

# On methods of spatial analysis for genotyped individuals

K Shimatani<sup>1</sup> and M Takahashi<sup>2</sup>

<sup>1</sup>The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato, Tokyo 106-8569, Japan; <sup>2</sup>Forest Tree Breeding Center, 3809-1, Ishi, Juo, Taga, Ibaraki 319-1301, Japan

Spatial autocorrelation methods have commonly been applied to individual-based spatial genetic studies, although their properties and the relations among the statistics have not been carefully examined. This paper first introduces a reformulation of widely used spatial statistics using point processes. When Moran's *I* statistics are applied to allele frequencies within an individual, the frequencies are no longer continuous variables but have only three discrete values and specific interpretations of Moran's *I* statistics and the number of alleles in common (NAC) can be expressed as the weighted sum of join-count statistics. The distributions of

minor genotypes are amplified in Moran's *I* depending on the allele frequency in the population, while NAC uses a constant weighting system. Under the point process framework, spatial analysis can be conducted on the common theoretical base, from individual locations to genetic distributions of different levels, (for example, genotype and allele). The methodology is demonstrated by application to field data for molecular ecological studies of *Fagus crenata* population dynamics.

*Heredity* (2003) 91, 173–180. doi:10.1038/sj.hdy.6800295

**Keywords:** join-count statistics; Moran's *I* statistics; NAC; point process; spatial autocorrelation; spatial genetics

## Introduction

Spatial autocorrelation analysis is currently used as a standard statistical technique for analysing individual-based spatial genetic structure from mapped data of genotyped individuals. The commonly used methods can be classified into two categories depending on the treatment of the genotypic data (Heywood, 1991). One method considers genotypes as nominal data and applies join-count statistics (ie, standard normal deviate, SND). The other, first transforms genotypic data into allele frequencies and applies Moran's *I* statistics to interval data. With this method, each individual is considered as a population, thus, its allele frequency is 1 if it homozygous for a specific allele, 0.5 if heterozygous, and 0 otherwise. Some studies have used the latter method (eg, Xie and Knowles, 1991; Geburek and Tripp-Knowles, 1994; Streiff *et al*, 1998; Ueno *et al*, 2000), others have used both (eg, Leonardi and Menozzi, 1996; Chung and Epperson, 2000), and some have calculated additional statistics, such as the number of alleles in common (NAC; Berg and Hamrick, 1995; Takahashi *et al*, 2000) and coancestry (Loisselle *et al*, 1995). In each case, pairs of individuals are classified into distance classes and statistics are calculated for every distance class.

Join-count statistics for short distance classes directly indicate whether a specific genotype is clustered and whether two genotypes are attracting or repulsing. In contrast, large positive Moran's *I* values for short

distance classes are generally interpreted as a tendency for neighbouring individuals to have a 'similar' allele frequency. When Sokal and Oden (1978) introduced spatial autocorrelation methods, Moran's *I* statistics were calculated for the allele frequencies of populations investigated. In this case, the allele frequencies are continuous variables, thus a general interpretation of correlations is feasible; positive correlations are present when the two variables tend to show similar values. However, at the individual level, the frequency can no longer be continuous but takes three discrete values.

In previous studies, when both Moran's *I* and join-count (and other spatial statistics) were calculated, the two statistics were not simultaneously interpreted. Although Epperson (1995) pointed out that Moran's *I* is a weighted sum of join-count statistics, no study has directly applied this relations to field data. In addition, some studies analysed the spatial distribution of individuals, although separately from genotypic distribution (eg, Berg and Hamrick, 1995; Ueno *et al*, 2000).

This paper introduces a methodology that uses join-count statistics, Moran's *I* statistics for within-individual frequencies, and other spatial statistics together with the spatial distribution of individuals. Beginning with a brief review of conventional spatial autocorrelation methods in genetics, the first part explains point processes, which have been commonly applied in individual-based spatial ecology and play an important role in simultaneous analysis of the spatial distribution of individuals and genotypes. The next part reformulates conventional spatial statistics using the language of point processes (Shimatani, 2002). The reformulated Moran's *I* for within-individual frequencies and NAC can be expressed as the weighted sum of the reformulated join-count

Correspondence: K Shimatani, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato, Tokyo 106-8569, Japan.  
E-mail: shimatan@ism.ac.jp

Received 31 May 2002; accepted 25 February 2003

statistics, thus providing an insight into the interpretation of autocorrelations between the three discrete values and clarifying the relation between the measures. The methodology is demonstrated by application to field data of a *Fagus crenata* population taken from Takahashi *et al* (2000). The final part discusses the biological implications of the analysis and the utility of the methods for population genetics and molecular ecology.

## Spatial statistics

### Conventional statistics for genetics

Suppose that the mapped data of genotyped individuals are given. Let  $\{X_i\}$  ( $i = 1, 2, \dots, n$ ) be the  $x$ - $y$  coordinates of individual  $i$ . Conventionally, spatial autocorrelation techniques are applied as follows. First, Euclidean distances are calculated between the individuals and divided into distance classes of width  $2\Delta$  as  $(0, 2\Delta]$ ,  $(2\Delta, 4\Delta]$ ,  $(4\Delta, 6\Delta]$ , ... . For each pair of individuals  $(i, j)$ , weight  $W_{i,j}[r]$  is given for discrete distances of  $r = \Delta, 3\Delta, 5\Delta, \dots$  depending on their interdistance  $\|X_i - X_j\|$  as

$$W_{i,j}[r] = \begin{cases} 1 & \text{if } r - \Delta < \|X_i - X_j\| \leq r + \Delta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

(1) *Moran's I statistics for within-individual frequencies*: Fix one locus and one allele, named  $A$ . Let  $a(i)$  be the allele frequency within individual  $i$ , namely,  $a(i) = 1$  if individual  $i$  is homozygous for allele  $A$ ,  $a(i) = 0.5$  if heterozygous for  $A$ , and  $a(i) = 0$  otherwise. Let  $\bar{a}$  be the (estimated) frequency of allele  $A$  in the given population. For each distance class  $(r - \Delta, r + \Delta]$ , Moran's  $I$  statistics (Cliff and Ord, 1981; Sokal and Oden, 1978) are applied to the frequency of allele  $A$  within an individual as

$$I_A[r] = \frac{\sum_{i,j=1}^n W_{i,j}[r](a(i) - \bar{a})(a(j) - \bar{a})}{V \sum_{i,j=1}^n W_{i,j}[r]} \quad (2)$$

where  $V$  is the variance of  $a(i)$ :

$$V = \sum_{i=1}^n (a(i) - \bar{a})^2 / n$$

(2) *Coancestry*: Using the same notations as above, some recent studies used the following equation called coancestry (with respect to allele  $A$ ) (Loisselle *et al*, 1995):

$$\rho_A[r] = \frac{\sum_{i,j=1}^n W_{i,j}[r](a(i) - \bar{a})(a(j) - \bar{a})}{\bar{a}(1 - \bar{a}) \sum_{i,j=1}^n W_{i,j}[r]} \quad (3)$$

If a population is Hardy-Weinberg equilibrium

$$V = \bar{a}^2(1 - \bar{a})^2 + 2\bar{a}(1 - \bar{a})(0.5 - \bar{a})^2 + (1 - \bar{a})^2(0 - \bar{a})^2 = \bar{a}(1 - \bar{a})/2$$

then,  $\rho_A(r) = I_A(r)/2$ .

