Heredity **84** (2000) 373–389 Received 2 June 1999, accepted 17 November 1999

Estimating multilocus linkage disequilibria

N. H. BARTON*
Institute of Cell, Animal and Population Biology, University of Edinburgh,
West Mains Road, Edinburgh EH9 3 JT, Scotland

The state of a diploid population segregating for two alleles at each of n loci is described by 2^{2n} genotype frequencies, or equivalently, by allele frequencies and by multilocus moments or cumulants of various orders. These measures of linkage disequilibrium cannot usually be determined, both because one cannot tell whether a gene came from the maternal or paternal gamete, and because such a large number of parameters cannot be estimated even from large samples. Simplifying assumptions must therefore be made. This paper sets out methods for estimating multilocus genotype frequencies which are appropriate for unlinked neutral loci, and for populations that are ultimately derived by mixing of two source populations. In such a hybrid population, all multilocus associations depend primarily on the number of loci involved that derive from the maternal genome, and the number derived from the paternal genome. Allele frequencies may differ across loci, and the contribution of each locus to multilocus associations may be scaled by the difference in allele frequency between source populations for that locus ($\delta p \leq 1$). For example, the cumulant describing the association between genes i, j, k from the maternal genome, and genes i, l from the paternal genome is $\kappa_{i,j,k,i^*l^*}, = \delta p_i^2 \, \delta p_j \, \delta p_k \, \delta p_l \, \kappa_{3,2}$. The state of the population is described by n allele frequencies; n divergences, δp; and by a symmetric matrix of cumulants, $\kappa_{J,K}$ $(J = 0, \ldots, n,\ K = 0, \ldots, n)$. Expressions for these cumulants under short- and long-range migration are given. Two methods for estimating the cumulants are described: a simple method based on multivariate moments, and a maximum likelihood procedure, which uses the Metropolis algorithm. Both methods perform well when tested against simulations with two or four loci.

Keywords: cytonuclear associations, hybrid population, linkage disequilibrium, Monte Carlo.

Introduction

Random mating and recombination break down associations between alleles at different genetic loci, leading populations towards "linkage equilibrium". Nonrandom associations ("linkage disequilibria") can be generated by a variety of evolutionary processes; their observation gives an indirect method for quantifying the rates of these processes. Thus, there has been considerable interest in estimating the strength of linkage disequilibria, motivated by varied aims: finding gene order and recombination rates (Hill & Weir, 1994); detecting genes responsible for human disease (Kaplan *et al*., 1995); detecting selection for particular gene combinations (Langley, 1977); and measuring patterns and rates of migration (Asmussen *et al*., 1987; Barton & Gale, 1993). This paper is concerned with understanding hybrid populations, in which the mixing of genetically distinct taxa can maintain strong associations even between unlinked genes (Li & Nei, 1974). Methods for estimating and interpreting pairwise associations are well worked out, both for nuclear–nuclear and nuclear–cytoplasmic associations (Hill, 1974a; Asmussen *et al*., 1987). However, data are usually available from several marker loci: here, we set out methods for combining data across loci, paying particular attention to the case where there are strong associations both between genes from the same gamete, and between genes inherited from different parents.

There are three key difficulties. First, estimates of associations between different pairs of loci are not independent of each other, especially when associations are strong. Secondly, we do not usually wish to estimate individual associations; in any case, there are far too many such associations with even a few loci. Rather, we wish to find some composite measure which gives information about the process responsible for generating linkage disequilibrium. (For example, random drift and recombination will generate a distribution whose parameters might be estimated from the variance in pairwise disequilibrium; e.g. Langley, 1977.) Thirdly,

*Correspondence. E-mail: n.barton@ed.ac.uk

unless samples are of zygotes freshly generated by random mating, there will be associations between genes from different parents; with diploids, these cannot be distinguished from associations between genes derived from the same gamete.

Most of the literature on estimating linkage disequilibrium has been concerned with the third problem, of resolving genotypes that are confounded in diploid data (for example, using the EM algorithm; Hill, 1974a, 1975; Dempster *et al.*, 1977; Long *et al.*, 1995; Slatkin & Excoffier, 1996). There is also a good understanding of the distribution of associations generated by random drift within a single population (Hill, 1974b,c). The primary motivation for this study is the estimation of associations generated by the mixing of populations. This raises particular difficulties, because there are often strong associations, both within and between genomes; it is also particularly simple, in that admixture models make straightforward predictions as to the structure of associations in a population.

The crucial requirement is for some model which specifies the frequencies of genotypes in terms of a reasonably small number of parameters. Given such a model, the parameters can be estimated by maximum likelihood. Such a model is also necessary for other statistical problems where genotype frequencies must be specified: for example, estimating paternities or mapping quantitative trait loci. Here, we show that many kinds of admixture lead to a model with a simple structure, and show how the parameters of this model can be estimated by maximum likelihood. First, however, we justify the choice of this model by discussing alternative methods which might be used to determine the structure of genotype frequencies.

The key difficulties arise because of the very large number of possible genotypes. For example, with two alleles at each of five loci, there are $2^5 = 32$ haplotypes and $3^5 = 243$ distinguishable diploid genotypes. Unless samples are extremely large, one cannot estimate even haplotype frequencies. Moreover, if gametes are not combined at random, the genetic structure is determined by the much larger number of diploid genotypes; yet, these cannot be distinguished without knowing which gene came from which parent. The very large number of genotypes rules out the most obvious method: to make a maximum likelihood estimate (MLE) of the genotype frequencies. Even in large samples, most diplotypes, and even most haplotypes, may be missing. Therefore, the MLE would be that only observed genotypes are actually present, implying complete associations between certain gene combinations. This can raise serious difficulties. For example, suppose that we wish to find whether some family was sired by a known male, or by some unknown father. If we suppose that genotypes that are absent in our sample are absent in the rest of the population, then there will be an undue tendency to assign paternity to implausible parental genotypes which happen to be represented amongst the known individuals.

The problem, therefore, is to describe the genotype frequencies in terms of a small number of parameters; the description must be simple enough to make calculations tractable, and must include the structure of the actual populations under study. A popular solution to this kind of problem has been to use the method of "maximum entropy" (Guiasu & Shenitzer, 1985; Phipps & Brill, 1995). The entropy of a sample is defined as $S = \sum_X g[\boldsymbol{X}] \log[g[\boldsymbol{X}]]$, where $g[\boldsymbol{X}]$ is the frequency of genotype $\boldsymbol{X}$. Given the constraint $\sum_X g[\boldsymbol{X}] = 1$, entropy is maximized when all genotypes are equally common. Given further constraints on allele frequencies, the maximum entropy is at linkage equilibrium. With constraints on pairwise linkage disequilibria, the maximum entropy solution defines the genotype frequencies in terms of the small number of pairwise coefficients. In the haploid case, the solution is $g[\boldsymbol{X}] = \exp[\sum_i \lambda_i X_\mathrm{i} + \sum_{i,j} \lambda_{i,j} X_i X_j]/Z$, where Z is a normalizing constant, and $X_i = 0$ or 1 defines the state of the *i*th locus. The λs are determined by the equations $\partial \ln(Z)/\partial\lambda_i = p_i$, $\partial \ln(Z)/\partial\lambda_{i,j} = p_i p_j + C_{i,j}$, where p_i is the allele frequency, and $C_{i,j}$ is the covariance (equivalent to the linkage disequilibrium) between loci i and j. This method has a tempting generality, but suffers two drawbacks. First, numerical solutions involve complicated transcendental equations, which become intractable in the diploid case. Secondly, there is no reason why evolutionary processes should lead populations to have "maximum entropy". The method has been justified as reflecting our ignorance of unknown frequencies (analogous to arguments justifying parsimony in phylogeny reconstruction; Sober, 1983). However, even though we may not know actual genotype frequencies, we may know the patterns of genotype frequencies likely to be produced by evolution: estimation procedures should be based on these patterns. Moreover, the parameter estimates are not interesting in themselves: they can only be interpreted in the light of some evolutionary model.

A related approach is that taken by Haber (1984) where a log-linear model is used to parameterize genotype frequencies. For two loci (labelled i, j), the frequency of the diploid combination $\{\{P_i, P_j\},\{P_{i^*}, P_{j^*}\}\}$ (where i and j label the alleles inherited from the mother, and i^*, j^* those from the father), is given as $\exp[\mu + \alpha_i + \alpha_j + \alpha_{i^*} + \alpha_{j^*} + \beta_{ij} + \beta_{i^*j^*} + \beta_{ii^*} + \beta_{jj^*} + \beta_{ij^*} + \beta_{ji^*}]$. Only first- and second-order interactions are considered, but the method could be extended to include higher-order interactions and multiple loci. A difficulty with this method is that, in order to determine the parameters uniquely, some arbitrary reference point needs to be assigned, which will not necessarily have a biological interpretation. For example, Haber (1984) suggests setting the sums over the indices to zero. With no higher-order interactions, the allele frequencies for locus i are then $p_1 = \exp[\mu + \alpha_1]$ and $p_2 = \exp[\mu - \alpha_1]$; however, the parameters μ, α do not bear a simple relation to allele frequencies in the presence of interactions. Another problem is that the parameters of the log-linear model measure net effects; the parameter β_{jj^*}, which measures deviations from random mating with respect to locus j, could be zero, but if some of the other five sets of two-way parameters, β, are nonzero, the covariance between alleles at locus j will not be zero. Finally, there is still the problem of estimating a large number of parameters; simply setting parameters for higher-order interactions to zero may cause erroneous estimates of lower-order interactions.

Here, we suppose that the population is subject to continuing immigration by adults, from two source populations with different allele frequencies. Mixing of divergent populations builds up associations between genes inherited from the same parent, and from different parents. In each generation, associations between genes inherited from different parents are eliminated by segregation and recombination, and then built up again by immigration. We assume that continued migration does not dissipate differences between populations; in reality, these may decay, or may be maintained by selection. However, the assumption of a short-term balance between recombination and migration is accurate even when allele frequencies are homogenizing, or when selection maintains them in the face of gene flow (Barton & Gale, 1993; Kruuk, 1997; Barton & Shpak, 2000a; Kruuk *et al.*, 1999).

It would be possible to fit a specific model of admixture, giving estimates of migration rates, allele frequencies in the source populations, and so on. However, the distribution of genotype frequencies depends on the detailed migration pattern, which is often unknown. In particular, migration at a low rate from divergent populations maintains a small fraction of genetically distinct individuals; a higher rate of immigration from less divergent populations may maintain the same genetic variance, but much smaller higher-order associations. We therefore aim to find a genetic structure which covers this range of admixture models; estimates of this general structure can then be compared with particular migration models.

