Genet. Res., Camb. (2000), **76**, *pp.* 51–62. *With* 4 *figures.* Printed in the United Kingdom © 2000 Cambridge University Press

The effect of subdivision on variation at multi-allelic loci under balancing selection

MIKKEL H. SCHIERUP[1]*, XAVIER VEKEMANS[2] AND DEBORAH CHARLESWORTH
[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 3 November 1999*)

Summary

Simulations are used to investigate the expected pattern of variation at loci under different forms of multi-allelic balancing selection in a finite island model of a subdivided population. The objective is to evaluate the effect of restricted migration among demes on the distribution of polymorphism at the selected loci at equilibrium, and to compare the results with those expected for a neutral locus. The results show that the expected number of alleles maintained, and numbers of nucleotide differences between alleles, are relatively insensitive to the migration rate, and differentiation remains low even under very restricted migration. However, nucleotide divergence between copies of functionally identical alleles increases sharply when migration decreases. These results are discussed in relation to published surveys of allelic diversity in MHC and plant self-incompatibility systems, and to the possibility of inferring ancient population genetic events and processes. In addition, it is shown that, for sporophytic self-incompatibility systems, it is not necessarily true in a subdivided population that recessive alleles are more frequent than dominant ones.

1. Introduction

Several different genetic systems in various organisms share the property that a large number of alleles are maintained by selection in the face of genetic drift. Such systems include homomorphic self-incompatibility in plants (Wright, 1939; Nasrallah *et al*., 1987; Anderson *et al*., 1986), major histocompatibility (MHC) systems in vertebrates (Klein 1979, 1986), sex-determination in some species of Hymenoptera (Crozier, 1977; Duchateau *et al*., 1994; Crozier & Pamilo, 1996) and mating-type genes in fungi (Raper, 1966; May & Matzke, 1995). Natural selection favouring high allelic diversity (broadly termed balancing selection) is either known (self-incompatibility systems) or inferred (MHC, and fungal incompatibility systems) to maintain and shape this variation. The understanding of how selection works in each type of system is both a fascinating evolutionary question and one of considerable practical interest.

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

The MHC loci have been investigated in great detail (recently reviewed by Hedrick, 1994; Hughes & Yeager, 1998*a*, *b*; Edwards & Hedrick, 1998). The system consists of several homologous loci. Large sequence differences have been found between alleles associated with different functional specificities (serotypes), with high ratios of replacements to silent differences (K_a/K_s values; see Hughes *et al*., 1990) in the functionally important peptide binding region of the proteins. Some of the polymorphisms also appear to date back to before species such as human and chimpanzee separated (Klein, 1986). It is therefore generally accepted that alleles at these loci are under balancing selection. Assuming per nucleotide mutation rates similar to other loci, the number of alleles and sequence divergence seem to require either negative frequency dependent selection or symmetrical overdominance (reviewed in Hughes & Yeager, 1998*b*), which are difficult to distinguish from sequence data alone (Takahata & Nei, 1990).

Homomorphic self-incompatibility (SI) systems of flowering plants can be classified into two types: gametophytic (GSI) and sporophytic (SSI) self-incompatibility. A mating in these systems is com-

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

patible only if the compatibility type of the stigma is different from that of the pollen. In GSI, the pollen phenotype is determined by the haploid pollen genotype, but in SSI by the genotype of the diploid paternal plant. In contrast to MHC, it is well established that frequency-dependent selection maintains these polymorphisms (Wright, 1939). The dynamics of alleles under GSI resemble those under symmetrical overdominance with a special form of stronger selection (formally equivalent to selection stronger than lethality of all homozygotes: see Clark, 1993; Vekemans & Slatkin, 1994). In species with SSI, frequent dominance between alleles makes the dynamics more complex (Cope, 1962). Dominance decreases the number of alleles maintained, and is in this sense equivalent to a weakening in the strength of selection (Schierup *et al*., 1997, 1998).

At the mating type genes in some fungi such as *Coprinus cinereus*, very high sequence and allelic diversity suggests that these loci are probably also subject to balancing selection (May *et al*., 1999). In these systems, mating individuals are haploids, and the control of mating behaviour usually involves multi-allelic loci (Casselton, 1997). Sex-determination in many species of Hymenoptera is also believed to involve a single multi-allelic locus. The systems differ between species, but in some species homozygotes are eliminated because they develop as diploid males that are killed by the workers (Crozier & Pamilo, 1996), equivalent to a symmetrical overdominance model with a selection coefficient of one.

The expected behaviour of systems such as these is well understood for the case of a panmictic population (e.g. MHC: Takahata, 1990; GSI: Wright, 1939; SSI: Schierup *et al*., 1997). Most species, however, do not constitute one panmictic population but rather show evidence of population structure. In order to interpret the results from experimental studies, we thus need to understand how population subdivision affects the amount and distribution of genetic diversity at multi-allelic loci under balancing selection. Taking into account population structure in models for neutral loci has been important for the interpretation of sequence data from humans and other species (Harpending *et al*., 1998; Harris & Hey, 1999), and hypotheses about migration patterns may be tested by applying simple models to data (Seielstad *et al*., 1998). However, results on the effect of population structure on neutral loci cannot be extrapolated to loci under balancing selection in subdivided populations, because migration and selection interact with one another (Clark, 1996; Schierup, 1998).

One application of empirical data has been to use the genealogy of alleles at loci under balancing selection to infer population genetic parameters in the distant past. This is assumed to be possible because such polymorphisms are expected to persist for very long times (Richman & Kohn, 1996; Vincek *et al*., 1997). The genealogical structure of functional alleles therefore reflects demographic factors over a longer time span than that of a neutral gene; Takahata (1990) thus defined the 'long-term effective population size' as the size of an ideal population that is consistent with the allelic genealogy observed. This quantity can be estimated from the number of extant allelic lineages, defined as the number of functionally different allelic types, at a given time in the past assuming an infinite alleles model, symmetrical overdominance with known selection coefficient, and no recombination. For GSI, where the strength of selection is known (see above), the existence of trans-specific polymorphism, combined with independent estimates of the divergence times of the species in question, makes an estimate of the long-term effective population size possible (Richman & Kohn, 1996). Richman *et al*. (1996) used this method to provide evidence for an ancient bottleneck in *Physalis crassifolia* and Vincek *et al*. (1997) used a similar argument to estimate the number of founders of Darwin's finches from MHC data. Population subdivision is, however, highly likely for the kinds of species to which these approaches were applied, and it is unclear how it will affect these inferences.

