The Genealogy of Sequences Containing Multiple Sites Subject to Strong Selection in a Subdivided Population

Magnus Nordborg[1] and Hideki Innan[2]

Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089-1340

Manuscript received August 7, 2002
Accepted for publication December 17, 2002

ABSTRACT

A stochastic model for the genealogy of a sample of recombining sequences containing one or more sites subject to selection in a subdivided population is described. Selection is incorporated by dividing the population into allelic classes and then conditioning on the past sizes of these classes. The past allele frequencies at the selected sites are thus treated as parameters rather than as random variables. The purpose of the model is not to investigate the dynamics of selection, but to investigate effects of linkage to the selected sites on the genealogy of the surrounding chromosomal region. This approach is useful for modeling strong selection, when it is natural to parameterize the past allele frequencies at the selected sites. Several models of strong balancing selection are used as examples, and the effects on the pattern of neutral polymorphism in the chromosomal region are discussed. We focus in particular on the statistical power to detect balancing selection when it is present.

COALESCENT theory is based on the realization that selective neutrality allows the separation of descent from state. This makes it possible to model samples (or populations) as random genealogies with superimposed neutral mutations (see Nordborg 2001). Neutrality is thus fundamental to this approach. Nonetheless, selection has been successfully incorporated in two very different ways.

One way is to construct a genealogical process that "leaves room" for selection by creating genealogies that contain "virtual" branches representing possible lines of descent. After mutations have been superimposed and the state transmitted through each branch is known, the genealogy is "pruned" by preferentially removing selectively inferior branches so that only "actual" lines of descent remain. The process known as "ancestral selection graph" (Krone and Neuhauser 1997; Neuhauser and Krone 1997) accomplishes this (see also Donnelly and Kurtz 1999; Neuhauser 1999; Slade 2000a,b, 2001; Fearnhead 2001). An important feature of this approach, which can be seen as a natural extension of the standard neutral coalescent, is that all selection coefficients must be scaled using the standard coalescent/diffusion scaling. This means that it is not possible to model arbitrarily strong selection this way: From a mathematical point of view, infinitely strong selection would require infinitely many virtual branches; from a practical point of view, simulation of strong selection becomes extremely inefficient because of the large number of virtual branches.

The second way, first described in the context of the coalescent by Kaplan *et al.* (1988), utilizes the fact that a polymorphic population can be thought of as subdivided into allelic classes, within which no selection occurs. Genealogies can then be modeled using existing models of geographic subdivision, with mutation (and, in a sense, recombination) taking the place of migration. It is necessary to know the past (relative) sizes of the "subpopulations," *i.e.*, of the allelic classes. Thus the approach may be seen as modeling genealogies *conditional* on the past frequencies of the selectively different alleles (Nordborg 1999, 2001). However, since it is in general not known how to obtain the *unconditional* process when the past allele frequencies are random variables, the approach has been used only for strong selection, when it may be reasonable to model the selective dynamics deterministically. Examples include balancing selection (see Hudson and Kaplan 1988; Takahata 1990; Hey 1991; Kaplan *et al.* 1991; Nordborg 1997, 1999; Takahata and Satta 1998; Kelly and Wade 2000; Schierup *et al.* 2001; Barton and Navarro 2002; Navarro and Barton 2002), positive selection (or "selective sweeps," see Kaplan *et al.* 1989; Hudson *et al.* 1994; Braverman *et al.* 1995; Simonsen *et al.* 1995; Kim and Stephan 2002; Przeworski 2002), and purifying selection (or "background selection," see Hudson and Kaplan 1995; Nordborg 1997; Campbell 1999).

The two approaches are in a sense complementary: Whereas the ancestral selection graph works only for

[1]*Corresponding author:* Molecular and Computational Biology, University of Southern California, SHS 172, 835 W. 37th St., Los Angeles, CA 90089-1340. E-mail: magnus@usc.edu

[2]*Present address:* Human Genetics Center, School of Public Health, University of Texas Health Science Center, 1200 Hermann Pressler, Houston, TX 77030.

weak selection, the conditional approach works only for strong selection. How to connect the two is not clear.

In this article we use the second, conditional approach, to show how the original formulations of Kaplan *et al.* (1988) and Hudson and Kaplan (1988) may be extended to model selection at multiple sites, with the possibility of different selection coefficients in different subpopulations (local adaptation). This has been done before (Takahata 1990; Kaplan *et al.* 1991; Nordborg 1997, 2001; Kelly and Wade 2000; Schierup *et al.* 2000; Barton and Navarro 2002); however, most treatments have considered the genealogy of a nonrecombining site linked to the selected loci. In contrast, we consider the genealogy of the entire region. This makes it possible to ask statistical questions about the pattern of polymorphism (see also Innan and Tajima 1999; Schierup *et al.* 2001; Kim and Stephan 2002; Przeworski 2002). In particular we are interested in the conditions under which we would expect to be able to detect balancing selection.

A BASIC MODEL

Consider a diploid, hermaphroditic population consisting of P patches, each of which harbors a constant, large number of adult individuals, N_k, $k = 1, 2, \ldots, P$. Let $N = \Sigma N_k$ be the total population size, and define $c_k = N_k/N$. The population reproduces in discrete, nonoverlapping generations according to a generalized Wright-Fisher model in the following manner. Each individual produces an (effectively) infinite number of gametes. Male gametes (*e.g.*, pollen) flows between the patches; let f_{kl} be the probability that a male gamete produced in patch k ends up in patch l. After migration, gametes unite randomly to form zygotes. The number of immature individuals in each patch is thus still effectively infinite, but only a finite number (N_k in the kth patch) reach adulthood. The probability that a given individual survives is determined by its genotype and the genotypic frequencies in the patch. Generalizations of this model are discussed below.

Forward dynamics at the selected loci: Let H be the number of different haplotypes with respect to the selected locus or loci (which are assumed to be linked). Label the haplotypes $1, 2, \ldots, H$, and the $G = H(H + 1)/2$ different genotypes by pairs of indices $\{i, j\}$, $i \le j = 1, 2, \ldots, H$, according to their haplotypic composition. Let $N_{ij,k}(t)$ be the number of adults with genotype $\{i, j\}$ in patch k in generation t. Note that whereas $N_{ij,k}(t)$ is a random variable, N_k is not. The frequency of genotype $\{i, j\}$ in patch k is

$$x_{ij,k}(t) = \frac{N_{ij,k}(t)}{N_k},$$

and the frequency of haplotype i in patch k is

$$y_{i,k}(t) = \frac{1}{2}\left(\sum_{j=1}^{i} x_{ji,k}(t) + \sum_{j=1}^{H} x_{ij,k}(t)\right).$$

Let $y'_{i,k}(t)$ be the frequency of haplotype i among the gametes produced in patch k, generation t. In general, these gamete frequencies will be functions of the adult genotype frequencies and the appropriate segregation, mutation, and recombination parameters. Let

$$y''_{i,k}(t) = \frac{\sum_{l=1}^{P} y'_{i,l}(t)\, c_l f_{lk}}{\bar{f}_k(t)}$$

be the frequency of haplotype i among the male gametes in patch k after migration (female frequencies are of course unaffected by migration), where

$$\bar{f}_k(t) = \sum_{j=1}^{H}\sum_{l=1}^{P} y'_{j,l}(t)\, c_l f_{lk}.$$

Zygotes are then formed by random union of gametes within patches, so the frequency of genotype $\{i, j\}$ among the zygotes in patch k is

$$x''_{ij,k}(t) = \begin{cases} y'_{i,k}(t)\, y''_{i,k}(t), & i = j, \\ y'_{i,k}(t)\, y''_{j,k}(t) + y'_{j,k}(t)\, y''_{i,k}(t), & i \neq j. \end{cases}$$

Let the relative viability of a zygote with genotype $\{i, j\}$ in patch k be $1 - w_{ij,k}(t)$, and define

$$x'''_{ij,k}(t) = \frac{x''_{ij,k}(t)\,[1 - w_{ij,k}(t)]}{\bar{w}_k(t)},$$

where

$$\bar{w}_k(t) = 1 - \sum_{i \le j} x''_{ij,k}(t)\, w_{ij,k}(t).$$

The next generation of adults in patch k is formed by drawing N_k individuals according to these "postselection" frequencies. Thus, conditional on the genotype frequencies among adults in generation t, the length-G vector

$$[N_{11,k}(t + 1) \quad N_{12,k}(t + 1) \quad \ldots \quad N_{HH,k}(t + 1)]$$

is multinomially distributed with parameters N_k and

$$[x'''_{11,k}(t) \quad x'''_{12,k}(t) \quad \ldots \quad x'''_{HH,k}(t)].$$

Genealogy of the surrounding segment: Consider a chromosomal segment that contains the selected locus or loci. Take a particular copy of this segment, sampled from the adult individuals in generation $t + 1$. With respect to geographic location, it belongs to one of the P patches, and with respect to the selected locus or loci, it belongs to one of the H haplotypes.

