Title: How good are predictions of the effects of selective sweeps on levels of neutral diversity?

Author: Brian Charlesworth

Affiliation:
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

Key words: Selective sweeps, background selection, gene conversion, crossing over, neutral variability, favorable mutations, *Drosophila melanogaster*

Running title: Predicting sweep effects

Corresponding author:

Name: Brian Charlesworth

Address: Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, King's Buildings, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK

Telephone: +44 131 650 5751

Email: Brian.Charlesworth@ed.ac.uk

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 coalescent events that occur during a sweep can be neglected, leading to simple expressions for the effects of sweeps on pairwise nucleotide site diversities and site frequency spectra. It is shown here that there is a significant probability that a pair of alleles sampled at the end of a sweep coalesce during the sweep. In the case of a favorable mutation with intermediate dominance, this probability is approximately equal to 0.135 in the absence of recombination. Expressions are derived for the expected reductions in pairwise neutral diversities caused by both single and recurrent sweeps in the presence of 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 high ratios of the recombination rate to the selection coefficient, the approximation that neglects within-sweep coalescence can be inaccurate by several orders of magnitude. 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 in levels of variability on X chromosomes and autosomes under selection at linked sites are discussed, and compared with data on a population of *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 twenty years later, it was shown that selection against recurrent deleterious mutations can also reduce variability, by the hitchhiking process known as background selection (Charlesworth *et al*. 1993). It is therefore preferable to use the term "selective sweep" (Berry and Kreitman 1993) for the hitchhiking effects of favorable mutations. There is now a large theoretical literature on both types of hitchhiking, recently reviewed by Walsh and Lynch (2018, Chap. 8) and Stephan (2019). With sufficiently weak selection, partially recessive deleterious mutations can also increase, rather than reduce, 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). The expected effects of all three processes on pairwise diversity at a neutral site can be described by the same general formula (Zhao and Charlesworth 2016), which is a consequence of the Price-Robertson equation for the change in mean of a trait induced by its covariance with fitness (Robertson 1968; Price 1970).

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 both the effects of both selective sweeps and background selection (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 Gillespie (2000); see Kern and Hahn (2018) and Jensen *et al*. (2019) for a recent discussion of this question.

As described by Campos and Charlesworth (2019) and Hartfield and Bataillon (2020), several important assumptions underlie a model of selective sweeps that is frequently used for inferring parameters of selection from population genomic data, which traces back to the work of Barton (1998, 2000). (For brevity, Campos and Charlesworth (2019) will be referred to as CC from now on.) The main aim of the present paper is to relax two assumptions made in this model.

The first is that the fixation of a favorable mutation happens so fast that non-recombinant alleles at a linked neutral site, sampled after the completion of the sweep, do not experience any coalescent events until the beginning of the sweep, when they all coalesce simultaneously. The second is that linkage is sufficiently tight that at most a single recombination event occurs during the sweep. (The original model of Maynard Smith and Haigh (1974) was completely deterministic, but allowed for multiple recombination events.)

These assumptions mean that the gene genealogy for a set of alleles sampled immediately after a sweep has a simple "star shape", which enables the site frequency spectrum at the neutral site to be calculated in a straightforward fashion (Barton 2000; Durrett and Schweinsberg 2004; Kim 2006). It also allows a formula for the reduction in pairwise diversity at a neutral site to be obtained from a few lines of algebra, using a coalescent process argument, which can be used to model recurrent sweeps by treating coalescent events caused by classical genetic drift and by sweeps as competing exponential processes (Barton 2000; Weissman and Barton 2012; Berg and Coop 2015; Elyashiv *et al*. 2016; Campos and Charlesworth 2019).

This approach is likely to be accurate for very strongly selected favorable mutations, with their short time to fixation compared with the expected neutral coalescent time of $2N_e$ generations, where N_e is the effective population size. Recent population genomic analyses suggest, however, that there may be important contributions from relatively weakly selected favorable mutations, which can take as long as 10% of the neutral coalescent time to become fixed (Elyashiv *et al*. 2016; Campos *et al*. 2017); some studies have even inferred much weaker selection, with correspondingly longer coalescent times (Sella *et al*. 2009; Keightley *et al*. 2016; Chen *et al*. 2020). With weak selection, coalescence during the sweep cannot necessarily be neglected. Indeed, it is shown here that, in the absence of recombination, the probability that a pair of alleles will coalesce during a sweep of a mutation with intermediate dominance is approximately $1 - e^{-2} \approx 0.135$, almost independently of the strength of selection and genetic system. Consistent with this result, Barton (1998, Figure 1) showed that neglecting coalescence during a sweep can lead to a substantial underestimation of the associated reduction in pairwise diversity at a neutral site, although the results derived here suggest that the underestimation is much greater than was found by Barton.

While a general representation of the dual processes of recombination and coalescence during a sweep is available (Kaplan *et al*. 1989), it does not provide simple formulae, even for pairwise nucleotide site diversity. In the present paper, a general analytical model of selective sweep effects on pairwise neutral diversity is developed, for the case of weak selection at a single

locus. This is based on a recent study of the expected time to fixation of a favorable mutation in a single population (Charlesworth 2020), which provides a general framework for analyzing both autosomal and sex-linked inheritance with arbitrary levels of inbreeding and dominance. The resulting formulae, which include a heuristic treatment of multiple recombination events, 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. 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 and without modelling 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).

Methods

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 kilobases 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 5 exons of 300 basepairs (bp) each, interrupted by 4 introns of 100bp. The lengths of the 5´and 3´ UTRs were 190bp and 280bp, 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 ½ 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 non-crossover 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: 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 $r_c z$, 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 towards 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 towards 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 Section S1 of File S1, 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. Diversity values were thus under-estimated 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 x 1.5. Furthermore, to correct for the effects of interference among co-occurring favourable 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 non-coding 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.

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. The codes for the computer programs used to implement the analytical models described below will be made available in the Supplementary Information, File S2, on Figshare on acceptance. The detailed statistics for the results of the computer simulations were provided in Files S2-S3 of Campos and Charlesworth (2019).

The effect of a single sweep on expected nucleotide site diversity

Theoretical results

All rates and times are 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 1931,1969; 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 $B_1 N_e$, 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.

Recombination is negligible in the final stochastic phase of a sweep, when A_2 is near fixation; in addition, the rate of coalescence is then close to the neutral value. Coalescence is virtually instantaneous in the initial stochastic phase, provided that the frequency of A_2 at the start of the deterministic phase is sufficiently small (Barton 1998). Accurate approximations for the effect of a single sweep on diversity can, therefore, usually be obtained by treating the the beginning and end of the deterministic phase as equivalent to that for the sweep as a whole, as discussed by Charlesworth (2020), and this is what is done here. The framework described by Hartfield and Bataillon (2020) can then be used, extended to include a more general model of selection and the possibility of BGS effects. This approach assumes that all evolutionary forces are weak, so that a continuous time scale approximation to a discrete generation model can be applied.

Let T_d be the duration of the deterministic phase, and write $q_1 = q(T_d) \approx 0$ and $q_2 = q(0) \approx 1$, where q_1 and q_2 are the frequencies of A_2 at the start and end of the deterministic phase, respectively. Expressions for q_1 and q_2 for a hard sweep in the absence of BGS are given by Charlesworth (2020); with BGS, the terms in N_e in the relevant expressions are each to be multiplied by B_2, as 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 $[B_1 q(T)]^{-1}$, 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 $\rho = 2N_e r$ 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 $1 - F$ are sufficiently large that their second-order terms cannot be neglected, we have $c \approx 1 - F$ (Nordborg 1997; Charlesworth and Charlesworth 2010, p.). The exact value of ϕ is determined by the mating system; in the case of self-fertilization, Equation 1 of Hartfield and Bataillon (2020) 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:

$$k_c(T) = [B_1 q(T)]^{-1} P_{nc}(T) P_{nr}(T) \tag{1}$$

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:

$$P_{nc}(T) = \exp - \int_0^T [B_1 q(\tau)]^{-1} \, \mathrm{d}\tau \tag{3}$$

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

The net probability that the pair of sampled alleles coalesce during the 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 (the sweep is “hard”), 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 non-linearity of the equation for the rate of change of q under selection. Following Charlesworth (2020), for the case of weak selection we can write the forward-in-time selection equation as:

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

where tilde is 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):

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

Similar expressions are available for the cases when $a = 0$ (complete recessivity) or $a + b = 0$ (complete dominance), as described by Charlesworth (2020).

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

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

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

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

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

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

The net expected pairwise coalescence time associated with the sweep is:

$$T_s = P_{c2}T_d + T_c \qquad (9a)$$

where

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

and $T(q)$ is given by Equations A1.

Results with only a single recombination event: The contribution of T_s to the mean coalescent time was neglected by Hartfield and Bataillon (2020), who only considered the contribution from the case when at most one of the two alleles has recombined onto the wild-type background without having previously coalesced, giving a coalescence time of B_1 if T_d is negligible. (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 return recombination event from A_1 into A_2.) From Equation 2, the probability of at least one recombination event is given by:

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

Using Equations 6, 10a and A1-A3, P_r can be expressed in terms of ρ, γ, a and b, but 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:

$$-\Delta\pi = B_1(1-P_r) - T_s \qquad (11a)$$

Equation 9 of Hartfield and Bataillon (2020) 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 approximation that ignores coalescence during the sweep (Barton 1998, 2000; Weissman and Barton 2012) :

$$-\Delta\pi \approx B_1 P_{nr}(T_d) \qquad (11b)$$

Explicit formulae for the components of the above equations, using Equation 6, are given in the Appendix. In the case of an autosomal locus with random mating and semi-dominant selection ($h = 0.5$), substitution of these formulae into Equation 11c yield the convenient and frequently used formula:

$$-\Delta\pi \approx B_1 \gamma^{-4\rho/\gamma} \qquad (11c)$$

***The importance of coalescence during a sweep*:** These results bring out the potential importance of considering coalescence during a sweep. Consider the case with incomplete dominance ($a \neq 0$). The probability of no coalescence 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 is approximately 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. Ignoring coalescence during the sweep therefore induces some error into estimates of the reduction in diversity (Stephan *et al*. 1992; Barton 1998).

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 probabilities 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 the supplementary information of 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 ½ T_d for large γ, yielding $T_s \approx 0.635\, T_d$. Table S1 compares the exact 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 towards 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 non-recombinant 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:

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

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 non-recombinant 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:

$$P_{nr}(q_1, q) = \exp -\rho \int_{q_1}^{q} \dot{x}^{-1}\{(1 - x) + x\}\, \mathrm{d}x$$

$$= \exp -\rho\, T(q) \qquad (13b)$$

However, this simple argument 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. $(P_r - P_{rs})\, T_r$ should thus be deducted from the left-hand side of Equation 12, to give:

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

and

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

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

Equations 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 compared with Equation 9 of Hartfield and Bataillon (2020).

Comparisons with simulation results

Numerical results for Equations 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; it was found that $n = 200$ points usually gave values that were close to those for a more exact method of integration; for the results in the figures in this section, 2000 points were used. Background selection effects are ignored here, so that B_1 and B_2 are set to 1. Simulation results for hard sweeps for an autosomal locus with varying rates of self-fertilization (S), described by Hartfield and Bataillon (2020), were used as a basis for comparison with the predictions based on Equations 11a and 14 (denoted by *C1* and *C2*, respectively), on Equations 9 of Hartfield and Bataillon (2020) (*C3*), which ignore the term in T_s, and on the approximation that ignores coalescence during the sweep, Equation 11b (*NC*). The simulation results for the mean reductions in diversity for an autosomal locus in a randomly mating

population with three difference dominance coefficients, together with the four theoretical predictions, are shown in Figure 1.

