Heredity **82** (1999) 561–573

Received 18 September 1998, accepted 5 January 1999

Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure

PETER E. SMOUSE*† & ROD PEAKALL‡

†Department of Ecology, Evolution and Natural Resources and Center for Theoretical & Applied Genetics, Cook College, Rutgers University, New Brunswick, NJ 08901-8551, U.S.A. and ‡Division of Botany and Zoology, Australian National University, Canberra ACT 0200, Australia

Population genetic theory predicts that plant populations will exhibit internal spatial autocorrelation when propagule flow is restricted, but as an empirical reality, spatial structure is rarely consistent across loci or sites, and is generally weak. A lack of sensitivity in the statistical procedures may explain the discrepancy. Most work to date, based on allozymes, has involved pattern analysis for individual alleles, but new PCR-based genetic markers are coming into vogue, with vastly increased numbers of alleles. The field is badly in need of an explicitly multivariate approach to autocorrelation analysis, and our purpose here is to introduce a new approach that is applicable to multiallelic codominant, multilocus arrays. The procedure treats the genetic data set as a whole, strengthening the spatial signal and reducing the stochastic (allele-to-allele, and locus-to-locus) noise. We (i) develop a very general multivariate method, based on genetic distance methods, (ii) illustrate it for multiallelic codominant loci, and (iii) provide nonparametric permutational testing procedures for the full correlogram. We illustrate the new method with an example data set from the orchid *Caladenia tentaculata*, for which we show (iv) how the multivariate treatment compares with the single-allele treatment, (v) that intermediate frequency alleles from highly polymorphic loci perform well and rare alleles poorly, (vi) that a multilocus treatment provides clearer answers than separate single-locus treatments, and (vii) that weighting alleles differentially improves our resolution minimally. The results, though specific to *Caladenia*, offer encouragement for wider application.

Keywords: *Caladenia*, multilocus allozyme analysis, spatial autocorrelation.

Introduction

Studies of spatial structure in plant populations can reveal the operation of key evolutionary processes. When spatial structure develops, it may influence the patterns of local breeding and evolution. Population genetic theory predicts that plant populations will exhibit local population genetic structure when gene flow is restricted (Wright, 1943, 1978). In support of these models, Turner *et al.* (1982) showed with computer simulations that under restricted gene flow and selective neutrality, the population (as a whole) develops a patchy distribution of genotypes. Subsequent computer studies have confirmed that positive spatial autocorrelation, declining with distance, develops quickly under restricted gene flow (e.g. Sokal & Wartenberg, 1983; Sokal *et al.*, 1989; Epperson, 1990, 1995a,b; Sokal & Jacquez, 1991).

Contrary to theoretical expectations, however, spatial structure is rarely consistent across loci or sites, and when found, is generally weak, with many studies showing minor spatial structure on a strictly microspatial scale (Heywood, 1991). Schnabel *et al.* (1991) have found spatial autocorrelation at short distances for some allozyme loci in *Gleditsia triacanthos* and *Maclura pomifera*; Perry & Knowles (1991) found similar patterns in *Acer saccharum*, and weak spatial structure was reported in *Pinus banksiana* (Xie & Knowles, 1991), *Quercus laevis* (Berg & Hamrick, 1995) and *Psychotria officinalis* (Loiselle *et al.*, 1995). On the other hand, Waser (1987) found no apparent spatial autocorrelation at five allozyme loci in *Delphinium nelsonii*, although reliable estimates of pollen and seed dispersal suggest that it should have existed. Similarly, no spatial autocorrelation for allozyme loci has been found in *Pinus contorta* (Epperson & Allard, 1989), *Picea mariana* (Knowles, 1991), *Psychotria nervosa*

*Correspondence: Department of Ecology, Evolution and Natural Resources, Rm 152A, Environmental and Natural Resource Sciences Building, Cook College, Rutgers University, 14 College Farm Road, New Brunswick, NJ 08901, U.S.A. E-mail: smouse@aesop.rutgers.edu

(Dewey & Heywood, 1988) or *Carpa procera* (Doligez & Joly, 1997).

Several factors may explain why spatial structure is often weaker than anticipated. The usual *post hoc* explanation is that gene flow must have been greater than expected and sufficient to minimize local spatial structure (e.g. Dewey & Heywood, 1988; Epperson & Allard, 1989; Knowles, 1991; Doligez & Joly, 1997). An alternative is that spatial structure varies with life history stage, and that we are sometimes sampling the wrong cohort. In *Cecropia obtusifolia*, for example, seedlings show marked microspatial structure (spaced out on a cm × cm scale), but spatial structure declines among saplings (spaced out on a m × m scale), and is absent among adults (spaced 10s of m apart) (Epperson & Alvarez-Buylla, 1997).

A lack of sensitivity in the statistical procedures may also be a problem. Epperson (1995c) has shown that joint-count statistics for pairs of individuals are more sensitive to spatial structure than is the more familiar Moran *I*-statistic (Cliff & Ord, 1981), based on allele frequencies of sets of individuals. He has pointed out that widely cited simulations (e.g. Sokal & Wartenberg, 1983) employed a different computational procedure for Moran's *I* than is typically used for empirical work in natural populations. The denominator of the published Moran statistic leads to an underestimate of the autocorrelation under restricted gene flow. Appropriate adjustments (Epperson, 1995c) indicate that natural and simulated levels of spatial structure match more closely than had previously been realized.

Most spatial structure studies to date are based on allozymes. However, a series of new PCR-based genetic markers is becoming widely used in plant studies, in part because they are more numerous and more informative than allozymes. These include microsatellites (or simple sequence repeat markers), henceforth SSRs (Jarne & Lagoda, 1996), RAPDs – random amplified polymorphic DNA markers (Welsh & McClelland, 1990; Williams *et al.*, 1990) – and AFLPs – amplified fragment length polymorphic markers (Vos *et al.*, 1995). SSR markers are single-locus, codominantly inherited systems, not unlike allozymes, but given the often large number of alleles per locus, allele frequency spectra are sometimes dominated by numerous low-frequency alleles. The information per locus is considerably greater than for allozymes, and in selfers exceeds that for any other markers (e.g. Rongwen *et al.*, 1995). The RAPD and AFLP procedures yield multilocus DNA profiles with dominantly inherited polymorphisms. By virtue of the essentially unlimited number of primer combinations available, however, both RAPDs (Peakall *et al.*, 1995) and AFLPs (Powell *et al.*, 1996) represent more variable assay batteries than allozymes.

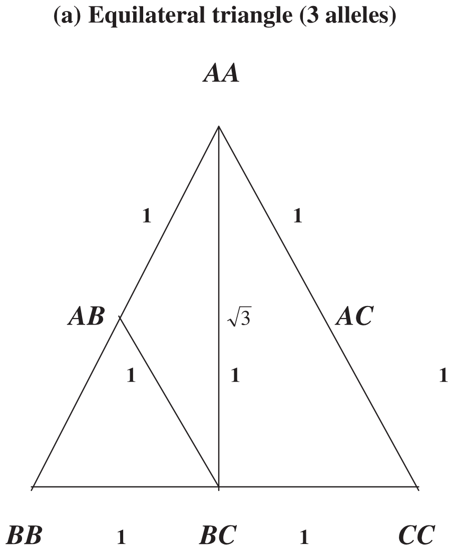
The procedures and publicly available software for investigating spatial genetic structure are not well designed for the large numbers of codominant alleles typical of many SSR loci, nor are they well designed to deal with multiple loci. Our purpose in this paper is to introduce a new approach to spatial genetic structure analysis that is applicable to multiallelic codominant, multilocus arrays, a method that can easily incorporate differential weighting of low-frequency alleles, if that is desired. The procedure is intrinsically multivariate, avoiding the need for allele-by-allele, locus-by-locus analysis, though such analyses can always be undertaken, if desired. By treating the genetic data set as a whole, we can strengthen the spatial signal by reducing the stochastic (allele-to-allele, and locus-to-locus) noise.