In this paper we extend Schierup's (1998) study, which investigated the number of alleles maintained by selection in a subdivided population, under either symmetrical overdominance or GSI. We here also include models of sporophytic self-incompatibility and the investigation of the effect of migration by examining expected levels of diversity at these loci, and sequence diversity within and between demes. Furthermore, we relax the assumption of strictly symmetrical selection on alleles by allowing for dominance among sporophytic SI alleles, and investigate how population structure affects the expected frequencies of alleles with different levels of dominance. We also study the effect on diversity within and between functionally different allelic lineages. For the case of SI, an allelic lineage consists of all alleles with identical specificity. Two types of genealogical processes must therefore be studied: the genealogy of functionally different allelic lineages in a sample, and the genealogy of all alleles (the entire gene genealogy).

2. Methods

(i) Notation and theory

A locus under various forms of balancing selection is modelled. Functionally different allelic lineages are assumed to arise by mutations according to the *infinite alleles* model. We use the notation H for the gene diversity for functionally different alleles. In addition,

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

we are interested in nucleotide sequence diversity at the locus, assuming the *infinite sites* model. We consider genealogical processes characterized in terms of pairwise coalescence times (denoted by T) and times until coalescence of all alleles in a sample (Tc). Under the infinite sites model, pairwise coalescence times are proportional to the expected number of nucleotide differences between pairs of alleles, and Tc is the expected age of the observed polymorphism in the system.

Following the notation of Nei (1987), for subdivided populations, we use the subscript T to denote the coalescence times of genes sampled from the total population, and subscript S for those sampled within single demes. Subscripts will be also used to distinguish whether we are referring to the genealogy of functionally different allelic lineages in a sample (subscript F), or to the genealogy of allele sequences having the same functional allelic class (i.e. within a given allelic lineage, subscript A). We denote the actual number of functionally different allelic lineages in a population by n_a. The notation $\langle X \rangle$ is used in the results to refer to the expectation from analytical theory (see below) of parameter X.

Takahata (1990) showed that the genealogy of allelic lineages has a structure similar to a neutral gene genealogy, with appropriate rescaling of time. For a panmictic population of size N, the expected coalescence time of all extant allelic lineages is given by:

$$Tc_F = 4Nf_s(1-1/n_e), \tag{1}$$

where n_e is the effective number of allelic lineages and f_s is a scaling factor depending on N, u and selection intensity. Expressions for f_s are available for symmetrical overdominance (Takahata, 1990) and for GSI (Vekemans & Slatkin, 1994).

Vekemans & Slatkin (1994) showed that the gene genealogy within an allelic lineage is like the gene genealogy for a neutral locus, but on a much shorter time scale. In a panmictic population, the expected pairwise coalescence time for different copies of the same functional allele is given by:

$$T_A = \frac{2N}{n_e\left(1+\frac{1}{2f_s}\right)} \tag{2}$$

(Takahata & Satta, 1998), which reduces to $2N/n_e$ for low u or low N.

Gene genealogies in subdivided populations. In a subdivided population, another partition of diversity is within and between demes. Here, analytical results for the gene genealogy are available only for the neutral case. Under the neutral finite island model of S demes each with N hermaphroditic individuals and a diploid migration rate m (Wright, 1931), the expectations for T_S and T_T, the pairwise coalescence times of neutral genes sampled, respectively, from within a deme, and from the total population are (following Slatkin, 1991):

$$T_S = 2NS, \tag{3}$$

$$T_T = 2NS + \frac{(S-1)^2}{2Sm}. \tag{4}$$

Assuming neutrality, expected nucleotide diversities within demes and in the total population are found by multiplying the corresponding coalescence times by twice the synonymous mutation rate per nucleotide. For comparison with empirical data, differentiation of demes can be quantified using Nei's (1977) statistic $G_{ST} = (H_T - H_S)/H_T$, where H_T and H_S are the expected gene diversity values in the total population and within demes, respectively. The expected value of G_{ST} at equilibrium in a neutral island model is

$$G_{ST} = \left(1 + 2N\left(\frac{S}{S-1}\right)\left(\frac{1}{(1-m)^2(1-u)^2} - 1\right)\right)^{-1}, \tag{5}$$

where u is the mutation rate for the locus of interest (Takahata & Nei, 1984).

(ii) *Assumptions of the models studied*

To study subdivided populations under the island model, we investigated by simulation (using methods outlined below) two main models of single-locus, multi-allelic balancing selection: symmetrical overdominance and self-incompatibility (gametophytic (GSI) or sporophytic (SSI)). In either kind of model, patterns of mating and zygote survival are determined by alleles S_1, S_2, S_{n_a} at the selected locus. In the GSI system, compatibility reactions are determined by matching either of the two alleles from the maternal individual with the allele of the paternal gamete; if either allele matches, the combination is incompatible. No other selection occurs. In the symmetrical overdominance model, a combination of maternal and paternal gametes is retained if their alleles do not match; otherwise the homozygous zygote is discarded with probability s, the selection coefficient against homozygotes. Three models of sporophytic self-incompatibility (SSI) described by Schierup *et al.* (1997) were also investigated. One model assumes complete co-dominance of all alleles (SSIcod in the notation of Schierup *et al.*, 1997). We also modelled a linear dominance hierarchy of alleles in both maternal and paternal functions (SSIdom of Schierup *et al.*, 1997), and an intermediate model with dominance in pollen and co-dominance in pistils (SSIdomcod). In models involving dominance, the extant alleles in a population were sorted into their dominance hierarchy ($S_1 < S_2 < \ldots < S_{n_a}$) for determination of the phenotypes of diploids. New allelic lineages were assumed to

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

arise according to the infinite alleles model with a mutation rate u per locus per generation. In cases with dominance, each new allelic lineage was assigned a random place in the dominance hierarchy.

To model migration, maternal gametes were assumed to remain in their demes of origin. A paternal gamete was assumed to have a probability of $(1-m_p)$ of coming from the same deme and a probability of m_p of coming from the population as a whole. Migration is thus haploid; this is approximately equal to a diploid migration rate $m = m_p/2$, which we shall therefore use throughout.

(iii) *Simulations and analyses*

Forward single-locus simulations were performed to examine the evolutionary dynamics of selected alleles in a population of S subpopulations or demes with N diploid individuals (total population size $N_t = SN$). For symmetrical overdominance and GSI, the methods were as described in Schierup (1998). Implementation of the SSI models follows Schierup *et al.* (1997). Each simulation run was started with $2N_t$ functionally different alleles in the population, and allowed to evolve for 60000 generations, to allow approximate mutation–selection–drift equilibrium to be reached. Beginning at generation 60000, one of three different recording schemes (see below) was started. Different runs were made for the different schemes because of large differences in run times depending on which process was being studied. Replicates were obtained by repeating the whole process many times.

