Heredity (2008) 101, 92–103
© 2008 Nature Publishing Group All rights reserved 0018-067X/08 $30.00
www.nature.com/hdy

ORIGINAL ARTICLE

Revealing cryptic spatial patterns in genetic variability by a new multivariate method

T Jombart, S Devillard, A-B Dufour and D Pontier
Laboratoire de Biométrie et Biologie Evolutive, UMR-CNRS 5558, Université de Lyon, Université Lyon 1, Villeurbanne Cedex, France

Increasing attention is being devoted to taking landscape information into account in genetic studies. Among landscape variables, space is often considered as one of the most important. To reveal spatial patterns, a statistical method should be spatially explicit, that is, it should directly take spatial information into account as a component of the adjusted model or of the optimized criterion. In this paper we propose a new spatially explicit multivariate method, spatial principal component analysis (sPCA), to investigate the spatial pattern of genetic variability using allelic frequency data of individuals or populations. This analysis does not require data to meet Hardy–Weinberg expectations or linkage equilibrium to exist between loci. The sPCA yields scores summarizing both the genetic variability and the spatial structure among individuals (or populations). Global structures (patches, clines and intermediates) are disentangled from local ones (strong genetic differences between neighbors) and from random noise. Two statistical tests are proposed to detect the existence of both types of patterns. As an illustration, the results of principal component analysis (PCA) and sPCA are compared using simulated datasets and real georeferenced microsatellite data of Scandinavian brown bear individuals (*Ursus arctos*). sPCA performed better than PCA to reveal spatial genetic patterns. The proposed methodology is implemented in the adegenet package of the free software R.

Heredity (2008) **101,** 92–103; doi:10.1038/hdy.2008.34; published online 30 April 2008

Keywords: landscape genetics; Moran's *I*; multivariate analysis; principal component analysis; spatial genetics

Introduction

Recently, growing attention is being devoted to taking landscape information into account in genetic studies (Manel *et al*., 2003). Among landscape features, space is most likely to influence the genetic structuring of a set of individuals or populations (Manel *et al*., 2004; Coulon *et al*., 2006). This structuring can exhibit different patterns, such as isolation by distance (Wright, 1943), clines (Haldane, 1948), metapopulations (Hanski and Simberloff, 1997; Kerth and Petit, 2005) and barriers to gene flow (Slatkin, 1985). Moreover, as technological advances have made obtaining spatial information easier, there is strong interest in including this information in the analysis of genetic data.

Exploiting the geographic dimension of genetic data is not new. Spatial information can be used *a posteriori* for graphical display purposes (for example, Bertranpetit and Cavalli-Sforza, 1991; Manel *et al*., 2004) or to measure spatial autocorrelation (for example, Sokal and Wartenberg, 1983; Sokal *et al*., 1986; Bertorelle and Barbujani, 1995; Smouse and Peakall, 1999). Such methods are useful descriptive tools to visualize, quantify and test spatial structure, but are not properly designed to investigate spatial patterns. For instance, ordinary ordination methods (Bertranpetit and Cavalli-Sforza, 1991) may reveal spatial patterns wherever they are obvious, but they are not constrained to do so. To investigate spatial structures other than the most evident, a method should be spatially explicit, that is, it should directly take spatial information into account as a component of the adjusted model or of the optimized criterion, thereby focusing on the part of the variability, which is spatially structured.

Such methods have been developed using different approaches. Within the analysis of molecular variance (AMOVA) framework (Excoffier *et al*., 1992), the spatial analysis of molecular variance (SAMOVA; Dupanloup *et al*., 2002) has proven useful for phylogeographic studies (Pramual *et al*., 2005; Tolley *et al*., 2006) to assess the spatial structure of a known number of populations. Within the Bayesian clustering framework, GENELAND (Guillot *et al*., 2005) and, more recently, a hierarchical Markov random field (HMRF) model (François *et al*., 2006) were proposed as improvements of STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003) by integrating geographic information to infer the number of populations and detect the genetic discontinuities among these populations (Coulon *et al*., 2006). Combining wombling and Bayesian assignment, Manel *et al*. (2007) proposed a method to detect genetic boundaries among multilocus genotypes. However, these approaches rely on a genetic model and require populations to meet Hardy–Weinberg equilibrium expectations (although the HMRF model allows inbreeding) and for linkage equilibrium to exist between loci (see Kaeuffer *et al*., 2007). This might be a problem as such expectations are unrealistic in many

Correspondence: Dr T Jombart, Laboratoire de Biométrie et Biologie Evolutive, UMR-CNRS 5558, Université Lyon 1—CNRS, 43 bd du 11 novembre 1918, Villeurbanne Cedex 69622, France.
E-mail: jombart@biomserv.univ-lyon1.fr

Received 12 February 2008; revised 19 March 2008; accepted 26 March 2008; published online 30 April 2008

cases and robustness of these methods have not been evaluated yet. Another, maybe more concerning, issue with these methods resides in the clustering approach itself: assigning individuals to groups is a likely inappropriate strategy when individuals are genetically structured as a cline. A last approach would be to use a Mantel correlogram (Legendre and Legendre, 1998, pp 736–738) to assess the variation of spatial autocorrelation in allelic frequencies across scales. However, this method is not wholly satisfying as it would only allow to detect spatial structuring, but would not permit to visualize the corresponding spatial patterns.

An appealing alternative for exploring genetic data is offered by reduced space ordination methods because their utilization is not contingent on a particular genetic model. Hardy–Weinberg equilibrium or linkage equilibrium are thus no longer required. Basically, these methods aim at summarizing strongly multivariate data into a few uncorrelated components, forming the so-called ‘reduced space’. For this summary to be meaningful, the components are chosen so as to reflect most of the variability in data, as defined by an optimized criterion (for example, variance among observations). Such methods can be applied on allelic frequency data to obtain a summary of the genetic variability among individuals or populations. A great illustration of such practice was offered by Menozzi *et al*. (1978), who used a principal component analysis (PCA; Pearson, 1901) to investigate the spatial patterns of the genetic variability, obtaining the well-known synthetic maps of human gene frequencies. More recently, PCA proved useful to correct for population stratification (Price *et al*., 2006) and to infer and test the number of subpopulations represented in a set of genotypes (Patterson *et al*., 2006). However, PCA can be criticized when applied to reveal spatial patterns. Indeed, this method finds synthetic variables on which the variance among genotypes is maximized, but does not take spatial information into account. PCA seeks genetic variability, not spatial structures; it is not a likely optimal method for revealing *cryptic spatial patterns*, that is, spatial patterns that are not associated to the highest genetic variation.

In this paper, we propose a new method, the spatial principal component analysis (sPCA), as a tool to investigate cryptic spatial patterns of genetic variability using georeferenced multilocus genotypes. Our method relies on a modification of PCA so that not only the variance between the studied entities (individuals or populations), but also their spatial autocorrelation is taken into account. The main results of the analysis are maps of entities scores allowing a visual assessment of the spatial genetic structures. The obtained scores reveal two types of patterns, which we define as *global* and *local* structures (*sensu* Thioulouse *et al*., 1995). Although both types express a fair amount of genetic variability, global structures display positive spatial autocorrelation whereas local ones display negative spatial autocorrelation. In other words, a global pattern would differentiate between two spatial groups or find a cline (or any intermediate state), whereas local scores would retrieve stronger genetic differences among neighbors than among random pairs of entities. As the studied entities can be genotypes or groups of genotypes (later referred to as ‘populations’, in a broad sense), global and local structures encompass a wide range of biological situations. For instance, global patterns of genotypes could indicate population patches in an island model, as well as cline wherever isolation by distance occurs. Local structures could arise when individuals from the same genetic pool are selected to avoid each other (repulsion) or to be attracted by individuals from other genetic pools. Similarly at the population level, global and local patterns may result from stratification (Price *et al*., 2006) or from adaptations to environmental variables that are inherently spatially structured (‘spatial dependence’; *sensu* Wagner and Fortin, 2005).

First, we explain how the spatial information is modeled explicitly through a connection network (Legendre and Legendre, 1998, pp 752–756) to be used in further computations. Then, we detail the meaning of Moran’s index of spatial autocorrelation (*I*; Moran, 1948, 1950) which is incorporated into the sPCA criterion, and show how it can identify global and local patterns in allelic frequency data. We then demonstrate how sPCA yields independent components optimizing the product of the variance and Moran’s *I*. As an aid to choose the sPCA scores to be interpreted, we developed two multivariate tests against the absence of global and local patterns. Our approach is illustrated and compared to PCA using simulated and real datasets. We conclude by discussing the prospective applicability of this method for the analysis of genetic data. The developed methodology is implemented in the adegenet package (Jombart, 2008) of the free software R (Ihaka and Gentleman, 1996; R Development Core Team, 2008).

Methods

Modeling spatial information

The first step of a spatially explicit method is to define how spatial information is introduced in the method. In sPCA, the detection of spatial structures uses the well-known Moran’s *I* (Moran, 1948, 1950), which relies on the comparison of the value of a quantitative variable (for example, allelic frequency) observed at one site (that is, individual or population) to the values observed at neighboring sites. This approach thus requires ‘neighboring sites’ to be defined. This is usually achieved by building a connection network (also called neighboring graph) which uses an objective criterion to define which entities are neighbors, and which are not. To simplify the definition of spatial structures provided in this paper, the term ‘neighbors’ is here restrained to immediate neighbors, that is, two vertices of the same edge of the connection network.