NC underestimates the reduction in diversity observed at the end of a sweep when the recombination rate is relatively high compared with the selection coefficient. In contrast *C1*, *C2* and *C3* agree quite well with most of the simulation results; *C2* tends to undestimate the effects by a small amount, with the higher rates of recombination. The predictions from Equation 12 are not shown, as they are close to those from *C1*. The underestimation of the sweep effect by *NC* is especially noticeable at high recombination rates and for $h = 0.1$; this can be seen most clearly in the lower panels of Figure 1, which display only the higher range of recombination rates in the simulations ($0.108 \leq r/s \leq 0.228$). This effect is far from trivial for $h = 0.1$ and $h = 0.5$; for example, with $h = 0.5$ and the highest value of ρ in Figure 1, the *C1, C2* and *C3* predictions are 0.048, 0.022, and 0.049 respectively, whereas the *NC* prediction is 0.0035.

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. Again, *C1* and *C3* mostly give very similar predictions, especially with $S = 0.95$, whereas *C2* predicts smaller effects that agree less well with the simulations at the higher recombination rates, as in the case of outcrossing. However, for $S = 0.95$, especially with $h = 0.5$, the simulations yield much larger sweep effects than any of the theoretical predictions. This presumably reflects the fact that random variation among individuals in the occurrence of selfing versus 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).

Inaccuracy of the NC approximation: Given the widespread use of the *NC* formula in attempts to infer the parameters of positive selection from patterns of neutral diversity, it is disconcerting that it is systematically somewhat inaccurate at high rates of recombination. The simulations show clearly that approximations that include coalescence during the sweep are needed, in addition to multiple coalescent events, unless the recombination rate is small relative to the selection coefficient. Figure 3 shows the decline in the reduction in diversity

predicted by the *C1* and *C2* approximations for an autosomal locus in a randomly mating population, as a function of the ratio of the recombination rate to the selection coefficient (r/s) for various combinations of dominance coefficients and strengths of selection. As has been found previously (Teshima and Przeworski 2006; Ewing *et al*. 2011; Hartfield and Bataillon 2020), the reduction in diversity increases with h for low values of r/s, but the values for each h converge as r/s increases.

As r/s approaches 1, the ratio of *C1* to *C2* becomes very large (27.0 for $h = 0.5$, $\gamma = 250$ and $r/s = 1.28$), so that it remains unclear how large a reduction in diversity is expected at high recombination distances. For sufficiently high r/s, *C2* can even become negative, indicating that it over-corrects for multiple recombination events. However, the results for recurrent sweeps described in the next section show that *C1* is not a satisfactory approximation for moderately high recombination rates, whereas *C2* performs quite well.

Multiple recombination events can be included in the purely deterministic model without approximations of the type used to obtain the *C2* formulae. Using the representation of this model in Barton (2000) and Charlesworth and Charlesworth (2010, pp.410-411), a general expression for the expected reduction in diversity relative to neutrality (ignoring BGS) is:

$$-\Delta\pi \approx \{\int_{q_1}^{q_2} \exp[-\rho T(q)]\ \mathrm{d}q\}^2 \qquad (15)$$

Using the expressions for $T(q)$ given by Equations A1, the integral in Equation 15 can be evaluated numerically for any desired selection scheme. For large r/s, Equation 15 would be expected to give a better approximation to the sweep effect than the standard *NC* approximation describe above. This is indeed the case, but the improvement is not sufficient to produce agreement with the simulation results, due to the neglect of coalescence during the sweep. For example, with random mating, $\gamma = 500$, $\rho = 114$ ($r/s = 0.228$) and $h = 0.1$, Equation 15 predicts a diversity reduction of 1.40×10^{-4}, the standard *NC* approximation predicts 1.24×10^{-5}, the *C2* approximation predicts 1.43×10^{-2}, and the simulations gave 2.55×10^{-2}. The larger reduction in diversity predicted by Equation 15 compared with the standard *NC* approximation presumably reflects the fact that a second recombination event can return a neutral site that has recombined onto the wild-type background into association with the sweeping allele, as modeled in CC, File S1. This effect must also contribute to the underestimation of diversity reductions by the *BC2* approximation.

Some insight into the properties of the unexpectedly large diversity reductions at high values of r/s under the *C1* approach can be obtained from examining an approximation for the case of autosomal inheritance with $h = 0.5$ and random mating, which is derived in the Appendix. 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 one minus the probability of recombination during the sweep, as given by Equation 11a with $T_s = 0$. From Equation A7, we have:

$$-\Delta\pi \approx 1 - e^{\alpha}\gamma^{-\alpha} - e^{\alpha}\gamma^{-\alpha}\alpha R\{2[(R-1)(R+1)]^{-1} - \sum_{i=2}^{\infty}[i(R+i)]^{-1}\} \quad (16)$$

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 16 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.

Figure 3 illustrates this effect, using numerical integration rather than Equation 16, although the two agree well for $0.4 \leq r/s \leq 1.28$. A doubling of γ causes an approximate halving 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. 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 versus recombination.

Table 1 gives details of several of the variables described above, for the case of random mating and $\gamma = 250$; other examples are shown in Table S1 of File S1. These results confirm the conclusion that there can be a substantial probability of coalescence during the sweep (P_{c1}, given by 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 probability of no recombination without conditioning on an absence of coalescence, $1 - P_r$, decreases much more slowly with r/s than P_{nr} This explains why the *NC* approximation for the reduction in diversity performs 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, both the conditional mean times to coalescence and to the first recombination event are quite small relative to the duration of the sweep, implying that these events must occur quite soon if they are to occur at all.

The effects of recurrent selective sweeps on nucleotide site diversity

Theoretical results

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.

As a first step, it is useful to note that Equation 9 of CC for the expected nucleotide site diversity immediately after a substitution, π_0, is equivalent to:

$$\pi_0 \approx \bar{P}_r \pi_1 + \bar{T}_s \qquad (17)$$

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 the mean pairwise coalescent time relative to the neutral value, $2N_e$); the overbars indicate expected values. 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 12 to take into account the lag time to coalescence of a pair of alleles separated by recombination, we can write:

$$\pi_0 \approx \bar{P}_r \pi_1 + \bar{T}_s + \bar{P}_{rs} \bar{T}_d \qquad (18)$$

If we use the approach of Equation 13 to model multiple recombination events, we obtain:

$$\pi_0 \approx \bar{P}_r \pi_1 + \bar{T}_s + \bar{P}_{rs} \bar{T}_d + (\bar{P}_r - \bar{P}_{rs}) \bar{T}_r \qquad (19)$$

A somewhat more accurate expression can be obtained 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 Equation 8b of the Appendix); this overestimates the time available for the increase in diversity, since a time greater than T_r is required for randomization to occur.

Following CC, it is assumed that the diversity relative to neutrality for a pair of alleles sampled at a random time point is approximately equal to the mean value of the diversity between successive sweeps, π. This is the chief quantity of interest, and can be found using Equations A10-A13 of the Appendix.

Comparisons with simulation results

The accuracy of Equations A10-A13 was tested using the simulation results from Figure 4 and Table S6 of 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 the Methods section of this paper; full details are given in CC. The effect of BGS on the rate of substitutions of favourable mutations for a given parameter set was calculated here 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 in practice. 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.

Equations A10-13 can be applied in two ways. First, the mean values of $\bar{P}_r$, $\bar{T}_s$, $\bar{T}_r$ and $\bar{T}_d$ 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 the Methods section. 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 4 shows the reductions in neutral diversity per gene obtained from the simulation results (red bars), the *C2* predictions using Equations A10-A12 (blue bars), the predictions from Equations 12 of CC (black bars), 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 the Methods section.

The most notable point is that, as was also found by CC, the standard recurrent sweep formula 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 A10-12 generally have a similar level of accuracy. However, in the presence of gene conversion, the latter predictions perform the best, and provide accurate predictions except for the lowest CO rate, especially in the absence of BGS. 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 shown 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 *C1* predictions were found to overestimate the effects of sweeps compared with the simulations, especially at high CO rates (because *C3* gives slightly larger values than *C1*, it was not used for the results discussed here). For example, for the case with gene conversion and no BGS in Figure 4, the *C1*predictions for reduction in diversity for relative CO rates of 0.5, 1.0,

1.5, 2.0 and 2.5, respectively, are 0.346, 0.291, 0.256, 0.231 and 0.212, whereas the corresponding simulation means are 0.315, 0.233, 0.185, 0.150 and 0.137.

As discussed in the section on single sweep effects, the failure of the *NC* approximation at high recombination rates reflects the fact that the unconditional probability of no recombination events during a sweep, $1 - P_r$, declines more slowly with increasing recombination rate than than the probability of no recombination conditioned on no coalescence, P_{nr}. This suggests that it might be possible to equate the rate of coalescence events induced by sweeps (denoted by S^{-1} in CC) to the sum of the products of the values of $1 - P_r$ with the rates of substitutions for all classes of sites, and to substitute this into Equation 7 of CC. The results of applying this method to calculate the mean reduction in diversity over all sites show that it indeed provides a much better fit to the simulation results than the standard *NC* approximation or the prediction from Equation 12 of CC, but generally does not provide as good a fit as the *C2* approximation, especially at high recombination rates. For example, in the case of gene conversion but no BGS shown in Figure 4, the simulation results gave diversity reductions of 0.315 and 0.137 for relative CO rates of 0.5 and 2.5, respectively; the corresponding reductions predicted by the present method are 0.296 and 0.179, the standard *NC* predictions are 0.183 and 0.084, and the *C2* predictions are 0.274 and 0.141. The inaccuracy of this approach partly comes from the fact that, like other results based on competing exponential processes of coalescence by sweeps and by drift (Kaplan *et al*. 1989; Stephan *et al*. 1992; Kim and Stephan 2000; Gillespie 2000), it treats sweeps as point events, allowing too much opportunity for drift-induced coalescent events between sweeps (p.293 of CC).

X chromosomes versus autosomes

Introduction

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 versus 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 (Nagylaki 1979). The values of the coefficients a and b in Equation 6 for 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, the values of s for the cases of male- and female-only effects with X-linkage are set 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 5 shows the reductions in diversity at the end of a sweep predicted by the *C1* and *C2* methods, for the case of an X-linked locus whose effective population size, kN_{e0}, is three-quarters of that for A, N_{e0}. As in Figures 1 and 2, gene conversion is absent. This case corresponds to a randomly mating population in which males and females have equal variances in reproductive success (Wright 1931). 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 inccuracies in the integrations. This expectation is confirmed by the results shown in Figure 5. 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).

For the *Drosophila* model (the upper panel), the comparison of the X-linked results with $h = 0.5$ with the autosomal results for $\gamma = 250$ in Figure 3 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 non-sex-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.

For the mammalian model (lower panel of Figure 5), the results are broadly similar, 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.

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 greater than 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 $k \neq 0.75$ are shown in Figures S1 and S2 in File S1, for the *Drosophila* and mammalian models, respectively. Comparing these with Figure 5, it can be seen that smaller k values cause somewhat larger X-linked sweep effects for all modes of selection and for both genetic systems. 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 A11-A13 (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 X is $2kN_{e0}$ instead of $2N_{e0}$, with k generally expected to be less than 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 non sex-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 rate of substitution are given in the final section of the Appendix. These

opposing effects of sex linkage implies that simple generalizations about the effects of sweeps on X-linked versus autosomal variability cannot easily be made, as will shortly be seen.

Figure 6 shows the 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 s, using the same gene structure that was used to generate the theoretical results for autosomes shown in Figure 4 (see Methods section). Here $k = 0.75$, so that potential effects of sexual selection or variance in female reproductive status are absent. Gene conversion was allowed in this case. 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).

