Molecular Ecology (2002) **11**, 2571–2581

A simulated annealing approach to define the genetic structure of populations

I. DUPANLOUP,* S. SCHNEIDER† and L. EXCOFFIER‡
Dipartimento di Biologia, Università di Ferrara, Via L. Borsari 46, 44100 Ferrara, Italy; †Laboratoire de Génétique et Biométrie, Département d'Anthropologie, Université de Genève, Switzerland; ‡Zoological Institute, University of Berne, Switzerland

Abstract

We present a new approach for defining groups of populations that are geographically homogeneous and maximally differentiated from each other. As a by-product, it also leads to the identification of genetic barriers between these groups. The method is based on a simulated annealing procedure that aims to maximize the proportion of total genetic variance due to differences between groups of populations (spatial analysis of molecular variance; SAMOVA). Monte Carlo simulations were used to study the performance of our approach and, for comparison, the behaviour of the Monmonier algorithm, a procedure commonly used to identify zones of sharp genetic changes in a geographical area. Simulations showed that the SAMOVA algorithm indeed finds maximally differentiated groups, which do not always correspond to the simulated group structure in the presence of isolation by distance, especially when data from a single locus are available. In this case, the Monmonier algorithm seems slightly better at finding predefined genetic barriers, but can often lead to the definition of groups of populations not differentiated genetically. The SAMOVA algorithm was then applied to a set of European roe deer populations examined for their mitochondrial DNA (mtDNA) HVRI diversity. The inferred genetic structure seemed to confirm the hypothesis that some Italian populations were recently reintroduced from a Balkanic stock, as well as the differentiation of groups of populations possibly due to the postglacial recolonization of Europe or the action of a specific barrier to gene flow.

Keywords: AMOVA, *Caproleus caproleus*, coalescent simulations, genetic barriers, simulated annealing, subdivided population

Received 13 June 2002; revision received 5 September 2002; accepted 5 September 2002

Introduction

Analysis of gene frequencies on a world and continental scale revealed that geographical distances are an important factor explaining a large portion of the genetic diversity of human populations (Barbujani & Sokal 1991; Excoffier *et al*. 1991; Cavalli-Sforza *et al*. 1994; Poloni *et al*. 1995, 1997). Under isolation-by-distance, genetic differences are inversely related to the amount of gene flow, which depends on geographical proximity between populations (Wright 1943; Morton *et al*. 1968; Malécot 1973). Migratory movements are not only a function of geographical distance, but are also influenced by the presence of particular ecological or cultural barriers. In the last few years, the concept of barriers and their effect on population differentiation has been discussed repeatedly in the literature (Barbujani & Sokal 1990, 1991; Sokal & Oden 1988; Sokal *et al*. 1989; Dupanloup de Ceuninck *et al*. 2000; Rosser *et al*. 2000). Several authors have tried to ascertain the impact of barriers on gene flow when their potential location is known from other evidence (Sokal & Oden 1988; Sokal *et al*. 1989; Dupanloup de Ceuninck *et al*. 2000). The goal was to test whether these barriers overlap with zones of rapid genetic change and to quantitatively evaluate their effect on the exchange of genes between populations.

Several techniques have been developed to detect the presence of *genetic* barriers, i.e. spatial areas where the rate of change of gene frequencies is particularly high, and to locate them (Barbujani *et al*. 1989; Barbujani & Sokal 1990; Stenico *et al*. 1998; Simoni *et al*. 1999). The barriers are detected using the observed genetic data, and once their

Correspondence: Isabelle Dupanloup. Fax: 39 532 24 9761; E-mail: dpi@unife.it

locations have been defined, they are compared with physical or cultural barriers. Available methods define zones of maximum genetic change either along a network connecting localities (the so-called Monmonier algorithm: see Monmonier 1973; Stenico *et al*. 1998; Simoni *et al*. 1999), or over interpolated allele-frequency surfaces (the so-called Wombling method: see Womble 1951; Barbujani *et al*. 1989).

We propose a new approach which indirectly detects genetic barriers in a sampling region but is especially designed to define groups of populations without the need for interpolation. When the sampling points are not regularly spaced in the region under study, the interpolation process leading to continuous allele frequency surfaces can sometimes introduce artefactual discontinuities (Sokal *et al*. 1999). In fact, our approach consists of defining groups of populations that are maximally differentiated from each other (i.e. those for which the proportion of total genetic variance due to differences between groups is maximum). As a by-product, these groups are separated from each other by a genetic barrier. In contrast to classical tests of genetic structure, in which groups of populations are defined a priori on the basis of physical, ecological, linguistic or cultural characters, our method enables one to find a group structure based solely on genetic data. Our approach is similar in spirit to that implemented in the program STRUCTURE proposed by Pritchard *et al*. (2000), which is a Bayesian clustering approach to assign individuals to populations. Their model assumes Hardy–Weinberg and linkage equilibria and attempts to define groups of individuals that minimize departures from these equilibria. In our case, a higher hierarchical level is considered: instead of defining groups of individuals, our goal is to define groups of populations. Moreover, we assign populations to groups with the constraint that they must be geographically adjacent and genetically homogeneous. Our approach also differs in that it can be applied to both genotypic and haplotypic data, and it makes no assumptions about Hardy–Weinberg equilibrium within populations, or about the linkage equilibrium between loci.

Below, we provide a description of this approach and evaluate its power using Monte Carlo simulations. For comparison purposes, we use the same simulated data to evaluate the performance of the Monmonier algorithm. We finally apply these two methods to the case of 18 roe deer (*Capreolus capreolus*) populations tested for mitochondrial HVRI sequence polymorphisms.

