MOLECULAR ECOLOGY

Molecular Ecology (2009) **18**, 4734–4756 doi: 10.1111/j.1365-294X.2009.04410.x

INVITED REVIEW

Statistical methods in spatial genetics

GILLES GUILLOT,* RAPHAËL LEBLOIS,† AURÉLIE COULON‡ and ALAIN C. FRANTZ§

*Department of Informatics and Mathematical Modelling, Technical University of Denmark, Copenhagen, Denmark, †Département Systématique et Évolution, Muséum National d'Histoire Naturelle, UMR7205 MNHN/CNRS, Paris, France, ‡Département Écologie et Gestion de la Biodiversité, Muséum National d'Histoire Naturelle, UMR 7179 MNHN/CNRS, Brunoy, France, §Department of Animal & Plant Sciences, University of Sheffield, Sheffield, UK

Abstract

The joint analysis of spatial and genetic data is rapidly becoming the norm in population genetics. More and more studies explicitly describe and quantify the spatial organization of genetic variation and try to relate it to underlying ecological processes. As it has become increasingly difficult to keep abreast with the latest methodological developments, we review the statistical toolbox available to analyse population genetic data in a spatially explicit framework. We mostly focus on statistical concepts but also discuss practical aspects of the analytical methods, highlighting not only the potential of various approaches but also methodological pitfalls.

Keywords: conservation genetics, habitat degradation, population ecology, population genetics – empirical, population genetics – theoretical, wildlife management

Received 17 June 2009; revision received 29 September 2009; accepted 1 October 2009

Introduction

As many ecological and evolutionary processes that influence genetic variation are mediated by space, the joint analysis of genetic and spatial information can lead to a better understanding of these processes. While this was acknowledged as early as the 1940s in the theoretical work of Dobzhansky & Wright (1941), Wright (1943) and Malécot (1948), the traditional population genetic studies were limited in spatial inference to tests of the effect of geographic distance. More recent methods allow testing for the influence of environmental features on gene flow and genetic structure. There is, however, no spatial genetic theory as yet, but rather a set of concepts and broad variety of methods that are combined in a rather *ad hoc* manner in different studies. Furthermore, spatial genetics has become such a rapidly evolving field that it is becoming increasingly difficult to keep abreast with the latest statistical developments.

In the present work, our aim was to review the statistical toolbox available to molecular ecologists aiming to detect, quantify and test the spatial structure of genetic variation. We mostly focus on statistical concepts but also discuss practical aspects of the analytical methods, highlighting both the potential and the pitfalls associated with analysing population genetic data in a spatially explicit framework. We introduce general methods for exploratory data analysis and present and discuss models of isolation by distance (IBD) as well as spatial clustering models. Throughout, we hope to show that, instead of trying to get rid of a spatial pattern, it is often fruitful to *model* it. By doing so, one can hope to make more accurate inferences by appropriately injecting *a priori* information and obtaining directly interpretable parameters. While we will try to emphasize that no approach can give definitive answers, we finish by discussing how information about genetic patterns can be used to gain a deeper insight into the underlying ecological and evolutionary mechanisms.

The subject of this review partially overlaps with landscape genetics, the branch of population genetics concerned with the effect of landscape features on genetic structure. Readers interested in this subject may refer to earlier reviews by Manel *et al.* (2003), Holdregger & Wagner (2006), Storfer *et al.* (2007) and Møller-Hansen & Hemmer-Hansen (2007). As spatial genetic studies have mostly been concerned with the study of neutral genetic variations, we chose to focus our review on the methodologies suitable for this type of analysis

Correspondence: Gilles Guillot;
E-mail: gigu@imm.dtu.dk

and do not address the problems linked to the spatial structure of genetic markers under selection. A list of relevant computer programs can be found in the Supporting Information (Table S1) and we refer readers further interested in this aspect to a recent review by Excoffier & Heckel (2006).

Exploratory data analysis

We start by focusing on methods that are essentially descriptive as they do not make explicit assumptions about past and on-going biological processes or even about spatial patterns that may influence genetic variation. As will be shown later, more formal assumptions on evolutionary or ecological processes should, in principle, increase the power of inference and ease of interpretation of the identified patterns.

Traditionally, the statistical units of analysis in a population genetic study have been groups of individuals or populations (Waples & Gaggiotti 2006). The exploratory methods were mostly designed with predefined populations in mind. However, for many species, population density can be continuous, with allele frequencies displaying smooth variation across space. In this context, predefined populations may make little sense. The more formal statistical approaches can either be applied to both types of data or they were specifically designed to focus on individual variation, which, again, can make their use more appropriate in spatial genetics.

Global statistics

Investigating dependence between geographic and genetic distances. There is a broad variety of methods to detect, quantify and test the spatial structure of genetic variation. Before implementing any of these, it can be useful to carry out a preliminary analysis to test for a statistical dependence between geographic and genetic distances (pairs consisting of individuals or *a priori* defined groups). This is usually carried out using a Mantel test, a permutational procedure used to test the statistical significance of matrix correlations (Sokal & Rohlf 1995, pp. 813–819). It returns a *P*-value for the empirical correlation coefficient between the geographic and the genetic distance matrices, with a significant correlation being indicative of spatial structure.

There can be uncertainty about the underlying causes of this statistical dependence. A significant *P*-value could be due, for example, to IBD (see section below) or to the presence of barriers to gene flow between populations that are otherwise globally panmictic. These two circumstances are of a very different nature. While the former reflects the intrinsic dispersal ability of the species, the latter results from landscape features reducing gene flow. Different combinations of these two effects with different spatial patterns can lead to similar *P*-values. Plotting pairwise geographic vs. genetic distances may give clues about the relative importance of these two factors (see Fig. 1 for a synthetic example and Box 2 for an example of combined effects of IBD and landscape features).

From Mantel test to empirical spatial autocorrelation analysis. The variation of genetic distance as a function of geographic distance can be analysed by the empirical spatial autocorrelation function, an exploratory tool widely used in geostatistics and environmental statistics. This function is defined for a sample $(x_1, \ldots, x_n)$ of a single quantitative variable x as

$$c(h) = \frac{1/n_h \sum_{C_h}(x_i - \bar{x})(x_j - \bar{x})}{1/n \sum_{i}(x_i - \bar{x})^2} \qquad (eqn\,1)$$

where C_h denotes the set of pairs of individuals separated by a distance of approximately h, n_h denotes the number of such pairs and n denotes the number of individuals. In practice, as it is more robust to sampling variance, the empirical variogram function defined as

$$\gamma(h) = \frac{1}{2n_h} \sum_{C_h}(x_i - x_j)^2 \qquad (eqn\,2)$$

is often preferred. To deal with genotypic data, which are most often treated as categorical variables, one can either work with allele frequencies obtained by averaging over individuals taken at (or around) a sampling site or work with an indicator variable for each allele (taking the value 0, 1 or 2 depending on the number of copies of this allele carried by the individual). The latter leads to Moran's *I* statistic, a widely used descriptor of spatial genetic structure. Its use for inference of dispersal characteristics will be discussed in the sequel (see also Hardy & Vekemans 1999; Rousset & Leblois 2007).

The functions defined above are widely used in environmental statistics because they bring insights into the spatial scale of variation of the process, in particular the characteristic distance at which statistical dependence disappears (this distance being known as the range in geostatistics) and also its direction of maximal rate of decrease. This method produces an out-of-sample prediction of the variable at hand (e.g. allele or haplotype frequency) and therefore a map of the variable on the whole study domain from a limited sample. Attempts to relate geostatistics to classical population genetics models (in particular mutational and dispersal models) can be found in Wagner *et al.* (2005) and Hardy &

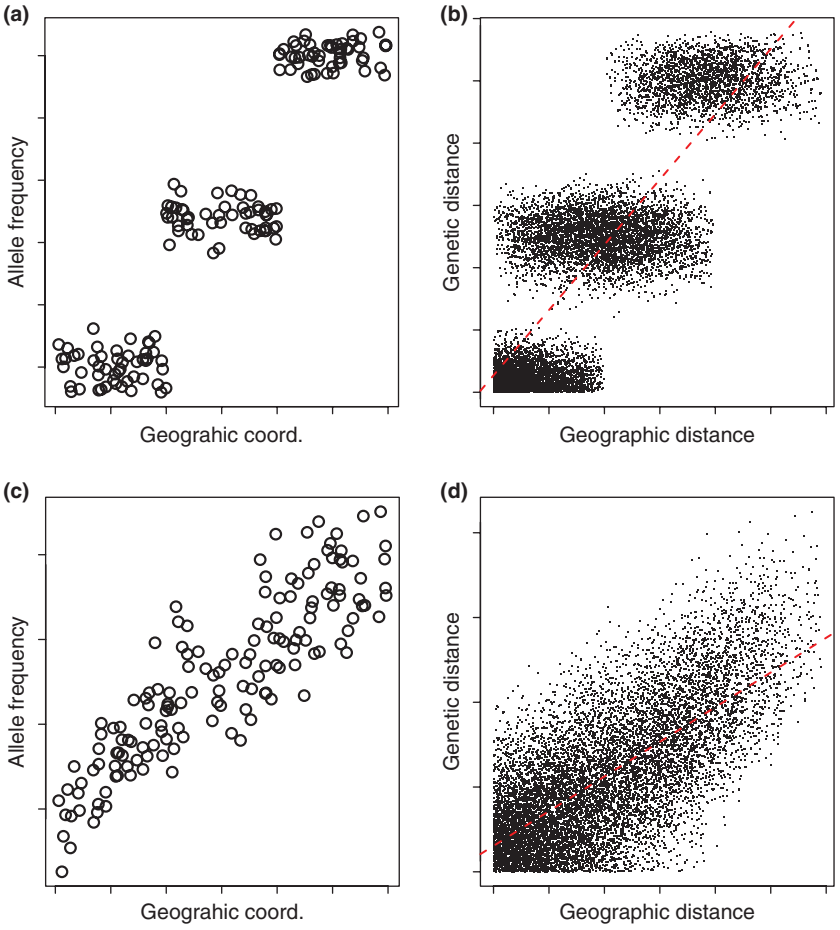


Fig. 1 Artificial examples of patterns of allele frequencies at a single locus across space in a one-dimensional habitat (a,c) and plots of the relationship between geographic and genetic distance as tested by a Mantel test (b,d). Panels (a,b): three panmictic populations separated by barriers that allow restricted gene flow between adjacent populations; at the scale considered, the intrinsic dispersal process is not affected by distances but mostly by the presence of barriers and the mutual locations of patches; panels (c,d) correspond to a genuine continuous population under isolation by distance. In both cases, a Mantel test would reject the hypothesis of global panmixia, although the ecological processes are of different natures.

Vekemans (1999) and examples of application can be found, e.g. in Monestiez *et al.* (1994), Le Corre *et al.* (1998) and Thompson *et al.* (2005).

It must be emphasized, however, that the interpretation of the parameters inferred from the variogram, and especially of its range in terms of dispersal distances, must be done with caution, in particular because variogram features might depend on the mutation process and can also be affected by the sampling scheme (see also conclusion on this point).

Statistical methods to quantify spatial genetic variations locally

The correlation coefficient between geographic and genetic distances, or the parameters of the variogram, are global statistics in the sense that their computation involves all the data points and that they reflect global properties of the sample over the whole study area. They make sense when the study area is homogeneous (e.g. in terms of landscape features and gene flow patterns). It is possible to get further insights into the details of genetic variation by computing statistics locally. This is the aim of methods collectively referred to as barrier detection methods.

While clustering methods (described in section below) look for homogeneous spatial domains, barrier detection methods try to identify areas of abrupt genetic discontinuities. As such, they complement clustering methods, as they can identify features that disrupt gene flow locally, along U-shaped patterns, for example, without creating distinct genetic units (see Fig. 2 for a synthetic illustration and Irwin *et al.* 2001; Joseph *et al.* 2008 for the related issue of ring species).

