Molecular Ecology Notes (2007) **7**, 747–756 doi: 10.1111/j.1471-8286.2007.01769.x

TECHNICAL ARTICLE

Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study

CHIBIAO CHEN,*† ERIC DURAND,† FLORENCE FORBES* and OLIVIER FRANÇOIS†
** INRIA Rhône-Alpes, avenue De l'Europe, 38334 Montbonnot, Saint Ismier cedex, France,* †*TIMC, Université Joseph Fourier, Institut National Polytechnique de Grenoble, avenue Felix Viallet, 38031 Grenoble cedex 1, France*

Abstract

On the basis of simulated data, this study compares the relative performances of the Bayesian clustering computer programs STRUCTURE, GENELAND, GENECLUST and a new program named TESS. While these four programs can detect population genetic structure from multilocus genotypes, only the last three ones include simultaneous analysis from geographical data. The programs are compared with respect to their abilities to infer the number of populations, to estimate membership probabilities, and to detect genetic discontinuities and clinal variation. The results suggest that combining analyses using TESS and STRUCTURE offers a convenient way to address inference of spatial population structure.

Keywords: Bayesian clustering, computer programs, population membership, relative performances, spatial assignment methods, spatial genetics

Received 27 December 2006; revision accepted 26 February 2007

Introduction

Bayesian clustering algorithms have recently emerged as a prominent computational tool for inferring population structure in population genetics and in molecular ecology (Beaumont & Rannala 2004). Bayesian clustering methods use genetic information to ascertain population membership of individuals without assuming predefined populations. They can assign either the individuals or a fraction of their genome to a number of clusters based on multilocus genotypes. The methods operate by minimizing Hardy–Weinberg and linkage disequilibria, and the assignment of each individual genotype to its population of origin is carried out probabilistically. The assignment can generally be achieved by using Markov chain Monte Carlo (MCMC) approaches. These particular assignment methods are useful when genetic data for potential source populations are not available, and they offer a powerful tool to answer questions of ecological, evolutionary, or conservation relevance (Manel *et al*. 2005).

Correspondence: O. François, TIMB, TIMC, Faculté de Médecine de Grenoble, F38706 La Tronche cedex, France. Fax: +33 456520 044; E-mail: olivier.francois@imag.fr

A recent study by Latch *et al*. (2006) compared the relative performance of three nonspatial Bayesian clustering programs, STRUCTURE (Pritchard *et al*. 2000), BAPS (Corander *et al*. 2003) and PARTITION (Dawson & Belkhir 2001). Latch *et al*. (2006) provided evidence that the three algorithms generally perform well at low levels of genetic differentiation (i.e. F_{ST} levels around 0.03–0.05). However, current developments of Bayesian clustering methods explicitly address the spatial nature of the problem of locating genetic discontinuities by including the geographical coordinates of individuals in their prior distributions (Wasser *et al*. 2004; Guillot *et al*. 2005; François *et al*. 2006). But comparative studies that evaluate the relative performance of these spatial algorithms are still lacking. In addition, neither the power to delineate population boundaries, nor the benefit of the inclusion of spatial coordinates into classical analyses has been explored systematically.

This study evaluates the relative performance of three spatial Bayesian clustering programs, namely GENELAND (Guillot *et al*. 2005), GENECLUST (François *et al*. 2006) and a new software named TESS. In order to assess the benefit of including geographical coordinates into more classical analyses, the performances of these programs are also compared to those of STRUCTURE.

The algorithms were originally described as follows. GENELAND is an R package, which estimates the number of populations in a study area, assigns individuals to their population of origin, and potentially detects immigrants, while taking into account uncertainty on the location of sampled individuals. It is based on a hidden partition model, in which the study area is divided into polygonal regions. The number of these regions, λ, controls the amount of spatial dependencies within the hidden partition. Low values of λ correspond to long-range dependencies and to the existence of a few clusters, whereas large values of λ correspond to weak spatial dependencies and fragmented populations.

GENECLUST is based on the concept of Hidden Markov Random Field (HMRF), which models the spatial dependencies at the cluster membership level. HMRF's are statistical models that can account for the fact that individuals from spatially continuous populations are more likely to share cluster membership with their close neighbours than with distant representatives. GENECLUST can detect significant geographical discontinuities in allele frequencies, and it can regulate the number of clusters. It can also reduce the number of loci required to obtain accurate assignments. GENECLUST uses a specific HMRF model which accounts for the dependencies at the cluster label level: the so-called Potts model, whose parameter ψ controls the importance given to spatial interactions. In general, the interaction parameter ψ is non-negative. For less than 10 populations (a common number), François *et al*. (2006) suggested to use values of ψ in the interval (0.5, 1). Values less than 0.5 often lead to the same output as obtained with $\psi = 0$, for which the statistical model used by STRUCTURE is recovered. GENECLUST is distributed as an R package running a FORTRAN-coded MCMC algorithm.

This study includes a new computer program, called TESS, which mainly implements the same statistical model as GENECLUST. Nevertheless, the algorithm used by TESS differs from GENECLUST in many regards, like data structures, Monte Carlo proposal kernels, and other numerical options which contribute to optimize program significantly (see the program documentation for details). TESS uses an input format compatible with the one used by STRUCTURE, and it contains an intuitive graphical user interface shell and a command-line engine. TESS provides facilities for creating and managing projects, which are coherent units grouping the input data, the algorithmic parameter settings, and the output results altogether. By interacting with the graphical interface, users can check their data, specify the parameter settings, run the MCMC algorithm, and visualize the results. In addition to the MCMC program, TESS includes an EM solver designed in the same spirit as the software FASTRUCT which is also able to compete with STRUCTURE favourably (Chen *et al*. 2006).