(3) *NAC*: Let  $nac(i, j)$  be the average number of alleles in common over the loci considered between individuals  $i$  and  $j$  (Surles *et al*, 1990). This genetic similarity index can be extended to spatial statistics as (Berg and Hamrick, 1995)

$$NAC[r] = \frac{\sum_{i,j=1}^n W_{i,j}[r] \cdot nac(i, j)}{\sum_{i,j=1}^n W_{i,j}[r]} \quad (4)$$

(4) *Join. count statistics*: Fix one locus and let  $m(i)$  be the genotype of individual  $i$ . Classify individuals by their genotypes of that locus, such as  $AA, AB, BB, AC, BC, \dots$  and define join-count statistics, for example, with

respect to  $AA-AA, AA-AB$  as

$$J_{AA-AA}[r] = \sum_{m(i)=AA, m(j)=AA} W_{i,j}[r] \quad (5a)$$

$$J_{AA-AB}[r] = \sum_{(m(i)=AA, m(j)=AB) \text{ or } (m(j)=AA, m(i)=AB)} W_{i,j}[r] \quad (5b)$$

These are equal to the observed number of joins with specific genotype(s) (in biallelic cases, there are an additional four statistics denoted as  $J_{AA-BB}[r], J_{AB-AB}[r], J_{AB-BB}[r], J_{BB-BB}[r]$ ). Subtracting the expected number of joins under the random distribution and dividing it by the square root of the variance, we obtain SND which approximately follow the normal distribution.

For any of these statistics, plotting values at  $r = \Delta, 3\Delta, 5\Delta, \dots$ , produces a correlogram, illustrating fine-scale spatial genetic structure.

### Point processes

A point process is a stochastic system that places points in the plane. If points are classified into several types, the system is called a multivariate point process. If each point has a mark (generally, a real number or a set of real numbers), the system is called a marked point process. In this paper, a point corresponds to an individual (tree), a type to the genotype, and a mark to its allele frequency within an individual or multilocus genotype. The details and brief introduction to terminology below are taken from Stoyan and Stoyan (1994), and Stoyan and Penttinen (2000), respectively.

Let  $\lambda$  be the density; the mean number of individuals per unit area. The *product density*,  $J(r)$ , is the probability density that there are individuals at two arbitrarily chosen points with interdistance  $r$ . If individuals are randomly distributed, the probability that an individual exists at each point is independently equal to  $\lambda$ , thus,  $J(r) = \lambda^2$ . The normalised product density,  $g(r) = J(r)/\lambda^2$ , is called the *pair correlation function*.  $g(r) > 1$  (or  $J(r) > \lambda^2$ ) for relatively small  $r$  means that the interdistance  $r$  is more frequent than a random point pattern, thus, there is clustering of individuals.

When individuals are classified into  $K$  types,  $K(K+1)/2$  product densities  $\{J_{k,l}(r)\}$  ( $1 \leq k \leq l \leq K$ ) can be considered; they express the probability density that there are type  $k$  and type  $l$  individuals at two arbitrarily chosen points of interdistance  $r$  ( $J_{k,l}(r)$  does not specify which type exists at which point). Let their normalised versions be denoted as  $g_{k,k}(r) = J_{k,k}(r)/\lambda_k^2$  and  $g_{k,l}(r) = J_{k,l}(r)/2\lambda_k\lambda_l$ , where  $\lambda_k$  refers to the density of type  $k$  individuals.

If each individual has a mark, let  $M$  denote the set of marks, and  $m(i)$  the mark of individual  $i$ . Let  $f(m_1, m_2)$  be a function on  $M \times M$ . For two arbitrarily chosen points with interdistance  $r$ , define a random variable that vanishes if there is no individual at one of the points and is equal to  $f(m(i), m(j))$  if individuals  $i$  and  $j$  exist. Let  $J_f(r)$  be the expected value of this random variable and define

$$g_f(r) = J_f(r)/J(r) \quad (6)$$

$g_f(r)$  can be interpreted as the conditional mean of  $f(m(i), m(j))$  given that  $\|X_i - X_j\| = r$ .

Suppose that we have a complete map of individuals with types or marks in a rectangular sampling plot with side lengths  $a$  and  $b$  ( $a < b$ ). Denote the data by  $\{X_i, m(i)\}$   $\{i = 1, 2, \dots, n\}$ , where  $m(i)$  refers to the type or mark.  $g(r), g_{k,k}(r), g_{k,l}(r)$ , and  $g_f(r)$  can be estimated as follows

(Penttinen *et al*, 1992; Stoyan and Stoyan, 1994, pp 284–293).

$$\hat{g}(r) = \sum_{i \neq j} \frac{w(\|X_i - X_j\| - r)}{\hat{\lambda}^2 2\pi r s(r)} \quad (7)$$

$$\hat{g}_{k,k}(r) = \sum_{m(i)=m(j)=k} \frac{w(\|X_i - X_j\| - r)}{\hat{\lambda}_k^2 2\pi r s(r)} \quad (8)$$

$$\hat{g}_{k,l}(r) = \sum_{\substack{m(i)=k \\ m(j)=l}} \frac{w(\|X_i - X_j\| - r)}{2\hat{\lambda}_k \hat{\lambda}_l 2\pi r s(r)} \quad (9)$$

$$\hat{g}_f(r) = \frac{\sum_{i \neq j} w(\|X_i - X_j\| - r) f(m(X_i), m(X_j))}{\sum_{i \neq j} w(\|X_i - X_j\| - r)} \quad (10)$$

Here,  $r < a$ ,  $\hat{\lambda} = n/ab$  is the estimated density,  $\hat{\lambda}_k$  is the estimated density of type  $k$  individuals,

$$w(z) = \begin{cases} 3/4\delta(1 - z^2/\delta^2) & \text{if } |z| < \delta \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

is the Epanechnikov kernel in which  $\delta$  is an arbitrarily fixed constant, and  $s(r) = ab - r(2a + 2b - r)/\pi$  is the edge correction factor. Generally, an estimator includes an edge correction. An exception is  $g_f(r)$  (equation (10)) in which the edge corrections in the numerator and the denominator cancel each other. The majority of ecological studies use Ripley's edge correction (Hasse, 1995), but the edge correction factors do not cancel and  $\hat{g}_f(r)$  must be calculated by a more complicated equation.

