Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance?

A. C. Frantz1*, S. Cellina2, A. Krier2, L. Schley2 and T. Burke1

1Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK; 2Direction des Eaux et Forêts, 16 rue Eugène Ruppert, L-2453 Luxembourg

Summary

1. Spatially explicit Bayesian clustering techniques offer a powerful tool for ecology and wildlife management, as genetic divisions can be correlated with landscape features. We used these methods to analyse the genetic structure of a population of European wild boar Sus scrofa with the aim of identifying effective barriers for disease management units. However, it has been suggested that the methods could produce biased results when faced with deviations from random mating not caused by genetic discontinuities, such as isolation by distance (IBD).

2. We analysed a data set consisting of 697 wild boar multilocus genotypes using spatially explicit (baps, geneland) and non-explicit (structure) Bayesian methods. We also simulated and analysed data sets characterized by different degrees of IBD, with and without genetic discontinuities.

3. When analysing the empirical data set, different programs did not converge on the same clustering solution and some clusters were difficult to explain biologically. Results from the simulated data showed that IBD, also present in the empirical data set, could cause the Bayesian methods to overestimate genetic structure. Simulated barriers were identified correctly, but the programs superimposed further clusters at higher IBD levels.

4. It was not possible to ascertain with confidence whether the clustering solutions offered by the various programs were an accurate reflection of population genetic structure in our empirical data set or were artefacts created by the underlying IBD pattern.

5. Synthesis and applications: We show that Bayesian clustering methods can overestimate genetic structure when analysing an individual-based data set characterized by isolation by distance. This bias could lead to the erroneous delimitation of management or conservation units. Investigators should be critical and suspicious of clusters that cannot be explained biologically. Data sets should be tested for isolation by distance and conclusions should not be based on the output from just one method.

Key-words: classical swine fever, landscape genetics, spatial genetic structure, Sus scrofa, translocation, wildlife diseases, wildlife forensics

Introduction

Bayesian clustering algorithms are prominent computational tools for inferring genetic structure in molecular ecology. Mostly, these methods probabilistically assign individuals to groups based on their multi-locus genotypes by minimizing Hardy–Weinberg and linkage disequilibria, without presuming pre-defined populations (Pearse & Crandall 2004). Recent advances explicitly address the spatial nature of the problem of locating genetic discontinuities by including the geographical coordinates of individuals in their prior distributions (e.g., Guillot et al. 2005; Corander, Sirén & Arjas 2008). These models offer a powerful tool to answer questions in ecology, conservation and wildlife management, as genetic discontinuities within populations can be correlated with landscape features.

There have been few studies to evaluate the performance of the various spatial Bayesian methods in biologically realistic scenarios and there is uncertainty about possible methodological biases. Of particular concern in this context is isolation by distance (IBD) – the regular increase in genetic differentiation among individuals with geographical distance due to limited dispersal – which has been shown to potentially confound the non-spatial Bayesian methods. For example, investigations of global human genetic diversity have highlighted that the structure algorithm (Pritchard, Stephens & Donnelly 2000) may detect non-existent clusters when geographical sampling...
is clumped along an IBD cline (Serre & Pääbo 2004; Rosenberg et al. 2005). Similarly, Schwartz & McKelvey (in press) have recently shown that STRUCTURE may superimpose artificial clusters on data sets solely characterized by an IBD cline, even when geographical sampling is conducted evenly over the more local scale typical of many ecological studies.

The issue as to whether the spatially explicit Bayesian algorithms are similarly biased, despite imposing additional spatial constraints on a clustering solution, has yet to be tested. Some authors have suggested that, when a data set is characterized by isolation by distance, taking the spatial context of individuals into consideration might improve the efficiency of the analysis (Frantz et al. 2006; Fontaine et al. 2007). However, Guillot et al. (2005) hypothesized that IBD (as well as other deviations from random mating not caused by genetic discontinuities) could negatively affect the performance of their GENELAND application. In particular, the program might overestimate genetic clustering of the data and not be capable of correctly detecting and locating a genuine genetic discontinuity. Consequently, while it has become possible to test hypotheses concerning barriers to dispersal and gene flow, the reliability of spatial Bayesian programs when analysing data sets characterized by an IBD pattern needs further clarification. This is important, as inappropriate clustering can lead to the erroneous delimitation of management or conservation units.

Here, our initial objective was to analyse the genetic structure of wild boar Sus scrofa L. in Wallonia, Luxembourg, and the Rhineland-Palatinate where there have been recent outbreaks of classical swine fever, a highly contagious viral disease of pigs (Schoos 2002). Prevention and control of infection in wild boar is of great importance, as diseased individuals represent a permanent threat to domestic pig populations and the large-scale culling necessary to control the disease on pig farms can cause major economic losses (Artois et al. 2002). We therefore aimed to use spatially explicit Bayesian clustering methods to correlate genetic discontinuities with landscape features and thereby identify geographical barriers to gene flow that could be used to effectively delimit management units (see Anonymous 1999). However, given the continuous distribution of the species in the study area, and considering that individual European wild boar rarely disperse beyond 20 km (e.g., Truvé & Lemel 2003), it was likely that the data set would exhibit an IBD pattern. Given the unknown reliability of the spatial methodologies when confronted with such data, the issue of artificial clustering needed to be taken seriously. We therefore simulated data sets characterized by different levels of IBD, with and without genetic discontinuities. The main objective of these simulations was to test whether the spatial Bayesian methods could overestimate genetic structure when faced with IBD and, consequently, fail to detect genuine genetic discontinuities.