Several algorithms, whose review is beyond the scope of the present paper, can be used to build a connection network (Legendre and Legendre, 1998, pp 752–756). Although other spatially explicit methods, such as SAMOVA (Dupanloup *et al*., 2002) or GENELAND (Guillot *et al*., 2005), impose a specific connection network (the Delaunay triangulation; Upton and Fingleton, 1985), sPCA can use any graph. This plasticity makes sPCA adaptable to various spatial distributions. For instance, Delaunay triangulation or Gabriel graph (Gabriel and Sokal, 1969) would be adapted to evenly distributed entities, whereas distance-based neighborhood would be more appropriate to aggregated distributions. Moreover, connection networks can be refined manually to include empirical knowledge of the spatial connectivity among entities. Once the connection network is defined, the spatial information is stored in a binary connection matrix **M**, which is symmetrical and its lines and columns correspond to the same biological entities (as in a distance matrix). The values of **M** are 1 if the two considered entities are connected, and 0 otherwise. This matrix is used in the computation of Moran's *I* and therefore in sPCA.

Measuring spatial autocorrelation of an allelic frequency

Let us consider one allelic frequency measured on *n* individuals or populations. Once the binary connection matrix **M** is obtained, the spatial autocorrelation of this frequency can be quantified using Moran's *I*. The general form of this index can be written using matrix notation (Cliff and Ord, 1981, p 119), where **x** is the vector of *n* centered allelic frequencies and *W* is the sum of all the terms of **M**:

$$I(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}}{W}\frac{n}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \tag{1}$$

The meaning of this index depends only on its first component; the effect of the second component $n/\mathbf{x}^{\mathrm{T}}\mathbf{x}$ (which is the inverse of the variance of **x**) is to scale the variable so that *I* only reflects its spatial structure, not its variability. In this paper we use a version of *I* in which **M** is standardized so that the rows sum to one (Cliff and Ord, 1973, p 13). Denoting by **L** the resulting matrix, (1) becomes:

$$I(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}}{n}\frac{n}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \tag{2}$$

The expected value when the frequency observed at a site is independent of its neighbors (the null value, denoted I_0) equals $-1/(n-1)$ under a nonparametric model of the *n*! possible permutations of the data (Cliff and Ord, 1973, pp 29 and 32). Note that if *I* is to be interpreted quantitatively, its range of variation, which depends on the connection network, should be taken into account (De Jong *et al*., 1984).

This index has a straightforward interpretation. Let *i* and *j* indicate a row and a column of **L** ($i=1, n$; $j=1, n$). The row *i* contains positive values if *i* and *j* are neighbors, and 0 otherwise. As the terms of row *i* sum to one, these values are weights. Hence, the lag vector **Lx** computes, for a given entity, the mean frequency of its neighbors (Anselin, 1996). It follows that $\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x}$ is the scalar product of the allelic frequencies and their lag vector ($\mathbf{x}^{\mathrm{T}}\mathbf{L}\mathbf{x} = \langle \mathbf{x} \mid \mathbf{L}\mathbf{x} \rangle$): the frequency observed for any entity is multiplied by the mean frequency of its neighbors, and the obtained values are added over all entities.

Two types of spatial structuring can be observed in individuals or populations, whenever allelic frequency observed among neighbors are more similar or more dissimilar than expected in a random spatial distribution. These cases are illustrated using 20 fictitious populations (Figure 1). Patches of similar allelic frequencies (Figure 1a) lead to a highly positive *I* because the allelic frequency observed in a population is positively correlated to the allelic frequency of its neighbors (Figure 1c). Conversely, different neighbor to neighbor allelic frequencies (Figure 1b) lead to a highly negative *I*, as the value taken by any population is negatively correlated to those taken by its neighbors (Figure 1d). These two patterns are global and local structures as defined by Thioulouse *et al*. (1995). In the sPCA context, we define global and local patterns as entities being more genetically similar (respectively dissimilar) to their immediate neighbors than expected in a random spatial distribution.

As we have shown, Moran's *I* can be used to numerically detect such patterns using the frequencies of a single allele. Now we tackle the following question: how to reveal these patterns using a complete set of alleles?

Spatial principal component analysis

Two different objectives arise when analyzing a set of georeferenced allelic frequencies. On the one hand, we would like to summarize the genetic variability among the biological entities (individuals or populations) into a few informative components. On the other hand, we would also like to reveal existing spatial patterns.

A convenient solution to the first problem is to use a centered PCA (Pearson, 1901; Menozzi *et al*., 1978). This method analyzes a table **X** of *p* centered allelic frequencies (displayed in columns) measured on *n* biological entities (rows). The allelic frequencies define Euclidian distances between the *n* entities in $\mathbf{R}^p$, the *p*-dimensional space of real numbers. Finding the line of closest fit through the *n* points (Pearson, 1901) is the same as finding an axis in $\mathbf{R}^p$ on which the projections of the *n* entities are as widely scattered as possible, that is, where the Euclidian distances between the entities are best preserved. To fulfill this property, PCA seeks a scaled vector **u** ($\|\mathbf{u}\|^2=1$) containing *p* loadings (one per allele) so that the entities scores onto this axis ($\boldsymbol{\phi}=\mathbf{X}\mathbf{u}$) have a maximum variance. This can be reformulated as the maximization of:

$$\| \mathbf{X}\mathbf{u} \|^2_{1/n} = \frac{1}{n}(\mathbf{X}\mathbf{u})^{\mathrm{T}}\mathbf{X}\mathbf{u} = \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} \tag{3}$$

where $\|\mathbf{X}\mathbf{u}\|^2_{1/n} = \mathrm{var}(\boldsymbol{\phi})$. The solution is given by the first eigenvector of $(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$, which yields scores whose variance is maximized and equates to the highest eigenvalue. The second objective can be tackled by testing the spatial autocorrelation of the PCA scores, to assess whether they display significant spatial structure (Wartenberg, 1985b). However, these scores are appropriate only to summarize genetic variability and are in no way designed to reveal spatial patterns. Thus, there is a need for a methodology summarizing the genetic diversity and revealing spatial structures at the same time.

sPCA encompasses these two objectives. This new method finds a few independent synthetic variables that no longer optimize the variance of the entities' scores (as in PCA), but the product of their variance and of Moran's *I*. sPCA is closely related to Wartenberg's multivariate spatial correlation (MSC; Wartenberg, 1985a), but MSC constrains all alleles to have the same variance. This has the undesirable effect of masking the variability of the most informative alleles. Our method is also linked to that of Thioulouse *et al*. (1995), which also focuses on global and local structures. However, their method differs from sPCA in two ways. Firstly, it introduces nonuniform row weights giving more importance to the entities with many neighbors, whereas sPCA gives equal weights to all entities. Secondly, Thioulouse *et al*. (1995)

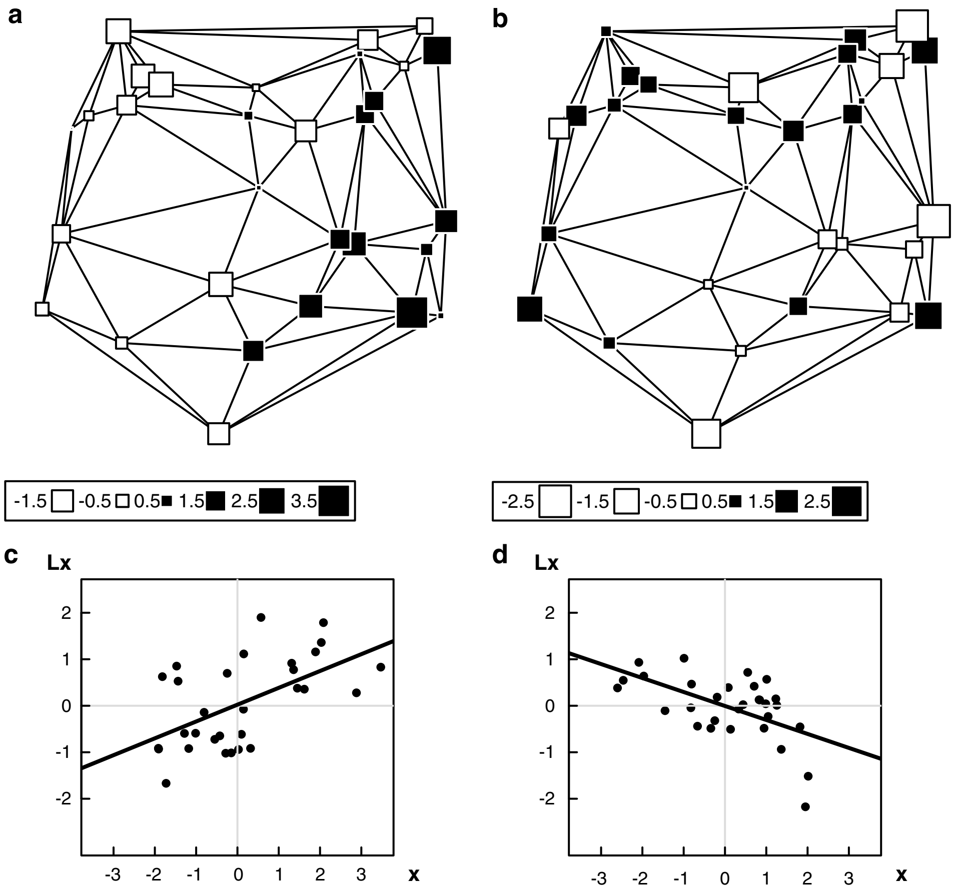

Figure 1 Illustration of global and local patterns of an allelic frequency for 20 fictitious populations overlying their sampling area. Each square represents the frequency of a population. Edges correspond to the connection network (Gabriel's graph). (**a**) Example of global structure, corresponding to $I(\mathbf{x})>I_0$. (**b**) Example of local structure, corresponding to $I(\mathbf{x})<I_0$. (**c**) Moran's scatterplot showing that in the global structure (a), the allelic frequency **x** of a population is positively correlated with the mean frequency of its neighbors, **Lx**. The line corresponds to the linear regression of **Lx** on **x**. (**d**) Conversely, the Moran's scatterplot associated with the local structure (**b**) shows that frequency **x** of a population is negatively correlated with the mean value of its neighbors, **Lx**.

