How Good Are Predictions of the Effects of Selective Sweeps on Levels of Neutral Diversity?

Brian Charlesworth[1]
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3FL, United Kingdom
ORCID ID: 0000-0002-2706-355X (B.C.)

ABSTRACT Selective sweeps are thought to play a significant role in shaping patterns of variability across genomes; accurate predictions of their effects are, therefore, important for understanding these patterns. A commonly used model of selective sweeps assumes that alleles sampled at the end of a sweep, and that fail to recombine with wild-type haplotypes during the sweep, coalesce instantaneously, leading to a simple expression for sweep effects on diversity. It is shown here that there can be a significant probability that a pair of alleles sampled at the end of a sweep coalesce during the sweep before a recombination event can occur, reducing their expected coalescent time below that given by the simple approximation. Expressions are derived for the expected reductions in pairwise neutral diversities caused by both single and recurrent sweeps in the presence of such within-sweep coalescence, although the effects of multiple recombination events during a sweep are only treated heuristically. The accuracies of the resulting expressions were checked against the results of simulations. For even moderate ratios of the recombination rate to the selection coefficient, the simple approximation can be substantially inaccurate. The selection model used here can be applied to favorable mutations with arbitrary dominance coefficients, to sex-linked loci with sex-specific selection coefficients, and to inbreeding populations. Using the results from this model, the expected differences between the levels of variability on X chromosomes and autosomes with selection at linked sites are discussed, and compared with data on a population of *Drosophila melanogaster*.

KEYWORDS selective sweeps; background selection; neutral variability; favorable mutations; *Drosophila melanogaster*

MAYNARD SMITH and Haigh (1974) introduced the concept of hitchhiking into population genetics, showing how the spread of a favorable mutation reduces the level of neutral variability at a linked locus. Nearly 20 years later, it was shown that selection against recurrent deleterious mutations can also reduce variability, by the hitchhiking process known as background selection (BGS) (Charlesworth *et al.* 1993). It is, therefore, preferable to use the term "selective sweep" (Berry *et al.* 1991) for the hitchhiking effects of favorable mutations. There is now a large theoretical and empirical literature on both types of hitchhiking, recently reviewed by Walsh and Lynch (2018) and Stephan (2019). With sufficiently weak selection, recurrent partially recessive deleterious mutations can sometimes increase variability at linked sites, because fluctuations in their frequencies due to genetic drift create associative overdominance (Zhao and Charlesworth 2016; Becher *et al.* 2020; Gilbert *et al.* 2020).

These theoretical studies have provided the basis for methods for inferring the nature and parameters of selection from population genomic data, recently reviewed by Booker *et al.* (2017). Several recent studies have concluded that the level of DNA sequence variability in a species is often much smaller than would be expected in the absence of selection (Corbett-Detig *et al.* 2015; Elyashiv *et al.* 2016; Campos *et al.* 2017; Comeron 2017), especially for synonymous sites in coding sequences, reflecting the effects of both selective sweeps and BGS. However, estimates of the parameters involved differ substantially among different studies. There is also an ongoing debate about the extent to which the level of genetic variability in a species is controlled by classical genetic drift, reflecting its population size, or by the effects of selection in removing variability. The possibility that the effects of selective sweeps dominate over drift was originally raised by Maynard Smith and Haigh (1974), and later advocated by Kaplan *et al.* (1989) and

doi: https://doi.org/10.1534/genetics.120.303734
Manuscript received September 28, 2020; accepted for publication October 22, 2020; published Early Online October 26, 2020.
Supplemental material available at figshare: https://doi.org/10.25386/genetics.13136012.
[1] Address for correspondence: Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, King's Bldg., Charlotte Auerbach Road, EH9 3FL, UK. E-mail: Brian.Charlesworth@ed.ac.uk

Gillespie (2002); see Kern and Hahn (2018) and Jensen *et al.* (2019) for recent discussions of this question.

The model of Maynard Smith and Haigh (1974) assumed that the trajectory of the selectively favored allele was purely deterministic. Kaplan *et al.* (1989) developed a representation of the dual processes of recombination and coalescence during a sweep, which allowed for stochastic effects on the frequency of the selected allele when it is either rare or very common. This approach enabled calculations of the effect of a sweep on both pairwise diversity and the site frequency spectrum, but did not provide simple formulae. Explicit formulae for the effect of a sweep on pairwise diversity that removed the assumption of a purely deterministic trajectory were derived by Stephan *et al.* (1992) using diffusion equations. Barton (1998, 2000) developed an alternative approach using a combination of branching processes and diffusion equations, from which the properties of a postsweep sample of n alleles could be calculated. Kaplan *et al.* (1989), Stephan *et al.* (1992), Wiehe and Stephan (1993), Barton (2000), Kim and Stephan (2000) and Gillespie (2002) also analyzed the effects of recurrent selective sweeps, treating coalescent events caused by classical genetic drift and by sweeps as competing exponential processes. All of these approaches assumed either a haploid population or an autosomal locus with semidominant fitness effects.

A great simplification in such calculations was achieved by the following approximation, proposed by Barton (1998, 2000) and extended by Durrett and Schweinsberg (2004)—see also Kim (2006) and Coop and Ralph (2012). This approach is based on two assumptions. The first is that the fixation of a favorable mutation happens so fast that nonrecombinant alleles at a linked neutral site, sampled after the completion of the sweep, effectively coalesce instantaneously. The second is that linkage is sufficiently tight that, at most, a single recombination event occurs during the sweep, placing a neutral site onto a wild-type background with which it remains associated throughout the sweep. These assumptions mean that the gene genealogy for a set of alleles sampled immediately after a sweep, and that failed to recombine onto the wild-type background, has a “star-like” shape. The reduction in diversity and site frequency spectrum at the neutral site can then be calculated in a straightforward fashion (Barton 2000; Durrett and Schweinsberg 2004; Kim 2006; Coop and Ralph 2012; Weissman and Barton 2012). This approximation provides the basis for detecting recent sweeps in the programs SweepFinder (Nielsen *et al.* 2005) and Sweed (Pavlidis *et al.* 2013). It can readily be incorporated into models of recurrent selective sweeps (Barton 2000; Weissman and Barton 2012; Berg and Coop 2015; Elyashiv *et al.* 2016; Campos *et al.* 2017; Campos and Charlesworth 2019), which has stimulated the development of methods for estimating the parameters of recurrent sweeps from population genomic data (Elyashiv *et al.* 2016; Campos *et al.* 2017; Campos and Charlesworth 2019).

This approach is likely to be accurate for favorable mutations that are sufficiently strongly selected that their time to fixation is short compared with the expected neutral coalescent time of $2N_e$ generations (where N_e is the effective population size), provided that the ratio of the recombination rate to the selection coefficient is sufficiently small. Conversely, when this ratio is large, a sweep will have a negligible effect on variability. There is a need to examine the properties of sweeps when the selection and recombination parameters do not meet these conditions, especially as recent population genomic analyses suggest that there may be important contributions from relatively weakly selected favorable mutations, which take as long as 10% or more of the neutral coalescent time to become fixed (Sella *et al.* 2009; Keightley *et al.* 2016; Chen *et al.* 2020). In such cases, the time to coalescence during the sweep cannot necessarily be neglected, and the assumption that a pair of nonrecombined alleles are identical in state leads to an underestimate of diversity at the end of the sweep, especially with very low rates of recombination. In contrast, coalescence during the sweep competes with recombination, so that calculating the probability that one of a pair of alleles recombines onto the wild-type background without including the probability that they have escaped prior coalescence underestimates the effect of a sweep (Barton 1998). More generally, when the assumption that the duration of a sweep is negligible compared with the neutral coalescent time is invalid, the mean coalescent time of a pair of alleles cannot accurately be calculated simply from the probability that they escape recombination onto the wild-type background.

The present paper describes a general analytical model of selective sweep effects on the mean time to coalescence of a pair of alleles at a linked neutral locus (which determines the expected pairwise neutral diversity), for the case of weak selection at a single locus, where the selection coefficient is sufficiently small that a differential equation can used instead of a difference equation. This is based on a recent study of the expected time to fixation of a favorable mutation in a single population (Charlesworth 2020), which provided a general framework for analyzing both autosomal and sex-linked inheritance with arbitrary levels of inbreeding and dominance. There are, of course, other statistics of importance for population genetic inferences, such as the effect of sweeps on site frequency spectra. Results on these are hard to obtain analytically without the use of the star phylogeny approximation (Barton 2000; Durrett and Schweinsberg 2004; Kim 2006; Coop and Ralph 2012), and are therefore not considered in this paper.

The resulting formulae, which include a heuristic treatment of multiple recombination events during a sweep, enable predictions of the effects on diversity of both a single sweep and recurrent selective sweeps, and allow for the action of BGS as well as sweeps. They apply to cases when the product of N_e and the strength of selection is sufficiently large that the expected trajectory of allele frequency change at the selected locus is close to the deterministic predictions, except for allele frequencies close to 0 or 1. Hartfield and Bataillon (2020) have recently presented similar results for an autosomal locus with coalescence during a sweep, in the case of a single sweep in the absence of BGS, but without modeling multiple recombination events. Only hard sweeps will be considered here, although it is straightforward to extend the models to soft sweeps by the approach of Berg and Coop (2015) and Hartfield and Bataillon (2020). The validity of the approximations is tested against computer simulations, including those of Campos and Charlesworth (2019) and Hartfield and Bataillon (2020). For the sake of brevity, these papers will be referred to as CC and HB, respectively.

Methods

Simulating the effect of a single sweep

The algorithm described by Equation 27 of Tajima (1990) was used to calculate the effects of a sweep on pairwise diversity at a neutral locus with an arbitrary degree of linkage to a selected locus with two alleles, A_1 and A_2, where A_2 is the selectively favored allele. A Wright–Fisher population with constant size N was assumed. The equations provide three coupled, forward-in-time recurrence relations for the expected diversities at the neutral locus for pairs of alleles carrying either A_1 or A_2, and for the divergence between A_1 and A_2 alleles. These are conditioned on a given generation-by-generation trajectory of allele frequencies at the selected locus, and assume an infinite sites model of mutation and drift (Kimura 1971).

The initial conditions for a simulation run were that a single A_2 allele was introduced into the population, with zero expected pairwise diversity at the associated neutral locus; the expected pairwise diversity among A_1 alleles and the divergence between A_1 and A_2 was equal to those for an equilibrium population in the absence of selection, $\theta = 4Nu$, where u is the neutral mutation rate. Since only diversities relative to θ are of interest here, θ was set to 0.001 in order to satisfy the infinite sites assumption for the neutral locus. The expected change in the frequency q of A_2 in a given generation for an assigned selection model was calculated using the standard discrete-generation selection formulation (see the section *Theoretical results—single sweep* for details of the models of selection). Binomial sampling using the frequency after selection and $2N$ as parameters was used to obtain the value of q in the next generation. Equation 27 of Tajima (1990) were applied to the old value of q in order to obtain the state of the neutral locus in the new generation.

This procedure was repeated generation by generation until A_2 was lost or fixed; only runs in which A_2 was lost were retained, and the value of the pairwise diversity among A_2 alleles at the time of its fixation was determined. This gives the expected diversity after a sweep conditional on a given trajectory, so that an estimate of the overall expected diversity relative to θ can be found by taking the mean over a large number of replicate simulations. It was found that 100 replicates were sufficient to produce a standard error of 2% or less of the mean. The value of N was chosen so that the selection coefficient s for a given value of the scaled selection parameter $\gamma = 2Ns$ was sufficiently small that terms of order s^2 could be neglected, to satisfy the assumptions of the model described in the section *Theoretical results—single sweep*.

Recurrent sweeps: simulation methods

For checking the theoretical predictions concerning recurrent sweeps, the simulation results described in CC were used. These involved groups of linked autosomal genes separated by 2 kb of selectively neutral intergenic sequence, with all UTR sites and 70% of nonsynonymous (NS) sites subject to both positive and negative selection, and the same selection parameters for 5′ and 3′ UTRs (see Figure 1 of CC). There were five exons of 300 basepairs (bp) each, interrupted by four introns of 100 bp. The lengths of the 5′and 3′ UTRs were 190 and 280 bp, respectively. The selection coefficients for favorable and deleterious mutations at the NS and UTR sites, and the proportions of mutations at these sites that were favorable, were chosen to match the values inferred by Campos *et al.* (2017) from the relation between the synonymous diversity of a gene and its rate of protein sequence evolution. Both favorable and deleterious mutations were assumed to be semidominant.

Five different rates of reciprocal crossing over (CO) were used to model recombination, which were chosen to be multiples of the approximate standard autosomal recombination rate in *Drosophila melanogaster*, adjusted by a factor of 1/2 to take into account the absence of recombinational exchange in males (Campos *et al.* 2017): 0.5×10^{-8}, 1×10^{-8}, 1.5×10^{-8}, 2×10^{-8}, and 2.5×10^{-8} cM/Mb, respectively, where 10^{-8} is the mean rate across the genome.

The simulations were run with and without BGS acting on both NS and UTR sites, and with and without noncrossover associated gene conversion events. Cases with gene conversion assumed a rate of initiation of conversion events of 1×10^{-8} cM/Mb for autosomes (after correcting for the lack of gene conversion in males), and a mean tract length of 440 bp, with an exponential distribution of tract lengths.

Recurrent sweeps at multiple sites: numerical predictions based on analytical formulae

A single gene is considered in the analytical models, so that a linear genetic map can be assumed, because there is a negligible frequency of double crossovers. The CO contribution to the frequency of recombination between a pair of sites separated by z basepairs is r_cz, where z is the physical distance between the neutral and selected sites and r_c is the CO rate CO per bp.

An important point regarding the cases with gene conversion should be noted here. CC stated that, because the simulation program they used (SLiM 1.8) modeled gene conversion by considering only events that are initiated on one side of a given nucleotide site, the rate of initiation of a gene conversion tract covering this site is one-half of that used in the standard formula for the frequency of recombination caused by gene conversion; see Equation 1 of Frisse *et al.* (2001). However, this statement is incorrect, because it overlooks the fact that the standard model of gene conversion assumes that there are equal probabilities of a tract moving toward and away from the site. If tracts are constrained to move in one direction, the net probability that a tract started at a random point moves toward a given site is the same as in the standard formula, for a given probability of initiation of a tract.

