Patterns of Neutral Diversity Under General Models of Selective Sweeps

Graham Coop[1] and Peter Ralph
Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95616

ABSTRACT Two major sources of stochasticity in the dynamics of neutral alleles result from resampling of finite populations (genetic drift) and the random genetic background of nearby selected alleles on which the neutral alleles are found (linked selection). There is now good evidence that linked selection plays an important role in shaping polymorphism levels in a number of species. One of the best-investigated models of linked selection is the recurrent full-sweep model, in which newly arisen selected alleles fix rapidly. However, the bulk of selected alleles that sweep into the population may not be destined for rapid fixation. Here we develop a general model of recurrent selective sweeps in a coalescent framework, one that generalizes the recurrent full-sweep model to the case where selected alleles do not sweep to fixation. We show that in a large population, only the initial rapid increase of a selected allele affects the genealogy at partially linked sites, which under fairly general assumptions are unaffected by the subsequent fate of the selected allele. We also apply the theory to a simple model to investigate the impact of recurrent partial sweeps on levels of neutral diversity and find that for a given reduction in diversity, the impact of recurrent partial sweeps on the frequency spectrum at neutral sites is determined primarily by the frequencies rapidly achieved by the selected alleles. Consequently, recurrent sweeps of selected alleles to low frequencies can have a profound effect on levels of diversity but can leave the frequency spectrum relatively unperturbed. In fact, the limiting coalescent model under a high rate of sweeps to low frequency is identical to the standard neutral model. The general model of selective sweeps we describe goes some way toward providing a more flexible framework to describe genomic patterns of diversity than is currently available.

THE high levels of genetic variation within natural populations have long fascinated population geneticists. One school of thought holds that a substantial proportion of this molecular polymorphism is neutral or very weakly deleterious (Kimura and Ohta 1971; Ohta 1973; Kimura 1983). For neutral polymorphism, the level of genetic diversity results from a balance between the introduction of alleles through mutation and their stochastic loss (Kimura and Crow 1964; Kimura 1969; Ewens 1972). Under the neutral theory of molecular evolution this stochasticity is thought to result mostly from genetic drift (Kimura 1983), the random resampling that occurs in finite populations, an effect that is exaggerated by fluctuating population size and large variation in reproductive success among individuals (see Charlesworth 2009, for a recent review). However, selection at linked sites may provide a major source of stochasticity as the dynamics of a neutral allele can be strongly influenced by the random genetic background on which selected alleles arise (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Charlesworth *et al.* 1995; Hudson and Kaplan 1995b).

In many species examined to date, levels of diversity are substantially lower in regions of low recombination, as found in multiple species of *Drosophila* (Aguade *et al.* 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Begun *et al.* 2007; Shapiro *et al.* 2007), *Caenorhabditis* (Cutter and Payseur 2003; Cutter and Choi 2010), humans (Hellmann *et al.* 2008; Cai *et al.* 2009), and *Saccharomyces cerevisiae* (Cutter and Moses 2011), but not in all species, *e.g.*, *Arabidopsis* (Nordborg *et al.* 2005; Wright *et al.* 2006). Moreover, levels of diversity are also lower in regions that *a priori* are expected to have a higher rate of functional mutations, *e.g.*, near genes and conserved elements (Cai *et al.* 2009; McVicker *et al.* 2009; Hernandez *et al.* 2011). Since the rate of neutral genetic drift is independent of recombination rate, this positive correlation between recombination rates and

doi: 10.1534/genetics.112.141861
Manuscript received May 8, 2012; accepted for publication June 1, 2012
Supporting information is available online at http://www.genetics.org/content/suppl/2012/06/19/genetics.112.141861.DC1.
[1] Corresponding author: Department of Evolution and Ecology and Center for Population Biology, One Shields Ave., University of California, Davis, CA 95616. E-mail: gmcoop@ucdavis.edu

diversity offers good evidence that linked selection plays a substantial role in the fate of alleles, especially in low-recombination regions. What is still far from clear is how different forms of linked selection contribute to this reduction and whether linked selection can explain the narrow observed range of genetic diversity across species with vastly different (census) population sizes (Lewontin 1974; Maynard Smith and Haigh 1974).

Models of the effect of linked selection have often been divided between those that propose the source of this linked selection to be either the purging of deleterious variation (background selection) or the selective sweep of beneficial alleles (hitchhiking). In this article we explore the consequences of a generalized model of hitchhiking on patterns on neutral diversity. We first review some of the key results of models of linked selection. Under the background selection model, genetic diversity is continuously lost from natural populations due to the removal of haplotypes that carry deleterious alleles (Charlesworth *et al.* 1995; Hudson and Kaplan 1995b). For strongly deleterious alleles, this continuous loss acts primarily to increase the rate of genetic drift at markers closely linked to loci with high deleterious mutation rates (Hudson and Kaplan 1995a; Nordborg *et al.* 1996). Therefore, this background selection model leads to a reduction in genetic diversity but no skew in the frequency spectrum. However, a skew toward rare neutral alleles can result if weakly deleterious mutations are incorporated into the model (Nordborg *et al.* 1996; Gordo *et al.* 2002).

On the other end of the spectrum, Maynard Smith and Haigh (1974) proposed that local levels of genetic diversity could be reduced by the hitchhiking effect. The hitchhiking effect results from the fact that when an initially rare, beneficial allele sweeps rapidly to fixation, it carries with it a linked region of the haplotype on which it arose. The size of the genomic region affected by a recent sweep is proportional to the ratio of the strength of selection to the rate of recombination (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1998), and so the reduction in levels of diversity is determined by the distribution of selection coefficients and the rate of sweeps per unit of the genetic map. Neutral alleles farther away from the selected site may not be pulled all of the way to fixation if recombination occurs during the sweep, which can lead to a transient excess of high-frequency derived alleles an intermediate distance away from the selected site after each sweep (Fay and Wu 2000; Przeworski 2002; Kim 2006). As neutral diversity levels slowly recover through an influx of new mutations after the sweep, there is a strong skew toward low-frequency derived alleles, a pattern that persists for many generations (Braverman *et al.* 1995; Przeworski 2002; Kim 2006). In a large population, the rate of sweeps could be high enough that hitchhiking dominates genetic drift as the source of stochasticity (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Gillespie 2000), an idea that has been termed genetic draft (Gillespie 2000).

Support for a hitchhiking model over the standard model of background selection is found in *Drosophila*, where there is a greater skew toward rare alleles at putatively neutral sites in regions of low recombination (Andolfatto and Przeworski 2001; Shapiro *et al.* 2007) and regions surrounding amino acid substitutions have lower levels of diversity (Andolfatto 2007; Macpherson *et al.* 2007; Sattath *et al.* 2011). However, in humans (and other species) there is no strong skew toward rare alleles in low-recombination regions (McVicker *et al.* 2009; Hernandez *et al.* 2011; Lohmueller *et al.* 2011), which combined with other evidence (Coop *et al.* 2009; Hernandez *et al.* 2011) suggests that full sweeps may have been rare and that background selection may be the main mode of linked selection, in humans and a number of other species.

Although the recurrent full-sweep model has been the subject of considerable theoretical investigation, it may actually be relatively rare for advantageous alleles to sweep rapidly all the way to fixation. Fluctuating environments (*e.g.*, Gillespie 1991; Kopp and Hermisson 2007, 2009a,b) and changing genetic backgrounds may often act to prevent alleles from achieving rapid fixation within the population (see Pritchard *et al.* 2010 for a recent discussion). For example, if multiple mutations affecting the adaptive phenotype segregate during the sweep, then it may be that no one of these alleles sweeps to fixation (Pennings and Hermisson 2006a,b; Chevin and Hospital 2008; Ralph and Coop 2010). Multiple alleles spreading rapidly from low frequency can lead to either a set of partial sweeps within the population or a soft sweep if the alleles are tightly linked. Furthermore, a similar effect can occur when selection acts on an allele present as standing variation, if the allele is present on multiple haplotypes when it starts to spread (Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski *et al.* 2005). The fact that, under these models, no single haplotype goes quickly to fixation acts to reduce the hitchhiking effect and alters the effect on the frequency spectrum.

The genome-wide effect of other modes of linked selection on patterns of diversity is relatively unexplored. One model that has been investigated is an infinitesimal model of directional selection, where the aggregated effect of selection over many loci can be a substantial source of stochasticity at linked and even unlinked sites (Robertson 1961; Santiago and Caballero 1995, 1998; Barton 2000). Fluctuating selection due to varying environments has also been shown to lead to reduced levels of diversity at linked neutral sites (Gillespie 1994, 1997; Barton 2000) and simulations of specific models of fluctuating selection have shown that the same reduction in diversity can result in a much smaller skew in the frequency spectrum than under the hitchhiking model (Gillespie 1994, 1997). However, as yet no coalescent model of the effect of recurrent incomplete sweeps has been developed.

Here is an outline of how we proceed. First, we develop a coalescent-based model of patterns of diversity surrounding a selected allele that sweeps into the population but not necessarily to fixation. We concentrate on the case of a very large population and sites that are partially linked to this selected locus. We find that if the initial rise of the selected allele is rapid, then the coalescent process is primarily affected by this stage and relatively insensitive to the subsequent dynamics of the selected allele. Using this intuition, we then develop a coalescent model of recurrent sweeps on patterns of neutral diversity in which selected alleles may reach only intermediate frequency. To test the approximations involved in the model we compare the results at several stages to simulations. Some of the implications of these results for interpretation of genome-wide diversity patterns are presented in the *Discussion*.

Results

Coalescent framework and assumptions

As first described by Hudson and Kaplan (1988) and Kaplan *et al.* (1988), patterns of neutral diversity at a neutral locus linked to a selected locus can be modeled by conditioning on the trajectory of the frequency of the selected allele through time and treating the two allelic classes as subpopulations within each of which the dynamics are neutral, with recombination moving lineages between the two (see also Barton and Etheridge 2004; Barton *et al.* 2004). Consider a locus under selection at which a derived allele D and an ancestral allele A segregate, and let the frequency of D at time t be denoted $X(t)$. We study the coalescent process at a neutral locus partially linked to our selected locus, with recombination occurring at rate r per generation between the selected and the neutral locus. Each ancestor on a given lineage in the coalescent process carried either the D or the A allele at the selected locus, which we refer to as the "type" of that lineage.

Throughout we assume that the diploid population size N is large and constant over time. For simplicity, we assume that the effective population size is $2N$ [*i.e.*, the neutral coalescence rate of a pair of lineages is $1/(2N)$] and that no more than two lineages coalesce at once in the absence of a selective sweep.

Suppose at time t that k_D and k_A of our lineages are of the derived and the ancestral type, respectively. There are $NX(t)$ individuals carrying the derived allele that could be progenitors of the k_D lineages, so the instantaneous rates of coalescence of pairs of lineages within the two allelic classes at time t are

$$\binom{k_D}{2}\frac{1}{2NX(t)} \quad \text{and} \quad \binom{k_A}{2}\frac{1}{2N(1-X(t))}, \quad \text{respectively.} \tag{1}$$

The total instantaneous rate of recombination is $(k_D + k_A)r$. If a recombination event occurs on a lineage at time t, it chooses to be of type D with probability $X(t)$ and chooses to be of type A otherwise.

We leave the dynamics of the selective sweeps that determine $X(t)$ fairly unspecified, and while stochasticity may play an important role in shaping the trajectories, in examples we usually treat $X(t)$ as nonrandom. As we want coalescences caused by a single selective sweep to occur at more or less the same time, we require that once the selected allele is introduced into the population, it increases in frequency rapidly, and that once the allele frequency leaves the boundary (*e.g.*, moves above 1%), it does not return (*e.g.*, drops below 1%) unless it does so on the way to loss (*e.g.*, hits 0 before returning to 1%). This condition implies that our model applies to alleles that are at least partially codominant, as fully recessive alleles spend appreciable time, behaving stochastically, at very low frequencies, which can lead to different coalescent dynamics at linked loci (Teshima and Przeworski 2006; Ewing *et al.* 2011).

Relation to previous models

We describe a simple approximation to the coalescent with recurrent sweeps that is inspired by similar approximations for a model of recurrent full sweeps. The approximation postulates two types of coalescent events: "neutral" events occurring at rate $1/2N$ between any pair of lineages and additional coalescent events, involving two or more lineages, due to selective sweeps. The first class of events can occur at any time, due to random resampling of lineages. The second class of events, the sweep-induced coalescent events, can involve more than two lineages, as we assume that lineages forced to coalesce by a sweep do so instantaneously on the relevant timescale. We assume that all such lineages coalesce into a single lineage and that the distribution of the number of such lineages is binomial, with a success probability that is a function of the trajectory taken by the selected allele and the recombination distance to that allele. This framework is a natural extension of similar approximations used for full sweeps (Barton 1998; Gillespie 2000; Kim and Stephan 2002; Durrett and Schweinsberg 2005; Nielsen *et al.* 2005).

