View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Selective sweeps under dominance and inbreeding

Citation for published version:
Hartfield, M & Bataillon, T 2020, 'Selective sweeps under dominance and inbreeding', *G3*, vol. 10, no. 1, g3.400919.2019. https://doi.org/10.1534/g3.119.400919

Digital Object Identifier (DOI):
10.1534/g3.119.400919

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
G3

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 22. Feb. 2020

G3: Genes|Genomes|Genetics Early Online, published on January 23, 2020 as doi:10.1534/g3.119.400919

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; Population Genetics.

© The Author(s) 2020. Published by the Genetics Society of America.

Abstract

A major research goal in evolutionary genetics is to uncover loci experiencing positive selection. One approach involves finding 'selective sweeps' patterns, which can either be 'hard sweeps' 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. These models 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 focused 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 in regions flanking the selected site, by raising the haplotype carrying the selected allele to high frequency (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 haplotype homozygosity (Sabeti *et al.* 2002; Kim and Nielsen 2004; Voight *et al.* 2006; Ferrer-Admetlla *et al.* 2014; Vatsiou *et al.* 2016), or an increase in high frequency 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 carry 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; Harris and DeGiorgio 2018, 2019). Soft sweeps have been reported 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; Laval *et al.* 2019), maize (Fustier *et al.* 2017), *Anopheles* mosquitoes (Xue *et al.* 2019), and pathogens including *Plasmodium falciparum* (Anderson *et al.* 2016) and HIV (Pennings *et al.* 2014; Williams and Pennings 2019). 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, finding that recessive beneficial mutations have markedly different signatures compared to those with other dominance values. 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 popu-

lations, with haplotypes freely mixing between individuals over generations. Different reproductive modes alter how alleles are inherited, potentially changing 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 than 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, making it likelier that deleterious mutations hitchhike to fixation with adaptive alleles (Hartfield and Glémin 2014), or that rare mutations are lost by drift due to competition between adaptive mutations (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; Price *et al.* 2018); *Caenorhabditis elegans* (Andersen *et al.* 2012); *Medicago truncatula* (Bonhomme *et al.* 2015); and *Microbotryum* fungi (Badouin *et al.* 2017). Soft sweeps have also been reported in soya bean (Zhong *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 selective sweep model that accounts for dominance and inbreeding via self–fertilisation. 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. 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).

Methods

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 $1-\sigma$. A derived allele arises at a locus, and we are interested in determining the population history of neutral regions that are linked to it, with a recombination rate r between them. We principally look at the case where the beneficial allele arises from previously–neutral standing variation, and subsequently look at a sweep arising from recurrent mutation. The derived allele initially segregates neutrally for a period of time, then 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 the population size is large and selection is large enough so that the beneficial allele's change in frequency can be modelled deterministically (i.e., $N_e hs \gg 1$ and $1/N_e \ll s \ll 1$). Table 1 lists the notation used in the analysis.

Our goal is to determine how the spread of the derived, adaptive allele affects genealogies at linked neutral regions. For a sweep originating from standing variation, we follow the approach of Berg and Coop (2015) and, looking backwards in time, break down the selected allele history into two phases. In the recent past is the 'sweep phase' where the derived allele was selectively favoured, with its frequency decreasing from 1 to p_0. Prior to that phase is the 'standing phase', which assumes that the derived allele is present at an approximate fixed frequency p_0. During both phases, a pair of haplotypes can either coalesce, or one of them recombines onto the ancestral background. A schematic is shown in Figure 1.

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
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	Probability of coalescence at time t
P_r	Probability of recombination at time t
P_m	Probability of mutation at time t
P_{NE}	Probability that neutral marker does not coalesce or recombine 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
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.

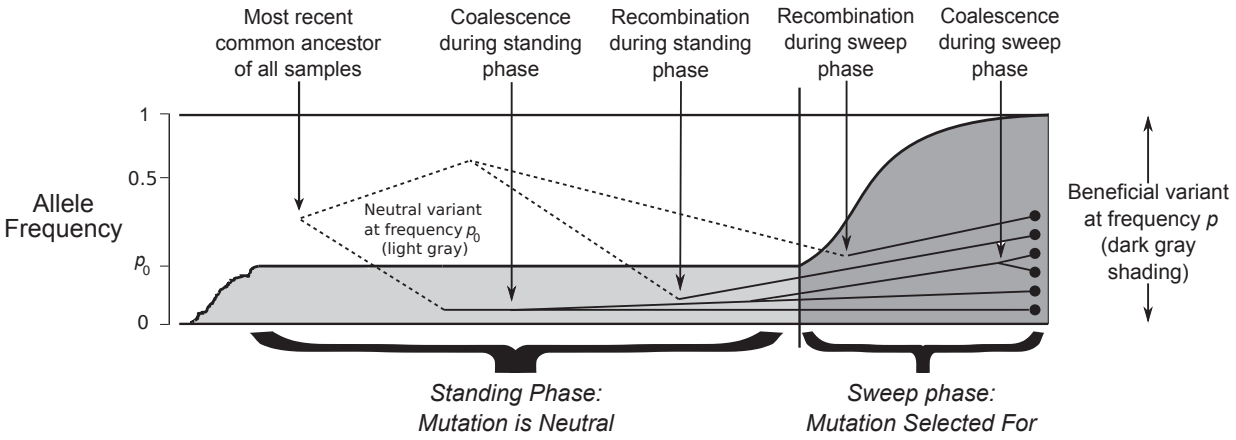

Figure 1. A schematic of the model. The history of the derived variant is separated into two phases; the 'standing phase' (shown in light gray), and the 'sweep phase' (shown in dark gray). Axis on the left-hand side show allele frequency on a log-scale. Dots on the right-hand side represent a sample of haplotypes taken at the present day, with lines representing their genetic histories. Solid lines represent coalescent histories for the derived genetic background; dotted lines represent coalescent histories for the ancestral, neutral background. Note the allele trajectory is an idealised version as assumed in the model.