Although BAPS newest version (Corander *et al*. 2003) also enables the analysis of spatial data, this program is not compared to the three others here. Yet there is no published description of the spatial model used by BAPS. A fair comparison would require information about model details and internal settings. To be consistent with Latch *et al*., we used STRUCTURE version 2.1 admixture model with correlated allele frequency (F-model). We used GENELAND version 1.0.5 correlated allele frequency model (GENELAND has no admixture model). We used GENECLUST version 0.1 (model without inbreeding) and with fixed values of the interaction parameter ψ. We used both TESS admixture and no-admixture models (TESS has no F-model). All simulations were performed using a 3.2GHz Xeon Intel processor with 2GB memory.

Methods

We compared spatial Bayesian clustering programs on the basis of three simulated scenarios: (i) five 'overlapping' islands, (ii) two islands with various geographical connectivity levels, and (iii) continuous variation in allele frequencies. The simulated scenarios were designed to evaluate the programs' abilities to correctly infer population structure and clinal variation in allele frequencies. These scenarios typically involved no genetic admixture, which translated into very similar behaviours of the admixture and no-admixture models of STRUCTURE and TESS. Therefore, the interpretation of Q coefficients as probabilities that a genome is correctly assigned to its population of origin may be better than an interpretation as the proportion of genome correctly assigned to its population of origin. However, distinguishing between the two concepts is generally subtle (and ambiguous), especially here because the admixture and no-admixture results agreed perfectly.

Finite island model

These simulations were based on previously published data by Latch *et al*. (2006) who used them for measuring the relative performance of nonspatial Bayesian clustering programs. Each simulated data set consisted of one population structured into five subpopulations differentiated at one of five F_{ST} levels from 0.01 to 0.05. Five hundred multilocus genotypes (100 per subpopulation) were randomly drawn from the five subpopulation allele frequency distributions across 10 codominant unlinked loci to form a single data set. Spatial coordinates were simulated using two-dimensional Gaussian distributions, so that the five subpopulations were organized in a star shape on a ring. The regions occupied by each subpopulation overlapped with the other regions geographically, so that these regions shared their adjacent areas. This process mimicked an instantaneous expansion to a large size following a bottleneck, such that the range of each subpopulation suddenly increased and gave rise to recent contact zones. The contact zones contained around 20% of all individuals, 10% of which could be considered as migrants because they lied within the range of a foreign island. A picture of the ground truth for one typical spatial pattern used in this simulation scenario is displayed in the Supplementary material (Fig. S1).

The programs STRUCTURE, GENELAND, GENECLUST and TESS provide estimates for the number of genetic clusters K and for individual assignment probabilities. STRUCTURE and TESS admixture model can also compute the proportion of the genome of each individual that can be assigned to the inferred clusters. This quantity was very close to the individual assignment probabilities in the simulated scenarios. To infer the number of populations, STRUCTURE's users usually proceed with successive runs by increasing the number of clusters, and they select the number of clusters with the highest likelihood. Because this method might not be always accurate, the ΔK measure was proposed to provide a better estimate of the true K (Evanno *et al*. 2005). GENELAND uses a Reversible Jump (RJ) MCMC algorithm that computes posterior probabilities for various K and that estimates K as the most likely value obtained at the end of the program run (Guillot *et al*. 2005). GENECLUST and TESS procedure is similar to STRUCTURE but these programs use an additional regularization feature which generally leads to a less ambiguous determination of K (François *et al*. 2006).

To evaluate the ability of the programs to correctly estimate the number of populations, we used 25 data sets with F_{ST}'s ranging from .01 to .05 (5 of each). For each data set, GENELAND, GENECLUST and TESS runs were performed for 12 000 sweeps (burn-in period: 2000 sweeps) with the maximum number of clusters fixed to $K_{max} = 6$. GENELAND initial number of nuclei was fixed to $\lambda = 100$ ($\lambda_{max} = 300$). TESS and GENECLUST interaction parameter was set to $\psi = 0.6$. The three programs were run 10 times for each data set, and the run with the highest likelihood was stored as the final result. STRUCTURE was also run for checking that TESS produced the same results when the parameter $\psi = 0$ was used. The average runtime of GENELAND for a single data set (consisting of 10 runs) was approximately 5 h. The average runtime for GENECLUST was approximately 3 h, and the runtime dropped to less than 15 min for TESS.

Assuming that the number of populations was correctly inferred, the relative performance of each program to correctly assign individuals to their population of origin was assessed from the same 25 data sets. All programs were started with $K_{max} = 5$, and they were run for 1200 sweeps each, including a 200 sweeps burn-in period. As recommended by its authors, GENELAND was run with fixed K. For TESS and GENECLUST, four values of the interaction parameter ranging from $\psi = 0.3$ to $\psi = 1.2$ were experimented. Misassignment rates and average membership probabilities (or proportions of genome belonging to the 'correct' subpopulation for TESS) were computed from the highest likelihood run over 10 runs. The average runtime of GENELAND for a single data set (including 10 runs) was around 33 min. The average runtime for GENECLUST was about 11 min, and the runtime dropped to less than 1 min and a half for TESS.

