HHS Public Access
Author manuscript
Annu Rev Ecol Evol Syst. Author manuscript; available in PMC 2023 April 21.

Published in final edited form as:
Annu Rev Ecol Evol Syst. 2021 November ; 52: 177–197. doi:10.1146/annurev-ecolsys-010621-044528.

Effects of Selection at Linked Sites on Patterns of Genetic Variability

Brian Charlesworth[1], **Jeffrey D. Jensen**[2]

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

[2]School of Life Sciences, Arizona State University, Tempe, Arizona 85281, USA

Abstract

Patterns of variation and evolution at a given site in a genome can be strongly influenced by the effects of selection at genetically linked sites. In particular, the recombination rates of genomic regions correlate with their amount of within-population genetic variability, the degree to which the frequency distributions of DNA sequence variants differ from their neutral expectations, and the levels of adaptation of their functional components. We review the major population genetic processes that are thought to lead to these patterns, focusing on their effects on patterns of variability: selective sweeps, background selection, associative overdominance, and Hill–Robertson interference among deleterious mutations. We emphasize the difficulties in distinguishing among the footprints of these processes and disentangling them from the effects of purely demographic factors such as population size changes. We also discuss how interactions between selective and demographic processes can significantly affect patterns of variability within genomes.

Keywords

hitchhiking; genetic recombination; background selection; selective sweeps; Hill–Robertson interference; associative overdominance

INTRODUCTION

A consistent pattern that has emerged from over 30 years of investigations of DNA sequence variability within species is that the level of selectively neutral or nearly neutral variability in a given region of the genome is usually positively correlated with its rate of genetic recombination (Begun & Aquadro 1992, Charlesworth & Campos 2014, Cutter & Payseur 2013), as illustrated in Figure 1. Regions of genomes that completely lack crossing over —sections of chromosomes near centromeres or telomeres and all or parts of Y and W chromosomes—can have levels of variability as low as one-tenth, or even less, of the

Brian.Charlesworth@ed.ac.uk .

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genome-wide average (B. Charlesworth et al. 2010). Such regions also often show evidence of reduced levels of adaptation, including low codon usage bias and increased levels of nonsynonymous nucleotide substitutions between related species. Similar patterns are also found in asexual and highly self-fertilizing species, in which recombination is either absent or ineffective (Bast et al. 2018, D. Charlesworth 2003, Cutter & Payseur 2013). In such cases, it is difficult to distinguish the consequences of reduced rates of recombination from the effects of confounding ecological factors, such as a liability to local extinction and recolonization events (D. Charlesworth 2003). We therefore focus on patterns within the genome of a single sexual, outcrossing species.

The generality of these observations implies that ubiquitous population genetic forces must be involved. Much effort has been expended in characterizing these forces and attempting to assess their relative roles. A major factor is hitchhiking, whereby selection at a given site in the genome influences the frequencies of variants at genetically linked sites. There are two main types of hitchhiking, selective sweeps (SSWs) (Berry et al. 1991, Maynard Smith & Haigh 1974) and background selection (BGS) (B. Charlesworth et al. 1993). A SSW involves the spread of a new, selectively favorable mutation. In the absence of recombination, any variants present on the chromosome in which the mutation arose will be carried along with it, and its fixation in the population wipes out variability at nearby sites. Recombination allows variants linked to the mutation to escape onto the wild-type background, reducing the loss of variability. BGS involves the reverse situation, in which new deleterious mutations that enter a population are eliminated by selection, together with any linked neutral variants. In the absence of recombination, the effective size of the population is multiplied by a factor approximately equal to the proportion of the population that carries one or more deleterious mutations, provided that selection is sufficiently strong that deleterious mutations have a high probability of becoming extinct.

There are two complementary ways of describing the general effects of selection on the variation and evolution at linked sites. First, selection at one locus causes heritable variation in fitness at other loci as a result of linkage disequilibrium (LD), reducing the apparent effective population size (N_e) that they experience, because of the well-known effect of variation in fitness in reducing N_e (B. Charlesworth 2009; Hill & Robertson 1966; Santiago & Caballero 1995, 1998). This reduces variability at these loci and reduces the fixation probabilities of new advantageous mutations, as both of these decrease with N_e (Kimura 1983); conversely, the fixation probabilities of deleterious mutations are increased. This reduction in N_e provides a useful heuristic, despite important qualifications that are described below. In the case of two or more selected loci, the effects on fixation probabilities are known as Hill–Robertson interference (Felsenstein 1974).

The second approach uses the finding that the expected change in the mean of a trait over one generation is proportional to the additive genetic covariance between the trait and fitness —the Price–Robertson equation (Crow & Nagylaki 1976, Price 1970, Robertson 1968). The additive covariance between the allelic state at a focal biallelic locus and fitness at a linked biallelic locus is equal to the product of the coefficient of LD between the two loci (D) and the average excess in fitness at the linked locus, a_w, of a selectively favored allele with frequency p (see the terms in the margins for definitions of the main symbols used here).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

With weak selection, the change in p over one generation is approximately equal to $p(1 - p)a_W$. If x is the allele frequency at the focal locus, its frequency change due to selection at the other locus is $\Delta x \approx Da_W$, and its allelic diversity, $2x(1 - x)$, is changed by approximately $2(1 - 2x)\Delta x$. Hitchhiking thus alters diversity by approximately $2(1 - 2x)Da_W$, an expression used in several analyses of the different forms of hitchhiking (Barton 2000, Santiago & Caballero 1995, Zhao & Charlesworth 2016). This result implies that variability may not necessarily be reduced by hitchhiking; as we describe below, the type of hitchhiking known as associative overdominance (AOD) causes an increase in variability (Zhao & Charlesworth 2016).

This general framework for understanding hitchhiking effects neither distinguishes between the effects of different types of selection nor explicitly predicts important quantities such as levels of variability and fixation probabilities. More detailed analyses are needed in order to make such predictions and to assess which processes are likely to be involved in causing observed patterns of variability. In this article, we focus on BGS, the related process of AOD, and SSWs. In addition, we discuss how demographic factors such as population subdivision and size changes must be considered jointly with the effects of selection in order to make sense of data on patterns of variability across genomes.

Selective processes other than hitchhiking can, however, also affect variability at linked sites. If two or more alleles are maintained for a long time by balancing selection, genetic drift and mutation cause them to diverge at closely linked neutral or nearly neutral sites, in much the same way as local populations diverge when migration is limited (B. Charlesworth et al. 1997, Hudson 1990). A similar effect occurs if alleles at a locus in a spatially subdivided population are maintained at different frequencies by spatially divergent selection acting on different local populations (Barton 1979, B. Charlesworth et al. 1997). Divergence at linked sites in a subdivided population can also arise from hybrid incompatibilities among two or more loci (Bengtsson 1985). These findings have led to the hope that signals of balancing or divergent selection can be identified by genome scans of patterns of variability (Fijarczyk & Babik 2015, Hoban et al. 2016). Considerable progress has been made in both theoretical and empirical studies of these processes, although many problems remain to be solved. For brevity, we do not discuss them further, although they must be considered for a full understanding of natural variability.

THE NEUTRAL NULL MODEL OF VARIABILITY

Before describing the effects of different types of hitchhiking and demographic changes, we briefly review the properties of the simplest null model: a neutral locus in a population of constant effective population size, N_e, whose value is determined by the nature of the genetic system under consideration (e.g., haploidy, mitochondrion, autosome, sex chromosome) together with the breeding system (e.g., discrete generations, overlapping generations, random mating, partial inbreeding) (B. Charlesworth 2009).

Consider a set of n haploid genomes sampled from the population. At a nucleotide site in a genomic region of interest, either the n copies of this region (alleles) include two or more variants (i.e., the site is polymorphic) or the alleles are all identical (i.e., it is monomorphic).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Under the infinite sites model of sequence evolution, variability is assumed to be sufficiently low that sites show at most two variants, one ancestral and one derived (Kimura 1971). We can then describe the state of the sample with respect to a sequence of m nucleotide sites by the variant frequencies at the polymorphic sites: f_i is the proportion of sites where the rarer variant is present in i copies (this is called the folded site frequency spectrum). A frequently used descriptor of the overall level of diversity for a genomic region consisting of m nucleotide sites is the pairwise nucleotide site diversity, π, given by the mean of $2i(n-i)/n$ over all sites, including sites where i is 0. Another way of quantifying variability uses the number of polymorphic sites in the sample, P. This yields the measure called Watterson's theta, defined as $\theta_w = P/(ma_n)$, where a_n is the harmonic series, $1 + 2^{-1} + \cdots + (n-1)^{-1}$. If the population size has been constant for a long time, the expectations of both π and θ_w are equal to $\theta = 4N_e u$, where u is the probability of a neutral mutation per base pair per generation. These fundamental results can be most easily derived using neutral coalescent theory (Figure 2a), as explained in more detail in section 1 of the Supplemental Material (see also Hudson 1990 and Wakeley 2008). Observed values may, of course, deviate considerably from these expectations.

