Genet. Res., Camb. (2001), **77**, *pp.* 67–81. *With* 6 *figures.* Printed in the United Kingdom © 2001 Cambridge University Press

Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population

RENAUD VITALIS[1,2]* AND DENIS COUVET[3]

[1]*Laboratoire de Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France*
[2]*Station Biologique dé la Tour du Valat, Arles, France*
[3]*CRBPO – Museum National d'Histoire Naturelle, Paris, France*

(*Received 8 May 2000 and in revised form 28 August 2000*)

Summary

Measures of association of genes at different loci (linkage disequilibrium) are widely used to determine whether the structure of natural populations is clonal or not, to map genes from population data, or to test for the homogeneity of response of molecular markers to background selection, for example. However, the usual definitions of parameters for gametic associations may not be suitable for all these purposes. In this paper, we derive the recursion equations for one- and two-locus identity probabilities in an infinite island model. We study the role of drift, gene flow, partial selfing and mutation model on the expected association of genes across loci. We define the 'within-subpopulation identity disequilibrium' as the difference between the joint two-locus probability of identity in state and the expected product of one-locus identity probabilities. We evaluate this parameter as a function of recombination rate, effective size, gene flow and selfing rate. Within-subpopulation identity disequilibrium attains maximum values for intermediate immigration rates, whatever the selfing rate. Moreover, identity disequilibrium may be very small, even for high selfing rates. We discuss the implications of these findings for the analysis of data from natural populations.

1. Introduction

It is commonly admitted that the extent to which genes are associated across loci may be useful for the analysis of population structure or for the detection of selection pressures. Therefore, it is important to determine the amount of gametic association for neutral genes expected in a population. There is, however, no real consensus on the relevance of different measures of gametic associations.

(i) *The mean linkage disequilibrium*

The pairwise gametic disequilibrium parameter between two loci is defined as the excess of coupling gametes over that expected under random association. Let P_v^u be the frequency of gametes carrying allele u at the first locus and allele v at the second locus. Let P^u and P_v be the frequencies of alleles u and v at the first and the second locus, respectively. Then, a general expression for linkage disequilibrium is given by $D_v^u = P_v^u - P^u \cdot P_v$. This parameter may also be viewed as a covariance since it is expressed as the difference between the joint frequency and the product of single frequencies of genes at two loci.

Extending Geiringer's (1944) theory, Bennett (1954) demonstrated that, in an infinite population, the linkage disequilibrium among any set of neutral loci always tends to zero. However, the rate of approach to equilibrium may be very slow if genes are tightly linked. Hill & Robertson (1966) provided the equations for the change in linkage disequilibrium between a pair of loci in a finite random mating population. They showed that, if none of the genes has an effect on fitness, the expected linkage disequilibrium is zero. This is true whatever the number of loci (Hill, 1974*a*).

* Corresponding author. Tel: +33 4 67 14 32 50. Fax: +33 4 67 14 36 22. e-mail: vitalis@isem.univ-montp2.fr

In an infinite partially selfing population, there is no statistical association between allele frequencies at linked loci at mutation–drift equilibrium (Bennett & Binet, 1956). While the value of linkage disequilibrium is therefore expected to be zero, it has been shown that selfing slows its decay in a similar way to linkage (Weir & Cockerham, 1973). However, Hastings (1984) showed a complex interaction of linkage and partial selfing for maintaining high linkage disequilibrium in the presence of selection.

In subdivided populations without selection the mean linkage disequilibrium eventually vanishes (Nei & Li, 1973: Ohta, 1982*a*, *b*). In such situations, with restricted gene flow among groups, linkage disequilibrium can only be transient (Nei & Li, 1973; Slatkin, 1975). Thus, in the absence of selection whatever the number of loci considered, in isolated or subdivided populations, whatever the mating system and the population size, the mean linkage disequilibrium is zero at mutation–drift equilibrium.

(ii) *The variance of linkage disequilibrium*

While the mean linkage disequilibrium asymptotically converges to zero, it has been shown that there may be a large variance among segregating lines, due to finite population size (Hill & Robertson, 1968; Sved, 1968; Ohta & Kimura, 1969; Hill, 1974*b*; Weir & Hill, 1980). Since its expectation is zero, the variance of linkage disequilibrium equals the mean squared disequilibrium. Sved (1971) further studied the correlation of genotype frequencies at two linked loci by comparison of one approach based on disequilibrium parameters and a second approach based on identity-by-descent probabilities. From his results, he suggested that the measures of linkage disequilibrium in finite populations might be used to give information about effective population size (Hill, 1981).

Ohta (1982*a*, *b*) investigated the consequence of drift and gene flow on the variance of linkage disequilibrium in a finite subdivided population. She defined various statistics for linkage disequilibrium, to account for within- and between-subpopulation effects, by analogy with *F*-statistics. She also determined the variance of these statistics as a function of moments of gametes and genes frequencies in a symmetric two-allele model (Ohta, 1982*b*) and in an infinite-allele model (Ohta, 1982*a*). Finally, she proposed a test to discriminate between epistatic selection and limited gene flow as explanations for a large observed linkage disequilibrium by comparison of appropriate variance components (Ohta, 1982*a*). Tachida & Cockerham (1986) further developed the analysis and attempted to clarify the definition of linkage disequilibrium parameters. They determined the variance–covariance structure of some estimators of these parameters as functions of probabilities of identity by descent (Weir & Cockerham, 1969).

(iii) *Identity disequilibrium within individuals*

Bennett & Binet (1956) showed that, in an infinite partially selfing population, even though genes are associated at random across loci at equilibrium, there is a positive association between the genotypic states at different loci, even if they are not linked. There is an excess of double homozygotes for two independent loci compared with the product of single homozygotes at these two loci (Bennett & Binet, 1956). This has also been proved to exist in a finite population as a consequence of the variation of inbreeding among individuals (Weir & Cockerham, 1973). This quantity, referred to as 'identity disequilibrium', 'genotypic association' or even 'equilibrium constant', is a decreasing function of recombination rate in an infinite mixed selfing and random mating population (Weir & Cockerham, 1973). Furthermore, for any particular recombination rate, there is an amount of selfing which maximizes this parameter. The variation of inbreeding among individuals within a population (identity disequilibrium within individuals) has been described as a function divided in two parts, one being viewed as a correlation among united gametes and the other being the effect of linkage (Weir & Cockerham, 1969). More generally, in partially inbreeding (finite) populations there is an association between the homozygosity of different loci, even if they are not linked (Charlesworth, 1991).

It is worth noticing that this feature has been recognized as a possible explanation for the heterozygosity–fitness correlations through associative overdominance in finite, inbred populations (Strobeck, 1979; Charlesworth, 1991). More recently, David (1999) proposed an extension of this theory for studying the association between phenotypic variance and heterozygosity at marker loci. Correlations between heterozygosity and both mean and variance of phenotypes could indeed be parsimoniously explained by inbreeding.

Other measures of association have also been defined (Hill, 1975; Ohta, 1980; Takahata, 1982). Indeed, describing the two-locus population structure in terms of identity probabilities naturally leads to the definition of 'within-subpopulation identity disequilibrium', as the excess of two-locus identity probabilities over the product of single-locus probabilities. We evaluate this parameter in a structured, partially selfing population. We show that this measure has interesting properties that should be useful in the context of inferring demographic parameters. We discuss the implications of these results for the analysis of data from natural populations.

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

2. The model

We consider an infinite island model of population structure (Wright, 1931) with gametic migration. This model assumes that the population is subdivided into an infinitely large number of diploid monoecious subpopulations. Each generation, individuals reproduce in each subpopulation and some offspring are produced by selfing. All individuals contribute with equal probability to the next generation. Recombination occurs between loci at rate r. Thereafter, the subpopulations exchange migrant gametes at a rate m. The proportion of pairs of genes that come from a single subpopulation in the previous generation is thus $(1-m)^2$. Since we consider gametic migration, we shall define the selfing rate s as the conditional probability that two homologous sets of genes of one individual are derived from the same parent, given they are both copies of genes from one subpopulation before migration. The overall amount of selfing is thus $(1-m)^2 s$. Locally, in a completely random mating subpopulation of N hermaphroditic individuals, $s = 1/N$. Genes are sampled after dispersal. We consider two mutation models. The infinite-allele model (IAM) provides expected values of the probabilities of identity by descent (IBD). Alternatively, the symmetric K-allele model (KAM) gives the expectations for the probabilities of identity in state (IIS).