Because the *C2* approximation gave the best fit to the simulation results for autosomes (Figure 4), only results for this method are displayed here, although similar general patterns are found with the *C1* and *NC* methods. As described in the Methods section, the values of the BGS parameter B_1 were obtained from estimates given by Charlesworth (2012), which include contributions from selectively constrained non-coding 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 over 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 $h = 0.5$ (upper panel of Figure 6) 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 influence of the lower

N_e for X than A 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 non-sex-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 S3 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 7 shows the values of the X/A ratio of diversities (R_{XA}) for different CO rates, obtained from the results shown in Figure 6. 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 towards $R_{XA} > 0.75$ for low CO rates, declining towards 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 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 non-sex-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. The comparable results without gene conversion are shown in Figure S4 of File S1. Again, the patterns are similar

to those found with gene conversion; 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 S5 and S6, 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 a 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

The results derived here for the effects of selective sweeps on pairwise neutral diversity (π) show that ignoring coalescent events that occur during the course of a sweep leads to inaccuracies in predictions concerning reductions in π, especially when the ratio of recombination rate to the homozygous selection coefficient (r/s) is relatively high, consistent with the findings of Barton (1998) and Hartfield and Bataillon (2020). For example, Figure 1 shows that, for a semi-dominant favorable mutation with $\gamma = 2N_e s = 500$, the assumption that coalescence occurs only at the start of the sweep (the *NC* prediction) gives an end-of-sweep reduction of 0.01 when $r/s \approx 0.2$, whereas the simulations and predictions that include coalescence during the sweep (*C1*-*C3*) give values that are approximately five times higher. Similarly, Figure 4 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, whereas the *C2* approximation derived here fits much better. As described earlier, with recurrent sweeps the *C1* and *C3* approximations greatly overestimate the diversity reductions 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 events, ($1 - P_r$), when taking coalescence into account (see Table S1 of File S1). Of course, if BGS effects are included, the errors with the *NC* approximation are reduced, since BGS can have a major effect on diversity. 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* and *C3* approximations ignore these events completely. Table S1 shows 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 sweeps that should, in principle, allow for multiple recombination events and coalescence, have been given by (Stephan *et al*. 1992) and Barton (1998) for the case of semi-dominant selection with autosomal inheritance and random mating. However, these do not yield significant improvements over the *NC* approximation for large r/s values. For example, for the example in Figure 1 with $h = 0.5$ and $\rho = 102$ ($r/s = 0.204$), Equation 19 of Stephan *et al*. (1992) predicts a reduction in diversity at the end of a sweep of 0.0063, Equation 7 of Barton (1998) gives a value of 0.0070, and the *NC* approximation gives 0.0056. In contrast, the simulation value of Hartfield and Bataillon (2020) is 0.040, and the *C1* and *C2* approximations are 0.048 and 0.025, respectively. The discrepancy apppears to arise from a neglect of stochastic effects on diversity early in the sweep (Stephan *et al*. 1992, p.244) or late (Barton 1998, p.126).

Even for recombination events within genes, relatively large r/s values are likely. For example, with the parameters used for Figure 4, $s = 1.25 \times 10^{-4}$ in a population with $N_e = 10^6$. With the standard effective autosomal CO rate of 1×10^{-8} per bp, but without gene conversion, the recombination rate between two sites 1kb 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 440bp, the recombination rate is 1.88×10^{-5}, and $r/s = 0.15$. Figure 3 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*; 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 (Hammer *et al*. 2010; Halligan *et al*. 2013) 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 4 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 (Kessler *et al*. 2020). 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 non-crossover 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 artefacts of their use of classical marker-based maps (Fiston-Lavier *et al*. 2010) or the Loess smoothing procedure applied to the 100kb 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) in connection with estimates of the relation between the rate of adaptive evolution of protein sequences and the CO rate. A diminishing returns relation between non-

coding site diversity for the Raleigh population of *D. melanogaster* and CO rate was found by Comeron *et al*. (2012) 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 artefactual, leading to a flatter relation between π and CO rate than truly exists. In addition, Comeron's non-coding 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 non-coding sites involve at least some sequences that are subject to selection. This could easily lead to a less than linear relation between π and CO rate. The question of the true empirical relationship between 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) is equal to 0.75 were shown in Figure 7). 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 greater than 1 or close to 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 close to 0.8 with the other dominance coefficients; for the other modes of selection it is close to 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 non-coding 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 equilbrium 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 $\beta\,(0 \leq \beta \leq 1)$ 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 both 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 A8a applied to X and A diversities relative to their final equilibrium values, R_{XA} at time T is given by:

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

This expression shows that, as expected, R_{XA} is equal to k when $T = 0$, and also when $T \gg k\beta$. 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 (see Section 1 of File S1).

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 ten-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 $T = 1$.

Equation 20 also sheds light on the effect of a population size change on the relative diversity values within X or A at different recombination rates. If β is set to 1, k can be interpreted as a measure of the ratio of π values for two different CO rates, and R_{XA} is equivalent to the ratio at a given time after a population size change. 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 towards low frequency variants, with Tajima's D values 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; 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 1 and 2. I also thank Josep Comeron for helpful discussions, and for sharing his analyses of Drosophila polymorphism data.

Literature Cited

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149-1152.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet.Res. 72: 123-134.

Barton, N. H., 2000 Genetic hitchhiking. Phil. Trans. R. Soc. B. 355: 1553-1562.

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.

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

Berry, A. J., and M. Kreitman, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. Genetics 129: 869-893.

Betancourt, A. J., Y. Kim and H. A. Orr, 2004 A pseudohitchhiking model of X vs. autosomal diversity. Genetics 168: 2261-2269.

Booker, T. R., B. C. Jackson and P. D. Keightley, 2017 Detecting positive selection in the genome. BMC Biology 15: 98.

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. 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.

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

Campos, J. L., 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.

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.

Charlesworth, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77: 153-166.

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.

Charlesworth, B., 2020 How long does it take to fix a favorable mutation, and why should we care? Am. Nat. 195: 753-771.

Charlesworth, B., J. L. Campos and B. C. Jackson, 2018 Faster-X evolution: theory and evidence from *Drosophila*. Mol. Ecol. 27: 3753-3771.

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.

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

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.

Comeron, J., R. Ratnappan and S. Bailin, 2012 The many landscapes of recombination in *Drosophila melanogaster*. PLos Genetics 8: e1002905.

Comeron, J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genetics 10: e1004434.

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.

Corbett-Detig, R. B., D. L. Hartl and T. B. Sackton, 2015 Natural selection constrains neutral diversity across a wide range of species. PloS Biology 13: e1002112.

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. Pop. Biol. 66: 129-138.

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 Genetics 12: e1006130.

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.

Ewing, G. B., and J. D. Jensen, 2016 The consequences of not accounting for background selection in demographic inference. Mol. Ecol. 25: 135-141.

Fiston-Lavier, A., N. D. Singh, M. Lipatov and D. A. Petrov, 2010 *Drosophila melanogaster* recombination rate calculator. Gene 463: 18-20.

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.

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.

Gillespie, J. H., 2000 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. 81: 310-316.

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: R67.61-68.

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: R18.11-17.

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*., 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PloS Genetics 9: e10003995.

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. Nature Genetics 42: 830-831.

Hartfield, M., and T. Bataillon, 2020 Selective sweeps under dominance and inbreeding. G3. Genes, Genomes, Genetics 10: 1063-1075.

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.

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.

Johri, P., Charlesworth B. and J. D. Jensen, 2020 Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics 215: 173-192.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitch-hiking" 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.

Kelly, J. K., 2007 Mutation-selection balance in mixed mating populations. J. Theoret. Biol. 246: 355-365.

Kern, A. D., and M. W. Hahn, 2018 The neutral theory in light of natural selection. Mol. Biol. Evol. 35: 1366-1371.

Kessler, M. D., D. P. Loesch, J. A. Perry, N. L. Heard-Costa, D. Taliun *et al*., 2020 De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proc. Natl. Acad. Sci. USA 117: 2560-2560.

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics 172: 1967-1978.

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. Pop. Biol. 2: 174-208.

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.

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.

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.

Messer, P. W., and D. A. Petrov, 2013 Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. USA 110: 8615-8620.

Nagylaki, T., 1979 Selection in dioecious populations. Ann. Hum. Genet. 14: 143-150.

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. Nature Reviews Genetics 14: 83-87.

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.
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.
Price, G. R., 1970 Selection and covariance. Nature 227: 520-521.
Rice, W. R., 1984 Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735-742.
Robertson, A., 1968 The spectrum of genetic variation, pp. 5-16, edited by R.C. Lewontin Syracuse University Press, Syracuse, NY.
Roze, D., 2009 Diploidy, population structure, and the evolution of recombination. Am. Nat. 174: S79-S94.
Sella, G., D. A. Petrov, M. Przeworski and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? Public Library of Science Genetics 6: e1000495.
Stephan, W., 2019 Selective sweeps. Genetics 211: 5-13.
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. Pop. Biol. 41: 237-254.
Teshima, K. M., and M. Przeworski, 2006 Directional positive selection on an allele of arbitrary dominance. Genetics 172: 713-718.
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. Pop. Biol. 61: 311-318.
Vicoso, B., and B. Charlesworth, 2009 Effective population size and the Faster-X effect: an extended model. Evolution 63: 2413-2426.
Walsh, J. B., and M. Lynch, 2018 Evolution and selection of quantitative traits. Oxford University Press, Oxford.
Weissman, D. B., and N. H. Barton, 2012 Limits to the rate of adaptive substitution in sexual populations. PLoS Genetics 8: e1002740.
Wilson Sayres, M. A., 2018 Genetic diversity on the sex chromosomes. Genome Biol. Evol. 10: 1064-1078.
Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97-159.
Wright, S., 1969 Evolution and the Genetics of Populations. Vol. 2. The Theory of Gene Frequencies. University of Chicago Press, Chicago, Illinois.
Zeng, K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity 110: 363-371.
Zhao, L., and B. Charlesworth, 2016 Resolving the conflict between associative overdominance and background selection. Genetics 203: 1315-1334.

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:

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

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

$$T(q) = \gamma^{-1}[q^{-1} + \ln(pq^{-1}p_2^{-1}) - 1] \qquad \text{(A1b)}$$

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

$$T(q) = \gamma^{-1}[p_2^{-1} + \ln(pq^{-1}p_2^{-1}) - p^{-1}] \qquad \text{(A1c)}$$

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

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

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

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

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

$$P_{nc}(q) = \exp - \{\gamma^{-1} \int_q^{q_2} x^{-2}(1-x)^{-2}\,\mathrm{d}x\}$$
$$= \exp - \{\gamma^{-1} \int_q^{q_2} [x^{-2} + 2x^{-1}(1-x)^{-1} + (1-x)^{-2}]\mathrm{d}x\}$$
$$= \exp - \gamma^{-1}[(q^{-1} - q_2^{-1}) + (p_2^{-1} - p^{-1}) + 2\ln(\tfrac{q_2 p}{q p_2})] \quad \text{(A2c)}$$

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

$$P_{nr}(q) = \exp - \{2\rho\gamma^{-1} \int_q^{q_2} x^{-1}(a+bx)^{-1}\,\mathrm{d}x\}$$
$$= \exp - \{2\rho\gamma^{-1} a^{-1} \ln\left[\tfrac{q_2(a+bq)}{q(a+bq_2)}\right]\} \quad \text{(A3a)}$$

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

$$P_{nr}(q) = \exp - \{2\rho\gamma^{-1} \int_q^{q_2} x^{-2}\,\mathrm{d}x\}$$
$$= \exp - \{2\rho\gamma^{-1}(q^{-1} - q_2^{-1})\} \quad \text{(A3b)}$$

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

$$P_{nr}(q) = \exp - \{2\rho\gamma^{-1} \int_q^{q_2} x^{-1}(1-x)^{-1}\,\mathrm{d}x\}$$
$$= \exp - \{2\rho\gamma^{-1} \ln(\tfrac{p_2 p}{q p_2})\} \quad \text{(A3c)}$$

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 ½ T_d.