Trace the genealogy of this segment one generation back in time. Each nucleotide in the segment in the current generation is a copy of the homologous nucleotide in some parental segment in the previous generation. In the absence of recombination, all nucleotides must have the same parental segment; otherwise, they may have different ones. However, rather than modeling nucleotides, it is convenient to think of the segment

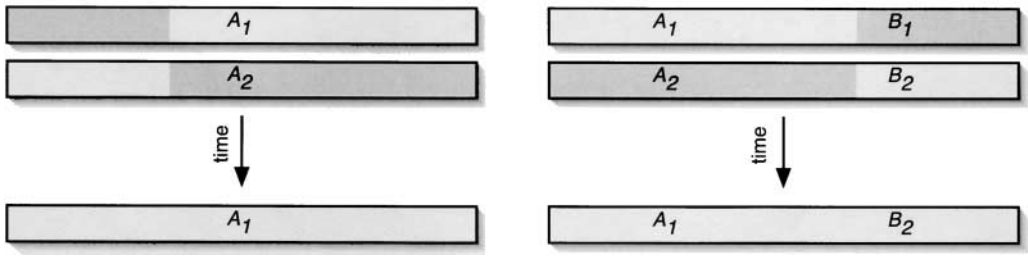

FIGURE 1.—Two examples of the effects of recombination on the single-generation genealogy of a segment. The colors denote ancestry (only). On the left, a segment with a single-locus A_1 haplotype was produced through recombination in an A_1/A_2 heterozygous individual. As a result, going back in time, one piece of the segment takes on the A_2 haplotype, whereas the other (containing the locus defining the haplotype) remains A_1. On the right, a segment with a two-locus A_1B_2 haplotype was produced through recombination in an A_1B_1/A_2B_2 doubly heterozygous individual, causing both pieces to change haplotype (going back in time).

abstractly as a continuous unit interval where each point may have a different genealogy. This makes mathematical and computational sense and does not entail any loss of biological generality (see, *e.g.*, NORDBORG 2001).

When tracing the genealogy of the segment back through the life cycle, the first thing to note is that selection cannot affect its state with respect to either patch or haplotype. The chosen segment must have been present in one of the gametes of the same type in the same patch before zygotes were formed and selection took place. However, migration can change the state of the segment with respect to patch, because the gamete may have been an immigrant, if it was male. The probability that a randomly chosen gamete of type i in patch k was male is $y''_{i,k}(t)/[y''_{i,k}(t) + y'_{i,k}(t)]$, and the probability that a male gamete of type i currently in patch k was produced in patch l is

$$\frac{y'_{i,l}(t)\, c_l f_{lk}}{\bar{f}_k(t)\, y''_{i,k}(t)} = \frac{y'_{i,l}(t)\, c_l f_{lk}}{\sum_{m=1}^{P} y'_{i,m}(t)\, c_m f_{mk}}. \qquad (1)$$

Note that migration always changes the state of the *entire* segment, because a gamete either does or does not migrate.

Having traced the segment through migration to the premigration gamete pool, the next step is to trace it through gamete production to the previous adult generation. Gamete production, *i.e.*, meiosis, can change the state of the segment with respect to haplotype both through mutation at one of the selected loci and through recombination inside the segment. If the gamete was a mutant, then the segment changes state to take on the haplotype of the parental segment. If the gamete was a recombinant, things are more complicated, because the segment then has two parental segments. At each breakpoint, the parentage of the offspring segment switches from one parental segment to the other. Going backward in time, the segment splits into two pieces (or sets of pieces if there was more than one breakpoint), one of which takes on the haplotype of the first parental segment while the other takes on the haplotype of the second parental segment (see Figure 1). Both parental segments consequently have to be followed if the genealogy of the original segment is to be traced farther back in time. As is demonstrated later, the backward transition probabilities through meiosis can readily be derived, but the expressions depend on the details of the model (*e.g.*, on the number of loci). In general, the probability that a gamete of type i in patch k resulted from a particular type of meiotic event in a particular genotype is a function of the length-G vector $[x_{ij,k}(t)]_{i\le j}$ and the recombination and mutation parameters.

Once the genotype and patch of the parental individual has been determined, all eligible individuals are equally likely to have been the parent. Segments can thus be seen as "picking" their parent randomly.

To trace the single-generation genealogy of n copies of the segment, note that, since infinitely many gametes are produced, and gametes unite randomly within patches, the fate of each segment is independent of the fates of all the other segments. In other words, each segment "picks" its parental segment (or segments, in case of recombination) independently of the other segments. Whenever two or more segments pick parental segments belonging to the same patch, haplotype, and genotype, two or more of them may pick the same one. Segments that pick the same parental segment are said to *coalesce*; if their genealogy is to be traced farther back in time, only the single segment needs to be followed.

Segments can of course pick only the same parental segment if they first pick the same parental individual. Since all individuals of the same genotype are equally likely to be picked, the probability that n segments all pick *different* parents, given that they all pick parents with genotype $\{i, j\}$ in patch k is

$$N_{ij,k}^{-n}(t) \prod_{m=0}^{n-1} (N_{ij,k}(t) - m) = \prod_{m=1}^{n-1} \left(1 - \frac{m}{N_{ij,k}(t)}\right). \qquad (2)$$

When two segments pick the same parental haplotype in the same parental individual, they coalesce with probability one-half if that individual is homozygous (so that there are two possible parental segments) and with probability one if it is heterozygous (so that only one segment is possible as parent).

COALESCENT APPROXIMATION

The purpose of the preceding section was to demonstrate that, conditional on the P length-G vectors $[N_{ij,k}$

$(t)]_{i\leq j}$, $k = 1, 2, \ldots, P$, for $t = 0, -1, \ldots$, *i.e.*, conditional on the genotype frequencies in all past generations, it is possible to model the genealogy of n segments sampled in generation $t = 0$ as a discrete-time Markov process running backward in time. However, the interesting process is the unconditional one, in which the genotype frequencies are governed by another discrete-time Markov process, running forward in time (as described above). Typically, we would like to know how the genealogy is affected by the *parameters* of that forward process (*e.g.*, the selection coefficients and population sizes), *not* how it is affected by a particular realization of it.

Clearly, the unconditional process could be studied through discrete-time simulation: One would simply simulate the genotype frequencies forward in time and then simulate the genealogy backward in time, conditional on those frequencies.

An alternative approach, which is taken here, is to use a coalescent/diffusion approximation and assume that the genotype frequencies can be treated as having evolved deterministically on the continuous timescale (Kaplan *et al.* 1988). It follows from the standard diffusion arguments of population genetics (see, *e.g.*, Neuhauser 2001) that this may be justified if selection is sufficiently strong relative to the inverse of the population size (or, in the present case, patch sizes). It should be stated clearly that this approach is not mathematically rigorous, but it is likely that it can be made rigorous for some scenarios and that it will be a reasonable approximation for many others.

We focus on strong balancing selection in the following because the approach is easiest to explain and justify for such models (with balancing selection, the model is very close to that of Barton and Navarro 2002). Balancing selection is here simply meant as any form of selection that tends to maintain all genotypes at constant, nonzero frequencies, which we denote $\hat{x}_{ij,k}$. In a finite population, the actual frequencies in any given generation will of course differ from these values. Similarly, in the diffusion approximation, the actual frequencies at a given point in time will differ from their expectations. The differences will tend to be smaller the stronger the selection. Note that nothing in the basic model described above limits how strong selection may be. Indeed, it is possible to assume infinitely strong selection (*i.e.*, the selection coefficients need not be scaled in the diffusion approximation), in which case

$$x_{ij,k}(t) = \hat{x}_{ij,k}, \quad \forall i, j, k, t, \tag{3}$$

and treating the genotype frequencies as constant is evidently justified. However, as discussed in Nordborg (1999), infinitely strong selection significantly complicates the algebra (because it causes deviations from Hardy-Weinberg equilibrium) without significantly affecting the results, and we therefore do not assume that selection is infinitely strong, but that it is nonetheless sufficiently strong for Equation 3 to hold to a sufficiently good approximation in what follows. Note that Equation 3 implies $x'''_{ij,k} = x_{ij,k} = \hat{x}_{ij,k}$, and similarly for the haplotype frequencies.