Materials and Methods

Monmonier algorithm

The first step in the Monmonier algorithm consists in connecting the sampled localities using a Delaunay network (Delaunay 1934; Brassel & Reif 1979), a graphical method for defining adjacent points on a map (see Fig. 1 for an example of such a Delaunay network). Genetic distances (estimated here as pairwise F_{ST} distances) are then computed between all pairs of localities that are connected by direct edges in the Delaunay network. A genetic barrier is initiated by tracing a perpendicular line across the edge with the highest associated genetic distance. It is then extended progressively across the adjacent edges associated to the highest genetic distances until the line reaches the border of the network, or until it closes a circle around one or more localities. An indirect output of the Monmonier algorithm is the definition of distinct groups of populations located on the two sides of the genetic barrier.

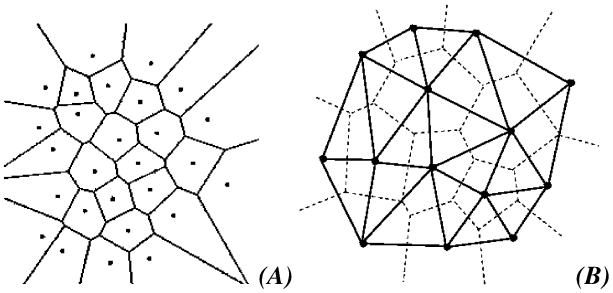

Fig. 1 Example of a Voronoi diagram (A) and a Delaunay triangulation (dark line) superimposed on the corresponding Voronoi diagram (dashed line) (B). Given a set of n points, the Delaunay triangulation is a set of lines connecting each point to its nearest neighbours. For the same set of points, the Voronoi diagram corresponds to the partition of the space into n convex polygons such that every point within a given polygon is closer to its central point than to the central point of any other polygon. Delaunay triangulation and Voronoi diagrams are related: edges of a Voronoi diagram are perpendicular bisectors of branches of a Delaunay network.

Spatial analysis of molecular variance (SAMOVA): partitioning the populations into genetically and geographically homogeneous groups

We start with an initial random partition of the n sampled populations into K groups (K is here assumed to be known). We then use a simulated annealing procedure to find the composition of the K groups and to maximize the F_{CT} index, which is the proportion of total genetic variance due to differences between groups of populations (see e.g. Excoffier *et al*. 1992).

Simulated annealing is an optimization technique that is applicable to a wide variety of problems. It is inspired by the process through which a metal cools and freezes into a crystalline structure with minimum energy (the annealing process). An algorithm mimicking this cooling process was initially proposed by Metropolis *et al*. (1953) to find

the equilibrium configuration of a collection of atoms at a given temperature. The connection between this algorithm and mathematical optimization procedures was then noted by Kirkpatrick *et al*. (1983), who proposed it as a general optimization technique for combinatorial and other problems. An advantage of simulated annealing is its ability to avoid becoming trapped at a local optimum. The algorithm uses a random search that not only accepts changes that decrease (or increase) a particular function to optimize, but also changes that lead to suboptimal solutions. It does this with a probability that decreases with the number of steps already performed in the optimization process. The underlying assumption is that, as times passes, we should get closer to a global optimum and be less prone to accept departure from that optimum. At the beginning of the process one thus tolerates frequent escapes from local optimum and accepts suboptimal solutions to explore the solution space more widely. The algorithm we call SAMOVA can be decomposed into the following steps.

Preliminary steps

1 A set of Voronoi polygons (Voronoi 1908) is constructed from the geographical location of the n sampled points (see Fig. 1).

2 An arbitrary partition of the n populations into K groups is initially chosen at random (in our case, each group, except one, is composed of a single population and the last group contains the populations not assigned to any other groups).

3 The genetic barrier(s) between the K groups are identified as edges of Voronoi polygons separating groups of populations.

4 The F_{CT} index associated to the K groups is computed.

Simulated annealing steps

5 We select an edge at random on a given barrier.

6 The two populations located on both sides of the selected edge are identified, and one population chosen at random is assigned to the group of the other population.

7 The genetic barrier is modified by updating the list of edges separating the newly defined groups of populations [the edge selected in step 6 is replaced by the edges surrounding the population whose group location has changed (see Fig. 2A)].

8 The new F_{CT} value (noted F_{CT}*) associated with the new partition is computed.

9 The new structure is accepted with probability

$$p = \begin{cases} 1 & \text{if} \quad F_{CT}^* \geq F_{CT} \\ e^{(F_{CT}^* - F_{CT})S^A} & \text{if} \quad F_{CT}^* < F_{CT} \end{cases}$$

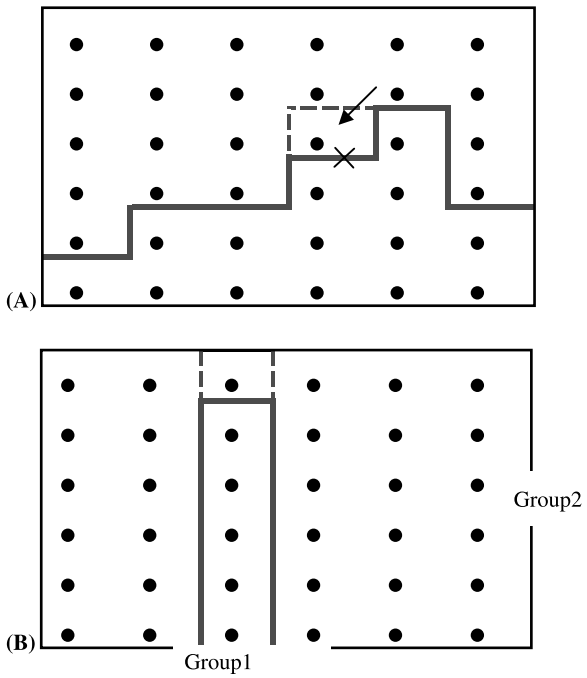

Fig. 2 Modification of the group structure under steps 5 and 6 of the simulated annealing algorithm. (A) The bold line delineates the limit between the two groups before the modification. The cross designates the edge selected at random under step 5. The arrow designates the population whose group location is to be changed under step 6. The dashed line(s) define the new barrier(s) between groups after step 6 has been performed. (A) Normal case leading to geographically adjacent groups. (B) Case in which the allocation of one population from group 1 to group 2 leads to the fragmentation of group 2 into two distinct sets of adjacent populations.

where S is the number of steps performed in the simulated annealing process, and A is an arbitrary constant controlling the speed of the cooling process. Steps 5–9 are then repeated 10 000 times. The constant A was set to 0.9158 such that the probability p defined above is equal to 1% if the difference between F_{CT} and F_{CT}* at the 10 000th iteration is equal to 0.001. It means that we have a probability of 1% to accept a slightly worse F_{CT} value at the end of the annealing process. To ensure that the final configuration of the K groups is not affected by a given initial configuration, the simulated annealing process is repeated 100 times, starting each time from a different initial partition of the n samples into the K groups. The configuration with the largest associated F_{CT} value after the 100 independent simulated annealing processes is retained as the best grouping of populations.