Probability of recombination during a sweep, for large values of *r/s*

A semi-dominant ($h = 0.5$) autosomal locus with random mating is assumed. Using the notation described in the main text, together with Equations A1a, A2a, and A3a, Equation 10 can be written as:

$$P_r(q) = Re^{\alpha}\gamma^{-\alpha}\int_{q_1}^{q_2} q^{R+\alpha-1}p^{-\alpha}\exp\ -(\alpha\, q^{-1})\ \mathrm{d}q \qquad \text{(A4)}$$

Assuming that $\gamma >> 1$ (so that $\alpha << 1$), and $R > 1$, the integrand can be approximated by:

$$f(q) = \ q^{R-1}p^{-\alpha}[1 - \alpha q^{-1}]$$

Expanding $p^{-\alpha}$ as a binomial series in powers of q, the integrand can be written as:

$$q^{R-1} - \alpha\, q^{R-2} \ + \ \alpha q^{R}[1 - \alpha q^{-1}] + \sum_{i=2}^{\infty}\frac{\alpha(\alpha+1)\ldots(\alpha+i-1)}{i!}[q^{R+i-1} - \alpha q^{R+i-2}] \quad \text{(A5a)}$$

Neglecting higher powers in α, this can be approximated by:

$$q^{R-1} - \alpha\, q^{R-2} + \ \alpha q^{R} + \alpha\sum_{i=2}^{\infty} i^{-1}\, q^{R+i-1} \qquad \text{(A5b)}$$

This expression can be integrated term by term to yield the following indefinite integral:

$$R^{-1}q^{R} - \alpha(R-1)^{-1}\, q^{R-1} + \ \alpha(R+1)^{-1}q^{R+1} + \alpha\sum_{i=2}^{\infty}[i(R+i)]^{-1}\, q^{R+i} \quad \text{(A6a)}$$

We can then use $q_1 = q_2 = 1/\gamma$ to obtain the definite integral in Equation A4, again neglecting higher order terms in α:

$$R^{-1} - \alpha(R-1)^{-1} + \alpha(R+1)^{-1} + \alpha\sum_{i=2}^{\infty}[i(R+i)]^{-1} \qquad \text{(A6b)}$$

Simplifying the first two terms in α in this expression, and multiplying by the terms outside the integral in Equation A4, we obtain the following expression for the probability of recombination during a sweep:

$$P_r \approx e^{\alpha}\gamma^{-\alpha}\{1 - 2\alpha R[(R-1)(R+1)]^{-1} + \alpha R\sum_{i=2}^{\infty}[i(R+i)]^{-1}\} \qquad \text{(A7)}$$

Increase in expected diversity of recombinant alleles at the start of a sweep

At the start of a sweep, the expected diversity (relative to the neutral expectation) is π_1. If there is sufficient recombination that recombinant alleles have an effective population size of B_1 relative to neutrality, their expected relative pairwise diversity relative to neutrality at a given time T since the start of the sweep is given by:

$$\pi(T) = 1 - (1 - \pi_1 B_1^{-1})\exp(-B_1^{-1}T) \qquad \text{(A8a)}$$

(Compare with Equation 11 of CC).

For sweep of type i, with duration T_{di} and mean time to the first recombination event T_{ri}, the expected diversity at the time of the recombination event is thus:

$$\begin{aligned}\pi(T_{ri}) &= B_1\{1 - (1 - \pi_1 B_1^{-1})\exp[-B_1^{-1}(T_{di} - T_{ri})]\} \\ &= B_1[1 - (1 - \pi_1 B_1^{-1})E_i] \qquad \text{(A8b)}\end{aligned}$$

If we assume that this is approximately the same as the mean coalescent time associated with multiply-recombinant haplotypes, the diversity at the end of a sweep of this type can be written as:

$$\begin{aligned}\pi_{0i} &\approx (P_{ri} - P_{rsi})B_1[1 - (1 - \pi_1 B_1^{-1})E_i] + P_{rsi}\pi_1 + T_{si} + P_{rsi}T_{di} \\ &= [P_{rsi} + (P_{ri} - P_{rsi})E_i\,]\,\pi_1 + (P_{ri} - P_{rsi})B_1(1 - E_i) + T_{si} + P_{rsi}T_{di} \qquad \text{(A9)}\end{aligned}$$

If the rate of occurrence of sweeps of type i is ω_i, their frequency among all sweeps is equal to $f_i = \omega_i / \omega$. The expected diversity at the start of a sweep is given by:

$$\pi_0 \approx [\bar{P}_{rs} + G]\ \pi_1 + (\bar{P}_r - \bar{P}_{rs})B_1 - GB_1 + \bar{T}_s\ + \bar{P}_{rs}\bar{T}_d \qquad \text{(A10a)}$$

where

$$G = \sum_i f_i(P_{ri} - P_{rsi})E_i \qquad \text{(A10b)}$$

From Equation 10 of CC, we also have:

$$\pi_1 = B_1[1 - (1 - \pi_0 B_1^{-1})A) \qquad \text{(A11a)}$$

where

$$A = \omega/(1 + B_1^{-1}) \qquad \text{(A11b)}$$

Furthermore, by Equation 10 of CC, the mean diversity over the interval between sweeps, π, is given by:

$$\pi = [1 - (1 - \pi_0 B_1^{-1})]B_1^2 \omega I(\omega, B_1) \qquad \text{(A12)}$$

where I is the integral defined in Equation 10 and section 5 of File S1of CC.

As in CC, these expressions allow π_0, π_1 and π to be determined explicitly. Let $C_1 = \bar{P}_{rs} + G$ and $C_2 = (\bar{P}_r - \bar{P}_{rs})B_1 - GB_1 + \bar{T}_s\ + \bar{P}_{rs}\bar{T}_d$. Using Equations A10 and A11, we have:

$$\pi_1 = [AC_2 + (1 - A)B_1]/(1 - AC_1) \qquad \text{(A13)}$$

Substituting this expression into Equation A10a yields an expression for π_0, which in turn allows π to be determined from Equation A12.

Rates of substitution of new favorable mutations for arbitrary genetic systems

Following Charlesworth *et al*. (2018), let N_H be the total number of haploid copies among breeding adults for a given genetic system, where $N_H = 2N$ for A, $N_H = 3N/2$ for X, and N is the number of breeding adults. Let the effective population size for the genetic system in question be kN_{e0}, where N_{e0} is the effective population size for A. Under the selection model described in the text (Equation 6), the probability of fixation of a new mutation is $4as(kN_{e0}/N_H)$ (Charlesworth 2020), and the rate of entry of new favorable mutations into the

population each generation is $N_H u p_a$, where p_a is the proportion of mutations that are favourable. The rate of substitution of favorable mutations per generation is thus $4N_H u p_a\,(as\ kN_{e0}/N_H) = 4kN_{e0}\,up_a\,(as) = 2up_a\,(ak\gamma_0)$, where γ_0 is the scaled selection coefficient for autosomal mutations. The expected number of substitutions over one unit of coalescent time for the given genetic system is thus $4kN_{e0} \times up_a\,(ak\gamma_0) = \pi p_a(ak\gamma_0)$, where $\pi = 4kN_{e0}u$ is the expected neutral nucleotide site site diversity with mutation rate u. The relevant formulae for a described in the text can be used to obtain the substitution rates for the genetic system of interest.

Table 1 Parameters describing the effect of a single sweep

$\gamma = 250$

h = 0.1, T_d = 0.111, approx. T_s = 0.0703, P_{nc} = 0.229

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
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} = 0.118

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
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 = 0.111, approx. T_s = 0.0705, P_{nc} = 0.118

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
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 a 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.

Figures and Figure Legends

Figure 1 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 ($2N_{e0}r$) for a randomly mating population. The results for three different values of the dominance coefficient (h) are displayed. A population size of 5000 is assumed, with a scaled selection coefficient ($\gamma = 2N_{e0}s$) of 500. The red circles are the mean values from computer simulations (Hartfield and Bataillon 2020); the filled blue circles are the *NC* predictions; the green triangles are the *C3* predictions from Hartfield and Bataillon (2020); the black crosses and blue squares are the *C1* and *C2* predictions, respectively. The upper panels show the full range of results corresponding to the simulation parameters; the lower panels show only the results for $\gamma \geq 54$ for an expanded Y axis.

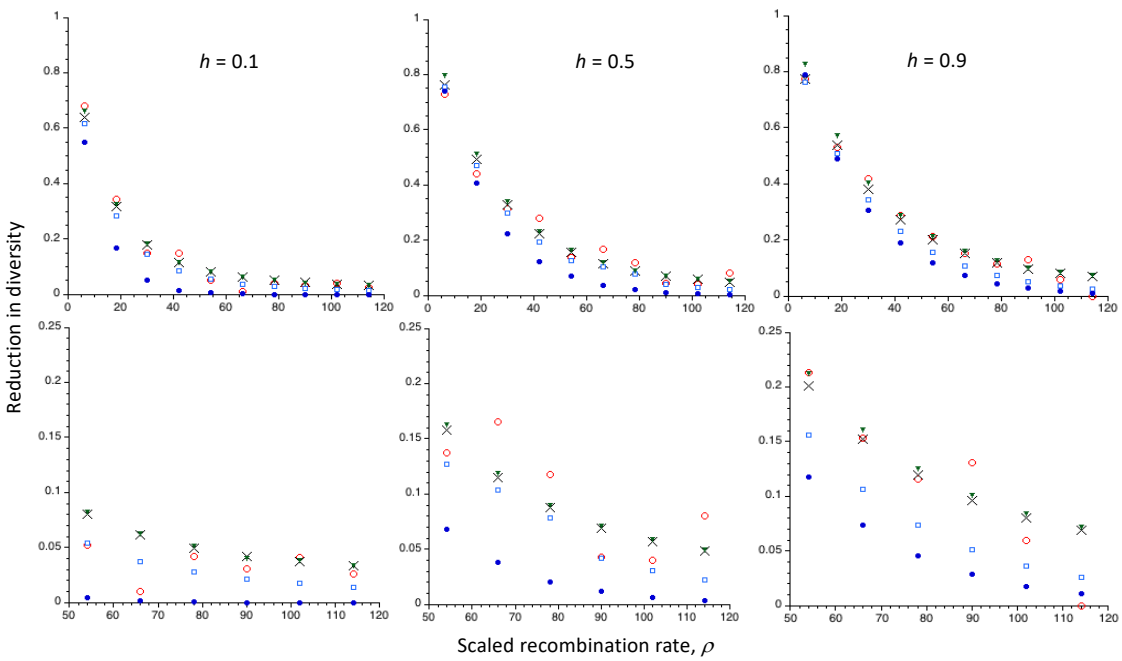


Figure 2_ The reduction in diversity (relative to the neutral value) at the end of a sweep for an autosomal locus, as a function of the scaled rate of recombination ($2N_{e0}r$) for a randomly mating population. The results for three different values of the dominance coefficient (h) and two different selfing rates are shown. A population size of 5000 is assumed, with a scaled selection coefficient ($\gamma = 2N_{e0}s$) of 500. The red circles are the mean values from computer simulations (Hartfield and Bataillon 2020); the filled blue circles are the *NC* predictions; the green triangles are the *C3* predictions from Hartfield and Bataillon (2020); the black crosses and blue squares are the *C1* and *C2* predictions, respectively.

