Invited Review paper

Title: Spatially Explicit Bayesian Clustering Models in Population Genetics

Authors: Olivier François* and Eric Durand*

Affiliation: *Grenoble IT, Université Joseph Fourier Grenoble, CNRS UMR 5525, TIMC-IMAG, Group of Computational and Mathematical Biology.

Corresponding author: Olivier François
Address: TIMC-IMAG, Group of Computational and Mathematical Biology, Faculty of Medicine, 38706 La Tronche France.
Tel: +33 (0)456 520 025
Fax:+33 (0)456 520 044
email: olivier.francois@imag.fr

Abstract: Geographic patterns of intraspecific genetic variation are central to many fields of evolutionary biology, molecular ecology, conservation biology and population genetics. This study 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: with or without admixture. In order to aid users 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 incorrect assessmnent of population genetic structure in many generic cases. In contrast, admixture models are robust to an absence of admixture in the sample. We also give statistical and conceptual reasons why spatially explicit models of admixture should used in addition to models without admixture.

Keywords: Spatial population structure – Spatial clustering models – Admixture – Software packages

Introduction

Statistical methods that can describe and quantify geographic patterns of intraspecific genetic variation are essential to many researchers (Endler 1977, Cavalli-Sforza et al 1994, Avise 2000). Inference about population genetic structure started around the 1960's with principal component analysis (PCA) and tree-based clustering algorithms (Cavalli-Sforza and Edwards 1965; Edwards and Cavalli-Sforza 1964). Those algorithms are descriptive methods making no assumptions about the biological processes that generated the data. Since the early epoch, the Bayesian revolution that occurred in population genetics has changed our ways to make such inferences (Beaumont and Rannala 2004). The Bayesian paradigm has fostered the emergence of several new model-based parametric methods, the most representative of which being implemented in the computer program STRUCTURE (Pritchard et al 2000). 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 assume 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 original publication, 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), inbreeding (Francois et al 2006; Gao et al 2007), migration (Zhang 2008), mutation (Shringarpure and Xing 2009), allele dominance (Falush et al 2007; see Bonin et al 2007), automate the choice of the number of cluster (Dawson and Belkhir 2001; Pella and Matsuda 2006; Huelsenbeck and Andolfatto 2007) and speed up the inference algorithm (Corander et al 2003; Tang et al 2005; Chen et al 2006; Wu et al 2006; Alexander et al 2009).

An important class of Bayesian clustering models improve 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 2007; Durand et al 2009) and BAPS5 (Corander et al 2008). Table 1 summarizes a variety of recent applications of those programs in molecular ecology, conservation genetics and evolutionary genetics. Although the programs are targeted at similar goals, they rely on models that substantially differ in their background hypotheses. 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.

Table 1. *Use of spatial clustering programs in 2008-2009 publications*

Reference	Taxon	Scale	Patterns	Programs
Garrick et al 2009	*Euphorbia lomelii* (magnoliopsida)	regional	clusters	geneland
Bizoux et al 2009	*Milicia excelsa* (magnoliopsida)	regional	ibd	tess
François et al 2008	*Arabidopsis thaliana* (magnoliopsida)	continental	cline/clusters	tess
Garcia-Gil et al 2009	*Pinus sylvestris* (pinopsida)	local	ibd	tess (geneclust)
Dvorak et al 2009	*Pinus oocarpa* (pinopsida)	regional	cluster/admixture	baps
Lavandero et al 2009	*Eriosoma lanigerum* (insect)	regional	clusters	structure baps
Orsini et al 2008	*Melitaea cinxia* (insect)	regional	clusters	baps
Holzer et al 2009	*Formica paralugubris* (insect)	local	clusters	structure geneland
Fuentes-contreiras et al 2008	*Cydia pomonella* (insect)	local	clusters/ibd	baps
Dépraz et al 2009	*Trochulus sericeus/hispidus* (gastropod)	regional	hybrid zone	tess
Spear and Storfer 2008	*Ascaphus truei* (amphibian)	regional	clusters	structure tess
Richmond et al 2009	*Plestiodon reynoldsi* (saurian)	regional	fragmentation/ibd	tess
Dudgeon et al 2009	Stegostoma fasciatum (Chondrichthyes)	regional	ibd clusters	structure tess
Galarza et al 2009	Littoral fish species	regional	clusters	geneland
McCairns and Bernatchez 2009	*Gasterosteus aculeatus* (actinopterygii)	regional	clusters	geneland
Durand et al 2009	*Fundulus heteroclitus* (actinopterygii)	continental	Cline contact zone	structure tess
Dionne et al 2008	*Salmo salar* (actinopterygii)	regional	clusters	geneland
Johansson et al 2008	*Sebastes caurinus* (actinopterygii)	local	ibd	tess
Dupont et al 2008	*Styela clava* (ascidian)	regional	dispersal clusters	structure baps geneland
Groombridge et al 2009	*Falco araea* (bird)	regional	Random mating	tess
Fedy et al 2008	*Lagopus leucura* (bird)	regional	Clusters	structure tess
Sahlsten et al 2008	*Bonasa bonasia* (bird)	regional	secondary contact	structure geneland
Coulon et al 2008	*Aphelocoma coerulescens* (bird)	regional	clusters	structure geneland
Lindsay et al 2008	*Dendroica chrysoparia* (bird)	local	admixture	structure baps tess
Barr et al 2008	*Vireo atricapilla* (bird)	local	Admixture clusters	structure baps tess
Devitt et al 2009	*Rangifer tarandus* (mammal)	regional	clusters	structure tess
Henry et al 2009	*Panthera tigris altaica* (mammal)	regional continental	clusters admixture	structure tess
Cullingham et al 2008	*Procyon lotor* (mammal)	regional	clusters admixture	structure tess
Gauffre et al 2008	*Microtus arvalis* (mammal)	local	ibd	structure tess geneland
Braaker and Heckel 2009	*Microtus arvalis* (mammal)	regional	clusters	structure geneland
Gardner-Santana et al 2009	*Rattus norvegicus* (mammal)	local	clusters fragmentation	structure geneland
Echenique-Diaz et al 2009	*Hipposideros turpis turpis* (mammal)	regional	admixture	tess
Yoshino et al 2008	*Rhinolophus cornutus pumilus* (mammal)	regional	admixture	structure geneland tess
Liu et al 2009	*Rhinopithecus bieti* (mammal)	regional	clusters fragmentation	geneland
Quéméré et al 2009	*Propithecus tattersalli* (mammal)	regional	clusters fragmentation	structure tess geneland
Guschanski et al 2008	*Gorilla Gorilla (*mammal)	regional	clusters	baps
Wang et al 2008	Humans	continental	clines clusters	structure tess
Tishkoff et al 2009	Humans	continental	clines clusters	structure tess

