Syst. Biol. 61(5):897–911, 2012

DOI:10.1093/sysbio/sys038
Advance Access publication on March 6, 2012

A Unifying Model for the Analysis of Phenotypic, Genetic, and Geographic Data

GILLES GUILLOT[1,*], SABRINA RENAUD[2], RONAN LEDEVIN[3], JOHAN MICHAUX[4], AND JULIEN CLAUDE[5]

[1]*Informatics Department, Technical University of Denmark, Richard Petersens Plads, 2800 Lyngby, Denmark;* [2]*Laboratoire de Biométrie et Biologie Evolutive UMR 5558, CNRS, Université Lyon 1, Université de Lyon, 69622 Villeurbanne, France;* [3]*Anthropological Institute and Museum, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland;* [4]*Centre de Biologie et de Gestion des Populations, INRA/IRD/CIRAD/SupAgro, CBGP, Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez cedex, France; and* [5]*Laboratoire de Morphométrie, ISE-M, UMR5554 CNRS/UM2/IRD, Université de Montpellier II, Montpellier 34095, France;*

**Correspondence to be sent to: Informatics Department, Technical University of Denmark, Richard Petersens Plads, 2800 Lyngby, Denmark; E-mail: gigu@im.dtu.dk.*

Received 21 December 2011; reviews returned 14 February 2012; accepted 25 February 2012
Associate Editor: Cécile Ané

Abstract.—Recognition of evolutionary units (species, populations) requires integrating several kinds of data, such as genetic or phenotypic markers or spatial information in order to get a comprehensive view concerning the differentiation of the units. We propose a statistical model with a double original advantage: (i) it incorporates information about the spatial distribution of the samples, with the aim to increase inference power and to relate more explicitly observed patterns to geography and (ii) it allows one to analyze genetic and phenotypic data within a unified model and inference framework, thus opening the way to robust comparisons between markers and possibly combined analyses. We show from simulated data as well as real data that our method estimates parameters accurately and is an improvement over alternative approaches in many situations. The power of this method is exemplified using an intricate case of inter- and intraspecies differentiation based on an original data set of georeferenced genetic and morphometric markers obtained on *Myodes* voles from Sweden. A computer program is made available as an extension of the R package Geneland. [Bayesian model; bio-geography; clustering; Markov chain Monte Carlo; molecular markers; morphometrics; Myodes; R package; spatial data.]

Species delimitation is of interest in conservation biology (delimitation and management of endangered species), epidemiology (detection of new pathogens), and evolutionary biology to describe, quantify, and understand mechanisms of speciation. Methodological advances in evolutionary biology have led to methods for species delimitation solely based on the variation of key genetic markers (e.g., DNA barcoding, Luo et al. 2011). Limits of these single-marker approaches are made evident by conflicts between different genes in a multi-marker approach (Rodríguez et al. 2010; Turmelle et al. 2011) or between genetic and phenotypic markers (Nesi et al. 2011). In this context of species or population delimitation, phenotypic data still remain of interest together with genetic markers.

Phenotypes such as size and/or shape of morphological structures are the product of numerous interacting nuclear genes (Klingenberg et al. 2001) and, as such, can provide a global estimate of the divergence between units. Furthermore, by being the target of selection, morphological variation can provide precious insights on the selection pattern contributing to shape the units. In the case of fossil lineages, it may even be the only information available to identify evolutionary and systematic units (Girard and Renaud 2011; Néraudeau 2011).

A rich toolbox is available to tackle these questions. Many methods work as partition clustering and aim at defining how many groups are represented in a sample and assign individuals following some optimality principles. These methods were initially developed to deal with continuous quantitative measurements. These classical clustering methods have been implemented in programs such as EMMIX (McLachlan et al. 1999), MCLUST (Fraley 1999), and MIXMOD (Biernacki et al. 2006). Such clustering and assignment methods were not commonly used in systematics until they were made more widely available through population genetic applications such as the computer program STRUCTURE (Pritchard et al. 2000) and related work (reviewed, e.g., by Excoffier and Heckel 2006). More recently, Hausdorf and Hennig (2010) and Yang and Rannala (2010) developed methods for delimiting species based on multi-locus data. While the approach of Hausdorf and Hennig (2010) method hinges on Gaussian clustering, that of Yang and Rannala (2010) is based on the coalescent and makes use of a user-specified guide tree. Methods for genetic data have also been extended to incorporate information about the spatial location of each sample—information rarely used although commonly available in data analysis in evolutionary biology—with the aim of increasing power of inferences and of relating more explicitly observed patterns to geography (Guillot et al. 2005, 2009).

These tools have been developed by different communities (evolutionists, population geneticists, and statisticians). Therefore, one still lacks a unified framework, and this constitutes a major impediment for combining various kinds of data. This is especially true for morphological markers that have not received as much attention as genetic markers for recognizing populations and species. There are therefore a few major gaps in the toolbox available to identify evolutionary units, namely there is to date: no method to analyze genetic data and phenotypic data under the same general paradigm (model and inference framework) and no method to incorporate spatial information in such phenotypic/genetic analysis.

The goal of the present paper is to fill these gaps. We propose a model to deal, in an integrated way, with georeferenced phenotypic and genetic data. We also provide a freely available computer program that implements this model and should ease data analysis in many respects. Given the complexity of the modeling and inferential task, our method is not based on an explicit evolutionary model (e.g., coalescent) but on a statistical model. This model is a parametrization that is not only general enough to capture some essential patterns in the data but also simple enough to be subject to a rigorous and accurate inference method. Briefly, our model assumes the existence of several clusters that display some kind of homogeneity. This model mimics more or less what would be expected from a population: homogeneity in terms of genetic and phenotypic variation and some geographical continuity. The existence of homogeneous clusters corresponds to the fact that some individuals have shared some aspects of their recent ecological or evolutionary history. This shared history is summarized by cluster-specific parameters that represent allele frequencies and means and variances of phenotypic traits. Because it is not based on an explicit evolutionary model, it does not require prior information, as for instance, a guide tree in the method of Yang and Rannala (2010). The statistical challenge in this context is to estimate the number of clusters and these cluster-specific parameters.

This article is organized as follows. First we provide a description of the model and inference machinery. Next we illustrate our method and test its accuracy on a large set of simulated data as well as on two published empirical data sets. Then we implement our method on an original data set of georeferenced genetic and morphometric markers to decipher the complex inter- and intraspecific structure of red-backed and bank voles *Myodes rutilus* and *M. glareolus* in Sweden. We conclude by discussing potential applications in a more general context.

METHOD

Overview

We assume that we have a data set consisting of n individuals sampled at sites $\mathbf{s} = (s_i)_{i=1,\ldots,n}$ (where s_i is the two-dimensional spatial coordinate of individual i), observed at some phenotypic variables denoted $\mathbf{y} = (y_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,q}}$, and/or some genetic markers denoted $\mathbf{z} = (z_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,l}}$. Our approach is able to deal with any combination of phenotypic and genetic data, including situations where only phenotypic or only genetic data are available as well as situations when each individual is observed through its own combination of phenotypic and genetic markers. As it will be shown below, our approach also encompasses the case where sampling locations are missing (or considered to be irrelevant). The only constraint that we impose at this stage is that if spatial coordinates are used, they must be available for all individuals. We assume that each individual sampled belongs to one of K different clusters and that variation in the data can be captured by cluster-specific location and scale parameters.

Prior and Likelihood Model for Phenotypic Variables

Denoting by p_i the cluster membership of individual i ($p_i \in \{1, \ldots, K\}$), we assume that, conditionally on $p_i = k$, y_{ij} is drawn from a parametric distribution with cluster-specific parameters. Independence is assumed within and across clusters conditionally on cluster membership. This means in particular that there is no residual dependence between variables not captured by cluster memberships. Implications of this assumption are discussed later. Although most of the analysis that follows would be valid for all families of continuous distribution, we assume in the following that the y values arise from a normal distribution. Each cluster is therefore characterized by a mean μ_{kj} and a variance σ^2_{kj} and our model is a mixture of multivariate-independent normal distributions (Frühwirth-Schnatter 2006). Following a common practice in Bayesian analysis (Gelman et al. 2004), we use the natural conjugate prior family on $(\mu_{kj}, 1/\sigma^2_{kj})$ for each cluster k and variable j. Namely, we assume that the precision $1/\sigma^2_{kj}$ (i.e., inverse variance) follows a Gamma distribution $\mathcal{G}(\alpha, \beta)$ (α shape, β rate parameter) and that, conditionally on σ_{kj}, the mean μ_{kj} has a normal distribution with mean ξ and variance σ^2_{kj}/κ. In the specification above, α, β, ξ, and κ are hyperparameters. Details about their choice are discussed in the Appendix and in the Supplementary Material (available from 10.5061/dryad.9mp902ck).

Prior and Likelihood Model for Genetic Data

We assume here a mixture of multinomial distributions. This is the model previously introduced by Pritchard et al. (2000) to model individuals with pure ancestries. Denoting frequency of allele a at locus l in cluster k by f_{kla} for diploid genotype data, we assume that

$$\pi(z_{ij} = \{a, b\} | p_i = k) = 2 f_{kla} f_{klb} \quad \text{whenever } a \neq b \tag{1}$$

$$\text{and } \pi(z_{ij} = \{a, a\} | p_i = k) = f^2_{kla}. \tag{2}$$

While for haploid data, we have

$$\pi(z_{ij} = a | p_i = k) = f_{kla}. \tag{3}$$

We also deal with dominant markers for diploid organisms with a modified likelihood (see Guillot and Santos 2010; Guillot and Carpentier-Skandalis 2011 for details). We assume independence of the various loci within and across clusters conditionally on cluster memberships. In particular, as with all other population genetic clustering models (including STRUCTURE), we do not attempt to model background linkage disequilibrium. Therefore, our model can handle nonrecombining DNA sequences (such as data obtained from

mitochondrial DNA (mtDNA), Y chromosomes, or tightly linked autosomal nuclear markers), provided data are reformatted in such a way that the various haplotypes are recoded as alleles of a single locus (but see also Discussion section). We assume that allele frequencies $f_{kl.}$ have a Dirichlet distribution. Independence of the vectors $f_{kl.}$ is assumed across loci. Regarding the dependence structure across clusters, we consider either independence (referred to as uncorrelated frequency model or UFM) or an alternative model (referred to as correlated frequency model or CFM) introduced by Balding and Nichols (1995, 1997). In this second model, allele frequencies also follow a Dirichlet distribution but now depending on some cluster-specific drift parameters. In this model, $f_{kl.}$ are assumed to follow a Dirichlet distribution $\mathcal{D}(\tilde{f}_{la}(1-d_k)/d_k, \ldots, \tilde{f}_{lA}(1-d_k)/d_k)$, where d_ks parametrize the speed of divergence of the various clusters and the $\tilde{f}_{la}$s represent the allele frequency in an hypothetical ancestral population. This model can be viewed as a heuristic and computationally convenient approximation of a scenario in which present-time clusters result from the split of an ancestral cluster some generations ago. It is also a Bayesian way of introducing correlation between clusters at the allele frequency level and, hence, to infer subtle differentiations that would have been missed by a model assuming independence of allele frequencies across clusters (Falush et al. 2003; Guillot 2008; Sirén et al. 2011).