Since no derivation of the formula of Frisse *et al.* (2001) appears to have been given, one is provided in File S1, section 1, which makes this point explicit (Equation S5 is equivalent to the formula in question). Gene conversion tract lengths are assumed to be exponentially distributed, with a mean tract length of d_g, and a probability of initiation r_g. It follows that the effective rates of initiation of gene conversion events (r_g) used in the theoretical calculations in CC should have been twice the values that were used there. Diversity values were thus underestimated by these calculations, because there was more recombination than was included in the predictions. The correct theoretical results for sweep effects are presented here.

The effects of selective sweeps on neutral sites within a gene were obtained by summing the expected effects of substitutions at each NS and UTR site in the gene on a given neutral site (synonymous site), assuming that every third basepair in an exon is a neutral site, with the other two (NS) sites being subject to selection, as described by Campos *et al.* (2017). This differs from the SLiM procedure of randomly assigning selection status to exonic sites, with a probability p_s of being under selection ($p_s = 0.7$ in the simulations used in CC). To correct for this, the overall rate of NS substitutions per NS site was adjusted by multiplication by 0.7×1.5. Furthermore, to correct for the effects of interference among co-occurring favorable mutations in reducing their probabilities of fixation, their predicted rates of substitution were multiplied by a factor of 0.95, following the procedure in CC.

In order to speed up the computations, mean values of the variables used to calculate the effects of sweeps on neutral diversity were calculated by thinning the neutral sites by considering only a subset of them, starting with the first codon at the 5′end of the gene. For the results reported here, 10% of all neutral sites were used to calculate the values of the variables. Comparisons with results from using all sites showed a negligible effect of using this thinning procedure.

Background selection effects on diversity for autosomes and X chromosomes for genes in regions with different CO rates were calculated as described in sections S9 and S10 of File S1 of CC, which included estimates of the effects of BGS caused by selectively constrained noncoding sequences as well as coding sequences, derived from (Charlesworth 2012). If gene conversion was absent, the correction factors for gene conversion used to calculate these effects were omitted.

Data availability statement

The author states that all data necessary for confirming the conclusions presented in the article are represented fully within the article. The author states that no new data or reagents were generated by this research. Details of some of the mathematical derivations are described in the Supplementary Information, File S1 on Figshare. The codes for the computer programs used to implement the analytical models described below are available in the Supplementary Information, File S2 on Figshare. The detailed statistics for the results of the computer simulations shown in Figure 3 were provided in Files S2–S3 of Campos and Charlesworth (2019). Supplemental material available at figshare: https://doi.org/10.25386/genetics.13136012.

Results

The effect of a single sweep on expected nucleotide site diversity

Theoretical results—single sweep: The aim of this section is obtain an expression for the mean coalescent time at a neutral site linked to a selected locus, at the time of fixation of the selectively favored allele; under the infinite sites model, this yields the expected pairwise diversity at the neutral site. All times are expressed on the coalescent timescale of $2N_e$ generations, where N_e is the neutral effective population size for the genetic system under consideration (autosomal or X-linked loci, random mating, or partial inbreeding). If we use N_{e0} to denote the value of N_e for a randomly mating population with autosomal inheritance, N_e for a given genetic system can be written as kN_{e0}, where k depends on the details of the system in question (Wright 1939, 1969; Crow and Kimura 1970; Charlesworth and Charlesworth 2010). For example, with an autosomal locus in a partially inbreeding population with Wright's fixation index $F > 0$, we have $k \approx 1/(1+F)$ under a wide range of conditions (Pollak 1987; Nordborg 1997; Laporte and Charlesworth 2002). In addition, following Kim and Stephan (2000) and CC, if BGS is operating, it is assumed that, for purely neutral processes, N_e can replaced by the quantity B_1N_e, where B_1 measures the effect of BGS on the mean neutral coalescent time of a pair of alleles. The effect of BGS on the fixation probabilities of favorable mutations is likely to be somewhat less than that for neutral processes, so that a second coefficient, B_2, should ideally be used as a multiplier of N_e, where $B_2 = B_1/\lambda$ ($\lambda \leq 1$). As discussed in CC, B_1 can be determined analytically for a given genetic model, whereas B_2 usually requires simulations, so it is often more convenient to use B_1 for both purposes, although this procedure introduces some inaccuracies.

As has been discussed in previous treatments of sweeps, there are two stochastic phases during the spread of a favorable mutation, A_2, in competition with a wild-type allele, A_1. A detailed analysis of these stochastic phases for the general model of selection used here is given by Charlesworth (2020). In the first phase, the frequency of A_2 is so low that it is subject to random fluctuations that can lead to the loss of A_2 from the population. Provided that the product of N_e and the selection coefficient for homozygotes for the favorable allele (s) is $>>1$, a mutation that survives this phase will enter the deterministic phase, where it has a negligible probability of loss, and in which its trajectory of allele frequency change is well approximated by the deterministic selection equation (Equation 6 below). When A_2 reaches a frequency close to 1, A_1 is now vulnerable to stochastic loss, so that there is a second stochastic phase. Formulae for the frequencies of A_2 at the boundaries of the two stochastic phases, q_1 and q_2, are given by Charlesworth (2020), together with expressions for the durations of the stochastic and deterministic phases. For mutations with intermediate levels of dominance, q_1, 1 – q_2 and the durations of the two stochastic phases are all of the order of $1/(2N_es)$, measured on the coalescent timescale of $2N_e$ generations.

If q_2 is close to 1, A_2 has only a small chance of encountering an A_1 allele, so that there is a negligible chance that a neutral site in a haplotype carrying A_2 will recombine onto a background recombination during the final stochastic phase. In addition, the rate of coalescence within haplotypes carrying A_2 is then close to the neutral value, and so does not greatly affect the mean time to coalescence of a pair of alleles sampled after the end of the sweep compared with neutral expectation. Under these conditions, the second stochastic phase has little effect on the mean coalescent time of the alleles compared with neutral expectation. Provided that the duration of the first stochastic phase on the coalescent time scale is $<<1$ (*i.e.*, q_1 is close to 0), this phase will also have a minimal impact on the mean coalescent time of such a pair of alleles. Accurate approximations for the effect of a single sweep on diversity can, therefore, usually be obtained by treating the beginning and end of the deterministic phase as equivalent to that for the sweep as a whole, as discussed by Charlesworth (2020).

The general framework presented in HB can then be used to determine the effect of a sweep on pairwise diversity, extended to include a more general model of selection as well as the possibility of BGS effects, and using analytical expressions for probabilities of coalescence and recombination during the sweep rather than numerical evaluations. This approach assumes that all evolutionary forces are weak (*i.e.*, second order terms in changes in allele frequencies and linkage disequilibrium can be neglected), so that a continuous time scale approximation can be applied.

Let T_d be the duration of the deterministic phase, defined as the period between frequencies q_1 and q_2 as given by Charlesworth (2020). With BGS, the terms in N_e in the relevant expressions are each to be multiplied by B_2, as was done in CC. For a pair of haplotypes that carry the favorable allele A_2 at the end of the sweep, the rate of coalescence at a time T back from this time point is $[B_1q(T)]^{-1}$, where $q(T)$ is the frequency of A_2 at time T. The rate at which a linked neutral site recombines from A_2 onto the wild-type background at time T is $\rho[1-q(T)] = \rho p(T)$, where $\rho = 2N_er$ is the scaled recombination rate and r is the absolute recombination rate between the selected and neutral loci. With inbreeding and/or sex-linkage, r differs from its random mating autosomal value, r_0, such that $r = cr_0$, where c is a function of the genetic system and mating system. For example, with autosomal inheritance with partial inbreeding, $c \approx 1 - 2F + \phi$, where ϕ is the joint probability of identity by descent at a pair of neutral loci (Roze 2009; Hartfield and Bataillon 2020). Unless both r_0 and F are sufficiently large that their second-order terms cannot be neglected, we have $c \approx 1 - F$ (Nordborg 1997; Charlesworth and Charlesworth 2010, p. 381). The exact value of ϕ is determined by the mating system; in the case of self-fertilization, Equation 1 of HB gives an expression for ϕ as a function of r_0 and the rate of self-fertilization, which is used in the calculations presented here.

Under these assumptions, the probability density function (p.d.f.) for a coalescent event at time T for a pair of alleles sampled at the end of the sweep is:

$$k_c(T) = [B_1q(T)]^{-1}P_{nc}(T)P_{nr}(T) \tag{1}$$

where $P_{nc}(T)$ is the probability of no coalescence by time T in the absence of recombination, and $P_{nr}(T)$ is the probability that neither allele has recombined onto the wild-type background by time T, in the absence of coalescence.

Similarly, the p.d.f. for the event that one of the two sampled haplotypes recombines onto the wild-type background at time T (assuming that r is sufficiently small that simultaneous recombination events can be ignored) is given by:

$$k_r(T) = 2\rho p(T)P_{nc}(T)P_{nr}(T) \tag{2}$$

We therefore have:

$$P_{nc}(T) = \exp - \int_0^T [B_1q(\tau)]^{-1}\, \mathrm{d}\tau \tag{3}$$

$$P_{nr}(T) = \exp - 2\rho \int_0^T p(\tau)\, \mathrm{d}\tau \tag{4}$$

The net probability that the pair of sampled alleles coalesce during the deterministic phase of the sweep is given by:

$$P_{c1} = \int_0^{T_d} k_c(T)\, \mathrm{d}T \tag{5a}$$

If it is assumed that haplotypes that have neither recombined nor coalesced during the sweep coalesce with probability one at the start of the sweep, there is an additional contribution to the coalescence probability, given by:

$$P_{c2} \;= P_{nc}(T_d)P_{nr}(T_d) \tag{5b}$$

The net probability of coalescence caused by the sweep is thus:

$$P_c = P_{c1} + P_{c2} \tag{5c}$$

These equations are simple in form, but getting explicit formulae is made difficult by the nonlinearity of the equation for the rate of change of q under selection. Following Charlesworth (2020), for the case of weak selection (when terms of order s^2 can be ignored) we can write the forward-in-time selection equation as:

$$\dot{q}\left(\tilde{T}\right) \approx \gamma p\left(\tilde{T}\right) q\left(\tilde{T}\right)\left[a + bq\left(\tilde{T}\right)\right] \tag{6}$$

where tildes are used to denote time measured from the start of the sweep; $\gamma = 2N_e s$ is the scaled selection coefficient for A_2A_2, assigning a fitness of 1 to A_1A_1 and an increase in relative fitness of s to A_2A_2. Here, a and b depend on the dominance coefficient h and fixation index F, the genetic and mating systems, and the sex-specificity of fitness effects (Glémin 2012; Charlesworth 2020). For example, for an autosomal locus, the weak selection approximation gives $a = F + (1 - F)h$ and $b = (1 - F)(1 - 2h)$.

For $a > 0$ and $a + b > 0$, corresponding to intermediate levels of dominance, integration of Equation 6 yields the following expression for the expectation of the duration of the deterministic phase, T_d (Charlesworth 2020):

$$T_d \approx \gamma^{-1}\left\{a^{-1}\ln\left[\frac{q_2(a+bq_1)}{q_1(a+bq_2)}\right] + (a+b)^{-1}\ln\left[\frac{p_1(a+bq_2)}{p_2(a+bq_1)}\right]\right\} \tag{7}$$

Here, $q_1 \approx 1/2a\gamma$ and $p_2 \approx 1/2(a + b)\gamma$ (Charlesworth 2020).

Similar expressions are available for the cases when $a = 0$ (complete recessivity) or $a + b = 0$ (complete dominance), as described by Charlesworth (2020); see Equations A1b and A1c, respectively.

Using Equation 6, we can write T as a monotonic function of q, $T(q)$. Substituting q for T and using the relation $dT = \dot{q}^{-1}$ dq. Equations 3, 4, and 5a then become:

$$P_{c1} = \int_{q_1}^{q_2} [B_1 \ \dot{q}q]^{-1} P_{nc}(q) P_{nr}(q)\,\mathrm{d}q \tag{8a}$$

$$P_{nc}(q) = \exp - \int_{q}^{q_2} \dot{x}^{-1}[B_1 x]^{-1}\mathrm{d}x \tag{8b}$$

$$P_{nr}(q) = \exp - 2\rho \int_{q}^{q_2} \dot{x}^{-1}(1 - x)\mathrm{d}x \tag{8c}$$

Explicit formulae for $P_{nc}(q)$ and $P_{nr}(q)$ are given in the Appendix (Equations A2 and A3).

Substituting q_1 for q in Equations 5b and 5c, Equation 5b can be written as:

$$P_{c2} = P_{nc}(q_1) P_{nr}(q_1) \tag{8d}$$

The net expected pairwise coalescence time associated with the sweep, T_s, includes a contribution from the case when no coalescence occurs until the start of the sweep, given by the product of P_{c2} and T_d, and a contribution from coalescent events that occur during the sweep, denoted by T_c. We have:

$$T_s = P_{c2} T_d + T_c \tag{9a}$$

where

$$T_c = \int_{q_1}^{q_2} \dot{q}^{-1}[B_1 q]^{-1} T(q) P_{nc}(q) P_{nr}(q)\,\mathrm{d}q \tag{9b}$$

and $T(q)$ is the time to reach frequency q of A_2, given by Equations A1.

Results with only a single recombination event: The possibility of recombination back onto the background of A_2, examined in CC, is ignored for the present, as is the possibility of a second recombination event from A_2 onto A_1. From Equation 2, the probability of at least one recombination event is given by:

$$P_r = 2\rho \int_{q_1}^{q_2} \dot{q}^{-1} p P_{nc}(q) P_{nr}(q)\,\mathrm{d}q. \tag{10}$$

Using Equations 6 and 10a and A1-A3, P_r can be expressed explicitly in terms of ρ, γ, a, and b, but the resulting expression has to be evaluated numerically.

