Genetics: Published Articles Ahead of Print, published on January 16, 2005 as 10.1534/genetics.104.036038

Variogram analysis of the spatial genetic structure of continuous populations using multilocus microsatellite data

AUTHORS

Helene H. Wagner, Rolf Holderegger, Silke Werth, Felix Gugerli, Susan E. Hoebee and Christoph Scheidegger

AFFILIATION

WSL Swiss Federal Research Institute, 8903 Birmensdorf, Switzerland

RUNNING HEAD

Variogram analysis of genetic structure

KEYWORDS

Clonal structure, gene diversity, geostatistics, *Lobaria pulmonaria*, molecular variance.

CORRESPONDING AUTHOR

Helene Wagner

WSL Swiss Federal Research Institute

Zürcherstrasse 111

8903 Birmensdorf

Switzerland

Phone: +41-1-739-2587

Fax: +41-1-739-2215

Email address: helene.wagner@wsl.ch

ABSTRACT

A geostatistical perspective on spatial genetic structure may explain methodological issues of quantifying spatial genetic structure and suggest new approaches to address them. We use a variogram approach to (*i*) derive a spatial partitioning of molecular variance, gene diversity, and genotypic diversity for microsatellite data under the infinite allele model (IAM) and the stepwise mutation model (SMM), (*ii*) develop a weighting of sampling units to reflect ploidy levels or multiple sampling of genets, and (*iii*) show how variograms summarize the spatial genetic structure within a population under isolation-by-distance. The methods are illustrated with data from a population of the epiphytic lichen *Lobaria pulmonaria*, using six microsatellite markers. Variogram-based analysis not only avoids bias due to the underestimation of population variance in the presence of spatial autocorrelation, but it also provides estimates of population genetic diversity and the degree and extent of spatial genetic structure accounting for autocorrelation.

INTRODUCTION

Methods for the analysis of spatial genetic structure have mostly been developed for single-locus, diploid genotypic data such as provided by isozymes (SMOUSE and PEAKALL 1999). In contrast to this latter marker type, microsatellite data also contain information on repeat numbers of individual gene copies. Microsatellite markers are often highly variable, and differences in allele size are interpreted in the light of alternative evolutionary models. Under the infinite allele model (IAM), any mutation is assumed to lead to a new allele, whereas under the stepwise mutation model (SMM), mutation is likely to increase or decrease the number of repeats at a microsatellite locus by one (BALLOUX and GOUDET 2002). Neither of these two extreme mutation models seems to fit perfectly to microsatellite loci, so that measures based on IAM and SMM are often reported together (BALLOUX and LUGON-MOULIN 2002) The difference between statistical measures (see below) under the two models is assumed to indicate the relative importance of mutation and drift (HARDY 2003).

Population genetic analyses are based on gene diversity under IAM (e.g., F_{ST}) and on molecular variance under SMM (e.g., R_{ST}). F_{ST} and R_{ST} quantify the differentiation of isolated populations assuming random mating within and restricted gene flow among populations. Both F_{ST} and R_{ST} can be adapted to pairwise comparisons, and Mantel tests are used to test the correlation with geographic distance between pairs of populations (HARDY and VEKEMANS 2002). However, limited gene movement can cause isolation-by-distance effects even within continuous populations. The resulting spatial genetic structure within a population can be summarized by kinship for IAM (LOISELLE *et al.* 1995) or relationship coefficients for SMM (STREIFF *et al.* 1998). Kinship and relationship coefficients assess the similarity of homologous alleles between individuals and may be expressed as a function of geographic distance. Statistical

tests for isolation by distance within continuous populations often involve either a Mantel test of Moran's *I* (or related correlation coefficients, e.g. SMOUSE and PEAKALL 1999) or join-count statistics (EPPERSON 2003).

When assessing genetic diversity, it may be necessary to exclude comparisons of gene copies within individuals if they cannot be assumed to be independent. For organisms with variable ploidy levels within populations such as *Taraxacum* sp. (MEIRMANS *et al.* 2003; VAN DER HULST *et al.* 2003), individuals with a high ploidy level will receive more weight in the estimation of the population genetic diversity than do, e.g., diploid individuals unless ploidy level is accounted for. A similar problem arises for clonal organisms, where the multiple sampling of ramets from the same genetic individuum (genet), can bias any measure of genetic structure of a population (BALLOUX *et al.* 2003; HÄMMERLI and REUSCH 2003; PARKS and WERTH 1993). This is commonly taken into account by retaining a single sample per genet, either assuming the center of a clonal patch to be its origin or randomly selecting one sample per genet (CHUNG and EPPERSON 2000; HÄMMERLI and REUSCH 2003; REUSCH *et al.* 1999). Both approaches may, however, lead to a considerable loss of information and increased error in the description of the spatial genetic structure within populations.

VEKEMANS and HARDY (2004) identified some important problems and common misuses of spatial analysis in population genetics. (*i*) The spatial genetic structure is often described in terms of a maximum distance to which such structure extends. The common practice of assessing the extent of spatial genetic structure by the distance at which a Moran's *I* correlogram reaches zero (e.g., EPPERSON 2003) is misleading, as this estimate depends strongly on the sampling design (VEKEMANS and HARDY 2004). (*ii*) The presence of non-random spatial genetic structure can be tested using Mantel permutation tests for a series of distance classes, and

a Bonferroni correction is applied to account for multiple tests. VEKEMANS and HARDY (2004) caution that while the uncorrected test is too liberal, the correction makes it too conservative, and argue that this approach should not be used to determine the scales of spatial genetic structure, as the null hypothesis is only the overall absence of spatial genetic structure. (*iii*) The amount of spatial genetic structure should not be assessed from the value (e.g. of Moran's *I*) for the first distance class, as this absolute value depends strongly on the sampling design (FENSTER *et al.* 2003; VEKEMANS and HARDY 2004). (*iv*) Estimating biological parameters, such as dispersal distances, is only valid if the observed spatial genetic structure represents a true isolation-by-distance pattern at dispersal-drift equilibrium (VEKEMANS and HARDY 2004), thus assuming that the patterning results only from limited dispersal, that it has reached a stationary phase, and that the scale of the study is appropriate (VEKEMANS and HARDY 2004).

Moran's *I*, Mantel tests and join-count statistics were borrowed from the general field of spatial statistics, originally developed, e.g., in geography, and adapted to population genetic data and questions as necessary. Other measures of spatial genetic structure, such as kinship or relationship coefficients, were developed specifically for genetic data and are little integrated with spatial statistical theory. However, many of the above problems are of general nature and not specific to population genetics. Specifically, variogram modeling as developed in geostatistics may provide explanations and alternatives for the problems raised by (VEKEMANS and HARDY 2004). The term variogram refers to a plot of the semivariance (see below) against distance. The well-known Geary's *c* correlogram is actually a standardized variogram (LEGENDRE and LEGENDRE 1998). Several population genetic measures and methods rely on the semivariance, namely the genetic distance measure by GOLDSTEIN *et al.* (1999) and the R_{ST} statistic (SLATKIN 1995). Nonetheless, variogram modeling is rare in population genetics.

PIAZZA and MENOZZI (1983) proposed a variogram of differences in allele frequencies between populations, and MONESTIEZ and GOULARD (1997) provided an application of multivariate geostatistical analysis to genetic data, but neither approach found much resonance in the population genetic literature.

WAGNER (2003, 2004) developed a formal integration of multivariate analysis and geostatistics in the context of plant community ecology. The crucial point of such an integration of spatial and non-spatial analysis is that the semivariance partitions the estimate of the population variance by distance class (WAGNER 2003). Hence, the semivariance can be used to partition the results of non-spatial analyses, such as population estimates of genetic diversity, by distance (multiscale ordination), and variograms can be interpreted in an ecologically more meaningful way.

This paper extends the spatial partitioning of variance to population genetic data and problems. The first section introduces key geostatistical concepts and methods and discusses the sensitivity of commonly used measures of autocorrelation and population variance. The methods section pursues three specific objectives: (*i*) to derive a spatial partitioning of measures of genetic diversity compatible with IAM and SMM, (*ii*) to develop a method for weighting sampling units to reflect different ploidy levels or multiple sampling of ramets within genets without data reduction, and (*iii*) to show how variogram modeling can be used for estimating population genetic parameters and summarizing the spatial genetic structure within populations. The methods are illustrated with a worked example (Appendix) and with an application to empirical microsatellite data from a population of the haploid, tree-colonizing (epiphytic) lichen *Lobaria pulmonaria*. We conclude with considerations for the robust estimation of the spatial genetic structure of continuous populations.