Prior Models for Cluster Membership

Spatial model.—We consider a statistical model known as colored Poisson–Voronoi tessellation (Møller and Stoyan 2009). Loosely speaking, this model assumes that each cluster area in the geographic domain can be approximated by the union of a few polygons. Most of the modeling ideas can be grasped from the examples shown in Figure 1. The polygons are assumed to be centered around some points that are generated by a homogeneous Poisson process (i.e., points located completely at random in the geographic domain). Formally, we denote by $(u_1, \ldots, u_m)$ the realization of this Poisson process. These points in $\mathbb{R}^2$ induce a Voronoi tessellation into m subsets $\Delta_1, \ldots, \Delta_m$. The Voronoi tile associated with point u_i is defined as $\Delta_i = \{s \in \mathbb{R}^2, \mathrm{dist}(s, u_i) < \mathrm{dist}(s, u_j) \forall j \neq i\}$. Each tile receives a cluster membership c_i (coded graphically as a color hence the terminology) at random sampled independently from a uniform distribution on $\{1, \ldots, K\}$. Denoting by D_k the union of tiles with color k, the set $(D_1, \ldots, D_K)$ defines a tessellation in K subsets. This model is controlled by the intensity of the Poisson process λ (the average number of points per unit area) and the number of clusters K. We place a uniform prior on $[0, \lambda_{\max}]$ and on $\{0, \ldots, K_{\max}\}$, respectively. This model is a flexible tool widely used in engineering to fit arbitrary shapes in a nonparametric way (Møller and Stoyan 2009). It offers a good trade-off between model complexity, realism, and computational efficiency. It is presumably most useful in situations of incipient allopatric speciation but examples of applications in other contexts can be found (e.g., in the studies of Coulon et al. 2006; Fontaine et al. 2007; Wasser et al. 2007; Hannelius et al. 2008; Joseph et al. 2008; Sacks et al. 2008; Galarza et al. 2009; Beadell et al. 2010; see also Guillot et al. 2009 for review and additional references). Lastly, we note that our approach relates to that of Hausdorf and Hennig (2003), who propose a test for clustering of areas of distribution. However, rather than testing clusteredness, our approach estimates these areas of distribution. To do that, we assume some clusteredness but without making strong assumptions about its intensity.

Nonspatial model.—If spatial coordinates are not available or thought to be irrelevant to the species at the spatial scale considered, then a nonspatial model can be used. The nonspatial modeling option considered here does not require to introduce any auxiliary point process as above, but for the sake of consistency, we use the same setting as in the paragraph above. We set $m = n$ and impose $(u_1, \ldots, u_n) = (s_1, \ldots, s_n)$. Here, the s_is are some known spatial coordinates or dummy points if this piece of information is missing. This model does not impose any spatial structure and corresponds to the model implemented in most nonspatial cluster programs, including the genetic clustering programs BAPS (Corander et al. 2003, 2004) and STRUCTURE (with the exception of the latest model presented by Hubisz et al. 2009).

Summary of Proposed Model

The parameters in our model are as follows: number of clusters K, rate of Poisson process λ, number of events (points) of the Poisson process m, events of Poisson process $\mathbf{u} = (u_1, \ldots, u_m)$, color of tiles (i.e., cluster membership of spatial partitioning subdomains) $\mathbf{c} = (c_1, \ldots, c_m)$, allele frequencies $\mathbf{f} = (f_{kla})$ (frequency of allele a at locus l in cluster k), genetic drift parameters $\mathbf{d} = (d_1, \ldots, d_K)$, allele frequencies in the ancestral population $\tilde{\mathbf{f}} = (\tilde{f}_{la})$, expectations of phenotypic variables $\boldsymbol{\mu} = (\mu_{kj})$, standard deviations of phenotypic variables $\boldsymbol{\sigma} = (\sigma_{kj})$ (note that $\boldsymbol{\sigma}$ is not a variance–covariance matrix (the phenotypic variables are assumed to be independent) but rather a set of scalar variances stored in a two-dimensional array. On top of this, we place a uniform prior on $[0, \lambda_{\max}]$ on λ, a uniform prior on $\{0, \ldots, K_{\max}\}$ on K, a Beta $\mathrm{B}(\delta_k, \delta_k))$ prior on d_k and a Gamma distribution $\mathcal{G}(g, h)$ on β.

The vector of unknown parameters is therefore $\boldsymbol{\theta} = (K, \lambda, m, \mathbf{u}, \mathbf{c}, \mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$. We also denote by $\boldsymbol{\theta}_{\mathrm{S}} = (\lambda, m, \mathbf{u}, \mathbf{c})$, $\boldsymbol{\theta}_{\mathrm{G}} = (\mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d})$, and $\boldsymbol{\theta}_{\mathrm{P}} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ the parameters of the spatial, genetic, and phenotypic parts of the model, respectively.

The hierarchical structure of the model is summarized on the graph shown in Figure 2. There are three blocks of parameters relative to the genetic, phenotypic, and geographic component of the model. Infor-

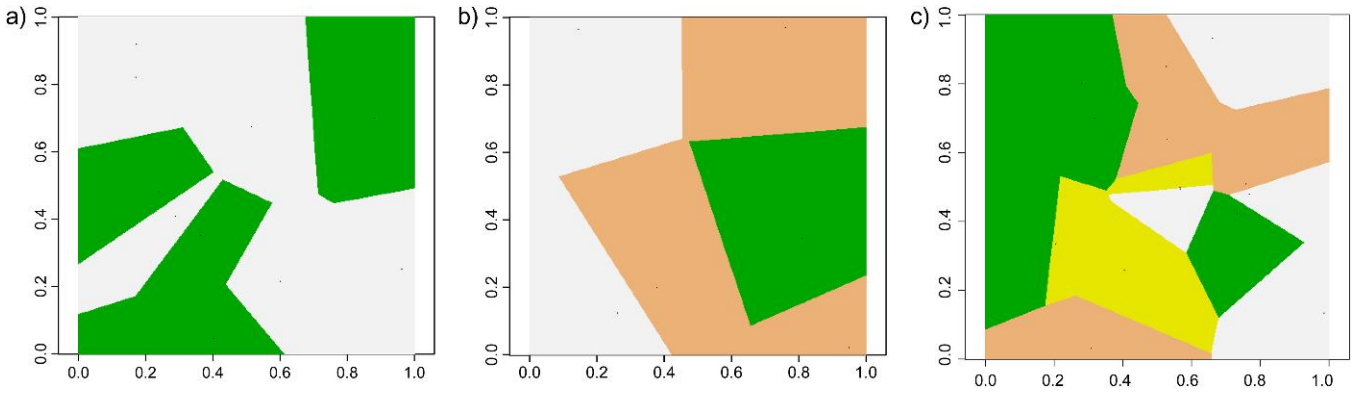

FIGURE 1. Examples of spatial clusters simulated from our prior model. The square represents the geographic study area. Membership of a geographical site to one of the K clusters is coded by color in the online version of this article, and by shading in print. From left to right: $K = 2, 3$ and 4. A given clustering depends on K and on the number, locations, and colors/shading (cluster memberships) of each polygon. If the prior placed on the number of polygons tends to favor low values, then each cluster tends to be made of one or only a few large areas. This is in sharp contrast with nonspatial Bayesian models which typically assume that clusterings with highly fragmented cluster areas are not unlikely.

mation propagates from data to higher levels of the model across the various nodes of the graph through probabilistic relationships specified between neighboring nodes. The structure of the global model can be summarized by the joint distribution of $\boldsymbol{\theta}$ and $(\mathbf{y}, \mathbf{z})$. By the conditional independence assumptions, we get

$$\begin{aligned}\pi(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\mathbf{z}|\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}_{\mathrm{P}})\pi(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{G}}). \end{aligned} \quad (4)$$

Each genetic or phenotypic marker brings one factor in the likelihood. Whether the clustering is driven by the genetic or the phenotypic data depends on the respective differentiation and on the number of markers of each kind.

Estimation of Parameters

Bayesian estimation and MCMC inference.—We are interested in the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. Note that this notation does not refer explicitly to the sample locations because, unlike genetic markers and phenotypic variables, locations are not considered as random quantities in our model. The model does, in fact, implicitly account for spatial information. The distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ is defined on a high-dimensional space and deriving properties analytically about this distribution is out of reach. We implement a Markov chain Monte Carlo (MCMC) strategy. This amounts to generating a sample of N correlated replicates $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. The initial state $\boldsymbol{\theta}_1$ is simulated at random from a distribution that does not matter in principle, a fact that has to be checked in practice by convergence monitoring tools (Gilks et al. 1996; Robert and Casella 2004). We always sample $\boldsymbol{\theta}_1$ from the prior and we check that starting from various random states does not affect the overall result provided a suitable number of burn-in iterations are discarded. In analyses reported below, the order of magnitude of N was 50,000–100,000 iterations with 20,000 burn-in iterations. See Appendix for details on the MCMC algorithm.

Estimation of the number of clusters.—Each simulated state $\boldsymbol{\theta}_i$ includes a simulated number of clusters K_i. The number of clusters is estimated as the most frequent value among the N simulated values $K_1, \ldots, K_N$ and we denote it by $\hat{K}$.