Scaling: Under the assumption that the genotype frequencies can be treated deterministically, it is possible to use standard arguments to find a continuous-time coalescent approximation for the discrete-time genealogical process described above. Note that the probability (2) can be rewritten

$$\prod_{m=1}^{n-1}\left(1 - \frac{m}{Nc_k\hat{x}_{ij,k}}\right) = 1 - \frac{\binom{n}{2}}{Nc_k\hat{x}_{ij,k}} + O\left(\frac{1}{N^2}\right). \tag{4}$$

Thus, a coalescence event occurs with probability $O(1/N)$ per discrete generation, and it is natural to turn the process into a continuous-time process by rescaling time in units of $O(N)$ generations and letting N go to infinity (while keeping c_k constant to ensure that all the patches become large). The standard scaling of $2N$ is used throughout.

The per-generation probabilities of migration, mutation, and recombination are also assumed to be $O(1/N)$, and the corresponding scaled parameters are introduced. Thus, it is assumed that the migration probability f_{kl} can be written

$$f_{kl} = \frac{\phi_{kl}}{4N} + O\left(\frac{1}{N^2}\right), \quad k \neq l,$$

where ϕ_{kl} is the migration *rate* (recombination and mutation are introduced below). Similarly, it is assumed that all the selection coefficients are $O(1/N)$ (but still large enough for Equation 3 to hold approximately; see above). Taken together, these assumptions ensure that, to $O(1/N)$,

$$x''_{ij,k} \approx x'_{ij,k} \approx \hat{x}_{ij,k},$$

and that, to the same order of approximation,

$$\hat{x}_{ij,k} \approx \begin{cases} \hat{y}^2_{i,k}, & i = j, \\ 2\hat{y}_{i,k}\hat{y}_{j,k}, & i \neq j. \end{cases} \tag{5}$$

To proceed farther we must consider specific genetic models.

Single-locus example: Consider a segment that contains a locus (or site) that is maintained polymorphic for two alleles by strong balancing selection in a subdivided environment. There are $H = 2$ "haplotypes," A_1 and A_2, and three genotypes, A_1A_1, A_1A_2, and A_2A_2. Let a_{ij} be the probability that allele A_i mutates to A_j during meiosis, and let r be the probability of a recombination event. It is assumed that these can be written

$$a_{ij} = \frac{\alpha_{ij}}{4N} + O\left(\frac{1}{N^2}\right), \quad i \neq j,$$

and

$$r = \frac{\rho}{4N} + O\left(\frac{1}{N^2}\right),$$

where α_{ij} and ρ are the mutation and recombination rates. This model is identical to previously published models (Hudson and Kaplan 1988; Kaplan *et al.* 1988, 1991; Hey 1991; Nordborg 1997), except that it considers the genealogy of a segment rather than of a point.

Consider the backward transitions for a single segment with haplotype i sampled from an individual in patch k, as before. Because the probabilities of migration, mutation, and recombination are all assumed to be small, it is clear that the segment is most likely to be a copy of a segment with the same haplotype in the same patch in the previous generation. Indeed, it is easy to show that the probability is $1 - O(1/N)$. Furthermore, given that its state with respect to haplotype and patch did not change, the probability that it was produced by an $\{i, j\}$ individual is $\hat{y}_{j,k} + O(1/N)$. It can also be shown that the segment is an immigrant of the same type from patch l with probability

$$\frac{1}{2} \cdot \frac{c_l \hat{y}_{i,l}}{c_k \hat{y}_{i,k}} f_{lk} + O\left(\frac{1}{N^2}\right),$$

a mutant haplotype $j \neq i$ from the same patch with probability

$$\frac{\hat{y}_{j,k}}{\hat{y}_{i,k}} a_{ji} + O\left(\frac{1}{N^2}\right),$$

and a recombinant from an $\{i, j\}$ individual in the same patch with probability

$$\hat{y}_{j,k} r + O\left(\frac{1}{N^2}\right).$$

All other transitions (*e.g.*, the segment is an immigrant mutant) have probability $O(1/N^2)$ or less.

As discussed above, it is simple to extend to a sample of n segments, because the single-generation transitions are mutually independent. Only the coalescence probabilities need to be determined. From Equation 4 a single coalescence event has probability $O(1/N)$ or less, which means that the probability of more than two segments coalescing in a single generation has probability $O(1/N^2)$ or less. It also means that the probability that a segment involved in a coalescence is a migrant, mutant, or recombinant is $O(1/N^2)$. The only coalescence event that has probability $O(1/N)$ is thus between two segments that do not change state with respect to haplotype or patch. For simplicity, consider the probability that $n =$ two segments with haplotype i in patch k coalesce. To do so, they must have been produced by individuals of the same genotype, by the same individual of that genotype, and by the same segment within that individual. The probability of this is

$$\hat{y}_{i,k}^2 \cdot \frac{1}{N c_k \hat{x}_{ii,k}} \cdot \frac{1}{2} + \hat{y}_{j,k}^2 \cdot \frac{1}{N c_k \hat{x}_{ij,k}} \cdot 1 + O\left(\frac{1}{N^2}\right) = \frac{1}{2 N c_k \hat{y}_{i,k}} + O\left(\frac{1}{N^2}\right),$$

where $j \neq i$, and Equation 5 has been used. Generally, it can be shown that the probability that a pair out of n segments of type i in patch k coalesces is

$$\frac{\binom{n}{2}}{2 N c_k \hat{y}_{i,k}} + O\left(\frac{1}{N^2}\right).$$

Given these transition probabilities, the limiting coalescent process can be derived using standard arguments. Segments belong to states with respect to haplotype and patch as before. Measure time in units of $2N$ generations, and let N go to infinity. Then, independently of all other segments, each segment with haplotype i in patch k migrates to patch l at rate

$$\frac{c_l \hat{y}_{i,l}}{2 c_k \hat{y}_{i,k}} \phi_{lk}/2,$$

mutates to haplotype $j \neq i$ at rate

$$\frac{\hat{y}_{j,k}}{\hat{y}_{i,k}} \alpha_{ji}/2,$$

and recombines with a j haplotype at rate $\hat{y}_{j,k}\rho/2$. If there are currently n segments with haplotype i in patch k, the total rate of migration to l is thus

$$n \frac{c_l \hat{y}_{i,l}}{2 c_k \hat{y}_{i,k}} \phi_{lk}/2,$$

etc. Similarly, each pair of segments with haplotype i in patch k independently coalesces at rate $1/(c_k \hat{y}_{i,k})$, so that the total rate for n such segments is

$$\frac{\binom{n}{2}}{c_k \hat{y}_{i,k}}.$$

Two-locus example: Consider a model with two loci each with two alleles, A_1/A_2 and B_1/B_2. There are $H = 4$ haplotypes: A_1B_1, A_1B_2, A_2B_1, and A_2B_2, which are numbered 1, . . . , 4 in the order listed, and 10 genotypes. Define the mutation probability at the B-locus, b_{ij}, analogously to a_{ij}. The recombination probability, r, is also defined as before, but let dr be the length of the part that lies between the two loci (which are loci in the strict sense of the word—*i.e.*, there is no recombination within them; they can be thought of as single-nucleotide polymorphisms, for example).