### Reformulation of statistics

Shimatani (2002) reformulated the spatial statistics (equations (2)–(5)) in the language of the point process. Using  $a(i)$  as a mark and  $f(m_1, m_2) = (m_1 - \bar{a})(m_2 - \bar{a})/V$  and  $f(m_1, m_2) = (m_1 - \bar{a})(m_2 - \bar{a})/\bar{a}(1 - \bar{a})$  as a function in equation (6), we obtain the reformulated Moran's  $I$  statistics for within-individual frequencies and the coancestry, respectively. Substituting these functions into (10), we obtain their estimators as

$$\hat{I}_A(r) = \frac{\sum_{i \neq j} w(r - \|X_i - X_j\|) (a(i) - \bar{a})(a(j) - \bar{a})}{\hat{V} \sum_{i \neq j} w(r - \|X_i - X_j\|)} \quad (12)$$

$$\hat{\rho}_A(r) = \frac{\sum_{i \neq j} w(r - \|X_i - X_j\|) (a(i) - \bar{a})(a(j) - \bar{a})}{\bar{a}(1 - \bar{a}) \sum_{i \neq j} w(r - \|X_i - X_j\|)} \quad (13)$$

where  $\hat{V} = \sum_{i=1}^n (a(i) - \bar{a})^2 / (n - 1)$  is the unbiased estimator of the variance of  $a(i)$ .

In the same way, let the mark set be (multilocus) genotypes and let the function be  $\text{nac}(i, j)$ . It induces the reformulated NAC, and its estimator is given as

$$N\hat{A}C(r) = \frac{\sum_{i \neq j} w(r - \|X_i - X_j\|) \text{nac}(i, j)}{\sum_{i \neq j} w(r - \|X_i - X_j\|)} \quad (14)$$

equations (12)–(14) express the expected values of  $(a(i) - \bar{a})(a(j) - \bar{a})/V$ ,  $(a(i) - \bar{a})(a(j) - \bar{a})/\bar{a}(1 - \bar{a})$ , and  $\text{nac}(i, j)$ , respectively, given that  $\|X_i - X_j\| = r$ .

Classifying individuals by the single-locus genotype, the estimators of product densities,  $\hat{J}_{k,l}(r)$ , which correspond to joint-count statistics, are given as

$$\hat{J}_{AA-AA}(r) = \sum_{m(i)=AA, m(j)=AA} w(r - \|X_i - X_j\|) / s(r) 2\pi r \quad (15a)$$

$$\hat{J}_{AA-AB}(r) = \sum_{\{m(i)=AA, m(j)=AB\} \text{ or } \{m(i)=AB, m(j)=AA\}} w(r - \|X_i - X_j\|) / s(r) 2\pi r \quad (15b)$$

which express the probability density that there are individuals of genotype  $AA-AA$ , and  $AA-AB$ , ... at two points of interdistance  $r$ , respectively.

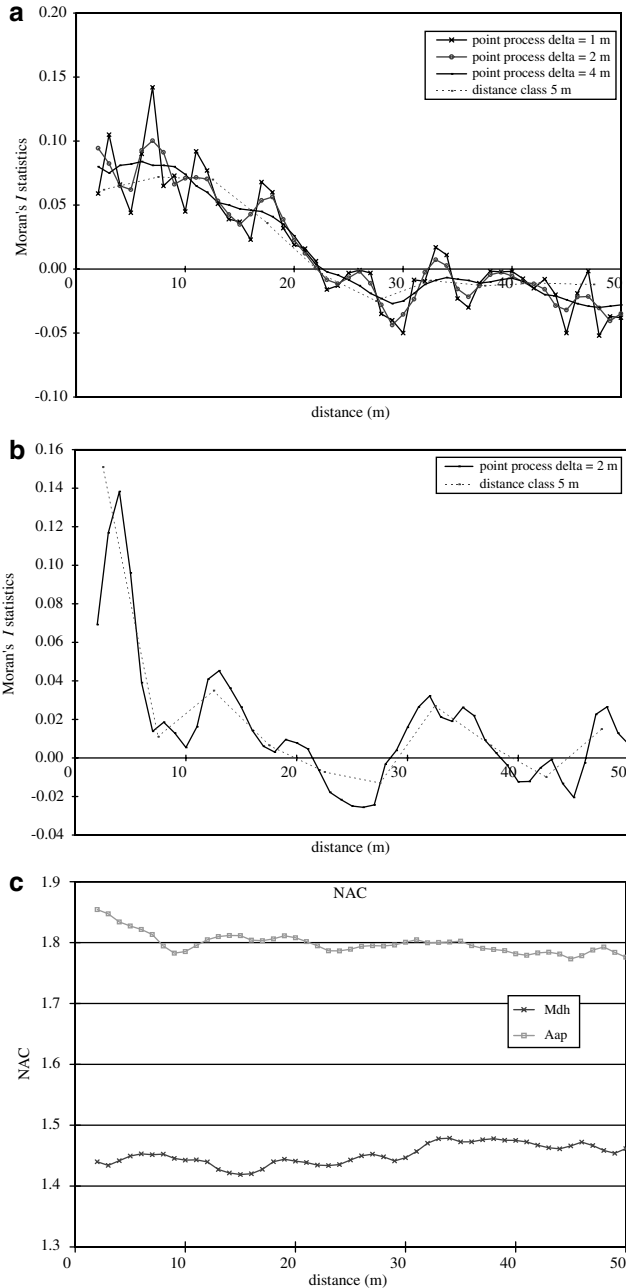
Conventional spatial autocorrelation methods and marked point processes have different mathematical backgrounds. When Sokal and Oden (1978) first introduced the former into population genetics, each population was considered as a lattice point, a set of populations formed an irregular lattice, and Moran's  $I$  statistics were calculated for allele frequencies of the populations. Later, the method was modified to be applicable to individual-based studies, in which an individual is treated as a population. Using this approach, individual locations are fixed and statistical analysis is primarily conducted to test whether the spatial genetic distribution on the given locations significantly differs from the random pattern. On the other hand, marked point processes investigate the spatial distribution of marked points. This approach assumes that genotyped individuals are distributed throughout a plane according to some stochastic system and equations (7)–(10) estimate functions associated with the hidden process from samples [thus, the unbiased estimator should be used in the variance term in equation (12)]. Hence, the point process contains potential for constructing stochastic models that can simultaneously explain individual locations and their genotypes (Shimatani, 2002).