During the sweep phase, the derived allele will also cause the spread of linked haplotypes that it appeared on. Over the course of the sweep, haplotypes are broken down by recombination; the total number of recombination events is proportional to $r\tau_{p_0}$, where τ_{p_0} is the fixation time of the beneficial allele, given an initial frequency p_0 (Maynard Smith and Haigh 1974). Dominance and self–fertilisation have different effects on τ_{p_0}, and therefore the number of fixing haplotypes. If p_0 is low ($\sim 1/2N$) then highly recessive or dominant mutations take longer to go to fixation (Glémin 2012), which can increase the number of recombination events. Dominance also affects the nature of the sweep trajectory. For example, recessive mutations spend more time at a low frequency compared to dominant mutations.

These different sweep trajectories can also affect the final sweep profile (Teshima and Przeworski 2006). Self–fertilisation leads to decreased fixation time of adaptive mutations through converting heterozygotes to homozygotes (Glémin 2012). Recombination is likelier to act between homozygotes under self-fertilisation, so its effective rate is reduced by a factor $1 - 2F + \Phi$, for $F = \sigma/(2 - \sigma)$ the inbreeding coefficient (Nordborg *et al.* 1996; Nordborg 2000) and Φ the joint probability of identity-by-descent at the two loci (Roze 2009, 2016; Hartfield and Glémin 2016), defined as:

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

Note that $1 - 2F + \Phi$ approximates to $1 - F$ (as $\Phi \approx F$), unless σ is close to one and r is high (approximately greater than 0.1).

During the standing phase, the amount of initial recombinant haplotypes that are swept to fixation depend on the relative rates of recombination and coalescence. The latter occurs with probability proportional to $1/2N_e$ for N_e the effective population size. Under self–fertilisation $N_e = N/(1+F)$ (Wright 1951; Pollak 1987; Charlesworth 1992; Caballero and Hill 1992; Nordborg and Donnelly 1997), so self–fertilisation increases the coalescence probability. This scaling factor will change if there is a large non-Poisson variation in offspring 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. Previous models tended to only determine how lineages recombine away from the derived background during the sweep phase, without considering how two lineages coalesce during the sweep phase. If lineages coalesce during the sweep, then the total number of unique recombination events, and hence the number of linked haplotypes, are reduced. Barton (1998) showed that these coalescent events are negligible only for very strong selection ($\log(Ns) \gg 1$; and B. Charlesworth, unpublished results). Hence, accounting for these coalescent events is important for producing accurate matches with simulation results.

Throughout, analytical solutions are compared to results from Wright-Fisher forward-in-time stochastic simulations that were ran using SLiM version 3.3 (Haller and Messer 2019). Results for outcrossing populations were also tested using coalescent simulations ran with *msms* (Ewing and Hermisson 2010). The simulation methods are outlined in Supplementary File S2.

Data Availability. File S1 is a *Mathematica* notebook of analytical derivations and simulation results. File S2 contains additional methods, results and figures. File S3 contains copies of the simulation scripts, which are also available from `https://github.com/MattHartfield/SweepDomSelf`. Supplemental material has also been uploaded to Figshare.

Results

Probability of events during sweep phase

We first look at the probability of events (coalescence or recombination) acting during the sweep phase for the simplest case of two alleles. Looking back in time following the fixation of the derived mutation, 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 $N_e = N/(1+F)$ 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). Equation 2

demonstrates how each event is differently influenced by p. In particular, the per–generation coalescence probability P_c can be small unless p is close to $1/2N$. The total probability that coalescence occurs during the sweep phase increases if the beneficial allele spends a sizeable time at low frequency, e.g., when it is recessive. The terms in Equation 2 can also be defined as functions of p.

We are interested in calculating (i) the probability P_{NE} that no coalescence or recombination occurs in the sweep phase; (ii) the probability $P_{R,Sw}$ that recombination acts on a lineage to transfer it to the neutral background that is linked to the ancestral allele, 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 do not coalesce or recombine over τ_{p_0} generations equals:

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

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. Also note that

we switch from a discrete–time calculation to a continuous–time calculation, which can give simplifying results. To calculate P_{NE} we insert the deterministic change in allele frequency p (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 $P_r(p') = 2r(1 - 2F + \Phi)(1 - p')$. $P_{R,Sw}$ is obtained by integrating this probability over the entire sweep from time 0 to τ_{p_0}:

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

where:

$$\begin{aligned} P_{R,p'} &= \exp\left[-\int_{p=1-\epsilon}^{p'} \frac{P_c(p) + P_r(p)}{\mathrm{d}p/\mathrm{d}t}\mathrm{d}p\right] \cdot P_r(p') \\ &= \exp\left[-\int_{p=1-\epsilon}^{p'} \frac{\frac{1+F}{2Np} + 2r(1-2F+\Phi)(1-p)}{\mathrm{d}p/\mathrm{d}t}\mathrm{d}p\right] \cdot [2r(1-2F+\Phi)(1-p')] \end{aligned} \tag{6}$$

Note that the exponential term of $P_{R,p'}$ is different from P_{NE} (Equation 3) since the upper integral limit is to p' rather than p_0. That is, it only covers part of the sweep phase. Equation 5 is evaluated numerically. In Supplementary File S2, we provide a 'star–like' analytical approximation to P_{NE} that assumes no coalescence during the sweep phase.

