

# Multilocus estimation of genetic structure within populations

Bryan K. Epperson\*

*126 Natural Resources Building, Michigan State University, East Lansing, MI 48824, USA*

Received 24 March 2003

## Abstract

Spatial structure of genetic variation within populations is well measured by statistics based on the distribution of pairs of individual genotypes, and various such statistics have been widely used in experimental studies. However, the problem of uncharacterized correlations among statistics for different alleles has limited the applications of multiallelic, multilocus summary measures, since these had unknown sampling distributions. Usually multiple alleles and/or multiple loci are required in order to precisely measure spatial structures, and to provide precise indirect estimates of the amount of dispersal in samples of reasonable size. This article examines the correlations among pair-wise statistics, including Moran  $I$ -statistics and various measures of conditional kinship, for different alleles of a locus. First the correlations are mathematically derived for random spatial distributions, which allow averages over alleles and loci to be used as more powerful yet exact test statistics for the null hypothesis. Then extensive computer simulations are conducted to examine the correlations among values for different alleles under isolation by distance processes. For loci with more than three alleles, the results show that the correlations are remarkably and perhaps surprisingly small, establishing the principle that then alleles behave as nearly independent realizations of space–time stochastic processes. The results also show that the correlations are largely robust with respect to the degree of spatial structure, and they can be used in a straightforward manner to form confidence intervals for averages. The results allow a precise connection between observations in experimental studies and levels of dispersal in theoretical models.

© 2004 Elsevier Inc. All rights reserved.

*Keywords:* Spatial autocorrelation; Spatial structure; Dispersal; Isolation by distance; Multilocus

## 1. Introduction

The spatial distribution of genetic variation within a population influences many processes of population genetics, including biparental inbreeding and the operation of some forms of natural selection. Experimental studies statistically characterize spatial structure of genetic variation, because of its inherent interest as well as to detect natural selection by contrasting spatial patterns for different genetic loci (Epperson and Clegg, 1986), and to indirectly measure levels of dispersal standardized by density (Chung et al., 1998). Several dozen studies have employed a particularly popular method based on Moran's  $I$ -statistic, calculated for pairs of genotypes converted into allele frequencies (reviewed in Heywood, 1991; Epperson, 2003a). Many studies have used various other pair-wise measures of

conditional kinship that are closely related to and often inter-convertible with  $I$ -statistics.

Pair-wise measures appear to capture most of the information in spatial patterns of genetic variation (Barton and Wilson, 1995). Moreover, the distribution of Moran's (1950)  $I$ -statistic is known under the null hypothesis of a random spatial distribution (Cliff and Ord, 1981). Under the sample-based assumption of sampling pairs of observed genotypes (without replacement), the distribution depends on the frequencies of genotypes in samples, not those in the population. The latter may depend on a stochastic process and are usually not known. Under the same null hypothesis the distributions of conditional kinship measures can be determined by using  $I$ -statistics and the fixation index (Epperson, 2003a).

A number of studies have characterized many of the properties of pair-wise statistics, usually directly in terms of  $I$ -statistics, in several stochastic processes, including isolation by distance (e.g., Sokal and Wartenberg, 1983; Doligez et al., 1998), with added selection

\*Fax: +1-517-432-1143.

E-mail address: [epperson@msu.edu](mailto:epperson@msu.edu).

(e.g. Sokal et al., 1989; Epperson, 1990), and properties have also been studied in several deterministic processes (e.g., Sokal, 1979). The sampling properties of  $I$ -statistics also have been thoroughly characterized under some of these processes (e.g., Sokal et al., 1997). Values observed in samples can be compared to those in isolation by distance models, and hence used to estimate dispersal (Epperson et al., 1999).

Both theoretical and experimental studies have shown that precise inferences usually require data for multiple alleles and/or multiple loci. When spatially distributed variables represent a spatial cross-section of stochastic space–time processes, they are subject to substantial stochastic noise. Averaging over multiple genetic variables reduces both stochastic variations and statistical noise in estimators. The problem of averaging over alleles of the same locus has been troublesome. Only the case for a locus with two alleles has been straightforward, because then the  $I$ -statistics for both are identical (Heywood, 1991). For a locus with more than two alleles, it has been widely recognized that the values are correlated, but the strengths of correlations were unknown. Averaging over loci per se does not pose any particular difficulty, as long as it is reasonable to expect that there is little or no linkage disequilibrium among loci; however, this is usually done either together with or subsequent to averaging over alleles. Thus the lack of knowledge of the correlations among values for different alleles has limited the utility or efficiency of multiallelic, multilocus summary statistics based on averages of  $I$ -statistics or conditional kinship.

Among the large number of experimental studies using  $I$ -statistics a wide range of approaches to the problem of correlations have been taken, including (with examples): 1. using all alleles for the average, but do no formal testing of the average (Peakall and Beattie, 1995); 2. discarding the least common allele for each locus—sometimes claiming that the  $I$ -statistics for the remaining alleles are independent (Campbell and Dooley, 1992); 3. using only the most common allele (Bacilieri et al., 1994), or otherwise using only one allele per locus (Maki and Masuda, 1993); 4. discarding the most common allele and using all remaining ones (Knowles, 1991); and 5. using mixtures of the foregoing. Still other studies have used different measures, most often multi-locus forms of conditional kinship measures, and re-sampling methods.

It should be noted that empirical measures of kinship are not necessarily the same as kinship measures in mathematical models of stochastic processes (Malécot, 1955; Harpending, 1973; Morton, 1973a). Various empirical measures include several developed by Morton (1973a, b, 1982), multiple-locus and multilocus forms of the measure of Cockerham (1969), which has been re-examined recently (Loiselle et al., 1995; Rousset, 2000; Hardy and Vekemans, 1999), and the sophisticated

measure of Smouse and Peakall (1999). Still other measures are “analysis of molecular variance”-type estimators designed for modern molecular data (e.g. Bertorelle and Barbujani, 1995), as well as measures based on the numbers of alleles in common (Berg and Hamrick, 1995) in multilocus genotypes of pairs of individuals. However, there does not appear to be significantly greater information in say DNA sequence data compared to that contained in allelic state, with respect to spatial distributions of genotypes within populations (Barton and Wilson, 1995).

Truly multilocus genotypic estimators average first over alleles and loci and then (at least for spatial analyses) over pairs of individuals, but for other summary statistics, such as averages of Moran’s  $I$  or averages of single allele measures of conditional kinship, averaging is first over pairs of individuals. Both approaches use essentially the same information unless there is linkage disequilibrium. If there is substantial linkage disequilibrium, then the multilocus kinship estimators will be affected and difficult to characterize, because there almost no theoretical results for spatially explicit multilocus distributions. The one two-locus study indicated that there is likely to be little disequilibrium at most spatial scales, even if loci are tightly linked (Epperson, 1995a).