Step 6 of the above-mentioned process should ensure that the inferred groups are composed of adjacent populations: samples that are neighbours (because they are on both sides of an edge of the Voronoi diagram) and initially members of different groups are then grouped together. However, despite this constraint, SAMOVA can sometimes

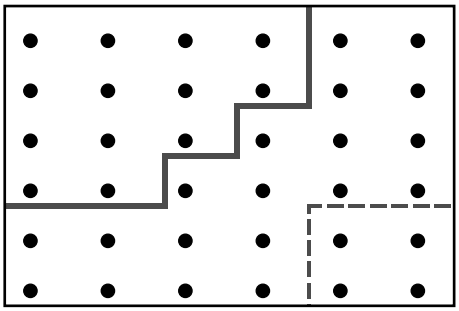

Fig. 3 Conditions of the simulations: we simulated 36 samples on a two-dimensional 6×6 stepping-stone grid; demes were arranged in two (or three) groups separated by the lines.

lead to the definition of groups in which all the populations are not geographically adjacent, as shown in Fig. 2B. It can thus end with a partition of two distinct sets of geographically adjacent population belonging to the same group.

Simulation study

We have performed two series of Monte Carlo simulations to evaluate the performance of the SAMOVA methodology, but also that of the Monmonier algorithm, whose performance to detect genetic barriers had never been assessed to our knowledge. Using a coalescent approach (Excoffier *et al*. 2000), we first reconstructed the genealogies of genes sampled from 36 demes arranged on a two-dimensional stepping-stone grid. The demes were arranged in two or three groups as shown in Fig. 3. We studied the effect of different (but constant) levels of gene flow between adjacent demes within or between groups (see Table 1, Table 2 and Table 3 for migration parameters).

Poisson-distributed mutations were then introduced onto the realized genealogical trees, assuming either finite-sites mutational models, for the simulations of DNA sequences (200 bp), or the stepwise-mutation model (SMM) for the simulation of microsatellite data at 5, 10 or 20 unlinked microsatellite loci. The size of the samples was always set to 20 haploid individuals, and the mutation parameter value $\theta = 2N\mu$ was set to 0.4 for the simulation of DNA sequences and θ was set to 1 per locus for the simulation of microsatellite data. These values were chosen such as to approximately reflect the mutation rates at work in mammals. The haploid deme size N was always set to 1000.

We also undertook a series of simulations based on a finite-islands model with 36 demes arranged into two

Table 1 Results of the simulation study under the stepping-stone model (one locus, 200 bp DNA sequence)

Cases	*Nm* intra	*Nm* inter	Method	F_{CT}*	Correct groups†	Stronger groups‡	Weaker groups§
Two groups							
1	1	0.1	M¶	0.192 [–0.003; 0.695]	0.085	0.523	0.394
			S**		0.023	0.977	0
2	1	0.01	M	0.573 [0.028; 0.950]	0.702	0.137	0.161
			S		0.531	0.469	0
3	10	0.1	M	0.209 [0.012; 0.789]	0.525	0.164	0.311
			S		0.383	0.617	0
4	10	0.01	M	0.636 [0.055; 0.965]	0.964	0.007	0.029
			S		0.924	0.076	0
5	100	0.1	M	0.209 [0.013; 0.766]	0.623	0.105	0.272
			S		0.572	0.428	0
6	100	0.01	M	0.635 [0.110; 0.970]	0.974	0.003	0.023
			S		0.960	0.040	0
Three groups							
7	1	0.1	M	0.229 [0.008; 0.730]	0.028	0.644	0.328
			S		0.002	0.998	0
8	10	0.1	M	0.253 [0.043; 0.767]	0.365	0.348	0.287
			S		0.134	0.866	0
9	10	0.01	M	0.700 [0.116; 0.963]	0.952	0.029	0.019
			S		0.668	0.332	0

*Mean F_{CT} value associated with the simulated group (minimum and maximum values are given within brackets). †Fraction of simulated groups retrieved of 1000 simulated cases. ‡Fraction of barriers retrieved with a larger F_{CT} than that associated with the simulated barrier. §Fraction of groups retrieved with a smaller associated F_{CT} than that associated with the simulated barrier. ¶Results obtained using Monmonier algorithm. **Results obtained using the SAMOVA algorithm.

Table 2 Results of the simulation study under the stepping-stone model. Simulation of 5, 10 or 20 independent microsatellite loci for different levels of gene flow within and between groups

Cases	*Nm* intra	*Nm* inter	No. of loci	Method	F_{CT}	Correct groups	Stronger groups	Weaker groups
Two groups								
1	1	0.1	5	M	0.182 [–0.011; 0.755]	0.115	0.456	0.429
				S		0.040	0.960	0
2			10	M	0.192 [0.018; 0.511]	0.249	0.335	0.416
				S		0.098	0.902	0
3			20	M	0.198 [0.038; 0.514]	0.478	0.163	0.359
				S		0.211	0.789	0
4	10	0.1	5	M	0.203 [-1×10^{-4}; 0.633]	0.541	0.135	0.324
				S		0.416	0.584	0
5			10	M	0.218 [0.025; 0.663]	0.810	0.035	0.155
				S		0.763	0.237	0
6			20	M	0.220 [0.078; 0.573]	0.955	0.007	0.038
				S		0.941	0.059	0
7	10	0.01	5	M	0.667 [0.059; 0.975]	0.987	0.002	0.011
				S		0.977	0.023	0
8			10	M	0.698 [0.175; 0.954]	0.999	0	0.001
				S		0.997	0.003	0
9			20	M	0.712 [0.341; 0.922]	1.000	0	0
				S		1.000	0	0

Footnotes as Table 1.

