Recombination, Balancing Selection and Phylogenies in MHC and Self-Incompatibility Genes

Mikkel H. Schierup, Anders M. Mikkelsen and Jotun Hein

Bioinformatics Research Center (BiRC), Department of Ecology and Genetics, University of Aarhus, 8000 Aarhus C., Denmark

Manuscript received May 23, 2001
Accepted for publication October 1, 2001

ABSTRACT

Using a coalescent model of multiallelic balancing selection with recombination, the genealogical process as a function of recombinational distance from a site under selection is investigated. We find that the shape of the phylogenetic tree is independent of the distance to the site under selection. Only the timescale changes from the value predicted by Takahata's allelic genealogy at the site under selection, converging with increasing recombination to the timescale of the neutral coalescent. However, if nucleotide sequences are simulated over a recombining region containing a site under balancing selection, a phylogenetic tree constructed while ignoring such recombination is strongly affected. This is true even for small rates of recombination. Published studies of multiallelic balancing selection, *i.e.*, the major histocompatibility complex (MHC) of vertebrates, gametophytic and sporophytic self-incompatibility of plants, and incompatibility of fungi, all observe allelic genealogies with unexpected shapes. We conclude that small absolute levels of recombination are compatible with these observed distortions of the shape of the allelic genealogy, suggesting a possible cause of these observations. Furthermore, we illustrate that the variance in the coalescent with recombination process makes it difficult to locate sites under selection and to estimate the selection coefficient from levels of variability.

LOCI under multiallelic balancing selection are the most polymorphic genes known in eukaryotes. These systems include gametophytic (Emerson 1939) and sporophytic (Kusaba *et al.* 1997) self-incompatibility systems in plants, incompatibility systems in fungi (May and Matzke 1995), and some of the MHC genes in vertebrates (Anderson *et al.* 1986; Hughes and Nei 1988). In each system a large number of alleles (20–150) are maintained at intermediate frequencies and nucleotide sequence variation among alleles often exceeds 30%.

The MHC data in particular have stimulated the analysis of models that are consistent with these striking levels of polymorphism. Overdominant selection with (close to) equal selection coefficient is sufficient (and appears necessary) to explain the data (Takahata and Nei 1990). With incompatibility systems, the polymorphism can be explained by the inherent selection (Vekemans and Slatkin 1994; Schierup *et al.* 1998).

Population genetics theory has successfully explained some aspects of these polymorphisms. Nevertheless, one important aspect of the pattern of polymorphism in these systems is not yet well understood. This is the shape of the phylogenetic tree of the alleles. Takahata (1990) showed for symmetrical overdominance that the allelic genealogy (*i.e.*, the phylogeny of functionally different alleles) can be approximated well by a Moran process with a constant number of allelic classes with equal death rates. Such a Moran process satisfies the assumptions of the neutral coalescent (Kingman 1982) with time scaled appropriately through a scaling factor f_S. This implies that the phylogenetic tree of alleles has the same expected shape as the neutral coalescent, differing only in the timescale. Extension to gametophytic self-incompatibility has shown that f_S is very large (>1000) for realistic population sizes and mutation rates to new specificities (Vekemans and Slatkin 1994). Uyenoyama (1997) characterized the shape of the phylogenetic tree through four ratios calculated from the branch lengths of the trees and scaled to have approximate means of one under the neutral coalescent with no recombination. She found by simulation that the values of these ratios for allelic genealogies of gametophytic self-incompatibility systems are (almost) independent of the overall sequence variability (*i.e.*, the mutation rate). Allelic genealogies in sporophytic self-incompatibility (Schierup *et al.* 1998) and fungal incompatibility (May *et al.* 1999) are also expected to have a shape close to the neutral coalescent, when measured through these ratios.

However, when these ratios are applied to real sequence data of functionally different alleles, they show significant deviations from coalescent expectations (Uyenoyama 1997; May *et al.* 1999; Richman and Kohn 1999; Table 1; for definition of ratios, see statistics). The main deviation is that the terminal branches are much longer than expected ($R_{SD} > 1$). This pattern of

Corresponding author: Mikkel H. Schierup, Bioinformatics Research Center (BiRC), Department of Ecology and Genetics, University of Aarhus, Ny Munkegade, Bldg. 540, 8000 Aarhus C., Denmark.
E-mail: mikkel.schierup@biology.au.dk

TABLE 1

Analysis of population data sets of self-recognition systems

System	Gene	Species	No. alleles	Pairwise divergence	R_{PT}	R_{ST}	R_{SD}	R_{BD}	PIST test of recombination[c]	R^2 test of recombination[d]	Reference
MHC[a]	*DRB1*	*Homo sapiens*	11	0.08	0.69	2.15	3.46	0.91	$P < 0.001$	NS	Bergstrom *et al.* (1998)
MHC[a]	*RT1.Ba*	*Rattus fuscipes greyii*	36	0.06	0.42	2.62	9.41	0.58	$P < 0.001$	$P < 0.05$	Seddon and Baverstock (1999)
Fungal SI[b]	*beta* 1	*Coprinus cinereus*	6	0.48	1.06	2.26	5.15	0.11	$P < 0.05$	NS	May *et al.* (1999)
Gametophytic SI[a]	S-RNAse	*Lycium andersonii*	22	0.70	0.54	2.40	7.20	0.32	NS	NS	Richman and Kohn (1999)
Gametophytic SI[b]	S-RNAse	*Physalis crassifolia*	28	0.54	0.63	1.98	3.00	0.99	NS	$P < 0.01$	Richman *et al.* (1996)
Gametophytic SI[b]	S-RNAse	*Solanum carolinense*	13	0.69	0.68	2.32	5.47	0.64	NS	$P < 0.05$	Richman *et al.* (1995)
Gametophytic SI[a]	S-RNAse	*Petunia inflata*	14	0.56	0.67	2.21	4.94	0.27	NS	NS	Wang *et al.* (2001)
Sporophytic SI[c]	*ARSK*	*Arabidopsis lyrata*	11	0.37	0.60	2.68	7.59	0.14	$P < 0.05$	NS	Schierup *et al.* (2001)
Sporophytic SI[b]	*SLG*	*Brassica oleracea*	23	0.14	0.45	2.74	5.15	0.58	$P < 0.01$	$P < 0.001$	Kusaba *et al.* (1997)
Sporophytic SI[b]	*SLG*	*Brassica campestris*	19	0.13	0.46	3.02	8.87	0.42	$P < 0.01$	$P < 0.001$	Kusaba *et al.* (1997)

[a] Sequences were downloaded from GenBank and aligned with ClustalX (Thompson *et al.* 1997). Trees were reconstructed using DNAdist with F81 model and Kitsch (Felsenstein 1995), and the four statistics were calculated from the branch lengths. Other reconstruction methods gave very similar results.

[b] Values of the four statistics were taken from the literature.

[c] This test is the informative sites test of Worobey (2001). The test was performed using the PIST 1.0 software, following closely the recommendations of Worobey (2001), including using PAUP* (Swofford 2000). NS, nonsignificant.