Theory shows that a population expansion, a recent SSW, and BGS all cause an excess of rare variants at neutral sites, compared with the equilibrium site frequency spectrum. This is because they cause the gene trees generated by the coalescent process to have longer external branches (the branches that have not yet experienced a coalescent event, looking back in time) relative to the total sizes of the trees, such that more mutations are represented only once in the sample than under strict neutrality (see Figure 2b), as explained in section 1 of the Supplemental Material. With a population expansion, the reason for this effect is simple; in the recent past, coalescent events take longer to occur because N_e is larger than it is in the more remote past. The effects of SSWs and BGS are discussed below. The expected value of π is then smaller than that of θ_w; conversely, population contractions, AOD, or balancing selection has the reverse effect (Wakeley 2008). Subdivision into partially isolated subpopulations also affects the site frequency spectrum in a way that depends on how the different subpopulations are sampled (Chikhi et al. 2018, Wakeley 2008). This greatly complicates inferences, and we largely evade this difficulty by focusing on organisms, such as *Drosophila*, for which it is reasonable to assume that the population is close to being panmictic.

BACKGROUND SELECTION AND ASSOCIATIVE OVERDOMINANCE

The Abundance of Deleterious Mutations in Natural Populations

Here, we review theory and evidence concerning the hitchhiking effects of purifying selection against deleterious mutations. Analyses of sequence data have shown that deleterious mutations are abundant in natural populations, both in coding sequences (Galtier & Rousselle 2020) and in functional noncoding sequences, such as the untranslated but transcribed regions adjacent to coding sequences, some components of introns, and many intergenic sequences (e.g., Casillas et al. 2007). For these mutations to be effective at causing BGS, they must be sufficiently strongly selected that their frequencies are controlled largely by selection rather than by genetic drift, thus requiring estimates of the strength

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of selection in order to assess the effects of BGS. Population genomic analyses of the distribution of fitness effects (DFEs) of new deleterious mutations for autosomal loci in randomly mating populations show that most of these mutations have small heterozygous selection coefficients (t, where $t = hs$), with a wide distribution about a mean so that $N_e\bar{t} \gg 1$, although the conclusions about the form of their distribution, and the means and variances of t, differ considerably among studies (Campos et al. 2017, Eyre-Walker & Keightley 2009, Galtier & Rousselle 2020, Johri et al. 2020, Kousathanas & Keightley 2013). But there is general agreement that a substantial fraction of deleterious mutations behave nearly deterministically; for example, in *Drosophila melanogaster*, only approximately 10% of new nonsynonymous mutations appear to be in the nearly neutral category ($4N_et < 1$), where variant frequencies are controlled largely by genetic drift (Johri et al. 2020, Kousathanas & Keightley 2013).

The Classical Background Selection Process

These results imply that the elimination of recurrent deleterious mutations by purifying selection may cause hitchhiking effects on linked neutral and weakly selected mutations (B. Charlesworth et al. 1993). Much progress has been made in developing models of these effects. The basic model of BGS assumes deterministic mutation–selection balance at the sites responsible, with sufficiently frequent recombination among them that Hill–Robertson interference can be ignored. For an autosomal locus in a randomly mating population of constant size, this assumption leads to the following prediction for the mean pairwise coalescence time at a focal neutral site surrounded by m sites subject to selection, measured relative to its value in the absence of selection (Hudson & Kaplan 1995, Nordborg et al. 1996). This quantity is commonly denoted by B,

$$B \approx \exp\left(-\sum_{i=1}^{m} \frac{u_i}{[1 + r_i(1 - t_i)/t_i]^2}\right). \qquad 1.$$

Here, u_i is the mutation rate to deleterious alleles at the ith selected nucleotide site, t_i is the selection coefficient against heterozygous carriers of mutations at this site, and r_i is the recombination frequency with the focal site.

If the assumptions of the infinite sites model are met (see above), B corresponds to the expected pairwise nucleotide site diversity (π) relative to its purely neutral expectation (π_0). The formula implies that, all else being equal, π increases as the recombination rate increases. In the absence of recombination, $B \approx f_0$, where $f_0 = \exp\left(-\Sigma_i(u_i/t_i)\right)$ is the equilibrium frequency of the mutation-free or least-loaded haplotype. This is equivalent to assuming that coalescence of a pair of alleles sampled from the population can occur only in a mutation-free background, because haplotypes carrying deleterious mutations are quickly eliminated from the population (D. Charlesworth et al. 1995, Hudson & Kaplan 1994, Nicolaisen & Desai 2013).

Equation 1 has been used in attempts to interpret observed levels of variability in genomic regions with differing local recombination rates (e.g., B. Charlesworth 1996, Comeron 2017,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Elyashiv et al. 2016, Hudson & Kaplan 1995) (Figure 1a), as well as observations of increases in diversity in intergenic sequences at increasing distances from coding sequences (Johri et al. 2020, McVicker et al. 2009, Pouyet et al. 2018). An example of the second type of pattern is shown in Figure 3a. The results strongly suggest that BGS contributes significantly to these observed patterns but is not the sole factor involved (Booker et al. 2017). For example, although 60% of the variance in diversity at putatively neutral noncoding sites in *D. melanogaster* can be explained by BGS (Comeron 2017) and the pattern in Figure 3a is consistent with BGS alone (Johri et al. 2020), BGS cannot explain the negative relationship between a gene's synonymous site diversity and the divergence of its protein sequence from that of a related species, as shown in Figure 3b (Campos et al. 2017).

In addition to its effects on nucleotide site diversity, BGS affects the shape of the gene tree connecting a sample of alleles from a population. It distorts the site frequency spectrum toward a higher frequency of rare variants than is expected under neutrality, because alleles in a sample from the population that carry a deleterious mutation at a particular site cannot coalesce with mutant-free alleles but must wait until the time of origin of the mutation for such coalescent events to be possible. This increases the contribution from the external branches of the gene tree relative to its total size (see section 1 of the Supplemental Material). This effect is stronger when N_et is relatively small, because the mutation can persist in the population for a longer time than it can with large N_et (D. Charlesworth et al. 1995, Hudson & Kaplan 1994, Nicolaisen & Desai 2013). This weakens the effect of BGS on π compared with the classical expression (B is increased), but θ_W is less affected.

Overall, BGS seems to cause only relatively small distortions in the site frequency spectrum in genomic regions with recombination rates typical of most outcrossing species of multicellular organisms, even at synonymous sites in coding sequences (Campos & Charlesworth 2019, B. Charlesworth et al. 1993, D. Charlesworth et al. 1995, Hudson & Kaplan 1994, Zeng 2013). Nevertheless, these distortions can result in substantial errors in estimates of patterns of population change based on purely neutral models, as we discuss below. With large samples of alleles (B. Charlesworth et al. 1993, D. Charlesworth et al. 1995, Hudson & Kaplan 1994, Cvijović et al. 2018), excesses of both low- and high-frequency-derived variants are clearly visible under BGS. The excess of high-frequency-derived variants arises because new neutral mutations sometimes arise on a background with a mean fitness that is higher than average and then hitchhike to a high frequency (B. Charlesworth et al. 1993, Cvijović et al. 2018).

Muller's Ratchet

This model of BGS breaks down when applied to regions of the genome where recombination is rare or absent, and selective interference among the deleterious mutations involved reduces the effective strength of selection. It also fails when $2N_es$ is approximately 5 or less, such that individual mutation frequencies experience significant stochastic fluctuations (B. Charlesworth et al. 1993, D. Charlesworth et al. 1995, Nordborg et al. 1996).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Muller's ratchet (Felsenstein 1974, Muller 1964) is widely assumed to be the paradigm for the first situation, but this is an oversimplification. Models of Muller's ratchet assume unidirectional mutation from wild-type to deleterious alleles, as well as an almost complete absence of recombination; they also usually assume that all mutations have similar fitness effects, although this assumption can be relaxed (Söderberg & Berg 2007). Under these assumptions, if a finite population that lacks recombination is founded from an initial equilibrium population, and the population size is sufficiently large that some least-loaded haplotypes are initially present, haplotypes carrying the current smallest number of deleterious mutations are successively and irreversibly lost from the population—the ratchet (Felsenstein 1974, Jain 2008). Each loss event is accompanied by the fixation of a deleterious mutation (Charlesworth & Charlesworth 1997). However, if reverse mutations can occur, as is the case for single-nucleotide mutations, these ultimately halt the ratchet, resulting in a statistical equilibrium under which the mean number of deleterious mutations per individual remains constant (B. Charlesworth et al. 2010).

There are two circumstances under which the ratchet is likely to accurately describe the behavior of a nonrecombining genomic region: when mutations are irreversible, such as deletions and complex sequence rearrangements, and when a nonrecombining genome, such as a new asexual lineage, has recently been formed, such that the system is far from the final equilibrium. Sequencing of the relatively long-established nonrecombining neo-Y chromosome of *Drosophila miranda*, formed by fusion of an autosome with the ancestral Y chromosome, has revealed that many genes have acquired loss-of-function mutations that are probably irreversible, such as frameshifts and deletions, and it seems likely that the ratchet has contributed to this process (Kaiser & Charlesworth 2010). Its role in other situations is plausible but hypothetical; for example, a role for the ratchet as well as other forms of Hill–Robertson interference in cancer progression has been proposed (McFarland et al. 2013).