The role of spatial data

A spatial variant of the island model was used in order to assess the sensitivity of Bayesian spatial clustering algorithms to simultaneous variation of genetic differentiation and spatial density within the data. Two hundred diploid genotypes were sampled at 20 unlinked loci from two subpopulations of equal effective size N. Alleles were simulated according to the infinite allele model with constant mutation rate $\theta = 4\,\mu N = 1$ at each locus (μ is the mutation rate per generation). Spatial coordinates were simulated from a geographical mixture of two independent two-dimensional Gaussian distributions. In subpopulation 1, the spatial data were sampled according to the standard Gaussian distribution centred at the origin N [(0, 0) Id)], while for population 2, the distribution was centred at distance D from the origin. The parameter D also measured the degree of population connectivity. In the simulations, D was increased from $D = 0.5$ (40% overlap) to $D = 4$ (2.2% overlap). For each D value, five data sets were created with levels of differentiation around $F_{ST} \approx 0.02$, and the parameter settings described in the previous paragraphs were applied without change to the three algorithms STRUCTURE, GENELAND and TESS. The scenario is relevant to two weakly differentiated subpopulations one of which underwent very recent massive migration resulting into a spatial contact zone. A second possible interpretation could be that the population underwent a recent fission resulting in two subpopulations which are now diverging both geographically and genetically. F-statistics were computed using Weir and Cockerham estimates (Weir & Cockerham 1984). GENECLUST was checked to perform similarly as TESS, but with runtimes significantly longer.

Discontinuous sampling along a directional cline

A frequently reported issue in the recent literature is that (nonspatial) Bayesian algorithms may be confounded by discontinuous spatial sampling (Serre & Pääbo 2004; Rosenberg *et al*. 2005). Because STRUCTURE puts a strong prior on the existence of clusters, it may be prone to errors when geographical sampling is discrete along clines. François *et al*. (2006) argued that GENECLUST (and TESS) can help dealing with the sampling issue by checking which clusters are robust to the inclusion of a spatially continuous prior distribution. We simulated clinal variation at 20 unlinked biallelic loci along one direction, for which the coordinate x varied from 0 to 2, and the allele frequencies at each locus varied linearly from 0.3 to 0.7 as a function of x. Three hundred genotypes were sampled from three geographical sites with individual coordinates around $x = 0.5, 1, 1.5$ (100 individuals at each site, s.d. ≈ 0.1) so that the spatial coordinates formed three nonoverlapping geographical clusters. This process was replicated to produce 10 data sets. Program runs of STRUCTURE, GENELAND ($\lambda = 100$), and TESS ($\psi = 0.6$) were performed using 1000 sweeps preceeded by 200 burn-in sweeps with $K_{max} = 3$ first and then with $K_{max} = 2$.

Large samples and starting configurations

Bayesian software practioners privilege the use of repeated short runs instead of long runs, although long runs are more consistent with the theory of Markov chains (Gelman *et al*. 1995). Short runs may be a reasonable strategy when parameter dimension is high because they provide a mean to explore a larger number of local optima. With short runs, the program outputs may be sensitive to the initial starting points. This is particularly true when the sample size increases as the programs need to compute an increased number of membership coefficients. To evaluate the impact of the starting configuration, we simulated genotypes at 20 biallelic loci (a/A), with the frequency of allele A equal to 0.4 in one subpopulation and equal to 0.6 in the other subpopulation (two subpopulations, $F_{ST} \approx 0.04$). The sample sizes were increased from 500 to 1000 and 2000 individuals. Spatial coordinates in population 1 were sampled from a uniform distribution over the square (0, 1) × (0, 1). Spatial coordinates in population 2 were sampled from a uniform distribution over the square (0.8, 1.8) × (0, 1). We generated five data sets using this model. TESS provides three options: pure MCMC runs, EM runs, and the possibility to combine both. We ran the TESS MCMC program for 2200 sweeps (including 200 sweeps as a burn-in period). These results were then compared to those obtained after 10 iterations of the EM algorithm (Dempster *et al*. 1977; Celeux *et al*. 2003) followed by 2000 sweeps of the MCMC algorithm. TESS was run with $K_{max} = 3$, and the interaction parameter was set to $\psi = 1.0$.

Results

Five-island model

Table 1 reports the average probability that an individual genome is correctly assigned to its population of origin for GENELAND, GENECLUST and TESS (total correct membership probability) when estimating the true number of populations. The results for TESS could also be interpreted as the average proportion of an individual genome correctly assigned to its population of origin by the admixture version. The programs could not discern five populations at an F_{ST} of 0.01, as they generally ended with a single cluster. GENECLUST and TESS ($\psi = 0.6$) detected the five populations at an F_{ST} of 0.02, and the correct number of populations was retrieved in each of the 10 runs. GENELAND was able to detect population STRUCTURE at F_{ST}'s greater or equal than 0.05. For F_{ST}'s within the interval 0.02–0.04, 28% of the runs were successful at detecting the number of populations. The total correct membership probabilities were lower for GENELAND than for GENECLUST and TESS. Results for STRUCTURE can be found in Latch *et al*. (2006).