Methods

STUDY AREA

We aimed to investigate the genetic structure of wild boar from Luxembourg, Wallonia, and the north-west of Rhineland-Palatinate (Fig. 1), an area covering approximately 19 500 km² with extensive forest cover accounting for 32 3% of the area of Wallonia (Perrin, Temmerman & Laitat 2000), 34 5% of Luxembourg (Rondeux 2005) and 42 6% of the Rhineland-Palatinate (von Rüden 2006). There are a number of potential barriers to wild boar dispersal in the study area (Fig. 1). The Moselle river valley has previously been identified as a dispersal-barrier for red deer Cervus elaphus L. (Frantz et al. 2006). A number of fenced four-lane motorways, all < 30 years old, dissected the study area (Fig. 1). Finally, about 10 years ago, wild boar with an abnormal behaviour suddenly appeared in two hunting areas in southern Luxembourg, from which the species had previously been absent (Fig. 1). Clandestine translocation was suspected, but we did not have a priori suspects among our samples.

LABORATORY PROCEDURES

Tissue samples (spleen, ear or muscle) were either frozen or stored in 70% ethanol. DNA extractions were performed following Whitlock et al. (2008). Genotyping was performed using 14 unlinked microsatellite loci (Hampton et al. 2004; S0002, S0005, S0026, S0090, S0097, S0155, S0226, Sw122, Sw240, Sw632, Sw857, SW911, Sw936, Sw951). The loci were amplified in three multiplexed Polymerase Chain Reactions (PCR). Information on which loci were amplified together and the detailed conditions for these reactions can be found in Supporting Information, Table S1 and Appendix S1. Reactions were performed using a DNA Engine Tetrad thermocycler (Bio-Rad, Hercules, California, USA). Fragments were separated using an ABI 3730 automated DNA sequencer (Applied Biosystems, Carlsbad, California, USA) and the data were analysed using GENEMAPPER version 3.5 (Applied Biosystems).

If one locus in a multiplex failed to amplify, the whole multiplex was re-genotyped. In the majority of cases, it was possible to amplify all the loci in this second round of amplification and these results were retained for the final data set. If a locus in a multiplex failed again, only genotypes with an identical score in the first and second amplification were retained. In order to assess genotyping errors, 40 samples (out of a total of 697) were chosen randomly from the data base, re-extracted and re-genotyped, while 10 of the initial extracts were genotyped in duplicate. Allelic mismatches were identified by comparing these 50 duplicate genotypes to the initial ones.

POPULATION GENETIC ANALYSES

We calculated observed (H0) and expected (H1) heterozygosities (Nei 1978) for each locus, as well as the average expected heterozygosity, using GENETIX 4.05-2 (Belkhir 2004). The data were tested for linkage disequilibrium using an exact test based on a Markov chain method as implemented in GENEPop 3.4 (Raymond & Rousset 1995). The same program was used to perform the exact tests of Guo & Thompson (1992) for deviations from Hardy–Weinberg (HW) genotypic proportions at each locus. The sequential Bonferroni technique was used to eliminate significance by chance (Rice 1989).

POPULATION GENETIC STRUCTURE

We used different Bayesian clustering methods to investigate the spatial genetic structure of the wild boar in the sampled region. Firstly, we used the spatial Bayesian clustering method implemented in program BAPS 4.14 (Corander, Strén & Arjas 2008). The program was run ten times for each of K = 2–10. Given the potential presence of relatives of wild boar suspected of clandestine translocation (see
above), we looked for unusually strongly differentiated clusters by using PHYLIP 3.66 (Felsenstein 2005) to construct a neighbour-joining tree with the Kullback–Leibler divergence matrix provided as output with BAPS. This matrix can be used as a measure of the relative genetic distance between the BAPS-inferred clusters. We also tried to identify suspects by visualizing the genetic relationship between the individuals in our data set using a factorial correspondence analysis (FCA) in the program GENETIX version 4.05.2.

Secondly, we analysed population genetic structure using a Bayesian model executed in a Markov chain Monte Carlo (MCMC) scheme and implemented in the GENELAND version 2.0.0 extension (Guillot, Morrier & Estoup 2005) of program R 2.4.1 (Ihaka & Gentleman 1996). The number of clusters was determined by running the MCMC iterations five times, allowing $K$ to vary, with the following parameters: 500 000 MCMC iterations, maximum rate of the Poisson process fixed to 100, uncertainty attached to the spatial coordinates fixed at 5 km, minimum $K = 1$, maximum $K = 10$, maximum number of nuclei in the Poisson–Voronoi tessellation fixed to 300. The Dirichlet model was used as a prior for all allele frequencies. After inferring the number of populations in the data set from these five runs, the MCMC was run 30 times with $K$ fixed to the inferred number of clusters, with the other parameters the same as above. The mean logarithm of the posterior probability was calculated for each of the 30 runs and the posterior probability of population membership for each pixel of the spatial domain was then computed for the three runs with the highest values.