The single-generation backward transition probabilities can be found in the same manner as in the single-locus model, but are in some cases more complicated. The probability that a segment of type i in patch k is an immigrant of the same type from patch l is

$$\frac{1}{2} \cdot \frac{c_l \hat{y}_{i,l}}{c_k \hat{y}_{i,k}} f_{lk} + O\left(\frac{1}{N^2}\right), \tag{6}$$

as before. The mutation probabilities depend on the number and type of mutations involved. Thus the probability that a segment of type $i = 1$ (A_1B_1) in patch k is a mutant type $j \neq i$ from the same patch is

$$\begin{cases} \dfrac{\hat{y}_{2,k}}{\hat{y}_{1,k}} b_{21} + O\left(\dfrac{1}{N^2}\right), & j = 2\ (A_1B_{21}), \\ \dfrac{\hat{y}_{3,k}}{\hat{y}_{1,k}} a_{21} + O\left(\dfrac{1}{N^2}\right), & j = 3\ (A_2B_1), \\ O\left(\dfrac{1}{N^2}\right), & j = 4\ (A_2B_2), \end{cases} \tag{7}$$

for example. The recombination probabilities depend on the genotype in which the recombination event would have taken place. Recombination in homozygotes or single heterozygotes can be treated as in the single-locus model, but recombination in double heterozygotes cannot. Consider again a segment of type $i = 1$ (A_1B_1) in patch k. The probability that it is a recombinant from an $\{1, j\}$ individual in the same patch is

$$\hat{y}_{j,k} r + O\left(\frac{1}{N^2}\right), \tag{8}$$

for $j = 1, 2, 3$. However, it can also be a recombinant from a $\{1, 4\}$ individual (the *cis*-heterozygote, A_1B_1/A_2B_2), *as long as the recombination did not take place between the two loci*. The probability of this event is thus

$$\hat{y}_{4,k}(1 - d) r + O\left(\frac{1}{N^2}\right). \tag{9}$$

Given such an event, the breakpoint can be anywhere in the flanking pieces (it could, for example, be uniformly distributed). The opposite is true if the segment is a recombinant from a $\{2, 3\}$ individual (the *trans*-heterozygote, A_1B_2/A_2B_1), because in this case the recombination must have taken place between the loci. The probability of this is

$$\frac{\hat{y}_{2,k}\hat{y}_{3,k}}{\hat{y}_{1,k}} dr + O\left(\frac{1}{N^2}\right). \tag{10}$$

All other events have probability $O(1/N^2)$ or less.

The extension to n segments and the conversion to continuous time can be done precisely as for the single-locus model. The complete transition rates are given in Table 1.

DETECTING SELECTION

In this section we use simulations to investigate the distribution of the pattern of polymorphism in regions that contain sites subject to balancing polymorphism. We are in particular interested in the power to detect selection under various models and assumptions about parameter values. The simulation software, which is written in C++ with a Mathematica front end, is available on request.

Symmetric single-locus model: Consider first a "classical" balancing selection model, in which selection maintains two different alleles at high frequencies in a random-mating population. The precise values of the equilibrium frequencies do not matter greatly (we assume that selection is completely symmetric so that the sizes of the two allelic classes, A_1 and A_2, are even), but the assumption that there is no subdivision does, as we see later. We also assume mutation between the two allelic classes is rare and symmetric (*i.e.*, A_1 mutates to A_2 at the same rate as A_2 mutates to A_1). This assumption is discussed further below as well.

Figure 2 shows a typical realization of this model. The time to the most recent common ancestor (MRCA) of the sample is extremely high at the selected site and decreases with distance. The reason for this is clear: The selected site itself cannot coalesce unless a mutation from one allelic class to the other occurs. Since mutations are rare, this means that, almost always, all members of a particular class will coalesce to the MRCA of that class long before a mutation occurs. Sites that are more distantly linked to the selected site can move between allelic classes by recombination and coalesce much faster.

The genealogical pattern shown in Figure 2 may result in a characteristic pattern of polymorphism, which may be detected in the data. Figure 3 shows the distribution of the amount of variation within and between allelic classes and of Tajima's D statistic along the chromosome in six realizations. It is evident that the presence of balancing selection usually causes a "peak" due to divergence between the two allelic classes. This well-known phenomenon is also reflected in Tajima's D, which is expected to be greater than zero in the presence of balancing selection or (many forms of) population subdivision (Tajima 1989). However, note the considerable randomness: The peaks are not always centered on the selected site; in Figure 3b, variation is inflated within one of the allelic classes as well as between them; and in Figure 3e, there is no peak at all. The pattern of polymorphism depends heavily on the history of mutations between the allelic classes, as well as on the history of recombination in the region.

We considered the power to detect selection using Tajima's D statistic. Table 2 gives the probability of observing a significantly positive value of this statistic under various assumptions. Several conclusions are clear. First, the power depends sensitively on the width of the window used to calculate D. A very small window will not have enough segregating sites to achieve statistical significance, whereas a large window will "drown" the peak in neutral noise. The optimal window width will depend on the ratio between θ and ρ, *i.e.*, the number of neutral mutations per recombination. If $\theta/\rho > 1$, balancing selection should usually be detected, but if $\theta/\rho \ll 1$, it usually will not be.

To make the numbers in Table 2 applicable to sequence data, it is necessary to make assumptions about θ and ρ per base. Assume that $\theta = 0.01$ per site, as is reasonable for *Drosophila melanogaster* (Przeworski *et al.* 2001). Then the regions simulated in Figure 3 and

TABLE 1

Backward transition rates in the two-locus model

Event	Present state			
	A_1B_1 in k	A_1B_2 in k	A_2B_1 in k	A_2B_2 in k
Migration to patch l	$\frac{1}{2}\frac{c_l\hat{y}_{1,l}}{c_k\hat{y}_{1,k}}\phi_{lk}/2$	$\frac{1}{2}\frac{c_l\hat{y}_{2,l}}{c_k\hat{y}_{2,k}}\phi_{lk}/2$	$\frac{1}{2}\frac{c_l\hat{y}_{3,l}}{c_k\hat{y}_{3,k}}\phi_{lk}/2$	$\frac{1}{2}\frac{c_l\hat{y}_{4,l}}{c_k\hat{y}_{4,k}}\phi_{lk}/2$
Mutation to A_1B_1	NA	$\frac{\hat{y}_{1,k}}{\hat{y}_{2,k}}\beta_{12}/2$	$\frac{\hat{y}_{1,k}}{\hat{y}_{3,k}}\alpha_{12}/2$	0
Mutation to A_1B_2	$\frac{\hat{y}_{2,k}}{\hat{y}_{1,k}}\beta_{21}/2$	0	NA	$\frac{\hat{y}_{2,k}}{\hat{y}_{4,k}}\alpha_{12}/2$
Mutation to A_2B_1	$\frac{\hat{y}_{3,k}}{\hat{y}_{1,k}}\alpha_{21}/2$	NA	0	$\frac{\hat{y}_{3,k}}{\hat{y}_{4,k}}\beta_{12}/2$
Mutation to A_2B_2	0	$\frac{\hat{y}_{4,k}}{\hat{y}_{2,k}}\alpha_{21}/2$	$\frac{\hat{y}_{4,k}}{\hat{y}_{3,k}}\beta_{21}/2$	NA
Recombination with A_1B_1	$\hat{y}_{1,k}\rho/2$	$\hat{y}_{1,k}\rho/2$	$\hat{y}_{1,k}\rho/2$	See below
Recombination with A_1B_2	$\hat{y}_{2,k}\rho/2$	$\hat{y}_{2,k}\rho/2$	See below	$\hat{y}_{2,k}\rho/2$
Recombination with A_2B_1	$\hat{y}_{3,k}\rho/2$	See below	$\hat{y}_{3,k}\rho/2$	$\hat{y}_{3,k}\rho/2$
Recombination with A_2B_2	See below	$\hat{y}_{4,k}\rho/2$	$\hat{y}_{4,k}\rho/2$	$\hat{y}_{4,k}\rho/2$
Recombination in A_1B_1/A_2B_2	$\hat{y}_{4,k}(1-d)\rho/2$	$\frac{\hat{y}_{1,k}\hat{y}_{4,k}}{\hat{y}_{2,k}}d\rho/2$	$\frac{\hat{y}_{1,k}\hat{y}_{4,k}}{\hat{y}_{3,k}}d\rho/2$	$\hat{y}_{1,k}(1-d)\rho/2$
Recombination in A_1B_2/A_2B_1	$\frac{\hat{y}_{2,k}\hat{y}_{3,k}}{\hat{y}_{1,k}}d\rho/2$	$\hat{y}_{3,k}(1-d)\rho/2$	$\hat{y}_{2,k}(1-d)\rho/2$	$\frac{\hat{y}_{2,k}\hat{y}_{3,k}}{\hat{y}_{4,k}}d\rho/2$
Coalescence (per pair)	$\frac{1}{c_k\hat{y}_{1,k}}$	$\frac{1}{c_k\hat{y}_{2,k}}$	$\frac{1}{c_k\hat{y}_{3,k}}$	$\frac{1}{c_k\hat{y}_{4,k}}$

NA, not applicable.