The Interference Selection Limit

Theoretical work has shown that, especially with the wide distributions of mutational effects suggested by the population genomic analyses mentioned above, Hill–Robertson interference occurs among the deleterious mutations responsible for BGS, greatly reducing the effective strength of selection against individual mutations and resulting in an interference selection limit (Comeron & Kreitman 2002, Good et al. 2014). It is also necessary to consider the reverse mutations that occur with base substitutions (which constitute the majority of mutations) in order to understand the long-term properties of a low-recombination genomic region. Simulations with reverse mutations show that, at statistical equilibrium under mutation, drift, and selection, the effect of BGS on nucleotide site diversity at neutral sites is greatly reduced in a low-recombination region and is accompanied by a large increase in the abundance of low-frequency neutral variants and in the frequencies of deleterious variants (B. Charlesworth et al. 2010, Comeron & Kreitman 2002, Kaiser & Charlesworth 2009). This explains why low-recombination genomic regions, such as the neo-Y chromosome of *D. miranda*, often have much greater variability than is predicted by Equation 1 (B. Charlesworth et al. 2010, Kaiser & Charlesworth 2009).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Associative Overdominance

Another consequence of linkage and selection is associative overdominance (AOD). In its original formulation, AOD involves apparent heterozygote advantage at biallelic neutral loci caused by associations either with a linked selected locus that is subject to a genuine heterozygote advantage or with linked loci subject to recurrent mutations to recessive or partially recessive ($h < 0.5$) deleterious alleles. Here, we discuss the effects of deleterious mutations in randomly mating populations, where genetic drift can cause nonrandom associations between alleles at selected loci and alleles at closely linked neutral loci. This causes homozygosity for alleles at the neutral loci to be correlated with homozygosity at the selected loci, creating apparent heterozygote advantage at the neutral loci (Ohta 1971). In contrast, AOD in partially inbreeding populations is caused mainly by associations between the frequencies of heterozygotes at different loci caused by the different levels of homozygosity of outbred versus inbred individuals and does not involve drift (D. Charlesworth 1991).

This apparent heterozygote advantage was originally thought to retard the loss of neutral variability by genetic drift (Ohta 1971), the opposite effect of BGS. This raises the question of how deleterious mutations can apparently either retard or enhance the loss of variability, depending on the parameter values, as has been found in computer simulations (Gilbert et al. 2020, Latter 1998, Palsson & Pamilo 1999, Zhao & Charlesworth 2016). This paradox is partially resolved by the fact that substitution of the apparent selection coefficients against homozygotes into the standard equation for allele frequency change shows that they do not affect allele frequencies at the neutral locus and hence cannot influence its diversity (Zhao & Charlesworth 2016). Nevertheless, according to the Price–Robertson equation, diversity at the neutral locus must be changed by the hitchhiking effects of the selected loci, as described in the Introduction.

A detailed analysis of the case of a single selected locus and a linked neutral locus shows that neutral variability is increased only when the deleterious mutations are partially recessive ($h < 0.5$) and the scaled selection coefficient $2N_e s$ is ≤ 2, such that genetic drift causes a substantial variance in the frequencies of deleterious mutations around their deterministic equilibrium values, especially when h is small. Otherwise, BGS operates and variability is reduced, not enhanced (Zhao & Charlesworth 2016). The intuitive basis for this result comes from the seemingly paradoxical fact that weakly selected recessive or partially recessive deleterious mutations have slightly longer mean sojourn times between their origination and their fixation or loss than do neutral mutations (Mafessoni & Lachmann 2015), whereas more strongly selected mutations have short sojourn times (see section 2 of the Supplemental Material for more details). When selection is sufficiently weak and mutations are sufficiently recessive, a linked neutral locus can therefore experience coalescence times that are longer than expected under neutrality. With multiple selected loci, simulations show that the conditions for AOD are somewhat relaxed compared with the analytical predictions for a single selected locus and confirm that a transition from BGS to AOD occurs as h and population size are reduced for a given strength of selection (Gilbert et al. 2020, Palsson & Pamilo 1999, Zhao & Charlesworth 2016).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

The requirement for a small value of $2N_es$ suggests at first sight that AOD is likely to operate only in small populations, such as laboratory populations or closed populations of domestic animals. Evidence for its action in such populations is indeed seen in their rates of loss of variability at molecular marker loci (Latter 1998, Zhao & Charlesworth 2016). However, even in large natural populations the small N_e of low-recombination regions of the genome, and the accompanying interference among loci under selection, could reduce $2N_es$ below the threshold needed for AOD, at least for weakly selected deleterious mutations. Given the evidence from the parameters of the DFE mentioned above, most sites in low-recombination regions are probably subject to BGS rather than AOD, such that the mean diversity at neutral sites is expected to be much lower than the genome-wide average. However, some sites will fall into the region where AOD operates. This indeed occurs in computer simulations, whose predictions compare well with population genomic data (Becher et al. 2020). The excess of low-frequency synonymous variants in low-recombination regions of several *Drosophila* species is much less than expected under BGS alone (Becher et al. 2020). In humans, there is a decrease in the skew of the site frequency spectrum toward low-frequency variants in regions with low recombination, compared with regions with slightly more recombination, whereas the skew decreases with increasing recombination rate elsewhere (Gilbert et al. 2020, Pouyet et al. 2018). Both of these patterns indicate the action of AOD.

SELECTIVE SWEEPS AND THEIR FOOTPRINTS ACROSS THE GENOME

Effects of a Single Selective Sweep

Although beneficial mutations are a comparatively small fraction of all new mutations (Bank et al. 2014), they are obviously important in evolution. There is great general interest in studying adaptation, as well as in identifying loci underlying desirable traits in domesticated and cultivated species or undesirable traits, such as drug resistance in viral and bacterial populations (e.g., Xia et al. 2009, Foll et al. 2014). Maynard Smith & Haigh (1974) significantly enhanced our understanding of the genomic consequences of such positive selection by analyzing the reduction in variability at a linked site as a new beneficial mutation spreads toward fixation in a population, a process commonly referred to as a selective sweep (SSW) (Berry et al. 1991).

If selection favoring a beneficial mutation is sufficiently strong, it reaches fixation much faster than under neutrality, greatly reducing variability in the genomic region surrounding the target of selection, whose physical size is inversely related to the local rate of recombination. After the completion of a sweep, neutral variability gradually recovers as new mutations enter the population over a period of the order of $2N_e$ generations (Wiehe & Stephan 1993), although the time during which the genomic patterns associated with a sweep are statistically identifiable is generally much shorter (Przeworski 2002).

The set of events that occurs during a sweep can be understood with the help of the coalescent process, as illustrated in Figure 2b for the case of a sample of four alleles. In a sample of n alleles taken after a beneficial mutation has become fixed, all alleles will carry the mutation. If we trace the ancestry of these alleles back in time, we can see that a given linked neutral site in one or more of them can be derived by recombination from a wild-type

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

allele that was still present in the population and whose ancestor was present before the sweep occurred. In contrast, all nonrecombinant alleles are derived from the single ancestral haplotype in which the beneficial mutation arose, and their expected time to coalescence is much shorter than in the absence of selection, causing a reduction in the expected value of π at the neutral site relative to π_0, the purely neutral value. Prior to the origin of the beneficial mutation, coalescence with respect to the neutral site for any extant alleles follows the standard neutral model.

These considerations suggest that the overall reduction in diversity caused by a sweep will be inversely related to the product of the time taken for the beneficial mutation to become fixed (T_f) and the recombination rate (r) between the selected and neutral sites, as this product determines the probability of occurrence of one or more recombination events during the sweep. For beneficial mutations that are sufficiently strongly selected that their trajectories of allele frequency change are close to those in an infinite population, T_f is inversely related to the selection coefficient s_a that measures the increase in fitness of heterozygotes for the mutation relative to wild-type homozygotes (B. Charlesworth 2020b, Haldane 1924). The reduction in diversity caused by a sweep, $\Delta = 1 - \pi/\pi_0$, should thus be inversely related to r/s_a. Furthermore, if T_f is sufficiently small, and there is only a short time T_s between the end of the sweep and the time of sampling, nonrecombinant alleles can coalesce only at the start of the sweep, creating a star-shaped genealogy, in which any new mutations are present only once in the sample (Figure 2b). The site frequency spectrum is thus expected to be skewed toward low-frequency variants compared with the standard neutral model (Barton 2000, Braverman et al. 1995).

Multiple extensions have been made to the theory of SSWs since the work of Maynard Smith & Haigh (1974) (for a recent overview, see Stephan 2019). In particular, Kaplan et al. (1989) used coalescent theory to develop a stochastic treatment that includes the initial phase when the beneficial mutation is rare and vulnerable to loss from the population. On the assumption that all nonrecombinant alleles coalesce at the start of the sweep, and that at most one recombination event occurs during the sweep, Barton (2000) derived a simple approximation for how a semidominant autosomal mutation affects the expected coalescence time for a pair of alleles sampled immediately after a sweep,

$$\Delta = 1 - \frac{\pi}{\pi_0} \approx (2N_e s_a)^{-4r/s_a}. \qquad 2.$$

This formula has been used in several methods for detecting sweeps, as has a different approximation due to Stephan et al. (1992).

The advent of high-throughput sequencing technologies has focused interest on characterizing patterns of variation associated with recent sweeps, potentially allowing them to be detected in population genomic data. As well as the reduced local variability and excess of low-frequency variants just mentioned, these patterns include high frequencies of derived variants around a recent sweep in a recombining genome region (Fay & Wu 2000) and an excess of LD during a sweep and a break in LD across the target of selection at the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

time of fixation of a beneficial mutation (Jensen et al. 2007, Kim & Nielsen 2004, McVean 2007, Stephan et al. 2006).