Processes with two classes of coalescent events have previously been developed to approximate a recurrent full-sweep model (Kaplan *et al.* 1989; Gillespie 2000; Durrett and Schweinsberg 2005). When the transition probabilities can be written in this binomial form, as they also are in the recurrent full-sweep models of Gillespie (2000) and Durrett and Schweinsberg (2005), the model is called a Λ-coalescent (Pitman 1999; Sagitov 1999). These also arise in neutral models where individuals have large variance in reproductive success (*e.g.*, Möhle and Sagitov 2001; Sargsyan and Wakeley 2008). As in other work, we present this model as an approximation not in the sense of asymptotic convergence, but rather as a simplification, which we show later is close enough to be useful. We make a number of simplifying assumptions and often do not make use of the most accurate analytical forms available, in an effort to maintain an intuitive form and description of the process obtained. In particular, Durrett and Schweinsberg (2004)

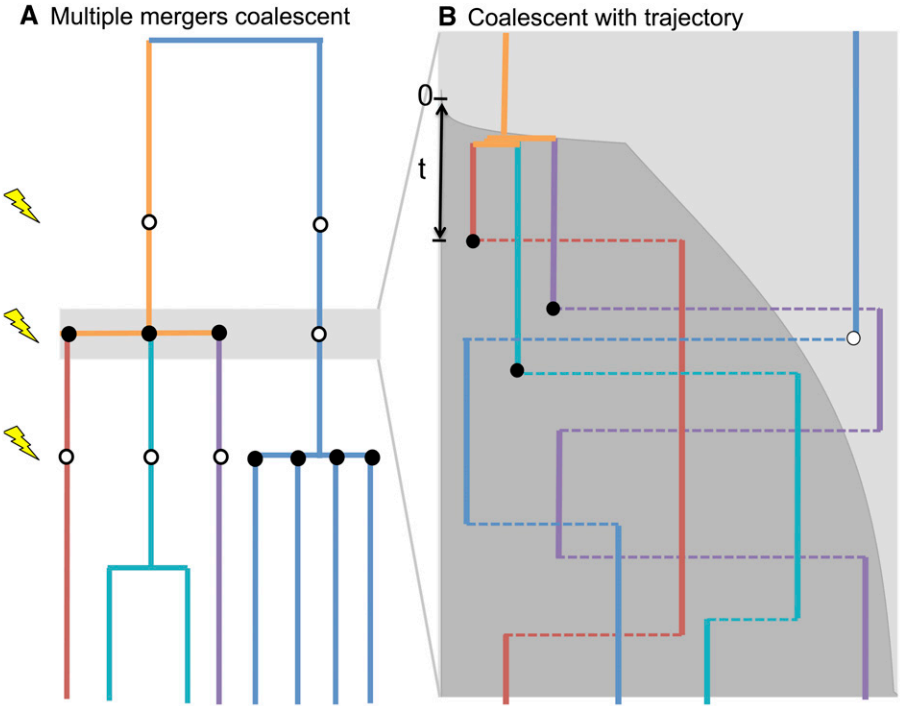

Figure 1 (A) An example of a multiple-merger coalescent genealogy. Eight alleles have been sampled in the present day, and we trace their lineages backward through time, up the page. Lightning bolts indicate the times when a selected allele has swept into the population. At each sweep, each lineage is either descended from the original carrier of the derived allele at the selected site (lineages marked with a black circle) or descended from some other ancestor (lineages marked with a white circle). (B) Zooming in on one sweep. The frequency of the derived allele, *D*, through time, *X*(*t*), is shown in dark gray. The four surviving lineages are shown in different colors as in A. Horizontal dashed lines depict recombination events in the history of a lineage. A circle indicates the oldest recombination event experienced by each of our lineages before the *D* allele arose, and the color of the circle indicates where the allele recombined onto the *D* background (black) or on to the *A* background (white). As we approach the time the selected allele arose, the three lineages found on the *D* background coalesce into a single lineage.

showed that a coalescent process with simultaneous multiple collisions could provide a better approximation to the coalescent process during a sweep, a direction we do not pursue (see also Barton 1998; Etheridge *et al.* 2006).

An approximation to the coalescent process during the sweep

Figure 1A shows an example of the relationships between different sampled individuals at a neutral locus in a finite population undergoing recurrent selective sweeps. At the times indicated by the lightning bolts, selective alleles sweep into the population at some locus linked to our neutral site. All lineages descended from the original carrier of the derived allele coalesce, nearly instantaneously on this timescale.

Figure 1B zooms in on one of these selective sweeps. The derived allele at the selected locus (*D*) arose τ generations ago. The five surviving ancestral lineages recombine on and off the *D* background, whose frequency through time is shown by the dark gray shading. Just after time 0 those lineages on the *D* background coalesce as *X* goes to zero (their coalescent rate, which is proportional to $1/X$, goes to infinity). We will show that the complexity of the process shown in Figure 1B can be approximated by a much simpler multiple-merger coalescent process suggested by Figure 1A, in which lineages coalesce “neutrally” at rate $1/(2N)$, and furthermore, each lineage flips a coin at each selective sweep to decide which type it is, and those that are of type *D* merge simultaneously.

Suppose that a derived allele at the selected locus (*D*) arose τ generations ago, at time 0. The selected mutation may still segregate within the population in the present day or may have gone to fixation or loss sometime before the present [in which case $X(\tau) = 1$ or 0, respectively]. First consider coalescences occurring very close to the origin of a selective mutation. A lineage can be type *D* at time 0 for one of two reasons: either it was of type *D* in the present day and not yet recombined off the *D* background or at the first recombination after the selected allele arose, the lineage chose to be of type *D*. The lineage of an individual drawn at random from the present-day population is therefore of type *D* at time 0 with probability

$$q = q(r,X) := X(\tau)e^{-r\tau} + r\int_0^{\tau} e^{-rt}X(t)dt. \qquad (2)$$

Here the integral is over *t*, the number of generations between the origin of *D* and the first subsequent recombination on a lineage (*t* is marked for the red lineage in Figure 1B). Note that although many recombination events may have occurred, since at each recombination event the lineage chooses a new type independently of its previous type, we need consider only the first after the sweep. If $\tau \gg 1/r$, the first term can be ignored, so we commonly assume that

$$q(r,X) = r\int_0^{\infty} e^{-rt}X(t)dt, \qquad (3)$$

as the allelic state of the sample has long been forgotten. Importantly, we can see that the dependence of *q* on *X* decays exponentially through time at rate *r*. Therefore, the fate of the selected allele more than a few multiples of *r* after it arose, including its presence or absence in the present day, will have little effect on *q*. Concretely, for two trajectories labeled 1 and 2, if $X_1(s) = X_2(s)$ for all $0 \le s \le T$, then regardless of subsequent differences in the trajectories, $|q_1 - q_2| \leq e^{-rT}$.

We can now approximate the rapid coalescence of lineages that are forced by the sweep by assuming that all lineages descended from the original carrier of the D allele coalesce *simultaneously* when the selected allele appears (a "multiple merger"). The lineages will actually coalesce at slightly different times, but the assumption the derived allele increases rapidly implies that this difference is small on the coalescent timescale [*i.e.*, $o(2N)$]. As each lineage takes part in this merger independently with probability q, the probability that i of k surviving lineages coalesce at time 0 is

$$\binom{k}{i} q^i (1-q)^{k-i}, \quad \text{for} \quad 2 \leq i \leq k, \tag{4}$$

reducing the number of lineages from k to $k - i + 1$.

This approximation assumes that each lineage makes an independent choice of whether to recombine off the sweep, which is equivalent to assuming that the coalescences caused by the sweep form a "star"-like tree, with no internal edges of nonzero length. Therefore, the approximation ignores dependencies between lineages induced by coalescent events earlier in the sweep and so is a poorer approximation for a large number of lineages. More sophisticated approximations have been developed to account for this dependency, which improve the properties for large samples (Barton 1998; Durrett and Schweinsberg 2004; Etheridge *et al.* 2006; Pfaffelhuber *et al.* 2006). However, we believe this approximation captures many of the important features.

The other component of our approximation is that at all times, all pairs of lineages coalesce at rate $1/(2N)$ regardless of their allelic background. This approximation ignores the fact that lineages that are currently on different backgrounds cannot coalesce and that lineages on the same background coalesce at a higher rate (see Equation 1).

We also note that although large changes in the allele frequency over a small number of generations represent a large number of children descended from a smaller number of ancestors, this will not cause rapid coalescence in a large population if the allele remains at intermediate frequencies. Concretely, consider a short time interval from generation t_1 to generation t_2, over which interval $X(t) \gg (t_2 - t_1)/N$. The chance that any coalescence occurs during this time interval on the derived background is small [$O((t_2 - t_1)/(X(t)N))$], regardless of how the frequency X changes. Therefore, large, sudden changes in allele frequencies will force coalescence on the derived background only if $X(t)$ is of order $1/N$ (and similarly for the ancestral background). For sites that are only partially linked to the selected locus, if recombination is moving the lineages across backgrounds at a sufficiently high rate compared to the neutral coalescent rate ($Nr \gg 1$), then two lineages in this subdivided model coalesce at a rate close to $1/2N$ (see Hudson and Kaplan 1988; Hey 1991; Nordborg 1997; Barton and Etheridge 2004 for a detailed discussion). As such, our approximation will therefore be worse close to the selected site, but is asymptotically correct for large r.

A simple trajectory: To build intuition, we first consider a simple trajectory, making further approximations to keep the results accessible, and compare the results to full coalescent simulations. Assume that D arises τ generations ago at a site at distance r from the neutral site under consideration, rapidly sweeps to frequency x, and remains close to this frequency for a time $\gg 1/r$. Under many models of directional selection, most of the time spent in reaching x is spent at low frequency, so that any recombination that occurs during this time will likely move a lineage to the ancestral type, and so only lineages that do not recombine during the initial sweep will coalesce. If we let t_x be the time it takes for the selected allele to sweep to x and assume $r\tau \gg 1$, then a simple approximation to $q(r, X)$ is therefore (with the subscript emphasizing dependence on x)

$$q(r,X) \approx q_x := xe^{-rt_x}. \tag{5}$$

If the initial increase of D is driven by additive selection of strength s with $Ns > 1$, then the initial trajectory of D will be logistic, and it is reasonable to take $t_x = \log(\alpha x/(1 - x))/s$, where $\alpha = 2N$ or $4Ns$ depending on whether s is of order 1 or $1/N$ [the latter case corresponding to the case where the selected allele has to rapidly achieve frequency $1/(Ns)$ to escape loss by drift]. Using q_x to approximate the probability that a lineage is caught by the sweep, the expected pairwise coalescent time is smaller by a factor of

$$\left(1 - q_x^2 e^{-\tau/(2N)}\right), \tag{6}$$

which can be found by considering whether a pair of lineages coalesce before, during, or after the sweep.

If rather than remaining near x, the selected allele continues to sweep to fixation—perhaps it is still under selection with strength $s_2 \gg r$—then $q_x \approx e^{-rt_x}$ because the selected allele has gone quickly to fixation as in a full sweep, and the only time for recombination is in the early phase of the trajectory t_x. On the other hand, if the allele became strongly deleterious ($s_2 < 0$ and $|s_2| \gg r$), then $q \approx 0$, because there is little chance of it contributing genetic material to the population. However, if selection subsequently experienced by D is weak ($|s_2| \ll r$), so that subsequent dynamics of the selected allele are sufficiently slow, then q and therefore the coalescent process are independent of the eventual fate of the selected allele. In summary, for q_x to be a good approximation to $q(r, X)$ and for the sweep to have an appreciable effect on the coalescent, we need $|s_2| \ll r < s$.