[d] This test followed Awadalla *et al.* (1999) closely, except that only sites with >30% frequency were included. Significance was assessed by 1000 permutations of the variable sites. NS, nonsignificant.

deviation is remarkably consistent over the four different kinds of systems, even though these are based on completely different molecular mechanisms. Two hypotheses have been put forward to explain this observation. Uyenoyama (1997) suggested that the enforced heterozygosity in gametophytic self-incompatibility (SI) leads to accumulation of recessive deleterious variants through sheltering. The probability of invasion and the retention time of a new specificity would then decrease over time because it would be selected against when forming heterozygotes with the specificity it arose from. Richman and Kohn (1999) suggested, on the basis of a statistical analysis of phylogenetic trees of alleles, that divergent alleles are preferentially maintained in gametophytic SI. However, these two hypotheses have not yet been quantitatively investigated theoretically. Either of them is not likely to play an equally strong role in each of the four distinct systems. For example, homozygotes can be formed in the MHC and sporophytic SI but not in gametophytic SI, and Uyenoyama's hypothesis quantitatively depends on the frequency of homozygotes.

Takahata's (1990) allelic genealogy is an infinite alleles model that treats alleles as entities that cannot be broken up by recombination. It is thus an important assumption for application of this theory to sequence data that recombination does not occur. However, intragenic recombination/gene conversion has been reported within genes of the MHC (Bergstrom *et al.* 1998), gametophytic self-incompatibility (Wang *et al.* 2001), sporophytic self-incompatibility (Awadalla and Charlesworth 1999; Schierup *et al.* 2001), and fungal incompatibility (May and Matzke 1995). In light of this it is important to investigate how violation of the no-recombination assumption affects genealogical inferences.

Here, the expected effects of recombination on the shape of the genealogy are investigated. We simulate a simple model of multiallelic selection with recombination using an extension of Hudson's (1983) algorithm for the coalescent with recombination. The main assumption is that variation in a single nonrecombining spot on the sequence is subject to selection. This spot could either be a single nucleotide site or a collection of adjacent nucleotides forming a specificity-determining region.

First, we investigate the genealogy as a function of the recombination distance from the spot under selection. At the spot under selection, we expect a neutral-shaped genealogy of alleles with an extended timescale, which depends on the number of allelic classes and the mutation rate to new specificities (Takahata 1990). Sufficiently far from the spot under selection we expect that genealogical trees of sequences are determined by the neutral coalescent. Our first question is how, as we move away from the spot under selection, the allelic genealogy transforms into the neutral coalescent. Does the shape

of the genealogical tree remain unaffected? How far, measured in recombination distance, is the effect of selection measurable? The last question follows HUDSON and KAPLAN (1988) and TAKAHATA and SATTA (1998).

Second, we quantify the shape of the "average" genealogical tree of a sample of whole sequences subject to recombination. With recombination a single genealogical tree does not normally describe the sequence variation since different parts of the sequence have different histories. Previous investigations of allelic genealogies are, however, based on phylogenetic trees, and it is therefore of interest to investigate the expected shape of a phylogenetic tree of sequences even when recombination occurs. Biases introduced by ignoring recombination can then be quantified. To investigate this question we simulate samples of nucleotide sequences assuming a given amount of recombination and a specified substitution model for the linked neutral nucleotides. We describe how much recombination is needed before the shape of the phylogenetic tree is distorted, compare these results with the published studies, and conclude that relatively small amounts of recombination are compatible with the deviations from the expected shape of genealogical trees observed in the data sets.

MODEL

The model is an extension of HUDSON's (1983) coalescent with recombination, here allowing for a simple form of symmetrical balancing selection. It is reminiscent of the process formulated by GRIFFITHS and MARJORAM (1996) except that mutation to a different specificity can happen in a single position of the sequence only. For simplicity, we define the site of selection to be at the left endpoint of the sequence (Figure 1).

There are n sequences sampled and the diploid population size is N. The continuous time approximation scales time in $2N$ generations. Recombination can happen with the same probability over the sequence determined by the overall recombination parameter $\rho = 4Nr$, which is the number of recombination events in a sequence in $4N$ generations, with r thus being the probability of a recombination event in a single sequence in a single generation.

To model strong balancing selection we assume that M distinct allelic classes are kept in equal frequencies in the population. An allelic class is also termed a specificity as in studies of self-incompatibility or the MHC. The turnover process of specificities follows TAKAHATA (1990), which describes it as a symmetric Moran process viewed back in time with an allelic turnover rate, Q. In other words, Q is the rate at which specificity lineages bifurcate in the population. Q depends on the mutation rate to new specificities and the selection coefficient (see TAKAHATA 1990). At a turnover event, each allelic class is equally likely to be lost and a new allelic class is then created to keep the number of specificities constant.

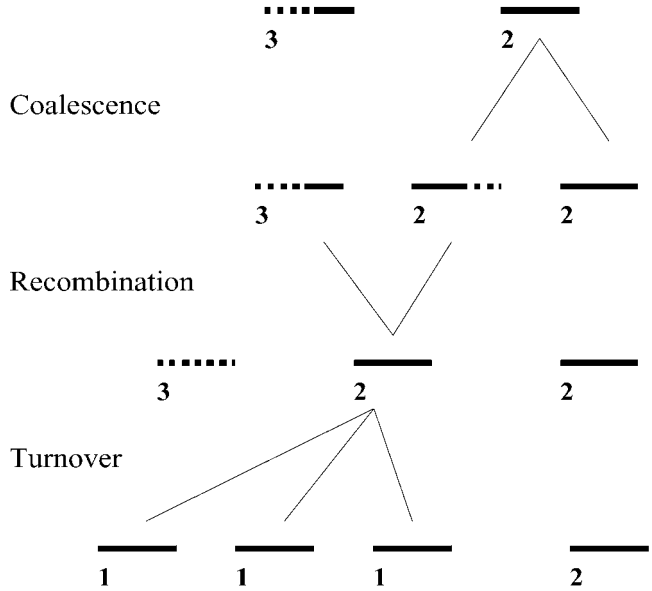

FIGURE 1.—The coalescent process with recombination and balancing selection (see text). Ancestral material, solid line; nonancestral material, dotted line. Bottom line shows a sample of four genes associated at their left end points with two different specificities (types), three copies of specificity 1, and one copy of specificity 2. The first event (counting from the bottom) is an allelic turnover event from type 1 to type 2, where type 1 changes to type 2. This leads to instant coalescence of all genes with type 1 and assignment of type 2 to the resultant gene. A new type (in this case type 3) is invented to keep the number of specificities constant. Initially this type does not carry any ancestral material. The second event is a recombination that splits a gene in two. The left ancestor keeps the same type (in this case type 2), whereas the right ancestor is assigned a random type among the other types present (here type 3). Type 3 now carries ancestral material and "trapped material" (see text). The third event is coalescence of two genes of the same type (in this case type 2). The left part of the sequence has thus found a most recent common ancestor. At least one further allelic turnover event and a subsequent coalescence event are necessary before the right part of the sequence finds a common ancestor.

Each of the n sampled sequences is associated at its left endpoint with one of the allelic classes. A given point at a given sequence can change its associated specificity if either the specificity is changed by an allelic turnover event or if recombination occurs between the selected site and the focal point. Two sequences can coalesce only when they have the same specificity (Figure 1, top).

Assume that there are M specificities in the population and that we sample n sequences of different specificities $n \leq M$. This corresponds to the situation where an investigator sequences only one copy of each specificity, but not all existing specificities are necessarily sampled. The coalescent process with recombination and selection can then be approximated by three independent exponentially distributed waiting times, namely

coalescence, recombination, and allelic turnover (see Figure 1). A sample of sequences is followed backward in time until all parts of each sampled sequence (the ancestral material) have found a most recent common ancestor.

