Genet. Res., Camb. (2000), **76**, *pp.* 63–73. *With* 6 *figures.* Printed in the United Kingdom © 2000 Cambridge University Press

The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population

MIKKEL H. SCHIERUP[1]*, DEBORAH CHARLESWORTH[1] AND XAVIER VEKEMANS[2]
[1] *Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK*
[2] *Laboratoire de Génétique et d'Ecologie Végétales, Université Libre de Bruxelles, 1850 Chaussée de Wavre, B-1160 Brussels, Belgium*

(*Received 2 June 1999 and in revised form 5 November 1999*)

Summary

The effect of multi-allelic balancing selection on nucleotide diversity at linked neutral sites was investigated by simulations of subdivided populations. The motivation is to understand the behaviour of self-recognition systems such as the MHC and plant self-incompatibility. For neutral sites, two types of subdivision are present: (1) into demes (connected by migration), and (2) into classes defined by different functional alleles at the selected locus (connected by recombination). Previous theoretical studies of each type of subdivision separately have shown that each increases diversity, and decreases the relative frequencies of low-frequency variants, at neutral sites or loci. We show here that the two types of subdivision act non-additively when sampling is at the whole population level, and that subdivision produces some non-intuitive results. For instance, in highly subdivided populations, genetic diversity at neutral sites may decrease with tighter linkage to a selected locus or site. Another conclusion is that, if there is population subdivision, balancing selection leads to decreased expected F_{ST} values for neutral sites linked to the selected locus. Finally, we show that the ability to detect balancing selection by its effects on linked variation, using tests such as Tajima's D, is reduced when genes in a subdivided population are sampled from the total population, rather than within demes.

1. Introduction

Models of loci under balancing selection recognize two genealogical processes with markedly different time scales. These are the genealogy of functionally different alleles (the genealogy of allelic lineages) and the genealogy of functionally equivalent alleles (the gene genealogy within allelic lineages). In panmictic population models, the genealogy of allelic lineages is expected to have a neutral-like branching pattern, with times rescaled because of the extremely long residence times of such alleles (Takahata, 1990). The genealogy within lineages is also very similar in structure to a neutral gene genealogy, but the time scale is much shorter (Takahata, 1990; Vekemans & Slatkin, 1994; Nordborg, 1997). Our previous paper explored the consequence of relaxing the assumption of panmixia (Schierup *et al.*, 2000). We found that, as intuitively expected, the diversity between allelic lineages at such selected loci is insensitive to population structure, unlike that of alleles at a neutral locus, whereas variation within allelic lineages is strongly affected by population structure, behaving like a neutral locus with a reduced effective population size.

Here we add recombination to the model and consider the expected patterns of sequence diversity at a neutral locus linked to the locus under balancing selection (hereafter termed the selected locus). The motivation for adding such complexity to the already complex situation of balancing selection in subdivided populations is two-fold. First, intragenic recombination appears to occur in both the major histocompatibility system (MHC) (Gyllensten & Erlich, 1991; Bergstrom *et al.*, 1998), sporophytic self-incompatibility (SI) systems (Awadalla & Charlesworth, 1999), and the b1 mating gene in the fungus *Coprinus cinereus* (Badrane & May, 1999), and therefore interpretation of the patterns of sequence diversity within such loci requires an understanding of

* Corresponding author. Department of Ecology and Genetics, University of Aarhus, Ny Munkegade, Building 540, DK-8000 Aarhus C., Denmark.

how they behave under strong, but not absolute, linkage. Since strong balancing selection leads to increased neutral diversity over a chromosomal region, as discussed next, it is of interest to study the size of this region and the extent to which diversity is altered at sites within the gene that are not under selection, and at linked loci. Secondly, such an understanding should help to evaluate the potential to detect balancing selection through analysis of marker genes, such as microsatellites, linked to putatively selected loci. For instance, the number and frequencies of alleles at linked loci have been used to infer selection on MHC loci (Paterson, 1998; Meagher & Potts, 1997).

If a neutral locus is linked to a locus under balancing selection, variants at the neutral locus will be associated with different allelic lineages at the selected loci. Since the allelic lineages at the selected locus are maintained at intermediate frequencies by balancing selection, the alleles at linked neutral loci can be viewed as being subdivided into allelic classes (corresponding to the selected lineages maintained in the population), and can move between classes by recombination. This subdivision is equivalent to the finite island model of subdivision but with the deme sizes being variable over time (due to random fluctuations in frequencies of allelic lineages at the selected locus: see Vekemans & Slatkin, 1994). There are thus two types of subdivision: into demes and into allelic classes. Each of these, on its own, is relatively well understood theoretically as regards its consequences for neutral genetic diversity. Neutral diversity in the total population is increased by population subdivision under the finite island model (Takahata, 1988; Slatkin, 1991), and in panmictic populations it is increased by linkage to a locus under balancing selection (Strobeck, 1983; Hudson & Kaplan, 1988). Since both population subdivision and linkage to selected allelic lineages can be thought of as types of subdivision, into demes and allelic classes, respectively, with genes moving between these classes by migration and recombination, respectively (Charlesworth *et al*., 1997; Nordborg, 1997), both nucleotide diversity and Tajima's D values for a neutral locus should increase when a set of genotypes is subdivided into demes or allelic classes. The purpose of this study is to investigate how the two types of subdivision interact to affect patterns of sequence and allelic diversity at linked neutral loci within and between subpopulations. We focus on qualitative patterns, which do not depend strongly on the parameter values chosen for the model, and we seek general patterns across the different models of balancing selection, in order to discuss the relevance to empirical data without relying heavily on details of the models studied. To investigate the effect of varying selection intensity, we mainly present results for gametophytic self-incompatibility (GSI) and for symmetrical overdominance with $s = 0{\cdot}2$, which represent models with strong and moderate selection, respectively.