Estimating cluster memberships.—A model assuming that individuals i and j belong, respectively, to Clusters 1 and 2 characterized by a mean phenotypic trait equal to 5 and 7 is essentially the same as a model assuming that individuals i and j belong, respectively, to Clusters 2 and 1 characterized by a mean phenotypic trait equal to 7 and 5. This trivial fact is due to the invariance of the likelihood under permutation of cluster labels and brings up a number of computational difficulties in the postprocessing of MCMC algorithm outputs known as the label-switching issue (Stephens 1997). In particular, it does not make sense to average values across the MCMC iterations. To deal with this, we implement the strategy described by Marin et al. (2005) and Guillot (2008). We consider the set of simulated $\boldsymbol{\theta}$ values restricted to the set of states such that $K = \hat{K}$. Then, working on this restricted set, we relabel each state in such a way that they "best look like" the modal state of the posterior distribution. Cluster memberships of each individual are estimated as the modal value in this relabeled sample. Then we estimate all cluster-specific parameters (mean phenotypic values and allele frequencies) by taking the average simulated value over the relabeled sample.

ANALYSIS OF SIMULATED DATA

We investigate here two new aspects of the model, namely its ability to cluster phenotypic data only and

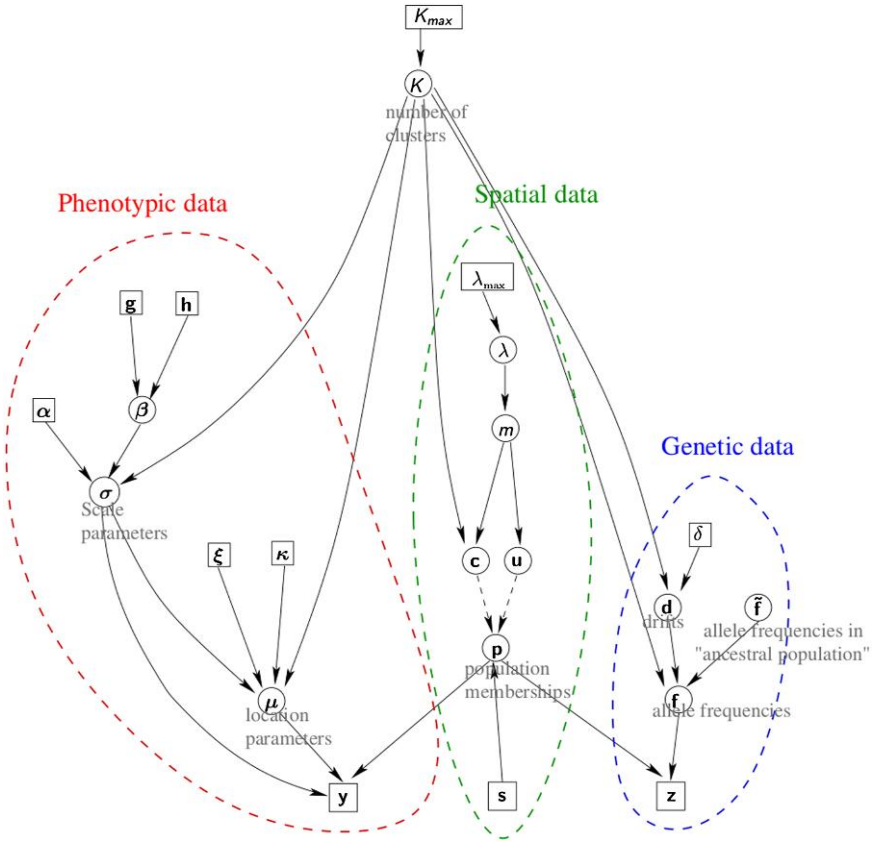

FIGURE 2. Graph of proposed model. Continuous arrows represent stochastic dependencies, dashed arrows represent deterministic dependencies. Boxes enclose data or fixed hyperparameters, circles enclose inferred parameters. Bold symbols refer to vector parameters. The dash-line rounded boxes enclose parameters relative to the phenotypic, geographic, and genetic parts of the model, respectively. The parameters of interest to biologists are the number of clusters K, the vector $\mathbf{p}$ that encode the cluster memberships, and possibly allele frequencies $\mathbf{f}$, mean phenotypic values μ, phenotypic variance σ^2 that quantify the genetic and phenotypic divergence between and within clusters. Other parameters can be viewed mostly as nuisance parameters.

phenotypic and genetic data jointly together with some spatial information.

Inference from Phenotypic Data Only

In this section, we present new results on the model for phenotypic data and focus on the spatial model option. We carried out simulations from our prior model and performed inferences as described in Estimation of Parameters section above. We produced data sets consisting of $n = 200$ individuals with $q = 5, 10, 20$, and 50 phenotypic variables. For each value of q, we produced 500 data sets with a uniform prior $\mathcal{U}(\{1, \ldots, 5\})$ on K. In real life, the range of value of the putative true K is largely unknown. To be as close as possible to this situation (i.e., pretend to ignore the true value of K), we carried out inference under a uniform $\mathcal{U}(\{1, \ldots, 10\})$ prior for K. We assessed the accuracy of inferences by computing the classification error (Fig. 3). Further details are provided in Supplementary Material.

We also wished to assess how our method performs compared with other computer programs implementing state-of-the-art methods. We therefore considered the R package MCLUST (Banfield and Raftery 1993; Fraley 1999) which is one of the most widely used and arguably most advanced programs to perform clustering. This program implements inference for Gaussian mixtures and as such deals solely with continuous quantitative data. It implements a nonspatial algorithm and in its default setting performs inference by likelihood maximization via the expectation maximization (EM) algorithm. It implements a wide class of submodels regarding the covariance structure of the data. In its default option (which we used), it performs model selection (covariance structure and number of clusters) by optimizing a Bayesian information criterion. We set the maximum number clusters to the $K_{\max} = 10$, that is, to the same value as in analyses with our method.

We stress here that the goal of this experiment is not to rank our method and MCLUST, as the two methods/programs differ in many important respects. They differ regarding the type of data handled (MCLUST is not aimed at genetic data and does not implement any spatial model) and the breadth of covariance structure

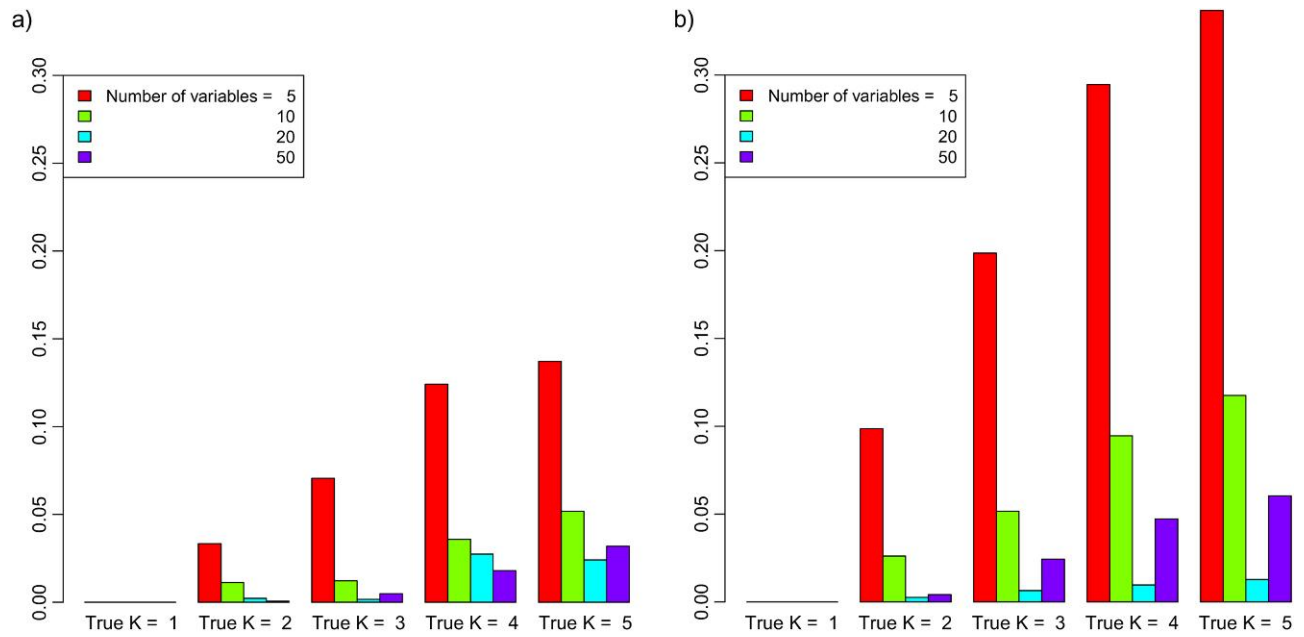

FIGURE 3. Classification error from simulated data. The variable plotted on the *y*-axis is the proportion of misclassified individuals (after correction for potential label switching issues). Each bar is obtained as an average over 500 data sets consisting of $n = 200$ individuals. Both methods are excellent at avoiding false positives (i.e., reporting $\hat{K} = 1$ when K=1) and have a clear ability to reduce the error rate when the number of variables increases. They seem to lose accuracy in the same fashion when they are given an increasingly difficult problem (i.e., when the true K increases) and have difficulty fully exploiting all the available information when the number of variables is large (cf. loss of accuracy for 50 variables compared with 20 variables). In the overall, under this type of simulated data, our method is typically twice as accurate as the competing method.

considered (our approach assumes conditional independence while MCLUST considers in excess of 10 types of covariance structures). It would be, therefore, difficult to design an efficient and fair comparison. Results are mostly given here to support the claim that our method compares with state-of-the-art methods and to assess the magnitude of improvement brought by the use of a spatial model in a best-case scenario when data are spatially structured (see also Discussion section). Most of the numerical results are summarized in Figure 3. To understand better how the method behaves as a function of the pairwise phenotypic differentiation between clusters, we also report the classification error as a function of the T^2 statistic in a Hotelling T test (Anderson 1984) on Figure 4. See also Supplementary Material for further details.

Inference from Phenotypic and Genetic Data Jointly

We illustrate here how combining phenotypic and genetic data can improve the accuracy of inferences compared with inferences carried out from one type of data only. To do so, we simulated 500 data sets consisting of two clusters each. There were 5 phenotypic variables and 10 codominant genetic markers. We investigated a broad range of phenotypic and genetic differentiation and it appears that, on average, combining the two types of data increases the accuracy of inference (Fig. 5).