Coalescence: The intensity of coalescence is given by $C_t = M\sum_{i=1}^{M}(n_i(n_i - 1)/2)$, where n_i is the number of copies of specificity i, since coalescences can only happen within a given specificity that each has an effective size of $2N/M$. If a coalescent event happens, an allelic class i is chosen with probability proportional to $n_i(n_i - 1)/2$, two random sequences from class i are merged into one ancestral sequence with the same specificity i, and n_i is decreased by 1. Note that initially, since we sample at most one sequence from each specificity, $n_i = 1$ for all specificities sampled and $C_0 = 0$. Thus, coalescence can only happen once recombination has shifted ancestral material to other specificities, making $n_i > 1$ for at least one i.

Recombination: The intensity of recombination R_t at a given point in time is determined by the amount of ancestral material to the sample plus any material "trapped" by blocks of ancestral material (Wiuf and Hein 1997). The amount of ancestral material is the total part of the sampled sequences that have not yet found a most recent common ancestor. The reason why nonancestral "trapped material" has to be counted in the intensity of recombination is that a recombination event there would distribute ancestral material onto two sequences rather than one, thus affecting the coalescence process. At time zero, $R_0 = n\rho/2$, *i.e.*, the number of sequences times their lengths. If recombination happens, the recombination point is picked uniformly over this length of sequence. A recombination event breaks up the sequence in a left and a right segment (Figure 1). The left segment retains its allelic class. The right segment is assigned an allelic class among the other existing classes. In the case of self-incompatibility, this class is chosen randomly among the $M - 1$ allelic classes distinct from the class of the recombining sequence. For overdominance with selection coefficient s, the present class is chosen with relative weight $1 - s$, corresponding to selection against homozygotes of strength s.

Allelic turnover: The intensity of allelic turnover is determined by Q, which is independent of the time t by definition. If an allelic turnover event happens, an allelic class, say i, is chosen randomly with equal probability among the M allelic classes. The n_i sequences from this class are then made to coalesce instantly and the resultant sequence has its specificity changed to one of the other $M - 1$ allelic classes at random (Figure 1). Viewed forward in time this corresponds to a new specificity arising by mutation followed by its (almost) immediate increase in frequency due to the strong selection favoring rare specificities. Then a new allelic class is introduced to keep the number constant. Initially, this new allelic class does not carry ancestral material, but recombination events can transfer ancestral material to the class (see Figure 1). Note that if this happens, the material between the point of selection and the left border of ancestral material is also added to the trapped material part of the recombination intensity because a recombination event here would change the specificity associated with the ancestral material and thus the coalescent history. If the allelic turnover rate is small, the coalescent process of the left endpoint of the sequence (the point under selection) is dominated by the allelic turnover process alone, and, according to Takahata (1990), the expected time to the most recent common ancestor is $D = M(M - 1)(1 - 1/n)/Q$, which is likely to be much longer than the neutral value of $D = 2(1 - 1/n)$ when $Q < 1$.

Since the three events are independent and exponentially distributed, the intensity of any event to happen is exponentially distributed with parameter $C_t + R_t + Q$ and given that an event happens, the probability that it is a coalescent event, say, is $C_t/(C_t + R_t + Q)$. The process is simulated from starting conditions by determining the time of the first event by drawing a random number from an exponential distribution with mean $1/(C_t + R_t + Q)$, then determining the type of the event, and finally updating the intensities of the three events according to the rules above. The process is continued until all parts of the sequences have found a common ancestor. For the left endpoint of the sequences the time until a common ancestor is primarily determined by the allelic turnover process, whereas recombination plays an increasingly important role the farther a point is away from the left endpoint.

The process results in a set of correlated trees relating the samples along the sequence. In contrast to the neutral coalescent with recombination these trees are not taken from the same distribution since their expected branch lengths depend on the distance to the left endpoint where specificities are determined. During a single stochastic realization of the process we stored all information on topology and branch length for each of these trees. From these we (a) investigate the coalescent process as a function of distance from the point of selection and (b) simulate and subsequently analyze nucleotide sequences under this process.

STATISTICS

To characterize the shape of the phylogenetic trees we used five quantities calculated from branch lengths. These are

- S, sum of the length of terminal branches;
- T, total length of all branches;
- D, time to the most recent common ancestor;
- P, average pairwise distance between two specificities;
- B, average length of basal branches emanating from the root.

From these, four ratios,

$$R_{PT} = \frac{2Pa_n}{T}, \quad R_{ST} = \frac{Sa_n}{T}, \quad R_{SD} = \frac{S(1 - 1/n)}{D},$$

and

$$R_{BD} = \frac{B(1 - 1/n)}{Db_n},$$

can be defined, where $a_n = \Sigma_{i=1}^{n-1} (1/i)$ and $b_n = 1/n + \Sigma_{i=2}^{n-1} (1/i^2)$ (Uyenoyama 1997).

Viewed as ratios of means, all four ratios are scaled to have an expected mean of one under the neutral coalescent, and simulations have shown that their means as ratios are also close to one (Uyenoyama 1997). The ratios have the advantage when applied to data that they are (almost) independent of the mutation rate and in many cases they have power to reject the hypothesis of a neutral coalescent process (Uyenoyama 1997; Schierup and Hein 2000). The most powerful statistic has generally been found to be R_{SD}, which measures the ratio of the length of the terminal branches to the height of the tree.

We also calculated the time between subsequent coalescence events. Under the neutral coalescent with i sequences, the mean waiting times F_i to the next coalescence are independent and exponentially distributed with mean $2/(i(i - 1))$. Thus, $G_i = F_i i(i - 1)/2$ are exponentially distributed with mean 1, and plotting G_i as a function of i can visualize systematic deviations from neutral expectations.

These measures can all be calculated from the branch length of the true trees over the sampled sequences in a single realization of the coalescent with recombination and balancing selection process. They can also be calculated from phylogenetic trees reconstructed from nucleotide sequences simulated under the model (see below).

Genealogical structure over a gene: The model was used to simulate genealogical histories. Each of n sampled genes was assigned a unique specificity among the M possibilities. Models were simulated and analyzed where either all specificities were sampled ($n = M$) or just a subset was sampled ($n < M$). One run of the program generates a set of trees with branch lengths over the set of genes. Such a set is a single outcome of the stochastic process and is termed a "history." We sampled a given history at 1000 points spaced as a logarithmic function of ρ and calculated the various statistics at each point. Mean and standard deviations for a given set of parameters were then found over many (>15,000) recorded histories. The statistics were then plotted as a function of the recombination distance from the site under selection. A total length of $\rho = 100$ was investigated, which means that 100 recombinations are expected between the endpoints of a gene in $4N$ generations.

Nucleotide sequences: Simulation of nucleotide sequences followed Schierup and Hein (2000). Neutral mutations can be added after the genealogies have been constructed under the coalescent model because the coalescent process and the neutral mutation process are independent. Mutations were added at rate m to the simulated genealogy by dividing the sequence length into L equally sized fragments corresponding to nucleotides. We used the simple Jukes-Cantor substitution model (Jukes and Cantor 1969) and assumed that nucleotides mutate independently. For a given position in the sequence, first a nucleotide is assigned to the most recent common ancestor (MRCA) with probabilities according to the equilibrium frequencies of nucleotides, which for the Jukes-Cantor model is 25% of each. The evolution of the nucleotide is then followed over the specific genealogical tree at this position. For a given branch of length l, the number of mutations is Poisson distributed with mean ml. Repeating this process for each nucleotide results in n aligned sequences of length L. We restricted analysis to the Jukes-Cantor model because we found previously that more complex substitution models have little effect on the expected values of the above quantities (Schierup and Hein 2000).