2. Methods

Forward two-locus simulations assuming a population of S demes, each with N individuals, giving a total population size of $N_t = SN$, were performed according to the finite island model with pollen, but not seed, migration. We assume a migration rate equivalent to migration of diploids at rate m, as described in the companion paper (Schierup *et al*., 2000). As in that paper, the model assumes a locus subject to balancing selection, at which new functional alleles arise at a rate u per generation, under the *infinite alleles* model (Kimura & Crow, 1964). The infinite alleles model is appropriate under the assumption, which is reasonable for self-incompatibility and the other systems of interest, that mutations causing changes in specificity may potentially occur at many different sites in the sequence, with each amino acid replacement that alters specificity producing a new type. Here, we concentrate on models of symmetrical overdominance and gametophytic self-incompatibility (GSI). To study the effects of linkage to the selected locus, a selectively neutral locus was added to the simulations, with a recombination frequency of r per generation with the selected locus. For study of the number of different alleles maintained at the neutral locus (see below), variation was introduced at a mutation rate of v per generation, also according to the infinite alleles model. Each run was started with $2N_t$ different alleles in the population at both the selected and neutral locus, and allowed to evolve for 50000 generations, at which time approximate mutation–selection–drift equilibrium had been reached. Gene genealogies at the neutral locus were then tracked by the method of Vekemans & Slatkin (1994), as described in the companion paper (Schierup *et al*., 2000). Simulations were run for a number of generations equal to three times the time needed for all genes at the neutral locus to coalesce. At this time the gene genealogy at the neutral locus is at an approximate equilibrium (tested by sampling at intervals). The genealogy of allelic lineages at the selected locus approaches equilibrium much more slowly than the neutral loci, because of these lineages' extremely long residence times, and is probably not in equilibrium for most parameter values we have studied. However, the number of allelic lineages at the selected locus is the relevant parameter for the behaviour of the genealogy of alleles at the neutral locus, and this approaches equilibrium quickly.

The following statistics were computed for the neutral locus:

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

(*a*) The number of different alleles and expected heterozygosity, both within demes and in the total population.

(*b*) The average pairwise coalescence times of alleles. The average of this quantity was computed within demes (T_S) and in the total population (T_T). Under the finite island model in the absence of selection or linkage to selected loci, the expected T_S and T_T values are given by:

$$T_S = 2N_t,$$

and

$$T_T = 2NS + [(S-1)^2]/2Sm,$$

respectively (Slatkin, 1991). Assuming the *infinite sites* model, the expected nucleotide diversities, π, are directly proportional to the pairwise coalescence times:

$$\pi_S = 2T_S x/G$$

and

$$\pi_T = 2T_T x/G,$$

where x is the mutation rate per gene and G is the number of nucleotides in the gene.

(*c*) The fixation index, F_{ST}. This quantifies genetic differentiation among demes, and can be computed as

$F_{ST} = (T_T - T_S)/T_T$ (Slatkin, 1991).

We compare this F_{ST} with the value of G_{ST} expected under the finite island model in the absence of selection:

$$G_{ST} = \left(1 + 2N\left(\frac{S}{S-1}\right)\left(\frac{1}{(1-m)^2(1-v)^2} - 1\right)\right)^{-1}$$

(Takahata & Nei, 1984). (1)

(*d*) Tajima's D statistic (Tajima, 1989) measures deviation in the topology of the gene genealogy from that expected under neutrality in a panmictic population. It is computed as

$$D = \frac{(2T_i - L_i/a_1)}{\sqrt{[e_1 L_i + e_2 L_i(K_i - 1)]}}, \quad i \in \{S, T\}, \qquad (2)$$

where L_i is the total length of the tree, i.e., the sum of all branch lengths in the gene genealogies, the subscript i denotes S or T depending on whether D is calculated from samples from the deme or total population, respectively, and a_1, e_1 and e_2 are constants given in Tajima (1989). We substituted L_i for the number of segregating sites, Sn_i, and T_i for the average number of nucleotide differences between two genes (i.e. the nucleotide diversity times the gene length) in Tajima's (1989) formula because $Sn_i = L_i x$ and, as just explained, $\pi_l = 2T_i xG$. D calculated from coalescence times is the value expected when the genealogy of the sample is known with certainty, i.e., when the number of segregating sites approaches infinity. This is because the standard deviation of the estimation of branch length is the square root of its expectation when the number of mutations follows a Poisson distribution. In practice, when the number of segregating sites exceeds 100, the approximation is fairly accurate (results not shown).

(*e*) The mean pairwise coalescence time of alleles sampled at the neutral locus, conditional on being associated with the same allelic lineage at the selected locus, T_A. This was obtained by computing, for each lineage at the selected locus, the average pairwise coalescence times for all alleles at the neutral locus associated with the lineage, and then taking the average over all allelic classes at the selected locus. We checked our values of T_A for a panmictic population for different recombination rates r, with those expected under the derivation of Takahata & Satta (1998) for a selectively neutral site partially linked to a site under symmetrical overdominance. For a given set of parameters the number of replicates was 100, except for the smallest migration rate ($Nm = 0{\cdot}005$), where only 20 replicates could be done within a reasonable time.

3. Results

(i) *Nucleotide diversity at the neutral locus within and between demes*

Fig. 1 shows mean coalescence times between randomly sampled pairs of alleles at the neutral locus, for several linkage distances from the selected locus (r) and migration rates among demes (Nm). Results are shown with symmetrical overdominance with $s = 0{\cdot}2$, and for selection caused by gametophytic self-incompatibility (GSI). Coalescence times are given for sampling at two levels, within demes (T_S, Fig. 1*a*) and in the total population (T_T, Fig. 1*b*). The migration axis also shows the numbers of alleles in the total population at the selected locus at equilibrium, when the mutation rate to new functional allelic specificities is $u = 10^{-6}$. These numbers depend on the migration rate (Schierup, 1998). The standard deviations were 25–40% of the mean values.