A GEOSTATISTICAL PERSPECTIVE

Geostatistical concepts and methods

Spatial autocorrelation and stationarity. Spatial autocorrelation refers to the common phenomenon that nearby observations tend to be more similar than distant ones. Positive spatial autocorrelation is assumed to result from any kind of spatial process, such as pollen flow or seed dispersal in plants. The observed spatial autocorrelation can be quantified for various purposes (FORTIN *et al.* 2001), such as: (*i*) testing for the presence of autocorrelation, e.g., in order to meet assumptions for estimating population characteristics, (*ii*) assessing the range of autocorrelation, i.e., the distance beyond which observations are spatially independent, (*iii*) fitting a theoretical model in order to summarize the observed spatial structure, (*iv*) inference about the underlying spatial process, such as dispersal distances and differences among populations. However, geostatistical analysis requires some assumption of stationarity, i.e., the structure of spatial autocorrelation must be the same throughout the study area. Specifically, it is common to assume weak stationarity, where the mean and the variance are constant and the autocorrelation only depends on the geographic distance between sampling units (BURROUGH 1995).

Correlograms and the empirical variogram. Geostatistics considers four statistical moments of a random variable: (*i*) its mean, (*ii*) variance, (*iii*) covariance, and (*iv*) semivariance (BURROUGH 1995). Spatial autocorrelation can be quantified based on covariance (Moran's *I*) or semivariance (empirical variogram and Geary's *c* correlogram). Correlograms are standardized through division by the sample variance (Moran's *I*) or population variance (Geary's *c*; CLIFF and ORD 1981):

$$\text{Moran's } I\text{: } I(r) = \frac{\text{covariance}}{\text{variance}} = \frac{\frac{1}{n_r}\sum_{a\neq b} x_{ab}^{(r)}(y_a - \bar{y})(y_b - \bar{y})}{\frac{1}{N}\sum_{a}(y_a - \bar{y})^2} \quad (1)$$

$$\text{Geary's } c\text{: } c(r) = \frac{\text{semivariance}}{\text{variance}} = \frac{\frac{1}{2n_r}\sum_{a\neq b} x_{ab}^{(r)}(y_a - y_b)}{\frac{1}{N-1}\sum_{a}(y_a - \bar{y})^2} \quad (2)$$

$$\text{Empirical variogram: } \gamma(r) = \text{semivariance} = \frac{1}{2n_r}\sum_{a\neq b} x_{ab}^{(r)}(y_a - y_b) \quad (3)$$

Given N samples, these coefficients are calculated based on pairs of samples a and b falling into a series of distance classes r. The Kronecker weight $x_{ab}^{(r)}$ for the pair of observations a and b takes the value $x_{ab}^{(r)} = 1$ if a pair of samples belongs to distance class r and $x_{ab}^{(r)} = 0$ otherwise, and n_r is the sum of the weights $x_{ab}^{(r)}$ for the given distance class, i.e., the number of unique pairs of gene copies a and b from two samples separated by a distance falling into distance class r. However, n_r decreases for large distance classes r, and bias may arise from the fact that only the observations from the edge of the sampled population can contribute to the estimates for larger distances. It is therefore customary to limit the description of the spatial structure to half the maximum distance between sampling units (CRESSIE 1993).

Figure 1 shows the empirical variogram (A) as well as Moran's I and Geary's c correlograms (B) of an artificial, spatially autocorrelated random variable. Geary's c correlogram is a rescaled version of the empirical variogram, and Moran's I correlogram resembles, but is not identical to, $1 - c(r)$ (LEGENDRE and LEGENDRE 1998). In the absence of spatial autocorrelation, the expected value of Geary's c is $E[c(r)] = 1$, whereas the expected values of Morans's I is

$E[I(r)] = -1 / (N - 1)$, which approaches zero for large sample sizes N (EPPERSON 2003; SOKAL and WARTENBERG 1983).

Variogram modeling. The autocorrelation structure can be modeled by fitting a theoretical variogram model to the empirical variogram. The elementary theoretical variograms suitable for modeling patterns due to a single, stationary spatial process are defined by the following parameters: (*i*) model family, such as exponential, spherical, or Gaussian, (*ii*) *nugget* variance, i.e., the variance among adjacent samples, (*iii*) *range*, or the distance beyond which observations are spatially independent, and (*iv*) *sill*, the constant variance among spatially uncorrelated samples (Figure 1A; ISAAKS and SRIVASTAVA 1989).

Sensitivity of measures of autocorrelation and population variance

Sensitivity to non-stationarity. The assumption of weak stationarity can be violated in several ways, including (*i*) non-stationarity of the mean in the population, e.g., in the presence of clinal structure, (*ii*) non-stationarity of the variance, e.g., if the variability of a microsatellite locus increases with increasing number of repeats, or (*iii*) anisotropy, where the autocorrelation structure depends on direction, e.g., if mean seed dispersal distances are larger than average in the predominant wind direction. Strictly speaking, the stationarity assumption concerns the underlying process and not the observed pattern, so that it cannot be tested directly (FORTIN *et al.* 2003). However, the empirical variogram can be used to check for problems with non-stationarity. A finite, constant variance will always result in the presence of a sill, whereas a continued increase of the semivariance with distance may indicate a spatial trend in the mean, possibly coupled with dependence of the variance on the mean. Separate empirical variograms can be calculated for different directions and compared in order to check for anisotropy. In

theory, the same visual inspections could be performed with correlograms, but only the variogram offers the possibility of modeling different components of variance and, thus, accounting for them. For instance, an exponential variogram function could be used to model the autocorrelation due to a stationary spatial process, and a linear variogram function could be used to model the increase in variance with distance due to a cline.

Sensitivity to the sampling design. The variogram can be interpreted as a distance-dependent estimate of the population variance (WAGNER 2003). The commonly used "unbiased" estimator of population variance, $\hat{V} = \sum (y - \bar{y})^2 / (N - 1)$, assumes independent, spatially uncorrelated observations, which would correspond to a strictly horizontal empirical variogram. In essence, this requires the assumption of a panmictic population with random dispersal, which is likely to be violated in most natural systems.

Spatial autocorrelation reduces the variance between closely-spaced pairs of observations. Here, we illustrate the consequences of estimating the population variance with a simple simulation. An artificial, autocorrelated variable was sampled in different ways and the estimates of the population variance, averaged over many replicate simulations, were compared to the true value. We compared three sampling strategies: (*i*) systematic sampling, with a spacing known to be larger than the range of spatial autocorrelation in order to obtain spatially uncorrelated, independent sampling units; (*ii*) random sampling, and (*iii*) stratified or clustered sampling, selecting groups of nearby locations so as to obtain an appropriate representation of short distance classes for spatial analysis. The criteria for comparison were *accuracy*, i.e, the absence of bias so that the mean of all replicate estimates is close to the true population value, and *precision*, i.e., low variability of replicate estimates (PALMER 1990). For details of the simulation experiment, see caption of Table 1.

The systematic and the random samples provided unbiased estimates, independent of sample size (Table 1). Precision increased with sample size, i.e., the standard deviation of the estimates was reduced. For small sample sizes, where the chances of randomly selecting autocorrelated samples were small, the systematic and the random samples reached a similar precision. With increasing sample size, however, the random samples provided a lower precision than the systematic samples. This effect was due to the increasing number of comparisons between autocorrelated samples, not their proportion. Parametric statistical tests assume spatially uncorrelated samples, which in this simulation corresponds to the systematic sampling design. For a spatially autocorrelated variable, the increased variability of estimates from a random sample may render such tests too liberal. This means that the actual probability of rejecting the null hypothesis when it is true may be larger than the stated significance level alpha.

On average, the clustered samples strongly underestimated the population variance (Table 1). This negative bias was reduced with increasing sample size, as more and more clusters of samples were selected, thus reducing the proportion of comparisons between autocorrelated pairs of samples. In fact, the variance of the estimates based on clustered samples was comparable to the variance for systematic samples with a five times smaller sample size, which can be explained by the sampling of clusters of five strongly autocorrelated locations. However, the clustered samples provided biased estimates, whereas the corresponding systematic samples were unbiased. Hence, the "unbiased" variance estimator may be negatively biased due to spatial autocorrelation. The magnitude of this systematic bias will depend on the spatial autocorrelation structure and the proportion of autocorrelated samples, which is a function of the spatial configuration of the sample rather than sample size. One may argue that the spatial autocorrelation structure is an inherent characteristic of a population. However, because the

estimate of the population variance depends on the sampling design, it should be based on independent samples.