The net expected pairwise coalescence time in the presence of BGS under this set of assumptions is given by $B_1P_r + T_s$. Under the infinite sites model (Kimura 1971), the expected reduction in pairwise nucleotide site diversity for alleles sampled at the end of the sweep, relative to its value in the absence of selection (θ), is given by:

$$-\boldsymbol{\Delta}\pi = B_1(1 - P_r) - T_s. \tag{11a}$$

Equation 9 of HB for the case of a hard sweep is equivalent to Equation 11a without the term in T_s. In addition, if T_s and the probability of coalescence during the sweep are both negligible, it is easily seen that $P_r \approx 1 - P_{nr}(T_d)$, yielding the following result for the star phylogeny approximation (Barton 1998, 2000; Durrett and Schweinsberg 2004; Weissman and Barton 2012):

$$-\boldsymbol{\Delta}\pi \approx B_1 P_{nr}(q_1) \tag{11b}$$

In the case of an autosomal locus with random mating and semidominant selection ($h = 0.5$), this yields the following convenient formula:

$$-\boldsymbol{\Delta}\pi \approx B_1 \gamma^{-4\rho/\gamma} \tag{11c}$$

As mentioned in the Introduction, this formula has been used in several methods for making inferences from population genomic data.

The importance of coalescence during a sweep: These results bring out the potential importance of considering coalescence during a sweep, as opposed to the coalescence of nonrecombined alleles at the start of a sweep. Consider the case with incomplete dominance ($a \neq 0$). The probability of no coalescence during the sweep conditional on no recombination, $P_{nc}(q_1)$, is given by Equation A2a with $q = q_1$, where $q_1 \approx (2a\gamma)^{-1}$ (Charlesworth 2020). Somewhat surprisingly, for large γ this expression becomes independent of a and γ, provided that $a^{-2} >> \gamma$, and approaches $e^{-2} \approx 0.135$, so that the probability of coalescence during a sweep in the absence of recombination is ~0.865 (see the Appendix). With low rates of recombination, there is thus a high probability of coalescence during the sweep itself, in contrast to what is assumed in Equation 11b and 11c. If such a coalescent event is not preceded by a recombination event, the mean coalescent time will thus be smaller than predicted by these Equations.

This raises the question of the magnitude of T_s in the more exact treatment. While Equation 9 can only be evaluated exactly by numerical integration, a rough estimate of T_s for the case of no recombination can be obtained as follows (this is the maximum value, as the terms involving the probability of no recombination must decrease with the frequency of recombination). By the above result for $P_{nc}(q_1)$, the first term in Equation 9 is approximately $e^{-2}T_d$. The second term is equivalent to the mean coalescent time associated with events during the sweep; by the argument presented in section S3 of File S1 in CC, this is approximately equal to the harmonic mean of $1/q$ between q_1 and q_2. Equation S10 of CC for this quantity can be generalized as shown in the Appendix, with the result that the expected coalescent time associated with the sweep (T_c) is approximately $\frac{1}{2}T_d$ for large γ, giving $T_s \approx 0.635T_d$.

Table 1 and Supplemental Material, Table S1 of File S1 compare the results from numerical integrations with this approximation; as expected from the assumptions made in deriving this approximation, it is most accurate when γ is large and a is not too close to 1. Overall, for low frequencies of recombination, T_s is a non-negligible fraction of T_d, but decreases toward zero with increasing rates of recombination, as would be expected.

Multiple recombination events: Finally, the problem of multiple recombination events needs to be considered. In principle, this problem can be dealt with on the lines of Equation 10, but this involves multiple integrals of increasing complexity as more and more possible events are considered. The following heuristic argument can be used instead. A first approximation is to assume that, if the frequency of recombination is sufficiently high, multiple recombination events are associated with a coalescent time equal to that of an unswept background, B_1. In contrast, a single recombinant event is associated with a mean coalescent time of $B_1 + T_d$, since the recombinant cannot coalesce with the nonrecombinant haplotype until the end of the sweep. If the probability of a single recombinant event is denoted by P_{rs}, Equation 11a is replaced by:

$$-\boldsymbol{\Delta}\pi = B_1(1 - P_r) - T_s - P_{rs}T_d \qquad (12)$$

P_{rs} is given by the probability of a recombination event that is followed by no further recombination events. This event requires both the recombinant A_1 haplotype (whose rate of recombination at an A_2 frequency of x is ρx) and the nonrecombinant A_2 haplotype (whose rate of recombination is $\rho[1 - x]$) to fail to recombine.

We thus have:

$$P_{rs} = 2\rho \int_{q_1}^{q_2} \dot{q}^{-1} p P_{nc}(q) P_{nr}(q) P_{nr}(q_1, q)\, \mathrm{d}q \qquad (13a)$$

where $P_{nr}(q_1, q)$, is the probability of no further recombination after an A_2 frequency of q, given by:

$$\begin{aligned} P_{nr}(q_1, q) &= \exp - \rho \int_{q_1}^{q} \dot{x}^{-1}\Big\{(1 - x) + x\Big\}\ \mathrm{d}x \\ &= \exp - \rho\,[T(q_1) - T(q)] \end{aligned} \qquad (13b)$$

However, Equation 12 ignores the fact that there is a time-lag until the initial recombination event, whose expectation, conditioned on the occurrence of the initial recombination event, is denoted by T_r. This lag contributes to the time to coalescence of multiple recombinant alleles, causing the reduction in diversity to be smaller than predicted by Equation 12b. The probability of multiple recombination events is (P_r – P_{rs}), so that a better approximation is to deduct (P_r – P_{rs})T_r from the left-hand side of Equation 12, giving:

$$-\boldsymbol{\Delta}\pi = B_1(1 - P_r) - T_s - P_{rs}T_d - (P_r - P_{rs})T_r \qquad (14a)$$

where

$$T_r = 2\rho P_r^{-1} \int_{q_1}^{q_2} \dot{q}^{-1} p T(q) P_{nc}(q) P_{nr}(q)\, \mathrm{d}q. \qquad (14b)$$

The integral for T_r can be expressed in terms of ρ, γ, a, and b, on the same lines as for Equation 10.

Equation 14 are likely to overestimate the effect of recombination on the sweep effect, as complete randomization of the sampled pair of haplotypes is unlikely to be achieved, whereas Equation 11a clearly underestimates it; Equation 12 should produce an intermediate prediction. The correct result should thus lie between the predictions of Equation 11 and Equation 14. When the ratio of the rate of recombination to the selection coefficient, r/s, is $<<1$, all three expressions agree, and predict a slightly smaller sweep effect than Equation 9 of HB.

Comparisons with simulation results

Numerical results for Equation 11 can be obtained by numerical integration of the formulae given in the Appendix. For speed of computation, Simpson's rule with $n + 1$ points was used here; this method approximates the integral of a function by a weighted sum of discrete values of the integrand over n equally spaced subdivisions of the range of the function (Atkinson 1989). It was found that $n = 200$ usually gave values that were close to those for a more exact method of integration; for the results in the figures in this section, $n = 2000$ was used. Background selection effects are

Table 1 Parameters describing the effect of a single sweep

γ = 250							
r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c/P_{c1}	T_r
h = 0.1, T_d = 0.111, approx. T_s = 0.0703, P_{nc} = 0.229							
0	0.771	1	0	0	0.0594	0.0770	—
0.04	0.354	0.226	0.595	0.346	0.0293	0.0669	0.0359
0.08	0.193	0.0571	0.796	0.244	0.0121	0.0562	0.0477
0.16	0.0895	0.0026	0.910	0.0645	0.0036	0.0392	0.0393
0.32	0.0471	0.0000	0.953	0.0027	0.0011	0.0244	0.0266
0.64	0.0313	0.0000	0.969	0.0000	0.0005	0.0017	0.0259
1.28	0.0237	0.0000	0.977	0.0000	0.0003	0.0013	0.0215
h = 0.5, T_d = 0.0883, approx. T_s = 0.0561, P_{nc} = 0.125							
0	0.875	1	0	0	0.0637	0.0728	—
0.04	0.491	0.413	0.457	0.335	0.0385	0.0690	0.0556
0.08	0.293	0.171	0.686	0.353	0.0207	0.0640	0.0528
0.16	0.131	0.0293	0.866	0.200	0.0071	0.0517	0.0476
0.32	0.0589	0.0009	0.941	0.0341	0.0019	0.0320	0.0407
0.64	0.0386	0.0000	0.962	0.0006	0.0008	0.0210	0.0339
1.28	0.0297	0.0000	0.971	0.0000	0.0004	0.0162	0.0279
h = 0.9, T_d = 0.118, approx. T_s = 0.0749, P_{nc} = 0.118							
0	0.882	1	0	0	0.0860	0.0972	—
0.04	0.531	0.485	0.412	0.307	0.0557	0.0929	0.0792
0.08	0.335	0.235	0.637	0.346	0.0323	0.0872	0.0757
0.16	0.160	0.0554	0.834	0.223	0.0122	0.0719	0.0690
0.32	0.0746	0.0031	0.926	0.0489	0.0034	0.0449	0.0582
0.64	0.0486	0.0000	0.954	0.0013	0.0014	0.0296	0.0451
1.28	0.0341	0.0000	0.969	0.0000	0.0008	0.0221	0.0324

T_d and T_s are the expected durations of the deterministic phase of the sweep and pairwise coalescent time associated with the sweep, respectively; P_{nc} is the probability of no coalescence during the sweep, in the absence of recombination; P_{nr} is the probability of no recombination during the sweep, in the absence of coalescence; P_{c1} is the probability of coalescence during the sweep; P_r is the probability of at least one recombination event during the sweep; T_c/P_{c1} is the mean time to coalescence during the sweep, conditioned on coalescence; T_r is the mean time to the first recombination event, conditioned on the occurrence of a recombination event. The approximate value of T_s is for the case of no recombination, and = $0.635T_d$ (see text).

ignored here, so that B_1 and B_2 are set to 1. Simulation results for hard sweeps for an autosomal locus with random mating were obtained using the algorithm of Tajima (1990) (see the *Methods* section), providing a basis for comparison with the predictions based on Equations 11a and 14 (denoted by *C1* and *C2*, respectively), and on the star phylogeny approximation that ignores coalescence of nonrecombined alleles during the sweep, Equation 11b (*NC*). The results are shown in Figure 1, with the reduction in diversity observed at the end of the sweep, $-\Delta\pi$, on a $\log_{10}$ scale, plotted against *r/s* on a $\log_2$ scale, with values of *r/s* increasing by a factor of two from 0.04 to 1.28. Figure S1 of File S1 shows the same results using linear plots with *r/s* between 0 and 0.32, with the addition of the values of $-\Delta\pi$ for $r/s = 0$. Since *C1* and Equation 9 of HB, which ignore the term in T_s in Equation 11a, gave very similar results, only results from *C1* are shown here.

One feature that is worth noting is that, with no recombination, the simulations and *C1*/*C2* formulae (which are identical and exact in this case) predict $-\Delta\pi$ values that are substantially <1, especially with the lower strengths of selection. The *NC* approximation predicts a complete reduction in diversity, since the probability of coalescence is 1, and the duration of the sweep is ignored (this can be seen most clearly in the linear plots in Figure S1). In contrast, *NC* underestimates $-\Delta\pi$ when *r/s* is sufficiently large; even for *r/s* as small as 0.16 there can be a very large ratio of the simulation value to the *NC* value, although the simulation value of $-\Delta\pi$ is then usually quite small (10% or less) for this value of *r/s*. For example, with $\gamma = 250$, $h = 0.5$, and $r/s = 0.16$, the simulation value of $-\Delta\pi$ was 0.0959 (SE 0.0019), whereas the *NC* value was 0.0293. Conversely, *C1* tends to overestimate $\Delta\pi$ for the higher values of *r/s*, reflecting the fact that it does not allow for multiple recombination events.

Overall, *C2* agrees quite well with most of the simulation results, especially for $h = 0.5$, but tends to underestimate $-\Delta\pi$ for $h = 0.9$, especially for large *r/s*, presumably because the relatively long period which A_2 spends at high frequencies means that a substantial proportion of multiple recombination events involve a return of a recombined neutral site back onto the A_2 background, For much larger *r/s* values than those shown here, *C2* can become negative, indicating that it overcorrects for multiple recombination events, but $-\Delta\pi$ is then very small, so that this effect is probably not biologically important. As has been found previously (Teshima and Przeworski 2006; Ewing *et al.* 2011; Hartfield and Bataillon 2020), $-\Delta\pi$ increases with *h* for low values of *r/s*, but the values for each *h* converge as *r/s* increases.

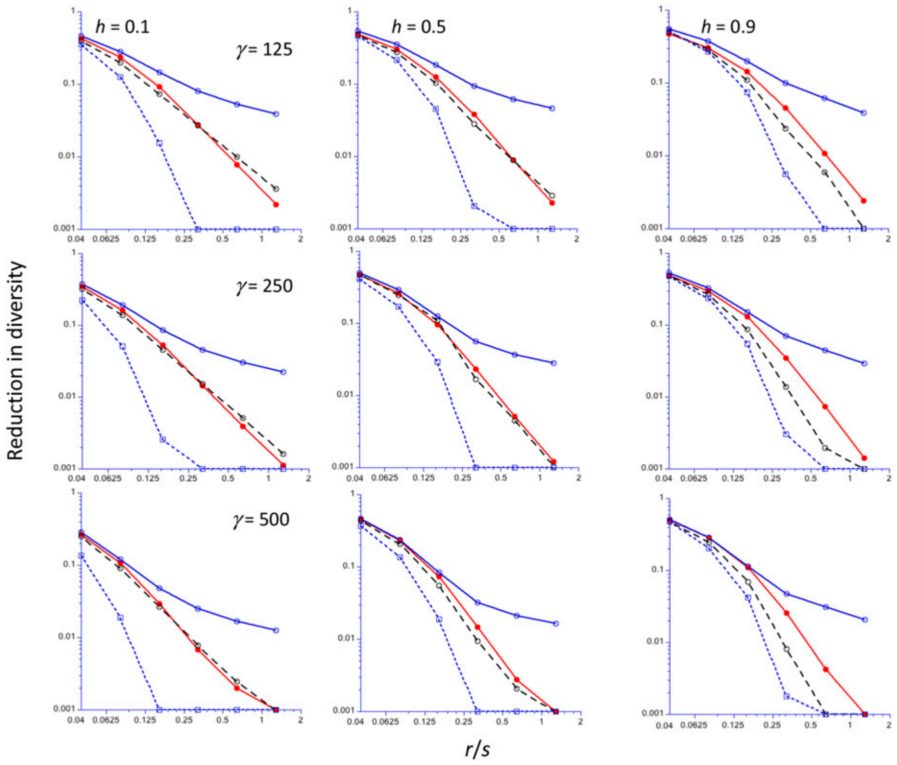

Figure 1 The reduction in diversity (relative to the neutral value) at the end of a sweep for an autosomal locus (y-axis, $\log_{10}$ scale), as a function of the ratio of the frequency of recombination (r) to the selection coefficient for homozygotes (s) (x-axis, $\log_2$ scale). A population size of 5000 is assumed, with three different values of the scaled selection coefficient ($\gamma = 2N_es$): 125 (top panel), 250 (middle panel), and 500 (bottom panel), and three different values of the dominance coefficient (h), increasing from left to right. The filled red circles are the mean values from computer simulations, using the algorithm of Tajima (1990); the open blue circles and black circles are the *C1* and *C2* predictions, respectively; the open blue squares are the *NC* predictions. Values of the reduction in diversity <0.001 have been reset to 0.001.