Within demes, pairwise coalescence times increase monotonically with decreasing r for all Nm values (Fig. 1*a*), showing that linkage to the selected locus produces the expected peak of diversity. For unlinked loci ($r = 0{\cdot}5$), T_S values are small and independent of Nm. These results are expected from theoretical treatments (Maruyama, 1971). At intermediate r values, T_S values are highest for low Nm. For samples within demes, the peak of nucleotide diversity (linearly

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

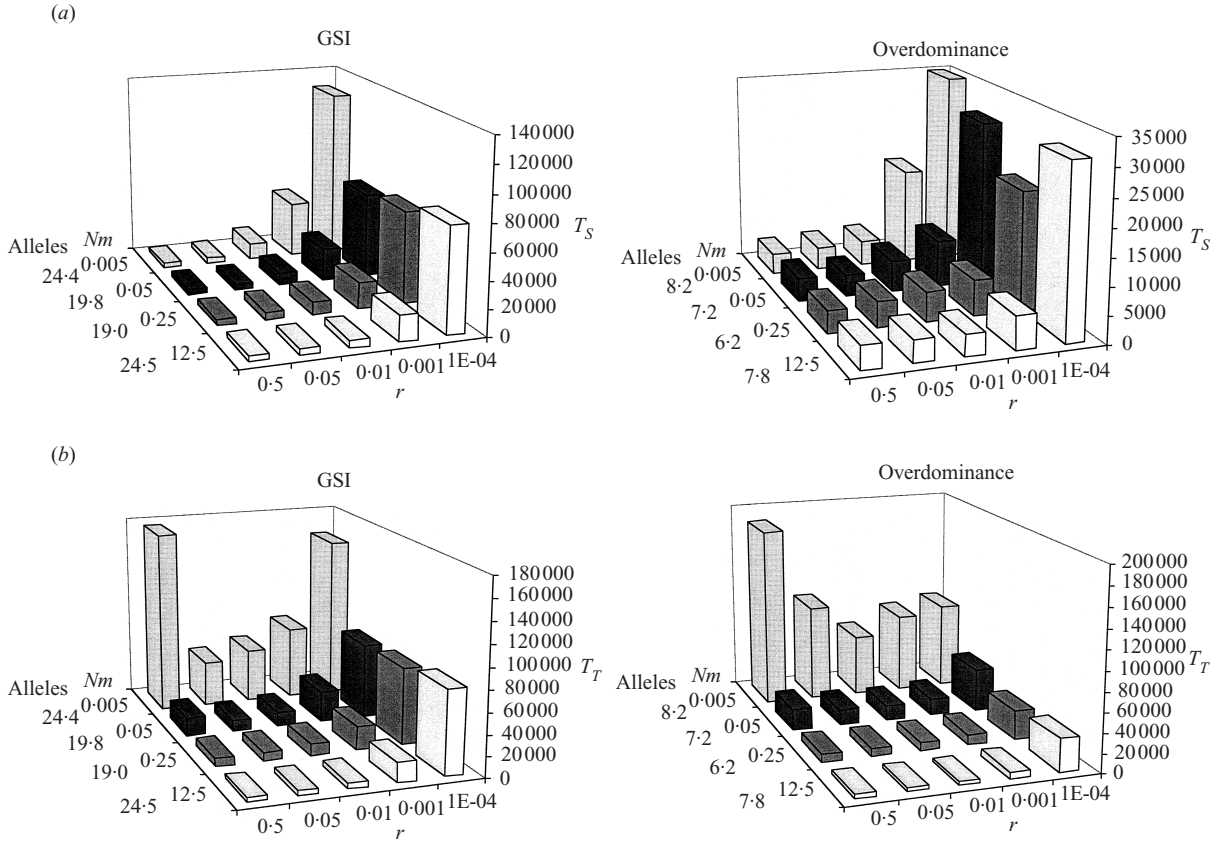

Fig. 1. Diversity at a neutral locus (shown as the pairwise coalescence times in generations) as a function of migration rate (Nm) and rate of recombination (r) to a locus under balancing selection. Results are shown for GSI and overdominance with $s = 0{\cdot}2$. There are 40 demes with 50 individuals. The number of functional alleles maintained with $u = 10^{-6}$ is also shown on the migration axis. (a) Results for sampling at the deme level, (b) results for sampling at the total population level. The number of replicates is 100 except for $Nm = 0{\cdot}005$ (20 replicates).

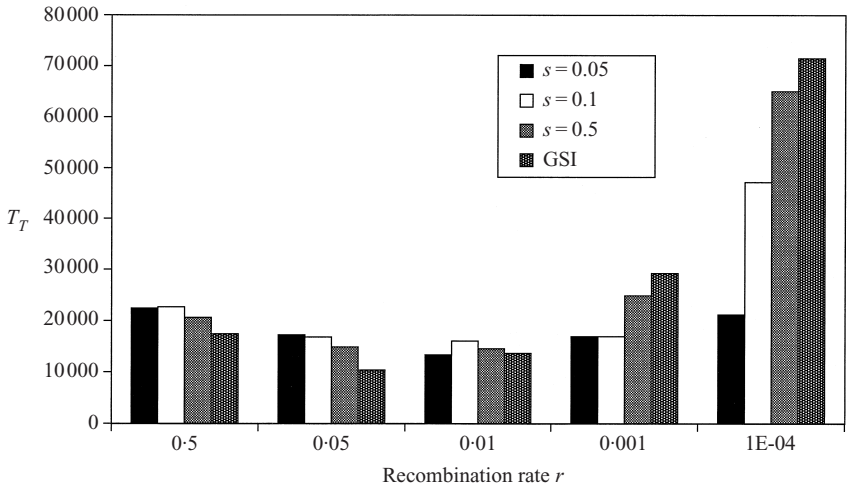

Fig. 2. Diversity (shown as the pairwise coalescence times in generations) as a function of recombination rate (r) for four different selection intensities: GSI, and overdominance with $s = 0{\cdot}05$, $s = 0{\cdot}1$, and $s = 0{\cdot}5$. There are 40 demes with 50 individuals, and a migration rate of $Nm = 0{\cdot}05$.

proportional to T_S) around the selected locus is thus broader in a more subdivided population than in a panmictic one.