Despite the differences in their mathematical background, in practice, the reformulated expressions (equations (12)–(15)) can be derived simply by replacing  $W_{ij}[r]$  with  $w(r - \|X_i - X_j\|)$  in equations (2)–(4). Hence, both equations exhibit a similar correlogram and a graph (Shimatani, 2002). Figures 1a and b compare  $I_A[r]$  and  $\hat{I}_A(r)$  for a *Fagus crenata* population (data from Takahashi *et al* (2000), Figure 2). The conventional statistics use a weight of either 1 or 0, whether the interdistance exactly falls into  $(r - \Delta, r + \Delta]$ , while in the point processes, weights are gradually decreased when interdistances diverge from  $r$ , and are fixed to 0 if it falls outside  $r \pm \delta$ . The latter approach enables us to calculate the values for any distance and draw a smooth curve illustrating spatial genetic patterns, more elegantly than the broken line of a correlogram. In addition, it is no longer necessary to fix arbitrarily the width  $2\Delta$  of the distance class; occasionally, a slight change of  $\Delta$  dramatically affects the statistics, especially for the first distance class. In contrast, although the point process formulation requires arbitrary fixation of the width  $\delta$  of the kernel function (equation (11)), it works at most on smoothing the curve (Stoyan and Stoyan, 1994, pp 284–290, see also Figure 1a).

### Weighted-sum expressions

Because mark  $a(i)$  takes only three discrete values,  $\hat{I}_A(r)$  and  $\hat{\rho}_A(r)$  can be reformulated under multivariate point processes. Fix a locus and an allele, pack all the alleles other than the fixed  $A$  into one group, and denote it by  $*$ . Index genotypes  $AA, A*, **$  as 1, 2, 3.  $\hat{I}_A(r)$  (and  $\hat{\rho}_A(r)$ ) can be expressed as the weighted sum of  $(\hat{J}_{K-L}(r)/\hat{J}(r))$  ( $1 \leq K \leq L \leq 3$ ) as in (Shimatani, 2002):

$$\hat{I}_A(r) = \sum_{1 \leq K \leq L \leq 3} S_{KL} \hat{J}_{K-L}(r) / \hat{J}(r) \quad (16)$$



**Figure 1** (a) and (b) The dashed lines indicate the correlograms of the conventional Moran's  $I$  statistics for within-individual frequencies (equation (2),  $\Delta = 2.5$  m) and the solid lines show the graphs of the point process functions [equations (12), (16) and (17),  $\delta = 1, 2, 4$  m in (a) and  $\delta = 2$  m in (b)] for *F. crenata* population AK from Takahashi *et al* (2000) for: (a) *Mdh-3* and (b) *Aap-1*. (c) Single locus  $\hat{N}\hat{A}\hat{C}(r)$  calculated by the biallelic approximation (equations (18) and (19),  $\delta = 2$  m) for *Mdh-3* (—×—) and *Aap-1* (—□—).

where  $S_{KL}$  denotes the  $K$ - $L$  component of the matrix

$$S = \frac{1}{\hat{V}} \cdot \begin{pmatrix} (1-\bar{a})^2 & (1-\bar{a})(0.5-\bar{a}) & (1-\bar{a})(0-\bar{a}) \\ & (0.5-\bar{a})^2 & (0.5-\bar{a})(0-\bar{a}) \\ & & (0-\bar{a})^2 \end{pmatrix} \quad (17)$$

(This relation was first pointed out in Epperson (1995) for the conventional Moran's  $I$  for within-individual fre-

quencies (equation (2)) and join-count statistics (equation (5)). Replacing  $\hat{V}$  with  $\bar{a}(1-\bar{a})$  involves the weighted-sum expression of  $\hat{\rho}_A(r)$ . Note that  $\hat{J}_{K-L}(r)/\hat{J}(r)$  does not contain the edge correction factors.

If one biallelic locus is considered,  $\hat{N}\hat{A}\hat{C}(r)$  can be written in the same form:

$$\hat{N}\hat{A}\hat{C}(r) = \sum_{1 \leq K < L \leq 3} T_{KL} \hat{J}_{K-L}(r) / \hat{J}(r) \quad (18)$$

where

$$(T_{KL}) = \begin{pmatrix} 2 & 1 & 0 \\ & 2 & 1 \\ & & 2 \end{pmatrix} \quad (19)$$

Even when the locus has more than two alleles, if the minor alleles have sufficiently small frequencies,  $\hat{N}\hat{A}\hat{C}(r)$  of that locus can be approximated by this biallelic form. Although  $\hat{I}_A(r)$  and  $\hat{N}\hat{A}\hat{C}(r)$  have the same form except for the weighting system, the weight matrix  $S$  of  $\hat{I}_A(r)$  is a variable of allele frequency  $\bar{a}$  in the population, whereas  $\hat{N}\hat{A}\hat{C}(r)$  uses a constant matrix. If the population is at the Hardy-Weinberg equilibrium,  $V = \bar{a}(1-\bar{a})/V_2$ , thus

$$S = 2 \cdot \begin{pmatrix} (1-\bar{a})/\bar{a} & (0.5-\bar{a})/\bar{a} & -1 \\ & (0.5-\bar{a})^2/\bar{a}(1-\bar{a}) & -(0.5-\bar{a})/(1-\bar{a}) \\ & & \bar{a}/(1-\bar{a}) \end{pmatrix}$$

Hence,  $S = S(\bar{a})$  varies depending on the allele frequency  $\bar{a}$ , for example, as:

$$S(0.7) = \begin{pmatrix} 0.86 & -0.57 & -2 \\ & 0.38 & 1.33 \\ & & 4.67 \end{pmatrix},$$

$$S(0.5) = \begin{pmatrix} 2 & 0 & -2 \\ & 0 & 0 \\ & & 2 \end{pmatrix},$$

$$S(0.3) = \begin{pmatrix} 4.67 & 1.33 & -2 \\ & 0.38 & -0.57 \\ & & 0.86 \end{pmatrix}$$

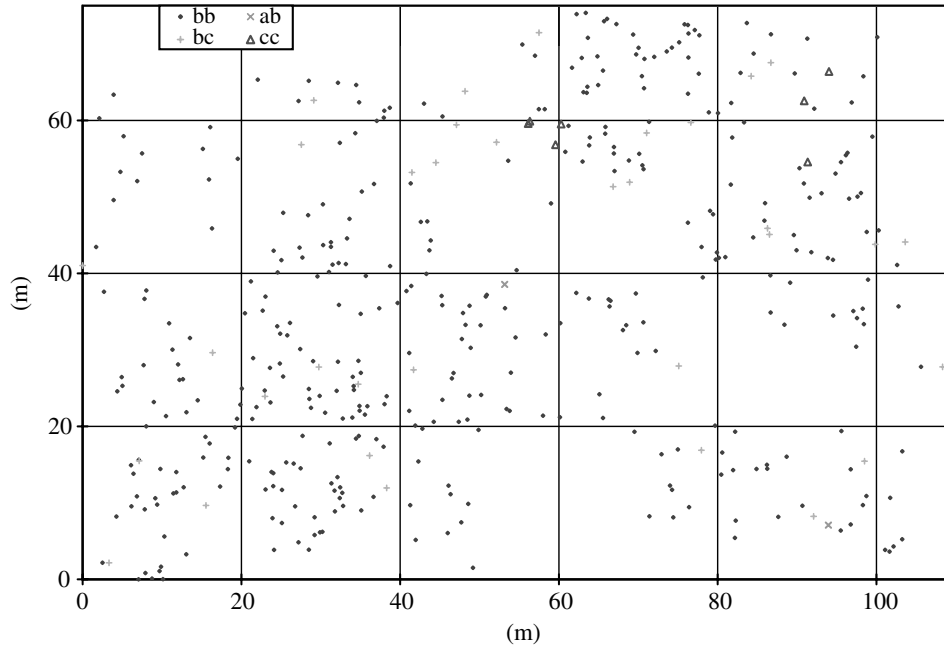
Hence, for the spatial autocorrelation of the three discrete variables, positive values are obtained when: (1) both genotypes (frequencies) are identical or (2) one is a heterozygote ( $A^*$ ) and the other is a homozygote ( $AA$ ) if  $\bar{a} < 0.5$  or  $**$ -type if  $\bar{a} > 0.5$ . Unlike the general cases in which positive correlations appear when two variables tend to have similar values, the autocorrelation of the three discrete variables involves more concrete and specific interpretation.