Table 2 correspond to 1 kb, and the optimal window size is usually 100 bp. Balancing selection affects very small regions. This realization calls into question the infinite-sites assumption for mutation, because, given the very long coalescence times expected under balancing selection, it is no longer reasonable to assume that each site is hit only once. Depending on the level of selective constraint, at most 1000 selectively neutral sites are in the region, and more likely 1000/3. When finite sites are taken into account, balancing selection be-

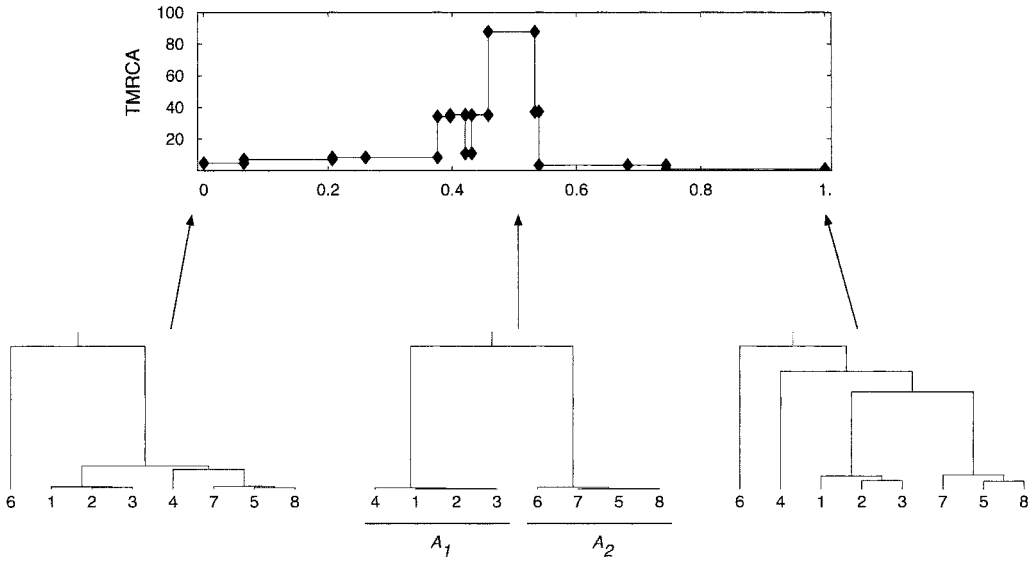

FIGURE 2.—An example of the genealogical pattern around a selected site (positioned in the center of the region, at 0.5). The sample size is $n = 8$, with 1–4 belonging to the A_1 allelic class and 5–8 belonging to the A_2 allelic class. The recombination rate, ρ, for the whole region is 2. The mutation rate at the selected site, α, is 0.01. Note that the trees for different regions are drawn on very different scales.

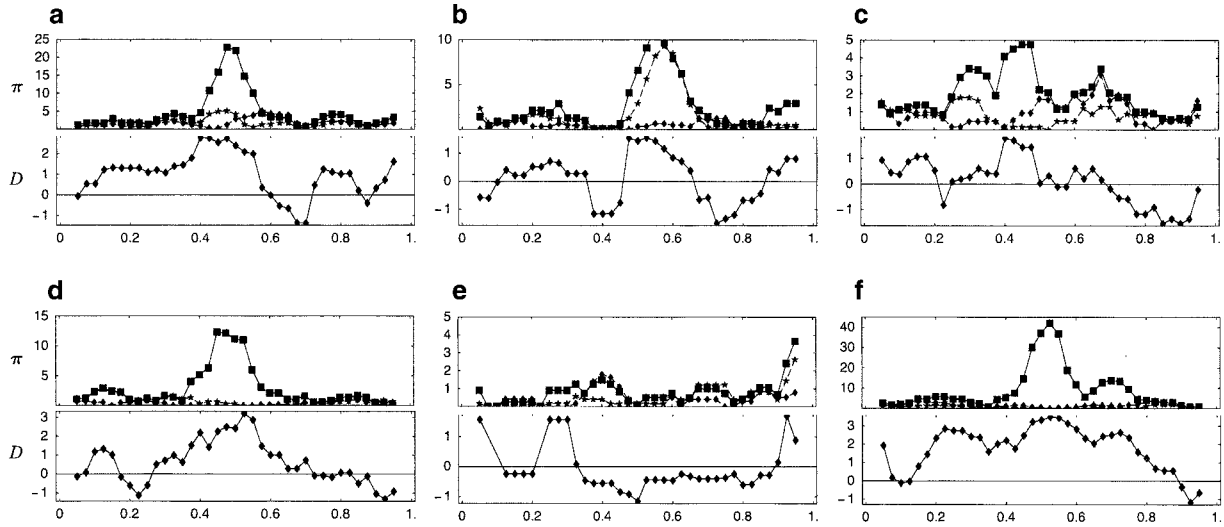

FIGURE 3.—Sliding-window analysis of the distribution of the average number of pairwise differences within and between the two different allelic classes (π_w and π_b, respectively) and of Tajima's D in six realizations of the symmetric balancing selection model described in the text. In the top, diamonds and stars connected by broken lines show the distributions of π_w for the two allelic classes, and squares connected by solid lines show the distribution of π_b. The bottom shows the distribution of Tajima's D. Simulations were carried out with $n = 24$ (12 in A_1 and 12 in A_2) and the infinite-sites recombination/mutation model with $\theta = \rho = 10$. A sliding-window of size 0.1 was moved with increments of 0.025.

comes harder to detect, because some of the "excess" variability simply results in repeat mutations (Table 2).

Loss-of-function mutations: Many cases of balancing selection, especially those that are due to a trade-off between resistance and cost of resistance to some parasite or pathogen, are likely to involve loss-of-function mutations (OLSON 1999; STAHL *et al.* 1999; JOHANSON *et al.* 2000; TIAN *et al.* 2002). This type of scenario fits well into the modeling framework presented here, but leads to very different predictions. Specifically, since mutation between the two allelic classes is essentially unidirectional (for example, if mutation at any of 100 different sites can lead to loss of function, then the total rate of loss-of-function mutations would be 1, assuming that the mutation rate per base pair is 0.01, while the rate of back mutation would still be 0.01), there is usually no ancient polymorphism, and no peak of polymorphism will develop. As is illustrated in Table 2, balancing selection is not likely to be detected in these cases. We return to this topic in the DISCUSSION.

Two loci: Figure 4 shows a typical realization of a classical symmetric two-locus model with all four haplotypes present at equal frequencies. The two selected sites are positioned at 0.4 and 0.6, respectively. Note that the topologies of the trees behave in the intuitively obvious way as we walk along the chromosome: Close to each selected site, the sample must coalesce in a manner determined by the allelic classes *at that site*. In

TABLE 2

The probability (%) of detecting balancing selection

	Width of window														
	Infinite sites					Finite sites (1000 bp)					Finite sites (333 bp)				
ρ	0.025	0.05	0.1	0.2	1	0.025	0.05	0.1	0.2	1	0.025	0.05	0.1	0.2	1
1	88.1	91.4	93.3	94.3	88.2	85.6	90.2	92.8	93.5	83.8	75.0	83.8	89.4	91.5	76.5
3	84.7	88.2	88.9	87.1	57.6	80.8	85.3	86.3	83.5	44.7	69.2	76.6	79.2	76.4	32.2
10	76.7	77.3	73.6	62.0	7.7	69.0	69.9	64.5	51.0	4.1	55.2	57.6	52.3	39.9	2.4
30	59.3	53.2	40.7	22.0	0.2	48.1	42.5	31.0	15.7	0.1	37.2	34.3	23.8	11.5	0.0
100	26.0	20.2	10.9	3.1	0.0	20.5	15.6	8.9	2.4	0.0	15.4	12.3	6.9	2.1	0.0
10[a]	3.7	6.1	6.8	5.2	0.3	3.7	6.0	6.8	5.1	0.3	3.2	5.6	6.8	5.1	0.2

Power was estimated from 5000 replicates (1000 for $\rho = 100$) using $\theta = 10$ and $n = 24$ (12 in each allelic class). Selection was deemed to have been detected if $D > 2.01$ (TAJIMA 1989) in a window of the specified size. In the finite-sites models, infinite-sites mutations (*i.e.*, random numbers in the unit interval) that were sufficiently close to each other were deemed to have affected the same site and simply reverted it to its previous state.