(i) *Definition of one-locus identity probabilities*

Let us define Q_0 as the probability that two homologous genes taken at one locus in a single individual are identical in state. Q_1 and Q_2 are the IIS probabilities of two genes at one locus, taken in distinct individuals, respectively in the same (Q_1) or different (Q_2) subpopulations. With local panmixia, $Q_0 = Q_1$. As we consider an infinite island model, the probability of coalescence for genes taken in distinct subpopulations is zero at equilibrium. Therefore, the equilibrium IIS probability for two genes taken in distinct subpopulations depends on the mutation model. In the IAM the IBD probability Q_2 is zero at equilibrium. In the KAM, the equilibrium IIS probability Q_2 depends on the number of allelic states: two genes can be IIS (and not IBD) with probability $= 1/K$, where K is the number of allelic states (Cockerham, 1984).

(ii) *Definition of two-locus identity probabilities with local random mating*

With local random mating, one needs to define three two-locus IIS probabilities (Whitlock *et al.*, 1993). In the following, a haplotype is a set of two genes sampled in one individual at two distinct loci which comes from a single gamete in the previous generation. The two loci need not be physically linked on a single chromosome. All two-locus probabilities are defined as IIS probabilities. However, the IAM provides expectations for IBD probabilities.

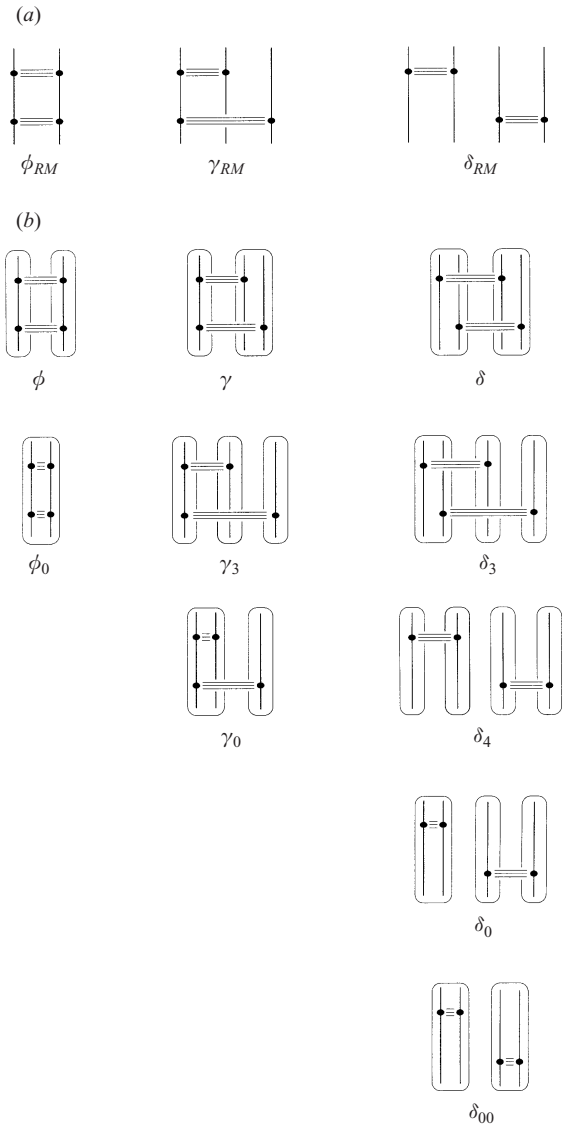

Fig. 1. Definition of two-locus probabilities for the probability of identity in state (IIS). (*a*) Random mating. Vertical lines represent sampled haplotypes, on which upper and lower positions of filled circles represent two loci. The symbol ≡ among pairs of homologous genes stands for identity in state. In the infinite allele model, these coefficients define the corresponding probabilities for the identity by descent (IBD). (*b*) Non-random mating. The IIS probabilities defined above may have different values according to the number of individuals from which the haplotypes are sampled. Other probabilities are thus defined, with each diploid individual represented as a box. Only the sampled haplotypes are shown. See the text for details.

ϕ_{RM} is the probability that two randomly sampled haplotypes are IIS at both loci. γ_{RM} is the probability that, among three sampled haplotypes, one pair is IIS at one locus and a distinct pair (with a single common haplotype) is IIS at the second locus. δ_{RM} is the

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

probability that, when four haplotypes are sampled, two are IIS at one locus and the other two are IIS at a second locus. These identity probability parameters are schematically defined in Fig. 1*a*.

As in the one-locus case, we can derive the recursion equations for two-locus IIS probabilities. Let us take the recursion for ϕ_{RM}, for example. With probability $(1-m)^2(1-1/N)$, two randomly sampled haplotypes come from different parents within the subpopulation in the previous generation. Now, these two sampled haplotypes may be two recombinant gametes, one recombinant and one parental gamete, or two parental gametes. In that case, the IIS probability involves respectively four, three or two haplotypes in the previous generation. Therefore, they are IIS with probability $[(1-r)^2\phi_{RM}+2r(1-r)\gamma_{RM}+r^2\delta_{RM}]$. With probability $(1-m)^2/N$, the two sampled haplotypes come from a single parent in the same subpopulation in the previous generation. With probability $[(1-r)^2+r^2]$ the haplotypes are both recombinant or non-recombinant gametes. The gametes may be copies of the same or of distinct haplotypes, and therefore are IIS with probability $(\phi_{RM}+1)/2$. With probability $[2r(1-r)]$ one haplotype is recombinant, and the other is non-recombinant. They are copies of distinct gametes and are IIS with probability Q_{RM} (with Q_{RM} defined as the one-locus IIS probability for a pair of genes taken at random within one subpopulation). Although rather more complicated, since they involve triplets and quadruplets of haplotypes, recursion equations for γ_{RM} and δ_{RM} can be derived similarly.

(iii) *Definition of two-locus identity probabilities with selfing*

With non-random mating, eg. partial selfing or dioecy, more IIS probabilities need to be defined, since IIS probabilities defined for pairs, triplets or quadruplets of haplotypes (respectively ϕ_{RM}, γ_{RM} and δ_{RM} in the random mating case) may now have different values whether these haplotypes are found in two, three or four individuals (Weir *et al*., 1980; Weir & Cockerham, 1969; Weir & Hill, 1980).

Now, ϕ is the probability that two haplotypes randomly sampled in two individuals are IIS at both loci. ϕ_0 is the corresponding probability when the two haplotypes are taken within a single individual.

γ is the probability that, when a first haplotype is chosen from one individual and two others are chosen from a distinct individual, the first haplotype is IIS to the second one at the first locus and IIS to the third haplotype at the second locus. γ_3 is the probability that, among three sampled haplotypes, each taken in a distinct individual, one pair is IIS at one locus, and a distinct pair (with a single common haplotype) is IIS at the second locus. γ_0 is the probability that, when two individuals are sampled, one is homozygous at the first locus and both carry IIS genes at the second locus.

δ is the probability that, when two pairs of haplotypes are taken among two individuals, one pair is IIS at the first locus and the other pair is IIS at the second locus. δ_3 is the probability that, when four haplotypes are taken among three individuals, one pair of haplotypes taken among a first pair of individuals is IIS at the first locus and another pair of haplotypes, taken among a second pair of individuals, is IIS at the second locus. δ_4 is the probability that, when four haplotypes are sampled among four distinct individuals, two are IIS at one locus, and the other two are IIS at a second locus. δ_0 is the probability that, when three distinct individuals are sampled, one individual is homozygous at the first locus, and two sampled haplotypes among the other two individuals are IIS at the second locus. δ_{00} is the probability that, for two randomly sampled individuals, one is homozygous at the first locus and the other is homozygous at the second locus. All these identity probability parameters are schematically defined in Fig. 1*b*.