The pattern is qualitatively different at the total population level (Fig. 1 *b*). When Nm is high, pairwise coalescence times again increase monotonically close to the selected locus. With low Nm, however, there is a minimum at intermediate linkage. The r value at which this minimum occurs depends on the value of Nm and on the strength of selection (comparing

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

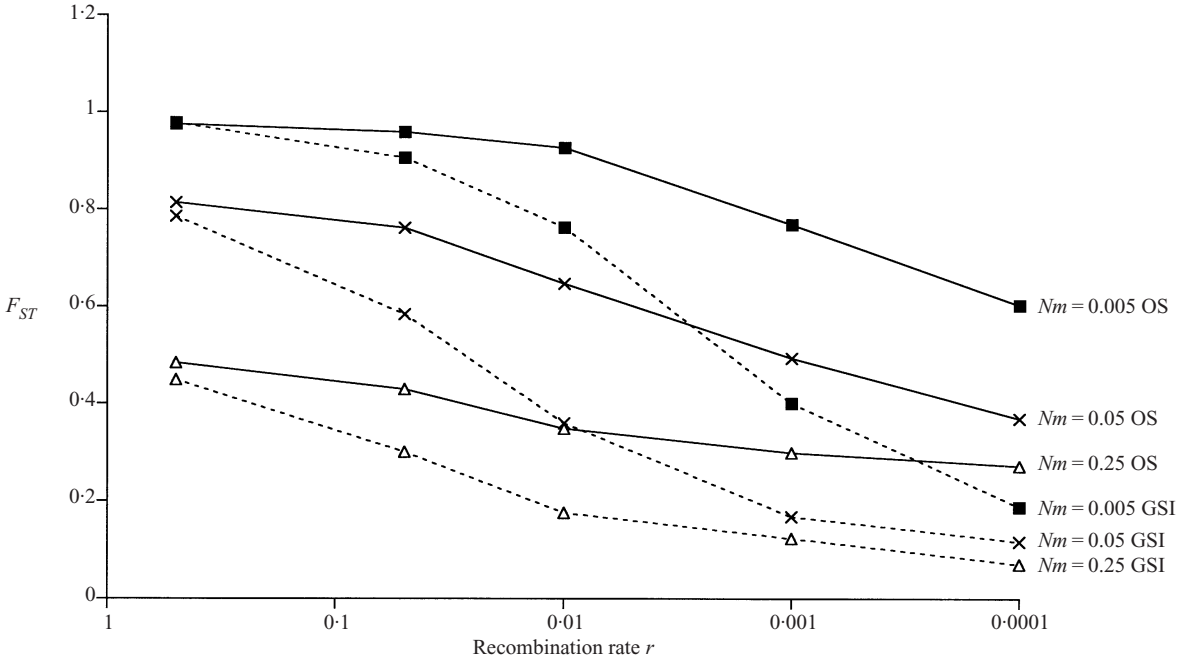

Fig. 3. F_{ST} as a function of recombination distance (r) to a locus under either overdominant selection (OS) with $s = 0{\cdot}2$ or GSI. Results are shown for three different migration rates. There are 40 demes with 50 individuals, and a mutation rate $u = 10^{-6}$ at the selected locus.

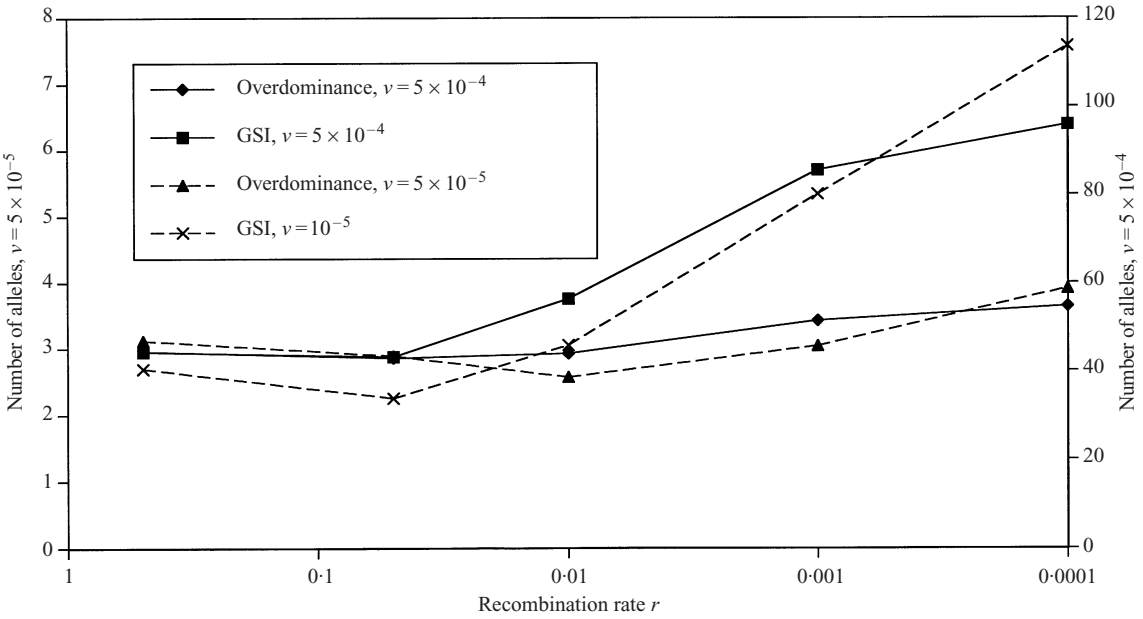

Fig. 4. The number of alleles maintained at the neutral locus as a function of the recombination rate r to either a locus under overdominant selection with $s = 0{\cdot}2$ or GSI. Results are shown for two mutation rates at the neutral locus, $v = 10^{-5}$ (primary y-axis, continuous lines) and $v = 5 \times 10^{-4}$ (secondary y-axis, dotted lines). There are 40 demes with 50 individuals and a mutation rate $u = 10^{-6}$ at the selected locus. $Nm = 0{\cdot}05$.

overdominance and GSI, with GSI representing stronger balancing selection than symmetrical overdominance with $s = 0{\cdot}2$: see Vekemans & Slatkin, 1994). With very low migration, the effect can be so large that expected pairwise coalescence times at loci as closely linked to the selected locus as with $r = 0{\cdot}001$ ($N_t r = 2$) can be lower than those between alleles at unlinked loci, i.e., the peak of diversity appears very narrow in the population as a whole. The decrease in T_T at intermediate levels of Nr is most dramatic for the GSI model (stronger selection). The effect of varying the selection coefficient is shown in more detail in Fig. 2, where the dependence of T_T values on r is shown for a single value of Nm (0·05), but with various selection