Comparison to simulation: To demonstrate this, we apply the same approximation to situations with different long-term behaviors. The code for these simulations and all simulations in the paper are contained in Supporting Information, File S1. We consider five different possible trajectory types. In all

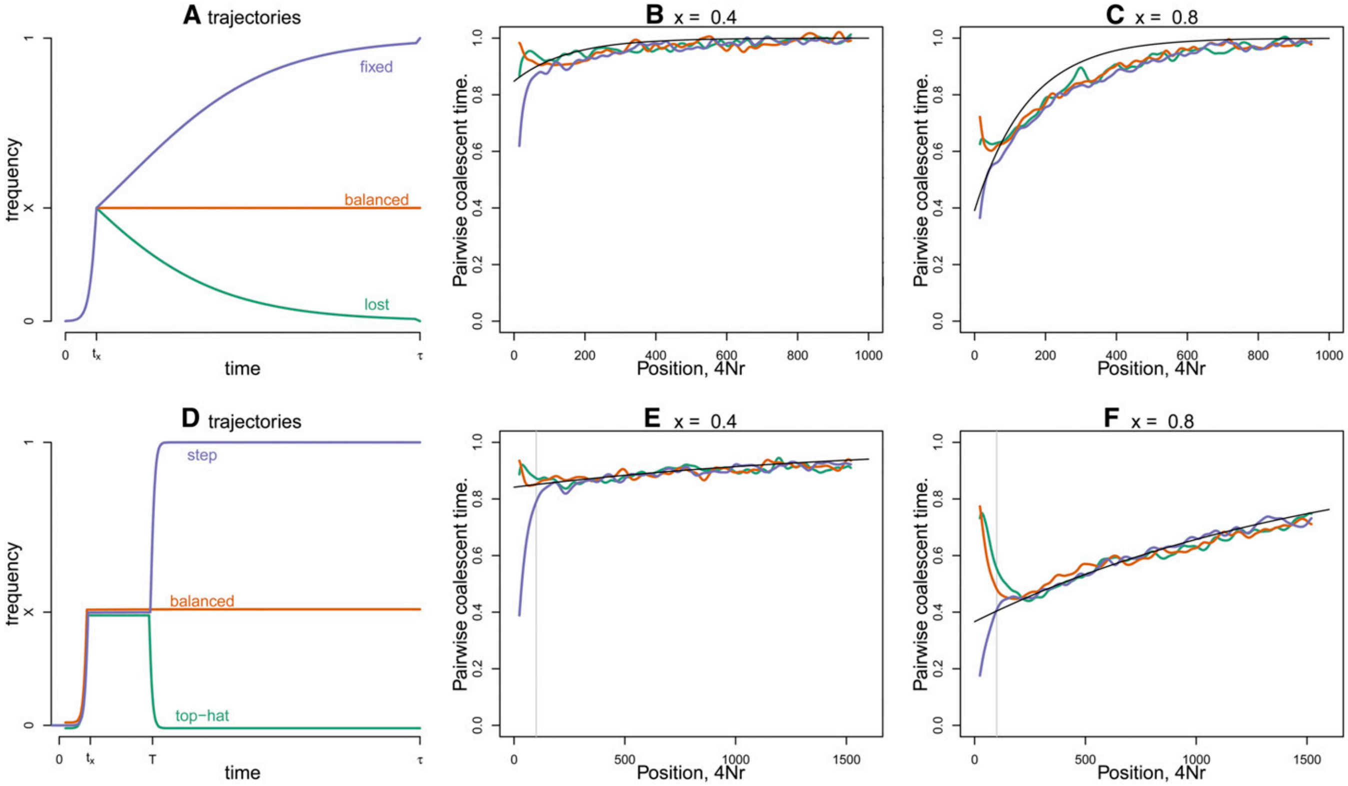

Figure 2 The effect of a single partial sweep. (A) Three possible trajectories followed by the *D* allele after it arises *τ* generations ago, as described in the text: blue is "fixed," green is "lost," and orange is "balanced." (B and C) Mean pairwise coalescent time against recombination distance away from a selected site that has experienced one of the three types of sweeps shown in A, with $x = 0.4$ and 0.8, respectively. The other parameters were $t_x/2N = 6.6 \times 10^{-3}$ and $\tau/2N = 0.05$. (D) Another three possible trajectories: green is "top hat" and blue is "step." (E and F) Pairwise coalescent time as in B and C, but using the trajectories shown in D. The other parameters were $t_x/2N = 6.1 \times 10^{-4}$, $\tau/2N = 0.1$, and $T/2N = 0.02$. The black line shows the approximation to the pairwise coalescent time of Equation 6. In E and F, the vertical line gray line marks $r = 1/T$.

cases, the initial rise of D was modeled as deterministic logistic growth begun at frequency $1/2N$ and adjusted to reach frequency x after t_x units of time. In the first case ("balanced"), the allele remains thereafter at frequency x. In the next two cases (Figure 2, A–C), after time t_x, allele D approaches either frequency 1 ("fixed") or frequency 0 ("lost") logistically, reaching frequency $1 - 1/2N$ (or $1/2N$, respectively) after the next τ time units. In the last two cases (Figure 2, D–F), the allele D remains at x for T generations and then proceeds logistically, in time t_x, either to frequency $1 - 1/2N$ ("step") or to frequency $1/2N$ ("top hat").

In each case, we used *mssel* [a modified version of *ms* (Hudson 2002) that allows an arbitrary trajectory, kindly supplied by Richard Hudson] to simulate genealogies for a recombining sequence surrounding a selected locus at which a selected allele performs one of the trajectories shown in Figure 2 . The average pairwise coalescence time from these simulations was calculated by dividing the pairwise genetic diversity by the mutation rate and is shown in Figure 2 at different distances from the selected locus, compared to the quantity predicted by Equation 6. Close to the selected site (*e.g.*, for $r < 1/T$ in Figure 2, E and F) the curves diverge, since the sites represented by the blue curves see a full sweep, reducing diversity close to the selected site, while those in the orange curves see a short-term balanced polymorphism and hence show a peak in polymorphism near the selected site. As we increase recombination distance away from the selected site, the three curves are in good agreement with the black line (Equation 6), indicating that our partial sweep model captures the main effect on diversity.

Our simple approximation describes diversity levels well at partially linked sites over a range of different scenarios and works well for a wider range of parameters (results not shown). We furthermore used Equation 4 to predict the effect of this simple partial sweep on the coalescent process of more than two lineages and found close agreement with further *mssel* simulations for various summaries of diversity such as the expected number of segregating sites (results not shown). Overall, these results confirm that for partially linked sites, the coalescent process is mostly determined by the initial rapid behavior of the selected allele.

A recurrent sweep coalescent model

We now consider patterns of diversity at a neutral locus affected by many different selected alleles that sweep into the population at the times of a homogeneous Poisson process with rate ν. We assume that the sweep rate is low enough that sweeps do not interfere with each other and return to discuss this assumption later. Each sweep occurs at some distance r from the neutral locus, and as it sweeps its frequency follows some particular trajectory $X(t)$, which together in Equation 3 determine q, the probability that a lineage at the neutral site is caught by the sweep. Rather than try to explicitly model randomness in these two components, at first we assume that each sweep independently chooses its value of q from a probability distribution with density $f(q)$. This model is exactly a Lambda coalescent, with $\Lambda(dq) = q^2\nu f(q)dq + \delta_0(dq)/2N$ (see Berestycki 2009, for a recent review), but we leave our discussion in terms of f to make the results more intuitive.

Following from our assumption that each lineage is affected by a given sweep independently with probability q, when there are k surviving lineages, the rate at which they coalesce to $k - i + 1$ lineages due to sweeps is

$$\nu\binom{k}{i}\int_0^1 q^i(1-q)^{k-i}f(q)dq. \tag{7}$$

This follows from our assumption that sweeps occur homogeneously through time and do not interfere with each other and with properties of marked Poisson processes. For ease of presentation we denote

$$I_{k,i} = \binom{k}{i}\int_0^1 q^i(1-q)^{k-i}f(q)dq. \tag{8}$$

Recall that under our model, the rate of coalescence of pairs of lineages due to genetic drift is $1/(2N)$, so that the rate at which the coalescent process with k lineages coalesces to $k - i + 1$ lineages is

$$\lambda_{k,i} = \binom{k}{2}\frac{1}{2N}\delta_{i,2} + \nu I_{k,i} \quad \text{for} \quad 2 \le i \le k, \tag{9}$$

where $\delta_{i,2} = 1$ if $i = 2$ and 0 otherwise. The total rate of coalescent events when there are k lineages is therefore

$$\lambda_k = \frac{1}{2N}\binom{k}{2} + \nu\sum_{i=2}^{k} I_{k,i} \quad \text{for} \quad k \ge 2, \tag{10}$$

and conditional on a coalescent event the probability that i lineages of k coalesce, reducing from k to $k - i + 1$ lineages, is

$$p_{k,k-i+1} = \frac{\lambda_{k,i}}{\lambda_k} = \frac{(1/2N)\binom{k}{2}\delta_{i,2} \; + \; \nu I_{k,i}}{(1/2N)\binom{k}{2} + \nu\sum_{i=2}^{k} I_{k,i}}, \quad \text{for} \quad 2 \le i \le k. \tag{11}$$

To simulate from this coalescent process we can simulate an exponential waiting time with rate λ_k, pick a number of lineages to coalesce using probabilities $p_{k,k-i+1}$, and run this process until we have a single lineage remaining.

Note that in deriving this process we have assumed that at all times, lineages also coalesce at a neutral rate $1/2N$. This can be justified by assuming that recombination moves lineages between backgrounds at a high enough rate to allow the effects of the partitioning of the population by segregating alleles to be ignored. Therefore, the approximation will break down if a typical neutral site, at any given time, is close enough (*e.g.*, within an r of order $1/N$) to an allele maintained at intermediate frequency by long-term balancing selection (*e.g.*, alleles maintained for timescales of order N). Further work is needed to refine the coalescent under those conditions, but our approximations should be suitable for a broad range of scenarios and genomic regions.

The coalescent process with homogeneous sweeps

It is natural to examine the case in which selective sweeps occur at a uniform rate along a sequence of total length L. We assume that this sequence recombines at rate r_{BP} per base each generation and that sweeps enter the population at a rate ν_{BP} per base each generation, so that the total rate of sweeps is $\nu = \nu_{\text{BP}}L$. We also assume that the sweeps are homogeneous; *i.e.*, the trajectory followed by the frequency of the derived allele, X, is independent of the distance between our neutral site and the site at which a sweep occurs.

We consider sweeps occurring along a very long chromosome and so take $L \to \infty$, but then the total rate of sweeps, $\nu = \nu_{\text{BP}}L$, also goes to infinity. To obtain a meaningful limit, we need that as $L \to \infty$ the rate of sweeps corresponding to each nonzero value of q converges to a finite value. Recall from (3) that the probability a lineage is caught up in a given sweep depends on the distance to the sweep (which is $r_{\text{BP}}\ell$ for a site ℓ bases away) and the trajectory X taken by the sweep and is given by $q(r_{\text{BP}}\ell, X) = r_{\text{BP}}\ell\int_0^\tau \exp(-r_{\text{BP}}\ell t)X(t)dt$. In a finite genome of length L, the probability distribution on values of q has density $f(q) = h_L(q)/L$, where $h_L(q) = \int_0^L \mathbb{P}_X\{q(r_{\text{BP}}\ell, X) \in dq\}d\ell$. Here $h_L(q)$ is the rate at which selective sweeps appear at location $r_{\text{BP}}\ell$ and whose trajectory X gives $q(r_{\text{BP}}\ell,\, X) = q$, integrated across the genome; and $f(q)$ is $h_L(q)$ normalized to integrate to 1, since $\int_0^1 h_L(q)dq = L$. The functions h_L converge for $q > 0$ as L becomes large as long as the probability that distant sweeps affect the focal site decays quickly enough. We therefore assume that $h_L(q)$ converges to a finite limit $h(q)$, *i.e.*, that the following exists:

$$h(q) = \lim_{L\to\infty} L\,f(q) \quad \text{for} \quad 0 < q \le 1. \tag{12}$$

This means that although the total rate of sweeps per generation is infinite, only a finite number happen close enough to our neutral site to potentially affect our coalescent process. Therefore, the rate at which k lineages coalesce down to $k - i + 1$ due to sweeps converges:

$$\nu_{\text{BP}}LI_{k,i} \to \nu_{\text{BP}}\binom{k}{i}\int_0^1 q^i(1-q)^{k-i}h(q)dq \quad \text{as} \quad L \to \infty. \tag{13}$$

If we take the trajectory X to be fixed, we can rewrite Equation 13 as

$$\begin{aligned}
&\nu_{\mathrm{BP}}\binom{k}{i}\int_0^1 q^i(1-q)^{k-i}h(q)dq\\
&=\nu_{\mathrm{BP}}\binom{k}{i}\int_0^\infty q(r_{\mathrm{BP}}\ell,X)^i(1-q(r_{\mathrm{BP}}\ell,X))^{k-i}d\ell\\
&=\frac{\nu_{\mathrm{BP}}}{r_{\mathrm{BP}}}\binom{k}{i}\int_0^\infty q(r,X)^i(1-q(r,X))^{k-i}dr,
\end{aligned}\tag{14}$$

which decouples the dependency of the rate of sweeps on the recombination rate r_{BP} from the trajectory X. If X is random, then we need to average over possible trajectories, and so we define

$$J_{k,i}=\binom{k}{i}{}_X\left[\int_0^\infty q(r,X)^i(1-q(r,X))^{k-i}dr\right],\tag{15}$$

where $\mathbb{E}_X[\cdot]$ denotes the average over possible trajectories. We assume that this integral is finite for $2 \le i \le k$; for further discussion of these points see *Appendix A*. Importantly, under our assumption that sweeps do not interfere with each other, $J_{k,i}$ does not depend on the recombination rate r_{BP} or the rate of sweeps ν_{BP}, but only on the dynamics of the selective sweeps X.