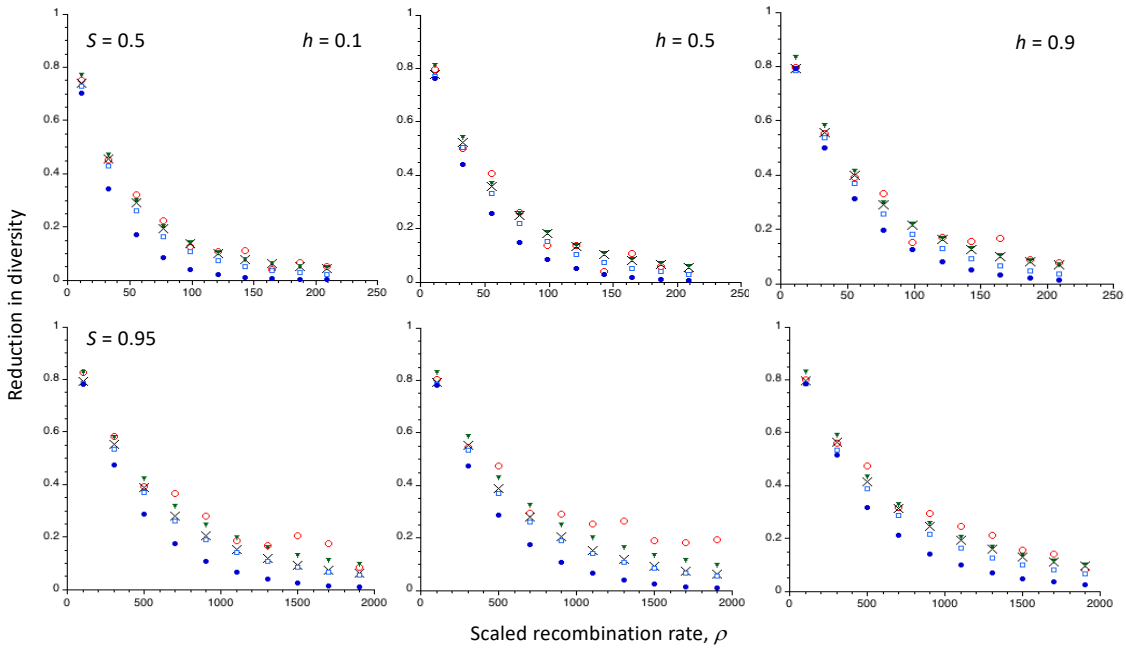


Figure 3 The reduction in diversity (relative to the neutral value) at the end of a sweep for an autosomal locus, as a function of the ratio of the recombination rate to the selection coefficient (r/s), for an autosomal locus in a randomly mating population with three different values of the dominance coefficient (h). A population size of 10^6 is assumed. The solid bars indicate results for $\gamma = 125$, the stippled bars are for $\gamma = 250$, and the hatched bars are for $\gamma = 500$.The colors red and blue indicate results for the *C1* and *C2* approximations, respectively.

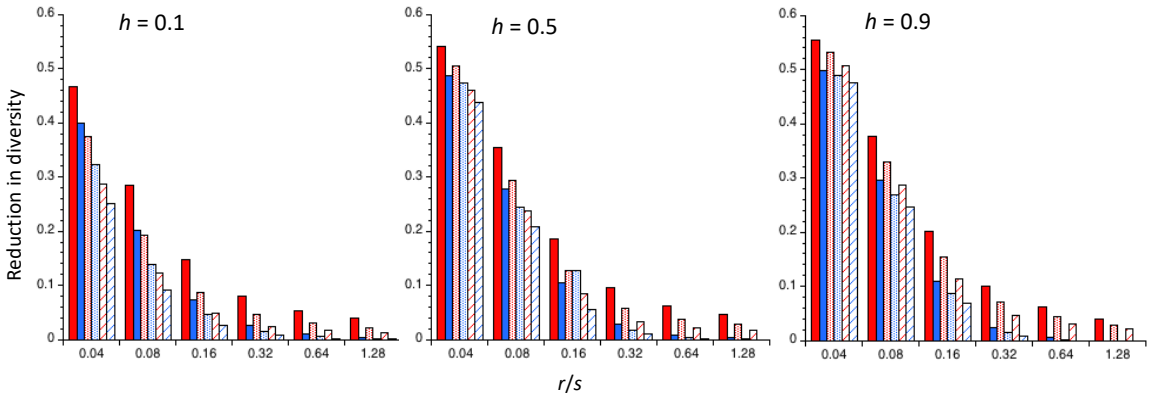


Figure 4 Comparisons with several different theoretical predictions of the mean values of the reduction in diversity obtained in computer simulations of recurrent sweeps with random mating, autosomal inheritance and semidominant favorable mutations (CC, Figure 4 and Table S6). The X axis shows the values 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 A10-1a3; the black bars are the predictions from Equation 12b of CC; the white bars are the predictions from the standard coalescent model of recurrent sweeps, assuming no coalescence within 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.

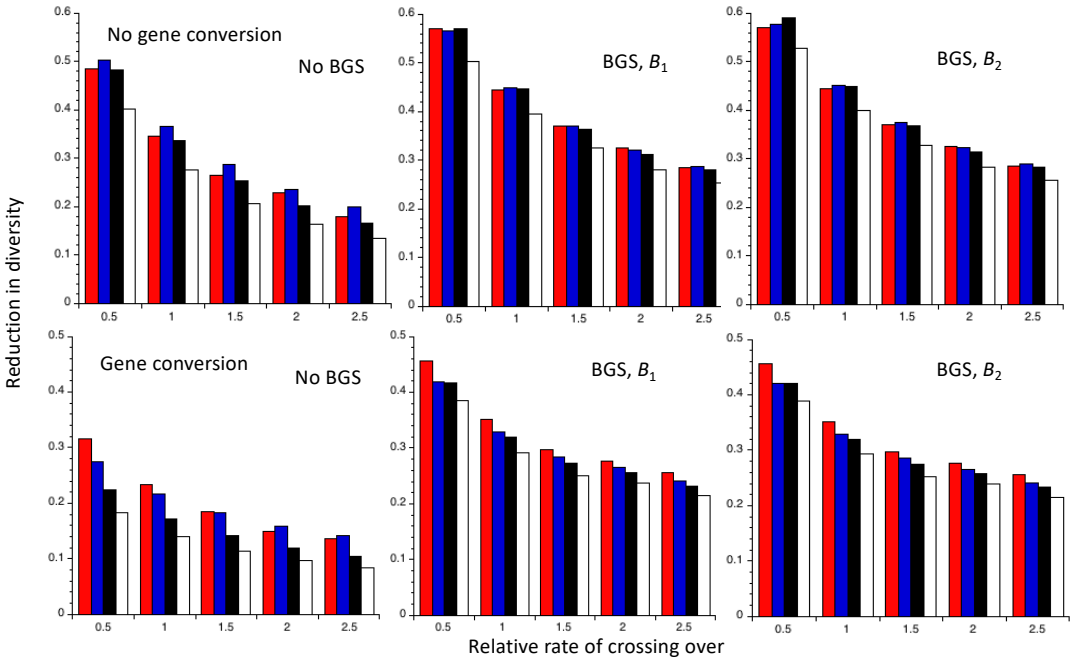


Figure 5 Reductions in diversity (relative to the neutral value) at the end of a sweep for an X-linked locus in a randomly mating population, as a function of the ratio of the autosomal value of the rate of crossing over to the homozygous selection coefficient (r/s), for three different dominance coefficients (h). The upper panel is for a *Drosophila* model, with no crossing over in males; the lower panel is for a mammalian model, with equal rates of crossing over for autosomes in both sexes. Gene conversion is absent. The red and blue colors denote the *C1* and *C2* predictions, respectively. The full bars denote results for mutations with equal effects in both sexes, with $\gamma = 250$. The stippled bars denote male-limited mutations and the hatched bars denote female limited mutations. For the sex-limited cases, $\gamma = 500$ to ensure comparability to autosomal mutations with $\gamma = 250$. A population size of 10^6 is assumed.

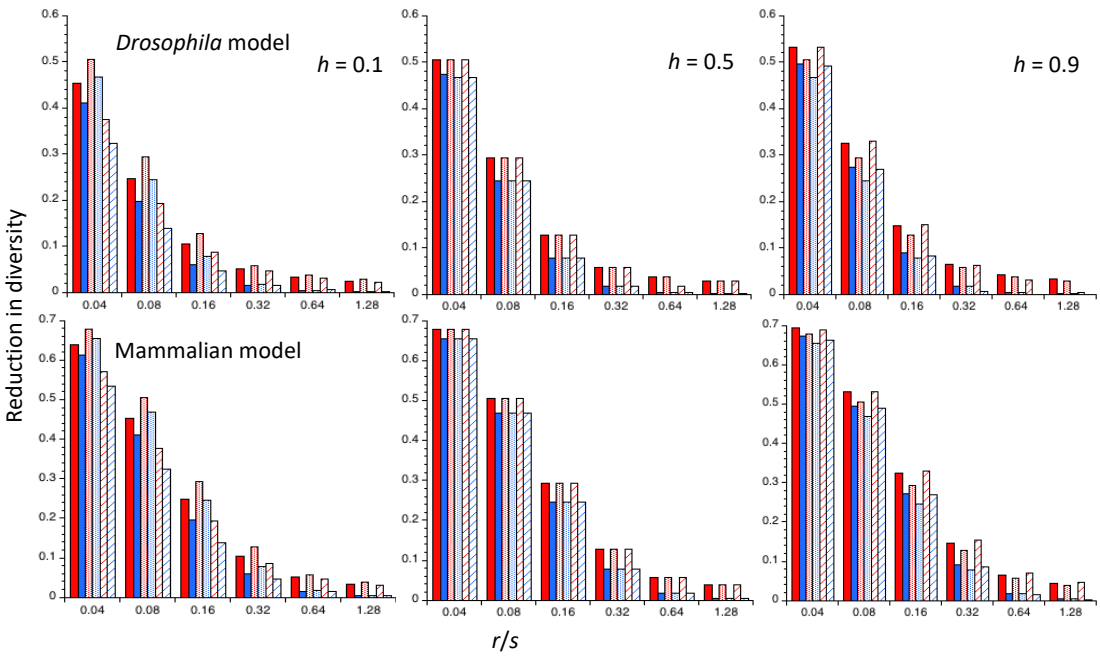


Figure 6 Reductions in diversity (relative to neutrality) under recurrent sweeps at autosomal and X-linked loci for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates 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 (using 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 black 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, γ = 500 to ensure comparability with the autosomal and non-sex-limited X-linked mutations. A population size of 10^6 is assumed.