(*a*) *Investigation of population genetic structure.* Simulations were run for 90000 further generations. Statistics were recorded at 100 time points (i.e. every 900 generations) to reduce the sampling variance within a single run. The average over these 100 time points constitutes one replicate. At the total population level, we computed the number of functionally different alleles (n_a) at the selected locus and, using the allele frequencies, the expected gene diversity (H_T). The same quantities were computed within each deme and averaged over the S demes (H_S). G_{ST} was then calculated as $G_{ST} = (H_T - H_S)/H_T$, using the averages of H_T and H_S over replicates. To check the program, simulations of a neutral locus were also performed and checked against (5) for neutral diversity.

(*b*) *Investigation of genealogies of allelic lineages.* In these runs, only genealogical information on different allelic lineages was recorded. The method of Takahata & Nei (1990) was used. This method assigns to each extant allele a vector that records the genealogical relationships among alleles (according to Maruyama & Nei, 1981) and a vector that records the times when the mutation occurred to create each new allelic type. After a number of generations equal to 5 times the expected coalescence time of all alleles in the GSI model (calculated from equation 10 of Vekemans & Slatkin, 1994), a steady-state genealogy of allelic lineages was assumed to have been reached. We computed the coalescence time of all allelic lineages (Tc_F) and the average pairwise coalescence time between lineages (T_F), both at the total population level and also within demes, averaged over the S demes. These coalescent times thus record the time since the mutation changing allelic specificity occurred. This is shorter than the time when the alleles had a common ancestor; however, the difference between these time points is expected to be very small for strong selection (Vekemans & Slatkin, 1994). Four statistics described by Uyenoyama (1997) that characterize the topological structure of genealogies of allelic lineages were also computed. These are ratios of coalescence times, scaled by functions of allele numbers, whose expectations equal one for a random sample of neutral genes. Since the topological structure of the genealogies of allelic lineages under symmetrical balancing selection and GSI is very similar to the neutral case (Takahata, 1990), these ratios can also be used to detect deviations from the expected topology of alleles in systems under balancing selection (Schierup *et al.*, 1998; Uyenoyama, 1997).

(*c*) *Investigation of gene genealogies.* A separate, much slower simulation program was used to track the genealogical process of all the alleles in the population according to the method of Vekemans & Slatkin (1994). A complete genealogical tree of all $2N_t$ alleles was built with information on the functional allelic lineage that each allele belongs to, the time of each coalescent event, and which alleles coalesced at each event. From this tree, pairwise coalescence times in the total population were computed for alleles sampled at random irrespective of their lineage (T_T), for alleles sampled from the same allelic lineage (T_A). Due to the slowness of this program, only a very restricted parameter space could be investigated.

3. Results

(i) *Numbers of alleles maintained and differentiation between sub-populations*

In a panmictic population, the number of functional alleles maintained under balancing selection depends strongly on the effective population size N and less strongly on the mutation rate to new specificities (Vekemans & Slatkin, 1994; Vekemans *et al.*, 1998). Schierup (1998) found that population subdivision can decrease the number of functional alleles maintained, and therefore the effective population size

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

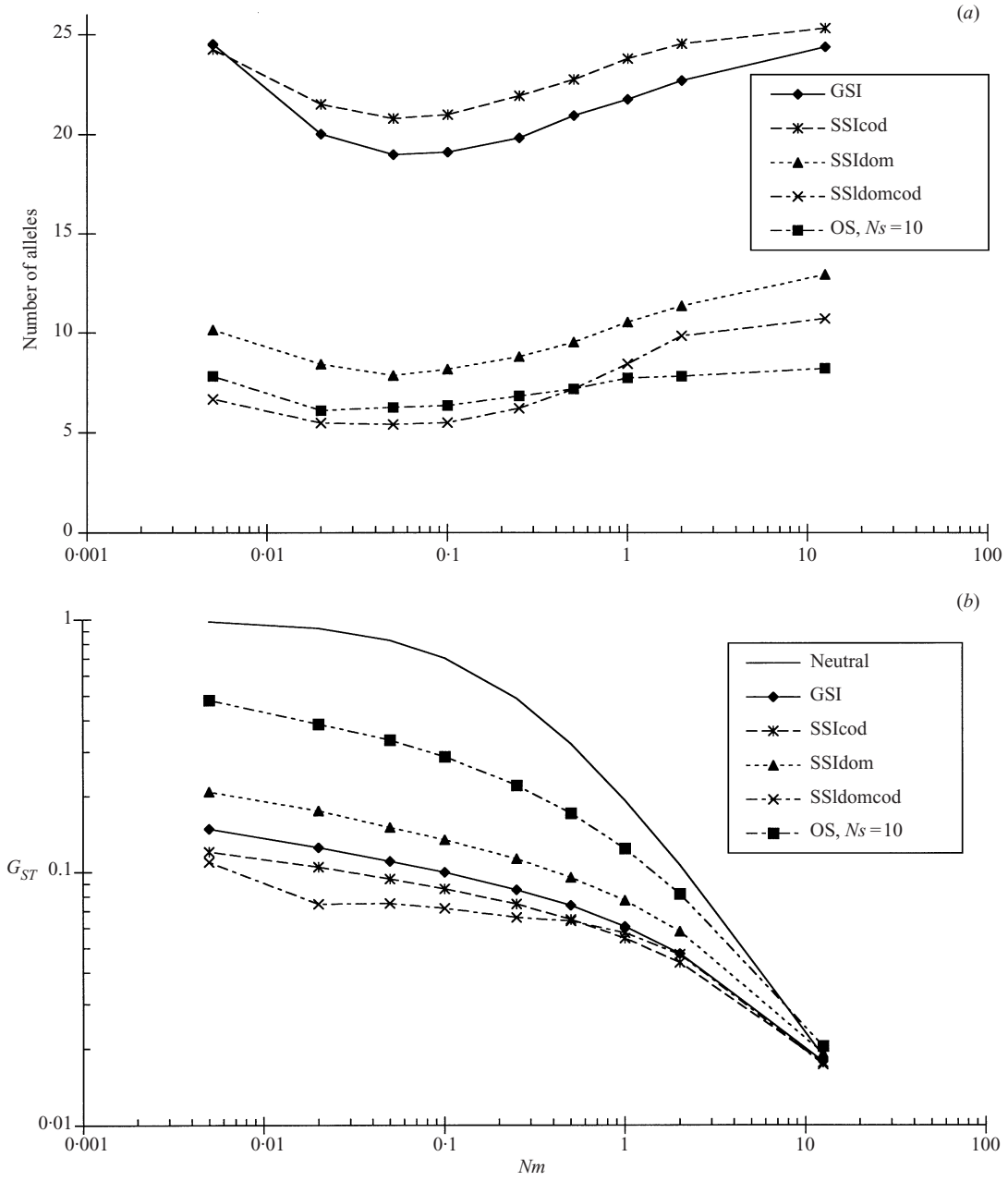

Fig. 1. The total number of functionally different alleles maintained (*a*) and G_{ST} (*b*) as a function of the scaled migration rate Nm for different models of balancing selection, including overdominant selection (OS). $N_t = 2000$, $u = 10^{-6}$, and there are 40 demes, each with 50 individuals.

defined for neutral alleles in a subdivided population (Nei & Takahata, 1993) does not predict the number of selected alleles maintained.

