MOLECULAR ECOLOGY RESOURCES

THE STATE OF THE FIELD | Full Access

Spatially explicit Bayesian clustering models in population genetics

OLIVIER FRANÇOIS, ERIC DURAND

First published: 16 August 2010 | https://doi.org/10.1111/j.1755-0998.2010.02868.x | VIEW METRICS

Check for full text

Olivier François, Fax: +33 (0)456 520 044; E-mail: olivier.francois@imag.fr

Abstract

This article reviews recent developments in Bayesian algorithms that explicitly include geographical information in the inference of population structure. Current models substantially differ in their prior distributions and background assumptions, falling into two broad categories: models with or without admixture. To aid users of this new generation of spatially explicit programs, we clarify the assumptions underlying the models, and we test these models in situations where their assumptions are not met. We show that models without admixture are not robust to the inclusion of admixed individuals in the sample, thus providing an incorrect assessment of population genetic structure in many cases. In contrast, admixture models are robust to an absence of admixture in the sample. We also give statistical and conceptual reasons why data should be explored using spatially explicit models that include admixture.

Introduction

Statistical methods that can describe and quantify geographic patterns of intraspecific genetic variation are essential to many types of researcher (Endler 1977; Cavalli-Sforza *et al.* 1994; Avise 2000). Inference about population genetic structure accelerated around the 1960s with the introduction of principal component analysis (PCA) and tree-based clustering algorithms (Edwards & Cavalli-Sforza 1964; Cavalli-Sforza & Edwards 1965). Those algorithms are descriptive methods making no assumptions about the biological processes that generated the data. More recently, the Bayesian revolution that has occurred in population genetics has changed our ways of making such inferences (Beaumont & Rannala 2004). The Bayesian paradigm has fostered the emergence of several new model-based parametric methods, the most influential of these being implemented in the computer program STRUCTURE (Pritchard *et al.* 2000b). STRUCTURE uses multilocus genotype data to describe population genetic structure. The method differs from other statistical procedures for estimating genetic subdivision, such as *F*-statistics or the analysis of molecular variance that quantify the divergence among predefined subpopulations (Wright 1951; Excoffier *et al.* 1992). Instead, STRUCTURE assumes that there are K (K is unknown) clusters, each of which is characterized by a set of allele frequencies at each locus. Since its publication in 2000, many modifications of the original models have been proposed. These modifications include the presence of genetic linkage (Falush *et al.* 2003; Hoggart *et al.* 2004; Corander & Tang 2007), inbreeding (François *et al.* 2006; Gao *et al.* 2007), migration (Zhang 2008), mutation (Shringarpure & Xing 2009), dominance (Falush *et al.* 2007; see Bonin *et al.* 2007), automate the choice of the number of cluster (Dawson & Belkhir 2001; Corander *et al.* 2003; Pella & Masuda 2006; Huelsenbeck & Andolfatto 2007) and speed up the inference algorithm (Tang *et al.* 2005; Chen *et al.* 2006; Wu *et al.* 2006; Alexander *et al.* 2009).

An important class of Bayesian clustering models improves STRUCTURE by including information on individual geographic coordinates. These models are currently implemented in the computer programs GENELAND (Guillot *et al.* 2005), TESS (Chen *et al.*

Volume 10, Issue 5
Special Issue: Advances in the Analysis of Spatial Genetic Data
September 2010
Pages 773-784

Figures | References | Related | Information

2007; Durand *et al.* 2009) and BAPS5 (Corander *et al.* 2008). Although the programs are targeted at similar goals, they rely on models that substantially differ in their background hypotheses. A recent review (Guillot *et al.* 2009) has referred to some of these methods in the context of a more general overview of the whole field of spatial statistical genetics. Our objective here is different because we aim at providing a much more accurate discussion of Bayesian clustering methods, which represent a rapidly developing area of research. More specifically, the objective of this review is to clarify the assumptions underlying spatially explicit Bayesian clustering models, to test their robustness to departures from their primary assumptions and to aid users in interpreting their program outputs.

Population structure and Bayesian clustering

Genetic structures and spatial scales

Spatially explicit Bayesian models address three major types of genetic structures that can appear, possibly at different geographical scales: genetic clusters, clines and patterns of isolation-by-distance. The first type, genetic clusters, can be viewed as genetically divergent groups of individuals that arise when gene flow is impeded by physical or behavioural obstacles. The term 'cline' is used in population genetics to refer to a large-scale spatial trend in allele frequencies or genetic diversity (Hartl & Clark 1997). Clines in allele frequencies may be the consequence of adaptation along an environmental gradient (Berry & Kreitman 1993), or of genetic admixture occurring in secondary contact zones (Barton & Hewitt 1985). The third category, described by Wright (1943) as isolation-by-distance, is the pattern of local genetic differences that can accumulate under geographically restricted dispersal. A classical model of isolation-by-distance is the equilibrium stepping-stone model in which regularly spaced subpopulations exchange migrants locally (Malécot 1948; Kimura & Weiss 1964). The equilibrium model implies a decrease in genetic correlation with distance, a phenomenon that also occurs in nonequilibrium populations (Slatkin 1993).

Clines, clusters and patterns of isolation-by-distance are not mutually exclusive genetic structures. A classic example of co-occurrence of these patterns is the internal genetic structure of the Yanomama, a tribal population from Venezuela and Northern Brazil (Ward 1972; Ward & Neel 1976; Smouse & Long 1992). The tribe is hierarchically organized in villages and dialect clusters, and several polymorphic loci show clinal variation within the tribe in distinct spatial directions. The proposed interpretation of these patterns is that those clines and clusters are the results of centrifugal range expansion at an earlier stage of the history of the tribe. Other examples of coexistence of the three geographical patterns are found in ring species (Irwin *et al.* 2005). In a ring species, two reproductively isolated forms are connected by a chain of intermediate subpopulations that encircle a geographic barrier. Isolation-by-distance and selection against hybrids can lead to well-differentiated genetic clusters that may be separated by a cline at the closure of the ring (Bensch *et al.* 2009).