Table 1 Five-island simulation. Probability that an individual genome is correctly assigned to its population of origin when the programs are started with an incorrect number of populations ($K_{max} = 6$). F_{ST} varies in the range (0.02, 0.04)

	F_{ST}		
	0.02	0.03	0.04
GENELAND	0.47	0.39	0.59
GENECLUST	0.82	0.91	0.96
TESS	0.83	0.92	0.96

Assuming that K was correctly estimated ($K = 5$), GENECLUST and TESS performed extremely well at low levels of genetic differentiation, and reached misassignment rates lower than 3.5% for F_{ST}'s greater or equal than 0.03. The rates decreased as F_{ST}'s increased (Fig. 1). Regarding missassignment scores, GENELAND was less efficient than GENECLUST and TESS, and reached a value around 9%. The misassignment value did not decrease for the highest F_{ST} levels, perhaps because GENELAND was unable to ascertain the complex boundaries that delineate the adjacent islands. In these simulations, about 10% of all individuals were geographically farther from their population of origin than from a foreign population. GENECLUST and TESS were particularly efficient at locating these migrants and reassigning them to the correct population whereas the performances of GENELAND were poorer in this respect. For TESS (and GENECLUST), the probability of an individual genome assigned to its correct population reached over 94% for an F_{ST} of 0.03 (see Fig. 2).

The ratio of success over 10 runs for GENELAND, GENECLUST and TESS is reported in Fig. S2 (Supplementary material). Further results for $\psi = 0.3$–1.2 are reported in a text file available from the Supplementary material. TESS performances were slighlty lower for $\psi = 0.3$ or $\psi = 1.2$ than for $\psi = 0.6$–0.9, suggesting that the latter range is better appropriate to deal with $K \approx 5$ populations. Tables S1–S4 (Supplementary material) give detailed results for all data sets and programs for $K_{max} = 5$ and $K_{max} = 6$.

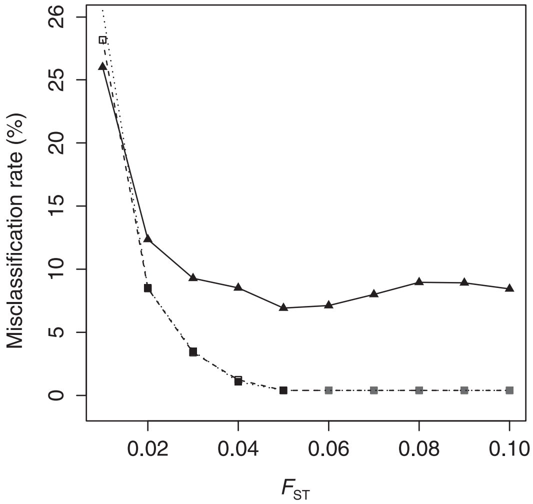

Fig. 1 Five-island model simulation results (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.01 to 0.10). Misassignment rates (percentage) for GENELAND, filled triangles; TESS, filled squares; GENECLUST, empty squares (hidden by the filled squares). The results are for $K_{max} = K = 5$, that is the number of populations is assumed to be correct. Results for GENECLUST are indistinguishable from those of TESS.

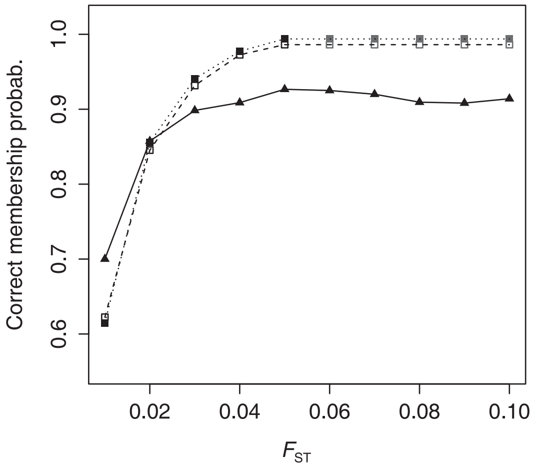

Fig. 2 Five-island model simulation results (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.01 to 0.10). Probability that a genome is correctly assigned to its population of origin by GENELAND, filled triangles; TESS, filled squares; GENECLUST, empty squares. The results are for $K_{max} = K = 5$, that is the number of populations is assumed to be correct.

Considering that GENECLUST and TESS performed similarly, and that TESS is approximately 10-fold faster than GENECLUST, no further results of the latter will be reported afterwards.

Sensitivity to spatial density

TESS provided the lowest rates of misassignment for distances between $D = 1$ and $D = 4$. This range corresponded to spatial overlap (or geographical admixture) varying from 2% to 30% (Fig. 3). TESS reached misassignment rates lower than 3.5% for overlaps less than 30%. STRUCTURE performed better than TESS when the percentage of overlap was equal to 40%. GENELAND was not able to detect population STRUCTURE when the overlap was greater than 30%. For overlap levels around 20%, the average misassignment rate over geneland successful runs was around 9%. Like the five-island simulation, TESS was significantly better than GENELAND at detecting individuals that did not lie close to their origin centre. The performances of GENELAND increased from an error rate of 9% to an error rate of 1% as the degree of connectivity decreased to its minimum. In this case, GENELAND performed similarly as TESS did. Figure 4, which displays the average probability that a genome is correctly assigned to its original cluster, supports the previous findings. GENELAND contrasting results for the largest D values were due to the fact that this program assigned a fraction of genome to a third inconsistent cluster. Detailed results are reported in Tables S5–S8 (Supplementary material).