As the allele frequency deviates from 0.5, more variation appears between the six weights, for instance,

$$S(0.9) = \begin{pmatrix} 0.22 & -0.89 & -2 \\ & 3.56 & 8 \\ & & 18 \end{pmatrix},$$

$$S(0.95) = \begin{pmatrix} 0.11 & -0.95 & -2 \\ & 8.53 & 18 \\ & & 38 \end{pmatrix},$$

$$S(0.99) = \begin{pmatrix} 0.02 & -0.99 & -2 \\ & 48.51 & 98 \\ & & 198 \end{pmatrix}$$



**Figure 2** Distribution of *Aap-1* genotypes in *F. crenata* individuals in population AK from Takahashi *et al* (2000).

$\hat{I}_A(r)$  (and coancestry  $\hat{\rho}_A(r)$ ) provides  $38/0.11 = 345$  times greater weight with **\*\*\_\*\*** joins than **AA-AA** joins if  $\bar{a} = 95\%$ , and  $198/0.02 = 9801$  times greater if  $\bar{a} = 99\%$ , meaning that  $\hat{I}_A(r)$  and  $\hat{\rho}_A(r)$  intensify the information on minor genotype(s). This contrasts with NAC, which always considers **AA-AA** as similar as **AB-AB** and **BB-BB**.

Instead of  $f(m_1, m_2) = (m_1 - \bar{a})(m_2 - \bar{a})/V$ , if product  $f(m_1, m_2) = m_1 m_2$  is used, the resulting function  $g_{m_1 m_2}(r)$  normalised by the square of the mean of marks is called the *mark correlation function* (Stoyan and Stoyan, 1994, pp 291–293; Stoyan and Peenttinen, 2000). This function has frequently been used in ecology where the mark refers to the sizes of trees (eg, Penttinen *et al.* 1992). If the discrete mark  $a(i)$  above is used,  $g_{m_1 m_2}(r)/\bar{a}^2$  can be estimated in the form:

$$\hat{g}_{m_1 m_2}(r)/\bar{a}^2 = 1/\bar{a}^2 \cdot \sum_{1 \leq K \leq L \leq 3} U_{KL} \hat{J}_{K-L}(r)/\hat{J}(r) \quad (20)$$

where

$$(U_{KL}) = \begin{pmatrix} 1 & 0.5 & 0 \\ & 0.25 & 0 \\ & & 0 \end{pmatrix} \quad (21)$$

This function is easier to interpret than  $\hat{I}_A(r)$ , and directly indicates whether allele *A* is clustering. Namely,  $g_{m_1 m_2}(r)/\bar{a}^2 \equiv 1$  if allele *A* is randomly distributed.  $g_{m_1 m_2}(r)/\bar{a}^2 < 1$  for small *r* suggests that the neighbouring individuals tend to be **AA** homozygotes (or **A\*** heterozygotes if  $\bar{a} < 0.5$ ), while  $g_{m_1 m_2}(r)/\bar{a}^2 > 1$  indicates that the neighbouring individuals tend not to share this allele. In contrast, Moran's *I* takes a product of  $(a(i) - \bar{a})(a(j) - \bar{a})$ , thus,  $\hat{I}_A(r) > 0$  suggests either or both, and cannot distinguish between the two cases. However, matrix *U* (equation (21)) includes three zero components, meaning that this measure ignores half of the spatial information.

The effects of the different weighting systems are demonstrated below.

### Field data

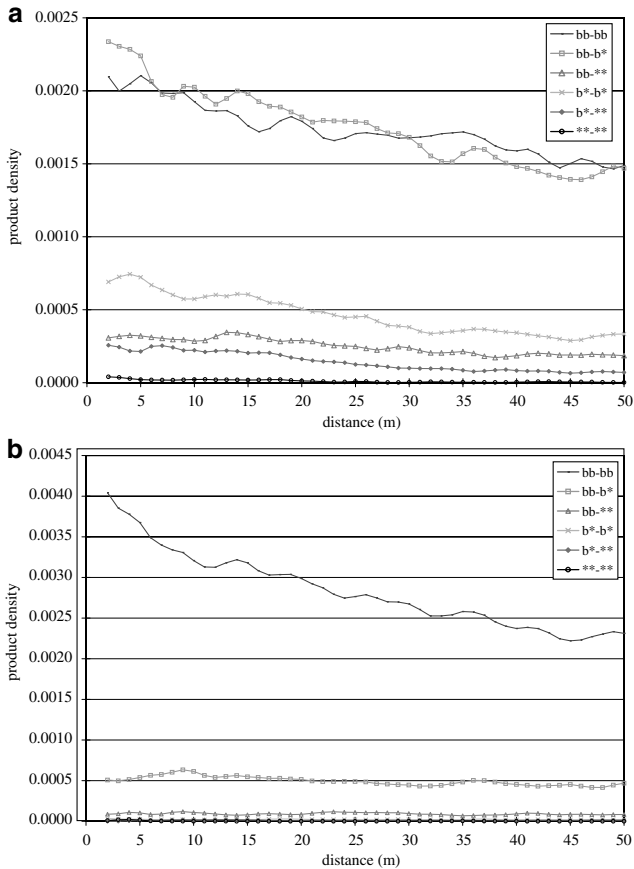
The above methodology is demonstrated by field data of *Fagus crenata* (population AK) from Takahashi *et al* (2000). *Fagus crenata* is widely distributed in cool temperate forest in Japan, especially in areas with abundant snowfall. This species is wind-pollinated, self-incompatible and has limited seed dispersal. The study site was once old-growth forest dominated by *F. crenata*, harvested approximately 80 years ago preserving some seed trees, then, naturally regenerated. The stand is currently covered with secondary *F. crenata* forest. Takahashi *et al* (2000) genotyped 486 individuals in the 0.77 ha plot for nine isozyme loci. This paper primarily uses the genetic data of two loci; *Mdh-3* (EC 1.1.1.37) and *Aap-1* (EC 3.4.11.2) (Figure 2). The details of the data are described in Takahashi *et al* (2000).