Bayesian clustering

One explanation of the great popularity of STRUCTURE in evolutionary applications is its ability to provide a description of clines and clusters by making use of multilocus genotypes obtained from a sample of individuals. In this perspective, isolation-by-distance may be viewed as an ubiquitous phenomenon that complicates the analysis. Under its generic name, the program includes many distinct models that fall into two broad categories: models with or without admixture.

The models without admixture assume that the sample is a mixture of K diverging subpopulations. Individuals are then probabilistically assigned to the K genetic clusters. The existence of clusters, and the cluster to which each individual belongs, can be assessed because population structure generates a Wahlund effect (an excess of homozogosity over the Hardy–Weinberg expectations for a single population) and linkage disequilibrium. These effects mean that the likelihood is larger if individuals are correctly assigned to subpopulations.

In contrast, the admixture models suppose that the data originate from the admixture of K putative parental populations that may be unavailable to the study. The K parental populations may have existed at unknown times in the past. In these models, the parameters of interest are the ancestry coefficients, also termed admixture proportions, computed for each individual in the sample. These coefficients are stored in a matrix, Q, which elements, q_{ik}, represent the proportion of individual is genome that

originates from the parental population *k*. The most often used option of STRUCTURE implements a variant of the admixture model with correlated allele frequencies (Falush *et al.* 2003). In the correlated allele frequency model, allele frequencies in parental populations have drifted away from frequencies in an ancestral population (Balding & Nichols 1995). In addition to enabling inference of population structure, the *Q* matrix is fundamental for correcting stratification in genome-wide association studies, one of its primary target (Pritchard *et al.* 2000a).

More specifically, the models of STRUCTURE describe the joint probability distribution of the data (the multilocus genotypes) and the parameters, which include all allele frequencies, latent clusters for each individual (without admixture) or allele (with admixture), and admixture proportions. The joint probability distribution decomposes into the product of two terms: the likelihood, a quantity that describes the probability of the data conditional on the parameters, and the prior distribution, which summarizes background information about the parameters. Posterior estimates for the parameters of interest are computed by updating the prior distribution based on the data using a Markov chain Monte Carlo (MCMC) algorithm. Spatial models, which will be described below, adopt the same individual-based likelihood framework, but they rely on very different priors distributions. The spatial clustering models fall into the same two categories as those implemented in STRUCTURE: with or without admixture. We present five distinct spatial Bayesian individual-based clustering models implemented in three software packages. While this review is focused on individual-based methods, we should mention that population-based methods, based on similar Bayesian principles, can also include spatial covariates in their prior distributions (Foll & Gaggiotti 2006; Faubet & Gaggiotti 2008; Orsini *et al.* 2008).

Preliminary clarifications

Before describing the Bayesian clustering approaches, it is useful to make a number of preliminary remarks. (i) Each program name hides a plethora of distinct models. For example, STRUCTURE encompasses (many) more than 16 different models depending on the choice of the admixture model (Pritchard *et al.* 2000b), the linkage model (Falush *et al.* 2003), the dominance model (Falush *et al.* 2007) or the use of population information (Hubisz *et al.* 2009). This means that we should clearly indicate which model we use in addition to the program. Here, unless mentioned, we refer to the default options of each program. (ii) A second distinction arises because of the release of upgraded versions of programs and program documentation. Consequently, comparisons of programs are only valid on short time scales, and references to program documentation may be more accurate than references to original publications. Our objective here is not to compare the relative performance of the models. For such comparisons, see (Latch *et al.* 2006; Chen *et al.* 2007). (iii) Some essential postprocessing methods have been developed separately from the models themselves. Examples include model selection methods to decide which number of cluster should be retained and utilities that average results over multiple program runs (CLUMPP, Jakobsson & Rosenberg 2007).

Spatially explicit Bayesian models

Models without admixture

The no-admixture model implemented in BAPS5 defines the neighbourhood of each individual based on a Voronoi tessellation of the study area (Corander *et al.* 2008). In a Voronoi tessellation, each individual sampling site, s_i, is surrounded by a cell made of points that are closer to s_i than to any other sampling site. The use of the sampling locations to define cells is natural unless the sampling locations are unrepresentative of the individual spatial distribution (François *et al.* 2006). Two sampling sites are neighbours if their cells share a common edge. BAPS5 imposes a prior distribution that puts more weight on geographically homogeneous partitions of the sample, although this geographic trend is not explicitly based on biological considerations. In the BAPS5 model, the important property of the prior distribution is to belong to the class of Markov random fields (Clifford 1990; Bishop 2006; François *et al.* 2006). In Markov field models, neighbouring individuals are more likely to be co-assigned to a cluster than individuals far apart. In addition, the correlation between cluster labels decreases with the distance between sampling sites, as expected under spatially restricted dispersal (Kimura & Weiss 1964). However, although the Markov property accounts for local dependencies, the model does not allow us to estimate the magnitude and scale of spatial correlations in presence of patterns of isolation-by-distance. The model without admixture is implemented through a greedy stochastic split and merge algorithm. The

algorithm is faster and requires less tuning than MCMC algorithms (Corander *et al.* 2008). In practice, the only parameter, a user of BAPS5 needs to tune, is the maximal number of cluster, K_{max}, to be explored by the program (but many other options are available).

In the without-admixture model of TESS, the prior distribution on cluster labels is similar to the model used in BAPS5 (Chen *et al.* 2007). Again, TESS builds a neighbourhood for each individual based on a Voronoi tessellation where each cell is centred on a sampled site. The prior distribution on cluster labels corresponds to the Potts model, which is widely used in epidemiology, image analysis and statistical physics. The Potts model is a statistical model where the state of each individual is influenced only by the states of its neighbours. In other words, neighbouring individuals are genetically closer to each other than to distant individuals (but again, the model does not allow us to infer the magnitude of spatial correlations in presence of isolation-by-distance). If sampling is geographically irregular, TESS can use a modified version of the Potts model in which the Delaunay graph is weighted by an inverse function of geometric distance so that long edges in the graph have virtually no influence.