Probability of coalescence from standing variation

The variant becomes advantageous at frequency p_0. We assume that p_0, and hence event probabilities, remain fixed 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 an exponentially-distributed waiting time with rate $P_c(p_0) = (1 + F)/(2Np_0)$. Alternatively, one of the two haplotypes could recombine onto the ancestral background with mean waiting time $P_r(p_0) = 2r_{eff}(1 - p_0)$. 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, Chapter 2). Hence the probability that recombination occurs instead of coalescence equals:

$$
\begin{aligned}
P_{R,Sd} &= \frac{P_r(p_0)}{P_c(p_0) + P_r(p_0)} \\
&= \frac{2r_{eff}(1-p_0)}{\frac{1+F}{2Np_0} + 2r_{eff}(1-p_0)} \\
&= \frac{2R(1-2F+\Phi)p_0(1-p_0)/(1+F)}{1+2R(1-2F+\Phi)p_0(1-p_0)/(1+F)} \\
&\approx \frac{2R(1-\sigma)p_0(1-p_0)}{1+2R(1-\sigma)p_0(1-p_0)}
\end{aligned} \tag{7}
$$

The probability of coalescence rather than recombination is $P_{C,Sd} = 1 - P_{R,Sd}$. Here $R = 2Nr$ is the population-scaled recombination rate. The final approximation arises as $(1-2F+\Phi)/(1+F) \approx (1-F)/(1+F) = (1-\sigma)$ if $\Phi \approx F$. This term reflects how increased homozygosity reduces both effective recombination and N_e, with the latter making coalescence more likely. In addition, it also highlights how the signature of a sweep from standing variation, as characterised by the spread of different initial recombinant haplotypes, is spread over an increased distance of $1/(1-\sigma)$ under self–fertilisation.

Effective starting frequency for a *de novo* mutation, and effective final frequency

When a new beneficial mutation appears as 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 effective p_0, especially for recessive mutations (Figure 2).

The effective final frequency of the derived allele $1-\epsilon$, at which its spread is no longer deterministic, can be obtained by setting $\epsilon = p_{0,A}(1-h)$; that is, by substituting H_l to $H_h = 1 - h + Fh$ in Equation 8. This final frequency is always used, even if $p_0 > 1/2N$. 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 Charlesworth (2020) for a fuller discussion of effective final frequencies and their impact on sweep fixation times.

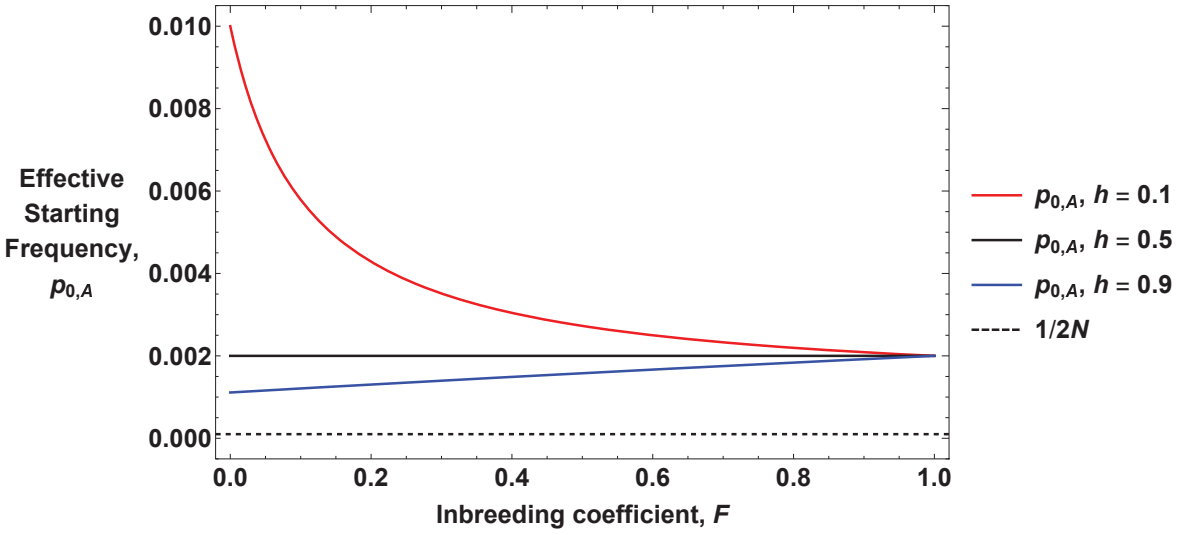

Figure 2. Examples of the effective starting frequency. Equation 8 is plotted as a function of F for different dominance values, as shown in the legend. Other parameters are $N = 5,000$, $s = 0.05$. The dashed line shows the actual starting frequency, $1/2N$.

Expected Pairwise Diversity

We 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, the two neutral sites could either coalesce, or one of them recombines onto the ancestral background. If coalescence occurs, since it does so in the recent past then it is assumed that no diversity exist between samples, i.e., $\pi \approx 0$ for π the average number of differences between two alleles (Tajima 1983). In reality there may be some residual diversity caused by appearance of mutations during the sweep phase; we do not account for these mutations while calculating π but will do so when calculating the site-frequency spectrum. Alternatively, if one of the two samples recombines onto the neutral background, they will have the same pairwise diversity between them 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 π_{SV}, relative to the background value π_0, equals:

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

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. Equation 9 is undefined for $h = 0$ or 1 with $\sigma = 0$; these cases can be derived separately.

Figure 3 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. It is unclear why this mismatch arises. One explanation could be that drift effects are magnified under self–fertilisation, which causes a quicker sweep fixation time than expected from deterministic spread, if conditioning on a sweep going to fixation. Although $p_{0,A}$ (Equation 8) captures these drift effects for rare alleles, there may be additional effects that are not accounted for. Under complete outcrossing, baseline diversity is restored (i.e., $\mathbb{E}(\pi_{SV}/\pi_0)$ 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. Sweeps caused by dominant and semidominant mutations result in a similar genetic diversity, so these cases may be hard to differentiate from diversity data alone.