used a globally standardized connection matrix instead of the row-standardized matrix **L**, and thus lost the meaning of the lag vector **Lx**.

sPCA seeks scaled axes **v** ($\|\mathbf{v}\|^2=1$) in $\mathbf{R}^p$ so that entity scores $\boldsymbol{\psi}=\mathbf{Xv}$ are both scattered and spatially autocorrelated. Similarly to the centered PCA (3), this relies on identifying the extreme values of a function (denoted C for 'criterion'):

$$C(\mathbf{v}) = \mathrm{var}(\mathbf{Xv})I(\mathbf{Xv}) = \frac{1}{n}(\mathbf{Xv})^{\mathrm{T}}\mathbf{LXv} = \frac{1}{n}\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{LXv} \qquad (4)$$

We show that the solution is given by the eigenvectors of the symmetric matrix $\frac{1}{2n}\mathbf{X}^T(\mathbf{L}+\mathbf{L}^{\mathrm{T}})\mathbf{X}$ associated with the highest and lowest eigenvalues (Supplementary Appendix A). As with PCA, other eigenvectors associated with less extreme eigenvalues display weaker structuring under the orthogonality constraint.

Although PCA and sPCA rely on a common approach, two major differences between these analyses must be underlined. Firstly, sPCA does not decompose the total variance into decreasing additive components. Instead, the product of the variance var($\boldsymbol{\psi}$) and the spatial autocorrelation $I(\boldsymbol{\psi})$ is separated into positive, null and negative components. Indeed, if the variance is always positive, the spatial autocorrelation can be positive as well as negative. Hence, while PCA focuses on the scores associated to the highest eigenvalues, sPCA encompasses two types of informative scores, both reflecting an aspect of the spatial patterning of the genetic variability. On the one hand, scores with a strong variance and a highly positive spatial autocorrelation (that is, global structures) correspond to highly positive eigenvalues. On the other hand, scores with a strong variance and a highly negative spatial autocorrelation (that is, local structures) correspond to highly negative eigenvalues. Note that these negative eigenvalues are thus useful tools to detect local patterns, and should not be ignored, as it was done in MSC (Wartenberg, 1985a).

Secondly, it makes no sense to compare a sPCA eigenvalue to the sum of all eigenvalues (as done in PCA) because this sum itself has no meaning: it can be low if there is no structure at all, as well as when there are strong global and local structures. Therefore, the percentage of total criterion associated to a given eigenvalue cannot be used as a rule to choose the structures to retain. However, as in other multidimensional methods, an abrupt decrease of the eigenvalues is likely to indicate the boundary between strong and weak structures (Legendre and Legendre, 1998). The interesting patterns are displayed graphically, and their spatial autocorrelation is measured using Moran's I. Note that it is meaningless to test the I of the sPCA scores, as is done in PCA (Wartenberg, 1985b) because the sPCA scores are already optimized regarding spatial autocorrelation.

Multivariate tests to detect global and local structuring

Sometimes, the sPCA eigenvalues may not clearly indicate if global and/or local structures should be interpreted. A first aid would be to assess if there are significant global and local patterns in the data. We developed two statistical tests (a global and a local test) to answer to these questions.

These tests rely on the spectral decomposition of the row-standardized connection matrix $\mathbf{L}$ into Moran's eigenvector maps (MEMs; Griffith, 1996; Dray *et al.*, 2006). These vectors are uncorrelated variables modeling different spatial structures; they were initially used in geography for spatial filtering purposes, that is, to remove spatial autocorrelation from the residuals of a statistical model (Griffith, 2000). In ecology, MEMs are used as explanatory variables in linear modeling approaches to model complex spatial patterns (Dray *et al.*, 2006; Griffith and Peres-Neto, 2006). Each of these spatial predictors is associated to a Moran's I and can therefore be characterized either as a global or a local pattern. We denote $\mathbf{E}^+$ the matrix whose columns are the global MEMs of $\mathbf{L}$, and $\mathbf{E}^-$ the matrix storing local MEMs. As there are always $(n-1)$ MEMs for n locations, these vectors can fully decompose a centered allelic frequency $\mathbf{x}$ into global and local spatial structures using simple linear regression. Note that this decomposition is not subject to multicollinearity troubles because MEMs are orthogonal: each MEM explains a different part of the variance of $\mathbf{x}$, which is measured by the corresponding coefficient of determination (R^2). This can be applied to the p centered allelic frequencies of matrix $\mathbf{X}$, yielding $p\mathbf{x}(n-1)$ coefficients of determination (one for each allele/MEM combination) which are stored separately for $\mathbf{E}^+$ and $\mathbf{E}^-$ (see detailed computations in Supplementary Appendix B). The R^2 of alleles with vectors of $\mathbf{E}^+$ are used in the global test, whereas R^2 computed with MEMs of $\mathbf{E}^-$ are used in the local test.

The basic idea behind our testing procedures is that if a global (respectively local) pattern exists among individuals (or populations), a large number of alleles is expected to be fairly correlated to at least one vector of $\mathbf{E}^+$ (respectively $\mathbf{E}^-$). To detect this, the mean R^2 across alleles is computed for each MEM. Denoting these means by t_j ($j=1, q$), a vector $\mathbf{t}$ containing all t_j is then obtained ($\mathbf{t}=[t_1 \ldots. t_j \ldots. t_q]^T$). To detect an eventual MEM with which all alleles would be significantly correlated, the test statistic used in both procedures is the maximum of $\mathbf{t}$ values, denoted max($\mathbf{t}$). The null hypothesis ($H0$) is that allelic frequencies of the individuals (or populations) are distributed at random on the connection network. Alternative hypotheses are that allelic frequencies of the studied entities display at least one global (respectively local) spatial structure. The distribution of max($\mathbf{t}$) under H_0 is obtained by a Monte Carlo procedure involving a large (say at least 999) number of permutations. For each permutation, the rows of $\mathbf{X}$ are randomized and max($\mathbf{t}$) is computed. In both tests, the P-value is defined as the relative frequency of permuted statistics equal to or higher than the initial value of max($\mathbf{t}$). We verified that the type I errors of both tests were correct using simulated datasets (see Supplementary Appendix B).

Illustrations

Simulated data: simple structures

This illustration compares the results of PCA and sPCA on three simulated datasets containing simple spatial structures: two global patterns (patches, cline) and one local structure (repulsion). For discontinuous patterns (patches and repulsion, Figures 2a, b, e and f), three populations of 500 diploid individuals were simulated using EASYPOP version 2.0.1 (Balloux, 2001), using a hierarchical island model to have different levels of genetic differentiation. Migration rate between populations 1 and 2 was set to 0.005 and to 0.002 between population 3 and the other two. Genotypes consisting in 20 microsatellite-like loci were obtained after 1000 generations using a KAM mutation model (that is, loci mutate to any new allelic state with the same probability) with a mutation rate of 0.0001 and 50 possible allelic states. For the continuous pattern (cline, Figures 2c and d), 4 populations of 500 diploid genotypes were simulated under an isolation-by-distance process. Spatial coordinates of the four populations were set in a two-dimensional space to (0, 0), (0, 2), (2, 0) and (2, 2). Dispersal distances were drawn from a negative exponential distribution with a mean of 1. EASYPOP computes the migration rates as $\exp(-r^*d_{ij})$, where d_{ij} is the distance between populations i and j, and r is the inverse of the dispersal distance. The other input parameters of this simulation were the same as in the simulation of discontinuous patterns. All analyzed data were obtained by randomly sampling genotypes from the created populations. Spatial coordinates were defined using the R software to create the various spatial structures.

Three datasets of 80 georeferenced genotypes were created: (1) two patches of 35 individuals from populations 1 and 2 with 10 individuals from population 3 randomly distributed; (2) 40 individuals from populations 1 and 2 forming a cline with 40 individuals from population 3 and 4 randomly distributed; (3) 30 individuals from population 3 distributed in repulsion among a total of 50 individuals of populations 1 and 2.