Allowing coalescent events due to drift, k lineages coalesce down to $k - i + 1$ at rate

$$\lambda_{k,i}=\frac{1}{2N}\binom{k}{2}\delta_{i,2}+\frac{\nu_{\mathrm{BP}}}{r_{\mathrm{BP}}}J_{k,i}\quad\text{for}\quad 2\le i\le k,\tag{16}$$

where $\delta_{i,2} = 1$ if $i = 2$ and is 0 otherwise. As Equation 16 follows from the simple change of variable in Equation 14, it will hold under any homogeneous sweep model where sweeps instantaneously (relative to a timescale of $2N$) force lineages to coalescence, with $J_{k,i}$ replaced by some constant that does not depend on r_{BP} or ν_{BP}. This result greatly generalizes that of Kaplan *et al.* (1989), who described a similar coalescent process for a full-sweep model.

We can see from Equation 16 that $2N\nu_{\mathrm{BP}}/r_{\mathrm{BP}}$ is the relevant compound parameter that in a general sweep model determines the rate of sweeps relative to neutral coalescent events. In small samples, sweep-induced coalescent events will dominate those due to drift if the population-scaled rate of sweeps per unit of the genetic map is much greater than one, provided that not all the $J_{k,i}$ are too small. We revisit this strong-sweep limit in the *Limiting processes* section.

The coalescent process with homogeneous partial sweeps

We now return to the setting above (in *A simple trajectory*), in which a simple trajectory rises quickly to frequency x, under which assumptions $q(r, X) \approx q_x$ (Equation 5). We suppose that the frequency x at which each sweep slows is chosen independently with probability density $g(x)$. It also seems reasonable to assume furthermore that t_x, the time it takes to reach frequency x, does not depend on x; we denote this time by t. This is approximately true for many models of directional selection, since selected alleles move quickly through intermediate frequencies. In this case, the rate at which k lineages coalesce to $k - i + 1$ is

$$\lambda_{k,i}\,\frac{1}{2N}\binom{k}{2}\delta_{i,2}+\binom{k}{i}\frac{\nu_{BP}}{t\,r_{BP}}\int_0^\infty\left(\int_0^1(xe^{-r})^i\,(1-xe^{-r})^{k-i}g(x)dx\right)dr,\tag{17}$$

suggesting that the important quantity, which acts as a coalescent timescaling, is $2N\nu_{\mathrm{BP}}/(t\ r_{\mathrm{BP}})$, with the distribution on x acting to control how many lineages are forced to coalesce with each sweep. If we determine t by a simple model of additive selection with selection coefficient s, the key parameter becomes $2N\nu_{\mathrm{BP}}s/(\log(Ns)r_{\mathrm{BP}})$.

This compound parameter, $2N\nu_{\mathrm{BP}}s/(\log(Ns)r_{\mathrm{BP}})$, is also the key parameter in full-sweep models (Kaplan *et al.* 1989; Stephan *et al.* 1992). However, since full sweeps require $x = 1$, if diversity is strongly reduced, then numerous lineages must merge at each sweep, which in turn leads to a strong skew toward rare alleles in the frequency spectrum. We will see that this relationship between the reduction in diversity and the skew in the frequency spectrum is substantially weakened under a partial sweep model when we allow $x \ll 1$.

Summaries of neutral genetic diversity

Level of neutral diversity: A key quantity of interest is the level of neutral nucleotide diversity, π, the number of differences between randomly sampled alleles at a neutral locus. Under an infinite-sites model of mutation, which we use here, the expectation of π, averaging across sites, is equal to the expected coalescent time of a pair of lineages multiplied by twice the mutation rate. If the mutation rate per generation at our neutral locus is μ, in the absence of sweeps, the level of diversity is $\mathbb{E}[\pi] = \theta$, where $\theta = 4N\mu$ is the population-size–scaled mutation rate, and the expectation is the average across sites. Note that θ is the level of diversity under the usual neutral model.

Under our model featuring both sweeps and drift,

$$\mathbb{E}[\pi]=\frac{\theta}{1\ +\ 2NI_{2,2}\nu},\tag{18}$$

so a key parameter is the population-scaled rate of sweeps $2N\nu$.

To examine the applicability of our approximations we again performed coalescent simulations with *mssel* for a selected locus at a fixed location experiencing recurrent sweeps. In this case, where selected alleles recurrently sweep into the population at a *fixed* genetic distance r, following our simple partial sweep trajectory again as characterized by q_x and $2N$, the nucleotide diversity is given by

$$\mathbb{E}[\pi]=\frac{\theta}{1+2N\nu x^2\ \exp(-2rt_x)}.\tag{19}$$

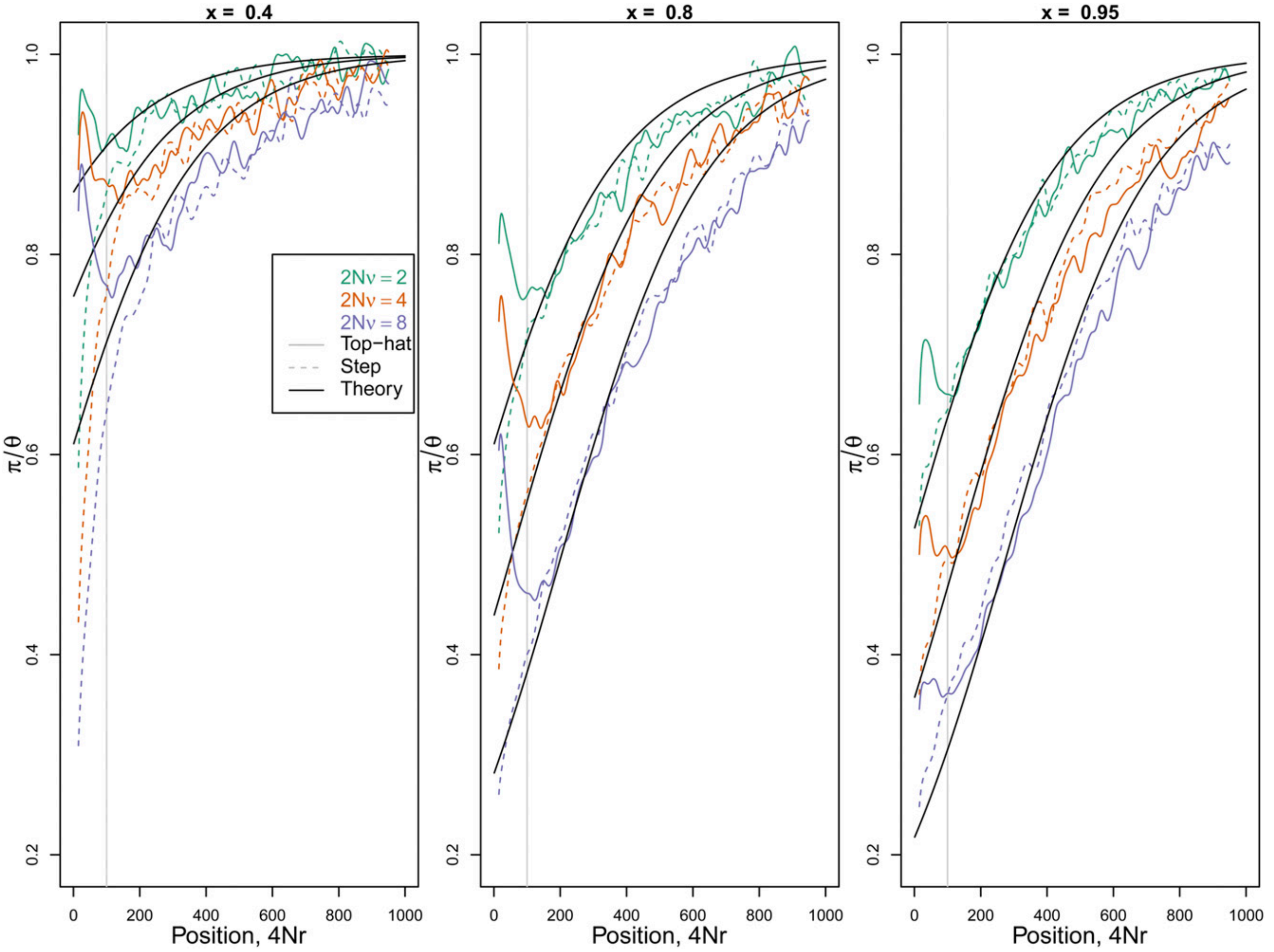

Figure 3 Reduction in diversity (π/θ) as a function of recombination distance from a site experiencing recurrent sweeps. The three panels are for different values of the frequency x that each sweep reached rapidly. The solid line is for recurrent top-hat trajectories and the dashed line for recurrent step trajectories The time that the trajectory pauses is $T/2N = 0.01$ and $t_x/2N = 0.003$ in both cases. The three colors correspond to three different population-scaled rates of sweeps: $2N\nu = 2$, 4, and 8. The vertical gray line marks recombination distance $r > 1/T$ from the selected locus, above which the dynamics subsequent to reach x should make little difference. The solid black lines give the prediction of (19).

We used two types of recurrent trajectory, the recurrent step and the recurrent top hat, as described earlier. For the recurrent top-hat trajectory, we simulated an exponential waiting time with mean ν between the end of one top hat and the start of the next (and similarly for the step case). In Figure 3 we show diversity levels moving away from the locus undergoing these two types of recurrent sweeps, as well as the analytical approximation given by Equation 19. Recall that in both types of trajectories the derived allele pauses at frequency x for time T, and therefore we expect that the fate of the allele will affect diversity at recombination distances $<1/T$. For distances $>1/T$, Equation 19 shows good agreement with our simulations, regardless of whether the recurrent sweeps go to loss or fixation. The approximation does not perfectly match our simulations, presumably because e^{-r2t_x} is an imperfect approximation to the probability of recombination during the sweep. Nevertheless, diversity levels generated by the two types of recurrent trajectory agree away from the selected site, which importantly confirms that only the initial rapid stage of the trajectory affects the coalescent process at partially linked sites.

