Theory of Sweeps from Standing Variation

Jeremy J. Berg[1,2,3] and Graham Coop[1,2,3]
[1] Graduate Group in Population Biology, University of California, Davis.
[2] Center for Population Biology, University of California, Davis.
[3] Department of Evolution and Ecology, University of California, Davis
To whom correspondence should be addressed: jjberg@ucdavis.edu, gmcoop@ucdavis.edu

Abstract

1 Introduction

Understanding the major processes which shape genetic variation within species and the signatures they leave behind in observable data (e.g. sequence data) is a major goal of evolutionary biology. One major process of interest is that of positive directional selection, whereby mutations are driven to high frequency, and perhaps to fixation, as a result of differential reproductive success of carriers relative to non-carriers. Alleles with selection coefficients strong enough that their rate of frequency change exceeds the typical rate of change under drift leave behind a characteristic pattern in DNA sequence data. First described by **?**, this "selective sweep" pattern can in some cases allow researchers to identify the individual regions of the genome that have been recent targets of positive selection, thus yielding insight into the genetic details of the adaptive process.

By now, the original model first studied by **?**, in which a single co-dominant mutation arises and is driven swiftly to fixation, is well studied *(cite a whole lot of people)* . but recent work has re-emphasized the fact that likely not all adaptation occurs via this simple model, and different models predict different diagnostic signatures (**??????**). A large class of these models, in there may be one or many beneficial mutations present on multiple haplotypes, go by the name of "soft sweeps"

Now commonly recognized phenomenon resulting from strongish natural selection is the selective sweep, which leaves characteristic signature in DNA data.

Historical focus on "hard sweep" model, (one allele, one haplotype). Widely varying opinions on whether we've found more or fewer of these than we expected to.

More recent interest in alternative modes of adaptation, and whether they might be detectable, e.g. multiple mutation sweeps, sweeps from standing variation.

Here we take a look at standing sweeps and the patterns they generate.

2 Results

2.1 Non-technical description

2.2 more technical

We consider two loci, labeled $\mathcal{N}$ and $\mathcal{B}$, which are separated on the chromosome by a recombination distance r. The ancestral allele at the $\mathcal{B}$ locus is b, and we imagine that a new allele, B, arises at the $\mathcal{B}$ locus and segregates at low frequency for some period of time (either due to neutral fluctuations, or because it is balanced at low frequency), before a change in the environment causes it to become beneficial and sweep to fixation. All variation at the $\mathcal{N}$ locus is assumed to be neutral, and its history can be therefore be determined by considering a structured neutral coalescent conditional on the behavior of the $\mathcal{B}$ locus.

Our aim is to describe some features of the genealogy at the $\mathcal{N}$ locus, and to use this understanding of the genealogies to build intuition regarding the process of a sweep from standing variation, as well

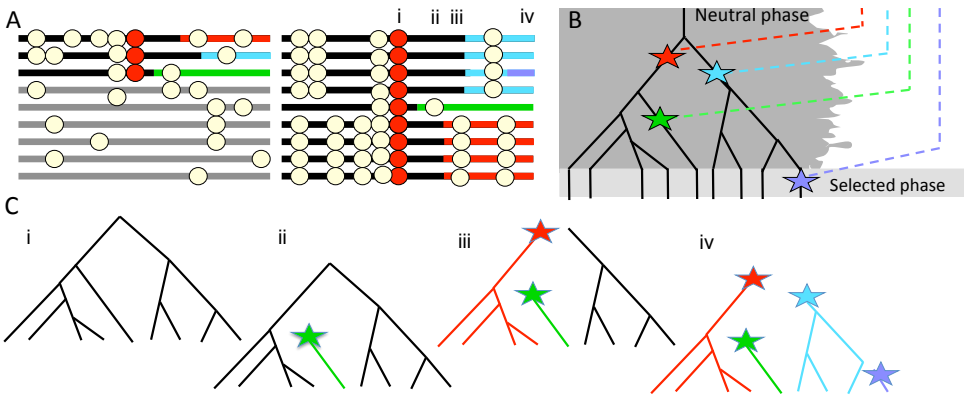


Figure 1: caption goes here

as to derive the patterns of DNA sequence diversity we expect to observe at the $\mathcal{N}$ locus following the conclusion of the standing sweep.

Our general approach is to break the history of the standing sweep into two periods, the first being the time during which the B allele is selectively favored and rising in frequency (we refer to this as the sweep phase), and the second being the period after the mutation has arisen but before the environmental shift causes it to become beneficial (we refer to this as the standing phase). Stated very briefly, our approach is to assume that selection is sufficiently strong that only recombination (i.e. no coalescence) occurs during the sweep phase, and to recognize that recombination events occurring during the standing phase can be treated as mutations in the infinite alleles model, which allows us to make use of a version of the Ewens Sampling Formula in a variety of useful ways.