ANALYSIS OF DATA FROM THE LITERATURE

Analysis of Iris Morphometric Data

Fisher's iris data set (Anderson 1935; Fisher 1936) gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *I. versicolor*, and *I. virginica*. We applied our method to the data transformed into log-shape ratios (see Claude 2008 and references therein). Since the data are not georeferenced, we used the nonspatial prior. We launched 10 independent MCMC runs. Seven of them return correctly $\hat{K} = 3$, the other three runs return $\hat{K} = 4, 5$, and 6, respectively. Ranking the runs according to the average posterior density, the best run corresponds to one of the seven runs that estimate K correctly (according to the number of actual species in the data set). This run achieves a classification error of 6% (see Fig. 6). MCLUST returns an estimate of K equal to 2 (raw data or log shape ratio data) and 50 out of 150 individuals are misclassified, thus failing to identify the three species of the data set.

AFLP Data of Calopogon *from Eastern North America and the Northern Caribbean*

The way our model deals with genetic data and the accuracy resulting from this method based on genetic

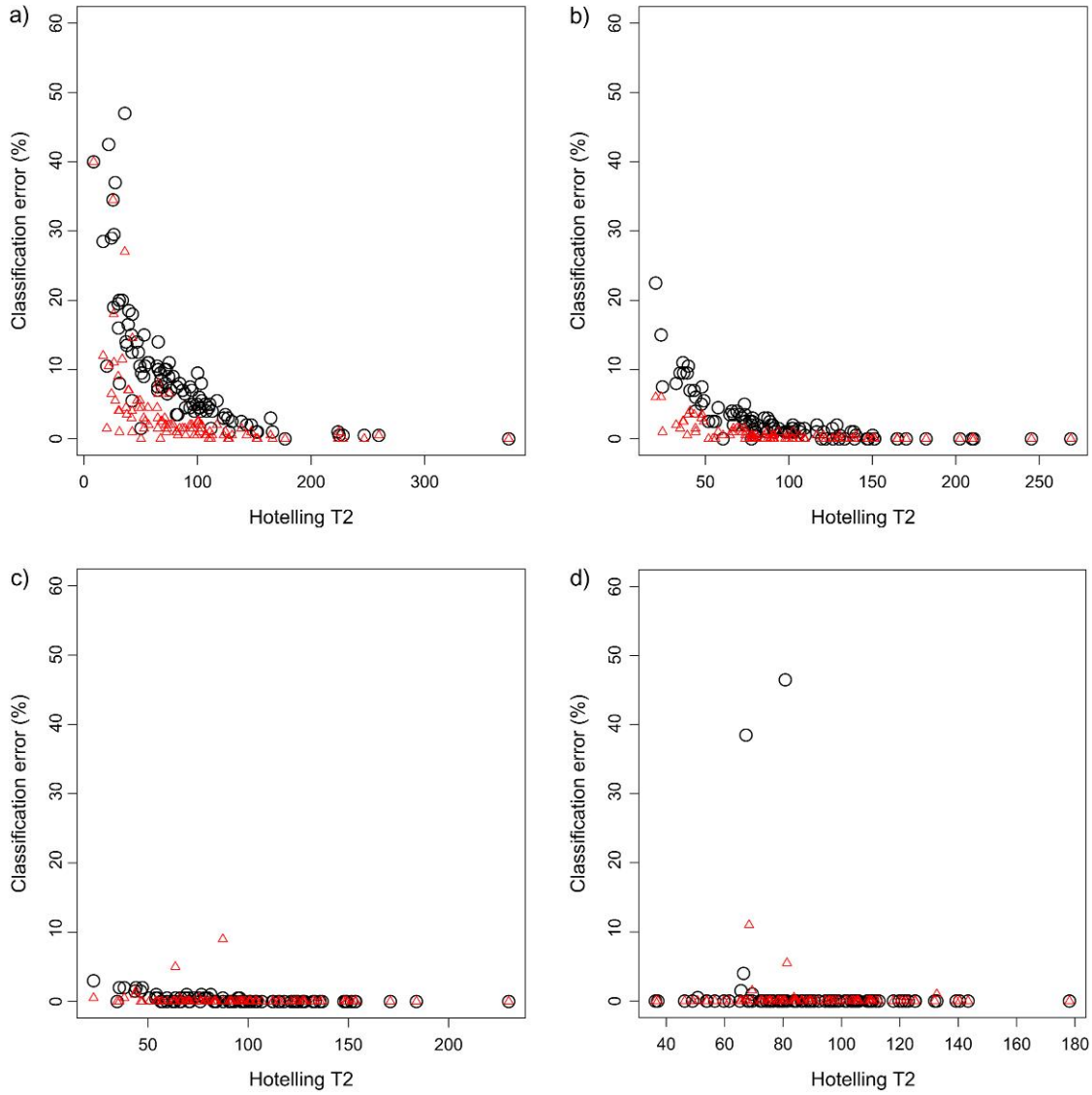

FIGURE 4. Classification error for simulated data sets consisting of $K = 2$ clusters as a function of the phenotypic differentiation between the clusters. The variable plotted on the y-axis is the proportion of misclassified individuals (after correction for potential label switching issues). The variable plotted on the x-axis is the Hotelling T statistic and assesses the magnitude of the phenotypic differentiation. Our method: △, MCLUST: ○.

data only has been investigated by Guillot et al. (2005, 2008), Guillot (2008), Guillot and Santos (2010), Safner et al. (2011), and Guillot and Carpentier-Skandalis (2011) and further discussion can be found in Guillot et al. (2009). However, to further illustrate the accuracy of our method when used with genetic data only, we study here a data set produced and first analyzed by Goldman et al. (2004).

This data set consists of 60 *Calopogon* samples genotyped at 468 AFLP markers. Goldman et al. (2004) identified the presence of five species (*C. barbatus, C. oklahomensis, C. tuberosus, C. pallidus,* and *C. multiflorus*) and two hybrids specimens (*C. tuberosus* × *C. pallidus* and *C. pallidus* × *C. multiflorus*). According to Goldman et al. (2004), *C. tuberosus* has been widely considered to have three varieties: var. *tuberosus*, var. *latifolius*, and var. *simpsonii*. In addition, the data set contains samples from two outgroups so that one could consider that the data set contains up to 11 distinct species.

We analyzed this data set under the same setting as the previous data set. Under the UFM, the estimated K ranges between 2 and 3 . The best run (in terms of average posterior density) corresponds to $\hat{K} = 3$. In this clustering, one cluster contains the samples of the *C. tuberosus* species, a second cluster merges the samples of the *C. barbatus, C. oklahomensis, C. pallidus,* and *C. multiflorus* species and the hybrids. The last cluster contains the samples from the two outgroups. Under the CFM, the estimated K ranges between 7 and 8. The best run (in terms of average posterior density)

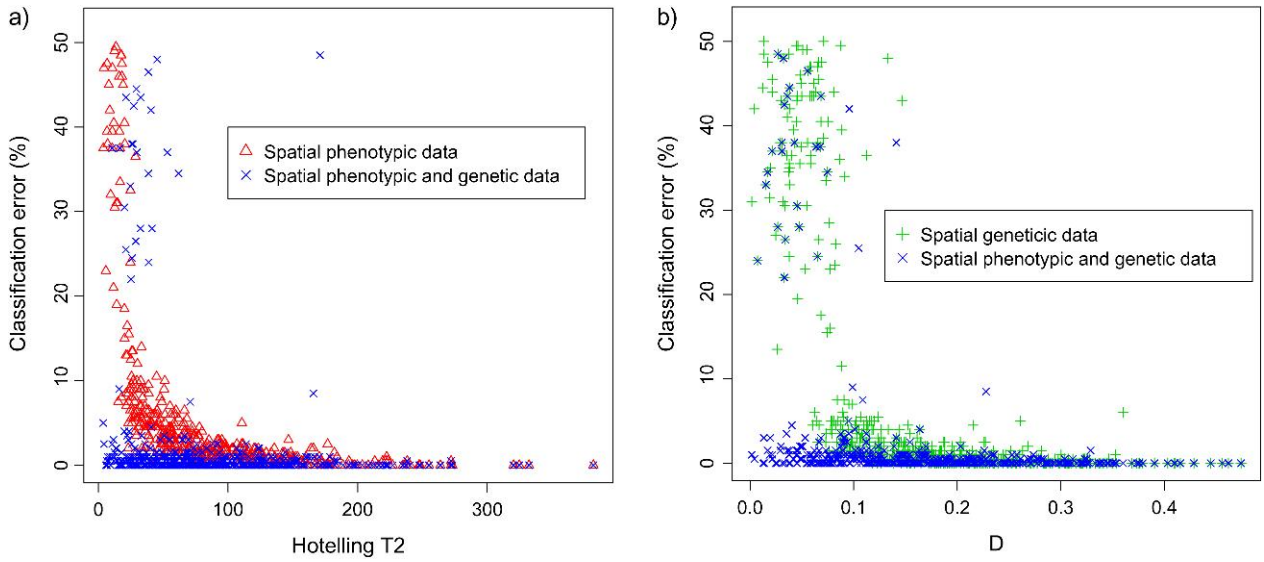

FIGURE 5. Classification error for 500 simulated data sets consisting of 200 individuals belonging to $K = 2$ clusters and recognized by $q = 5$ quantitative variables and $l = 10$ codominant loci. The variable plotted on the y-axis is the proportion of misclassified individuals using our method (after correction for potential label switching issues).

corresponds to $\hat{K} = 8$. It clusters the individuals of the various species as follows: *C. oklahomensis*/*C. multiflorus*/*C. barbatus*/*C. pallidus*, *C. tuberosus* × *C. pallidus* and *C. pallidus* × *C. multiflorus*/*C. tuberosus tuberosus* except three samples/the three *C. tuberosus tuberosus* previous samples/two extra clusters for the outgroups.