Finally, program STRUCTURE version 2.2 (Pritchard, Stephens & Donnelly 2000) was also used to investigate the genetic structure of the wild boar. The first step of the analysis consisted of estimating $K$, the number of subpopulations or clusters. Ten independent runs of $K = 1–10$ with 200 000 MCMC iterations and a burn-in period of 100 000 were performed, using the model with correlated allele frequencies and assuming admixture. ALPHA, the Dirichlet parameter for the degree of admixture, was allowed to vary between runs. For each value of $K$, the log-likelihood values were averaged and standard deviation calculated. We tried to infer the appropriate number of clusters by calculating the $\Delta K$ statistic (Evanno, Regnaut & Goudet 2005). After placing samples into the cluster for which they showed the highest percentage of membership ($q$), averaging $q$ over the 10 runs, we plotted the STRUCTURE-assigned individuals on a map of the study region to assess geographical congruence of the clusters.

SIMULATIONS OF ISOLATION-BY-DISTANCE DATA SETS

Data sets characterized by IBD were simulated using the program MUTANT TRACKER 0.211a (Wilkins 2004). This coalescent program simulates genealogies of samples drawn from a continuous habitat with limited gene flow, allowing the user to specify the geographical location from which each sample is drawn, as well as the effective population density and the dispersal rate of the simulated population. The location of each lineage is tracked explicitly backwards in time and the lineages move by a Gaussian random walk. Two lineages coalesce if at a particular time they are within a certain distance of each other – the inverse of the population density specified by the user. Mutations were simulated using a stepwise model.
We performed simulations based on the geographical coordinates of the 678 autochthonous individuals (see Results) in the empirical data set. However, we additionally performed simulations based on 678 artificial sampling locations generated so that they were spread evenly over our study area. This was done to exclude the possibility that heterogeneous spatial sampling in the empirical data set caused the algorithms to overestimate genetic clustering (Serre & Pääbo 2004; Schwartz & McKelvey in press). The homogenous set of coordinates was generated using GENELAND version 2.0-0, by setting the limits of the geographical domain to correspond to the boundaries of our study area. We performed simulations based on 14 loci using two different densities: 0·5 and 3 individuals per unit area (equivalent to 1 km\(^2\)). While the latter choice might appear excessively high, population densities of wild boar in Western Europe have increased dramatically over the last 30 years. For example, over 4000 wild boar were hunted in Luxembourg (area 2586 km\(^2\)) in 2004 (Schley et al. 2008). For both densities, we simulated three independent data sets for dispersal distances ranging from 0·1 to 7, increasing distances incrementally by 0·1 units. For eight dispersal distance, chosen to cover a range of IBD levels, the data sets were analysed using the Bayesian clustering methods.

We assessed the level of IBD in the simulated and empirical data sets by performing individual-based statistical correlation analyses between a measure of genetic kinship and the (log-transformed) pairwise spatial distances using SPAGeDi: 1·2 (Hardy & Vekemans 2002). The linear regression slope, \(b\), of this relationship offers a convenient measure of the degree of spatial genetic structuring (Hardy & Vekemans 2002). As suggested by Vekemans & Hardy (2004), Loiselle's kinship coefficient \(F_i\) was chosen as a pairwise estimator of genetic relatedness, as it is a relatively unbiased estimator with low sampling variance. The standard error and significance of the linear regression slope were calculated by jackknifing (over loci) and by 10 000 permutations of genotypes. The standard error and significance of the estimates were calculated by jackknifing (over loci) and by 10 000 permutations of genotypes between populations, respectively.

**Results**

In total, 697 DNA samples were obtained from Wallonia (227 samples), Luxembourg (289 samples) and the German federal state of Rhineland-Palatinate (181 samples; Fig. 1). It was possible to generate a full 14-locus profile for 692 of the 697 samples (see Supporting Information, Table S1). In order to assess the reliability of the profiles, 700 genotypes (and 1400 alleles) were compared, i.e. 50 samples (13·9%) typed in duplicate at 14 loci each. Duplicate genotypes always corresponded to the initial genotype and no allelic dropout was observed. We therefore expect an error rate of less than 1/1400 = 17·14 × 10\(^{-4}\) per allele, too small to affect our results.

The mean expected heterozygosity in the complete data set was \(H_e(\pm SD) = 0·615 \pm 0·222\). We expected the collective data set to exhibit signs of a Wahlund effect, either through the presence of distinct genetic clusters, or due to IBD in the study area. The global sample of microsatellite loci did indeed show a highly significant deficit of heterozygotes as compared to Hardy–Weinberg expectations (\(P < 0·001\); see Supporting Information, Table S1). Similarly, when analysing the whole data set, 29 pairs of the unlinked loci deviated from linkage disequilibrium at \(P < 0·05\) before Bonferroni correction, and nine pairs after.

**POPULATION GENETIC STRUCTURE**

Taking spatial information into account, BAPS 4·14 gave a probability of \(> 0·999\) of there being four genetic clusters in the study area. The inferred clusters were geographically coherent (Fig. 2a). The German samples to the southeast of Luxembourg formed a genetic cluster (cluster ‘East’), as did the Belgian samples to the west of Luxembourg (cluster ‘West’). The third cluster was distributed over a larger area from Belgium, across Luxembourg, to the northern part of the Rhineland-Palatinate (cluster ‘Central’). The fourth cluster, however, was formed by 19 samples collected in the south of Luxembourg (cluster ‘Suspect’), roughly in the area where the translocations were suspected to have occurred.