Sequences were simulated with a single allelic turnover rate at the site under selection but with different levels of recombination over the sequence. Again, sequences were initially assigned unique specificities, equivalently to sampling sequences with different specificities only, as is done in published studies of these systems. Each set of sequences was subsequently run though DNAdist and Kitsch programs of PHYLIP (Felsenstein 1995), which results in an inferred phylogenetic tree reconstructed on the basis of a distance matrix and restricted by a molecular clock. This is clearly not an appropriate method when recombination occurs, but the purpose here is to investigate the bias created in doing so. From the branch length of the reconstructed trees the various statistics were recorded. Several combinations of parameters n, M, s, and Q were investigated. The program for simulations was written in C and can be accessed through http://www.birc.dk/~mheide.

RESULTS

Genealogical structure over gene: Figure 2 shows results for four of the basic quantities for two different allelic turnover rates to new specificities [$Q = 0.01$ (solid line) and $Q = 0.1$ (dotted-dashed line)] for the sample size $n = 30$ genes and $M = 30$ specificities. The values of each quantity for $\rho = 0$ are as expected from Takahata's (1990) theory (marked on y-axis), which predicts coalescence times proportional to the square number of specificities and to $1/Q$ (*e.g.*, $D = M(M - 1)(1 - 1/n)/Q$). Selection can be seen to greatly increase expected coalescence times close to the site under selection, but as ρ increases, each quantity approaches the value expected under Kingman's (1982) coalescent

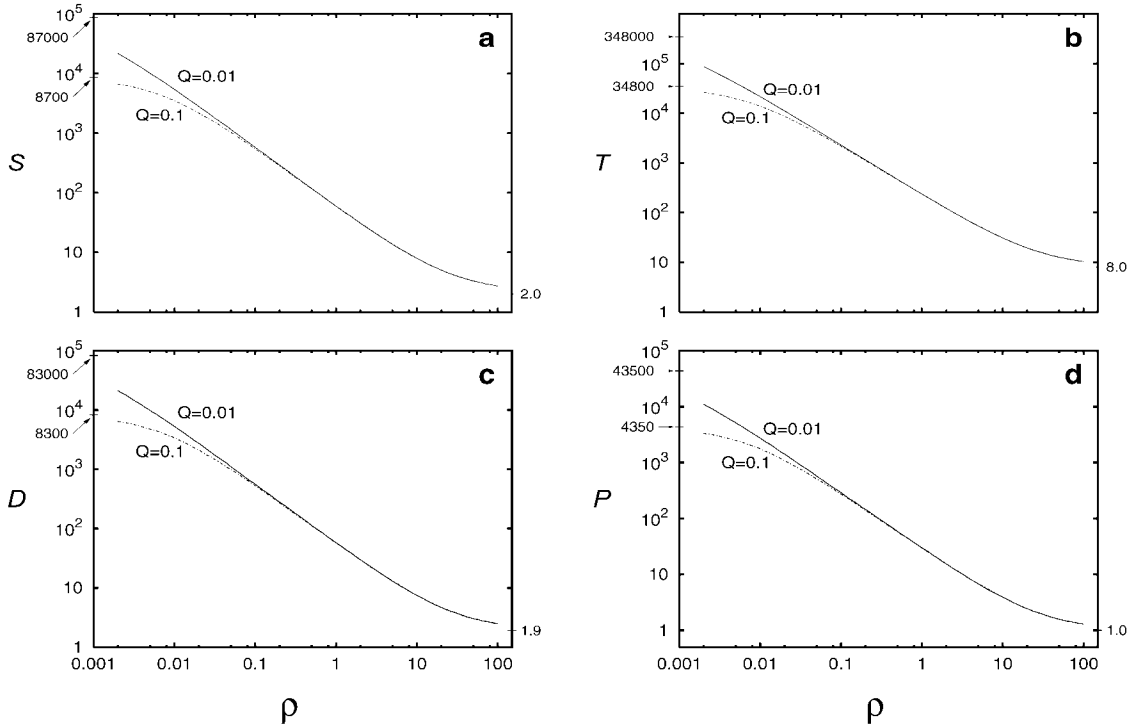

FIGURE 2.—Four tree statistics as a function of recombination rate from the site under balancing selection. Solid lines, results for $Q = 0.01$; dotted-dashed lines, results for $Q = 0.1$. $M = n = 30$, self-incompatibility model ($s = 1$). Predictions from Takahata's allelic genealogy are marked on the y-axis and those from Kingman's coalescent are marked on the right vertical axis. Results plotted are means of 15,000 histories. (a) The length of terminal branches, S; (b) the total length of the tree, T; (c) the height of the tree, D; and (d) the average pairwise distance, P.

(marked on right vertical axis). Each of the four quantities shows, as expected, a monotonic decrease with distance from the point of selection. We stress two observations. First, even though the two values of Q correspond to a 10-fold difference in f_S, the graphs become (almost) indistinguishable when $\rho > 0.02$. Thus, differences in selection intensities only have an effect extremely close to the point of selection (corresponding perhaps only to a couple of nucleotides). Second, the recombination distance needed to approach the neutral coalescent depends approximately linearly on the number of allelic classes as shown previously for the pairwise divergence times (TAKAHATA and SATTA 1998). This means that, whereas for $\rho = 1$ the effect of selection on levels of diversity is negligible for $M = 2$ (shown by HUDSON and KAPLAN 1988), the effect of selection is still appreciable for $\rho = 10$ when $M = 30$.

Figure 3 shows the four ratios for the same runs as in Figure 2. For $\rho = 100$, ratios from all models are expected to converge to a value very close to one as expected from the neutral coalescent. Similarly, the ratios are expected to converge to one for $\rho = 0$ (TAKAHATA 1990). Figure 3 shows that the ratios are to a good approximation constant for any value of the recombination distance from the site under selection, meaning that only the timescale changes while the phylogenetic tree retains the same shape.

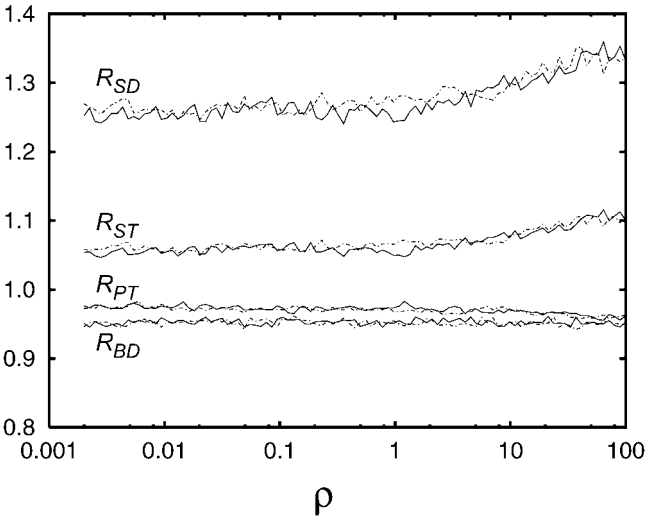

FIGURE 3.—The four ratios calculated from the same runs as described in Figure 2. The four ratios as a function of distance from the selected site (15,000 histories) are shown. R_{ST} and R_{SD} measure the length of the terminal branches relative to the height and length of the tree, respectively. R_{PT} measures the average pairwise difference relative to the total length of the tree, and R_{BD}, measures the length of the basal branches relative to the height of the tree.