2.2.1 Sweep Phase

Looking backward in time, we let $X(t)$ be the frequency of the B allele at the $\mathcal{B}$ locus at time t in the past, where $t = 0$ is the moment of fixation (i.e. $X(0) = 1; X(t) < 1 \ \forall \ t > 0$). Recalling that r is the per generation recombination rate between the $\mathcal{B}$ and $\mathcal{N}$ loci, the probability that a single lineage sampled at the $\mathcal{N}$ locus at $t = 0$ fails to recombine off of the selected background in generation t is $1 - r(1 - X(t))$. If we let τ_f be the generation in which the environmental change occurred, marking the boundary between the sweep phase and the standing phase (i.e. $X(\tau_f) = f$), then the probability that a single lineage manages to recombine off the selected background at any point during the course of the sweep phase is given by

$$P_{NR} = \prod_{t=0}^{\tau_f} 1 - r(1 - X(t)) \approx \exp\left(-r \int_0^{\tau_f} (1 - X(t))\mathrm{d}t\right) \tag{1}$$

for $r \ll 1$. We set $\mathcal{T}_{(s,f)} = \int_0^{\tau_f} (1 - X(t))\mathrm{d}t$, so that the probability that a lineage manages to recombine off the selected background during the course of the sweep from frequency f can be written $e^{-r\mathcal{T}_{(s,f)}}$. If the effect of our beneficial allele on relative fitness is strictly additive, such that heterozygotes enjoy a selective advantage of $s/2$ and heterozygotes an advantage of s, then the trajectory of the beneficial allele through the population can be approximated deterministically by the logistic function, such that

$$\mathcal{T}_{(s,f)} = \frac{ln\left(\frac{N_e - 1}{f} - N_e + 1\right)}{s}. \textit{(this is the 1/2N approx)} \tag{2}$$

We assume that the sweep occurs fast enough that the probability of coalescence during the sweep is essentially zero. Therefore, each lineage either recombines off the B background, or fails to do so, independently of all other lineages, so that the probability that i out of n lineages fail to escape off the sweeping background is

$$P_{NR}(i; n) = \binom{n}{i} P_{NR}^i (1 - P_{NR})^{n-i}. \tag{3}$$

This binomial approximation has been made by a number of authors in the context of hard sweeps (**?**), and more accurate approximations have been developed (**?**). However, as long as selection is strong, the sample is not too large, and τ_f is not too long, then this approximation should be adequate. Other, more accurate approximations could certainly be incorporated into our framework, but we stick with this simple form for the sake of clarity of presentation.

2.2.2 Standing Phase

Looking backward in time, having originally sampled n lineages at the $\mathcal{N}$ locus at $t = 0$, we arrive at the beginning of the standing phase at time τ_f with $n - i$ lineages linked to the non-beneficial b background at the $\mathcal{B}$ locus (which has a frequency of $1 - f$), and i lineages linked to the beneficial B background (which has a frequency of f).

We will argue that an understanding of patterns of neutral diversity at the $\mathcal{N}$ locus following a standing sweep can be obtained principally by tracking the histories of only the lineages that are found on the **B** background at the beginning of the standing phase, and assuming that lineages which recombine off of the **B** background do not re-enter it. This is obviously a coarse approximation to the true process, but as we argue below, it captures many of the major features sufficiently well for our purposes.

The Coalescent Process at the $\mathcal{B}$ Locus In attempting to construct the genealogy of the $\mathcal{B}$ locus backward in time, consider that in the first generation of the standing phase, the probability that any pair of B coalesces with one another is $1/(2Nf)$, while the probability that any pair of b alleles coalesce with one another is $1/(2N(1-f))$. In general, the pairwise coalescent probabilities for pairs of lineages T generations back into the standing phase are $1/(2NX(\tau_f + T))$ for the B alleles, and $1/(2N(1 - X(\tau_f + T)))$ for the b alleles, where the frequency $X(\tau_f + T)$ may be equal to, greater than, or less than f, due to the random sampling effects of genetic drift. Further, coalescent events which may have occurred during the intervening T generations contain information about the distribution of $X(\tau_f + T)$.

A number of researchers have studied the behavior of this process (**?**), either conditional on the frequency of the allele in a sample or in the population. **?** has shown that the expected time to the first coalescent event is $2Nf/\binom{i}{2}$ in the absence of other information, e.g. as to whether the allele is ancestral or derived. However, the distribution of coalescence times is no longer exponential. The variance of the time between coalescent events is increased relative to the exponential as a direct result of the fact that the frequency may increase or decrease from f before a given coalescent event is reached. Further, in contrast to the standard coalescent, there is non-zero covariance between subsequent coalescent intervals, as a result of the information contained about how the frequency of the allele has changed, and thus about the rate at which subsequent coalescent events occur. Lastly, if the allele is known to be either derived or ancestral the coalescent times have a more complicated expectation, as the allele is in expectation either decreasing or increasing in frequency backward in time due to the conditioning on loss or fixation respectively.

Despite these complications we have found assuming that lineages coalesce at a rate $\binom{i}{2}/(2Nf)$ and that coalescent time intervals are independent, i.e. that the allele frequency does not drift from f, is not a bad approximation when $f \ll 1$ regardless of whether the allele is ancestral or derived. In Supp. Figures XXX-XXX we show some comparisons of the coalescent process embedded in a drifting allele frequency and this approximation.

The main reason for using this approximation is that, in conjunction with a separation of timescales (i.e. an assumption that no **b** alleles coalesce until after all **B** alleles have), it allows us to work with a simple, well understood caricature of the true process that describes the genealogy at the selected site with reasonable accuracy. Conditional on this simplified coalescent process, we can study the process of recombination events occurring between the $\mathcal{B}$ and $\mathcal{N}$ loci to understand the distribution of the genetic variation at the $\mathcal{N}$ locus that will hitchhike along with the beneficial allele once the sweep begins.