The level of diversity under homogeneous sweeps: Under the model in which sweeps occur homogeneously along an infinite sequence, with coalescent rates given by Equation 16, the level of nucleotide diversity is given by

$$\mathbb{E}[\pi] = \frac{\theta}{2N\nu_{\mathrm{BP}}J_{2,2}/r_{\mathrm{BP}} + 1}. \tag{20}$$

These results generalize previous results by Kaplan *et al.* (1989) and Stephan *et al.* (1992), who found a relationship of the form (20) for a model of homogeneous recurrent full sweeps. In fact, since Equation 20 follows only from the assumption that the rate and characteristics of sweeps are independent of their location along the genome (see Equation 14), this relationship between diversity, the density of selective targets, and recombination rate will hold for a wide

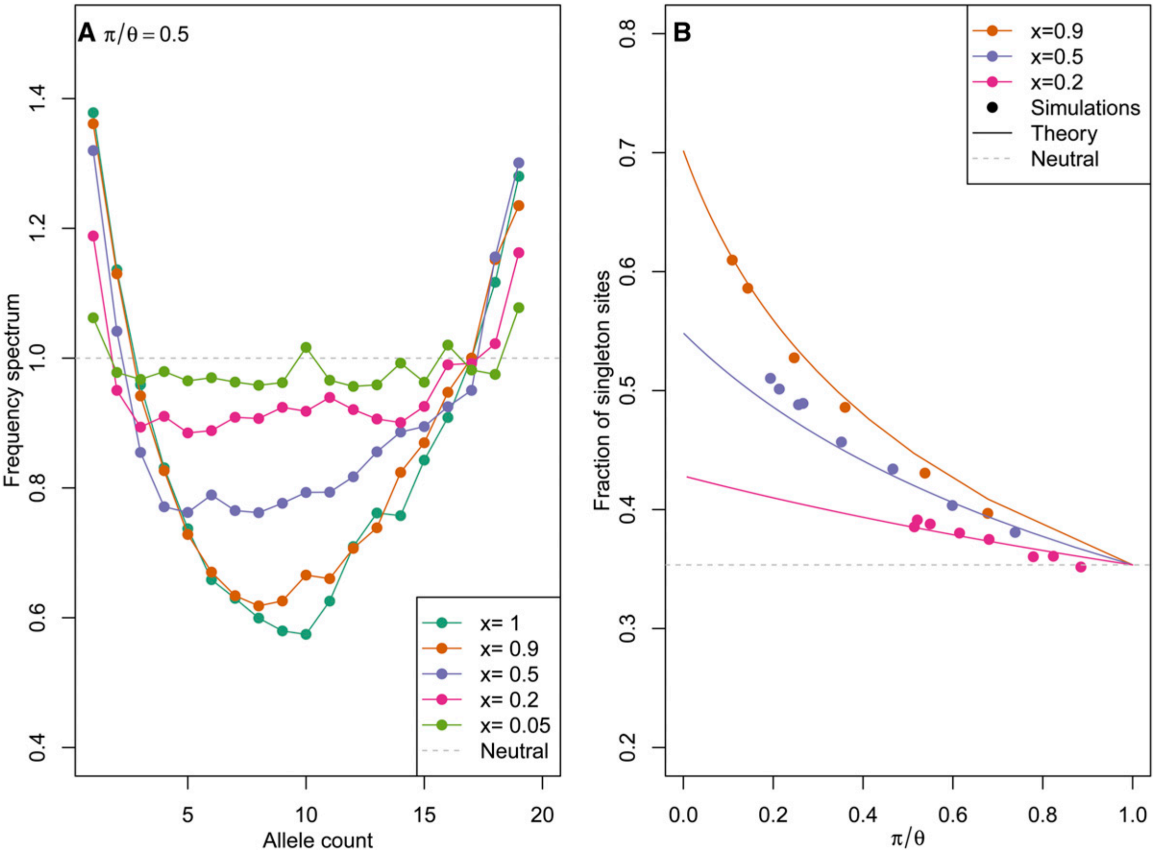

Figure 4 Properties of the frequency spectrum with sweeps occurring at a fixed genetic distance. Coalescent rates are given by Equation 9, with q fixed to $q_x = xe^{-t_x r}$ and $t_x r$ = 0.6, across a range of x. (A) The percentage of segregating sites found at frequency $1 \le k \le 20$, relative to the neutral expectation (*i.e.*, $F_{20,k}/F^{\mathrm{N}}_{20,k}$). In these simulations the rate of sweeps $N\nu$ has been fixed to result in a 50% reduction in diversity. The dashed gray line gives the neutral expectation. (B) The mean number of singletons divided by mean number of segregating sites, from *mssel* simulations with a sample size of 10 at a neutral site a distance $2Nr = 200$ from a selected site. The selected allele performs a recurrent top-hat trajectory (with N = 10,000 and $t_x/2N = 0.003$, giving $rt_x = 0.6$, and pausing $T/2N = 0.01$) to frequency $x = 0.2$, $x = 0.5$, or $x = 0.9$ across a range of $2N\nu$. Note the span of π/θ is smaller in the low-x simulations as the effect on diversity of a given $2N\nu$ is smaller. Solid lines show the analytical approximation for $\mathbb{E}[T_1]/\mathbb{E}[T_{\mathrm{tot}}]$ of *Appendix B*. The dashed gray line gives the neutral value of the expected proportion of singletons $1/\sum_{j=1}^{n-1} 1/j$.

variety of homogeneous recurrent-sweep models including the homogeneous full-sweep model.

Frequency spectrum: We now study the effects of recurrent partial sweeps on other properties of neutral diversity at a locus besides pairwise nucleotide diversity and compare our calculations to simulation.

Two commonly studied properties of a sample of neutral diversity at a locus are the expected number of segregating sites in a sample of size n and the expected number of singletons in a sample of size n. Under the infinite-sites assumption, these are respectively equal to the mutation rate multiplied by the expected total length of the genealogical tree of the sample (which we denote T_{tot}) and by the mutation rate multiplied by the expected total length of the terminal branches (T_1). We provide recursions that allow easy calculation of both $\mathbb{E}[T_{\mathrm{tot}}]$ and $\mathbb{E}[T_1]$ in *Appendix B*.

We also look more generally at the frequency spectrum of segregating alleles, which is, in a sample of n individuals, the proportion of segregating sites at which k derived alleles are found, for each $1 \le k \le n$. Let $F_{n,k}$ denote the expected proportion of segregating sites in a sample of size n at which exactly k samples carry the derived allele under an infinite-sites model of mutation. $F_{n,k}$ is equal to the expected time in the coalescent tree spent on branches that subtend exactly k tips (those on which mutation would lead to a site segregating at k of the n samples), divided by $\mathbb{E}[T_{\mathrm{tot}}]$. Under neutrality (Kingman's coalescent), this quantity is $F^{\mathrm{N}}_{n,k} = (1/k)/\sum_{j=1}^{n-1}(1/j)$. It is not so easy to find an explicit general expression under the coalescent model with sweeps that we study, but for the case $k = 1$ we have described in *Appendix B* how to compute $\mathbb{E}[T_1]/\mathbb{E}[T_{\mathrm{tot}}]$, and the general case can be found from simulation of the coalescent process.

Figure 4A shows the ratio of $F_{n,k}/F^{\mathrm{N}}_{n,k}$, estimated by direct simulation of our coalescent process. The rates are given by Equation 9, with q fixed to $q_x = xe^{-t_x r}$, and $t_x r = 0.6$ (and various x). To make the simulations comparable, the population-scaled rate of sweeps $2N\nu$ was adjusted such that $\pi/\theta = 1/2$ in each, *i.e.*, to obtain a 50% reduction in diversity due to sweeps. We see that for partial sweeps at a fixed site,

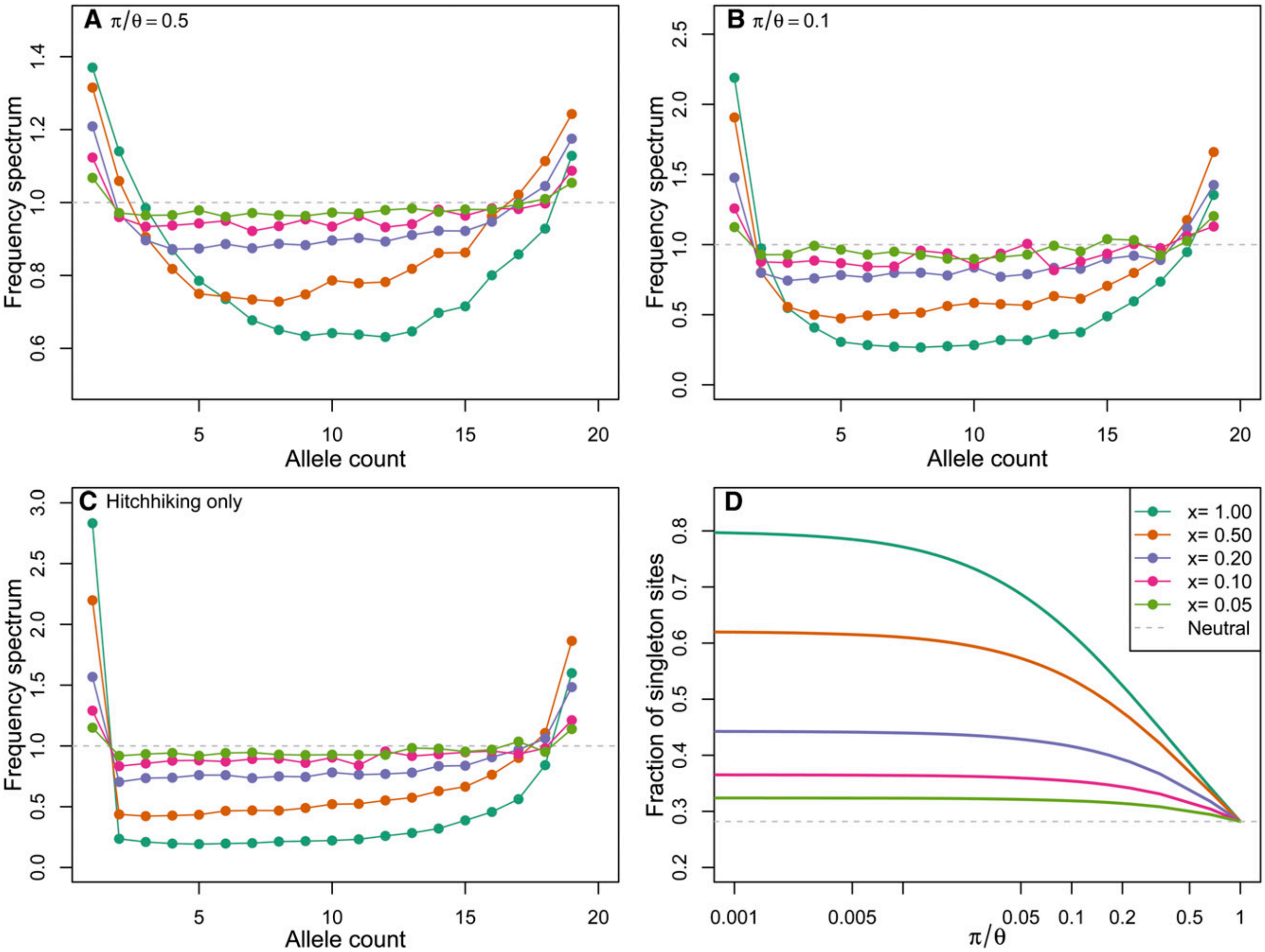

Figure 5 Properties of the frequency spectrum under a spatially homogeneous model of sweeps using the coalescent process with rates given by Equation 17. Simulations were performed for a sample size of 20. For a particular x we adjusted the value of $N\nu_{BP}/(tr_{BP})$ to achieve the specified reduction in π. (A and B) The percentage of segregating sites found at frequency $1 \leq k \leq 20$, relative to the neutral expectation for sweeps. In each panel the reduction in diversity, π/θ is fixed. (C) The same quantities as in A and B, but for the case where there is no genetic drift, and sweeps are the only stochastic force affecting allele frequencies. (D) The fraction of segregating sites that are singletons, for different x, as a function of π/θ, calculated using recursions for $\mathbb{E}[T_1]/\mathbb{E}[T_{\text{tot}}]$ (*Appendix B*).

across a range of x, the frequency spectrum is skewed toward rare alleles and away from intermediate-frequency alleles.

To test the degree to which our coalescent matches the full model, in Figure 4B we compare the mean proportion of singleton sites under our coalescent model to that found from simulation with *mssel*. We simulated a recurrent top-hat trajectory of the frequency at a selected locus as before and used this trajectory with *mssel* to simulate the neutral coalescent at a nonrecombining locus a distance r away from this selected locus. We used the three values $x = 0.9$, 0.5, and 0.2 for the intermediate frequency the allele reached and in each case varied the rate of sweeps, ν. Each combination of ν and x gives a point in Figure 4B, plotted at its mean reduction in diversity (π/θ) and the mean number of singletons divided by the mean number of segregating sites. These are compared to the analytical values of $\mathbb{E}[T_1]/\mathbb{E}[T_{\text{tot}}]$ computed using Equations B1 and B3, with coalescent rates given by Equation 9, using a constant $q = xe^{-rt_x}$ and Equation 20 to find the reduction π/θ. There is good agreement between the simulations and the analytical results, showing that our simplified process approximates the properties of the full coalescent process at a single site reasonably well.