Data were analyzed using the R software, especially the *ade4* package for multivariate analysis (Chessel *et al.*, 2004; Dray *et al.*, 2007), *spdep* for spatial methods (Bivand, 2007) and *adegenet* for genetic data handling, sPCA and global/local tests (Jombart, 2008). The same procedure was applied to each dataset: first, data were analyzed by PCA, using Moran's I test to detect spatial structuring in the PCA scores; second, data were analyzed by sPCA using global and local tests (with 9999 permutations) as an aid to select the structures to be interpreted. All connection networks were defined using the Delaunay triangulation (Upton and Fingleton, 1985), a common graph that underlies several other methods (Dupanloup *et al.*, 2002; Guillot *et al.*, 2005; François *et al.*, 2006). The patches of the first dataset were retrieved by the first PCA scores (Figure 2a), which were significantly autocorrelated ($I=0.228$, $P=0.0005$). However, these patches appeared more clearly on the first global scores of sPCA

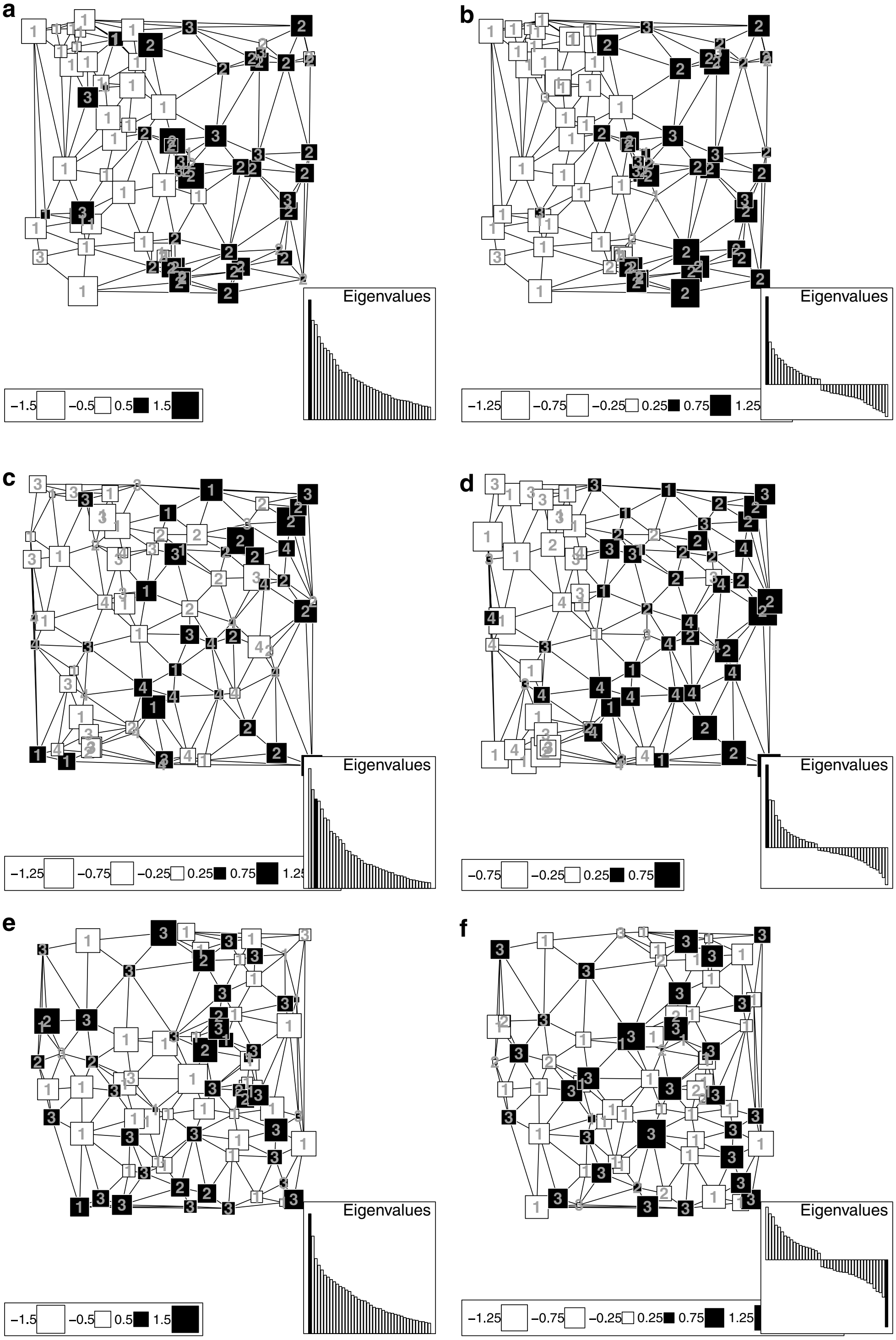

Figure 2 Analyses of simple global and local structures among 80 genotypes from three different populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a genotype and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which genotypes belong. (**a**, **b**) Two patches with random noise. (**c**, **d**) A cline with random noise. (**e**, **f**) Repulsion with random noise. (**a**, **c**, **e**) First PCA scores. (**b**, **d**) First global scores of sPCA. (**f**) First local scores of sPCA.

(Figure 2b). The global test confirmed the existence of global pattern (max(**t**) = 0.0166, $P = 0.0011$), whereas the local test did not detect any local structure (max(**t**) = 0.0140, NS). In the second dataset, PCA overlooked the cline, showing a weak spatial pattern on the third principal component (Figure 2c; $I = 0.128$, $P = 0.022$), whereas sPCA completely retrieved it (Figure 2d). Note that the first global structure is indeed a cline—and not patches—because genotypes situated in the middle of the distribution have less extreme scores (smaller squares). The global test confirmed the presence of global structure (max(**t**) = 0.0161, $P = 0.0038$), whereas the local test detected no local pattern (max(**t**) = 0.0117, NS). In the third dataset, the local pattern (repulsion among genotypes from population 3) was not identified by PCA (Figure 2e; $I = -0.054$, NS). On the contrary, the first local score of sPCA revealed this pattern: large black squares (population 3) are rarely found as neighbors and tend to be surrounded by white ones (genotypes from other populations) more often than at random. The global test did not detect any global pattern (max(**t**) = 0.0132, NS), whereas the local test was significant (max(**t**) = 0.0174, $P = 0.0008$).

Simulated data: complex structures in individuals

This illustration compares the results of PCA and sPCA using a simulated dataset in which different structures are mixed. Four populations of 500 diploid individuals were simulated using EASYPOP, following a hierarchical island model. Migration rate between populations 1 and 2 was set to 0.005, and to 0.002 for other populations. Otherwise, all parameters were those used in the previous illustration. A random sample of 80 genotypes was obtained, with unequal sample sizes (from population 1 to 4, sizes were 30, 30, 10, 10). Spatial coordinates were defined so that: (1) the 60 individuals from populations 1 and 2 were structured as a cline; (2) the 10 individuals from population 3 were distributed randomly; (3) the 10 individuals from population 4 were structured in repulsion.

This dataset was analyzed as previously, first by PCA and then by sPCA. The Delaunay triangulation was employed to model the spatial connectivity among genotypes. Two axes were retained for PCA (Figures 3a and c). No clear spatial pattern was revealed by PCA. The cline between populations 1 and 2 seemed split between the first (Figure 3a) and the second scores (Figure 3c), whereas the local structure induced by individuals from population 4 does not appear clearly on either axis. The Moran's I tests did not detect significant autocorrelation in either scores ($I = 0.081$, NS; $I = -0.014$, NS). On the contrary, sPCA revealed both structures. The first global and local scores were retained (Figures 3b and d). The global scores clearly differentiated populations 1 and 2, even if it was not clear whether this global structure consisted in two patches or in a cline (Figure 3b). The global

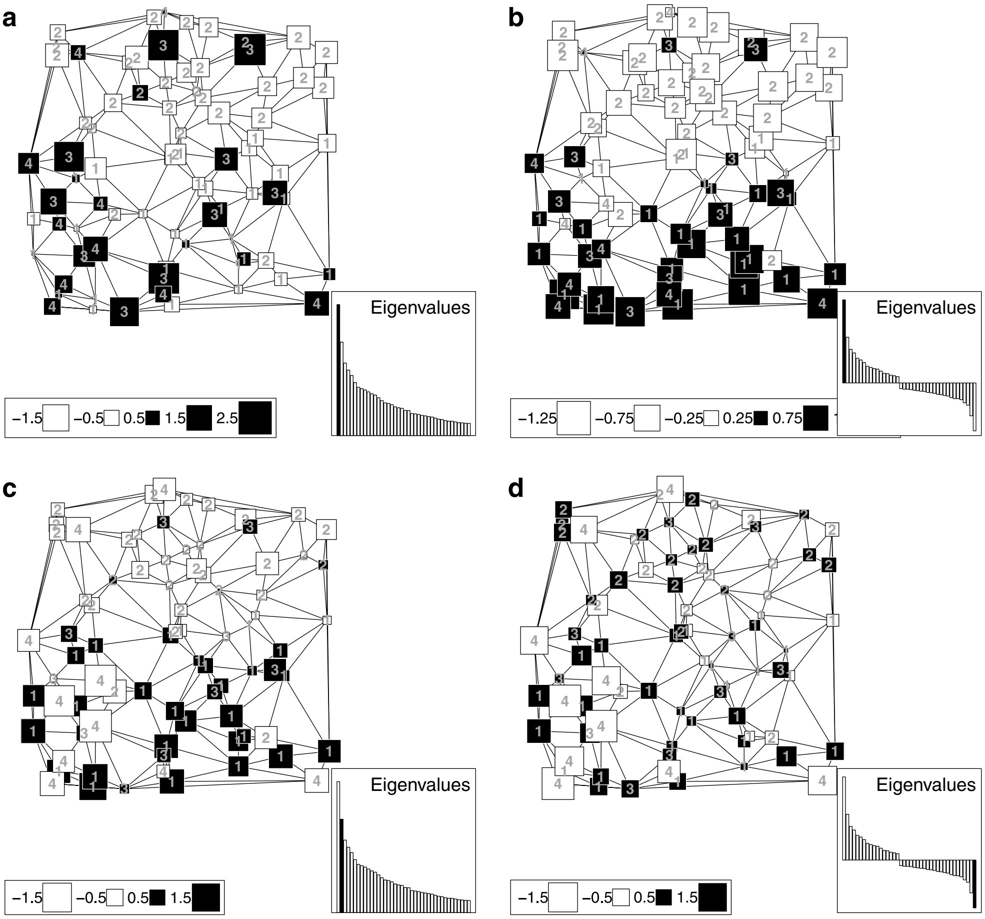

Figure 3 Analyses of complex global and local structures among 80 genotypes from four different populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a genotype and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which genotypes belong. (**a**) First PCA scores. (**b**) First global scores of sPCA. (**c**) Second PCA scores. (**d**) First local scores of sPCA.

test confirmed that a global structure existed ($\max(\mathbf{t}) = 0.0200$, $P = 0.0005$). The first local score clearly emphasized to genetic differences of individuals from population 4 (large white squares) from their immediate neighbors (other populations). The local test was consistently significant ($\max(\mathbf{t}) = 0.0270$, $P = 0.0001$).