In contrast, the prior distribution on cluster labels implemented in GENELAND is based on a biologically motivated probabilistic model inspired by landscape genetics (Manel *et al.* 2003; Guillot *et al.* 2005). GENELAND attempts to detect genetic boundaries, considering that these boundaries separate *K* random mating subpopulations. So it differs from BAPS and TESS, which instead attempt to minimize the Walhund effect by incorporating local dependencies in their model. Unlike BAPS5 or TESS, Voronoi cells in GENELAND are not associated with individuals, but with 'territories'. Each territory can group several individuals within a single Voronoi cell. The geographic locations of the cells as well as their number are considered as parameters of the model and are estimated using an MCMC algorithm. The number of cells is controlled by a fixed parameter (λ) that influences both the posterior estimates and the convergence rate of the algorithm.

Models with admixture

The admixture model of BAPS5 searches for admixture events given some source populations (Corander & Marttinen 2006). The model assumes that every source population has been sampled before inferring potential admixture events. Thus, the admixture model is by itself not spatially explicit. Note that only if the admixture event was recent are the parental populations or closely related populations likely to be sampled. With this assumption in mind, Corander & Marttinen (2006) recommend starting the analysis by partitioning the sampled individuals using their mixture model. To compute admixture proportions, BAPS5 runs a Monte Carlo algorithm over allele frequency parameters while using an optimization algorithm for inferring the admixture coefficients at each iteration.

TESS implements a spatially explicit admixture model that does not require that the source populations have been sampled (Durand *et al.* 2009). Individual ancestry proportions are estimated by incorporating spatial trends and spatial autocorrelation in the prior distribution of the *Q* matrix. In TESS, individual admixture proportions are allowed to vary over space, and the variation is decomposed into effects at both regional and local scales. In this model, trend surfaces account for clines in all geographic directions, and autocorrelated residuals account for isolation-by-distance. More importantly, the parameters that specify the shape of the clines are also estimated from the data, as well as the magnitude of spatial autocorrelation (using noninformative prior distributions on those hyper-parameters). The models – implemented in an MCMC algorithm – have the potential to simultaneously detect clines and clusters by examining the inferred variation of admixture proportions. Table 1 summarizes the main features of the computer programs discussed in this study.

Table 1. Summary of four Bayesian clustering software packages and their underlying model

Software	Admixture	Parental populations	Rationale	Prior distribution	Algorithm	Choice of K
STRUCTURE	Yes	Not required	Infers ancestry coefficients Minimizes departures from HW and LD disequilibria	Non spatial	MCMC	Multiple runs ln $P(D \mid K)$

Software	Admixture	Parental populations	Rationale	Prior distribution	Algorithm	Choice of K
GENELAND	No	Not relevant	Delineates populations under Hardy–Weinberg equilibrium	Colored Voronoi tiling	RJMCMC	Single run Reversible jump
BAPS	Yes	Required or provided by the mixture model	Seeks spatially smooth and genetically homogeneous clusters	Inspired from Markov Random fields (mixture) Non spatial (admixture)	Stochastic optimization	Multiple runs Split and merge
TESS	Yes	Not required	Models spatial trends and autocorrelation	Markov random field (mixture) Log-Gaussian random field (admixture)	MCMC	Multiple runs Information theoretic criterion

MCMC, Markov chain Monte Carlo.

Choosing the number of clusters

Distinct approaches have been proposed to estimate the number of cluster in each model. To avoid errors and misuses, it is important to remark that K_{max} does not have the same meaning in each program. In BAPS5 and GENELAND, K_{max} represents a bound on the number of clusters to be explored by the algorithm. In TESS (and STRUCTURE), K_{max} (like K) is a fixed value, and the models have to be run for a range of values of K_{max} (or K).

To estimate the number of clusters, STRUCTURE relies on ln $P(D|K)$: the logarithm of the probability of the data given K. This statistic can be viewed as a penalized measure of fit based on a Gaussian approximation of the model deviance. Typically, STRUCTURE is run for several values of K, ln $P(D|K)$ is computed for each of them and plotted against K. The value of K is then chosen to correspond to the point at which the curve plateaus. The ΔK criterion of Evanno *et al.* (2005) aims to automate this process.

To decide which values of the number of clusters are best supported by the genetic data, GENELAND estimates the posterior probabilities of each K via a reversible jump algorithm. The algorithm visits each value of K between 1 and K_{max} within a single long run. It can increase or decrease K of one unit by splitting an existing cluster or by merging two existing clusters. Similarly, BAPS5 implements a split and merge algorithm that allows K to be automatically estimated. Because, in large dimensions, the posterior distribution is likely to be multimodal, the MCMC algorithms may visit only the vicinity of a few modes of the posterior distribution. Thus, it may be necessary to run those programs several times. TESS computes the deviance information criterion (DIC) to choose K (Spiegelhalter *et al.* 2002): a generalization of the Akaike information criterion for hierarchical models (Akaike 1974). DIC is a measure of model fit penalized by an estimate of model complexity. In practice, this criterion is similar to that used by STRUCTURE because ln $P(D|K)$, up to a factor ½, has been proposed as an alternative to the calculation of the DIC (Gelman *et al.* 2003). To choose K_{max} (and K) or any internal model, TESS can be run for distinct values of K_{max}. In practice, we suggest plotting DIC against K_{max} and choosing the values of K_{max} that correspond to a plateau of the curve (Durand *et al.* 2009).

The number of genetic clusters detected by Bayesian clustering algorithms does not necessarily correspond to the number of biologically meaningful populations in our sample (cf. Waples & Gaggiotti 2006). For example, inference of population structure can be biased by the choice of a particular sampling strategy (Schwartz & McKelvey 2009). For PCA and STRUCTURE, the ability to

detect population structure also depends on the sample size and on the number of markers (Patterson *et al.* 2006; Fogelqvist *et al.* 2010). In particular, finer structure might be detected with a larger sample size. The choice of a particular value of *K* in Bayesian clustering models will also depend on the information contained in the data, rather than on biological grounds. We ought to be aware that when we determine an optimal value of *K*, it is optimal only for the particular model we are using. Because the models differ in their prior assumptions, there is no reason why values of *K* should be congruent in every model (see Discussion). It follows that choosing *K* based on a consensus of outputs may not always be justified. Resorting to model selection criteria to choose *K* by comparing distinct models overcomes these issues.