Discontinuous sampling along a cline

For the three programs, runs performed with $K = 3$ led to inconsistent results across successive replicates. STRUCTURE produced estimates of membership coefficients around 1/3 for all individuals. GENELAND and TESS split the population into two groups, but the location of the boundary between the two clusters exhibited significant spatial variation from one run to another. For $K = 2$, Figs 5, 6 and 7 display estimates of the membership probabilities along the cline

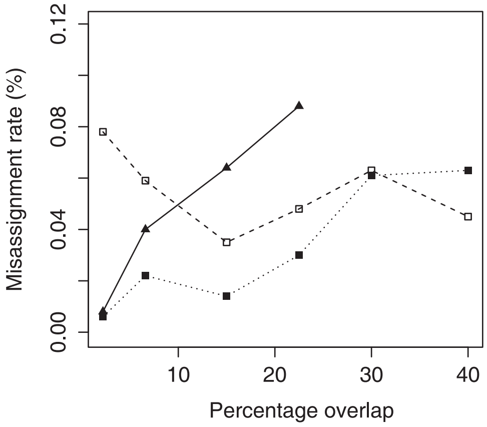

Fig. 3 Two-island model simulation results (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical overlap between the two subpopulations varies from 2% to 40%. Misassignment rates (percentage) for GENELAND, filled triangles; TESS, filled squares; STRUCTURE, empty squares as a function of density overlap. The results are for $K_{max} = 3$, that is the number of populations is unknown.

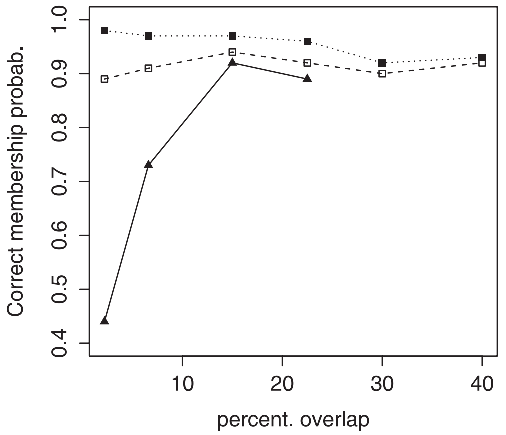

Fig. 4 Two-island model simulation results (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical overlap between the two subpopulations varies from 2% to 40%. Probability that a genome is correctly assigned by GENELAND, filled triangles; TESS, filled squares; STRUCTURE, empty squares as a function of density overlap. The results are for $K_{max} = 3$, that is the number of populations is unknown.

for one typical data set over 10 replicated runs after correcting for label switching. A nonlinear regression curve provided a quasi-linear estimate for the coefficient membership of STRUCTURE, while these curves provided an exaggerated variation around the central cluster for TESS and GENELAND. The same kinds of results were obtained for all simulated data sets (not reported). These curves provide evidence that the three programs were able to detect clinal variation even when the level of differentation was low. STRUCTURE yielded the best performance when ascertaining clinal variation.

Starting values

For the largest sizes (2000 individuals), the 20 preliminary EM steps increased the number of successful runs of TESS MCMC program significantly. Table 2 reports the ratio of successful runs for the five simulated data sets. This ratio increased from 28% to 66% when the EM steps were used beforehand. For these runs, the misassignment rate was lower than 10%, and the average proportion of genome correctly assigned to its cluster of origin was greater than 90%. Surprisingly, the EM steps did not provide obvious benefit unless the sample size was increased to 2000 individuals. These results and others that were reported in the Supplementary materials (Tables SM2 and SM4) suggest that the EM steps should be used in combination to MCMC runs preferentially.

Discussion

This study compared three distinct Bayesian assignment approaches corresponding to four computer programs which all compute probabilities that each individual genome originates from one of K populations. The three approaches could be classified according to their distinctive specificities as follows. The first one (STRUCTURE) is a nonspatial clustering method. The second one (GENELAND)

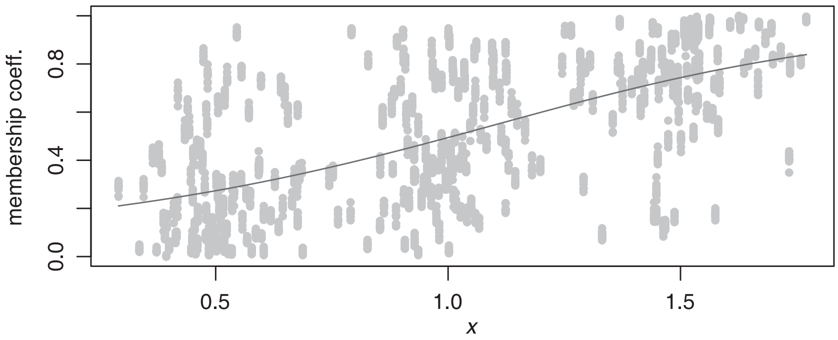

Fig. 5 Clinal variation simulation results for STRUCTURE. Three nonoverlapping samples (300 individuals) were simulated with individuals genotyped at 20 unlinked biallelic markers. Allele frequencies were varied from 0.3 to 0.7 linearly along the cline. Membership probability in one population as a function of the location along the cline (x).