Simulated data: complex structures in populations

This illustration had the same objective as the previous one, but involved populations rather than individuals. Three populations of 500 diploid individuals were simulated using EASYPOP, following an island model with a migration rate of 0.01. Other parameters of simulations were the same as in previous illustrations. Subpopulations (16) were created by taking random samples of 30 individuals from a given population; 10 subpopulations were thus obtained from population 1 and 2, and 6 were drawn from population 3. Spatial coordinates of subpopulations were defined so that: (1) the 20 subpopulations from populations 1 and 2 were structured in two patches; (2) the 6 subpopulations from population 3 were distributed following a local pattern.

After transforming the data into allelic frequencies for each subpopulation, a PCA and an sPCA were performed. The Delaunay triangulation was used to model the spatial connectivity among subpopulations. The PCA eigenvalues showed that two strongly structured axes were to be retained (Figures 4a and c). This was likely due to the fact that the variability among subpopulations was essentially an interpopulation variability: only two axes are required to differentiate three populations. The first PCA scores displayed a significant spatial structure (Figure 4a; $I = 0.265$, $P = 0.0096$), but it was merely as a by-product: it simply differentiated the population 1 from the two others. Similarly, the second PCA scores differentiated the population 3 from the others (Figure 4c), but these scores were not spatially structured ($I = 0.031$, NS). The sPCA eigenvalues clearly showed that one global and one local axes were to be retained (Figures 4b and d). The first global scores (Figure 4b) found the two patches of subpopulations from populations 1 and 2, giving rather low values to the scores of population 3 (small squares). The global test detected the existence of spatial pattern ($\max(\mathbf{t}) = 0.131$, $P = 0.0065$). The local scores highlighted the differences between subpopulations from population 3 with the neighboring subpopulations (Figure 4d). The local test was also significant ($\max(\mathbf{t}) = 0.133$, $P = 0.0053$).

Scandinavian brown bear data

The Scandinavian brown bear dataset illustrated other methods such as the wombling approach of Manel *et al.*

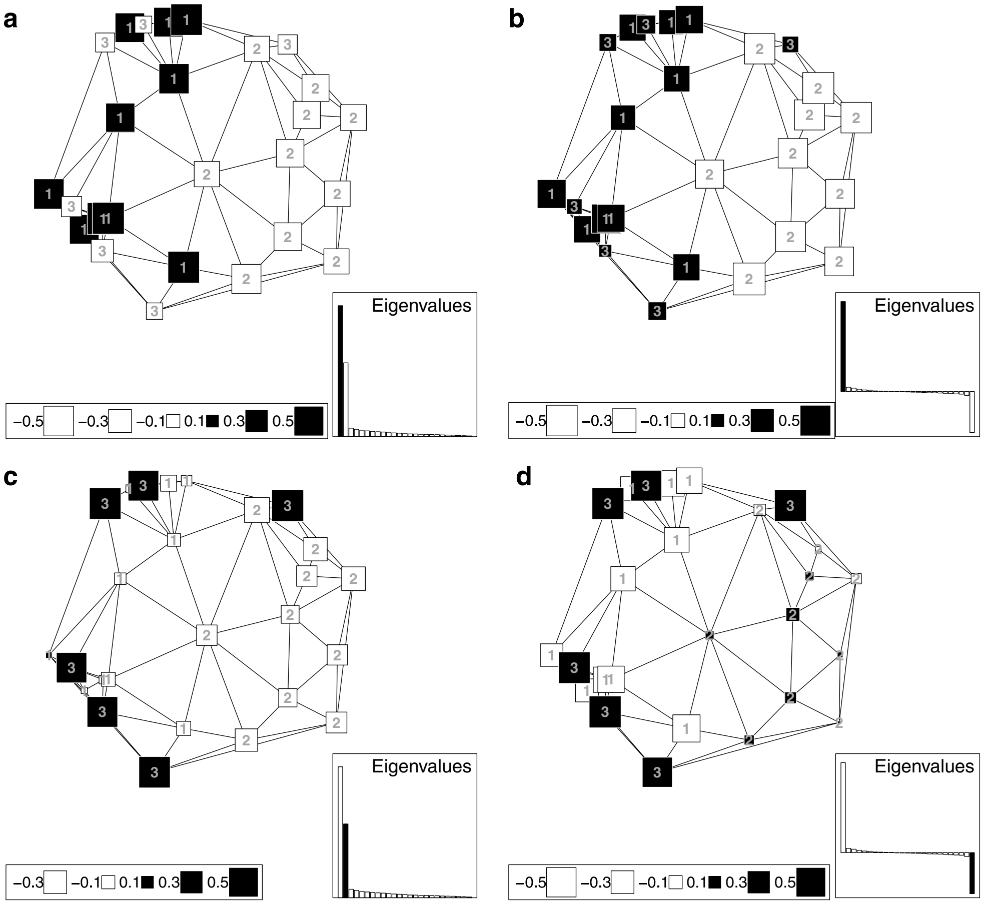

Figure 4 Analyses of complex global and local structures among 16 subpopulations from three populations by principal component analysis (PCA) and spatial PCA (sPCA). Each square represents the score of a subpopulation and is positioned by its spatial coordinates. The eigenvalues corresponding to the displayed scores are filled in black on the screeplots. Numbers indicate the population to which subpopulations belong. (**a**) First PCA scores. (**b**) First global scores of sPCA. (**c**) Second PCA scores. (**d**) First local scores of sPCA.

(2007). These data contain the georeferenced genotypes of 964 brown bears sampled over 200 000 km^2 and typed for 18 microsatellite markers. A more complete description of the data will be found in Waits *et al.* (2000). Former studies stressed the need for identifying management units (MUs) among Scandinavian brown bear for conservation purposes (Waits *et al.*, 2000). Using density indicators, Swenson *et al.* (1998) suggested four different MUs. There seems to be a general agreement that the southernmost group is strongly differentiated from all the others because of different lineages (Manel *et al.*, 2004). Nonetheless, the number of MUs to be considered is still discussed: using microsatellites, Waits *et al.* (2000) confirmed the four groups suggested previously (Swenson *et al.*, 1998), whereas more recent studies found only three MUs (Manel *et al.*, 2007), considering northern individuals as from one MU instead of two. Expressed in terms of sPCA, different MUs would appear as global structures, each global score potentially differentiating between two MUs. Thus, these data seem appropriate to illustrate how sPCA can identify several spatial groups.

First, a centered PCA was performed on the allelic frequencies of the individuals. The first three scores were significantly positively autocorrelated, but only the first