Recombination Between $\mathcal{B}$ and $\mathcal{N}$ We will again rely on the condition that $f \ll 1$, and assume that any lineage at the $\mathcal{N}$ locus that recombines off of the background of our beneficial allele will not recombine back into that background before it is removed by mutation. Under these assumptions, recombination events which move lineages at the $\mathcal{N}$ locus from the beneficial background onto the non-beneficial background can be viewed as events on the genealogy at the $\mathcal{B}$ locus which occur at

rate $r\,(1-f)$ per lineage. Rescaling time by $2Nf$, an understanding of the genealogy at the $\mathcal{N}$ locus can therefore be found by considering the competing poisson processes of coalescence at rate $\binom{i}{2}$, and recombination at total rate $2Nirf(1-f)$

If we are interested in the number and size of different recombinant clades at a given recombination distance from the selected site (colored clades in Figure 2.2B & C) this a direct analogy of the infinitely-many allele model (). In the infinite alleles model, every mutation event creates a new allele, while in our process every recombination event creates a new recombinant lineage (and potentially a distinct haplotype, depending on the configuration of mutations it carries). Further, a sample from the infinite alleles process can be found by simulating the coalescent, scattering mutations down on the genealogy, and then assigning each lineage to be of a type corresponding to the mutation that sits lowest above it in the genealogy (see Figure 2.2B). Equivalently, we can create a sample under the infinite allele model by simulating the mutational and coalescent processes simultaneously, 'killing' lineages whenever they first encounter a mutation and assigning all tips sitting below the mutation to be of the same allelic type (see Figure 2.2C).

Given the direct analogy to the infinite alleles model under our set of approximations, the number and frequency of the various recombinant lineage classes at a given distance from the selected site can be found using the Ewens' Sampling Formula (ESF).The population-scaled mutation rate in the infinitely-many alleles model ($\theta/2 = 2N\mu$), in our model, is replaced by the rate of recombination out of the selected class ($R_f/2 = 2Nrf(1-f)$). If i lineages sampled at the moment of fixation fail to recombine off of the beneficial background during the course of the sweep, then the probability that these i lineages coalesce into a set of k recombinant lineages is

$$p_{ESF}(k \mid R_f, n) = S(i,k)\frac{R_f^k}{\prod_{\ell=1}^{i-1}(R_f+\ell)} \tag{4}$$

where $S(i,k)$ is an unsigned Stirling number of the first kind

$$S(i,k) = \sum_{i_1+\cdots+i_k=i} \frac{i!}{k! i_1 \dots i_k} \tag{5}$$

These recombinant lineages partition our sample up between themselves, such that each lineage has some number of descendants in our present sample $\{i_1, i_2, \dots, i_k\}$, where $\sum_{j=1}^{k} i_j = i$. Conditional on k the probability of a given sample configuration is

$$p(\{i_1, i_2, \dots, i_k\} \mid k, i) = \frac{i!}{k! i_1 \cdots i_k S(i,k)} \tag{6}$$

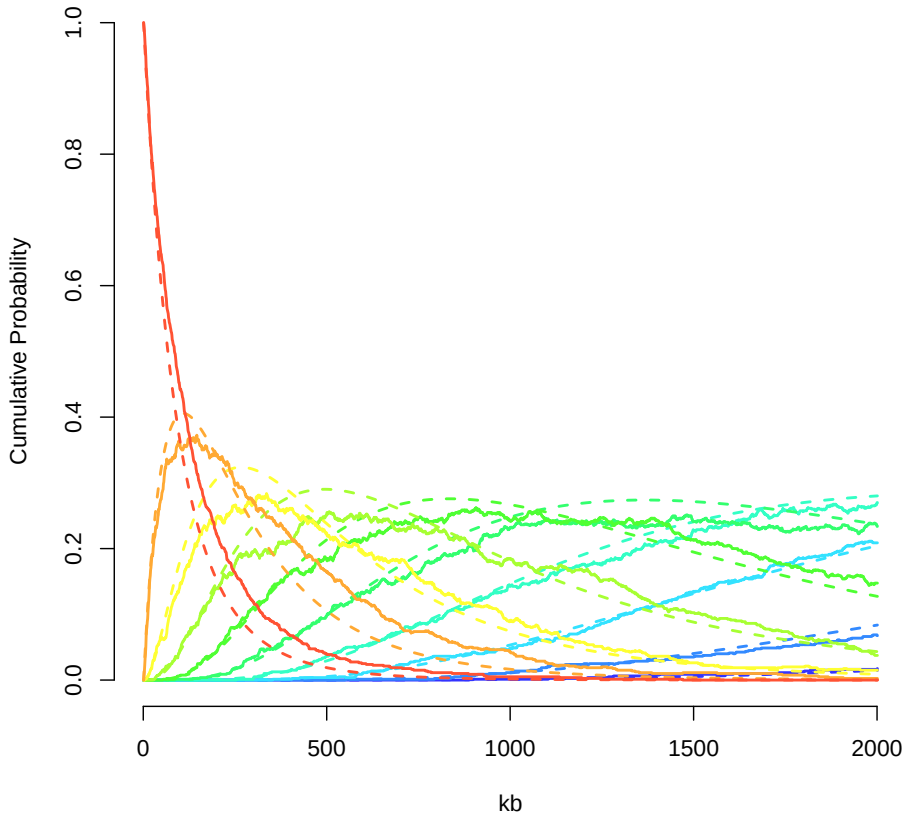
Note this does not depend on R_f, which gives the classic result that the number of alleles is sufficient statistic for R_f (i.e. the partition is not needed to estimate R_f).