Robustness of models

In this section, we evaluate the robustness of the models to departures from their basic assumptions, we test the Bayesian clustering programs under three distinct scenarios: (i) A scenario of recent divergence (or fission) of five subpopulations. (ii) A fusion scenario in which source populations are lost, but the relative proportions of each individual genome originating in each source population are variable across space. (iii) A spatially realistic scenario of the colonization of Europe from two refugia.

Models with admixture are robust in diverging subpopulations

In first series of simulations, we consider a simulated data set that consists of recently diverged genetic clusters (Latch *et al.* 2006). The simulation process mimics an instantaneous fission of a large reference population, such that the clusters are created by drawing a random set of founders from the reference population. One dataset is created in which F_{ST} = 2% and another in which F_{ST} = 3%. Spatial coordinates are then associated with each individual such that the individuals group into geographically coherent partially connected units (Chen *et al.* 2007). In this simulation scenario, five genetic clusters are represented in the sample and contribute to the global gene pool with no admixture.

For the data set with an F_{ST} = 3%, the models without admixture correctly inferred the number of clusters, and misclassification rates were less than 12% (Table 2). Though producing larger error rates, admixture models still exactly detected that there were five clusters in the data. As expected, BAPS5 had the most accurate model for the admixture analysis of those simulated data because its reference populations were inferred with its mixture version (no admixture). For the data set with an F_{ST} = 2%, all but one model without admixture correctly inferred the number of cluster (Table S1). The misclassification rates were lower for GENELAND and TESS than for STRUCTURE. Admixture models correctly detected five clusters, but they produced noisier estimates of ancestry proportions than for the previous data set. Overall, spatial models performed better than the nonspatial model, and thus including spatial information was beneficial to the analysis.

Table 2. Data set without admixture (five clusters, F_{st} = 0.03). Selected number of clusters (*K*), *average* value and standard deviation of missclassification rates or fraction of genome incorrectly assigned (*p*). For STRUCTURE and TESS, K_{max} was varied from 2 to 8, and for each K_{max}, 10 independent runs of 10 000 sweeps were performed (700 000 sweeps allocated to STRUCTURE and TESS). Because GENELAND and BAPS5 infer *K* automatically, their maximum number of cluster was set to K_{max} = 8, and 10 independent runs of 100 000 sweeps were performed (allocating 1 000 000 sweeps to each program)

	STRUCTURE			GENELAND			BAPS5			TESS		
	K	*P*	SD	*K*	*P*	SD	*K*	*P*	SD	*K*	*P*	SD
Without admixture	5	0.081	0.001	5	0.126	0.023	5	0.039	0.032	5	0.044	0.001
Admixture	5	0.215	0.001	N.A.			5	0.044	0.030	5	0.213	0.018

Models without admixture are not robust to fusion events

To simulate admixture, we assume two weakly differentiated parental populations (A,B) in migration-drift equilibrium, and an instantaneous admixture event. To include a spatial framework, spatial coordinates are associated to each individual in each population along a longitudinal axis. Then, the fraction of an individual's genome originating in population A is proportional to its distance to A (Durand *et al.* 2009). As a consequence, the individual coefficients of ancestry vary continuously along a longitudinal gradient (Fig. 1).

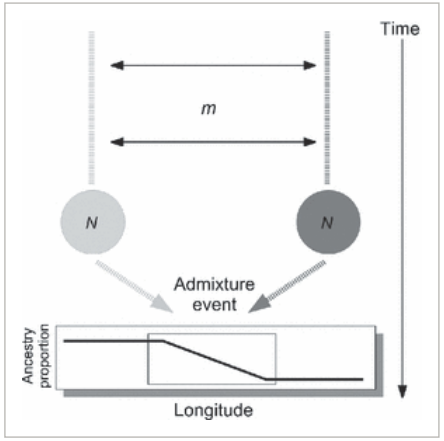

Figure 1

Open in figure viewer | PowerPoint

Schematic representation of the fusion scenario. Two weakly differentiated populations admixed in a recent past, creating a cline in allele frequencies and variable admixture proportions along a longitudinal gradient.

For these simulations, none of the models without admixture was able to uncover population structure, all leading to the inference of a single cluster in the sample. With the exception of the models of BAPS, which assumes known source populations, the situation was most favourable to admixture algorithms (Fig. 2). For an F_{ST} of 4% between the ancestral populations, both STRUCTURE and TESS admixture models performed very well. The Pearson correlation between the estimated and the true coefficients was greater than 90%. The benefit of including spatial information is visible when the ancestral level of differentiation was decreased, as STRUCTURE failed to detect the cline (Fig. 2B).

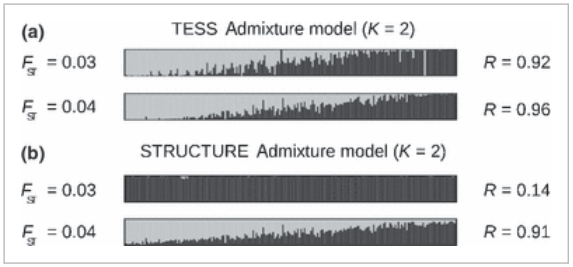

Figure 2

Open in figure viewer | PowerPoint

Inferred admixture coefficients for data sets generated under a pure fusion scenario with two parental populations. (a) Admixture model implemented in TESS. The correlation coefficient *R* between estimated and true admixture coefficients is greater than 90% for both data sets. (b) Admixture model implemented in STRUCTURE. For an F_{ST} of 3% between the parental populations, the cline is not uncovered; otherwise, the estimates are similar to TESS. The data sets contain *n* = 400 genotypes at *L* = 100 diploid loci. For each program, we performed 100 independent runs of 10 000 sweeps

with $K = 2$ and we kept the 10 runs that had the lowest DIC or lnP(D|K) values. We averaged the outputs of these 10 runs using CLUMPP.