Population structure and Bayesian clustering

Genetic structures and spatial scales. Spatially explicit Bayesian models address three major types of genetic structures that can appear at possibly different geographical scales: genetic clusters, clines, and patterns of isolation-by-distance. Genetic clusters can be viewed as genetically divergent groups of individuals that arise when gene flow is impeded by physical or behavioral obstacles. In population genetics, the concept of cline refers to a large-scale spatial trend in the variation of allele frequencies or genetic diversity (Hartl and Clark 1997). Clines in allele frequencies may be the consequence of adaptation along an environmental gradient (Berry and Kreitman 1993), or of genetic admixture occurring in secondary contact zones (Barton and Hewitt 1985). Introduced by Wright (1943), isolation-by-distance is the accumulation of local genetic differences 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 and Weiss 1964). The equilibrium model implies a decrease of genetic correlation with distance, a phenomenon that also occurs in non-equilibrium populations (Slatkin 1993).

Clines, clusters and patterns of isolation-by-distance are not mutually exclusive genetic structures. A classic example of co-occurence of these patterns is the internal genetic structure of the Yanomama, a tribal population from Venezuela and Northern Brazil (Ward 1972; Ward and Neel 1976; Smouse and 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 for ring species (Irwin 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 (Bensh 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 at the level of each genome. Isolation-by-distance may be viewed as an ubiquituous phenomenon that complicates the analysis of genetic variation. 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 results of the mixture of K diverging subpopulations. Individuals are then probabilistically assigned to the K genetic clusters. Assignment is conducted in order to minimize the Wahlund effect, that predicts departures from Hardy-Weinberg and linkage equibrilium caused

by population substructure. 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 population may be ancestral to the sample 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 i's 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 addition to enabling inferences of population structure, the Q matrix is fundamental for correcting stratification in genome-wide association studies, one of its primary target (Pritchard et al 2000).

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 parameter, and the prior distribution which summarizes background information about the parameter. Posterior estimates for the parameters of interest are computed by updating the prior distribution based on the data and a Markov chain Monte Carlo (MCMC) algorithm. Spatial explicit programs, that will be described below, adopt the same individual-based likelihood framework, but they rely on very different priors distributions.

Spatially explicit Bayesian models

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 need to mention that population-based methods, based on similar Bayesian principles, can also include spatial covariates in their prior distributions (Foll and Gaggiotti 2006; Faubet and Gaggiotti 2008).

Preliminary clarifications. Before describing apparently related Bayesian clustering approaches, it is useful to make a number of preliminary remarks that can help to better understand the differences between the programs. 1) Each program name hides a plethora of distinct models. For example, STRUCTURE encompasses (much) more than 16 different models depending on the choice of the admixture model (Pritchard et al 2000), 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 which program we use. Here, unless mentioned, we refer to the default options of each program. 2) A second distinction is between models and their computer implementation. Computer implementations are often changing, and the changes generally lead to upgraded versions of programs and program documentations. Because of the accelerated process of successive releases, comparisons of programs are only valid on short time scales, and references to program documentations may be more accurate than references to original publications. Our objective here is not to compare the relative performances of the presented models. For such comparisons, see (Latch et al 2006; Chen et al 2007). 3) Some essential post-processing methods do not belong to the models themselves. For a particular data analysis, examples include model selection methods to decide which number of cluster should be retained, and utilities that deal with label switching and multimodality issues in averaging results over multiple program runs (Jakobsson and Rosenberg 2007).

Models without admixture. The no-admixture model implemented in BAPS5 defines the neighborhood of each individual based on a Voronoi tessellation of the study area (Corander et al 2008; François et al 2006). In this graphical representation, pairs of neighbors correspond to two adjacent cells centered on sampling sites. BAPS5 models spatial dependencies within the prior distribution of individual cluster labels, assuming that this distribution writes as a product of functions of particular subgraphs – called cliques and separators. The definition of the model, as a prior distribution that puts more weights on geographically homogeneous partitions of the sample, is purely statistical and not based on biological considerations. According to its mathematical definition, the prior of BAPS5 models the spatial autocorrelation of cluster labels, and the decrease of such correlation with distance on theVoronoi tessellation. 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 can tune is the maximal number of cluster, K_{max}, to be explored by the program.

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. Unlike BAPS5, Voronoi cells in GENELAND are not associated to individuals, but to "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 that influences both the posterior estimates and the convergence rate of the algorithm. A distinction between BAPS5 and GENELAND is that

the model assumes the presence of K Hardy-Weinberg clusters in the sample. It differsfrom the other models which make no such assumptions, and instead attempt to minimize the Walhund effect by including possible statistical sources of departure in their prior distributions. To make inferences, GENELAND implements a Reversible Jump MCMC algorithm visiting all values of K in a prescribed range from 1 to K_{max}. During this single long run, the territorial cells are split or merged to eventually delineate population boundaries.