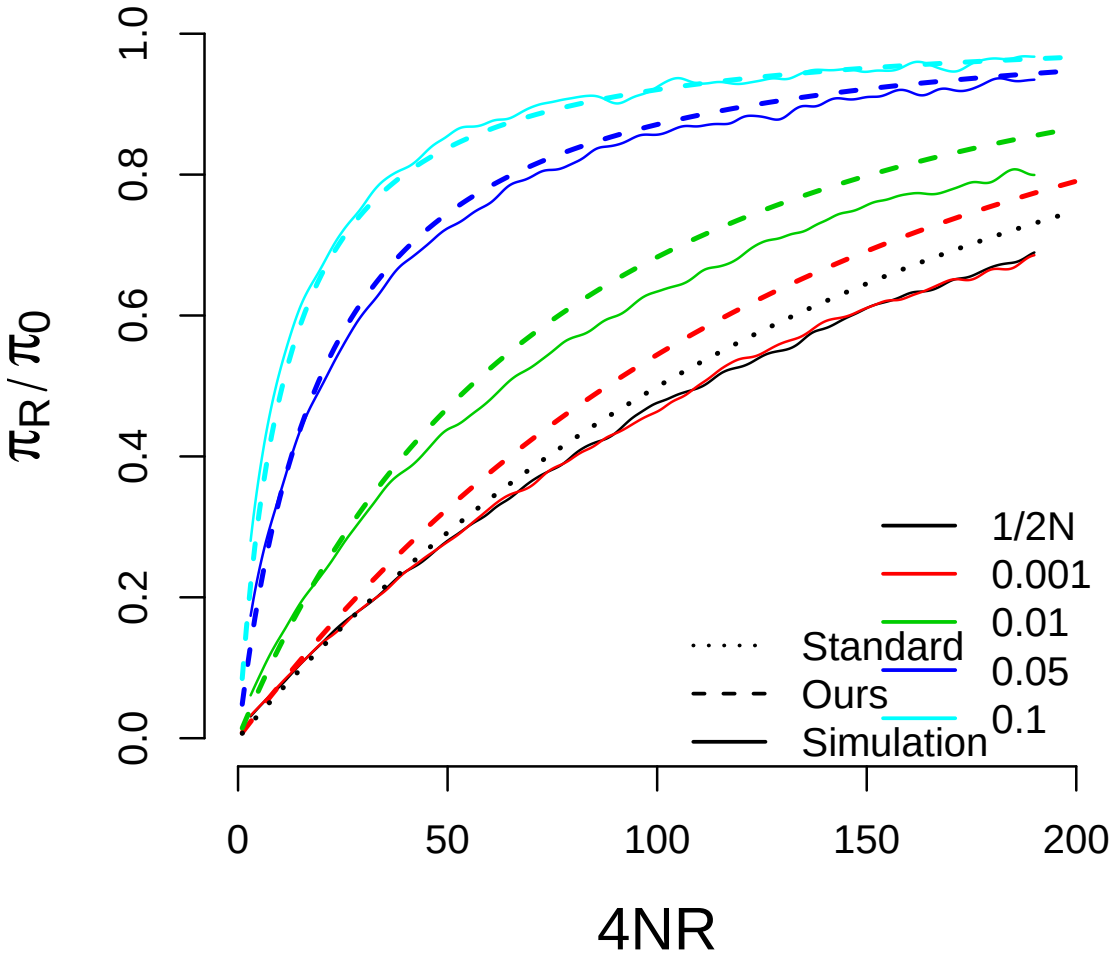
2.2.3 Patterns of neutral diversity surrounding standing sweeps

Given this approximate model of the coalescent with a sweep from standing variation we can now calculate basic summaries of variation in the region surrounding the sweep. We will assume that the per base pair mutation rate per generation is μ. We will ignore mutations over the time-scale of our shrunken coalescent tree, and assume that all diversity comes from mutations that occurred prior to the sweep, or equivalently that this part of the genealogy contributes negligibly to the total amount of time in the genealogy. This corresponds to an assumption that $2N\mu \gg 2N\mu f$, in line with our previous assumption that $f \ll 1$. If this is the case we simply consider patterns of diversity in our sample at a given site by considering properties of the recombinant lineages in our sample, which correspond to alleles drawn independently from a neutral population prior to the start of our sweep.

For example, excluding recombination during the sweep for a moment, the expected pairwise coalescent time for lineages a distance r away from the sweep is

$$\mathbb{E}(T_2) \approx \frac{1}{1+4Nrf(1-f)} \times 0 + \frac{4Nrf(1-f)}{1+4Nrf(1-f)} \times 2N \tag{7}$$


Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
kb



where the two terms correspond to the contribution from failing to recombine during the standing phase and so coalescing very rapidly, and alternately to one or both lineages escaping from the beneficial background and coalescing $2N$ generations ago.

Now incorporating recombination during the sweep our expected pairwise coalescent time a distance r away from the sweep is

$$\mathbb{E}(T_2) \approx \left(1 - \frac{1}{1 + 4Nrf(1-f)} P_{NR}^2\right) \times 2N \tag{8}$$

as to avoid (near) instantaneous coalescence our pair of lineages could either recombine either during the sweep or during the standing phase. The expected level of pairwise diversity as we move away from a sweep is given by $2\mu\mathbb{E}(T_2)$ In Figure XXX we show this approximation, and coalescent simulations done using ms.