Wombling methods. In its initial formulation (Womble 1951; Barbujani *et al.* 1989), the Wombling method produced a map of the norm of the gradient of allele frequencies (an index quantifying the local variability of allele frequencies), highlighting areas of abrupt changes in allele frequencies. Its main drawback is that it does not provide a frame of reference to assess the relative importance of the observed break. The recent extensions that attempt to reformulate the method in a more rigorous statistical framework (Bocquet-Appel & Bacro 1994;

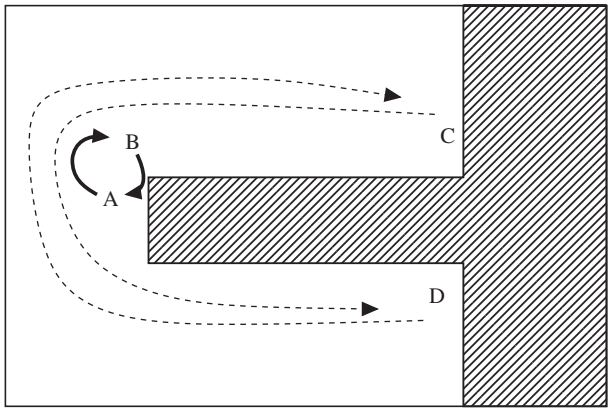


Fig. 2 Schematic example of a physical obstacle (dashed area) that limits gene flow (e.g. mountain, urban area) but does not split the study area into nonconnected pieces. Letters A–D represent individuals. Gene flow is not possible through the obstacle but along the obstacle. Despite the presence of an obstacle, there is no reason to expect clusters. However, in case of weak isolation by distance, one might expect some differentiation between individuals located in distant areas on opposite sides of the barrier (C and D, as opposed to A and B, which are close to each other and not separated by the barrier) and therefore some kind of spatial genetic discontinuities.

Cercueil *et al.* 2007; Crida & Manel 2007; Manel *et al.* 2007) still rely strongly on user-specified parameters. The accuracy of these extensions also needs to be assessed with simulated data. More formal statistical methods implemented in a nongenetic context could be easily extended to genetic data (Banerjee & Gelfand 2006; Liang *et al.* 2008).

Monmonier's algorithm. Monmonier's (1973) algorithm is a related approach that tries to identify pairs of neighbouring predefined population units that display relatively large genetic differentiation. It employs an *ad hoc* strategy that does not follow a clear statistical or biological rationale. We stress that, although the method has been widely used, the accuracy and the influence of the choice of some important parameters have rarely been assessed in a controlled setting, i.e. with simulated data. One exception is a study by Dupanloup *et al.* (2002), which clearly showed that populations need to be fairly strongly differentiated for the algorithm to correctly infer the location of genetic discontinuities (although accuracy improved when the number of loci was increased). Furthermore, it has been suggested that the output of the method depends on the geographical sampling design (Rollins *et al.* 2006). Importantly, it requires *a priori* definition of the number of genetic groups believed to be present in a data set; a piece of information that frequently is not available.

Ordination methods

Ordination methods are exploratory methods aimed at finding proximities between high-dimensional objects by summarizing information in low-dimensional space. The most popular of those methods is the principal components analysis (PCA). The method aims to summarize information contained in p possibly correlated variables by creating p synthetic uncorrelated variables that can be ordered by decreasing information content (see Jombart *et al.* (2009) for a recent review). It is usually expected that most of the information can be captured by a small number of these new variables. PCA can be applied to allele frequencies computed from *a priori* defined populations or to individual genotypes. Patterson *et al.* (2006) showed recently that PCA can be successfully used to detect population structure in particular large data sets consisting of thousands of SNPs where implementing a more complex method (e.g. Bayesian clustering discussed in Clustering methods section below) becomes impractical. Hannelius *et al.* (2008), Lao *et al.* (2008) and Novembre *et al.* (2008) also showed that the inferred genetic structure might reflect geographic features (pairwise geographic distances of the samples). The efficiency of these methods is dependent, however, on the ability of users to interpret the synthetic variables and a small number of those synthetic variables might fail to capture enough information. Note also that Reich *et al.* (2008) and Novembre & Stephens (2008) discussed how certain spatial patterns of PCA maps can result from IBD gradients and pointed out the need for great care in interpreting the inferred patterns in terms of past migration/colonization processes.

Isolation by distance

Overview

Natural populations are often nonrandom mating units because reproduction occurs preferentially between geographically close individuals and because inter-generational individual dispersal distances are usually small compared with the area delimiting the population: a phenomenon that leads to IBD. IBD models have often been used in empirical studies to quantify dispersal from genetic data, in particular as an alternative to demographic data that can be difficult and time consuming to collect. The aim of these analyses was to investigate aspects of reproductive, demographic and migratory functioning of populations. Questions of particular interest include the study of local adaptation (Petit *et al.* 2001; Prugnolle *et al.* 2005; Loiseau *et al.* 2009), the quantification of dispersal ability for the

design of conservation area or for the management of pest species (Olsen *et al.* 2003; Gonzales-Suarez *et al.* 2009). It has also recently been highlighted that IBD can be a confounding factor in population genetic analyses based on some panmictic populations, as illustrated by Leblois *et al.* (2006) for the detection of population size variations and discussed in the Clustering methods section for the detection of genetic discontinuities. Although methods to detect and quantify the effect of IBD are ubiquitous in molecular ecology studies, the precise assumptions they rely on are often not well understood. This relates presumably to the fact that the literature on the subject spans a long period of time and can be of a quite theoretic nature. For this reason, before discussing practical methods, we try to give a short but hopefully self-contained overview of the theoretic aspect of IBD models.

Population genetics models

Historical perspective: Wright's original isolation by distance model and Kimura's stepping stone model. Wright (1943) considered initially a model in which individuals are distributed randomly and uniformly over space and where mating occurs between neighbouring individuals (separated by a small distance). Later, Wright (1946) extended his model to allow dispersal according to a Gaussian distribution. In a fashion analogous to Wright–Fisher island models where differentiation is governed by the product of population sizes and migration rate, the central parameter of interest in this model is the neighbourhood size defined (up to a factor 2π) as $D\sigma^2$, where D is the density of genes (or haploid individuals) and σ^2 the mean-squared dispersal distance (the noncentred second-order moment of the dispersion function) of the Gaussian distribution. Wright showed that, under such models, individuals living nearby tend to be genetically more similar than those living further apart and that the increase in genetic differentiation with geographic distance strongly depends on the neighbourhood size.

Besides these continuous models, Kimura (1953) proposed a model known as 'stepping stone', where populations or demes (and not individuals) are located at the nodes of a grid and dispersal occurs mainly between adjacent subpopulations at rate m but where long-range dispersal between distant populations also occasionally occurs independently of the distance at rate m_∞. Kimura & Weiss (1964) showed that under localized dispersal (i.e. $m_\infty \ll m \ll 1$), genetic differentiation increases with geographic distance and that this increase depends on the product of deme size N and the local dispersal rate m. As long-range dispersal events can be viewed (for the purpose of reasoning) as mutations (introducing an unrelated allele in the host population), the neighbourhood size $D\sigma^2$ in a stepping-stone model is equal to Nm.

More general dispersal models. Data on dispersal distributions in natural populations suggest that dispersal is of limited spatial magnitude mostly with occurrence of rare long-distance dispersal events (leptokurtic distributions) (Bateman 1950; Portnoy & Wilson 1993; Clark *et al.* 1999; see also Endler 1977; Rousset 2004 for reviews). However, the particular shapes of dispersal distributions are expected to be highly diverse. General methods that would not rely on specific dispersal distributions, such as the normal distribution or a 'stepping stone' dispersion, are therefore expected to be more robust than others when applied to real data sets. Most of the recent IBD analyses are based on the so-called infinite lattice model. This model considers populations or individuals distributed on a lattice with spatially homogeneous demographic parameters, i.e. homogeneous population sizes or density and dispersal and was first formulated by Malécot (1950). This model is compatible with any arbitrary dispersal distribution with finite first-, second- and third-order moments. The Wright–Fisher island model and the stepping stone model are particular cases of this model that consider dispersal to be uniform or restricted to adjacent populations respectively.

Malécot (1975), Nagylaki (1976) and Rousset (1997) computed probabilities of gene identity as functions of demographic and mutational parameters of the lattice model for general forms of dispersal distributions. Those aspects are reviewed by Nagylaki (1989) and Rousset (2004). In these analyses, the effective density on the lattice and the second moment of the axial dispersal distance distribution (or the mean-squared parent–offspring dispersal distance) σ^2, are of particular importance. Note that σ^2 is not, as too often considered, the variance of the dispersal distribution, but a more useful interpretation is that σ^2 is a measure of the speed at which two lineages derived from a common ancestor move away from each other generation by generation (Rousset 2004). $D\sigma^2$ can thus be viewed as a simple measure of spatial genetic structure. Under an island model and a stepping stone model, σ^2 equals $+\infty$ and m respectively. Let us define a_r as

$$a_r = (Q_0 - Qr)/(1 - Q_0), \qquad (eqn\,3)$$

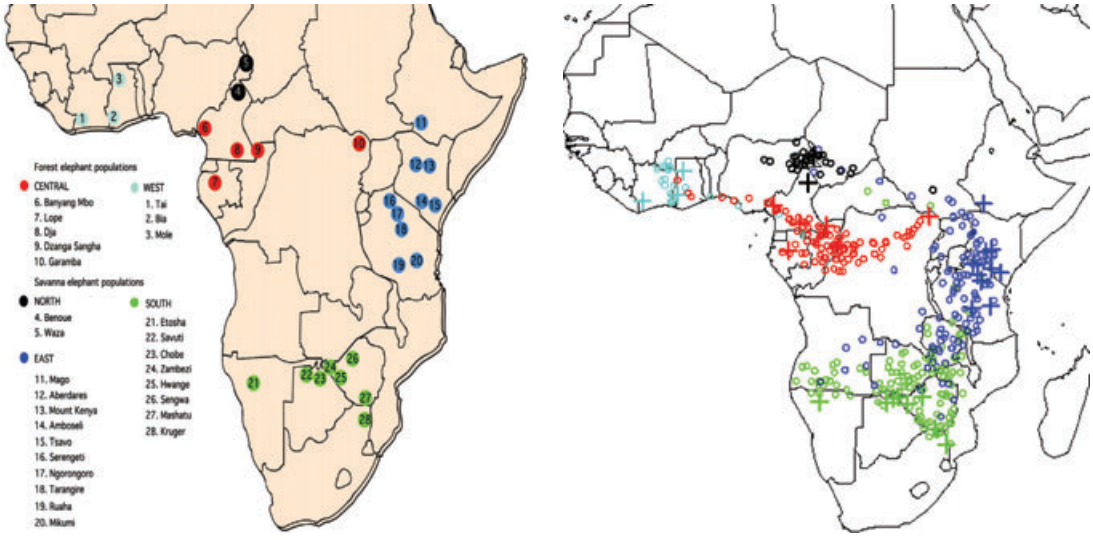
where Q_r is the probability of identity by descent between two genes separated by a geographic distance r ($r \geq 0$). The main result of the lattice model analysis is the linear relationship between a_r and the geographic distance in one dimension or its logarithm in two dimension:

Box 1: Using isolation-by-distance patterns to perform spatially continuous assignment

Random genetic drift under IBD tends to produce smooth spatial variations of allele frequencies. Inferred maps of allele frequencies can be used to perform geographically explicit individual assignments. Wasser *et al.* (2004) and Wasser *et al.* (2007) developed a method that jointly estimates such maps and estimates the unknown geographic origin of a DNA sample by comparing its alleles with estimated allele frequencies. Rather than simply assigning individuals to predefined populations, the method can, in principle, assign individuals to any spatial location whose inferred allele frequencies best explains the genotype of the sample. Using this method, Wasser *et al.* (2007) showed that a large shipment of contraband ivory originated from a narrow region centred on Zambia. The accuracy of the assignment depends on the accuracy of the allele frequency map implicitly generated during the inference step, which in turn depends on the size of the training data set and on how much allele frequencies characterize a given region.

Pope *et al.* (2007) found that the individual spatial assignments generated by the method proposed by Wasser *et al.* (2004) could give ambiguous results (many possible locations). This might result from: (i) a lack of differentiation in the data; (ii) uncertainty about allele frequencies due in particular to the use of data with individuals continuously sampled over space; (iii) departure of data from the underlying statistical model; (iv) overparametrization compared with sample size; (v) MCMC convergence flaw. Pope *et al.* (2007) devised a simpler method based on the same rationale. They used their method to compare the movement of individual badgers before and after a culling operation performed in the context of bovine tuberculosis (*Mycobacterium bovis*) control. Even though they showed that the badgers moved, on average, further post- than pre-cull, it yet remains to be seen how accurate Pope *et al.*'s method is in the assignment of individuals to specific geographic localities.

In a study in human genetics, modelling allele frequencies as a linear function of spatial coordinates as the synoptic scale, Amos & Manica (2006) were capable of assigning individuals with an accuracy of 1200 miles. Novembre & Stephens (2008) proposed a method based on a PCA suitable for large SNPs data that predict spatial origin through a linear regression on the first two principal components.