Overall, we have defined 10 two-locus identity probabilities. Along with one-locus identity probabilities Q_0 and Q_1 for each locus, all these 14 parameters are necessary and sufficient to describe the two-locus genetic structure of a single, partially selfing population. As in the one-locus case, the probability of coalescence for two-locus pairs of genes in distinct subpopulations is zero at equilibrium, in an infinite island model. Therefore the equilibrium IIS probabilities for two-locus pairs of genes taken in distinct subpopulations depend also on the mutation model. In the IAM, these probabilities are zero (IBD probabilities). In the KAM, IIS two-locus probabilities are simple products of one-locus probabilities, i.e. products of $1/K$'s over pairs of loci.

We derived the recursion equations for each of these 14 parameters, in the IAM and the KAM. Some details are given in the Appendix. With panmixia, $\phi = \phi_0$, $\gamma = \gamma_3 = \gamma_0$ and $\delta = \delta_3 = \delta_4 = \delta_0 = \delta_{00}$, and therefore the system reduces to the three two-locus IIS probabilities defined for the random mating case above.

It may not be possible to know from genotypic data, when pairs of genes are taken at two loci, whether the genes within each individual belong to the same haplotype ('coupling genes', or 'genes in phase') or not ('genes in repulsion'). Therefore, we define the composite identity probability Φ as the probability that a pair of genes taken at two distinct loci within an individual is IIS to a pair of genes taken at homologous loci in a second, distinct, individual:

$$\Phi = \frac{\phi+2\gamma+\delta}{4}. \quad (1)$$

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

(iv) *Definition of identity disequilibrium parameter*

Ohta (1980) defined a measure of association among amino acid sites as the excess probability of simultaneous identity over that expected from random combination of the identity at the two loci, a quantity equivalent to the covariance of non-identity (heterozygosity) at two loci within populations (Avery & Hill, 1979; Hedrick, 1987).

We define an analogous parameter, hereafter referred to as 'within-subpopulation identity disequilibrium', η_S, as the difference between the joint two-locus identity probability among two randomly chosen individuals in a population, and the expected product of one-locus identity probabilities:

$$\eta_{S,ij} = \Phi_{ij} - \delta_{4ij}. \quad (2)$$

There are some advantages to considering this parameter. First, its definition is a straightforward function of identity probabilities. Second, as further developed in a companion paper (Vitalis & Couvet, in press), statistics can easily be derived to estimate this quantity.

3. Results

(i) *The within-subpopulation identity disequilibrium as a measure of association*

Although the within-subpopulation identity disequilibrium may be viewed as the covariance for the probability of identity across a pair of loci, it is better defined as a function of probabilities of identity. This parameter depends on some demographic parameters of the population model (effective size, migration rate), but also on the mutation model considered. Indeed, it is quite sensitive to the mutation rate, and especially to the number of allelic states that can arise at each locus. A 'standardized' parameter can be defined as

$$\eta'_{S,ij} = \frac{\Phi_{ij} - \delta_{4ij}}{(1-Q_{2i})(1-Q_{2j})}. \quad (3)$$

Similar standardized parameters have already been defined (Ohta, 1980; Takahata, 1982; Hedrick, 1987).

Fig. 2 shows η_S and η'_S for various models of mutation (IAM and KAM). For large Nr, both η_S and η'_S tend to zero, indicating that two-locus identity probabilities converge to the expected product of one-locus identities. But the within-subpopulation identity disequilibrium remains significantly greater than zero for increasing values of Nr, as the mutation model approaches the IAM. Indeed, the within-subpopulation identity disequilibrium attains its maximum value when the potential number of allelic states at both loci is infinite. More generally, the within-subpopulation identity disequilibrium increases as the number of allelic states becomes large, especially when linkage is tight. Conversely, the looser the linkage, the smaller the discrepancy between IAM and KAM. In other words, drift creates associations more easily when the number of allelic states is high.

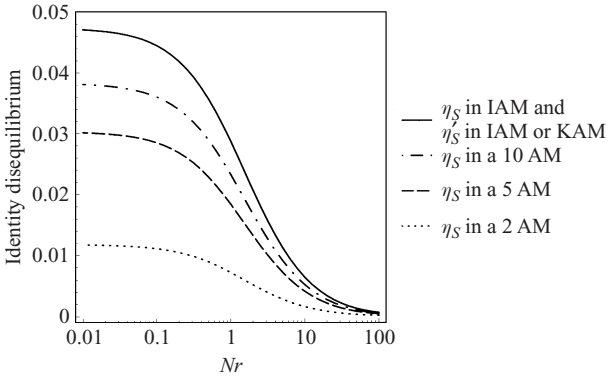

Fig. 2. Expected within-subpopulation identity disequilibrium as a function of Nr for various mutation models, with $\mu = 10^{-6}$. The other population parameters are $N = 200$ and $Nm = 1$. Random mating is assumed ($Ns = 1$). Note the logarithmic scale on the abscissa.

Fig. 2 further shows that η'_S does not depend on the mutation model, whatever the range of other population parameters in the model (proportion of selfing, recombination rate, effective population size, migration rate). Ohta (1980) and Takahata (1982) found an analogous result for a single, random mating population. Therefore the standardized within-subpopulation identity disequilibrium is closely related to linkage and nearly independent of the properties of the mutation model.

(ii) *The effect of selfing*

Golding & Strobrek (1980) derived the expression for the squared linkage disequilibrium in a single, finite, partially selfing population. They found that, for $N\mu < 1$, squared linkage disequilibrium may increase with selfing if $Nr > 1$, or decrease if $Nr < 1$. For $N\mu < 1$, squared linkage disequilibrium increases with selfing, whatever the range of Nr. Considering the effect of migration, we found in turn that the same holds as long as $Nm < N\mu < 1$. Only when $Nm > N\mu$ does the squared linkage disequilibrium increase with the selfing rate, whatever the range of Nr.

Fig. 3 depicts the expected value of within-subpopulation identity disequilibrium as a function of Nr, for a range of s values. It is shown that within-subpopulation identity disequilibrium increases with decreasing values of Nr. Moreover, increasing the proportion of selfing always increases the expected within-subpopulation identity disequilibrium. Contrary to a single population model (Golding &

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

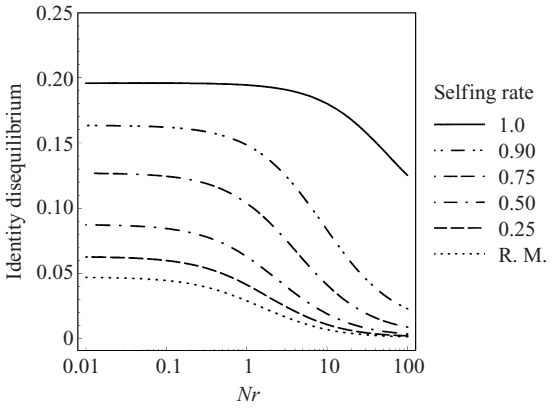

Fig. 3. Expected within-subpopulation identity disequilibrium as a function of Nr for various selfing rates and random mating (R.M.). The other population parameters are $N = 200$ and $Nm = 1$. The mutation model is an IAM with $\mu = 10^{-6}$. Note the logarithmic scale on the abscissa.

Strobeck, 1980), the measure of association among loci depends on the recombination rate when $s = 1$. One would expect that, for completely selfing organisms, the equilibrium value of a measure of association of genes among loci is not influenced by the proportion of recombination. But as noted before, the overall amount of selfing is $(1-m)^2 s$. Consequently, when $m \neq 0$, the effective rate of selfing is less than one. This is the reason why, even for $s = 1$, the within-subpopulation identity disequilibrium slightly decreases with increasingly larger values of Nr.