Analysis of *Myodes* Vole Data

Data and Statistical Analysis

We now study an original data set of georeferenced genetic and phenotypic markers of the voles of the genus *Myodes* in Sweden. This data set has several interesting and complex aspects useful for investigating the efficiency of our method. (i) Fennoscandia has been recognized as a zone where the mtDNA of the northern red-backed vole *M. rutilus* introgressed its southern relative, the bank vole *M. glareolus* (Tegelström 1987). This makes the delimitation of these two species impossible based on common mitochondrial markers. (ii) The bank vole is further characterized by intraspecific lineages (Deffontaine et al. 2009). Two of them are documented in Sweden (Razzauti et al. 2009), providing a complex case for disentangling intra- and interspecific structure. (iii) Both genetic and morphological data are available for this model to confront the structure provided by the two kinds of markers and explore their combination.

The data set consists of 182 individuals. These individuals were genotyped at 14 microsatellite loci (Lehanse 2010). The phenotypic data set corresponds to a subsample of 69 individuals (Ledevin 2010). We used measurements of the third upper molar shape for which phenotypic differentiation has been evidenced at the phylogeographic scale (Deffontaine et al. 2009; Ledevin, Michaux, et al. 2010). The two-dimensional outline was manually registered from numerical pictures, starting from a comparable starting point among teeth (Ledevin, Michaux, et al. 2010). For each molar, the outline is described by the Cartesian coordinates of 64 points sampled at equally spaced intervals along the outline. These 64 landmarks are strongly correlated and therefore carry redundant information. To summarize this information into a lower number of variables and decrease the intensity of correlation between variables, we first performed an elliptic Fourier transform (EFT, Kuhl and Giardina 1982). The EFT provides shape variables standardized by size, the Fourier coefficients that weight the successive functions of the EFT, namely the harmonics. A study of the successive contribution of each harmonic to the description of the original outline showed that considering the first 10 harmonics offered a good compromise between the number of variables and the efficient description of the outline (Ledevin, Michaux, et al. 2010). Then we performed a principal component analysis of the Fourier coefficients and retained the scores on the first five principal components, which together account for more than 80% of the variance (PC1 = 26.6%, PC2 = 21.6%, PC3 = 15.2%, PC4 = 7.4%, and PC5 = 6.5%). These scores were used as phenotypic data input (the $\mathbf{y}$ data matrix) to our clustering method.

We analyzed this data set with our model first under the UFM allele frequency prior then under the CFM prior. For each allele frequency prior, we fed the model with five types of data combination: using the georeferenced phenotypic data under the spatial model (PS), using the phenotypic data under the nonspatial model (PnS), using the georeferenced genetic data under the spatial model (GS), using the genetic data under the nonspatial model (GnS), using the georeferenced phe-

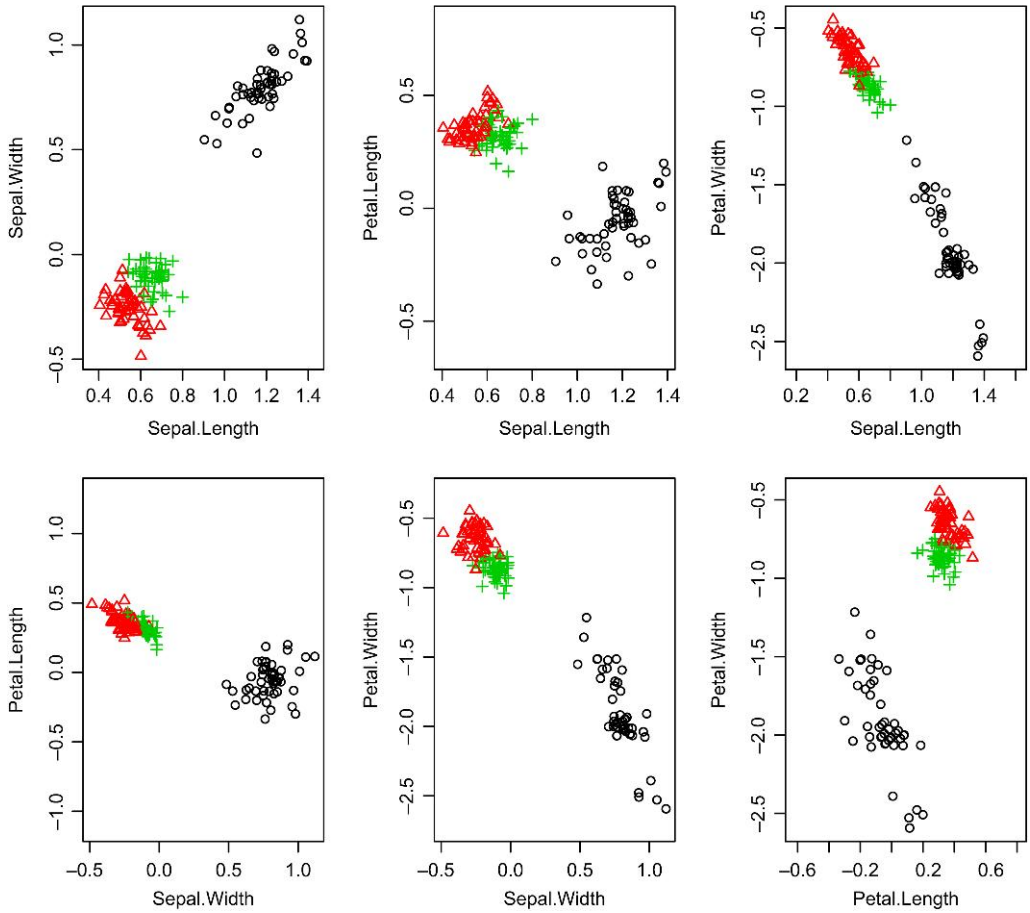

FIGURE 6. Pairs plots of Fisher's Iris data (transformed into log shape ratios). Symbol shapes indicate individual species estimated by our method. The true number of species (three) is correctly estimated. Only 6 out of 150 individuals are misclassified.

notypic and genetic data under the spatial model (PGS). In each case, we performed 10 independent MCMC runs of 100,000 iterations discarding the first 10,000 iterations as burn-in.

RESULTS

For each type of analysis, we observed an excellent congruence across the 10 independent MCMC runs. The UFM and the CFM model provide qualitatively similar results with a tendency of the CFM model to return slightly larger estimates of K. Although the CFM option has been shown to detect finer differentiation than the UFM option (see analysis of AFLP data above), a detailed analysis and interpretation of the fine scale structure inferred by the CFM model would require extended data analysis, including some extra data still under production. We therefore focus on the results obtained under the UFM option.

In the analysis based on georeferenced phenotypic data (PS), we inferred two clusters with one cluster in the far North of Sweden (Fig. 7a), all remaining samples belonging to the other cluster. These clusters correspond to the interspecific differentiation between the red-backed vole to the North and the bank vole to the South. Analyzing these data without spatial information (PnS), we also inferred also two clusters (Fig. 7b). The areas occupied by the two clusters under the PS and the PnS analyses match in the sense that they both correspond to a top North versus South dichotomy with a region of marked transition estimated to be along the same line in Swedish Lapland with a SW–NE orientation. In the PnS analysis, the clusters display a large amount of spatial overlap with a regular North to South cline. In the analysis based on georeferenced genetic data (GS), we inferred the presence of four clusters. The most northern cluster corresponds to the samples identified as belonging to the top North cluster in the phenotypic clustering and hence to the Northern red-backed vole (Fig. 7c). The three other clusters correspond to the intraspecific structure within the bank vole. This hierarchical pattern of inter- and intraspecific differences is confirmed by estimates of interpopulation differentiation provided by Fst values. The far North population attributed to the red-backed vole appears as strongly differentiated from all other populations (N Sweden vs. NE Sweden: $F_{ST} = 0.15$; N vs. Central Sweden: $F_{ST} = 0.19$; N Sweden vs. South Sweden: $F_{ST} = 0.17$). In comparison, the differentiation is of smaller magnitude

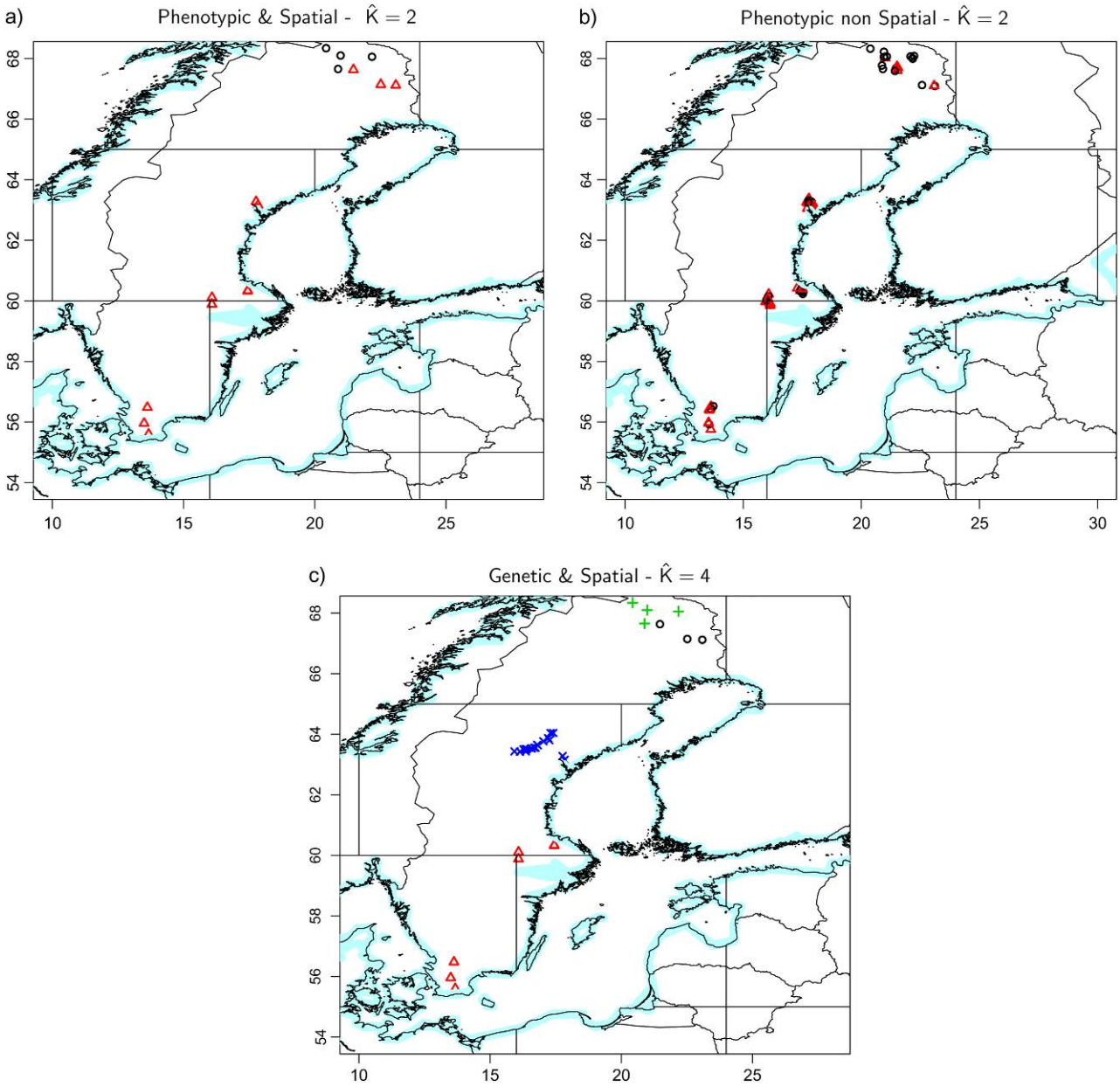

FIGURE 7. Population structure inferred from the Fenno–Scandia bank vole data. The upper titles refer to the number of clusters detected with the various combinations of data.