Table 3 Results of the simulation of group fission and island model (one locus, 200 bp DNA sequences)

Cases	Divergence time	Method	F_{CT}	Correct groups	Stronger groups	Weaker groups
Two groups						
1	5*N* generations	M	0.213 [0.028, 0.796]	0.675	0.071	0.254
		S		0.682	0.318	0
2	10*N* generations	M	0.362 [0.070, 0.908]	0.906	0.016	0.078
		S		0.918	0.082	0
3	20*N* generations	M	0.530 [0.086, 0.943]	0.989	0.001	0.010
		S		0.994	0.006	0

Footnotes as Table 1.

groups supposed to have diverged some T generations ago. Within these two groups made up of 23 and 13 demes, respectively, each deme exchanges genes with all other demes at constant and high rates ($Nm = 10$). Group separation time was fixed at either $T = 5N$, $10N$ or $20N$ generations before present, where N is the size of each deme. We further assume that after the group fission no further gene flow occurred between the two groups, and that before group fission all demes exchanged genes at the same rate as within group after the fission. In that case, genetic data were simulated as samples of DNA sequences.

For each set of demographic parameters, we performed 1000 coalescent-based simulations. The Monmonier and the SAMOVA algorithms were applied after each simulation to attempt to recover the simulated groups of demes.

The molecular distances between pairs of sequences necessary for the computations of F_{CT} values (in the case of the SAMOVA algorithm) and F_{ST} distances (in the case of the Monmonier algorithm) were computed as pairwise differences for the DNA sequences, and as sums of squared size differences for microsatellite data (Michalakis & Excoffier 1996).

Results

Simulation study

Stepping-stone model: 1 locus. The average, minimum and maximum values of the F_{CT} indices associated with the simulated barriers are shown in Table 1. As expected, F_{CT}

mean values are larger for smaller intergroup migration rates. They are also slightly larger when the samples are separated in three groups than in two. Overall, the Monmonier algorithm performs better than the SAMOVA algorithm for finding the simulated groups, especially when the amount of gene flow within group is small (cases 1 and 2). We note, however, that when the SAMOVA algorithm does not find the predefined groups, it always finds groups of populations that are more differentiated than the simulated groups. This suggests that the SAMOVA algorithm is able to identify maximally differentiated groups, whereas the Monmonier algorithm is better at finding genetic barriers between sets of populations. This different behaviour indeed reflects their conceptual difference.

The performance of both algorithms is very good only when there is a very low level of gene flow between groups ($Nm = 0.01$) and when gene flow within groups is at least 1000 times larger than gene flow between groups (cases 4 and 6, with > 90% success in recovering the correct groups). When one of these two conditions is not met, the success rate drops sharply (cases 2 and 5, with success rates in the range 50–70%). When both conditions are not met (cases 1 and 3), the performance of both methods is lower, especially when the level of gene flow within group is relatively low (case 1). When three groups are simulated, the same conclusions remain valid (cases 7–9 in Table 1).

Stepping-stone model: 5, 10 and 20 microsatellite loci. The results of the multilocus simulations are given in Table 2. The mean F_{CT} value computed by taking into account the difference in repeat number between alleles (Michalakis & Excoffier 1996) is found to be very close to those obtained from DNA sequences at one locus. We observe, however, that the mean F_{CT} value increases with the number of simulated microsatellite loci. As expected, the simulated groups are identified in a larger number of cases when the number of loci is increased with both the Monmonier and the SAMOVA algorithms. However, when the level of gene flow between groups is only 10-fold lower than that within groups (case 1), the correct groups are found in < 50% of the simulations even with 20 loci. As in the single-locus case described earlier, good results are obtained only when the *Nm* value between group is ≤ 0.01 or lower. Note also that the Monmonier algorithm quite often finds a genetic barrier associated with less differentiated groups than those simulated (cases 1–6) when the *Nm* value between groups is > 0.01; this is true even with 20 loci.

Island model with group divergence. Under these simulation conditions (see Table 3), both the Monmonier and SAMOVA algorithms identify the simulated structure in a larger number of cases than under the stepping-stone model for roughly similar F_{CT} values. In that case, however, the performance of both the Monmonier and SAMOVA algorithms appears very close, with even a slight advantage to the SAMOVA method. Note that ≈ $10N$ generations of complete divergence between groups are necessary to identify groups correctly from a single DNA sequence locus.

Application to roe deer populations of Europe

The European roe deer (*Capreolus capreolus*) is widespread in Europe but its genetic structure has been strongly affected by two historical events: the recolonization of Europe from glacial refugia at the end of the Pleistocene (Randi *et al*. 1998; Wiehler & Tiedemann 1998; Vernesi *et al*. 2002) and, in more recent times, habitat fragmentation and restocking for hunting purposes (Randi *et al*. 1998; Vernesi *et al*. 2002). To investigate in greater detail its genetic structure, we applied SAMOVA and Monmonier algorithms to define groups and find the location of the most important genetic barriers in the distribution map of 18 roe deer populations tested for mitochondrial DNA (mtDNA) HVR1 sequence polymorphisms (Vernesi *et al*. 2002). Figure 4 shows the location of the 18 samples considered and the final allocation of roe deer populations into groups for the two algorithms. Table 4 gives the composition of the corresponding groups of populations inferred by the two algorithms, their associated fixation indices, and their significance evaluated by permuting the populations without considering their geographical position (Excoffier *et al*. 1992).