These results can be better understood by examining the underlying allele trajectories, using logic described by Teshima and Przeworski (2006) (Figure 4). For outcrossing populations, recessive mutations spend most of the sojourn time at

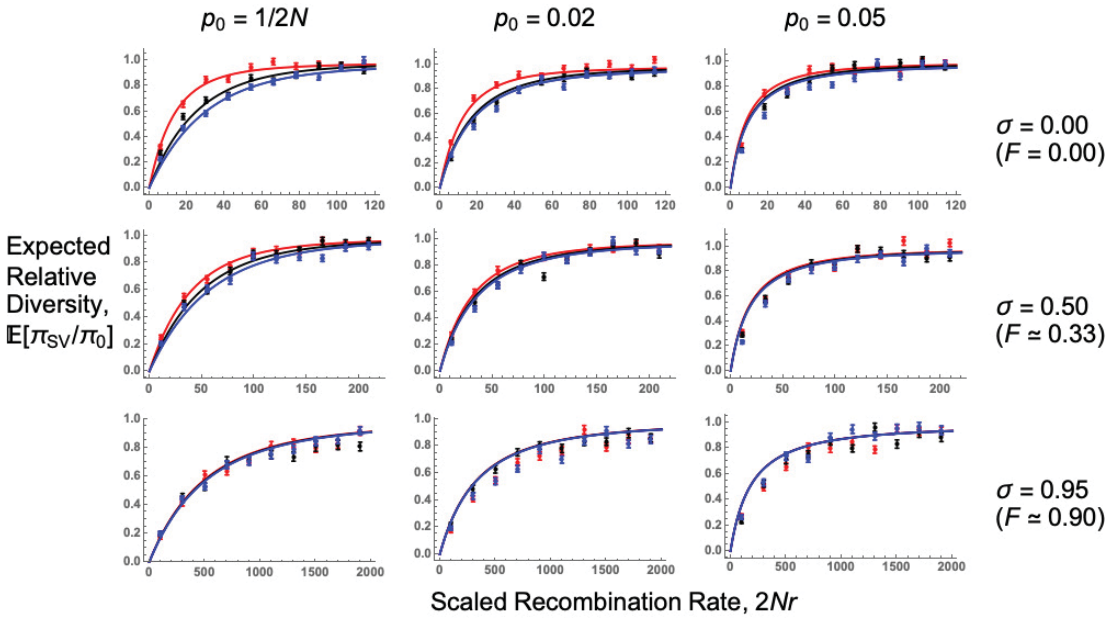

Figure 3. Expected relative pairwise diversity following a selective sweep. Plots of $\mathbb{E}(\pi_{SV}/\pi_0)$ as a function of the recombination rate scaled to population size $2Nr$. Lines are analytical solutions (Equation 9), points are forward-in-time simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$ (note μ is scaled by N, 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.

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, so most recombination events are between haplotypes that carry the derived allele. Hence, there is a reduced chance for linked neutral alleles to recombine onto the ancestral background.

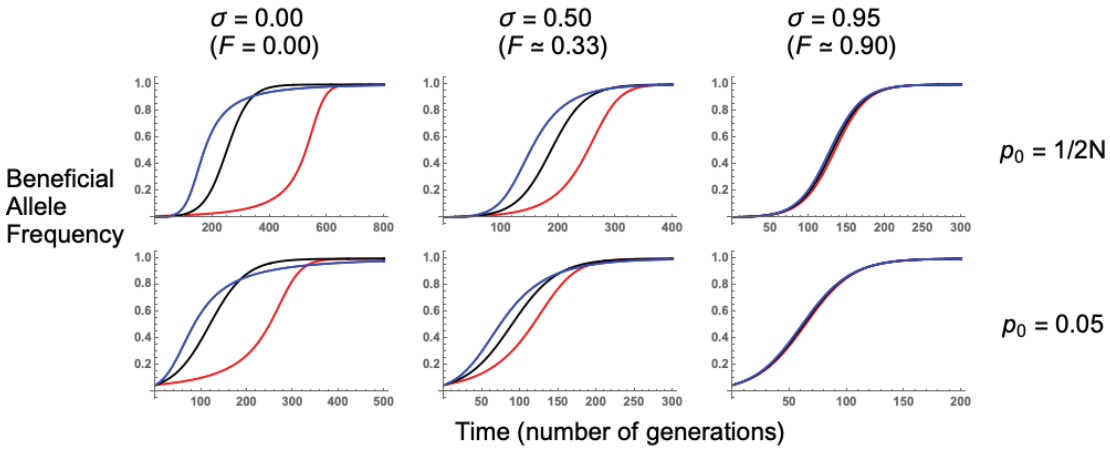

Figure 4. Beneficial allele trajectories. These were obtained by numerically evaluating the negative of Equation 4 forward in time. $N = 5,000$, $s = 0.05$, and h equals either 0.1 (red lines), 0.5 (black lines), or 0.9 (blue lines). Values of p_0 and self-fertilisation rates σ used are shown for the 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.

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 on starting at an elevated frequency.

An analytical approximation can be obtained by using the 'star-like' result for P_{NE} (described in Supplementary Files S1, S2). In this case the expected pairwise diversity approximates to:

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

Note that Equation 10 instead uses the probability of coalescence during the standing phase, $P_{C,sd} = 1 - P_{R,sd}$. This approximation reflects similar formulas for diversity following soft sweeps in haploid 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. In Supplementary File S2 we demonstrate that this equation overestimates the relative diversity following a selective sweep. This mismatch arises since the star-like assumption of no coalescence during the sweep phase is only accurate for very strongly selected mutations (Barton 1998; B. Charlesworth, unpublished results). Hence it is important to consider coalescence during the sweep phase to accurately model selective sweeps that do not have an extremely high selection coefficient.

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, which uses the star-like approximation. Figure 5 plots the SFS (Equation A12 in Supplementary File S2)

alongside simulation results. Analytical results fit the simulation data well after including an adjusted singleton class, which accounts for recent mutations that arise on the derived background during both the standing and sweep phases (Berg and Coop 2015). Including this new singleton class improves the model fit, but there remains a tendency for analytical results to underestimate the proportion of low- and high-frequency classes ($l = 1$ and 9 in Figure 5), and overestimate the proportion of intermediate-frequency classes. Additional inaccuracies could have arisen due to the use of the star-like approximation, which assumes that there is no coalescence during the sweep phase.