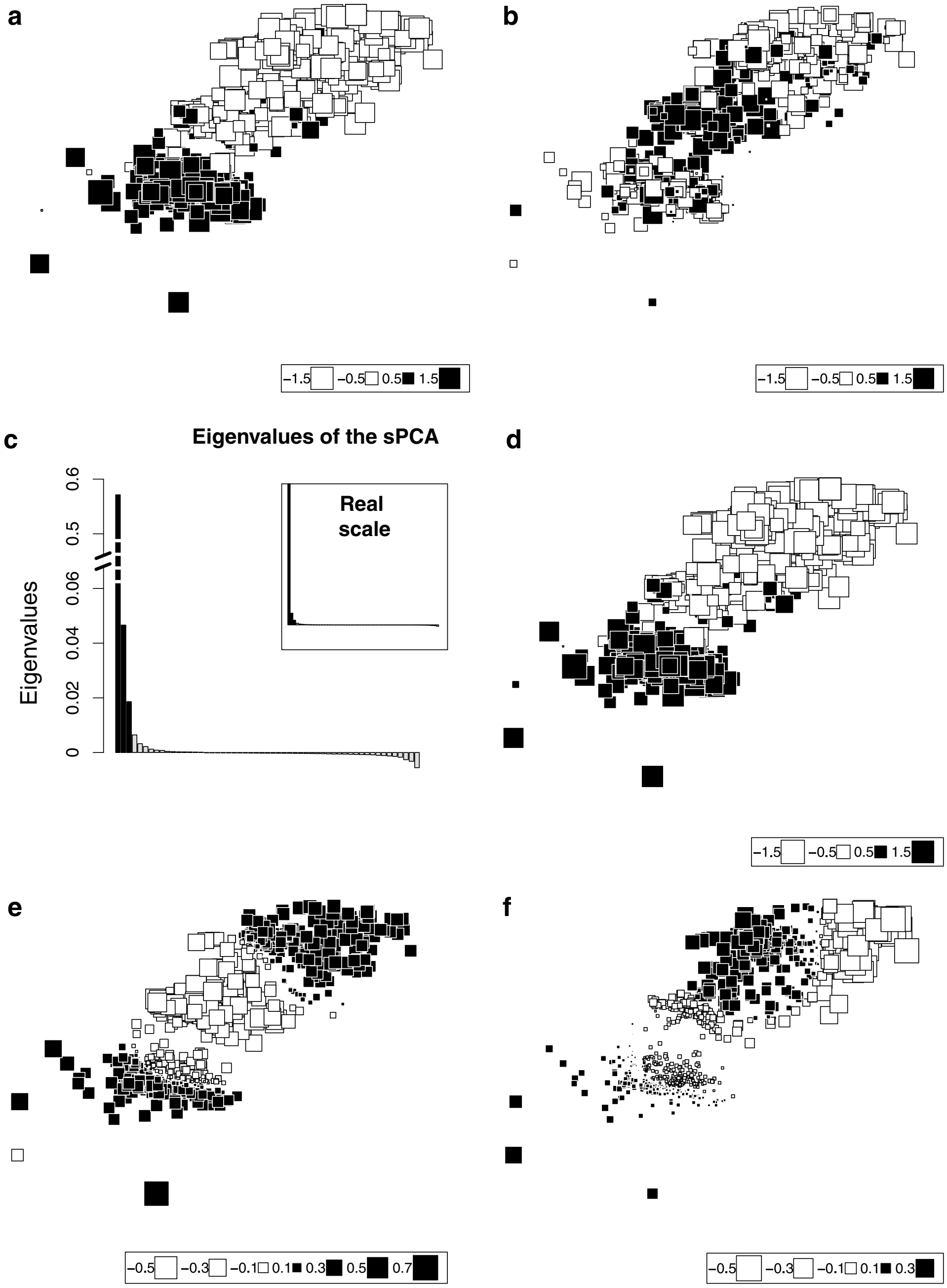

Figure 5 Analyses of Scandinavian brown bears data. (**a**) First and (**b**) second scores of the centered principal component analysis (PCA), displaying significant spatial structure ($I = 0.647$, $P = 0.001$; $I = 0.125$, $P = 0.001$). (**c**) Screeplot of spatial PCA (sPCA). The first eigenvalue (see real scale screeplot, top-right box) was truncated to better appreciate the others. Retained structures are filled in black. (**d**–**f**) First, second and third global scores of sPCA. Autocorrelation statistics were respectively: $I = 0.686$, $I = 0.147$, and $I = 0.122$ ($I_0 \approx 0$).

two had biologically meaningful I values ($I=0.647$, $P=0.001$; $I=0.125$, $P=0.001$; $I_0 \approx 0$). The first PCA scores differentiated the southern MU from all the others (Figure 5a). The second PCA scores (Figure 5b) were more difficult to interpret, but seemed to correspond in part to the middle MU identified in previous studies (Swenson *et al*., 1998; Manel *et al*., 2004). The third scores displayed a very small spatial autocorrelation, and the associated map was not interpretable (result not shown).

Second, we proceeded to sPCA. We used a distance-based connection network because the spatial distribution was fairly aggregated; such a graph ensures that genotypes inside aggregates have more neighbors than outliers. The threshold distance between any two neighbors was chosen as the minimum distance so that no individual was excluded from the graph. We call the resulting graph a *minimum distance neighboring graph*. The first sPCA eigenvalue was strikingly large compared to the others, but with no doubt the first three eigenvalues and corresponding scores were to be retained (Figure 5c). The first scores revealed the same pattern as in PCA (Figure 5d) and separated individuals from the southern MU from all the others, like in previous studies. This pattern was associated to a strong spatial autocorrelation ($I=0.686$). The second sPCA scores ($I=0.147$) clearly differentiated individuals from the 'middle' subpopulation (Waits *et al*., 2000) from the others (Figure 5e). By combining the first two global scores we thus recovered the three subpopulations found in previous studies (for example, Manel *et al*., 2007). But more interestingly, our analysis retrieved an additional weaker structure: undoubtly the third global scores ($I=0.122$) showed an east–west differentiation among northern individuals (Figure 5f). Contrary to the first pattern (Figure 5d), this structure does not show sharp boundaries between patches, but rather progressive changes from one patch to another, suggesting an isolation-by-distance process or progressive introgression. This may be the reason why a method based on boundary detection (Manel *et al*., 2007) overlooked this structure. The global test confirmed the existence of at least one global pattern ($\max(\mathbf{t})=0.0533$, $P=0.0001$) without detecting local structuring ($\max(\mathbf{t})=0.0043$, NS).

Discussion

We propose a spatially explicit multivariate method, sPCA, as a new tool to explore georeferenced multilocus genotypes and, therefore, to try to understand how geographical and environmental features structure genetic information. Although ordinary centered PCA yields scores that summarize the genetic variability among considered entities (individuals or populations), sPCA adds the constraint that the provided scores should be spatially autocorrelated and, thus, focuses on the spatial pattern of genetic variability. Two types of patterns are discriminated: global and local structures, corresponding respectively to large positive and large negative eigenvalues. Maps of sPCA scores are used to visually assess these patterns. As an aid to the interpretation of sPCA results, two Monte Carlo tests are proposed to detect the existence of global and local patterns. Simulated data illustrated that sPCA can retrieve simple structures (patches, clines, repulsion) as well as more complex patterns among genotypes or populations, and performs better in this task than PCA. The global and local tests efficiently detected the existing patterns, with a reliable type I error, and can therefore be used to assess which kind of pattern should be interpreted. sPCA also retrieved already known patterns in Scandinavian brown bear dataset, as well as more cryptic structures, which were overlooked by another method (Manel *et al*., 2007), but were biologically expected (Swenson *et al*., 1998).

Several points relative to the method should be discussed. Firstly, the spatial information is integrated using a connection network. This widely used approach allows taking virtually any type of spatial information into account. Contrary to other spatially explicit methods (Dupanloup *et al*., 2002; Guillot *et al*., 2005), we do not impose a specific connection network; one would have to choose from existing algorithms, and refine it according to what is known about the ecological connectivity in the system. It is important to keep in mind that sPCA is not intended to study the spatial connectivity among the considered entities; it aims at finding spatial structuring given that connectivity.

Secondly, sPCA is proposed mainly as an exploratory tool. For this purpose, our approach seems relevant as it is a reduced space ordination method; no assumptions are made about the data model. It is thus free, for instance, from modeling constraints like Hardy–Weinberg equilibrium assumptions, which are often violated when considering markers involved in selection processes. This is in contrast to, for instance, STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003), which assumes both Hardy–Weinberg equilibrium and linkage equilibrium. Nonetheless, further investigations should be devoted to link sPCA to existing population genetics models. Indeed, the ability of spatial autocorrelation based methods (of which sPCA is one) for inferring genetic processes has been a controversial topic (Sokal and Wartenberg, 1983; Sokal *et al*., 1989; Slatkin and Arter, 1991a, b), but useful studies have shown that Moran's I can be linked to population genetics models (Hardy and Vekemans, 1999). Similarly, a recent study demonstrated that the number of significant eigenvalues of PCA can be directly related to the number of subpopulations in a set of genotypes (Patterson *et al*., 2006). Such development with sPCA would surely enhance the interpretation of the provided results.

Thirdly, the efficiency of sPCA in different population genetics scenario remains to be investigated further, as it was done with spatial autocorrelation. For instance, we did not tackle the relative power of the analysis to reveal patterns due to directional selection (Epperson, 1990) or isolation by distance (Barbujani, 1987; Epperson, 1995). The influence of other parameters, such as the connection network or the level of genetic differentiation, should also be evaluated. These topics as well as comparisons of sPCA to other methods will be investigated using simulations in a next paper.

To conclude, we have shown that sPCA can be used and is useful at the scale of individuals as well as at a population scale. This suggests that our method could be an appropriate tool in different domains. As sPCA can be performed on data from individuals with no *a priori* knowledge of the studied system, our method should become a useful tool in landscape genetics studies (Manel *et al*., 2003), to link the revealed genetic patterns to landscape features and to explain genetic discontinuities in terms of environmental, behavioral or physiological barriers. Indeed, the sPCA scores can be correlated to other variables or included as dependent or independent variables in models, as long as their spatial autocorrelation is taken into account (Anselin, 2002). Moreover, sPCA can assess the genetic structuring of a set of fragmented populations, which seems especially relevant in conservation biology where this is common. It is particularly important to identify the most isolated populations, when introducing new individuals to maintain genetic diversity or to predict the spatial spread and maintenance of an introduced disease to control pest species. In these cases, sPCA may help to develop appropriate management and surveillance strategies for a disease. Therefore, the proposed method can be seen as a versatile tool for investigating the genetic structuring of set of individuals or populations, within different contexts.

Acknowledgements

This project would not have been possible without the help of Daniel Chessel, who provided insightful discussions about the method and whose comments greatly improved earlier versions of this manuscript. We thank the three anonymous reviewers and the associated editor whose comments also improved a former version of this manuscript. We also thank Jon E Swenson, Stéphanie Manel and Pierre Taberlet for providing the data collected by the Scandinavian Brown Bear Research Project. Finally, we are very grateful to John O'Brien, Nicolas Perrin and Christian Biémont for their helpful comments.

References

Anselin L (1996). Spatial analytical perspectives on GIS. In: Fisher M, Scholten H, Unwin D (eds). *The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association*. Taylor and Francis: London. pp 111–125.

Anselin L (2002). Under the hood. Issues in the specification and interpretation of spatial regression models. *Agric Econ* **27**: 247–267.