### Application to field data

Figure 1 illustrates the reformulated Moran's *I* statistics for within-individual frequencies  $\hat{I}(r)$ , equations (12), (16) and (17) with respect to allele *b* (frequency = 0.80) of *Mdh-3* and allele *b* (frequency = 0.94) of *Aap-1*, and (single-locus)  $\hat{N}\hat{A}\hat{C}(r)$  of each locus for the *Fagus crenata* population. The biallelic approximation (equations (18) and (19)) was used for  $\hat{N}\hat{A}\hat{C}(r)$  because the third alleles have frequencies of only 0.2–0.3%.

For the *Mdh-3* locus, the monotonically decreasing  $\hat{I}_b(r)$  suggests spatial genetic structure whereas  $\hat{N}\hat{A}\hat{C}(r)$  shows no clear tendency. For *Aap-1*, both  $\hat{I}_b(r)$  and  $\hat{N}\hat{A}\hat{C}(r)$  indicate a trend of decreasing up to 10 m but  $\hat{I}_b(r)$  takes its maximum at 4 m. The application of equations (16)–(21) leads us to examine as to what caused the differences between the two functions and between the two loci, and reveals details of the spatial genetic patterns.

Figure 3 illustrates the six product density functions  $\{\hat{J}_{K-L}(r)\}$  ( $1 \leq K \leq L \leq 3$ ) (equation (15)). Dividing  $\hat{J}_{K-L}(r)$



**Figure 3** The six-product density functions  $\{\hat{J}_{K-L}(r)\} (1 \leq K \leq L \leq 3, \delta = 2\text{ m})$  of the *F. crenata* population for: (a) *Mdh-3* and (b) *Aap-1*.

by their sum ( $=\hat{J}(r)$ ), the conditional probabilities that a randomly selected pair have genotypes  $K$  and  $L$ , given that their interdistance is  $r$ , are obtained (Figure 4). Multiplying  $\hat{J}_{K-L}(r)/\hat{J}(r)$  by weight

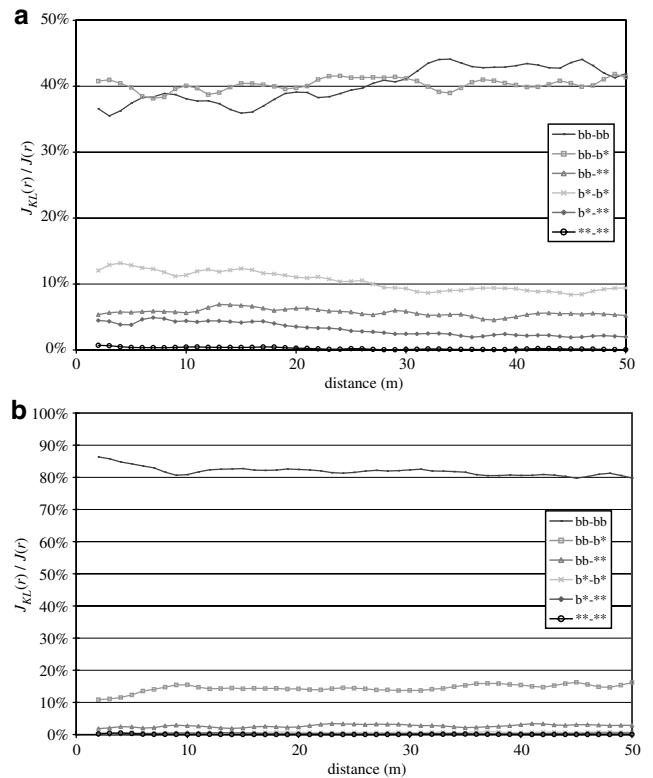
$$S_{Mdh-3} = \begin{pmatrix} 0.48 & -0.73 & -1.94 \\ & 1.10 & 2.93 \\ & & 7.81 \end{pmatrix} \text{ or}$$

$$S_{Aap-1} = \begin{pmatrix} 0.1 & -0.76 & -1.62 \\ & 5.79 & 12.33 \\ & & 26.29 \end{pmatrix}$$

involves the six curves and their sum is equal to  $\hat{I}_b(r)$  (Figure 5). Changing the weight matrix to  $T$  (equation (19)), another six curves are obtained whose sum is  $\hat{N}\hat{A}\hat{C}(r)$  (Figure 6).

### Mdh-3

There appear to be a great number of  $bb-bb$  and  $bb-b^*$  joins for short distances but not as many for long distances (Figure 3a). This is largely because of the clustering of trees themselves rather than the clustering of genotypes on the trees. In fact, compared with the long interdistances, the ratios  $\hat{J}_{bb-bb}(r)/\hat{J}(r)$  and  $\hat{J}_{bb-b^*}(r)/\hat{J}(r)$  are smaller for short distances (Figure 4a); instead, ratios for  $b^*-b^*$ ,  $b^*-**$ , and  $**-**$  are greater for short distances than for long distances. This means that arbitrarily chosen neighbouring trees are expected to be  $bb-bb$  or  $bb-b^*$  because allele  $b$  is in the majority, that the



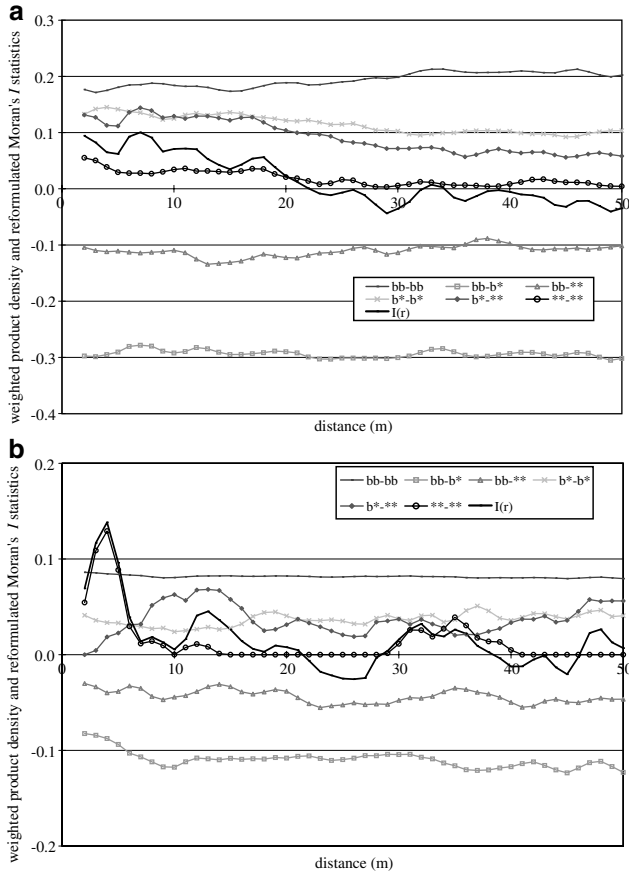
**Figure 4**  $\hat{J}_{K-L}(r)/\hat{J}(r)$  ( $1 \leq K \leq L \leq 3$ ) for (a) *Mdh-3* and (b) *Aap-1* in a *F. crenata* population.

probabilities of selecting  $b^*-b^*$ ,  $b^*-**$ , or  $**-**$  are small because these genotypes are in the minority, and that minor genotypes are chosen more frequently for neighbouring pairs than separated pairs.