The two approaches show the genetic peculiarities of several Italian populations, but lead to quite different groups of populations. The SAMOVA approach suggests the association of the Ligurian sample with the samples from southeastern Europe. This result is in agreement with the hypothesis that the Ligurian population was reintroduced in recent times from a Balkanic source (Vernesi *et al*. 2002). The SAMOVA approach also suggests the existence of two other groups in central Italy. To explain it, Vernesi *et al*. (2002) have proposed the existence of two glacial refugia in Italy (one in the southern Alps and one in the Siena-Castel Porziano area), or, alternatively, the presence of a very efficient barrier to gene flow (the Arno River), which would have prevented short-range migrations between populations from these regions.

The Monmonier algorithm gives a quite different view. It confirms the genetic peculiarities of the Ligurian sample and that of the populations in the central Italian region. However, the association of these populations in the same group does not lead to a significant F_{CT} value (Table 4, row 1). The Monmonier algorithm also suggests the separation of western European (including northwest Italy) and eastern European samples (see Table 4 and Fig. 4). This result could be due in part to the differentiation of populations

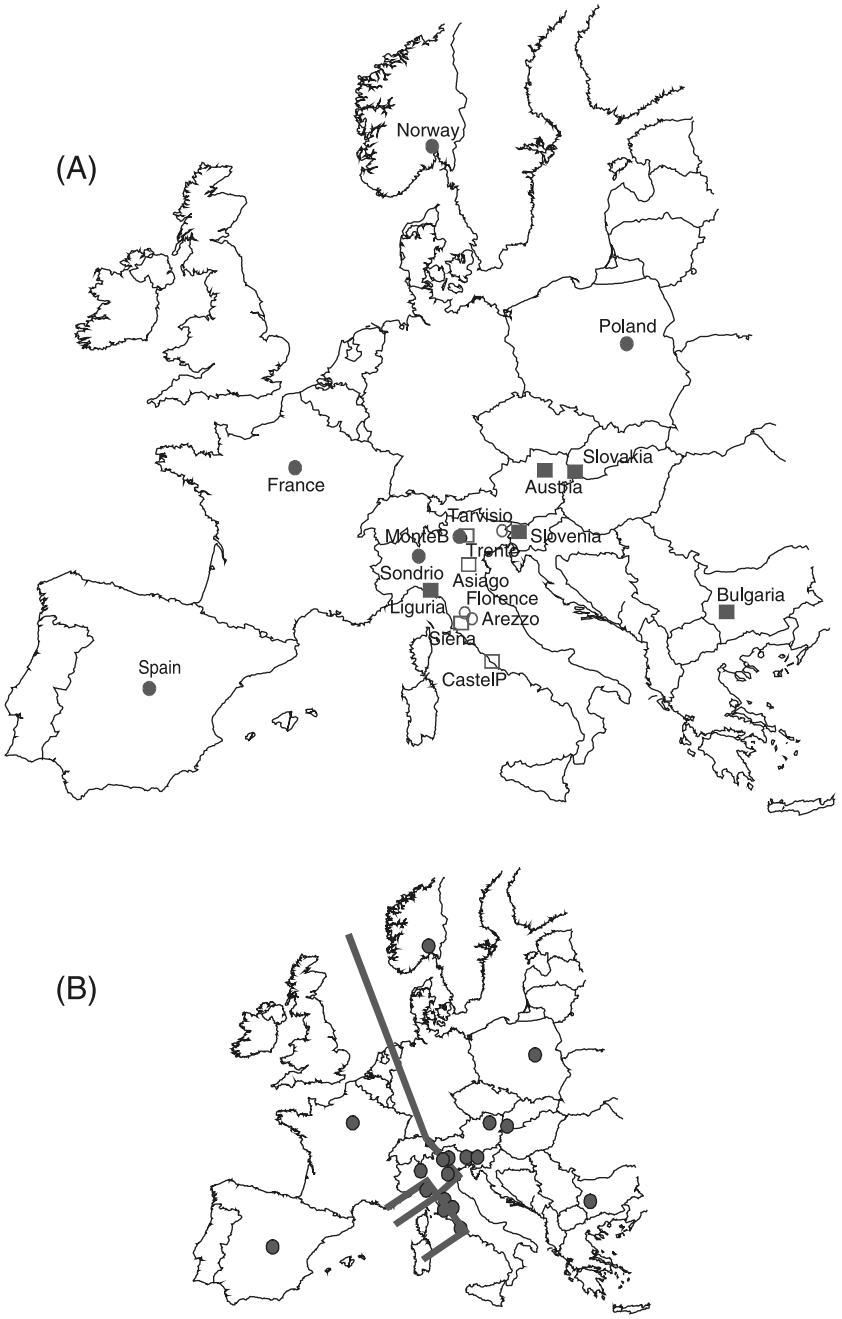

Fig. 4 Distribution of the 18 roe deer populations tested for mtDNA sequences polymorphisms (Vernesi *et al*. 2002) and group structure defined by SAMOVA (samples with the same symbol belong to the same group) (A) as well as genetic barriers detected by Monmonier (B). See also Table 4.

that were restricted to the Balkans and the Iberian regions during the last glacial maximum.

Overall, the F_{CT} values associated with groups defined using the SAMOVA algorithm are much higher than those defined using the Monmonier algorithm (Table 4), in agreement with smaller F_{SC} values representing the extent of differentiation between populations within groups. The larger extent of differentiation between groups associated with a greater homogeneity of populations within groups supports the view that the grouping found using the SAMOVA algorithm is more reliable than that obtained using the Monmonier algorithm.