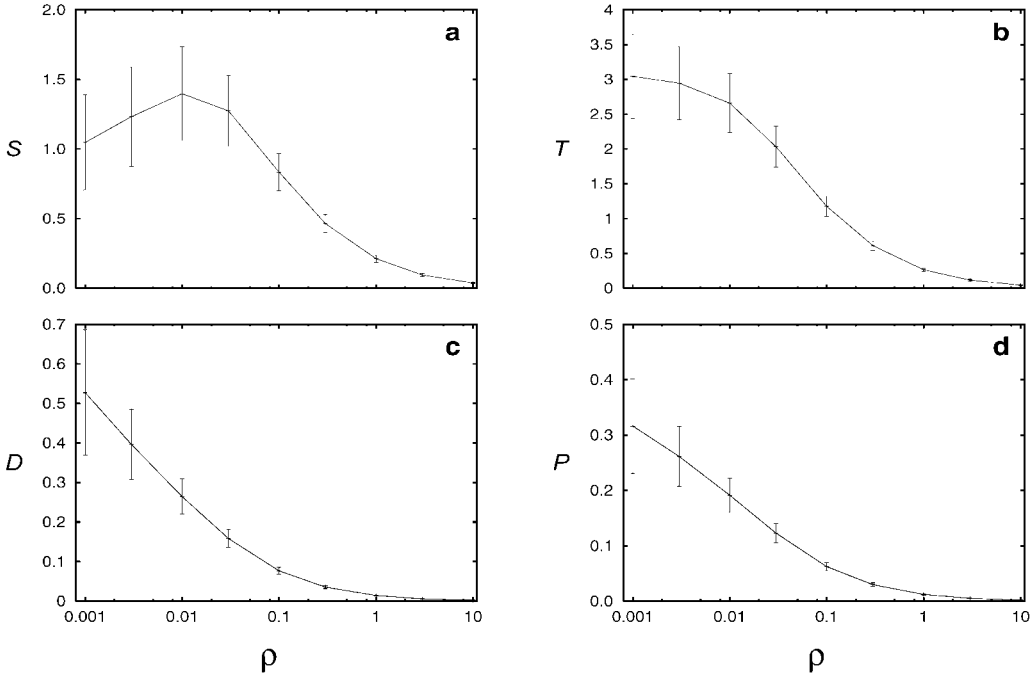

FIGURE 4.—The effect on tree statistics of ignoring recombination as a function of ρ, for a sample of 30 sequences from 30 different allelic classes. Selection is at one end of the sequence with turnover rate for specificities $Q = 0.1$. Self-incompatibility model ($s = 1$) is shown. Sequences are 10,000 bp long, $m = 0.0001$. Results (±standard deviation) are based on 15,000 replicates. (a) The length of terminal branches, *S*; (b) the total length of the tree, *T*; (c) the height of the tree, *D*; and (d) the average pairwise distance, *P*.

Several different combinations of parameters *M*, *n*, *Q* (including cases where $n < M$, *i.e.*, not all specificities sampled), and selection coefficients against homozygotes were investigated and found to yield the same conclusions (results not shown).

Expected phylogenetic tree for sequence sets with selection and recombination: Again, we present results for 30 sequences sampled, each from a distinct specificity ($M = 30$), and the turnover rate to a new specificity was $Q = 0.1$. The recombination rate here is the value of ρ used when simulating the sequences and not the recombination distance from the site under selection as in the previous section. The idea here is to show how various amounts of recombination will affect inferences that are based on sequences from different specificities.

Figure 4 shows *S* (the length of terminal branches), *T* (the length of the tree), *D* (the height of the tree), and *P* (the average pairwise difference) as a function of ρ. All statistics except *S* show a monotonic decrease as a function of ρ. This is expected when *Q* is constant, because parts of the sequence are less influenced by selection when ρ increases, lowering the average (compare Figure 1). The statistic *S*, the length of the terminal branches, increases slightly for small recombination rates, indicating that the shape of the genealogical tree is changed when recombination occurs. Figure 5 shows this to be true. Even very small rates of recombination ($\rho \approx 0.1$) have a large effect on the four ratios. The terminal branches are relatively longer than expected (when compared with the height and total length of the tree, *i.e.*, both R_{SD} and $R_{ST} > 1$), and the average pairwise distance is smaller than expected from the total branch length ($R_{PT} < 1$). This pattern is similar to that previously reported for neutrally evolving sequences (SCHIERUP and HEIN 2000) except that here the effect is observed for much smaller recombination rates. For neutrally evolving sequences the cause of the pattern is that recombination makes sequences more equidistant, and attempting to reconstruct a phylogenetic tree will make it appear star-like (SCHIERUP and HEIN 2000). The same phenomenon occurs here but at smaller recombination rates, because selection increases the overall timescale of the genealogical process, which amounts to an increase in the effective recombination rate close to the site under selection.

Again, the inferences drawn from these results were found to be very robust to changing values of the different parameters, including modeling different strengths of balancing selection (results not shown).

The scaled internode distances, G_i, reveal a more detailed picture of the shape of the genealogy. Figure 6 shows these as functions of the coalescence events for four different recombination rates. For $\rho = 0$, the line is horizontal as expected from theory (TAKAHATA 1990). When ρ increases, the recent coalescence times are too long relative to the coalescence times close to the root. This again reflects the long terminal branches (Figure 5).

Variation over a single set of sequences: Figure 2d shows the average expected diversity *P* as a function of the recombination distance from the spot under selection. However, the variation around these means is enormous, mainly because of the inherent variation in the coalescent with recombination process. To visualize this variance we chose three random data sets for each of three different amounts of recombination. Sequence diversity was then calculated in a sliding window (Figure