Our technical objectives are: (i) to develop a multiple-character, spatial autocorrelation analysis that is very general, based on genetic distance methods, (ii) to illustrate that treatment for multiallelic codominant loci, and (iii) to provide nonparametric permutational testing procedures for the full correlogram. We will illustrate these new methods with an example allozyme data set from *Caladenia tentaculata*, and in that context, we will address the following four questions. (i) How does the multiallele treatment compare with separate single-allele treatment, and which alleles convey most of the information, the common or rare alleles? (ii) How do different loci perform, and how is that related to allelic richness and the overall level of polymorphism? (iii) Is the multilocus treatment more powerful than are separate single-locus treatments, and do we lose information with the overarching treatment? (iv) What is the advantage (if any) of weighting alleles differentially, in terms of their respective frequencies?

Mathematical and statistical methods

Genetic distance measures

We begin by defining the genetic distance between a pair of individuals for a multiallelic codominant locus, such as one would have with either allozymes or microsatellite (SSR) markers; it is a slight modification of the scheme shown in Peakall *et al.* (1995). Consider a trio of codominant alleles (A, B, C) and a sextet of diploid genotypes, as shown in Fig. 1(a). The homozygotes (AA, BB, CC) are represented by the vertices of an equilateral triangle, and the linear distances between these vertices are scaled to be '2' units apart, measured along the side between adjacent vertices. The heterozygotes (AB, AC, BC) are positioned midway between the respective heterozygotes, and are a linear distance of '1' from each homozygote. From the geometry of the triangle, it is also clear that the linear


(a) Equilateral triangle (3 alleles)

(b) Additive scoring system (3 alleles)

Genetic $\underline{Y}_j$ - vector	Codominant genotype					
	AA	*AB*	*AC*	*BB*	*BC*	*CC*
y_A	2	1	1	0	0	0
y_B	0	1	0	2	1	0
y_C	0	0	1	0	1	2

Fig. 1 Additive distance metric for multiallelic, codominant loci; (a) three-allele (A, B, C) equilateral triangle, with linear distances indicated; (b) equivalent vector representation of a three-allele (A, B, C) equilateral triangle.

distance between heterozygotes sharing a single allele (say *AB* and *AC*) is 1, and that between any heterozygote and the opposite vertex homozygote (e.g. *AB* to *CC*) is √3. The squared distances between the various genotypes are thus 1 (*AB* to *AA*, *BB*, *AC* or *BC*; *AC* to *AA*, *CC*, *AB* or *BC*; *BC* to *BB*, *CC*, *AB* or *AC*), 3 (*AB* to *CC*, *AC* to *BB*, *BC* to *AA*), or 4 (*AA* to *BB*, *AA* to *CC*, *BB* to *CC*). The distance between any genotype and itself is clearly '0'. We will need *squared* distances for what follows, but the scoring scheme is 'additive'.

We can see the connection between the linear scoring convention and squared genetic distance by defining these distances in terms of paired vectors of three linear genetic variables, defined as shown in Fig. 1(b). The squared distance between any two genotypes is one-half the Euclidean distance between their respective vectors,

$$d_{ij}^2 = \frac{1}{2}\sum_{k=1}^{K} y_{ik} \quad y_{jk})^2, \tag{1}$$

where the subscript $k = 1, \ldots, K$ indexes the genetic (scoring) character. Just to illustrate, consider the two vectors for *AA* and *BC*. Equation 1 takes the form

$$d_{AA\text{ vs. }BC}^2 = \tfrac{1}{2}\Big[(2 \quad 0)^2 + (0 \quad 1)^2 + (0 \quad 1)^2\Big] = 3, \tag{2}$$

the same result we had from the triangle. We can extend to four alleles (equilateral tetrahedron or a vector of length 4), five alleles (equilateral pentahedron or a vector of length 5), and so on. The only thing that is new is the distance between two nonoverlapping heterozygotes, for example *AB* vs. *CD*, which is $d = \sqrt{2}$, in linear form, and $d^2 = 2$, in squared form. We are led to an interindividual distance matrix of the form shown in Table 1.

Unequal character weights

It has been suggested (Epperson, 1995c) that the information available from rare alleles is greater than

Table 1 Unweighted squared genetic distances for a four-allele, codominant locus

	AA	*AB*	*AC*	*AD*	*BB*	*BC*	*BD*	*CC*	*CD*	*DD*
AA	0	1	1	1	4	3	3	4	3	4
AB	1	0	1	1	1	1	1	3	2	3
AC	1	1	0	1	3	1	2	1	1	3
AD	1	1	1	0	3	2	1	3	1	1
BB	4	1	3	3	0	1	1	4	3	4
BC	3	1	1	2	1	0	1	1	1	3
BD	3	1	2	1	1	1	0	3	1	1
CC	4	3	1	3	4	1	3	0	1	4
CD	3	2	1	1	3	1	1	1	0	1
DD	4	3	3	1	4	3	1	4	1	0

that from common alleles, and that it should thus be weighted differentially. We can weight inversely by frequency in such a way that the squared distances shown in Table 1 still obtain for equifrequent alleles, as a special case. We require a simple change in the distance metric (eqn 1) to incorporate allele-specific weights:

$$d_{ij}^2 = \sum_{k=1}^{K} W_{kk} \; y_{ik} \quad y_{jk}\big)^2, \tag{3}$$

where the weights are inversely proportional to the allele frequencies (p_k) and the total number of alleles, and are given by

$$W_{kk} = (2Kp_k)\;^{1}, \tag{4}$$

with the subscript $k = 1, \ldots, K$ again indexing the genetic characters. A few examples from the four-allele case should suffice to show the pattern:

$$d^2_{AA \text{ vs. } BB} = \frac{1}{2K}\left[\frac{4}{p_A} + \frac{4}{p_B}\right] \qquad d^2_{AA \text{ vs. } AB} = \frac{1}{2K}\left[\frac{1}{p_A} + \frac{1}{p_B}\right]$$

$$d^2_{AB \text{ vs. } AC} = \frac{1}{2K}\left[\frac{1}{p_B} + \frac{1}{p_C}\right]$$

$$d^2_{AA \text{ vs. } BC} = \frac{1}{2K}\left[\frac{4}{p_A} + \frac{1}{p_B} + \frac{1}{p_C}\right] \tag{5}$$

$$d^2_{AB \text{ vs. } CD} = \frac{1}{2K}\left[\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_D}\right].$$

The weighting scheme gives the rare alleles more impact, on the premise that spatial proximity for two rare genotypes should carry more weight than proximity of two common genotypes.

The only difficulty with this weighting scheme is that the data used to assess spatial autocorrelation will probably also have to be used to establish the allele frequencies. We recommend substantial sample sizes, but even with large sample sizes, the precise frequencies of the rare alleles are not well established. The usual estimates of ($1/p_k$) are biased upwards for rare alleles, and the bias increases as the true frequency decreases, so we will follow Smouse & Chakraborty (1986) and Xu *et al.* (1994) in recommending less biased estimates, in this case

$$\tilde{p}_K^{\;\;1} = \frac{2N+1}{n_k + \frac{1}{K}}, \tag{6}$$

where n_k is the tally of the kth allele in the study population, and N is the total sample size. Unequal weighting of characters has no impact on the subsequent computations, as long as we use eqn (3). Whether weighting actually helps to detect autocorrelation is an empirical matter, to which we will return for the *Caladenia* illustration.

Multilocus treatment

Whether we have used weighted or unweighted distances, we can define an $N \times N$ genetic distance matrix, **D**, for a single locus, using the appropriate elements extracted from Table 1, or from the analogous treatment in eqn (5), a matrix that takes the form

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12}^2 & \ldots & d_{1N}^2 \\ d_{12}^2 & 0 & \ldots & d_{2N}^2 \\ \ldots & \ldots & \ldots & \ldots \\ d_{1N}^2 & d_{2N}^2 & \ldots & 0 \end{bmatrix}. \tag{7}$$

To obtain a multilocus distance, we simply add across loci. For the ith and jth individuals,

$$d_{ij}^2\ (L \text{ separate loci}) = \sum_{l=1}^{L} d_{ij}^2\ (l\text{th locus}), \tag{8}$$

where each locus is separately scored. We assume here that there are no missing genetic data for particular individuals, i.e. that the data set is complete. We could, with some sort of 'missing value' technique, adjust for holes in the data set, but that is beyond the scope of this effort. All later aspects of the analysis amount to manipulations of the elements of **D**, so the choice of metric is the aspect that really matters. Alternative distance measures can be envisaged and defined for these and other sorts of genetic markers, and we have developed alternatives for dominant/recessive RAPDs (Huff *et al.*, 1993) and haplotypic markers, such as seen with mtDNA assay (Excoffier *et al.*, 1992). We can even introduce some other weighting scheme. Suffice it, however, that as long as our chosen distance metric provides Euclidean closure, all the computations that follow are invariant with respect to that choice of metric.