This approach is closely related to the simple practice of classifying hybrid individuals into various classes of cross (Arnold, 1997). If hybridization is rare or recent, then there may be distinct classes of parental and F_1 genotypes. Backcrosses can be identified because they are homozygous only for genes from a single taxon; given enough markers, the number of generations of backcrossing could be estimated (Goodman *et al.*, 1999). However, if hybridization has been extensive, individuals can no longer be unambiguously assigned (at least, without an inordinate number of markers). The classification can then be misleading: for example, putatively F_1 genotypes may be more likely to be generated by chance than by an actual cross between parental genotypes. One could still describe a population as some mixture of parentals, F_1s, backcrosses and F_2s. However, this mix is not uniquely determined by genotype frequencies. We discuss below the relation between our description of genotypic structure, and classification by hybrid status.

The analysis falls into three sections. First, it is shown that models of admixture, involving both short- and long-range migration, lead to associations among genes which depend on the number of genes involved, and on how many of each kind come from each parent. Associations involving particular loci are proportional to the divergence in allele frequency for that locus; this makes it simple to rescale, so as to allow for variation in associations across loci. Secondly, an algorithm is presented for making maximum likelihood estimates of the within- and between-genome associations, and for testing hypotheses as to their magnitude. Finally, the accuracy of this algorithm is demonstrated against simulated datasets. Procedures for analysing multilocus admixture models, and for maximum likelihood estimation, are implemented in MATHEMATICA 3.0, and are available from http://helios.bto.ed.ac.uk/evolgen/.

Models of admixture

Suppose that proportions m_1, m_2 migrate in from two source demes, with allele frequencies $p_{1;i}$, $p_{2;i}$ at locus i. At equilibrium, the allele frequency in the deme of interest is $p_i = (m_1 p_{1;i} + m_2 p_{2;i})/(m_1 + m_2)$. The notation can therefore be simplified by letting $p_{1;i} = p_i - U\delta p_i$, $p_{2;i} = p_i + V\delta p_{\mathrm{i}}$, where the migration rates are $m_1 = mU$, $m_2 = mV$, and where $\delta p_i = (p_{2;i} - p_{1;i})$. Thus, m is the total immigration rate, and U, V are the proportions of migrants from each source ($U + V = 1$). Note that U is determined by the ratio of migration rates, and so is the same for all loci; it is not necessarily equal to the mean of the p_i. If alternative alleles are fixed in the source demes, then $\delta p_i = 1$, and $p_i = U$. However, if $\delta p_i < 1$, then the equilibrium allele frequencies p_i may vary across loci.

We must now determine the associations among loci at equilibrium. Following Turelli & Barton (1994), we define these as multivariate cumulants across sets of genes. Consider loci with two alleles, labelled by $X_i = 0$ or 1; in a diploid, the copies inherited from the mother and father are labelled X_i, X_i^*, respectively, and the full genotype is defined by the pair of vectors $\{\boldsymbol{X}, \boldsymbol{X}^*\}$. Deviations from the population mean are defined by $\zeta_i = (X_i - p_i)$, where $p_i = \mathrm{E}[X_i]$. The pairwise association between genes at i and j, inherited from the mother, can be described by the covariance between X_i, X_j, denoted $C_{\{i,j\}} \equiv \mathrm{E}[\zeta_i \zeta_j]$. Similarly, the pairwise association between the gene at i inherited from the mother, and the gene at locus j inherited from the father, can be described by $C_{\{i,j^*\}} \equiv \mathrm{E}[\zeta_i \zeta_j^*]$, where $p_{i^*} = \mathrm{E}[X_i^*]$. (The coefficient $C_{\{i,i^*\}}$ describes the deficit of heterozygotes at locus i). Higher-order associations are described in a similar way to multilocus moments (Barton & Turelli, 1991.) The genotype frequencies can be reconstructed as a linear combination of the various multilocus moments. With two loci, for example, the frequency of the double homozygote {{1, 1}, {1, 1}} is:

$$\begin{aligned} &g[\{\{1,1\},\{1,1\}\}] \\ &\quad = p_i p_j p_{i^*} p_{j^*} + p_{i^*} p_{j^*} C_{\{i,j\}} + p_j p_{j^*} C_{\{i,i^*\}} \\ &\qquad + p_j p_{i^*} C_{\{i,j^*\}} + p_i p_{j^*} C_{\{j,i^*\}} + p_i p_{i^*} C_{\{j,j^*\}} + p_i p_j C_{\{i^*,j^*\}} \\ &\qquad + p_{j^*} C_{\{i,j,i^*\}} + p_j C_{\{i,i^*,j^*\}} + p_{i^*} C_{\{i,j,j^*\}} \\ &\qquad + p_i C_{\{j,i^*,j^*\}} + C_{\{i,j,i^*,j^*\}}. \end{aligned} \tag{1}$$

The frequency of other genotypes is found by replacing p_i by q_i, and changing the sign of every coefficient C_U, whenever an allele $X_i = 1$ is replaced by $X_i = 0$ in (1). Thus, the 16 genotype frequencies are determined by four allele frequencies, six pairwise associations, four three-way associations, and one four-way association. The frequencies of more complex genotypes could be described in a similar way, by use of higher-order moments.

The population can also be described in terms of multilocus cumulants rather than moments (Turelli & Barton, 1994). Cumulants are polynomial functions of the moments; the second- and third-order cumulants are identical to the central moments, but the fourth- and higher-order moments differ. For example, $\kappa_{\{i,j,k,l\}} = C_{\{i,j,k,l,\}} - C_{\{i,j\}} C_{\{k,l\}} - C_{\{i,k\}} C_{\{j,l\}} - C_{\{i,l\}} C_{\{j,k\}}$. Cumulants can be thought of as describing the association amongst a set of genes, over and above those expected from the lower-order associations amongst the various subsets. They are defined so as to be additive. Thus, if an additive trait is constructed as the sum of contributions of all the genes $(\sum_i (X_i + X_i^*))$, then its kth cumulant is the sum of all cumulants amongst sets of k genes. We choose cumulants to describe admixture models because higher-order cumulants are small in admixture models; all cumulants of the same kind are of similar magnitude (see below); and at Hardy–Weinberg, all cumulants involving genes inherited from different parents are zero. (This is not the case for moments: for example, $C_{\{i,j,k^*,l^*\}} = C_{\{i,j\}} C_{\{k^*,l^*\}}$ for a population in Hardy–Weinberg proportions.)

Genotype frequencies are readily derived from eqn 1 by expressing moments in terms of cumulants. For example:

$$\begin{aligned} &g[\{\{1,1\},\{1,1\}\}] \\ &\quad = (p_i p_j + \kappa_{\{i,j\}})(p_{i^*} p_{j^*} + \kappa_{\{i^*,j^*\}}) \\ &\qquad + (p_j p_{j^*} + \kappa_{\{j,j^*\}})(p_i p_{i^*} + \kappa_{\{i,i^*\}}) + (p_j p_{i^*} + \kappa_{\{j,i^*\}}) \\ &\qquad \times (p_i p_{j^*} + \kappa_{\{i,j^*\}}) + (p_{j^*} \kappa_{\{i,j,i^*\}} + p_j \kappa_{\{i,i^*,j^*\}} + p_{i^*} \kappa_{\{i,j,j^*\}} \\ &\qquad + p_i \kappa_{\{j,i^*,j^*\}}) + \kappa_{\{i,j,i^*,j^*\}} \quad 2 p_i p_j p_{i^*} p_{j^*}. \end{aligned} \tag{2}$$

The same rules apply for finding the frequency of other genotypes as above.

These expressions can readily be extended to allow for multiple alleles, provided that there are ultimately just two sources of migrants, which are in linkage equilibrium. Suppose that within deme 2, allele α_i at locus i is at frequency $p_{2;\alpha_i}$; similarly for deme 1. Now, regardless of their actual allelic state, label alleles that derive from either of the two demes as $X_i = 0$ or 1. Then, the frequency of an individual entirely composed of alleles derived from deme 2 is given by eqns 1 or 2; and the chance that this individual carries alleles α_i is the product of the $p_{2;\alpha_i}$s. The net frequency of some allelic combination is given by a sum over all origins of those alleles, multiplied by their probability, given that origin. This sum will simplify if certain alleles are found only in one or other source.

The algorithms which define the effects of recombination and random mating within demes are given in Turelli & Barton (1994); migration simply involves a linear mixture of moments. The algorithms can be found at http://helios.bto.ed.ac.uk/evolgen/. Here, we apply them to find the associations which would be found at equilibrium, in a balance between migration and recombination. We assume no linkage between loci, which is the case of most practical interest. (Linkage introduces considerable difficulties, because associations will decrease in a complicated way with recombination rate.) The expressions below assume symmetry across the two sexes; they also apply to cytonuclear disequilibria, provided that evolutionary processes apply equally to both sexes.

Below, we give the equilibrium solutions for two extreme cases. First, migrants might come from source populations which are in linkage equilibrium. This will necessarily apply if the sources are fixed for alternative alleles, and so we refer to this case as "long-range migration". Secondly, migrants might come from neighbouring demes with different allele frequencies, but with the same levels of association. This is an approximation to a stepping-stone model, where migration is between neighbours which are in a similar state. It ignores the diffusion of linkage disequilibrium, which tends to reduce associations in the centre, and increase them at the edge. (This is the "quasi-linkage equilibrium" approximation of Barton (1986) and Kruuk *et al.* (1999).) We compare the "short-range migration" approximation with exact simulations below.

Assuming symmetry across the sexes, pairwise associations under these two approximations are:
short-range migration:

$$\kappa^{\text{short}}_{\{i,j\}} = 3mUV\bar{q}\delta p_i\delta p_j, \quad \kappa^{\text{short}}_{\{i,j^*\}} = mUV\delta p_i\delta p_j;$$

long-range migration

$$\kappa^{\text{long}}_{\{i,j\}} = \kappa^{\text{short}}_{\{i,j^*\}}\frac{(1 \quad \frac{m}{3})}{(1+m)}, \quad \kappa^{\text{long}}_{\{i,j^*\}} = \kappa^{\text{short}}_{\{i,j^*\}}. \tag{3}$$

Associations between the two copies of an allele at the same locus, $\kappa_{\{i,i^*\}}$, which describe heterozygote deficit, are given by replacing j^* by i^* in the expression for $\kappa_{\{i,j^*\}}$ above. The influx of linkage disequilibrium from the source populations assumed under the model of short-range migration increases associations within genomes by a factor $(1+m)/(1-m/3)$, but does not affect associations between genomes. With both kinds of migration, the associations between genomes are smaller because they are broken down in every generation by segregation. All the expressions in eqn 3 are for associations measured immediately after dispersal; among zygotes, there will be no associations between maternal and paternal genomes. Associations within genomes will be reduced to $\kappa^*_{\{i,j\}} = (\kappa_{\{i,j\}} + \kappa_{\{i,j^*\}})/2$ by random segregation and mating.