The prior distribution on cluster labels in the without-admixture model of TESS is similar to the model used in BAPS5 (François et al 2006; Chen et al 2007). TESS builds a neighborhood for each individual based on a Voronoi tessellation where each cell is centered on a sampled individual. The prior distribution on cluster labels corresponds to a Potts model, which is widely used in epidemiology, image analysis and statistical physics (François et al 2006). The Potts model is a special case of a Markov random field, a statistical model for the spatial correlation of individual cluster labels. Markov random fields have the property that the state of each individual is influenced only by the states of its neighbors. In other words, neighboring individuals are genetically closer to each other than to distant individuals. The intensity of the spatial dependencies is controlled by an hyperparameter, ψ. The implemented value of ψ corresponds to a critical value in the Potts model, below which no spatial organization can a priori be observed and above which K spatially structured clusters can coexist. Simulations in Francois et al (2006) show, that when sampling is regular and for 2 - 6 clusters in the data, ψ is around 0.5 – 0.7. If sampling is geographically irregular, TESS can use a modified version of the Potts model in which the neighborhood graph is weighted by an inverse function of geometric distance, so that long edges in the graph have virtually no influence.

Models with admixture. Starting from an initial partition of the sampled individuals in K clusters, the admixture model of BAPS5 searches for admixture events between predefined clusters (Corander and Martinen 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 and Martinen (2006) recommend to start the analysis by partitioning the sampled individuals with their without-admixture model. To compute admixture proportions, BAPS5 runs an optimization algorithm that maximizes the posterior distribution of admixture coefficients conditional on allele frequencies estimated in the K parental populations.

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. The priors are defined as hidden regression models with autocorrelated residuals. The regression models include spatial effects both at regional and local scales using a weighted Voronoi tessellation. In this approach, the regression is part of the modeling process. Trend surfaces account for clines in all directions, and autocorrelated residuals account for isolation-by-distance. Including spatial information in the prior distribution on the admixture proportions can also provide posterior estimates that have been corrected for genealogical correlation between individuals (Durand et al 2009). 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 2 summarizes the main features of the computer programs discussed in this study.

Table 2: *Summary of 4 Bayesian clustering software packages and their model assumptions.*

Software	**Admixture Model**	**Parental populations**	**Rationale**	**Prior distribution**	**Algorithm**	**Choice of K**
STRUCTURE	Yes	Not required	Estimates admixture proportions, Minimizes departures from HW and LD disequilibria	Non-informative	MCMC	Multiple runs ln P(D|K)
GENELAND	No	Not relevant	Delineates populations under Hardy-Weinberg equilibrium	Colored Voronoi tiling	RJMCMC	Single run Reversible Jump
BAPS	Yes	Required	Seeks spatially smooth and genetically homogeneous clusters	Inspired from Markov Random Field (no admixture) Non-informative (admixture)	Stochastic optimization	Single run Split and merge
TESS	Yes	Not required	Models spatial trends and autocorrelation	Markov Random Field (no admixture) Log-Gaussian Random Field (admixture)	MCMC	Multiple runs Information theoretic criterion DIC

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} has not 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 cluster STRUCTURE relies on a statistical criterion, denoted

lnP(D|K), that computes the logarithm of the probability of the data for each run. From a statistical point of view, this criterion is a penalized measure of fit based on a Gaussian approximation of the model deviance. Typically, STRUCTURE is run for several values of K, and lnP(D|K) is computed for each run. In practice, it is recommended to plot lnP(D|K) against K, and to choose the value of K that corresponds to a plateau of the lnP(D|K) curve. The ΔK criterion of Evanno et al (2005) aims at automating 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 its 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. Although theoretically attractive, reversible jump MCMC have been criticized for having poor mixing properties in large dimensional problems (Green and Richardson, 1997)

Similarly to GENELAND, BAPS5 split and merge algorithm allows K to be automatically estimated. Each split and merge move of BAPS5 algorithm is accepted if it leads to an increase of the posterior distribution of cluster labels. In large dimensions, the posterior distribution is likely to be multimodal, and the split and merge algorithm may be stuck in local optima. Thus, it may be necessary to run the program several times.

For choosing K, TESS computes the deviance information criterion (DIC; 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. The values of K, or more generally the models that receive the most support from the data are those with the lowest values of the DIC. The computation of DIC is actually similar to the computation of lnP(D|K) in STRUCTURE. In reality, STRUCTURE uses an approximation of lnP(D|K) that, up to a factor ½, was also proposed for DIC (Gelman et al 2004). To choose K_{max} (and K) or any internal model, TESS can be run for distinct values of K_{max}. In practice, we suggest to plot the DIC against K_{max}, and choose the values of K_{max} that correspond to a plateau of the DIC curve (Durand et al 2009).

Even though we could give biologically meaningful definitions of populations (Waples and Gaggiotti 2004), the number of genetic clusters detected by Bayesian clustering algorithms does not thoroughly inform the number of such populations in our sample. For example, inference of population structure can be biased by the choice of a particular sampling strategy (Schwartz and 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 2009). Thus finer structure is expected to be detected with a larger sample size. The choice of a particular value of K in Bayesian clustering models is done on the basis of the information

contained in the data, and not 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). In addition, choosing K based on a consensus of outputs may not always be justified. Resorting to model selection criteria to choose K overcomes these issues (Johnson and Omland 2004).

Robustness of models

To evaluate the robustness of the models to departures from their basic assumptions, we test the Bayesian clustering programs under three distincts scenarios: 1) A scenario of recent divergence (or fission) of 5 subpopulations, 2) A fusion scenario in which source populations are lost, but the relative proportions of each individual genome originating in each source population is variable accross space. 3) A spatially realistic scenario of the colonization of Europe from 2 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 mimicks 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. By repeating the sampling of founder individuals, data sets with two distinct levels of genetic differentiation are created (F_{ST} = 2 – 3%). Spatial coordinates are then associated to 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.