among bank vole populations (NE vs. C: $F_{ST} = 0.07$; NE vs. S: $F_{ST} = 0.07$; C vs. S: $F_{ST} = 0.06$). Analyzing these data without spatial information (GnS), we inferred four clusters whose locations match tightly those obtained under analysis GS (results not shown). In the joint analysis of georeferenced phenotypic and genotypic data (PGS), we obtained results similar to those obtained with georeferenced genetic data (results not shown).

DISCUSSION

Summary of Approach Proposed

Main features.—We have proposed the first method to date for analyzing georeferenced phenotypic and genetic data within a unified inferential framework, opening the way to combined analyses and robust comparison between markers. Our method takes as input any combination of phenotypic and genetic individual data and these data can be optionally georeferenced. Analyses can be run on phenotypic and genetic data separately or jointly. The main outputs of the method are estimates of the number of homogeneous clusters and of cluster memberships of each individual. If analyses are made on georeferenced data, the method also provides an estimate of the spatial location of each cluster which can be displayed graphically in form of continuous maps (see program documentation for details on such graphic representation).

Our approach is based on an explicit statistical model. This contrasts with model-free methods such as PAM (Kaufman and Rousseeuw 1990) which, roughly speaking, attempts to cluster individuals in order to maximize some homogeneity criterion. Although such methods are fast and presumably robust to departure from specific model assumptions, they are expected to behave poorly compared with methods based on an explicit model that fits the data to a reasonable extent. This claim is supported by the recent study of Safner et al. (2011) in the case of spatial genetic clustering methods. In addition, because model-free methods do not rely on an explicit model, their output might be difficult to interpret or relate to biological processes.

Main results from simulation study and analysis of classic data sets.—All numerical results obtained here demonstrate the good accuracy of our method and its efficiency for identifying species and/or populations boundaries. It is excellent at avoiding false positives (i.e., at reporting $\hat{K} = 1$ when $K = 1$) and has a clear ability to reduce the error rate when the number of variables increases. The method loses accuracy when it is given a difficult problem (i.e., when the true K is large). For a fixed number of iterations, it also has increasing difficulty to exploit fully all the available information when the number of variables is large (cf. loss of accuracy for 50 variables compared with 20 variables), presumably due to loss of numerical efficiency in the MCMC algorithm. We also noted that MCLUST is subject to similar difficulties for large number of clusters and/or large number of variables presumably due to the existence of multiple maxima of the likelihood. In our method, this problem can be resolved to a certain extent by longer MCMC runs, an aspect not investigated in detail here. Overall, our method offers a notable improvement over the nonspatial penalized maximum likelihood method of MCLUST used under its default set of options. One factor responsible for this improvement could be that our method exploits spatial information while MCLUST does not. Results from the section Analysis of Classic Data of the Clustering Literature, where our method still provides better results than MCLUST even though the data are nonspatial, suggests this is not the sole factor. This might relate to model selection, which is the second major difference between the two methods considered (Bayes vs. penalized maximum likelihood), more so given that MCLUST considers a broad family of covariance structure, whereas our method assumes conditional independence.

We also stress that the numerical values characterizing the accuracy of our method are biased because the model used to analyze the data matches exactly the model that generated them. This situation is a best-case scenario that is unlikely to be met in real-life cases. However, our results are informative about the potential of the method and evaluations of the Iris data suggest a certain robustness of these results (see also analysis of crab morphometric data in Supplementary Material).

As a final note, we warn the reader unfamiliar with clustering methods against overly pessimistic interpretation of Figure 4. From this figure, it seems that the methods lose accuracy very quickly as the "phenotypic differentiation" decreases and are, in general, not very efficient.

This is because detecting a hidden structure is a much harder statistical problem than testing the significance of a differentiation between two known clusters (the former involving many more parameters and hence uncertainty than the latter). More details are given in Power to Test the Significance of a Known Structure Versus Power to Detect a Hidden Structure section of Supplementary Material.

Genetic and phenotypic data can trace different evolutionary histories, for instance, phylogenetic divergence for neutral genetic markers and adaptation for a morphological structure (Renaud et al. 2007; Adams et al. 2009).

Confronting the structure provided by different markers emerges more and more as a way to get a comprehensive view of the dynamics and processes of differentiation among and within species. Our method, by providing a unified inferential framework for analyzing different kind of data, including phenotypic data, appears as a significant improvement for valid confrontation between data sets. Furthermore, in situations when genetic and phenotypic patterns are suspected to coincide, making inference from genetic and phenotypic data jointly has the potential to increase the power to detect boundaries between evolutionary units at different levels (populations, species).

Analysis of the Calopogon *AFLP data set.*—The ability of our model under the CFM prior to detect and classify species is excellent. This data set has been reanalyzed by Hausdorf and Hennig (2010) who carried out a comparison of STRUCTURE, STRUCTURAMA, a method known as "field of recombination" (Doyle 1995) and a hybrid method mixing sequentially multidimensional scaling and model-based Gaussian clustering. The STRUCTURE program and the field of recombination method were not able to detect any structure. STRUCTURAMA identified only three clusters and misclassifies 44% of the samples. The hybrid method of Hausdorf and Hennig (2010) identifies five clusters but misclassifies 15% of the samples. Our method under the CFM prior also identifies five clusters but misclassifies only 5% of the samples. Under the UFM model, the results we obtain are highly consistent with those obtained with the CFM.

We also refer the reader to the Supplementary Material where we analyze AFLP data of *Veronica (pentasepalae)* from the Iberian Peninsula and Morocco produced and first analyzed by Martínez-Ortega et al. (2004). The results we report there confirm the excellent performance of our method compared with the four methods investigated by Hausdorf and Hennig (2010). Finally, all the analysis carried out in the present

article show that concerns of Hausdorf and Hennig (2010) against methods for dominant markers based on Hardy–Weinberg equilibrium were not grounded, provided the dominant nature of AFLP markers is taken into account at the likelihood level as we did. We suspect that the poor performances of STRUCTURE observed by Hausdorf and Hennig (2010) relate to the procedure used to estimate *K* (Evanno et al. 2005), as noted earlier by Waples and Gaggiotti (2006).

The Myodes *data set*.—We confronted clustering hypotheses using various data subsets with or without spatial data and with or without genetic markers or morphometric variables. This sheds new light on the population structure of *Myodes*. The pattern of phenotypic and genetic differentiation reveals a complex pattern of contact between species and populations. The northernmost area corresponds to the narrow zone of possible overlap between *M. glareolus* and its close northern relative *M. rutilus*. Both species are difficult to recognize based on external phenotypic characters and impossible to identify based on common mitochondrial markers because of the introgression of *M. rutilus* mtDNA into the northern fringe of *M. glareolus* distribution. The northern cluster detected by our method corresponds most probably to the occurrence of the northern red-backed vole *M. rutilus* that tends to differ in molar shape from its relative *M. glareolus* (Ledevin, Quéré, et al. 2010).

The two analyses based on phenotypic data with and without spatial information lead to slightly different results, the former suggesting the presence of an abrupt phenotypic discontinuity in the North, whereas the latter suggests clinal variation (Fig. 7a,b). In the absence of model-fit criteria to assess the value of these two maps, we can only speculate. We note however that these maps are congruent concerning the location of the main area of transition between the clusters and that the analysis based on spatial information is graphically more efficient at displaying the location of this transition. Molar shape in bank vole has been shown to display large variation even within populations due to wear and developmental factors (Guérécheau et al. 2010; Ledevin, Quéré, et al. 2010). This may render even clear-cut interspecific boundaries difficult to detect. Our georeferenced method may greatly help to make such signal emerge despite the intrinsic variability in the phenotypic markers. This suggests that our method could be viewed as an efficient generalization of the methods of Womble (1951) and Bocquet-Appel and Bacro (1994) aimed at detecting abrupt changes.

Regarding the additional clusters detected based on genetic data, the location of two of them suggests that they correspond to bank vole lineages already known in this region based on mtDNA data. Indeed, after the last ice age, Sweden has been recolonized by different populations separated several hundreds of thousand years ago coming from the South and from the North of Fennoscandia (Jaarola et al. 1999; Razzauti et al. 2009). Our new data therefore confirm the existence of two different bank vole lineages in Sweden based on mtDNA and now nuclear DNA markers. The existence of a fourth cluster located in Central Sweden strongly suggests that the contact zone between these two main lineages is situated in this latter region. Its origin may be attributed to hybridization between animals of the two genetic lineages. The discovery of this last cluster is new and it was never detected previously using only mtDNA marker.

Combining phenotypic and genetic data in a joint analysis (PGS) did not allow us to detect any extra structure (map not shown) possibly because beyond the interspecific phenotypic difference corresponding to the differentiation between far-North and the rest of Sweden, a cline in molar shape exists through Sweden that is roughly congruent with the genetic clusters (data not shown). It shows that the confrontation between data sets may be as informative as a joint analysis, by providing clues about the hierarchical pattern of differentiation. Morphometric clusters revealed only interspecific differences between red-backed and bank voles, whereas based on microsatellite data, both inter- and intraspecific levels of differentiation emerged as separate clusters. The structure of genetic differentiation corroborates this interpretation. The interspecific differentiation of the top North cluster from the rest of Sweden is indeed much stronger than the intraspecific differentiation among the bank vole populations from North-East, Central, and South Sweden. Combining both data types allows us to interpret the complex phylogeographic structure of this species and helps to distinguish differences between true species and populations within a species.