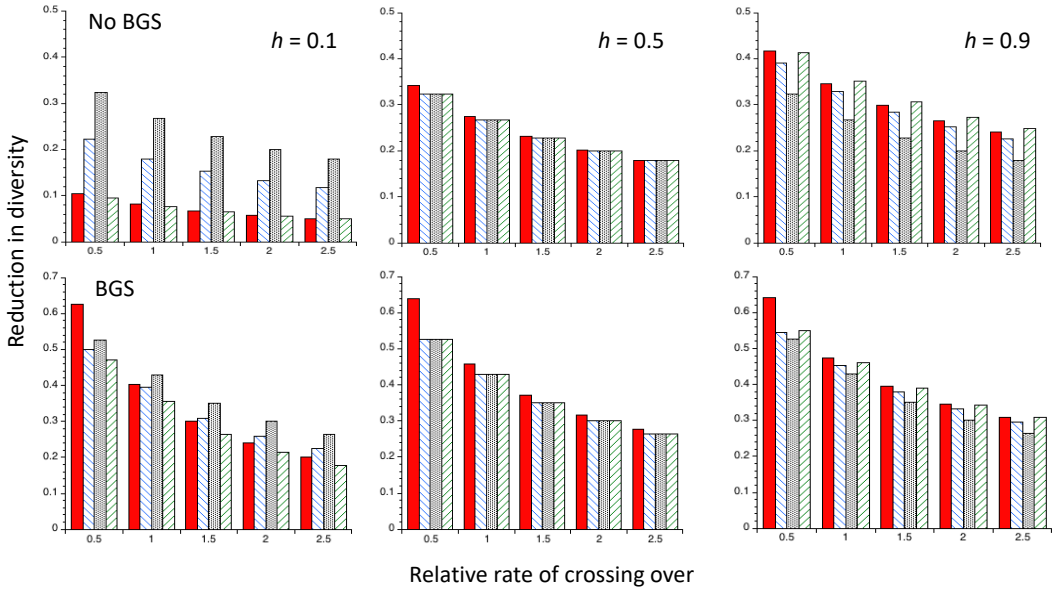
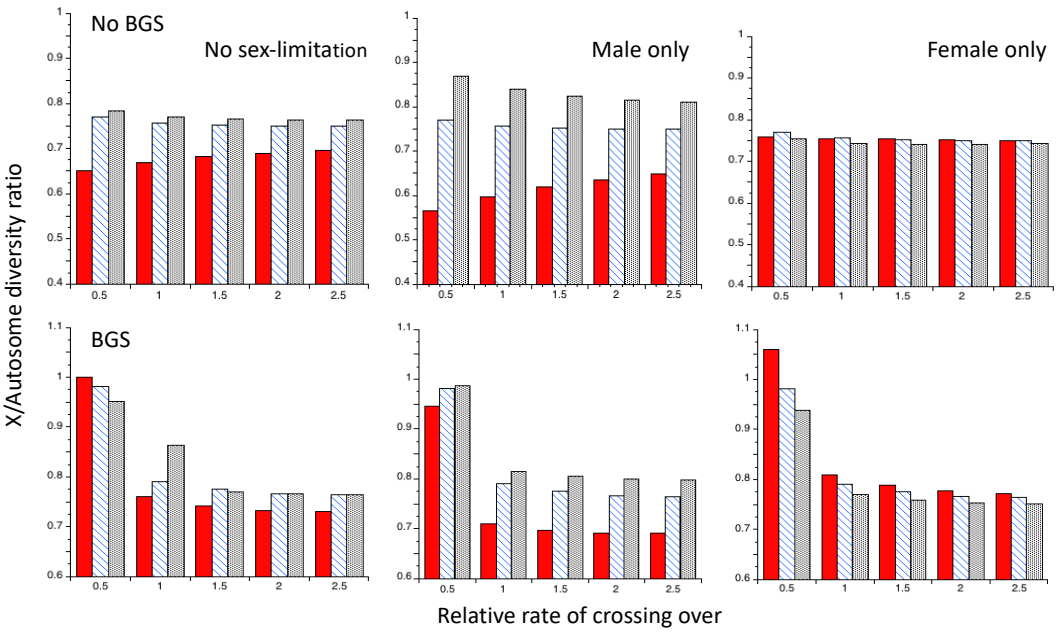


Figure 7 The ratios of X chromosome to autosome nucleotide site diversities (R_{XA}) for the *Drosophila* model under recurrent sweeps, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of 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 dominance coefficient of favorable mutations), the hatched blue bars are for $h = 0.5$ and the black stippled bars are for $h = 0.9$. The other details are as for Figure 6.



Section S1

Derivation of Frisse et al. formula for the effect of gene conversion

Consider two sites (1 and 2, with 1 to the left of 2) separated by z basepairs; let the rate of initiation of a gene conversion tract be r_g and the mean tract length by d_g. Assume an exponential distribution of tract lengths, with rate parameter $\lambda = 1/d_g$. If a tract is initiated to the left of 1, it has a probability of ½ of moving towards 1. If it is initiated at distance y from 1, the probability that it includes 1 but falls short of 2, resulting in the conversion of 1 but not 2, is given by:

$$\lambda \int_{y}^{y+z} \exp(-\lambda x)\ \mathrm{d}x = \exp(-\lambda y)[1 - \exp(-\lambda z)] \qquad \text{(S1)}$$

The net probability of conversion of 1 from this class of event is then given by:

$$\tfrac{1}{2} r_g \int_0^{\infty} \exp(-\lambda y)\,[1 - \exp(-\lambda z)]\ \mathrm{d}y = \tfrac{1}{2} r_g d_g \left[1 - \exp\left(-z/d_g\right)\right] \qquad \text{(S2)}$$

We also need to consider the class of events where a tract is initiated between sites 1 and moves the right including 2. If it starts at a distance u from 1, the probability that it includes 2 is given by:

$$\lambda \int_{z-u}^{\infty} \exp(-\lambda x)\ \mathrm{d}x = \exp[-\lambda(z-u)] \qquad \text{(S3)}$$

The net probability of a tract of this class converting site 2 but not 1 is:

$$\tfrac{1}{2} r_g \int_0^{z} \exp[-\lambda(z-u)]\ \mathrm{d}u = \tfrac{1}{2} r_g d_g [1 - \exp(-z/\,d_g)] \qquad \text{(S4)}$$

The same argument applies to tracts moving from right to left, so the net probability of recombination between sites 1 and 2, due to one site but not the other being included in a conversion tract, is:

$$2 r_g d_g [1 - \exp(-z/\,d_g)] \qquad \text{(S5)}$$

Effect of the β parameter on R_{XA}

It is sufficient to consider the partial derivative of R_{XA} with respect to $1/\beta$, as given by Equation 19 of the main text. We have:

$$\frac{\partial R_{XA}}{\partial \beta^{-1}} \propto \left(1 - \alpha\, e^{-\beta^{-1}T}\right) k^{-1} e^{-k^{-1}\beta^{-1}T} \; - \left(1 - \alpha\, e^{-k^{-1}\beta^{-1}T}\right) e^{-\beta^{-1}T}$$

$$= k^{-1} e^{-k^{-1}\beta^{-1}T} - e^{-\beta^{-1}T} \; + \alpha(1 - k^{-1}) e^{-(1+k^{-1})\beta^{-1}T} \qquad \text{(S6)}$$

As T tends to 0, this expression approaches $(1 - \alpha)(k^{-1} - 1)$. Provided that $k < 1$, this is positive if $\alpha < 1$ (a population expansion), and negative if $\alpha > 1$ (a population contraction). The product of $\exp(\beta T)$ and the derivative is equal to:

$$k^{-1} e^{-(k^{-1}-1)\beta^{-1}T} - 1 \; + \alpha(1 - k^{-1}) e^{-k^{-1}\beta^{-1}T}$$

For $T >> k\beta$ and $k < 1$, this quantity approaches -1, which implies that the derivative must be negative; a negative value will be reached at smaller values of T, the larger α. In the case of a population expansion, this corresponds to a smaller increase in population size. The effect of β on R_{XA} for a given value of k is therefore dependent on the time since the start of the expansion.

Section S2

Table S1 Results for a single sweep of an autosomal locus in a randomly mating population with $N = 10^6$

$\gamma = 125$

$h = 0.1$, $T_d = 0.169$, approx. $T_s = 0.107$, $P_{nc} = 0.284$

r/s	P_c	P_{nr}	P_r	P_{rs}	T_s	T_c/P_c	T_r
0.04	0.427	0.354	0.472	0.324	0.0607	0.102	0.0902
0.08	0.280	0.125	0.685	0.304	0.0311	0.0898	0.0821
0.16	0.154	0.0157	0.842	0.137	0.0112	0.0682	0.0698
0.32	0.0849	0.0002	0.915	0.0154	0.0038	0.0447	0.0566
0.64	0.0534	0.0000	0.945	0.0001	0.0017	0.0302	0.0456
1.28	0.0408	0.0000	0.960	0.0000	0.0009	0.0022	0.0369

$h = 0.5$, $T_d = 0.154$, approx. $T_s = 0.0978$, $P_{nc} = 0.118$

r/s	P_c	P_{nr}	P_r	P_{rs}	T_s	T_c/P_c	T_r
0.04	0.558	0.463	0.388	0.292	0.0607	0.113	0.0956
0.08	0.373	0.214	0.602	0.331	0.0428	0.104	0.0906
0.16	0.198	0.0457	0.797	0.216	0.0177	0.0852	0.0820
0.32	0.101	0.0021	0.899	0.0482	0.0568	0.0559	0.0698
0.64	0.0655	0.0000	0.935	0.0014	0.0024	0.0368	0.0570
1.28	0.0277	0.0000	0.953	0.0000	0.0014	0.0277	0.0454

$h = 0.9$, $T_d = 0.169$, approx. $T_s = 0.0617$, $P_{nc} = 0.109$

r/s	P_c	P_{nr}	P_r	P_{rs}	T_s	T_c/P_c	T_r
0.04	0.586	0.523	0.358	0.274	0.0874	0.133	0.112
0.08	0.401	0.274	0.569	0.326	0.0547	0.124	0.107
0.16	0.108	0.0748	0.775	0.234	0.0237	0.103	0.0962
0.32	0.101	0.0056	0.892	0.0629	0.0568	0.0073	0.0795
0.64	0.0664	0.0000	0.935	0.0025	0.0028	0.0423	0.0593
1.28	0.0439	0.0000	0.959	0.0000	0.0013	0.0299	0.0406

$\gamma = 500$

$h = 0.1$, $T_d = 0.0695$, approx. $T_s = 0.0441$, $P_{nc} = 0.192$

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
0.04	0.274	0.136	0.701	0.338	0.0132	0.0417	0.0314
0.08	0.123	0.0188	0.874	0.175	0.0044	0.0334	0.0267
0.16	0.0496	0.0004	0.951	0.0262	0.0011	0.0216	0.0216
0.32	0.0257	0.0000	0.975	0.0004	0.0003	0.0130	0.0175
0.64	0.0175	0.0000	0.983	0.0000	0.0002	0.0090	0.0145
1.28	0.0136	0.0000	0.987	0.0000	0.0001	0.0069	0.0123

$h = 0.5$, $T_d = 0.0497$, approx. $T_s = 0.0316$, $P_{nc} = 0.129$

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
0.04	0.433	0.370	0.519	0.368	0.0200	0.0407	0.0319
0.08	0.230	0.137	0.753	0.361	0.0096	0.0380	0.0300
0.16	0.0848	0.0188	0.913	0.177	0.0027	0.0306	0.0270
0.32	0.0335	0.0004	0.969	0.0234	0.0006	0.0178	0.0232
0.64	0.0223	0.0000	0.978	0.0002	0.0003	0.0117	0.0197
1.28	0.0178	0.0000	0.983	0.0000	0.0002	0.0093	0.0166

$h = 0.9$, $T_d = 0.0695$, approx. $T_s = 0.0441$, $P_{nc} = 0.124$

r/s	P_{c1}	P_{nr}	P_r	P_{rs}	T_s	T_c / P_{c1}	T_r
0.04	0.484	0.453	0.460	0.336	0.0335	0.0612	0.0525
0.08	0.281	0.205	0.694	0.361	0.0180	0.0579	0.0504
0.16	0.116	0.0421	0.879	0.212	0.0059	0.0479	0.0463
0.32	0.0469	0.0018	0.951	0.0389	0.0015	0.0289	0.0399
0.64	0.0340	0.0000	0.968	0.0008	0.0007	0.0198	0.0321
1.28	0.0252	0.0000	0.979	0.0000	0.0004	0.0154	0.0242

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 a 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.