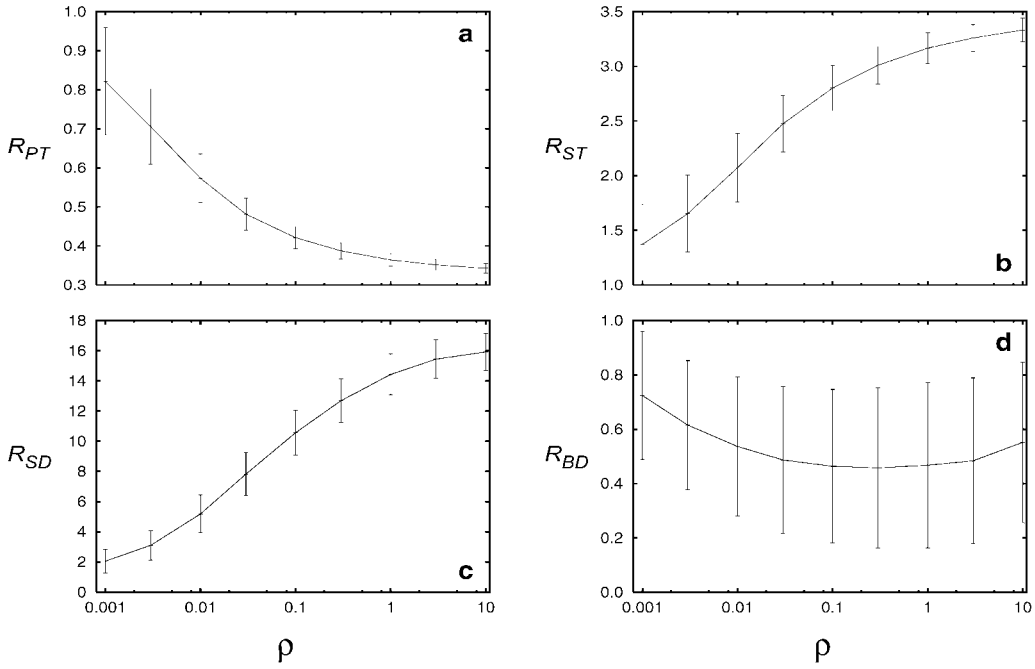

FIGURE 5.—The four ratios calculated from the same runs as described in Figure 4. Ratios were calculated for each history and then averaged over the 15,000 histories. (a) R_{PT}; (b) R_{ST}; (c) R_{SD}; and (d) R_{BD}.

7). Variation between runs is very large as can be seen by comparing the three replicates for a given value of ρ (Figure 7). Increasing the rate of recombination makes it easier to see where selection is acting. When $\rho = 0.01$ it appears virtually impossible to pinpoint the spot under selection (which by definition is at the left endpoint) from sequence diversity pattern and even for $\rho = 0.1$ there are spurious peaks of diversity separated from the site of selection by low diversity regions. This reflects that the coalescent with recombination process determines the history of blocks of nucleotides and when $\rho < 0.1$ the size of such blocks is large. Therefore, detection of selection through regions of overall hypervariability of both synonymous and nonsynonymous substitutions is likely to be successful only if $\rho > 1$ for the sequence under study.

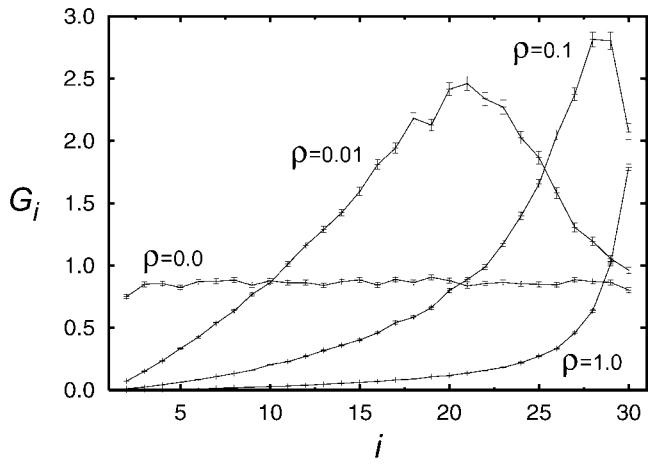

FIGURE 6.—Internode distances G_i, $i = 2, \ldots, 30$ for sets of 30 sampled sequences with different expected recombination rates ρ. Calculated on the same runs as in Figure 4.

DISCUSSION

We have investigated a very simple model of multiallelic balancing selection with recombination. The allelic genealogy and the neutral coalescent have the same genealogical structure, differing only in timescale (TAKAHATA 1990). We have shown to a close approximation that the same genealogical shape is found at any recombination distance from the spot under selection. Intuitively, this makes good sense: For a point in a sequence some distance from the spot of selection, say $\rho = 1$, recombination shifts a given nucleotide among the different specificities in much the same way as a change in specificity caused by an allelic turnover event but at a higher rate. The process is therefore similar to an allelic genealogy with increasing allelic turnover rate as one moves farther away from the site under selection. To be able to compare with experimental data of these systems, we simulated sequences under the same model and reconstructed phylogenies. Results show that very little recombination significantly changes the shape of the inferred genealogy.

Limitations of the model: Some of the simplifications in the model need consideration.

a. The approximation of a fixed number of allelic classes to simulate balancing selection under mutation-selection-drift balance ignores random fluctuations in the number of specificities. Previous investigations have found this approximation to be accurate

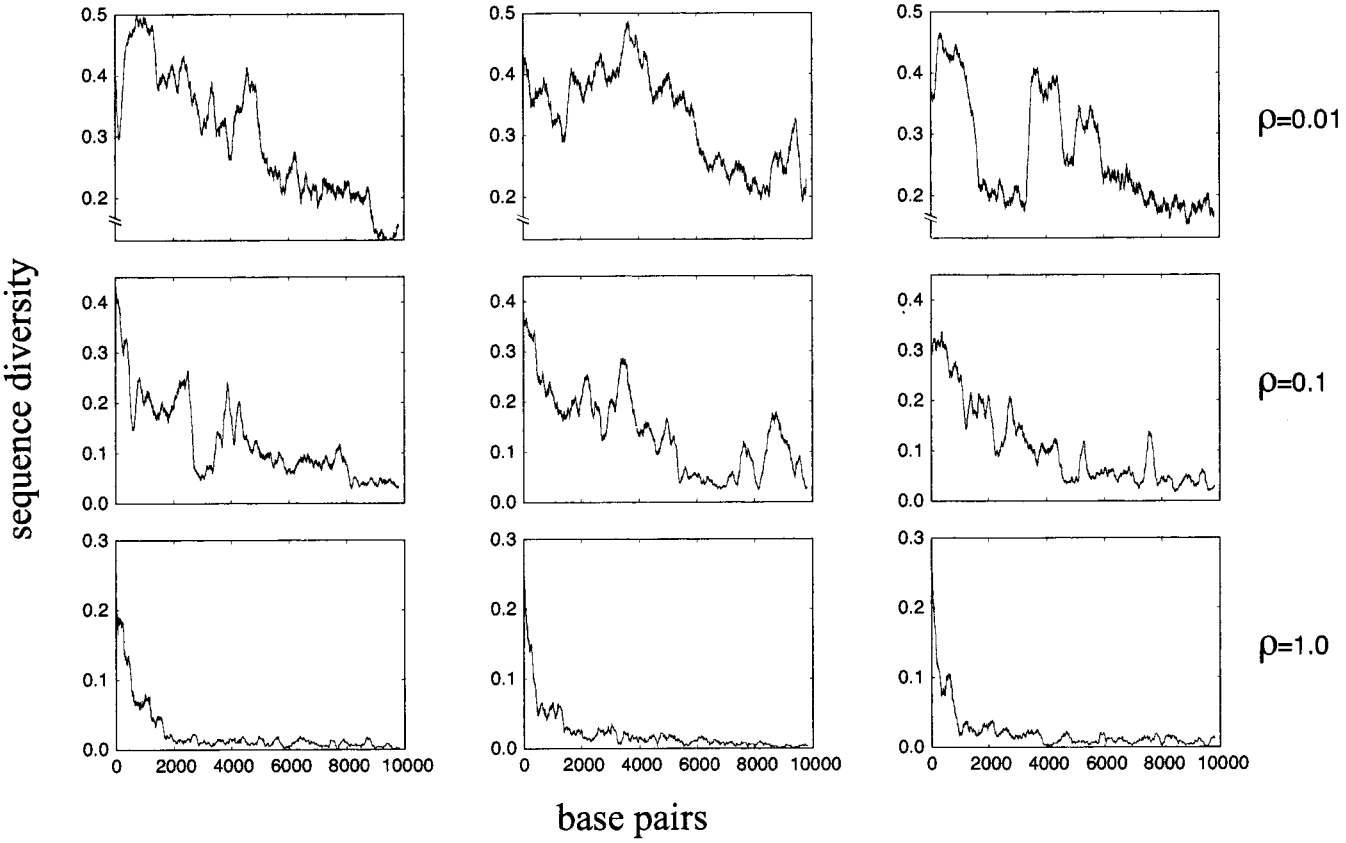

FIGURE 7.—Sliding window diversity plots for three randomly chosen simulation runs of 30 sequences for each of $\rho = 0.01$, $\rho = 0.1$, and $\rho = 1$. Sequences were 10,000 bp long, $m = 0.0001$. Window size is 200 bp.