Table 1 and Table S1 show details of some of the relevant variables, obtained by numerical integration. They confirm the conclusion that there can be a substantial probability of coalescence during the sweep, as given by P_{c1} in Equation 8a; this probability decreases much more slowly with r/s than does the probability of no recombination in the absence of coalescence (P_{nr}). In parallel with this behavior of P_{c1}, the unconditional probability of no recombination, 1 – P_r, decreases much more slowly with r/s than P_{nr}. This explains why the *NC* approximation for the reduction in diversity performs rather poorly at high r/s values. The results also show that the probability of a single recombination event (P_{rs}, given by Equation 13a) becomes very small compared with the probability of at least one recombination event (P_r, given by Equation 10a) as r/s increases, so that neglecting the effects of multiple recombination events leads to errors in predicting sweep effects on diversity. For high values of r/s, the conditional mean times to coalescence and to the first recombination event are both small relative to the duration of the sweep, implying that these events must occur quite soon if they are to occur at all.

To illustrate the approximations further, both Tajima and HB simulation values of $-\Delta\pi$ for an autosomal locus in a randomly mating population with three difference dominance coefficients and $\gamma = 500$, together with the theoretical predictions, are shown in Figure S2 of File S1. These confirm the general conclusions from Figure 1, despite the fact that the HB simulation results seem to be considerably noisier than the Tajima results, sometimes showing a nonmonotonic relation between $-\Delta\pi$ and the recombination rate.

Figure 2 displays results for selfing rates of 0.5 and 0.95, corresponding to F values of 0.3333 and 0.9048, respectively. The reduction in diversity is plotted against the scaled recombination rate for an autosomal locus with outbreeding; the scaled effective recombination rate with inbreeding is much smaller than this, as described above. Here, only the HB simulation results are shown, as the Tajima method cannot give an exact representation of the system with nonrandom mating. As before, *C1* and the approximation given by Equations 9 of HB mostly give very similar predictions, whereas *C2* predicts smaller effects that agree slightly less well with the simulations at the higher recombination rates. However, for $S = 0.95$, especially with $h = 0.5$, the simulations yield considerably larger sweep effects at relatively high recombination rates than any of the theoretical predictions. This discrepancy presumably reflects the fact that random variation among individuals in the occurrence of selfing *vs.* outcrossing events means that individuals sampled in a given generation differ in the numbers of generations of selfing in their ancestral lineages, and, hence, in the extent to which recombination and selection have interacted to cause departures from neutral expectations (Kelly 2007). This is not taken into account in the formula used to correct for the effects of selfing on the effective rate of recombination (Equation 1 in HB).

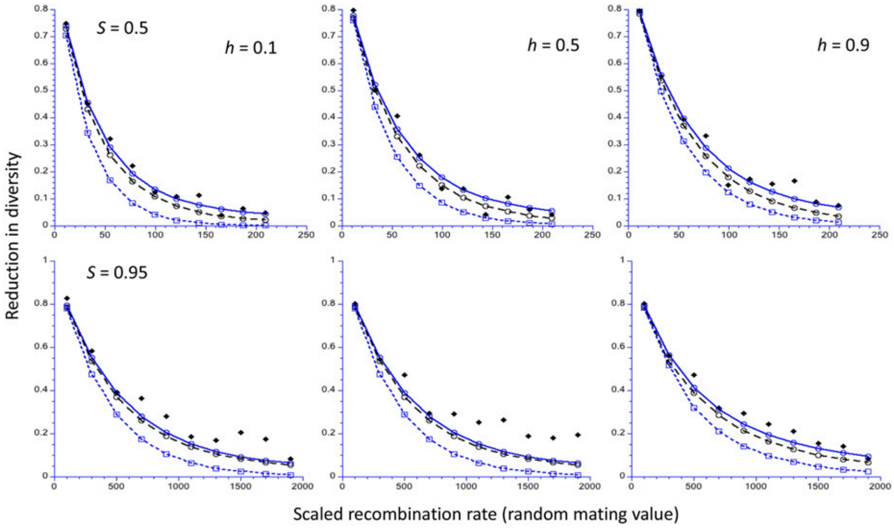

Figure 2 The reduction in diversity (relative to the neutral value) at the end of a sweep for an autosomal locus, as a function of the scaled rate of recombination ($2N_er$) for a randomly mating population. The results for two different selfing rates, *S*, are shown in the upper and lower panels, respectively, together with three different values of the dominance coefficient, *h*. A population size of 5000 is assumed, with a scaled selection coefficient for A_2A_2 homozygotes in a randomly mating population ($\gamma = 2N_es$) of 500. The filled black lozenges are the mean values from the computer simulations of Hartfield and Bataillon (2020); the open blue circles and open black circles are the *C1* and *C2* predictions, respectively; the open blue squares are the *NC* predictions.

Inaccuracy of the NC approximation: Given the widespread use of the star phylogeny assumption in methods for detecting recent sweeps and inferring the parameters of positive selection, described in the Introduction, it is disconcerting that the *NC* approximation is systematically somewhat inaccurate with respect to pairwise diversity at relatively large values of r/s. Some insights into this effect can be obtained from examining an approximation for the case of autosomal inheritance with $h = 0.5$ and random mating, which is derived in section 2 of File S1.

If we write $R = 4r/s$ and $\alpha = 2/\gamma$, and assume that $R > 1$ and $\gamma >> 1$, the reduction in diversity after a sweep is close to 1 minus the probability of recombination during the sweep, as given by Equation 11a with $T_s = 0$. From Equation S9 of File S1, we have:

$$-\Delta\pi \approx 1 - e^{\alpha}\gamma^{-\alpha} - e^{\alpha}\gamma^{-\alpha}\alpha R\left\{2[(R-1)(R+1)]^{-1} - \sum_{i=2}^{\infty}[i(R+i)]^{-1}\right\}. \qquad (15)$$

This series converges quite slowly when *R* is large, but the formula agrees well with the results of numerical integration even for r/s as low as 0.4, provided that γ is sufficiently large. It tends to break down for high values of *R* (>10), especially for relatively small γ. Equation 15 implies that, paradoxically, *larger* values of γ lead to *smaller* values of the diversity reduction for a given r/s, as can be seen as follows. For large γ, $1 - e^{\alpha}\gamma^{-\alpha} \approx \alpha \ln(\gamma)$ for large γ, which is nearly proportional to α. In addition, the term in braces is positive for sufficiently large *R*, and its product with $\alpha e^{\alpha}\gamma^{-\alpha}$ is also nearly proportional to α.

This effect can be seen in Figure 1 and Figure S1. A doubling of γ results in substantially smaller values of the diversity reduction for a given value of r/s, as expected from the above properties of Equation 15. This applies to both the *C1* and *C2* predictions, as well as the simulations. Thus, contrary what is predicted by the *NC* approximation, the effect of a sweep on diversity is negatively related to the scaled strength of selection, for a given value of r/s. The intuitive interpretation of this finding is that weaker selection prolongs the duration of a sweep, allowing more opportunity for coalescence *vs.* recombination.

The effects of recurrent selective sweeps on nucleotide site diversity

Theoretical results—recurrent sweeps: The approach of CC for determining the effects of recurrent sweeps, which was based on the *NC* approximation, can be modified to apply to the more general case considered here. It is assumed that adaptive substitutions occur at a total rate ω per $2N_e$ generations, such that the times between substitutions follow an exponential distribution with rate parameter ω (this rate includes any effects of BGS in reducing the probability of fixation of favorable mutations). By summing up over all relevant nucleotide sites that contribute to the effect of sweeps at a focal neutral site, weighting each selected site by its rate of adaptive substitution (which may differ according to the class of site subject to selection), and then dividing by ω, we can define expected values of P_r, P_{rs}, T_s, T_r, and T_d for a given neutral site (expected values are denoted by overbars in what follows).

As a first step, it is useful to note that Equation 8 of CC for the expected nucleotide site diversity immediately after a substitution, π_0, is equivalent to:

$$\pi_0 \approx \bar{P}_r\pi_1 \qquad (16)$$

where π_1 is the expected nucleotide site diversity at the time of the initiation of a new substitution. Both π_0 and π_1 are measured relative to the neutral diversity θ, and hence are equivalent to mean pairwise coalescent times relative to the neutral value, $2N_e$. This expression assumes that there is at most a single recombination event, and that a pair of alleles that have been separated by recombination onto the A_1 and A_2 backgrounds have a coalescent time equivalent to that for a pair of alleles that are sampled at the start of the sweep.

If we apply the argument leading to Equation 11a to take into account the lag time to coalescence of a pair of alleles separated by recombination, we obtain the *C1* approximation:

$$\pi_0 \approx \bar{P}_r \pi_1 + \bar{T}_s \tag{17}$$

If we use the approach of Equations 14 for modeling multiple recombination events, we obtain the *C2* approximation:

$$\pi_0 \approx \bar{P}_r \pi_1 + \bar{T}_s + \overline{P_{rs} T_d} + \overline{(P_r - P_{rs}) Tr} \tag{18}$$

A somewhat more accurate expression can be found by noting that, under the assumption that multiple recombination events cause randomization between A_1 and A_2 haplotypes, so that coalescence occurs at rate B_1^{-1}, diversity will increase from its value at the start of a sweep over a time interval that is approximately the same as the difference between the sweep duration and the time of the first recombination event (see File S1, section 3); this overestimates the time available for the increase in diversity, since a time greater than T_r is required for randomization to occur.

Equations S12–S15 enable us to find the expected diversity, π, for a pair of alleles sampled at a random point in time. This is done by assuming that such a time point falls between two sweeps, and that the length of the interval T separating the two sweeps follows an exponential distribution with parameter ω. Conditional on T, the time τ from a random sample to the first of the two sweeps is a uniform variate on the interval T, with p.d.f. equal to $1/T$. The expected diversity at time τ (on the coalescent timescale) is given by the equivalent of Equation 9 in CC:

$$1 - \pi(\tau)(B_1\theta)^{-1} \approx \left[1 - \pi_0(B_1\theta)^{-1}\right]\exp\left(-B_1^{-1}\tau\right) \tag{19}$$

The overall expected diversity is thus given by:

$$1 - \pi(B_1\theta)^{-1} \approx \left[1 - \pi_0(B_1\theta)^{-1}\right]\omega \int_0^\infty T^{-1}\exp(-\omega T) \times \int_0^T \exp\left(-B_1^{-1}\tau\right)\mathrm{d}\tau \,\mathrm{d}T \tag{20}$$

This expression is identical with Equation 10 of CC, so that their Equation 12 for π can be used, but with a more precise interpretation of the meaning of π.

Comparisons with simulation results

The accuracy of Equations S12–S15 was tested using the simulation results from Figure 4 and Table S6 in CC. These simulations modeled a group of 70 linked genes with properties similar to those of typical *D. melanogaster* autosomal genes, and provided values of the mean nucleotide site diversity at synonymous sites under the assumption that they are selectively neutral. The genetic model and parameters of the simulations are summarized in *Methods*; full details are given in CC. The effect of BGS on the rate of substitutions of favorable mutations for a given parameter set was calculated by multiplying the rate in the absence of sweeps by the value of B_1 for neutral sites obtained from simulations; use of B_2 instead of B_1 made little difference. The corresponding theoretical predictions were obtained for a single gene with the structure described in the Methods section, on the assumption that sweep effects decay sufficiently fast with distance from the selected site that each gene can be treated independently; this is probably not entirely accurate for the lowest rate of crossing over.

Equations S12–15 can be applied in two ways. First, the mean values of the relevant quantities across all neutral sites can be determined, and substituted into these equations—the procedure used in CC for the older method of predicting recurrent sweep effects. Second, the values of these statistics for individual neutral sites can be used to predict π, and the mean of π taken across all neutral sites in a gene, as described in *Methods*. The latter procedure is more accurate statistically, and is used here for the *C2* predictions; in practice, the two methods yield similar results for the parameter sets used here.

Figure 3 shows the reduction in neutral diversity per gene obtained from the simulation results (red bars), the *C2* predictions using Equations A12–A15 (blue bars), the predictions from Equation 12 of CC (black bars) that use the *NC* approximation, and the *NC*-based formula for recurrent sweep effects that assumes competing exponential processes of coalescence due to drift and selection, respectively (Equation 7 of CC) (white bars). For the last two calculations, the *NC* assumption with the deterministic sweep duration (T_d) was used to calculate the effect of a sweep, using the first method described above for estimating the mean effect over neutral sites. Further details of the calculations are given in *Methods*.