Next, we used realistic simulations to generate data from a scenario that mimics the postglacial re-colonization of Europe for many taxa, implying the co-occurrence of clusters, clines and local patterns of isolation-by-distance in the data (Hewitt 2000). The simulation takes place in a two-dimensional nonequilibrium stepping-stone model defined on a lattice of demes covering Europe. The parameters used in this simulation are described in (Durand *et al.* 2009). In short, Europe is colonized from two distant southern refugia, one in the Iberian Peninsula and the other close to the Black Sea (Fig. 3). The simulation involves genetic divergence between parental populations (~600 generations), range expansion (during ~500 generations) and secondary contact that occurred ~500 generation ago. The simulation was performed with the program SPLATCHE (Currat *et al.* 2004). For these data, the admixture models implemented in TESS and STRUCTURE inferred $K = 3$ (Fig. 4), which corresponded to a east-west cline and one cluster that arose from a founder effect in Scandinavia. Both models detected spatial variation of ancestry coefficients in an area that unambiguously corresponds to the contact zone. In contrast, models without admixture and BAPS admixture models detected four clusters, corresponding to artificial genetic discontinuities located both sides of the contact zone.

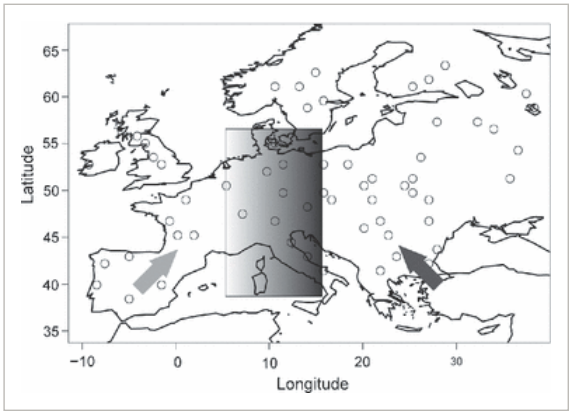

Figure 3

Schematic representation of a realistic secondary contact scenario implemented with SPLATCHE. Two waves of expansion started from two distinct southern refugia ~1000 generations ago and the two waves met in central Europe ~500 generation ago. For details on the simulation, see (Durand *et al.* 2009). Three individuals are genotyped at each of the 60 sample locations represented by empty circles (20 microsatellite loci).

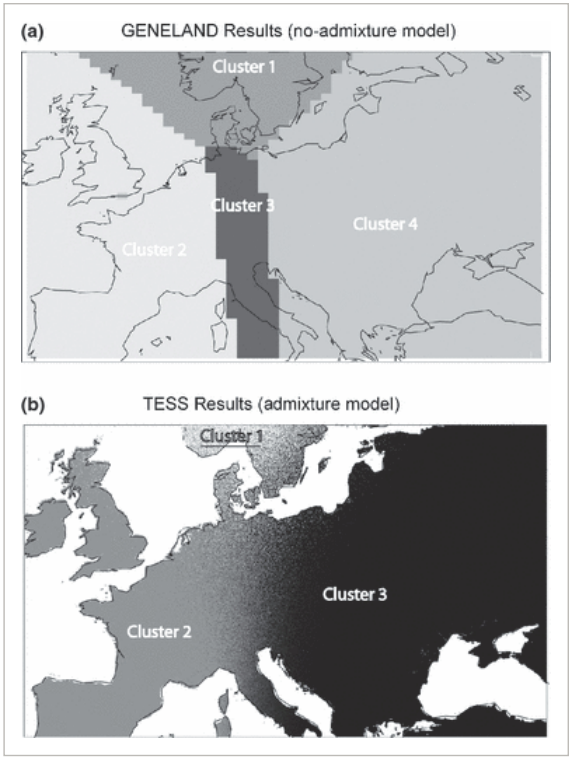

Figure 4

Secondary contact in Europe. (a) Clusters inferred by the model without admixture implemented in GENELAND. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm chooses $K = 4$ clusters. (b) Posterior prediction of admixture proportions inferred by TESS (admixture model). The DIC leads to select a model with $K = 3$ parental populations. For TESS, we varied K from 2 to 8, performing 100 runs of length 10 000 for each value of K; we kept the 10 runs that obtained the lowest DICs. For GENELAND, we performed 10 independent runs each of length 100 000 sweeps.

Discussion

Model assumptions

In models without admixture, the sample is assumed to consist of K genetically divergent groups of individuals, and the analysis uses the genetic data to classify each individual in a sample into a specific group. The models may be appropriate if we have prior knowledge on reproductive isolation or on a fragmented habitat. Thus, in applications of spatial Bayesian models, a frequent focus is on detecting genetic discontinuities associated with barriers to gene flow or habitat loss and fragmentation (e.g. Spear and Storfer 2008; Fedy *et al.* 2008; Quéméré*et al.* 2009; Gardner-Santana *et al.* 2009; Richmond *et al.* 2009; Dudgeon *et al.* 2009). In models without admixture, the allele frequencies are assumed to be constant over space within each cluster. Consequently, in the presence of clines, the sample may be either wrongly classified as a single homogeneous population as in our simulations of recent admixture or partitioned into geographic regions where the allele frequencies stay approximately constant as in Fig. 4B and in Fig. S1. In the latter case, the results of the program may confound the detection of actual boundaries.

In admixture models, individuals are not given a cluster label (Note that the terminology of 'clustering' can be misleading here). In fact the 'clusters' detected by the models are interpretable as source populations that had diverged in the past, had reached

equilibrium and could have been brought into contact again at a later date. Examples of spatial admixture analyses are found in (François *et al.* 2008; Lindsay *et al.* 2008; Cullingham *et al.* 2009; Dépraz *et al.* 2009; Henry *et al.* 2009). In admixture models, the allele frequencies are less constrained than in mixture models, because there is no assumption that there are *K* random mating populations in the sample. As a consequence, the model can detect geographic clines in allele frequencies and ancestry coefficients as in Fig. 2. Spatial models of admixture are useful in this respect, because they explicitly take the spatial dependencies into account at local and global scales and improve inference (Durand *et al.* 2009). In summary, admixture models are more flexible than models without admixture, and they may be more useful in interpreting population structure resulting from fission and fusion events and for correcting biases in association studies (Pritchard *et al.* 2000a; Falush *et al.* 2003). Furthermore, the likelihood framework of Bayesian clustering models make no explicit assumptions about the timing of divergence or admixture events.