(a) Map of Africa showing the collection sites divided into five regions: West Africa (cyan), Central forest (red), and Central (black), South (green) and East (blue) savanna. (b) Estimated locations of elephant tissue and faecal samples from across Africa when assignments are allowed to vary anywhere within the elephants' range. All tissue and scat samples ($n = 399$) were successfully amplified at seven or more loci. Sampling locations are indicated by a cross and are colour coded according to actual broad geographic region of origin: West Africa, Central forest, and Central, South and East savanna [colour coded as in (a)]. Assigned location of each individual sample is shown by a circle and is colour coded according to its actual region of origin. The closer each circle is to crosses of the same colour, the more accurate is that individual's assignment (figures and caption reprinted from Wasser *et al.* 2004).

$$a_r \approx \frac{r}{2D\sigma^2} + A_1 \quad \text{in one dimension} \qquad (eqn\,4)$$

and

$$a_r \approx \frac{\ln r}{2\pi D\sigma^2} + A_2 \quad \text{in two dimensions,} \qquad (eqn\,5)$$

where D is the effective density of genes over the sampled area, σ^2 the mean-squared parent–offspring distance, and $A1$ and $A2$ are constant terms that depend on the shape of the dispersal distribution, but not on population sizes or mutation rates (see Rousset 1997, 2004 for details on these terms). Note that these equations relate the slope of the regression between a_r and the geographic distance to the neighbourhood size in a straightforward manner.

Neglecting A_1 and A_2 leads to inaccurate approximations of *F*-statistics in terms of the model parameters. It also explains why it is often considered that the neighbourhood size $D\sigma^2$ solely determines the whole genetic patterns under IBD. It must be emphasized that the above linear relationship is based on approximations that are only valid for simultaneously large r and small $r\mu$, where μ is the mutation rate, an important constraint that is often forgotten (Rousset 1997). In practice, the linear relationship will be reasonably accurate for intermediate geographic distance with $\sigma \leq r \leq 0.2\sigma/\sqrt{2\mu}$ in one dimension and $\sigma \leq r \leq 0.56\sigma/\sqrt{2\mu}$ in two dimensions (Rousset 1997, 2004). At a shorter distance, the shape of the dispersal distribution will have a strong influence on the increase in differentiation with geographic distance and at larger scales, mutation rate will break the linear relationship (Leblois *et al.* 2003; Rousset 2004).

From discrete to continuous populations. As advocated in the work of Wright (1943) mentioned earlier, there are not necessarily discrete subpopulations or demes (i.e. geographically localized panmictic units) in natural populations and individuals can often be continuously distributed over space. The absence of demic structure can be achieved in a lattice model by assuming subpopulation sizes equal to one. This case can be viewed as an approximation for continuous populations when density regulation (e.g. by competition) is strong enough to keep a constant local density (Felsenstein 1975; Malécot 1975; Slatkin 1989; Rousset 2000). It would be even more realistic to assume that individuals could settle at any location in space. This would induce spatial and temporal density heterogeneities. Such models with continuous distribution of individuals have been formulated (Wright 1943, 1946; Malécot 1967; Nagylaki 1974; Barton *et al.* 2002); however, they did not rely on a well-defined set of biological assumptions and led to incoherent results (Maruyama 1972; Felsenstein 1975). A recent study by Robledo-Arnuncio and Rousset (2009) solved some of those problems in a continuous model of IBD with demographic fluctuations and allowed robust definitions of effective dispersal and density parameters in terms of demographic parameters. Their results showed that the expected linear relationship between a_r and geographic distance holds for continuous IBD models with fluctuating density. They also showed that, similar to the case of a lattice model with one individual per node, the slope is given by the effective $D_e\sigma_e^2$ parameter.

Inference methods under isolation by distance using genetic distances

Inference of demographic parameters under isolation by distance. Slatkin (1993) developed the first method that took demographic parameters explicitly into account when analysing genetic data under IBD. The method is based on a plot of estimates of M defined as $(1/F_{ST} - 1)/2$ against geographical distance on a log–log scale. An estimate of the number of migrants per generation Nm is given by the intercept of this regression. Using M instead of other genetic distances has two main interests : (i) M is roughly independent of mutation rates and sample design; (ii) the value of M computed for a pair of samples distant of d is related to Nm through the simple formula $M \approx Nm/d$ under a stepping stone model (Slatkin 1991, 1993). These features allow easy comparison between different samples.

This method allows quantitative inferences of dispersal under stepping stone models but not under more general IBD models because the simple relation between M and the number of migrants does not hold any longer. The most general inference method is derived from the linear relationship between a_r and the geographic distance with a slope solely determined by the product $D\sigma^2$ at a local geographical scale. Rousset (1997, 2000) proposed to use the inverse of the regression slope between estimates of a_r computed from genetic data collected at a local scale and the geographic distance in one dimension, or its logarithm in two dimensions, to estimate $D\sigma^2$.

Connections between empirical autocorrelation functions and theoretical isolation-by-distance models. Connections between empirical autocorrelation methods and theoretical IBD models have been investigated by numerical simulations with the aim of quantifying IBD (e.g. Sokal & Wartenberg 1983; Epperson 1995; Epperson & Li 1997). Many empirical studies, especially on plant populations, have used this approach (reviewed by

Heywood 1991; Epperson 1993) but most of them only describe patterns in a qualitative way, making comparison among studies difficult. Epperson (1995, 2005, 2007) developed quantitative inference methods based on numerical simulations. However, as detailed by Rousset (2008), such methods are not fully adequate because: (i) they are based on numerical simulations considering that the neighbourhood size, $D\sigma^2$, is the only dispersal parameter that shapes spatial genetic structure, whereas the above IBD analyses show that differentiation under IBD is not only a function of $D\sigma^2$ but also depends on mutation processes as well as complex features of the dispersal distribution; and (ii) they use genetic statistics that are not independent of sampling scheme and mutation rates, introducing complications for comparison among studies.

There are many reasons, however, to consider other differentiation statistics (e.g. different from a_r) for inferences under IBD. In any case, to allow appropriate analyses and especially robust inference, the relationship between the statistic used and the demographic parameters of the model should be well defined; and the latter relationship should be to some extent robust to mutation processes and sampling design. Examples of different statistics design and tests can be found in Hardy & Vekemans (1999) and Hardy (2003) for plant populations and dominant markers, as well as in Watts *et al.* (2007) for an improvement of Rousset's method for populations with large $D\sigma^2$ values (i.e. weak IBD pattern). One can also choose to use biased statistics with small variance to test for IBD with more power and then use unbiased statistics to make demographic inferences. Detailed discussion on the different statistics to use and on the relationship between spatial autocorrelation analyses, population genetic models and the neighbourhood size can be found in Hardy & Vekemans (1999), Vekemans & Hardy (2004) and Rousset (2008).

Maximum likelihood inference of demographic parameters under isolation by distance

Likelihood methods are theoretically more powerful than methods based on summary statistics because they use the whole information present in the data. The migration matrix model, with one migration rate parameter for each population pair, as implemented in the MIGRATE program (Beerli & Felsenstein 2001), theoretically allows inference under IBD. However, as pointed out by Beerli (2006), it will be practically inaccurate as it appears difficult to make inference under a model with more than four subpopulations because of the high number of parameters in the migration matrix model. There has been one attempt to use MIGRATE with IBD data but analyses of both real and simulated data overestimated dispersal (i.e. σ^2 one order of magnitude higher than the expected value), whereas the regression method described above gave results close to expectations (Leblois 2004).

More recently, a likelihood-based method specifically developed to infer demographic parameters under a one-dimensional IBD model gave interesting insights on the behaviour of such likelihood-based methods (Rousset & Leblois 2007). As expected, difficulties arise from the inherent dependency of the likelihood on all parameters of the model, especially nuisance parameters and those for which information on genetic data is limited. Simulation tests showed that: (i) likelihood-based inferences of $D\sigma^2$ are slightly more precise than inferences with the regression method when all assumptions of the likelihood model are verified; but also (ii) other parameters (e.g. shape of the dispersal distribution, total number of subpopulations and mutation rates as well as population sizes and migration rates inferred separately) cannot be inferred with good precision from classical genetic samples (i.e. single time sampling of independent genotypic markers). Finally, maximum likelihood inferences from data sets simulated under a model different from the model of analyses always lead to less robust results than those obtained with the regression method (Rousset & Leblois 2007).

Testing for isolation by distance on real data sets

As already mentioned, the presence of an IBD pattern (i.e. positive correlation between genetic and geographic distances, corresponding to finite $D\sigma^2$) is usually inferred using Mantel tests (Mantel 1967). However, there is often low power to detect IBD with a Mantel test using typical sample size (e.g. hundred individuals sampled at the adequate geographical scale so as to avoid biases discussed by Leblois *et al.* (2003) because estimates of the differentiation statistics have high variance and $D\sigma^2$ often show large values in natural populations, both factors leading to weak correlation between genetic and geographic distances. Note also that IBD could theoretically be detected by testing for HW disequilibrium (evidence from simulated data are reported, e.g. by Frantz *et al.* (2009). However, the power of such tests has not been investigated yet and some empirical data sets suggest that Hardy–Weinberg equilibrium (HWE) will often not be rejected even in the presence of strong IBD (Sumner *et al.* 2001; Winters & Waser 2003; Broquet *et al.* 2006; Watts *et al.* 2007).

It would be useful to provide confidence intervals for a measure of IBD that can be related to dispersal and density parameters, such as $D\sigma^2$. Along this line, Leblois *et al.* (2003) used ABC bootstrap (DiCiccio & Efron 1996) to compute confidence intervals on the slope of

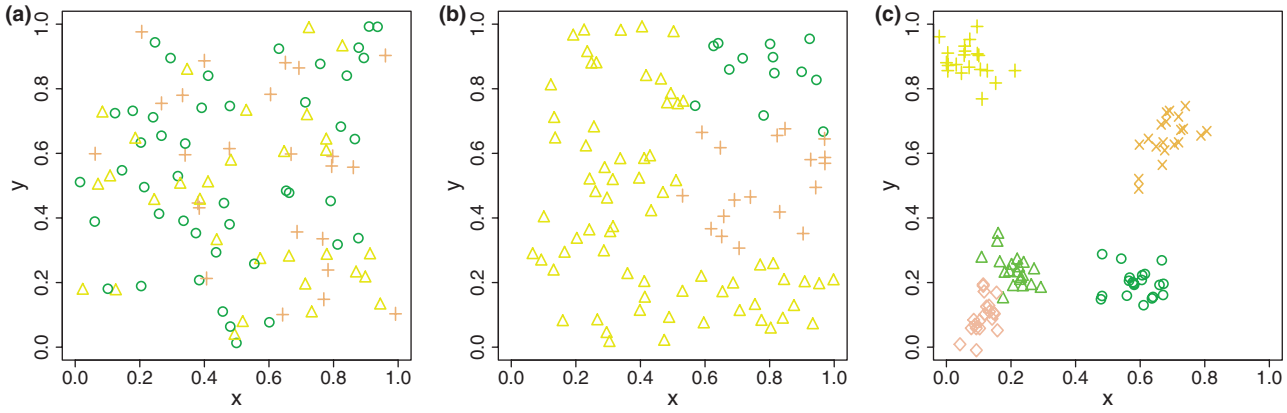


Fig. 3 Artificial examples of spatial patterns for some putative panmictic clusters. (a) Complete spatial randomness; (b) pattern typical of a continuously populated species with clusters separated by spatially simple shaped boundaries; (c) pattern typical of a species with large variations of population density across space (e.g. due to habitat fragmentation). The colour symbol indicates cluster memberships of individuals.

the linear relationship between a_r and geographic distance, but they show that the intervals computed with this procedure are always too narrow due to an overestimated lower bound. Finally, maximum likelihood should allow model test between IBD and Wright–Fisher models but this, to our knowledge, has never been done.