Most of these pair-wise statistics can be derived from Moran’s  $I$ , using estimates of Wright’s (1965) fixation index (Barbujani, 1987; Hardy and Vekemans, 1999; Rousset, 2000). The sampling distribution theory was originally developed by Cliff and Ord (1973) for Moran’s (1950)  $I$ -statistics, thus providing a starting point. Results can be first developed for Moran’s  $I$ -statistics and then extended to the other kinship measures. Moran’s  $I$ -statistic for a spatial distribution of diploid genotypes is typically calculated as follows. For each allele  $A$ , for each individual  $i$ , all genotypes are converted into gene frequencies. Homozygotes  $AA$  are assigned the value  $X_i = 1.0$ , heterozygotes for  $A$  are assigned 0.5, and all other genotypes have value zero. Then the grand mean (also equal to the allele frequency in the sample),  $X$ , is subtracted,  $Z_i = X_i - X$ . In other words,  $Z_i$  is the mean-adjusted frequency of gene  $A$  in the  $i$ th spatially located sample individual. Moran’s  $I$ -statistic is written:

$$I = \frac{n \sum_i \sum_j w_{ij} Z_i Z_j}{W \sum_i Z_i^2}, \quad (1)$$

where  $n$  is the number of sampled individuals, each  $w_{ij}$  is a spatial proximity weight defined for the spatial locations of individuals  $i$  and  $j$ , and  $W$  is the sum of all weights (Cliff and Ord, 1981). Most commonly, weights are chosen in order to form mutually exclusive and exhaustive distance classes of pairs, or “joins,” of individual locations. Then, for a specific distance

class  $k$ :

$$I(k) = \frac{n \sum_i \sum_j \delta_{ij}(k) Z_i Z_j}{W_k \sum_i Z_i^2}, \quad (2)$$

where the  $\delta_{ij}(k)$  indicate the inclusion (1) or exclusion (0) of the pair (join) of individuals  $i$  and  $j$  in distance class  $k$  (Sokal and Oden, 1978; Cliff and Ord, 1981). One may focus on statistics for distance classes (Eq. (2)), but recognize that the results can be extended to statistics weighted by measures of spatial proximity (Eq. (1)).

The distribution of  $I$  can be complicated by the fact that it is a ratio, but it has been determined under two forms of null hypotheses of spatially random distributions of types (here genotypes), by Cliff and Ord (1973). The first is to assume that the values (genetic values), independent of spatial locations, have normal distributions that are also known. Both assumptions are not met in the usual experimental setting for spatial distributions of diploid genotypes. The second, used here, is that the denominator of  $I$  is fixed, determined by the sample, hence this form is readily applied to experiments. Moreover, the relationship between  $I$  and the single allele measure of genetic correlation or conditional kinship,  $R$  (Cockerham, 1969), is straightforward. In fact, Moran's  $I$ -statistic as applied to genetic data is of a form suggested by Malécot in 1955 (Malécot, 1955). The denominator of  $I$  is equal to the sample variance,  $\sigma^2$ , of genetic values times  $nW_k$ , and  $R = (I\sigma^2)/(q(1-q))$ , where  $q$  is the allele frequency in the sample. An alternative expression of this relationship is  $R = I(1+F)/2$ , where  $F$  is the allele-specific maximum likelihood estimator of the fixation index (Brown, 1970) (see also Hardy and Vekemans, 1999; Rousset, 2000). The fixation index is determined by the sample, hence the distribution of  $R$  under the null hypothesis  $H_0$  is equal to that of  $I$  multiplied by the constant  $(1+F)/2$ . Moreover, averages of  $R$ , taken either over genes (if there is no linkage disequilibrium) or over individuals first, are simply weighted averages (by constants  $(1+F)/2$ ) of  $I$ -statistics, under the null hypothesis.

$$I_a(k) = \frac{2n[q^2 n_{AAAA} - q(0.5-q)n_{AAAA} - q(1-q)n_{Aaaa} + (0.5-q)^2 n_{AaAa} + (0.5-q)(1-q)n_{Aaaa} + (1-q)^2 n_{aaaa}]}{W_k [q^2 n_{AA} + (0.5-q)^2 n_{Aa} + (1-q)^2 n_{aa}]} \quad (3)$$

Hence the results in this article are extended to single and multiple gene versions of  $R$ .

In this article, first methods are developed for characterizing the covariances and correlations of Moran's  $I$ -statistics, and by extension  $R$  statistics, for different alleles of a locus, under the null hypothesis of a

random distribution. This allows estimation of the standard error under the null hypothesis and tests of significance based on the average value across alleles and loci. In addition, the values of the correlations are characterized under processes of isolation by distance. Fortunately, and perhaps surprisingly, the results will show that the correlations do not depend much (if at all) on the degree of isolation by distance. Hence, under a wide range of circumstances they can be used to estimate standard errors of and form confidence intervals for multiallelic, multiple-locus estimators of genetic correlations (i.e. without assuming the null hypothesis), without having to posit the degree of spatial autocorrelation, which could otherwise involve a degree of circularity. In sum, it becomes possible to analyze spatial genetic structure based on multiallelic multiple locus data, using estimators that have known distributions and relationships to process parameters of simulated theoretical models.

## 2. Methods

### 2.1. Analysis of covariances and correlations under the null hypothesis

Moran's  $I$ -statistic for converted genotypes is a function of the so-called join counts and the average value (which equals the allele frequency in the entire sample). To illustrate, first consider the two allele case with three genotypes, AA, Aa, aa, which occur with numbers  $n_{AA}$ ,  $n_{Aa}$ , and  $n_{aa}$  in a sample of total size  $n$ . Let the frequency of the a allele be denoted as  $q$  [note that  $q = (n_{aa} + n_{Aa}/2)/n$ ], and  $n_{AAAA}$ ,  $n_{AAAa}$ ,  $n_{AAaa}$ ,  $n_{AaAa}$ ,  $n_{Aaaa}$ ,  $n_{aaaa}$  be the numbers of AA × AA, AA × Aa, AA × aa, Aa × Aa, Aa × aa, aa × aa pairs or joins, respectively, for a given distance class  $k$ . Then, the Moran  $I$ -statistic,  $I_a(k)$ , for allele a and distance class  $k$  is:

(Epperson, 1995b). For loci with arbitrary numbers of alleles, there are three important classes of joins, those that involve aa, those that involve  $\bar{a}a$  (where  $\bar{a}$  means any allele that is not allele a), and those that involve neither (joins with  $\bar{a}\bar{a}$  only). It can be shown that:

$$I_a(k) = \frac{2n[q^2 \sum_{\bar{a}} n_{\bar{a}\bar{a}\bar{a}\bar{a}} - q(0.5-q) \sum_{\bar{a}} n_{\bar{a}\bar{a}\bar{a}\bar{a}} - q(1-q) \sum_{\bar{a}} n_{\bar{a}\bar{a}\bar{a}\bar{a}} + (0.5-q)^2 \sum_{\bar{a}} n_{\bar{a}\bar{a}\bar{a}\bar{a}} + (0.5-q)(1-q) \sum_{\bar{a}} n_{\bar{a}\bar{a}\bar{a}\bar{a}} + (1-q)^2 n_{aaaa}]}{W_k [q^2 \sum_{\bar{a}} n_{\bar{a}\bar{a}} + (0.5-q)^2 \sum_{\bar{a}} n_{\bar{a}\bar{a}} + (1-q)^2 n_{aa}]} \quad (4)$$

where each summation is over all alleles ( $\bar{a}$ ) that are not allele  $a$  (Epperson, 2003a).