Correlograms imply division by the sample or population variance (see equations 1 and 2). Because of this, it follows that (*i*) the actual values of Moran's $I(r)$ and Geary's $c(r)$ depend on the spatial configuration of the sample, and (*ii*), for a stationary process, $I(r)$ reaches a value slightly below zero, and $c(r)$ a value above one, for distances beyond the range of spatial autocorrelation. The exact deviation cannot be predicted without knowing the spatial autocorrelation structure and the details of the sampling design. This is not accounted for by subtracting $E[I(r)] = -1 / (N - 1)$ for Moran's $I(r)$. On the other hand, an empirical variogram can be used to estimate the real population variance accounting for autocorrelation, usually by fitting a theoretical variogram model. Hence, the above exemplified problem of Moran's I and Geary's c can be avoided.

DEVELOPMENT OF METHODS

Definition of Genetic Variograms

This paragraph shows how variance-based measures of genetic distance can be estimated from pairwise comparisons, so that variograms can be defined that provide an estimate of genetic diversity as a function of geographic distance.

Variogram of molecular variance. The univariate definition of a variogram (equation 3) can be extended to multivariate data, in which case y_a and y_b are not two observations of the allele size y_l of a single locus l, but vectors $\mathbf{Y}_a$ and $\mathbf{Y}_b$ of two observations of the number of repeats at L loci. The empirical semi-variance $\hat{\gamma}(r)$ becomes half the squared Euclidean distance between $\mathbf{Y}_a$ and $\mathbf{Y}_b$ and is equal to the sum of the empirical semi-variances $\hat{\gamma}_l(r)$ of the number of repeats y_l (WAGNER 2003):

$$\hat{\gamma}(r) = \frac{1}{2n_r}\sum_{a<b}\sum_{l=1}^{L} x_{ab}^{(r)}\left(y_{la} - y_{lb}\right)^2 = \frac{1}{n_r}\sum_{a<b}\sum_{l=1}^{L} x_{ab}^{(r)}\hat{\gamma}_l(a,b) = \sum_l \hat{\gamma}_l(r). \qquad (4)$$

More generally, a multivariate variogram can be defined as a weighted average $\bar{\gamma}(r)$ of the component variograms $\hat{\gamma}_l(r)$, with equation 4 as the special case of $w_l = 1$:

$$\bar{\gamma}(r) = \sum_l w_l \hat{\gamma}_l(r) \qquad (5)$$

Most often, the variograms of the L loci will be weighted by $w_l = 1/L$.

Under SMM, genetic diversity is related to differences in allele size, where allele size y_{la} is defined as the number of repeats of gene copy a at locus l. The molecular variance of a single locus l with k alleles can be defined as (RENWICK *et al.* 2001):

$$\hat{V}_l = \frac{N}{N-1}\sum_k p_{lk}\left(y_{lk} - \bar{y}_l\right)^2 = \frac{1}{N-1}\sum_a \left(y_{la} - \bar{y}_l\right)^2 ,$$

$$\text{where } \bar{y}_l = \sum_k p_{lk} y_{lk} = \frac{1}{N}\sum_a y_{la} \qquad (6)$$

Equations 4 and 6 provide a distance-dependent estimate $\hat{V}(r)$ of the molecular variance $\hat{V}$, averaged over *L* loci, which can be used as a within-population analogue to R_{ST} to investigate isolation-by-distance effects within a continuous population:

$$\hat{V}(r) = \sum_{a<b}\sum_{l=1}^{L} \frac{w_l x_{ab}^{(r)}}{2n_r}\left(y_{la} - y_{lb}\right)^2 . \qquad (7)$$

The statistical significance of a departure of $\hat{V}(r)$ from its expected value under the null hypothesis of no spatial autocorrelation can be tested in a Mantel permutation test (LEGENDRE and LEGENDRE 1998). If the alternative hypothesis is positive spatial autocorrelation at short distances, a one-sided test with a progressive Bonferroni correction can be applied, where the significance level for the k^{th} distance class is α/k (LEGENDRE and LEGENDRE 1998; LICHSTEIN *et al.* 2002).

Variogram of gene diversity. The analysis of the genetic structure of a locus *l* under the IAM is often based on join-count statistics. The proportion of unlike joins between observations is equivalent to the sum of the variograms of a set of dummy variables z_k, where z_{ka} = 1 if gene copy *a* is of allele *k*, and z_{ka} = 0 otherwise:

$$\hat{\gamma}_l(r) = \sum_k \sum_{a<b} \frac{x_{ab}^{(r)}}{2n_r}\left(z_{lka} - z_{lkb}\right)^2 . \qquad (8)$$

Due to the inherent correlation between the dummy variables, $\hat{\gamma}_l(ab)$ will equal 1 if gene copies *a* and *b* are different alleles, and 0 if they are the same allele.

Gene diversity or expected heterozygosity of a locus is a key parameter in population genetics under IAM. Gene diversity H_l is the probability that two gene copies sampled with replacement differ at locus *l*. The unbiased estimator of gene diversity, $\hat{H}_l$, at locus *l* for a sample of *N* gene copies of *k* different alleles is (NEI 1978):

$$\hat{H}_l = \frac{N}{N-1}\left(1 - \sum_k p_k^2\right) \tag{9}$$

The variogram of multi-locus gene diversity $\hat{H}$ can thus be defined as:

$$\hat{H}(r) = \sum_{l=1}^{L} w_l \hat{\gamma}_l(r) = \sum_{a<b} \sum_{l=1}^{L} \frac{w_l x_{ab}^{(r)}}{2n_r} (z_{lka} - z_{lkb})^2 . \tag{10}$$

$\hat{H}(r)$ provides a within-population analogue to F_{ST}. As with $\hat{V}(r)$, the significance of an observed autocorrelation in $\hat{H}(r)$ can be tested with a Mantel permutation test.

Variogram of genotypic diversity. Genotypic diversity measured by Simpson's diversity *D* is similar to single-locus gene diversity $\hat{H}_l$, but instead of allele *k*, the multilocus genotype *g* is used, so that *D* is the probability of sampling two individuals of different multilocus genotypes. The unbiased estimator of genotypic diversity is:

$$\hat{D} = \frac{N}{N-1}\left(1 - \sum_g p_g^2\right) \tag{11}$$

The variogram of genotypic diversity $\hat{D}$ (Simpson diversity) is obtained by coding each multilocus genotype by a dummy variable z_g, which takes the value $z_g = 1$ if individual *a* is of genotype *g* and $z_g = 0$ if it is not. For a haploid organism, the analysis is based on gene copies, whereas for diploid organisms, genotype coding would normally reflect diploid genotypes. The

variogram of genotypic diversity, $\hat{D}(r)$, estimates the probability of sampling two individuals of different multilocus genotypes as a function of their distance in space and is calculated as:

$$\hat{D}(r) = \sum_{a<b} \sum_{g} \frac{x_{ab}^{(r)}}{2n_r} \left(z_{ga} - z_{gb}\right)^2 . \qquad (12)$$

Accounting for ploidy levels and clonality

This paragraph introduces a weighting scheme that accommodates different ploidy levels or multiple sampling of genets. If the different gene copies of the same diploid or polyploid organism are not assumed to be independent, e.g. due to inbreeding, one may want to restrict comparisons to gene copies from different individuals. In organisms with various ploidy levels, one may want to give equal weight to each individual independent of its ploidy level. Both problems can be solved by modifying the weights $x_{ab}^{(r)}$:

$$x'^{(r)}_{ab} = \begin{cases} x_{ab}^{(r)} \dfrac{1}{N_i} \dfrac{1}{N_j} & i \neq j \\ 0 & i = j \end{cases}, \qquad (13)$$

where N_i is the number of gene copies of individual i with gene copy a and N_j is the number of gene copies of individual j with gene copy b. The same type of weighting can be applied to account for multiple sampling of genets in clonal organisms. In that case, N_i is the number of gene copies from genet i etc.

The permutation test needs to be adapted so that instead of permuting gene copies, the individuals or genotypes are permuted.