(iii) *The effect of migration*

Within-subpopulation identity disequilibrium is a unimodal function of the migration rate (Fig. 4). Within-subpopulation identity disequilibrium attains a maximum value for intermediate migration rates. For small migration rates, since migration introduces new gametes in fairly monomorphic subpopulations, it inflates the disequilibrium among two-locus IIS probabilities. However, as the migration rate increases and converges towards unity, the whole population approaches panmixia and the within-subpopulation identity disequilibrium tends to a small value for a large (infinite) population. Similar results were obtained for the variance of the gametic component of linkage disequilibrium in an island model (Tachida & Cockerham, 1986). While the mode of the curve depends on Nm rather than m, the maximum value of within-subpopulation identity disequilibrium decreases with N. Furthermore, the effect of migration is more pronounced for tightly linked loci.

Interestingly, this leads to the observation that, for a given value of Nm, the within-subpopulation identity disequilibrium could be used to infer the effective population size and migration rate. This point will be further developed in a companion paper (Vitalis & Couvet, in press).

(iv) *Joint effect of selfing and migration*

Fig. 5 shows the expected within-subpopulation identity disequilibrium as a function of Nr and Nm in random mating and complete selfing situations. Within-subpopulation identity disequilibrium increases with selfing, whatever the range of Nr or Nm. There are values of Nm for which within-subpopulation identity disequilibrium is always expected to be very low. One can deduce that when $Nm < 10^{-2}$ or $Nm > 10^2$, no within-subpopulation identity disequilibrium shall be found. With random mating, significant within-subpopulation identity disequilibrium is expected when Nr is smaller than one. With

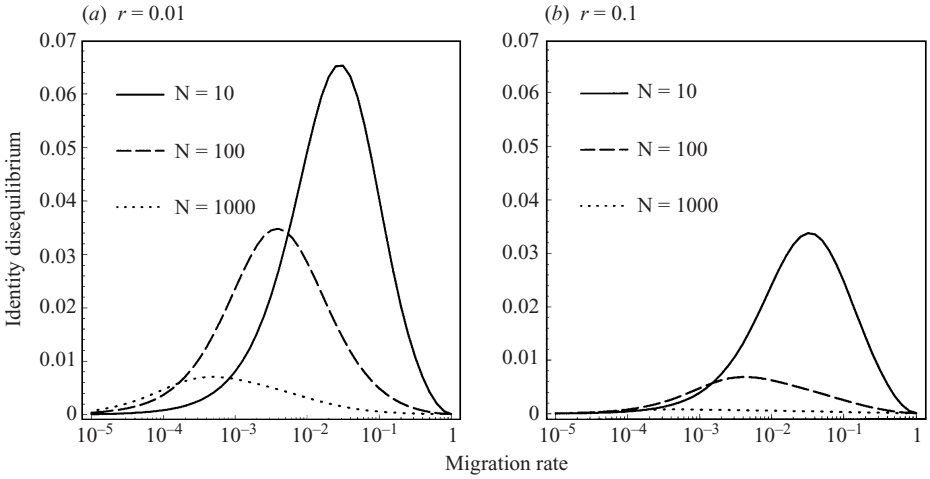

Fig. 4. Expected within-subpopulation identity disequilibrium as a function of migration rate, for various population sizes ($N = 10$, 100 and 1000) and recombination rates: (*a*) $r = 0{\cdot}01$ and (*b*) $r = 0{\cdot}1$. Random mating is assumed. The mutation model is an IAM, with $\mu = 10^{-6}$. Note the logarithmic scale on the abscissa.

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

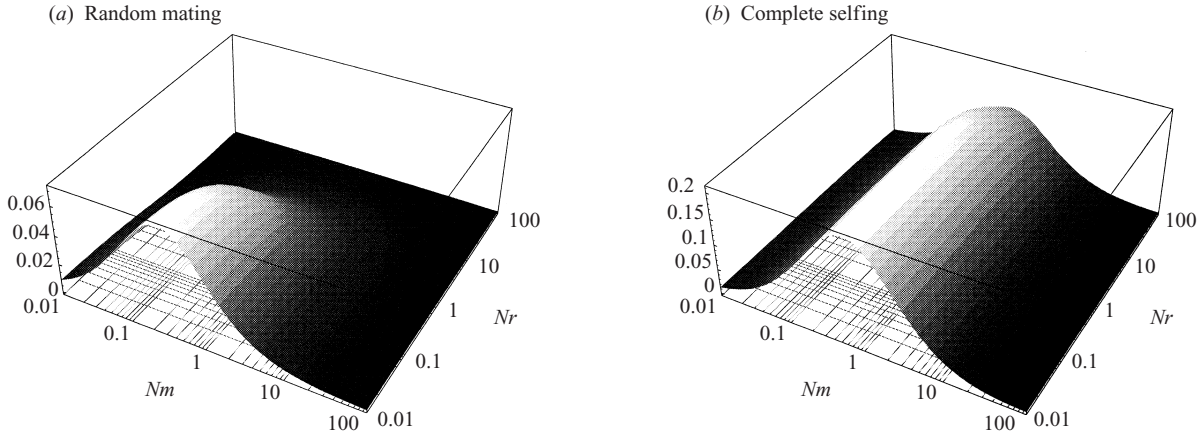

Fig. 5. Expected within-subpopulation identity disequilibrium as a function of Nm and Nr in two contrasted situations: (*a*) random mating ($Ns = 1$) and (*b*) complete selfing ($s = 1$). The population size is $N = 200$. The mutation model is an IAM, with $\mu = 10^{-6}$. Note the logarithmic scale on the x- and y-axes.

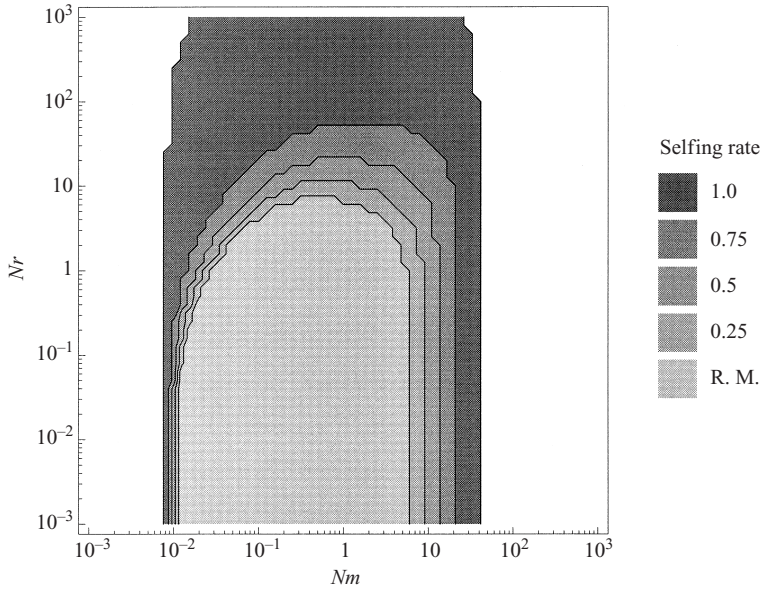

Fig. 6. Diagrammatic representation of the range of Nr and Nm necessary and sufficient to maintain significant ($> 0{\cdot}01$) within-subpopulation identity disequilibrium, for various selfing rates and random mating (R.M.). Each (continuous) region draws the parameter space in which the within-subpopulation identity disequilibrium is greater than 1%. Darker areas always include lighter ones. Note the logarithmic scale on the x- and y-axes.

complete selfing, within-subpopulation identity disequilibrium is asymptotically always maintained for $10^{-2} < Nm < 10^{2}$. Consequently, within the appropriate range of Nm, within-subpopulation identity disequilibrium is maintained when $Nr(1-s) < 1$, as for squared linkage disequilibrium in the single population case (Golding & Strobeck, 1980).