These various expected signatures have been utilized in several tests for recent sweeps (reviewed by Bank et al. 2014, Booker et al. 2017, Stephan 2019). For example, Kim & Stephan (2002) developed a composite-likelihood ratio test, comparing the probability that the site frequency spectrum in a given genomic region was drawn from a neutral model versus a SSW model. Jensen et al. (2005) extended this approach, showing that violations of demographic equilibrium (i.e., population size changes) could result in high false-positive rates, and proposed an additional goodness-of-fit test to evaluate the fit of a SSW model. Nielsen et al. (2005) extended the composite-likelihood framework to use a null model, rather than the standard neutral model, based on the observed frequency spectrum across the genome as a whole; DeGiorgio et al. (2016) incorporated mutation rate variation, and information from LD patterns was added by Pavlidis et al. (2013).

These methods all assume a hard sweep (Figure 4), in which positive selection has acted on a new or rare variant and has driven it to fixation in the population, resulting in the fixation of a single haplotype in a region where recombination has failed to occur around the target of selection; methods for detecting ongoing or incomplete sweeps have also been developed (e.g., Ferrer-Admetlla et al. 2014). Two other alternatives to the hard sweep model must also be considered, which, although different from each other, are often collectively called soft sweeps because of their similar expected patterns of variation. In a soft SSW, multiple nonrecombinant haplotypes can be present after the initial wild-type state has been replaced (Figure 4). This can happen in two ways. First, positive selection can act on a variant that was initially present on multiple genetic backgrounds in the initial population, for example, a neutral or weakly deleterious variant that was segregating when a shift in selective pressure caused it to become beneficial (Hermisson & Pennings 2005). Second, either there is a high rate of input of beneficial mutations at a given locus or the locus represents a large mutational target. In this case, multiple mutations with identical fitness effects could arise on different genetic backgrounds, each remaining in the population after the wild-type state has been replaced (Pennings & Hermisson 2006).

A soft sweep can generate LD across the target of selection at the time of fixation (rather than only on opposite sides of the target) and produces a pattern of intermediate- and high-frequency variants close to the target of selection (rather than only in flanking regions) (Figure 4). This creates the problem that a region flanking a hard sweep (with strong LD and high-frequency-derived alleles) could be misinterpreted as a soft SSW (Schrider et al. 2015), as could gene conversion events that move the beneficial allele onto ancestral haplotypes. Furthermore, the soft sweep outcome requires an unusual situation to occur. For selection on standing variation to result in the continued presence of multiple haplotypes, the initial frequency of the advantageous allele must be considerable; if the hypothesized standing variants are rare, stochastic loss of all but one of the haplotypes carrying the beneficial variant is likely (Orr & Betancourt 2001). Generalizing this argument, Jensen (2014) argued that hard sweeps are the most likely outcome across much of the biologically relevant parameter space; for counterarguments, see Hermisson & Pennings (2017).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Several statistical tests aim to detect hitchhiking patterns associated with soft sweeps. For example, Garud et al. (2015) proposed a haplotype-based statistic based on the expectation that hard sweeps result in the fixation of a single haplotype close to the target of selection, whereas soft sweeps result in multiple common haplotypes. Schrider & Kern (2016) developed a machine learning–based approach using multiple summary statistics that utilizes sliding windows across a genome to classify patterns of variation in each region as being the result of a hard sweep (with no recombination), linkage to a hard sweep, a soft sweep, linkage to a soft sweep, or neutrality.

Finally, SSW-like effects are not limited to beneficial mutations. Conditional on its fixation, the fixation time of a semidominant deleterious mutation is the same as that of an advantageous one with the same selection coefficient (Maruyama & Kimura 1974). The fixation of a deleterious mutation by genetic drift can thus generate the same sweep effect as the fixation by selection of a beneficial mutation (Johri et al. 2021a). A recent fixation of a neutral mutation is also associated with reduced variability at closely linked sites (Tajima 1990). Although the probabilities of fixation differ greatly between beneficial, neutral, and deleterious mutations, the far higher rates of occurrence of the last two suggest that recent fixations of such mutations could significantly affect the levels and patterns of variation at nearby sites when scanning genomes for fixation events (as was done by Sattah et al. 2011 and Elyashiv et al. 2016). However, these effects are likely to be highly localized to a few dozen base pairs except in low-recombination regions, as weak selection is required (Mafessoni & Lachmann 2015).

Recurrent Selective Sweeps

All the approaches described above suffer from the difficulty that many selective events may have occurred so far back in time that they leave little trace in the population statistics that we have just discussed. Given the evidence from population genomic studies that a substantial fraction of nucleotide differences between related species in functionally important regions of the genome reflect past positive selection (Booker et al. 2017), it is important to examine the expected effects of recurrent SSWs distributed across the genome.

The most commonly used underlying theory was developed by Kaplan et al. (1989), Wiehe & Stephan (1993), and Kim & Stephan (2000). It assumes that the reduction in diversity (relative to the neutral value) caused by a single sweep, Δ, is equal to the probability of coalescence of a pair of nonrecombinant alleles, as is assumed in Equation 2. If SSWs at a given selected site j occur at rate ω_j, causing an expected reduction in diversity Δ_j at a focal neutral site, then the rate of coalescence at the focal site caused by sweeps is equal to $\Sigma_j \omega_j \Delta_j$, ignoring any selective interference among the different beneficial mutations involved. With a BGS effect of B, the rate of coalescence due to drift is $1/(2BN_e)$; under this competing coalescent model, the net rate of coalescence $\approx (2BN_e)^{-1} + \Sigma_j \omega_j \Delta_j$ (Kim & Stephan 2000). The mean pairwise coalescence time is the reciprocal of this expression, which yields the following expected value of π at the focal site relative to the neutral value,

$$\frac{\pi}{\pi_0} \approx \frac{1}{(2B)^{-1} + 2N_e \Sigma_j \omega_j \Delta_j}. \qquad 3.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Equation 3 has been used in several attempts to estimate the rate of occurrence of SSWs and the strength of selection on beneficial mutations. These attempts exploit patterns such as the relationship between π and recombination rate (Wiehe & Stephan 1993), the relationship between π and recent nonsynonymous substitutions at nearby sites (Sattah et al. 2011), and the negative relationship between the value of synonymous site π for a gene and its level of protein sequence divergence from a related species, as seen in Figure 3b (Campos et al. 2017), as well as a composite-likelihood method that uses a combination of several such statistics (Elyashiv et al. 2016). The negative relationship between the proportion of rare variants in a sample and the level of recombination over the lower part of the range of recombination rates, as seen in Figure 1b, is suggestive of the effects of recurrent SSWs, as is the larger proportion of rare variants on the X chromosome compared with the autosomes, although a full quantitative analysis of these patterns is lacking (Campos et al. 2014).

These methods agree in suggesting that diversity at putatively neutral sites close to functional sites is significantly affected by recurrent substitutions. However, there are disagreements about the frequency at which sweeps occur and the strength of selection—indeed, the assumption that these patterns involve only the fixation of beneficial mutations as opposed to deleterious or neutral mutations is not entirely correct. In addition, both the competing coalescent model and the expressions for Δ described above are only approximate. Their use could lead to errors in estimates of sweep parameters, especially if the assumption of semidominance used in Equation 2 is relaxed (Campos & Charlesworth 2019, B. Charlesworth 2020a, Hartfield & Bataillon 2020). More work is needed before firm conclusions can be drawn.

THE INTERPLAY BETWEEN DEMOGRAPHIC CHANGES AND SELECTION AT LINKED SITES

These theoretical expectations for the effects of BGS and SSWs can be used to investigate the power and false-positive rates of the statistical tests based on them, particularly under demographic models relevant to natural populations, given that population size changes can produce genomic signatures similar to those of selection at linked sites. For example, the coalescent trees generated by a population size bottleneck, followed by rapid expansion, resemble those caused by a SSW, as most of the alleles in a sample coalesce at the time of the bottleneck (Barton 2000). Similarly, as already noted, both BGS and recent population growth generate a site frequency spectrum with an excess of low-frequency variants (B. Charlesworth et al. 1993, D. Charlesworth et al. 1995, Ewing & Jensen 2016, Nicolaisen & Desai 2013, Zeng 2013), especially in low-recombination genomic regions under the interference selection regime (Becher et al. 2020, B. Charlesworth et al. 2010, Kaiser & Charlesworth 2009).

In order to solve this problem, selective effects are often assumed to be locus specific, whereas demographic effects are genome wide. Population bottleneck⁄expansion patterns in individual genomic regions are then attributed to sweeps, whereas a similar genome-wide pattern is used to infer population history (Galtier et al. 2000). However, this is problematic because, under neutrality, population bottlenecks inflate the variance of commonly used

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

summary statistics, such that many more outlier regions are generated than in a stationary population and can be mistaken for localized sweeps (Teshima et al. 2006, Thornton & Jensen 2007). Furthermore, any outlier detection approach relies on several specific, and often unsubstantiated, assumptions: (*a*) that selected regions reside in the tails of the distributions across the genome of test statistics (which is not always true, because this depends on the type and strength of selection as well as on the population's demographic history), (*b*) that recent sweeps have indeed occurred (any model, including neutrality, will have outliers for any given test statistic), and (*c*) that sweeps are rare (if they were common, an outlier approach would fail to recognize the regions as unusual).