The covariance matrix

The distance matrix **D** can be used to compute an inter-individual covariance matrix, **C**, which is what we will need for the autocorrelation analysis. The matrix **C** takes the form

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1N} \\ c_{21} & c_{22} & \ldots & c_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ c_{N1} & c_{N2} & \ldots & c_{NN} \end{bmatrix}, \tag{9}$$

where the diagonal terms measure the squared Euclidean distance from each individual genotype to the multivariate centroid of the genetic space,

$$c_{ii} = \sum_{m=1}^{M} W_{mm}(y_{im} \quad \bar{y}_m)^2, \tag{10}$$

where $m = 1, \ldots, M$ now indexes all the characters (alleles) of the multilocus set, and where the weights are either ½ or are given by eqn (4). Similarly, the inter-individual covariance terms provide a measure of the tendency of the ith and jth individuals to vary in the same (multidimensional) genetic direction from the centroid,

$$c_{ij} = \sum_{m=1}^{M} W_{mm}(y_{im} \quad \bar{y}_m)\; y_{jm} \quad \bar{y}_m)^2. \tag{11}$$

There is a convenient duality between genetic distance matrices and covariance matrices that makes our task simple. For the ith and jth individuals, Gower (1966) has shown that we can derive the squared distance, from eqns (10) and (11), via

$$d_{ij}^2 = c_{ii} + c_{jj} \quad 2c_{ij}. \tag{12}$$

In whatever way we choose to define our genetic characters, and whatever we choose as a weighting scheme, we can always convert the covariance matrix, **C**, into a corresponding genetic distance matrix, **D**, where the characters are defined in the same fashion. It is more convenient for us, however, to compute **D** directly, as described earlier, and then to effect a *back-translation* to **C**, using the 'centring' technique described by Gower (1966). Collapsing several algorithmic steps into one equation, we define c_{ij} as

$$c_{ij} = \left[\quad d_{ij}^2 + \left(\sum_{j=1}^{N} d_{ij}^2 + \sum_{i=1}^{N} d_{ij}^2 \right) \bigg/ N \quad \left(\sum_{i \neq j}^{N} d_{ij}^2 \right) \bigg/ N^2 \right] \bigg/ 2 \text{ for all } i \text{ and } j. \tag{13}$$

The first summation is over all N elements of the ith row; the second summation is over all N elements of the jth column; the third summation is over all N^2 elements of the matrix **D**, including the implicit diagonal zeroes. We have now converted our genetic characters, embedded in the multiallelic, multilocus distance matrix, **D**, into a corresponding genetic covariance matrix, **C**.

The $\mathbf{X}^{(h)}$-matrices

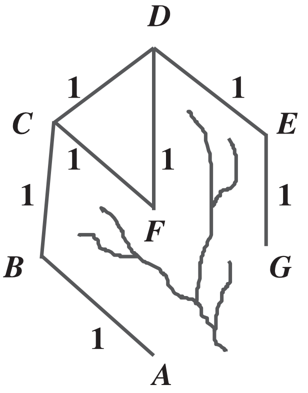
Corresponding to the covariance matrix, **C**, among individuals, we also need to define a set of corresponding $N \times N$ matrices for the spatial distances, $\mathbf{X}^{(h)}$, between pairs of individuals separated by h steps (or 'lags', as they are sometimes called in geospatial literature),

$$\mathbf{X}^{(h)} = \begin{bmatrix} x_{11}^{(h)} & x_{12}^{(h)} & \cdots & x_{1N}^{(h)} \\ x_{21}^{(h)} & x_{22}^{(h)} & \cdots & x_{2N}^{(h)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1}^{(h)} & x_{N2}^{(h)} & \cdots & x_{NN}^{(h)} \end{bmatrix}, \tag{14}$$

where $x_{ij}^{(h)} = 1$ for all pairs of individuals (i and j) that are h spatial distance classes apart (h lags apart), and $x_{ij}^{(h)} = 0$ otherwise. The diagonal element is the number of nonzero pairs involving the ith individual at distance (lag) h. We have a separate matrix for each class (h) of spatial separation. The essence of spatial autocorrelation analysis is to compare the elements of **C** with those of $\mathbf{X}^{(h)}$, for sets of paired observations that are h steps apart ($h = 1, \ldots, H$).

To consolidate the idea of a 'distance class', consider the seven-individual hypothetical population in Fig. 2(a). For purposes of illustration, let us suppose that genetic propagule (seed or pollen) movement is animal-mediated, and that the animal vector cannot cross water. The wavy line in Fig. 2(a) represents the stream network, so genetic movements (in our hypothetical example) must accommodate the drainage. Although the "line-of-sight" distances between points A and F, B and F, E and F, and F and G are small, those direct connections are less relevant than the travel paths that would have to be taken by the animal vectors, all routed through positions C or D. Although the physical distances between the adjacent (immediately connected) points are not precisely equal, we arbitrarily denote them as all being members of the first distance class ($h = 1$). That is to say, all such pairs are said to be 'one step (lag) apart'. The appropriate form for $\mathbf{X}^{(1)}$ is shown in Fig. 1(b). Individual A is adjacent only to B, so $x_{AA}^{(1)} = 1$, and $x_{AB}^{(1)} = 1 = x_{BA}^{(1)}$. On the other hand, individual B is adjacent to both A and C, so $x_{BB}^{(1)} = 2$, whereas $x_{BA}^{(1)} = 1 = x_{AB}^{(1)}$ and $x_{BC}^{(1)} = 1 = x_{CB}^{(1)}$. Because individual B is connected to both A and C at distance ($h = 1$), it is counted twice. We continue in this fashion for points C through G. For two-step connections ($h = 2$), A vs. C, B vs. D, B vs. F, C vs. E, F vs. E, D vs. G, the analogous treatment is shown in the matrix $\mathbf{X}^{(2)}$ of Fig. 1(b). Individuals A and G are only involved in two-step distances once each, so the Ath and Gth diagonal elements are '1', but all the other individuals

(a) Seven - position (*A* – *G*) spatial diagram



(b) $\mathbf{X}^{(h)}$ - matrices for *h* = 1, 2, 3, 4

$$\mathbf{X}^{(1)} = \begin{array}{|ccccccc|l} A & B & C & D & E & F & G & \\ \mathbf{1} & 1 & 0 & 0 & 0 & 0 & 0 & A \\ 1 & \mathbf{2} & 1 & 0 & 0 & 0 & 0 & B \\ 0 & 1 & \mathbf{3} & 1 & 0 & 1 & 0 & C \\ 0 & 0 & 1 & \mathbf{3} & 1 & 1 & 0 & D \\ 0 & 0 & 0 & 1 & \mathbf{2} & 0 & 1 & E \\ 0 & 0 & 1 & 1 & 0 & \mathbf{2} & 0 & F \\ 0 & 0 & 0 & 0 & 1 & 0 & \mathbf{1} & G \end{array} \qquad \mathbf{X}^{(2)} = \begin{array}{|ccccccc|l} A & B & C & D & E & F & G & \\ \mathbf{1} & 0 & 1 & 0 & 0 & 0 & 0 & A \\ 0 & \mathbf{2} & 0 & 1 & 0 & 1 & 0 & B \\ 1 & 0 & \mathbf{2} & 0 & 1 & 0 & 0 & C \\ 0 & 1 & 0 & \mathbf{2} & 0 & 0 & 1 & D \\ 0 & 0 & 1 & 0 & \mathbf{2} & 1 & 0 & E \\ 0 & 1 & 0 & 0 & 1 & \mathbf{2} & 0 & F \\ 0 & 0 & 0 & 1 & 0 & 0 & \mathbf{1} & G \end{array}$$