Discussion

The AMOVA approach (Excoffier *et al*. 1992) has been widely used for the hierarchical analysis of the genetic diversity in a set of sampled populations. A physical, ecological, linguistic or cultural criterion is often used to define a priori groups of populations on which a test of genetic structure is applied. Where no obvious criterion exists for the definition of groups of populations, the investigation of the genetic structure in a set of populations may be difficult. The goal of the method we have presented here

Table 4 Fixation indices corresponding to the groups of populations inferred by Monmonier and SAMOVA algorithms for the 18 *Capreolus capreolus* populations tested for mitochondrial HVR1 sequences

	Method	Group composition	F_{SC}	F_{ST}	F_{CT}
Two groups	M	1. Castelporziano + Siena + Liguria 2. Other populations	0.519**	0.594**	0.156^{NS}
	S	1. Liguria + Bulgary 2. Other populations	0.506**	0.665**	0.323*
Three groups	M	1. Liguria 2. Castelporziano + Siena 3. Other populations	0.454**	0.635**	0.331**
	S	1. Liguria + Bulgary + Slovenia 2. Castelporziano + Siena 3. Other populations	0.428**	0.629**	0.351**
Four groups	M	1. Liguria 2. Castelporziano + Siena 3. Western group (see Fig. 4) 4. Eastern group (see Fig. 4)	0.398**	0.588**	0.316**
	S	1. Liguria + Austria + Slovenia + Bulgaria + Slovakia 2. Castelporziano + Siena + Trento + Asiago 3. Arezzo + Florence + Tarvisio 4. Other populations	0.295**	0.590*	0.418**

*$P < 0.01$; **$P < 0.001$.

is to allow one to define the strongest structure of populations in genetic terms. As a by-product, it also leads to the identification of genetic barriers between the inferred groups and represents thus an alternative to other methods (Monmonier algorithm, wombling) for finding such barriers.

Here, we also present the first evaluation of the behaviour of the Monmonier algorithm, which has been applied several times in a human population genetics context (e.g. Stenico *et al*. 1998; Simoni *et al*. 1999). Despite its simplicity, this algorithm allows the identification of the simulated barriers in a larger number of cases than does the SAMOVA method, but it also leads to the identification of weaker genetic barriers in an important fraction of the cases. This spurious behaviour is probably due to the directional and incremental nature of the algorithm, which makes it very sensitive to local minima. The local and strong differentiation of some populations may indeed initiate or extend barriers that do not necessarily lead to maximally differentiated groups on a more global scale. In contrast, the ability of SAMOVA to tolerate suboptimal solutions allows it to ultimately find a global maximum, and to avoid becoming trapped at local optimum.

The roe deer example illustrates some other features of the SAMOVA algorithm. The ability of our new method to define groups in which *all* the populations are not geographically adjacent (i.e. the Ligurian sample is associated with samples from southeastern Europe) can thus allow one to identify recent reintroduction events, which is particularly important for conservation genetics purposes. We further tested this ability by new simulations in which 36 populations tested at 10 microsatellites were arranged according to the first configuration shown in Fig. 5. In this configuration, populations on the left and right of the grid exchange migrants at a large rate, as if on a cylinder. We show in Table 5 that SAMOVA recognizes the simulated genetic structure in > 81% of cases. This result indicates the strong capacity of SAMOVA to detect a quite complex structure, which cannot be identified by Monmonier's algorithm as it must cluster geographically adjacent localities.

Another difference between the two methods is the fact that the configuration defined for small number (K) of groups of populations are not necessarily preserved with larger K-values (compare in Table 4, the composition of the first group when two, three or four groups are desired). The goal of SAMOVA is indeed to find the strongest group structure for a given number of groups of populations, and, unlike the Monmonier algorithm, it does not incrementally add barriers one after the other until the desired number of groups has been reached.

The genetic structure identified by our new method seems to confirm the Balkan origin of the Ligurian population proposed by Vernesi *et al*. (2002). The identification of two groups of populations in Italy is in agreement with the hypothesis concerning the existence of two former glacial refugia in Italy, or alternatively, of the action of a specific natural barrier to gene flow (possibly the Arno River). These results obtained from the analysis of a single locus and from a small number of populations are, however, preliminary, and would need to be confirmed with a larger number of (nuclear) markers.

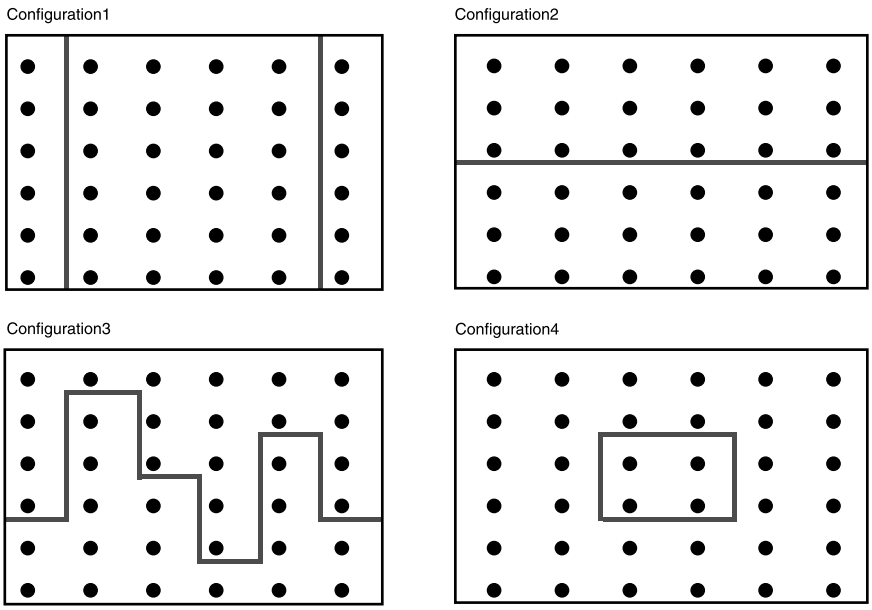

Fig. 5 Four other conditions of simulations: 36 samples on a two-dimensional 6×6 stepping-stone grid arranged in two groups separated by lines. The spatial configuration of the barriers are simple (configuration 2) or complex (configuration 3) and its effect on the behaviour of SAMOVA and Monmonier is tested (see text). In configuration 1, samples on the left and right sides of the grid exchange migrants at a large rate, i.e. are defined as members of the same group which is thus divided in two by the populations of the other group.