Selective and demographic effects may thus be confounded. Multiple studies have sought to quantify true- and false-positive rates for tests for recent sweeps. For statistics aiming to detect hard sweeps, Crisci et al. (2013) evaluated the performance of the commonly used methods of Nielsen et al. (2005) and Pavlidis et al. (2013) (described above) under various nonequilibrium demographic models. The approach by Nielsen et al. (2005) had low false-positive rates under a variety of neutral bottleneck models (generally under 5%) but true-positive rates for identifying sweeps were also low (generally under 10%), implying that their method simply lacks power to distinguish bottlenecks from sweeps. Conversely, the method of Pavlidis et al. (2013) had greater true-positive rates (up to 50% under certain models) at the expense of many false positives (approaching 90% under extreme bottleneck models). Therefore, under some demographic histories (such as recent bottlenecks; Poh et al. 2014), SSWs cannot be distinguished from genome-wide events, although in other situations the same statistics perform well and sweeps are identifiable.

Harris et al. (2018) similarly investigated the performance of tests for soft sweeps, examining the claims of Garud et al. (2015) and Schrider & Kern (2017) for frequent genome-wide soft sweeps in *D. melanogaster* and humans, respectively. Owing to the lack of appropriate null models, the validity of these claims was questioned. The top outlier regions of the *D. melanogaster* genome in Garud et al. (2015) were consistent with soft sweeps but also with hard sweeps or neutrality, and the population's demographic history alone could have generated the empirically observed haplotype structure attributed to genome-wide positive selection effects. Similarly, the approach of Schrider & Kern (2017) had a high false-positive rate for detecting soft sweeps under virtually every model examined, and the number of sweeps detected across human populations was consistent with that expected from the false-discovery rate.

This misinference can act in the other direction—demographic models can be misinferred through a lack of consideration of selective effects. Population history is often estimated using intronic or synonymous sites, which are affected by selection at nearby directly selected exonic sites and may experience direct selection themselves (Machado et al. 2020, Parmley & Hurst 2007). Patterns of variation attributed to a neutral demographic model may therefore be the result of either purifying selection or positive selection (Messer & Petrov 2013, Zeng 2013). Ewing & Jensen (2016) showed that the skew toward low-frequency variants caused by BGS effects can lead to the inference of rapid population growth, even under a constant population size. Johri et al. (2021b) examined the widely used demographic estimators *fastsimcoal2* (Excoffier et al. 2013) and the multiply sequential Markovian

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

coalescent (MSMC) (Li & Durbin 2011, Schiffels & Durbin 2014) and found misinferred demographic histories in the presence of BGS, even after masking functional regions. This misinference was amplified as the strength of purifying selection and the density of directly selected sites increased. Notably, the pattern of past changes in N_e inferred by MSMC frequently had a characteristic shape indicating ancient decline and recent growth (as has been observed in studies of several species), even in a sample from a population at strict neutral equilibrium. These findings imply a need for caution when inferring population size changes if selection could be acting at linked sites; similar concerns have been raised about the effects of population structure (Chikhi et al. 2018).

Conversely, changes in population size can strongly affect the estimated values of the BGS parameter B, even when its true value remains constant, because differences in N_e affect the rate at which variability responds to a population size change (Johri et al. 2021b). In addition, regions of the genome that differ in N_e because of recombination rate differences may exhibit different degrees of skew in their variant frequencies caused by their different rates of response to population size changes rather than by differences in the effects of SSWs or BGS. There is evidence for this effect in *Drosophila*, in which the skew toward low-frequency variants can increase rather than decrease with the recombination rate over some of the range of recombination rates, as shown in Figure 1b (Becher et al. 2020, Campos et al. 2014). This observation seems paradoxical at first sight, as recombination is expected to counteract hitchhiking effects.

These findings indicate the need for an appropriate null model, incorporating the effects of population processes that are certain to be constantly occurring across the genome (e.g., genetic drift under a realistic demographic history, purifying selection and BGS, and mutation and recombination rate variation), in order to accurately quantify the role of processes hypothesized to operate episodically or locally, such as positive selection (Comeron 2017, Johri et al. 2020). Johri et al. (2020) used an approximate Bayesian approach, which used statistics such as π, the site frequency spectrum, LD, and between-species divergence, to obtain the first joint estimate of the DFE together with population history. Their simulations showed that stepwise inferences, in which a demographic model is first estimated and then used to obtain an estimate of the DFE (as in the DFE-alpha method of Eyre-Walker & Keightley 2009), are prone to error when synonymous sites are under selection, because fitting a demographic model that corrects for the associated skew in variant frequencies leads to misinference of the DFE. Specifically, the excess of rare variants leads to populations being incorrectly inferred as growing rapidly. Consistent with this effect, they found less evidence for growth in a Zambian population of *D. melanogaster* than did earlier studies, and a substantial amount of weak purifying selection at functional sites was detected. Although a model with strong, frequent beneficial mutations was rejected, addition of a component of rare, weakly selected beneficial mutations could not be rejected (but did not improve the fit).

Recent efforts have also been made to extend inference methods beyond the standard neutral coalescent process described here (Figure 2), which assumes that N_e is so large, and the variance in the number of successful progeny per individual is so small, that at most one coalescent event can occur per generation (Hudson 1990, Wakeley 2008). While these assumptions are probably appropriate for many commonly studied organisms (including mammals, birds, and *Drosophila*), a wide variety of species, including many plants, marine spawners, and pathogens, exhibit large progeny number distributions better represented by multiple-merger coalescent models, in which more than one coalescent event can occur in a given generation (Irwin et al. 2016, Tellier & Lemaire 2014). This feature alone can radically alter expected levels and patterns of variation under neutrality, and its neglect could lead to serious misinference of both selection and demography. Recent theoretical and computational efforts have identified patterns that may allow population growth and neutral multiple-merger coalescence to be distinguished (Eldon et al. 2015, Matuszewski et al. 2018), and tests for SSWs under these alternative coalescent models have recently been proposed (Sackman et al. 2019). Such model development is feasible (Harris & Jensen 2020), given that SSWs themselves involve a localized multiple-merger coalescent event (Figure 2). Further theoretical studies of these alternative coalescent models are needed for analyses of organisms such as viruses (Irwin et al. 2016).

CLOSING THOUGHTS

The theoretical underpinnings of genetic hitchhiking models, and the development of statistical inference approaches for detecting and quantifying hitchhiking effects in genomic data, have proliferated over the past few decades. These advances have made it increasingly clear that the effects of selection at linked sites need to be considered for a full understanding of the levels and patterns of variation in natural populations. Both genetic drift, as modulated by population history, and purifying selection (with the associated effects of BGS) are pervasive factors determining the fates of new mutations, consistent with the neutral theory of molecular evolution (Kimura 1983). However, there is also strong evidence that positive selection plays an important role in between-species sequence divergence at functional sites. Through its associated SSW effects, positive selection modulates both the amount of DNA sequence variability and the shape of the site frequency spectrum, especially at sites within or close to coding sequences and certain types of noncoding sequences.

It remains a formidable challenge to estimate the individual contributions of these evolutionary processes accurately, but two things seem clear. First, ignoring one process in order to estimate another (e.g., neglecting direct and indirect effects of selection when estimating population history or neglecting population history when estimating the effects of selection) can lead to serious misinference. Second, any analysis of population genomic data must use a proper null model, capturing the roles of mutation and recombination rate variation, purifying selection and BGS in and around functional elements, and using the underlying history of population size change, structure, and migration. Encouragingly, theoretical and computational approaches are emerging that appear to be capable of jointly estimating the parameters of such an evolutionary null model. It may therefore become possible to accurately characterize the expected effects of processes such as hard versus soft sweeps, as well as complete versus incomplete sweeps, on levels and patterns of variability across genomes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Deborah Charlesworth, Parul Johri, Susanne Pfeifer, and Michael Whitlock for helpful comments and discussion. We also thank Parul Johri and José Campos for providing the data for Figure 3a and Figure 3b, respectively. J.D.J. was supported by National Institutes of Health grant R35GM139383-01.

Glossary

$\mathbf{N_e}$	the effective population size, such that the rate of coalescence of a pair of neutral alleles is $1/(2N_e)$
D	the coefficient of linkage disequilibrium, equivalent to the covariance between the allelic states of a pair of biallelic loci
$\mathbf{a_w}$	the average excess, measured by the difference in fitness between the carriers of the alleles at a biallelic locus
$\boldsymbol{\pi}$	the mean value of the pairwise nucleotide site diversity over a set of nucleotide sites
$\boldsymbol{\theta_w}$	Watterson’s estimator of the scaled mutation rate from the number of polymorphic sites
$\boldsymbol{\theta = 4N_e u}$	the scaled neutral mutation rate, where u is the mutation rate per nucleotide site
s	the reduction in fitness to homozygotes for a deleterious mutation, relative to the fitness of wild-type homozygotes
t = hs	the reduction in fitness to heterozygotes for a deleterious mutation; h is the dominance coefficient
B	the ratio of the mean coalescence time at a neutral site under BGS, relative to its value without selection
r	the frequency of recombination between a pair of loci

LITERATURE CITED

Bank C, Foll M, Ferrer-Admetlla A, Ewing G, Jensen JD. 2014. Thinking too positive? Revisiting current methods in population genetic selection inference. Trends Genet 30:540–46 [PubMed: 25438719]

Barton NH. 1979. Gene flow past a cline. Heredity 43:333–39

Barton NH. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. B 355:1553–62

Bast J, Parker DJ, Dumas Z, Jalvingh KM, Van PT, et al. 2018. Consequences of asexuality in natural populations: insights from stick insects. Mol. Biol. Evol 35:1668–77 [PubMed: 29659991]