	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

Table 3. *Data set without admixture (5 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).*

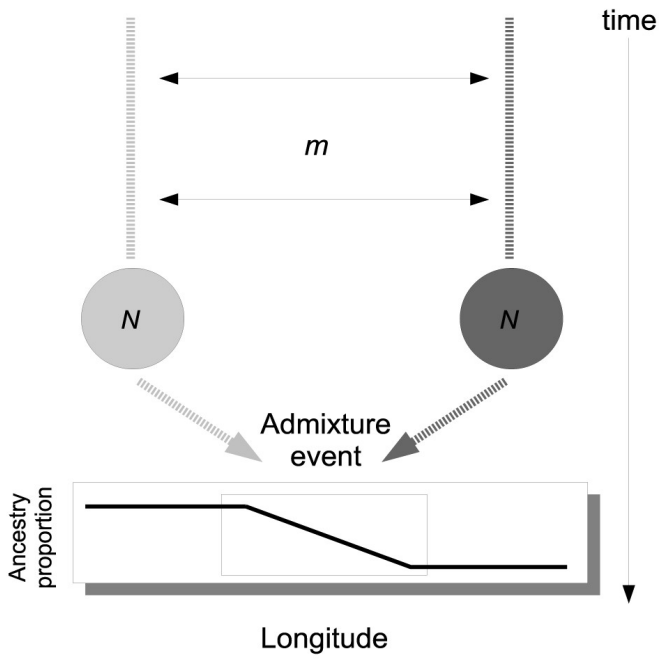

Figure 1. *Schematic representation of the fusion scenario. Two weakly differentiated populations admixed in a recent past, creating a cline in allele frequencies and variable admixture proportions along a longitudinal gradient.*

For the data set with an F_{ST} = 3%, the models without admixture correctly infer the number of cluster, and misclassification rates are less than 12% (Table 3). Though producing larger error rates, admixture models still exactly detect that there are five clusters in the data. As expected, BAPS5 has the most accurate admixture model in this simulation study because parental populations are present in the data set. For the data set with an F_{ST} = 2%, all but one models without admixture correctly infer the number of cluster (Supp Table 1). The misclassification rates are lower for GENELAND and TESS than for STRUCTURE. Admixture models still correctly detect 5 clusters, but they produce noisier estimates of ancestry proportions than for the previous data set. Overall, spatial models perform better than the aspatial model, and thus including spatial priors is beneficial to the analysis.

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 then we create an instantaneous admixture event. To include a spatial framework, we associate spatial coordinates

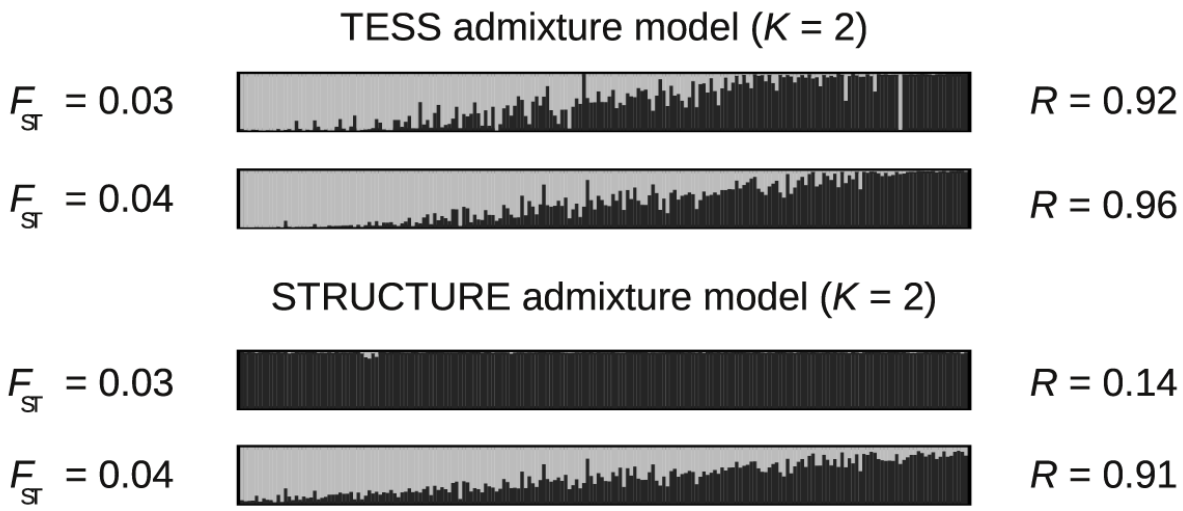

Figure 2. *Inferred admixture coefficients for data sets generated under a pure fusion scenario with two parental populations. A) Admixture model implemented in TESS. The correlation coefficient R between estimated and true admixture coefficients is greater than 90% for both data sets. B) Admixture model 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.*

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 (Figure 1).

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

Then we use realistic simulations to generate data from a scenario that mimics the post-glacial recolonization of Europe for many taxa, implying the co-occurrence of clusters, clines and local patterns of isolation-by-distance in the data (Hewitt et al 2000). The simulation takes place in a two-dimensional non-equilibrium stepping-stone model defined on a lattice of demes covering