Summary and conclusion on isolation by distance

When working on data from natural populations, many factors are uncontrolled and inferences are based on highly simplified models. One major concern is thus the robustness of the analyses based on a given model when some assumptions do not hold because it determines what can be estimated. As an example, the regression method of Rousset (1997, 2000) has been thoroughly tested using simulated data sets with regard to mutational and sampling factors (Leblois *et al.* 2003) and with regard to demographic fluctuations in space and time (Leblois *et al.* 2004). All those tests showed that inference of $D\sigma^2$ using this method is robust to mutational processes and that, in numerous realistic conditions, the method estimates the local and actual demographic parameters with good precision (e.g. within a factor of two). Moreover, comparison between independent demographic and genetic estimates of $D\sigma^2$ on the same populations showed reasonable agreement (i.e. within a factor of two), on 10 different data sets (Rousset 1997, 2000; Sumner *et al.* 2001; Fenster *et al.* 2003; Winters & Waser 2003; Broquet *et al.* 2006; Watts *et al.* 2007). These results go against the belief in population genetics (Lewontin 1974; Slatkin 1987; Whitlock & McCauley 1999) that observed genetic structure is often not consistent with expectations from theoretical population genetic models and that inference from genetic data thus cannot give accurate estimates of demographic parameters in natural populations. Furthermore, all those results suggest that the lattice model predicts rather well the local increase in differentiation with distance for natural populations with limited dispersal and that the regression method is fairly robust to various demographic and mutational factors when adequately used.

For ecologists, one relevant caveat of the above analyses is that there is no method to infer dispersal parameters other than $D\sigma^2$ from genetic data under IBD using *F*-statistics or similar analyses such as autocorrelation analyses. Other dispersal parameters that could be of great interest are, for example, maximal dispersal distance, long dispersal rates or more generally a finer characterization of the dispersal distribution independently of the density parameter. The important question is then whether genetic data *per se* contain enough information to infer more detailed features of dispersal. Some recent likelihood analyses suggest that there is some information in typical genotypic samples but not enough to allow precise estimations of the shape of dispersal, total population size or to separate density from dispersal parameter estimates (Rousset & Leblois 2007).

Clustering methods

Nonspatial vs. spatial models

An important body of work has been concerned with variations of allele frequencies due to random drift induced by lack of gene flow. This problem has been investigated prominently by the use of Bayesian cluster-

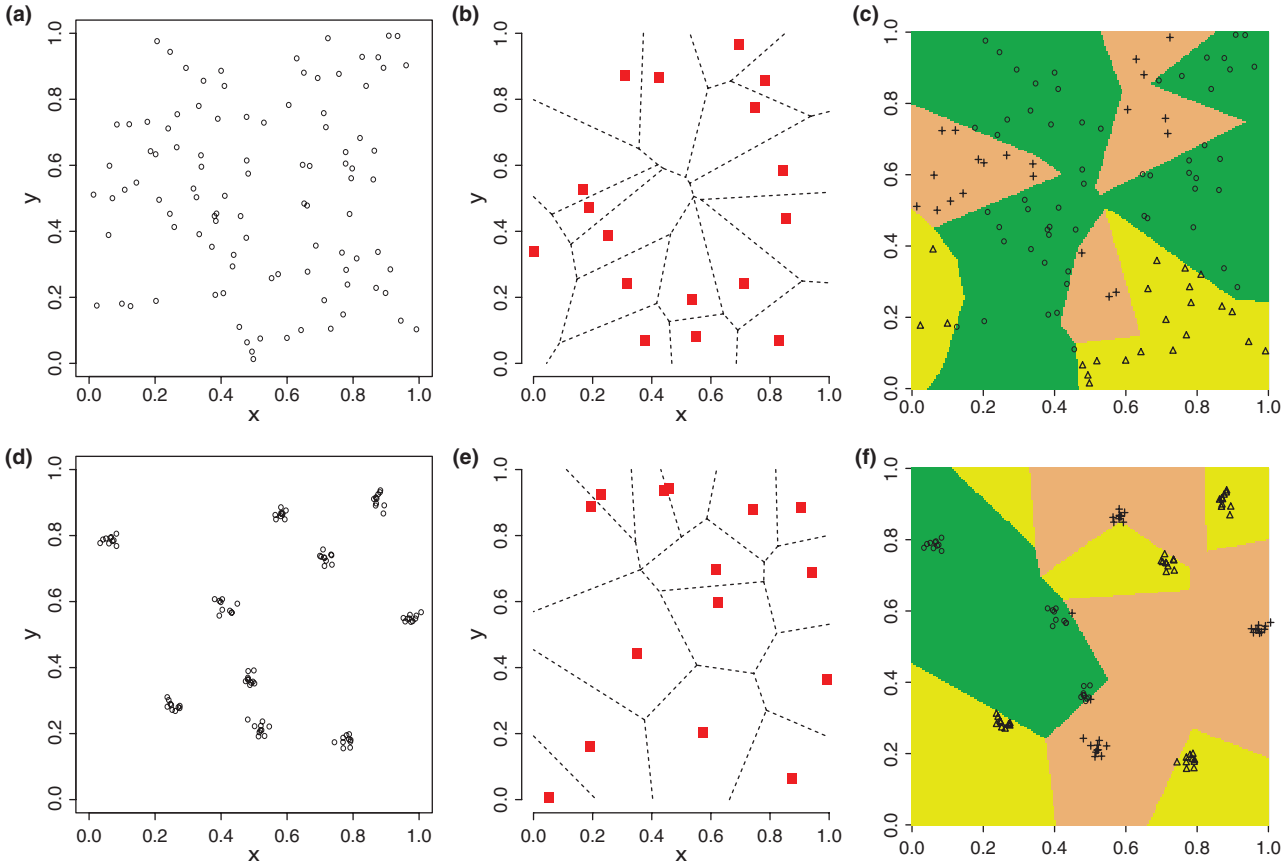


Fig. 4 Examples of simulated spatial pattern for cluster membership as modelled by a free Voronoi tessellation: (a) locations of individuals here assumed to be regularly sampled over space; (b) some artificial auxiliary points (often referred to as nuclei) are introduced to represent the locations of some polygons approximating the domain of the sought-after clusters. Note that the number and location of nuclei do not coincide with those of the sampling sites. Each nucleus defines a polygon (consisting of all points in the spatial domain closer to this nucleus than to any other nucleus) and these polygons define collectively a tessellation of the spatial domain known as Voronoi tessellation; (c) making inference about the domain of the cluster amounts to infer the location of the nuclei and the cluster membership of the polygons (often referred to as colour in stochastic geometry and represented as such here). Bottom row: (d–f) similar to (a–c) for some individuals irregularly sampled over the spatial domain. Note that the tessellation relies on the auxiliary points only and the location of the sampled individuals does not play any role. This is the core of the model underlying the spatial algorithm in GENELAND.

ing models that try to infer populations or clusters of individuals that fit some genetic criteria that define them as distinct groups. The detailed assumptions underlying these models vary, but the simplest versions assume that: (i) each individual's genome has its origin in a single cluster (no-admixture); (ii) the genotypic proportions in each sought-after cluster are at HWE, with loci at linkage equilibrium (HWLE); and (iii) a minimum of between-cluster genetic differentiation is present (cluster-specific allele frequencies). This scenario corresponds to the situation where the data set at hand consists of a few well-differentiated panmictic clusters and is the model underlying the no-admixture option of the STRUCTURE program pioneered by Pritchard *et al.* (2000). While this earlier version of the program also included an option to account for hybrid individuals (admixture options), subsequent developments include a model to deal with allele frequencies displaying correlation between clusters, a model to deal with linked loci (Falush *et al.* 2003) and a scheme to account for dominant markers (Falush *et al.* 2007).

Clustering individuals from genotypes is a task that strictly speaking does not require the use of any spatial information. The models mentioned above only make use of the genetic information at hand. They make the implicit assumption that individual cluster membership does not display any particular spatial pattern, a scenario corresponding to the very specific situation where the putative factor limiting gene flow would be completely spatially unstructured. Such a pattern could be, for instance, due to assortative mating or some other kind of behavioural barrier, or to specialization to dif-

ferent types of sympatric hosts in the case of parasite species (see, e.g. Martel *et al.* 2003 about host plant-mediated sympatric speciation). Except in those very particular cases, complete spatial randomness is not realistic. Some artificial examples of putative spatial patterns are given in Fig. 3 where a pattern of complete spatial randomness is shown in the left panel. Importantly, a model making no particular assumption about spatial patterns would consider all three patterns as *a priori* equally likely. Two types of alternative models are described in the next sections.

Spatial model of cluster membership based on a tiling of the continuous spatial domain

Motivations. The presence of physical barriers to dispersal is an important factor that limits gene flow. For example, Rieseberg *et al.* (2009) report that about one-half of the studies published in *Molecular Ecology* related genetic population structure to the presence of a barrier. There can be physical barriers of human origin such as roads (Coulon *et al.* 2006; Riley *et al.* 2006; see also Gauffre *et al.* 2008 on this question), urban areas and areas of human activity (for species living in a linear habitat (see, e.g. Monaghan *et al.* 2001; Yamamoto *et al.* 2005 for the effect of dams on fresh water fishes; see also Su *et al.* 2003 for the effect of the Great Wall of China on certain plant species). Barriers of natural origin have also been widely reported. They include, for example, climate conditions (Stenseth *et al.* 2004; Pilot *et al.* 2006), oceanographic features (Fontaine *et al.* 2007; Galarza *et al.* 2009), vegetation cover (Sacks *et al.* 2008) and waterways (Coulon *et al.* 2006).

Model. Generally, while the exact location of dispersal barriers is unknown, it is reasonable to expect that they have a simple spatial shape. At least, this was the rationale for the development of the model underlying the GENELAND program (Guillot *et al.* 2005a,b, 2008; Guillot 2008). The central assumption of this model is that the spatial domain occupied by the inferred clusters can be approximated by a small number of polygons. While this approach can be formalized in various ways (see Stoyan *et al.* 1995; Lantuéjoul 2002; Møller & Stoyan 2009 for overviews), in GENELAND, the model assumed is the Voronoi tessellation (see Fig. 4 for graphical examples). Polygons are assumed to be centred on some artificially introduced auxiliary points (referred to as *nuclei*). Making inference about clusters domains (and thus about cluster memberships of individuals) amounts to inferring the location and cluster memberships (thought of as colours) of these polygons.

This component of the model is referred to as *free* Voronoi tessellation, as the polygons are constructed independently of the sampling sites. Related models also based on polygons have been used by Blackwell (2001) to model the territories of clans of badgers *Meles meles* (with nongenetic data) and by Wasser *et al.* (2007) where the polygons are used to model preferential spatial sampling and/or variations of population density in space. Polygon-based spatial cluster domains impose some spatial consistency (in the sense that they avoid complete spatial randomness) and allow the interpolation of cluster membership, i.e. the prediction of values outside the set of sampling sites. In the case of the GENELAND program, specific features include schemes to account for uncertainty in spatial coordinates, to account for null alleles and estimate their frequency at each locus, and the computation of the posterior distribution of all inferred parameters. The latter feature allows one to assess the relative confidence one might place on each value of K and to get a detailed probability map of assignments to evaluate the degree of uncertainty of the estimated cluster memberships.

Spatial model of cluster membership based on a graph

Motivations. Models based on free tessellations as described above might have some drawbacks if the putative spatial domains do not have simple shapes. This might be the case in instances of assortative mating that are only weakly spatially structured. For instance, in studies of human genetic variation at small spatial scales, gene flow may be better explained by social relationships than geographic distance (see, e.g. Britton *et al.* 2008 for a development of this idea in epidemiology). If a complex spatial pattern is expected it can be more fruitful to make assumptions about the statistical distribution of cluster membership of individuals that are spatial neighbours, rather than making generic assumptions about the spatial domains occupied by the sought-after clusters.

A family of models designed exactly for this purpose has long been used in statistical physics and image analysis (see Hurn *et al.* 2003 and references therein). The key idea, central to the methods used for image de-noising, is that neighbouring pixels in an image are more likely to share the same colour than a set of pixels taken at random. By analogy, this idea can be extrapolated to mean that individuals that are neighbours are more likely to belong to the same cluster than individuals taken at random over the whole sampling area. This approach was taken in the model proposed by François *et al.* (2006) and implemented in

the softwares GENECLUST (Ancelet & Guillot 2006) and TESS (a simplified version of GENECLUST proposed by Chen *et al.* 2007) and in the model proposed by Corander *et al.* (2008) and implemented under the spatial option of the BAPS program. A model based on a similar rationale has also been used by Vounatsou *et al.* (2000) to map haplotype frequencies. This idea involves two steps: (i) defining what 'neighbours' means, which often entails some kind of arbitrariness (especially when the sites sampled are not regularly spaced) and (ii) modelling how likely 'neighbours' are to belong to the same clusters.

Model. Towards step (i), both François *et al.* (2006) and Corander *et al.* (2008) use a Voronoi tessellation. However, in contrast with Guillot *et al.* (2005a), this tessellation is not constructed independently of the sampling sites but built *on* them. This model is therefore referred to as *constrained* Voronoi tessellation (see Figs S1–S3 for examples). Loosely speaking, two sampling sites are considered to be neighbours if there is no other sampling site 'around a straight line' that joins them. This has a number of drawbacks which will be discussed later.