The most notable point is that, as was also found by CC, the last method is consistently the least accurate, especially at low CO rates in the presence of gene conversion but without BGS. In the absence of gene conversion, the predictions from Equation 12 of CC and Equations A12–15 generally have a similar level of accuracy. However, in the presence of gene conversion, the latter equations perform best, and provide fairly accurate predictions, except for the lowest CO rate. This comparative failure of predictions based on the *NC* approach presumably reflects the fact that it considerably underestimates the effects of sweeps in the presence of recombination, as seen in Figure 1. Given that gene conversion is pervasive in genomes, and contributes substantially to recombination rates over short physical distances, the approach developed here should provide the most accurate predictions of recurrent sweep effects, despite its heuristic basis. The inaccuracy of the predictions based on the competing exponential process model, introduced by Kaplan *et al.* (1989) and further

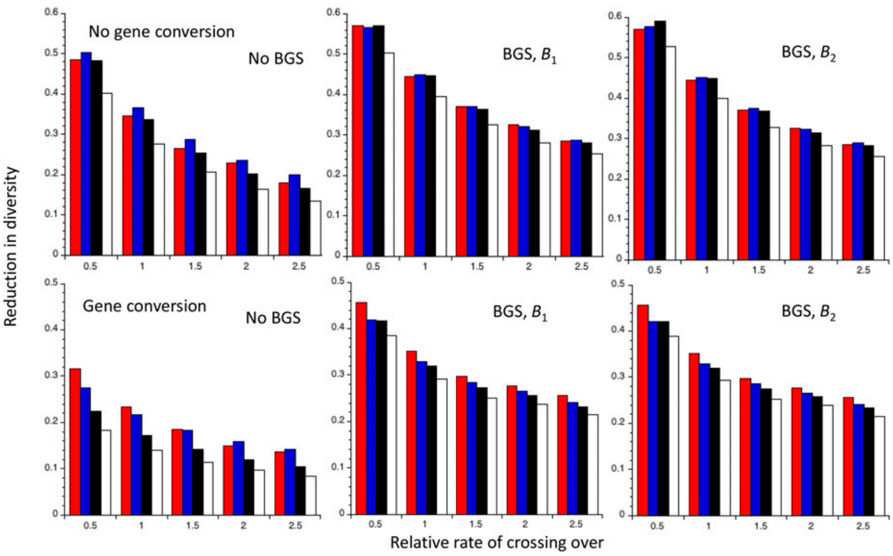

Figure 3 Comparisons with several different theoretical predictions of the mean values of the reduction in diversity obtained in computer simulations of recurrent sweeps with random mating, autosomal inheritance and semidominant favorable mutations, described in CC. The X axis shows the values of the rate of crossing over, expressed relative to the mean value for *Drosophila melanogaster*. The red bars are the mean values of the simulation results for neutral (synonymous sites) in a group of 70 genes; the blue bars are the predictions from Equations S12–S15; the black bars are the predictions from Equation 12b of CC; the white bars are the predictions from the coalescent model of recurrent sweeps that assumes competing exponential processes of coalescence by drift and sweeps (Equation 7 of CC). Cases with and without gene conversion (upper and lower panels, respectively), and in the absence or presence of background selection (BGS), using either B_1 or B_2 to predict fixation probabilities when BGS is acting, are shown.

analyzed by Stephan *et al.* (1992), Kim and Stephan (2000), and Gillespie (2002), is probably due to the fact that it treats sweeps as point events, allowing too much opportunity for drift-induced coalescent events between sweeps (CC, p. 293).

X chromosomes vs. autosomes

There has been considerable interest in comparing the properties of variability of sequences on X or Z chromosomes with those on autosomes (A), since these may shed light on questions such as the relative importance of BGS *vs.* selective sweeps in shaping genome-wide patterns of variability, and on the causes of the apparently faster rates of adaptive evolution on the X or Z chromosome (Charlesworth *et al.* 2018; Wilson Sayres 2018). It therefore seems worth revisiting this question in the light of the models of selective sweeps developed here, which can easily be applied to sex-linked loci. The findings extend those of Betancourt *et al.* (2004), who considered only the case of selection acting equally on the two sexes and used the equivalent of the *NC* model described above.

As noted by Betancourt *et al.* (2004), there are important differences in the theoretical expectations for taxa such as *Drosophila* and Lepidoptera, in which autosomal recombinational exchange is absent in the heterogametic sex, and taxa such as mammals and birds, where recombination is absent between the X (Z) and Y (W), but occurs on autosomes and pseudoautosomal regions in the heterogametic sex. In the first type of system, the sex-averaged effective rate of recombination (which controls the rate of breakdown of linkage disequilibrium) between a pair of X- or Z-linked genes is 4/3 times that for an autosomal pair with the same rate of recombination in the homogametic sex, due to the fact that the X or Z spends 2/3 of its time in the homogametic sex and 1/3 of its time in the heterogametic sex, whereas an autosome spends half of its time in the heterogametic sex where it cannot recombine (Langley *et al.* 1988). In the second type of system, the ratio of sex-averaged rates is 2/3 (Betancourt *et al.* 2004). These two systems will be referred to here as the "*Drosophila*" and "mammalian" models, respectively. For brevity, only male heterogamety is considered here; the results for female heterogamety can be obtained by interchanging male and female.

The effect of a single sweep on X-linked diversity

It is straightforward to use the framework leading to Equation 9 to examine the effect of a single sweep on variability for an X-linked locus. In this case, it is necessary to model the effects of sex differences in the effects of a mutation on male and female fitnesses, since these greatly affect the evolutionary trajectories of favorable X-linked mutations (Rice 1984; Charlesworth *et al.* 1987; Charlesworth 2020). Three extreme cases are considered here: no sex-limitation of fitness effects (so that the homozygous selection coefficient, s, is the same for males and females), male-only fitness effects, and female-only fitness effects. Random mating is assumed throughout.

For simplicity, the dominance coefficient h is assumed to be independent of sex. For the autosomal case with weak selection, a single s that is given by the mean of the male and female fitness effects is sufficient to describe the system (Wright and Dobzhansky 1946). The values of the coefficients a and b in Equation 6 for X-linkage and the three types of sex-dependent fitness effects can be obtained from the expressions in Box 1 of Charlesworth (2020). With no sex-limitation and random mating, $a = (2h + 1)/3$ and $b = 2(1–2h)/3$; with male-only selection, $a = 1/3$ and $b = 0$; with female-only selection, $a = 2h/3$ and $b = 2(1 - 2h)/3$. In order to ensure comparable strengths of selection for X and A with the same patterns of relation between gender and fitness, the values of s for the cases of male- and female-only effects with X-linkage are set equal to twice the corresponding autosomal s without sex-limitation, compensating for the fact that the effective s for a sex-limited autosomal mutation is only one-half of the selection coefficient in the affected sex.

Figure 4 shows the reductions in diversity at the end of a sweep predicted by the *C1* and *C2* methods for the *Drosophila* model, together with the results of simulations using the algorithm of Tajima (1990), for the case of an X-linked locus whose effective population size, kN_{e0}, is three-quarters that of A, N_{e0}. This case corresponds to a randomly mating population in which males and females have equal variances in reproductive success (Wright 1939). With $h = 0.5$ and $k = 0.75$, all three types of sex-specific selection on X-linked loci have similar evolutionary dynamics, provided that the selection coefficients are adjusted as described above (Charlesworth 2020). No differences among their sweep effects are thus to be expected, apart from small deviations reflecting numerical inaccuracies in the integrations. This expectation is confirmed by the results shown in Figure 4. As before, the *C1* approximation predicts much larger effects than *C2* at high r/s values; the *NC* approximation predicts even smaller effects than *C2* (results not shown).

The comparison of the X-linked results with $h = 0.5$ with the autosomal results for $\gamma = 250$ in Figure 1 confirms the expectation that the diversity reductions are the same for the two genetic systems, when s is adjusted appropriately. In addition, female-limited X-linked mutations have the same effects as female-limited autosomal mutations for all h values, again as expected from their similar dynamics. For a given r/s, male-limited selection gives the largest reduction in X-linked diversity when $h = 0.1$, which is substantially larger than the autosomal and female-limited values for the same adjusted selection strength. This is expected from the slow initial rates of increase in the frequencies of partially recessive autosomal or female-limited X-linked mutations (Haldane 1924; van Herwaarden and van der Wal 2002; Teshima and Przeworski 2006; Ewing *et al.* 2011; Charlesworth 2020). With $h = 0.9$, the differences between the various cases are relatively small, with male-limited X-linked mutations having the smallest effects, and nonsex-limited and female-limited mutations having almost identical effects. The sweep effects decrease more rapidly with r/s for X than for A, as expected from the higher effective recombination rate for X.

Figure S3 in File S1 shows comparable results for the mammalian model. The results are broadly similar to those for the *Drosophila* model, the main difference being that the sweep effects are always larger than for the corresponding *Drosophila* cases, as would be expected from the fact that the effective rate of recombination on the X chromosome is half the *Drosophila* value. In this case, the X sweep effects decrease more slowly with r/s than the A effects. Figure S4 shows the results for both the *Drosophila* and mammalian models on a linear scale.

The X/A ratio of N_e values in the absence of selection at linked sites (k) may differ from three-quarters. Sex differences in these variances cause k values that differ from 0.75 (Caballero 1995; Charlesworth 2001; Vicoso and Charlesworth 2009), with higher male than female variances leading to $k > 0.75$, and lower male than female variances having the opposite effect. Male–male competition for mates is likely to cause a higher male than female variance in fitness, so that k can be >0.75 with male heterogamety. In contrast, female heterogamety with sexual selection leads to $k < 0.75$. Some examples of the reductions in diversity at the end of a sweep with $k \neq 0.75$, using the *C2* predictions, are shown for the mammalian model in Figure S5 in File S1. Comparing these with Figure S3, it can be seen that smaller k values cause somewhat larger X-linked sweep effects for all modes of selection. The effects are, however, relatively small, and unlikely to be detectable in most datasets. This pattern is presumably caused by the fact that smaller k means that coalescence during a sweep occurs more rapidly relative to recombination.

The effects of recurrent selective sweeps on X-linked diversity

Expressions for the effects of recurrent selective sweeps on X-linked neutral diversity can be obtained using the appropriate modifications to Equation 17 (the *C1* approximation) and Equations S12–S15 in File S1 (the *C2* approximation). There is an important factor that leads to differences between A and X sweep effects, additional to those considered in the previous section. This is the fact that the expected coalescent time for the X is $2kN_{e0}$ instead of $2N_{e0}$, with k generally expected to be <1 (Charlesworth 2001; Vicoso and Charlesworth 2009). The parameter ω that appears in the equations for sweep effects is the expected number of substitutions over $2kN_{e0}$ generations, so that a smaller value of k for a given rate of substitution per generation implies a smaller ω value. An alternative way of looking at this effect is to note that $k < 1$ implies a faster rate of genetic drift for X than A; other things being equal, coalescent events induced by drift are then more frequent for X than for A, relative to coalescent events induced by selection (Betancourt *et al.* 2004).

A countervailing factor is that the rates of substitution per generation of favorable X-linked mutations are expected to be higher than for comparable autosomal mutations with male-limited or nonsex-limited selection, given sufficiently small h values (the condition is $h < 0.5$ with $k = 0.75$) (Charlesworth *et al.* 1987, 2018; Vicoso and Charlesworth 2009); expressions for the rates of substitution are given in File S1, section 4. These opposing effects of sex linkage implies that simple generalizations about the effects of sweeps on X-linked *vs.* autosomal variability cannot easily be made, as will shortly be seen.

Figure 5 shows the *C2* predictions for the reductions in diversity relative to neutral expectation for X-linked and autosomal loci under the *Drosophila* model, as a function of the ratio of the autosomal effective CO rate to a value of 10^{-8} per

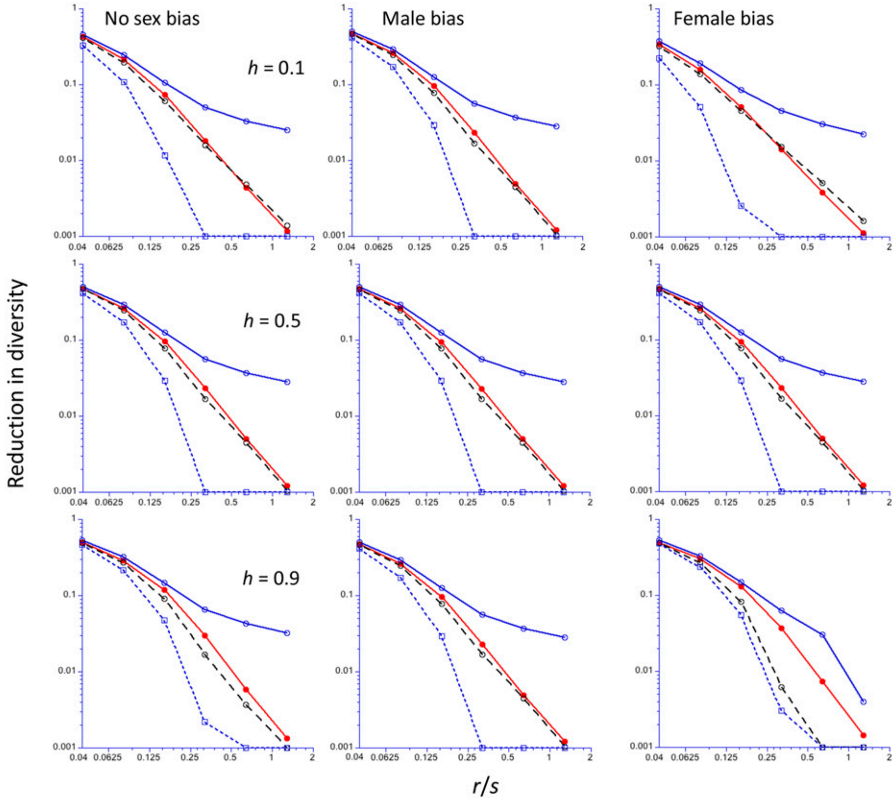

Figure 4 The reduction in diversity (relative to the neutral value) at the end of a sweep for an *x*-linked locus (*y*-axis, $\log_{10}$ scale), as a function of the ratio of the frequency of recombination (*r*) to the selection coefficient for homozygotes (*s*) (*x*-axis, $\log_2$ scale). The *Drosophila* recombination model is assumed; N_e for the X chromosome is three-quarters of that for the autosomes. The results for mutations with no sex limitation are shown in the left-hand panels; those for male-limited and female-limited mutations are shown in the middle and right-hand panels, respectively. A population size of 5000 is assumed, with a scaled selection coefficient for an autosomal mutation in a randomly mating population ($\gamma = 2N_es$) of 250 for the cases with no sex-limitation. For the sex-limited cases, $\gamma = 500$ to ensure comparability to sex-limited autosomal mutations. Results for three different values of the dominance coefficient (*h*) are shown, with *h* increasing from the top to bottom panels. The filled red circles are the mean values from computer simulations, using Tajima's algorithm; the open blue circles and black circles are the *C1* and *C2* predictions, respectively; the open blue squares are the *NC* predictions. Values of the reduction in diversity <0.001 have been reset to 0.001.

basepair for *D. melanogaster*, using the same gene structure that was used to generate the theoretical results for autosomes shown in Figure 3 (see *Methods*). Gene conversion was allowed, with the same rate of initiation as crossovers. The CO and gene conversion rates for the X chromosome were set by multiplying the corresponding autosomal effective rates by 0.75, so that the effective recombination rates for the X chromosome and A are equal, following the procedure used in empirical comparisons of diversity levels on the X and A (Campos *et al.* 2014). Here $k = 0.75$, so that potential effects of sexual selection or variance in female reproductive status are absent.