Fig. 1 shows the effect of the migration rate on the number of functionally different alleles (n_a) in the overall population (Fig. 1 *a*) and on G_{ST} (Fig. 1 *b*). The simulations for which these results are shown assumed $S = 40$ demes, each with $N = 50$ individuals, and $u = 10^{-6}$. For all models there is an intermediate level of subdivision (at a migration rate m_{min}) that maintains the smallest number of different alleles. The models differ in the relative magnitude of the decrease in n_a compared with a panmictic population with the same N_t, and in the value of m_{min}. The relative effect is larger when selection is weaker (the decrease in n_a for overdominance is 26 %; for SSIcod it is 18 %), or with dominance (SSIdom: 40 %; and SSIdomcod: 50 %). The minimum should disappear with still weaker selection since, under neutrality, n_a increases monotonically with decreasing migration (Nagylaki, 1985).

With balancing selection, differentiation between populations as measured by G_{ST} is significantly reduced compared with the case of a neutral locus (Fig. 1 *b*). This has two causes. First, balancing

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

Table 1. *The 'effective' migration rates for different models and* $N_t = 2000$, $u = 10^{-6}$, $S = 40$

	Effective migration rate				
		SSI			
Pollen: Ovules:	GSI	cod Co-dominant Co-dominant	dom Dominant Dominant	domcod Dominant Co-dominant	Overdominance
0·25	0·2627	0·2653	0·2436	0·2694	0·2284
0·005	0·0513	0·0590	0·0375	0·0674	0·0168
0·001	0·0383	0·0459	0·0269	0·0573	0·0095
0·0001	0·0273	0·0348	0·0181	0·0387	0·0051

First column shows the actual migration rate m used in the simulations.

selection increases within-deme diversity H_S, relative to total diversity, because alleles are kept in more equal frequencies than for a neutral locus, thus decreasing the numerator in the expression for G_{ST}. Secondly, and more importantly, an incoming migrant allele that is not already present in a deme will be selected for, increasing its chance of invasion compared with that of a neutral allele, i.e. the *effective* migration rate m_e is higher. Including both causes of the decrease in G_{ST}, we can use (5) to calculate m_e for the selected locus as the migration rate that, for a neutral locus under the finite island model, would yield the same value of G_{ST}, The effective migration rates thus calculated from the simulation results are shown for four actual migration rates (m) in Table 1.

At the lowest migration rate ($Nm = 0{\cdot}005$), m_e values are between 50 (overdominance, $Ns = 10$) and almost 400 (SSIdomcod) times higher than the actual rates of movement of gametes. The reduction in differentiation increases with the strength of selection for the symmetrical models (overdominance, GSI and SSIcod), and for these models strong differentiation is very unlikely. It appears that for SI in general an effective migration rate less than 2% is not expected under any realistic amount of dispersal (i.e. $Nm > 0{\cdot}001$). In the sporophytic models, even though the introduction of dominance decreases the strength of selection, the SSIdomcod model leads to slightly less differentiation than the model without dominance (SSIcod). The SSIdomcod model thus leads to the smallest level of differentiation even though it does not maintain the largest number of alleles (Fig. 1*a*). This will be discussed later.

(ii) *Coalescence times of allelic lineages*

The expected coalescence time or depth of the genealogical tree Tc_F and pairwise coalescence time $\langle T_F \rangle$ of extant allelic lineages determine the amount of neutral sequence diversity expected at loci with balancing selection, assuming that there is no recombination between lineages (for consideration of recombination see Schierup *et al.*, 2000). Table 2 shows results for three migration rates, for a small population of $N = 40$, $S = 5$ and $u = 5 \times 10^{-6}$, for which it is computationally possible to make enough simulation runs to observe values of Tc_F and T_F in 200 independent replicate simulations. The two quantities are affected similarly by differences in m, so only the results for Tc_F are shown. This behaves similarly to the number of functional alleles, and is considerably (25–50%) smaller at the intermediate migration rate ($Nm = 0{\cdot}06$) than in panmictic populations, in sharp contrast to the expectation for a neutral locus, where coalescence times increase monotonically with decreasing migration (Slatkin, 1991; Nei & Takahata, 1993). With the parameter values of our simulations, the relative decrease in Tc_F is larger than the decrease in the total number of alleles. In a panmictic population under symmetrical balancing selection, $\langle Tc_F \rangle$ is proportional to the square of the number of alleles (Takahata, 1990; Schierup *et al.*, 1998). This explains why Tc_F decreases relatively more than the total number of alleles at intermediate levels of migration. Comparing values of Tc_F for the deme and total population level, it can be seen that Tc_F depends on the total number of alleles rather than the number of alleles in the deme. This is because, even for $Nm = 0{\cdot}01$, migration is common on the time scale of coalescence of functional alleles, which amounts to 100–800 N generations (Table 2). Thus, each allelic lineage is expected to have migrated many times before coalescence occurs. It was possible to investigate coalescence times only for very small populations, but it is likely that the reduction at intermediate migration rates would be larger in larger populations, where the effect on allele numbers is higher (see Fig. 1*a*).