Balloux F (2001). EASYPOP (version 1.7): a computer program for population genetics simulations. *J Hered* **92**: 301–302.

Barbujani G (1987). Autocorrelation of gene frequencies under isolation by distance. *Genetics* **117**: 777–782.

Bertorelle G, Barbujani G (1995). Analysis of DNA diversity by spatial autocorrelation. *Genetics* **140**: 811–819.

Bertranpetit J, Cavalli-Sforza L (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. *Ann Hum Genet* **55**: 51–67.

Bivand R (2007). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-9.

Chessel D, Dufour A-B, Thioulouse J (2004). The ade4 package-I-one-table methods. *R News* **4**: 5–10.

Cliff A, Ord J (1973). *Spatial Autocorrelation*. Pion: London.

Cliff A, Ord J (1981). *Spatial Processes. Model & Applications*. Pion: London.

Coulon A, Guillot G, Cosson J-F, Angibault J, Aulagnier S, Cargnelutti B *et al*. (2006). Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Mol Ecol* **15**: 1669–1679.

De Jong P, Sprenger C, van Veen F (1984). On extreme values of Moran's *I* and Geary's *c*. *Geogr Anal* **16**: 17–24.

Dray S, Dufour A-B, Chessel D (2007). The ade4 package—II: Two-table and *K*-table methods. *R News* **7**: 47–54.

Dray S, Legendre P, Peres-Neto P (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecol Modell* **196**: 483–493.

Dupanloup I, Schneider S, Excoffier L (2002). A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* **11**: 2571–2581.

Epperson B (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* **124**: 757–771.

Epperson B (1995). Spatial distribution of genotypes under isolation by distance. *Genetics* **140**: 1431–1440.

Excoffier L, Smouse P, Quattro J (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: applications to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.

Falush D, Stephens M, Pritchard J (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

François O, Ancelet S, Guillot G (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Gabriel K, Sokal R (1969). A new statistical approach to geographic variation analysis. *Syst Zool* **18**: 259–278.

Griffith D (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Can Geogr* **40**: 351–367.

Griffith D (2000). A linear regression solution to the spatial autocorrelation problem. *J Geogr Syst* **2**: 141–156.

Griffith D, Peres-Neto P (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**: 2603–2613.

Guillot G, Estoup A, Mortier F, Cosson JF (2005). A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–1280.

Haldane J (1948). The theory of a cline. *J Genet* **48**: 277–284.

Hanski I, Simberloff D (1997). Metapopulation biology: ecology, genetics and evolution. In: Hanski I, Gilpin M (eds). *The Metapopulation Approach, Its History, Conceptual Domain, and Application to Conservation*. Academic Press. pp 5–26.

Hardy O, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.

Ihaka R, Gentleman R (1996). R: A language for data analysis and graphics. *J Comput Graph Stat* **5**: 299–314.

Jombart T (2008). *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* (doi:10.1093/bioinformatics/btn129; e-pub ahead of print, 8 April 2008).

Kaeuffer R, Réale D, Coltman DW, Pontier D (2007). Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* **99**: 374–380.

Kerth G, Petit E (2005). Colonization and dispersal in a social species, the Bechstein's bat (*Myotis bechtsteinii*). *Mol Ecol* **14**: 3943–3950.

Legendre P, Legendre L (1998). *Numerical ecology, Developments in Environmental Modelling*. Elsevier Science B.V.: Amsterdam.

Manel S, Bellemain E, Swenson J, François O (2004). Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. *Mol Ecol* **13**: 1327–1331.

Manel S, Berthoud F, Bellemain E, Gaudel M, Luikart G, Swenson JE *et al*. (2007). A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Mol Ecol* **16**: 2031–2043.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol Evol* **18**: 189–197.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Moran P (1948). The interpretation of statistical maps. *J R Stat Soc Ser B* **10**: 243–251.

Moran P (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**: 17–23.

Patterson N, Price A, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: 2074–2093.

Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Philos Mage* **2**: 559–572.

Pramual P, Kuvangkadilok C, Baimai V, Walton C (2005). Phylogeography of the black fly *Simulium tani* (Diptera: Simuliidae) from Thaïland as inferred from mtDNA sequences. *Mol Ecol* **14**: 3989–4001.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Pritchard J, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Slatkin M (1985). Gene flow in natural populations. *Annu Rev Ecol Syst* **16**: 393–430.

Slatkin M, Arter H (1991a). Spatial autocorrelation methods in population genetics. *Am Nat* **138**: 499–517.

Slatkin M, Arter H (1991b). Spatial autocorrelation methods in population genetics. *Am Nat* **138**: 522–523.

Smouse P, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561–573.

Sokal R, Jacquez G, Wooten M (1989). Spatial autocorrelation analysis of migration and selection. *Genetics* **121**: 845–855.

Sokal R, Smouse P, Neel J (1986). The genetic structure of a tribal population, the Yanomama Indians. XV. Patterns inferred by autocorrelation analysis. *Genetics* **114**: 259–287.

Sokal R, Wartenberg D (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**: 219–237.

Swenson J, Sandegren F, Soderberg F (1998). Geographic expansion of an increasing brown bear population: evidence for presaturation dispersal. *J Anim Ecol* **67**: 819–826.

Thioulouse J, Chessel D, Champely S (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environ Ecol Stat* **2**: 1–14.

Tolley K, Burger M, Turner A, Matthee C (2006). Biogeographic patterns and phylogeography of dwarf chameleons (*Bradypodion*) in an African biodiversity hotspot. *Mol Ecol* **15**: 781–793.

Upton G, Fingleton B (1985). *Spatial Data Analysis by Sample. Vol. 1: Point Pattern and Quantitative Data*, Volume 1 of Spatial Data Analysis by Example. Wiley: New York.

Wagner H, Fortin M-J (2005). Spatial analysis of landscapes: concepts and statistics. *Ecology* **86**: 1975–1987.

Waits L, Taberlet P, Swenson J, Sandegren F, Franzen R (2000). Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Mol Ecol* **9**: 421–431.

Wartenberg D (1985a). Multivariate spatial correlations: a method for exploratory geographical analysis. *Geogr Anal* **17**: 263–283.

Wartenberg D (1985b). Spatial autocorrelation as a criterion for retaining factors in ordinations of geographic data. *Math Geol* **17**: 665–682.

Wright S (1943). Isolation by distance. *Genetics* **28**: 114–138.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

Appendix A : rational of the spatial Principal Component Analysis

In this appendix, the following notations are used :

$\mathbf{X}$ is the n-by-p table of centred allelic frequencies, where rows are observations (on individuals or populations) and columns are alleles.

$\mathbf{L}$ is a row-standardized connection matrix associated to the connection network among genotypes or populations.

$\mathbf{v}$ refers to any scaled vector of p loadings, so that $\|\mathbf{v}\|^2 = 1$.

$\boldsymbol{\alpha}$ refers to any vector of n row scores obtained by linear combinaison of the columns of $\mathbf{X}$ so that : $\boldsymbol{\alpha} = \mathbf{X}\mathbf{v}$.

$C(\mathbf{v})$ is a quantity serving as a score criterion in the analysis, defined as : $C(\mathbf{v}) = var(\mathbf{X}\mathbf{v})I(\mathbf{X}\mathbf{v}) = \frac{1}{n}(\mathbf{X}\mathbf{v})^T\mathbf{L}\mathbf{X}\mathbf{v} = \frac{1}{n}\mathbf{v}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{v}$.

$\mathbf{w}$ refers to any scaled vector of p loadings provided by the sPCA, thus verifying $\|\mathbf{w}\|^2 = 1$.

$\boldsymbol{\psi}$ refers to any vector of n row scores obtained by linear combinaison of the columns of $\mathbf{X}$ so that : $\boldsymbol{\psi} = \mathbf{X}\mathbf{w}$.

The purpose of the spatial Principal Component Analysis is to find the extrema of :

$$C(\mathbf{v}) = \frac{1}{n}\mathbf{v}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{v} = \frac{1}{n}\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha}$$

which we rewrite, posing $\mathbf{A} = \frac{1}{n}\mathbf{X}^T\mathbf{L}\mathbf{X}$:

$$C(\mathbf{v}) = \mathbf{v}^T\mathbf{A}\mathbf{v}$$

The solution to this problem is well-known when $\mathbf{A}$ is symmetric. This is, however, not the case because $\mathbf{L}$ is not symmetric itself. To solve this problem, we seek a symmetric matrix $\mathbf{B}$ so that :

$$C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$$

The expression $\mathbf{v}^T\mathbf{A}\mathbf{v}$ is a scalar, so $\mathbf{v}^T\mathbf{A}\mathbf{v} = (\mathbf{v}^T\mathbf{A}\mathbf{v})^T = \mathbf{v}^T\mathbf{A}^T\mathbf{v}$. Thus we have :

$$\begin{aligned}
\mathbf{v}^T\mathbf{A}\mathbf{v} &= \frac{1}{2}(\mathbf{v}^T\mathbf{A}\mathbf{v} + \mathbf{v}^T\mathbf{A}^T\mathbf{v}) \\
&= \frac{1}{2}(\mathbf{v}^T(\mathbf{A} + \mathbf{A}^T)\mathbf{v}) \\
&= \mathbf{v}^T(\frac{1}{2}(\mathbf{A} + \mathbf{A}^T))\mathbf{v}
\end{aligned}$$

where $(\mathbf{A} + \mathbf{A}^T)$ is symmetric because $(\mathbf{A} + \mathbf{A}^T)^T = \mathbf{A}^T + \mathbf{A}$. As a consequence, $C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$ with :

$$\mathbf{B} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) = \frac{1}{2n}\mathbf{X}^T(\mathbf{L} + \mathbf{L}^T)\mathbf{X}$$

