doi: 10.1534/genetics.114.165571.

Wilson, B. A., Pennings, P. S., and Petrov, D. A. 2017. Soft selective sweeps in evolutionary rescue. *Genetics*, 205(4): 1573–1586, doi: 10.1534/genetics.116.191478.

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

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

Yang, Z., Li, J., Wiehe, T., and Li, H. 2018. Detecting recent positive selection with a single locus test bipartitioning the coalescent tree. *Genetics*, 208(2): 791–805, doi: 10.1534/genetics.117.300401.