We can extend this idea of conditioning on the number of lineages that escape the sweep to calculate the expected total time in the genealogy as we move away from the $\mathcal{B}$ locus. Conditional on k independent lineages escaping the sweep, the expected total time in the genealogy is $2N \sum_{\ell=1}^{k-1} 1/\ell$, the standard result for a neutral coalescent with k lineages (**?**). Ignoring for a moment recombination during the sweep phase, the probability that k lineages escape the sweep is the probability of k alleles in a sample of n under the

ESF, $p_{ESF}(k \mid R_f, n)$ (under our approximation). The expected time in the genealogy a distance r away from the selected site is therefore

$$\mathbb{E}(T_{TOT}) \approx 2N \sum_{k=2}^{n} p_{ESF}(k \mid R_f, n) \sum_{j=1}^{k-1} 1/j. \tag{9}$$

Reincorporating recombination during the sweep phase, the probability that k distinct lineages have recombined off of the beneficial background between the two phases together is

$$\sum_{m=0}^{k} \binom{n}{m} P_{NR}^{m} (1 - P_{NR})^{n-m} p_{ESF}\left(k - m \mid R_f, n - m\right), \tag{10}$$

because if m recombinant lineages are generated during the sweep, then the remaining $k-m$ recombinant lineages have to come from recombination events in the standing phase (we take $p_{ESF}\left(0 \mid R_f, 0\right) = 1$, representing the extreme case where $k = n$ and all lineages manage to recombine during the sweep phase). The expected total time in the genealogy is therefore

$$\mathbb{E}(T_{TOT}) \approx 2N \sum_{m=0}^{k} \binom{n}{m} P_{NR}^{m} (1 - P_{NR})^{n-m} p_{ESF}\left(k - m \mid R_f, n - m\right) \sum_{\ell=1}^{k-1} 1/\ell \tag{11}$$

and the expected number of segregating sites can be found by taking μ times this. In Figure XXX we show this approximation, and coalescent simulations done using ms.

We can also obtain an expression for the frequency spectrum at sites surrounding a standing sweep. To break the problem into approachable components, we first condition on an absence of recombination during the sweep phase, and a fixed number k recombinant families which are created by coalescence and recombination in the standing phase (both of these will be relaxed momentarily). Each of the k recombinant families represents an independent draw from the population frequency prior to the **B** allele entering the population. Borrowing a trick from **?** (equation 14 of their paper), if we condition on j out of k recombinant lineages carrying a derived allele, then we can obtain the probability that l of the n lineages sampled at fixation carry the derived allele by summing over all possible partitions of the n tips into k families such that the j recombinant ancestors carrying the derived mutation have exactly l descendants in the present day as

$$p(l \mid j,\ k,\ i = n) = \sum_{\substack{i_1 + \cdots + i_j = l \\ i_{j+1} + \cdots + i_k = n - l}} p(\{i_1, \ldots, i_k\} \mid k, n) = \frac{\binom{n}{l}}{\binom{k}{l}} \frac{S(l, j) S(n - l, k - j)}{S(n, k)} \tag{12}$$

Next, we write $q_{j,k}$ to denote the probability that j out of the k recombinant families carry the derived mutation. For our purposes, we will assume this distribution follows that of the standard neutral coalescent expectation (i.e. $q(j \mid k) = \frac{1/j}{\sum_{\ell=1}^{k-1} 1/\ell}$ gives the probability of j derived alleles in a sample of k, conditional on segregation), although one could easily use an empirical frequency spectrum measured from genome-wide data, as in (**?**). The probability that the derived allele is present in l out of n sampled lineages, conditional on there having been k recombinant families, is then

$$p(l \mid k, i = n) = \sum_{j=1}^{k} p(l \mid j,\ k,\ n) q(j \mid k). \tag{13}$$

Summing over the distribution of k given by (4), we obtain an expression for the frequency spectrum conditional on no recombination in the sweep phase as

$$p(l \mid i = n) = \sum_{k=2}^{n} p_{ESF}(k \mid R_f, n) \sum_{j=1}^{k-1} q(j \mid k) p(l \mid j,\ k,\ n) \tag{14}$$

When we allow for recombination during the sweep, this expression becomes more complex, but the same logic can be followed and write

$$p(l \mid n) = \sum_{i=0}^{n} P_{NR}(i \mid n) \sum_{k=1}^{i} P_{ESF}(k \mid R_f, n - S) \sum_{j=1}^{k+n-i-1} q(j \mid k + n - i) \sum_{g=0}^{j \wedge l \wedge (n-i)} H(g \mid j, k, n - i) p(l - g \mid j - g, k, n - i) \tag{15}$$

where A∧B denotes min(A,B) and

$$H(g \mid j, k, n-i) = \frac{\binom{j}{g}\binom{n-i}{g}}{\binom{k+n-i}{j}} \tag{16}$$

gives the probability that g out of j derived alleles are found on singleton recombinants created during the sweep, given that there are $n-i$ singletons, and k recombinant families created during the standing phase. Here $n-i$ lineages recombine out during the selected phase, while the remaining i lineages are partitioned into k families due to recombination and coalescence in the standing phase. Out of the $n-i$ singleton lineages, g of them carry the derived allele, while the remaining $j-g$ copies of the derived allele which existed just prior to the arrival of the beneficial allele give rise to $l-g$ derived alleles in the present day, resulting in l out of n sampled lineages carrying the derived allele.