Future Extensions

Our method is based on an assumption of independence of the phenotypic variables within each cluster. This does not amount to independence between these variables globally. Indeed, the fact that phenotypic variables are sampled with cluster-specific parameters does include a correlation (similar to the dependence structure assumed in a linear mixed model). However, our method does not deal with residual dependence not accounted for at the cluster level such as that generated by allometry. Results from simulations and classic datasets suggest that this can be partially dealt with by preprocessing the data (e.g., transforming raw data into log-shape ratio). Several other procedures may be applied for avoiding or reducing problems with covariation among phenotypic variable. For example, working on principal components rather than on raw data may help in this task. Procedures such as the Burnaby approach (Burnaby 1966) may also allow removal of covariance structures due only to growth or other confounding factors that the user may wish to filter out. A more rigorous approach would be to allow the variables to covary within clusters, which would also allow one to quantify these covariations.

Potential Applications

Evolutionary biology has been flooded by molecular data in the recent years. However, efficient methods to deal with phenotypic data alone are still needed when this type of data is the only available. This includes the important case of fossil data. We note that in systematic paleontology, the methods used are often simpler than those discussed in the present paper and chosen as a matter of tradition in the field rather than on objective basis. Implementing our method in a free and user-friendly program should help provide more objective methods in this context.

Our method was specifically tailored for biometric/morphometric measurements that are typically obtained from a few tens of phenotypic variables. The method proposed is therefore computer intensive and not expected to be well suited for large data sets, such as expression data produced in functional genomics. However, in situations where the scientist is able to select some variables of particular interest and reduce the dimensionality of the model (as we did for our analysis of the *Myodes* molar shape data), our method could be used and play a role in the emerging field of landscape genomics (Schwartz et al. 2010).

The submodel for genetic data used here was presented and discussed in detail by Guillot et al. (2005) and Guillot (2008). It has been used mostly to analyze variation and structure in neutral nuclear markers (Guillot et al. 2009) and has proved useful to detect and quantify fine-scale structure typical of landscape genetics studies. The novel possibility brought here to combine it with morphometric data might popularize this genetic model among scientists interested in larger spatial and temporal scales typical of phylogeography. In the latter field, the use of mtDNA is common. As noted earlier, the analysis of such non-recombining DNA sequence data using our method is technically possible and meaningful by recoding the various observed haplotypes as different alleles of the same locus. We stress that this approach is an expedient which incurs a considerable loss of information and that our approach should not be viewed as a substitute to those that model the genealogy of genes (including the mutational process) explicitly. Extending our model to deal with nonrecombining DNA in a more rigorous way is a natural direction for future work.

Our method for the combined analysis of phenotypic and genetic data can be used to assess the relative importance of random genetic drift and directional natural selection as causes of population differentiation in quantitative traits and to assess whether the degree of divergence in neutral marker loci predicts the degree of divergence in quantitative traits (Merilä and Crnokrak 2001). Furthermore, our method should be useful in the study of hybrid zones where, as noted by Gay et al. (2008), comparing clines of neutral genetic markers with clines of traits known to be under selection also indicates the extent to which specific traits or their underlying genes are under selection.

Lastly, because phenotypic and genetic markers may reflect different evolutionary or demographic history, combined analyses can help to understand the hierarchy between evolutionary units (species and populations) as shown in the *Myodes* example.

Computer Program Availability

The algorithm presented here is available as the R package GENELAND (version $\geq$ 4.0.0). Information will be found on the program homepage: http://www2.imm.dtu.dk/~ gigu/Geneland/.

Supplementary Material

Supplementary Material can be found in the Dryad data repository (doi: 10.5061/dryad.9mp902ck).

Funding

This work has been supported by the French National Research Agency (project EMILE, ANR-09-BLAN-0145-01) and the Danish Centre for Scientific Computing (2010-06-04). This study was partially funded by EU grant FP7-261504 EDENext and is catalogued by the EDENext Steering Committee as EDENext027 (http://www.edenext.eu).

Acknowledgments

The first author is most grateful to Cino Pertoldi for discussions that prompted him to develop the model for morphometric data. Our work benefitted from discussions with Jean-Marie Cornuet and comments of Andrew J. Crawford. Part of the original data of the *Myodes* analysis belong to Bernard Lehanse's Master thesis (genetic data). We thank him for sharing these data with us. We are also grateful to Montse Martínez-Ortega and Doug Goldman for making their data available to us. The contents of this publication are the sole responsibility of the authors and don't necessarily reflect the views of the European Commission.

References

Adams D.C., Berns C.M., Kozak K.H., Wiens J.J. 2009. Are rates of species diversification correlated with rates of morphological evolution? Proc. R. Soc. Lond. Biol. Sci. Ser. B 276:2729–2738.

Anderson E. 1935. The irises of the Gaspe peninsula. Bull. Am. Iris Soc. 59:2–5.

Anderson T. 1984. An introduction to multivariate statistical analysis. Probability and mathematical statistics. 2nd ed. New York: Wiley.

Balding D., Nichols R. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 96:3–12.

Balding D., Nichols R. 1997. Significant genetic correlation among Caucasians at forensic DNA loci. Heredity. 78:583–589.

Banfield J.D., Raftery A.E. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics. 49:803–821.

Beadell J.S., Hyseni C., Abila P.P., Azabo R., Enyaru J.C.K., Ouma J.O., Mohammed Y.O., Okedi L.M., Aksoy S., Caccone A. 2010. Phylogeography and population structure of *Glossina fuscipes fuscipes* in Uganda: implications for control of tsetse. PLoS Neglected Trop. Dis. 4.

Biernacki C., Celeux G., Govaert G., Langrognet F. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. Comput. Stat. Data Anal. 51:587–600.
Bocquet-Appel J., Bacro J. 1994. Generalized wombling. Syst. Biol. 43:442–448.
Burnaby T.P. 1966. Growth-invariant discriminant functions and generalized distances. Biometrics. 22:96–110.
Claude J. 2008. Morphometrics with R. New York: Springer.
Corander J., Waldmann P., Martinen P., Sillanpää M. 2004. Baps2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics. 20:2363–2369.
Corander J., Waldmann P., Sillanpää M. 2003. Bayesian analysis of genetic differentiation between populations. Genetics. 163:367–374.
Coulon A., Guillot G., Cosson J., Angibault J., Aulagnier S., Cargnelutti B., Galan M., Hewison A. 2006. Genetics structure is influenced by landscape features. Empirical evidence from a roe deer population. Mol. Ecol. 15:1669–1679.
Deffontaine V., Ledevin R., Fontaine M.C., Quéré J.-P., Renaud S., Libois R., Michaux J.R. 2009. A relict bank vole lineage highlights the biogeographic history of the Pyrenean region in Europe. Mol. Ecol. 18:2489–2502.
Doyle J. 1995. The irrelevance of allele tree topologies for species delimitation and a non-topological alternative. Syst. Biol. 20:574–588.
Evanno G., Regnault S., Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14:2611–2620.
Excoffier L. Heckel G. 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7:745–758.
Falush D., Stephens M., Pritchard J. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567–1587.
Fisher R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics. 7:179–188.
Fontaine M., Baird S., Piry S., Ray N., Tolley K., Duke S., Birkun A., Ferreira M., Jauniaux T., Llavona A., Östürk B., Östürk A., Ridoux V., Rogan E., Sequeira M., Siebert U., Vikingson G., Bouquegneau J., Michaux J. 2007. Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters. BMC Biol. 5.
Fraley C. 1999. Mclust: software for model-based cluster analysis. J. Classification. 12:297–306.
Frühwirth-Schnatter S. 2006. Finite mixture and Markov switching model. Series in Statistics. New York: Springer.
Galarza J., Carreras-Carbonell J., Macpherson E., Pascual M., Roques S., Turner G., Ricod C. 2009. The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. Proc. Natl. Acad. Sci. U.S.A. 106:1473–1478.
Gay L., Crochet P., Bell D., Lenormand T. 2008. Comparing genetic and phenotypic clines in hybrid zones: a window on tension zone models. Evolution. 62:2789–2806.
Gelman A., Carlin J., Stern H., Rubin D. 2004. Bayesian data analysis. New York: Chapman and Hall.
Gilks W., Richardson S., Spiegelhalter D. editors. 1996. Markov chain Monte Carlo in practice. Interdisciplinary Statistics. New York: Chapman and Hall.
Girard C., Renaud S. 2011. The species concept in a long-extinct fossil group, the conodonts. Comptes Rendus Palevol. 10:107–115.
Godsill S. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graph. Stat. 10:230–248.
Goldman D.H., Jansen R.K., Van Den Berg C., Leitch I.J., Fay M.F., Chase M.W. 2004. Molecular and cytological examination of calopogon *(Orchidaceae, Epidendroideae)*: circumscription, phylogeny, polyploidy, and possible hybrid speciation. Am. J. Bot. 91:707–723.
Guérécheau A., Ledevin R., Henttonen H., Deffontaine V., Michaux J.R., Chevret P., Renaud S. 2010. Seasonal variation in molar outline of bank voles: an effect of wear? Mammalian Biol. 75:311–319.
Guillot G. 2008. Inference of structure in subdivided populations at low levels of genetic differentiation. The correlated allele frequencies model revisited. Bioinformatics. 24:2222–2228.
Guillot G., Carpentier-Skandalis A. 2011. On the informativeness of dominant and co-dominant genetic markers for Bayesian supervised clustering. Open Stat. Probab. J. 3:7–12.
Guillot G., Estoup A., Mortier F., Cosson J. 2005. A spatial statistical model for landscape genetics. Genetics. 170:1261–1280.
Guillot G., Leblois R., Coulon A., Frantz A. 2009. Statistical methods in spatial genetics. Mol. Ecol. 18:4734–4756.
Guillot G., Santos F. 2010. Using AFLP markers and the Geneland program for the inference of population genetic structure. Mol. Ecol. Resour. 10:1082–1084.
Guillot G., Santos F., Estoup A. 2008. Analysing georeferenced population genetics data with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. Bioinformatics. 24:1406–1407.
Hannelius U., Salmela E., Lappalainen T., Guillot G., Lindgren C., von Döbeln U., Lahermo P., Kere J. 2008. Population substructure in Finland and Sweden revealed by a small number of unlinked autosomal SNPs. BMC Genet. 9.
Hausdorf B., Hennig C. 2003. Biotic element analysis in biogeography. Syst. Biol. 52:717–723.
Hausdorf B., Hennig C. 2010. Species delimitation using dominant and codominant multilocus markers. Syst. Biol. 59:491–503.
Hubisz M., Falush D., Stephens M., Pritchard J.K. 2009. Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9:1322–1332.
Jaarola M., Tegelström H., Fredga K. 1999. Colonization history in Fennoscandian rodents. Biol. J. Linn. Soc. 68:113–127.
Joseph L., Dolman G., Donnellan S., Saint K., Berg M., Bennett A. 2008. Where and when does a ring start and end? Testing the ring-species hypothesis in a species complex of Australian parrots. Proc. R. Soc. Lond. Ser. B. 275:2431–2440.
Kaufman L., Rousseeuw P.J. 1990. Finding groups in data: an introduction to cluster analysis. New York: Wiley.
Klingenberg C.P., Leamy L.J., Routman E.J., Cheverud J.M. 2001. Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. Genetics. 157: 785–802.
Kuhl F.P., Giardina C.R. 1982. Elliptic Fourier features of a closed contour. Comput. Graph. Image Process. 18:236–258.
Ledevin R. 2010. La dynamique évolutive du campagnol roussâtre *(Myodes glareolus)*: structure spatiale des variations morphométriques. [dissertation]. Lyon (France): Université Lyon 1.
Ledevin R., Michaux J.R., Deffontaine V., Henttonen H., Renaud S. 2010. Evolutionary history of the bank vole *Myodes glareolus*: a morphometric perspective. Biol. J. Linn. Soc. 100:681–694.
Ledevin R., Quéré J.-P., Renaud S. 2010. Morphometrics as an insight into processes beyond tooth shape variation in a bank vole population. PLoS One. 5:e15470.
Lehanse B. 2010. Étude génétique d'une zone de contact en Suède entre deux lignées de campagnols roussâtres *Myodes glareolus* [Master's thesis]. Liège (Belgium): Université de Liège.
Luo A., Zhang A., Ho S.Y., Xu W., Shi W., C. S.L., Zhu C. 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. BMC Genomics. 12.
Marin J., Mengersen K, Robert C. 2005. Bayesian modelling and inference on mixtures of distributions. Handbook of Statistics. Vol. 25. New York: Elsevier Sciences.
Martínez-Ortega M.M., Delgado L., Albach D.C., Elena-Rossello J.A., Rico E. 2004. Species boundaries and phylogeographic patterns in cryptic taxa inferred from AFLP markers: *Veronica subgen. Pentasepalae (Scrophulariaceae)* in the Western Mediterranean. Syst. Bot. 29:965–986.
McLachlan G.J., Peel D., Basford P., Adams K.E. 1999. The EMMIX software for the fitting of mixtures of normal and t-components. J. Stat. Softw. 4:1–4.
Merilä J., Crnokrak P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. J. Evol. Biol. 14:892–903.
Møller J., Stoyan D. 2009. Tessellations in the sciences: virtues, techniques and applications of geometric tilings. Stochastic Geometry and Random Tessellations. New York: Springer.
Néraudeau. 2011. The species concept in palaeontology: ontogeny, variability, evolution. Comptes Rendus Palevol. 10:71–75.
Nesi N., Nakoumé E., Cruaud C., Hassanin A. 2011. DNA barcoding of African fruit bats (*Mammalia, Pteropodidae*): the mitochondrial genome does not provide a reliable discrimination between