Robustness of models

In scenarios of diverging populations, genetic groups are the results of random drift. Although we set the level of differentiation to low values, the models without admixture detected population structure accurately, and spatially explicit programs performed better where there was indeed spatial structure (Latch *et al.* 2006; Chen *et al.* 2007). Models with admixture incorrectly assigned a nonnegligible fraction of individual genomes to wrong clusters. However, the admixture models usually inferred the number of cluster correctly, and their results actually suggest that the levels of admixture in the sample were low. Not surprisingly, models without admixture failed to uncover population structure in scenarios of fusion of two weakly differentiated populations, leading to the erroneous conclusion that the sample is genetically homogeneous. When $K = 2$, the admixture models implemented in STRUCTURE and TESS revealed themselves efficient at detecting the cline, in which the allele frequencies vary along a longitudinal gradient. The failure of the admixture model of BAPS5 occurred because this model requires the presence of nonadmixed individuals in the sample; whereas in this case, no close descendant of the parental populations was sampled. Under spatially realistic scenarios, in which a species colonizes Europe from two southern refugia and exhibits a contact zone in the centre of the area, models without admixture identified a cluster in Scandinavia, but they partitioned the continental cline into three artificial compartments, producing spurious delineations that could be misinterpreted as genetic discontinuities.

Model checking and model choice

Our short simulation study does not answer two fundamental questions that might be asked of a particular data set: which models are best supported by our particular data and do spatial models provide a better description of the sample than nonspatial ones? Systematic answers to these two questions have perhaps been hindered by the hegemony of STRUCTURE in population genetic analyses. Here, we argue that it is possible to answer these questions by the techniques of model checking and model choice (Gelman *et al.* 2003).

One way to check an inferred population structure is by visualizing the posterior distribution on individual cluster labels or ancestry coefficients. In models with admixture, point estimates of the ancestry coefficients are routinely reported in a graphical way. In models without admixture, a graphical representation is obtained by the plotting marginal distributions of the inferred sample partition (the membership probabilities). For models without admixture, Dawson & Belkhir (2009) suggested to improve visualization of the posterior distribution on sample partitions by associating co-assignment probabilities with the height of nodes in a hierarchical clustering tree structure, a format which is easy to view and interpret. Another way to check an inferred population structure is by applying an independent inference method, like PCA, which has recently re-gained in popularity owing to its ease of use and its speed in analysing large genomic datasets. In addition, PCA can be modified to account for spatial autocorrelation (Jombart *et al.* 2008). The results of PCA can provide a useful validation of Bayesian clustering outputs in particular if admixture models are used (Patterson *et al.* 2006; McVean 2009). Model checking can also be performed by simulating replicates from the posterior predictive distribution (Gelman *et al.* 2003). In this setting, model checking is performed to test whether a previously fitted model can reproduce the observed data or not. When applying techniques of model checking, an important point to keep in mind is that models are wrong (but some are useful). For example, it is rather unlikely that any sample contains *K* random mating subpopulations as we assume in models without admixture (such an ideal partitioning of a sample into

Hardy–Weinberg equilibrium clusters can nevertheless be useful for understanding population structure). Model checking is a way to explore and understand differences between data and models, to improve them, but not to reject them. With Bayesian clustering models, posterior simulations of multilocus genotypes can be easily generated given the estimated assignment probabilities and the allele frequencies in each cluster. Hoggart *et al.* (2004) proposed that model checking can be performed by computing the percentage of variance explained by the first PCs of simulated genotypes and by comparing the distribution of these values to those computed from the data.

Although the Bayesian models considered here are based on a common likelihood framework, they make different assumptions. Thus, even though we would be able to find values of K that optimally describe our data set for each algorithm, those optimal values of K could still disagree with each other. The choice of K and more generally the decision of which models are best supported by the data can be addressed on the basis of information theoretic criteria, like the deviance information criterion (Fig. 5; François *et al.* 2008; Garcia-Gil *et al.* 2009; Keller *et al.* 2010). Like the more familiar Akaike information criterion values, lower values of DIC can be used to indicate better models in the sense that they are parsimonious and explain the data well. For example, Durand *et al.* (2009) used DIC to choose between three distinct prior distributions on ancestry coefficients for data from the killifish *Fundulus heteroclitus.* One of the tested models was nonspatial and equivalent to the uncorrelated allele frequency admixture model of STRUCTURE. According to the DIC, there were five clusters in the sample with the nonspatial model. The five clusters were checked to be almost identical to those obtained with the default options of STRUCTURE, for which the ΔK criterion also selected five clusters. A spatially explicit admixture version of TESS obtained lower DIC scores than the nonspatial model, indicating that a cline better described the data than the five clusters inferred by STRUCTURE. Where models have similar levels of support, model averaging – using for example the program CLUMPP (Jakobsson & Rosenberg 2007) – can also produce robust estimates of cluster membership or ancestry coefficients.

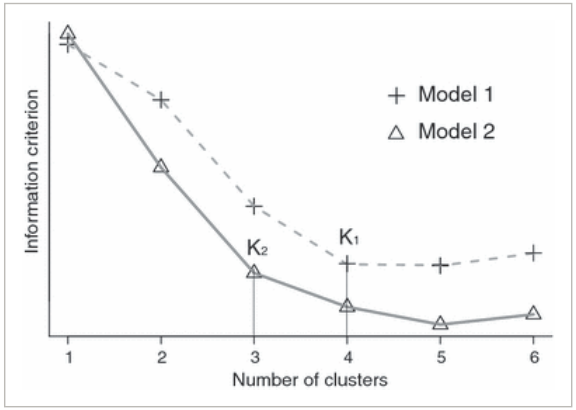

Figure 5

Open in figure viewer | PowerPoint

Choice of K and model selection based on an information theoretic criterion (DIC or a variant). In Model 1, the values of the criterion plateaus at $K_1 = 4$ whereas in Model 2, the plateau starts at $K_2 = 3$. Because the values of the criterion are smaller in Model 2 than in Model 1, we choose Model 2 with three clusters.