2.3 Patterns of Haplotypic Variation

We have described a model for the marginal genealogical and recombinational history at a site some distance from a standing sweep. It is of interest to consider patterns of haplotypic diversity we expect to observe along the sequence for a single realization of a standing sweep, or in other words the *joint* history of the sequence in the region surrounding beneficial allele. If we again follow the custom of first ignoring recombination during the sweep phase We want to be able to describe patterns of haplotypic variation surrounding standing sweeps. In particular, we are interested in the usefulness of haplotype statistics to distinguish three different types of "full sweep": the classical hard sweep, a sweep including multiple mutations, and a sweep from standing variation.

We continue our infinitely-many haplotypes perspective, in which we imagine that every piece of DNA present in the sample taken at fixation can be labelled/colored according to which chromosome in the ancestral, pre-sweep population it traces its ancestry to. Under this convention, right at the selected site, all chromosomes descending from either a hard sweep or a standing sweep will be of the same "color", due to their shared common ancestry, while for multiple mutation soft sweeps each set of chromosomes tracing to a different mutational origin of the beneficial allele will be of a different color. For all kinds of sweeps, the number of distinctly colored haplotypes increases with distance from the selected site due to recombination events which cause transitions along the sequence in the haplotype partition scheme.

With this perspective, most of the information about the parameters of the sweep (including that of which type of sweep it is), are contained jointly in the set of transitions in the haplotype partition scheme as one moves away from the selected site, along with the list of their locations along the sequence. As we are considering the haplotype partition largely in the context of inference, it is worth briefly acknowledging that by envisioning detailed knowledge of haplotype identities and transitions around a selected site, we are assuming access to a level of data that is quite unlikely to be attainable in any real population. The following results therefore constitute a sketch of the upper bound on the ability to distinguish the different kinds of sweeps. In cases where the ancestral genetic diversity is low, the problem is considerably worse.

Below, we consider the construction of this haplotype partition as a process along the sequence for each of the three types of sweeps, beginning with either a single haplotype at the selected site (hard and standing cases), or potentially more than one (in the multiple mutation case), and transitioning as we move away through states where there are two, three, four, etc. haplotypes, until we eventually transition through to the state where all sampled chromosomes are found on independent haplotypes.

The Classic Hard Sweep Case Under the binomial approximation for the classic hard sweep case, as used by **??**, the expected distance in base pairs to the transition from one haplotype to two is $\frac{1}{n\mathcal{T}_s r_{BP}}$, and in general the expected distance between the transition from i haplotypes to $i+1$ haplotypes is $\frac{1}{(n-i+1)\mathcal{T}_s r_{BP}}$. Further, conditional on the value of $\mathcal{T}_s$, the distances between subsequent haplotype partitions are roughly independent. In this approximation, all transitions result from one individual from the core haplotype becoming a singleton due to a recombination event that occurred before the multiple merger at the beginning of the sweep, and thus all of the information about the sweep is contained in the joint distribution on the distances at which the transitions occur. Each of the transitions from grey to black in Figure 2.3 represents an independent recombination event, and each grey haplotype is randomly drawn from the diversity prior to the sweep.

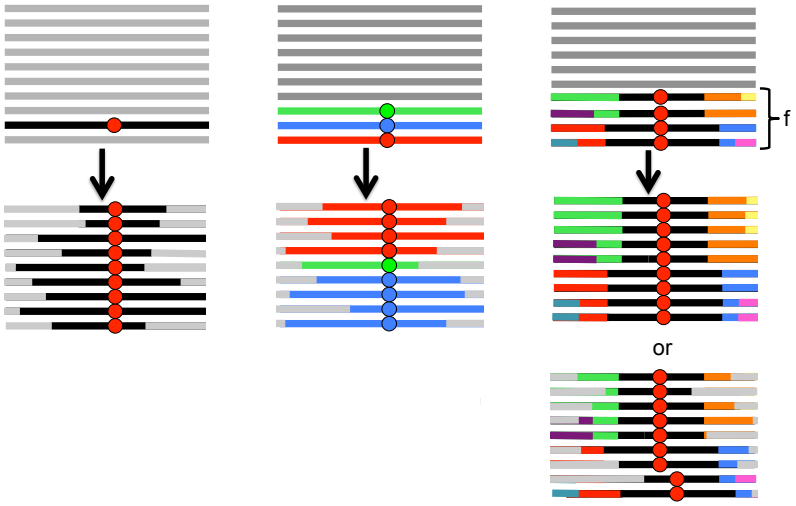


Figure 2: caption goes here

The majority of our information for identifying full sweeps therefore comes from the presence, and slow decay, of a single haplotype over long distances. The fact that recombination is slowly peeling off singleton haplotypes as we move away from the sweep results in the signal of an excess of singletons and high frequency derived alleles (**?**) occurring in genomic run, a signal which forms the basis for a number of popular sweep detection algorithms **???**. Another consequence of this behavior is that immediately following a full sweep, there is a much reduced level of pairwise linkage disequilibrium across the selected site (**??**), as the star-like tree at the selected site renders recombinations on either side of the sweep effectively independent (**?**).