The effect of increased selfing is summarized in Fig. 6. Grey areas draw the range of Nr and Nm values for which $\eta'_S > 0{\cdot}01$. Darker areas always include lighter ones. Higher selfing rates maintain identity disequilibrium in a larger range of conditions. In the special case of complete selfing, with $10^{-2} < Nm < 10^{2}$, η'_S is asymptotically always greater than 1%.

4. Discussion

The two-locus genetic structure of a partially selfing metapopulation has been characterized by means of appropriate definitions of probabilities of identity of genes, for one and two loci (Fig. 1). A simple measure of association of genes across loci has been defined, as the difference between the joint probability of identity at two loci and the expected product of one-locus probabilities at these loci. This quantity has been referred to as the 'within-subpopulation identity disequilibrium' and is equivalent to the 'identity excess' of Ohta (1980) and to the covariance of non-identity defined by Avery & Hill (1979) (see also

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

Hedrick, 1987). This parameter, denoted η_S, could also be viewed as the covariance for the probability of identity of genes, across pairs of loci. Finally, we derived the expected value of this parameter in an infinite island model of population structure, with some emphasis on the consequence of mutation model, reproductive system, dispersal, population size and recombination.

(i) *The effect of the mutation model*

Within-subpopulation identity disequilibrium depends on the mutation model (Fig. 2). However, for low mutation rates, the 'standardized' definition of within-subpopulation identity disequilibrium, η'_S, has been shown to be independent of the mutation model (IAM vs KAM) and of the number of allelic states in the KAM (Fig. 2). A similar property has already been discussed for *F*-statistics by Crow & Aoki (1984) and Rousset (1996). This result is particularly important in the perspective of comparing estimates across independent pairs of loci. Although the question of estimation will be addressed elsewhere (Vitalis & Couvet, in press), this result also suggests that, since they have the same expectation, estimates could be pooled over pairs of loci.

Analogous results were obtained for the analysis of other measures of gametic associations. For example, since the squared linkage disequilibrium has been shown to depend on allele frequencies, some authors defined another measure of disequilibrium as the ratio of squared linkage disequilibrium over the product of variance in gene frequencies at the two loci. This quantity has been referred to as 'correlation' (Hill & Robertson, 1968) or 'standardized linkage disequilibrium' (Ohta, 1982*a*). Various definitions have been proposed, depending on whether ratios of expectations or expectations of ratios were taken (Ohta, 1989; Takahata, 1982; Hedrick, 1987). However, the difference between the parameters and the statistics that estimate those parameters has not always been clearly defined. This led to some controversy on the dependence of some measures of association on the underlying gene frequencies (see Hedrick, 1987; Lewontin, 1988). Here, our purpose was to define a parameter which, in a given model of population structure, does not depend strongly on any 'nuisance' parameter such as the mutation rate, or the number of (possible) allelic states.

(ii) *The conditions for the maintenance of identity disequilibrium*

Identity disequilibrium was found to be a unimodal function of migration rate. (Figs. 4, 5). All else being equal, there is a value of *Nm* which maximizes the identity disequilibrium. However, the value of identity disequilibrium increases with smaller effective sizes. Consequently, it is suggested that both drift within subpopulations and gene flow among subpopulations shape the distribution of identity disequilibria. Taking advantage of this result, we propose in a companion paper (Vitalis & Couvet, in press) a method-of-moments framework to infer both effective population size and migration rates from one- and two-locus identity functions.

Selfing always increases identity disequilibrium (Fig. 3). Furthermore, we showed that, within the appropriate range of *Nm*, identity disequilibrium is maintained when $Nr(1-s) < 1$ (Figs. 5, 6). Golding & Strobeck (1980) obtained the same result for the squared linkage disequilibrium in a partially selfing finite population.

(iii) *Implications for data analysis*

There is still much debate in the literature on whether selfing maintains linkage disequilibrium or not. However, linkage disequilibrium can only be transient, and depends on other ecological parameters, such as the dispersal pattern. Indeed, a recent AFLP study in the selfing species *Arabidopsis thaliana* revealed a low proportion of significant pairwise linkage disequilibrium among markers (Miyashita *et al*., 1999).

With the advance in molecular techniques, assessing the genetic structure of microbial populations has also received growing interest (Lenski, 1993). It has been argued that many micro-organisms may have a clonal population structure (Tibayrenc *et al*., 1990). Linkage disequilibrium, within or among classes of molecular markers, is taken as evidence of clonal, rather than sexual, reproduction. For the agent of human malaria, *Plasmodium falciparum*, Tibayrenc *et al*. (1991) reported contradictory results. Indeed, linkage disequilibrium analyses failed to prove whether the genetic population structure deviated from panmixia (Tibayrenc *et al*., 1991). The interpretation of linkage disequilibrium data may depend on how one analyses the data and on the sampling scale (Lenski, 1993; Maynard Smith *et al*., 1993). Improvement of standard methods to estimate inbreeding coefficients revealed significant departure from Hardy–Weinberg equilibrium among zygotes of the parasite in Tanzania (Hill *et al*., 1995). In Papua New Guinea, high levels of heterozygote deficits (as compared with Hardy–Weinberg proportions) were also found among *P. falciparum* zygotes, in addition with the lack of any linkage disequilibrium across loci (Paul *et al*., 1995). Tibayrenc & Lal (1996) found these two results to be incompatible. They argued that low levels of outcrossing could not prevent non-random gametic

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

associations. However, as we have shown in this paper, high inbreeding and gametic equilibrium may jointly occur, since the extent of identity disequilibrium depends also on gene flow. Finally, within a highly polymorphic region of chromosome 5 in *P. falciparum*, linkage disequilibrium declines with increasing distance map (Conway *et al*., 1999). Furthermore, Anderson *et al*. (2000) recently suggested that highly significant deficits of heterozygous parasite oocysts within mosquitoes could be explained by the presence of null alleles and the subsequent mis-scoring of genotypes. Therefore, there may be enough outcrossing to prevent gametic disequilibrium across unlinked markers. Genetic studies of *P. falciparum* natural populations should be greatly improved with the development of genome-wide linkage maps (Su *et al*., 1999).

However, gene mapping from population-based studies of linkage disequilibrium may be very difficult since the relationship between the recombination fraction and the physical distance is obscured by other population factors, such as genetic drift, reproductive system or population structure (Hill & Weir, 1994; Devlin & Rish, 1995). Instead, if the true map is established from genetic crosses, measures of within-subpopulation identity disequilibria conditioned on recombination rates may be used to infer population parameters of interest, such as the effective size or the immigration rate (Vitalis & Couvet, in press).

The analytical theory we presented here predicts the expected value of η'_S under the assumptions of our model. That expectation is unconditional on any locus being polymorphic. Yet Slatkin (1994) found, through simulations of stationary populations, that there may be a large number of statistically significant linkage disequilibria (Fisher's exact test performed only on pairs of polymorphic loci) even though the population parameters are such that the expected (unconditional) squared linkage disequilibrium is low. Therefore, our results may underestimate the within-subpopulation identity disequilibrium which is likely to be estimated in natural situations from only those loci that are polymorphic. This question deserves further attention.

Finally, we should indicate that other identity disequilibria could be considered. Indeed, within-individual identity disequilibrium can be defined as the difference between the joint homozygosity at two loci and the product of single probabilities for each locus. The analysis of the expected joint distribution in a neutral model of these different disequilibrium parameters could be used, for example, to build neutrality tests for molecular markers.

We thank K. Dawson, B. Godelle and I. Olivieri for helpful comments on a previous draft of this manuscript. We thank W. G. Hill and two anonymous referees for constructive comments. R.V. acknowledges financial support from the Training and Mobility of Researchers programme 'FRAGLAND' of the European Commission, coordinated by I. Hanski, and from the Fondation Sansouire. This is contribution number 2000-079 of the Institut des Sciences de l'Évolution de Montpellier.