[a] Loss-of-function model, with rates of loss and reversion equal to 1 and 0.01, respectively.

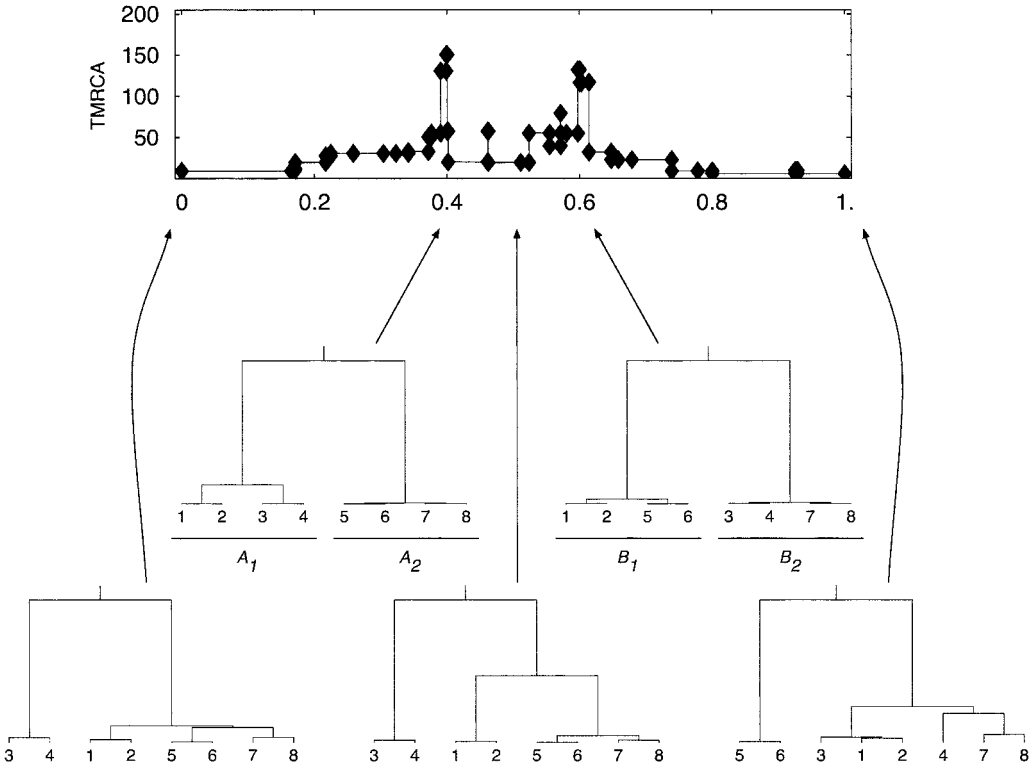

FIGURE 4.—An example of the genealogical pattern around two selected sites (located at positions 0.4 and 0.6). The sample size is $n = 8$, with 1–2 belonging to the A_1B_1 haplotypic class, 3–4 belonging to the A_1B_2 haplotypic class, 5–6 belonging to the A_2B_1 haplotypic class, and 7–8 belonging to the A_2B_2 haplotypic class. As in Figure 2, we have $\rho = 2$ and $\alpha = 0.01$ (for each selected site).

a sample that contains the "complementary" haplotypes (A_1B_1 and A_2B_2 or A_1B_2 and A_1B_2), sites located between the selected sites can coalesce only if there are at least two recombination events.

Figure 5 illustrates the effect of the distance between the selected sites on the distribution of π_w, π_b, and Tajima's D along the chromosome. If the distance between the sites is sufficiently great, there may be two distinct peaks (Figure 5, a and b); otherwise, only a single peak may be visible (Figure 5c). Power studies analogous to those in Table 2 indicate that the probability of rejecting neutrality using Tajima's D depends sensitively on the positioning, numbers, and sizes of the windows used (not shown). Obviously, the best strategy is to use a window size that captures each peak, but since the number of selected sites is not known *a priori*, this may be difficult to implement in practice. However, regions containing multiple sites subject to balancing selection are considerably less likely to be missed (see also NAVARRO and BARTON 2002).

Subdivision: The behavior of balancing selection models with population subdivision is very complicated, but the example shown in Figure 6 illustrates the main points. In addition to the structure imposed by the allelic and haplotypic classes, there is now population structure as well. Which structure turns out to be pre-

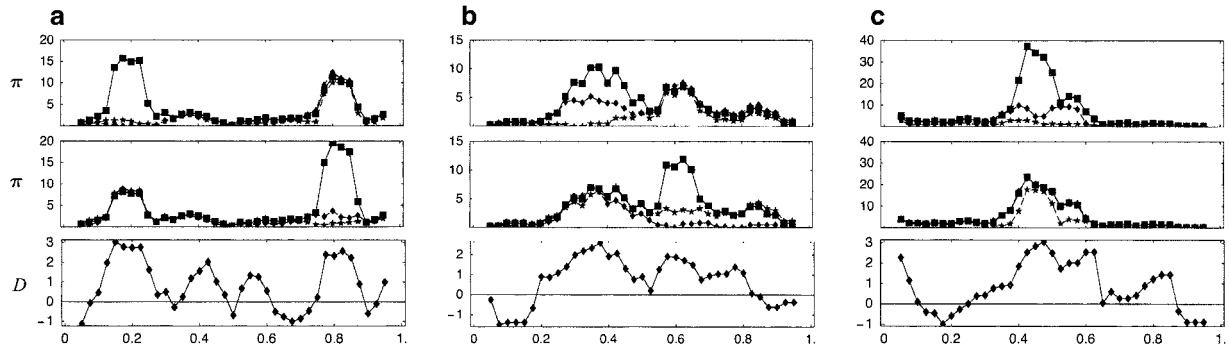

FIGURE 5.—Sliding-window analysis of the distribution of π_w, π_b, and of Tajima's D in three different realizations of the symmetric two-locus model. In the top, diamonds and stars connected by broken lines represent the distributions of π_w for the A-locus, and squares with solid lines represent the distributions of π_b. The middle shows the same for the B-locus. The bottom shows the distribution of Tajima's D. Simulations were carried out with $n = 24$ (6 A_1B_1, 6 A_1B_2, 6 A_2B_1, and 6 A_2B_2) and $\theta = \rho = 10$. The selected sites were located at (a) 0.2 and 0.8, (b) 0.4 and 0.6, and (c) 0.45 and 0.55. Sliding-window parameters were the same as in Figure 3.

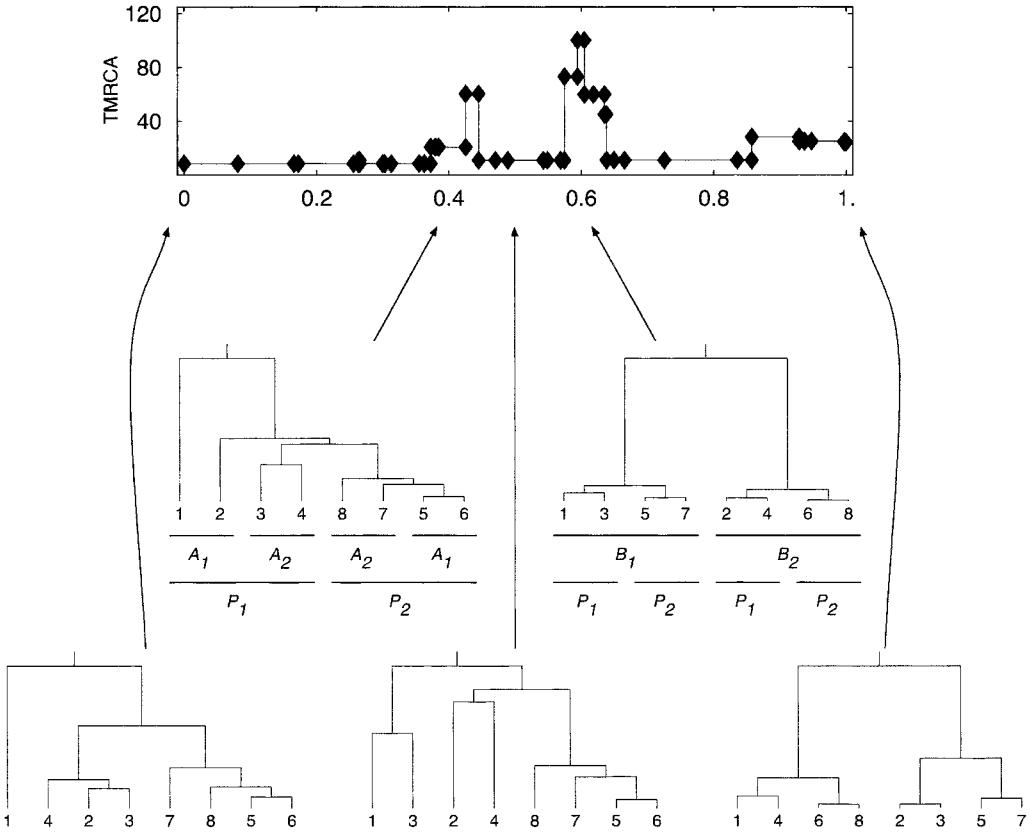

FIGURE 6.—An example of the genealogical pattern around two selected sites in a subdivided population. The parameters are as in Figure 4, except that the population is divided into two patches of equal size, connected with symmetric migration at rate $\phi = 0.1$. The structure of the sample is shown.