Modeling of genetic variograms

Expected shape of spatial genetic structure. Theoretical models of isolation by distance predict that in a two-dimensional space and if certain conditions are met, kinship or relationship coefficients between individuals, as well as pairwise F_{ST} or R_{ST}, vary approximately linearly with the logarithm of distance (HARDY 2003; HARDY and VEKEMANS 1999; ROUSSET 1997). Thus, with some assumptions concerning the drift-dispersal-mutation equilibrium and the dispersal function, the observed spatial genetic structure can be quantified to infer gene dispersal parameters (VEKEMANS and HARDY 2004). The general approach, as described by VEKEMANS and HARDY (2004), is to estimate the probability of identity in state as a function of the spatial distance between individuals. Because this function depends on the variability and thus the mutation rate of the locus, it needs to be standardized, for instance by reference to random genes from a sample of individuals (ROUSSET 2000; ROUSSET 2002). The standardized values $F(r)$ for each distance class r are regressed against spatial distance (one-dimensional case) or against the logarithm of distance (two-dimensional case) to estimate the slope parameter $\hat{b}_F$. As $\hat{b}_F$ is negative and depends somewhat on the sampling design, VEKEMANS and HARDY (2004) proposed quantifying spatial genetic structure by a new statistic $Sp = \hat{b}_F / (1 - F_N)$, where F_N is the relatedness of immediate neighbors competing for the same resources and may be estimated by $F_{(1)}$, the value of $F(r)$ for the first distance class. If the observed, two-dimensional spatial genetic structure results solely from isotropic limited gene dispersal, a dispersal-drift equilibrium has been reached, and the sampling scale is appropriate for the dispersal distance of the organism, dispersal parameters can be estimated from Sp (VEKEMANS and HARDY 2004).

Exponential variogram model: Assuming an exponential relationship, the spatial genetic structure of a population can be summarized by fitting an exponential variogram model (Figure 1):

$$\gamma(r) = C_0 + C_1\left[1 - e^{-\frac{3r}{b}}\right], \qquad (14)$$

where C_0 is the nugget variance, or the proportion of the variance that is not spatially structured, and C_1 is the spatially structured variance component (LEGENDRE and LEGENDRE 1998). The sill $C = C_0 + C_1$ provides an estimate of the population variance based on spatially independent samples, i.e., accounting for spatial autocorrelation. The relative size of the nugget provides an estimate of F_N: $1 - C_0 / C = C_1 / C = \hat{F}_N$. This can be set to $F_{(1)}$ by fitting a fixed-nugget model, constraining the nugget variance to the observed semivariance for the first distance class.

The exponential model approaches the sill C asymptotically. Therefore, the range or slope parameter b indicates the practical range of the exponential variogram, i.e., the distance at which the curve reaches 95 % of the sill (JOURNEL and HUIJBREGTS 1978). It can be shown that $b = -3 / \hat{b}_F$. Directional dispersal or migration may lead to anisotropy, where the genetic structure depends on direction. If there is reason to expect anisotropy, directional variograms can be fitted, providing estimates of the slope parameter b for different compass directions.

A confidence interval for the slope parameter b may be estimated using the permutation method for the confidence interval for the matrix regression coefficient proposed by MANLY (1997). The residuals of the exponential model are randomly permuted many times to obtain the reference distribution of the correlation of the residuals with distance. A series of exponential models with varying range parameters is derived and the critical values are determined at which

the correlation of the residuals of these new models with distance is as strong as for the α/2 and the (1-α)/2 quantiles of the reference distribution. The two critical values provide the lower and upper limits of the confidence interval for the range parameter *b*.

APPLICATION TO THE GENETIC STRUCTURE OF *LOBARIA PULMONARIA*

Model organism: *Lobaria pulmonaria* is a foliose epiphytic lichen species of humid temperate and boreal regions of the northern hemisphere and cooler parts of the tropics (YOSHIMURA 1971). This clonal and recombinant species (WALSER *et al.* 2004), which produces both vegetative and sexual diaspores, is considered endangered in most parts of Central Europe (WIRTH *et al.* 1996) and in other industrialized regions. It is used as an indicator of ecological continuity (ROSE 1992) in natural forests and traditionally managed agro-forestry landscapes, such as wooded pastures and chestnut orchards (SCHEIDEGGER *et al.* 2002).

Data: We studied the spatial genetic structure of a continuous population of *L. pulmonaria* from the Swiss Jura Mountains. A hierarchical random sample of 461 thalli was collected from a pasture-woodland landscape. In a first step, 100 circular plots of 1 ha were randomly selected from the wooded parts of the study area. Within each plot, all suitable trees exceeding 5 cm in diameter at breast height were searched for *L. pulmonaria*. A maximum of 24 thalli were randomly selected from different trees in each of the 24 plots where the lichen was present. If there were fewer than 24 colonized trees, multiple thalli were sampled from the same tree, and if there were fewer than 24 thalli in a plot, every thallus found was included. This results in a heterogeneous data set that could exhibit spatial autocorrelation at varying scales.

DNA extraction and fragment length determination at six microsatellite loci (*LPu03, LPu09, LPu15, LPu16a, LPu20a, LPu27a*) specific to the haploid mycobiont using an ABI 3100-

automated sequencer (Applied Biosystems) followed WALSER *et al.* (2003). Allele assignment was performed using GENOTYPER 2.5 software (Applied Biosystems).

Statistical analyses: Omni-directional variograms of molecular variance $\hat{V}(r)$ and gene diversity $\hat{H}(r)$ were calculated according to equations 7 and 10, giving equal weight to each of the six loci. The first distance class of $r = 0$ contained pairs of thalli from the same tree. The lag distance was 50 m, starting with 1 – 50 m. The last distance class contained all sample comparisons at distances larger than 450 m. Autocorrelation was tested per distance class using a one-sided Mantel test with 500 permutations of the thalli and a progressive Bonferroni correction of $\alpha = 0.05/k$ for the k^{th} distance class up to the first non-significant value.

A second set of variograms $\hat{V}'(r)$ and $\hat{H}'(r)$ was calculated weighting each thallus by the number of occurrences of its multi-locus genotype within the population using modified weights $x'^{(r)}_{ab}$ (equation 13). Autocorrelation was again tested per distance class using a one-sided Mantel test with 500 permutations of the multi-locus genotypes, using the same settings as above.

An isotropic and four directional variograms of genotypic diversity $\hat{D}(r)$ were also calculated according to equation 12 giving equal weight to all samples.

Exponential variogram models were fitted to all variograms, using the weighted least squares algorithm (CRESSIE 1993) that minimizes the following expression:

$$\sum_r n_r \left(\frac{\hat{\gamma}(r)}{\gamma(r; C_0, C_1, b)} \right)^2 , \qquad (15)$$

where $\gamma(r; C_0, C_1, b)$ is the fitted semivariance for distance class r based on the exponential model with parameters C_0, C_1 and b.

All calculations were performed in R (IHAKA and GENTLEMAN 1996). The exponential variograms were fitted using the R library 'GSTAT' (PEBESMA and WESSELING 1998; PEBESMA in press).

Results and discussion: Based on the six microsatellite markers, we found 92 multi-locus genotypes of the haploid mycobiont of *L. pulmonaria*. All but nine multi-locus genotypes occurred in single 1-ha plots, and only one was spread over more than 210 m. The probability of origin by recombination was below 0.003 for all recurrent multilocus genotypes, suggesting that they arose from clonal propagation. Weighting for recurrent genotypes drastically reduced effective sample size from 461 thalli to 92 multilocus genotypes. For recurrent genotypes, pairwise comparisons were distributed over several distance classes: the first distance class contained an equivalent of 36.8 pairs (instead of 1,588 for all samples) and the other distance classes up to 450 m comprised 47.6 to 148.6 pairs (instead of more than 2000).

The spatial genetic structure of the studied *L. pulmonaria* population consisted of two patterns caused by clonal reproduction (variogram of genotypic diversity, $\hat{D}(r)$) and sexual reproduction (variograms of molecular variance, $\hat{V}(r)$, and gene diversity, $\hat{H}(r)$; Figure 2). Weighting for recurrent genotypes reduced the autocorrelation of the first distance class and the range estimate both for molecular variance and for gene diversity (Table 2). After weighting for clones, the range parameters *b* of the fitted exponential variogram models were smaller for molecular variance, $\hat{V}'(r)$, and gene diversity, $\hat{H}'(r)$, than for genotypic diversity, $\hat{D}(r)$, suggesting larger dispersal distances or several steps of clonal dispersal (Table 2). On the other hand, genotypic diversity showed a higher degree of autocorrelation for the first distance class, $\hat{F}_1$, which consisted of pairs of samples from the same tree. The fitting of directional variograms

for genotypic diversity revealed that spatial genetic structure extended further in the main wind direction (WSW – ENE) than in the other directions (Table 2).

The conventional estimates of the population variance, $\hat{V}$, $\hat{H}$ and $\hat{D}$, slightly underestimated the variance for spatially independent samples, i.e., the total sill *C* for all three measures of diversity (Table 2). In this specific example, however, weighting for recurrent genotypes largely compensated this bias.

