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

2 Results

Consider two loci, labeled $\mathcal{N}$ and $\mathcal{B}$, which are separated on the chromosome by a recombination distance r. We imagine that an allele 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. Alleles at the $\mathcal{N}$ locus are neutral, and its history can be determined by considering a structured neutral coalescent conditional on 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 on the process of a sweep from standing variation, as well as to derive expectations for a range of population genetic summary statistics in a sample taken at the conclusion of a standing sweep.

The key insight for this paper is that the recombination events that occur between the $\mathcal{N}$ and $\mathcal{B}$ loci before the sweep has begun can be treated as mutations on the genealogy of the $\mathcal{B}$ locus, and thus under the assumption of constant population size, we can use the Ewens Sampling Formula to calculate expectations for a variety of quantities of interest in the region surrounding the selected site.

(In this paper, we will use the phrase 'beneficial allele' to refer to the allele that becomes beneficial during the selected phase and the phrase 'non-beneficial allele' to refer to the complementary allele which becomes disfavored, regardless of whether the effect of selection is actually being felt during the part of the process being considered. We will use the phrase 'sweep phase' to refer to the part of the history during which the beneficial allele is increasing in frequency due to positive selection, and the phrase 'standing phase' to refer to the part of the history that preceeds the onset of selection.)

2.0.1 Sweep Phase

Looking backward in time, we let $X(t)$ be the frequency of the beneficial allele at the $\mathcal{B}$ locus at time t in the past, where $t = 0$ is the moment of fixation (i.e. $X(0) = 1; X(t) < 1 \ \forall \ t > 0$). 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 be the first generation (forward in time) in which the allele at the $\mathcal{B}$ locus is beneficial (i.e. $X(\tau) = f$), then the probability that a single lineage manages to recombine off the selected background at any point during the course of the sweep is given by

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

for $r \ll 1$. We set $\mathcal{T}_{(s,f)} = \int_0^{\tau} (1 - X(t))\mathrm{d}t$, 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 $e^{-r\mathcal{T}_{(s,f)}}$.

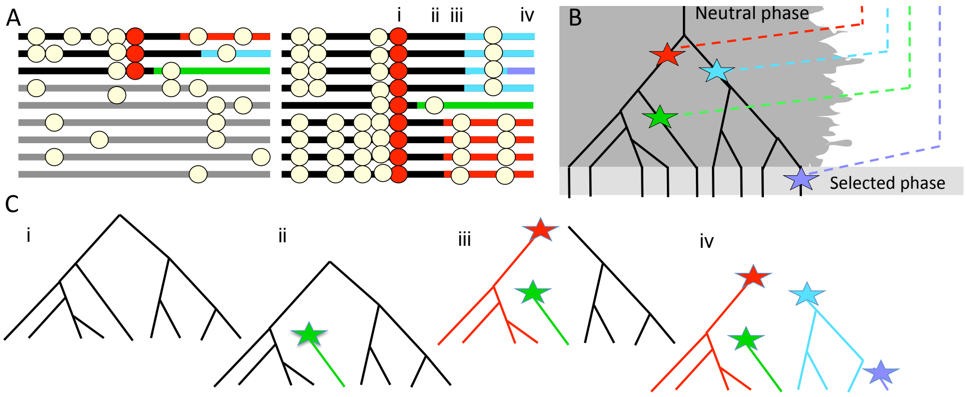


Figure 1: caption goes here

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, 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 makes an independent choice of whether it recombines out of the sweep, so that the probability that i out of n lineages fail to escape off the sweep 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 the population is large, the sample is not too large, and τ 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.

2.0.2 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 with $n - i$ lineages linked to the non-beneficial background at the $\mathcal{B}$ locus (which has a frequency of $1 - f$), and i lineages linked to the beneficial background (which has a frequency of f). We will argue that an understanding of patterns of neutral diversity at the $\mathcal{N}$ locus following the sweep can be obtained principally by considering the genealogy at the $\mathcal{B}$ locus, and then considering the effect of recombination events which occur in between the $\mathcal{B}$ and $\mathcal{N}$ loci conditional on this genealogy. The following paragraphs are therefore structured as follows: first, we describe an approximation to the coalescent process at the $\mathcal{B}$ locus conditional on being at a frequency f at the beginning of the standing phase. Next, we describe an approximation to the process of recombination events which move alleles at the $\mathcal{N}$ locus from the beneficial background onto the non-beneficial background at the $\mathcal{B}$ locus. Finally, we combine these two processes along with the recombination process during the sweep phase and a separation of timescales argument to give an approximate description of the full genealogy at the $\mathcal{N}$ locus.