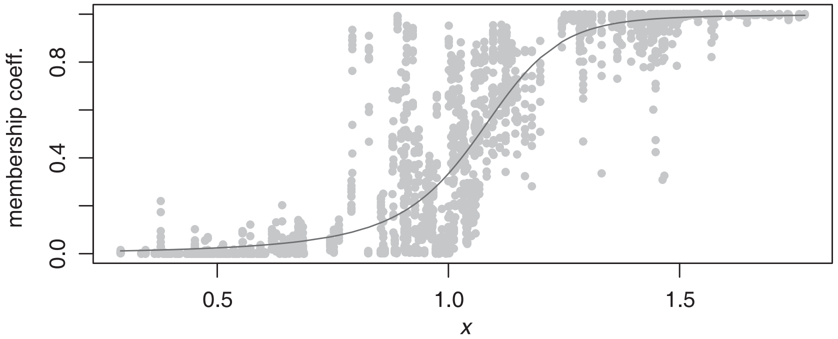

Fig. 6 Clinal variation simulation results for GENELAND. Three nonoverlapping samples (300 individuals) were simulated with individuals genotyped at 20 unlinked biallelic markers. Allele frequencies were varied from 0.3 to 0.7 linearly along the cline. Membership probability in one population as a function of the location along the cline (x).

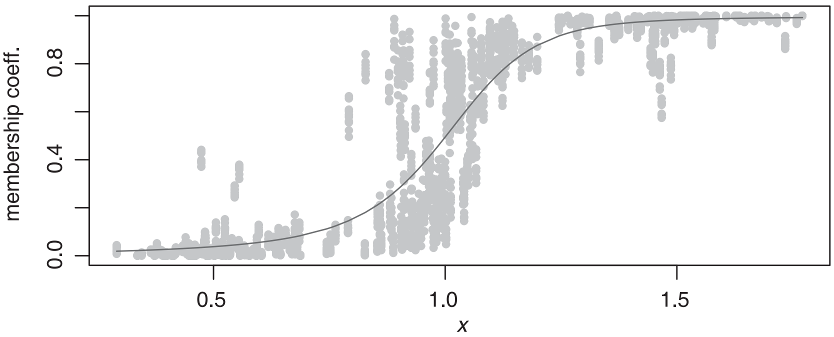

Fig. 7 Clinal variation simulation results for TESS-GENECLUST. Three nonoverlapping samples (300 individuals) were simulated with individuals genotyped at 20 unlinked biallelic markers. Allele frequencies were varied from 0.3 to 0.7 linearly along the cline. Membership probability in one population as a function of the location along the cline (*x*).

Table 2 The MCMC algorithm compared to the combination EM + MCMC in TESS. Success rates over 10 runs for data sets with two populations, 20% spatial overlap, 2000 individuals, $F_{ST} \approx 0.04$

Data set	1.	2.	3.	4.	5.
MCMC	0.42	0.11	0.40	0.33	0.24
EM + MCMC	0.74	0.67	0.71	0.62	0.73

attempts to group individuals in geographical components within a partition model. The third one (TESS/GENECLUST) assumes proximate interactions between individuals, and spatial dependencies are prescribed at the individual level directly. Manel *et al*. (2005) claimed that for species in which individuals are continuously distributed, methods other than assignments methods are more appropriate to address questions such as the location of genetic discontinuities or the detection of geographical barriers. For example they referred to nonparametric approaches like spatial autocorrelation methods (Epperson 2003). This study provided evidence that assignment methods are actually well-suited to deal with the identification of genetic discontinuities and migrant individuals. Spatial Bayesian clustering is as reliable as nonspatial Bayesian clustering programs are, and this is particularly true when the number of polymorphic loci available to the study is limited.

Number of populations

The ability of STRUCTURE to correctly infer the number of subpopulations has been previously investigated by Evanno *et al*. (2005) and Latch *et al*. (2006). The additional simulations carried out here provided evidence that the current practice (e.g. Rosenberg *et al*. 2002), which iteratively examines increasing values of the number of populations unless the likelihood reaches a plateau, is generally efficient. For the five-island model at levels of genetic differentiation $F_{ST} \geq 0.03$, Latch *et al*. (2006) reported that STRUCTURE correctly inferred *K*, but that this program could not detect more than one population at an $F_{ST} = 0.01$, and could not discern all five populations at an $F_{ST} = 0.02$. The iterative examination of *K*'s applied to TESS and GENECLUST proved to be more efficient than for STRUCTURE, because both algorithms regulate the number of clusters without the need of any additional statistical criterion (like an information criterion for example). The situation was less favourable to GENELAND which frequently overestimated the true number of distinct clusters. This overestimation issue seems an inherent drawback of the original RJMCMC method (Richardson & Green 1997), which is believed to exist even for lower dimensional search spaces. Although the existence of 'ghost' clusters was reported by Guillot *et al.* we were not able to correctly measure the impact of these fictive clusters on the assignment results.