Figure 4 studied the effect on the frequency spectrum of recurrent sweeps at a fixed distance from a neutral site; in Figure 5 we study the frequency spectrum under the coalescent process with sweeps occurring homogeneously along the genome. Figure 5, A and B, shows the same quantities as Figure 4A, for simulations of the homogeneous partial-sweep coalescent process with a fixed value of x, using rates given by Equation 17, and $2N\nu_{\text{BP}}/(tr_{\text{BP}})$ chosen so that π is 50% and 10% of its value under neutrality, respectively. In Figure 5C, there is no genetic drift and only sweeps force coalescence, *i.e.*, $N = \infty$, and so we do not need to specify $2N\nu_{\text{BP}}/(tr_{\text{BP}})$ as it acts only as a timescaling. In Figure 5D we show our analytic calculation of $\mathbb{E}[T_1]/\mathbb{E}[T_{\text{tot}}]$ as a function of the reduction in π caused by selective sweeps.

The skew in the frequency spectrum depends strongly on the frequency x reached by the selected allele. Sweeps to low frequencies lead to a much smaller distortion for the same reduction in π. Therefore, the strong relationship between the reduction in π and the skew in the frequency spectrum under a model of full sweeps is much weaker if the sweeps do not go to fixation.

Intriguingly, sweeps that go to intermediate frequency can lead to a greater proportion of high-frequency–derived alleles than under a full-sweep model. While a single, recent full sweep leads to high-frequency derived alleles through hitchhiking (Fay and Wu 2000), under a recurrent full-sweep model these alleles are then fixed in the population by subsequent sweeps and drift (Kim 2006) and therefore removed from the frequency spectrum. Further work would be needed to understand the intuition for the excess of high-frequency derived alleles under a recurrent partial-sweep model.

Summaries of the frequency spectrum: In Figures 4 and 5, we saw that regardless of whether sweeps occur at a fixed distance from our neutral site or homogeneously along the sequence, as we increase the rate of sweeps the frequency spectrum becomes further skewed toward rare derived alleles at the expense of intermediate-frequency alleles. Here we provide evidence that this will hold for any set of parameter values. Tajima's D and Fu and Li's D (Tajima 1989; Fu and Li 1993) are two common ways of detecting deviations away from the frequency spectrum expected under a neutral model with a constant population size. Negative values of Tajima's D can be thought of as indicating a deficit of intermediate-frequency alleles, and Fu and Li's D indicates an excess of singleton alleles. Durrett and Schweinsberg (2005) proved that in large samples, both of these summary statistics are negative under a multiple-mergers coalescent model of full sweeps, as long as λ_k, the total coalescent rate when there are k lineages, satisfies

$$\sum_{k=2}^{\infty}\left(\lambda_k - \binom{k}{2}\right)\frac{\log(k)}{k^2} < \infty. \tag{21}$$

See equation 4.5 in Durrett and Schweinsberg (2005). Informally, this condition requires that the total coalescent rate is not too much higher than the neutral coalescent rate when there are a large number of lineages. Their methods were not specific to their situation but hold for all multiple-merger coalescent models satisfying Equation 21. As above, we argued that a generalized-sweep model can be approximated by a multiple-merger coalescent, and therefore it seems that reasonable generalized-sweep models will, at least for large samples, have a frequency spectrum that is skewed toward singletons at the expense of intermediate-frequency alleles (a notable exception is the "low-frequency" limit we discuss below).

Limiting processes

Before we move to discuss the implications of these results for data analysis there are two limiting processes that merit our attention. The first is when the rate of sweeps is sufficiently high to dominate genetic drift as a source of stochasticity. The second limit results when sweeps very rarely achieve high frequency in the population, in which case the resulting coalescent model is identical to the standard neutral coalescent, despite the fact that much of the stochasticity may be driven by sweeps.

The rapid sweep limit: A surprising conclusion from the homogeneous model and Equation 16 is that if all coalescences come from "selective" events, then the frequency spectrum does not depend on the density of selective targets or on the recombination rate (although the number of segregating sites certainly does). This effect can be seen in Figure 5D as the fraction of singleton sites plateaus when the reduction in π is large; *i.e.*, when the population-scaled rate of sweeps per unit of recombination is high, $\nu_{\mathrm{BP}}/r_{\mathrm{BP}} \gg 1/2N$. The easiest way to see this is to take $N \to \infty$ while keeping the rate of sweeps and their trajectory dynamics fixed, so that in a sample of fixed size the coalescence rate from Equation 16 converges to $\lambda_{k,i} \to \nu_{\mathrm{BP}}/r_{\mathrm{BP}}J_{k,i}$, where $J_{k,i}$ does not depend on ν_{BP}, r_{BP}, or N. In this limit, ν_{BP} and r_{BP} affect the process only by a timescaling, do not affect the transition probabilities of Equation 11, and so do not affect the frequency spectrum. Diversity in this limit behaves as

$$\mathbb{E}[\pi] = \frac{2\mu r_{\mathrm{BP}}}{\nu_{\mathrm{BP}}J_{2,2}} \tag{22}$$

(assuming, as usual, that μ is sufficiently small); *i.e.*, nucleotide diversity increases linearly with the recombination rate, if neither ν_{BP} nor $J_{2,2}$ varies across recombination environments. Similar limits can also be derived by letting $N \to \infty$ under the more general coalescent process with rates given by Equation 7.

For this limit to be a reasonable approximation for a sample of size k in a population of size N, we need the rate of neutral coalescences to be much smaller than the rate of selective coalescences; *i.e.*, $\binom{k}{2} \ll N\nu_{\mathrm{BP}}/r_{\mathrm{BP}}\sum_{i=2}^{k}J_{k,i}$. In sufficiently large samples, $\binom{k}{2}$ will be large enough that the coalescence rate due to genetic drift will be appreciable, at least until the number of lineages surviving back in time declines. From a technical standpoint, this is related to the question of whether the coalescent process "comes down from infinity" (for a review see Berestycki 2009).

The low-frequency limit: As noted in our discussion of Figure 5, the frequency spectrum may be close to neutral in appearance even with large reductions in π if selected alleles sweep only to low frequency. In fact, by taking a limit (satisfying certain conditions) in which sweeps occur frequently, but each sweep has a small probability of causing coalescence, we can recover Kingman's coalescent.

We illustrate this limit by taking $\nu \to \infty$ and allowing $f(q)$ to depend on ν in such a way that as $\nu \to \infty$, $I_{k,\ell}/I_{k,2} \to 0$ for all $3 \leq \ell \leq k$, and that $\nu\ \ I_{k,2} \to \binom{k}{2}\gamma$, for some $0 < \gamma < \infty$. As shown in *Appendix C*, a sufficient condition for this is that $\lim_{\nu\to\infty}\nu\int_0^1 q^2 f(q)dq$ is finite. In this limiting case, the rate of coalescence is

$$\lambda_k = \binom{k}{2}\left(\gamma + \frac{1}{2N}\right), \tag{23}$$

so the limiting model behaves exactly as the standard neutral coalescent but with an effective population size of

$$N_e = \frac{2N}{2N\gamma + 1}. \tag{24}$$

Note that the limiting coalescent process does not satisfy condition (21) of Durrett and Schweinsberg (2005) and that Tajima's D and Fu and Li's H will have mean equal to zero at all sample sizes, as is natural since the limiting process is just the neutral (Kingman's) coalescent.

In the case of our simple partial-sweep coalescent this limit would occur if the frequency x reached by sweeps is taken to zero as the rate of sweeps grows at least as $1/x^2$. The simple homogeneous full-sweep coalescent process obviously cannot be taken to this limit as there is a proscribed set of $J_{k,\bullet}$, which features a nontrivial amount of coalescence involving more than pairs of lineages.

Interference: In both limits discussed above the population-scaled rate of sweeps has to be very high. In the first limit the rate of sweeps has to be high enough to dominate the rate of neutral coalescence, and in the second limit the rate of sweeps has to be high enough to compensate for the fact that any one sweep is very unlikely to cause coalescence. The requirement of a high rate of sweeps implies that interference between the sweeps may occur, thus violating our assumption that the sweeps are independent. Investigations of the effect of such interference on the signal of hitchhiking have shown that interference reduces the impact of any one sweep on patterns of polymorphism (Kim and Stephan 2003; Chevin *et al.* 2008), although to interfere, the sweeps must begin at very similar times at loci separated by a low recombination rate. This suggests that a very high rate of sweeps is needed indeed before interference will have an appreciable impact on the hitchhiking effect, as would occur in the homogeneous sweep model if ν_{BP}/r_{BP} is very large. The limits we describe above require only that the population-size–scaled rate of sweeps ($N\nu$ or $N\nu_{BP}$) be high, and therefore it is possible to keep the *per generation* rate of sweeps sufficiently low to avoid the effect of interference. Further work is needed to investigate coalescent models under such high rates of sweeps and could be useful in understanding genealogical processes in organisms with low or no recombination that also experience strong selection pressures.

Discussion

The prevailing view of adaptation in a population genetics setting is based on a lone selected allele racing from its introduction into the population to fixation, carrying with it a chunk of the chromosome on which it arose. This cartoon has been a very useful prop for developing tests to identify genes underlying recent adaptations and for interpreting genome-wide patterns of polymorphism. However, it seems likely that such full sweeps constitute only a small proportion of the selected loci whose frequency changes in response to adaptation (see Pritchard *et al.* 2010, for a recent discussion). If we are to develop a better understanding of the full impact of linked selection on patterns of diversity, we need to develop a richer and more flexible set of models.

The work in this article was motivated by models in which the external environment or the genetic background varies on a fast enough timescale that new alleles rarely reach fixation before selective pressures change, either slowing their advance or reversing their trajectory. We laid out an approximation to the coalescent process under such a model and showed that, while the initial rapid stage of the trajectory will strongly affect the coalescent process, subsequent slower dynamics of the selected alleles have a much smaller effect. We then extended this idea to a recurrent-sweep model, approximating the dynamics by a multiple-merger coalescent. While some of our results are fairly general, to provide a more intuitive sense we have often employed simple allele-frequency trajectories and made other approximations. Nonetheless, we expect more realistic models to give rise to qualitatively equivalent results.

Each sweep we consider consists of a single allele at a locus rising on a single haplotype from very low frequency into the population. This contrasts with many other soft-sweep models, under which a sweep starts on multiple haplotypes, because multiple different alleles initially segregated at the locus (Hermisson and Pennings 2005), or as a result of multiple mutations occurring after selection pressures switched (Pennings and Hermisson 2006a,b; Ralph and Coop 2010), or because the adaptive allele was previously neutral and present on multiple haplotypes (Innan and Kim 2004; Przeworski *et al.* 2005). It is likely that recurrent models of such soft sweeps could be approximated through coalescent models with simultaneous multiple collisions (Schweinsberg 2000), to model the simultaneous rise of multiple haplotypes. This seems to be a fruitful area of future work as it would substantially extend our understanding of the effects of a broad family of recurrent-sweep models on genomic patterns of diversity.

We have also ignored the effect of background selection. To a first approximation, the effect of background selection can be modeled as an increase in coalescence rate, which would be a minor modification to Equations 9 and 16. This would alter the predicted relationship between diversity and recombination (Innan and Stephan 2003) given by Equation 20 and would offer a simple way to model the genealogical effects of both general models of hitchhiking and background selection.

Table 1 Estimates of sweep parameters from the relationship between diversity and recombination

	θ	$2N\nu_{\mathrm{BP}}J_{2,2}$	ν_{BP} across a range of x			
			$x = 1.0$	$x = 0.5$	$x = 0.2$	$x = 0.05$
Human	0.0017	6×10^{-11}	3.0×10^{-12}	1.2×10^{-11}	7.5×10^{-11}	1.2×10^{-9}
D. mel	0.025	7.3×10^{-9}	3.6×10^{-12}	1.5×10^{-11}	9.1×10^{-11}	1.5×10^{-9}

The estimate for humans was taken from Hellmann *et al.* (2008), who fitted a curve of the form of Equation 20. The estimate from *Drosophila melanogaster* (*D. mel*) was obtained from fitting Equation 20 to the synonymous polymorphism and sex-averaged recombination rates of Shapiro *et al.* (2007) (kindly provided by Peter Andolfatto; see Sella *et al.* 2009 for details), using nonlinear least squares via the nls() function in R. These estimates were converted into estimates of the rate of sweeps per generation per base pair (ν_{BP}, last four columns) under the simple partial-sweep trajectory model where $J_{2,2} = x^2/t_x$, assuming $t_x = 1000$ generations (equivalent to a selection coefficient of ~0.01) and that $N = 10^6$ in *D. mel* and $N = 10^4$ in humans.