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

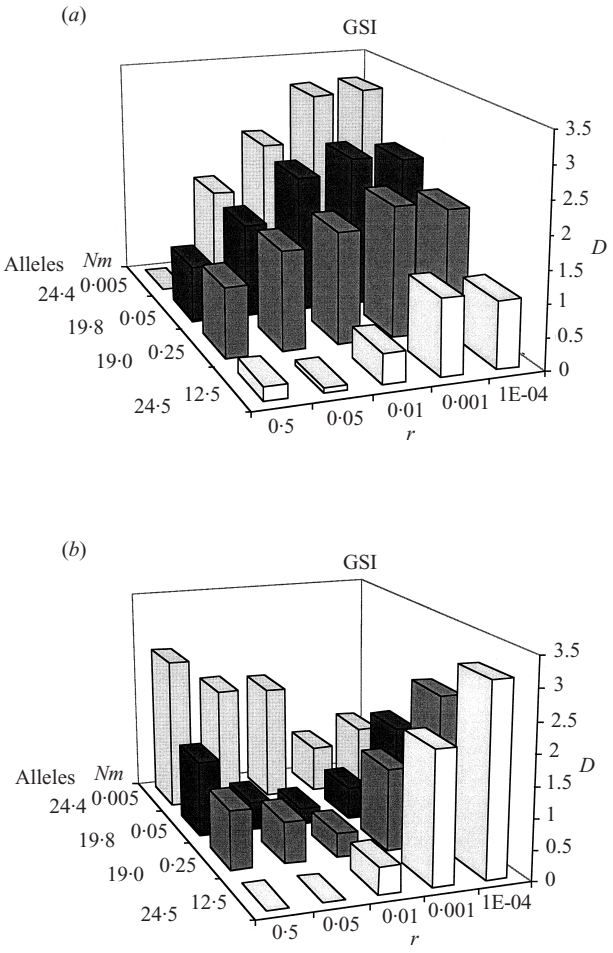

Fig. 5. Tajima's D as a function of migration rate (Nm) and recombination rate (r) to a locus under balancing selection. Results are shown for the GSI model. There are 40 demes with 50 individuals. The number of functional alleles maintained with $u = 10^{-6}$ is also shown on the migration axis. (*a*) Results for sampling at the deme level, (*b*) results for sampling at the total population level. The number of replicates is 100 except for $Nm = 0{\cdot}005$ (20 replicates).

coefficients. The decrease in T_T is largest for strong selection, but reaches closer to the selected locus for weaker selection. Identical simulations were run for the three models of sporophytic SI involving dominance among alleles (see Schierup *et al*., 2000), and qualitatively similar patterns of coalescence times with Nm and r were found.

Fig. 3 shows the effects on F_{ST}, calculated from the coalescence times shown in Fig. 1. In the accompanying paper (Schierup *et al*., 2000), we show that population differentiation at a locus subject to balancing selection is low even when migration is very restricted. Linkage to a selected locus also has a substantial effect on patterns of differentiation. Even when the neutral locus is unlinked, F_{ST} is expected to be lower than the expectation from (1), because the selected locus is in linkage disequilibrium even with an unlinked locus, in genotypes migrating into a deme. This is also illustrated in Fig. 3 where F_{ST} values are slightly lower for stronger selection (comparing GSI with overdominance for $r = 0{\cdot}5$ and $Nm = 0{\cdot}25$ or $0{\cdot}05$). However, the decrease in F_{ST} for an unlinked locus compared with expectation for a neutral locus was never found to exceed 10%. For a given level of subdivision, F_{ST} decreases monotonically with decreasing r, and the decrease is larger with stronger selection. For intermediate migration rates ($Nm =$ $0{\cdot}05$ and $0{\cdot}25$), the effect of the selected locus on the neutral one is noticeable even for relatively loose linkage ($r = 0{\cdot}01$), provided that selection is sufficiently strong, as can be seen in the case of GSI.

(ii) *The number of alleles and diversity*

The same set of simulations also yielded allele numbers and diversity (expected heterozygosity) values at the neutral locus under the infinite alleles model. Two mutation rates ($v = 5 \times 10^{-4}$ and $v = 10^{-5}$) were chosen for study, to represent reasonable values for microsatellites or isozymes, respectively (Amos *et al*., 1996; Nei, 1987). Fig. 4 shows the number of alleles for the total population, for $Nm = 0{\cdot}05$. This suffices to illustrate the main similarities and differences compared with those for coalescence times. Comparison of Figs. 1 and 4 shows that the numbers of alleles and T_T co-vary closely, both showing minima at intermediate linkage. The minimum number of alleles is found at a higher recombination rate in the model with stronger selection (GSI) because the effect of associative overdominance is larger. When measured as the number of alleles maintained, the peak of diversity around the selected locus is less dramatic but it spans a somewhat broader region than was observed with coalescence times.

(iii) *Tajima's* D *within demes and in the total population*

Several statistics have been proposed to test for deviations of nucleotide sequences from expectations in a panmictic population under the neutral model. We focus on Tajima's D because it is widely applied, and it has been shown to be among the most powerful statistics while relying on relatively few assumptions (Fu & Li, 1993; Simonsen *et al*., 1995). Tajima's D is the difference between the estimates of $4Nu$ calculated from the pairwise nucleotide differences, and the number of segregating sites, Sn, standardized by the standard deviation of the difference, and should therefore have an expectation of zero with standard deviation one in an ideal population; simulations show that the expectation is slightly below zero (Tajima, 1989). Instead of Sn and the number of pairwise nucleotide differences we used the total times in the genealogy and the pairwise coalescence times as

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

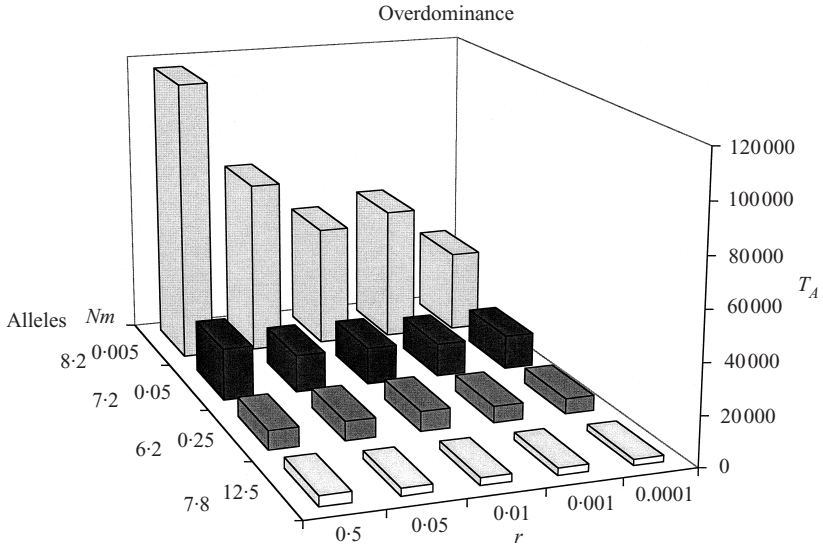

Fig. 6. The diversity for alleles at the neutral locus associated with the same functional allele at a locus under overdominant selection ($s = 0{\cdot}2$) as a function of migration rate (Nm) and recombination rate (r). Results are for sampling at the total population level. There are 40 demes with 50 individuals. The number of functional alleles maintained with $u = 10^{-6}$ is also shown on the migration axis. The number of replicates is 100 except for $Nm = 0{\cdot}005$ (20 replicates).