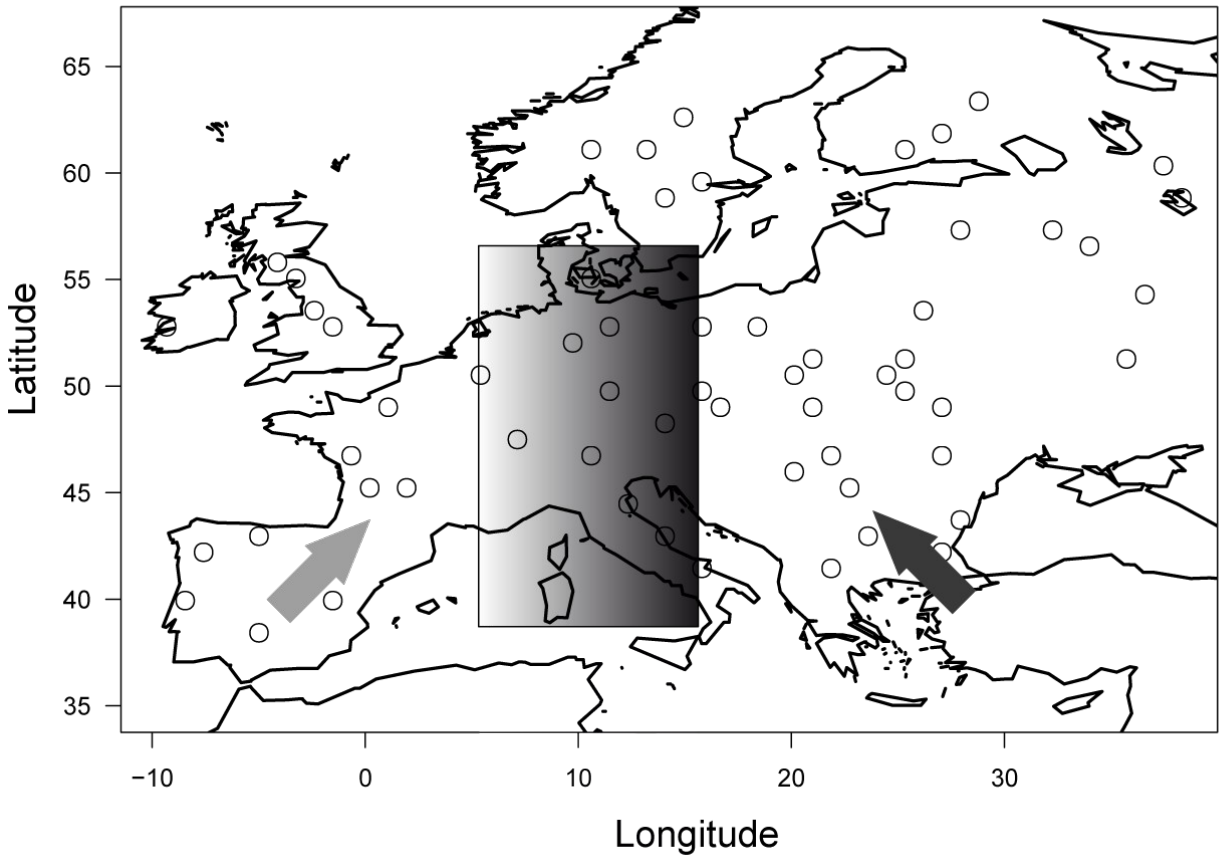

Figure 3. *Schematic representation of a realistic secondary contact scenario implemented with SPLATCHE. Two waves of expansion started from 2 distinct southern refugia ~1,000 generations ago and the two waves met in central Europe ~500 generation ago. For details on the simulation, see (Durand et al 2009). Three individuals are genotyped at each of the 60 sample locations represented by empty circles (20 microsatellite loci).*

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 (Figure 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 is performed with the program SPLATCHE (Currat et al 2004). For these data, the admixture models implemented in TESS and STRUCTURE infer K = 3 (Figure 4), which corresponds to a cline and one cluster that arose from a founder effect in Scandinavia. Both models detect spatial variation of ancestry coefficients in an area that unambiguously corresponds to the contact zone. In contrast, models without admixture detect 4 clusters, corresponding to artificial genetic discontinuities located both sides of the contact zone.