Toward step (ii), François *et al.* (2006) considered a so-called Potts model. It is used to inject some information about how likely 'neighbours' are to belong to the same cluster. In this model, the log probability of an individual belonging to a cluster, given the cluster membership of its neighbours, is proportional to the number of neighbours belonging to this cluster. In statistical physics, the proportionality coefficient ψ is referred to as the interaction parameter. Quoting Ripley (1991): 'One problem with using Markov random field priors is that their parameters are not immediately interpretable […]'. It has usually no phenomenological interpretation and should be viewed as a statistical way of injecting the information that complete spatial randomness is unlikely.

In the model proposed by François *et al.* (2006) and reused partly by Chen *et al.* (2007), the interpretation of the interaction parameter is even more challenging. Indeed, the model is defined on a graph which vertices depend on the sampling scheme and, therefore, the interpretation of the parameter changes with the sampling scheme. Estimating the number of clusters in this model is a difficult task and there is no solution implemented and validated as of today (see discussion below). Corander *et al.* (2008) proposed an alternative model in which the probability of a given colouring model does not have a closed expression but depends on the local properties of the graph. This model does not belong to the mainstream spatial statistics toolbox and is difficult to visualize and interpret.

The correct inference of population genetic structure

Problems related to the estimation of K. One of the most challenging problems for clustering models is the correct estimation of the number of clusters K (here we refer to the correct number in terms of statistical inference; see Waples & Gaggiotti 2006 for a discussion of its biological meaning). In STRUCTURE, the estimation of the number of clusters K is based on an approximation of its posterior distribution obtained from Markov chain Monte Carlo (MCMC) runs with different putative K values proposed by the user. The program needs to be run for each value of K and the corresponding approximate values of posterior probabilities need to be compared which involves some subjective appreciation. Evanno *et al.* (2005) have suggested that this procedure lacked accuracy and proposed an alternative strategy. According to Waples & Gaggiotti (2006), however, this alternative strategy brings little improvement (see also Latch *et al.* 2006 for an assessment of the accuracy of Bayesian clustering programs).

The estimation of K within a formal statistical model and algorithm has been first proposed by Dawson & Belkhir (2001) and implemented in the PARTITION program. Corander *et al.* (2003) and Corander *et al.* (2004) tackled the problem with a slightly different algorithm within a similar model. Subsequent developments of the BAPS program include a scheme to use information about known clusters (also referred to as baseline clusters) (Corander *et al.* 2006), to estimate admixture coefficients (Corander & Marttinen 2006) and to use linked loci (Corander & Tang 2007). In the latest version of the BAPS program, the estimation of K is based on a Monte Carlo maximization of the posterior distribution. An alternative model and algorithm to estimate K in the no-admixture model has also been proposed by Pella & Masuda (2006) and implemented in the STRUCTURAMA program by Huelsenbeck & Andolfatto (2007).

GENECLUST does not formally address the question of estimating K. Instead its strategy consists in fixing K to a large value and counting the number of nonempty clusters at the end of the run. Chen *et al.* (2007) introduced one further simplification of the global algorithm by skipping the estimation of the interaction parameter in the TESS program. The claim by Chen *et al.* (2007) that this omission provided accurate estimates of K has been strongly criticized and it has been shown that inferences in Tess can be highly inaccurate (Guillot 2009a,b). Furthermore, it has been reported that setting the interaction parameter ψ to 0 in GENECLUST and TESS produced results comparable with those of the program STRUCTURE. This is inaccurate because the inference of K is carried out with different algorithms in the two programs (see also, e.g. Cullingham *et al.* 2009 for empirical evidence).

Guillot *et al.* (2005a) reported some concerns regarding the estimation of K in GENELAND. These were related to the use of a two-step algorithm (a first run to estimate K, a second run with fixed K) and some other algorithm weaknesses. These issues were solved in later versions of the program (Guillot 2008; Guillot *et al.* 2009), where the inference of the optimal number of K is based on a single-step strategy (see also Dawson & Belkhir 2009) that includes a review of some more specific clustering models (e.g. where sought-after clusters are families).

Checking compliance with modelling assumptions and the need for model selection methods. After having run a clustering model, there is no straightforward way of confirming that the data set consists of the inferred number of HWLE clusters. In addition, in contrast to domesticated populations (see, e.g. Rosenberg *et al.* 2001, population membership in wildlife can generally not be independently validated and might not even exist. At the very least, one can check that the inferred clusters comply with the HWLE assumptions and that the allele frequencies of the inferred genetic clusters are significantly differentiated from each other. If the clusters are not under HWLE, one can look for common sources of discrepancy, including residual Wahlund effects (undetected clusters due, e.g. to low differentiation), strong IBD (see also below) or other forms of (spatially or nonspatially structured) departure from random mating. More generally, there is a need for statistical methods allowing to select among various models implemented in the clustering programs.

Decision from the output of several runs or programs. Several authors have noted that different clustering algorithms can infer different solutions for the optimal partitioning of a data set. For example, Rowe & Beebee (2007) reported noncongruent outputs of BAPS, GENELAND and STRUCTURE when analysing genetic data of Natterjack toads *Bufo calamita* in Great Britain. This phenomenon can arise from differences in the underlying models, in the statistical estimators or in approximations in the algorithm used to compute this estimator. It is in general difficult to disentangle the relative effects of these three sources of disagreement. It is important to bear in mind that all programs discussed here are based on MCMC and are hence prone to convergence issues. This means that outputs of the programs might not be in certain cases the exact solution of the mathematical equations but an approximation which quality remains unknown. An efficient strategy to check that program outputs are not subject to this kind of error is to run a large number of long runs and to check that those runs give similar outputs. Note that the comparison of different runs must take the possible swap or switch of cluster labels into account. This problem is known as the label switching issue in the analysis of mixture models and can be addressed by the computer programs CLUMPP (Jakobsson & Rosenberg 2007) and PARTITIONVIEW (Dawson & Belkhir 2009).

Given that different clustering algorithms can produce different solutions, it is good practice to analyse genetic data with more than one method. If the outputs coincide, it is suggestive of the presence of a strong genetic signal (at least if the data set is not characterized by IBD; see below). If the various outputs do not coincide, one can speculate that departure from modelling assumptions interplay with the usual MCMC convergence issues. In this latter case, we warn against ignoring the nonconvergence symptoms and choosing the clustering solution that most conforms to some *a priori* expectation (see Frantz *et al.* 2009 for an example of how inferences based on a single algorithm could lead to different conclusions compared with a consensus based on several programs).

Isolation by distance, the cline vs. clusters dilemma and the optimization of spatial sampling scheme. Effect of isolation by distance and the cline vs. clusters dilemma. Another confounding factor of the clustering algorithms is IBD. All the models make sense fully only at a scale that is small enough to ignore its effect. At larger spatial scales, any species is affected by IBD and assuming within-cluster panmixia becomes inappropriate.

Several authors have studied how clustering models behave for organisms whose mating is restricted by distance. Frantz *et al.* (2009), Schwartz & McKelvey (2009) and Guillot & Santos (2009) reported that clustering models are affected by IBD regardless of the cluster membership prior used (spatial or nonspatial). The general effect reported is that the presence of clinal variations tends to be interpreted as the presence of clusters and a number of clusters larger than one is generally inferred, even though no barrier to gene flow was present. Guillot & Santos (2009) observed that this effect is weak in case of weak IBD. If the presence of IBD is suspected, it is important to test compliance with HWLE globally and for each inferred cluster. Plots of genetic distances against geographic distances coloured according to cluster memberships (as in McRae *et al.* 2005; Rosenberg *et al.* 2005; Fontaine *et al.* 2007) can be a great aid in assessing whether genetic variations are explained by distance alone or whether other factors are involved. An implementation of this method is presented in Box 2. While it is recommended to implement this method, it may lack power for data sets with continuous spatial sampling where only a small number of clusters are inferred. We also note that this method relies on a visual check and that there is a need for a

formal statistical method in this context. In practice, it can be difficult indeed to distinguish between genuine and artificial genetic clusters in data sets characterized by IBD (see, for example Frantz *et al.* (2009).

What do spatial clustering models have to say about IBD per se*?*. The view that most organisms are subject to IBD is widely accepted (see Lawson-Handley *et al.* 2007; Schwartz & McKelvey 2009 and references therein). The use of a spatially dependent prior of cluster membership (as in Guillot *et al.* 2005a; François *et al.* 2006) amounts to injecting in the model the property that in average (over all possible clusterings) the genetic similarity decreases continuously with distance. However, conditionally on the cluster membership variables, these models assume the existence of several clusters at Hardy–Weinberg equilibrium. As this condition is not fulfilled for organisms whose mating is restricted by distance only, spatial clustering models are not more suitable than nonspatial clustering models for analysing organisms under IBD.

How can spatial sampling schemes be optimized?. The optimization of spatial sampling schemes is notoriously a difficult question. One aspect of the problem is that the optimal sampling depends on the true pattern which is not known in advance. One has therefore to base the decision on average expected features. Guillot & Santos (2009) noted that the spatial sampling scheme (regular or irregular, see Storfer *et al.* 2007; Schwartz & McKelvey 2009 for examples) plays little role in the way clustering models are affected by IBD. Inferences were found to be accurate regardless of the sampling scheme used if the data set consists genuinely of HWLE clusters and inaccurate if these clusters were subject to IBD. In traditional geostatistical studies, it is often recommended to sample the whole study domain but in such a way that the various scales are investigated (containing both close and distant pairs of sampling sites; see Diggle & Lophaven 2006; Diggle & Ribeiro 2007). The recent study by Schwartz & McKelvey (2009) concludes with similar recommendations.

Summary and discussion on clustering models

While genetic clustering of individuals is a task that, strictly speaking, does not require the use of spatial information, in natural populations most barriers to gene flow are in some sense related to variables structured in space. The advantage of using spatial vs. nonspatial clustering models is potentially the ability to get more accurate results, in particular when analysing data sets consisting of a small number of loci, or characterized by low levels of genetic differentiation (Guillot *et al.* 2005a; Frantz *et al.* 2006; Fontaine *et al.* 2007; Dudaniec *et al.* 2008; Lecis *et al.* 2008). Existing models have been used to address questions in ecology relating to habitat specialization, habitat fragmentation, metapopulation dynamics (Coulon *et al.* 2008; Orsini *et al.* 2008), epidemiology (see Section 'Using spatial genetics methods to investigate disease spread' in the Supporting Information), colonization patterns of endangered or introduced species (Dudaniec *et al.* 2008; Janssens *et al.* 2008; Lecis *et al.* 2008), population management (Zannèse *et al.* 2006; Fuentes-Contreras *et al.* 2008) and forensics (Frantz *et al.* 2006).

Although other modelling techniques would be possible, existing programs are all based on the constrained (BAPS, GENECLUST and TESS) or free (GENELAND) Voronoi tessellation. The advantage of the latter is to be independent of the sampling scheme and to provide an (out-of-sample) spatial prediction of population spread. This is possible thanks to the use of auxiliary points (the nuclei). This involves some extra computational burden but has one important advantage: it provides directly a map of cluster domains without subjective decision incurred by manually post-processing estimated individual cluster memberships. This map of posterior probabilities of cluster membership can be interpreted as a map of admixture coefficients and can be used, e.g. for the study of secondary contact zones (Sacks *et al.* 2008).

We hope to have illustrated that the partitioning of individuals into spatial domains based on genetic data is not a straightforward problem. The relative ease with which the various clustering programs can be utilized should not lead to a blind faith in an output best summarized by the *incipit* to chapter 3 of Chilès & Delfiner (1999): *Once a map is drawn, people tend to accept it*. Instead, one should be critical of results and consider whether the underlying assumption of a model might have been violated when analysing the data set at hand. A case in point is IBD: as the clustering methods assume the data set to consist of a number of panmictic clusters, the presence of IBD in the data set can lead to the identification of spurious clusters. It is apparent that future spatial clustering models need to control for the effect of IBD.

We urge researchers to use various Bayesian clustering approaches to investigate the spatial genetic structure in order to evaluate the robustness and reliability of the inferred results. Different estimates of the number of clusters can be obtained using slightly different models, and even using slightly different algorithms under similar models. The results of similar models and algorithms should agree in order to have confidence in the proposed clustering solution. We can only emphasize again not to ignore instances of nonconvergence, as they might point towards spurious results. It might be

helpful to consider whether all the inferred clusters can be explained in a biologically meaningful way. However, the biological interpretation of the presence of genetic discontinuities (and even sometimes their absence) can be challenging. This is discussed further below.