Section S3

Figure S1 Reductions in diversity (relative to the neutral value) at the end of a sweep for an X-linked locus in a randomly mating population for the *Drosophila* model (no crossing over in males), as a function of the ratio of the autosomal value of the rate of crossing over to the homozygous selection coefficient (r/s), for three different dominance coefficients (h). Gene conversion is absent. The upper panel is for the case when N_e for the X chromosome is half of that for the autosomes; in the lower panel, N_e is the same for both X and A. The red and blue colors denote the *C1* and *C2* predictions, respectively. The full bars denote results for mutations with equal effects in both sexes, with γ = 250. The stippled bars denote male-limited mutations, and the hatched bars denote female-limited mutations. For the sex-limited cases, γ = 500 to ensure comparability with the autosomal mutations. A population size of 10^6 is assumed.

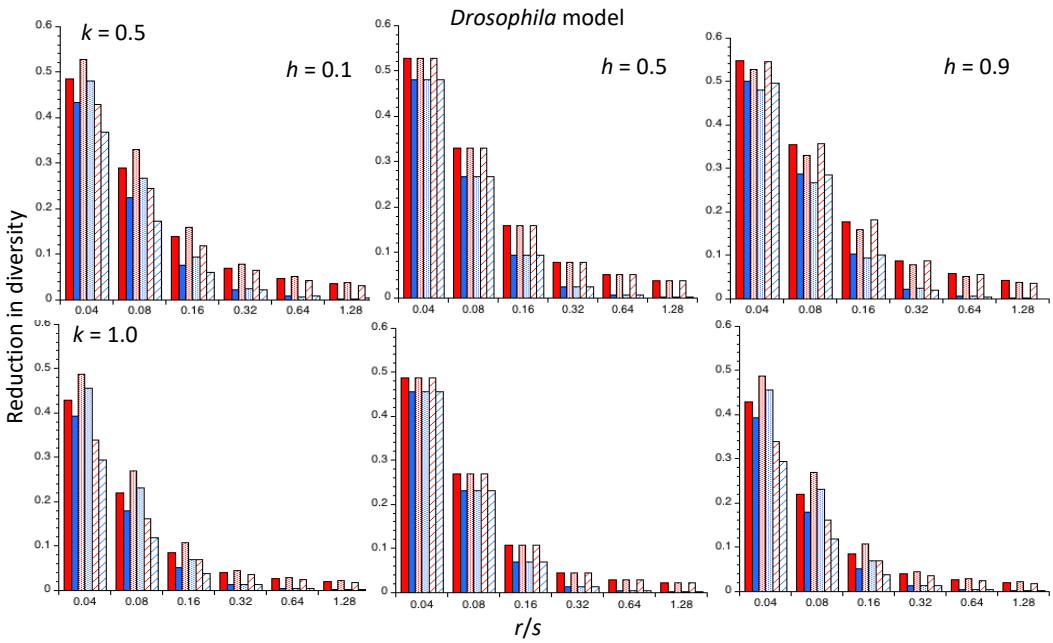


Figure S2 Reductions in diversity (relative to the neutral value) at the end of a sweep for an X-linked locus in a randomly mating population for the mammalian model (crossing over on the autosomes in both sexes), as a function of the ratio of the autosomal value of the rate of crossing over to the homozygous selection coefficient (r/s), for three different dominance coefficients (h). Gene conversion is absent. The upper panel is for the case when N_e for the X chromosome is half of that for the autosomes; in the lower panel, N_e is the same for both X and A. The red and blue colors denote the *C1* and *C2* predictions, respectively. The full bars denote results for mutations with equal effects in both sexes, with $\gamma = 250$. The stippled bars denote male-limited mutations, and the hatched bars denote female-limited mutations. For the sex-limited cases, $\gamma = 500$ to ensure comparability with the autosomal and non-sex-limited X-linked mutations. A population size of 10^6 is assumed.

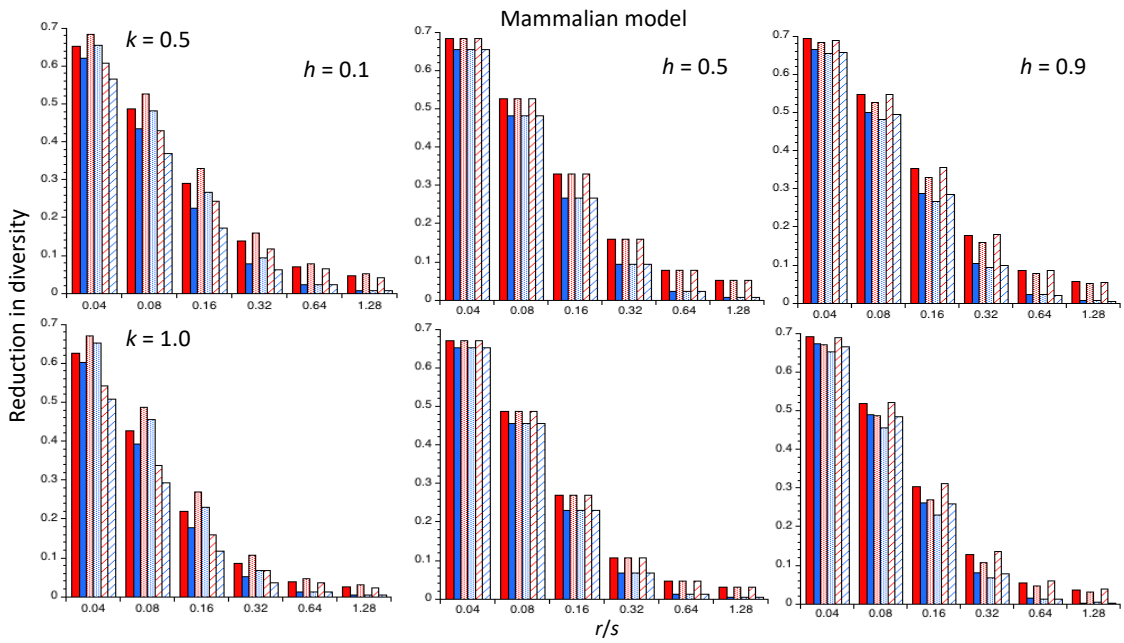


Figure S3 Reductions in diversity (relative to neutrality) under recurrent sweeps at autosomal and X-linked loci for the *Drosophila* model, using the *C2* theoretical predictions with no gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over were chosen to give the same sex-averaged effective recombination rates as for autosomes). The upper panel is for cases without BGS; the lower panel is for cases with BGS (using 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 black bars are for male-limited mutations 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 autosomal and non-sex-limited X-linked mutations with $\gamma = 250$. A population size of 10^6 is assumed.

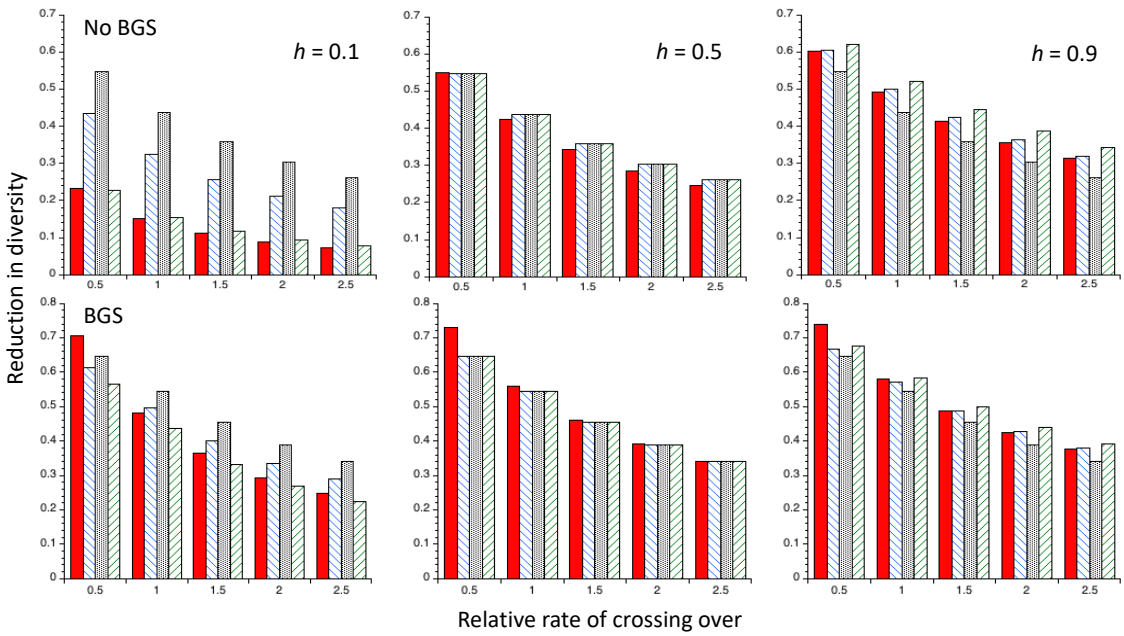


Figure S4 The ratios of X chromosome to autosome nucleotide site diversities under recurrent sweeps for the *Drosophila* model, using the *C2* theoretical predictions with no gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over were chosen to give the same sex-averaged effective rates as for the autosomes). 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 dominance coefficient of favorable mutations), the hatched blue bars are for h = 0.5 and the black stippled bars are for h = 0.9. The other details are as for Figure S3.

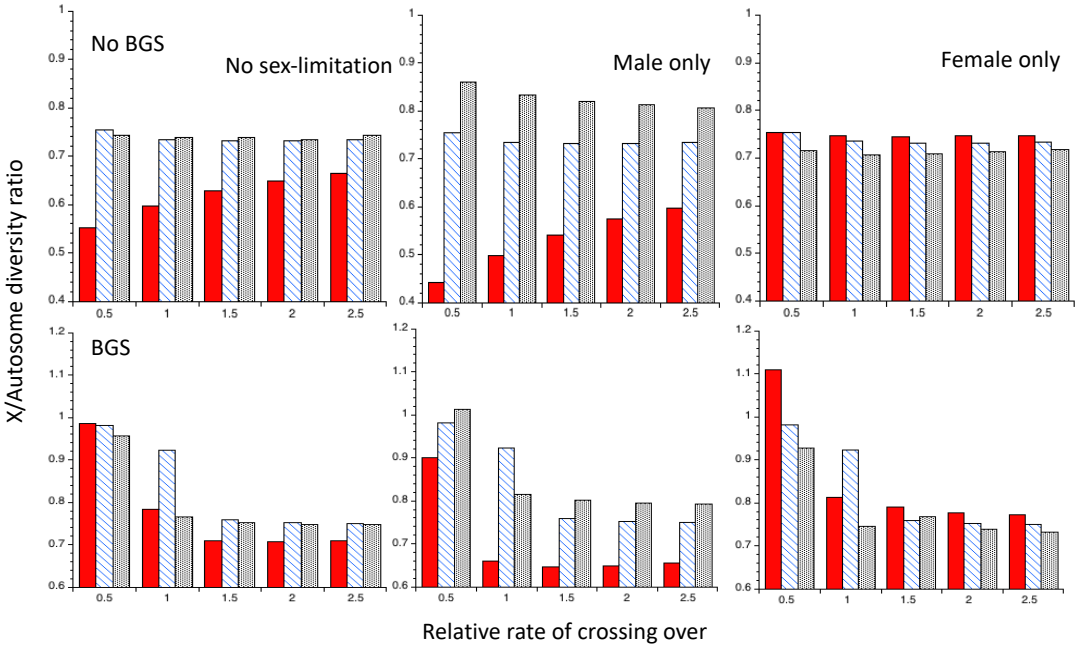


Figure S5 Reductions in diversity (relative to neutrality) under recurrent sweeps at autosomal and X-linked loci for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as for the autosomes). BGS is assumed to be present, with the parameters described in the main text. The upper panel is for the case when N_e for the X chromosome is half of that for the autosomes; in the lower panel, N_e is the same for both X and A. The filled red bars are for autosomal mutations, the hatched blue bars are for X-linked mutations with no sex-limitation, the stippled black bars are for male-limited mutations, and the hatched green bars are for female-limited mutations. For the sex-limited cases, γ = 500 to ensure comparability with the autosomal and non-sex-limited X-linked mutations. A population size of 10^6 is assumed.