The population dendrogram based on the Kullback–Leibler divergence matrix showed that the longest branch separated the suspects from the remaining clusters (Fig. 3a). The PCA showed that there were a number of genetic profiles...
that differed substantially from the majority of the samples in the data set (Fig. 3b). These samples corresponded to the BAPS-inferred cluster in the south of Luxembourg. Two alleles were found frequently in the 19 genetic profiles that made up this cluster, that were not present in any of the 678 remaining profiles: allele 167 at locus SW911 in 13 individuals and allele 243 at locus S0097 in six individuals. Eighteen of these individuals originated from the region where illegal translocation has been suspected, while one individual was sampled a little farther north (Figs 1, 2a). We therefore concluded that these 19 individuals were probably non-autochthonous and related to individuals illegally translocated 10 years earlier.

When repeating the spatial BAPS analysis with the suspect individuals omitted (henceforth referred to as the truncated data set), we obtained slightly different results, as the program still gave a probability > 0.999 of there being four genetic clusters.
in the study area. The same clusters were obtained as in the analysis with the complete data set (Fig. 2b), with the exception that cluster West was split into two: one cluster immediately to the west of Luxembourg (‘West One’), and one west of the motorway (‘West Two’). Further analyses were performed using GENELAND and STRUCTURE to test for the robustness of this result.

We firstly used the GENELAND method on the complete data set. The five initial GENELAND runs suggested the presence of three genetic clusters in the study area (Supporting Information, Fig. S1a). However, fixing the number of populations to three, the modal assignments of individuals in the run with the highest mean logarithm of posterior probability were not geographically coherent, as the inferred clusters consisted of geographically distinct groups (Fig. 2c). The second and third best runs gave rise to similar clusters (results not shown). When performing the GENELAND analysis with the truncated data set, the five initial runs suggested the presence of only two genetic clusters in the study area (Supporting Information, Fig. S1a). The modal assignments in the three best-supported runs roughly corresponded to the STRUCTURE cluster West and a combination of Central and East (Fig. 2d).

The spatially explicit Bayesian programs did not converge on a clustering solution. When using the non-spatial program STRUCTURE 2.2, the highest value for $\Delta K$, the rate of change in the log probability of the data between successive clusters, was obtained for $K = 3$ (Supporting Information, Fig. S1b). However, the corresponding clusters were largely overlapping and not geographically coherent (Supporting Information, Fig. S2).

### SIMULATIONS OF ISOLATION-BY-DISTANCE DATA SETS

We chose the data sets generated for eight dispersal distances to test the performance of the Bayesian clustering methods (Table 1). We first simulated IBD data sets that did not contain any genetic discontinuities (i.e. barriers). All three Bayesian programs overestimated clustering at the higher levels of IBD (i.e. $h \leq -0.01$), where one or more loci deviated from Hardy-Weinberg proportions at $P < 0.05$ after Bonferroni correction (Table 1a). Overall, no program appeared to substantially outperform the others at avoiding the inference of artificial clusters. For each combination of dispersal and density, the three replicates and all three programs broadly gave similar number of clusters. In the case of STRUCTURE, however, by applying the correction of Evanno, fewer clusters were inferred than with the other two methods.

### Table 1. Characteristics of simulated data sets and the number of genetic clusters inferred using three Bayesian programs. Simulated data sets were characterized either by (a) isolation by distance alone (one cluster) or (b) contained one barrier to gene flow (two clusters).

<table>
<thead>
<tr>
<th>Dispersal (unit area)</th>
<th>IBD slope</th>
<th>$\Delta K$</th>
<th>HW</th>
<th>Density: 0·5 individual unit area$^1$</th>
<th>$\Delta K$</th>
<th>HW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1·5</td>
<td>-0·076</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>-0·017</td>
</tr>
<tr>
<td>1·5</td>
<td>-0·079</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>-0·021</td>
</tr>
<tr>
<td>1·5</td>
<td>-0·077</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>-0·021</td>
</tr>
<tr>
<td>2·0</td>
<td>-0·053</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>-0·013</td>
</tr>
<tr>
<td>2·0</td>
<td>-0·068</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>-0·014</td>
</tr>
<tr>
<td>2·0</td>
<td>-0·054</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>-0·012</td>
</tr>
<tr>
<td>2·5</td>
<td>-0·044</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>-0·008</td>
</tr>
<tr>
<td>2·5</td>
<td>-0·034</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>10</td>
<td>-0·008</td>
</tr>
<tr>
<td>2·5</td>
<td>-0·037</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>-0·008</td>
</tr>
<tr>
<td>3·0</td>
<td>-0·041</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>-0·006</td>
</tr>
<tr>
<td>3·0</td>
<td>-0·034</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>10</td>
<td>-0·006</td>
</tr>
<tr>
<td>3·0</td>
<td>-0·029</td>
<td>9</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>-0·006</td>
</tr>
<tr>
<td>3·9</td>
<td>-0·022</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>-0·004</td>
</tr>
<tr>
<td>3·9</td>
<td>-0·018</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-5</td>
<td>-0·004</td>
</tr>
<tr>
<td>3·9</td>
<td>-0·018</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>-2</td>
<td>-0·004</td>
</tr>
<tr>
<td>4·4</td>
<td>-0·014</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>-1</td>
<td>-0·002</td>
</tr>
<tr>
<td>4·4</td>
<td>-0·011</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-0·003</td>
</tr>
<tr>
<td>4·4</td>
<td>-0·014</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>-0·003</td>
</tr>
<tr>
<td>6·0</td>
<td>-0·008</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-0·002</td>
</tr>
<tr>
<td>6·0</td>
<td>-0·007</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-0·001</td>
</tr>
<tr>
<td>6·0</td>
<td>-0·006</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-0·002</td>
</tr>
<tr>
<td>6·8</td>
<td>-0·006</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-0·001</td>
</tr>
<tr>
<td>6·8</td>
<td>-0·007</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-0·001</td>
</tr>
</tbody>
</table>