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 instantaneous pairwise coalescence rate of beneficial alleles is $1/(2Nf)$, such that the total rate of coalescence of beneficial alleles in the sample for the first generation is $\binom{i}{2}/(2Nf)$. If the beneficial allele were held fixed at this frequency f over a long period into the past, then the genealogy of our i beneficial lineages would simply be a neutral coalescent with pairwise coalescent rate $1/(2Nf)$ (assuming that $f \gg 1/(2N)$ so that the standard coalescent assumptions hold). Such a fixed frequency could result from a beneficial allele that was balanced at low frequency by strong, constant selection. If instead of being held steady at a frequency f, the allele is allowed to drift neutrally, this constant coalescent rate no longer holds, and the problem becomes considerably more complicated.

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 fact that early coalescent events contain information 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.

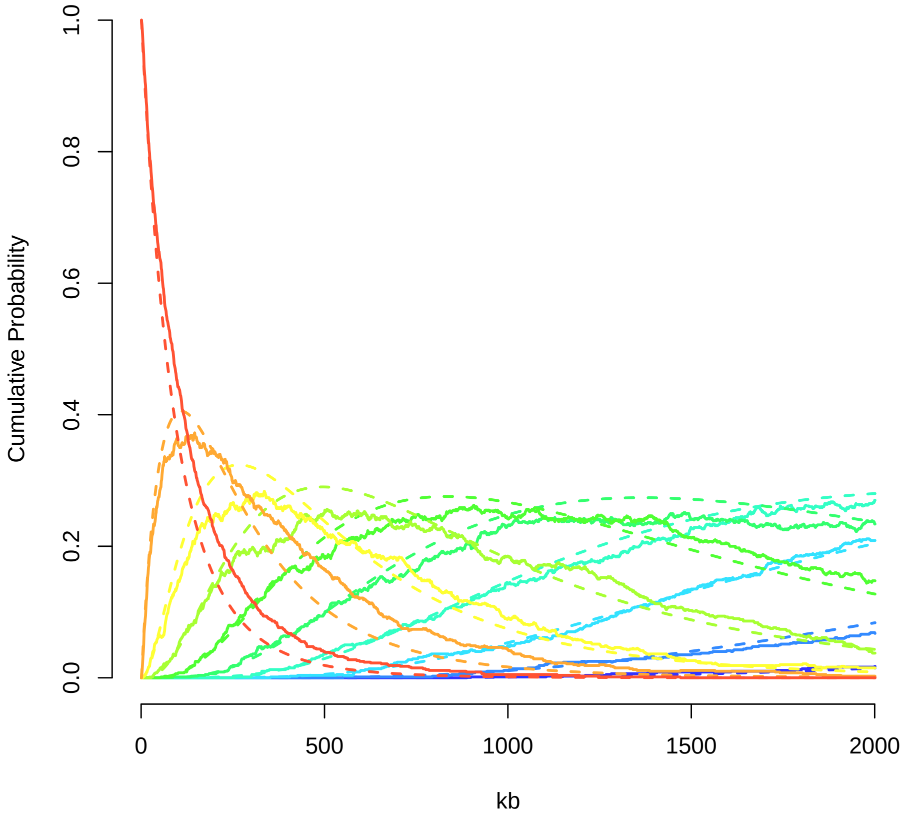
Despite these complications we have found assuming that lineages coalesce at a rate $\binom{i}{2}/(2Nf)$ and that coalescent time intervals are independent, i.e. that the allele frequency does not drift from f, is not a bad approximation when $f \ll 1$ regardless of whether the allele is ancestral or 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 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 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 (an potentially a distinct haplotype, depending on the configuration of mutations). 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 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 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 a Stirling number of the first kind

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

These recombinant lineages partition our sample up between themselves, such that each lineage has some number 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).

(Formally, we consider a limit in which f tends to zero while N tends to infinity such that $2Nf$ is held constant at some value much larger than one. This results in a separation of timescales where both coalescent and recombination events on the background of the beneficial allele occur nearly instantaneously relative to events on the background of the non-beneficial allele.)

2.0.3 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 neglibably 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 set of assumptions that $f \ll 1$. If this is the case we simply consider patterns of diversity in our sample at a 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 a distance r away from our sweep is

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

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. In Figure XXX we show this approximation, and coalescent simulations done using ms.

Now incorporating recombination during the sweep our expected pairwise coalescent time a distance r away from our 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 recombine during either the sweep or standing phases. The expected level of pairwise diversity as we move away from a sweep is given by $2\mu\mathbb{E}(T_2)$

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

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

In Figure XXX we show this approximation, and coalescent simulations done using ms.

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

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

as if we generate m recombinant lineages during our sweep, then the remaining $k-m$ recombinant lineages have to come from recombination events in the standing phase (we set $p_{ESF}\left(0, R_f, 0\right) = 1$, 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.