The four ratios derived by Uyenoyama (1997) for describing the shape of the genealogy of allelic lineages

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

Table 2. *The number of functional alleles* n_a, *and the coalescence times* Tc_F, *for different levels of migration*

Model:	Undivided		$Nm = 0{\cdot}06$		$Nm = 0{\cdot}006$	
	n_a	Tc_F	n_a	Tc_F	n_a	Tc_F
Total population level						
GSI	$8{\cdot}49 \pm 0{\cdot}77$	699 ± 4363	$7{\cdot}29 \pm 0{\cdot}14$	450 ± 295	$8{\cdot}36 \pm 0{\cdot}24$	607 ± 431
SSIcod	$9{\cdot}28 \pm 0{\cdot}64$	762 ± 435	$8{\cdot}27 \pm 0{\cdot}13$	485 ± 284	$9{\cdot}63 \pm 0{\cdot}25$	657 ± 371
SSIdom	$4{\cdot}57 \pm 0{\cdot}56$	239 ± 148	$3{\cdot}78 \pm 0{\cdot}09$	152 ± 108	$4{\cdot}36 \pm 0{\cdot}19$	160 ± 86
SSIdomcod	$4{\cdot}70 \pm 0{\cdot}70$	145 ± 97	$3{\cdot}69 \pm 0{\cdot}06$	58 ± 35	$3{\cdot}99 \pm 0{\cdot}09$	62 ± 43
Overdominance ($s = 0{\cdot}5$)	$4{\cdot}42 \pm 0{\cdot}10$	342 ± 197	$3{\cdot}77 \pm 0{\cdot}12$	253 ± 199	$4{\cdot}19 \pm 0{\cdot}18$	260 ± 199
Deme level						
GSI	—	—	$5{\cdot}78 \pm 0{\cdot}05$	436 ± 295	$4{\cdot}92 \pm 0{\cdot}03$	551 ± 386
SSIcod	—	—	$6{\cdot}50 \pm 0{\cdot}04$	468 ± 276	$5{\cdot}63 \pm 0{\cdot}03$	628 ± 370
SSIdom	—	—	$3{\cdot}19 \pm 0{\cdot}03$	142 ± 95	$2{\cdot}72 \pm 0.03$	114 ± 67
SSIdomcod	—	—	$3{\cdot}34 \pm 0{\cdot}04$	57 ± 35	$3{\cdot}22 \pm 0{\cdot}02$	60 ± 42
Overdominance ($s = 0{\cdot}5$)	—	—	$3{\cdot}12 \pm 0{\cdot}04$	237 ± 195	$2{\cdot}48 \pm 0{\cdot}03$	202 ± 158

Results are shown for alleles sampled at the total population level, and for alleles sampled at the deme level, respectively. Mean values are shown $\pm$ SD.

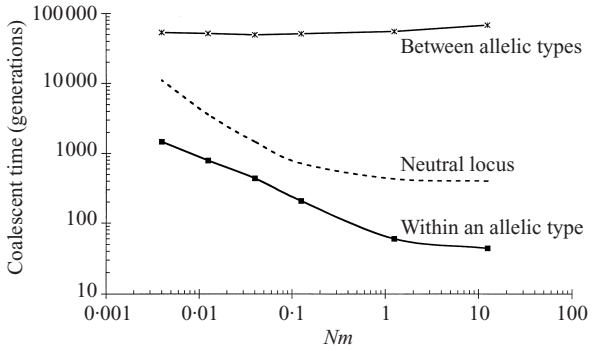

Fig. 2. The pairwise coalescence times for randomly chosen alleles T_T, alleles belonging to the same allelic lineage T_A, and the expected pairwise coalescence times (in generations) for random neutral genes (calculated from Slatkin, 1991). The GSI model is used, and there are 5 demes with 40 individuals. $u = 5 \times 10^{-6}$.

were all very close to the expected value of 1 under a neutral gene genealogy, deviating by 25% at most (results not shown). This deviation is small compared with the standard deviation of the ratios and in comparison with the deviations reported from analysis of data sets (Uyenoyama, 1997; Schierup *et al*., 1998).

(iii) *Pairwise coalescence times of copies of the same allelic lineage*

Fig. 2 shows mean coalescence times of pairs of alleles sampled irrespective of their functional status (T_T) and of pairs of alleles sampled from within sets of alleles from the same allelic lineage (T_A). The same parameter set as Table 2 was used in these simulations and the GSI model is shown as an illustration; results for other models are similar and are not shown. These samples are at random with respect to demes. Fig. 2 also shows $\langle T_T \rangle$ values for pairs of neutral genes sampled at random under the same migration rate (calculated from (3) above). For T_T, the results are dominated by the large coalescence times between alleles from different lineages and, as for Tc_F, minima occur at intermediate migration rates. Within a given allelic lineage, however, coalescence times T_A increase rapidly with decreasing migration in a pattern that resembles that expected for a neutral locus. T_A is expected to be smaller than $\langle T_T \rangle$ for a neutral locus, because the average number of copies of genes within an allelic class is $2N_t/n_a$ compared with $2N_t$ for a neutral locus, and n_a is between seven and nine for this parameter set (Table 2).

These results can be understood by considering separately two types of migration events: (1) immigration of an allele not present in the deme, and (2) immigration of an allele already present in the deme. The high effective migration rate (Table 1) is caused by immigration events of type (1) which merely homogenizes the population as a whole. However, for the coalescence times of genes within allelic lineages, only the second type of immigration contributes. Thus, alleles of any specific lineage experience population subdivision with extinction and recolonization. Extinction–recolonization dynamics may either increase or decrease differentiation, compared with a similar population without these dynamics (Wade & McCauley, 1988). Under the 'migrant pool' type of colonization where new colonists are drawn randomly from the total population, extinction–recolonisation dynamics enhances differentiation (as measured by F_{ST}) when $K < 2Nm + 0{\cdot}5$, where K is the number of colonists (see also Whitlock & McCauley, 1990). In our case of independent pollen grain migration, which fits the migrant pool model, $K = 0{\cdot}5$ for a single colonizing copy of an allele. Since $Nm > 0$, this implies that the extinction–recolonization dynamics

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

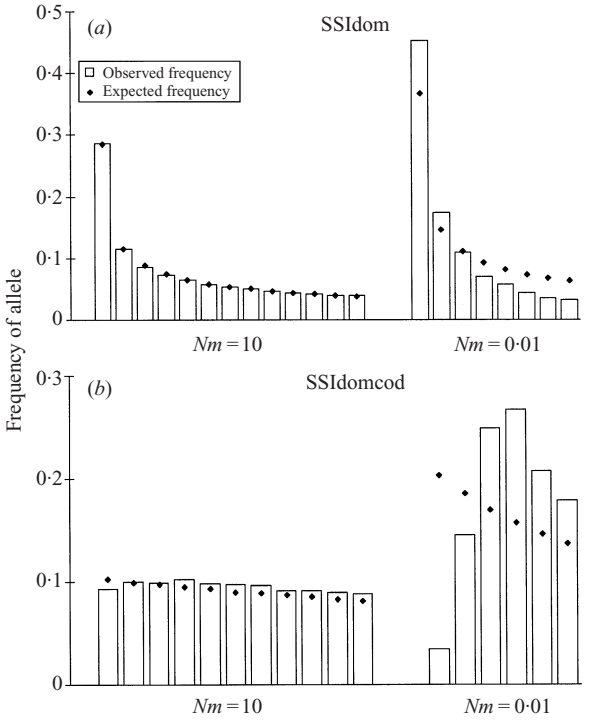

Fig. 3. Comparison of frequencies of alleles at the total population level in sporophytic systems with dominance. Observed values from simulations are the columns, and expected values (calculated from Schierup *et al*., 1997) are the diamonds. Alleles are ordered according to their dominance hierarchy with the most recessive allele first. The number of classes corresponds to the average number of alleles maintained for the given parameter values. Results are shown for weak ($Nm = 5$) and strong ($Nm = 0.005$) subdivision. (*a*) The SSIdom model. (*b*) The SSIdomcod model. There are 40 populations, each with 50 individuals, and $u = 10^{-6}$.

always increases differentiation, and explains why the proportional increase in T_A (in particular for $Nm > 0{\cdot}1$) is larger for $\langle T_T \rangle$ than for a neutral gene (Fig. 2).