Hard sweeps in either outcrossers or partial selfers are characterised by a large number of singletons and highly-derived variants (Figure 5), which is a typical selective sweep signature (Braverman *et al.* 1995; Barton 1998; Kim and Stephan 2002). As the initial frequency p_0 increases, so does the number of intermediate-frequency variants (Figure 5). 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 4). 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 5, bottom row). This is an example of how signatures of sweeps from standing variation are extended over an increased recombination distance of around $1/(1 - \sigma)$, as demonstrated by Equation 7.

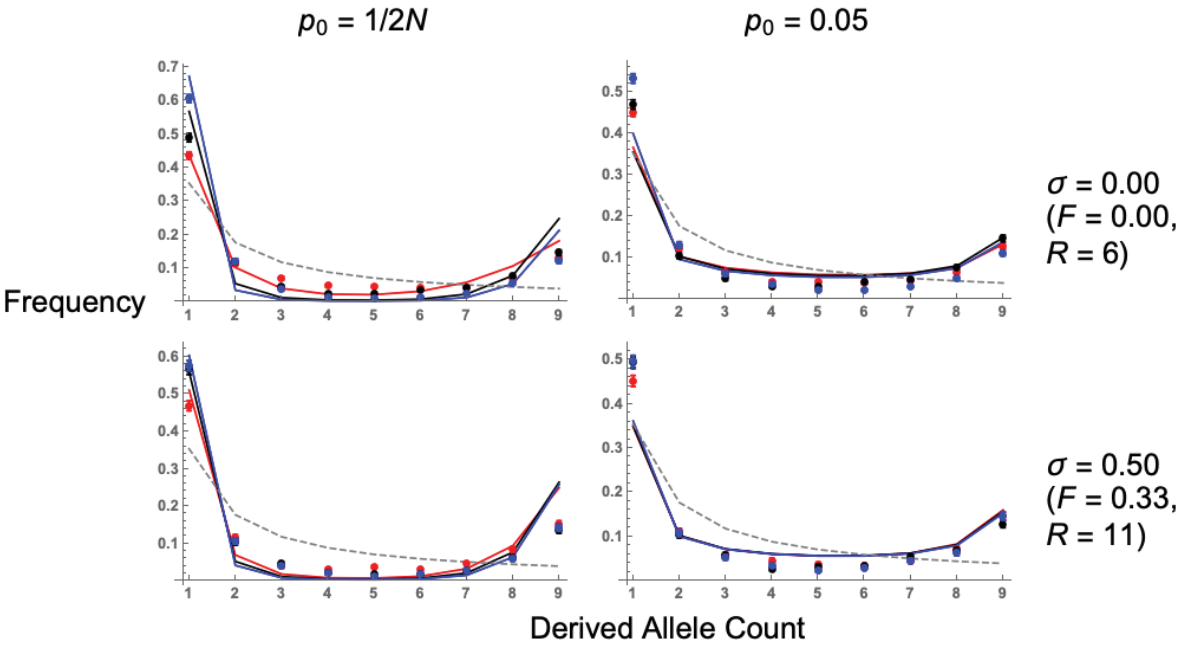

Figure 5. Expected site frequency spectrum, in flanking regions to the adaptive mutation, following a selective sweep. Lines are analytical solutions (Equation A12 in Supplementary File S2), points are simulation results. $N = 5,000$, $s = 0.05$, $4N\mu = 40$, and dominance coefficient $h = 0.1$ (red 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.

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. Below we derive the expected relative diversity between two alleles following this type of soft sweep, and outline

the SFS for more than two samples in Supplementary File S2.

In this model, derived alleles arise from recurrent mutation and are instantaneously beneficial (i.e., there is no 'standing phase'). During the sweep phase, lineages can escape the derived background by recombination, or if they are derived from a mutation event. If the beneficial allele is at frequency p then the probability of being descended from an ancestral allele by mutation is $P_m(p) = 2\mu_b(1-p)/p$, for μ_b the mutation probability (Pennings and Hermisson 2006b). Denote the probability of a lineage experiencing recombination or mutation during this sweep phase by $P_{R,sw}$, $P_{M,sw}$ respectively. In both these cases the expected diversity present at linked sites is π_0. If none of these events arise with probability P_{NE}, then remaining lineages can either coalesce, or they arise from independent mutation events. If they coalesce then they have approximately zero pairwise diversity between them; alternatively, they have different origins and thus exhibit the same pairwise diversity π_0 as the neutral background. Let $P_{M,sd}$ denote the probability that mutation occurs at the sweep origin, as opposed to coalescence.

Following this logic, the expected relative diversity for a sweep arising from recurrent mutation equals (with additional details in Supplementary File S1):

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

π_M denotes the diversity around a soft sweep from recurrent mutation. $P_{R,sw}$, P_{NE} are similar to the equations used when modelling a sweep from standing variation. They are both modified to account for additional beneficial 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:

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

and:

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

Note that Equation 14 has an upper integral limit of $p_{0,A}$, as opposed to a general p_0 used in the sweep from standing variation model, reflecting that there is no standing phase.

$P_{M,sw}$ is the mutation probability during the sweep phase, and is similar to Equation 13 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):

$$
\begin{aligned}
P_{M,Sd} &= \frac{P_m(p_{0,A})}{P_c(p_{0,A}) + P_m(p_{0,A})} \\
&= \frac{\frac{2\mu_b(1-p_{0,A})}{p_{0,A}}}{\frac{1+F}{2Np_{0,A}} + \frac{2\mu_b(1-p_{0,A})}{p_{0,A}}} \\
&= \frac{2\Theta_b(1-p_{0,A})}{1+F+2\Theta_b(1-p_{0,A})} \qquad (15)
\end{aligned}
$$

where $\Theta_b = 2N\mu_b$. The coalescence probability is $1 - P_{M,Sd}$. Equation 15 implies that self–fertilisation makes it more likely for beneficial mutations to coalesce at the start of a sweep, rather than arising from independent mutation events. Hence the signatures of soft sweeps via recurrent mutation will be weakened under inbreeding.