dominant will depend on the relative magnitudes of the parameters. In the example shown in Figure 6, the pattern of coalescence differs between the two selected sites. At the *B*-locus (located at 0.6), the first stage of coalescence is within allelic class within patches, followed by coalescence within allelic class between the two patches. Finally coalescence occurs between two allelic classes. On the other hand, at the *A*-locus (located at 0.4), there is a cluster of the four sequences in patch 2, and the longest branch is between sequence 1 (A_1 in patch 1) and the others. The genealogical pattern is highly variable between realizations.

Local adaptation: An important motivation for the model described above is to consider local adaptation. A balance between migration and selection seems much more likely to maintain polymorphism than does "pure" balancing selection. Because local adaptation leads to a deficit of heterozygotes at the selected locus or loci, the effect on linked neutral variation may be much greater. Local adaptation should thus be much easier to detect using polymorphism data.

Figure 7 shows an example of a single-locus model of local adaptation. The frequency of A_1 is assumed to be 0.9 in patch 1 and 0.1 in patch 2 (and conversely for A_2). The effects of local adaptation are even more dramatic when multiple sites are involved. Figure 8 shows an example of a two-locus model where A_1B_1 is favored in patch 1 and A_2B_2 is favored in patch 2. Note that the effects of selection extend throughout the simulated region, completely obscuring the peak around the *B*-locus. Multiple linked sites involved in local adaptation can, in principle, "lock up" entire chromosomal regions in complementary haplotypes.

DISCUSSION

We have shown how the "structured coalescent" described by NORDBORG (1997) may be combined with the "ancestral recombination graph" of GRIFFITHS and MARJORAM (1997) to yield a genealogical model for sequences that contain multiple sites subject to strong selection in a subdivided population. We refer to the combined model as the "structured ancestral recombination graph" (SARG). Albeit complex, the SARG is highly suitable for simulation, just like the standard coalescent.

Robustness of the SARG: We described the SARG as a limiting approximation to a specific model of pollen flow in an outcrossing hermaphroditic plant species, but the SARG is much more general. As is the case for the standard coalescent, phenomena such as selfing, separate sexes, or sex linkage, etc., can readily be incorporated (although formally proving convergence is likely to be both difficult and tedious; see NORDBORG and KRONE 2002).

The hardest problem from a mathematical point of view is also the most interesting from a biological point of view; namely, when is it reasonable to treat selection

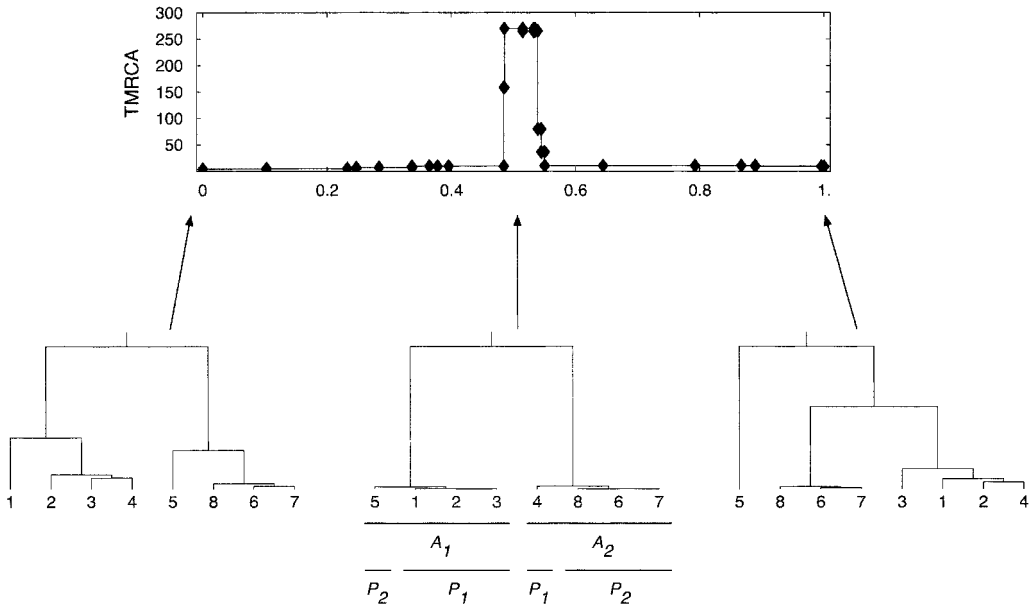

FIGURE 7.—An example of the genealogical pattern around a site involved in local adaptation (located at position 0.5). The selected site is located at position 0.5. The two equal-sized patches are connected with symmetric migration at rate $\phi = 0.1$; the rest of the parameters are as in Figure 2. The sample size is $n = 8$, with 1–3 belonging to A_1 and 4 belonging to A_2 in patch 1, and 5 belonging to A_1 and 6–8 belonging to A_2 in patch 2.

as population structure? For strong balancing selection, it is plausible to argue that the approximation is a good one (KAPLAN *et al.* 1988). In particular, when local adaptation is involved, it is easy to imagine very strong selection. However, to treat allele frequencies as constant in a model of local adaptation, it is also necessary to assume that migration is strong so that a deterministic migration-selection balance results. In such a model, there would be no signs of subdivision when looking at neutral markers unless these were sufficiently closely linked to the adaptively important sites. This may be the case in strong clines and even in some hybrid zones (although the approximation will certainly break down if the number of selected loci becomes too large; see BARTON and NAVARRO 2002).

Detecting balancing selection: Our motivation for deriving the SARG was that we wanted to simulate sequence data from regions containing sites subject to balancing selection. It has long been known that balancing selection may create a peak of increased polymor-