The covariance of  $I_A(k)$  and  $I_B(k)$ , for alleles denoted A and B (to avoid confusion of A,  $a$  and  $\bar{a}$ , etc. in the multiallele case), is the expected value of the product  $I_A(k)I_B(k)$  minus the product of the expected values of  $I_A(k)$  and  $I_B(k)$ . A correlation coefficient is the covariance divided by the product of the standard errors of  $I_A(k)$  and  $I_B(k)$ . It follows that the correlations are functions of the actual numbers of joins, the expected values of products of numbers of joins, the products of expected numbers of joins, the frequencies of genotypes and alleles, the sample size, and three other parameters that are determined by the spatial distributions of sample locations. The latter are:

$$S_0(k) = \sum_{(2)} \delta_{ij}(k) \tag{5}$$

or  $W(k)$ , twice the number of joins total for distance class  $k$ ,

$$S_1(k) = \frac{1}{2} \sum_{(2)} (\delta_{ij}(k) + \delta_{ji}(k))^2 \tag{6}$$

which is four times the number of joins ( $2W(k)$ ) and,

$$S_2(k) = \sum_{i=1}^n \left( \sum_{j=1}^n \delta_{ij}(k) + \sum_{j=1}^n \delta_{ji}(k) \right)^2 \tag{7}$$

(Cliff and Ord, 1981).

As noted in the introduction, the most useful null hypothesis is that for “randomization,” in part because it does not require the true mean of a stochastic process (Cliff and Ord, 1973), which is usually unknown if not unknowable. The randomization null hypothesis means that the distributions of pairs of genotypes is that produced by randomly sampling (without replacement) from the sample genotypes without respect to their spatial locations, and it is also termed “non-free” sampling. The expected values of products of join counts under the non-free null hypothesis were derived in an earlier article (Epperson, 2003b). It was shown that there are seven distinctive types of products of joins (Table 1). We will use single-letter notations for diploid genotypes, in order to manage the expressions for the products of joins, and only four genotypes, denoted B, W, C, and D, are needed. The seven classes of types of products of joins are typified by BB BB, BW BW, BB BW, BB WW, BW BC, BW CC, and BW CD, respectively, and for example BW CD denotes the product of joins between genotypes B and W and joins between genotypes C and D. The method is completely general for arbitrary numbers of genotypes E, F, G, etc., because for example the product of EE joins with FG joins (or vice versa) is in the same class as BB WC (Epperson, 2003b).

Table 1

Types of products of joins, indicating the appropriate one of the seven equations in the appendix, for the case of four genotypes, B, W, C, and D

Type of join	Type of join									
	BB	BW	BC	BD	WW	WC	WD	CC	CD	DD
BB	A1	A3	A3	A3	A4	A6	A6	A4	A6	A4
BW	—	A2	A5	A5	A3	A5	A5	A6	A7	A6
BC	—	—	A2	A5	A6	A5	A7	A3	A5	A6
BD	—	—	—	A2	A6	A7	A5	A6	A5	A3
WW	—	—	—	—	A1	A3	A3	A4	A6	A4
WC	—	—	—	—	—	A2	A5	A3	A5	A6
WD	—	—	—	—	—	—	A2	A6	A5	A3
CC	—	—	—	—	—	—	—	A1	A3	A4
CD	—	—	—	—	—	—	—	—	A2	A3
DD	—	—	—	—	—	—	—	—	—	A1

For “non-free” sampling the expected number of joins for any distance class  $k$  is

$$\mu_{BB}(k) = \frac{S_0(k)n_B^{(2)}}{2n^{(2)}} \tag{8}$$

and

$$\mu_{BW}(k) = \frac{S_0(k)n_B n_W}{n^{(2)}} \tag{9}$$

for  $B \neq W$ . Throughout what follows we will assume that where different letters are used, it means that they are different genotypes. Here  $n_B$  is the number of times that genotype B occurs among all  $n$  locations,  $n_B^{(2)} = n_B(n_B - 1)$ , and  $n^{(2)} = n(n - 1)$ ,  $n^{(3)} = n(n - 1)(n - 2)$ ,  $n^{(4)} = n(n - 1)(n - 2)(n - 3)$ , and similarly for the  $n_B^{(3)}$  etc. (Cliff and Ord, 1981). To simplify exposition we make the minor assumption that there are at least four of all genotypes in the sample, so that the  $n_B^{(4)}$  are defined. The seven equations for the expected values of products of join counts are given in the appendix.

It is worth noting that the products typically involve  $S_2(k)$ , a second-order measure determined by distribution of sample locations and the definition of distance class  $k$ . Equations for the variances of join counts are also given in the appendix. Together these results provide all that is needed to characterize the covariances and correlations of Moran’s  $I$ -statistics for different alleles of the same locus, under the non-free sampling null hypothesis of random spatial distributions.

The expected products (and subtracting off the product of expected values, then the covariances) between Moran’s  $I$ -statistics for alleles A and B at a locus, for a given distance class  $k$ ,  $I_A$  and  $I_B$ , were calculated by multiplying the two appropriate expressions for Eq. (4) (right-hand side), i.e. by summing terms of products of functions of allele frequencies and expected values of products of join counts. The latter were computed using appendix equations (A.1)–(A.7).

A wide range of allele frequencies were input, and genotypic frequencies were calculated according to Hardy–Weinberg proportions, which appropriately reflects the lack of mating by proximity and inbreeding in a randomly distributed population under the null hypothesis (other forms of inbreeding, e.g. selfing, would violate Hardy–Weinberg). Because these expressions involve values of  $S_0(k)$ ,  $S_1(k)$  and  $S_2(k)$ , the values may depend on the array of sample locations and distance class. In the results reported here calculations were done on a  $100 \times 100$  grid. Correlations were obtained by dividing by the standard errors (Appendix equations (A.8) and (A.9)). The method of calculation is completely general. A computer program was written to calculate the values of correlations of  $I_A$  and  $I_B$ , for any sample [any set of weights,  $\delta_{ij}(k)$ ], with any number of alleles, and this program is available upon request from the author. The program takes advantage of the fact that, regardless of the number of alleles, all products of joins can be placed into seven classes (Table 1) that determine the coefficient (multiplier of the allele frequencies function) and the powers of the genotypic frequency terms. It is not reasonable to write here the equations for the covariances (or correlations), because there are large numbers of terms. For example, the smallest case of interest is that of three alleles, for which there are six genotypes, 21 types of joins, and  $21^2 = 441$  terms of products of joins. For four alleles there are 10 genotypes, 55 types of joins, and 3025 products of joins. Instead, the computer program first finds the one appropriate form of seven possible equations (see Table 1), by recognizing the type of products of joins. The algorithm also simultaneously counts the actual numbers of joins and  $S_0(k)$ ,  $S_1(k)$ , and  $S_2(k)$ , and calculates the genotypic frequencies.

## 2.2. Simulations of genetic isolation by distance

Full details of the simulation program, which uses Monte Carlo methods to simulate stochastic generations of life cycles, were described previously (Epperson, 1990). Briefly, each simulated population consisted of 10,000 individuals with diploid single-locus genotypes located on a  $100 \times 100$  square lattice. Each simulation was initialized with a randomly sampled distribution of genotypes, with expected proportions following Hardy–Weinberg proportions. Simulations were replicated in sets of 100. Sets differed according to the amount of simulated dispersal (Table 2), and in the numbers of alleles (three, four or five) and their initial frequencies (Table 3). For each multiplicity of alleles, a set of simulations was conducted with initially even frequencies of all alleles. Allele frequencies changed very little during the course of a simulation. In addition, four sets of simulations were conducted with uneven frequencies of the alleles (Table 3). Together, the sets represent a

Table 2  
Dispersal parameters in various sets of simulations

	Simulation dispersal model <sup>a</sup>				
	1	2	3	4	5
$N_f$	1	25	49	1	625
$N_m$	9	25	49	225	625
$N_e$	4.2	25.1	50.2	115.2	632.4

<sup>a</sup>  $N_f$  and  $N_m$  are the numbers of nearest female and male individuals from which parents of an offspring are randomly chosen, and  $N_e$  is effectively Wright's neighborhood size.