Conclusion

Summary

Current statistical methods rely mostly on IBD and clustering models and allow one to relate genetic variations to demographic and migratory processes, which can, in turn, be related to some aspects of the surrounding landscape. Recent methods comply with a broad variety of spatial sampling schemes, particularly with continuous spatial sampling, allowing more efficiently investigation of how spatial patterns of genetic variations relate to environmental features. On the other hand, the current IBD or clustering models are still based on idealized assumptions and investigations of the relationship between genetic variations and landscape features often rely on descriptive statistics that make it challenging to choose among several competing explanations. Below, we outline several research directions in terms of models or statistical inference techniques that might help to improve methods in these respects.

Geostatistical models

The use of geostatistical ideas in population genetics has probably been hampered by the difficulty of dealing with multivariate and categorical data. In particular, it has long been a difficult problem to capture the spatial variation of categorical variables with a parsimonious model. This difficulty has now been partially alleviated by hierarchical models (e.g. as in Wasser *et al.* 2004), and the relative weak informativeness of a single locus can be compensated by the increased availability of data sets including a large number of loci. Although some difficulties can be anticipated regarding the validity of spatial autocorrelation models (Matheron 1987, 1993; Chilès & Delfiner 1999), one important direction for future work could be to try to further relate demographic and genetic parameters in a genetic model to parameters of geostatistical models.

Marked point process models

Traditional spatial autocorrelation analysis relies on the assumption that the location of sampling sites is not informative about the process. This is the case, for example, if one measures temperature at different sites with either regularly spaced coordinates (nodes of a regular grid) or sampled at random over the study domain. In this case, the coordinates arise as a choice of the scientist and do not reflect any intrinsic property of the process under study. By contrast, sampling designs in ecology are often dependent on the species density. This dependence can arise if some prior knowledge about population density is explicitly used in the sampling design (e.g. one avoids to sample low population density areas) or as a consequence of lower probability to capture individuals in areas of low population density. In these cases, sampling locations can no longer be considered as independent from genotypes but should be considered as part of the process to be analysed and modelled. In this context, standard results about correlation functions and how they can be interpreted are no longer valid. The situation where the variable (the multi-locus genotype in our context) and the location of measurements are both informative is addressed by so-called marked point-process models (see Cressie 1994; Schlather *et al.* 2004 and references therein). Schlather *et al.* (2004) proposed some methodological developments to analyse data in exactly this situation. Unfortunately for population geneticists, these methods only consider a single quantitative variable. Some useful and more context-specific modelling suggestions can be found in Shimatani (2002), Shimatani & Takahashi (2003) and Shimatani (2004).

Bridging the gap between clustering models and isolation-by-distance models

While restricted dispersal leading to local genetic drift and differentiation can be studied by IBD models, differentiation induced by barriers to gene flow is addressed by clustering models. In many studies, the two factors interplay and one factor can act as a confounding factor in the assessment of the other one. There is therefore a need for models allowing one to assess and quantify in a unified conceptual and inferential framework the effect of restricted dispersal and barriers on gene flow.

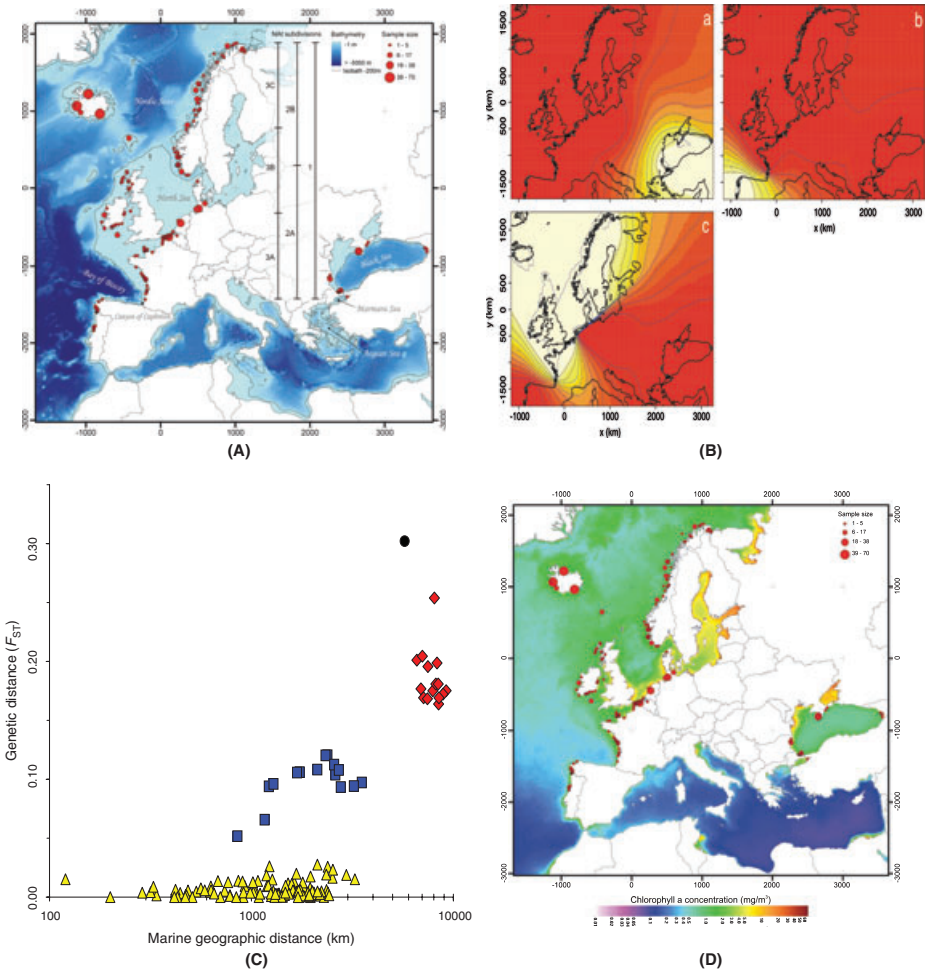
Assessing how plausible genetic structure is under different scenarios

While some clustering results can be difficult to interpret biologically, it is also possible that clusters do not match any environmental feature at all (Zannèse *et al.* 2006; Sahlsten *et al.* 2008). In this context, it can be useful to analyse simulated data to assess the validity of the inferred clusters. For example, Frantz *et al.* (2009) used various Bayesian clustering methods to infer the genetic structure of a continuously distributed population of wild boar

Box 2: Disentangling the effect of isolation by distance and of barriers to gene flow: an example in seascape genetics

Fontaine *et al.* (2007) conducted a study on the harbour porpoise *Phocoena phocoena*, a highly mobile cetacean, using a data set of 752 individuals sampled across Europe (panel A) and genotyped at 10 microsatellite loci. The authors found evidence for the existence of at least three distinct genetic clusters (panel B). Given the spatial scale considered, they suspected a confounding effect of IBD that might have been exacerbated by the irregular spatial sampling.

Pairwise F_{ST}s are plotted against marine geographic distances, pairs of sites assigned to the same cluster are represented by triangles, while pairs of sites assigned to different clusters were represented by either a square or a diamond depending on the cluster combination (panel C). This revealed that for a given class of spatial distances, genetic distances are typically much larger for pairs of sites belonging to distinct clusters than for pairs of sites belonging to the same cluster. This idea could be formalized a step further by applying a statistical test of equality of coefficients in the regression plot. This strongly suggests that differentiation is not solely caused by IBD but also by the existence of a hidden variable. Further analysis revealed that the locations of genetic barriers coincide with areas of changes in environmental characteristics (availability of nutrient panel D). This study provided for the first time evidence that cryptic environmental processes have a major impact on the genetic and demographic structure of cetaceans.

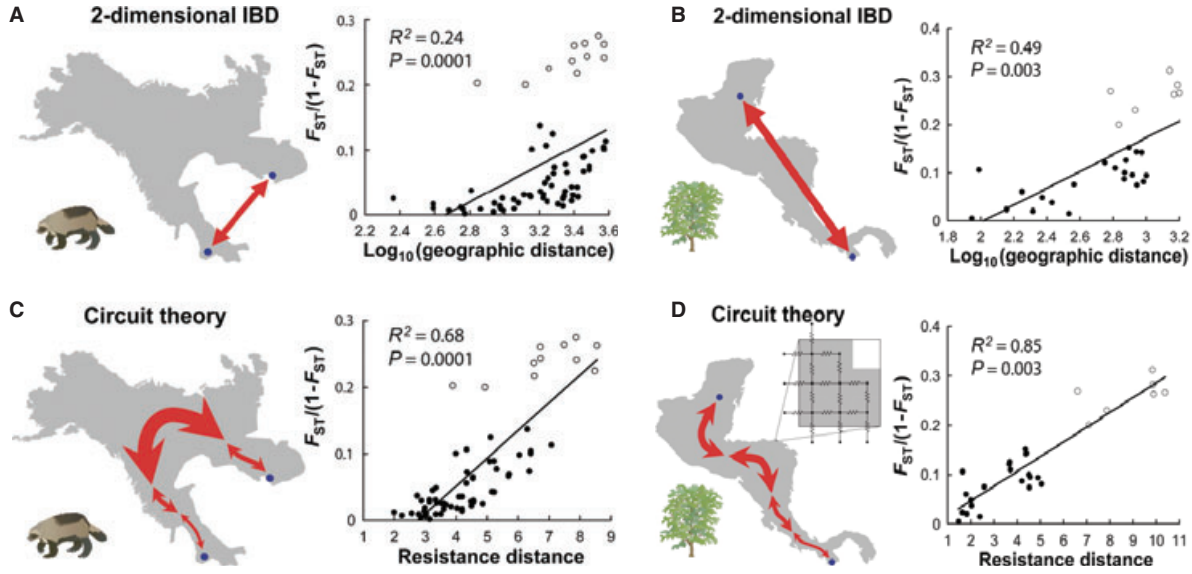


(*Sus scrofa*). Different algorithms inferred different clustering solutions and, in some instance, there was no ecological explanation for the inferred genetic discontinuities. As it had been reported previously that deviations from random mating that were not caused by genetic discontinuities might bias results, the authors

Box 3: Resistance distance

Even if some of the models discussed in the identity by distance section include the possibility of nonhomogeneous density, they all assume that the landscape itself is homogeneous and in particular opposes the same resistance to dispersal everywhere. This assumption can be problematic, for example, if dispersal pathways are restricted to narrow corridors in certain areas. In this case, straight line geographical distances might not reflect the underlying ecological process during dispersal events and distances integrating heterogeneities in dispersal pathways might be more relevant. To address this issue, McRae (2006) suggested to use the resistance distance. In analogy with circuit theory where current does not flow along a single one-dimensional path but across the whole material, this distance is defined as the effective resistance that would oppose a conductive material displaying a topology similar to that of the study area. Studies on mahogany trees and wolverines showed empirically that resistance distance better correlated with genetic distances than the usual two-dimensional straight line distance.

Generally, the use of the resistance distance might help to reveal patterns of IBD in heterogeneous landscapes that would not have appeared with the use of Euclidean distances. Pairwise plots of genetic distances vs. 2D Euclidean (straight line) distances (A, C) and resistance distances (B, D). Open circles indicate pairs including the most southern site (Figures reprinted from McRae & Beier 2007).



analysed simulated data sets characterized by different levels of IBD, with and without barriers to gene flow. This approach confirmed that the inferred clusters might be artefacts and the authors were unable to make firm conclusion as to the presence of barriers to gene flow in their study area.

Gauffre *et al.* (2008) recently investigated the population genetic structure of common voles (*Microtus arvalis*) in an agricultural landscape in France using individual-based Bayesian clustering methods. Contrary to expectations, a motorway in the study area was not associated with a spatial genetic discontinuity. Simulating genotype data based on coalescence theory and plausible scenarios of genetic drift, the authors showed that, despite simulating a complete dispersal barrier, the effective population size of their study population was too large for the populations to have diverged substantially since the construction of the motorway.