(iv) *The effect of dominance in sporophytic self-incompatibility systems*

In models of sporophytic self-incompatibility with dominance, a dominant allele is more likely than a recessive one to invade a deme where it is not already present. On the other hand, in panmictic populations, recessive alleles reach higher frequencies and should therefore form a larger proportion of migrants (Sampson, 1974). Fig. 3 shows frequencies of alleles at each dominance level (most recessive allele first) in one set of simulations for the SSIdom and SSIdomcod models (Fig. 3*a*, *b*, respectively). The figure shows high and low migration rates, using the same parameters as in Fig. 1. The expected frequencies in a panmictic population for the same numbers of alleles are shown by the diamonds (calculated according to Schierup *et al*., 1997, appendix A). Fig. 3 shows that in the SSIdom model subdivision slightly increases the relative frequency of recessive alleles, whereas in the SSIdomcod model, recessive alleles become less common than dominant alleles, sometimes greatly so.

4. Discussion

(i) *Differentiation under balancing selection*

In a subdivided population, multi-allelic loci under strong balancing selection are expected to show very different distributions of sequence diversity from those of neutral loci. Substantial variation is maintained by balancing selection within demes, with increasing selection pressure in small populations (Vekemans *et al*., 1998), and migrant alleles are selected for if they are not already present in the deme. Both these effects decrease G_{ST}values, which measure the ratio of between-deme diversity to total diversity. This is consistent with empirical observations of very limited differentiation among populations for incompatibility loci in plants and fungi (Lawrence *et al*., 1993; Richman *et al*., 1995; Zambino *et al*., 1997). However, in none of these studies was G_{ST} estimated from neutral loci as well, emphasizing a need for studies where neutral loci are used as reference loci to loci under balancing selection.

The magnitude of the reduction in G_{ST} depends on the selection intensity. Contrasting the pattern of differentiation between loci of interest and neutral loci may therefore be used to test for the operation of balancing selection, and may provide information on the selection intensity at the putatively selected locus. Fig. 4 shows the reduction in G_{ST} for a locus under symmetrical balancing selection, as a function of the selection intensity parameter Ns, for different levels of subdivision. When G_{ST} for a neutral locus is below 0·1, only strong selection, of the order of $Ns = 100$, is likely to be detectable, but when the neutral G_{ST} is greater than 0·1 even $Ns = 10$ should be detectable.

There are at present few data sets to which this approach can be applied. Boyce *et al*. (1997) compared G_{ST} between five MHC loci and three microsatellites unlinked to the MHC, in bighorn sheep. They found similar values of G_{ST} (0·20) for both types of loci. In Chinook salmon populations, G_{ST} values were also similar (about 0·1) for both a polymorphic MHC locus and for unlinked microsatellites (Miller & Withler, 1997; Small *et al*., 1998). From Fig. 4, it thus seems likely that $Ns < 10$ in both these species. When neutral reference loci are absent, comparison among different MHC loci may also potentially yield insight. Comas *et al*. (1998) assayed four MHC loci in Basque populations and found F_{ST} values of 0·020 for HLA-C and 0·010-0·013 for HLA-A, HLA-B and DRB1. Though this difference among loci is not significant it agrees with HLA-C being the locus with the smallest

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

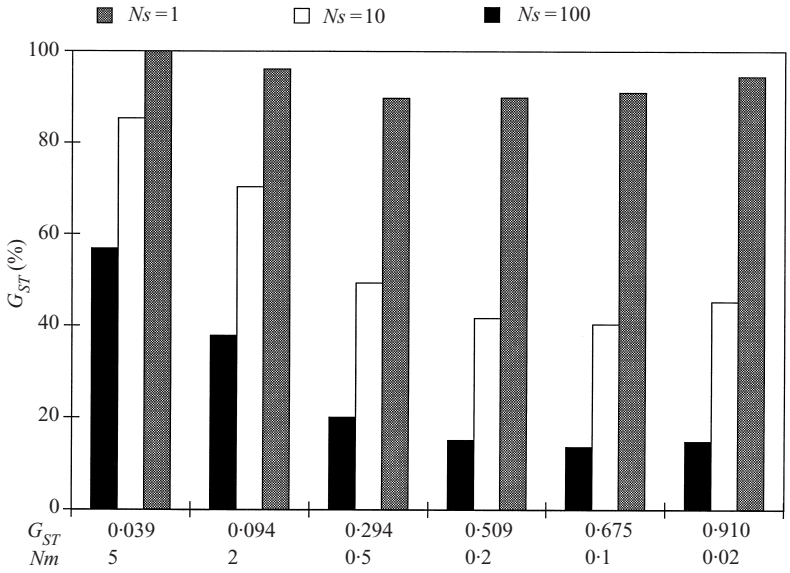

Fig. 4. Values of G_{ST} expressed as a percentage of the value expected for a neutral locus according to (5). Results are for a locus under symmetrical overdominant selection, with $Ns = 1$, $Ns = 10$ and $Ns = 100$, as a function of Nm. Shown at the x-axis is also G_{ST} for a neutral locus calculated from Nm using (5).

estimated selection coefficient ($Ns = 674$ for HLA-C and $Ns = 2371$–6301 for the other loci: see Takahata *et al.*, 1992), and also the smallest number of alleles. However, the general level of differentiation is so low that even these relatively large differences in selection coefficients may be hard to detect (judging by Fig. 4).