As described in *Methods*, the values of the BGS parameter B_1 were obtained from estimates given by Charlesworth (2012), which include contributions from selectively constrained noncoding sequences as well as coding sequences; comparable values were obtained in the more detailed analyses of Comeron (2014). As described in CC, the BGS effect parameter B_1 for the X chromosome with a relative crossing rate of 0.5 was set to a relatively high value (0.549 instead of 0.449) to correct for the relatively low gene density in this regions of the *D. melanogaster* X chromosome, whereas the values for the rates of 1, 1.5, 2, and 2.5 assumed normal gene densities, giving B_1 values of 0.670, 0.766, 0.818, and 0.852, respectively. This results in a relatively weak effect of selection in reducing diversity for the lowest CO rate compared with the autosomes, where the B_1 values were set to 0.538, 0.733, 0.813, 0.856, and 0.883 for the relative CO rates of 0.5, 1, 1.5, 2, and 2.5, respectively. Male and female mutation rates were assumed to be equal, in view of the lack of strong evidence for a sex difference in mutation rates in *Drosophila* (Charlesworth *et al.* 2018). For convenience, B_2 was assumed to be equal to B_1.

As expected, in the absence of BGS the X results for $h = 0.5$ are the same for the three types of sex-specific fitness effects. The X effects are slightly smaller than the A effects for low recombination rates, reflecting the reduced rate of substitution on the coalescent timescale of the lower N_e for X than A, which was described above. With $h = 0.1$, the lower rates of substitution of A mutations and female-limited X mutations greatly reduce their sweep effects, but the effects for nonsex-limited and male-limited X mutations are much larger than for A mutations. With $h = 0.9$, male-limited X mutations have the weakest effects, while the other three classes of mutations have quite similar effects, with female-limited mutations having the largest effects. Similar general patterns are seen with BGS, with much smaller differences between X and A than in the absence of BGS, except for the lowest CO rate, where the relatively small BGS effect for the X causes it to have a much smaller reduction in diversity compared with A. Comparable results with no gene conversion are shown in Figure S6 of File S1. The general patterns are quite similar to those with gene conversion, but with a greater sensitivity to the CO rate with $h = 0.1$ and no sex-limitation or male-limitation, especially with no BGS.

Figure 6 shows the values of the X/A ratio of diversities (R_{XA}) for different CO rates, obtained from the results shown

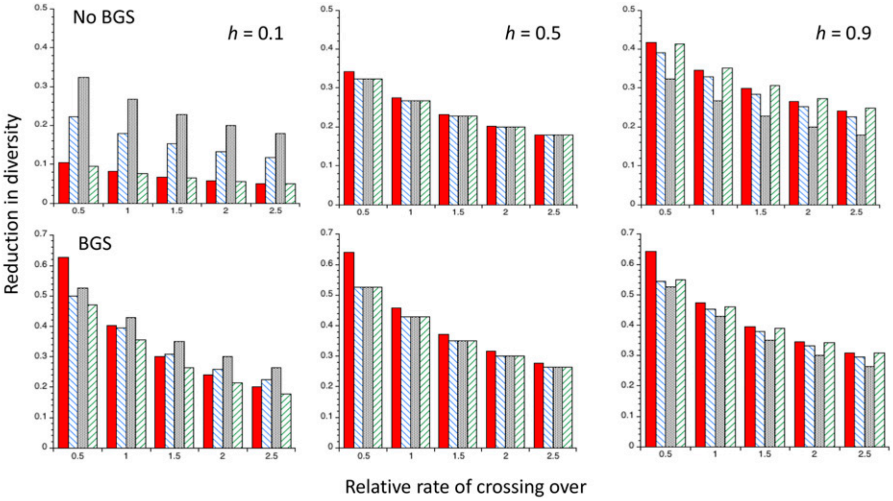

Figure 5 Reductions in diversity (relative to neutrality) under recurrent sweeps at autosomal and X-linked loci for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as for the autosomes). N_e for the X chromosome is three-quarters of that for the autosomes. The upper panel is for cases without BGS; the lower panel is for cases with BGS (with the parameters described in the main text). The filled red bars are for autosomal mutations, the hatched blue bars are for X-linked mutations with no sex-limitation, the stippled gray bars are for male-limited X-linked mutations, and the hatched green bars are for female-limited X-linked mutations. For the sex-limited cases, $\gamma = 500$ to ensure comparability with the autosomal and nonsex-limited X-linked mutations.

in Figure 5; Figure S7 shows comparable results with no gene conversion. First, consider the case when BGS effects are absent. With $h = 0.5$, R_{XA} is always close to the neutral expectation of 0.75 for all modes of selection; this is also true with female-limited selection for all three dominance coefficients, with a slight tendency toward $R_{XA} > 0.75$ for low CO rates, declining toward 0.75 as the CO rate increases. With $h = 0.1$ and no sex-limitation, it can be seen that $R_{XA} << 0.75$ for the lowest CO rates, approaching 0.7 at the highest rate. With $h = 0.1$ and male-limitation, R_{XA} increases with the CO rate, but remains well below 0.7 even at the highest rate. With $h = 0.9$ and no sex-limitation, R_{XA} is slightly larger than 0.75 for the lowest rate of crossing, approaching 0.75 as the rate increases; with male-limited selection, $R_{XA} > 0.85$ at the lowest CO rate, and $R_{XA} \approx 0.8$ at the highest rate.

The presence of BGS greatly alters these patterns; the lower gene density for the X in the region with the lowest CO rate causes R_{XA} values of 0.9 or more for all three modes of selection and dominance coefficients. R_{XA} even exceeds 1 for female-limited selection with $h = 0.1$ and the lowest CO. BGS causes a much steeper decline in R_{XA} with the CO rate than in its absence (when it can even increase), especially with $h = 0.1$ and female-limited selection. The contrast between the presence and absence of BGS and the male-limited and nonsex-limited cases with $h = 0.1$ is especially striking. However, if the B_1 value of 0.449 for a normal gene density is used for the lowest CO rate, X-linked diversity is considerably increased and R_{XA} is correspondingly reduced; for example, with $h = 0.1$ and no sex-limitation, $R_{XA} = 0.849$ instead of 1.001. Again, the patterns without gene conversion are similar to those found with gene conversion; however, with $h = 0.1$ and no sex-limitation or male-limitation, R_{XA} is much more sensitive to the CO rate than when gene conversion is acting.

The effects of differences in k are shown in Figures S8 and S9, for the case with both BGS and gene conversion. Under the substitution model used here, a larger k is associated with a faster rate of substitution, countering the small effect of the size of an individual sweep described above. Comparisons with Figure 6 show that there tend to be somewhat larger effects of X-linked sweeps with the larger values of k. These translate into noticeably larger values of the X/A diversity ratios, but a reduced sensitivity of these ratios to the CO rate.

Discussion

General considerations

As described in the introduction, a widely used simplification for calculating the effect of a selective sweep on nucleotide site diversity at a linked neutral site is the "star phylogeny" assumption that alleles sampled at the end of a sweep, and which have not recombined onto a wild-type background, coalesce instantaneously. Their mean coalescent time (relative to the purely neutral value) for a pair of alleles can then be equated to the probability that one of them undergoes a recombination event that transfers the neutral site onto the wild-type background (Wiehe and Stephan 1993; Barton 1998, 2000; Durrett and Schweinsberg 2004).

The results presented here show that this often leads to inaccuracies in predictions concerning the mean coalescent time for a pair of swept alleles, especially when the ratio of recombination rate to the homozygous selection coefficient (r/s) is relatively high, consistent with previous findings (Barton 1998; Hartfield and Bataillon 2020), as can be seen in Figure 1 and Figure S1. Similarly, Figure 3 shows that with recurrent sweeps, gene conversion, and no BGS, the *NC* approximation and its modification by CC considerably underpredict the effects of sweeps compared with simulations,

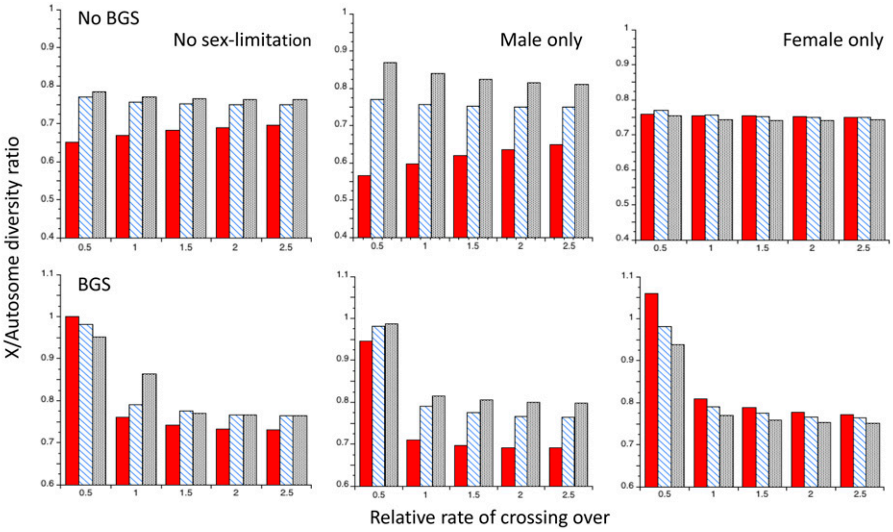

Figure 6 The ratios of X chromosome to autosome nucleotide site diversities (R_{XA}) for the *Drosophila* model under recurrent sweeps, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as for the autosomes). N_e for the X chromosome is three-quarters of that for the autosomes ($k = 0.75$). The upper panel is for cases without BGS; the lower panel is for cases with BGS. The filled red bars are for $h = 0.1$, the hatched blue bars are for $h = 0.5$ and the stippled gray bars are for $h = 0.9$. The other details are as for Figure 5.

whereas the *C2* approximation derived here fits much better. The *C1* approximation greatly overestimates the diversity reduction at high recombination rates.

This inaccuracy of the *NC* approximation reflects the fact that the probability of no recombination in the absence of coalescence (P_{nr}) used in the *NC* approximation (Equation 11b) declines much faster with increasing recombination rate than does the true probability of no recombination, $(1 - P_r)$ (see Table 1 and Table S1 of File S1). The theory for large r/s is, however, not entirely satisfactory, as the *C2* approximation derived here uses a heuristic approach to modeling the effects of multiple recombination events, while the *C1* approximation ignores these events. Table 1 and Table S1 show that the probability of a single recombination event is often less than half the net probability of a recombination event when $r/s \geq 0.08$, so that multiple events cannot then be ignored. As described above, the true expected reduction in diversity probably lies between the *C1* and *C2* predictions. Ideally, both multiple recombination events and within-sweep coalescent events should be included in the model without the approximations used here. This was done by Kaplan *et al.* (1989), but no simple formula can be obtained by this approach.

Other stochastic treatments of the mean coalescent time associated with a sweep that should, in principle, allow for multiple recombination events and coalescence within the sweep, have been given by Stephan *et al.* (1992) and Barton (1998) for the case of semidominant selection with autosomal inheritance and random mating. However, these are not necessarily very accurate. For the example in Figure 1 with $h = 0.5$, $\gamma = 250$ and $r/s = 0.64$, Equation 18 of Stephan *et al.* (1992) predicts a reduction in diversity at the end of a sweep of 0.013, whereas the simulations and *C2* approximation give values of 0.0051 and 0.0044, respectively. After some simplification (and equating Barton's ε to q_1), Equation 16 of Barton (1998) gives a predicted reduction in pairwise diversity in the absence of recombination of approximately $1 - T_d - 2\gamma^{-1}\ln(\gamma)$. For $\gamma = 250$ (with $T_d = 0.088$), this is equal to 0.868, compared with the simulation and *C1* and *C2* values of 0.925. In addition, Figure 3 of Barton (1998) shows a slightly faster than linear increase in mean coalescent time with increased r/s when $r/s < 0.5$, in contrast to the approximately exponential decline in $-\Delta\pi$ seen in Figure S1, corresponding to a diminishing returns relation for π.

It is important to note that, even for recombination events within genes, relatively large r/s values are likely. For example, with the parameters used for Figure 3, $s = 1.25 \times 10^{-4}$ in a population with $N_e = 10^6$. With the standard sex-averaged autosomal CO rate for *D. melanogaster* of 1×10^{-8} per bp, but without gene conversion, the recombination rate between two sites 1 kb apart is 1×10^{-5}, so that $r/s = 0.08$. With gene conversion at an effective rate of initiation of 1×10^{-8} per bp and a mean tract length of 440 bp, the recombination rate is 1.88×10^{-5}, and $r/s = 0.15$. Figure 1 shows that, for $r/s = 0.16$, $h = 0.5$, and $\gamma = 250$, the predicted reduction in diversity at the end of sweep for *C2* is approximately 86% of that for *C1* and 114% of the simulation value (this is not significantly different from the *C2* result at the 1% level); the *NC* approximation predicts a reduction that is approximately 27% of the *C2* value.