Three-way associations are given by similar expressions:
short-range migration:

$$\kappa^{\text{short}}_{\{i,j,k\}} = \quad \frac{7}{3}mUV(U \quad V)\delta p_i\delta p_j\delta p_k,$$

$$\kappa^{\text{short}}_{\{i,j^*,k^*\}} = \quad mUV(U \quad V)\delta p_i\delta p_j\delta p_k; \tag{4}$$

long-range migration:

$$\kappa^{\text{long}}_{\{i,j,k\}} = \kappa^{\text{short}}_{\{i,j,k\}}\frac{(1 \quad \frac{3m}{7})}{(1+\frac{m}{3})}, \quad \kappa^{\text{long}}_{\{i,j^*,k^*\}} = \kappa^{\text{short}}_{\{i,j^*,k^*\}}.$$

Finally, we give the four-way associations:
short-range migration:

$$\kappa^{\text{short}}_{\{i,j,k,l\}} = \frac{15}{7}mUV(1 \quad 3UV(1+m))\delta p_i\delta p_j\delta p_k\delta p_l,$$

$$\begin{aligned}\kappa^{\text{short}}_{\{i,j^*,k^*,l^*\}} &= \kappa^{\text{short}}_{\{i,j,k^*,l^*\}} \\ &= mUV(1 \quad 3UV(1+m))\ \delta p_i\delta p_j\delta p_k\delta p_l;\end{aligned} \tag{5}$$

long-range migration:

$$\kappa^{\text{long}}_{\{i,j,k,l\}} = \frac{mUV((1+m)^2(15 \quad 7m) \quad 3(15+82m \quad 24m^2 \quad 10m^3+m^4)UV)}{(1+m^2)(7+m)} \times \delta p_i\delta p_j\delta p_k\delta p_l,$$

$$\kappa^{\text{long}}_{\{i,j^*,k^*,l^*\}} = mUV\left(1 \quad 3\frac{(1+4m \quad m^2)}{(1+m)}UV\right)\delta p_i\delta p_j\delta p_k\delta p_l,$$

$$\kappa^{\text{long}}_{\{i,j,k^*,l^*\}} = mUV\left(1 \quad \frac{(3+13m+5m^2+3m^3)}{(1+m)^2}UV\right) \times \delta p_i\delta p_j\delta p_k\delta p_l.$$

Expressions for higher-order associations become more complicated, but share a similar structure. In particular, all associations involving a set of loci U are given by the product $\Pi_{i\in U}\delta p_i$, multiplied by a factor independent of the loci involved. This is important, because it gives a simple way of scaling out variation across loci, and because it allows associations between genes at different loci, $\kappa_{\{i,j^*\}}$, which are not directly observable, to be estimated from observations on heterozygote deficit, $\kappa_{\{i,i^*\}}$. It also allows the effects of immigration from near and far to be distinguished: if δp is small, then associations among many genes become small, whereas if δp is close to 1 (its maximum value), then all higher-order associations are proportional to mUV, and may be large. This is reflected in a leptokurtic distribution of the number of alleles from one or other taxon.

The strength of associations generated by admixture does not depend directly on allele frequency. It is proportional to mUV, which is the harmonic mean of the immigration rates mU, mV. If $\delta p_i < 1$ (for example, with short-range migration), then allele frequencies might vary across loci without directly affecting the associations between them.

Associations between genes inherited from different parents depend primarily on the numbers of genes involved. For example, eqn 5 shows that $\kappa_{\{i,j,k,l^*\}}$, $\kappa_{\{i,j^*,k^*,l^*\}}$ are identical to $\kappa_{\{i,j,k^*,l^*\}}$ for short-range migration, and with long-range migration, are extremely close over the whole range of migration rates (Figs 1 and 2). Similar calculations show that all fifth- and sixth-order cross-genome cumulants are similar; agreement becomes closer as migration rates become asymmetric (i.e. $U \ll 1/2$ or $\gg 1/2$). This similarity between different classes of association arises because higher-order associations are almost entirely eliminated by segregation: the nth-order association decreases by a factor 2^{n-1} every generation. Therefore, associations are close to those built up by admixture in the current generation, which is the same for all nth-order associations.

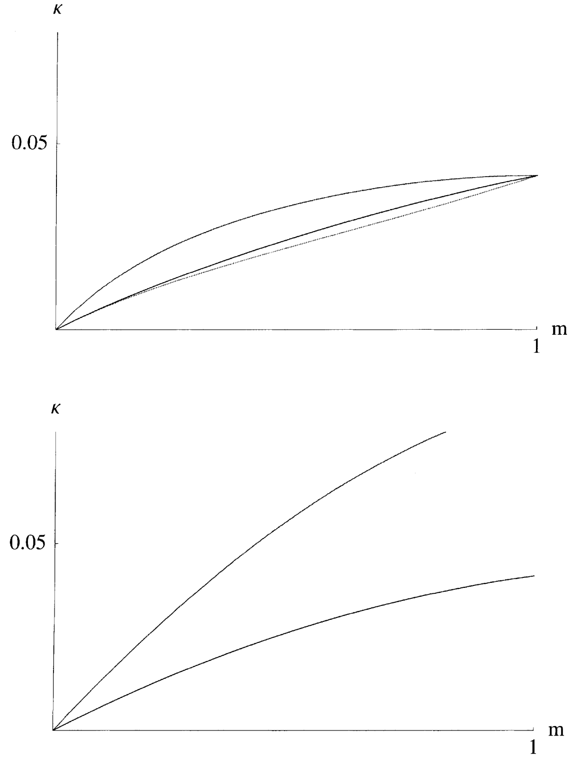

Fig. 1 Comparisons between the three kinds of four-way association, $\kappa_{\{i,j,k^*,l^*\}}$, $\kappa_{\{i,j,k,l^*\}}$, $\kappa_{\{i,j,k,l\}}$, (bottom to top) plotted against migration rate, m; $U = 0.1$. Values are from eqn 5, scaled relative to $\delta p_i\ \delta p_j\ \delta p_k\ \delta p_l$. (a) Long-range migration, (b) short-range migration. With short-range migration, (b), $\kappa_{\{i,j,k^*,l^*\}}$ is identical to $\kappa_{\{i,j,k,l^*\}}$. Note that with short-range migration in a stepping-stone model, (b), $m < 1/2$, and $\delta p \ll 1$; therefore, the actual associations may be much smaller than with long-range migration, even though the scaled values shown here are larger.

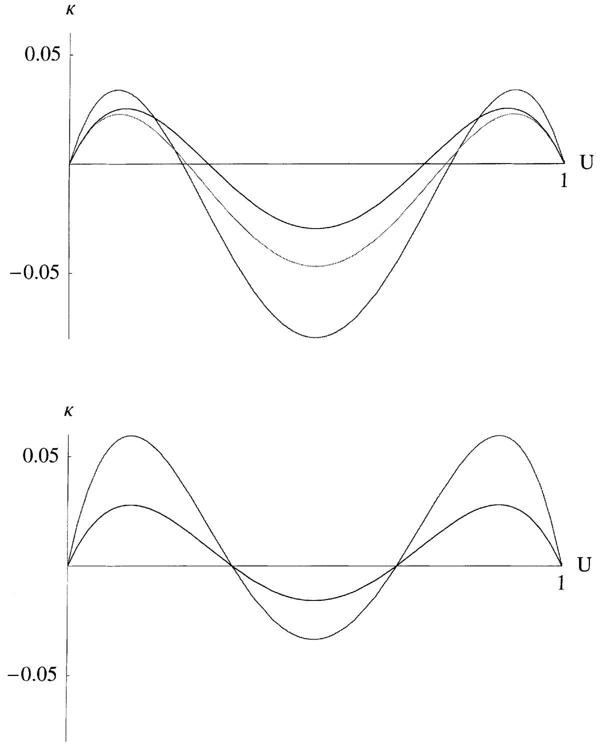

Fig. 2 Comparisons between the three kinds of four-way association, $\kappa_{\{i,j,k^*,l^*\}}, \kappa_{\{i,j,k,l^*\}}, \kappa_{\{i,j,k,l\}}$ (bottom to top at $p = 0.1$), plotted against allele frequency, U; $m = 1/2$. Values are from eqn 5, scaled relative to $\delta p_i\ \delta p_j\ \delta p_k\ \delta p_l$. (a) Long-range migration, (b) short-range migration. With short-range migration, (b), $\kappa_{\{i,j,k^*,l^*\}}$ is identical to $\kappa_{\{i,j,k,l^*\}}$.

Calculating genotype frequencies

In admixture models with random mating and with just two sources of immigrants, genotype frequencies can be specified in terms of allele frequencies, p_i; relative divergence across loci, δp_i; Kth-order within-genome associations, $\kappa_{0,K}$; and between-genome associations involving J genes from one parent, and K from the other, $\kappa_{J,K}$. With n loci, and symmetry across the sexes, there are $(n - 1)$ within-genome coefficients; $n(n - 1)/2$ between-genome coefficients, $\kappa_{J,K}$; n allele frequencies, p_i; and $(n - 1)$ divergences, δp_i. Therefore, the 3^n genotype frequencies are determined by $(n^2 + 5n - 4)/2$ parameters. (Note that the δp_i only give the relative importance of each locus, and could be multiplied by an arbitrary constant; they are therefore associated with $(n - 1)$ degrees of freedom.) For example, with four loci, the 81 distinguishable diploid genotypes are specified by just 16 parameters; with five loci, the 243 distinguishable genotypes depend on 23 parameters. The similarity between all nth-order cross-genome associations (Figs 1 and 2) suggests a further simplification: to equate all cross-genome cumulants involving the same number of loci. However, this can lead to negative genotype frequencies: even the small changes needed to force equality among cross-genome associations of the same order can force genotype frequencies outside their valid range.

If the set of genes U all come from the same parent, $\kappa_U = \kappa_{0,|U|}(\Pi_{i\in U}\delta p_i)$, where $|U|$ is the number of genes in U; if a set U come from one parent, and U^* from the other, then $\kappa_{U,U^*} = \kappa_{|U|,|U^*|}(\Pi_{i\in U}\delta p_i)(\Pi_{i\in U^*}\delta p_{i^*})$. The genotype frequency (1) can then be written compactly as follows:

$$\begin{aligned}
&g[\{\{1,1\},\{1,1\}\}]\\
&\quad = p_i p_j p_{i^*} p_{j^*} + (p_{i^*} p_{j^*} \delta p_i \delta p_j + p_i p_j \delta p_{i^*} \delta p_{j^*})C_{0,2}\\
&\qquad + (p_i \delta p_j + \delta p_i p_j)(p_{i^*} \delta p_{j^*} + \delta p_{i^*} p_j^*)C_{1,1}\\
&\qquad + (p_{j^*} \delta p_i \delta p_j \delta p_{i^*} + p_j \delta p_i \delta p_{j^*} \delta p_{i^*}\\
&\qquad + p_{i^*} \delta p_i \delta p_j \delta p_{j^*} + p_i \delta p_j \delta p_{i^*} \delta p_{j^*})C_{1,2}\\
&\qquad + \delta p_i \delta p_j \delta p_{i^*} \delta p_{j^*} C_{2,2}\\
&\quad = \mathbf{\Gamma} \cdot \mathbf{C} \cdot \mathbf{\Gamma}^{*\mathrm{T}}
\end{aligned}$$

where $\mathbf{\Gamma} = (p_i p_j(p_i \delta p_{\mathrm{j}} + \delta p_{\mathrm{i}} p_j)\delta p_i \delta p_j)$,

$$\mathbf{\Gamma}^* = (p_{i^*} p_{j^*}(p_{i^*} \delta p_{j^*} + \delta p_{i^*} p_j^*)(\delta p_{i^*} \delta p_{j^*})), \tag{6}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \mathrm{C}_{0,2} \\ 0 & \mathrm{C}_{1,1} & \mathrm{C}_{1,2} \\ \mathrm{C}_{2,0} & \mathrm{C}_{2,1} & \mathrm{C}_{2,2} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \kappa_{0,2} \\ 0 & \kappa_{1,1} & \kappa_{1,2} \\ \kappa_{2,0} & \kappa_{2,1} & \kappa_{2,2} + \kappa_{0,2}^2 + 2\kappa_{1,1}^2 \end{pmatrix},$$

where $C_{J,K}$ represents the multilocus moment, and $\kappa_{J,K}$ the multilocus cumulant.