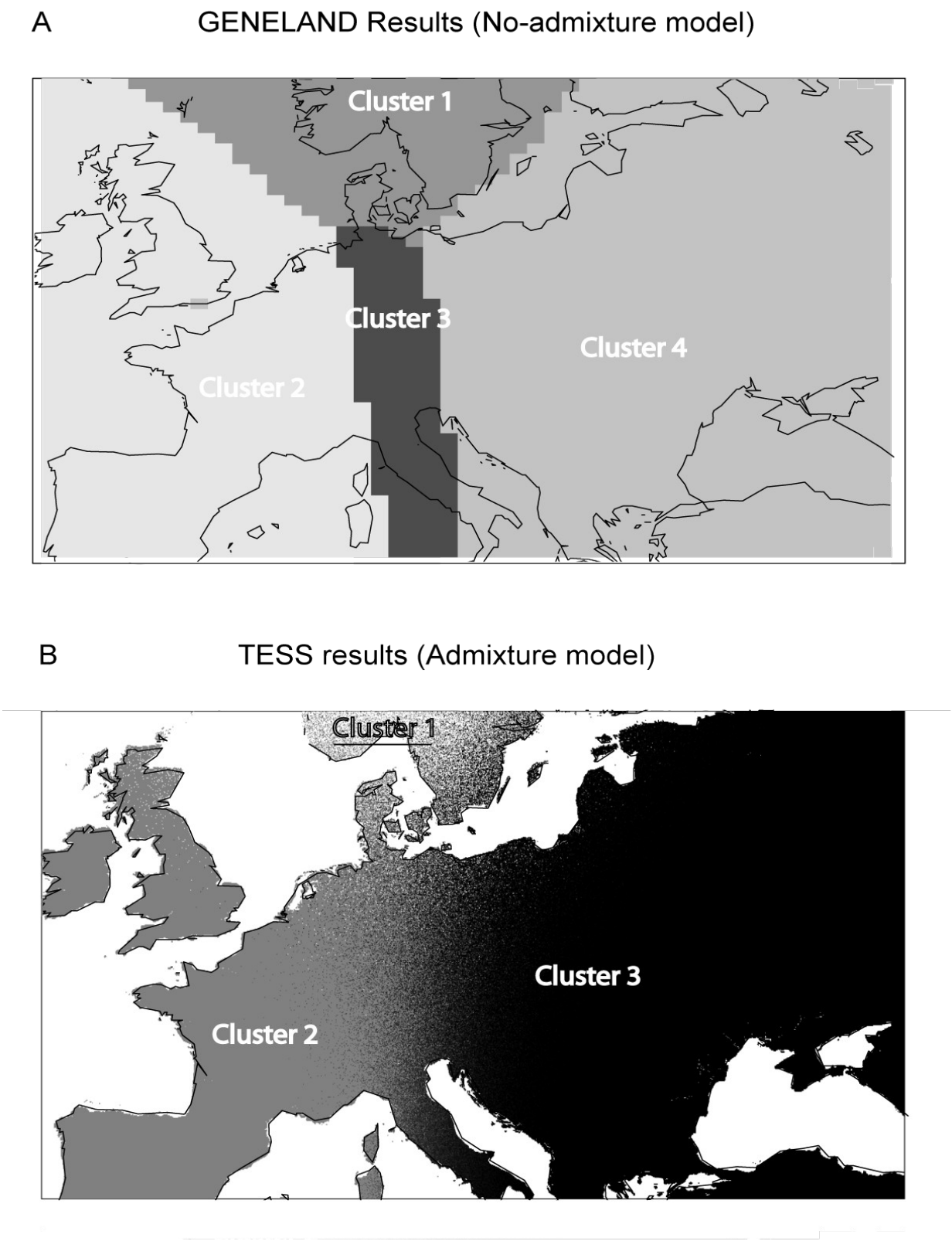

Figure 4. *Secondary contact in Europe. A) Clusters inferred by the model without admixture implemented in GENELAND. The RJMCMC algorithm chooses K = 4 clusters. B) Posterior prediction of admixture proportions inferred by TESS (admixture model). The DIC leads to select a model with K = 3 parental populations. 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.*

Applications of spatially explicit Bayesian clustering programs

Table 1 summarizes recent applications published during the last year (Spring 2008 – Summer 2009) showing the importance of spatial Bayesian clustering methods to evolutionary biologists and molecular ecologists. The published studies cover a large spectrum of taxa. Most of these studies have been conducted at a regional scale spanning a significant range of the species habitat. Only a few studies have been conducted at a continental scale, often for model species (Tishkoff et al 2009). The analyses make approximately equal use of TESS and GENELAND although the two programs are seldom applied to the data simultaneously. Due to its more recent publication date, the spatial version of BAPS has perhaps not been given enough time to become popular among users. In relation to landscape genetics (Manel et al 2003), a frequent focus is on detecting genetic discontinuities associated with barriers to gene flow or habitat loss and fragmentation (Quéméré et al 2009; Gardner-Santana et al 2009; Galarza et al 2009). In studies that attempt to locate genetic discontinuities, the spatial models are used without admixture options. Note that when STRUCTURE is used in those cases, the program is applied under its default admixture model. Other applications focus on detecting contact zones (Durand et al 2009), hybrid zones (Dépraz et al 2009), random mating populations (Goombridge et al 2009) or patterns of isolation-by-distance (Gauffre et al 2008). When more than one model is used, the studies often report consensual results, but there are interesting exceptions. Using spatial models, Lavarando 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 (2009) detected two clusters with STRUCTURE and TESS, but this was not consistent with the other methods which returned hardly interpretable results. Using STRUCTURE, Stahlen 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, and, at the same time, a well-differentiated cluster in Scandinavia (François et al 2008). For this data set, the result was similar to STRUCTURE for K = 3, but STRUCTURE further stratified the cline into four smaller clusters (Nordborg et al 2005).

Discussion

Model assumptions. In models without admixture, the sample consists of K genetically divergent groups of individuals, and each genome is classified into a specific group. Thus the models may be appropriate if we have prior knowledge on reproductive isolation or on a fragmented habitat. In these models, the variability of allele frequencies is constrained over space, because the frequencies are assumed to be constant within each cluster. This implies that, in the presence of clines, the sample may be either considered 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 Figure 4B). In the latter case, the results of the program may confound the detection of actual boundaries.

In admixture models, individuals' genomes 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 had been brought into contact again at a later date. In these models, the allele frequencies are less constrained, 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 Figures 1 and 2). Spatial models of admixture are useful in this respect, since they include prior distributions that explicitly take these spatial dependencies into account at local and global scales (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 of fission and fusion events and for correcting biases in association studies (Falush et al 2003; Pritchard et al 2000b). Remark that 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 detect population structure accurately, and there is a visible benefit of using spatially explicit programs (Chen et al 2007; Latch et al 2006). Models with admixture incorrectly assign a non-negligible fraction of individual genomes to wrong clusters. However, the admixture models ly infer the number of cluster correctly, and their results actually suggest that the levels of admixture in the sample are low.Not surprisingly, models without admixture fail 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 reveal themselves efficient at detecting the cline, in which the allele frequencies vary along a longitudinal gradient. The failure of the admixture model of BAPS5 explains as this model requires the presence of non-admixed individuals in the sample, but no close descendant of the parental populations are sampled. Under spatially realistic scenarios, in which a species colonizes Europe from two southern refugia and exhibits a contact zone in the center of the area, models without admixture identify a cluster in Scandinavia, but they partition the European cline into three artificial compartments, thus producing spurious delineations that could be misinterpreted as genetic discontinuities.