Conclusions

There are many cases where the inference of population structure can benefit from the modelling of the various geographic scales at which spatial genetic variation arises. Models can account for local dispersal that generates patches of covarying allele frequencies by including spatial autocorrelation (Epperson & Li 1996). In addition, they can also account for global trends in allele frequencies and admixture proportions created by range expansions and secondary contact at regional scales (Durand *et al.* 2009). Answering the 'with or without admixture' question, we urge users of Bayesian clustering programs to run admixture models on their data, because these models are more flexible and more robust than models without admixture. We suggest running more

than one model, and using statistical model selection, for example based on information-theoretic criteria, to decide which results should be retained. We also suggest that these results may not necessarily correspond to a consensus of program outputs. Using spatial models, Lavandero *et al.* (2009) and Barr *et al.* (2008) detected biologically meaningful clusters in cases where STRUCTURE failed to detect any population structure. Yoshino *et al.* (2008) detected two clusters in their data using STRUCTURE and TESS, but this was not consistent with the other methods which returned less interpretable results. Using STRUCTURE, Sahlsten *et al.* (2008) detected a cline in Scandinavian populations of *Bonasa bonasia* which seemed more plausible than the genetic boundaries found by GENELAND. Using the spatial admixture model of TESS in *Arabidopsis thaliana*, we detected a cline of variation at the scale of Europe. For these data, STRUCTURE further stratified the cline into smaller clusters (Nordborg *et al.* 2005). As in the case of a PCA, we should keep in mind that Bayesian clustering models are tools for exploring the data (Patterson *et al.* 2006; McVean 2009; François *et al.* 2010). Because their assumptions make an obvious simplification of the biological reality and because several demographic scenarios can result in similar clustering outputs, genealogical interpretations of those outputs remain difficult. Efforts to develop improved model-based clustering methods are still necessary.

Acknowledgments

The authors warmly thank Oscar Gaggiotti and Richard Nichols. They are also grateful to Lounes Chikhi, Frederic Austerlitz and an anonymous reviewer for useful comments, and to Jukka Corander for clarifications on BAPS. OF is supported by the Agence Nationale de la Recherche grant BLAN06-3146282 MAEV and by the IXXI Institute of Complex Systems.

Supporting Information

Figure S1 Inference of membership probabilities and ancestry coefficients in a simulation of a multilocus cline. In the simulation, 200 individuals are genotyped at seven bi-allelic loci at which allele frequencies display regular logistic variation along a longitudinal axis. (A) BAPS5 used with or without its admixture model partition the cline in four clusters. (B) STRUCTURE and TESS find two clusters and the interpolated ancestry coefficients mimic the cline.

Table S1 Relative performance of algorthims for simulated data with low levels of differentiation.

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

Filename	Description
MEN_2868_sm_FigureS1.pdf 341.8 KB	Supporting info item
MEN_2868_sm_TableS1.pdf 53 KB	Supporting info item

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

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716–723.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Avise JC (2000) *Molecular Markers, Natural History and Evolution*, 2nd edn. Chapman & Hall, New York, NY.

Google Scholar | Check for full text

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Barr KR, Lindsay DL, Athrey G *et al.* (2008) Population structure in an endangered songbird: maintenance of genetic differentiation despite high vagility and significant population recovery. *Molecular Ecology*, **17**, 3628–3639.

PubMed | Web of Science® | Google Scholar | Check for full text

Barton N, Hewitt G (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

PubMed | Web of Science® | Google Scholar | Check for full text

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251–261.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Bensch S, Grahn M, Müller N *et al.* (2009) Genetic, morphological, and feather isotope variation of migratory willow warblers show gradual divergence in a ring. *Molecular Ecology*, **18**, 3087–3096.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Berry A, Kreitman M (1993) Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. *Genetics*, **134**, 869–893.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Bishop CM (2006) *Pattern Recognition and Machine Learning*. Springer, New-York.

Google Scholar | Check for full text

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology*, **16**, 3737–3758.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Cavalli-Sforza LL, Edwards AWF (1965) Analysis of human evolution. In: *Genetics Today*. Proceedings of the XI International Congress of Genetics, The Hague, The Netherlands, September, 1963 (ed SJ Geerts), vol. **3**, pp. 923–933. Pergamon Press, Oxford.

Google Scholar | Check for full text

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.

Google Scholar | Check for full text

Chen C, Forbes F, François O (2006) FASTRUCT: model-based clustering made faster. *Molecular Ecology Notes*, **6**, 980–984.

Web of Science® | Google Scholar | Check for full text

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

Web of Science® | Google Scholar | Check for full text

Clifford P (1990) Markov random fields in statistics. In: *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley* (eds GR Grimmett, DJA Welsh), pp. 19–32. Oxford University Press, Oxford, UK.

Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

Web of Science® | Google Scholar | Check for full text

Cullingham CI, Kyle CJ, Pond BA *et al.* (2009) Differential permeability of rivers to raccoon gene flow corresponds to rabies incidence in Ontario, Canada. *Molecular Ecology*, **18**, 43–53.

PubMed | Web of Science® | Google Scholar | Check for full text

Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.

Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Dépraz A, Hausser J, Pfenninger M (2009) A species delimitation approach in the *Trochulus sericeus/hispidus* complex reveals two cryptic species within a sharp contact zone. *BMC Evolutionary Biology*, **9**, 171.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Dudgeon CL, Broderick D, Ovenden JR (2009) IUCN classification zones concord with, but underestimate, the population genetic structure of the zebra shark *Stegostoma fasciatum* in the Indo-West Pacific. *Molecular Ecology*, **18**, 248–261.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Durand E, Jay F, Gaggiotti OE *et al.* (2009) Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, **26**, 1963–1973.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. In: *Phenetic and Phylogenetic Classification* (eds VH Heywood, J McNeill), pp. 67–76. Systematics Association pub. no. 6, London.

Google Scholar | Check for full text

Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, New Jersey.

PubMed | Google Scholar | Check for full text

Epperson B, Li T (1996) Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10528–10532.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Faubet P, Gaggiotti OE (2008) A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics*, **178**, 1491–1504.