$$\mathbf{X}^{(3)} = \begin{array}{|ccccccc|l} A & B & C & D & E & F & G & \\ \mathbf{2} & 0 & 0 & 1 & 0 & 1 & 0 & A \\ 0 & \mathbf{1} & 0 & 0 & 1 & 0 & 0 & B \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 1 & C \\ 1 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & D \\ 0 & 1 & 0 & 0 & \mathbf{1} & 0 & 0 & E \\ 1 & 0 & 0 & 0 & 0 & \mathbf{2} & 1 & F \\ 0 & 0 & 1 & 0 & 0 & 1 & \mathbf{2} & G \end{array} \qquad \mathbf{X}^{(4)} = \begin{array}{|ccccccc|l} A & B & C & D & E & F & G & \\ \mathbf{1} & 0 & 0 & 0 & 1 & 0 & 0 & A \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 1 & B \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 0 & C \\ 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0 & D \\ 1 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & E \\ 0 & 0 & 0 & 0 & 0 & \mathbf{0} & 0 & F \\ 0 & 1 & 0 & 0 & 0 & 0 & \mathbf{1} & G \end{array}$$

Fig. 2 Translation of spatial field sample into connection matrices for an example of a hypothetical seven-individual collection in the headwaters of a small drainage: (a) portrayal of available single-step connections, disallowing traverse of the stream system; (b) the connection matrices, $\mathbf{X}^{(h)}$ for first ($h = 1$), second ($h = 2$), third ($h = 3$) and fourth ($h = 4$) distance classes. The matrix for $h = 5$ is not shown, but has '1's in all four corners.

are involved twice each in two-step connections, so their diagonal elements are each '2'. The three-and four-step connections are shown in $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(4)}$, respectively. There is also a single five-step connection, that between *A* and *G*, so $\mathbf{X}^{(5)}$ (not shown) has '1's in all four corners and '0's everywhere else. In this fashion, we can formally describe the connections between individuals separated by '*h* distance classes', for $h = 1, 2, 3, \ldots$, and so on.

The autocorrelation coefficient

For all pairs of individuals that are '*h* steps apart', we compute a correlation coefficient, according to the formula shown below. The two sets of genetic variables involved in a pairwise comparison of the *i*th and *j*th individuals are the same, however, with only their spatial location being different, so we describe the correlation as a *spatial autocorrelation* of individuals *h* steps apart (or 'at lag *h*').

$$r^{(h)} = \left(\sum_{i \neq j}^{N} x_{ij}^{(h)} c_{ij}\right) \Bigg/ \left(\sum_{i=1}^{N} x_{ii}^{(h)} c_{ii}\right), \tag{15}$$

where the numerator is the sum of all $N(N-1)$ off-diagonal 'element-by-element' products of $\mathbf{C}$ and $\mathbf{X}^{(h)}$, and the denominator is the sum of all N diagonal 'element-by-element' products of those same two matrices. The coefficient $r^{(h)}$ is a proper correlation coefficient, with a mean of '0' when there is no autocorrelation, and bounded by $[-1, +1]$, closely related to Moran's $-I^{(h)}$ coefficient, except that the $x_{ii}^{(h)}$ terms for Moran's coefficient are either '1' or '0', depending on whether the *i*th individual is (or is not) involved in pairs at distance class *h*, with no allowance for how many times the *i*th individual is paired with other individuals at that distance. Although the two coefficients have the same qualitative behaviour, we prefer $r^{(h)}$, because it is both completely general (multivariate) and a proper correlation. We compute an $r^{(h)}$-value for each matrix $\mathbf{X}^{(h)}$, and construct a 'correlogram' from the full series ($h = 1, \ldots, H$) of lags.

Sampling from the null distribution

Having computed an array of autocorrelations, one per distance class, we turn next to the question of how to test their departures from the null hypothesis that each is '0'. If there is no pattern of spatial genetic pattern within the site, then the null hypothesis should be that each of the correlations is drawn from a distribution of mean 0. If we represent the collection of H autocorrelations by a vector of length H, denoted $\mathbf{R} = [r^{(1)}, r^{(2)}, \ldots, r^{(H)}]^{\mathrm{T}}$, then the null hypothesis is that $\mathbf{R} = \mathbf{0}$. If we conclude from analysis of the data that the actual correlations at close distances are positive, we can also expect those at greater distances to be negative. The internal geometry of the covariance matrix $\mathbf{C}$ virtually guarantees it. We thus need to treat the vector $\mathbf{R}$ as a unit, representing all of the spatial inter-relationships in the data set.

To determine the null distribution for $\mathbf{R}$, we provisionally accept the (null hypothesis) view that genetic affinity and spatial affinity are not related at all, and we randomize spatial positions of the N separate individuals. That amounts to permuting ID numbers (the rows and corresponding columns) in the matrix $\mathbf{C}$, while holding those of the $\mathbf{X}^{(h)}$-matrices constant, effectively permuting whole genotypes amongst the spatial positions occupied. For the mth permutational shuffle, we compute a new array of estimated $r^{(h)}$-values, packed into a vector of length H, denoted $\mathbf{R}_m$. We permute $\mathbf{C}$ many times, say $(M - 1) = 999$, each time extracting the estimated vector of r-values, packed into $(M - 1)$ sample $\mathbf{R}_m = \{r_m^{(h)}\}$-vectors, all drawn from the same null distribution with no autocorrelation. On the premise that the genetic data were themselves drawn from that same null distribution, we use the estimated vector, $\mathbf{R}_G = \{r_G^{(h)}\}$, drawn from the actual positions of the individuals, as the Mth random realization.

To test the autocorrelation for a particular distance class, say the hth, we tally the M random estimates of $r^{(h)}$, extracting an estimated mean and standard deviation and generating an empirical confidence interval around that null-hypothesis mean. When testing for a positive autocorrelation (one-tailed test), we can simply tally the number of random $r^{(h)}$-values that are at least as large as that actually seen, i.e. $\hat{\alpha} = \mathrm{Tally}(r_m^{(h)} \geq r_G^{(h)})/M$. Because the actual value is also included in the random set, the tail probability will always be $\hat{\alpha} \geq M^{\ 1}$. If a two-tailed test is in order, because significant negative $r^{(h)}$-values are also deemed to be relevant, then we compute

$$t_h^2 = \frac{\left[r_G^{(h)} \quad \bar{r}_{\bullet}{}^{(h)}\right]^2}{s_h^2}, \tag{16}$$

where $\bar{r}_{\bullet}{}^{(h)}$ is the average of the randomized estimates, and where s_h^2 is their estimated variance. Substituting individual $r_m^{(h)}$-values into eqn (16), we compute an empirical null distribution for $t_{h,m}^2$, against which to compare the actual value, $t_{h,G}^2$, with $\hat{\alpha} = \mathrm{Tally}(t_{h,m}^2 \geq t_{h,G}^2)/M$. Alternatively, we can compare $r_G^{(h)}$ with a confidence interval spanning a desired fraction (say 95% or 99%) of the random values, $r_m^{(h)}$. No distributional assumptions are required for this nonparametric test.

It is customary (Morrison, 1976), when testing large numbers of correlations (several distance classes and many separate analyses of single alleles or single loci), to employ a Bonferroni probability criterion, assuming that all the test criteria are independent. Because we have included all loci and all alleles in a single genetic affinity measure, there is no need for a separate test of each allele and each locus, and the multiple-character analysis (measuring overall genetic affinity) is what we want. Moreover, the separate tests of H different $r_G^{(h)}$-values cannot be independent, given the internal Euclidean closure of the distance matrix, $\mathbf{D}$, and the consequent closure of the covariance matrix, $\mathbf{C}$. For an overall assessment of spatial autocorrelation, we test the null hypothesis, $\mathbf{R} = \mathbf{0}$, against the alternative hypothesis, $\mathbf{R} \neq \mathbf{0}$. If spatial autocorrelation is positive at some distances, it must be negative at others; the correlations at different lags are *not* independent. We need a test that allows for both types of simultaneous departures from the null model. We prefer a multivariate analogue of the univariate t^2-test in eqn (16).