Becher H, Jackson BC, Charlesworth B. 2020. Patterns of genetic variability in genomic regions with low rates of recombination. Curr. Biol 30:94–100 [PubMed: 31866366]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Begun D, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rate in Drosophila melanogaster. Nature 356:519–20 [PubMed: 1560824]

Bengtsson BO. 1985. The flow of genes through a genetic barrier. In Evolution: Essays in Honour of John Maynard Smith, ed. Greenwood PJ, Harvey PH, Slatkin M, pp. 31–42. Cambridge, UK: Cambridge Univ. Press

Berry AJ, Ajioka JW, Kreitman M. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics 129:1111–17 [PubMed: 1686006]

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. BMC Biol 15:98 [PubMed: 29084517]

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphism. Genetics 140:783–96 [PubMed: 7498754]

Campos JL, Charlesworth B. 2019. The effects on neutral variability of recurrent selective sweeps and background selection. Genetics 212:287–303 [PubMed: 30923166]

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol 31:1010–28 [PubMed: 24489114]

Campos JL, Zhao L, Charlesworth B. 2017. Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. PNAS 114:E4762–71 [PubMed: 28559322]

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol. Biol. Evol 24:2222–34 [PubMed: 17646256]

Charlesworth B 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res 68:131–50 [PubMed: 8940902]

Charlesworth B 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet 10:195–205 [PubMed: 19204717]

Charlesworth B 2020a. How good are predictions of the effects of selective sweeps on levels of neutral diversity? Genetics 216:1217–38 [PubMed: 33106248]

Charlesworth B 2020b. How long does it take to fix a favorable mutation, and why should we care? Am. Nat 195:753–71 [PubMed: 32364783]

Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. 2010. Genetic recombination and molecular evolution. Cold Spring Harb. Symp. Quant. Biol 74:177–86

Charlesworth B, Campos JL. 2014. The relations between recombination rate and patterns of molecular evolution and variation in *Drosophila*. Annu. Rev. Genet 48:383–403 [PubMed: 25251853]

Charlesworth B, Charlesworth D. 1997. Rapid fixation of deleterious alleles by Muller’s ratchet. Genet. Res 70:63–73 [PubMed: 9369098]

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–303 [PubMed: 8375663]

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res 70:155–74 [PubMed: 9449192]

Charlesworth D 1991. The apparent selection on neutral marker loci in partially inbreeding populations. Genet. Res 57:159–75

Charlesworth D 2003. Effects of inbreeding on the genetic diversity of plant populations. Philos. Trans. R. Soc. B 358:1051–70

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141:1619–32 [PubMed: 8601499]

Chikhi L, Rodríguez W, Grusea S, Santos P, Boitard S, Mazet O. 2018. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. Heredity 120:13–24 [PubMed: 29234166]

Comeron JM. 2017. Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Philos. Trans. R. Soc. B 372:20160471

Author Manuscript

Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics 161:389–410 [PubMed: 12019253]

Crisci JL, Poh Y-P, Mahajan S, Jensen JD. 2013. The impact of equilibrium assumptions on tests of selection. Front. Genet 4:235 [PubMed: 24273554]

Crow JF, Nagylaki T. 1976. The rate of change of a character correlated with fitness. Am. Nat 110:207–13 [PubMed: 29513548]

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet 14:262–72 [PubMed: 23478346]

Cvijović I, Good BH, Desai MM. 2018. The effect of strong purifying selection on genetic diversity. Genetics 209:1235–78 [PubMed: 29844134]

DeGiorgio M, Huber CD, Hubisz M, Nielsen R. 2016. SweepFinder2: increase in sensitivity, robustness and flexibility. Bioinformatics 32:1895–97 [PubMed: 27153702]

Eldon B, Birkner M, Blath J, Freund F. 2015. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics 199:841–56 [PubMed: 25575536]

Elyashiv E, Sattah S, Hu TT, Strutovsky A, McVicker G, et al. 2016. A genomic map of the effects of linked selection in Drosophila. PLOS Genet 12:e1006130 [PubMed: 27536991]

Ewing GB, Jensen JD. 2016. The consequences of not accounting for background selection in demographic inference. Mol. Ecol 25:135–41 [PubMed: 26394805]

Excoffier L, Dupanloup I, Huerta-Saánchez E, Sousa V, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLOS Genet 9:e1003905 [PubMed: 24204310]

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive mutations in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol 26:2097–108 [PubMed: 19535738]

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–13 [PubMed: 10880498]

Felsenstein J 1974. The evolutionary advantage of recombination. Genetics 78:737–56 [PubMed: 4448362]

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol. Biol. Evol 31:1275–91 [PubMed: 24554778]

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. Mol. Ecol 24:3529–45 [PubMed: 25943689]

Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, et al. 2014. Influenza virus drug resistance: a time-sampled population genetics perspective. PLOS Genet 10:e1004185 [PubMed: 24586206]

Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981–87 [PubMed: 10835415]

Galtier N, Rousselle M. 2020. How much does N_e vary among species? Genetics 216:559–72 [PubMed: 32839240]

Garud NR, Messer PW, Buszbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLOS Genet 11:e1005004 [PubMed: 25706129]

Gilbert KJ, Pouyet F, Excoffier L, Peischl S. 2020. Transition from background selection to associative overdominance promotes diversity in regions of low recombination. Curr. Biol 30:101–7 [PubMed: 31866368]

Good BH, Walczak AM, Neher RA, Desai MM. 2014. Genetic diversity in the interference selection limit. PLOS Genet 10:e1004222 [PubMed: 24675740]

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

Harris RB, Jensen JD. 2020. Considering genome scans for selection as coalescent model choice. Genome Biol. Evol 12:871–77 [PubMed: 32396636]

Harris RB, Sackman A, Jensen JD. 2018. On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLOS Genet 14:e1007859 [PubMed: 30592709]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Hartfield M, Bataillon T. 2020. Selective sweeps under dominance and inbreeding. G3 Genes Genomes Genet 10:1063–75

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–52 [PubMed: 15716498]

Hermisson J, Pennings PS. 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol 8:700–16

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res 8:269–94 [PubMed: 5980116]

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, et al. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am. Nat 188:379–97 [PubMed: 27622873]

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol 7:1–45

Hudson RR, Kaplan NL. 1994. Gene trees with background selection. In Non-Neutral Evolution: Theories and Molecular Data, ed. Golding B, pp. 140–53. London: Chapman & Hall

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. Genetics 141:1605–17 [PubMed: 8601498]

Irwin KK, Matuszewski S, Vuilleumier S, Ormond L, Shim H, et al. 2016. On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity 117:393–99 [PubMed: 27649621]

Jain K 2008. Loss of least-loaded class in asexual populations due to drift and epistasis. Genetics 179:2125–34 [PubMed: 18689884]

Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. Nat. Commun 5:5281 [PubMed: 25345443]

Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–10 [PubMed: 15911584]

Jensen JD, Kim Y, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176:2371–79 [PubMed: 17565955]

Johri P, Charlesworth B, Howell EK, Lynch M, Jensen JD. 2021a. Revisiting the notion of deleterious sweeps. Genetics In press. 10.1093/genetics/iyab094

Johri P, Charlesworth B, Jensen JD. 2020. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics 215:173–92 [PubMed: 32152045]

Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD. 2021b. The impact of purifying and background selection on the inference of population history: problems and prospects. Mol. Biol. Evol 38:2986–3003 [PubMed: 33591322]

Kaiser VB, Charlesworth B. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet 25:9–12 [PubMed: 19027982]

Kaiser VB, Charlesworth B. 2010. Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. Genetics 185:339–48 [PubMed: 20215466]

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitch-hiking" effect revisited. Genetics 123:887–99 [PubMed: 2612899]

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–24 [PubMed: 15280259]

Kim Y, Stephan W. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 155:1415–27 [PubMed: 10880499]

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–77 [PubMed: 11861577]

Kimura M 1971. Theoretical foundations of population genetics at the molecular level. Theor. Popul. Biol 2:174–208 [PubMed: 5162686]

Kimura M 1983. The Neutral Theory of Molecular Evolution Cambridge, UK: Cambridge Univ. Press

Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193:1197–208 [PubMed: 23341416]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Latter BDH. 1998. Mutant alleles of small effect are primarily responsible for the loss of fitness with slow inbreeding in *Drosophila melanogaster*. Genetics 148:1143–58 [PubMed: 9539431]

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–96 [PubMed: 21753753]

Machado HE, Lawrie DS, Petrov DA. 2020. Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. Genetics 214:511–28 [PubMed: 31871131]

Mafessoni F, Lachmann D. 2015. Selective strolls: Fixation and extinction in diploids are slower for weakly selected mutations than for neutral ones. Genetics 201:1581–89 [PubMed: 26500260]

Maruyama T, Kimura M. 1974. A note on the speed of gene frequency changes in reverse directions in a finite population. Evolution 28:161–63 [PubMed: 28563034]

Matuszewski S, Hildebrandt ME, Achaz G, Jensen JD. 2018. Coalescent processes with skewed offspring distributions and nonequilibrium demography. Genetics 208:323–38 [PubMed: 29127263]

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res 23:23–35 [PubMed: 4407212]

McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. 2013. Impact of deleterious passenger mutations on cancer progression. PNAS 110:2910–15 [PubMed: 23388632]

McVean G 2007. The structure of linkage disequilibrium around a selective sweep. Genetics 175:1395–406 [PubMed: 17194788]

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLOS Genet 5:e1000471 [PubMed: 19424416]