PubMed | Web of Science® | Google Scholar | Check for full text

Fedy BC, Martin K, Ritland C *et al.* (2008) Genetic and ecological data provide incongruent interpretations of population structure and dispersal in naturally subdivided populations of white-tailed ptarmigan (*Lagopus leucura*). *Molecular Ecology*, **17**, 1905–1917.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Fogelqvist J, Niittyvuopio A, Ågren J *et al.* (2010) Cryptic population genetic structure: the number of inferred clusters depends on sample size. *Molecular Ecology Resources*, **10**, 314–323.

PubMed | Web of Science® | Google Scholar | Check for full text

Foll M, Gaggiotti OE (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

François O, Blum MGB, Jakobsson M *et al.* (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

François O, Currat M, Ray N *et al.* (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, DOI: 10.1093/molbev/msq010.

PubMed | Web of Science® | Google Scholar | Check for full text

Gao HS, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, **176**, 1635–1651.

PubMed | Web of Science® | Google Scholar | Check for full text

Garcia-Gil MR, François O, Kamruzzahan S *et al.* (2009) Joint analysis of spatial genetic structure and inbreeding in a managed population of Scots pine. *Heredity*, **103**, 90–96.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Gardner-Santana LC, Norris DE, Fornadel CM *et al.* (2009) Commensal ecology, urban landscapes, and their influence on the genetic characteristics of city-dwelling Norway rats (*Rattus norvegicus*). *Molecular Ecology*, **18**, 2766–2778.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Gelman A, Carlin JB, Stern HS *et al.* (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Guillot G, Leblois R, Coulon A *et al.* (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734–4756.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Hartl DL, Clark AG (1997) *Principles of Population Genetics*, 3rd edn. Sinauer Associates, Inc., Sunderland, MA

Web of Science® | Google Scholar | Check for full text

Henry P, Miquelle D, Sugimoto T *et al.* (2009) In situ population structure and ex situ representation of the endangered Amur tiger. *Molecular Ecology*, **18**, 3173–3184.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Hewitt G (2000) The genetic legacy of the quaternary ice ages. *Nature*, **405**, 907–913.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Hoggart C, Shriver M, Kittles R *et al.* (2004) Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, **74**, 965–978.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Irwin DE, Bensch S, Irwin JH *et al.* (2005) Speciation by distance in a ring species. *Science*, **307**, 414–416.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Jombart T, Devillard S, Dufour AB *et al.* (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Keller SR, Olson MS, Silim S *et al.* (2010) Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera*. *Molecular Ecology*, **19**, 1212–1226.

PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

Web of Science® | Google Scholar | Check for full text

Lavandero B, Miranda M, Ramírez CC *et al.* (2009) Landscape composition modulates population genetic structure of *Eriosoma lanigerum* (Hausmann) on *Malus domestica* Borkh in central Chile. *Bulletin of Entomological Research*, **99**, 97–105.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Lindsay DL, Barr KR, Lance RF *et al.* (2008) Habitat fragmentation and genetic diversity of an endangered, migratory songbird, the golden-cheeked warbler (*Dendroica chrysoparia*). *Molecular Ecology*, **17**, 2122–2133.

PubMed | Web of Science® | Google Scholar | Check for full text

Malécot G (1948) *Les Mathématiques de l'Hérédité*. Masson, Paris.

Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**, e1000686.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

Web of Science® | Google Scholar | Check for full text

Pritchard J, Stephens M, Rosenberg NA *et al.* (2000a) Association mapping in structured populations. *The American Journal of Human Genetics*, **67**, 170–181.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Quéméré E, Louis EE Jr, Ribéron A *et al.* (2009) Non-invasive conservation genetics of the critically endangered golden-crowned sifaka (*Propithecus tattersalli*): high diversity and significant genetic differentiation over a small range. *Conservation Genetics*, DOI: 10.1007/s10592-009-9837-9.

Web of Science® | Google Scholar | Check for full text

Richmond JQ, Reid DT, Ashton KG *et al.* (2009) Delayed genetic effects of habitat fragmentation on the ecologically specialized Florida sand skink (*Plestiodon reynoldsi*). *Conservation Genetics*, **10**, 1281–1297.

Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

Web of Science® | Google Scholar | Check for full text

Shringarpure S, Xing EP (2009) mStruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, **182**, 575–593.

PubMed | Web of Science® | Google Scholar | Check for full text

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

PubMed | Web of Science® | Google Scholar | Check for full text

Smouse PE, Long JC (1992) Matrix correlation analysis in anthropology and genetics. *American Journal of Physical Anthropology*, **35**, 187–213.

Web of Science® | Google Scholar | Check for full text

Spear SF, Storfer A (2008) Landscape genetic structure of tailed frogs in protected versus managed forests. *Molecular Ecology*, **17**, 4642–4656.

PubMed | Web of Science® | Google Scholar | Check for full text

Spiegelhalter SD, Best NG, Carlin BP *et al.* (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Web of Science® | Google Scholar | Check for full text

Tang H, Peng J, Wang P *et al.* (2005) Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, **28**, 289–301.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

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

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Ward RH (1972) The genetic structure of a tribal population, the Yanomama Indians. V. Comparisons of a series of genetic networks. *Annals of Human Genetics*, **36**, 21–43.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Ward RH, Neel JV (1976) The genetic structure of a tribal population, the Yanomama Indians. XIV. Clines and their interpretation. *Genetics*, **82**, 103–121.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Wright S (1943) Isolation by distance. *Genetics*, **28**, 139–156.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Wu B, Liu N, Zhao H (2006) PSMIX: an R package for population structure inference via maximum likelihood method. *BMC Bioinformatics*, **7**, 317.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Yoshino H, Armstrong KN, Izawa M *et al.* (2008) Genetic and acoustic population structuring in the Okinawa least horseshoe bat: are intercolony acoustic differences maintained by vertical maternal transmission. *Molecular Ecology*, **17**, 4978–4991.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Zhang Y (2008) Tree-guided Bayesian inference of population structures. *Bioinformatics*, **24**, 965.

CAS | PubMed | Web of Science® | Google Scholar | Check for full text

Citing Literature