Table 1. Continued

| Dispersal (unit area) | IBD slope | \( F_{ST} \) | baps | geneland | \( \ln(X | K) \) | \( \Delta K \) | baps | geneland | \( \ln(X | K) \) | \( \Delta K \) |
|----------------------|-----------|-------------|------|-----------|----------------|--------|------|-----------|----------------|--------|
| (b)                  |           |             |      |           |                 |        |      |           |                 |        |
| 1·5                  | -0·101    | 0·083       | 10   | 10        | 10              | 3      | -0·022| 0·029     | 9               | 8 (7)  |
| 1·5                  | -0·099    | 0·103       | 10   | 10        | 10              | 2      | -0·022| 0·024     | 8               | 8      |
| 1·5                  | -0·080    | 0·139       | 10   | 10        | 10              | 2      | -0·022| 0·021     | 8               | 10     |
| 2·0                  | -0·073    | 0·078       | 10   | 10        | 10              | 2      | -0·014| 0·017     | 4               | 4      |
| 2·0                  | -0·066    | 0·067       | 10   | 10        | 10              | 2      | -0·011| 0·018     | 3               | 5      |
| 2·0                  | -0·056    | 0·114       | 10   | 10        | 10              | 2      | -0·013| 0·016     | 3               | 5      |
| 2·5                  | -0·049    | 0·076       | 10   | 10        | 10              | 2      | -0·008| 0·012     | 2               | 2      |
| 2·5                  | -0·045    | 0·066       | 10   | 10        | 10              | 2      | -0·008| 0·014     | 2               | 1      |
| 2·5                  | -0·049    | 0·063       | 10   | 9         | 10              | 2      | -0·009| 0·013     | 2               | 2      |
| 3·0                  | -0·035    | 0·061       | 10   | 9         | 10              | 2      | -0·006| 0·013     | 2               | 2      |
| 3·0                  | -0·028    | 0·036       | 10   | 6         | 10              | 2      | -0·006| 0·013     | 2               | 2      |
| 3·0                  | -0·027    | 0·057       | 8    | 5         | 8               | 2      | -0·005| 0·008     | 1               | 2      |
| 3·9                  | -0·021    | 0·050       | 4    | 4         | 6               | 2      | -0·003| 0·006     | 1               | 1      |
| 3·9                  | -0·020    | 0·050       | 4    | 4         | 5               | 2      | -0·004| 0·008     | 1               | 1      |
| 4·4                  | -0·026    | 0·055       | 4    | 4         | 6               | 2      | -0·004| 0·009     | 1               | 1      |
| 4·4                  | -0·014    | 0·038       | 3    | 3         | 4               | 2      | -0·003| 0·007     | 1               | 1      |
| 4·4                  | -0·017    | 0·028       | 3    | 3         | 3               | 2      | -0·002| 0·005     | 1               | 1      |
| 4·4                  | -0·013    | 0·044       | 2    | 2         | 2               | -      | -0·003| 0·010     | 1               | 1      |
| 6·0                  | -0·010    | 0·022       | 2    | 2         | 2               | -      | -0·002| 0·007     | 1               | 1      |
| 6·0                  | -0·008    | 0·027       | 2    | 2         | 2               | -      | -0·002| 0·004     | 1               | 1      |
| 6·0                  | -0·007    | 0·027       | 2    | 2         | 2               | -      | -0·001| 0·004     | 1               | 1      |
| 6·8                  | -0·006    | 0·018       | 2    | 2         | 2               | -      | -0·001| 0·006     | 1               | 1      |
| 6·8                  | -0·007    | 0·034       | 2    | 2         | 2               | -      | -0·001| 0·004     | 1               | 1      |
| 6·8                  | -0·007    | 0·017       | 2    | 2         | 2               | -      | -0·001| 0·004     | 1               | 1      |

With a few exceptions at the higher levels of IBD (Table 1a), geneland always assigned individuals to each of the inferred clusters and did not infer the presence of ‘ghost populations’. The clusters inferred by the spatial methods were always geographically coherent, and sometimes, similarly to the empirical data, could have been explained biologically due to the presence of roads or rivers at their boundaries (e.g., Fig. 4a,b). In the case of STRUCTURE, however, the coherence of the inferred clusters decreased with decreasing levels of IBD (e.g., Fig. 4c,d).