Knowledge of the expected effects of multiple recombination events for large r/s is even more important for modeling recurrent sweep effects on intergenic sequences, which is needed for interpreting the observed pattern of increased intergenic sequence variability as a function of the distance from a gene in both mammals (Halligan *et al.* 2010; Hammer *et al.* 2010; Booker 2018) and *Drosophila* (Johri *et al.* 2020). An improved analytical treatment of this problem is desirable. At present, the use of the *C2* approximation seems to provide the best option for dealing with recurrent sweeps, other than by numerical solutions using the results of Kaplan *et al.* (1989) or simulations of the type performed by Messer and Petrov (2013) and Johri *et al.* (2020).

Relations between synonymous site diversity and recombination rate

The main purpose of this paper is to explore some general principles rather than to attempt to fit models to data, but it is obviously of interest to examine the relations between the theoretical predictions for the *Drosophila* model of recurrent selective sweeps described above and the relevant empirical evidence. As Figure 3 shows, despite the caveats discussed about, the analytical results derived here for selective sweep effects using the *C2* approximation should provide better predictions concerning the relation between synonymous site diversity and local recombination rate than those discussed in CC. The basic expectation of a diminishing returns relation between synonymous site diversity and CO rate described in CC remains, however, unchanged.

Based on the empirical plots of this relationship provided in Campos *et al.* (2014), it was concluded by CC that the observed relation between synonymous π and CO rate in a Rwandan population of *D. melanogaster* was too steep and close to linear to be explained by models that include both selective sweep and BGS effects, or by either of these processes on their own. One possibility is that CO events are mutagenic, as indicated by recent studies of human *de novo* mutations (Halldorsson *et al.* 2019). This is, however, hard to reconcile with the lack of evidence for a correlation between silent site divergence and CO rate in *D. melanogaster*, outside the noncrossover genomic regions where divergence tends to be *higher* than average, presumably reflecting the effect of reduced N_e due to selection at linked sites (Haddrill *et al.* 2007; Campos *et al.* 2014). This observation is, however, not conclusive, since the recombination landscape in *D. melanogaster* is substantially different from that in its close relative *D. simulans*, with less suppression of crossing over near telomeres and centromere (True *et al.* 1996), so that current estimates of CO rates may not reflect the evolutionarily significant values.

Another possibility is that the nearly linear relationships between described by Campos *et al.* (2014) are artifacts of their use of classical marker-based maps (Fiston-Lavier *et al.* 2010) or the Loess smoothing procedure applied to the 100 kb window estimates of CO rates obtained by the SNP-based map of Comeron *et al.* (2012). Smoothing may cause relative low values of π associated with very high CO rates to be wrongly assigned to much lower CO rate, as noted by Castellano *et al.* (2016). A diminishing returns relation between noncoding site diversity for the Raleigh population of *D. melanogaster* and CO rate was found by Comeron (2014) when using the raw CO rate estimates; a similar pattern is seen in the Rwandan population (J.M. Comeron, personal communication). However, the use of the raw estimates is open to the objection that the extreme CO values may simply be artifactual, leading to a flatter relation between π and CO rate than truly exists. In addition, Comeron's noncoding π values for the Rwandan population are substantially lower than the synonymous site values of Campos *et al.* (2014), similar to what has been found in studies of other populations (Andolfatto 2005; Haddrill *et al.* 2005), suggesting that the noncoding sites involve at least some sequences that are subject to selection. This could lead to a less than linear relation between π and CO rate. The question of the true empirical relationship between recombination rate and neutral or nearly neutral variability in *Drosophila* needs further exploration before firm conclusions can be drawn.

Differences between X chromosomes and autosomes

As discussed by CC, the differences between X chromosomes and autosomes in their levels of neutral diversity, and the relations between these and CO rates, need to be interpreted in terms of models of the effects of selection at linked sites. A pattern seen in several analyses of *D. melanogaster* datasets is that the relation between silent or synonymous site diversity for the X and CO rate is considerably weaker than that for the autosomes (Langley *et al.* 2012; Campos *et al.* 2014; Comeron 2014).

The expected difference between X and A is seen mostly clearly by plotting the ratio of X to A diversity values (R_{XA}) against the CO rate, adjusted to give the same effective rate for X and A genes. The results for the case when R_{XA} in the absence of selection (k) = 0.75 were shown in Figure 6. The contrast between the cases with and without BGS is striking. Without BGS, for each mode of selection the ratio either slightly increases with CO rate ($h = 0.1$) or is constant, or nearly constant ($h = 0.5$ or $h = 0.9$). With BGS, there is a strong decline in R_{XA} from the lowest relative rate of CO (0.5), with values >1 or ≈ 1, and the standard rate (1.0). The value at the highest relative CO rate (2.5) varies according to the dominance coefficient and mode of selection. With male-only selection, $R_{XA} \approx 0.7$ with $h = 0.1$ but is ≈ 0.8 with the other dominance coefficients; for the other modes of selection it is ≈ 0.75. Given the evidence that the X in *Drosophila* is deficient in genes with male-biased expression, but enriched in female-biased genes (Parsch and Ellegren 2013), the results for male-biased genes are probably the least relevant. The data on synonymous site diversity in the Rwandan population of *D. melanogaster* in Figure S2 of Campos *et al.* (2014), based on Loess smoothing of the raw recombination estimates of Comeron *et al.* (2012), show that R_{XA} takes the values of 1, 0.84. 0.74, and 0.73 for relative CO rates of 0.5, 1.0, 1.5, and 2.0, respectively. Use of the raw estimates of CO rates and diversity at noncoding sites for the same population gives a qualitatively similar pattern (J.M. Comeron, personal communication).

Effect of a change in population size on R_{XA}

It has been shown previously that a change in population size can cause R_{XA} to deviate from its equilibrium value, k (Hutter *et al.* 2007; Pool and Nielsen 2007, 2008), reflecting the fact that the rate of response of neutral diversity to a change in population size is faster with smaller N_e. This raises the question as to whether the observed pattern of relationship between R_{XA} and the CO rate that has just been discussed could be explained by such a change, rather than by the differential effect of selection at linked sites on X and A diversity values. An approximate answer to this question can be obtained with a purely neutral model, in which the population size changes from a initial equilibrium value, but k remains constant during the process of change. In addition to k, we need to specify the relation between the rate of recombination and diversity. This can be done by introducing a variable β ($0 \leq \beta \leq 1$), which is equal to the ratio of the equilibrium diversity for a given effective rate of recombination to its value at the maximum recombination rate in the study. On the null hypothesis that there is no differential effect on R_{XA} of selection at linked sites, the same β should apply to X and A diversities with the same effective recombination rate.

The most extreme effect of a population size change on R_{XA} will come from a step change in population size, since this minimizes the ability of diversity values to track the population size. Consider a model in which the time T since the start of the expansion is scaled relative to the final autosomal N_e at the highest recombination rate; let the ratio of final to initial effective population sizes be R_N; and write $\alpha = 1–1/R_N$. Using the equivalents of Equation S10a applied to X and A diversities relative to their final equilibrium values, R_{XA} at time T is given by:

$$R_{XA}(T) = k\frac{\left[1-\alpha\ \exp\left(-k^{-1}\beta^{-1}T\right)\right]}{\left[1-\alpha\ \exp\left(-\beta^{-1}T\right)\right]} \tag{20}$$

This expression shows that, as expected, R_{XA} is equal to k when $T = 0$, and also when $T >> k\beta$. For a population expansion (so that $0 < \alpha < 1$), for intermediate T values we have $R_{XA} > k$, provided that $k < 1$. The converse is true for a population contraction, for which $\alpha < 0$. Thus, regardless of the value of β, a population expansion will temporarily increase R_{XA} above its equilibrium value. The extent of this increase for a given T is affected by β, but the direction and magnitude of the effect of β varies with T. For small T, a smaller value of β causes a larger value of R_{XA}; the reverse is true for large T (File S1, section 5).

It follows that no simple predictions are possible as to whether low rates of recombination are associated with larger values of R_{XA} than high rates, after a population expansion of the kind indicated by the data on the Rwandan population of *D. melanogaster*. However, numerical examples suggest that the effects of differences in β are at best modest when $k = 0.75$. For example, with $R_N = 10$ (a 10-fold increase in population size), the values of R_{XA} for $\beta = 0.25$ and $\beta = 1$ are 0.892 and 0.858, respectively at $T = 0.1$. By $T = 0.2$, the respective R_{XA} values are 0.869 and 0.885, and by $T = 0.5$ they are 0.801 and 0.884. There is therefore only a brief interval of time in which the lower β value (which corresponds to the lowest recombination rate considered above) is associated with R_{XA} substantially larger than that with the higher value (which corresponds to the higher recombination rate considered). With $\beta = 1$, $R_{XA} > 0.85$ persists until $T = 1$.

Instead of comparing X and A, we can compare the ratio of diversity values for two different regions of the same chromosome with different CO rates, with the left-hand side of Equation 20 representing this ratio at a given time after a population size change, where k now represents the effect of recombination rate differences on the equilibrium level of diversity (β is set =1, since we are now longer comparing X and A). Equation 20 shows that a population expansion will reduce the differentials between regions with different k values, whereas a contraction will enhance them. For example, with $R_N = 10$, $k = 0.5$, and $T = 0.1$, the diversity ratio becomes 0.709 instead of 0.5. Given the distortion of the site frequency spectrum at synonymous sites on the autosomes in the Rwandan population toward low frequency variants, with Tajima's D values at synonymous sites of approximately –0.2 (Campos *et al.* 2014), there has probably been a recent population expansion, which may have weakened the relation between diversity and CO rate compared with an equilibrium population. Further theoretical investigations of the interaction between such demographic effects and effects of selection at linked sites are needed if reliable inferences concerning both demography and selection are to be obtained (Messer and Petrov 2013; Zeng 2013; Ewing and Jensen 2016; Comeron 2017; Lange and Pool 2018; Becher *et al.* 2020; Johri *et al.* 2020).

Acknowledgments

I thank Matthew Hartfield for many useful discussions of this problem and for supplying me with the simulation results used in Figures 2 and S2. I also thank Josep Comeron for helpful discussions, and for sharing his analyses of *Drosophila* polymorphism data. I thank Nick Barton, Matthew Hartfield and an anonymous reviewer for their comments on the original manuscript, which have led to significant improvements.

Literature Cited

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–1152. https://doi.org/10.1038/nature04107

Atkinson, K. E., 1989 *Introduction to Numerical Analysis*, John Wiley, New York, NY.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133. https://doi.org/10.1017/S0016672398003462

Barton, N. H., 2000 Genetic hitchhiking. Phil. Trans. R. Soc. B 355: 1553–1562. https://doi.org/10.1098/rstb.2000.0716

Becher, H., B. C. Jackson, and B. Charlesworth, 2020 Patterns of genetic variability in genomic regions with low rates of recombination. Curr. Biol. 30: 94–100. https://doi.org/10.1016/j.cub.2019.10.047

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

Berry, A. J., J. W. Ajioka, and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 129: 1111–1117.

Betancourt, A. J., Y. Kim, and H. A. Orr, 2004 A pseudohitchhiking model of X *vs.* autosomal diversity. Genetics 168: 2261–2269. https://doi.org/10.1534/genetics.104.030999

Booker, T. R., 2018 Inferring parameters of the eistribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare. G3 (Bethesda) 10: 2317–2326. https://doi.org/10.1534/g3.120.401052

Booker, T. R., B. C. Jackson, and P. D. Keightley, 2017 Detecting positive selection in the genome. BMC Biol. 15: 98. https://doi.org/10.1186/s12915-017-0434-y

Caballero, A., 1995 On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics 139: 1007–1011.

Campos, J. L., and B. Charlesworth, 2019 The effects on neutral variability of recurrent selective sweeps and background selection. Genetics 212: 287–303. https://doi.org/10.1534/genetics.119.301951

Campos, J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth, 2014 The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. 31: 1010–1028. https://doi.org/10.1093/molbev/msu056

Campos, J. C., 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. https://doi.org/10.1073/pnas.1619434114

Castellano, D., M. Coronado-Zamora, J. L. Campos, A. Barbadilla and A. Eyre-Walker, 2016 Adaptive evolution is substantially impeded by Hill–Robertson interference in Drosophila. Mol. Biol. Evol. 33: 442–445. https://doi.org/10.1093/molbev/msv236

Charlesworth, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77: 153–166. https://doi.org/10.1017/S0016672301004979

Charlesworth, B., 2012 The role of background selection in shaping patterns of molecular evolution and variation: evidence from the Drosophila *X* chromosome. Genetics 191: 233–246. https://doi.org/10.1534/genetics.111.138073

Charlesworth, B., 2020 How long does it take to fix a favorable mutation, and why should we care? Am. Nat. 195: 753–771. https://doi.org/10.1086/708187

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

Charlesworth, B., J. A. Coyne, and N. H. Barton, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130: 113–146. https://doi.org/10.1086/284701

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

Charlesworth, B., J. L. Campos, and B. C. Jackson, 2018 Faster-X evolution: theory and evidence from *Drosophila*. Mol. Ecol. 27: 3753–3771. https://doi.org/10.1111/mec.14534

Chen, J., S. Glémin, and M. Lascoux, 2020 From drift to draft: how much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? Genetics 214: 1005–1018. https://doi.org/10.1534/genetics.119.302869

Comeron, J., R. Ratnappan, and S. Bailin, 2012 The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8: e1002905. https://doi.org/10.1371/journal.pgen.1002905

Comeron, J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 10: e1004434. https://doi.org/10.1371/journal.pgen.1004434

Comeron, J. M., 2017 Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Phil. Trans. R. Soc. B 372: 20160471. https://doi.org/10.1098/rstb.2016.0471

Coop, G., and P. Ralph, 2012 Patterns of neutral diversity under general models of selective sweeps. Genetics 192: 205–224. https://doi.org/10.1534/genetics.112.141861

Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton, 2015 Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13: e1002112. https://doi.org/10.1371/journal.pbio.1002112

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*, Harper and Row, New York.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138. https://doi.org/10.1016/j.tpb.2004.04.002

Elyashiv, E., S. Sattah, T. T. Hu, A. Strutovsky, G. McVicker *et al.*, 2016 A genomic map of the effects of linked selection in *Drosophila*. PLoS Genet. 12: e1006130. https://doi.org/10.1371/journal.pgen.1006130

Ewing, G. B., J. Hermisson, and P. Pfaffelhuber, 2011 Selective sweeps for recessive alleles and other modes of dominance. J. Math. Biol. 63: 399–431. https://doi.org/10.1007/s00285-010-0382-4

Ewing, G. B., and J. D. Jensen, 2016 The consequences of not accounting for background selection in demographic inference. Mol. Ecol. 25: 135–141. https://doi.org/10.1111/mec.13390

Fiston-Lavier, A., N. D. Singh, M. Lipatov, and D. A. Petrov, 2010 - *Drosophila melanogaster* recombination rate calculator. Gene 463: 18–20. https://doi.org/10.1016/j.gene.2010.04.015

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69: 831–843. https://doi.org/10.1086/323612

Gilbert, K. J., F. Pouyet, L. Escoffier, and S. Peischel, 2020 Transition from background selection to associative overdominance promotes diversity in regions of low recombination. Curr. Biol. 30: 101–107.e3. https://doi.org/10.1016/j.cub.2019.11.063

Gillespie, J. H., 2002 Genetic drift in an infinite population: the pseudohitchiking model. Genetics 155: 909–919.