The interpretation of population genomic patterns

Models in which selective sweeps do not always sweep to fixation have a much wider spectrum of predictions than the recurrent full-sweep model. Three broad correlations that have been used to argue for the prevalence of linked selection and used to potentially discriminate between models invoking background selection or full sweeps are (1) correlations between neutral diversity and the recombination rate, (2) correlations between the frequency spectrum and the rate of recombination, and (3) correlations between putatively adaptive divergence and neutral diversity. We now describe some of the implications of our results for understanding these patterns in population genomic data.

Correlation between recombination and diversity

One of the earliest and most compelling pieces of evidence for the role of linked selection in the fate of neutral alleles is a positive correlation between recombination and levels of diversity at putatively neutral sites (factoring out substitution rates as a proxy for differences in mutation rate). This pattern is consistent with both full sweeps and background selection, as both predict positive, albeit differently shaped, relationships (Innan and Stephan 2003). The shape of the diversity–recombination curve under a homogeneous rate of partial sweeps is identical to that of the full-sweep model and more generally similar for a broad class of homogeneous sweep models. In fact, the relationship under a homogeneous model depends only on $2N\nu_{\mathrm{BP}}J_{2,2}$, as seen in Equation 20.

To illustrate this point, in Table 1 we present estimates of $2N\nu_{\mathrm{BP}}J_{2,2}$ for humans and *Drosophila melanogaster*, assuming a model with drift and a homogeneous rate of selective sweeps across the genome, and from Equation 20 and data from Shapiro *et al.* (2007) and Hellmann *et al.* (2008). Along with these estimates, Table 1 also shows the implied rate of sweeps per generation per base pair, ν_{BP}, under the simple partial-sweep model, for a variety of values of x. These rates are surely overestimates and are intended for illustrative purposes only, as they ignore the effect of other forms of linked selection, *e.g.*, background selection.

The strength of the relationship between diversity levels and recombination varies dramatically between the two species, as indicated by the very different estimates of $2N\nu_{\mathrm{BP}}J_{2,2}$ (note that the estimates of ν_{BP} are similar due to the 100-fold difference in N). In *Drosophila* the positive relationship between recombination and diversity is strong (*e.g.*, Aguade *et al.* 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Begun *et al.* 2007; Shapiro *et al.* 2007), but in humans the relationship seems to be weaker and is and complicated by other confounding factors (Payseur and Nachman 2002; Hellmann *et al.* 2003, 2005, 2008; Cai *et al.* 2009). However, we should be cautious in the biological interpretation of this difference, as in humans diversity is usually estimated in large windows (much of which will be noncoding and far from genes), while in *Drosophila* neutral diversity levels are usually estimated from synonymous sites in individual genes. What is needed is a comparative analysis that studies these patterns at the same genomic scale and accounts for the profound differences in the density of functional targets among species.

The fact that the diversity–recombination curve plateaus rapidly in humans is strong evidence that linked selection does not affect the average neutral site in regions of high recombination. Technically, this could also occur if the density of selective targets ν_{BP} decreases approximately linearly with recombination rate; however, this option does not seem likely *a priori*.

Although in *D. melanogaster* this curve is still concave, it does not appear to flatten completely in high-recombination regions (*e.g.*, Sella *et al.* 2009), suggesting that linked selection is an important source of stochasticity even in these regions. At face value the concave nature of the curve suggests that both genetic drift and linked selection contribute to stochasticity, as $N\nu_{\mathrm{BP}} \gg r_{\mathrm{BP}}$ would lead to an almost linear relationship across the observed range of recombination rates (see Equation 22). However, a model with effectively no genetic drift can produce a concave curve and fit the observed data if $\nu_{\mathrm{BP}}J_{2,2}$ is not constant across recombination environments, *e.g.*, if sweeps occur at a moderately higher rate or achieve higher frequency in high-recombination regions. Neither of these two options seems particularly unlikely, suggesting that we still have little unambiguous evidence favoring genetic drift as an important source of stochasticity in *Drosophila*.

The frequency spectrum

The recurrent full-sweep model predicts a strong positive relationship between the reduction in neutral diversity and the skew toward rare alleles (Braverman *et al.* 1995; Kim

2006), a pattern not predicted under models of strong background selection. This relationship has been used to test between full sweeps and background selection models, although note that as we discussed in *Limiting processes*, this relationship is not expected if all coalescence comes from selective sweeps. Under our simple trajectory model, the distortion of the frequency spectrum is primarily determined by the frequencies that sweeps achieve. Therefore, although a lack of a strong skew in the frequency spectrum is consistent with a low rate of full sweeps, it cannot be used to rule out a high rate of partial sweeps. A lack of a genomic relationship between the frequency spectrum and recombination rate is therefore not grounds for rejecting sweeps as a force in shaping genetic diversity in favor of a model of background selection. Our results suggest that recurrent partial sweeps to low frequency in regions of high recombination in *D. melanogaster* and in the low-recombination regions in humans may be a major source of stochasticity in allele frequencies.

Correlation between divergence and polymorphism

Attention has recently focused on examining the correlation between neutral diversity and amino acid substitutions (or other putatively functional changes) between recently separated species. If a reasonable fraction of amino acid substitutions are driven by new mutations sweeping to fixation, then levels of diversity should dip on average around amino acid substitutions. This relationship has been tested for by looking for a positive correlation between diversity levels and amino acid substitution rates (Andolfatto 2007; Macpherson *et al.* 2007; Cai *et al.* 2009; Haddrill *et al.* 2011) or for a dip in diversity levels around a large set of aggregated amino acid substitutions (Hernandez *et al.* 2011; Sattath *et al.* 2011). If the density of functional sites is properly controlled for, these types of correlations between amino acid substitutions and neutral diversity are not expected under a (simple) model of background selection. Such correlations have been detected in *Drosophila* (Macpherson *et al.* 2007; Sattath *et al.* 2011) but in humans the dip in diversity around nonsynonymous substitutions seems to result from the dip in diversity levels around genes, an observation that seems inconsistent with a high rate of strong full sweeps (Hernandez *et al.* 2011). Similarly, it has been observed that the highest F_{ST} signals between human populations are not associated with strongly reduced haplotypic diversity (Coop *et al.* 2009).

The fact that selected alleles in the partial-sweep coalescent model do not have to sweep all the way to fixation partially decouples the rate of fixation of adaptive alleles from their effects on patterns of diversity within populations. Therefore, the strength of the positive relationship between substitution rates and diversity depends on the fate of alleles that sweep into the population. For example, this positive relationship may be weak and a poor predictor of the total reduction in diversity, if the majority of adaptive alleles that initially sweep into the population are eventually lost (*e.g.*, as can be the case for major-effect alleles in polygenic models of adaptation, see Lande 1983; Chevin and Hospital 2008).

Concluding thoughts

In this article, we have concerned ourselves with patterns of diversity at a single neutral site. However, partial sweeps also have a strong effect on linkage disequilibrium and haplotype diversity, a signature that has been exploited in scans for selection (*e.g.*, Hudson *et al.* 1994; Sabeti *et al.* 2002; Voight *et al.* 2006). One simple case that we can immediately describe is the low q limit (*Limiting processes*). In that limit, the coalescent is equivalent to the standard neutral model and as such the decay of linkage disequilibrium will be the same as in the standard neutral model with an N_e given by Equation 24. A natural way to extend this exploration would be the genealogical framework developed by McVean (2007) that has recently been extended to a multiple-mergers coalescent by Eldon and Wakeley (2008).

We will soon have polymorphism data across a broad range of taxa that will differ dramatically in selection regimes, recombination rates, genome size, and population size, allowing a much fuller picture of how these various factors interplay to shape genome-wide levels of polymorphism. The results presented here, however, suggest that we will continue to struggle to distinguish between modes of selection, as relaxing the assumptions of various models can generate a broad range of overlapping predictions.

Despite that, our results suggest a promising way forward, since a broad range of sweep models can be captured by simple parameterizations of multiple-merger coalescence processes. Importantly, this would allow parameter inference under a general model of linked selection, rather than focusing on a limited number of specific models. For example, we could estimate the rate that selection forces different numbers of lineages to coalesce [parameterized by $\nu f(q)$] as a function of recombination rates and the density of selective targets. As the multiple-mergers coalescent model is easily simulated under, it may be readily incorporated into many of our existing genealogical inference frameworks. It is likely that parameters of such models could be estimated very precisely from genome-wide data, allowing us to concentrate on what these high-level summaries of polymorphism tell us about linked selection across genomic environments and species. Such inferences may be important if we wish to move beyond documenting the presence of linked selection toward describing the genealogical process in species where selection is a major source of stochasticity.

Acknowledgments

We thank Yaniv Brandvain, Chuck Langley, Molly Przeworski, Josh Schraiber, Alisa Sedghifar, and Guy Sella for helpful conversations and comments on previous drafts of this article. We thank the two anonymous reviewers and the

Editor for helpful feedback. This work is supported by a Sloan Fellowship and funds from University of California at Davis (to G.C.) and a National Institutes of Health National Research Service Award postdoctoral fellowship (to P.R.).

Literature Cited

Aguade, M., N. Miyashita, and C. H. Langley, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster.* Genetics 122: 607–615.

Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 17: 1755–1762.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster.* Genetics 158: 657–665.

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

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355: 1553–1562.

Barton, N. H., and A. M. Etheridge, 2004 The effect of selection on genealogies. Genetics 166: 1115–1131.

Barton, N. H., A. M. Etheridge, and A. K. Sturm, 2004 Coalescence in a random background. Ann. Appl. Probab. 14(2): 754–785.

Begun, D., A. Holloway, K. Stevens, L. Hillier, Y. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5: e310.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* Nature 356: 519–520.

Berestycki, N., 2009 Recent progress in coalescent theory, pp. 1–193 in *Ensaios Matematicos*, Vol. 16. arXiv:0909.3985.

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

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

Cai, J. J., J. M. Macpherson, G. Sella, and D. A. Petrov, 2009 Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 5: e1000336.

Charlesworth, B., 2009 Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10: 195–205.

Charlesworth, D., B. Charlesworth, and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632.

Chevin, L. M., and F. Hospital, 2008 Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645–1660.

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

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li *et al.*, 2009 The role of geography in human adaptation. PLoS Genet. 5: e1000500.

Cutter, A. D., and J. Y. Choi, 2010 Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. Genome Res. 20: 1103–1111.

Cutter, A. D., and A. M. Moses, 2011 Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. Mol. Biol. Evol. 28: 1745–1754.

Cutter, A. D., and B. A. Payseur, 2003 Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol. Biol. Evol. 20: 665–673.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138.

Durrett, R., and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stoch. Proc. Appl. 115: 1628–1657.

Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178: 1517–1532.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger, 2006 An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. 16: 685–729.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3: 87–112.

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

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Gillespie, J., 1994 Alternatives to the neutral theory, pp. 1–17 in *Non-Neutral Evolution. Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, London/New York.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene 205: 291–299.

Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155: 909–919.

Gordo, I., A. Navarro, and B. Charlesworth, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics 161: 835–848.

Haddrill, P. R., K. Zeng, and B. Charlesworth, 2011 Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. Mol. Biol. Evol. 28: 1731–1743.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. 72: 1527–1535.

Hellmann, I., K. Prufer, H. Ji, M. C. Zody, S. Paabo *et al.*, 2005 Why do human diversity levels vary at a megabase scale? Genome Res. 15: 1222–1231.

Hellmann, I., Y. Mang, Z. Gu, F. Li, P. de la Vega *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 18: 1020–1029.

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

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.

Hey, J., 1991 A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. 39: 30–48.

Hudson, R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster.* Genetics 136: 1329–1340.

Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics 120: 831–840.

Hudson, R. R., and N. L. Kaplan, 1995a Deleterious background selection with recombination. Genetics 141: 1605–1617.

Hudson, R. R., and N. L. Kaplan, 1995b The coalescent process and background selection. Philos. Trans. R. Soc. Lond. B Biol. Sci. 349: 19–23.

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

Innan, H., and W. Stephan, 2003 Distinguishing the hitchhiking and background selection models. Genetics 165: 2307–2312.

Kaplan, N. L., T. Darden, and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics 120: 819–829.

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

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

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

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

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kimura, M., and J. F. Crow, 1964 The number of alleles that can be maintained in a finite population. Genetics 49: 725–738.

Kimura, M., and T. Ohta, 1971 Protein polymorphism as a phase of molecular evolution. Nature 229: 467–469.

Kopp, M., and J. Hermisson, 2007 Adaptation of a quantitative trait to a moving optimum. Genetics 176(1): 715–719.

Kopp, M., and J. Hermisson, 2009a The genetic basis of phenotypic adaptation I: fixation of beneficial mutations in the moving optimum model. Genetics 182(1): 233–249.