Table 3  
Models of unequal allele frequencies for simulations

Model	Allele				
	1	2	3	4	5
3u	0.1	0.3	0.6	—	—
4u	0.1	0.1	0.3	0.5	—
5u <sup>1</sup>	0.1	0.1	0.1	0.1	0.6
5u <sup>2</sup>	0.02	0.02	0.05	0.05	0.86

wide range of types of allele constellations and elucidated several principles applicable to more general situations. They also allow contrasts that separate effects of allele frequency per se from possible effects of number of alleles. There was little reason to consider cases with more than five alleles, in part because it would force low allele frequencies. Five dispersal models were conducted, which together represent a very wide range of dispersal levels (Table 2). Either or both the female and male parents of an offspring were chosen at random (using two Uniform (0,1) pseudo-random numbers to choose the two coordinates for each parent) generally from one of the nearest  $N_f$  and  $N_m$  (respectively) neighbors including self. Thus, each individual within the group of size  $N_f$  and  $N_m$  had equal chances of being the female or male parent, respectively. This may be considered unrealistic for many species, in which the probability of dispersal decays with distance. However, it is justified by the fact that the form of the dispersal curve has little effect on spatial structure, rather it is the standardized neighborhood size that matters (e.g., Rohlf and Schnell, 1971). In total, 35 sets, each requiring about 4 h of cpu on a Sun Ultraspark, or 3500 space–time simulations were run for loci with multiple alleles. Each simulation was run for 200 generations, by which time the simulations have obtained a quasi-stationary state (e.g., Sokal and Wartenberg, 1983). In addition, for each allele model 100 of the randomly generated initial surfaces were examined as realizations of the randomization null hypothesis, and these are denoted by “dispersal model”  $N_e = 10,000$  (random mating).

3. Results

3.1. Correlations under the null hypothesis

The covariances and correlations were calculated for several quite different distance classes, ranging from a first distance class of near-neighbors (0–1.5 times the distance between adjacent individual locations) to near maximum distances, 137.5–138.5 times the distance between adjacent locations (Table 4). These distance classes varied widely in values of  $S_0(k)$  and  $S_2(k)$  (it is unnecessary to consider  $S_1(k)$  because it is equal to  $2S_0(k)$  in the case of unweighted  $I$ -statistics). Naturally, the covariances varied widely for the different cases. More importantly and perhaps surprisingly, there were no discernible differences in the correlations among these cases. There were some small differences (0.001 or less), but these were always smaller than computational error. [The computation precision errors were evidenced by building several symmetries into the calculations, redundancies of types of pairs of alleles, and calculating

both the correlation with allele A first and B second and the correlation for the reverse (B,A). Errors were minuscule for moderate values of  $S_0(k)$  and  $S_2(k)$ , and still small for larger ones. They result from summing large numbers of products of large numbers, beyond computer language precision levels.] Inspection of the equations, with such terms as  $n_B^{(3)}$ , etc., indicates that the correlations are unlikely to be completely independent of  $S_0(k)$  and  $S_2(k)$ , but algebraic proof of this is infeasible because of the huge numbers of terms. Nonetheless, the correlations were essentially invariant over a wide range of values of  $S_0(k)$ ,  $S_2(k)$  and the ratio  $S_2(k)/S_0(k)$  (Table 4). Moreover, any differences that were observed are small compared to stochasticity in correlations in realizations of a randomization process (see below).

Because of the lack of effect of distance class, only the correlations for distance class one are shown in Table 5. For the allele models with equal allele frequencies,  $q$ , the correlations are 0.250, 0.111, 0.062 respectively for three alleles ( $q = 0.33$ ), four alleles (0.25), and five alleles (0.20). They are large only for loci with three alleles. For the models with alleles having unequal frequencies the correlations are also always positive, but they are large only for pairs of relatively high frequency alleles. Moreover, these results strongly indicate that for a locus with more than two alleles, the number of alleles is unimportant, and that only allele frequencies matter. For example, correlations (0.048) for the 1, 2 pair of alleles ( $q = 0.1, 0.3$ ) in the three allele case are identical to those for the “1”, 2 pair with the same allele frequencies in the four allele case (Table 5). All such

Table 4  
Distance classes analyzed for correlations under the null hypothesis

Distance class	$S_0$	$S_2$	$S_2/S_0$
0–1.5	78,804	2,497,968	32
4.5–5.5	262,216	28,119,680	107
9.5–10.5	490,508	100,450,256	205
19.5–20.5	851,120	313,403,104	368
49.5–50.5	1,396,720	827,825,824	593
99.5–100.5	235,272	50,731,776	216
137.5–138.5	80	800	10

Table 5  
Correlations of Moran’s  $I$ -statistics for converted genotypes for two alleles, for distance class one, for sets of simulated populations with various allele and dispersal models and under the null hypothesis

Allele model	Alleles <sup>a</sup>	Allele freqs.	Dispersal model						Average	10,000	$H_0$
			4.2	25.1	50.2	115.2	637.4				
3e	“1”, “2”	0.33, 0.33	0.256	0.303	0.210	0.320	0.301	0.278	0.154	0.250	
3u	1,2	0.10, 0.30	0.159	0.051	0.280	0.063	0.052	0.121	–0.097	0.048	
	1,3	0.10, 0.60	0.291	0.306	0.175	0.234	0.220	0.245	0.050	0.166	
	2,3	0.30, 0.60	0.667	0.638	0.571	0.585	0.669	0.625	0.618	0.643	
4e	“1”, “2”	0.25, 0.25	0.218	0.105	0.060	0.028	0.162	0.115	0.064	0.111	
4u	1,2	0.10, 0.10	0.095	0.139	–0.009	–0.067	0.150	0.062	–0.121	0.012	
	“1”,3	0.10, 0.30	–0.133	0.081	–0.041	–0.035	–0.039	–0.033	–0.056	0.048	
	“1”,4	0.10, 0.50	0.083	0.101	0.112	0.098	0.089	0.097	–0.015	0.111	
5e	3,4	0.30, 0.50	0.408	0.465	0.608	0.372	0.430	0.457	0.285	0.429	
	“1”, “2”	0.20, 0.20	0.023	0.036	0.051	0.058	0.031	0.040	–0.006	0.062	
5u <sup>1</sup>	“1”, “2”	0.10, 0.10	–0.035	0.007	–0.096	0.019	0.176	0.014	–0.013	0.012	
	“1”,5	0.10, 0.60	0.171	0.181	0.058	0.132	0.282	0.165	0.079	0.167	
5u <sup>2</sup>	1,2	0.02, 0.02	0.102	–0.101	–0.065	–0.026	0.026	–0.013	–0.059	0.000	
	“1”, “3”	0.02, 0.05	–0.018	–0.002	–0.015	–0.081	0.074	–0.008	0.025	0.001	
	“1”,5	0.02, 0.86	0.233	0.196	0.264	0.055	0.213	0.192	0.239	0.125	
	3,4	0.05, 0.05	–0.029	0.053	0.045	–0.041	–0.044	–0.003	–0.072	0.003	
	“3”,5	0.05, 0.86	0.366	0.366	0.398	0.402	0.378	0.382	0.384	0.324	

<sup>a</sup>Quotation marks means either that any one of alleles with same frequencies, in the case of the expected value under the null hypothesis, or the average of correlations found for alleles with the same initial frequencies, in the case of simulated processes.