(ii) *Long-term effective population size*

The method of estimation of long-term effective population size from data from loci under balancing selection outlined by Richman & Kohn (1996) relies on the height and shape of the genealogy of allelic lineages. We investigated through simulations how subdivision affects these quantities. Subdivision decreases the height of the genealogy, Tc_F, unless $Nm <$ 0·01. However, the shape is close to the expectation in a panmictic population when measured by four ratios of coalescence times proposed by Uyenoyama (1997). These results are qualitatively different from the expectation for a neutral locus (Slatkin, 1991), where the coalescence time of all alleles increases monotonically with increasing subdivision, and where the shape of the tree departs from the panmictic case because the terminal branches become relatively shorter due to rapid coalescences within demes. For a species that is subdivided (or was subdivided in the past), the roughly neutral shape of the genealogy implies that the long-term effective population size relevant for the total diversity at a neutral locus will be substantially underestimated by Richman & Kohn's (1996) method. The same effect of subdivision was previously found for the short-term effective population size (Schierup, 1998), which determines the number of functional alleles maintained. In a subdivided population, estimates of the short-term population size are more sensitive to how individuals are sampled than are estimates of the long-term effective size, because the number of functional alleles in a sample from a deme depends much more on the migration rate than does Tc_F.

For MHC data in human populations, Ayala (1995) used essentially the same approach as Richman & Kohn (1996), assuming a selection coefficient, s, against homozygotes of between 0·01 and 0·03. He estimated that, of 113 alleles presently segregating at the DRB1 locus, 60 alleles diverged more than 6 mya, and suggested a long-term effective population size (N_e) exceeding 100000. He suggested, in close agreement with Takahata (1993), that the human N_e has recently decreased to the 10000 estimated from neutral loci that have much shorter coalescence times (Nei, 1987). However, due to the very different effect of subdivision on T_T values for these two types of loci, estimates of N_e cannot easily be compared. We would predict that, since the human population appears less subdivided today than in the past (Harris & Hey, 1999), it is likely that long-term N_e values from MHC loci underestimate the N_e of the population ancestral to modern humans.

(iii) *Variation within allelic lineages*

Pairwise coalescence times for alleles from the same lineage are much more sensitive to population structure than those between functionally different lineages. MHC alleles are often defined from sequence data rather than phenotypic tests. Similarly, in SI systems, S-allele types cannot be determined without large-scale crossing, so each sequence is sometimes

https://doi.org/10.1017/S0016672300004535 Published online by Cambridge University Press

taken to represent a different allelic lineage. The ability to define functionally different allelic lineages from sequences rests on the assumption that alleles with the same specificity have much shorter coalescence times than alleles of different specificities. This should be true in a panmictic population (Vekemans & Slatkin, 1994; Takahata & Satta, 1998). Extreme subdivision may, however, to some extent invalidate this approach since relatively larger coalescence times within allelic lineages can be expected (Fig. 2). Few studies have sequenced several copies of alleles with the same specificity. Walker *et al.* (1996) found two synonymous differences (in 400 bp) between copies of the same functional GSI allele from two widely separated *Papaver rhoeas* populations. However, Matsushita *et al.* (1996) found identical sequences of three independent S24 alleles from different populations of *Brassica campestris*, a species with an SSI system. In the fungus *Coprinus cinereus*, we computed based on a study by Badrane & May (1999) that 384–586 nucleotide differences (out of 1965 bp) occur in pairwise comparisons of 7 functionally different lineages at the b1–2 mating locus. Replicate alleles sampled world-wide were sequenced within two of these types and they found 0–4 nucleotide differences in pairwise comparisons between seven alleles of the first functonal type and 0–5 differences among four alleles from the other functional type. This provides clear evidence for balancing selection maintaining the different types and indicates that population subdivision is limited. May *et al.* (1999) used these data to estimate the scaling factor f_s (see above) by comparing diversity within and between allelic types. However, it is clear that such an approach is very sensitive to the extent of population subdivision.

(iv) *The effect of dominance in sporophytic self-incompatibility systems*

The interaction between population subdivision and dominance was briefly investigated in two models of sporophytic self-incompatibility analysed previously (Schierup *et al.*, 1997, 1998). The effect of restricted migration is qualitatively different depending on whether dominance is present in both pollen and pistils (SSIdom), in which case recessive alleles reach slightly higher frequencies, or only in pistils (SSIdomcod), in which case recessive alleles may be strongly disfavoured. The difference arises from the models' very different dynamics (Schierup *et al.*, 1997). In the SSIdomcod model, the substitution process of alleles is directional with respect to dominance, with dominant alleles invading most easily, and recessive alleles being easily lost. Restricted migration further favours dispersal of dominant alleles, which are likely to be present in more demes than recessive alleles (albeit at a lower frequency in each deme). This effect prevents differentiation more strongly than in the absence of dominance (see Fig. 1*b* and Table 1). In the SSIdom model the substitution process of alleles is non-directional with respect to dominance and therefore subdivision dynamics is not expected to change allele frequencies greatly, since each dominance class contributes almost the same number of successful migrants. Thus, in a subdivided population, unlike a panmictic one, it is not always true that recessive alleles are expected to be the commonest.

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 suggestions, and the Department of Computer Sciences, University of Aarhus for computing facilities.

References

Anderson, M. A., Cornish, E. C., Mau, S.-L., Williams, E. G., Hoggart, R., Atkinsson, A., Bönig, I., Greg, B. & Simpson, R. (1986). Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in *Nicotiana alata*. *Nature* **321,** 38–44.

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.

Boyce, W. M., Hedrick, P. W., MuggliCockett, N. E., Kalinowski, S., Penedo, M. C. T. & Ramey, R. R. (1997). Genetic variation of major histocompatibility complex and microsatellite loci: a comparison in bighorn sheep. *Genetics* **145,** 421–433.

Casselton, L. A. (1997). Molecular recognition in fungal mating. *Endeavour* **21,** 159–163.

Clark, A. G. (1993). Evolutionary inferences from molecular characterisation of self-incompatibility alleles. In *Mechanisms of Molecular Evolution* (ed. N. Takahata & A. G. Clark), pp. 79–108. Sunderland, MA: Sinauer Associates.

Clark, A. G. (1996). Population genetic aspects of gametophytic self-incompatibility. *Plant Species Biology* **11**, 13–21.

Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A. & Bertrandpetit, J. (1998). HLA evidence for the lack of genetic heterogeneity in Basques. *Annals of Human Genetics* **62**, 123–132.

Cope, F. W. (1962). The effects of incompatibility and compatibility on genotype proportions in populations of *Theobroma cacao* L. *Heredity* **17**, 183–195.

Crozier, R. H. (1977). Evolutionary genetics of the Hymenoptera. *Annual Review of Entomology* **22,** 263–288.

Crozier, R. H. & Pamilo, P. (1996). *Evolution of Social Insect Colonies: Sex Allocation and Kin Selection*. Oxford: Oxford University Press.