From the collection of M null hypothesis $\mathbf{R}_m$-vectors (including the actual vector, $\mathbf{R}_G$), we compute a null mean vector,

$$\overline{\mathbf{R}} = \sum_{m=1}^{M} \mathbf{R}_m / M, \tag{17}$$

and a covariance matrix,

$$\mathbf{S}_R = \begin{bmatrix} \mathbf{s}_{11} & \mathbf{s}_{12} & \cdots & \mathbf{s}_{1H} \\ \mathbf{s}_{12} & \mathbf{s}_{22} & \cdots & \mathbf{s}_{2H} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{s}_{1H} & \mathbf{s}_{2H} & \cdots & \mathbf{s}_{HH} \end{bmatrix}. \tag{18}$$

We invert this covariance matrix, and compute

$$T_m^2 = \left[\mathbf{R}_m \quad \overline{\mathbf{R}}\right]^{\mathrm{T}} \mathbf{S}_R^{\ 1} \left[\mathbf{R}_m \quad \overline{\mathbf{R}}\right] \tag{19}$$

which we compare with the distribution of similarly computed values from the random vectors to determine how often a random (null-hypothesis) autocorrelation vector deviates as far from the multivariate centroid as does the observed vector, with $\hat{\alpha} = \mathrm{Tally}(T_m^2 \geq T_G^2)/M$. This multiclass test is patterned on the classic T^2-test (Hotelling, 1951), but we rely on permutational testing, rather than parametric theory. The T^2-criterion measures departures from the $\mathbf{R} = \mathbf{0}$ hypothesis for all distance classes simultaneously, and thus provides a global test of spatial genetic clustering that allows for all of the interdependencies within the spatial and genetic data sets.

An illustration with *Caladenia tentaculata*

Background

In a detailed study, Peakall & Beattie (1996) have examined the ecological and genetic consequences of exclusive pollination by sexually attracted male thynnine wasps in the orchid *Caladenia tentaculata* (Schldl). Pollination in this species is typical of the many Australian orchids, exploiting the reproductive behaviour of thynnine wasps by production of pheromone-like fragrances and presentation of labellum structures that mimic the female. Pollination occurs when male wasps attempt copulation (pseudocopulation) with the labellum. After pollination, wasps immediately leave the patch, rather than visiting additional plants within the patch. As a consequence of this behaviour, pollen movements approximate a linear (rather than leptokurtotic) distribution, with a mean dispersal distance of 17 m (max. = 58 m). This is among the largest mean pollen dispersal distances known in a herbaceous plant (Peakall & Beattie, 1996).

Despite extensive pollen flow in *C. tentaculata*, an analysis of five allozyme loci revealed significant genetic clustering within a 20 × 40 m quadrat. The existence of spatial structure, in the presence of homogenizing pollen flow, may be a consequence of restricted seed dispersal instead. Although *C. tentaculata*'s minute seeds are wind-dispersed, most of the seed probably fall close to the parent, where the chances of seedling establishment may be enhanced by the presence of mycorrhizal fungi that are required for germination.

Here, we revisit the data set of Peakall & Beattie (1996), demonstrating various features of our new method of spatial autocorrelation analysis. First, we use the method to compare the spatial structure of single alleles, comparing the results for rare and polymorphic alleles, along with those of a multiallelic analysis for the same locus. We then compare multiallele analyses for the different loci, relating the results to the levels of polymorphism for single loci, and contrast the results with those of a full multilocus analysis. Finally, we examine the impact of differential weights for common and rare alleles, to see whether weighting clarifies the situation by emphasizing the detectable pattern from the rarer alleles.

Sampling and genetic analysis

This spatial autocorrelation analysis focused on a 20 × 40 m plot, embedded within a larger population of *C. tentaculata*, a plot orientated in such a way that a maximum number of orchids was contained within it. All flowering plants within the plot were sampled, and allozyme analysis of 16 enzyme systems was performed, yielding 22 putative loci. Of these 22 loci, only the five most polymorphic loci were used for the subsequent analysis: *Mr*, *Pgm*, *Mdh-1*, *Got-3* and *Pgi-2* (Peakall & Beattie, 1996). Table 2 shows the allele frequencies of the loci, listed in order of decreasing heterozygosity (H_e). Because our multilocus genetic distance is not currently programmed to deal with missing genotypes, the final data set used for this paper consists of only the 384 plants for which the genotypes were known for all loci. Thus, the allele frequencies reported here differ slightly from those published in Peakall & Beattie (1996). Figure 3 shows the distribution of the 384 plants within the 20 × 40 m plot.

Microgeographical distance classes of 1 m were used in all analyses, so that the first distance class included all distances in the (0,1) interval, (1,2) interval, and so on. Tests of significance were computed for each distance class by comparing the observed value of $r_G^{(h)}$ with those obtained from 999 spatial permutations of the 384 sampled plants. The observed data were added as the 1000th permutation, on the null premise that the actual data were genetically random with respect to spatial position. The 1000 sets of pseudocorrelations were then sorted, and a 95% confidence interval was constructed from the 975th value and 25th value, respectively. We assessed significance of a single correlation with the two-tailed t_h^2-criterion, as computed from eqn (16), and that of the correlogram as a whole with the two-tailed T^2-criterion, eqn (19). We would remind the reader at this juncture that if we wished to test all of these separate t_h^2 criteria, as though they were independent, a Bonferroni correction would be in order, but because our preference is for a single test of the total correlogram, via T^2 (see below), we will ignore that nuance here.

Table 2 Allele frequencies and expected heterozygosities (gene diversities) within the plot of *Caladenia tentaculata* for the five polymorphic allozyme loci

Genetic locus	Number of alleles	Allele *1*	Allele *2*	Allele *3*	Allele *4*	Expected heterozygosity
Mr	4	0.283	0.441	0.185	0.090	0.683
Pgm	4	0.067	0.706	0.069	0.158	0.467
Mdh-1	3	0.208	0.709	0.084	–	0.447
Got-3	2	0.816	0.184	–	–	0.300
Pgi-2	4	0.021	0.017	0.946	0.016	0.103

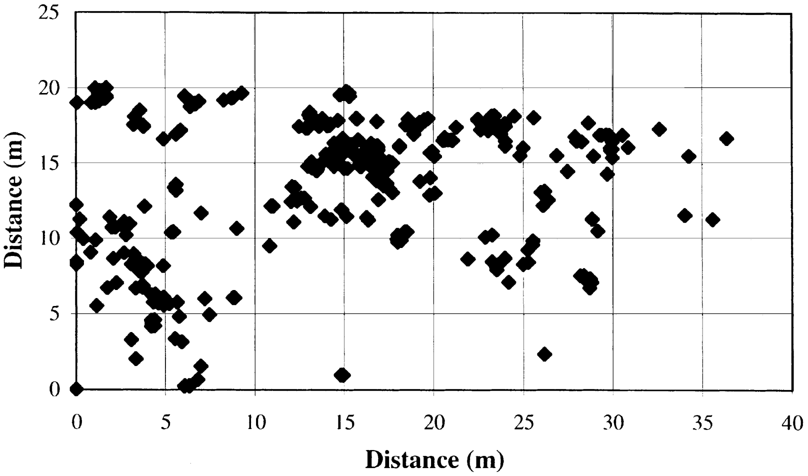


Fig. 3 Distribution of 384 *Caladenia tentaculata* plants sampled within a 20 m × 40 m plot.