Appendix

(i) *Expected values of one- and two-locus identity in the IAM*

Recursion equations for the probabilities of identity by descent (IBD) can generally be written in the form

$$\mathbf{Q}^{t+1} = \mathbf{A}\mathbf{Q}^{t} + \mathbf{B},$$

where $\mathbf{Q}$ is a vector of identity probabilities, $\mathbf{A}$ is a transition matrix, and $\mathbf{B}$ is a vector of coalescent terms. Since two-locus identity probabilities are partial functions of one-locus identity probabilities, we will consider $\mathbf{Q}$, $\mathbf{A}$ and $\mathbf{B}$ as partitioned vectors and matrices. Let

$$\mathbf{Q} = \left(\frac{\mathbf{Q_1}}{\mathbf{Q_2}}\right),$$

with $\mathbf{Q_1} = (Q_{0i}, Q_{0j}, Q_{1i}, Q_{1j})^{\mathrm{T}}$ being the sub-vector of one-locus identity probabilities, for loci i and j, and $\mathbf{Q_2} = (\phi_0, \phi, \gamma, \gamma_3, \gamma_0, \delta, \delta_3, \delta_4, \delta_0, \delta_{00})^{\mathrm{T}}$ being the sub-vector of two-locus identity probabilities, among loci i and j, as defined in the main text. Accordingly, the transition matrix $\mathbf{A}$ is composed of sub-matrices $\mathbf{A_1}$, $\mathbf{A_2}$ and $\mathbf{A_3}$

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{A_1} & \mathbf{0} \\ \hline \mathbf{A_2} & \mathbf{A_3} \end{array}\right)$$

and so follows the vector $\mathbf{B}$

$$\mathbf{B} = \left(\frac{\mathbf{B_1}}{\mathbf{B_2}}\right).$$

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

Considering the recursion equations for one-locus identity probabilities, the mean change over one generation is given by

$$\mathbf{A_1} = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D_1 M_1} \quad \text{and} \quad \mathbf{B_1} = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D_1 C_1}$$

where $\nu_i = (1-\mu_i)^2$ is the probability that, at locus i, neither gene in a pair has mutated; $\mathbf{D_1} = (1-m)^2\,\mathbf{I}$ is a 4×4 matrix containing the probabilities that both genes in a random pair are resident ($\mathbf{I}$ is the identity matrix); $\mathbf{C_1} = (\frac{s}{2}, \frac{s}{2}, \frac{1}{2N}, \frac{1}{2N})^{\mathrm{T}}$ is a vector containing the conditional probabilities that they have coalesced in the previous generation given that they are resident; $\mathbf{M_1}$ is a 4×4 matrix containing the conditional probabilities that two genes in a pair were alike in the previous generation given that they are resident:

$$\mathbf{M_1} = \begin{Bmatrix} \frac{s}{2} & 0 & 1-s & 0 \\ 0 & \frac{s}{2} & 0 & 1-s \\ \frac{1}{2N} & 0 & 1-\frac{1}{N} & 0 \\ 0 & \frac{1}{2N} & 0 & 1-\frac{1}{N} \end{Bmatrix}.$$

The recursion equations for the two-locus identity probabilities as functions of one- and two-locus probabilities are given by

$\mathbf{A_2} = \nu_i \nu_j \mathbf{D_2 M_2}$

$\mathbf{A_3} = \nu_i \nu_j \mathbf{D_2 M_3}$,

with the probabilities that joint pairs of genes have coalesced in the previous generation given by

$\mathbf{B_2} = \nu_i \nu_j \mathbf{D_2 C_2}$,

where $\mathbf{D_2}$ is a 10×10 diagonal matrix containing the probabilities that both pairs of genes were resident in the previous generation:

$\mathbf{D_2} = \mathrm{diag}[(1-m)^2, (1-m)^2, (1-m)^3, (1-m)^3, (1-m)^4, \ldots, (1-m)^4]$.

$\mathbf{M_2}$ is a 4×10 matrix for the conditional one-locus probabilities that, in one of the two pairs, genes were alike (the dots in matrix $\mathbf{M_2}$ signify that the first and second columns are identical, as are the third and fourth columns); $\mathbf{C_2}$ is a vector containing the conditional probabilities that joint pairs of genes have coalesced in the previous generation, given that they were in the same population; and $\mathbf{M_3}$ is a 10×10 matrix for the conditional two-locus probabilities that genes in both pairs were alike in the previous generation, given that they have not migrated:

$$\mathbf{M_2} = \begin{Bmatrix} sr(1-r) & \cdots & 0 & \cdots \\ \frac{r(1-r)}{N} & \cdots & 0 & \cdots \\ \frac{s}{4N} & \cdots & \frac{1-s}{2N} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{s}{4N} & \cdots & \frac{\alpha}{4} & \cdots \\ \frac{(1-s)^2}{4N(N-1)} + \frac{s^2}{4N} & \cdots & \frac{s(1-s)}{N} + (1-s)^2\frac{N-2}{2N(N-1)} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{s}{4N} & \cdots & \frac{\alpha}{4} & \cdots \\ \frac{s^2}{4} & \cdots & \frac{s(1-s)}{2} & \cdots \end{Bmatrix}, \quad \mathbf{C_2} = \begin{Bmatrix} \frac{s}{2}[(1-r)^2 + r^2] \\ \frac{(1-r)^2 + r^2}{2N} \\ \frac{s}{4N} \\ \frac{1}{4N^2} \\ \frac{s}{4N} \\ \frac{(1-s)^2}{4N(N-1)} + \frac{s^2}{4N} \\ \frac{1}{4N^2} \\ \frac{1}{4N^2} \\ \frac{s}{4N} \\ \frac{s^2}{4} \end{Bmatrix}$$

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

and

$$\mathbf{M}_3 = \left\{ \begin{array}{ccccc}
\frac{s}{2}[(1-r)^2+r^2] & (1-s)(1-r)^2 & 2r(1-r)(1-s) & 0 & 0 \\
\frac{(1-r)^2+r^2}{2N} & (1-r)^2\frac{N-1}{N} & 2r(1-r)\frac{N-1}{N} & 0 & 0 \\
\frac{s}{4N} & \frac{s}{2}(1-r)\left(\frac{N-1}{N}\right) & \frac{s}{2}\left(\frac{N-1}{N}\right) & (1-s)(1-r)\frac{N-2}{N} & \frac{1-s}{N} \\
\frac{1}{4N^2} & (1-r)\frac{N-1}{2N^2} & \frac{N-1}{2N^2} & (1-r)\frac{(N-1)(N-2)}{N^2} & \frac{N-1}{N^2} \\
\frac{s}{4N} & (1-r)\frac{1-s}{2N} & \frac{1-s}{2N} & (1-r)\frac{(1-s)(N-2)}{N} & \frac{\alpha}{2} \\
\frac{s^2}{4N} & \frac{\beta}{4(N-1)} & \frac{\beta}{2(N-1)} & (1-s)\frac{N-2}{N-1}\alpha & \frac{2s(1-s)}{N} \\
\frac{s}{4N^2} & \frac{\alpha}{4N} & \frac{\alpha}{2N} & \frac{N-2}{2N^2}[N\alpha+2(1-s)] & \frac{\alpha}{N} \\
\frac{1}{4N^3} & \frac{N-1}{2N^3} & \frac{N-1}{N^3} & \frac{2(N-1)(N-2)}{N^3} & \frac{2(N-1)}{N^3} \\
\frac{s}{4N^2} & \frac{(1-s)}{2N^2} & \frac{(1-s)}{N^2} & (1-s)\frac{2(N-2)}{N^2} & \frac{\alpha}{N} \\
\frac{s^2}{4N} & \frac{(1-s)^2}{2N(N-1)} & \frac{(1-s)^2}{N(N-1)} & (1-s)^2\frac{2(N-2)}{N(N-1)} & \frac{2s(1-s)}{N}
\end{array} \right.$$