Detecting genetic discontinuities

Determining what constitutes a natural break in continuous populations and delineating evolutionary significant units that form population subdivision are major objectives of population genetics and evolutionary biology. All programs performed well in this respect, as they were able to assign a large ratio of individuals to their population of origin at extremely low F_{ST} values. For the five-island simulations, the average proportion of an individual genome for TESS reached 98% for $F_{ST} = 0.04$. These results compare favourably to the 92% obtained with STRUCTURE for $F_{ST} = 0.05$ (Latch *et al*. 2006). The average total correct membership probability for GENELAND reached 91% for $F_{ST} = 0.04$. For F_{ST}'s greater or equal than 0.04, the average ratios of misassignment were lower than 1.1% for TESS, and lower than 8.5% for GENELAND. The spatial Bayesian clustering programs achieved even better performances when the five islands did not overlap spatially (not reported).

Individual-centred programs aim at detecting immigrants among samples analysed at various multi-allelic markers, using the fact that these immigrants will present different multilocus genotypes than expected for native individuals (Excoffier & Heckel 2006). A natural question is to know under what realistic conditions can Bayesian methods detect contemporary migration. The answer to this question strongly depends on the level of geographical admixture present in the population, that is the fact that individuals from separate ancestral origins share a fraction of the inhabited area.

When the degree of connectivity (spatial overlap) was higher than 30% ($D \leq 1$), the loss of spatial structure confounded GENELAND systematically. When the shape of the contact zone became very irregular or inconsistent, TESS still performed well, but STRUCTURE outperformed the spatial algorithms. At the other extreme ($D = 4$), the simulation study suggested that GENELAND performs well with no migrants and when genetic discontinuities correspond to crossing straight lines (see also Coulon *et al*. 2006). This can be seen in the case of maximal separation between the two islands, where GENELAND provided perfect assignment. Similar results were observed for the five-island simulations (not reported). Hence the power of GENELAND does not seem to rest on its ability to detect recent immigrants as claimed in (Guillot *et al*. 2005) and (Excoffier & Heckel 2006). Instead, the program seems more efficient at detecting potential zones of contact between populations without recent migrants, and living in geographical territories separated by simply shaped boundaries. In this case, GENELAND may be efficient even when the loss of connectivity is recent. In addition, GENELAND performances increased slightly when the F-model, which assumes an instantaneous fission from an ancestral population, was not used. This observation suggests that the above-mentioned limitation might come from a conflict between the spatial continuity assumption implemented through the prior distributions of the program and the fission model.

The island data with moderate levels of connectivity supported the hypothesis that TESS is superior to the other Bayesian clustering programs to detect a very recent contact zone between two weakly differentiated populations, and to identify which individuals crossed the boundary. For F_{ST}'s in the range (0.011, 0.018), TESS produced error rates twice lower than those produced by STRUCTURE, while GENELAND was hardly able to detect the correct structure (10 data sets, not reported). In particular, these results indicated that TESS could find a higher ratio of individuals from one population present into the second population than the other programs could. In this respect, TESS may have a greater ability to detect migrants.

Detecting clinal variation

The three Bayesian assignment methods revealed themselves efficient at detecting continuous variation in allele frequencies although spatial sampling was not continuous. In this respect, STRUCTURE performed better than the spatial algorithms, because it produced an estimation of the cline closer to the actual variation in allele frequencies. Nevertheless, we observed that TESS produced the best-estimated curve when using two loci instead of 10 (not reported). Although the cline may not be apparent when examining the results of a single run, averaging membership coefficients over several (high likelihood) runs is an efficient strategy to detect this type of variation. In general, the label-switching issue can be solved using the software CLUMPP (Jakobsson & Rosenberg 2006) before averaging.

The results presented here are consistent with those presented by François *et al*. 2006) which suggested that the degree of clustering might be diminished by use of higher prior levels of spatial clustering (i.e. higher ψ or lower λ). These results weakened some previous claims by Serre & Pääbo (2004) which alerted users to re-examine STRUCTURE results because discrete sampling during the experimental design might confound clustering algorithms. Although this effect may exist, our experimental results suggested that it actually has minor impact on dense regular sampling.

Program user interfaces

All programs have convenient user interfaces. Although being familiar with the R language may be more demanding than using the graphical interfaces of STRUCTURE and TESS, R offers additional data analysis and graphical functionalities that can facilitate the pre- and postprocessing of MCMC outputs. In addition, R is perfectly adapted to implement computational algorithms like MCMC, because the language can interact with foreign languages like C and FORTRAN (see Excoffier & Heckel 2006 for a different opinion). Nevertheless, the typical TESS runtimes are about 30-fold shorter than the GENELAND runtimes, and they are 10-fold shorter than the GENECLUST runtimes for moderately large data sets. In addition, running a few EM steps before starting the MCMC program can improve the final results when the sample sizes are very large.

The discussion is summarized in Table 3, which presents an objective ranking of the three Bayesian assignment approaches based on simulated data. The performances of GENECLUST were similar to TESS but with a lower computational speed. One may warn hurried readers that this summary should not be taken too literally. For example, the performance of spatial methods would increase if the weight on the spatial data were adjusted from the data (i.e. lower values of ψ when the connectivity is obviously high). Nevertheless, this summary suggests that combining analyses using STRUCTURE and TESS, like recently carried out in (Rosenberg *et al*. 2006), offers a convenient way to address the issue of detecting spatial population structure and locating discontinuities in allele frequencies.