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. PNAS 110:8615–20 [PubMed: 23650353]

Muller HJ. 1964. The relation of recombination to mutational advance. Mutat. Res 1:2–9

Nicolaisen LE, Desai M. 2013. Distortions in genealogies due to purifying selection and recombination. Genetics 195:221–30 [PubMed: 23821597]

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–75 [PubMed: 16251466]

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. Genet. Res 67:159–74 [PubMed: 8801188]

Ohta T 1971. Associative overdominance caused by linked detrimental mutations. Genet. Res 18:277–86 [PubMed: 5158298]

Orr HA, Betancourt AJ. 2001. Haldane’s sieve and adaptation from the standing genetic variation. Genetics 157:875–84 [PubMed: 11157004]

Palsson S, Pamilo P. 1999. The effects of deleterious mutations on linked neutral variation in small populations. Genetics 153:475–83 [PubMed: 10471727]

Parmley JL, Hurst LD. 2007. How do synonymous mutations affect fitness? Bioessays 29:515–19 [PubMed: 17508390]

Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol 30:2224–34 [PubMed: 23777627]

Pennings PS, Hermisson J. 2006. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol 23:1076–84 [PubMed: 16520336]

Poh Y-P, Domingues V, Hoekstra HE, Jensen JD. 2014. On the prospect of identifying adaptive loci in recently bottlenecked populations. PLOS ONE 9:e110579 [PubMed: 25383711]

Pouyet F, Aeschbacher S, Thiery A, Excoffier L. 2018. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. eLife 7:e36317 [PubMed: 30125248]

Price GR. 1970. Selection and covariance. Nature 227:520–21 [PubMed: 5428476]

Przeworski M 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–89 [PubMed: 11901132]

Robertson A 1968. The spectrum of genetic variation. In Population Biology and Evolution, Proceedings of the International Symposium, June 7–9, 1967, Syracuse, ed. Lewontin RC, pp. 5–16. Syracuse, NY: Syracuse Univ. Press

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Sackman AM, Harris RB, Jensen JD. 2019. Inferring demography and selection in organisms characterized by skewed offspring distributions. Genetics 211:1019–28 [PubMed: 30651284]

Santiago E, Caballero A. 1995. Effective size of populations under selection. Genetics 139:1013–30 [PubMed: 7713405]

Santiago E, Caballero A. 1998. Effective size and polymorphism of linked neutral loci in populations under selection. Genetics 149:2105–17 [PubMed: 9691062]

Sattah S, Elyashiv E, Kolodny O, Rinott Y, Sella G. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLOS Genet 7:e1001302 [PubMed: 21347283]

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet 46:919–25 [PubMed: 24952747]

Schrider DR, Kern AD. 2016. S/HIC: robust identification of soft and hard sweeps using machine learning. PLOS Genet 12:e1005928 [PubMed: 26977894]

Schrider DR, Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the human genome. Mol. Biol. Evol 34:1863–77 [PubMed: 28482049]

Schrider DR, Mendes F, Hahn MW, Kern AD. 2015. Soft shoulders ahead: Spurious signatures of soft and partial sweeps result from linked hard sweeps. Genetics 200:267–84 [PubMed: 25716978]

Söderberg RJ, Berg OG. 2007. Mutational interference and the progression of Muller's ratchet when mutations have a broad range of deleterious effects. Genetics 177:971–86 [PubMed: 17720933]

Stephan W 2019. Selective sweeps. Genetics 211:5–13 [PubMed: 30626638]

Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172:2647–63 [PubMed: 16452153]

Stephan W, Wiehe THE, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol 41:237–54

Tajima F 1990. Relationship between DNA polymorphism and fixation time. Genetics 125:447–54 [PubMed: 2379822]

Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Mol. Ecol 23:2637–52 [PubMed: 24750385]

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genome scans for selective sweeps? Genome Res 16:702–12 [PubMed: 16687733]

Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175:737–50 [PubMed: 17110489]

Wakeley J 2008. Coalescent Theory: An Introduction Greenwood Village, CO: Roberts & Co.

Wiehe THE, Stephan W. 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data. Mol. Biol. Evol 10:842–54 [PubMed: 8355603]

Xia Q, Guo Y, Zheng Z, Li D, Xuan Z, et al. 2009. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). Science 326:433–36 [PubMed: 19713493]

Zeng K 2013. A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity 110:363–71 [PubMed: 23188176]

Zhao L, Charlesworth B. 2016. Resolving the conflict between associative overdominance and background selection. Genetics 203:1315–34 [PubMed: 27182952]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

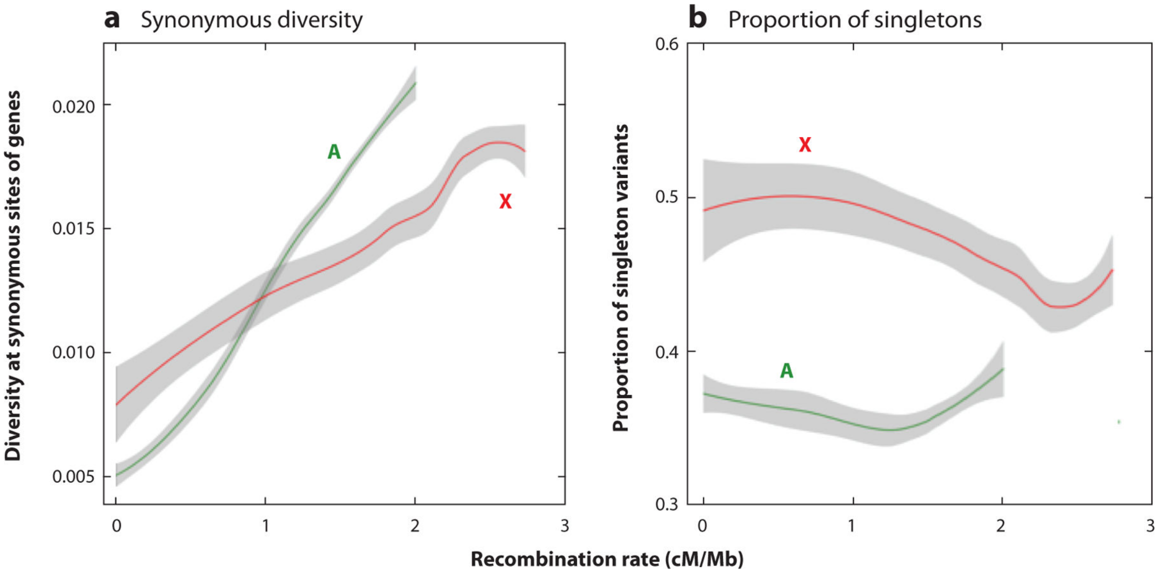

Figure 1.

(*a*) The *y*-axis shows the pairwise diversity per nucleotide site at synonymous sites of genes in a sample of 17 haploid genomes of *Drosophila melanogaster* from a Rwandan population (Campos et al. 2014). (*b*) The *y*-axis shows the proportion of singleton variants (those present as a single copy in the sample). The *x*-axis of panels *a* and *b* displays the estimated rate of crossing over per megabase for each gene, corrected for the absence of crossing over in males. The plots are Loess regression fits with 95% confidence intervals (*gray shading*). The green curves are for autosomal genes (A), and the red curves are for X-linked genes (X). Panel *a* shows that diversity increases with the rate of crossing over experienced by a gene, whereas panel *b* shows that the proportion of singletons has a complex relationship with the rate of crossing over, although on autosomes it tends to decline with the rate of crossing over in the lower part of the range of crossing over rates. The difference between the proportions of singletons on the X chromosome and autosomes is striking and is suggestive of stronger hitchhiking effects on the X chromosome (Campos et al. 2014).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

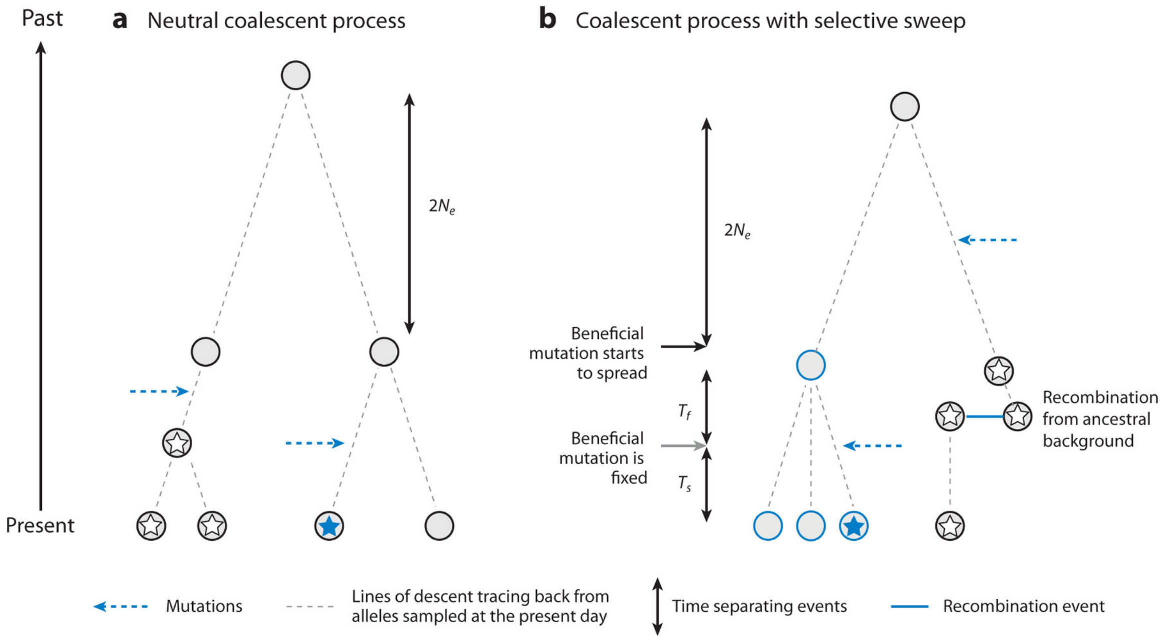
Past
a Neutral coalescent process
b Coalescent process with selective sweep
2N_e
2N_e
Beneficial mutation starts to spread
T_f
Beneficial mutation is fixed
T_s
Recombination from ancestral background
Present
Mutations
Lines of descent tracing back from alleles sampled at the present day
Time separating events
Recombination event

Figure 2.