The frequency of other genotypes is found by replacing p_i by q_i and δp_i by $-\delta p_i$ whenever an allele $X_i = 1$ is replaced by $X_i = 0$ in (1). In general, with n loci, $\mathbf{C}$ is the matrix of scaled moments (with $C_{1,0} \equiv 0$, $C_{0,0} \equiv 1$), and $\mathbf{\Gamma}$ is a vector indexed $0, \ldots, n$, whose ith element is the sum of products of $(n - i)$ distinct allele frequencies and the remaining i distinct δps. The complete matrix of $2^n \times 2^n$ diploid genotypes can be written as $\mathbf{T} \cdot \mathbf{C} \cdot \mathbf{T}^{*\mathrm{T}}$, where $\mathbf{T}$ is the $2^n \times (n + 1)$ matrix, each row being a permutation of the vector $\mathbf{\Gamma}$ defined in (6). For example, with two loci:

$$\mathbf{G} = \begin{pmatrix} g[\{\{0,0\},\{0,0\}\}] & g[\{\{0,0\},\{0,1\}\}] & \cdots & \cdots \\ g[\{\{0,1\},\{0,0\}\}] & g[\{\{0,1\},\{0,1\}\}] & \cdots & \cdots \\ \cdot & \cdot & \cdots & \cdots \\ \cdot & \cdot & \cdots & \cdots \end{pmatrix}$$

$$= \mathbf{T} \cdot \mathbf{C} \cdot \mathbf{T}^{*\mathrm{T}},$$

where

$$\mathbf{T} = \begin{pmatrix} q_i q_j & (\quad q_i \delta p_j \quad \delta p_i q_j) & \delta p_i \delta p_j \\ q_i p_j & (q_i \delta p_j \quad \delta p_i p_j) & \delta p_i \delta p_j \\ p_i q_j & (\quad p_i \delta p_j + \delta p_i q_j) & \delta p_i \delta p_j \\ p_i p_j & (p_i \delta p_j + \delta p_i p_j) & \delta p_i \delta p_j \end{pmatrix}. \tag{7}$$

Expressed in this form, genotype frequencies can be calculated efficiently, since $\mathbf{T}$ depends only on allele frequencies, and $\mathbf{C}$ depends only on the moments or cumulants. Thus, when exploring the statistical fit of alternative moments, $\mathbf{T}$ need not be recalculated.

The model discussed so far has assumed migration from just two source demes. More complex models, with migration from several demes, will retain the same structure only under special conditions. The linkage disequilibrium after migration is a linear sum of the disequilibria contributed by each source. Therefore, the effects of each locus will scale by a factor δp_i only if the deviations in allele frequency of all the sources from the target deme are proportional to this quantity. That is plausible if there is mixing between two taxa, via some network of demes, with the δp_i corresponding to the difference in allele frequency at locus i between the parental taxa. For example, consider a cline. If there is diffusion from neighbouring demes, plus a lower influx from the parental taxa, then associations will scale with δp_i provided that the gradient of the cline at some locus is proportional to the divergence between the taxa — which will be the case for neutral models. In general, however, admixture between large numbers of demes could produce any distribution of genotype frequencies.

A simple method for estimating multilocus moments

Suppose that we wish to estimate the pairwise linkage disequilibria, $C_{0,2}$, $C_{1,1}$ from a set of genotype frequencies; assume for the moment that the divergences are $\delta p_i = 1$. The coefficient $C_{0,2}$ is the usual measure of pairwise linkage disequilibrium between genes derived from the same gamete, whereas $C_{1,1}$ is the measure of the deficit of heterozygotes. One possibility would be to use the variance of an additive trait, $z \equiv \sum_i (X_i + X_i^*)$, where X is the vector giving the states of each gamete. This variance is just $\mathrm{var}(z) = \sum_i \sum_j (C_{\{i,j\}} + C_{\{i,j^*\}} + C_{\{i^*,j\}} + C_{\{i^*,j^*\}}) = 2(\sum_i p_i q_i + n(n \quad 1)C_{0,2} + n^2 \times C_{1,1})$, since $C_{\{i,i\}} = p_i q_i$. The associations within and between genomes can be disentangled by estimating $C_{\{i,i^*\}}$ from the deficit of heterozygotes at individual loci. Because we assume that between-genome associations are the same whether they involve the same or different loci ($C_{\{i,i^*\}} = C_{\{i,j^*\}} = C_{1,1}$), this allows the component of $\mathrm{var}(z)$ attributable to cross-genome associations to be separated from that attributable to associations within genomes (see Barton & Gale, 1993; Kruuk, 1997).

This approach extends to give a simple way of estimating the whole matrix of moments, $C_{j,k}$, from a set of 3^n diploid genotype frequencies. Each locus is described by the value $X_i + X_i^* = 0$, 1 or 2. However, the underlying genotypes {0, 1} and {1, 0} cannot be distinguished in heterozygotes: this makes it impossible to estimate directly all the multilocus moments. Nevertheless, provided that allele frequencies are known, and are the same amongst male and female gametes, two useful quantities can be defined for each locus. First, the additive effect is given by $z_i \equiv \zeta_i + \zeta_i^* = X_i + X_i^* \quad 2p_i$. Secondly, the deviation from Hardy–Weinberg proportions is described by $\phi_i \equiv \zeta_i \zeta_i^*$, which takes values $\{p_i^2, \quad p_i q_i, q_i^2\}$ for $X_i + X_i^* = \{0, 1, 2\}$. Next, find the average value

$M_{a,b}$ of all those terms which contain a factors z_i, and b factors ϕ_i, all with distinct indices: $M_{a,b} \equiv \langle \Pi_{i \in A}(\zeta_i + \zeta_i^*) \Pi_{j \in B}(\zeta_j \zeta_j^*) \rangle$, where A, B are sets of distinct and nonoverlapping indices, and < > indicates an average over all $(n!/a!b!(n \quad a \quad b)!)$ such terms. (For example, with four loci, $M_{2,2} = (z_1 z_2 \ \phi_3 \phi_4 + z_1 \ z_3 \ \phi_2 \phi_4 + \cdots)/6$.) If the genotype frequencies have the form given by the admixture model, with $\delta p = 1$, then the expectation of $M_{a,b}$ is $\mathrm{E}[M_{a,b}] = \sum_{J+K=a} \binom{a}{j} C_{J+b,K+b}$. (Note that because the $M_{a,b}$ have been defined to include only terms with distinct indices, moments such as $\mathrm{C}_{\{i,i\}} = p_i \ q_i$ do not arise; the expression for $\mathrm{E}[M_{a,b}]$ therefore does not include allele frequencies.)

Taking all possible $M_{a,b}$ provides a set of linear equations which uniquely determine all possible $C_{j,k}$. For example, for three loci, the moments $C_{j,k}$ are given in terms of the expectations of $M_{a,b}$ as follows:

$$\mathbf{C} = \mathrm{E}\left[\begin{pmatrix} 1 & 0 & \frac{1}{2}M_{2,0} \quad M_{0,1} & \frac{1}{2}M_{3,0} \quad \frac{3}{2}M_{1,1} \\ 0 & M_{0,1} & \frac{1}{2}M_{1,1} & \frac{1}{2}M_{2,1} \quad M_{0,2} \\ \frac{1}{2}M_{2,0} \quad M_{0,1} & \frac{1}{2}M_{1,1} & M_{0,2} & \frac{1}{2}M_{1,2} \\ \frac{1}{2}M_{3,0} \quad \frac{3}{2}M_{1,1} & \frac{1}{2}M_{2,1} \quad M_{0,2} & \frac{1}{2}M_{1,2} & M_{0,3} \end{pmatrix}\right]. \tag{8}$$

It would be straightforward to find unbiased estimators for the $\mathrm{M}_{a,b}$, based on small samples. However, because the degree of bias is negligible compared with the standard deviation of the estimates, the additional complication is not warranted.

If the δp_i vary across loci, the same method can be used, provided that z_i is scaled relative to δp_i, and ϕ_i relative to δp_i^2. Then, $M_{a,b} \equiv \Pi_{i \in A}(\zeta_i + \zeta_i^*)/\delta p_i \ \Pi_{j \in B}(\zeta_j \zeta_j^*)/\delta p_i^2$. Clearly, this method fails if any of the δp_i are zero. In itself, this poses no difficulty, because such irrelevant loci could be deleted. However, the high weight contributed by loci with weak associations ($\delta p \ll 1$) suggests that the method may be statistically inefficient if there is wide variation in the δp_i.

This method gives satisfactory estimates of individual coefficients. However, it is unsatisfactory as an estimator of the overall genotypic structure, described by the full set of moments, because individual genotype frequencies may be negative. Even if a population has the structure required by the admixture model, samples from that population will vary from that structure, and estimates of allele frequencies and moments made using eqn 8 may not give valid genotype frequencies.

Maximum likelihood estimation

Expressions for the genotype frequencies are linear combinations of the multilocus moments. The log likelihood ($\log(L)$) of the parameters (p_i, δp_i, $\kappa_{j,k}$) is given by summing $\log(g[X])$ over all observed genotypes, X. The maximum likelihood estimate (MLE) is found by maximizing $\log(L)$, subject to the constraint that all $g[X]$ must be non-negative, including those genotypes which did not happen to arise in the sample. This is a nontrivial task, because a very large number of constraints are involved. It can be simplified in two ways. First, allele frequencies can be set to their observed values; this considerably speeds the calculation, because calculating the matrix **T** (a function of the allele frequencies) is otherwise the limiting step. The observed allele frequencies are not quite the same as their MLE in the presence of linkage disequilibrium; indeed, with moderate linkage disequilibrium (e.g. with $m > 0.1$ below), the sampled allele frequencies may be incompatible with the true **C**s that generated the sample, leading to negative genotype frequencies. It will therefore be necessary to fit allele frequencies as well as linkage disequilibria in the simulations below, because we wish to compare with those true values. However, that may not be necessary in making estimates from real data.