Single-allele vs. multiallele treatment

Classical spatial autocorrelation analysis of genetic frequencies is executed one allele at a time, and our treatment is multiallelic. There is nothing to prevent us, however, from examining each allele separately, and we illustrate with the *Mr* locus. Figure 4 shows the correlograms for each of the *Mr* alleles, for a comparison with the correlogram for a multiallelic treatment of the locus. Mr^2 showed the strongest spatial structure, with $r^{(1)} = 0.175$, and with $r^{(h)}$ remaining significantly positive out to 7 m. Mr^1 shows the same trend, and Mr^4 shows a similar (but weaker) trend. Interestingly, no significant spatial structure was detected for Mr^3, although it occurred at higher frequencies than Mr^4 (Table 2). In general, we are interested in the demographic (propagule flow) processes causing autocorrelation. For those purposes, individual alleles are best viewed as replication, and their differences are best viewed as the stochastic consequences of random genetic sampling of alleles, scattered across a spatial landscape. The multiallele treatment shows the essential pattern of spatial genetic affinity, allows for the correlations among alleles, extracts all the information, and (by virtue of averaging across these discrete genetic variables) 'smoothes out the bumps' in the correlogram. The variation among alleles (the 'noise') is reduced and the pattern of genetic affinity (the 'signal') is clear.

Single-locus vs. multilocus analysis

We show a series of separate, unweighted correlograms for each locus and one for the multilocus treatment in Fig. 5. All five loci showed positive values of $r^{(h)}$ in the first three to seven distance classes, and (with the exception of *Pgi-2*) significant spatial structure was apparent for the shortest distance classes, as shown by the values of the $r^{(h)}$-values that exceeded the upper 97.5% confidence limits. The most striking result was that for *Mr*, with $r^{(h)}$ remaining significantly positive out to 7 m (i.e. seven steps, or seven 'lags'). Oscillation of the correlogram between positive and negative values is apparent for *Mr*, beyond 7 m, and to a lesser extent for *Got-3* as well, consistent with a pattern of strong microspatial structure. The multilocus correlogram shows a similar pattern to that for *Mr*, but with a somewhat smoother curve. The point at which the curve first crosses the x-axis provides an estimate of patch size (Sokal & Wartenberg, 1983), which

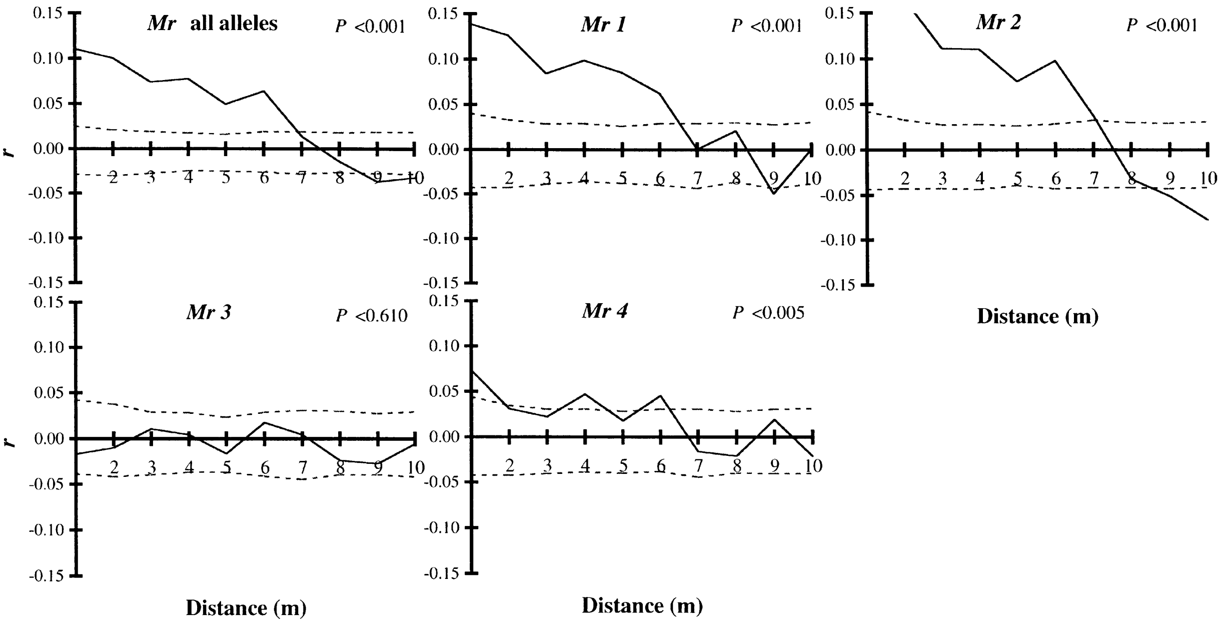


Fig. 4 Unweighted correlograms (solid lines) for each of the *Mr* alleles and the corresponding multiallelic (single-locus) correlogram, with 95% null hypothesis confidence regions indicated by dotted lines; probabilities of total correlogram test criterion t^2 (for individual alleles) or T^2 (for the full locus) are shown in the top right corner.

is between 2 and 7 m. The overall departure from the null hypothesis of no spatial genetic clustering is resoundingly rejected for the multilocus treatment.

With the exception of *Got-3*, the strength of autocorrelation decreased with decreasing heterozygosity (cf. Table 2), ranging from high in *Mr* ($H_e = 0.68$) to low and not statistically different from zero in *Pgi-2* ($H_e = 0.10$), showing that the strength of the spatial signal increases with the level of polymorphism, rather than the number of alleles *per se*. Although *Pgi-2* exhibited four alleles, the most common had a frequency close to 0.95, our arbitrary criterion for exclusion of a

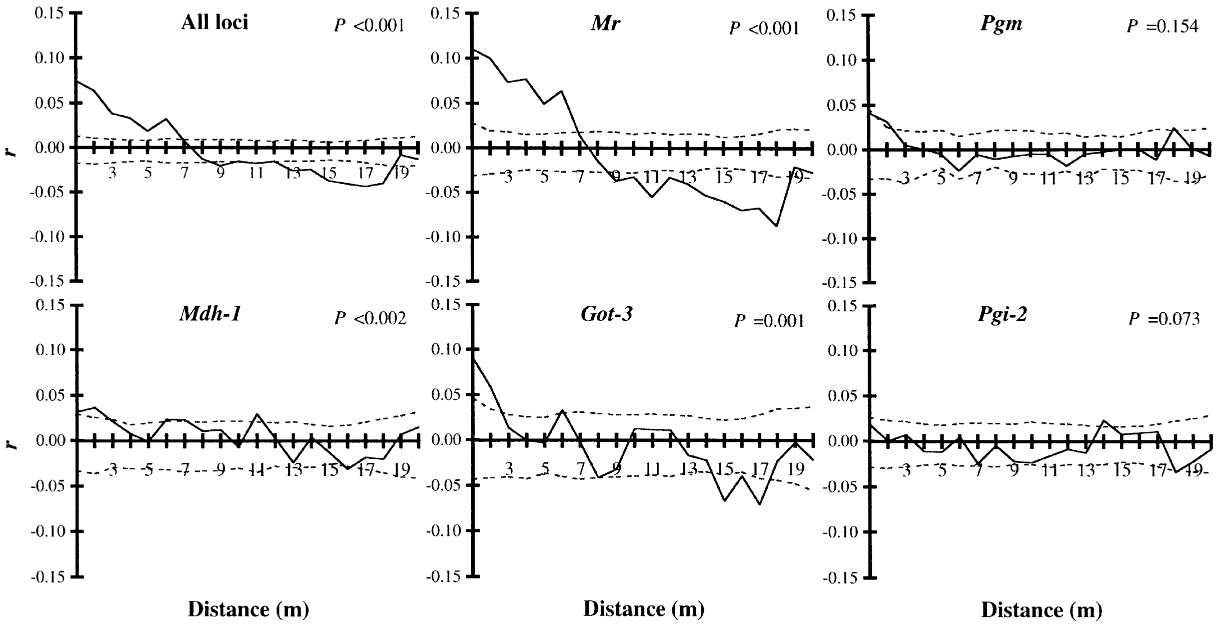


Fig. 5 Unweighted correlograms (solid lines) for each locus separately, and the corresponding multilocus correlogram, with 95% null hypothesis confidence regions indicated by dotted lines; probabilities of total correlogram T^2 are shown in the top right corner.

locus from analytical consideration (Table 2). None of the individual alleles showed any pattern for this locus (not shown). There is little to be gained by attempting to detect spatial pattern for a locus that shows almost no variation!