Kopp, M., and J. Hermisson, 2009b The genetic basis of phenotypic adaptation II: the distribution of adaptive substitutions in the moving optimum model. Genetics 183(4): 1453–1476.

Lande, R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity 50: 47–65.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lohmueller, K., A. Albrechtsen, Y. Li, S. Kim, T. Korneliussen *et al.*, 2011 Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7(10): e1002326.

Macpherson, J. M., G. Sella, J. C. Davis, and D. A. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics 177: 2083–2099.

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

McVean, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics 175: 1395–1406.

McVicker, G., D. Gordon, C. Davis, and P. Green, 2009 Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5: e1000471.

Möhle, M., and S. S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Ann. Appl. Probab. 29: 1547–1562.

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

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

Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 The effect of recombination on background selection. Genet. Res. 67: 159–174.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3: e196.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.

Payseur, B. A., and M. W. Nachman, 2002 Natural selection at linked sites in humans. Gene 300: 31–42.

Pennings, P., and J. Hermisson, 2006a Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23: 1076–1084.

Pennings, P. S., and J. Hermisson, 2006b Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2: e186.

Pfaffelhuber, P., B. Haubold, and A. Wakolbinger, 2006 Approximate genealogies under genetic hitchhiking. Genetics 174: 1995–2008.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.

Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20: R208–R215.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.

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

Ralph, P., and G. Coop, 2010 Parallel adaptation: One or many waves of advance of an advantageous allele? Genetics 186: 647–668.

Robertson, A., 1961 Inbreeding in artificial selection programmes. Genet. Res. 2: 189–194.

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

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125.

Santiago, E., and A. Caballero, 1995 Effective size of populations under selection. Genetics 139: 1013–1030.

Santiago, E., and A. Caballero, 1998 Effective size and polymorphism of linked neutral loci in populations under directional selection. Genetics 149: 2105–2117.

Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor. Popul. Biol. 74: 104–114.

Sattath, S., E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella, 2011 Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLoS Genet. 7: e1001302.

Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5: 1–50.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5: e1000495.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. Proc. Natl. Acad. Sci. USA 104: 2271–2276.

Stephan, W., T. Wiehe, and M. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. 41: 237–254.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

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

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

Wright, S. I., J. P. Foxe, L. DeRose-Wilson, A. Kawabe, M. Looseley *et al.*, 2006 Testing for effects of recombination rate on nucleotide diversity in natural populations of *Arabidopsis lyrata*. Genetics 174: 1421–1430.

Communicating editor: J. Wakeley

Appendices

Appendix A: $J_{k,i}$ for a Generalized Trajectory

Recall that we defined in Equation 13

$$J_{k,i} = \binom{k}{i}\mathbb{E}_X\left[\int_0^\infty q(X,r)^i(1-q(X,r))^{k-i}dr\right], \quad 2 \le i \le k, \tag{A1}$$

so that the rate at which the coalescent process having k lineages coalesces down to i lineages from selective events is $\nu_{\mathrm{BP}}/r_{\mathrm{BP}}J_{k,i.}$ The quantity $q(X, r)$ is the pathwise Laplace transform of the process X, defined in Equation 3, and consequently

$$1-q(X,r) = \int_0^\infty re^{-rt}(1-X(t))dt. \tag{A2}$$

It is useful to note that by changing the order of integration,

$$\begin{aligned} J_{k,i} &= \binom{k}{i}\mathbb{E}_X\left[\int_0^\infty\left(\int_0^\infty\cdots\int_0^\infty\prod_{j=1}^{i}X(t_j)\prod_{\ell=i+1}^{k}(1-X(t_\ell))r^k\exp\left(-r\sum_{j=1}^{k}t_j\right)dt_1\cdots dt_k\right)dr\right] \\ &= k!\binom{k}{i}\mathbb{E}_X\left[\int_0^\infty\cdots\int_0^\infty\frac{\prod_{j=1}^{i}X(t_j)\prod_{j=i+1}^{k}(1-X(t_j))}{\left(\sum_{j=1}^{n}t_j\right)^{k+1}}dt_1\cdots dt_k\right] \end{aligned} \tag{A3}$$

for $2 \le i \le k$, as long as the integral is finite. In the case of a pair of lineages $i = 2$ and this simplifies to

$$J_{2,2} = 2\mathbb{E}_X\left[\int_0^\infty\int_0^\infty\frac{X(\tau-t_1)X(\tau-t_2)}{(t_1+t_2)^3}dt_1dt_2.\right] \tag{A4}$$

To briefly explore the conditions for J to be finite, we suppose that X leaves zero as a power of t; *i.e.*, $X(t) \sim t^\alpha$ for some $\alpha > 0$, for small t. We see that $J_{k,2}$ increases as α increases; *i.e.*, the rate of sweeps is larger the more rapidly X leaves zero. In this case, $q(r) \sim Cr^{-\alpha}$ for large r, where C is a constant. Then since

$$\begin{aligned} J_{k,2} &= \lim_{L\to\infty}\binom{k}{2}\int_0^L q(r)^2(1-q(r))^{k-2}dr \\ &\le \lim_{L\to\infty}\binom{k}{2}\int_0^L q(r)^2dr, \end{aligned}$$

it can be seen that $J_{k,2}$ is *infinite* if $\alpha \le 1/2$, in the limit of an infinite genome. More generally, if X leaves zero more quickly than $\sqrt{t}$ (which may be biologically unrealistic), then sweeps occurring arbitrarily far away along the genome will cause coalescences.

Appendix B: Recursions to Find $\mathbb{E}[T_{\text{tot}}]$ and $\mathbb{E}[T_1]$

Two properties of interest are the expected total amount of time in the genealogy at a neutral locus ($\mathbb{E}[T_{\text{tot}}]$) and the expected total amount of time in terminal branches ($\mathbb{E}[T_1]$).

We first derive the expected total time in the genealogy. Recall that if the coalescent process has k lineages, then it waits an exponentially distributed amount of time with mean $1/\lambda_k$ and then jumps to a smaller number of lineages chosen with probabilities according to $p_{k,\ell}$, with λ_k and $p_{k,\ell}$ given in Equations 10 and 11. Therefore, if we let $G_{n,k}$ be the probability that the coalescent process that starts from n lineages ever visits the state with k lineages, then

$$\mathbb{E}[T_{\text{tot}}] = \sum_{k=2}^{n} \frac{k}{\lambda_k} G_{n,k}. \tag{B1}$$

By conditioning on the last state visited before dropping to k lineages, we can see that $G_{n,k}$ satisfies the recursion

$$G_{n,k} = \sum_{i=k+1}^{n} G_{n,i}\, p_{i,k}, \quad \text{for} \quad k < n, \tag{B2}$$

with $G_{n,n} = 1$. This recursion is of upper triangular form, so is easily solvable, which together with (B1) allows us to compute $\mathbb{E}[T_{\text{tot}}]$.

We now turn to the expected total time in terminal branches, *i.e.*, those branches on which mutations will lead to singletons. Note that, since all lineages are exchangeable, $\mathbb{E}[T_1] = n$ times the mean time until a particular lineage—say, the first one—coalesces with any other. To find this, let $S_{n,k}$ be the probability that at some point there are k lineages and that one of those k lineages is the original first lineage, still not coalesced with any others. Then the mean time until the first lineage coalesces is $\sum_{k=2}^{n}(1/\lambda_k)S_{n,k}$, and hence

$$\mathbb{E}[T_1] = n \sum_{k=2}^{n} \frac{1}{\lambda_k} S_{n,k}. \tag{B3}$$

As above, we can get a solvable recursion for $S_{n,k}$ by conditioning on the last coalescent event before reaching k lineages. If the coalescent process jumps from ℓ to k lineages, then the probability that a given lineage is not part of this coalescent event is $(k-1)/\ell$, and hence

$$S_{n,k} = \sum_{\ell=k+1}^{n} S_{n,\ell} p_{\ell,k} \frac{k-1}{\ell} \quad \text{for} \quad k < n, \tag{B4}$$

and $S_{n,n} = 1$. The recursion is also easily solvable, which lets us obtain $\mathbb{E}[T_1]$.

Appendix C: More on the Low q Limit

We want to arrange things so that asymptotically, all coalescent events affect only two lineages. We illustrate this limit by taking $\nu \to \infty$ and allowing $f(q)$ to depend on ν in such a way that as $\nu \to \infty$, $I_{k,\ell}/I_{k,2} \to 0$ for all $3 \le \ell \le k$, and that $\nu\ I_{k,2} \to \binom{k}{2}\gamma$, for some $0 < \gamma < \infty$. Since this model is a Lambda coalescent with $\Lambda(dq) = q^2 \nu f(q)dq + \delta_0(dq)/2N$, if we rescale time by a factor of C, a necessary and sufficient condition is that $C\Lambda$ converges weakly to a point mass at 0.

To emphasize the dependence of f on ν we write $f(q) = f_\nu(q)$ and $I_{k,\ell} = I_{k,\ell}(\nu)$. We want to find a simple condition under which the proportion of coalescences involving more than two lineages goes to zero, *i.e.*, that $I_{k,\ell}(\nu)/I_{k,2}(\nu) \to 0$ as $\nu \to \infty$ if $\ell > 2$. Fix k, and suppose for convenience that $f(q) = 0$ for all $q > 1 - \varepsilon$, for some $\varepsilon > 0$. Then

$$\epsilon^k \int_0^1 q^\ell f_\nu(q)dq < \int_0^1 q^\ell (1-q)^{k-\ell} f_\nu(q)dq < \int_0^1 q^\ell f_\nu(q)dq,$$

so that $I_{k,\ell}(\nu)/I_{k,2}(\nu) \to 0$ if and only if

$$\frac{\int_0^1 q^\ell f_\nu(q)dq}{\int_0^1 q^2 f_\nu(q)dq} \to 0.$$

Using Jensen's inequality,

$$\frac{\int_0^1 q^\ell f_\nu(q)dq}{\int_0^1 q^2 f_\nu(q)dq} \le \frac{\left(\int_0^1 q^2 f_\nu(q)dq\right)^{\ell/2}}{\int_0^1 q^2 f_\nu(q)dq}$$
$$= \left(\int_0^1 q^2 f_\nu(q)dq\right)^{(\ell-2)/2},$$

so if $\int_0^1 q^2 f_\nu(q)dq \to 0$, this will be achieved. By the same result,

$$\frac{I_{k,2}(\nu)}{\nu\binom{k}{2}\int_0^1 q^2 f_\nu(q)dq} \to 1,$$

so that, rescaling time by a factor C_ν, if

$$\nu C_\nu \int_0^1 q^2 f_\nu(q)dq \to \gamma \quad \text{as} \quad L \to \infty,$$

then $\nu C_\nu I_{k,2} \to \binom{k}{2}\gamma$ for all k. In this limit, the rate at which a pair of lineages coalesces converges and does not depend on the number of lineages present.

Ideally, we would illustrate this with a stochastic model for X. However, the formula requires the model to be analytically tractable to a degree satisfied by no population genetics models that we could think of, and it is much easier to make a concrete choice of $f(q)$. Consider the case where $f(q)$ is the density of a Beta(1, M) distribution. The mean of this distribution is $1/(1 + M)$. In that case

$$I_{k,\ell} = \binom{k}{\ell}\int_0^1 q^\ell (1-q)^{k-\ell+M-1}\, M dq = M\binom{k}{\ell}\Big/\binom{k+M-1}{\ell}, \tag{C1}$$

so that as $M \to \infty$,

$$MI_{k,2} = \binom{k}{2}\frac{2M^2}{(M+k-1)(M+k-2)} \xrightarrow{L\to\infty} 2\binom{k}{2},$$

so if $\nu = M$, then $\gamma = 2$. We can furthermore check that

$$\frac{I_{k,\ell}}{I_{k,2}} = \frac{\binom{k}{\ell}}{\binom{k}{2}} \frac{\ell!(k+M-\ell-1)!}{2!(k+M-3)!} \sim \frac{1}{M^{\ell-2}} \xrightarrow{M\to\infty} 0 \tag{C2}$$

so that this simple case satisfies our limit.

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2012/06/19/genetics.112.141861.DC1

Patterns of Neutral Diversity Under General Models of Selective Sweeps

Graham Coop and Peter Ralph

DOI: 10.1534/genetics.112.141861

File S1

R Scripts

File S1 is available for download at http://www.genetics.org/content/suppl/2012/06/19/genetics.112.141861.DC1 as a compressed file.