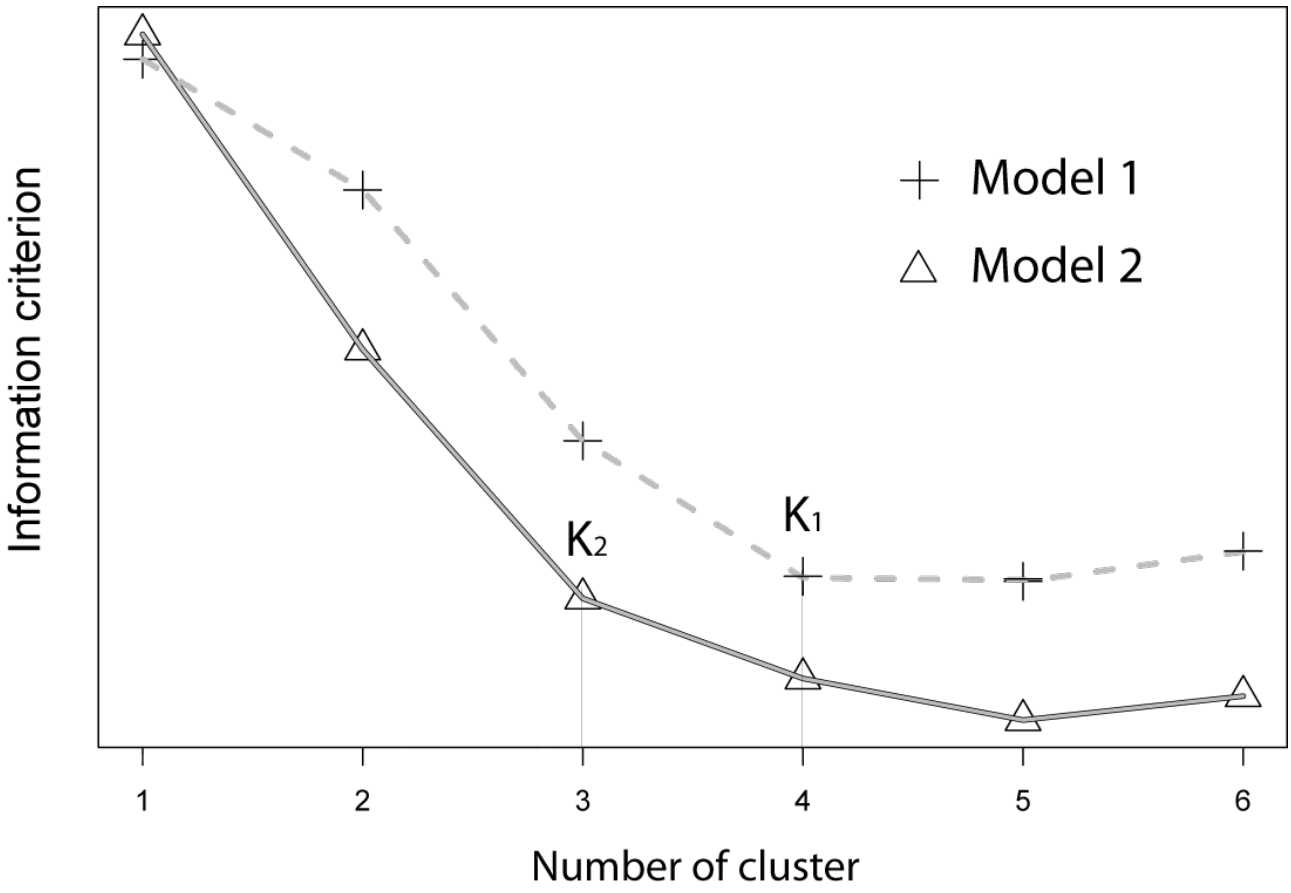

Figure 5. *Choice of K and model selection based on an information theoretic criterion (DIC or a variant). In Model 1, the values of the criterion plateaus at $K_1 = 4$ whereas in Model 2, the plateau starts at $K_2 = 3$. Because the values of the criterion are smaller in Model 2 than in Model 1, we choose Model 2 with 3 clusters.*

Model checking and model choice. Our short simulation study does not answer two fundamental questions in face of a particular data set. These questions are: Which models are best supported by our particular data? Do spatial models provide a better description of the sample than aspatial ones? Systematic answers to these 2 questions have perhaps been hindered by the hegemony of STRUCTURE in population genetic analyses. Because a variety of clustering models are now available, it becomes important to address them from the statistical viewpoint. Here we argue that one possibility it to address them by the techniques of model checking and model choice (Johnson and Omland 2004).

One 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 analyzing 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 admixture proportions (Patterson et al 2006; McVean 2009). Model checking also can be performed by simulating replicates from the posterior predictive distribution (Gelman et al 2004). In this setting, model checking is done to test whether a previously

fitted model can reproduce the observed data or not. 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. As proposed by Hoggart et al (2004), checking models 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 (Figure 5). Like AIC, lower values of DIC indicate better models. 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 equivalent to the uncorrelated allele frequency admixture model of STRUCTURE. According to the DIC, there was 5 clusters in the sample with the aspatial model. The 5 clusters were checked to be almost identical to those obtained with the default options of STRUCTURE, for which the ΔK criterion also selected 5 clusters. A spatially explicit admixture version of TESS obtained lower DIC scores than the aspatial model, indicating that a cline better described the data than the 5 clusters inferred by STRUCTURE. Where models have similar levels of support, model averaging – using the program CLUMPP (Jakobsson and Rosenberg 2007) – can also produce robust estimates of membership or ancestry coefficients.

Conclusions. There are many cases where the inference of population structure can benefit from the modeling of the various geographic scales at which spatial genetic variation arises. Models can account for local dispersion that generate patches of covarying allele-frequencies by including spatial autocorrelation (Epperson and 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 to run more than one model, and to use 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. Like for PCA, we should keep in mind that Bayesian clustering models are tools for exploring the data (Patterson et al 2006; McVean

2009). Because their assumptions make an obvious simplification of the biological reality, and because several demographic scenarios can result in similar clustering, genealogical interpretations of their outputs remain difficult. Efforts to develop improved model-based clustering methods are still necessary.

Acknowlegments. The authors warmly thank Oscar Gaggiotti and Richard Nichols. They are also grateful to Lounes Chikhi and Frederic Austerlitz for useful comments. OF is supported by the ANR grant BLAN06-3146282 MAEV and by the Institute of Complex Systems, IXXI.

References

1. Alexander DH, Novembre J, Lange K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 1655-64.

2. Alpermann TJ, Beszteri B, John U, Tillmann U, Cembella AD (2009) Implications of life-history transitions on the population genetic structure of the toxigenic marine dinoflagellate *Alexandrium tamarense*. *Molecular Ecology*, 18, 2122-2133.

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

4. Avise JC (2000) *Molecular Markers, Natural History and Evolution, Second Edition*. Chapman & Hall.

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

6. 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.

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

8. Bensch S, Grahn M, Müller N, Gay L, Akesson S (2009) Genetic, morphological, and feather isotope variation of migratory willow warblers show gradual divergence in a ring. *Molecular Ecology*, in press doi: 10.1111/j.1365-294X.2009.04210.x.

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

10. Bizoux JP, Daïnou K, Bourland O et al (2009) Spatial genetic structure in *Milicia excelsa (Moraceae)* indicates extensive gene dispersal in a low-density wind-pollinated tropical tree. *Molecular Ecology*, in press.