Both spatial Bayesian methods correctly identified and located the genetic discontinuity that was simulated to bisect the study area (Table 1b, Fig. 4e,f), but each superimposed further clusters at the higher levels of IBD (e.g., Table 1b, Fig. 4g,h). Also, no program identified a barrier in the higher-density data sets characterized by the longest dispersal distances (Table 1b). By applying the Evanno correction to the STRUCTURE results, the correct number of clusters, \( K = 2 \), was inferred in all but two cases (Table 1b). However, plotting the model assignments on a map suggested that STRUCTURE would not have been very efficient at pinpointing the precise location of the barrier (e.g., Fig. 4i), especially given the degree of overlap between the clusters at the lower levels of IBD (e.g., Fig. 4j).

We also simulated IBD data sets using artificial sampling locations generated so as to be evenly spread over our study area. We performed simulations at a density of three individuals per unit area only. These results confirmed the previous findings that, at higher levels of IBD, all three programs overestimated the number of genetic clusters (Table 2), even while correctly locating a genetic discontinuity (Fig. 5). However, comparison of Tables 1 and 2 suggests that the programs inferred fewer artificial clusters when the study area was sampled homogenously.

We found evidence for significant isolation by distance when analysing the truncated empirical data set as a whole (\( b \pm \text{SE} = -0·010 \pm 0·001; P < 0·001 \)). The overall magnitudes of population differentiation for the empirical clusters were fairly low (baps: \( F_{ST} \pm \text{SE} = 0·032 \pm 0·004 \); geneland: \( F_{ST} \pm \text{SE} = 0·014 \pm 0·003 \); structure: \( F_{ST} \pm \text{SE} = 0·025 \pm 0·004 \)) and, comparing the degrees of differentiation for dispersal distances 4·4 for density 0·5 and 2·0 for density 3 (comparable slopes), were similar to the corresponding values of the artificial clusters inferred in the IBD-only data sets (Supporting Information, Table S2). Similarly, the pairwise \( F_{ST} \) values for the empirical clusters (Supporting Information, Table S3) were within the range of values obtain for the pairwise comparisons of those same IBD-only data sets (Supporting Information, Table S2).

**Discussion**

**POPULATION GENETIC STRUCTURE**

We identified a distinct cluster of individuals located in the same area where an introduction was suspected. Because the genetic profiles of most of these individuals contained private
Fig. 4. Examples of individual modal assignment to clusters when analysing simulated data sets with three Bayesian methods. Different symbols represent different clusters. Figures (a) to (d) show results from analyses of IBD-only data sets (i.e. $K = 1$), with the data set in the other figures containing a simulated barrier dissecting the study area approximately at $6^\circ 06'$ (i.e. $K = 2$). (a) Assignments using the spatial model in baps 4.1.4, analysing simulated data with a density of 3 individuals unit area$^{-1}$ and a dispersal of 2·0 units, slope of $b = -0·014$, optimal $K = 3$ (b) GENELAND, density 3, dispersal 2·0, $b = -0·012$, $K = 4$ (c) STRUCTURE, density 0·5, dispersal 1·5, $b = -0·079$, $\Delta K = 4$ (d) STRUCTURE, density 0·5, dispersal 4·4, $b = -0·014$, $\ln(X | K) = 3$ (e) BAPS and (f) GENELAND analysis, density 0·5, dispersal 6·0, $b = -0·010$, $K = 2$ (g) BAPS and (h) GENELAND analysis, density 0·5, dispersal 3·9, $b = -0·021$, optimal $K = 4$ (i) STRUCTURE, density 0·5, dispersal 2·0, $b = -0·073$, $\Delta K = 2$ (j) STRUCTURE, density 0·5, dispersal 6·0, $b = -0·010$, $\ln(X | K) = 2$. 

alleles and the genetic cluster that they formed was substantially differentiated from the clusters formed by autochthonous individuals, we consider it unlikely that it resulted from genetic drift in an isolated population. We could not assess the likelihood of population membership here (see for example Frantz et al. 2006; Frantz & Krier 2007), but consider that there was, nevertheless, convincing genetic evidence for illegal introduction. Clandestine translocations by private individuals will, by their very nature, not follow quarantine guidelines and need to be prevented.

The Bayesian programs did not converge on the same clustering solution. When analysing the truncated data set, spatial baps inferred the presence of four clusters, compared to three populations inferred by structure and two by geneland. The issue of the results not agreeing has been reported fairly frequently for comparisons of structure and the non-spatial algorithm in baps, with the latter having a tendency to overestimate genetic structure (e.g., Latch et al. 2006; Rowe & Beebee 2007). baps is based on identifying populations with different allele frequencies, rather than partitioning individuals into clusters in Hardy–Weinberg equilibrium. The non-spatial baps algorithm takes weak stochastic fluctuations in the allele frequencies as evidence of genetic structure, allowing the number of clusters to increase without firm support (Corander, Sirén & Arjas 2008). Empirical studies appear to confirm that the incorporation of a spatial prior in baps reduces this bias and generates estimates of \( K \) that are comparable to those obtained with structure (Frantz et al. 2006; Robinson, Waits & Martin 2007).