Figure 6 compares $\mathbb{E}(\pi_{SV}/\pi_0)$ in the standing variation case, and $\mathbb{E}(\pi_M/\pi_0)$ 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 4). Second, the two models exhibit different behaviour close to the selected locus (R close to zero). The recurrent mutation model has non–zero diversity levels, while the standing variation model exhibits zero diversity. As R increases, diversity eventually becomes higher for the standing variation case compared to the recurrent mutation case. We can heuristically determine when this transition occurs as follows. Assume a large population size but weak recombination and mutation rates. Hence, it is unlikely that any events occur during the sweep phase, so $P_{R,sw}$, $P_{M,sw} \approx 0$ and $P_{NE} \approx 1$. Then the expected relative diversity (Equation 11) equals $P_{R,sd}$ for a sweep from standing variation, and $P_{M,sd}$ for one from recurrent

mutation. To find the recombination rate R_{lim} at which a sweep from recurrent mutation yields higher diversity than one from standing variation, we find the R value needed to equate the two probabilities, giving:

$$\begin{aligned} R_{Lim} &= \frac{\Theta_b}{p_0(1-2F+\Phi)} \\ &\approx \frac{\Theta_b}{p_0(1-F)} \end{aligned} \tag{16}$$

The last approximation arises as $\Phi \approx F$. 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 16 is generally accurate when sweeps from standing variation have higher diversity than sweeps with recurrent mutations (Figure 6, bottom row), but becomes inaccurate for $h = 0.1$ in outcrossing populations, as some events are likely to occur during the sweep phase. In Supplementary File S2 we show how similar results apply to the SFS.

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

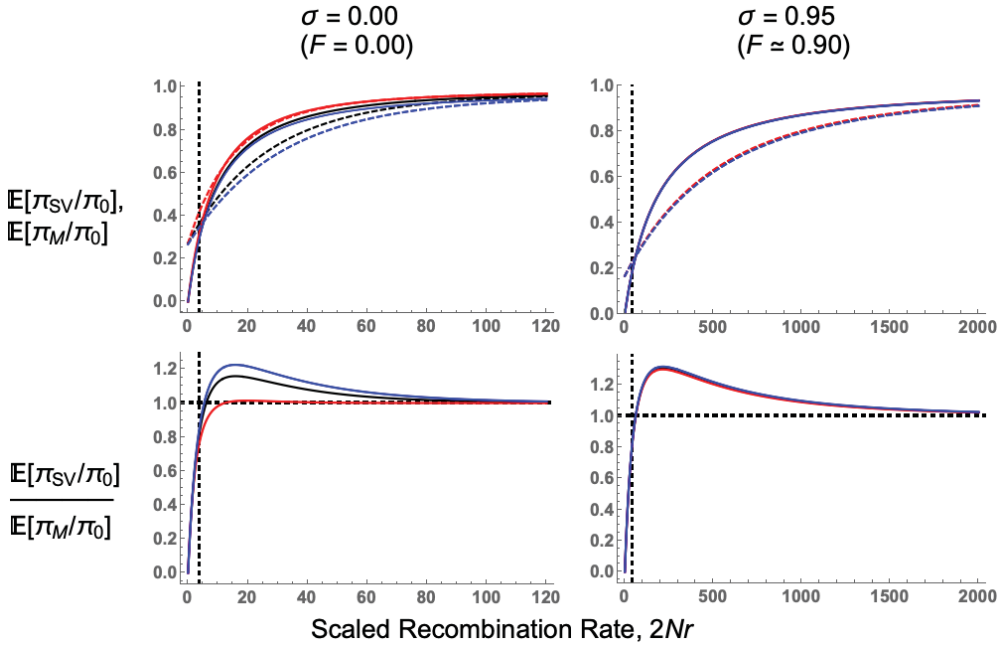

Figure 6. Comparing sweeps from recurrent mutation to those from standing variation. Top row: comparing relative diversity following a soft sweep, from either standing variation (Equation 9 with $p_0 = 0.05$, solid lines) or recurrent mutation (using Equation 11 with $\Theta_b = 0.2$, dashed lines). $N = 5,000$, $s = 0.05$, 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 to one from recurrent mutation. Parameters for each panel are as in the respective plot for the top row. Vertical dashed black line indicates R_{Lim} (the approximate form of Equation 16); horizontal dashed line in the bottom-row plots show when the ratio equals 1. Note the different $x-$axis between left- and right-hand panels. Results are also plotted in Section F of Supplementary File S1.

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 6).

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 5, 6). 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 more 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 $(1-\sigma)$ (Equation 7). Hence intermediate-frequency variants,

which could provide evidence of adaptation from standing variation, will be spread out over longer genomic regions (this result can be seen in the site–frequency spectrum results, Figure 5). The elongation of sweep signatures means sweeps from standing variation can be easier to detect in selfing organisms than in outcrossers. Conversely, sweeps from recurrent mutation will have weakened signatures under self–fertilisation. This result is due to a reduced effective population size, making it likelier that lineages trace back to a common ancestor rather than independent mutation events.

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 3, 4). With increased self-fertilisation, dominance has a weaker impact on sweep signatures as most mutations are homozygous (Figure 4). 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 5). 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 3, 4). Yet different dominance levels can affect sweep signatures if the beneficial allele is reintroduced by recurrent mutation (Figure 6). Hence if one wishes to understand how dominance affects sweep signatures, it is also important to consider which processes underlie observed patterns of genetic diversity.

These results also demonstrate that the effects of dominance on sweeps are

not necessarily intuitive. For example, both highly dominant and recessive mutations have elongated fixation times compared to co–dominant mutations (Glémin 2012). Based on this intuition, one could expect both dominant and recessive mutations to both produce weaker sweep signatures than co-dominant ones. In practice, dominant mutations have similar sweep signatures to co–dominant mutations (Figures 3, 5), and recessive sweeps could produce similar signatures as sweeps from standing variation (Figure 5). Dominance also has a weaker impact on sweeps from standing variation (Figures 3, 5).