More generally, before investing time and resources into a spatial genetic study, it might be a worthwhile exercise to analyse genotype data simulated under various realistic migration–mutation–drift scenarios to assess whether it is *a priori* likely that the empirical data set consists of two or more clearly differentiated subpopulations. One potential problem with simulating genetic drift is that information on the effective size of the study population is required. Both Frantz *et al.* (2009) and Gauffre *et al.* (2008) got around this problem by simulating multiple data sets spanning a range of feasible values and then selecting those data sets that most closely

resemble a specific parameter in their empirical data set: in the former study the degree of IBD and in the latter study the observed heterozygosity. Programs such as EASYPOP (Balloux 2001), MS (Hudson 2002), METASIM (Strand 2002), GENELAND (Guillot & Santos 2009), IBDSIM (Leblois *et al.* 2009) offer different complementary tools to simulate data. The recently introduced program DIYABC (Cornuet *et al.* 2008) allows one to infer parameters for one or more scenarios and compute bias and precision measures for given scenarios and therefore offers a flexible way to investigate and compare competing evolutionary hypotheses formulated from the clues gathered from more classical models.

Injecting more biological knowledge into statistical models

We have noted in this review that most quantitative approaches are based on idealized and abstract models that leave little possibility of injecting some biological knowledge about the species, population or area under study. For instance, landscape genetics studies for terrestrial species, marine species and fresh water fish are carried out with the same statistical tools despite the obvious differences, e.g. in terms of habitat and dispersal processes (see Galindo *et al.* 2006; Kalinowski *et al.* 2008 for exceptions and Selkoe *et al.* 2008 for discussions about incorporation of ecological and oceanographic information into seascape genetics study). There is generally a need for more realistic models that include species- and context-specific knowledge. Designing such models would not necessarily comply with the common usual requirement of a known and well-defined prior and likelihood function. Approximate Bayesian computation methods (ABC; see, e.g. Sisson & Fan 2009 for a review) open an avenue for carrying out inference in this kind of situations.

Need for model selection tools

The anticipated increase in models and programs in the next few years will strengthen the need for tools for model selection. This is a notoriously difficult problem in statistics and the case of ecological and genetic models is particularly delicate as most of them are essentially descriptive (as opposed to predictive) and there is no obvious choice as to which criterion should be chosen to compare models.

New challenges brought up by landscape genomics

So far, most landscape genetics studies have focused on the analysis of neutral genetic variations. However, the rapid technological advents of sequencing technologies will probably allow the development of landscape genomics. This emerging subdiscipline can be defined as the genome-wide identification of molecular markers potentially under selection and involved in adaptation to different environments. In addition to the existing issue of disentangling the effects of selection and drift, one will have to face the challenge set by massive data sets typical of genomics.

Acknowledgements

Part of this work has been carried out when G. Guillot was at Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, Oslo, Norway, with funding from the Research Council of Norway under the INTEGRATE project. This work was also partly supported by Agence Nationale de la Recherche with grant no. NT05-4-42230.

References

Amos W, Manica A (2006) Global genetic positioning: evidence for early human population centers in coastal habitats. *Proceedings of the National Academy of Sciences*, **103**, 820–824.

Ancelet S, Guillot G (2006) GENECLUST program documentation. Tech Rep., Institut National de la Recherche Agronomique.

Balloux F (2001) EASYPOP (version 1.7), a computer program for the simulation of population genetics. *Journal of Heredity*, **92**, 301–302.

Banerjee S, Gelfand A (2006) Bayesian wombling. *Journal of the American Statistical Association*, **101**, 1487–1501.

Barbujani G, Oden N, Sokal R (1989) Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

Barton N, Depaulis F, Etheridge A (2002) Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, **61**, 31–48.

Bateman A (1950) Is gene dispersion normal? *Heredity*, **4**, 353–363.

Beerli P (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–345.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, **98**, 4563–4568.

Blackwell P (2001) Bayesian inference for a random tessellation process. *Biometrics*, **57**, 502–507.

Bocquet-Appel J, Bacro J (1994) Generalized wombling. *Systematic Biology*, **3**, 316–329.

Britton T, Deijfen M, Lagerås N, Lindholm M (2008) Epidemics on random graphs with tunable clustering. *Journal of Applied Probability*, **45**, 743–756.

Broquet T, Johnson C, Petit E, Burel F, Fryxell J (2006) Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. *Molecular Ecology*, **15**, 1689–1697.

Cercueil A, François O, Manel S (2007) The genetical bandwidth mapping: a spatial and graphical representation

of population genetic structure based on the Wombling method. *Theoretical Population Biology*, **71**, 332–341.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Chilès J, Delfiner P (1999) *Geostatistics*. Wiley, London.

Clark J, Silman M, Kern R, Macklin E, Hillerislambers J (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, **80**, 1475–1494.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multi-locus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Tang J (2007) Bayesian analysis of population structure based on linked molecular information. *Mathematical Biosciences*, **205**, 19–31.

Corander J, Waldmann P, Sillanpää M (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Corander J, Waldmann P, Martinen P, Sillanpää M (2004) **baps2**: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. *Fishery Bulletin*, **104**, 550–558.

Corander J, Sirén J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**, 111–129.

Cornuet J, Santos F, Beaumont M *et al.* (2008) Inferring population history with DIYABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Coulon A, Guillot G, Cosson J *et al.* (2006) Genetics structure is influenced by landscape features. Empirical evidence from a roe deer population. *Molecular Ecology*, **15**, 1669–1679.

Coulon A, Fitzpatrick J, Bowman R *et al.* (2008) Congruent population structure inferred from dispersal behavior and intensive genetic surveys of the threatened Florida Scrub-Jay *aphelocomacœrulescens*. *Molecular Ecology*, **17**, 1685–1701.

Cressie N (1994) *Statistics for Spatial Data. Series in Probability and Mathematical Statistics*. Wiley, London.

Crida A, Manel S (2007) Wombsoft: an R package that implements the Wombling method to identify genetic boundaries. *Molecular Ecology Notes*, **7**, 588–591.

Cullingham C, Kyle C, Pond B, Rees E, White BN (2009) Differential permeability of rivers to tacoon gene flow corresponds to rabies incidence on Ontario, Canada. *Molecular Ecology*, **18**, 43–53.

Dawson K, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.

Dawson K, Belkhir K (2009) An agglomerative hierarchical approach to visualisation in Bayesian clustering problems. *Heredity*, **103** (1), 32–45.

DiCiccio T, Efron B (1996) Bootstrap confidence intervals (with discussion). *Statistical Science*, **11**, 189–228.

Diggle P, Lophaven S (2006) Bayesian geostatistical design. *Scandinavian Journal of Statistics*, **33**, 53–64.

Diggle P, Ribeiro P (2007) *Model-based geostatistics*. Springer Verlag, New York.

Dobzhansky T, Wright S (1941) Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. *Genetics*, **26**, 23–51.

Dudaniec RY, Gardner MG, Donnellan S, Kleindorfer S (2008) Genetic variation in the invasive avian parasite, *Philornis downsi* (Diptera, Muscidae) on the Galapagos archipelago. *BMC Ecology*, **8**, Article No. 13.

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

Endler J (1977) *Geographical variation, speciation, and clines*. Princeton University Press, Princeton, New Jersey.

Epperson B (1993) Recent advances in correlation studies of spatial patterns of genetic variation. *Evolutionary Biology*, **27**, 95–155.

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

Epperson B (2005) Estimating dispersal from short distance spatial autocorrelation. *Heredity*, **95**, 7–15.

Epperson B (2007) Plant dispersal, neighborhood size and isolation by distance. *Molecular Ecology*, **16**, 3854–3865.

Epperson B, Li T (1997) Gene dispersal and spatial genetic structure. *Evolution*, **51**, 672–681.

Evanno G, Regnault S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Review Genetics*, **7**, 745–758.

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.

Falush D, Stephens M, Pritchard J (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Felsenstein J (1975) A pain in the torus: some difficulties with models of isolation by distance. *American Naturalist*, **109**, 359–368.

Fenster C, Vekemans X, Hardy O (2003) Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (leguminosae). *Evolution*, **57**, 995–1007.

Fontaine M, Baird S, Piry S *et al.* (2007) Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in old world waters. *BMC Biology*, **5**, 30, doi:10.1186/1741–7007-5-30.

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

Frantz AC, Tigel Pourtois J, Heuertz M *et al.* (2006) Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. *Molecular Ecology*, **15**, 3191–3203.

Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

Fuentes-Contreras E, Espinoza JL, Lavandero B, Ramirez CC (2008) Population genetic structure of codling moth (Lepidoptera: Tortricidae) from apple orchards in central Chile. *Journal of Economic Entomology*, **101**, 190–198.

Galarza J, Carreras-Carbonell J, Macpherson E *et al.* (2009) The influence of oceanographic fronts and early-life-history traits

on connectivity among littoral fish species. *Proceedings of the National Academy of Sciences*, **106**, 1473–1478.

Galindo H, Olson D, Palumbi S (2006) A coupled oceanographic-genetic model predicts population structure of Caribbean corals. *Current Biology*, **16**, 1622–1626.

Gauffre B, Estoup A, Bretagnolle V, Cosson J (2008) Spatial genetic structure of small rodent in a heterogeneous landscape. *Molecular Ecology*, **17**, 4616–4629.

Gonzales-Suarez MFR, Hedrick PW, Aurioles-Gamboa D, Gerber L (2009) Isolation by distance among California sea lion populations in Mexico: redefining management stocks. *Molecular Ecology*, **18**, 1088–1099.

Guillot G (2008) Inference of structure in subdivided populations at low levels of genetic differentiation. The correlated allele frequencies model revisited. *Bioinformatics*, **24**, 2222–2228.

Guillot G (2009a) On the inference of spatial structure from population genetics data using the Tess program. *Bioinformatics*, **25** (14), 1796–1801.

Guillot G (2009b) Response to ''Comment on the inference of spatial structure from population genetics data''. *Bioinformatics*, **25** (14), 1805–1806.

Guillot G, Santos F (2009) A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, **9** (4), 1112–1120.

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

Guillot G, Mortier F, Estoup A (2005b) GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 708–711.

Guillot G, Santos F, Estoup A (2008) Analysing georeferenced population genetics data with GENELAND: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics*, **24**, 1406–1407.

Guillot G, Santos F, Estoup A (2009) *Population genetics analysis using R and GENELAND*. Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo. http://www2.imm.dtu.dk/g˜igu/Geneland/Geneland-Doc.pdf.

Hannelius U, Salmela E, Lappalainen T *et al.* (2008) Population substructure in Finland and Sweden revealed by a small number of unlinked autosomal SNPs. *BMC Genetics*, **9**, 54, doi: 10.1186/1471-2156-9-54.

Hardy O (2003) Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Molecular Ecology*, **12**, 1577–1588.

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.

Heywood J (1991) Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics*, **22**, 335–355.

Holdregger R, Wagner H (2006) A brief guide to landscape genetics. *Landscape ecology*, **2006**, 793–796.

Hudson R (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huelsenbeck J, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. *Genetics*, **175**, 1787–1802.

Hurn M, Husby O, Rue H (2003) A tutorial in image analysis. In: *Spatial Statistics and Computational Methods* (ed Moller J) pp. 87–141. Lecture Notes in Statistics. Springer Verlag, New York.

Irwin D, Bench S, Price T (2001) Speciation in a ring. *Nature*, **409**, 333–337.

Jakobsson M, Rosenberg N (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Janssens X, Fontaine MC, Michaux JR *et al.* (2008) Genetic pattern of the recent recovery of *European otters* in southern France. *Ecography*, **31**, 176–186.

Jombart T, Dufour A, Pontier D (2009) Genetic markers in the playground of multivariate analysis. *Heredity*, **102**, 330–341.

Joseph L, Dolman G, Donnellan S, Saint K, Berg M, Bennett A (2008) Where and when does a ring start and end? Testing the ring-species hypothesis in a species complex of Australian parrots. *Proceedings of the Royal Society of London, Series B*, **275**, 2431–2440.

Kalinowski ST, Meeuwig MH, Narum SR, Taper ML (2008) Stream trees: a statistical method for mapping genetic differences between populations of freshwater organisms to the sections of streams that connect them. *Canadian Journal of Fishery and Aquatic Sciences*, **65**, 2752–2760.

Kimura M (1953) ''Stepping-stone'' model of population. *Annual Report of the National Institute of Genetics Japan*, **3**, 62–63.

Kimura M, Weiss G (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–576.

Lantuéjoul C (2002) *Geostatistical Simulation*. Springer Verlag, Berlin.

Lao O, Lu T, Nothnagel M *et al.* (2008) Correlation between genetic and geographic structure in Europe. *Current Biology*, **18**, 1241–1248.

Latch E, Dharmarajan G, Glaubitz J, Rhodes O, Jr (2006) Relative performance of Bayesian clustering softwares for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.

Lawson-Handley L, Manica A, Goudet J, Balloux F (2007) Going the distance: human population genetics in a clinal world. *Trends in Genetics*, **23**, 432–439.