$$\left. \begin{array}{ccccc}
(1-s)r^2 & 0 & 0 & 0 & 0 \\
r^2\frac{N-1}{N} & 0 & 0 & 0 & 0 \\
\frac{s}{2}r\frac{N-1}{N} & (1-s)r\frac{N-2}{N} & 0 & 0 & 0 \\
r\frac{N-1}{2N^2} & r\frac{(N-1)(N-2)}{N^2} & 0 & 0 & 0 \\
r\frac{1-s}{2N} & r\frac{(1-s)(N-2)}{N} & 0 & 0 & 0 \\
\frac{\beta}{4(N-1)} & (1-s)\frac{N-2}{N-1}\alpha & (1-s)^2\frac{(N-2)(N-3)}{N(N-1)} & (1-s)^2\frac{N-2}{N(N-1)} & \frac{(1-s)^2}{4N(N-1)} \\
\frac{\alpha}{4N} & \frac{N-2}{2N^2}[N\alpha+2(1-s)] & (1-s)\frac{(N-2)(N-3)}{N^2} & (1-s)\frac{N-2}{N^2} & \frac{1-s}{4N^2} \\
\frac{N-1}{2N^3} & \frac{2(N-1)(N-2)}{N^3} & \frac{(N-1)(N-2)(N-3)}{N^3} & \frac{(N-1)(N-2)}{N^3} & \frac{N-1}{4N^3} \\
\frac{(1-s)}{2N^2} & (1-s)\frac{2(N-2)}{N^2} & (1-s)\frac{(N-2)(N-3)}{N^2} & \frac{N-2}{2N}\alpha & \frac{s(N-1)}{4N^2} \\
\frac{(1-s)^2}{2N(N-1)} & (1-s)^2\frac{2(N-2)}{N(N-1)} & (1-s)^2\frac{(N-2)(N-3)}{N(N-1)} & 2s(1-s)\frac{N-2}{2N} & s^2\frac{N-1}{4N}
\end{array} \right\},$$

with α and β defined such as

$N\alpha = (1-s)+s(N-1)$

$N\beta = (1-s)^2+s^2(N-1)^2.$

(ii) *Expected values of one- and two-locus identity in the KAM*

The K-allele model provides the probabilities of identity in state (IIS). It is assumed that all genes at locus i have the same probability μ_i of mutating to any of the K allelic states. Recursion equations for the probability of identity in state in a K-allele model can also be written in the form

$$\mathbf{Q}^{t+1} = \mathbf{M}'\mathbf{Q}^t + \mathbf{C}',$$

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

with $\mathbf{M}'$ and $\mathbf{C}'$ defined, as for the IAM, as

$$\mathbf{M}' = \left(\begin{array}{c|c} \mathbf{M}'_1 & \mathbf{0} \\ \hline \mathbf{M}'_2 & \mathbf{M}'_3 \end{array}\right)$$

$$\mathbf{C}' = \left(\frac{\mathbf{C}'_1}{\mathbf{C}'_2}\right).$$

Note that $\mathbf{C}$ no longer contains only terms of coalescence, but also the probabilities that some genes that were different in state in the preceding generation become IIS after mutation. As there is a finite number of allelic states K_i at the ith locus, and as the unconditional probability of mutation is μ_i, each allele can mutate to another state with probability $\mu_i/(K_i-1)$. Thus, genes which were identical in state in the previous generation are still identical in state with probability $\nu'_i = (1-\mu_i)^2+\mu^2/(K_i-1)$. Alternatively, genes which were in different allelic states in the previous generation can become identical in state with probability $\omega_i = (1-\nu'_i)/(K_i-1)$. For example, all terms of the form

$$Q_{1i}^{t+1} = \nu_i \frac{1+Q_{0i}^t}{2}$$

in the IAM become, in the KAM,

$$Q_{1i}^{t+1} = \frac{\nu'_i + Q'^t_{0i}}{2},$$

where Q'_{0i} and Q'_{1i} are defined as the conditional IIS probabilities for pairs of genes, after mutation, given the IIS probabilities Q_{0i} and Q_{1i} before mutation. For pairs of genes taken h steps apart in the hierarchy ($h = 0$ for genes taken within individuals or $h = 1$ for genes taken among individuals)

$$\begin{aligned} Q'_{hi} &= \nu'_i Q_{hi} + \omega_i(1-Q_{hi}) \\ &= \rho_i Q_{hi} + \omega_i, \end{aligned}$$

with $\rho_i = \nu'_i - \omega_i$. More generally, we need to rewrite recursion equations by multiplying each term of IAM recursions by ρ_i instead of ν_i, and replacing Q_{hi}'s by Q'_{hi}'s. Consequently, the recursion equations for the one-locus IIS probabilities are given by

$$\mathbf{M}'_1 = \mathbf{D}_1 \mathbf{M}_1 \begin{Bmatrix} \rho_i & 0 & 0 & 0 \\ 0 & \rho_j & 0 & 0 \\ 0 & 0 & \rho_i & 0 \\ 0 & 0 & 0 & \rho_j \end{Bmatrix}$$

and

$$\mathbf{C}'_1 = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D}_1 \mathbf{C}_1 + \mathbf{D}_1 \mathbf{M}_1 \begin{Bmatrix} \omega_i \\ \omega_j \\ \omega_i \\ \omega_j \end{Bmatrix} + (\mathbf{I}-\mathbf{D}_1) \begin{Bmatrix} \frac{1}{K_i} \\ \frac{1}{K_j} \\ \frac{1}{K_i} \\ \frac{1}{K_j} \end{Bmatrix}.$$

Note that the last term in the last equation represents the fact that two genes at locus i which were in different subpopulations in the previous generation have a probability $1/K_i$ to be identical in state, since the number of subpopulations is assumed to be infinitely large.

The natural extension for two-locus IIS identity probabilities is as follows. If we now denote by Q_i some one-locus identity probability, and Q_{ij} some two-locus identity probability, all terms of the form $Q_{ij}^{t+1} = \nu_i \nu_j (1+Q_i^t+Q_j^t+Q_{ij}^t)/4$ in the IAM, become, in the KAM, $Q_{ij}^{t+1} = (\nu'_i \nu'_j + Q'^t_i + Q'^t_j + Q'^t_{ij})/4$, where Q'_i and Q'_j are defined as in the previous section dedicated to the one-locus case, and Q'_{ij} represents some conditional IIS probability of a pair of genes, after mutation, given the IIS probability Q_{ij} before mutation.

$$\begin{aligned} Q'_{ij} &= \nu'_i \nu'_j Q_{ij} + \omega_i \omega_j (1-Q_{ij}) \\ &= (\nu'_i \nu'_j - \omega_i \omega_j) Q_{ij} + \omega_i \omega_j. \end{aligned}$$

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

Again, we need to rewrite recursion equations by multiplying each term of IAM recursions by appropriate factors. Consequently, the recursion equations for the two-locus probabilities of identity in state are now given by

$$\mathbf{M}_2' = \mathbf{D}_2\mathbf{M}_2 \begin{Bmatrix} \nu_j\rho_i & 0 & 0 & 0 \\ 0 & \nu_i\rho_j & 0 & 0 \\ 0 & 0 & \nu_j\rho_i & 0 \\ 0 & 0 & 0 & \nu_i\rho_j \end{Bmatrix} + \mathbf{D}_2'\mathbf{M}_2'' \begin{Bmatrix} \frac{\rho_i}{K_j} & 0 & 0 & 0 \\ 0 & \frac{\rho_j}{K_i} & 0 & 0 \\ 0 & 0 & \frac{\rho_i}{K_j} & 0 \\ 0 & 0 & 0 & \frac{\rho_j}{K_i} \end{Bmatrix}$$