Secondly, the δp_i may not need to be estimated. They may be known *a priori*, from knowledge of the source demes. If they are not, then they can be chosen so as to fit the average pairwise association for each locus. The covariance between the diploid genotype (scored as 0, 1 or 2) at loci i and j is $\mathrm{cov}[X_i + X_{i^*}, X_j + X_{j^*}] \equiv \hat{C}_{i,j} = 2(C_{\{i,j\}} + C_{\{i,j^*\}})$, and must equal $\alpha \delta p_i \delta p_j$, where α depends on migration rates, etc. Therefore, the sum of the covariances involving locus i is:

$$\sum_{j \neq i} \hat{C}_{i,j} \equiv \hat{C}_{i,\bullet} = \alpha \delta \pi_\iota (\delta \pi_s \quad \delta \pi_\iota), \psi \theta \varepsilon \sigma \varepsilon \, \delta \pi_s \equiv \sum_\iota \delta \pi_\iota . \tag{9}$$

Given some α, eqn 9 determines the δp_i which will fit the observed $\hat{C}_{i,\bullet}$. The choice of α is arbitrary, because the δp_i can be increased, and the $C_{J,K}$ decreased, so as to leave the predicted linkage disequilibria and genotype frequencies unaltered. One possibility is to suppose that the migration rates are in proportions U:V equal to the mean allele frequencies in the sample, and then take the largest δp_i which still allow valid allele frequencies in the hypothetical source populations (i.e. $0 < p_i - U\delta p_i$, $1 > p_i + \mathrm{V}\delta p_i$). Heterogeneity across loci can be tested by comparing the likelihood of this choice with setting $\delta p_i = 1$.

Once the p_i and δp_i are chosen, the likelihood must be maximized with respect to the $\kappa_{J,K}$, subject to constraints on genotype frequencies. For populations close to Hardy–Weinberg and linkage equilibrium, simple Newton–Raphson maximization is adequate.

However, when linkage disequilibria are strong, this often leads to solutions with negative frequencies of unobserved genotypes. We therefore deal with such cases using the Metropolis algorithm (Metropolis *et al.*, 1954). A random change is made to parameter k; this is drawn from a symmetrical uniform distribution, with maximum value $\pm\Delta_k$. If the new parameter set is valid, and if it increases the likelihood, it is accepted. If it is valid, but decreases the likelihood by a factor $\theta < 1$, then it is accepted with probability $\theta^{1/T}$. The size of random perturbations is optimized by increasing Δ_k slightly if a change in parameter k is accepted, and decreasing it if the change is rejected. This procedure, applied to each of the parameters in turn, generates a random walk with a probability distribution $L^{1/T}$. The parameter T determines how closely this distribution clusters around the optimum (or optima); it is analogous to a temperature, in that with large values of T, the system wanders randomly over a wider range, whereas with small T, it "freezes" to some local optimum. The global optimum can be found by starting at some high T, and gradually cooling to $T = 0$ ("simulated annealing"; Kirkpatrick *et al.*, 1983). Setting $T = 1$ gives a distribution proportional to the likelihood, which can be used to generate support limits on the parameters.

Likelihood provides a natural way of comparing nested hypotheses (Edwards, 1972; Mangel & Hilborn, 1996). One can find in turn the likelihood that the population is in Hardy–Weinberg and linkage equilibrium; that gametes are combined at random, but that there are associations between genes derived from the same gamete, $\kappa_{0,2}$; that there are also higher-order associations within gametes, $\kappa_{0,K}$ ($K > 2$); that there are pairwise associations between genes inherited from different parents, $\kappa_{1,1}$; and finally, that all associations within and between genomes contribute. One might also compare the likelihood that all allele frequencies, p_i, are equal, or that the contribution of each locus to linkage disequilibrium, δp_i, is the same. In order to assess the relative plausibility, one must trade an increase in likelihood against the number of parameters fitted. One approach is to treat the increase in log likelihood obtained by fitting v parameters as a statistic, which approaches a $\frac{1}{2}\chi_v^2$ distribution in large samples. However, this asymptotic result may not be accurate when samples are small, and when estimates are bounded by constraints. It would be possible (though extremely tedious) to find the exact distribution of the likelihood ratio statistic by simulation, or by bootstrap resampling. An alternative approach treats the likelihood itself as the criterion for inference; a plot of log likelihood against the parameters gives a measure of their relative plausibility. A difference in log likelihood of 2 units, which corresponds to one hypothesis being $e^2 = 7.4$ times as likely as the other, can be used as a conventional threshold for acceptance. When hypotheses differ by several degrees of freedom, it is convenient to take thresholds from the $\frac{1}{2}\chi_v^2$ distribution, without applying a significance test as such. An alternative procedure is the Akioke information criterion (Mangel & Hilborn, 1996), under which a model is to be preferred if it has a larger value of $\log(L) - 2v$, which trades each degree of freedom against 1/2 a unit of $\log(L)$. Finally, one could use the Metropolis algorithm, with $T = 1$, to generate a random walk proportional to the likelihood. The marginal distribution of each parameter then gives its likelihood, weighting the other parameters by their likelihood. This procedure amounts to Bayesian inference with uniform prior. The method described here is thus compatible with varied statistical philosophies.

We illustrate this approach using data on the genotypes of 37 toads sampled from the hybrid zone between *Bombina bombina* and *B. variegata* in Croatia (sample 1063 from MacCallum *et al.*, 1998). The toads were scored for four unlinked and diagnostic allozymes (IDH, AK, MDH and LDH). The numbers of each genotype observed are given in Table 1, and are compared with those expected under several hypotheses. The log likelihood that the population is in Hardy–Weinberg and linkage equilibrium is −45.27. On the assumption that all loci are equivalent ($\delta p_i = 1$), a significant improvement is achieved by allowing pairwise associations within genomes ($\log(L) = -36.69$; $\kappa_{0,2} = 0.035$). There is a further significant gain in allowing pairwise associations between genomes, representing a deficit of heterozygotes ($\log(L) = -34.12$; $\kappa_{0,2} = 0.015$, $\kappa_{1,1} = 0.035$). There is little further gain in fitting all the other higher-order associations ($\log(L) = -30.15$, giving an increase in $\log(L)$ of 3.97 for an additional 11 d.f.). Allowing associations to vary across loci, by fitting δp_i, gives a marginal improvement ($\log(L) = -31.31$ with $\kappa_{0,2} = 0.021$, $\kappa_{1,1} = 0.013$; $\delta p = \{0.37, 1.12, 1.18, 2.03\}$). This increase in log (likelihood) of 2.81 is not significant when compared with the asymptotic $\frac{1}{2}\chi_3^2$ distribution ($P = 13\%$). Figure 3 shows how the likelihood depends on $\kappa_{0,2}$ and $\kappa_{1,1}$. The most accurate estimate is of the net covariance between loci ($\kappa_{0,2} + \kappa_{1,1}$); this is reflected in the relative closeness of the contours as both $\kappa_{0,2}$, $\kappa_{1,1}$ increase to top right. The hypothesis that $\kappa_{1,1} = 0$ can be rejected: the contour corresponding to 2 units of log(likelihood) (second down from the peak) does not cross the horizontal axis (see Table 1). However, the hypothesis that $\kappa_{0,2} = 0$ cannot be rejected: that is, gametes might be in linkage equililbrium, provided that there is a strong enough heterozygote deficit.

Table 1 Numbers of each genotype observed in the sample of *Bombina* from site 1063, compared with numbers expected under various hypotheses. The third column gives expectations at Hardy–Weinberg and linkage equilibrium. The next three columns show the MLE for within-genome associations; pairwise associations within and between genomes; and for all orders of association. The last three columns show the same, but with the δp_i allowed to vary across loci. The estimated pairwise associations, log likelihoods and residual degrees of freedom are shown below each column. Genotypes at each locus are represented as homozygotes for *B. bombina* alleles (0); heterozygotes (1); or homozygotes for *B. variegata* alleles (2). In each case, allele frequencies were fitted by maximum likelihood

Genotype	Observed number	HW, LE	$\kappa_{0,2}$	$\kappa_{0,2}, \kappa_{1,1}$	All $\kappa_{j,k}$	$\delta p, \kappa_{0,2}$	$\delta p, \kappa_{0,2}, \kappa_{1,1}$	δp, all $\kappa_{j,k}$
{0, 0, 0, 0}	9	2.02	5.46	6.54	7.77	4.76	5.99	8.65
{1, 0, 0, 0}	4	2.60	3.71	3.07	2.98	4.07	5.64	3.27
{0, 0, 1, 0}	2	1.94	2.18	1.70	1.32	1.76	2.00	1.56
{1, 0, 0, 1}	1	2.34	2.35	1.34	0.65	1.62	0.57	0.61
{0, 1, 1, 0}	1	0.83	0.52	0.70	0.47	0.76	0.62	0.25
{2, 0, 1, 0}	4	0.80	0.59	0.68	0.62	0.53	0.67	0.53
{1, 1, 0, 1}	1	1.00	0.85	0.66	0.86	1.06	0.94	0.90
{1, 0, 1, 1}	1	2.25	2.25	1.23	1.17	2.01	1.60	1.51
{1, 0, 0, 2}	2	0.53	0.31	0.30	0.25	0.14	0.34	0.56
{2, 1, 0, 1}	1	0.32	0.29	0.34	0.52	0.31	0.28	0.35
{1, 1, 1, 1}	1	0.96	1.23	0.88	1.50	1.85	0.99	1.07
{1, 0, 2, 1}	1	0.54	0.42	0.44	0.53	0.38	0.46	0.65
{1, 0, 1, 2}	1	0.51	0.40	0.35	0.40	0.30	0.46	0.75
{0, 1, 2, 1}	2	0.18	0.13	0.19	0.27	0.29	0.25	0.19
{2, 1, 1, 1}	2	0.31	0.47	0.52	0.96	0.61	0.35	0.94
{1, 1, 1, 2}	1	0.22	0.29	0.31	0.56	0.43	0.48	0.86
{0, 2, 2, 1}	1	0.02	0.02	0.05	0.10	0.079	0.05	0.059
{0, 1, 2, 2}	1	0.04	0.04	0.07	0.10	0.14	0.16	0.11
{2, 1, 1, 2}	1	0.07	0.16	0.23	0.21	0.18	0.19	0.23
$\kappa_{0,2}$		0	0.035	0.015	0.035	0.038	0.021	0.033
$\kappa_{1,1}$		0	0	0.035	0.026	0	0.013	0.030
log(L)		−45.27	−36.69	−34.12	−30.15	−34.31	−31.31	−28.44
Residual d.f.		76	75	74	63	72	71	60

Testing against simulated datasets

In order to test these estimation procedures, samples of 100 individuals were taken from a population subject to immigration from two source demes, fixed for alternate alleles ("long-range migration"). Samples were taken after migration, at which point the population would be in linkage disequilibrium and would deviate from Hardy–Weinberg. Table 2 shows the results of fitting various models to 100 replicate datasets, assuming two unlinked loci at $U = 0.2$ or $V = 0.5$, and total migration rates $m = 0$, 0.1, 0.2, 0.3. (Note that with just two loci, the value of δp_i is confounded with the values of the coefficients, and so may arbitrarily be set to 1.) For this problem, convergence of the Metropolis algorithm seems relatively fast. In all cases, parameters were changed 400 times each at $T = 2$, 1 and then 200 times each at $T = 0.5$, 0. Repeating this algorithm usually gave values of log(L) which differed by less than 0.1. In each case, allele frequencies were fitted by maximum likelihood. Initially, allele frequencies were set at their sampled values; if that did not give positive genotype frequencies, then allele frequencies were set equal, at a value which would give positive frequencies. This method sometimes failed to find a valid starting point (see below).