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 have non–zero diversity at the selected locus, whereas a sweep from standing variation exhibits approximately zero diversity. Hence a sweep from recurrent mutation shows intermediate-frequency variants closer to the beneficial locus, compared to sweeps from standing variation (Figures 6 and C in Supplementary File S2). Further from the selected locus, a sweep from standing variation exhibits greater variation than one from recurrent mutation, due to recombinant haplotypes being created during the standing phase. Equation 16 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; from this equation, we see that the size of this region increases under higher self-fertilisation. Hence it may be easier to differentiate between these two sweep scenarios in self–fertilising

organisms.

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. Selective sweep signatures in data tend to be asymmetric, and this effect will exacerbate this asymmetry (Chevin *et al.* 2008). Further investigations of selective sweep patterns across long genetic distances will prove to be a rich area of future research.

Finally, we have assumed a fixed population size, and that sweeps from standing variation arose from neutral variation. The resulting signatures could differ if the population size has changed over time (Wilson *et al.* 2014), if populations are structured (Zheng and Wiehe 2019), or if the beneficial allele was previously deleterious (Orr and Betancourt 2001). Both issues could also affect our ability to discriminate between soft and hard sweeps.

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; Pavlidis *et al.* 2013; DeGiorgio *et al.* 2016; 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), or detect patterns of introgression (Setter *et al.* 2019).

Acknowledgments. We would like to thank Sally Otto for providing information on the elevated effective starting frequency of beneficial mutations; Brian Charlesworth on providing advice on modelling selective sweeps, sharing unpublished results, and providing comments on the manuscript; Ben Haller for answering questions about SLiM; Nick Barton, Jeffrey Ross-Ibarra and other anonymous referees for providing feedback on the manuscript. 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., J. P. Gerke, J. A. Shapiro, J. R. Crissman, R. Ghosh, *et al.*, 2012 Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. **44**: 285–290.

Anderson, T. J. C., S. Nair, M. McDew-White, I. H. Cheeseman, S. Nkhoma, *et al.*, 2016 Population parameters underlying an ongoing soft sweep in southeast asian malaria parasites. Mol. Biol. Evol. **34**: 131–144.

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

Barrett, R. D. H. and D. Schluter, 2008 Adaptation from standing genetic variation. Trends Ecol. Evol. **23**: 38–44.

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

Berg, J. J. and G. Coop, 2015 A coalescent model for a sweep of a unique standing variant. Genetics **201**: 707–725.

Billiard, S., M. López-Villavicencio, B. Devier, M. E. Hood, C. Fairhead, *et al.*, 2011 Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. Biol. Rev. Camb. Philos. Soc. **86**: 421–442.

Boitard, S., C. Schlötterer, and A. Futschik, 2009 Detecting selective sweeps: A new approach based on hidden markov models. Genetics **181**: 1567–1578.

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

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

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

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

Campos, J. L. and B. Charlesworth, 2019 The effects on neutral variability of recurrent selective sweeps and background selection. Genetics **212**: 287–303.

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

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

Charlesworth, B., 2020 How long does it take to fix a favorable mutation, and why should we care? Am. Nat. **Early Online**.

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

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

DeGiorgio, M., C. D. Huber, M. J. Hubisz, I. Hellmann, and R. Nielsen, 2016 SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics **32**: 1895–1897.

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

Elyashiv, E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, *et al.*, 2016 A genomic map of the effects of linked selection in Drosophila. PLoS Genet. **12**: e1006130.

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

Ewing, G., J. Hermisson, P. Pfaffelhuber, and J. Rudolf, 2011 Selective sweeps for recessive alleles and for other modes of dominance. J. Math. Biol. **63**: 399–431.

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

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol. **31**: 1275–1291.

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

Fujito, N. T., Y. Satta, T. Hayakawa, and N. Takahata, 2018 A new inference method for detecting an ongoing selective sweep. Genes Genet. Syst. **93**: 149–161.

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

Fustier, M. A., J. T. Brandenburg, S. Boitard, J. Lapeyronnie, L. E. Eguiarte, *et al.*, 2017 Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. Mol. Ecol. **26**: 2738–2756.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. PLoS Genet. **11**: e1005004.

Garud, N. R. and D. A. Petrov, 2016 Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. Genetics **203**: 863–880.

Glémin, S., 2012 Extinction and fixation times with dominance and inbreeding. Theor. Popul. Biol. **81**: 310–316.

Glémin, S. and J. Ronfort, 2013 Adaptation and maladaptation in selfing and outcrossing species: New mutations versus standing variation. Evolution **67**: 225–240.

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

Haller, B. C. and P. W. Messer, 2019 SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. **36**: 632–637.

Harris, A. M. and M. DeGiorgio, 2018 Identifying and classifying shared selective sweeps from multilocus data. bioRxiv p. 446005.

Harris, A. M. and M. DeGiorgio, 2019 A likelihood approach for uncovering selective sweep signatures from haplotype data. bioRxiv .

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

Harris, R. B., A. Sackman, and J. D. Jensen, 2018b On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. PLoS Genet. **14**: e1007859.

Hartfield, M., T. Bataillon, and S. Glémin, 2017 The evolutionary interplay between adaptation and self-fertilization. Trends Genet. **33**: 420–431.

Hartfield, M. and S. Glémin, 2014 Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. Genetics **196**: 281–293.

Hartfield, M. and S. Glémin, 2016 Limits to adaptation in partially selfing species. Genetics **203**: 959–974.

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

Hermisson, J. and P. S. Pennings, 2005 Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics **169**: 2335–2352.

Hermisson, J. and P. S. Pennings, 2017 Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol. **8**: 700–716.

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

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

Igic, B. and J. R. Kohn, 2006 The distribution of plant mating systems: study bias against obligately outcrossing species. Evolution **60**: 1098–1103.

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

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

Jarne, P. and J. R. Auld, 2006 Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution **60**: 1816–1824.

Jensen, J. D., 2014 On the unfounded enthusiasm for soft selective sweeps. Nat. Commun. **5**.

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

Karasov, T., P. W. Messer, and D. A. Petrov, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PLoS Genet. **6**: e1000924.