Recently, some empirical studies have analysed data using both structure and geneland. For example, Fontaine et al. (2007) and Latch et al. (2008) found good congruence between the two algorithms. Both Rowe & Beebee (2007) and Coulon et al. (2008) reported that, overall, geneland inferred credible clustering solutions comparable, but not identical, to structure. In the latter study, however, the geneland analysis had a high occurrence of ‘ghost’ populations. The authors, therefore, based their choice of \( K \) on the number of clusters that had individuals assigned to them in the second round of analyses. Performing more runs and analysing the outputs following the protocol outlined in Coulon et al. (2008) might have helped solve this problem.

Few simulation studies have as yet compared the clusters identified by spatial and non-spatial algorithms. One exception is the work by Chen et al. (2007). However, these authors do not provide information on the frequency with which the tested programs inferred the correct number of clusters.
Moreover, Chen et al. simulated genetic clusters without admixture, used extremely short MCMC runs and did not simulate panmictic populations to test whether the programs enforced substructure where it did not exist. Finally, the maximum number of clusters in the analysis was limited to six, while the data set comprised five clusters; setting the maximum number of clusters to a larger value might in fact have led to the inference of a larger number of clusters.

A certain amount of non-convergence between different Bayesian clustering methods thus appears to be relatively normal. Additionally, however, the same algorithm produced different solutions when analysing our full and the truncated data sets. Both types of non-convergence created interpretation problems: barriers identified by one program were not supported by another. This was the case with the river Moselle and the motorway to the west of Luxembourg. While Frantz et al. (2006) found evidence for the river Moselle acting as a barrier for red deer, the section of the motorway in question only opened in 1988, which might be too recent to cause population genetic structure through the effects of genetic drift and mutation. Finally, there was no obvious biological explanation for a putative barrier located between the Luxembourg samples and the Belgian samples located to the west of Luxembourg. Indeed, both areas are connected by a fairly extensive network of wildlife corridors (Baghli, Moes & Walzberg 2007). The $F_{ST}$ values obtained for the actual data were not very informative in deciding whether the inferred clusters were genuine, as they fitted into the range of values obtained for the artificial clusters inferred from the IBD data sets.

It is not entirely clear what caused the non-convergence of clustering methods in our study. One possible explanation is that programs produced different solutions because of differences in the underlying algorithms, as is the case for structure and baps. Similarly to structure, geneland assumes HW and linkage equilibria within genetic clusters. However, while structure was run assuming correlated allele frequencies, geneland was run assuming independent allele frequencies (following Guillot et al. 2005). The former model often improves clustering when populations are closely related, but can also increase the risk of overestimating $K$ (Falush, Stephens & Pritchard 2003), which might explain the differences observed between these two algorithms. Alternatively, it is possible that all the clusters identified represented genuine genetic sub-divisions, but that the level of differentiation between them was too low for the clustering to be consistent (e.g., Latch et al. 2006). Finally, given the IBD pattern in our data set, the clustering solution could be an artefact superimposed on the data (e.g., Guillot et al. 2005).

The spatial Bayesian clustering methods do not take isolation-by-distance clines explicitly into account. Rather, by including a spatial prior, all genetic structures are not a priori equally likely, but the joint probability that any two individuals belong to the same cluster decreases with the geographical distance between them (Guillot et al. 2005; Corander, Sirén & Arjas 2008). Similar to STRUCTURE, the spatial algorithms still assume that the inferred population genetic clusters are panmictic units. When we simulated a weak IBD pattern, the loci did not deviate substantially from HW proportions and no clusters were imposed. These results are in line with Guillot et al. (2005), who stated that, when analysing one or several panmictic populations, GENELAND did not enforce spatial substructure when it did not exist. However, a stronger IBD pattern led to deviations from random mating in our simulated data. This probably caused the overestimation of genetic clustering by the Bayesian programs.

In a recent study, Coulon et al. (2006) used GENELAND in an attempt to assess the effect of landscape features on the genetic structure of roe deer Capreolus capreolus. While STRUCTURE did not find evidence for any sub-structuring, GENELAND inferred the presence of two biologically feasible clusters. However, the degree of genetic differentiation between the two clusters was very low ($F_{ST} = 0.008$) and the study population was characterized by an IBD pattern. It is therefore possible that the two clusters inferred using GENELAND were artefacts, despite a plausible biological explanation for their presence.

We confirmed the conclusion of Schwartz & McKelvey (in press), that the non-spatial STRUCTURE algorithm can overestimate the number of genetic clusters when analysing data sets characterized by isolation by distance. As in our analyses presented here, these conclusions were based on a relatively limited number of simulations. However, the fact that two independent studies found the software to be similarly biased reinforces this finding. Schwartz & McKelvey (in press) found that different sampling schemes (at a local scale) affected the optimal number of clusters inferred by STRUCTURE. When we simulated genetic samples that had a more homogenous spatial distribution, all three methods superimposed clusters at higher levels of IBD, showing that overestimation of the number of genetic clusters was an inherent bias of these Bayesian programs. However, these biases might be less severe than when samples are more unevenly distributed. Further simulations are required to assess the extent of the bias of spatial Bayesian methods under different sampling schemes and IBD clines. It would also be desirable to test the performance of the spatial methods when sampling has been discontinuous along an IBD cline on a larger continental or global scale (sensu Serre & Pääbo 2004).