Maximum likelihood estimation of all the cumulants (second row of each table) gave results close to the true values; there is little bias, through $\kappa_{1,1}$ is slightly underestimated for $U = 0.5$, $m = 0.2$, 0.3. The standard deviation of estimates of $\kappa_{0,2}$ averages 0.0252 for $U = 0.5$, and 0.0194 for $U = 0.2$; within-genome associations can thus be detected reliably for low migration rates ($m = 0.1$). The standard deviation of estimates of $\kappa_{1,1}$ averages 0.0182 for $U = 0.5$, and 0.0156 for $U = 0.2$; between-genome associations can therefore be reliably detected for $m = 0.2$ at $U = 0.2$ and 0.5. The standard deviation is expected to scale approximately with the square root of sample size.

The simple method of eqn 8 performs comparably with maximum likelihood estimation: standard deviations are similar, and again, the only evidence of bias is a

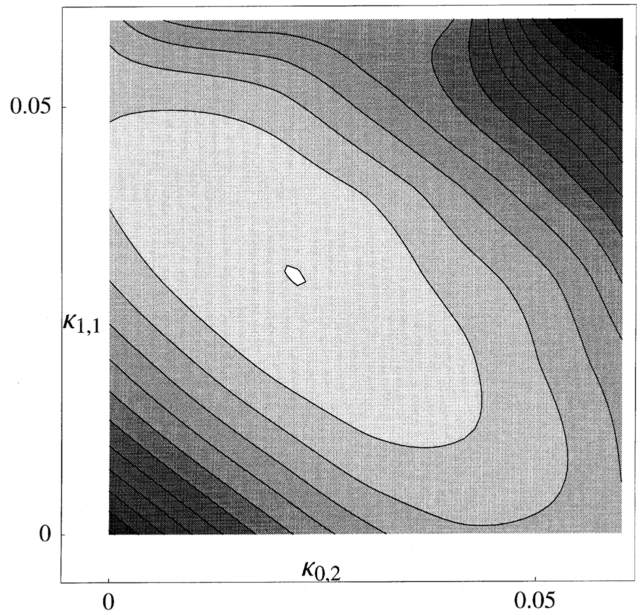

Fig. 3 The log likelihood surface for data from site 1063, plotted as a function of pairwise within- and between-genome association, $\kappa_{0,2}$, $\kappa_{1,1}$. All other associations are set to zero; δp_i are fixed at 1, and allele frequencies are fitted by maximum likelihood. Contours of log likelihood are drawn at unit intervals.

slight underestimation of $\kappa_{1,1}$ for $p = 0.5$ and large m. However, this method does not provide a general method of hypothesis testing, and may not give valid predictions of genotype frequencies: the cumulants estimated in this way, combined with the sampled allele frequencies, often lead to negative genotype frequencies.

Throughout, the distribution of log(L) fits reasonably well with the expected $\frac{1}{2}\chi^2_v$ distribution, which has mean and variance $v/2$ (where v is the number of residual degrees of freedom). For example, one can compare the log likelihood, fitting all coefficients (second row), with that obtained by setting coefficients equal to their true values (fifth row). The difference in mean log likelihood is close to the expected value of 2 (e.g $m = 0.3$, $p = 0.2$; Δlog(L) = 1.82). However, in a few cases (all for $p = 0.2$), fitting more parameters apparently led to a lower likelihood. This occurs when the Metropolis algorithm fails to find the global optimum; in principle, the difficulty could be solved by cooling more slowly from a higher temperature. This problem only arose in 2/800 cases when the likelihoods of the true values of all the cumulants were compared with the likelihood, fitting all the cumulants (fifth vs. second rows in Table 2). The problem was more frequent (30/800 cases) when the likelihood of the pairwise cumulants ($\kappa_{1,1}$, $\kappa_{0,2}$) was compared with the likelihood, fitting all cumulants (fourth vs. second rows in Table 2). This is because here, both likelihoods involve stochastic optimization using the Metropolis algorithm.

Hardy–Weinberg and linkage equilibrium was rejected in all cases for $m > 0.1$; in three cases for $m = 0$, $U = 0.5$; in 79 cases for $m = 0.1$, U and in four, 74 cases for $m = 0$, 0.1, $U = 0.2$ (second vs. first rows; 2 d.f.). Thus, where the population was in fact in Hardy–Weinberg and linkage equilibrium ($m = 0$), the true hypothesis was rejected ~5% of the time, whereas for $m \geq 0.1$, Hardy–Weinberg and linkage equilibrium was rejected in most cases.

The number of replicates in which the true values of the cumulants were rejected, at the 5% level, was close to expectation. The hypothesis that the pairwise associations $\kappa_{1,1}$, $\kappa_{0,2}$ equal their true values was rejected in 7, 5, 8, 5 cases for $m = 0$, 0.1, 0.2, 0.3 and $U = 0.5$, and 3, 8, 5, 3 cases for $U = 0.2$ (second vs. fourth rows; 2 d.f.). The hypothesis that the all cumulants equal their true values was rejected in 3, 5, 7, 5 cases for $m = 0$, 0.1, 0.2, 0.3 and $U = 0.5$, and 4, 5, 3, 4 cases for $U = 0.2$ (second vs. fifth rows; 2 d.f.). This agreement is some what surprising, given that the numbers of most genotypes are small, and the maximum likelihood estimate is often bounded by constraints on genotype frequencies.

Table 3 shows results of simulations with four unlinked loci; the migration rate is $m = 0$ or 0.3, and $U = 0.2$ or 0.5. As before, allele frequencies are fitted by maximum likelihood, and all loci are equivalent (i.e. $\delta p = 1$). (The case where pairwise associations were fixed at their true values and the remainder fitted is not now shown.) The simple method of eqn 8 (third row of Table 3) performs well, in that the mean of estimates of $\kappa_{1,1}$, $\kappa_{0,2}$ is close to the true value, and the standard deviation across replicates is similar to that of maximum likelihood estimates. However, the matrix of moments estimated using eqn 8 usually gives negative genotype frequencies when combined with sample allele frequencies; this reflects the strong constraints on the moments with four loci. With high migration ($m = 0.3$), maximum likelihood estimates of $\kappa_{1,1}$, $\kappa_{0,2}$ are on average ~10% lower than the true values; with no migration, there is a small positive bias for $\kappa_{0,2}$ (0.0136 for $U = 0.2$, 0.0094 for $U = 0.5$); a small positive bias for $\kappa_{1,1}$ with $U = 0.2$ (+0.0041); and a small negative bias for $\kappa_{1,1}$ with $U = 0.5$ (−0.0051).

Likelihoods now agree poorly with asymptotic theory; this is to be expected, as there are now stronger constraints on the parameters, and because the genotype frequencies are lower. The mean residual log(L) are now substantially smaller than the expected value, $v/2$ (first column of Table 3). Comparison of hypotheses also shows smaller differences in log(likelihood) than expected for $m = 0.3$. For example, with $U = 0.5$, the improvement in log(likelihood) obtained by fitting 13 cumulants

Table 2 Each table shows results from 100 replicate samples of 100 individuals. These were taken after migration from populations with the same allele frequencies at two unlinked loci ($p_1 = p_2 = 0.2$ or 0.5). Immigration is from two source demes, fixed for alternate alleles ($\delta p_1 = \delta p_2 = 1$); the total rate of migration is $m = 0.1, 0.2, 0.3$. Populations with $m = 0$ are in linkage equilibrium and Hardy–Weinberg proportions. For each of the eight sets of replicates, several alternative models were fitted. Log likelihoods were calculated for each, relative to a perfect fit; the mean and variance of these are shown, together with the number of residual degrees of freedom. Asymptotically, log(L) should follow a $\frac{1}{2}\chi_v^2$ distribution, with mean and variance $v/2$. The mean (SD) of each coefficient is shown; this was either fitted by maximum likelihood, estimated from the sample (eqn 8); or fixed at its true value. HW, LE denotes the null model of linkage equilibrium and Hardy–Weinberg proportions. The next three rows show estimates fitting all cumulants, $\kappa_{J,K}$; setting all cumulants to their values estimated from the sample by eqn 8; fixing pairwise cumulants to their true values, but fitting the remainder; and fixing all cumulants at their true values. The last row gives the true values in the population from which the replicates were sampled. In all cases, allele frequencies were fitted by maximum likelihood. In some cases, no valid starting point could be found: the number of such cases is listed, and these were excluded from the analysis

	log(L)											
	Mean	Variance	v	No. invalid	$\kappa_{0,2}$	(SD)	$\kappa_{1,1}$	(SD)	$\kappa_{1,2}$	(SD)	$\kappa_{2,2}$	(SD)
$m = 0$, $U = 0.2$												
HW, LE	−3.15	2.87	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.43	1.04	2	0	0.0024	(0.0178)	−0.0012	(0.0103)	−0.0002	(0.0033)	−0.0005	(0.0029)
Sample $\kappa_{J,K}$	−2.01	7.98	2	0	−0.0001	(0.0210)	−0.0007	(0.0118)	−0.0003	(0.0030)	−0.0008	(0.0024)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.32	1.96	4	0	0		0		0.0002	(0.0033)	0.0004	(0.0030)
True $\kappa_{J,K}$	−3.15	2.87	6	0	0		0		0		0	
True model					0		0		0		0	
$m = 0.1$, $U = 0.2$												
HW, LE	−10.04	27.68	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.27	1.17	2	0	0.0420	(0.0204)	0.0166	(0.0154)	0.0091	(0.0080)	0.0057	(0.0050)
Sample $\kappa_{J,K}$	−1.28	1.33	2	0	0.0414	(0.0199)	0.0163	(0.0153)	0.0087	(0.0073)	0.0056	(0.0046)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.39	2.46	4	0	0.0422		0.0160		0.0092	(0.0051)	0.0069	(0.0045)
True $\kappa_{J,K}$	−3.39	3.07	6	0	0.0422		0.0160		0.0096		0.0068	
True model					0.0422		0.0160		0.0096		0.0068	
$m = 0.2$, $U = 0.2$												
HW, LE	−21.26	64.94	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.36	0.82	2	0	0.0754	(0.0195)	0.0291	(0.0181)	0.0171	(0.0100)	0.0085	(0.0057)
Sample $\kappa_{J,K}$	−1.37	0.95	2	1	0.0745	(0.0190)	0.0277	(0.0160)	0.0167	(0.0092)	0.0088	(0.0041)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.30	2.18	4	0	0.0747		0.0320		0.0189	(0.0049)	0.0106	(0.0047)
True $\kappa_{J,K}$	−3.33	2.86	6	0	0.0747		0.0320		0.0192		0.0113	
True model					0.0747		0.0320		0.0192		0.0113	
$m = 0.3$, $U = 0.2$												
HW, LE	−37.61	113.42	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.34	0.71	2	0	0.1023	(0.0200)	0.0466	(0.0188)	0.0276	(0.0103)	0.0120	(0.0055)
Sample $\kappa_{J,K}$	−1.19	0.84	2	0	0.1013	(0.0187)	0.0457	(0.0184)	0.0278	(0.0103)	0.0126	(0.0055)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.13	1.11	4	0	0.0997		0.0480		0.0293	(0.0055)	0.0142	(0.0052)
True $\kappa_{J,K}$	−3.16	1.92	6	0	0.0997		0.0480		0.0288		0.0142	
True model					0.0997		0.0480		0.0288		0.0142	