GENERAL DISCUSSION

Advantages of variogram analysis

A geostatistical perspective on spatial genetic structure can provide explanations to many of the issues raised by VEKEMANS and HARDY (2004) and suggest new approaches to address them. First, the sampling design has a strong influence on the absolute values of Moran's *I* or other coefficients of relatedness, thus limiting comparability between studies, and on the distance at which these measures reach their expected value in the absence of spatial structuring, so that this distance only provides a somewhat arbitrary estimate of the extent of spatial genetic structure (VEKEMANS and HARDY 2004). This problem affects the analysis of kinship structure with Moran's *I* or relationship coefficients, where empirical values for larger distances tend to be slightly below zero, whereas in theory, negative kinship coefficients are not allowed (BARBUJANI 1987). Our simulation experiment showed that this effect is not simply due to sample size, but relates to the inclusion of autocorrelated samples in the estimation of the population variance, which is commonly used as a reference for rescaling correlograms and other measures of relatedness. Variogram modeling, on the other hand, provides an estimate of the population

variance accounting for spatial autocorrelation, and its model parameters are scaled by this corrected estimate.

Second, VEKEMANS and HARDY (2004) suggested that if a plot of $F(r)$ against distance r, e.g., a Moran's I correlogram, decreases steadily until some distance x and shows no further trend, this distance may be interpreted as the extent of spatial genetic structure. In geostatistical terms, this means that if the variogram represents a stationary spatial process as indicated by the presence of a sill, the range b can be estimated. Rather than visually identifying a critical distance at which the sill is reached, one would fit an exponential variogram model and estimate the practical range, where the curve reaches 95 % of the sill. This provides an estimate of the extent of spatial genetic structure. A confidence interval for the range parameter can be constructed using a method developed for matrix regression coefficients by MANLY (1997).

Third, the interpretation of Moran's I as a correlation coefficient or of other measures as the absolute degree of kinship or relationship is jeopardized by the dependence of the empirical values on the sampling design through an implicit rescaling (see above). The proposed empirical variograms, however, have direct interpretations independent of the sampling design, as they provide distance-dependent estimates of molecular variance $\hat{V}$, gene diversity $\hat{H}$,and genotypic diversity $\hat{D}$, providing within-population analogues to the population pairwise R_{ST} and F_{ST} statistics.

The variogram of molecular variance, $\hat{V}(r)$, for microsatellite data has a straight-forward interpretation as the variance in the number of repeats expected for samples at a given distance in geographic space. It corresponds directly to a plot of (half) the sum of squared size differences as used in AMOVA of microsatellite data (SCHNEIDER *et al.* 2000) or of $D_0/2$, where

D_0 is the average squared difference in repeat numbers for two alleles drawn from the same population (GOLDSTEIN *et al.* 1995).

The variogram of genotypic diversity, $\hat{D}(r)$, has a straight-forward interpretation as the probability of sampling two different multi-locus genotypes as a function of their spatial distance. The variogram of gene diversity, $\hat{H}(r)$, can directly be interpreted as the probability of sampling two different alleles given their distance in geographic space, averaged over different loci. For a single locus, it corresponds exactly to a plot of the proportion of unlike links against distance, but the variogram definition is computationally simpler than the explicit coding of links. This method could easily be adapted to dominant markers such as RAPDs, ISSRs or AFLPs.

Fourth, VEKEMANS and HARDY (2004) proposed a new statistic for quantifying spatial genetic structure, $Sp = - b_F / F_N$, which arguably is more robust than either of the two component measures that are sensitive to the sampling design (see above). Rather than taking the ratio of two potentially biased quantities, variogram modeling accounts for the source of this potential bias by estimating the variance between uncorrelated samples. Hence, the variogram parameters nugget variance, range and sill can be directly compared between studies. Furthermore, F_N can be estimated in two different ways. If the first distance class contains the direct neighboring samples, thus representing the smallest possible distance, the semivariance for this distance class can be used as an estimate of F_N (fixed nugget effect model). Alternatively, if the first distance class also contains not directly adjacent samples, the nugget effect needs to be fitted, providing an estimated of F_N.

Robust estimation of spatial genetic structure

Weighting for clonality and ploidy levels. The lichen example illustrated the importance of distinguishing between spatial patterns of clonality and of genetic diversity resulting from sexual recombination. Specifically, it is crucial to account for clonal patterns when analyzing patterns of genetic diversity within a population. The confounded pattern does not represent the average of the two component patterns, but their multiplication, so that the degree and extent of spatial genetic structure may be severely overestimated.

For diploid organisms, the weighting results in measures similar to the kinship coefficient by LOISELLE *et al.* (1995) and the relationship coefficient of STREIFF *et al.*(1998), as links within individuals are excluded. The weighting proposed here is more general and can equally be applied to organisms with variable ploidy levels and extended to the correlation coefficient *r* by SMOUSE and PEAKALL (1999) or to join-count statistics (EPPERSON 2003). The proposed weighting of clones solves the problem of arbitrary resampling of recurrent genotypes, which may bias the analysis of spatial genetic structure (HÄMMERLI and REUSCH 2003; REUSCH *et al.* 1999). Whether to weight for recurrent genotypes or not will depend on the research question (e.g., dispersal distances vs. distances between mates) and the type of organism under study (e.g., clonal organisms with physically connected or detached ramets).

Deviation from exponential relationship. Simulations showed that under isolation by distance on a two-dimensional grid, Moran's *I* typically drops from positive values at short distances to negative values at intermediate distances before reaching values just below zero for larger distances in the absence of a cline (EPPERSON 2003). In the *L. pulmonaria* example, the variograms of molecular variance and gene diversity showed evidence for such a humped distribution. This type of non-monotonic autocorrelation structure is often encountered in

geostatistical analysis and may arise from a periodic structure (PYRCZ and DEUTSCH 2003) or as a sampling artifact (JOURNEL and HUIJBREGTS 1978; PALMER and WHITE 1994). It may be modeled by a dampened sine hole-effect model (JOURNEL and HUIJBREGTS 1978; LEGENDRE and LEGENDRE 1998). In addition, a clinal structure may cause a linear change of variance with distance, which can be modeled by a linear variogram model. Anisotropic variograms can help identifying clinal structure, e.g., in the case of directional migration.

Robust variogram estimation. Testing of differences between spatial patterns from different populations is notoriously difficult, as the hypothesis concerns the underlying process and not its observed realization (FORTIN *et al.* 2003). Many measures of spatial genetic structure suffer from a high sampling variance (VEKEMANS and HARDY 2004). Thus, rigorous statistical testing requires a large number of replicate populations, each with a large internal sample. Several robust variogram estimators exist (CRESSIE and HAWKINS 1980; CRESSIE 1993). While CAVALLI-SFORZA (1984) used a robust variogram estimator based on the median variance, the *modulus variogram* (CRESSIE and HAWKINS 1980) has been shown to perform better than the basic variogram estimator in situation with at least 50 % nugget variance (CRESSIE 1993). We will perform simulations to assess to what degree robust variogram estimators may help reducing sample size within populations.

Conclusions

Most measures of spatial genetic structure are rescaled with reference to random samples from the population. This reference is itself estimated from the data set and subject to bias unless spatial autocorrelation is accounted for. Such bias limits the interpretation of absolute values of various measures of spatial genetic structure and poses problems to the comparison

between studies and to the estimation of biological parameters (VEKEMANS and HARDY 2004). Variogram modeling, on the other hand, estimates its reference value accounting for spatial autocorrelation, thus providing parameter estimates that are comparable between studies. Furthermore, the proposed variograms of molecular variance, gene diversity, and genetic diversity are directly interpretable without rescaling, as they provide a partitioning of genetic diversity by the distance between samples. While this paper focused on microsatellite data as interpreted either under IAM or SMM, the approach may easily be adapted to other types of genetic data. The formal integration with variograms makes the theory and tools of geostatistics available for population genetics, which may help to address some important challenges in bridging the gap between empirical studies of spatial genetic structure and theoretical approaches to isolation by distance.

ACKNOWLEDGMENTS

This research is part of a project funded by the Swiss National Science Foundation (SNF) under the NCCR Plant Survival. We thank Magnus Nordborg and two anonymous reviewers for their helpful comments on an earlier version of this manuscript.

REFERENCES

BALLOUX, F., and J. GOUDET, 2002 Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Mol. Ecol. **11:** 771-783.

BALLOUX, F., and N. LUGON-MOULIN, 2002 The estimation of population differentiation with microsatellite markers. Mol. Ecol. **11:** 155-165.

BALLOUX, F., L. LEHMANN and T. DE MEEUS, 2003 The population genetics of clonal and partially clonal diploids. Genetics **164:** 1635-1644.