A second main conclusion from our simulations was that the spatial Bayesian programs could identify a simulated barrier correctly but, at higher levels of IBD, the programs could erroneously infer the presence of further clusters. In other words, in some cases sub-structure detected by the spatial programs could accurately reflect a genetic discontinuity, while other clusters could be artefacts caused by some other deviation from random mating. An illustration of this may have been provided by Zannèse et al. (2006): when analysing genetic data from roe deer using GENELAND, the authors inferred the presence of three genetic clusters in their study area. Two of these clusters corresponded well to two main population units inferred using environmental and morphological data, while there was no obvious biological explanation for a smaller third cluster. The authors report a significant global heterozygosity deficiency in the whole sample, which could be the result of a Wahlund effect or due to IBD.

We simulated a fairly impermeable barrier that had been present for 100 generations. Nevertheless, in the data sets characterized by a density of three individuals per unit area, the Bayesian programs did not identify a barrier when simulating large dispersal distances. The question as to how powerful the various methods are in detecting barriers that have only been present for shorter periods and/or that are more permeable requires further investigation. Our limited simulations suggested that the $\Delta K$-corrected STRUCTURE results produced the fewest artificial clusters, and even helped to infer the correct number of clusters in some data sets simulated to contain a barrier. However, for both types of simulated data sets (barrier/no barrier), there was a significant amount of overlap between the various clusters, which would have made it difficult to pinpoint the precise location of a barrier and to distinguish between genuine and artificial clusters. Moreover, it is unclear whether the $\Delta K$ statistic would have inferred the correct result if more than one barrier had been simulated. Finally, the $\Delta K$ statistic represents the second-order rate of change of the likelihood function with respect to $K$, and as such cannot evaluate $K = 1$ (Evanno, Regnaut & Goudet 2005). Further simulations under biologically realistic scenarios are required to assess the performance of this ad hoc statistic.

Conclusions

We show that Bayesian clustering programs can overestimate genetic structure in data sets characterized by isolation by distance. This bias could lead to the erroneous delimitation of management or conservation units. Simulations suggested that the strength of IBD in our empirical data set was just high enough to cause artificial clustering. Some clusters in the empirical data set could be explained biologically, but there were inconsistencies between programs. It was also possible that some clusters were genuine, while others were artefacts. It was thus not possible to ascertain with confidence whether the clustering solutions offered by the various programs were an accurate reflection of population genetic structure in our empirical data set or artefacts created by an IBD pattern. Wild boar certainly are very mobile and a study over a similar scale in Australia did not identify any population structure (Cowled et al. 2006).

Our results suggest that it is very important to test a data set for isolation by distance before applying the Bayesian methods evaluated here. This should give an indication as to
whether the results might be biased and help to assess the validity of biologically non-feasible clusters. We also strongly agree with Pearse & Crandall (2004) that different Bayesian clustering approaches should be used to investigate the spatial genetic structure in a data set. On the one hand, we did find that the spatial methods could produce very similar, but wrong, clustering solutions (compare, for example, Fig. 4g,h), implying that convergence of results between different methods is no guarantee that results are correct. On the other hand, we found that BAPS and GENELAND produced different solutions when analysing the empirical data, and that our conclusions might have been very different had we only used one program.

The main aim of our simulations was to test whether the clusters derived from the empirical data could be artefacts. We therefore performed relatively limited simulations in a context that was specific in terms of number of microsatellite loci, samples size, sample location and simulated population densities. Nevertheless, we believe that our results confirm the warnings of previous authors that deviations from random densities. Nevertheless, we believe that our results confirm the warnings of previous authors that deviations from random

Acknowledgments
We are indebted to Hubertus Becker, Nico Bonanni, Erhard Ginter, Dr Dieter Hoff, Metzgerei Schmitt in Mandern, Andreas Michel, and the local services of the Ministry of the Walloon Region, General Directorate of Natural Resources and Environment, Nature and Forest Division (forest districts of Bièvre, Bertrix, Bouillon, Florenville, Habay-la-Neuve, Laroché, Neufchâteau, Saint-Hubert, Saint-Vith and Vielsalm) for providing us with tissue samples. We are indebted to Hubertus Becker, Nico Bonanni, Erhard Ginter, Dr Dieter Hoff, Metzgerei Schmitt in Mandern, Andreas Michel, and the local services of the Ministry of the Walloon Region, General Directorate of Natural Resources and Environment, Nature and Forest Division (forest districts of Bièvre, Bertrix, Boutrix, Bouillon, Florenville, Habay-la-Neuve, Laroché, Neufchâteau, Saint-Hubert, Saint-Vith and Vielsalm) for providing us with tissue samples. Roger Butlin and various anonymous referees provided helpful comments on earlier versions of the manuscript.

References


Received 16 August 2008; accepted 15 December 2008

Handling Editor: E J Milner-Gulland

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Inference of the number of genetic clusters in the study area: posterior distribution of the number of populations estimated using GENELAND and mean of log-likelihood values obtained using the program STRUCTURE.

Fig. S2. Map of individual STRUCTURE assignments at K = 3.

Appendix S1. Details of the polymerase chain reaction conditions

Table S1. Properties of the microsatellite loci used in this study

Table S2. Genetic differentiation between clusters inferred from IBD-only simulated data sets with IBD levels comparable to the empirical data set

Table S3. Pairwise *F*~ST~ values for clusters obtained when analysing the empirical data set

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