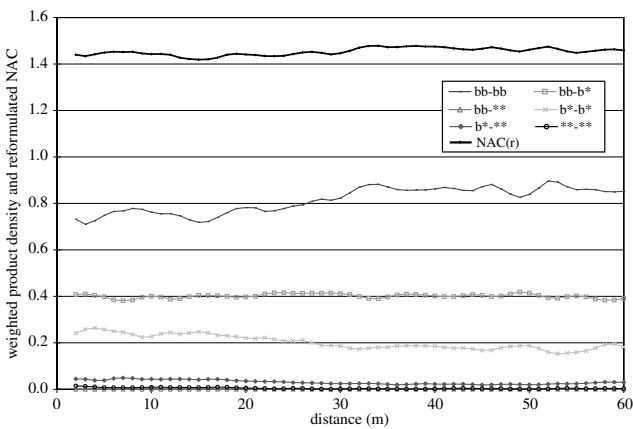
When the six percentage functions  $\hat{J}_{K-L}(r)/\hat{J}(r)$  are multiplied by matrix  $S_{Mdh-3}$ , components  $b^*-b^*$ ,  $b^*-**$ , and  $**-**$  are amplified;  $1.10/0.48-7.81/0.48 \approx 2-16$  times more than  $bb-bb$ , and their tendencies of monotonic decrease become apparent, resulting in the monotonically decreasing  $\hat{I}_b(r)$  (Figure 5a). On the other hand,  $\hat{N}\hat{A}\hat{C}(r)$  uses the matrix consisting of only  $\{0, 1, 2\}$  (equation (19)). All the curves stay almost constant, and  $\hat{N}\hat{A}\hat{C}(r)$  reveals no clear spatial structure for *Mdh-3* (Figure 6). The mark correlation function (Figure 7) also suggests that minor allele  $c$  is clustering while major allele  $b$  is not.  $\hat{I}_b(r)$  intensified the former characteristics, visualizing the spatial structure in the *Mdh-3* locus, while  $\hat{N}\hat{A}\hat{C}(r)$  with no amplification does not express this pattern.

### Aap-1

The allele frequency is highly biased to the major allele  $b$ . This involves large  $\hat{J}_{bb-bb}(r)$  and  $\hat{J}_{bb-b^*}(r)$ , while the other four nearly overlap the horizontal axis (Figure 3b). Taking their percentages, Figure 4b shows that up to 10 m,  $\hat{J}_{bb-bb}(r)/\hat{J}(r)$  monotonically decreases while  $\hat{J}_{bb-b^*}(r)/\hat{J}(r)$  monotonically increases.  $\hat{N}\hat{A}\hat{C}(r)$  reflects only the two major curves, resulting in the clear spatial structure up to 10 m (Figure 1c). On the contrary,  $\hat{I}_b(r)$  (Figure 1b) indicates the strongest spatial structure at 4 m. This contrasting result was caused by the spatial pattern of  $**-$  trees; although there is a sharp peak at 4 m, it has not become visible until the weight matrix  $S_{Aap-1}$



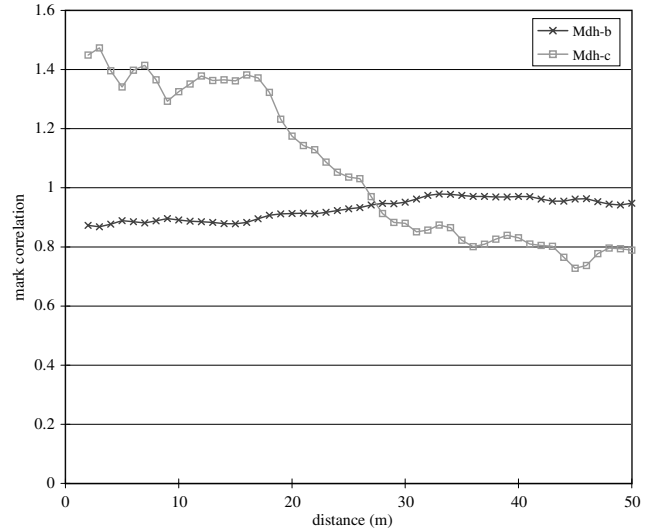
**Figure 5** The decomposition of  $\hat{I}_b(r)$  into the six components (equations (16) and (17)) for: (a) *Mdh-3* and (b) *Aap-1* in a *F. crenata* population.



**Figure 6** The decomposition of  $\hat{NAC}(r)$  calculated by the biallelic approximation into six components (equations (18) and (19)) for *Mdh-3* in a *F. crenata* population.

intensified **\*\*\_\*\*** joins  $26.29/0.1 = 263$  times more than *bb-bb* joins (Figure 5b).

In summary,  $\hat{NAC}(r)$  and  $\hat{I}_b(r)$  involve similar decreasing graphs, especially if calculated by 5 m distance classes because it averaged 0-5 m and the peak at 4 m has disappeared (Figure 1b). However, the former is mostly because of the spatial pattern of the major allele whereas in the latter, the minor alleles make a greater contribution.



**Figure 7** The mark correlation functions with respect to two alleles at *Mdh-3* (equations (20) and (21)).

## Discussion

Individual-based spatial genetic studies have applied Moran's *I* statistics for interval data to three discrete variables, which can be written as a weighted sum of the six joint-count statistics (Epperson, 1995). There have been two evaluations of the utility of this statistics and two directions in developing spatial statistics. Epperson (1995) suggested classifying individuals by their genotypes and applying full joint-count statistics to ensure the high resolution of spatial data. In contrast, Smouse and Peakall (1999) aimed to develop a statistic that summarizes spatial patterns and has as much statistical power as possible for testing the randomness of genetic patterns. Because full joint-count statistics reflect the original genotypic information, they can reveal various characteristics of spatial pattern, whereas information loss is inevitable for summarized statistics. On the contrary, even the summarized autocorrelations tend to be sensitive to stochastic variance (Slatkin and Arter, 1991), and joint-count statistics, which must be calculated from smaller sample sizes than their weighted sum, should be more vulnerable to stochastic effects. The two approaches have contrasting advantages and disadvantages. However, they can complement each other if summarised statistics can be decomposed into components. In fact, the statistics introduced in Smouse and Peakall (1999) are defined for multilocus genotypes as well as decomposable into loci and alleles, therefore, one can check which allele/locus largely contribute to the summarised statistics and which have little influence. This paper has extended this approach and shown that by changing the values of the weighting matrix, both Moran's *I* for within-individual frequencies and NAC are decomposable into joint-count statistics. Hence, under the spatial analysis proposed here, joint-count statistics (simplified by packing the other alleles into type \*) are used to examine the roles of each genotype in the summarized statistics, and thus play relatively complementary roles, which differs from the method suggested in Epperson (1995) in which many joint-count statistics should play a central role. Instead, we may accumula-

tively analyse spatial genetic patterns from the genotypic level to the allele level, and possibly to the single-locus and multilocus level, and examine their relations in the hierarchy.