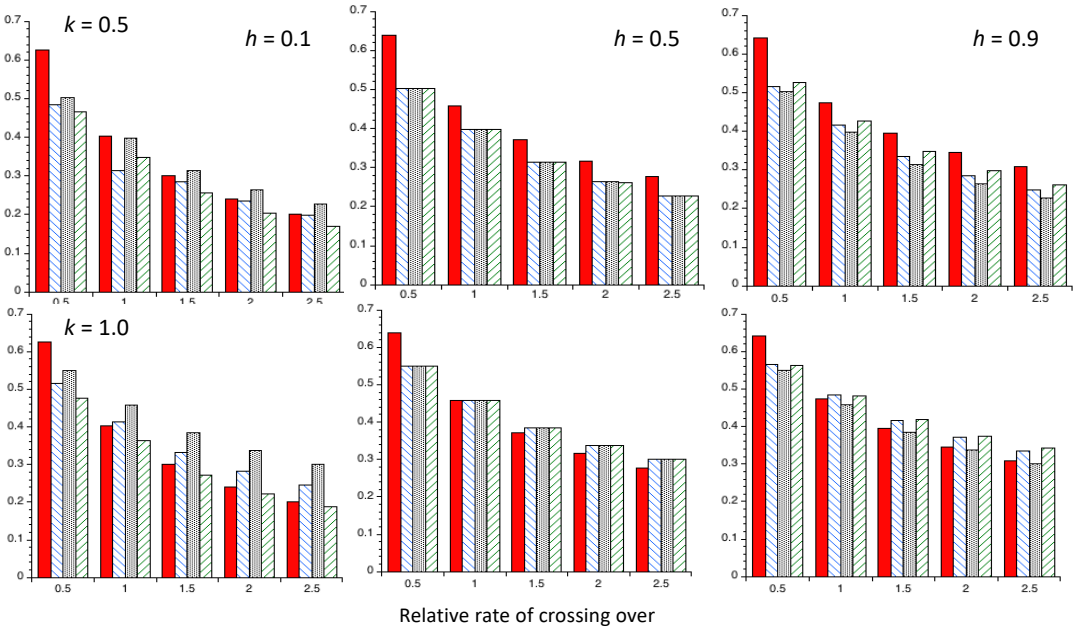


Figure S6 The ratios of X chromosome to autosome nucleotide site diversities for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as the autosomal rates). BGS is assumed to be present. The upper panel is for the case when N_e for the X chromosome is half that for the autosomes; in the lower panel, N_e is the same for both X and A. The filled red bars are for h = 0.1 (the dominance coefficient of favorable mutations), the hatched blue bars are for h = 0.5 and the black stippled bars are for h = 0.9. The other details are as for Figure S5.

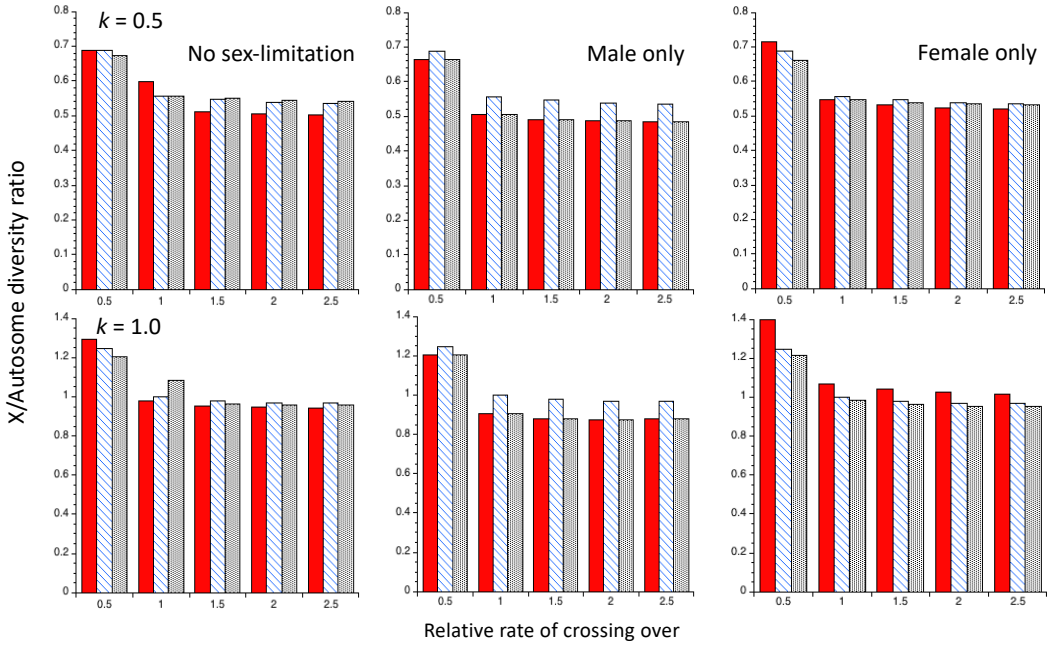


Figure S5 Reductions in diversity (relative to neutrality) under recurrent sweeps at autosomal and X-linked loci for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as for the autosomes). The upper panel is for the case when N_e for the X chromosome is half of that for the autosomes; in the lower panel, N_e is the same for both X and A. The filled red bars are for autosomal mutations, the hatched blue bars are for X-linked mutations with no sex-limitation, the stippled black bars are for male-limited mutations, and the hatched green bars are for female-limited mutations. For the sex-limited cases, γ = 500 to ensure comparability with the autosomal and non-sex-limited X-linked mutations. A population size of 10^6 is assumed.

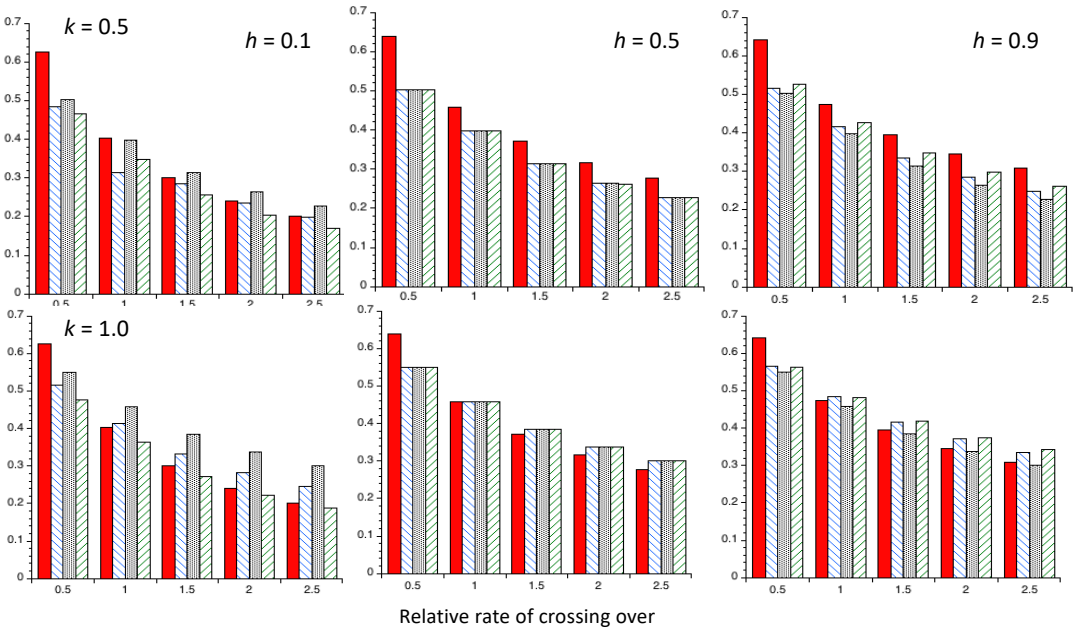
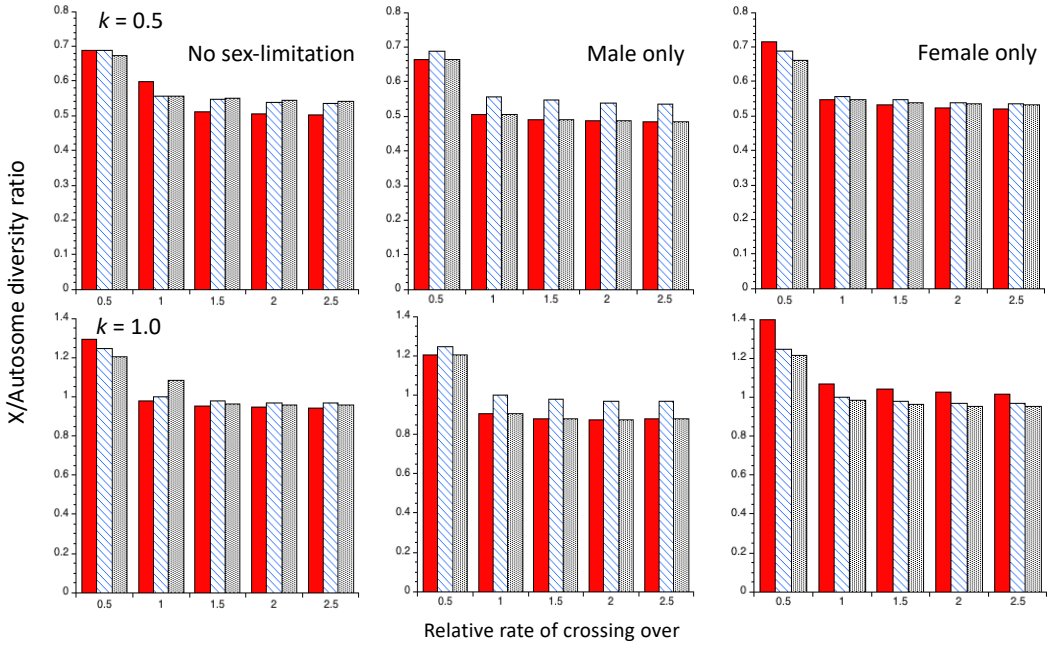



Figure S6 The ratios of X chromosome to autosome nucleotide site diversities for the *Drosophila* model, using the *C2* theoretical predictions with gene conversion and five different rates of crossing over relative to the autosomal standard value (the X-linked rates of crossing over and gene conversion were chosen to give the same sex-averaged effective rates as the autosomal rates). BGS is assumed to be present. The upper panel is for the case when N_e for the X chromosome is half that for the autosomes; in the lower panel, N_e is the same for both X and A. The filled red bars are for h = 0.1 (the dominance coefficient of favorable mutations), the hatched blue bars are for h = 0.5 and the black stippled bars are for h = 0.9. The other details are as for Figure S5.