BARBUJANI, G., 1987 Autocorrelation of gene frequencies under isolation by distance. Genetics **117:** 777-782.

BURROUGH, P. A., 1995 Spatial aspects of ecological data, pp. 213-251 in *Data Analysis in Community and Landscape Ecology*, edited by R. H. G. JONGMAN, C. J. F. TER BRAAK and O. F. R. VAN TONGEREN. Cambridge University Press, Cambridge.

CAVALLI-SFORZA, L. L., 1984 Isolation by distance, pp. 229 - 248 in *Human population genetics. The Pittsburgh Symposium*, edited by A. CHAKRAVARTI. Van Nostrand Reinhold, New York.

CHUNG, M. G., and B. K. EPPERSON, 2000 Clonal and spatial genetic structure in *Eurya emarginata* (Theaceae). Heredity **84:** 170-177.

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

CRESSIE, N. A. C., 1993 *Statistics for Spatial Data*. Wiley, New York.

CRESSIE, N., and D. M. HAWKINS, 1980 Robust estimation of the variogram. J. Int. Ass. Math. Geol. **12:** 115-125.

EPPERSON, B. K., 2003 *Geographical Genetics*. Princeton University Press, Princeton.

FENSTER, C. B., X. VEKEMANS and O. J. HARDY, 2003 Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). Evolution **57:** 995-1007.

FORTIN, M.-J., M. R. T. DALE and J. VER HOEF, 2001 Spatial analysis in ecology, pp. 2051-2058 in *The Encyclopedia of Environmetrics*, edited by A. H. EL-SHAARAWI and W. W. PIEGORSCH. John Wiley and Sons Ltd.

FORTIN, M. J., B. BOOTS, F. CSILLAG and T. K. REMMEL, 2003 On the role of spatial stochastic models in understanding landscape indices in ecology. Oikos **102:** 203-212.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995 An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463-471.

GOLDSTEIN, D. B., G. W. ROEMER, D. A. SMITH, D. E. REICH, A. BERGMAN *et al.*, 1999 The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics **151:** 797-801.

HÄMMERLI, A., and T. B. H. REUSCH, 2003 Genetic neighbourhood of clone structures in eelgrass meadows quantified by spatial autocorrelation of microsatellite markers. Heredity **91:** 448-455.

HARDY, O. J., 2003 Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. Mol. Ecol. **12:** 1577-1588.

HARDY, O. J., and X. VEKEMANS, 1999 Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. Heredity **83:** 145-154.

HARDY, O. J., and X. VEKEMANS, 2002 SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes **2:** 618-620.

IHAKA, R., and R. GENTLEMAN, 1996 R: a language for data analysis and graphics. J. Comp. Graph. Stat. **5:** 299-314.

ISAAKS, E. H., and R. M. SRIVASTAVA, 1989 *Applied Geostatistics*. Oxford University Press, New York.

JOURNEL, A. G., and C. J. HUIJBREGTS, 1978 *Mining Geostatistics*. Academic Press, New York.

LEGENDRE, P., and L. LEGENDRE, 1998 *Numerical Ecology*. Elsevier, Amsterdam.

LICHSTEIN, J. W., T. R. SIMONS, S. A. SHRINER and K. E. FRANZREB, 2002 Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. **72:** 445-463.

LOISELLE, B. A., V. L. SORK, J. NASON and C. GRAHAM, 1995 Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. J. Bot. **82:** 1420-1425.

MANLY, B. F. J., 1997 *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London.

MEIRMANS, P. G., E. C. VLOT, J. C. M. DEN NIJS and S. B. J. MENKEN, 2003 Spatial ecological and genetic structure of a mixed population of sexual diploid and apomictic triploid dandelions. J. Evol. Biol. **16:** 343-352.

MONESTIEZ, P., and M. GOULARD, 1997 Analysing spatial genetic structures by multivariate geostatistics: study of wild populations of perennial ryegrass (*Lolium perenne*), pp. 1197-1208 in *Geostatistics Wollongong '96*, edited by E. Y. BAAFI and N. A. SCHOFIELD. Kluwer, Dordrecht.

NEI, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583-590.

PALMER, M. W., 1990 The estimation of species richness by extrapolation. Ecology **71:** 1195-1198.

PALMER, M. W., and P. S. WHITE, 1994 Scale dependence and the species-area relationship. Am. Nat **144:** 717-740.

PARKS, J. C., and C. R. WERTH, 1993 A study of spatial features of clones in a population of bracken fern, *Pteridium aquilinum* (Dennstaedtiaceae). Am. J. Bot. **80:** 537-544.

PEBESMA, E., and C. G. WESSELING, 1998 Gstat, a program for geostatistical modelling, prediction and simulation. Comput. Geosci. **24:** 17-31.

PEBESMA, E. J., in press Multivariable geostatistics in S: the gstat package. Comput. Geosci.

PIAZZA, A., and P. MENOZZI, 1983 Geographic variation in human gene frequencies, pp. 444 - 450 in *Numerical Taxonomy*, edited by J. FELSENSTEIN. Springer, Berlin.

PYRCZ, M., and C. DEUTSCH, 2003 The whole story on the hole effect. GAA Newsletter **18:** 3-5.

RENWICK, A., L. DAVISON, H. SPRATT, J. P. KING and M. KIMMEL, 2001 DNA dinucleotide evolution in humans: Fitting theory to facts. Genetics **159:** 737-747.

REUSCH, T. B. H., W. HUKRIEDE, W. T. STAM and J. L. OLSEN, 1999 Differentiating between clonal growth and limited gene flow using spatial autocorrelation of microsatellites. Heredity **83:** 120-126.

ROSE, F., 1992 Temperate forest management: its effects on bryophyte and lichen floras and habitats, pp. 211-233 in *Bryophytes and Lichens in a Changing Environment*, edited by J. W. BATES and A. FARMER. Clarendon Press, Oxford.

ROUSSET, F., 1997 Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics **145:** 1219-1228.

ROUSSET, F., 2000 Genetic differentiation between individuals. J. Evol. Biol. **13:** 58-62.

ROUSSET, F., 2002 Inbreeding and relatedness coefficients: what do they measure? Heredity **88:** 371-380.

SCHEIDEGGER, C., P. CLERC, M. DIETRICH, M. FREI, U. GRONER *et al.*, 2002 *Rote Liste der gefährdeten Arten der Schweiz: Baum- und Erdbewohnende Flechten*. BUWAL, Bern.

SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 *ARLEQUIN Version 2.000: A Software for Population Genetic Data Analysis*. Genetics and Biometry Laboratory, University of Geneva, Geneva.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457-462.

SMOUSE, P. E., and R. PEAKALL, 1999 Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity **82:** 561-573.

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

STREIFF, R., T. LABBE, R. BACILIERI, H. STEINKELLNER, J. GLÖSSL *et al.*, 1998 Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. Mol. Ecol. **7:** 317-328.

VAN DER HULST, R. G. M., T. H. M. MES, M. FALQUE, P. STAM, J. C. M. DEN NIJS *et al.*, 2003 Genetic structure of a population sample of apomictic dandelions. Heredity **90:** 326-335.

VEKEMANS, X., and O. J. HARDY, 2004 New insights from fine-scale spatial genetic structure analyses in plant populations. Mol. Ecol. **13:** 921-935.

WAGNER, H. H., 2003 Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. Ecology **84:** 1045-1057.

WAGNER, H. H., 2004 Direct multiscale ordination with canonical correspondence analysis. Ecology **85:** 342-351.

WALSER, J. C., C. SPERISEN, M. SOLIVA and C. SCHEIDEGGER, 2003 Fungus-specific microsatellite primers of lichens: application for the assessment of genetic variation on different spatial scales in *Lobaria pulmonaria*. Fungal Genet. Biol. **40:** 72-82.

WALSER, J. C., F. GUGERLI, R. HOLDEREGGER, D. KUONEN and C. SCHEIDEGGER, 2004 Recombination and clonal propagation in different populations of the lichen *Lobaria pulmonaria*. Heredity.

WIRTH, V., H. SCHÖLLER, P. SCHOLZ, G. ERNST, T. FEUERER *et al.*, 1996 Rote Liste der Flechten (Lichenes) der Bundesrepublik Deutschland. Schriftenr. Vegetationskunde **28:** 307-368.

YOSHIMURA, I., 1971 The genus *Lobaria* of Eastern Asia. J. Hattori Bot. Lab. **34:** 231-364.