Table 6  
Two examples illustrating the stochasticity in correlations of  $I$ -statistics between alleles in sets of 100 simulated populations

Allele	Allele				
	1	2	3	4	5
1		−0.085	−0.063	−0.065	0.270
2	0.100		−0.164	−0.002	0.116
3	0.085	−0.111		0.172	0.171
4	−0.064	0.083	−0.019		0.128
5	0.104	0.102	0.003	0.320	

One example is dispersal with  $N_e = 4.2$  (above diagonal) and the other has  $N_e = 115.2$  (below diagonal). Both have the same allele model, with five alleles with unequal allele frequencies (model 5u<sup>1</sup>). In both cases, alleles 1, 2, 3, 4 have the same frequencies (0.1) and allele 5 is most common (0.6). Thus for example all analogous pairs of alleles (e.g. 1:5, 2:5, 3:5, and 4:5) have the same expected correlations.

comparisons yield identical or nearly identical (perhaps within calculation error) correlations. Finally, in cases where there is one major allele, one moderate-frequency allele, and the remaining alleles all in low frequency, the larger correlations involve the major allele. For example, in the four-allele model, apart from correlations involving allele 4, the largest is 0.048. In the two five-allele models, if allele 5 is omitted the largest correlations are 0.012 and 0.003, respectively.

### 3.2. Correlations under isolation by distance

The correlations in simulations, although variable among sets, show no strong systematic trends with the degree of isolation by distance (Table 5). For example, the averages across all allele models are 0.17, 0.17, 0.15, 0.12 and 0.19, respectively, for dispersal models 1–5. Moreover, as with the values under the null hypothesis, there were no apparent systematic trends among distance classes, thus only the results for distance class one are shown in Table 5. It appears that the differences among simulation sets are mostly due to stochasticity. The level of stochasticity is detectable as differences among pairs of different alleles with the same paired allele frequencies. Two examples are shown in Table 6. For example, for  $N_e = 4.2$  (above diagonal), values range from 0.116 to 0.270 for the four rare allele–major allele pair-wise combinations. For  $N_e = 115.2$  (below diagonal), the analogous range is 0.003–0.320. Moreover, the average values over all five non-random mating dispersal simulation sets are very close to the values under the null hypothesis, and so are the values for random simulated surfaces (Table 5).

## 4. Discussion

The results utilize the fact that Moran's  $I$ -statistics for genotypes converted to gene frequencies can be ex-

pressed in terms of join-counts of pairs of genotypes (Epperson, 1995b, 2003a). The covariances and correlations among  $I$ -statistics for different alleles of the same locus are functions of the expected values of the products of the join-counts (Epperson, 2003b). Under the null hypothesis, albeit the algebraic expressions of correlations among  $I$ -statistics for different alleles typically involve huge numbers of terms, algorithms for computing them can take advantage of the fact that the terms can be placed into seven classes. For the wide range of studied values of  $S_0(k)$  and  $S_2(k)$ , and the ratio  $S_2(k)/S_0(k)$ , covering any likely contemplated in experiments, the correlations were invariant. In other words, the results should apply to usual experiments, largely robust to the details of the actual spatial distribution of sample point locations and distance class designations. Although it appears unlikely that complete invariance would hold for all  $S_0(k)$  and  $S_2(k)$ , the size of the expressions makes infeasible any algebraic proof of this. In contrast, the correlations depend strongly on allele frequencies, and these are given by the data in experiments. Knowledge of the correlations under the null hypothesis allows for the first time proper statistical tests of significance based on the mean value of pair-wise spatial correlations over alleles, as is discussed in detail below.

Remarkably, and perhaps surprisingly, among the studied models of isolation by distance the correlations among  $I$ -statistics for different alleles were unaffected by the spatial autocorrelations of genotypes. This is useful for testing alternate hypotheses, and for generating standard errors for mean  $I$ -statistics. It implies that we do not need to know the spatial autocorrelation in order to obtain standard errors of summary measures of autocorrelation, thus escaping from a circularity. It should also be noted that the fixation index, caused by mating by proximity, ranged widely among the simulations, from near zero to about 0.33 (Epperson, 1990), suggesting that it has little or no effect on the correlations of  $I$ -statistics. Further, although there were no strong or even evident systematic trends of the correlations with degree of isolation by distance, there was considerable variation among sets of simulations. This variation is attributable to the substantial stochastic variability that was observed among redundancies built into the models, such as variations among analogous pairs of alleles within sets of simulations, which are caused solely by stochastic variation in the spatial distributions. It should not be surprising that correlations among autocorrelation statistics have considerably high levels of stochasticity. Different simulation runs within the same set can be considered to correspond to different loci, with similar arrays of allele frequencies, and subject to the dispersal characteristics of the same population. The correlations of autocorrelations are generally positive, indicating that in the spatial

distribution of a locus, when one allele exhibits higher autocorrelations, compared to the average across loci, other alleles of the locus also tend to have higher than average autocorrelations. It should be noted that the correlations over a smaller number of loci (e.g. 10) would exhibit even greater stochasticity than that observed in the sets, each of which contained 100 simulations.

The results also indicate that correlations of statistics for different alleles do not depend directly on the number of alleles, as long as it is greater than two. Rather they depend strongly on the frequencies of alleles. The effects of allele frequencies are also fairly simply expressed. Correlations are largest for pairs of moderate- to high-frequency alleles (e.g. both with frequency greater than ca. 0.3), moderate in value for common–rare pairs of alleles, and very small for pairs of low-frequency alleles (both less than ca. 0.2). Thus only loci with less than five alleles have substantial correlations, unless there is a very common allele. However, many isozyme loci fit this profile. In contrast, many microsatellite loci have much larger numbers of alleles and none with very high-frequency, although some do have high-frequency alleles (e.g., [Rajora et al., 2000](#)).

Because Moran's  $I$ -statistic is asymptotically normal-distributed ([Cliff and Ord, 1981](#)), the joint distribution of a set of values for different alleles should depend only on the means, variances and correlations. The average,  $\bar{I}$ , over  $m$  alleles of a locus has a variance given by the standard formula for an (unweighted) average (e.g., [Feller, 1957](#)):

$$\sigma^2(\bar{I}) = \left(\frac{1}{m}\right)^2 \left[ \sum_{A=1}^m \sigma^2(I_A) + 2 \sum_{A,B} \text{Cov}(I_A, I_B) \right] \quad (10)$$

where the second summation is over all alleles A and B such that  $A < B$ ,  $\sigma^2(I_A)$  is the variance of  $I_A$  and the  $\text{Cov}(I_A, I_B)$  are stochastic covariances for alleles A and B. For tests of significance of  $\bar{I}$ , the values of correlations,  $\text{Corr}(I_A, I_B)$ , under the null hypothesis are used, i.e. set  $\text{Cov}(I_A, I_B) = \text{Corr}(I_A, I_B)\sigma(I_A)\sigma(I_B)$ . The statistic  $(\bar{I} - E(\bar{I}))/\sigma(\bar{I})$  has an asymptotically standard normal distribution, where  $E(\bar{I})$  is the expected value of  $\bar{I}$ , which is equal to the average expected value for each allele. Moreover, averages could be further taken over loci, by assuming that the correlations are zero for genes of different loci (i.e., stochastic linkage disequilibrium coefficients are zero). Eq. (10) can also be considered to sum over loci as well as alleles, so that  $m$  is taken as a function of numbers of alleles for all loci.