Epomophorus gambianus and *Micropteropus pusillus*. Comptes Rendus Biologies. 334:544–554.
Pritchard J., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.
Razzauti M., Plyusnina A., Sironen T., Henttonen H., Plyusnin A. 2009. Analysis of Puumala hantavirus in a bank vole population in northern Finland: evidence for co-circulation of two genetic lineages and frequent reassortment between strain. J. Gen. Virol. 90:1923–1931.
Renaud S., Chevret P., Michaux J. 2007. Morphological vs. molecular evolution: ecology and phylogeny both shape the mandible of rodents. Zool. Script. 36:525–535.
Richardson S., Green P. 1997. On Bayesian analysis of mixtures with an unknown number of components. J.R. Stat. Soc. Ser. B. 59:731–792.
Robert C., Casella G. 2004. Monte Carlo statistical methods. 2nd ed. New York: Springer.
Rodríguez F., Pérez T., Hammer S.E., Albornoz J., Domínguez A. 2010. Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus *Rupicapra*). BMC Evol. Biol. 10:222.
Sacks B., Bannasch D. L., Chomel B. B., Ernst H. 2008. Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. Mol. Biol. Evol. 25:1354–1395.
Safner T., Miller M., McRae B., Fortin M., Manel S. 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. Int. J. Mol. Sci. 12:865–889.
Schwartz M.K., Luikart G., McKelvey K.S., Cushman S.A. 2010. Spatial complexity, informatics, and wildlife conservation chap. In: Cushman S.A., Huettmann F., editors. Landscape genomics: a brief perspective. New York: Springer. p. 165–174.
Sirén J., Marttinen P., Corander J. 2011. Reconstructing population histories from single nucleotide polymorphism data. Mol. Biol. Evol. 28:673–683.
Stephens M. 1997. Discussion of the paper by Richardson and Green "On Bayesian analysis of mixtures with an unknown number of components". J. R. Stat. Soc. Ser. B. 59:768–769.
Tegelström H. 1987. Transfer of mitochondrial DNA from the northern red-backed vole (*Clethrionomys rutilus*) to the bank vole (*C. glareolus*). J. Mol. Evol. 24:218–227.
Turmelle A.S., Kunz T.H., Sorenson M.D. 2011. A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. Mol. Ecol. 20:357–375.
Waples R., Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15:1419–1439.
Wasser S., Mailand C., Booth R., Mutayoba B., Kisamo E., Stephens M. 2007. Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. Proc. Natl. Acad. Sci. U.S.A. 104: 4228–4233.
Womble W. 1951. Differential systematics. Science. 28:315–322.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107:9264–9269.

Appendix: Detail of MCMC Inference Algorithm

Overview

The vector of unknown parameters is $\boldsymbol{\theta} = (K, \lambda, m, \mathbf{u}, \mathbf{c}, \mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ which can be decomposed into $\boldsymbol{\theta}_S = (\lambda, m, \mathbf{u}, \mathbf{c})$, $\boldsymbol{\theta}_G = (\mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d})$, and $\boldsymbol{\theta}_M = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ blocks of parameters of the spatial, genetic, and phenotypic data, respectively. We alternate block updates of Metropolis–Hastings or Gibbs type and also transdimensional updates involving changes of K and of parts of other parameters. The updates of blocks of parameters that do not involve phenotypic data are described in Guillot et al. (2005) and Guillot (2008). We describe below updates involving phenotypic data.

Joint Updates of $(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma})$

We update jointly $\mathbf{c}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ as follows. We propose a new vector $\mathbf{c}^*$ by picking two clusters at random and reassigning some individuals of one of those two clusters to the other one at random. Then we propose $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ by sampling from the full conditional distribution $\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{y}, \mathbf{c}^*)$. The Metropolis–Hastings ratio is

$$\begin{aligned} R &= \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \\ &= \frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}, \mathbf{c}^*|\mathbf{y})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2, \mathbf{c}|\mathbf{y})}\frac{q(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c})}{q(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*)}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \\ &= \frac{\pi(\mathbf{c}^*|\mathbf{y})}{\pi(\mathbf{c}|\mathbf{y})}\frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*, \mathbf{y})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c}, \mathbf{y})}\frac{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c}, \mathbf{y})}{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*, \mathbf{y})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \\ &= \frac{\pi(\mathbf{c}^*|\mathbf{y})}{\pi(\mathbf{c}|\mathbf{y})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \end{aligned} \tag{A.1}$$

Interestingly, the latter expression does not depend on $(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2^*})$, which in principle would allow us to decide whether a new state $\boldsymbol{\theta}^*$ is accepted prior to proposing $(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2^*})$. Unfortunately, expression (A.1) cannot be used, as $\pi(\mathbf{c}|\mathbf{y})$ is not known analytically under the present model. The ratio in equation (A.1) has therefore to be written as

$$\begin{aligned} R = &\frac{\pi(\mathbf{y}|\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}, \mathbf{c}^*)}{\pi(\mathbf{y}|\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2, \mathbf{c})}\frac{\pi(\mathbf{c}^*)}{\pi(\mathbf{c})} \\ &\frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2)}\frac{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{y}, \mathbf{c})}{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{y}, \mathbf{c})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})}, \end{aligned} \tag{A.2}$$

which involves only analytically known expressions.

Joint Updates of $(K, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma})$

We take the same strategy as Guillot et al. (2005). The algorithm follows ideas of Richardson and Green (1997). It consists in updating K by proposing to split a cluster into two clusters or merge two clusters in a way that complies with the spatial constraints and multivariate nature of the model. Since we use the natural prior conjugate family for parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}^*$, the full conditional $\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{y}, K^*, \mathbf{c}^*)$ is available and can be used as proposal distribution as advocated for example by Godsill (2001). The acceptance ratio takes essentially the same form as in equation (A.2), although it is now a genuine transdimensional move.

Detail on Hyperparameters

Although we do not use exactly the same prior structure as Richardson and Green (1997), we follow largely these authors. We take $\xi_j = \sum_i y_{ij}$, $h_j = \kappa_j = 2/R_j^2$, where R_j is the range of observed values of the jth phenotypic variable. $\beta_j|g_j, h_j, \sim \mathcal{G}(g_j, h_j)$. We also set $\alpha_j = 2$ and $g_j = 1/2$. Since $E[1/\sigma^2] = \alpha/\beta$, β represents $2/E[1/\sigma^2]$. Also $1/2h$ represents the prior mean of beta.