APPENDIX: WORKED EXAMPLE

Example data

The example data set consists of two artificial variables x_1 and x_2 that describe the fragment lengths x of two loci in $N = 6$ haploid individuals A - F along a transect t. There are three multi-locus genotypes g with differing frequencies:

	t	x_1	x_2	g
A	1	1	4	1
B	2	1	4	1
C	2	2	4	2
D	3	2	1	3
E	3	2	1	3
F	4	2	1	3

Spatial partitioning of molecular variance

The basic elements of spatial covariance are calculated as:

$$\hat{\gamma}_l(a,b) = \frac{1}{2}\left(y_{la} - y_{lb}\right)^2 .$$

For instance, the comparison of gene copy A to gene copies B and D provides:

$$\hat{\gamma}(A,B) = \frac{1}{2}\sum_i \left(x_{iA} - x_{iB}\right)^2 = \frac{1}{2}\left((1-1)^2 + (4-4)^2\right) = 0$$

$$\hat{\gamma}(A,D) = \frac{1}{2}\sum_i \left(x_{iA} - x_{iD}\right)^2 = \frac{1}{2}\left((1-2)^2 + (4-1)^2\right) = 5$$

The semivariance $\hat{\gamma}_i(a,b)$ for each pair of gene copies is tabulated in the following matrix:

$\hat{\gamma}(a,b)$	*A*	*B*	*C*	*D*	*E*	*F*
A	–	0	0.5	5	5	5
B	0	–	0.5	5	5	5
C	0.5	0.5	–	4.5	4.5	4.5
D	5	5	4.5	–	0	0
E	5	5	4.5	0	–	0
F	5	5	4.5	0	0	–

Matrix of distances *r*: Common geostatistical analysis omits distances of $r = 0$, so that an object is never compared to itself. For organisms such as the epiphytic lichen *Lobaria pulmonaria*, however, individuals may share the same two-dimensional geographic coordinates if they grow on the same tree. Therefore, it may be important to distinguish between different individuals separated by a distance of zero in two-dimensional space and the comparison of an individual with itself:

r	*A*	*B*	*C*	*D*	*E*	*F*
A	–	1	1	2	2	3
B	1	–	0	1	1	2
C	1	0	–	1	1	2
D	2	1	1	–	0	1
E	2	1	1	0	–	1
F	3	2	2	1	1	–

Variogram of molecular variance: The empirical variogram of molecular variance is calculated using equation 7, as is illustrated here for distance class $r = 2$ based on unique pairs only (upper or lower triangle of matrices $\gamma(a,b)$ and r):

$$\hat{V}(2) = \frac{1}{2n_2} \sum_{a,b|r(ab)\approx 2} \sum_i \left(x_{1a} - x_{1b}\right)^2 = \frac{1}{n_2} \sum_{a,b|r(ab)\approx 2} \gamma(a,b) = \frac{1}{4}\left[5 + 5 + 5 + 4.5\right] = 4.875\,.$$

The variance estimates $\hat{V}(r)$ for all distance classes *r* are listed below. A weighted average of the molecular variance per distance class $\hat{V}(r)$, weighted by n_r, the number of pairs of gene copies, provides the global variance $\hat{V}$.

h	$\hat{V}(r)$	n_r
0	0.25	2
1	2.4375	8
2	4.875	4
3	5	1
$\sum_r \hat{V}(r) n_r / n$	2.967	15

The total number of pairwise comparisons, *n*, is given by: $n = \sum_r n_r = \frac{N(N-1)}{2}$.

Variogram of gene diversity

For the variogram of gene diversity, a dummy variable z_{lk} is defined for each allele *k* at each locus *l*. The semivariance is calculated from the matrix of dummy variables following equation 8:

z_{lk}	z_{11}	z_{12}	z_{24}	z_{21}
A	1	0	1	0
B	1	0	1	0
C	0	1	1	0
D	0	1	0	1
E	0	1	0	1
F	0	1	0	1

$\hat{\gamma}(a,b)$	*A*	*B*	*C*	*D*	*E*	*F*
A	–	0	1	2	2	2
B	0	–	1	2	2	2
C	1	1	–	1	1	1
D	2	2	1	–	0	0
E	2	2	1	0	–	0
F	2	2	1	0	0	–

For each pair of observations *a* and *b*, the semivariance $\hat{\gamma}(a,b)$ equals the number of loci at which they differ. The variogram of gene diversity is:

r	$\hat{H}(r)$	n_r
0	0.5	2
1	0.875	8
2	1.75	4
3	2	1
$\sum_r \hat{H}(r) n_r / n$	1.133	15

Variogram of genotypic diversity

For the variogram of genotypic diversity following equation 12, a dummy variable z_g is defined for each genotype g, and the semivariance is calculated from the matrix of dummy variables:

z_g	z_1	z_2	z_3
A	1	0	0
B	1	0	0
C	0	1	0
D	0	0	1
E	0	0	1
F	0	0	1

$\hat{\gamma}(a,b)$	A	B	C	D	E	F
A	–	0	1	1	1	1
B	0	–	1	1	1	1
C	1	1	–	1	1	1
D	1	1	1	–	0	0
E	1	1	1	0	–	0
F	1	1	1	0	0	–

As a computational shortcut, the same result can be obtained by setting all values $\hat{\gamma}(a,b) > 0$ to 1.

The variogram of genotype diversity is:

r	$\hat{D}(r)$	n_r
0	0.5	2
1	0.625	8
2	1	4
3	1	1
$\sum_r \hat{D}(r) n_r / n$	0.496	15

Weighting for recurrent genotypes

Each gene copy *a* receives a weight w_a which is inverse to the number of gene copies of the same multilocus genotype g_a. According to equation 13, the matrix of weights $w_a w_b$ is:

	w_a	*A*	*B*	*C*	*D*	*E*	*F*	$\sum_b w_a w_b$
w_b		1/2	1/2	1	1/3	1/3	1/3	
A	1/2	–	–	1/2	1/6	1/6	1/6	1
B	1/2	–	–	1/2	1/6	1/6	1/6	1
C	1	1/2	1/2	–	1/3	1/3	1/3	2
D	1/3	1/6	1/6	1/3	–	–	–	2/3
E	1/3	1/6	1/6	1/3	–	–	–	2/3
F	1/3	1/6	1/6	1/3	–	–	–	2/3

If G= 3 is the number of genotypes, the sum of all weights is:

$$\sum w_a w_b = G(G-1) = 6 .$$

The variogram of molecular variance between genotypes is derived as:

$$\hat{V}(2) = \frac{1}{w_2} \sum_{a,b|r(ab)\approx 2} w_a w_b \cdot \gamma(a,b) = \frac{1}{1/6 + 1/6 + 1/6 + 1/3}\left[\frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 5 + \frac{1}{3}\cdot 4.5\right] = 4.8$$

r	$\hat{V}(r)$	w_r
0	0.5	1/2
1	3.28	3/2
2	4.8	5/6
3	5	1/6
$\sum_r \hat{V}(r) w_r / w$	3.33	3

TABLES

TABLE 1

Accuracy and precision of estimates of the population variance for different sampling designs.

	N	*Systematic*	*Random*	*Clustered*
Estimated population variance	10	1.019	1.026	0.584
	20	1.003	1.013	0.757
	50	1.011	1.003	0.904
	100	1.007	0.997	0.969
Standard deviation of estimates	10	0.477	0.478	0.810
	20	0.288	0.330	0.616
	50	0.151	0.192	0.401
	100	0.024	0.131	0.284
Proportion of autocorrelated pairs	10	0.00	0.016	0.448
	20	0.00	0.016	0.218
	50	0.00	0.016	0.089
	100	0.00	0.016	0.048

The autocorrelated data were generated by dividing an ordered vector of 500 random values from a standard normal distribution into groups of five consecutive values. The entire groups and the values within each group were reordered at random, representing a hypothetical transect where always five neighboring locations would show very similar values, with random steps between groups. Three types of samples were taken from the transect: (*i*) a random sample from all locations, (*ii*) a systematic sample selecting every fifth location, and (*iii*) a clustered sample, where entire groups of five neighboring locations were selected at random. Data simulation and sampling were repeated 1000 times for each sample size of 10, 20, 50, or 100.

TABLE 2

Variogram parameters for *Lobaria pulmonaria*.