Similarly, because the variances and correlations are essentially unaffected by structure under isolation by distance, the standard errors of  $\bar{I}$  are likewise unaffected. Thus confidence intervals for  $\bar{I}$  can be computed in the usual way, under the much more relaxed assumption of isolation by distance. That the variances

of allele-specific  $I$ -statistics are largely unaffected by isolation by distance has been shown in a number of studies (e.g., [Epperson and Li, 1997](#); [Epperson et al., 1999](#)). Moreover, the variances and expected values of  $I$  are unaffected by allele frequency ([Epperson, 1995b](#); [Epperson et al., 1999](#)), unless it is reduced to ca. 0.02 ([Epperson, 2003a](#)). There appears to be little compelling reason to weight alleles in calculating averages of  $I$ -statistics. The lack of gain of statistical power by weighting alleles by their frequency was also observed by [Smouse and Peakall \(1999\)](#).

In total, the results of this and other studies inform methods of analyzing spatial distributions of diploid multilocus genotypes. First, it should be pointed out that very rare alleles, e.g. ones represented in only one or a few individuals, should be discarded as uninformative and potentially biased. If only one allele remains, then that locus is uninformative. If only two alleles remain, then one should be discarded as providing nearly identical information as the other.

If there are three or more non-rare alleles, then there are three potential courses of action. The first is simply to ignore the correlations among alleles, which seems a reasonable proposition if there are four or more alleles that are not in high frequency. In all studied cases the correlations are positive under  $H_0$ , and this means that ignoring them would generally lead to an underestimate of the variance of  $\bar{I}$ . The maximum error occurs for a diallelic locus, where the correlation is 1.0, and if is not taken into account in Eq. (10), the estimated variance of the average of two  $I$  values would be half as large as it should be. To examine the effects for other cases, it will be assumed that the variances for  $I$  are the same for all  $m$  alleles. Then the ratio of the variance of  $\bar{I}$  ignoring correlations to the correct variance (Eq. (10)) is  $m \div [m + 2\sum \text{Corr}(I_A, I_B)]$ , where again the summation is over all pairs of alleles A and B such that  $A < B$ . For the models with equal allele frequencies the ratios are 0.67, 0.75, and 0.80, respectively for three, four and five alleles, indicating that the variance is underestimated by 20–33%, a substantial error. Interestingly, the variances would be correct if the sum of variances were divided by  $m(m-1)$  rather than  $m^2$ . This implies that  $2\sum \text{Corr}(I_A, I_B) = m/(m-1)$  (where the sum is over  $A < B$ ). For the other models,  $3u$ ,  $4u$ ,  $5u^1$  and  $5u^2$ , the ratios are 0.64, 0.72, 0.77 and 0.73. Note these are different from and somewhat smaller than those for the corresponding even-frequency cases, indicating that the error is increased. However, the differences are small, and little bias would be created if the sum of variances were divided by  $m(m-1)$  rather than by  $m^2$ , as in the equal-frequency cases. Nonetheless, in all of these cases, unless a correction is made, the variance will be seriously biased by ignoring the correlations. However, some markers may have as many as 25 or even 50 alleles, and then the bias would be negligible.

The second method would be for experimenters to find in the Table 5 the closest fit of allele frequencies to the data, or use the calculation formula (appendix equations (A.1)–(A.9)), and input the correlations into Eq. (10). This should be considered an approximation because the correlations are stochastic. It provides an unbiased estimate of the standard error of  $\bar{I}$ , although it requires a fair amount of calculation.

A third procedure may be recommended in that it is simple to implement. It is reasonable to consider this option when allele frequencies are not very even, especially where one allele has high frequency (say 0.6 or greater). This procedure is to omit from the average the most common allele, and ignore remaining correlations, all of which will then be small. For examples, if the highest frequency allele is removed in models 3u, 4u, 5u<sup>1</sup> and 5u<sup>2</sup>, the percent errors of ignoring correlations are 0.046, 0.035, 0.018, and 0.002, i.e. negligible. Note that the estimated variance is increased, because the sum of variances is divided by  $(m-1)^2$  rather than  $m^2$ . However, it appears that fairly little information is lost by throwing out the most common allele, because of the correspondence to the correct variance as shown above in the discussion of the approach of ignoring all correlations and not first removing the most common allele. Its spatial information is highly correlated with one or another of the other alleles. This also has been observed in our datasets in various experimental studies. Very often I-correlograms for high-frequency alleles appear to be “mimicked” by those for some low-frequency alleles. Finally, it is worth noting that a potential fourth alternative procedure, removing rare alleles (Campbell and Dooley, 1992), generally will not make the remaining alleles uncorrelated nor the estimated variance unbiased. For example, if one of the rarest alleles is omitted for models 3u, 4u, 5u<sup>1</sup> and 5u<sup>2</sup>, the variance estimates for the average not only are increased (reflecting loss of information) they still underestimate the true variance by 22–31%.

In summary, the simplest recommendations for creating per locus averages are: 1. very rare alleles should be discarded; 2. if there are more than 15 or 20 alleles, the correlations can be ignored; and 3. for smaller numbers of alleles, all alleles can be kept and the estimated variance of the average can be calculated in the usual way (i.e., by treating the allele-specific values as being independent and ignoring the correlations), and then multiplying the result by  $m/(m-1)$ . The latter should closely approximate the correct variance under conditions likely met in experimental studies. It is equally valid for both testing the null hypothesis, and for forming confidence intervals when population structure is determined by isolation by distance processes. Further, it can be verified by matching the constellation of pairs of allele frequencies to the closest ones displayed in Table 5, and checking if

$2\sum \text{Corr}(I_A, I_B)$  is close to  $m/(m-1)$ . Finally, the per locus averages can be further averaged over loci, and the assumption that they are independent can usually be assumed unless there are strong reasons to expect substantial linkage disequilibrium.