Glémin, S., 2012 Extinction and fixation times with dominance and inbreeding. Theor. Pop. Biol. 18: 310–316. https://doi.org/10.1016/j.tpb.2012.02.006

Haddrill, P. R., B. Charlesworth, D. L. Halligan and P. Andolfatto, 2005 Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Gen. Biol. 6: R67. https://doi.org/10.1186/gb-2005-6-8-r67

Haddrill, P. R., D. L. Halligan, D. Tomaras and B. Charlesworth, 2007 Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Gen. Biol. 8: R18. https://doi.org/10.1186/gb-2007-8-2-r18

Haldane, J. B. S., 1924 A mathematical theory of natural and artificial selection. Part I. Trans. Camb. Philos. Soc. 23: 19–41.

Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eory *et al.*, 2010 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9: e10003995.

Halldorsson, B. V., G. Palsson, O. A. Stefansson, H. Jonnson, M. T. Hardarson *et al.*, 2019 Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363: eaau1043. https://doi.org/10.1126/science.aau1043

Hammer, M. F., A. E. Woerner, F. L. Mendez, J. C. Watkins, M. P. Cox *et al.*, 2010 The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat. Genet. 42: 830–831. https://doi.org/10.1038/ng.651

Hartfield, M., and T. Bataillon, 2020 Selective sweeps under dominance and inbreeding. G3 (Bethesda) 10: 1063–1075. https://doi.org/10.1534/g3.119.400919

Hutter, S., H. P. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide nucleotide polymorphism data. Genetics 177: 469–480. https://doi.org/10.1534/genetics.107.074922

Jensen, J. D., B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch *et al.*, 2019 The importance of the neutral theory in 1968 and

50 years on: a response to Kern and Hahn 2018. Evolution 73: 111–114. https://doi.org/10.1111/evo.13650

Johri, P., B. Charlesworth, and J. D. Jensen, 2020 Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics 215: 173–192. https://doi.org/10.1534/genetics.119.303002

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

Keightley, P. D., J. L. Campos, T. R. Booker, and B. Charlesworth, 2016 Inferring the site frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. Genetics 203: 975–984. https://doi.org/10.1534/genetics.116.188102

Kelly, J. K., 2007 Mutation-selection balance in mixed mating populations. J. Theor. Biol. 246: 355–365. https://doi.org/10.1016/j.jtbi.2006.12.030

Kern, A. D., and M. W. Hahn, 2018 The neutral theory in light of natural selection. Mol. Biol. Evol. 35: 1366–1371. https://doi.org/10.1093/molbev/msy092

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics 172: 1967–1978. https://doi.org/10.1534/genetics.105.048447

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 155: 1415–1427.

Kimura, M., 1971 Theoretical foundations of population genetics at the molecular level. Theor. Popul. Biol. 2: 174–208. https://doi.org/10.1016/0040-5809(71)90014-1

Lange, J. D. and Pool, J. E., 2018 Impacts of recurrent hitchhiking on divergence and demographic inference in *Drosophila*. Genome Biol. Evol. 10: 1882–1891. https://doi.org/10.1093/gbe/evy142

Langley, C. H., E. A. Montgomery, R. H. Hudson, N. L. Kaplan, and B. Charlesworth, 1988 On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. 52: 223–235. https://doi.org/10.1017/S0016672300027695

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533–598. https://doi.org/10.1534/genetics.112.142018

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

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35. https://doi.org/10.1017/S0016672300014634

Messer, P. W., and D. A. Petrov, 2013 Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. USA 110: 8615–8620. https://doi.org/10.1073/pnas.1220835110

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. https://doi.org/10.1101/gr.4252305

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics 146: 1501–1514.

Parsch, J., and H. Ellegren, 2013 The evolutionary causes and consequences of sex-biased gene expression. Nat. Rev. Genet. 14: 83–87. https://doi.org/10.1038/nrg3376

Pavlidis, P., D. Zivkovic, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30: 2224–2234. https://doi.org/10.1093/molbev/mst112

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

Pool, J. E., and R. Nielsen, 2007 Population size changes reshape genomic patterns of diversity. Evolution 61: 3001–3006. https://doi.org/10.1111/j.1558-5646.2007.00238.x

Pool, J. E., and R. Nielsen, 2008 The impact of founder events on chromosomal variability in multiply mating species. Mol. Biol. Evol. 25: 1728–1736. https://doi.org/10.1093/molbev/msn124

Rice, W. R., 1984 Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735–742. https://doi.org/10.1111/j.1558-5646.1984.tb00346.x

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

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet. 6: e1000495.

Stephan, W., 2019 Selective sweeps. Genetics 211: 5–13. https://doi.org/10.1534/genetics.118.301319

Stephan, W., T. H. E. Wiehe, and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. 41: 237–254. https://doi.org/10.1016/0040-5809(92)90045-U

Tajima, F., 1990 Relationship between DNA polymorphism and fixation time. Genetics 125: 447–454.

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

True, J. R., J. M. Mercer, and C. C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 142: 507–523.

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. https://doi.org/10.1006/tpbi.2002.1576

Vicoso, B., and B. Charlesworth, 2009 Effective population size and the Faster-X effect: an extended model. Evolution 63: 2413–2426. https://doi.org/10.1111/j.1558-5646.2009.00719.x

Walsh, J. B., and M. Lynch, 2018 *Evolution and selection of quantitative traits*, Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198830870.001.0001

Weissman, D. B., and N. H. Barton, 2012 Limits to the rate of adaptive substitution in sexual populations. PLoS Genet. 8: e1002740. https://doi.org/10.1371/journal.pgen.1002740

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model and its application to DNA polymorphism data. Mol. Biol. Evol. 10: 842–854.

Wilson Sayres, M. A., 2018 Genetic diversity on the sex chromosomes. Genome Biol. Evol. 10: 1064–1078. https://doi.org/10.1093/gbe/evy039

Wright, S., 1939 *Statistical genetics and evolution*, Hermann and Cie, Paris.

Wright, S., 1969 Evolution and the Genetics of Populations. Vol. 2. The Theory of Gene Frequencies. University of Chicago Press, Chicago, Illinois.

Wright, S., and T. Dobzhansky, 1946 Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. Genetics 31: 125–156.

Zeng, K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity 110: 363–371. https://doi.org/10.1038/hdy.2012.102

Zhao, L., and B. Charlesworth, 2016 Resolving the contrast between associative overdominance and background selection. Genetics 203: 1315–1334. https://doi.org/10.1543/genetics.1116.188912

Communicating editor: R. Nielsen

Appendix

Explicit formulae for coalescence and recombination probabilities

When $a > 0$ and $a + b > 0$ (*i.e.*, excluding cases of random mating with complete recessivity or dominance), the time between a given frequency q of A_2, and its frequency q_2 at the end of the deterministic phase of the sweep is given by:

$$\begin{aligned} T(q) &= \gamma^{-1} \int_q^{q_2} x^{-1}(1-x)^{-1}(a+bx)^{-1}\,\mathrm{d}x \\ &= \gamma^{-1} \int_q^{q_2} \left\{x^{-1} + (1-x)^{-1}\right\}(a+bx)^{-1}\mathrm{d}x \\ &= \gamma^{-1}\left\{a^{-1}\ln\left[\frac{q_2(a+bq)}{q(a+bq_2)}\right] + (a+b)^{-1}\ln\left[\frac{p(a+bq_2)}{p_2(a+bq)}\right]\right\} \end{aligned} \tag{A1a}$$

When a tends to 0 and b tends to 1, corresponding to random mating with complete recessivity, we have:

$$T(q) = \gamma^{-1}\left[q^{-1} - q_2^{-1} + \ln\left(q_2 p q^{-1} p_2^{-1}\right)\right]. \tag{A1b}$$

When a tends to 1 and b tends to -1, corresponding to random mating with complete dominance, we have

$$T(q) = \gamma^{-1}\left[p_2^{-1} - p^{-1} + \ln\left(q_2 p q^{-1} p_2^{-1}\right)\right]. \tag{A1c}$$

Similarly, for $a > 0$ and $a + b > 0$ Equation 8b can be written as:

$$\begin{aligned} P_{nc}(q) &= \exp - \left\{\gamma^{-1} \int_q^{q_2} x^{-2}(1-x)^{-1}\ (a+bx)^{-1}\,\mathrm{d}x\right\} \\ &= \exp - \left\{a^{-1}\gamma^{-1}\left[\left(q^{-1} - q_2^{-1}\right) + a^{-1}b\ln\left(\frac{q(a+bq_2)}{q_2(a+bq)}\right)\right] + T(q)\right\} \end{aligned} \tag{A2a}$$

When a tends to 0 and b tends to 1 (complete recessivity), we have:

$$\begin{aligned} P_{nc}(q) &= \exp - \left\{\gamma^{-1} \int_q^{q_2} x^{-3}(1-x)^{-1}\mathrm{d}x\right\} \\ &= \exp - \left\{\gamma^{-1} \int_q^{q_2} \left[x^{-3} + x^{-2} + x^{-1} + (1-x)^{-1}\right]\mathrm{d}x\right\} \\ &= \exp - \gamma^{-1}\left[\frac{1}{2}\left[\left(q^{-2} - q_2^{-2}\right) + \left(q^{-1} - q_2^{-1}\right) + \ln\left(\frac{q_2 p}{q p_2}\right)\right]. \right. \end{aligned} \tag{A2b}$$

When a tends to 1 and b tends to -1 (complete dominance), we have:

$$\begin{aligned} P_{nc}(q) &= \exp - \left\{ \gamma^{-1} \int_{q}^{q_2} x^{-2}(1-x)^{-2}\, \mathrm{d}x \right\} \\ &= \exp - \left\{ \gamma^{-1} \int_{q}^{q_2} \left[x^{-2} + 2x^{-1}(1-x)^{-1} + (1-x)^{-2} \right] \mathrm{d}x \right\} \\ &= \exp - \gamma^{-1} \left[\left(q^{-1} - q_2^{-1} \right) + \left(p_2^{-1} - p^{-1} \right) + 2\ \ln\left(\frac{q_2 p}{q p_2} \right) \right] \end{aligned} \tag{A2c}$$

Equation 8c for $a > 0$ and $a + b > 0$ can be written as:

$$\begin{aligned} P_{nr}(q) &= \exp - \left\{ 2\rho\gamma^{-1} \int_{q}^{q_2} x^{-1}(a+bx)^{-1}\, \mathrm{d}x \right\} \\ &= \exp - \left\{ 2\rho\gamma^{-1} a^{-1} \ln\left[\frac{q_2(a+bq)}{q(a+bq_2)} \right] \right\} \end{aligned} \tag{A3a}$$

When a tends to 0 and b tends to 1, this becomes:

$$\begin{aligned} P_{nr}(q) &= \exp - \left\{ 2\rho\gamma^{-1} \int_{q}^{q_2} x^{-2} \mathrm{d}x \right\} \\ &= \exp - \left\{ 2\rho\gamma^{-1} \left(q^{-1} - q_2^{-1} \right) \right\}. \end{aligned} \tag{A3b}$$

When a tends to 1 and b tends to -1, we have:

$$\begin{aligned} P_{nr}(q) &= \exp - \left\{ 2\rho\gamma^{-1} \int_{q}^{q_2} x^{-1}(1-x)^{-1}\, \mathrm{d}x \right\} \\ &= \exp - \left\{ 2\rho\gamma^{-1} \ln\left(\frac{q_2 p}{q p_2} \right) \right\} \end{aligned} \tag{A3c}$$

Approximate probability of no coalescence conditional on no recombination

For simplicity, only the case of intermediate dominance ($a \neq 0$, $b \neq -1$) will be considered. With large γ, we can write $q_1 \approx 1/(2a\gamma)$, $p_2 \approx 1/[2(a + b)\gamma]$, and $q_2 = p_1 \approx 1$. We can then use Equations A1a and A2a with $q = q_1$. $T(q_1)$ and the multiplicand of b in the exponent are of order $\ln(\gamma)/\gamma$, provided that $a^{-2} << \gamma$. The leading term in the exponent is the product of $-1/(2a\gamma)$ and $1/q_1$, which is approximately equal to -2 under this condition, implying that $P_{nc}(q_1) \approx e^{-2} = 0.135$.

Harmonic mean of q during a sweep

The integral of $1/q$ between q_1 and q_2 is equivalent to the terms in braces in the exponents in the first lines of Equations A3, with $q = q_1$. The harmonic mean of q is given by taking the reciprocal of this integral, after division by T_d. From the above result, the integral is approximately 2 for large γ, so that the harmonic mean of q is approximately equal to $\frac{1}{2}T_d$.