Kern, A. D. and D. R. Schrider, 2018 diploS/HIC: An updated approach to classifying selective sweeps. G3 **8**: 1959–1970.

Kim, Y. and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167**: 1513–1524.

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

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

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

Laval, G., E. Patin, P. Boutillier, and L. Quintana-Murci, 2019 A genome-wide approximate bayesian computation approach suggests only limited numbers of soft sweeps in humans over the last 100,000 years. bioRxiv p. 2019.12.22.886234.

Long, Q., F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow, *et al.*, 2013 Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat. Genet. **45**: 884–890.

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

Maynard Smith, J., 1976 What determines the rate of evolution? Am. Nat. **110**: 331–338.

Maynard Smith, J. and J. Haigh, 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**: 1395–1406.

Messer, P. W. and D. A. Petrov, 2013 Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. **28**: 659–669.

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

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15**: 1566–1575.

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

Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. B **263**: 1033–1039.

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

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

Pavlidis, P., D. Živković, A. Stamatakis, and N. Alachiotis, 2013 SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. Mol. Biol. Evol. **30**: 2224–2234.

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

Pennings, P. S. and J. Hermisson, 2006b Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genet. **2**: e186.

Pennings, P. S., S. Kryazhimskiy, and J. Wakeley, 2014 Loss and Recovery of Genetic Diversity in Adapting Populations of HIV. PLoS Genet. **10**: e1004000.

Peter, B. M., E. Huerta-Sanchez, and R. Nielsen, 2012 Distinguishing between selective sweeps from standing variation and from a *De Novo* mutation. PLoS Genet. **8**: e1003011.

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

Price, N., B. T. Moyers, L. Lopez, J. R. Lasky, J. G. Monroe, *et al.*, 2018 Combining population genomics and fitness QTLs to identify the genetics of local adaptation in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **115**: 5028–5033.

Pritchard, J. K. and A. Di Rienzo, 2010 Adaptation - not by sweeps alone. Nat. Rev. Genet. **11**: 665–667.

Przeworski, M., G. Coop, and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution **59**: 2312–2323.

Rettelbach, A., A. Nater, and H. Ellegren, 2019 How linked selection shapes the diversity landscape in *Ficedula* flycatchers. Genetics **212**: 277–285.

Roze, D., 2009 Diploidy, population structure, and the evolution of recombination. Am. Nat. **174**: S79–S94.

Roze, D., 2016 Background selection in partially selfing populations. Genetics **203**: 937–957.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

Schoen, D. J., M. T. Morgan, and T. Bataillon, 1996 How Does Self-Pollination Evolve? Inferences from Floral Ecology and Molecular Genetic Variation. Philos. Trans. R. Soc. Lond. B Biol. Sci. **351**: 1281–1290.

Schrider, D. R. and A. D. Kern, 2016 S/HIC: Robust identification of soft and hard sweeps using machine learning. PLoS Genet. **12**: e1005928.

Schrider, D. R. and A. D. Kern, 2017 Soft sweeps are the dominant mode of adaptation in the human genome. Mol. Biol. Evol. **34**: 1863–1877.

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

Setter, D., S. Mousset, X. Cheng, R. Nielsen, M. DeGiorgio, *et al.*, 2019 Volcanofinder: genomic scans for adaptive introgression. bioRxiv p. 697987.

Sheehan, S. and Y. S. Song, 2016 Deep learning for population genetic inference. PLoS Comput. Biol. **12**: e1004845.

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

Stephan, W., 2019 Selective sweeps. Genetics **211**: 5–13.

Tajima, F., 1983 Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics **105**: 437–460.

Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16**: 702–712.

Teshima, K. M. and M. Przeworski, 2006 Directional positive selection on an allele of arbitrary dominance. Genetics **172**: 713–718.

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

Uecker, H., 2017 Evolutionary rescue in randomly mating, selfing, and clonal populations. Evolution **71**: 845–858.

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

Vatsiou, A. I., E. Bazin, and O. E. Gaggiotti, 2016 Detection of selective sweeps in structured populations: a comparison of recent methods. Mol. Ecol. **25**: 89–103.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti, 2013 Detecting natural selection in genomic data. Annu. Rev. Genet. **47**: 97–120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4**: e72.

Vy, H. M. T., Y.-J. Won, and Y. Kim, 2017 Multiple Modes of Positive Selection Shaping the Patterns of Incomplete Selective Sweeps over African Populations of *Drosophila melanogaster*. Mol. Biol. Evol. **34**: 2792–2807.

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

Williams, K.-A. and P. S. Pennings, 2019 Drug resistance evolution in HIV in the late 1990s: hard sweeps, soft sweeps, clonal interference and the accumulation of drug resistance mutations. bioRxiv p. 548198.

Wilson, B. A., P. S. Pennings, and D. A. Petrov, 2017 Soft selective sweeps in evolutionary rescue. Genetics **205**: 1573–1586.

Wilson, B. A., D. A. Petrov, and P. W. Messer, 2014 Soft Selective Sweeps in Complex Demographic Scenarios. Genetics **198**: 669–684.

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

Xue, A. T., D. R. Schrider, A. D. Kern, and Ag1000G Consortium, 2019 Discovery of ongoing selective sweeps within *Anopheles* mosquito populations using deep learning. bioRxiv p. 589069.

Yang, Z., J. Li, T. Wiehe, and H. Li, 2018 Detecting recent positive selection with a single locus test bipartitioning the coalescent tree. Genetics **208**: 791–805.

Zheng, Y. and T. Wiehe, 2019 Adaptation in structured populations and fuzzy boundaries between hard and soft sweeps. PLoS Comput. Biol. **15**: e1007426.

Zhong, L., Q. Yang, X. Yan, C. Yu, L. Su, *et al.*, 2017 Signatures of soft sweeps across the Dt1 locus underlying determinate growth habit in soya bean [*Glycine max (L.) Merr.*]. Mol. Ecol. **26**: 4686–4699.