when selection is strong (TAKAHATA 1990; SCHIERUP *et al.* 1997, 1998).

b. Instant coalescence of all members of an allelic class at an allelic turnover event demands that the time for invasion of a new allele is negligible on the timescale of the allelic turnover process. Since invasion of a new successful specificity is expected within a few generations and the timescale of allelic turnover process is $2N/Q$ generations, this approximation is accurate if $N > 1000$ (assuming $Q < 1$).

c. Equal probability of each allelic type at sampling assumes that allelic classes in the total population are at their deterministic frequency, *i.e.*, again that N is large.

d. A single point at the left border of the sequence determines specificity: This is likely the most restrictive assumption if results are compared to many of the data sets of Table 1. It most closely resembles the situation of a specificity-determining domain, like a peptide-binding region as in the MHC. It may also be a good approximation to diversity in adjacent introns and 5′ and 3′ noncoding regions of these systems.

In cases where specificity is determined by nucleotides dispersed in the sequence of interest, it is presently unclear how good an approximation our model is. We have not investigated models of this situation mainly because recombination between such nucleotides would then contribute to the creation of new specificities and a very complex mutation process would have to be modeled. The current understanding of the determination of specificities does not allow us to choose among the many possible models of such interaction of sites. However, we believe that the qualitative effects we observe would also be preserved under many realistic models where several interspersed positions determine specificity.

Comparison with experimental data sets: With these limitations in mind, we compare our results with observed values of the four ratios in different incompatibility systems (Table 1). Clearly, the observed pattern is very close to the results of simulations when sequences are allowed to recombine at a very low absolute rate. In fact, values of Table 1 correspond to recombination rates in the range of $\rho = 0.001$–0.1 in Figure 4.

Indirect evidence that gene exchange occurs between alleles has been reported in each of the four types of self-recognition analyzed in Table 1. In sporophytic SI, AWADALLA and CHARLESWORTH (1999) found a decay in linkage disequilibrium with distance in Brassica SLG

alleles (which show similar diversity as SRK alleles; see Kusaba *et al.* 1997; Nishio and Kusaba 2000), and signs of recombination were also found in the *Arabidopsis lyrata* SRK orthologue (Schierup *et al.* 2001; P. Awadalla, M. H. Schierup, B. K. Mable and D. Charlesworth, unpublished results). In gametophytic SI, Wang *et al.* (2001) recently reported evidence for recombination in *Petunia inflata*. In gametophytic SI in general, the diversity is so great that it is difficult to determine whether recombination has happened because multiple mutations have happened in most variable sites. In fungal incompatibility of *Coprinus cinereus*, May and Matzke (1995) report evidence for recombination in the gene region. Finally, gene conversion has been reported to occur in the MHC exons (Bergstrom *et al.* 1998; Takahata and Satta 1998).

To further investigate whether recombination has happened in the data sets of Table 1 we also applied two recent tests of recombination. The informative sites test (Worobey 2001) is designed for sequences with high levels of variation. The R^2 test of recombination tests whether there is a significant correlation between the R^2 measure of linkage disequilibrium and the distance between base pairs (Awadalla *et al.* 1999). Results of both tests should be treated cautiously because it is not yet clear how the presence of selection may bias results. Yet, they are probably the best available tests and there are no *a priori* reasons for false-positive results. However, it is likely that none of the tests are particularly powerful. When applied to the data sets of Table 1, the two tests yield some evidence for recombination in each type of system. Evidence for recombination is weakest in gametophytic SI. Whether this is because recombination is absent (or very rare) in gametophytic SI or whether the diversity of these systems is too high for the tests to be powerful is presently unclear.

We hypothesize that if recombination occurs at sufficiently high rates then it is a possible explanation for the "long terminal branches" (Uyenoyama 1997) observed in genealogies of different SI systems (Table 1), and recombination should be kept in mind as an alternative to the mechanisms' "sheltering of deleterious alleles" (Uyenoyama 1997) or "preferential retention of divergent lineages" (Richman and Kohn 1999) previously discussed as possible explanations for the long terminal branches in the different systems.

Recombination rates in the range $\rho = 0.001$–0.1 are normally too small to investigate by direct methods. Assuming 20 sequences are sampled, $\rho = 0.1$ corresponds to 0.3 recombination events in the whole sequence set during $2N$ generations. As an example, in *Drosophila melanogaster*, $\rho = 0.1$ corresponds to just 1.2 bp of a gene in a region of normal recombination if we assume the effective population size is $N = 10^6$ and the per generation recombination rate is 2×10^{-8} between adjacent nucleotides. Thus it is possible that recombination can be severely reduced in self-incompatibility genes (Casselman *et al.* 2000) but still have a large effect on the inferred genealogical tree. Indeed, even $\rho = 0.1$ is unlikely to be detected through direct observation of segregation. The reason for the large effect of such low rates of recombination on the structure of the allelic genealogy is that balancing selection slows the coalescent process of the alleles (Figures 2 and 3) and thereby extends the time interval where recombination may have an effect. In this sense, the closer to the selected site, the higher the effective recombination rate, which is expected to be f_S times the neutral expectation very close to the selected site. Thus, recombination in the ancestral material to the sampled genes is expected to happen with an inflated frequency close to the site under selection.

It remains to be determined whether recombination rates in the different self-recognition systems are on the order of $\rho = 0.001$–0.1 that one would expect if recombination alone should explain the long terminal branches in these systems. Because of the strong selection, $\rho = 0.001$ and 0.01 corresponds to $\sim$40 and 150 recombination events, respectively, in the history of the sample (values recorded in the simulations). Such levels of recombination should in principle be detectable for neutrally evolving sequences, but, for balancing selection, the selected position acts as an apparent "recombination hot spot" as discussed above. This invalidates application of current estimation methods of recombination rates to the data sets of Table 1. Even for a neutrally evolving gene, estimation of the recombination rate is already a formidable task (see, *e.g.*, Wall 2000).