Table 5 Effects of the spatial configuration of the groups on the performance of SAMOVA and the Monmonier algorithm

Configuration	*Nm* intra	*Nm* inter	Method	F_{CT}	Correct groups	Stronger groups	Weaker groups
1	1	0.01	M	0.545 [0.104; 0.926]	0	0.020	0.980
			S		0.814	0.186	0
2	1	0.1	M	0.240 [0.015; 0.775]	0.376	0.209	0.415
			S		0.225	0.775	0
3	1	0.1	M	0.100 [−0.011; 0.382]	0.018	0.663	0.319
			S		0	1.000	0
4	1	0.1	M	0.140 [−0.031; 0.656]	0.217	0.361	0.422
			S		0.055	0.945	0

Minimum and maximum values of F_{CT} are shown within brackets. Data from 10 unlinked microsatellite loci were simulated in 36 samples of 20 haploid individuals drawn from demes arranged under the stepping-stone models shown in Fig. 5.

Finding the correct number of groups

Our new approach requires the a priori definition of the number (K) of groups of populations to identify. Because our method is based on the maximization of the proportion of total genetic variance due to the differences between groups (F_{CT}), it is reasonable to think that this proportion may vary with the parameter K for the same set of populations. We thus expect that F_{CT} should increase with K because of the reduction of the proportion F_{SC} of variance due to differences between populations within each group. This is because in the classical relationship $(1 - F_{ST}) = (1 - F_{SC})(1 - F_{CT})$ the F_{ST} index should be quite insensitive to K, but as K increases, the number of populations becomes smaller, leading to a decrease of F_{SC} and a corresponding increase of F_{CT}. In order to see if we can recover the true number of groups from the data, we simulated the genetic diversity among 36 populations at 10 microsatellites under the hypothesis that either 2 or 3 groups exist. In both cases we applied the SAMOVA algorithm searching for 2, 3, 4 or 5 groups. Table 6 details the mean value and the standard deviations of the F_{CT} indices associated with either 2, 3, 4 or 5 groups. We find that the largest mean F_{CT} value is associated with the correct number of simulated groups, suggesting that F_{CT} has some power to retrieve the unknown number of groups. We note, however, that the mean F_{CT} values inferred for different numbers of groups are obviously not significantly different from each other, especially when the difference between groups is weak. We give also in Table 6 the fraction of simulations for which the largest F_{CT} index is obtained for the correct number of groups. By using SAMOVA successively on the

Table 6 Mean F_{CT} values inferred by the SAMOVA algorithm when searching for either 2, 3, 4 or 5 groups of populations

Correct no. of groups	*Nm* intra	*Nm* inter	F_{CT} simulated barrier	Average F_{CT} 2 groups	Average F_{CT} 3 groups	Average F_{CT} 4 groups	Average F_{CT} 5 groups	Correct inference*
2	1	0.1	0.192 (0.092)	0.281 (0.075)	0.278 (0.071)	0.276 (0.069)	0.276 (0.067)	0.406
	10	0.1	0.218 (0.098)	0.224 (0.093)	0.220 (0.091)	0.216 (0.089)	0.211 (0.086)	0.775
	10	0.01	0.698 (0.134)	0.698 (0.134)	0.692 (0.135)	0.682 (0.135)	0.672 (0.136)	0.989
3	1	0.1	0.233 (0.082)	0.320 (0.084)	0.322 (0.079)	0.299 (0.075)	0.322 (0.071)	0.083
	10	0.1	0.271 (0.085)	0.307 (0.101)	0.308 (0.098)	0.298 (0.094)	0.286 (0.088)	0.288
	10	0.01	0.747 (0.098)	0.691 (0.100)	0.753 (0.095)	0.749 (0.096)	0.743 (0.098)	0.701

Standard deviations of F_{CT} are shown within parentheses. Data from 10 unlinked microsatellite loci were simulated in 36 samples of 20 haploid individuals drawn from demes arranged under the stepping-stone model shown in Fig. 3. Either 2 or 3 groups were simulated, as reported above. *Fraction of simulations with the larger F_{CT} index corresponding to the correct number of simulated groups.

same datasets with different *K*, and using the F_{CT} index as a test statistic, we were able to identify the correct number of groups from 8.3 to 98.9% of the cases. We note that the identification of the correct number of groups depends critically on the degree of differentiation between groups and the absence of isolation-by-distance within groups, and should increase with the number of available loci. However, further work is clearly needed to infer the correct number of groups as the F_{CT} index does not seem particularly efficient in that respect.

Isolation by distance as a nuisance parameter

Simulations show that our new approach always allows one to identify the simulated group structure or another configuration associated with a larger F_{CT} index. The non-identification of weaker configurations seems to indicate that our simulated annealing strategy is suitable for maximizing the proportion of total genetic variance due to differences between groups of populations. Under a stepping-stone model of population and group differentiation, our new method performs less well than the Monmonier algorithm for the identification of the correct group structure, particularly when only one locus is used (Table 1). However, the reverse situation is observed when we simulate a fission of population into groups that do not exchange migrants (Table 3). This difference in behaviour reflects the conceptual difference of these two methods: the Monmonier algorithm is better at finding genetic barriers between sets of populations, whereas the SAMOVA algorithm finds maximally differentiated groups.

This difference could be due to the presence of a pattern of isolation-by-distance in the simulations performed under a stepping-stone model. Our simulations show indeed that under a stepping-stone model the ability of our method to identify the true structure is strongly dependent on the amount of migrants exchanged between populations within groups. In cases 2, 4 and 6 of Table 1 (for a *Nm* intergroup value of 0.01) the SAMOVA approach identifies the correct groups in 53.1% (*Nm* intragroup = 1), 92.4% (*Nm* intragroup = 10) and 96.0% (*Nm* intragroup = 100), respectively. This result shows that when the effect of isolation-by-distance within group is suppressed by increasing the amount of gene flow (with *Nm* intragroup >> 1), the behaviour of SAMOVA is much more satisfactory. This result is in keeping with a former study on the use the F_{CT} index to measure the genetic differentiation associated with cultural barriers (Dupanloup de Ceuninck *et al*. 2000). Monte-Carlo simulations indeed showed that this statistic could reveal significant differences between groups of populations in the absence of any genetic boundary, but in the presence of isolation-by-distance (Dupanloup de Ceuninck *et al*. 2000).