Diversity measure	Weighting	Variogram modeling			Conventional estimation
		$\hat{F}_N = \frac{C_1}{C_0 + C_1}$	$b = -\frac{3}{\hat{b}_F}$	$C = C_0 + C_1$	$\hat{V}, \hat{H}, \hat{D}$
Molecular variance V	All samples	0.93	121.5	25.1	24.3
	Weighted for clones	0.66	58.2	25.7	25.5
Gene diversity H	All samples	0.87	135.9	0.67	0.64
	Weighted for clones	0.55	69.3	0.68	0.68
Genotypic diversity D	All samples	0.71	106.5	1.00	0.98
	WSW - ENE	-	158.7	1.00	-
	WNW - ESE	-	77.7	1.00	-
	SSW - NNE	-	89.7	0.99	-
	NNW - SSE	-	110.7	1.00	-

Estimated parameters of the exponential variogram model fitted to the variograms of molecular variance and gene diversity for a population of *Lobaria pulmonaria* ($N = 461$) assessed with six microsatellite markers, with and without accounting for recurrent genotypes, and for genotypic diversity (clonal structure), with and without accounting for compass direction. The relatedness between immediate neighbors, F_N, is estimated from the autocorrelation of the first distance class of thalli taken from the same tree. The range parameter b denotes the distance at which the curve reaches 95% of the sill and provides an estimate of the extent of spatial genetic structure. The total sill C estimates the population diversity (molecular variance, gene diversity or genotypic diversity) accounting for spatial autocorrelation, whereas the conventional estimators $\hat{V}$, $\hat{H}$ *and* $\hat{D}$ do not account for autocorrelation and are, therefore, susceptible to bias.

FIGURE LEGENDS

FIGURE 1: Empirical variogram (A) and correlograms (B) for an artificial, spatially autocorrelated random variable simulated on a grid of 30 x 30 cells. Each symbol denotes the semivariance $\gamma(r)$ (circles), Geary's $c(r)$ (squares), or Moran's $I(r)$ (triangles) calculated from all pairs of samples falling into each distance class r. The value for the last distance class of each series contains all pairs separated by more than 20 units and is drawn at the mean of the respective distances. In Figure 1A, the solid line represents the fitted exponential variogram model. The dotted line (sill) indicates the population variance as estimated accounting for autocorrelation. The dashed line indicates the practical range, where the curve reaches 95 % of the sill. The intercept (nugget variance) is the variance component that is not spatially structured. In Figure 1B, the solid line indicates the expected value of Moran's $I(r)$, which is very close to zero, whereas the dotted line marks the expected value of Geary's $c(r)$, which equals one.

FIGURE 2: Variogram of gene diversity $\hat{H}(r)$ for a population of *Lobaria pulmonaria*. Each symbol denotes the mean semivariance over six microsatellite loci averaged over all pairs of thalli within each distance class. The semivariance is unweighted (circles) or weighted for recurrent genotypes (squares). The lines indicate the corresponding fitted exponential models, the dashed line the non-spatial model. Values below the dashed line correspond to positive, values above to negative autocorrelation. Filled symbols indicate statistically significant positive autocorrelation based on a one-sided Mantel permutation test with progressive Bonferroni correction.

FIGURES

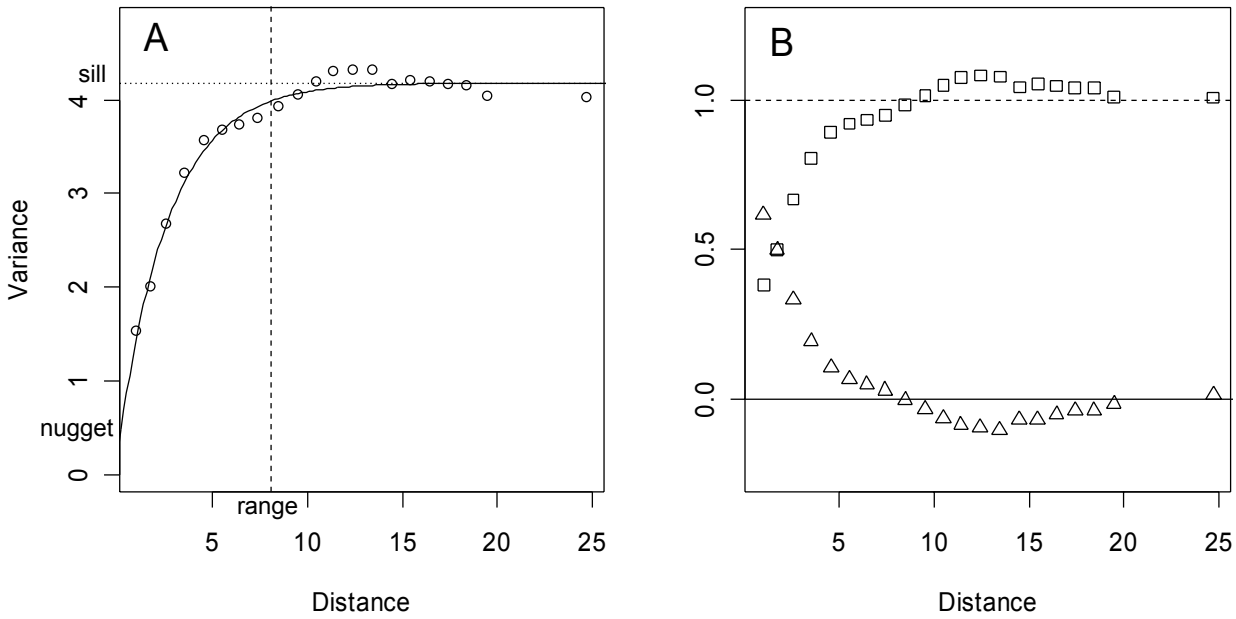

Figure 1

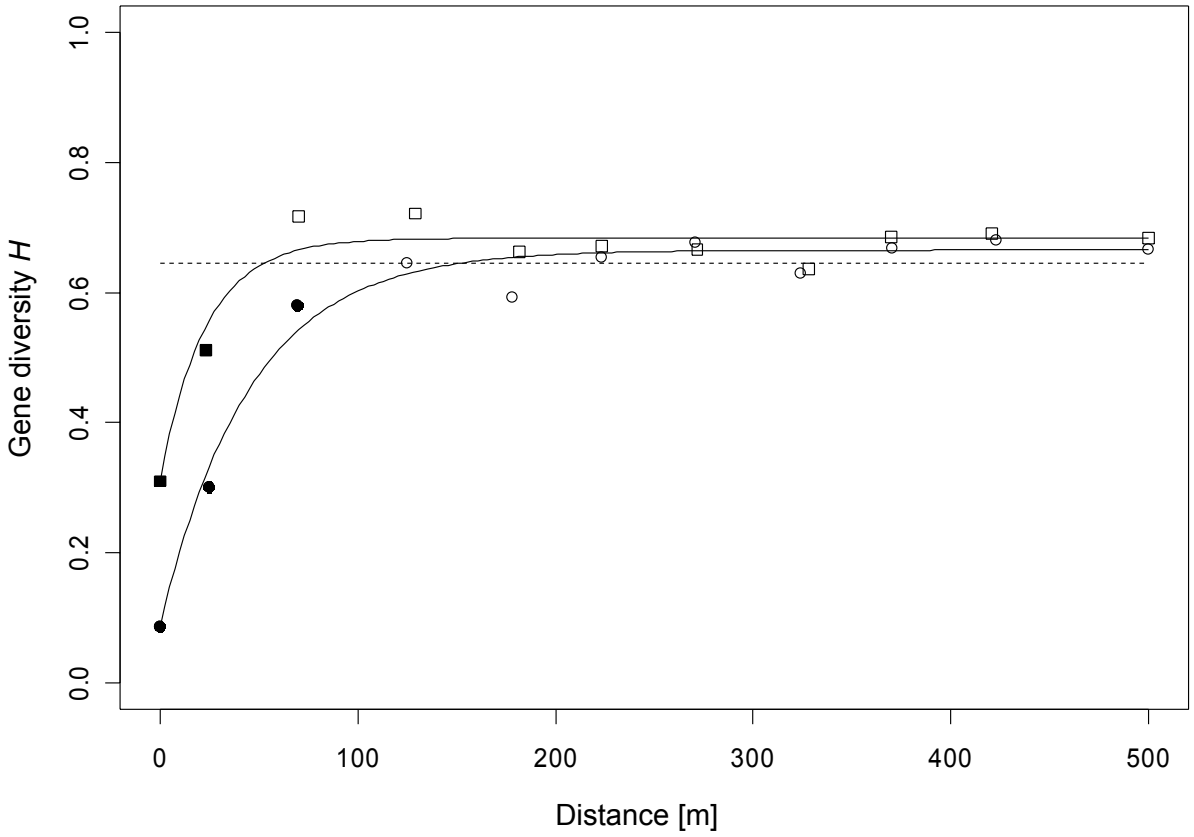
Gene diversity *H*
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Distance [m]

Figure 2