A further feature of each of the systems of Table 1 is that the alignments of alleles contain hypervariable regions. For example, in sporophytic SI of Brassicaceae three such regions have been described (Dwyer *et al.* 1991), five regions have been described in gametophytic SI of Solanaceae (Sims 1993), and these regions have been suggested to be the main targets for selection through determination of specificity. In MHC, the hypervariable region corresponds to the peptide-binding region, which is believed to be the target of selection. However, in self-incompatibility systems, the site of selection is unknown. We have found that selection is very difficult to locate through hypervariability unless recombination rates are unrealistically high ($\rho > 1$, see Figure 7). The reason is the very large variance in the time to the most recent common ancestor over a sequence in the coalescent with recombination process. Stronger evidence for selection in hypervariable regions would be that only the nonsynonymous substitution rate is elevated but this is difficult to test because the amount of synonymous substitution in most of these systems is very close to saturation.

If recombination is indeed happening at a rate $\rho >$

0.01 in some self-recognition systems, then some of the conclusions based on allelic phylogenies of self-incompatibility systems and MHC systems should be carefully reconsidered. The level of *trans*-specific evolution (TSE; *i.e.*, polymorphism shared between species) is very high in both MHC (Ayala 1995) and self-incompatibility (Richman and Kohn 1999; Nishio and Kusaba 2000). There is no doubt that some of allelic lineages diverged prior to speciation, but, since ignoring recombination leads to longer terminal branches in the inferred tree, one may greatly overestimate the number of such *trans*-specific lineages. If, *e.g.*, a molecular clock is applied, this implies that more lineages appear to coalesce in the common ancestor of the species. In MHC, sequence data from introns (Bergstrom *et al.* 1998) have shown that *trans*-specific polymorphism of DRB1 in humans and chimpanzee is significantly smaller than estimated from the exon data (Ayala 1995), in good agreement with evidence for gene conversion in the exons (Bergstrom *et al.* 1998). Estimates of TSE from self-incompatibility systems may be similarly affected. This has implications for methods of paleogenetics (Takahata *et al.* 1992), where the number of *trans*-specific lineages can be used to estimate long-term evolutionary parameters.

A final consequence of recombination is that the intensity of selection is very difficult to estimate. Figure 2 showed that stronger selection affects only a very minor part of the sequence because the rest of the sequence "escapes" the balancing selection through recombination. Thus, that selection *is* acting can be inferred from an increased level of polymorphism but the strength and location of selection cannot be determined with accuracy when recombination occurs.

M.H.S. thanks D. Charlesworth and P. Awadalla for continued discussions. X. Vekemans, F. B. Christiansen, Marcy K. Uyenoyama, and two anonymous referees made useful suggestions about the manuscript. T. Christensen is thanked for computer programming. The study was supported by grants nos. 9701412 and 1262 from the Danish Natural Sciences Research Council and by the Basic Research in Computer Science Centre of the Danish National Research Foundation.

LITERATURE CITED

Anderson, M. A., E. C. Cornish, S.-L. Mau, E. G. Williams, R. Hoggart *et al.*, 1986 Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in *Nicotiana alata*. Nature **321:** 38–44.

Awadalla, P., and D. Charlesworth, 1999 Recombination and selection at Brassica self-incompatibility loci. Genetics **152:** 413–425.

Awadalla, P., A. Eyre-Walker and J. M. Smith, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286:** 2524–2525.

Ayala, F. J., 1995 The myth of Eve: molecular biology and human origins. Science **270:** 1930–1936.

Bergstrom, T. F., A. Josefsson, H. A. Erlich and U. Gyllensten, 1998 Recent origin of HLA-DRB1 alleles and implications for human evolution. Nat. Genet. **18:** 237–242.

Casselman, A. L., J. Vrebalov, J. A. Conner, A. Singhal, J. Giovannoni *et al.*, 2000 Determining the physical limits of the Brassica S locus by recombinational analysis. Plant Cell **12:** 23–33.

Dwyer, K. G., M. A. Balent, J. B. Nasrallah and M. E. Nasrallah, 1991 DNA sequences of self-incompatibility genes from *Brassica campestris* and *B. oleracea*: polymorphism predating speciation. Plant Mol. Biol. **16:** 481–486.

Emerson, S., 1939 A preliminary survey of the *Oenothera organensis* population. Genetics **24:** 524–537.

Felsenstein, J., 1995 *PHYLIP (Phylogeny Inference Package) Version 3.572.* Distributed over the Worldwide Web, Seattle.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kusaba, M., T. Nishio, Y. Satta, K. Hinata and D. Ockendon, 1997 Striking sequence similarity in inter- and intra-specific comparisons of class I *SLG* alleles from *Brassica oleracea* and *Brassica campestris*: implications for the evolution and recognition mechanism. Proc. Natl. Acad. Sci. USA **94:** 7673–7678.

May, G., and E. Matzke, 1995 Recombination and variation at the a mating-type of Coprinus-cinereus. Mol. Biol. Evol. **12:** 794–802.

May, G., F. Shaw, H. Badrane and X. Vekemans, 1999 The signature of balancing selection: fungal mating compatibility gene evolution. Proc. Natl. Acad. Sci. USA **96:** 9172–9177.

Nishio, T., and M. Kusaba, 2000 Sequence diversity of SLG and SRK in Brassica oleracea L. Ann. Bot. **85** (Suppl. A): 141–146.

Richman, A. D., and J. R. Kohn, 1999 Self-incompatibility alleles from Physalis: implications for historical inference from balanced genetic polymorphisms. Proc. Natl. Acad. Sci. USA **96:** 168–172.

Richman, A. D., T.-H. Kao, S. W. Schaeffer and M. K. Uyenoyama, 1995 S-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). Heredity **75:** 405–415.

Richman, A. D., M. K. Uyenoyama and J. R. Kohn, 1996 S-allele diversity in a natural population of *Physalis crassifolia* (Solanaceae) (ground cherry) assessed by RT-PCR. Heredity **76:** 497–505.

Schierup, M. H., and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Schierup, M. H., X. Vekemans and F. B. Christiansen, 1997 Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. Genetics **147:** 835–846.

Schierup, M. H., X. Vekemans and F. B. Christiansen, 1998 Allelic genealogies in sporophytic self-incompatibility systems in plants. Genetics **150:** 1187–1198.

Schierup, M. H., B. K. Mable, P. Awadalla and D. Charlesworth, 2001 Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. Genetics **158:** 387–399.

Seddon, J. M., and P. R. Baverstock, 1999 Variation on islands: major histocompatibility complex (MHC) polymorphism in populations of the Australian bush rat. Mol. Ecol. **8:** 2071–2079.

Sims, T. L., 1993 Genetic regulation of self-incompatibility. Crit. Rev. Plant Sci. **12:** 129–167.

Swofford, D. L., 2000 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.* Sinauer Associates, Sunderland, MA.

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

Takahata, N., and M. Nei, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:** 967–978.

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

Takahata, N., Y. Satta and J. Klein, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130:** 925–938.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and

D. G. Higgins, 1997 The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24:** 4876–4882.

Uyenoyama, M. K., 1997 Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics **147:** 1389–1400.

Vekemans, X., and M. Slatkin, 1994 Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137:** 1157–1165.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wang, X., A. L. Hughes, T. Tsukamoto, T. Ando and T.-h. Kao, 2001 Evidence that intragenic recombination contributes to allelic diversity of the S-RNAse gene at the self-incompatibility(S) locus in *Petunia inflata*. Plant Physiol. **125:** 1012–1022.

Wiuf, C., and J. Hein, 1997 On the number of ancestors to a DNA sequence. Genetics **147:** 1459–1468.

Worobey, M., 2001 A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. **18:** 1425–1434.

Communicating editor: N. Takahata