explained in Section 2. Selection is detectable by this test, because it changes the frequency distribution of polymorphic variants. Balancing selection increases variant frequencies so that the number of pairwise nucleotide differences tends to exceed the value expected from the neutral theory, given the number of segregating sites, producing positive Tajima's D values (Tajima, 1989). Positive values are also expected for a subdivided population sampled at the total population level (Simonsen *et al.*, 1995). In both cases, Tajima's D is positive because the subdivision (into demes or allelic classes) separates the alleles into a number of classes within which coalescences occur rapidly, but between which coalescences occur more slowly (Nordborg, 1997).

Fig. 5 shows how Tajima's D values for the neutral locus are affected by Nm and r, for the GSI model. When alleles are sampled within demes (Fig. 5*a*), Tajima's D increases with decreasing values of r for all migration rates (with overdominance the results are similar and are not shown). Linkage to the selected locus increases Tajima's D in a broader region when migration is low than when it is frequent, in line with the pattern for pairwise coalescence times (Fig. 1*a*). Standard deviations (not shown) were up to 1·5 units, so the standard errors for $Nm = 0{\cdot}005$, where only 20 replicates were made, are quite large.

When alleles are sampled from the population as a whole, however, this is no longer true. As expected, Tajima's D for loci unlinked to the selected locus increases monotonically with decreasing migration (population subdivision). Similarly, if the population is panmictic, D increases with increasing linkage to the selected locus (subdivision into allelic lineages at the selected locus). However, in a subdivided population, greatly decreased Tajima's D values can arise for genes closely linked to the selected locus. In other words, Tajima's D is much less likely to detect selection in a subdivided population, unless the neutral locus is extremely tightly linked to the selected locus ($N_t r \ll 1$). When Ns was decreased, Tajima's D values were, not surprisingly, less likely to detect the selected locus (results not shown), paralleling the reduced effect of the selected locus on nucleotide diversity with weaker selection.

(iv) *Nucleotide diversity within allelic classes*

In a panmictic population with complete linkage, mean pairwise coalescence times within allelic lineages, T_A, are expected to be much smaller than between allelic lineages, because the coalescence process within a lineage should behave approximately like a neutral coalescence process with an effective population size determined by the harmonic mean over time of the number of copies of the allelic lineage (Vekemans & Slatkin, 1994; Takahata & Satta, 1998). Population subdivision, however, increases diversity within allelic lineages, even though coalescence times between lineages are largely unaffected (see Schierup *et al.*, 2000).

Pairwise coalescence times of genes at the neutral locus within allelic lineages at the selected locus are shown in Fig. 6 for the overdominance model, for various recombination frequencies. For the high migration rate ($Nm = 12{\cdot}5$), T_A decreases as r decreases, deviating less than 10% from theoretical expectations for a panmictic population according to

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

Takahata & Satta (1998) (results not shown). With restricted migration, T_A also decreases with decreasing r. Overall, subdivision increases diversity within allelic classes for any level of linkage, with the relative increase being of similar magnitude.

4. Discussion

The primary purpose of this study was to investigate the qualitative effects of a locus under balancing selection on variation at linked neutral loci, allowing for population subdivision. Several genetic systems, in particular self-recognition systems, with these characteristics are known, which motivate the explicit modelling. The model extends the results of Charlesworth *et al.* (1997) who investigated a two-deme model with a locus with two alleles maintained by balancing selection at equal frequency in each deme. Here, population subdivision into many demes amplifies the effects of the polymorphism at the selected locus, but the same qualitative effects are expected as in a two-deme model. The model of subdivision studied is clearly a very simple one and we investigated only a limited set of parameters in our simulations. We therefore focus on qualitative patterns and their implications for detecting selection at a locus using data from linked markers. These qualitative results should be robust to deviations from the model analysed, such as the number of subpopulations, and migration of diploids, rather than pollen.

Within demes, both π and Tajima's D increase with increasing linkage to the selected locus. The peak of nucleotide diversity extends over a greater map distance the greater the subdivision. Two factors can account for this result. First, fewer allelic lineages are maintained in each deme at the selected locus when Nm is small, leading to stronger selection for each lineage. Secondly, the deme effective size decreases with decreasing Nm for neutral genes linked to a selected locus. This is because selection leads to non-conservative migration (Nagylaki, 1982), which invalidates Maruyama's (1971) invariance principle for the case of strict neutrality in the island model. This lowers the effective recombination rate, $N_t r$. The relative contributions of these two factors are not easy to disentangle, in quantitative terms.

Our results show, however, that the two types of subdivision, into demes and allelic classes, do not act additively on π and Tajima's D values when alleles are sampled from the total population. When both types of subdivision are simultaneously present, the increase in π is smaller than with each separately, and the ability to detect either subdivision or balancing selection through effects on the site frequency spectra (e.g., using Tajima's D test) is greatly reduced. Unlike panmictic populations, when a population is physically subdivided, π estimated from a total population sample reaches a minimum value for an intermediate level of linkage. The reason for this is that two opposing forces affect π. Linkage to a selected allele increases π (associative overdominance: see Frydenberg, 1963; Strobeck, 1983; Charlesworth *et al.*, 1997). Subdivision also increases π through an increased effective population size of the whole population (Nei & Takahata, 1993). Linkage to an allelic lineage at the selected locus, however, increases the effective migration rate through hitchhiking, as evidenced by the decrease in F_{ST} with linkage (Fig. 3). This decreases π relative to the unlinked case. At intermediate linkage, the decrease in π caused by hitchhiking outweighs the increase caused by associative overdominance. The parallel pattern in Tajima's D has the same basic cause. Tajima's D can be interpreted as measuring the relative number of old versus new mutations in the genealogy (Fu, 1997). In a subdivided population, many of the most recent coalescences happen within demes on a fast time-scale (Nordborg, 1997). Similarly, coalescence is fast among neutral variants within a given selected allele class. Thus, with only one of these types of subdivision, long branches with many 'old' mutations connect demes (under physical subdivision) and different selected alleles (under balancing selection). With both types of subdivision, however, neutral alleles associated with any given selected allele are likely to be present in different demes, and neutral alleles in particular demes are likely to be associated with the different selected alleles, so the genealogy will display fewer old mutations, and will thus have a closer to neutral frequency spectrum of variants, and low Tajima's D if data are analysed from the set of demes as a whole.