$$\mathbf{M}_3' = (\nu_i\nu_j - \omega_i\omega_j)\,\mathbf{D}_2\mathbf{M}_3$$

$$\mathbf{C}_2' = \mathbf{D}_2\mathbf{M}_2 \begin{Bmatrix} \nu_j\omega_i \\ \nu_i\omega_j \\ \nu_j\omega_i \\ \nu_i\omega_j \end{Bmatrix} + \mathbf{D}_2\mathbf{M}_3 \begin{Bmatrix} \omega_i\omega_j \\ \vdots \\ \omega_i\omega_j \end{Bmatrix} + \nu_i\nu_j\mathbf{D}_2\mathbf{C}_2 + \mathbf{D}_2'\mathbf{M}_2'' \begin{Bmatrix} \frac{\omega_i}{K_j}+\frac{\omega_j}{K_i} \\ \vdots \\ \frac{\omega_i}{K_j}+\frac{\omega_j}{K_i} \end{Bmatrix} + \left(\frac{\nu_i}{K_j}+\frac{\nu_j}{K_i}\right)\mathbf{D}_2'\mathbf{C}_2''$$

$$+ (\mathbf{I} - \mathbf{D}_2 - \mathbf{D}_2') \begin{Bmatrix} \frac{1}{K_iK_j} \\ \vdots \\ \frac{1}{K_iK_j} \end{Bmatrix}$$

with the additional vectors and matrices $\mathbf{D}_2'$, $\mathbf{M}_2''$ and $\mathbf{C}_2''$ defined as

$$\mathbf{D}_2' = \mathrm{diag}[0, 0, 2m(1-m)^2, 2m(1-m)^2, 2m(1-m)^2(2-m), \ldots, 2m(1-m)^2(2-m)]$$

$$\mathbf{M}_2'' = \begin{Bmatrix} 0 & \cdots & 0 & \cdots \\ 0 & \cdots & 0 & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{Ns+1}{8N} & \cdots & \frac{(N-1)+N(1-s)}{4N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{Ns+1}{8N} & \cdots & \frac{(N-1)+N(1-s)}{4N} & \cdots \\ \frac{s}{4} & \cdots & \frac{1-s}{2} & \cdots \end{Bmatrix} \quad \text{and} \quad \mathbf{C}_2'' = \begin{Bmatrix} 0 \\ 0 \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{Ns+1}{8N} \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{Ns+1}{8N} \\ \frac{s}{4} \end{Bmatrix}.$$

References

Anderson, T. J. C., Paul, R. E. L., Donelly, C. A. & Day, K. P. (2000). Do malaria parasites mate non-randomly in the mosquito midgut? *Genetical Research* **75**, 285–296.

Avery, P. J. & Hill, W. G. (1979). Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. *Genetics* **91**, 817–844.

Bennett, J. H. (1954). On the theory of random mating. *Annals of Eugenics* **18**, 311–317.

https://doi.org/10.1017/S0016672300004833 Published online by Cambridge University Press

Bennett, J. H. & Binet, F. E. (1956). Association between Mendelian factors with mixed selfing and random mating. *Heredity* **10**, 51–56.
Charlesworth, D. (1991). The apparent selection on neutral marker loci in partially inbreeding populations. *Genetical Research* **57**, 159–175.
Cockerham, C. C. (1984). Drift and mutation with a finite number of allelic states. *Proceedings of the National Academy of Sciences of the USA* **81**, 530–534.
Conway, D. J., Roper, C., Oduola, A. M. J., Arnot, D. E., Krelsner, P. G., Grobusch, M. P., Curtis, C. F. & Greenwood, B. M. (1999). High recombination rate in natural populations of *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the USA* **96**, 4506–4511.
Crow, J. F. & Aoki, K. (1984). Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the USA* **81**, 6073–6077.
David, P. (1999). A quantitative model of the relationship between phenotypic variance and heterozygosity at marker loci under partial selfing. *Genetics* **153**, 1463–1474.
Devlin, B. & Rish, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.
Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics* **15**, 25–57.
Golding, G. B. & Strobeck, C. (1980). Linkage disequilibrium in a finite population that is partially selfing. *Genetics* **94**, 777–789.
Hastings, A. (1984). Maintenance of high disequilibrium in the presence of partial selfing. *Proceedings of the National Academy of Sciences of the USA* **81**, 4596–4598.
Hedrick, P. W. (1987). Gametic disequilibrium: proceed with caution. *Genetics* **117**, 331–341.
Hill, W. G. (1974*a*). Disequilibrium among several linked neutral genes in finite populations. I. Mean changes in disequilibrium. *Theoretical Population Biology* **5**, 366–392.
Hill, W. G. (1974*b*). Disequilibrium among several linked neutral genes in finite populations. II. Variances and covariances of disequilibria. *Theoretical Population Biology* **6**, 184–198.
Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.
Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics* **38**, 209–216.
Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.
Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.
Hill, W. G. & Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics* **54**, 705–714.
Hill, W. G., Babiker, H. A., Ranford-Cartwright, L. C. & Walliker, D. (1995). Estimation of inbreeding coefficients from genotypic data on multiple alleles, and their application to estimation of clonality in malaria parasites. *Genetical Research* **65**, 53–61.
Lenski, R. E. (1993). Assessing the genetic structure of microbial populations. *Proceedings of the National Academy of Sciences of the USA* **90**, 4334–4336.
Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics* **120**, 849–852.
Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences of the USA* **90**, 4384–4388.
Miyashita, N. T., Kawabe, A. & Innan, H. (1999). DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* **152**, 1723–1731.
Nei, M. & Li, W.-H. (1973). Linkage disequilibrium with the island model. *Genetics* **101**, 139–155.
Ohta, T. (1980). Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigene families. *Genetical Research* **36**, 181–197.
Ohta, T. (1982*a*). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **79**, 1940–1944.
Ohta, T. (1982*b*). Linkage disequilibrium with the island model. *Genetics* **101**, 139–155.
Ohta, T. & Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.
Paul, R. E. L., Packer, M. J., Walmsley, M., Lagog, M., Ranford-Cartwright, L. C., Paru, R. & Day, K. P. (1995). Mating patterns in malaria parasite populations of Papua New Guinea. *Science* **269**, 1709–1711.
Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation process. *Genetics* **142**, 1357–1362.
Slatkin, M. (1975). Gene flow and selection in a two-locus system. *Genetics* **81**, 787–802.
Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.
Strobeck, C. (1979). Partial selfing and linkage: the effect of a heterotic locus on a neutral locus. *Genetics* **92**, 305–315.
Su, X.-Z., Ferdig, M. T., Huang, Y., Huynh, C. Q., Liu, A., You, J., Wooton, J. C. & Wellems, T. E. (1999). A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286**, 1351–1353.
Sved, J. A. (1968). The stability of linked systems of loci with a small population size. *Genetics* **59**, 543–563.
Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.
Tachida, H. & Cockerham, C. C. (1986). Analysis of linkage disequilibrium in an island model. *Theoretical Population Biology* **29**, 161–197.
Takahata, N. (1982). Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genetical Research* **39**, 63–77.
Tibayrenc, M. & Lal, A. (1996). Self-fertilization, linkage disequilibrium, and strain in *Plasmodium falciparum*. *Science* **271**, 1300–1301.
Tibayrenc, M., Kjellberg, F. & Ayala, F. J. (1990). A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proceedings of the National Academy of Sciences of the USA* **87**, 2414–2418.
Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Brenière, S. F., Dardé, M.-L. & Ayala, F. J. (1991). Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proceedings of the National Academy of Sciences of the USA* **88**, 5129–5133.
Vitalis, R. & Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* (in press).
Weir, B. S. & Cockerham, C. C. (1969). Group inbreeding with two linked loci. *Genetics* **63**, 711–742.
Weir, B. S. & Cockerham, C. C. (1973). Mixed self and random mating at two loci. *Genetical Research* **21**, 247–262.

Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.

Weir, B. S., Avery, P. J. & Hill, W. G. (1980). Effect of mating structure on variation in inbreeding. *Theoretical Population Biology* **18**, 369–429.

Whitlock, M. C., Phillips, P. C. & Wade, M. J. (1993). Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. *Evolution* **47**, 1758–1769.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.