We can also obtain an expression for the frequency spectrum at sites surrounding a standing sweep. To break the problem into approachable components, we 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 beneficial allele at the $\mathcal{B}$ locus entering the population. 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 (such that $q(j \mid k) = \frac{1/j}{\sum_{\ell=1}^{k-1} 1/\ell}$ 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 $p(l \mid\ k,\ i = n) = \sum_{j=1}^{k} p(l \mid j,\ k,\ n)q(j \mid k)$. 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

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

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

$$p(l \mid n) = \sum_{i=0}^{n} P_{NR}(i \mid n) \sum_{k=1}^{i} P_{ESF}(k \mid R_f, n{-}S) \sum_{j=1}^{k+n-i-1} q(j \mid k{+}n{-}i) \sum_{i=1}^{l\wedge(n-i)} H(g \mid j, k, n{-}i)p(l{-}g \mid j{-}g, k, n{-}i) \tag{14}$$

where A∧B denotes min(A,B) and

$$H(g \mid j, k, n-i) = \frac{\binom{j}{g}\binom{n-i}{g}}{\binom{k+n-i}{j}} \tag{15}$$

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.

2.1 Patterns of Haplotypic Variation

(Write some nice intro here)

We have described a model for the marginal genealogical and recombinational history at a site some distance from a standing sweep. It is of some 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

Under this model, for both hard sweeps and standing sweeps, all chromosomes at the selected site trace their ancestry to the individual ancestral chromosome on which the beneficial mutation originally arose, and thus there is only a single haplotype at the selected site. As we imagine moving along the sequence away from the selected site, we would eventually encounter a transition from a single haplotype to two haplotypes (with some number of chromosomes partitioned into each haplotype group), and a subsequent transition from two haplotypes to three haplotypes (which arises from splitting one of the two existing haplotypes into two subgroups). If we consider moving further and further out from the selected site, we would eventually reach the point where there were n distinct haplotypes in a sample of n chromosomes, where each subsequent increase in the number of haplotypes as we move away from the selected site results from a partitioning of one existing haplotype into two sub-haplotypes.

In principle, then, all of the information about a given sweep is contained in the distances at which transitions in the partition scheme occur, as well as the path the sequence takes through partition space (e.g. whether the first transition creates a partition of {6,4} or {7,3}, etc.)

These transitions arise due to recombination events in on the (unknown) genealogy as we move away from the selected site. Continuing to ignore neutral mutations which have occurred since the mutation(s) which gave rise to the sweep, all of the information about the sweep is contained jointly in the distances

The transitions

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, the multiple mutation soft sweep, and the soft sweep from standing variation. We adopt the convention of querying the number and partition of haplotypes which are present within a window with the left boundary at the position of the selected mutation, and the right boundary some distance away. We will argue that the majority (but not all) of the information which can be used to distinguish these types of sweeps can be accessed by considering how the number and partition of haplotypes changes as the right boundary of the window is increased from zero until we reach the point at which all chromosomes are present on distinct haplotypes (the point beyond which there is no further information about the sweep).

Further, we adopt an infinite haplotypes perspective, in which every recombination event which transfers genetic material from the background of the ancestral allele to the beneficial alleles results in the creation of a distinct haplotype. In this model, 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 beneficial locus, 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 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 distinguishing the different kinds of sweeps. In cases where the ancestral genetic diversity is low, the problem is even worse.

(Note that there is also additional information present in the locations and frequencies, and correlations between mutations which have occurred on the background of the beneficial allele. We will focus for the most part on the haplotype partition scheme as defined solely by recombination events, and then comment briefly on the role of new mutations occurring on the beneficial background.)

The Classic Hard Sweep Case

In the case of a classic hard sweep, we consider the construction of the haplotype partition as a process along the sequence, beginning with a single haplotype at the selected site, 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.

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 simply $\frac{1}{n\mathcal{T}_s r_{BP}}$, and in general the expected distance between the transition from i haplotypes to $i+1$ haplotypes is $\frac{1}{(n-i+1)\mathcal{T}_s r_{BP}}$. Under the binomial approximation, all transitions simply result from one individual from the core haplotype becoming a singleton due to a recombination event which 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.

In fact, 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 results in transitions in the partition scheme that result in a class of additional non-singleton haplotypes, and their eventual breakdown due to further recombination events lower in the genealogy.

Standing Sweep Case

In the case of the standing sweep, the dynamics of the partition scheme in the vicinity of the selected site are dominated by events occurring during the standing phase, while at greater distances the partition scheme is dominated by recombination events occurring in the sweep phase.

2.2 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.

2.2.1 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.

Gene conversion –minor point in discussion too.

3 Discussion

4 Acknowledgements

5 Methods

6 Supplementary materials