Duchateau, M. J., Hoshiba, H. & Velthuis, H. H. W. (1994). Diploid males in the bumble-bee *Bombus terrestris:* sex determination, sex alleles and viability. *Entomologia Experimentalis et Applicata* **71**, 263–269.

Edwards, S. V. & Hedrick, P. W. (1998). Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends in Ecology and Evolution* **13**, 305–311.
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.
Hedrick, P. W. (1994). Evolutionary genetics of the major histocompatibility complex. *American Naturalist* **143**, 945–964.
Hughes, A., Ota, T. & Nei., M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Molecular Biology and Evolution* **76**, 515–524.
Hughes, A. L. & Yeager, M. (1998*a*). Natural selection and the evolutionary history of major histocompatibility complex loci. *Frontiers in Bioscience* **3**, 510–516.
Hughes, A. L. & Yeager, M. (1998*b*). Natural selection at major histocompatibility complex loci of vertebrates. *Annual Reviews of Genetics* **32**, 415–435.
Klein, J. (1979). The major histocompatibility system of the mouse. *Science* **203**, 516–521.
Klein, J. (1986). *Natural History of the Major Histocompatibility Complex*. New York: Wiley.
Lawrence, M. J., Lane, M. D., O'Connell, S. & Franklin-Tong, V. E. (1993). The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. V. Cross-classification of the S-alleles of samples from three natural populations. *Heredity* **71**, 581–590.
Maruyama, T. & Nei, M. (1981). Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics* **98**, 441–459.
Matsushita, M., Watanabe, M., Yamakawa, S., Takayama, S., Isogai, A. & Hinata, K. (1996). The *SLG*s corresponding to the same *S*24-haplotype are perfectly conserved in three different self-incompatible *Brassica campestris* L. *Genes & Genetic Systems* **71**, 255–258.
May, G. & Matzke, E. (1995). Recombination and variation at the a mating-type of *Coprinus cinereus*. *Molecular Biology and Evolution* **12**, 794–802.
May, G., Shaw, F., Badrane, H. & Vekemans, X. (1999). The signature of balancing selection: fungal mating compatibility gene evolution. *Proceedings of the National Academy of Sciences of the USA* **96**, 9172–9177.
Miller, K. M. & Withler, R. E. (1997). MHC-diversity in Pacific salmon: population structure and trans-specific allelism. *Hereditas* **127**, 83–95.
Nagylaki, T. (1985). Homozygosity, effective number of alleles, and interdeme differentiation in subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **82**, 8611–8613.
Nasrallah, J. B., Kao, T.-H., Chen, T.-H., Goldberg, M. & Nasrallah, M. E. (1987). Amino-acid sequence of glycoproteins encoded by three alleles of the *S*-locus of *Brassica oleraceae*. *Nature* **326**, 617–619.
Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics* **41**, 225–233.
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.
Raper, J. R. (1966). *The Genetics of Sexuality in Higher Fungi*. New York: The Ronald Press.
Richman, A. D. & Kohn, J. R. (1996). Learning from rejection: the evolutionary biology of single-locus incompatibility. *Trends in Ecology & Evolution* **11**, 497–502.
Richman, A. D., Kao, T.-H., Schaeffer, S. W. & Uyenoyama, M. K. (1995). S-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). *Heredity* **75**, 405–415.
Richman, A. D., Uyenoyama, M. K. & Kohn, J. R. (1996). Allelic diversity and gene genealogy at the self-incompatibility locus in the Solonaceae. *Science* **273**, 1212–1216.
Sampson, D. R. (1974). Equilibrium frequencies of sporophytic self-incompatibility alleles. *Canadian Journal of Genetics and Cytology* **16**, 611–618.
Schierup, M. H. (1998). The number of self-incompatibility alleles in a finite, subdivided population. *Genetics* **149**, 1153–1162.
Schierup, M. H., Vekemans, X. & Christiansen, F. B. (1997). Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. *Genetics* **147**, 835–846.
Schierup, M. H., Vekemans, X. & Christiansen, F. B. (1998). Allelic genealogies in sporophytic self-incompatibility systems in plants. *Genetics* **150**, 1187–1198.
Schierup, M. H., Charlesworth, D. & Vekemans, X. (2000). The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. *Genetical Research* **76**, 63–73.
Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**, 278–280.
Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167–175.
Small, M. P., Beacham, T. D., Withler, R. E. & Nelson, R. J. (1998). Discriminating coho salmon (*Oncorhynchus kisutch*) populations within the Fraser River, British Columbia, using microsatellite DNA markers. *Molecular Ecology* **7**, 141–155.
Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences of the USA* **87**, 2419–2423.
Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution* **10**, 2–22.
Takahata, N. & Nei, M. (1984). Fst and Gst statistics in the finite Island model. *Genetics* **107**, 501–504.
Takahata, N. & Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978.
Takahata, N. & Satta, Y. (1998). Footprints of intragenic recombination at HLA loci. *Immunogenetics* **47**, 430–441.
Takahata, N., Satta, Y. & Klein, J. (1992). Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* **130**, 925–938.
Uyenoyama, M. K. (1997). Genealogical structure among alleles regulating self-incompatibility in Angiosperms. *Genetics* **147**, 1389–1400.
Vekemans, X. & Slatkin, M. (1994). Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165.
Vekemans, X., Schierup, M. H. & Christiansen, F. B. (1998). Mate availability and fecundity selection in multi-allelic self-incompatibility systems in plants. *Evolution* **52**, 19–29.
Vincek, V., O'hUigin, C., Satta, Y., Takahata, N., Boag, P. T., Grant, P. R., Grant, B. R. & Klein, J. (1997). How large was the founding population of Darwin's finches? *Proceedings of the Royal Society of London*, *Series B* **264**, 111–118.

Wade, M. J. & McCauley, D. E. (1988). Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution* **42**, 995–1005.

Walker, E. A., Ride, J. P., Kurup, S., FranklinTong, V. E., Lawrence, M. J. & Franklin, F. C. H. (1996). Molecular analysis of two functional homologues of the S-3 allele of the *Papaver rhoeas* self-incompatibility gene isolated from different populations. *Plant Molecular Biology* **30**, 983–994.

Whitlock, M. C. & McCauley, D. E. (1990). Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution* **44**, 1717–1724.

Wright, S. (1931). Evolution in mendelian populations. *Genetics* **16**, 97–158.

Wright, S. (1939). The distribution of self-sterility alleles in populations. *Genetics* **24**, 538–552.

Zambino, P., Groth, J. V., Lukens, L., Garton, J. R. & May, G. (1997). Variation at the b mating type locus of *Ustilago maydis*. *Phytopathology* **87**, 1233–1239.