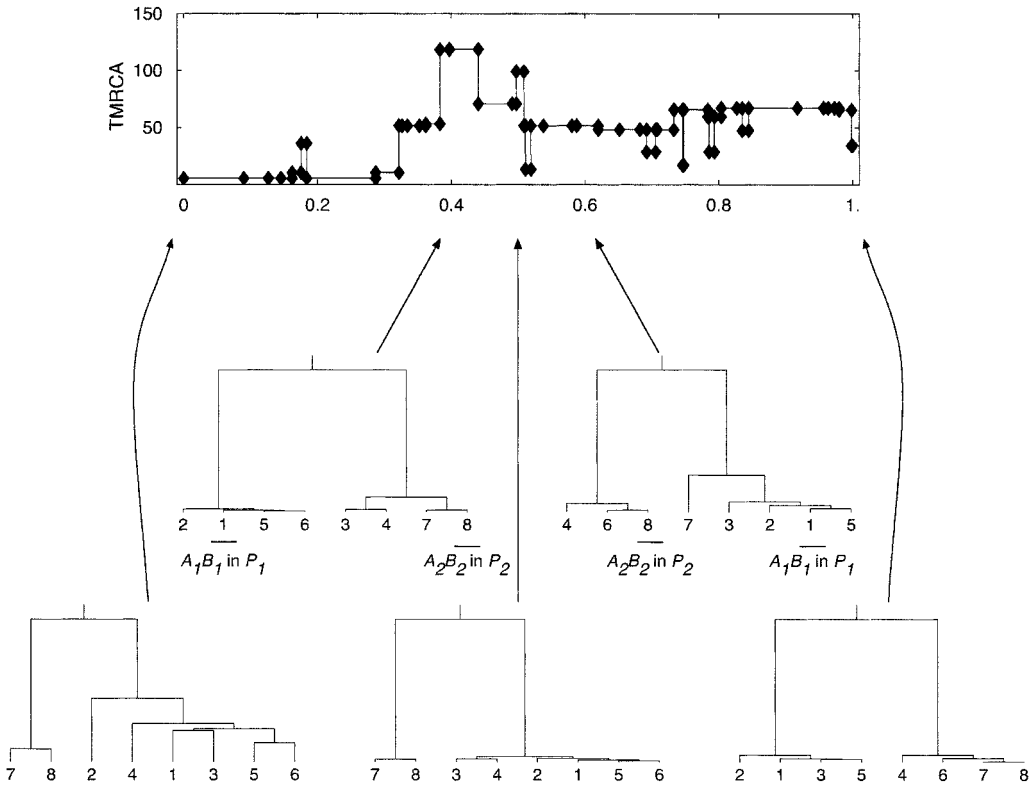

FIGURE 8.—An example of the genealogical pattern around a pair of sites involved in local adaptation (located at 0.4 and 0.6). The parameters are as in previous figures, except that the frequencies of A_1B_1, A_1B_2, A_2B_1, and A_2B_2 are 0.7, 0.1, 0.1, and 0.1 in patch 1 and 0.1, 0.1, 0.1, and 0.7 in patch 2, respectively. The sample size is $n = 8$, with one sequence from each haplotype in each patch.

phism centered around the selected site (Hudson and Kaplan 1988). Numerous articles have been published about the *expected* levels of polymorphism surrounding such a site (*e.g.*, Nordborg *et al.* 1996; Kelly and Wade 2000; Schierup *et al.* 2000; Barton and Navarro 2002). These kinds of results are of limited value for data analysis, because the pattern of polymorphism surrounding any particular balanced polymorphism will reflect the random history of this region and will usually be very far from expectations.

Using simulations, we have focused on the question of whether we should expect balancing selection to be detectable or not. Our original motivation for asking this question is that many years of population genetics research in Drosophila have failed to uncover strong evidence for balancing selection. Is this because balancing selection is indeed rare, perhaps limited to highly unusual cases, like the MHC and plant self-incompatibility loci (Kreitman and Akashi 1995; Hudson 1996), or is it simply very difficult to detect because recombination and gene conversion effectively destroys the evidence (Andolfatto and Nordborg 1998)? The latter view is supported by several recent observations of balancing selection in *Arabidopsis thaliana* (Stahl *et al.* 1999; Tian *et al.* 2002), where recombination is expected to be less effective because of selfing (Nordborg 1997).

Our simulations also tend to support the view that balancing selection might be difficult to detect in outcrossing organisms. For $\theta/\rho = 1$, as may be typical for Drosophila, taking finite sites into account, power to detect balancing selection seems to be $\sim$50% (Table 2). Note that this is power in the evolutionary sense, not in the usual "sampling" sense: It is not the case that a different sample would detect balancing selection; rather it is the case that a large fraction of existing balancing polymorphisms are expected to be undetectable (because of the particular evolutionary history of the polymorphism). Power will of course also depend on sampling and on the statistical method used, but perhaps less than we would think. It should be noted that our simulations do not take gene conversion into account: This could decrease power further (Andolfatto and Nordborg 1998).

Finally, it seems clear that we should not normally expect to be able to see a typical signal of balancing selection when selection maintains a polymorphism between loss-of-function and functional alleles. The reason is simply that loss-of-function alleles are created far too rapidly to be ancient. When loss-of-function alleles nonetheless appear to be ancient, as is the case for some disease-resistance loci in *A. thaliana* (Stahl *et al.* 1999; Tian *et al.* 2002), this may be an indication that not all loss-of-function alleles are created equal: It is notable that these cases involve complete deletions, and it may be that such deletions are preferable to point mutations that may lead to incorrectly folded proteins, for example. This could reduce the rate of loss-of-function mutations sufficiently for ancient polymorphism to be maintained.

We thank A. Navarro and an anonymous reviewer for comments on the manuscript. M.N. thanks Peter Donnelly, Bob Griffiths, Dick Hudson, Steve Krone, Tom Kurtz, Paul Marjoram, Claudia Neuhauser, Gesine Reinert, Simon Tavaré, and Carsten Wiuf for many, many conversations about selection.

LITERATURE CITED

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Barton, N. H., and A. Navarro, 2002 Extending the coalescent to multilocus systems: the case of balancing selection. Genet. Res. **79:** 129–139.

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 polymorphism. Genetics **140:** 783–796.

Campbell, R. B., 1999 The coalescent time in the presence of background fertility selection. Theor. Popul. Biol. **55:** 260–269.

Donnelly, P., and T. G. Kurtz, 1999 Genealogical processes for Fleming-Viot models with selection and recombination. Ann. Appl. Probab. **9:** 1091–1148.

Fearnhead, P., 2001 Perfect simulation from population genetic models with selection. Theor. Popul. Biol. **59:** 263–279.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, New York.

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

Hudson, R. R., 1996 Molecular population genetics of adaptation, pp. 291–309 in *Adaptation*, edited by M. R. Rose and G. V. Lauder. Academic Press, San Diego.

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, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

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

Innan, H., and F. Tajima, 1999 The effect of selection on the amounts of nucleotide variation within and between allelic classes. Genet. Res. **73:** 15–28.

Johanson, U., J. West, C. Lister, S. Michaels, R. Amasino *et al.*, 2000 Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science **290:** 344–347.

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.

Kaplan, N. L., R. R. Hudson and M. Iizuka, 1991 The coalescent process in models with selection, recombination and geographic subdivision. Genet. Res. **57:** 83–91.

Kelly, J. K., and M. J. Wade, 2000 Molecular evolution near a two-locus balanced polymorphism. J. Theor. Biol. **204:** 83–101.

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

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. **51:** 210–237.

Navarro, A., and N. H. Barton, 2002 The effects of multilocus balancing selection on neutral variability. Genetics **161:** 849–863.

Neuhauser, C., 1999 The ancestral graph and gene genealogy under frequency-dependent selection. Theor. Popul. Biol. **56:** 203–214.

Neuhauser, C., 2001 Mathematical models in population genetics,

pp. 153–178 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. J. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics **145:** 519–534.

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

Nordborg, M., 1999 The coalescent with partial selfing and balancing selection: an application of structured coalescent processes, pp. 56–76 in *Statistics in Molecular Biology and Genetics* (IMS Lecture Notes-Monograph Series, Vol. 33), edited by F. Seillier-Moiseiwitsch. Institute of Mathematical Statistics, Hayward, CA.

Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. J. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Nordborg, M., and S. M. Krone, 2002 Separation of time scales and convergence to the coalescent in structured populations, pp. 194–232 in *Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malécot*, edited by M. Slatkin and M. Veuille. Oxford University Press, Oxford.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser. B **263:** 1033–1039.

Olson, M. V., 1999 When less is more: gene loss as an engine of evolutionary change. Am. J. Hum. Genet. **64:** 18–23.

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

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 291–298.

Schierup, M. H., D. Charlesworth and X. Vekemans, 2000 The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. Genet. Res. **76:** 63–73.

Schierup, M. H., A. M. Mikkelsen and J. Hein, 2001 Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genes **159:** 1833–1844.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Slade, P. F., 2000a Most recent common ancestor probability distributions in gene genealogies under selection. Theor. Popul. Biol. **58:** 291–305.

Slade, P. F., 2000b Simulation of selected genealogies. Theor. Popul. Biol. **57:** 35–49.

Slade, P. F., 2001 Simulation of "hitch-hiking" genealogies. J. Math. Biol. **42:** 41–70.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman and J. Bergelson, 1999 Dynamics of disease resistance at the *Rpm1* locus of *Arabidopsis*. Nature **400:** 667–671.

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

Takahata, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

Takahata, N., and Y. Satta, 1998 Footprints of intragenic recombination at HLA loci. Immunogenetics **47:** 430–441.

Tian, D., H. Araki, E. Stahl, J. Bergelson and M. Kreitman, 2002 Signature of balancing selection in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **99:** 11525–11530.

Communicating editor: W. Stephan