m = 0, U = 0.5												
HW, LE	−2.99	2.85	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−0.86	0.54	2	0	0.0017	(0.0309)	−0.0004	(0.0180)	0.0010	(0.0064)	−0.0014	(0.0067)
Sample $\kappa_{J,K}$	−0.86	0.55	2	3	0.0014	(0.0315)	−0.0002	(0.0182)	0.0010	(0.0063)	−0.0015	(0.0067)
True $\kappa_{0,2}, \kappa_{1,1}$	−1.86	1.33	4	0	0.0000		0.0000		0.0010	(0.0065)	0.0001	(0.0066)
True $\kappa_{J,K}$	−2.99	2.85	6	0	0.0000		0.0000		0.0000		0.0000	
True model					0		0		0		0	
m = 0.1, U = 0.5												
HW, LE	−9.69	22.60	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.03	0.84	2	0	0.0653	(0.0273)	0.0226	(0.0184)	−0.0007	(0.0063)	0.0018	(0.0063)
Sample $\kappa_{J,K}$	−1.04	0.84	2	0	0.0650	(0.0270)	0.0228	(0.0185)	−0.0007	(0.0062)	0.0018	(0.0061)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.06	1.94	4	0	0.0659		0.0250		−0.0008	(0.0063)	0.0031	(0.0062)
True $\kappa_{J,K}$	−3.03	2.61	6	0	0.0659		0.0250		0.0000		0.0025	
True model					0.0659		0.0250		0.0000		0.0025	
m = 0.2, U = 0.5												
HW, LE	−25.26	61.41	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.14	1.35	2	0	0.1181	(0.0231)	0.0459	(0.0191)	−0.0002	(0.0067)	−0.0018	(0.0068)
Sample $\kappa_{J,K}$	−1.15	1.36	2	0	0.1180	(0.0230)	0.0460	(0.0191)	−0.0001	(0.0068)	−0.0016	(0.0066)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.23	2.68	4	0	0.1167		0.0500		−0.0002	(0.0068)	−0.0006	(0.0059)
True $\kappa_{J,K}$	−3.33	4.02	6	0	0.1167		0.0500		0.0000		−0.0006	
True model					0.1167		0.0500		0.0000		−0.0006	
m = 0.3, U = 0.5												
HW, LE	−43.06	77.88	6	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−1.11	1.03	2	0	0.1534	(0.0194)	0.0692	(0.0173)	−0.0013	(0.0077)	−0.0072	(0.0070)
Sample $\kappa_{J,K}$	−1.09	1.11	2	3	0.1537	(0.0198)	0.0685	(0.0175)	−0.0011	(0.0077)	−0.0068	(0.0065)
True $\kappa_{0,2}, \kappa_{1,1}$	−2.06	1.99	4	0	0.1558		0.0750		−0.0011	(0.0080)	−0.0077	(0.0052)
True $\kappa_{J,K}$	−3.24	2.61	6	0	0.1558		0.0750		0.0000		−0.0074	
True model					0.1558		0.0750		0.0000		−0.0074	

Table 3 Results from simulations with four unlinked loci, presented as in Table 2

	log(L)											
	Mean	Variance	v	No. invalid	$\kappa_{0,2}$	(SD)	$\kappa_{1,1}$	(SD)	$\kappa_{1,2}$	(SD)	$\kappa_{2,2}$	(SD)
m = 0, U = 0.2												
HW, LE	−23.22	19.09	76	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−22.30	16.29	63	0	0.0136	(0.0103)	0.0041	(0.0083)	0.0032	(0.0022)	0.0017	(0.0012)
Sample $\kappa_{J,K}$	−49.57	1118.41	63	37	0.0012	(0.0105)	−0.0013	(0.0093)	−0.0002	(0.0013)	−0.0001	(0.0010)
True $\kappa_{J,K}$	−23.22	19.09	76	0	0.0000		0.0000		0.0000		0.0000	
True model					0.0000		0.0000		0.0000		0.0000	
m = 0.3, U = 0.2												
HW, LE	−129.46	655.94	76	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−17.82	13.16	63	0	0.0894	(0.0172)	0.0419	(0.0150)	0.0198	(0.0074)	0.0081	(0.0032)
Sample $\kappa_{J,K}$	−20.24	9.50	63	96	0.0989	(0.0183)	0.0444	(0.0167)	0.0270	(0.0087)	0.0128	(0.0041)
True $\kappa_{J,K}$	−19.71	17.15	76	0	0.0997		0.0480		0.0288		0.0142	
True model					0.0997		0.0480		0.0288		0.0142	
m = 0, U = 0.5												
HW, LE	−42.46	32.02	76	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−36.91	26.73	63	0	0.0094	(0.0124)	−0.0051	(0.0095)	0.0006	(0.0024)	0.0005	(0.0019)
Sample $\kappa_{J,K}$	−37.91	31.70	63	89	0.0002	(0.0150)	0.0003	(0.0112)	0.0007	(0.0026)	−0.0003	(0.0026)
True $\kappa_{J,K}$	−42.46	32.02	76	0	0.0000		0.0000		0.0000		0.0000	
True model					0.0000		0.0000		0.0000		0.0000	
m = 0.3, U = 0.5												
HW, LE	−172.23	433.71	76	0	0		0		0		0	
Fit all $\kappa_{J,K}$	−25.94	17.78	63	0	0.1446	(0.0136)	0.0688	(0.0171)	0.0002	(0.0050)	−0.0048	(0.0045)
Sample $\kappa_{J,K}$	−25.89	17.64	63	84	0.1557	(0.0129)	0.0740	(0.0176)	−0.0002	(0.0061)	−0.0080	(0.0053)
True $\kappa_{J,K}$	−29.82	20.33	76	0	0.1558		0.0750		0.0000		−0.0074	
True model					0.1558		0.0750		0.0000		−0.0074	

is on average 3.89, with variance 6.35; this compares with the asymptotic expectation of 6.5 for each. Examination of the cumulative distribution of $\Delta\log(L)$ shows that it is shifted to the left by ~2.6 units for $U = 0.5$, $m = 0.3$, and ~4.2 units for $U = 0.2$, $m = 0.3$. Correspondingly, the true matrix of cumulants is rejected less often than expected from asymptotic theory: at $P = 5\%$, it is rejected 0/100, 1/100 times at $U = 0.2$, 0.5 and $m = 0.3$. For populations in linkage equilibrium ($m = 0$), and $U = 0.5$, agreement is better: the mean and variance of $\Delta\log(L)$ are 5.55, 5.52, close to the expectation of 6.5. With $U = 0.2$, the mean and variance of $\Delta\log(L)$ are 0.93, 6.78; the mean is again lower than the expectation of 6.5. The true cumulants are rejected in 4/100 cases for $U = 0.5$, and in 0/100 cases for $U = 0.2$.

Discussion

If linkage disequilibria are generated by the mixing of two source populations, either directly or across a cline, they take a simple form: the magnitude of the association between a set of loci depends primarily on the number of each kind of allele derived from the maternal genome, and the number derived from the paternal genome (J, K, say). The state of the population can therefore be described by the allele frequencies, and by a matrix giving the various associations. These associations can be described either in terms of multilocus moments, $C_{j,k}$, or multilocus cumulants, $\kappa_{j,k}$ ($j = 0\ldots\mathrm{n}$, $k = 0\ldots\mathrm{n}$). There is a simple relation between the matrix of 2^{2^n} diploid genotype frequencies, $\mathbf{G}$, and the matrix of moments, $\mathbf{C} : \mathbf{G} = \mathbf{\Gamma} \cdot \mathbf{C} \cdot \mathbf{\Gamma}^{\mathrm{T}}$, where $\mathbf{\Gamma}$ is a matrix which depends only on the allele frequencies (eqn 7). This method extends to allow for different allele frequencies in the male and female gamete pools, and for different degrees of divergence across loci.

Although the method described here is motivated by models of neutral admixture, it may apply to other cases in which different biallelic loci are equivalent: for example, where selection acts on an additive trait determined by unlinked loci. If all loci are interchangeable, such that all genotypes with the same numbers of '+' alleles contributed by the maternal and paternal gametes (n,n*) are equally frequent, then the population can be described by the joint distribution of (n, n*). This symmetrical polygenic model has been investigated by Kondrashov (1984), Barton (1992), Doebeli (1996), Shpak & Kondrashov (1999) and Barton & Shpak (2000b); the diploid case has been treated by Kondrashov & Kondrashov (1999). This description is equivalent to that set out here in terms of a matrix of moments or cumulants. However, current theory is restricted to cases where loci are fully interchangeable. It may be that variation across loci (for example, in their effects on an additive trait) might be described in the same way as variation in the divergence, δp, in admixture models. It may also be that the full model could be approximated by a description in terms of pairwise moments or cumulants (c.f. Turelli & Barton, 1994). Such possibilities warrant further investigation.

The matrix of moments, $C_{J,K}$ can be calculated as a linear combination of various products of additive and dominance effects associated with each locus (eqn 8). This method performs as well as maximum likelihood estimation, and is much faster; it also avoids the uncertainties of the Monte Carlo algorithm used here to maximize likelihood. However, this simple method does not allow the plausibility of different hypotheses to be compared, and more seriously, often leads to sets of allele frequencies and moments that predict negative genotype frequencies. In contrast, maximum likelihood estimation allows flexible comparison of nested hypotheses, and necessarily gives estimates consistent with positive genotype frequencies.