Some previous studies have separately analysed individual distribution and genes by Ripley's  $K$ -function of point processes (Berg and Hamrick, 1995) or Morishita's index of dispersion ( $I_s$ ) (Ueno *et al.*, 2000), and by spatial autocorrelation based on the lattice theory, respectively. The introduction of point processes involves fundamental conceptual changes. First, all of the spatial analyses are established on a common theoretical base, from the spatial distribution of individual trees to the spatial distribution of genetic variation, at different levels indicated above. More importantly, the lattice theory fixes individual locations whereas the point process treats them as random variables, and the goal is no longer to test the spatial randomness but includes the construction of stochastic models that can simultaneously explain individual distributions and their genotypes (Shimatani, 2002). It is currently quite common for ecologists to use genetic markers as tools for ecological studies, called molecular ecology. The marked point process provides appropriate analytical tools when we examine populations for ecological purposes such as forest dynamics, by means of genetic markers, which may involve new insights into population genetics.

Takahashi *et al.* (2000) suggested that the presence of the spatial genetic structure represented by Moran's  $I$  statistics is a result of regeneration from limited seed trees because then offspring surrounding the mother tend to share alleles inherited from the mother. Moreover, applying point process models in which the genetic structure was represented by the average of the reformulated Moran's  $I$  statistics, Shimatani (2002) quantitatively estimated the number of seed trees as moderately limited (eg, 35 trees/ha) rather than very limited (eg, 10 trees/ha), suggesting advance reproduction of harvested adults. Takahashi *et al.* (2000) also illustrated a map of genotypic distribution for the *Pgi-1* locus in which a preserved tree with a minor allele  $d$  is surrounded by young trees with this allele (this observation can be quantitatively assessed by the application of the decomposition analysis introduced here to the *Pgi-1* locus, which actually indicated the clustering of heterozygotes  $c^*$ , mostly  $cd$ ). This paper also shows the clustering of minor alleles for *Mdh-3* and *Aap-1* (in fact, seven  $cc$ -trees are clumped into two patches, see Figure 2).

Spatial patterns of minor alleles reflect the founder effect: regeneration from a limited number of seed trees. For the above three loci,  $\hat{I}_A(r)$ 's property of intensifying minor allele' information worked appropriately. On the contrary, the *Dia-1* locus has four alleles, and two minor alleles,  $a$  and  $c$ , are separately distributed in the plot. In such cases, because  $\hat{I}_A(r)$  specifies one allele and packs all the others together,  $\hat{I}_A(r)$  for any allele cannot effectively illustrate the spatial pattern at the locus while  $NAC(r)$  may. The advantages and disadvantages of each function should be extensively examined to characterize spatial patterns more adequately and to construct stochastic models.

In conclusion, spatial analysis for mapped, genotyped individuals should not rely on one statistic and the simultaneous use of several point process functions is recommended. If the locus is close to biallelic, the above demonstration suggests fixing an allele with sufficient

frequency, packing all the other alleles into one group, calculating the six product density functions, providing weights depending upon the function, drawing their curves, and then examining the spatial genetic pattern. The comprehensive application of point process functions provides analytical tools for spatial data of genotyped individuals with population genetics and molecular ecology.

## Acknowledgements

We thank two anonymous referees for critically commenting on the manuscript.

## References

- Berg EE, Hamrick JL (1995). Fine-scale genetic structure of a Turkey oak forest. *Evolution* **49**: 110–120.
- Chung MG, Epperson BK (2000). Clonal and spatial genetic structure in *Eurya emarginata* (Theaceae). *Heredity* **84**: 170–177.
- Cliff AD, Ord JK (1981). *Spatial Processes: Models and Applications*. Pion Ltd: London.
- Epperson B (1995). Fine-scale spatial structure: correlations for individual genotypes differ from those for local gene frequencies. *Evolution* **49**: 1022–1026.
- Geburek T, Tripp-Knowles P (1994). Genetic architecture in bur oak, *Quercus macrocarpa* (Fagaceae), inferred by means of spatial autocorrelation analysis. *Plant Syst Evol* **189**: 63–74.
- Hasse P (1995). Spatial pattern analysis in ecology based on Ripley's  $K$ -function: Introduction and methods of edge correction. *J Veg Sci* **6**: 575–582.
- Heywood JS (1991). Spatial analysis of genetic variation in plant populations. *Annu Rev Ecol Syst* **22**: 335–355.
- Leonardi S, Menozzi P (1996). Spatial structure of genetic variability in natural stands of *Fagus sylvatica* L. (beech) in Italy. *Heredity* **77**: 359–368.
- Loisselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* **82**: 1420–1425.
- Pettinen A, Stoyan D, Henttonen HM (1992). Marked point processes in forest statistics. *For Sci* **38**: 806–824.
- Shimatani K (2002). Point processes for fine-scale spatial genetics and molecular ecology. *Biom J* **44**: 325–352.
- Slatkin M, Arter HE (1991). Spatial autocorrelation methods in population genetics. *Am Nat* **138**: 499–517.
- Smouse PE, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561–573.
- Sokal RR, Oden NL (1978). Spatial autocorrelation in biology. 1. Methodology. *Biol J Linn Soc* **10**: 199–228.
- Stoyan D, Stoyan H (1994). *Fractals, Random Shapes and Point Fields*. John Wiley and Sons: Chichester.
- Stoyan D, Penttinen A (2000). Recent applications of point process methods in forestry statistics. *Statist Sci* **15**: 61–78.
- Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glössl J, Kremer A. (1998). Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Mol Ecol* **7**: 317–328.
- Surles SE, Arnold J, Schnabel A, Hamrick JL, Bongarten BC (1990). Genetic relatedness in open-pollinated families of two leguminous tree species, *Robinia pseudoacacia* L. and *Gleditsia triacanthos* L. *Theor Appl Genet* **80**: 49–56.
- Takahashi M, Mukouda M, Koono K (2000). Differences in genetic structure between two Japanese beech (*Fagus crenata* Blume) stands. *Heredity* **84**: 103–115.
- Ueno S, Tomaru N, Yoshimaru H, Manabe T, Yamamoto S. (2000). Genetic structure of *Camellia japonica* L. in an old-growth evergreen forest, Tsushima, Japan. *Mol Ecol* **9**: 647–656.
- Xie CY, Knowles P (1991). Spatial genetic structure within natural populations of jack pine (*Pinus banksiana*). *Can J Bot* **69**: 547–551.