Le Corre V, Roussel G, Zanetto A, Kremer A (1998) Geographical structure of gene diversity in *Quercus petraea* (matt.) liebl. III. Patterns of variation identified by geostatistical analyses. *Heredity*, **80**, 464–473.

Leblois R (2004) *Estimation de paramètres de dispersion en populations structurées à partir de données génétiques*. PhD thesis, Montpellier SupAgro.

Leblois R, Estoup A, Rousset F (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a 'continuous' population under isolation by distance. *Molecular Biology and Evolution*, **20**, 491–502.

Leblois R, Estoup A, Rousset F (2004) Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, **166**, 1081–1092.

Leblois R, Estoup A, Streiff R (2006) Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? *Molecular Ecology*, **15**, 3601–3615.

Leblois R, Estoup A, Rousset F (2009) IBDSIM: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, **9**, 107–109.

Lecis R, Ferrando A, Ruiz-Olmo I, Manas S, Domingo-Roura X (2008) Population genetic structure and distribution of introduced American mink (*Mustela vison*) in Spain, based on microsatellite variation. *Conservation Genetics*, **9**, 1149–1161.

Lewontin R (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Liang S, Banerjee S, Carlin B (2008) Bayesian Wombling for spatial point processes. *Biometrics*, doi:10.1111/j.1541–0420-2009-01203.x.

Loiseau L, Richard M, Garnier S *et al.* (2009) Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology*, **18**, 1331–1340.

Møller J, Stoyan D (2009) Stochastic geometry and random tessellations. In: *Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings* (eds van de Weijgaert R, Vegter G, Icke V, Ritzerveld J). Springer-Verlag, Berlin.

Møller-Hansen M, Hemmer-Hansen J (2007) Landscape genetics goes to sea. *Journal of Biology*, **6**, doi:10.1186/jbiol59.

Malécot G (1948) *Les mathématiques de l'hérédité*. Masson, Paris.

Malécot G (1950) Quelques schémas probabilistes sur la variabilité des populations naturelles. *Annales de l'Université de Lyon A*, **13**, 37–60.

Malécot G (1967) Identical loci and relationship. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4 (eds Le Cam LM, Neyman J), pp. 317–332. University of California Press, Berkeley, California.

Malécot G (1975) Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology*, **8**, 212–241.

Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**, 189–197.

Manel S, Berthoud F, Bellemain E *et al.* (2007) A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Molecular Ecology*, **10**, 2031–2043.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Martel C, Réjasse F, Rousset F, Bethenod M, Bourguet D (2003) Host-plant-associated genetic differentiation in Northern French populations of the Eurpean corn borer. *Heredity*, **90**, 141–149.

Maruyama T (1972) Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, **70**, 639–651.

Matheron G (1987) Suffit-il pour une covariance d'être de type positif? *Sciences de la terre, série informatique géologique*, **26**, 51–66.

Matheron G (1993) Une conjecture sur la covariance d'un ensemble aléatoire. In *Cahiers de Géostatistique*, Vol. 3, pp. 107–113. Centre de Géostatistique de l'École des Mines de Paris, Fontainebleau.

McRae B (2006) Isolation by resistance. *Evolution*, **60**, 1551–1561.

McRae B, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, **104**, 19885–19890.

McRae B, Beier P, Huynh L, DeWald L, Keim P (2005) Habitat barriers limit gene flow and illuminate historical events in a wide ranging carnivore, the *American puma*. *Molecular Ecology*, **14**, 1965–1977.

Monaghan M, Spaak P, Robinson C, Ward J (2001) Genetic differentiation of baetis alpinus pictet *Ephemeroptera baetidae* in fragmented alpine streams. *Heredity*, **86**, 395–403.

Monestiez P, Goulard M, Charmet G (1994) Geostatistics for spatial genetic structures: study of wild population of wild perennial ryegrass. *Theoretical and Applied Genetics*, **88**, 33–41.

Monmonier M (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis*, **5**, 245–261.

Nagylaki T (1974) The decay of genetic variability in geographically structured populations. *Proceedings of the National Academy of Sciences*, **71**, 2932–2936.

Nagylaki T (1976) The decay of genetic variability in geographically structured populations. II. *Theoretical Population Biology*, **10**, 70–82.

Nagylaki T (1989) Gustave malécot and the transition from classical to modern population genetics. *Genetics*, **122**, 253–268.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Novembre J, Johnson T, Bryc K *et al.* (2008) Genes mirror geography within Europe. *Nature*, **456**, 98–101.

Olsen J, Miller S, Spearman W, Wenburg J (2003) Patterns of intra- and inter-population genetic diversity in Alaskan coho salmon: implications for conservation. *Conservation Genetics*, **4**, 557–569.

Orsini L, Corander J, Alasentie A, Hanski I (2008) Genetic spatial structure in a butterfly metapopulation correlates better with past than present demographic structure. *Molecular Ecology*, **17**, 2629–2642.

Patterson N, Price A, Reich D (2006) Population structure and eigen analysis. *PLoS Genetics*, **2**, 2074–2093.

Pella J, Masuda M (2006) The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fishery and Aquatic Sciences*, **63**, 576–596.

Petit C, Fréville H, Mignot A *et al.* (2001) Gene flow and local adaptation in two endemic plant species. *Biological Conservation*, **100**, 21–34.

Pilot M, Jedrzejewski W, Branicki W *et al.* (2006) Ecological factors influence population genetic structure of European grey wolves. *Molecular Ecology*, **14**, 4533–4553.

Pope L, Pope BR, Wilson G *et al.* (2007) Genetic evidence that culling increases badger movement: implications for the spread of bovine tuberculosis. *Molecular Ecology*, **16**, 4919–4929.

Portnoy S, Wilson M (1993) Seed dispersal curves: behavior of the tails of the distribution. *Evolutionary Ecology*, **7**, 25–44.

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

Prugnolle F, Theron A, Pointier J *et al.* (2005) Dispersal in a parasitic worm and its two hosts: consequence for local adaptation. *Evolution*, **59**, 296–303.

Reich D, Price A, Patterson N (2008) Principal component analysis of genetic data. *Nature Genetics*, **40**, 491–492.

Rieseberg L, Vines T, Kane N (2009) Editorial and Retrospective 2008. *Molecular Ecology*, **18**, 1–20.

Riley S, Pollinger J, Sauvajot R, Bromley EY, Fuller T, Wayne R (2006) A southern California freeway is a physical and social

barrier to gene flow in carnivores. *Molecular Ecology*, **15**, 1733–1741.

Ripley B (1991) The use of spatial models as image priors. In: *Spatial Statistics and Imaging: Papers from the Research Conference on Image Analysis and Spatial Statistics held at Bowdoin College, Brunswick, Maine, Summer 1988, 426 pp* (ed. Possolo A), Lecture Notes–Monograph Series, Volume 20, pp. 309–340. Institute of Mathematical Statistics, Hayward, California.

Robledo-Arnuncio JJ, Rousset F (2009) Isolation by distance in a continuous population under stochastic demographic fluctuations. *Journal of Evolutionary Biology*, in press.

Rollins L, Woolnough A, Sherwin W (2006) Population genetic tools for pest management: a review. *Wildlife Research*, **33**, 251–261.

Rosenberg N, Burke T, Elo K *et al.* (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, **159**, 699–713.

Rosenberg N, Saurabh S, Ramachandran S, Zhao C, Pritchard J, Feldman M (2005) Clines, clusters, and the effect of study design on the influence of human population structure. *Public Library of Science, Genetics*, **1**, 660–671.

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. *Journal of Evolutionary Biology*, **13**, 58–62.

Rousset F (2004) *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton, New Jersey.

Rousset F (2008) Demystifying Moran's *I*. *Heredity*, **100**, 231–232.

Rousset F, Leblois R (2007) Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Molecular Biology and Evolution*, **24**, 2730–2745.

Rowe G, Beebee T (2007) Defining population boundaries: use of three Bayesian approaches with the microsatellite data from British natterjack toads (*Buffo calamita*). *Molecular Ecology*, **16**, 795–796.

Sacks B, Bannasch DL, Chomel BB, Ernst H (2008) Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. *Molecular Biology and Evolution*, **25**, 1354–1395.

Sahlsten J, Thörngren H, Höglund J (2008) Inference of hazel grouse population structure using multilocus data: a landscape genetic approach. *Heredity*, **101**, 475–482.

Schlather M, Ribeiro P, Diggle P (2004) Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society, Series B*, **66**, 79–93.

Schwartz M, McKelvey K (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

Selkoe K, Henzler C, Gaines S (2008) Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries*, **9**, 363–377.

Shimatani K (2002) Point processes for fine-scale spatial genetics and molecular ecology. *Biometrical Journal*, **44**, 323–332.

Shimatani K (2004) Spatial molecular ecological models for genotyped adults and offspring. *Ecological Modelling*, **174**, 401–410.

Shimatani K, Takahashi M (2003) On methods of spatial analysis for genotyped individuals. *Heredity*, **91**, 173–180.

Sisson SA, Fan Y (2009) Approximate Bayesian computation. *ISBA Bulletin*, **16**, 6–8.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.

Slatkin M (1989) Population structure and evolutionary progress. *Genome*, **31**, 196–202.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Sokal RR, Rohlf FJ (1995) *Biometry*. Freeman, San Francisco, California.

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

Stenseth N, Shabbar A, Chan K *et al.* (2004) Snow conditions may create an invisible barrier for lynx. *Proceedings of the National Academy of Sciences*, **101**, 10632–10634.

Storfer A, Murphy M *et al* (2007), Putting the ''landscape'' in landscape genetics. *Heredity*, **98**, 128–142.

Stoyan D, Kendall W, Mecke J (1995) *Stochastic Geometry and its Applications*. John Wiley and Sons, Chichester.

Strand A (2002) METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Molecular Ecology Notes*, **2**, 373–376.

Su H, Qu LHK, Zhang Z, Wang J, Chen Z, Gu H (2003) The Great Wall of China: a physical barrier to gene flow? *Heredity*, **90**, 212–219.

Sumner J, Estoup A, Rousset F, Moritz C (2001) Neighborhood size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology*, **10**, 1917–1927.

Thompson L, van Manen F, King T (2005) Geostatistical analysis of allele presence patterns among American black bears in eastern North Carolina. *Ursus*, **16**, 59–69.

Vekemans X, Hardy O (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, **13**, 921–935.

Vounatsou P, Smith T, Gelfand A (2000) Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics*, **1**, 177–189.

Wagner HH, Holderegger R, Werth S, Gugerli F, Hoebee S, Scheidegger C (2005) Variogram analysis of the spatial genetic structure of continuous populations using multilocus microsatellite data. *Genetics*, **169**, 1739–1752.

Waples R, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for indentifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wasser S, Shedlock A, Comstock K, Ostrander E, Mutayoba B, Stephens M (2004) Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences*, **101**, 14847–14852.

Wasser S, Mailand C, Booth R, Mutayoba B, Kisamo E, Stephens M (2007) Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences*, **104**, 4228–4233.

Watts P, Rousset F, Saccheri I, Leblois R, Kemp S, Thompson D (2007) Compatible genetic and ecological estimated of dispersal rates in insect (*Coeagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology*, **16**, 737–751.

Whitlock M, McCauley C (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. *Heredity*, **82**, 117–125.

Winters J, Waser P (2003) Gene dispersal and outbreeding in a philopatric mammal. *Molecular Ecology*, **12**, 2251–2259.

Womble W (1951) Differential systematics. *Science*, **28**, 315–322.

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

Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39–59.

Yamamoto S, Morita K, Koizumi I, Maekawa K (2005) Genetic differentiation of white-spotted charr *Salvelinus leucomaenis* populations after habitat fragmentation: spatial–temporal changes in gene frequencies. *Conservation Genetics*, **5**, 529–538.

Zannèse A, Morellet N, Targhetta C *et al.* (2006) Spatial structure of roe deer populations: towards defining management units at a landscape scale. *Journal of Applied Ecology*, **43**, 1087–1097.

Supporting information

Additional supporting information may be found in the online version of this article:

Table S1 List of main computer programs for spatial statistical analysis and spatially explicit simulation of genetic data

Fig. S1 Examples of simulated spatial patterns for population membership obtained by a constrained Voronoi tiling (induced by the sampling sites) and a Markov random field colouring.

Fig. S2 Same as Fig. S1 with irregularly spaced sampling sites.

Fig. S3 Modal assignments of individual coyotes to different clusters using the (a) STRUCTURE algorithm ($K = 6$) and the (b) spatially explicit GENELAND ($K = 8$) algorithm.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.