Weighted vs. unweighted analysis

One possible reason for the absence of autocorrelation for essentially invariant loci is that there are few useful comparisons in the data set, so that even if strong spatial structure were to exist, it would be difficult to detect. One possible solution to this problem is to weight the pairwise genetic distances inversely by allele frequency, as described in the Methods section. We evaluated this option by using the codominant weighting scheme in eqn (5). The analytical results of inverse frequency weighting are shown in Fig. 6. Empirically, weighting had little effect on the outcome, except for *Mdh-1* (compare *Mdh-1* in Figs 5 and 6). For *Mdh-1*, the weighted analysis indicated stronger autocorrelation than was evident from the unweighted analysis, with $r^{(1)}$ substantially above the 97.5% confidence interval and $r^{(h)}$ remaining high and mostly significant out to 12 m. The weighted results for *Mdh-1* reinforced the finding of strong and significant spatial autocorrelation for the other loci. Nevertheless, the overall interpretation from the weighted analysis is identical to that from the unweighted analysis. With the multilocus approach, we have the advantages of considerable averaging. Weighting creates no particular problems, but it does not seem to offer much additional resolution.

Extensions and generalizations

Other types of genetic markers

Although our genetic distance metric in Table 1, and its weighted analogue, eqn (5), are designed specifically for codominant diploid allozyme or SSR loci, there is nothing that restricts us to such markers. With a proper distance metric, we can examine any type of genetic marker available. Elsewhere, we (Huff *et al.*, 1993) have shown how to compute genetic distances for dominant/recessive markers, such as are generated by RAPD (Welsh & McClelland, 1990; Williams *et al.*, 1990) and AFLP (Vos *et al.*, 1995) methods. Both RAPDs and AFLPs can be used in any species, without DNA sequence knowledge. By virtue of the essentially unlimited number of primer combinations available, both RAPDs (Peakall *et al.*, 1995) and AFLPs represent more variable assay batteries than allozymes. With band presence scored as ‘1’ and its absence as ‘0’, the genetic distance between two individuals is a simple tally of the ‘1’ vs. ‘0’ (different bands) tally for those individuals. Unequal weighting for the different loci is easy to add if desired. Given a proper distance metric, of course, everything else is the same.

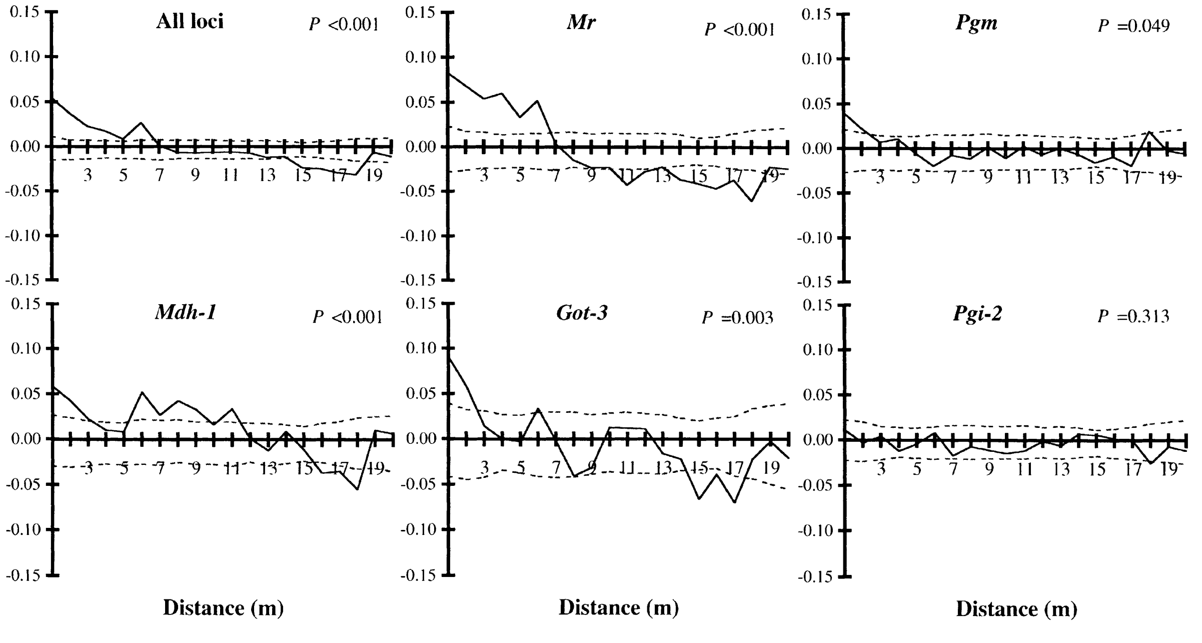


Fig. 6 Weighted correlograms (solid lines) for each locus separately, and the corresponding multilocus correlogram, with 95% null hypothesis confidence regions indicated by dotted lines; probabilities of total correlogram T^2 are shown in the top right corner.

Relationship to competing methodological alternatives

The distance-based autocorrelation methods described here are related to innovations by earlier workers, and a few words are in order to put this current effort into broader perspective. Sokal *et al.* (1986) adapted a distance-based version of autocorrelation analysis. More recently, Sokal (see Burgman & Williams, 1995) has formalized that treatment and labelled the outcome a 'Mantel Correlogram', which is basically a comparison of the genetic distance matrix, **D**, with a matrix of counter-variables, patterned after $\mathbf{X}^{(h)}$. The method is very similar to that described here, based on a correlational coefficient, but uses **D** and **X**, rather than **C** and **X**.

Bertorelli & Barbujani (1995) have introduced distance-based versions of both Moran's *I*- and Geary's *c*-coefficients, labelled AIDA (autocorrelation indices for DNA analysis) coefficients, for the case of linked haplotypes, using either the phenetic or phyletic metrics that we introduced elsewhere (Excoffier *et al.*, 1992). Barring a difference in denominator, their treatment is a binary haploid analogue of the multistate diploid treatment introduced here.

The patterns of spatial structure in *C. tentaculata* revealed by our new method are qualitatively similar to those in Peakall & Beattie (1996), whose multilocus treatment followed a technique developed by J. Nason (see Loiselle *et al.*, 1995), based on an estimate of Wright's relatedness coefficient ρ. The estimated value, denoted r_{ij}, but defined differently from our eqn (15), is a frequency-weighted average over all alleles and loci. The difference here is that we have allowed for the covariances between alleles, which follow from the fact that the allele frequencies at any one locus must sum to unity. Epperson (1995c) has developed a 'joint count' distance, a frequency-weighted measure that reduces each locus to 2-allele form. Each of these methods represents a laudable attempt to allow for inherent information differences that arise with alleles of different frequencies. Our treatment, embodied in eqns (3)–(5), is a more general codominant weighting scheme. Analogues for dominant/recessive loci, as well as for haplotypic marker sets, are easily constructed and fairly obvious. As mentioned earlier, given the choice of distance metric, everything else stays the same.

Beyond technical particulars, we have introduced a generic multivariate method for spatial autocorrelation analysis of genetic data on individuals, requiring only a proper distance metric (and perhaps a weighting scheme). This adds a method for spatial analysis to the battery of tests that can be accessed routinely from the interindividual genetic distance matrix, joining such standard tools as AMOVA (Excoffier *et al.*, 1992; Huff *et al.*, 1993; Peakall *et al.*, 1995) and Matrix Correlation Analysis (Mantel, 1967; Sneath & Sokal, 1973; Smouse *et al.*, 1986; Smouse & Long, 1992). The primary value of generic methods is that they can be used on different sorts of data, a feature that will increase in importance as new genetic methodologies are added to our assay battery. With a truly multivariate approach, we have the additional advantages of averaging over stochastically varying systems. For most purposes, we do not really need a separate test for every allele, or even a separate test for each locus. What we need is one test that addresses the question: 'Are genotypes in close physical proximity any more similar than those with greater physical separation?'. A multivariate approach provides that, both in the form of the vector $\mathbf{R}_G$, and its associated multivariate test criterion, T_G^2.