In reality, all coalescence does not occur instantaneously at the base of a classic hard sweep, especially when the sample is even moderately large (i.e. $n > 20$), and thus there is an opportunity for recombination events to occur on internal branches of the genealogy **??**. This causes transitions in the partition scheme that result in a class of additional non-singleton haplotypes, which eventually break apart farther out from the selected site due to additional recombination events lower down in the genealogy.

Sweep with multiple mutations. **??** investigated the case of multiple mutations at a single selected locus contributing to a sweep. This sort of sweep can occur if there are multiple mutations present at mutation-selection balance prior to the onset of selection, or if they arise sequentially during the sweep. The simple signal of hitchhiking in this case is complicated, as rather than a single allele sweeping to fixation, the multiple alleles all sweep to somewhat intermediate frequencies, partitioning the same into multiple core haplotypes. For example, in Figure 2.3B, the sampled haplotypes have been partitioned by multiple mutations during the sweep into three distinct haplotypes at the selected site. Parameterizing the mutation rate toward beneficial alleles at the locus as μ_B, **?** found that for small samples and strong selection, the partitioning of lineages at the selected site by different mutations was well approximated by the Ewens Sampling formula with parameter $4N_e\mu_B$. Clearly, if a particular stochastic realization of this process results in only a single mutation, then we simply have a classic hard sweep, as above.

The tree under each of these mutations will be a very short and reasonably star-like, such that there will be a single haplotype associated with each mutation close to the selected site. Aside from the presence of multiple haplotypes in this core region, the breakdown in the haplotype partition moving away from the selected locus is relatively similar to that under the hard sweep. Again assuming the

sweep took $\mathcal{T}_s$ generations, the distance between the i^{th} and $i+1^{st}$ haplotype transitions is exponential with expectation $\frac{1}{(n-i+1)\mathcal{T}_s r_{BP}}$.

Allelic differences between ancestral haplotypes at neutral sites will now be found at intermediate frequency in the sample, such that the frequency spectrum can be skewed close to the center of the sweep (**?**). We also anticipate the generation of considerable linkage disequilibrium both between sites on the same side, as well as opposite sides of the sweep due to the correlation induced by the partitioning at the core region**?**. Whether these patterns are visible depends on the magnitude of neutral diversity in the population before the sweep. If the distance between neutral polymorphisms is $\ll 1/(\mathcal{T}_s r_{BP})$ then there will likely be multiple neutral alleles revealing the haplotype partitioning at the selected site, while if this condition is not met, we are unlikely to see the signal of the sweep at all.

Standing Sweep Case The situation for a standing sweep is considerably more complicated than either the hard sweep or the multiple mutation sweep. Recall that τ_f gives the duration of the sweep phase, while the time from the beginning of the standing phase back to the common ancestor of the beneficial alleles is approximately $4N_e f$. Therefore, if $4N_e f \ll \tau_f$ (i.e. selection is very weak, or the mutation is still quite rare when selection switches on), most of the recombination events that affect the haplotype partition occur during the sweep phase, and so the pattern is very similar to that of a hard sweep.

The alternate extreme, where $4N_e f \gg \tau_f$, occurs when selection is strong or the mutation was segregating at a relatively high frequency prior to the onset of selection. In this parameter regime, the sweep phase occurs fast enough that recombination events from this period tend to be far away from the selected site, and the structure of the haplotype partition is dominated by the standing phase.

While an analytical treatment of this process is beyond our reach, we can still imagine constructing the haplotype partition as a process along the sequence, which yields some insight. If the total time in the coalescent tree in the neutral phase at the selected site is T_{tot}, then the distance to the first recombination is $\sim \exp(r_{BP} T_{TOT})$. Using the standard approximation to the constant in Watterson's θ, the expected length scale over which a single haplotype should persist away from the selected site is $\approx 1/(2N_e R_f log(n-1))$. This recombination partions the haplotypes according to the infinite alleles model for single mutation placed at random on the geneaology at the selected site (e.g. the green recombinant moving to the left in Figure 2.3C). Moving on down the sequence we then generate the next distance to a recombination along the sequence, again from $\sim \exp((r_{BP} T_{TOT})$. We once again uniformly simulate a position on the tree for this new recombination, however, this time only a recombination on some of the branches would result in another colored haplotype being introduced into the sample, e.g. a recombination that falls on the same branch as the previous one. If the recombination falls in a place that doesn't alter the configuration we ignore it, otherwise we split our sample configuration again. For example, the red recombinant in Figure 2.3C does not alter the sample configuration and so can be ignored. While it does recombine the selected allele off of its original black haplotype, this recombinant is not visible in our contemporary sample (likewise the light blue recombinant in Figure 1XX is not visible). We iterate this procedure moving away from the selected site, generating exponential distances to the next recombination, placing the recombination down, updating the configuration if needed, until we reach the point that every colored haplotype is a singleton. We then repeat this procedure on the other side of the selected site using the same underlying tree at the selected site.

An equivalent way to describe this process is to simulate distances to the next recombination that alters the configuration, given the tree. To do this we consider the total time in the tree where a recombination would alter the configuration. Numbering these recombinations out from the selected site, we start at the selected site $i = 0$, with $T_0 = T_{TOT}$ and generate a distance to the next recombination $\sim \exp(r_{BP} T_0)$. We place the recombination on the tree, then prune the tree of branches where no further change in sample configuration could result in a new colored haplotype. We then set T_i to the total time in this pruned tree and carry on this process till we have pruned the entire tree (at which point the distance to the next visible recombination would be infinite as we are left only with singleton haplotypes).