One motivation for this work was the need to combine information about pairwise linkage disequilibria across loci. When linkage disequilibria are weak, and samples are large, estimates of pairwise linkage disequilibrium should be approximately independent, even when pairs of loci overlap (e.g. $\mathrm{E}[C_{\{i,j\}}\ \mathrm{C}_{\{\mathrm{i,k}\}}] = 0$). The standard error of an estimate of the average pairwise disequilibrium should then decrease with the square root of the number of pairs of loci involved. However, with strong linkage disequilibria, estimates become strongly correlated, and so the standard error of estimates of pairwise associations should decrease more slowly with the number of loci. Figure 4 shows the standard deviation of estimates of $\kappa_{1,1}$, $\kappa_{0,2}$ as a function of m,

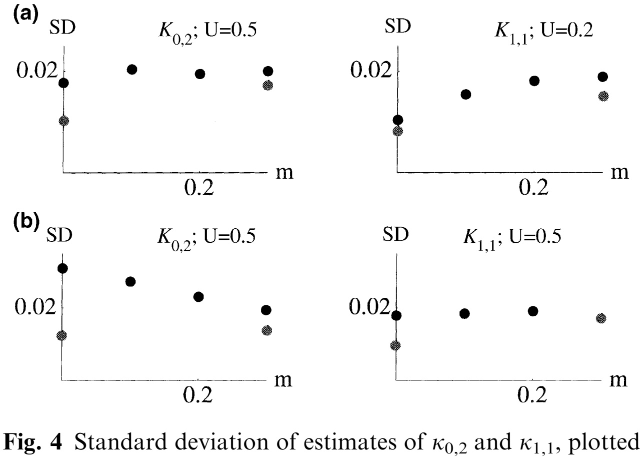

Fig. 4 Standard deviation of estimates of $\kappa_{0,2}$ and $\kappa_{1,1}$, plotted against migration rate, m. The lower grey circles show results for four loci, the upper solid circles for two loci. Each point is calculated from 100 replicate samples of 100 individuals (Tables 2 and 3). (a) $U = 0.2$, (b) $U = 0.5$.

for two and four loci, and for $U = 0.2$, 0.5. At linkage equilibrium ($m = 0$; left of Fig. 4), the standard error is on average smaller by a factor 0.57 with four loci, compared with two loci; this is somewhat higher than $1/\sqrt{6} = 0.41$, the factor expected if the six pairwise associations with four loci contribute independent information. In contrast, the standard error with $m = 0.3$ (i.e. with strong linkage disequilibrium) is reduced only by a factor 0.84 as the number of loci increases from two to four.

The methods described here give estimates of associations of all orders. In principle, higher-order associations can give additional information: for example, short- and long-range migration can be distinguished by the rate at which associations decrease with the number of loci involved (see above). However, estimates of higher-order linkage disequilibria have high sampling errors. With two loci (Table 2) the third- and fourth-order cumulants $\kappa_{1,2}$, $\kappa_{2,2}$ are small for $U = 0.5$, and so cannot be distinguished from zero with samples of 100 individuals. When allele frequencies are asymmetric ($U = 0.2$), higher-order associations are stronger, and become detectable for $m > 0.2$ (Table 2). A more fundamental problem is that the values of higher-order associations are strongly constrained by the requirement that genotype frequencies be positive, and therefore depend on both allele frequencies and on pairwise associations. It therefore makes little sense to make separate estimates of associations of different orders. If possible, an evolutionary model described by parameters such as migration rates should be fitted.

Often, data from hybrid populations are analysed by classifying each individual on the basis of its genotype (see Arnold, 1997). Where hybridization is rare, so that only a few distinct classes of hybrid are present, this is appropriate. However, there are several difficulties with this approach. First, loci may not be diagnostic (i.e. fixed for alternative alleles in different taxa). Even a low level of polymorphism makes it difficult to distinguish between hybrids and parental genotypes. Secondly, even if loci are initially diagnostic, continued backcrossing into a population causes polymorphism to build up within the native gene pool, making it impossible to distinguish recent hybrids merely by the presence of marker alleles. Recent hybrids can be distinguished by the presence of multiple alleles typical of the immigrant population — in other words, by linkage disequilibrium between introgressing alleles (Goodman *et al.*, 1999). If hybrids always cross with mates from a large native population, then individuals could in principle be classified by backcross generation. The kth backcross would carry a fraction 2^{-k} of introgressed alleles; if this fraction is higher than the background level of polymorphism, and if enough loci are scored, then the backcross generation could be inferred. However, such assignment is inaccurate even with many loci (Boecklen & Howard, 1997), and is impossible when hybrids mate with each other to produce F_2s or complex backcrosses. Because the proportion of backcross hybrids in a population increases with time, t, as 2^t, matings between hybrids soon become likely.

An alternative approach which is often used is to classify individuals by the fraction of their alleles which derive from each parental taxon, rather than attempting to infer their detailed ancestry (Arnold, 1997). Although this is a convenient and simple way of summarizing multilocus data, the method may be misleading if loci are not strictly diagnostic. Even if marker loci are diagnostic, much information is lost. An improvement would be to classify individuals according to both the number of alleles derived from one reference taxon, and by the number of loci that are heterozygous. This is equivalent to a description in terms of a matrix of moments, estimated using eqn 8. However, eqn 8 is more flexible in that it can allow for nondiagnostic loci through variations in allele frequencies and divergences, δp.

The methods described here give a flexible framework for describing hybrid populations in terms of linkage disequilibria of various orders. The strengths of pairwise linkage disequilibria can then be used to infer quantities such as rates of gene flow and the degree of assortative meeting. The likelihood of the full set of genotype frequencies can be calculated, which allows more elaborate hypotheses to be addressed. For example, one could find the likelihood that offspring were sired by sampled individuals of known genotype, rather than by some farther from the unsampled population. These methods make possible a variety of statistical analyses of hybrid populations.

Acknowledgements

This work was supported by grant MMI09726 from the BBSRC/EPSRC, and by the Darwin Trust of Edinburgh. I am grateful to W. G. Hill, L. Kruuk and M. Orive, and to the referees, for their helpful comments on the manuscript.

References

ARNOLD, M. 1997. *Natural Hybridization and Introgression*. Princeton University Press, Princeton, NJ.

ASMUSSEN, M. A., ARNOLD, J. AND AVISE, J. C. 1987. Definition and properties of disequilibrium statistics for associations between nuclear and cytoplasmic genotypes. *Genetics*, **115**, 755–768.

BARTON, N. H. 1986. The effects of linkage and density-dependent regulation on gene flow. *Heredity*, **57**, 415–426.

BARTON, N. H. 1992. On the spread of new gene combinations in the third phase of Wright's shifting balance. *Evolution*, **46**, 551–557.

BARTON, N. H. AND GALE, K. S. 1993. Genetic analysis of hybrid zones. In: Harrison, R. G. (ed.) *Hybrid Zones and the Evolutionary Process*, pp. 13–45. Oxford University Press, Oxford.

BARTON, N. H. AND SHPAK, M. 2000a. The effects of epistasis on the structure of hybrid zones. *Genet. Res.* (in press).

BARTON, N. H. AND SHPAK, M. 2000b. The stability of symmetrical polygenic models. *Theor. Pop. Biol.* (in press).

BARTON, N. H. AND TURELLI, M. 1991. Natural and sexual selection on many loci. *Genetics*, **127**, 229–255.

BOECKLEN, W. J. AND HOWARD, D. J. 1997. Genetic analysis of hybrid zones: number of markers and power of resolution. *Ecology*, **78**, 2611–2616.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, **39**, 1–38.

DOEBELI, M. 1996. A quantitative genetic competition model for sympatric speciation. *J. Evol. Biol.*, **9**, 893–910.

EDWARDS, A. W. F. 1972. *Likelihood.* Cambridge University Press, Cambridge.

GOODMAN, S. J., BARTON, N. H., SWANSON, G. ABERNETHY, K. AND PEMBERTON, J. M. 1999. Introgression through rare hybridisation: a genetic study of a hybrid zone between red and sika deer (genus *Cervus*), in Argyll, Scotland. *Genetics*, **152**, 355–371.

GUIASU, S. AND SHENITZER, A. 1985. The principle of maximum entropy. *Math. Intelligencer*, **7**, 42–48.

HABER, M. 1984. Log-linear models for linked loci. *Biometrics*, **40**, 189–198.

HILL, W. G. 1974a. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.

HILL, W. G. 1974b. Disequilibria among several linked neutral genes in finite populations I. Mean change in disequilibria. *Theor. Pop. Biol.*, **5**, 366–392.

HILL, W. G. 1974c. Disequilibria among several linked neutral genes in finite populations II. Variances and covariances of disequilibria. *Theor. Pop. Biol.*, **6**, 143–148.

HILL, W. G. 1975. Tests for association of gene frequencies at several loci in random mating diploid populations. *Biometrics*, **51**, 881–888.

HILL, W. G. AND WEIR, B. S. 1994. Maximum likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.*, **54**, 705–714.

KAPLAN, N. L., HILL, W. G. AND WEIR, B. S. 1995. Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.*, **56**, 18–32.

KIRKPATRICK, S., GELATT, C. D. AND VECCHI, M. P. 1983. Optimization by simulated annealing. *Science*, **220**, 671–680.

KONDRASHOV, A. S. 1984. On the intensity of selection for reproductive isolation at the beginnings of sympatric speciation. *Genetika*, **20**, 408–415.

KONDRASHOV, A. S. AND KONDRASHOV, F. A. 1999. Interactions among quantitative traits in the course of sympatric speciation. *Nature*, **400**, 351–354.

KRUUK, L. E. B. 1997. *Barriers to Gene Flow: A* Bombina *(Fire-bellied Toad) Hybrid Zone and Multilocus Cline Theory*. Ph.D. Thesis, University of Edinburgh.

KRUUK, L. E. B., BAIRD, S. J. E., GALE, K. S. AND BARTON, N. H. 1999. The effect of endogenous and exogenous selection on multilocus clines. *Genetics*, **153**, 1959–1971.

LANGLEY, C. H. 1977. Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster*. In: Christiansen, F. B. and Fenchel, T. M. (eds) *Measuring Selection in Natural Populations*, pp. 265–273. Springer Verlag, Berlin.

LI, W. H. AND NEI, M. 1974. Stable linkage disequilibrium without epistasis in subdivided populations. *Theor. Pop. Biol.*, **6**, 173–183.

LONG, J. C., WILLIAMS, R. C. AND URBANEK, M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.*, **56**, 799–810.

MACCALLUM, C. J., NURNBERGER, B., BARTON, N. H. AND SZYMURA J. M. 1998. Habitat preference in the *Bombina* hybrid zone in Croatia. *Evolution*, **52**, 227–239.

MANGEL, M. AND HILBORN, R. 1996. *The Ecological Detective*. Princeton University Press, Princeton, NJ.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. AND TELLER, E. 1954. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1095.

PHIPPS, T. E. AND BRILL, M. H. 1995. Bayesian entropy and inference. *Physics Essays*, **8**, 615–625.

SHPAK, M. AND KONDRASHOV, A. S. 1999. Applicability of the hypergeometric phenotypic model to haploid and diploid populations. *Evolution*, **53**, 600–604.

SLATKIN, M. AND EXCOFFIER, L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity*, **76**, 377–383.

SOBER, E. 1983. Parsimony in systematics: philosophical issues. *Ann. Rev. Ecol. Syst.*, **14**, 335–358.

TURELLI, M. AND BARTON, N. H. 1994. Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? *Genetics*, **138**, 913–941.