(*a*) Neutral coalescence and (*b*) coalescence with a selective sweep. The dashed black lines represent lines of descent tracing back from alleles sampled at the present day (*bottom*, *gray circles*). The coalescence of two alleles into an ancestral allele from which they are descended is indicated by the merger of the pair of lines connecting them to the ancestor. The horizontal dashed blue arrows represent occurrences of mutations at different sites in the sequence. (*a*) The neutral coalescent process for a sample of four alleles. The first mutation (*white star*) occurred after the first coalescent event (looking back in time), such that its frequency is 1⁄2; the second mutation (*blue star*) occurred before the first coalescent event on its branch of the tree, such that its frequency is 1⁄4. The double-headed arrow to the right of the tree indicates the expected time to the last coalescent event ($2N_e$ generations). (*b*) Coalescence for a neutral locus linked to a site that has experienced a selective sweep, which finished T_s generations ago. The blue outlines indicate alleles carrying the beneficial mutation, all of which coalesced at the start of the sweep and whose duration is T_f generations. The solid black and gray arrows indicate the times of spread and fixation of this mutation, respectively. The solid blue line indicates a recombination event, such that the neutral site in question traced its ancestry to a wild-type background at the selected locus; its expected time to coalescence with the ancestor of all the nonrecombinant alleles counting back from the start of the sweep is $2N_e$ generations. The white stars indicate a mutation that arose in an ancestor of the recombinant allele; the blue star indicates a mutation that arose at a different site in an ancestor of a nonrecombinant allele. Both mutations have a frequency of 1⁄4 in the sample.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

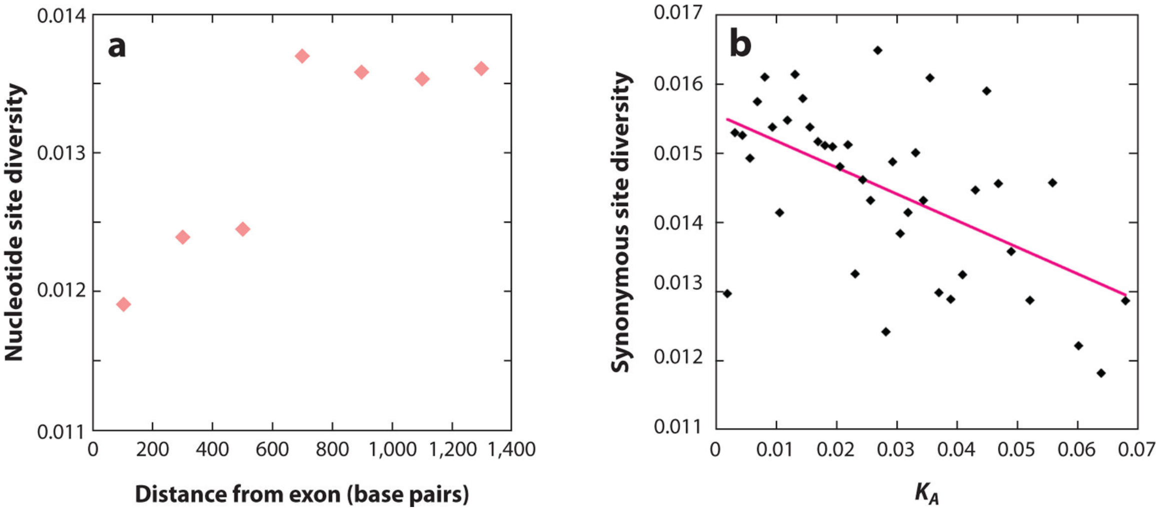

Figure 3.

(*a*) The *y*-axis is the mean nucleotide site diversity in 200-bp sliding windows of intergenic sequence, and the *x*-axis is the distance of the middle of each window from the 5’ end of the exon. The data are for 94 single-exon genes sequenced in 76 haploid genomes from *Drosophila melanogaster* individuals sampled in Zambia; noncoding sites under strong selective constraints have been masked (Johri et al. 2020). The Pearson correlation coefficient is $r = 0.88$, $p < 0.01$. (*b*) The points represent the mean synonymous site diversities of sets of autosomal genes from the Rwandan population of *D. melanogaster* used in Figure 1, grouped into 40 bins with respect to their divergence at nonsynonymous sites from the related species *D. yakuba* (K_A). The solid pink line is the least-squares linear regression of diversity on K_A ($y = 0.0156 – 0.0385x$, $r = − 0.563$, $p < 0.001$).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

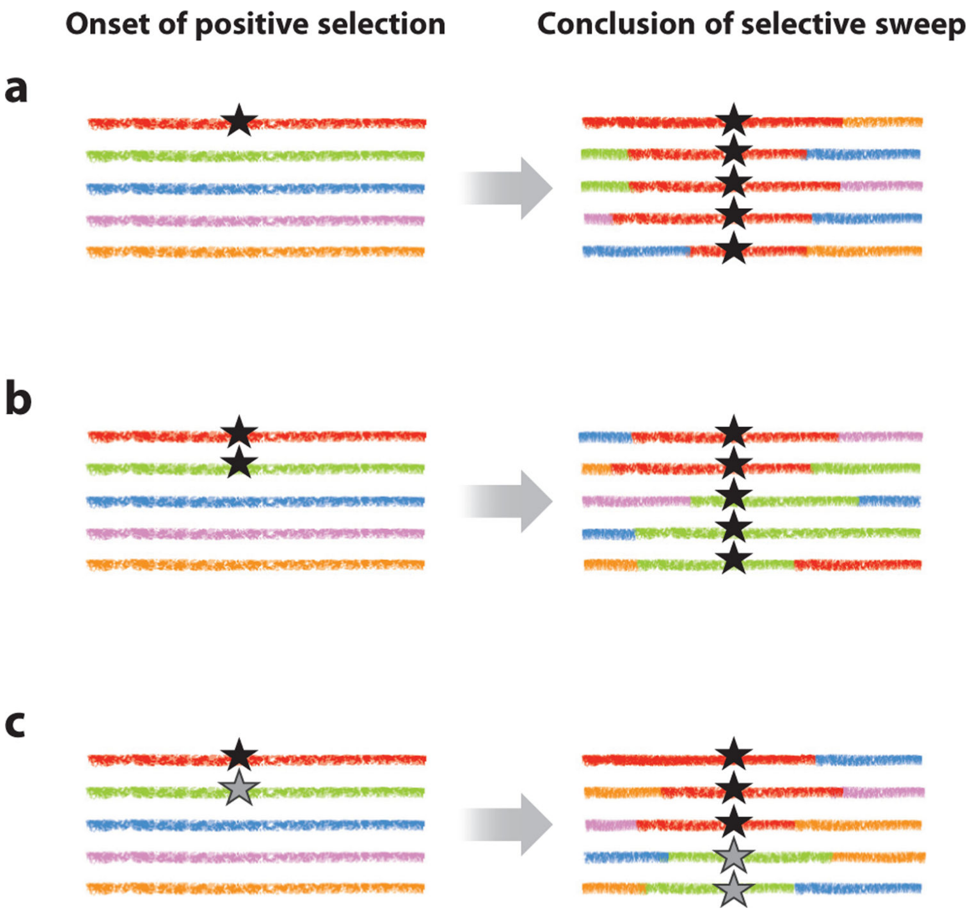

Figure 4.

Diagram of three model realizations at two timepoints for five chromosomes sampled from a population subject to a selective sweep. Each colored line represents a unique chromosome-wide haplotype segregating in the population at the onset of selection, carrying a unique combination of polymorphic variants. The black and gray stars indicate distinct new mutations that are subject to positive selection. The left-hand column represents the state of the population at the onset of a selective sweep, and the right-hand column represents the state at the end of the sweep, when all individuals carry a beneficial variant. All models are characterized by hitchhiking effects due to associations with the beneficial mutation, as well as by breaks in these associations at various distances from the selected site caused by recombination events. (*a*) Hard sweep, i.e., selection on a rare variant. This is characterized by fixation of a single haplotype close to the target of selection (*red haplotype*). (*b*) Soft sweep in which selection acts on a common variant that was formerly neutral or deleterious. (*c*) Soft sweep in which selection acts on two independently occurring, beneficial variants with the same selection coefficient (*black* and *gray stars*). Both soft selective sweep models are characterized by multiple haplotypes segregating immediately around the target of selection (the *red* and *green haplotypes*, on which the beneficial variant was previously segregating neutrally or on which two beneficial variants arose independently).