It should be pointed out that pair-wise statistics can also be used as indirect but precise estimators of levels of dispersal. Isolation by distance quickly results in a spatial distribution that is highly stable, persisting for long periods (e.g., Sokal and Wartenberg, 1983). Spatial autocorrelations remain constant during this so-called “quasi-stationary phase,” and the magnitude of Moran’s  $I$  for converted genotypes for short distance classes decreases monotonically with total dispersal, as measured for example by Wright’s neighborhood size (Epperson and Li, 1997). Two alternatives are to use only the statistics for the first distance class or use entire correlograms. The former is more straightforward, and can be nearly as powerful as is the latter (Oden, 1984), in part because specialized tests for correlograms and Bonferroni type corrections for multiple tests are overly conservative. In either case, generally samples of a few thousand genotypes in total are sufficient to detect and precisely measure autocorrelation, even when dispersal is high and autocorrelation is weak (Epperson et al., 1999). That is, the average for distance class one, averaged over  $m$  total alleles (following the above guidelines for which alleles to include in the average), for a sample of  $n$  individuals, will usually be adequate when  $nm$  is a few thousand. An example is five loci, with three alleles each, assayed for 150 individuals, which is well within the size of many experimental systems. In some cases one or two microsatellite loci could be sufficient. For example, in a sample of ca. 100 seedlings of *Pinus strobus* assayed for five microsatellites, having 2–12 alleles each with a total of 27 alleles, the average value of  $I$ -statistics for the first distance class was 0.020 (Walter and Epperson, 2004). The value was highly statistically significant, and even the weak spatial structure expected for such a dense, outbreeding population with wind-dispersed seed and pollen was detected as being significantly different from a random spatial distribution. Interestingly, the results also showed that spatial structure for one of the loci (Rps 50), which was by far the most variable with 12 alleles, was significantly different from that observed for the other loci. The values were  $-0.012$  for Rps 50 and  $0.048$  for the other loci combined. Such a difference might be caused by a higher mutation rate for Rps 50, and it would not have been detected if only the total multilocus average were computed.

As noted in the introduction, the results in this article characterizing the correlations for  $I$ -statistics for alleles of the same locus can be applied to several other measures. Most other measures are based on the single-allele measure of genetic relatedness of two individuals  $i$

and  $j$  given by  $(q_i - q)(q_j - q)/q(1 - q)$  (Morton, 1975). For example, an average may be found for a set of pairs of individuals:

$$R = \frac{\sum_{i,j} (q_i - q)(q_j - q)}{kq(1 - q)}, \quad (11)$$

where  $q_i$  is the converted frequency of the allele in individual  $i$  and  $q$  is the average frequency in the entire sample. The summation is over all  $k$  pairs of  $n$  sample genotypes in the set (here distance class). It can be easily shown that  $R = (I\sigma^2)/q(1 - q)$ , or alternatively  $R = I(1 + F)/2$ , where  $F$  is the fixation index specific to the allele. Some multilocus estimators are formed as the average value of  $R$ , as defined in Eq. (11), averaged over alleles and/or loci, forming  $\underline{R}$  (e.g., Loiselle et al., 1995). If  $F$  is not constant among alleles, then  $\underline{R}$  is the average,  $\underline{I}$ , only where each  $I$  is weighted by  $(1 + F)/2$ . Since the  $F$  are fixed in the sampling, the values of  $(1 + F)/2$  are constants (not random variables) and the correlations among the  $R$  equal those among the  $I$ . The correlations, together with an equation analogous to Eq. (10), but allowing for weights, can be used to find the variance of  $\underline{R}$ . If  $F$  can be considered constant among alleles then  $\underline{R} = \underline{I}(1 + F)/2$  and the variance of  $\underline{R}$  equals  $[(1 + F)^2/4]$  times the variance of  $\underline{I}$ .

Other multilocus kinship estimators, here denoted  $\underline{R}$ , first sum over all alleles and loci for each pair of multilocus genotypes, and then over pairs of individuals (e.g., Smouse and Peakall, 1999), opposite the order for averaged  $I$ -statistics and  $\underline{R}$  as defined above. If there is linkage disequilibrium then the loci are not independent, and this will affect the distribution, including the variance, of  $\underline{R}$ . The results of this study will not apply. In principle, the distribution could be developed in terms of pairs of multilocus genotypes. In practice this would be inordinately complex, because the number of genotypes would be large and the numbers of products of joint counts needed increases with the fourth power of the number of genotypes. If linkage disequilibrium is negligible, as would often be expected, then  $\underline{R}$  is again a weighted average of the  $I$ -statistics, and its variance can be calculated using the present methods.

## Appendix

The expected value of  $n_{BB}(k)n_{BB}(k)$  under non-free sampling is

$$\frac{1}{4} \left[ \frac{S_1(k)n_B^{(2)}}{n^{(2)}} + \frac{(S_2(k) - 2S_1(k))n_B^{(3)}}{n^{(3)}} + \frac{(S_0(k)^2 + S_1(k) - S_2(k))n_B^{(4)}}{n^{(4)}} \right] \quad (A.1)$$

and that for  $n_{BW}(k)n_{BW}(k)$  is

$$\frac{1}{4} \left[ \frac{2S_1(k)n_B n_W}{n^{(2)}} + \frac{(S_2(k) - 2S_1(k))n_B n_W (n_B + n_W - 2)}{n^{(3)}} + \frac{4(S_0^2(k) + S_1(k) - S_2(k))n_B^{(2)}n_W^{(2)}}{n^{(4)}} \right] \quad (A.2)$$

(Cliff and Ord, 1981).

The expected value for the remaining products were given in Epperson (2003b), and for  $n_{BB}(k)n_{BW}(k)$

$$E\{n_{BB}(k)n_{BW}(k)\} = \frac{1}{2} \left[ \frac{n_B^{(2)}n_W}{n^{(3)}}(S_2(k) - 2S_1(k))/2 + \frac{n_B^{(3)}n_W}{n^{(4)}}(S_0^2(k) - S_2(k) + S_1(k)) \right]; \quad (A.3)$$

for  $n_{BB}(k)n_{WW}(k)$

$$E\{n_{BB}(k)n_{WW}(k)\} = \frac{1}{4} \frac{n_B^{(2)}n_W^{(2)}[S_0^2(k) - S_2(k) + S_1(k)]}{n^{(4)}}; \quad (A.4)$$

for  $n_{BW}(k)n_{BC}(k)$

$$E\{n_{BW}(k)n_{BC}(k)\} = \frac{n_B n_W n_C}{4n^{(3)}}(S_2(k) - 2S_1(k)) + \frac{n_B^{(2)}n_W n_C}{n^{(4)}}(S_0^2(k) - S_2(k) + S_1(k)); \quad (A.5)$$

for  $n_{BW}(k)n_{CC}(k)$

$$E\{n_{BW}(k)n_{CC}(k)\} = \frac{n_B n_W n_C^{(2)}(S_0^2(k) - S_2(k) + S_1(k))}{2n^{(4)}}; \quad (A.6)$$

for  $n_{BW}(k)n_{CD}(k)$

$$E\{n_{BW}(k)n_{CD}(k)\} = \frac{n_B n_W n_C n_D (S_0^2(k) - S_2(k) + S_1(k))}{n^{(4)}}. \quad (A.7)$$

The variances,  $\sigma_{BB}^2(k)$ , under non-free sampling are

$$\sigma_{BB}^2(k) = \frac{1}{4} \left[ \frac{S_1(k)n_B^{(2)}}{n^{(2)}} + \frac{(S_2(k) - 2S_1(k))n_B^{(3)}}{n^{(3)}} + \frac{(S_0^2(k) + S_1(k) - S_2(k))n_B^{(4)}}{n^{(4)}} \right] - \mu_{BB}^2(k), \quad (A.8)$$

where  $\mu_{BB}(k)$  is given by text Eq. (8), for joins between identical types, and

$$\begin{aligned} \sigma_{BW}^2(k) &= \frac{1}{4} \left[ \frac{2S_1(k)n_{BNW}}{n^{(2)}} + \frac{(S_2(k) - 2S_1(k))n_{BNW}(n_B + n_W - 2)}{n^{(3)}} \right. \\ &\quad \left. + \frac{4(S_0^2(k) + S_1(k) - S_2(k))n_B^{(2)}n_W^{(2)}}{n^{(4)}} \right] - \mu_{BW}^2(k) \end{aligned} \quad (\text{A.9})$$

for  $B \neq W$ , where  $\mu_{BW}(k)$  is given by text Eq. (9) for joins between different types.