Acknowledgements

The authors wish to thank S. Kaufman, S. Krauss and a pair of anonymous reviewers for copious commentary. P.E.S. was supported by the New Jersey Agricultural Experiment Station, NJAES/USDA-32106, and R.P. was supported by the Australian Research Council Large Grants Scheme and the Australian National University. This work was initiated while the senior author was a Visiting Fellow in the Division of Botany and Zoology, Australian National University, and he wishes to thank the faculty, staff and students for their collegiality during his visit.

References

BERG, E. E. AND HAMRICK, J. L. 1995. Fine-scale genetic structure of a Turkey oak forest. *Evolution*, **49**, 110–120.

BERTORELLI, G. AND BARBUJANI, G. 1995. Analysis of DNA diversity by spatial autocorrelation. *Genetics*, **140**, 811–819.

BURGMAN, M. A. AND WILLIAMS, M. R. 1995. Analysis of the spatial pattern of arthropod fauna of jarrah (*Eucalyptus marginata*) foliage using a mantel correlogram. *Aust. J. Ecol.*, **20**, 455–457.

CLIFF, A. D. AND ORD, J. K. 1981. *Spatial Processes. Models and Applications*. Pion Ltd., London.

DEWEY, S. E. AND HEYWOOD, J. S. 1988. Spatial genetic structure in a population of *Psychotria nervosa* I. Distribution of genotypes. *Evolution*, **42**, 834–838.

DOLIGEZ, A. AND JOLY, H. I. 1997. Genetic diversity and spatial structure within a natural stand of a tropical forest tree species, *Carpa procera* (Meliaceae), in French Guiana. *Heredity*, **79**, 72–82.

EPPERSON, B. K. 1990. Spatial autocorrelation of genotypes under directional selection. *Genetics*, **124**, 757–771.

EPPERSON, B. K. 1995a. Spatial distributions of genotypes under isolation by distance. *Genetics*, **140**, 1431–1440.

EPPERSON, B. K. 1995b. Spatial structure of two-locus genotypes under isolation by distance. *Genetics*, **140**, 365–375.

EPPERSON, B. K. 1995c. Fine-scale spatial structure – correlations for individual genotypes differ from those for local gene frequencies. *Evolution*, **49**, 1022–1026.

EPPERSON, B. K. AND ALLARD, R. W. 1989. Spatial autocorrelation analysis of the distribution of genotypes within populations of lodgepole pine. *Genetics*, **121**, 369–377.

EPPERSON, B. K. AND ALVAREZ-BUYLLA, E. R. 1997. Limited seed dispersal and genetic structure in life stages of *Cecropia obtusifolia*. *Evolution*, **51**, 275–282.

EXCOFFIER, L., SMOUSE, P. E. AND QUATTRO, J. M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction sites. *Genetics*, **131**, 479–491.

GOWER, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

HEYWOOD, J. S. 1991. Spatial analysis of genetic variation in plant populations. *Ann. Rev. Ecol. Syst.*, **22**, 335–355.

HOTELLING, H. 1951. A generalized T test and measure of multivariate dispersion. Proc. Second Berkeley Symp. Math. Stat. Prob., pp. 23–41. University of California Press, Berkeley.

HUFF, D. R., PEAKALL, R. AND SMOUSE, P. E. 1993. RAPD variation within and among populations of outcrossing buffalograss (*Buchloë dactyloides* (Nutt.) Engelm). *Theor. Appl. Genet.*, **86**, 927–934.

JARNE, P. AND LAGODA, P. J. L. 1996. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol*, **11**, 424–429.

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, R., SORK, V. AND NASON, J. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.*, **82**, 1420–1425.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.*, **27**, 209–220.

MORRISON, D. F. 1976. *Multivariate Statistical Methods*, 2nd edn. McGraw-Hill, New York.

PEAKALL, R. AND BEATTIE, A. J. 1996. Ecological and genetic consequences of pollination by sexual deception in the orchid *Caladenia tentaculata*. *Evolution*, **50**, 2207–2220.

PEAKALL, R., SMOUSE, P. E. AND HUFF, D. R. 1995. Evolutionary implications of allozyme and RAPD variation in diploid populations of Buffalograss (*Buchloë dactyloides* (Nutt.) Engelm.). *Mol. Ecol.*, **4**, 135–147.

PERRY, D. J. AND KNOWLES, P. 1991. Spatial genetic structure within three sugar maple (*Acer saccharum* Marsh.) stands. *Heredity*, **66**, 137–142.

POWELL, W., MORGANTE, M., ANDRE, C., HANAFEY, M., VOGEL, J., TINGEY, S. AND RAFALSKI, A. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.*, **2**, 225–238.

RONGWEN, J., AKKAYA, M. S., BHAGWAT, A. A., LAVI, U. AND CREGAN, P. B. 1995. The use of microsatellite DNA markers for soybean genotype identification. *Theor. Appl. Genet.*, **90**, 43–48.

SCHNABEL, A., LAUSHMAN, R. H. AND HAMRICK, J. L. 1991. Comparative genetic structure of two co-occurring tree species, *Maclura pomifera* (Moraceae) and *Gleditsia triacanthos* (Leguminosae). *Heredity*, **67**, 357–364.

SMOUSE, P. E. AND CHAKRABORTY, R. 1986. The use of restriction fragment length polymorphisms in paternity analysis. *Am. J. Hum. Genet.*, **38**, 918–939.

SMOUSE, P. E. AND LONG, J. C. 1992. Matrix correlation analysis in anthropology and genetics. *Yearbook Phys. Anthropol.*, **35**, 187–213.

SMOUSE, P. E., LONG, J. C. AND SOKAL, R. R. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.*, **35**, 627–632.

SNEATH, P. H. A. AND SOKAL, R. R. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

SOKAL, R. R. AND JACQUEZ, G. M. 1991. Testing inferences about microevolutionary processes by means of spatial autocorrelation. *Evolution*, **45**, 152–168.

SOKAL, R. R. AND WARTENBERG, D. E. 1983. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, **105**, 219–237.

SOKAL, R. R., SMOUSE, P. E. AND NEEL, J. V. 1986. The genetic structure of a tribal population, the Yanomama Indians. XV. Patterns inferred by autocorrelation analysis. *Genetics*, **114**, 259–287.

SOKAL, R. R., JACQUEZ, G. M. AND WOOTEN, M. C. 1989. Spatial autocorrelation analysis of migration and selection. *Genetics*, **121**, 845–855.

TURNER, M. E., STEPHENS, J. C. AND ANDERSON, W. W. 1982. Homozygosity and patch structure in plant populations as a result of nearest-neighbor pollination. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 203–207.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., VAN-DE-LEE, T., HORNES, M. *ET AL.* 1995. AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.*, **23**, 4407–4414.

WASER, N. M. 1987. Spatial genetic heterogeneity in a population of the montane perennial plant *Delphinium nelsonii*. *Heredity*, **58**, 249–256.

WELSH, J. AND MCCLELLAND, M. 1990. Fingerprinting genomes using PCR with arbitrary primers. *Nucl. Acids Res.*, **18**, 7213–7218.

WILLIAMS, J. G. K., KUBELIK, A. R., LIVAK, K. J., RAFALSKI, J. A. AND TINGEY, S. V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.*, **18**, 6531–6535.

WRIGHT, S. 1943. Isolation by distance. *Genetics*, **28**, 114–138.

WRIGHT, S. 1978. *Evolution and the Genetics of Populations*, vol. 4, *Variability Within and Among Natural Populations*. University of Chicago Press, Chicago.

XIE, C. Y. AND KNOWLES, P. 1991. Spatial genetic substructure within natural populations of jack pine (*Pinus banksiana*). *Can. J. Bot.*, **69**, 547–551.

XU, S., KOBAK, C. AND SMOUSE, P. 1994. Constrained least squares estimation of mixed population stock composition from mtDNA haplotype frequency data. *Can. J. Fish. Aquat. Sci.*, **51**, 417–425.