As we have discussed above, marginally at a given site the partitioning of haplotypes (integrating out the unobserved genealogy) is given by the ESF. It is unclear to us how to couple together the partitioning at two sites in an analytically tractable way, i.e. how to calculate the probability of how the sample configuration $\{i_1, i_2, \ldots, i_k\}$, is further subdivided by recombinations at a site further along the sequence.

Obviously this can be calculated by resorting to brute force integration over the set of all unobserved trees that are consistent with the partition at the first site, but it seems like a more elegant solution could be available. Given the general interest in the ESF motivated by exchangeable partitions and clustering algorithms, such a coupling may be of wider interest.

Under our approximation, each of our coloured haplotypes represents an independent draw from the haplotypes present before the sweep, such that allelic differences we see between these ancestral haplotypes are the alleles that segregate after the sweep. As such, we should expect LD to be created between variants on the same haplotypic background. As haplotype configurations should extend over distances $\sim 1/(r_{BP}N_e f)$ this LD will extend over distances unusual compared to background LD if $N_e f \ll N_e$. The colored haplotypic patitioning on either side of the selected site is correlated due to the underlying dependence on the genealogy at the selected site, thus there can be LD across the selected site–unlike the full sweep model. But in contrast to the soft sweeps model we have to wait until we get a sufficient distance from the selected site in order for recombination to have occurred, to have variation to observe the LD. We have explored approximating the pairwise LD around the sweep using the coalescent approach of **?**, however, we found that no simple expressions were forthcoming from that approach (principally because we have not been able to couple two sites together other than by brute force).

If τ_f is on the same order as $4N_e f$ then the distances to recombination events in the two phases will be on the same genomic scale, but recombinations in the selected phase occur first moving backward in time and so obscure the patterns of the sample described above (see grey bars in Figure 2.3Cii). As before the physical distance along the sequence to each recombination during the selected phase is independently $\sim \exp(r_{BP}\tau_f)$, such that the core region swept clean of diversity to be $\sim \exp\left(R_f T_{tot} + r_{bp} n t_f\right)$. The ability to see the colored haplotype structure will depend strongly on the exact magnitude of τ_f as compared to $N_e f$, and is likely to be highly stochastic across replicates.

2.4 Inference Concerns

Some authors have developed methods to infer the existence of standing sweeps (**?**) or multiple hit soft sweeps (**?**). Here, we apply our model of a standing sweep to understand the limitations to inference of standing sweeps and to explore issues of potential confounding between standing and multiple hit soft sweeps.

If fN_e is not a lot smaller than N_e then the resulting sweep will be very hard to distinguish from neutral patterns of diversity, much like the difficulty of identifying weak full sweeps. Obviously we can still potentially identify the allele as a putative signal of selection, if we also identify it as an unusually large change in allele frequencies between populations. In that case the lack of a haplotype signal may alter us that the allele potentially swept from standing variation. In the absence of data from populations where the allele has not swept, we are restricted to identifying alleles where $fN_e + \tau_f \ll N_e$. In that case the core region swept clean of variation should usually be distinguishable from background fluctations in diversity (controling for mutation rate, e.g. via divergence data). However, the trough in diversity will be smaller than an equivalent classic sweep with $t_S \approx \tau_f$ if $f \gg 1/(2N_e)$. A sweep consisting of multiple adaptive alleles will not result in a region swept clear of diversity, as there is no one core haplotype.

Approaches that are based on looking for an unusually long persistance of haplotypes REFS should have much better power to identify all three kinds of sweep. If θ_B is not too large, or equivalently if $N_e f$ is on the order of t_f or smaller, then we should have multiple pairs (or higher) of haplotypes that match each other out to a genetic distance of $\sim 1/s$. Obviously the cluster of similar haplotypes is larger in the case of a classic sweep, and so it will likely be hard to design a test that has roughly uniform power over all three types of sweep (particularly the type of sweep from standing variation).

One key signal of a full sweep has been the prescence of singleton (or low frequency) recombinants off the sweep, leading to high frequency derived variants and contributing to the excess of singleton alleles. If the majority of our singleton recombinants arose

2.4.1 Are s and f independently inferable?

In our standing sweep model, the information indicating that the sweep began from standing variation comes primarily from an observation that the genealogy at the selected site is shrunken by a factor f, while the information about the sweep comes form of the external branches of this genealogy being

slightly longer than expected under this rescaling. The task of an inference procedure designed to identify standing sweeps is then to determine both that this shrunken genealogy exists (and that it is smaller than would be expected under neutrality), and that its external branches are too long, given the times in the rest of the tree.

Under our simplified model,

(points to hit)

Are s and f distinguishable or both inferable. –Weak selection maybe everyone recoms out –Large f maybe no signal of sweep? –What fraction of the singleton recombinants come from selected vs stand phases? –Do you get to see the singleton recombinants from the sweep. Total coalescent time in sweep phase vs total time in standing phase?

Comparison to multiple mutations. –What makes the two cases distinguishable –Code up? Pennings and Hermisson.

Time since to sweep. –does this mess things up? –scale of recombination and coalescence leading up to sweep. –Maybe just have this discussion.

3 Discussion

Gene conversion –minor point in discussion too.

4 Acknowledgements

5 Methods

6 Supplementary materials