These results demonstrate that, if one wishes to infer the operation of balancing selection, the sampling strategy is very important if the species in question is, or has recently been, subdivided. Furthermore, no species conforms exactly to the symmetrical island model with equal migration. In particular, migration rates and deme sizes are expected to vary over time. This might cause some of the non-linear characteristics discussed above for sampling at the total population level to arise even in a sample from a single deme, and the potentiality for detecting balancing selection by Tajima's D may thus be reduced even for the best possible sampling strategy. The magnitude of such effects remains to be investigated.

(i) *Implications for neutral sites within a locus with balanced polymorphism*

Recombination and/or gene conversion probably occur within genes in both the MHC (Bergstrom *et al.*, 1998) and *Brassica* sporophytic SI system (Awadalla & Charlesworth, 1999), and interpretations of patterns of nucleotide diversity must therefore

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

incorporate this, together with the possibility of population subdivision.

In MHC (HLA) loci in humans, high levels of polymorphism are thought to be maintained by balancing selection acting on the peptide binding region of these proteins (Hughes & Nei, 1988; Ayala, 1995; Salamon *et al*., 1999). Recently, silent nucleotide diversity has been estimated for several MHC regions at various recombination distances to the peptide binding regions, both within and between allelic classes defined either by serotypes or inferred from nucleotide sequence similarity (Takahata & Satta, 1998; Satta *et al*., 1998). Takahata & Satta (1998) calculated the expected diversity within lineages as a function of recombination for a simple population genetic model of symmetrical overdominance and random mating. When this model is fitted to the data, it is difficult to account for the high diversity within lineages under any reasonable value of the recombination rate (Takahata & Satta, 1998). Our results show that the inclusion of subdivision may account for some of this discrepancy, because within-lineage diversity can be considerably increased without affecting the between-lineage diversity. However, the extent of subdivision of the human population during the time span relevant for the polymorphism at the HLA loci is at present unclear (Harpending *et al*., 1998; Harris & Hey, 1999).

In the sporophytic SI system of Brassicaceae, a number of functionally different alleles have been sequenced from the two loci involved in self-incompatibility: SLG and SRK (Kusaba *et al*., 1997). The regions of the sequence that are involved in the recognition have not been identified with any degree of certainty, so it is not yet possible to ask about diversity at sites at different linkage distances from these parts of the sequence. However, it is known that levels of polymorphism throughout the two genes are high, even several hundred base pairs from the S-domain hypervariable regions, which are found in all SI loci that have been studied (Sims, 1993; Kusaba *et al*., 1997) and may include the specificity-determining sites (Awadalla & Charlesworth, 1999). In the SRK locus, diversity is high in the kinase region, which is separated from the S-domain by a large intron and is probably not involved in recognition functions. In this gene, diversity is found in both exons and introns. The allele sequences currently available are taken from cultivated strains of the species studied, so that subdivision of the population from which they originated is quite likely and this may influence the diversity patterns. Therefore, it will be valuable to obtain comparable results from natural populations, and to compare randomly sampled alleles from within individual populations, as well as between them. Furthermore, since there is evidence for recombination both within the SLG gene, and also between the SLG and SRK loci (Awadalla & Charlesworth, 1999), it would be worthwhile to test whether intron diversity declines with distance from the SRK locus S-domain. At present, however, not enough introns have been sequenced for this to be possible. Intron data from self-incompatibility genes in species with gametophytic self-incompatibility systems would also be valuable.

(ii) *Implications for linked neutral loci*

It is often assumed that greater diversity should be found at markers linked to the MHC than at loci elsewhere in the genome. In an early study of *Mus musculus*, Nadeau *et al*. (1982) found that isozymes on chromosome 7, where the MHC is located, had significantly higher diversity and more alleles than loci on other chromosomes. Since the mice analysed were sampled world-wide and *M. musculus* populations are highly subdivided (Dallas *et al*., 1998), this is unexpected unless the effective size of the metapopulation is extremely small, or the loci studied are extremely closely linked to MHC. A more detailed study using microsatellites found more alleles at loci within 1 cM of the MHC than at loci 2–5 cM from the MHC (Meagher & Potts, 1997). Applying our finding that a detectable peak of diversity is unlikely under any level of subdivision to extend further than a linkage distance corresponding to $N_t r = 10$, and assuming that the peak of diversity in *M. musculus* extends a minimum of 0·2–0·5 cM, we would then have to infer a maximum N_e value of about 2000–5000 for the *M. musculus* population studied, and possibly much lower if it is subdivided. Since the origin of the inbred strains used was not stated it is difficult to know whether this number represents the world-wide population size or a regional one.

In Soay sheep sampled from a single population, Paterson (1998) reported increased heterozygosities for two microsatellites located within, and at a distance of 2·6 cM from the polymorphic class II *DRB* locus of the MHC, compared with three microsatellites at distances of 5·5, 15·8 and 17·2 cM. Furthermore, Watterson's test (Watterson, 1978) detected deviations from neutrality at the two closely linked loci but not at the more distant loci. The increased polymorphism for the microsatellite as far as 2·6 cM away from DRB, together with significant linkage disequilibrium between this microsatellite and DRB, suggests that the effective size of this population is also very low (less than 100).