Hence, we can find the extrema of $C(\mathbf{v}) = \mathbf{v}^T\mathbf{B}\mathbf{v}$ using the existing solution (Harville, 1997, p533-534). It is shown that if $\mathbf{w}_1$ and $\mathbf{w}_r$ are the eigenvectors of $\mathbf{B}$ associated to λ_1 and λ_r, respectively the highest and lowest eigenvalues of $\mathbf{B}$, then :

$$\lambda_r = \mathbf{w}_r^T\mathbf{B}\mathbf{w}_r \leq \mathbf{v}^T\mathbf{A}\mathbf{v} \leq \mathbf{w}_1^T\mathbf{B}\mathbf{w}_1 = \lambda_1$$

and so :

$$\lambda_r = var(\boldsymbol{\psi}_r)I(\boldsymbol{\psi}_r) \leq C(\mathbf{v}) \leq var(\boldsymbol{\psi}_1)I(\boldsymbol{\psi}_1) = \lambda_1$$

Appendix B : diagram of the multivariate tests of global and local structuring and associated computations

Computations of the test statistic

The testing procedure (Figure 1) is the same for both multivariate tests (global or local structuring). The matrix of allelic frequencies $\mathbf{X}$ (n individuals or genotypes ; p alleles) is first centred and scaled. The obtained matrix is denoted $\mathbf{Y}$. The matrix $\mathbf{E}$ is obtained like in Griffith (1996) and Dray et al. (2006) by the eigen analysis of the connection matrix associated to the connection network between genotypes (or populations). Its columns $\mathbf{e}_j$ ($j = 1, q$) are the centred and scaled Moran's eigenvector maps (MEMs), which model either global or local structures (Dray et al. 2006). For the purpose of our tests, $\mathbf{E}$ contains 'global MEMs' ($\mathbf{E} = \mathbf{E}+$) for the test of global structuring and 'local MEMs' ($\mathbf{E} = \mathbf{E}-$) for the local test.

The coefficients of determinations (R^2) are computed after linear regression of each allele on each MEM, giving a matrix S containing p x q values of R^2 computed as :

$$\mathbf{S} = \frac{(\mathbf{Y}^T\mathbf{E}) \bullet (\mathbf{Y}^T\mathbf{E})}{n^2}$$

where $\mathbf{Y}^T$ is the transposed matrix of $\mathbf{Y}$ and where '$\bullet$' denotes the Hadamard product.

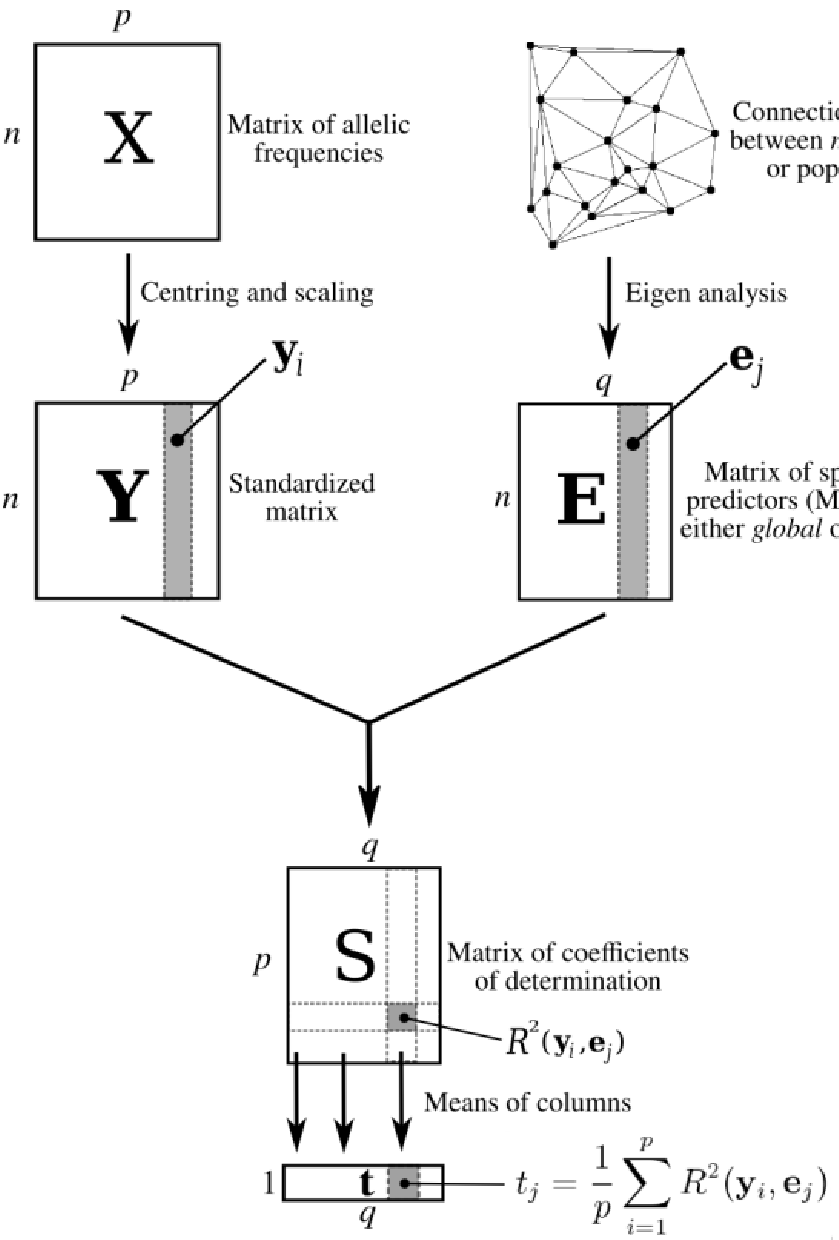
p
n
X
Matrix of allelic frequencies
Connection network between n genotypes or populations
Centring and scaling
Eigen analysis
$\mathbf{y}_i$
p
n
Y
Standardized matrix
$\mathbf{e}_j$
q
n
E
Matrix of spatial predictors (MEMs): either global or local
q
p
S
Matrix of coefficients of determination
$R^2(\mathbf{y}_i, \mathbf{e}_j)$
Means of columns
1
t
q
$t_j = \frac{1}{p} \sum_{i=1}^{p} R^2(\mathbf{y}_i, \mathbf{e}_j)$

FIG. 1 – Diagram of the testing procedure

As MEMs are orthogonal vectors, the coefficients of determination of alleles obtained from regression onto $\mathbf{E}+$ are independent from of those obtained with $\mathbf{E}-$. Practically, this implies that the test statistics in global and local tests are independent, in the sense that the value of the first is not conditionned by the value of the second (and reciprocally). Note that this is also true for the reference distributions of both statistics.

The squared correlations are then averaged for each MEM, giving a value t_j for the j^{th} allele defined by :

$$t_j = \frac{1}{p}\sum_{i=1}^{p} R^2(\mathbf{y}_i, \mathbf{e}_j)$$

The value of t_j is 'large' (respectively 'small') when alleles are in average 'strongly' (respectively 'weakly') correlated to the j^{th} MEM. The $(n-1)$ values of t_j are stored in the vector $\mathbf{t}$, directly computed as :

$$\mathbf{t} = \frac{1}{p}\mathbf{1}_p^T\mathbf{S}$$

where $\mathbf{1}_p$ is the p-dimensional vector whose components are all 1. The test statistic is defined as the maximum of all components of $\mathbf{t}$, max$(\mathbf{t})$.

Computation of the reference distribution

The 'reference distribution' is defined as the distribution of the test statistic (max$(\mathbf{t})$) under the null hypothesis H_0 that allelic frequencies of the genotypes (or populations) are distributed at random on the connection network. The alternative hypothesis H_1 depends on the type of test :

- global test : allelic frequencies of the genotypes (or populations) display at least one global structure
- local test : allelic frequencies of the genotypes (or populations) display at least one local structure

In both tests, the reference distribution is approximated by a Monte Carlo procedure involving a large number of permutations (at least 999) of the rows of $\mathbf{Y}$ (genotypes or populations), the test statistic being computed from each permutation. The p-value is computed as the relative frequency of permuted statistics equal to or higher than the initial value of max$(\mathbf{t})$.

Assessment of type I error

The actual type I error of both tests was assessed using simulated datasets of allelic frequencies with random spatial coordinates. We used datasets with different number of observations (25, 50, 100, 200) and different number of alleles (50, 100, 150). For each size of dataset, 200 simulations were performed and the results were pooled across simulations for each test, yielding a total of 2400 simulations per test. The actual type I error was measured as the relative frequency of reject of H_0 given different nominal α levels (Table 1). The estimated type I error was always very close to the chosen α level.

Nominal α level	Observed type I error (global test)	Observed type I error (local test)
0.1	0.0925	0.1042
0.05	0.0425	0.0512
0.01	0.0075	0.0062

TAB. 1 – Estimations of actual type I errors of the global and local tests assessed through simulations, for three nominal α levels. 2400 simulations of datasets with different size (see text) were performed for each test.

References

Dray S, Legendre P, Peres-Neto P (2006) Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). *Ecological Modelling* **196** : 483-493.

Griffith D (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer* **40** : 351-367.

Harville D (1997) *Matrix algebra from a statistician's perspective.* Springer, New York.