Additional simulations with a different spatial configuration of the groups of populations (Fig. 5, Table 5) indicate that the capacity of SAMOVA to detect the simulated group structure is strongly dependent on this configuration. Groups with complex spatial structure are more difficult to identify as they also imply a stronger level of isolation-by-distance and the isolation of some populations.

The sensitivity of SAMOVA to isolation-by-distance is more pronounced when only one locus is used, as it performs much better when more loci are analysed (compare, for example, case 3 in Table 1 with cases 4, 5 and 6 in Table 2). With one locus, the stochasticity of the coalescent process is very large, apparently leading to important genetic differences other than those intended under the simulations. From these results and the more general observation that the genetic diversity observed at a single locus represents just one realization of an evolutionary process with a large stochastic component, the application of our method to single locus data should be interpreted with caution, and its application to multilocus data should be preferred, whenever possible.

Acknowledgements

Many thanks to Guido Barbujani and Giorgio Bertorelle for stimulating discussion on this manuscript. We are grateful to Cristiano Vernesi and Giorgio Bertorelle for the access to the roe deer population database. This work was supported by a Swiss NSF grant 31-054059.98 to LE, and a Swiss NSF grant for advanced researchers to ID. A program to perform the computations described is available from ID upon request.

References

Barbujani G, Oden NL, Sokal RR (1989) Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Science*, **87**, 1816–1819.

Barbujani G, Sokal RR (1991) Genetic population structure of Italy. I. Physical and cultural barriers to gene flow. *Annals of Human Genetics*, **48**, 398–411.

Brassel KE, Reif D (1979) A procedure to generate Thiessen polygons. *Geographical Analysis*, **11**, 289–303.

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

Delaunay B (1934) Sur la sphère vide. *Bulletin of the Academy of Sciences of the USSR*, **7**, 793–800.

Dupanloup de Ceuninck I, Schneider S, Langaney A, Excoffier L (2000) Inferring the impact of linguistic boundaries on population differentiation: application to the Afro-Asiatic-Indo-European case. *European Journal of Human Genetics*, **8**, 750–756.

Excoffier L, Harding R, Sokal RR, Pellegrini B, Sanchez-Mazas A (1991) Spatial differentiation of RH and GM haplotype frequencies in sub-Saharan Africa and its relation to linguistic affinities. *Human Biology*, **63**, 273–297.

Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, **91**, 506–509.

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

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Malécot G (1973) Isolation by distance. In: *Genetic Structure of Populations* (ed. Morton NE), pp. 72–75. University of Hawaii Press, Honolulu.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, **142**, 1061–1064.

Monmonier MS (1973) Maximum–difference barriers: an alternative numerical regionalization method. *Geographical Analysis*, **3**, 245–261.

Morton N, Miki C, Yee S (1968) Bioassay of population structure under isolation by distance. *American Journal of Human Genetics*, **20**, 411–419.

Poloni ES, Excoffier L, Mountain JL, Langaney A, Cavalli-Sforza LL (1995) Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. *Annals of Human Genetics*, **59**, 43–61.

Poloni ES, Semino O, Passarino G, *et al.* (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics*, **61**, 1015–1035.

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

Randi E, Pierpaoli M, Danilkin A (1998) Mitochondrial DNA polymorphism in populations of Siberian and European roe deer (*Capreolus pygargus* and *C. capreolus*). *Heredity*, **80**, 429–437.

Rosser ZH, Zerjal T, Hurles ME *et al.* (2000) Y-Chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics*, **67**, 1526–1543.

Simoni L, Gueresi P, Pettener D, Barbujani G (1999) Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Human Biology*, **71**, 399–415.

Sokal RR, Oden NL (1988) Genetic changes across language boundaries in Europe. *American Journal of Physical Anthropology*, **76**, 337–361.

Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim J, Vaudor A (1989) Genetic differences among language families in Europe. *American Journal of Physical Anthropology*, **79**, 489–502.

Sokal RR, Oden NL, Thomson BA (1999) A problem with synthetic maps. *Human Biology*, **71**, 1–13.

Stenico M, Nigro L, Barbujani G (1998) Mitochondrial lineages in Ladin-speaking communities of the eastern Alps. *Proceedings of the Royal Society of London, Series B — Biological Sciences*, **265**, 555–561.

Vernesi C, Pecchioli E, Caramelli D, Tiedemann R, Randi E, Bertorelle G (2002) The genetic history of natural and reintroduced roe deer (*Capreolus capreolus*) in the Alps and in Central Italy, as inferred from mitochondrial DNA sequences. *Molecular Ecology*, **11**, 1285–1297.

Voronoi MG (1908) Nouvelles applications des paramètres continus à la théorie des formes quadratiques, deuxième mémoire, recherche sur les parallélloedres primitifs. *Journal of Reine Angewandte Mathematik*, **134**, 198–287.

Wiehler J, Tiedemann R (1998) Phylogeography of the European roe deer *Capreolus capreolus* as revealed by sequence analysis of the mitochondrial control region. *Acta Theriologia Supplement*, **5**, 187–197.

Womble WH (1951) Differential systematics. *Science*, **114**, 315–322.

Wright (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Isabelle Dupanloup is a population geneticist mainly interested in genetic data analysis with a focus on human populations. Stefan Schneider is a mathematician with interests in population genetics and computer programming. Laurent Excoffier is a population geneticist with a long-standing interest in estimating population structure from molecular data, in relationship with the settlement history of the populations.