11. 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.

12. Cavalli-Sforza LL, Edwards AWF (1965). Analysis of human evolution. pp. 923–933 in Genetics Today. Proceedings of the XI International Congress of Genetics, The Hague, The Netherlands, September, 1963, volume 3, ed. S. J. Geerts, Pergamon Press, Oxford.

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

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

15. 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.

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

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

18. Coulon A, Fitzpatrick JW, Bowman R, et al. (2008) Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma coerulescens*). *Molecular Ecology*, 17, 1685-1701.

19. 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.

20. 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.

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

22. 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.

23. Dionne M, Caron F, Dodson JJ, et al. (2008). Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology*, 17, 2382-2396.

24. Dvorak WS, Potter KM, Hipkins VD, Hodge GR (2009) Genetic Diversity and Gene Exchange in Pinus oocarpa, a Mesoamerican Pine with Resistance to the Pitch Canker Fungus (*Fusarium circinatum*). *International Journal of Plant Sciences* 170, 609–626.

25. 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.

26. Dupont L, Viard F, Dowell MJ, et al. (2009) Fine- and regional-scale genetic structure of the exotic ascidian *Styela clava* (*Tunicata*) in southwest England, 50 years after its introduction. *Molecular Ecology*, 18, 442-453.

27. Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, 26, 1963-1973.

28. Echenique-Diaz LM, Yokoyama J, Takahashi O, Kawata M (2009) Genetic structure of island populations of the endangered bat *Hipposideros turpis turpis*: Implications for conservation. *Population Ecology*, 51, 153-160.

29. Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. pp. 67–76 in Phenetic and Phylogenetic Classification, ed. V. H. Heywood and J. McNeill. Systematics Association pub. no. 6, London.

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

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

32. 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.

33. 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, 479–91.

34. 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.

35. 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.

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

37. Fedy BC, Martin K, Ritland C, Young J (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.

38. Fogelqvist J, Niittyvuopio A, Ågren J, et al. (2009) Cryptic population genetic structure: the number of inferred clusters depends on sample size. *Molecular Ecology Resources* (in press)

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

40. François O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random

fields in spatial population genetics. *Genetics*, 174, 805–816.

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

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

43. Galarza JA, 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 USA*, 106, 1473-1478.

44. 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.

45. 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.

46. Garcia-Gil MR, Francois 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.

47. Garrick RC, Nason JD, Meadows CA, Dyer RJ (2009) Not just vicariance: Phylogeography of a Sonoran desert euphorb indicates a major role of range expansion along the Baja peninsula. *Molecular Ecology*, 18, 1916-1931.

48. Gauffre B, Estoup A, Bretagnolle V, et al. (2008) Spatial genetic structure of a small rodent in a heterogeneous landscape. *Molecular Ecology*, 17, 4619-4629.

49. Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis, Second Edition.* Chapman & Hall/CRC.

50. Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B*, 59 731-792.

51. Groombridge JJ, Dawson DA, Burke T, Prys-Jones R, Brooke MD, Shah N (2009) Evaluating the demographic history of the Seychelles kestrel (*Falco araea*): Genetic evidence for recovery from a population bottleneck following minimal conservation management. *Biological Conservation*, 142, 2250-2257.

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

53. Hartl DL, Clark AG (1997) *Principles of Population Genetics, Third Edition*. Sinauer Associates, Inc.

54. 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.

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

56. 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.

57. Holzer B, Keller L, Chapuisat M (2009) Genetic clusters and sex-biased gene flow in a unicolonial Formica ant. *BMC Evolutionary Biology*, 9, 69.

58. 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.

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

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

61. 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.

62. Johnson JB, Omland KS (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution,* 19, 101-108.

63. Johansson ML, Banks MA, Glunt KD, et al. (2008) Influence of habitat discontinuity, geographical distance, and oceanography on fine-scale population genetic structure of copper rockfish (*Sebastes caurinus*). *Molecular Ecology*, 17, 3051-3061.

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

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

66. Latch E, Dharmarajan G, Glaubitz J, Rhodes O (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.

67. Lavandero, B, Miranda, M, Ramírez CC, Fuentes-Contreras E (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.

68. 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.

69. Liu ZJ, Ren BP, Wu RD, et al. (2009) The effect of landscape features on population genetic structure in Yunnan snub-nosed monkeys (*Rhinopithecus bieti*) implies an anthropogenic genetic discontinuity. *Molecular Ecology*, 18, 3831-3846.

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

71. 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.

72. McCairns RJS, Bernatchez L (2008) Landscape genetic analyses reveal cryptic population structure and putative selection gradients in a large-scale estuarine environment. *Molecular Ecology*, 17, 3901-3916.

73. McDevitt AD, Mariani S, Hebblewhite M, et al.(2009) Survival in the Rockies of an endangered hybrid swarm from diverged caribou (*Rangifer tarandus*) lineages. *Molecular Ecology*, 18, 665-679.

74. McVean G (2009) A genealogical interpretation of Principal Components Analysis. *PLoS Genetics* 5, e1000686.

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

76. 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.

77. Patterson N, Price A, Reich D (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190.

78. 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.

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

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

81. Quéméré E, Louis Jr EE, 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*, in press

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

83. Rosenberg NA, Saurabh S, Ramachandran S, et al. (2005) Clines, clusters, and the effect of study design on the influence of human population structure. *PLoS Genetics*, 1, 660–671.

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

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

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

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

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

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

90. Tishkoff SA, Reed FA, Friedlaender FR, et al. (2009). The genetic structure and history of Africans and African Americans. *Science* 324: 1035-1044.

91. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture : Analytical and study design considerations. *Genetic epidemiology*, 28, 289–301.

92. Wang S, Lewis CM, Jakobsson M, et al. (2007) Genetic variation and population structure in native Americans. *PLoS Genetics*, 3, e185.

93. 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.

94. 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.

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

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

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

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

99. 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.

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