In the HLA-H pseudogene in humans, Grimsley *et al*. (1998) reported a more than 10-fold greater nucleotide diversity than is generally found in the genome. However, Tajima's *D* was not significant. The HLA-H locus is located approximately 100 kb from the highly polymorphic HLA-A locus. Grimsley *et al*. (1998) concluded that the high HLA-H diversity

https://doi.org/10.1017/S0016672300004547 Published online by Cambridge University Press

is most probably caused by balancing selection at the HLA-A, rather than being caused by selection at HLA-H before it became a pseudogene. Assuming a recombination rate of 0·1 cM/Mb (Satta *et al*., 1998), HLA-H would be at a distance $r = 0{\cdot}0001$–$0{\cdot}001$ from HLA-A, a 10-fold increase in diversity compared to an unlinked locus would require a human effective population size of 500–5000, depending on the selection intensity at HLA-A (see Fig. 1). This calculation is based on the assumption of panmixia. Past and current subdivision of human populations would decrease the population size estimate. Most other estimates are higher than this (the minimum reported by Satta *et al*. (1998) is 10000). This discrepancy implies that at least part of the increased diversity at the HLA-H locus may have to be attributed to direct selection in the past.

In conclusion, it appears that the ability to detect the presence of a multi-allelic locus under balancing selection in a species with population structure requires sampling at the deme, rather than the total population level. If sampling is both within and between populations, and if data are available from reference loci unlinked to the gene of interest (to help provide evidence about population structure), F_{ST} analysis may offer a better way to detect balancing selection than tests based on the frequency spectrum of alleles such as Watterson's and Tajima's D tests. It is clear, however, that the overall effect of multi-allelic balancing selection on variation in flanking regions is limited in a subdivided population, unless effective population sizes are low. There is, however, evidence that this may be the case in several empirical studies reviewed here. At present, the studies are few and scattered, and a bias may exist against publishing studies with no apparently interesting effects, so firm conclusions about the details of selection, even in the well studied MHC region, are not yet possible.

The study was supported by a post-doctoral grant from the Carlsberg Foundation to M. H. S., a NERC of Great Britain Senior Research fellowship to D. C., and by a travel grant from the Belgian National Fund for Scientific Research to X. V. We thank P. Awadalla, B. Charlesworth and G. T. McVean for discussions, two anonymous reviewers for helpful comments, and the Department of Computer Sciences, University of Aarhus for computing facilities.

References

Amos, W., Sawcer, S. J., Feakes, R. W. & Rubinsztein, D. C. (1996). Microsatellites show mutational bias and heterozygote instability. *Nature Genetics* **13**, 390–391.

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

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

Badrane, H. & May, G. (1999). The divergence–homogenization duality in the evolution of the b1 mating type gene of *Coprinus cinereus*. *Molecular Biology and Evolution* **16**, 975–986.

Bergstrom, T. F., Josefsson, A., Erlich, H. A. & Gyllensten, U. (1998). Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nature Genetics* **18**, 237–242.

Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**, 155–174.

Dallas, J., Bonhomme, F., Boursot, P. & Britton-Davidian, J. (1998). Population genetic structure in a Robertsonian race of house mice: evidence from microsatellite polymorphism. *Heredity* **80**, 70–77.

Frydenberg, O. (1963). Population studies of a lethal mutant in *Drosophila melanogaster*. I. Behaviour in populations with discrete generations. *Hereditas* **50**, 89–116.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Fu, Y. X. & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Grimsley, C., Mather, K. A. & Ober, C. (1998). HLA-H: a pseudogene with increased variation due to balancing selection at neighbouring loci. *Molecular Biology and Evolution* **15**, 1581–1588.

Gyllensten, U. B. & Erlich, H. A. (1991). Shared epitopes among HLA class II alleles: gene conversion, common ancestry and balancing selection. *Immunology Today* **12**, 411–414.

Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R. & Sherry, S. T. (1998). Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the USA* **95**, 1961–1967.

Harris, E. E. & Hey, J. (1999). X chromosome evidence for ancient human histories. *Proceedings of the National Academy of Sciences of the USA* **96**, 3320–3324.

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

Hughes, A. L. & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–70.

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

Kusaba, M., Nishio, T., Satta, Y., Hinata, K. & Ockendon, D. (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. *Proceedings of the National Academy of Sciences of the USA* **94**, 7673–7678.

Maruyama, T. (1971). An invariant property of a structured population. *Genetical Research* **18**, 81–84.

Meagher, S. & Potts, W. K. (1997). A microsatellite-based MHC genotyping system for house mice (*Mus musculus*). *Hereditas* **127**, 75–82.

Nadeau, J. H., Collins, R. L. & Klein, J. (1982). Organization and evolution of the mammalian genome. 1. Polymorphism of *H*-2 linked loci. *Genetics* **102**, 583–598.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159-172.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M. & Takahata, N. (1993). Effective population-size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution* **37**, 240–244.
Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.
Paterson, S. (1998). Evidence for balancing selection at the major histocompatibility complex in a free-living ruminant. *Journal of Heredity* **89**, 289–294.
Salamon, H., Klitz, W., Esteal, S., Gao, X., Erlich, H., Fernandez-Vina, M., Trachtenberg, E. A., McWeeney, S. K., Nelson, M. P. & Thomson, G. (1999). Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics* **152**, 393–400.
Satta, Y., Li, Y.-J. & Takahata, N. (1998). The neutral theory and natural selection in the HLA region. *Frontiers in Bioscience* **3**, 459–467.
Schierup, M. H. (1998). The number of self-incompatibility alleles in a finite, subdivided population. *Genetics* **149**, 1153–1162.
Schierup, M. H., Vekemans, X. & Charlesworth, D. (2000). The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research* **76**, 51–62.
Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.
Sims, T. M. (1993). Genetic regulation of self-incompatibility. *Critical Reviews of Plant Sciences* **12**, 129–167.
Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167–175.
Strobeck, C. (1983). Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics* **103**, 545–555.
Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research* **52**, 213-222.
Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and transspecies evolution of polymorphism. *Proceedings of the National Academy of Sciences of the USA* **87**, 2419–2423.
Takahata, N. & Nei, M. (1984). Fst and Gst statistics in the finite Island model. *Genetics* **107**, 501–504.
Takahata, N. & Satta, Y. (1998). Footprints of intragenic recombination at HLA loci. *Immunogenetics* **47**, 430-441.
Vekemans, X. & Slatkin, M. (1994). Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165.
Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.