## References

- Bacilieri, R., Labbe, T., Kremer, A., 1994. Intraspecific genetic structure in a mixed population of *Quercus petraea* (Matt.) Lelbl and *Q. robur* L. *Heredity* 73, 130–141.
- Barbujani, G., 1987. Autocorrelation of gene frequencies under isolation by distance. *Genetics* 117, 777–782.
- Barton, N.H., Wilson, I., 1995. Genealogies and geography. *Philos. Trans. R. Soc. London B* 349, 49–59.
- Berg, E.E., Hamrick, J.L., 1995. Fine-scale genetic structure of a turkey oak forest. *Evolution* 49, 110–120.
- Bertorelle, G., Barbujani, G., 1995. Analysis of DNA diversity by spatial autocorrelation. *Genetics* 140, 811–819.
- Brown, A.H.D., 1970. The estimation of Wright's fixation index from genotypic frequencies. *Genetica* 41, 399–406.
- Campbell, D.R., Dooley, J.L., 1992. The spatial scale of genetic differentiation in a hummingbird-pollinated plant: comparison with models of isolation by distance. *Am. Nat.* 139, 735–748.
- Chung, M.Y., Chung, G.M., Chung, M.G., Epperson, B.K., 1998. Spatial genetic structure in populations of *Cymbidium goeringii* (Orchidaceae). *Genes Genet. Systems* 73, 281–285.
- Cliff, A.D., Ord, J.K., 1973. *Spatial Autocorrelation*. Pion, London.
- Cliff, A.D., Ord, J.K., 1981. *Spatial Processes*. Pion, London.
- Cockerham, C.C., 1969. Variance of gene frequencies. *Evolution* 73, 72–84.
- Doligez, A., Baril, C., Joly, H.I., 1998. Fine-scale spatial genetic structure with nonuniform distribution of individuals. *Genetics* 148, 905–919.
- Epperson, B.K., 1990. Spatial autocorrelation of genotypes under directional selection. *Genetics* 124, 757–771.
- Epperson, B.K., 1995a. Spatial structure of two-locus genotypes under isolation by distance. *Genetics* 140, 365–375.
- Epperson, B.K., 1995b. Fine-scale spatial structure: correlations for individual genotypes differ from those for local gene frequencies. *Evolution* 49, 1022–1026.
- Epperson, B.K., 2003a. *Geographical Genetics*. Princeton University Press, Princeton, NJ.
- Epperson, B.K., 2003b. Covariances among join-count spatial autocorrelation measures. *Theor. Popul. Biol.* 64, 81–87.
- Epperson, B.K., Clegg, M.T., 1986. Spatial autocorrelation analysis of flower color polymorphisms within substructured populations of morning glory (*Ipomoea purpurea*). *Am. Nat.* 128, 840–858.
- Epperson, B.K., Li, T.-Q., 1997. Gene dispersal and spatial genetic structure. *Evolution* 51, 672–681.
- Epperson, B.K., Huang, Z., Li, T.-Q., 1999. Spatial genetic structure of multiallelic loci. *Genet. Res. Camb.* 73, 251–261.
- Feller, W., 1957. *An Introduction to Probability Theory and its Applications*. Wiley, New York.
- Hardy, O.J., Vekemans, X., 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* 83, 145–154.
- Harpending, H.C., 1973. Discussion of relationship of conditional kinship to a priori kinship, following paper by N. E. Morton. In: Morton, N.E. (Ed.), *Genetic Structure of Populations*. University of Hawaii Press, Honolulu, pp. 78–79.
- Heywood, J.S., 1991. Spatial analysis of genetic variation in plant populations. *Annu. Rev. Ecol. Syst.* 22, 335–355.
- Knowles, P., 1991. Spatial genetic structure within two natural stands of black spruce [*Picea mariana* (Mill) B.S.P.]. *Silvae Genet.* 40, 13–19.
- Loiselle, B.A., Sork, V.L., Nason, J., Graham, C., 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* 82, 1420–1425.
- Maki, M., Masuda, M., 1993. Spatial autocorrelation of genotypes in a gynodioecious population of *Chionographis japonica* var. *kurohimensis* (Liliaceae). *Int. J. Plant Sci.* 154, 467–472.
- Malécot, G., 1955. Remarks on the decrease of relationship with distance. Following paper by M. Kimura. *Cold Spring Harbor Symp. Quant. Biol.* 20, 52–53.
- Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23.
- Morton, N.E., 1973a. Kinship and population structure. In: Morton, N.E. (Ed.), *Genetic Structure of Populations*. University of Hawaii Press, Honolulu, pp. 66–69.
- Morton, N.E., 1973b. Kinship bioassay. In: Morton, N.E. (Ed.), *Genetic Structure of Populations*. University of Hawaii Press, Honolulu, pp. 158–163.
- Morton, N.E., 1975. Kinship, fitness and evolution. In: Salzano, F.M. (Ed.), *The Role of Natural Selection in Human Evolution*. North-Holland, Amsterdam, pp. 133–154.
- Morton, N.E., 1982. Estimation of demographic parameters from isolation by distance. *Hum. Hered.* 32, 37–41.
- Oden, N.L., 1984. Assessing the significance of a spatial correlogram. *Geogr. Anal.* 16, 1–16.
- Peakall, R., Beattie, A.J., 1995. Does ant dispersal of seeds in *Sclerolaena diacantha* (Chenopodiaceae) generate local spatial genetic structure? *Heredity* 75, 351–361.
- Rajora, O.P., Rahman, M.H., Buchert, G.P., Dancik, B.P., 2000. Microsatellite DNA analysis of genetic effects of harvesting in old-growth eastern white pine (*Pinus strobus*) in Ontario, Canada. *Mol. Ecol.* 9, 339–348.
- Rohlf, F.J., Schnell, G.D., 1971. An investigation of the isolation-by-distance model. *Am. Nat.* 105, 295–324.
- Rousset, F., 2000. Genetic differentiation between individuals. *J. Evol. Biol.* 13, 58–62.
- Smouse, P.E., Peakall, R., 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* 82, 561–573.
- Sokal, R.R., 1979. Ecological parameters inferred from spatial correlograms. In: Patil, G.P., Rosenzweig, M.L. (Eds.), *Contemporary Quantitative Ecology and Related Econometrics*. International Cooperative Publishing House, Fairland, MD, pp. 167–196.
- Sokal, R.R., Oden, N.L., 1978. Spatial autocorrelation in biology. 1. Methodology. *Biol. J. Linn. Soc.* 10, 199–228.
- Sokal, R.R., Wartenberg, D.E., 1983. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* 105, 219–237.
- Sokal, R.R., Jacquez, G.M., Wooten, M.C., 1989. Spatial autocorrelation analysis of migration and selection. *Genetics* 121, 845–855.
- Sokal, R.R., Oden, N.L., Thomson, B.A., 1997. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biol. J. Linn. Soc.* 60, 73–93.
- Walter, R., Epperson, B.K., 2004. Microsatellite analysis of spatial structure among seedlings in populations of *Pinus strobus* (Pinaceae). *Am. J. Bot.*, in press.
- Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19, 395–420.