Program availability

The TESS program is available from the Internet at the following URL: http://www-timc.imag.fr/Olivier.Francois/tess.html

Table 3 Program performance ranking for standard questions addressed by the three programs STRUCTURE, TESS and GENELAND (1 means best). The performances of GENECLUST were similar to TESS but with a lower computational speed

	Software ranking		
	1	2	3
Estimating the number of pop.	TESS	STRUCTURE	GENELAND
Assignment			
high geographical admixture	STRUCTURE	TESS	GENELAND
moderate geographical admixture	TESS	STRUCTURE	GENELAND
no geographical admixture	GENELAND TESS		STRUCTURE
Identifying recent migrants	TESS	GENELAND STRUCTURE	
Detecting clinal variation	STRUCTURE	TESS GENELAND	
Computational speed	STRUCTURE TESS		GENELAND

Acknowledgements

We are grateful to Emily Latch who kindly communicated the five-island data. We thank Fred Mortier for discussions about GENELAND internal parameter settings, and Noah Rosenberg for suggestions about the TESS program. We are also acknowledgeable to one reviewer for her/his constructive comments that contributed to improve the manuscript. This work was conducted while Chibiao Chen visited TIMB, and it is supported by grants from the Agence Nationale de la recherche project 'Modèles Aléatoires pour l'Evolution et le Vivant'.

References

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

Celeux G, Forbes F, Peyrard N (2003) EM procedures using mean fieldlike approximations for Markov model-based image segmentation. *Pattern Recognition*, **36**, 131–144.

Chen C, Forbes F, François O (2006) FASTRUCT: model-based clustering made faster. *Molecular Ecology Notes*, **6**, 980–983.

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

Coulon A, Guillot G, Cosson J *et al*. (2006) Genetic structure is influenced by lansdcape features. Empirical evidence from a roe deer population. *Molecular Ecology*, **15** (6), 1669–1679.

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

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Epperson BK (2003) *Geographical Genetics*. Princeton University Press, New Jersey.

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

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

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

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis*. Chapman & Hall, London.

Guillot G, Estoup A, Mortier F, Cosson J (2005) A spatial statistical model for landscape genetics. *Genetics*, **170** (3), 1261–1280.

Jakobsson M, Rosenberg NA (2006) *CLUMPP: CLUster Matching and Permutation Program*. Available from URL: http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html.

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

Manel S, Gaggiotti O, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.

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

Richarson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59** (4), 731–792.

Rosenberg N, Pritchard J, Weber J *et al*. (2002) Genetic structure of human populations. *Science*, **298**, 2981–2985.

Rosenberg N, Mahajan S, Ramachandran S, Zhao C, Pritchard J, Feldman M (2005) Clines, clusters, and the effect of study design on the influence of human population structure. *PLoS Genetics*, **1** (6), e70.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C *et al*. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genetics*, **2** (**12**), e215.

Serre D, Pääbo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, **14**, 1679–1685.

Wasser S, Shedlock A, Comstock K, Ostrander E, Mutayoba B, Stephens M (2004) Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences, USA*, **101** (**41**), 14847–14852.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38** (6), 1358–1370.

Supplementary material

The following supplementary material is available for this article:

Figure S1. Ground truth for a typical 5 island data set (500 individuals, 20% spatial overlap). Spatial coordinates for each genetic cluster were simulated from standard Gaussian distributions. Each color represents a population of origin in the simulation.

Figure S2. Five island simulation results (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.01 to 0.10). Number of successes over ten replicated runs for GENELAND: filled triangles, TESS: filled squares, GENECLUST: empty squares (hidden by the filled squares). The results are for $K_{max} = K = 5$, which means that the number of populations is assumed to be correct.

Table S1. Five island model (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.02 to 0.05). Individual assignment data for GENELAND and GENECLUST. The results are for $K_{max} = K = 5$, which means that the number of populations is assumed to be correct.

Table S2. Five island model (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.02 to 0.05). Individual assignment data for TESS. The results are for $K_{max} = K = 5$. The number of populations is assumed to be correct.

Table S3. Five island model: Detecting the correct number of populations (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.02 to 0.05). Individual assignment data for GENELAND and GENECLUST. The results are for K_{max} = 6, that is the number of populations is assumed to be unknown.

Table S4. Five island model: Detecting the correct number of populations (500 individuals, 20% spatial overlap, 10 unliked codominant loci, F_{ST} ranging from 0.02 to 0.05). Individual assignment data for TESS. The results are for K_{max} = 6, that is the number of populations is assumed to be unknown.

Table S5. Two island model: Individual assignment data for STRUCTURE obtained for various levels of geographical distance (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical connectivity between the two subpopulations varies from 2% to 40%. The results are for K_{max} = 3, that is the number of populations is unknown.

Table S6. Two island model: Individual assignment data for GENELAND obtained for various levels of geographical distance (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical connectivity between the two subpopulations varies from 2% to 40%. The results are for K_{max} = 3, that is the number of populations is unknown.

Table S7. Two island model: Individual assignment data for TESS obtained for various levels of geographical distance (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical connectivity between the two subpopulations varies from 2% to 40%. The results are for K_{max} = 3, that is the number of populations is unknown.

Table S8. Two island model: Summary. Individual assignment data for STRUCTURE, GENELAND and TESS, averages obtained using 5 data sets for various levels of geographical distance (200 individuals, 20 unliked codominant loci, F_{ST} around 0.02). The geographical connectivity between the two subpopulations varies from 2% to 40%. The results are for K_{max} = 3, that is the number of populations is unknown.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/
10.1111/j.1471-8286.2007.01769.x
(This link will take you to the article abstract).

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