Genet. Res., Camb. (2006), **88**, *pp.* 27–44. © 2006 Cambridge University Press
doi:10.1017/S0016672306008305 Printed in the United Kingdom

Spatial genetic pattern in the land mollusc *Helix aspersa* inferred from a 'centre-based clustering' procedure

ANNIE GUILLER[1]*, ALAIN BELLIDO[2], ALAIN COUTELLE[3] AND LUC MADEC[4]
[1] *Laboratoire de Parasitologie Pharmaceutique* (*CNRS UMR 6553*), *Faculté des Sciences Pharmaceutiques et Biologiques, 35043 Rennes, France*
[2] *Station biologique de Paimpont* (*CNRS UMR 6553*), *35380 Paimpont, France*
[3] *Département des Sciences de la Terre* (*CNRS UMR 6538*), *29287 Brest, France*
[4] *CNRS UMR 6553, Campus de Beaulieu, 35042 Rennes, France*

(*Received 10 February 2006 and in revised form 17 May 2006*)

Summary

The present work provides the first broad-scale screening of allozymes in the land snail *Helix aspersa*. By using overall information available on the distribution of genetic variation between 102 populations previously investigated, we expect to strengthen our knowledge on the spread of the invasive *aspersa* subspecies in the Western Mediterranean. We propose a new approach based on a centre-based clustering procedure to cluster populations into groups following rules of geographical proximity and genetic similarity. Assuming a stepping-stone model of diffusion, we apply a partitioning algorithm which clusters only populations that are geographically contiguous. The algorithm used, which is actually part of leading methods developed for analysing large microarray datasets, is that of the *k*-means. Its goal is to minimize the within-group variance. The spatial constraint is provided by a list of connections between localities deduced from a Delaunay network. After testing each optimal group for the presence of spatial arrangement in the genetic data, the inferred genetic structure was compared with partitions obtained from other methods published for defining homogeneous groups (i.e. the Monmonier and SAMOVA algorithms). Competing biogeographical scenarios inferred from the *k*-means procedure were then compared and discussed to shed more light on colonization routes taken by the species.

1. Introduction

Taking into account both genealogy and geography to describe gene variation within and among populations has been demonstrated to be of great interest in understanding population genetic dynamics and speciation processes (Templeton, 1998; Hewitt, 2001). However, such time and space dimensions are not always both available, especially when variation is based on empirical data such as allozymes. Notwithstanding restricted information on the existence of alleles through evolutionary time, we focused here on allozyme variation to define the genetic structure of populations of the land mollusc *Helix aspersa*, since it was the only source of genetic information available over all the hundreds of populations sampled. Within the context of the evolutionary history of the species, we propose a new approach based on a clustering procedure specially designed to define groups of populations that are geographically and genetically homogeneous.

The Mediterranean snail *H. aspersa*, characterized by a conspicuous shell polymorphism especially in North Africa, has long been the subject of extensive studies that have led to the recognition of several endemic forms (Taylor, 1913). Whereas the origin of the *maxima* form is enigmatic because its range is unknown, the subspecies *aspersa* has become a typically anthropochorous form widespread throughout the world in many zones that have Mediterranean, temperate and even subtropical climates. In an attempt to trace back the spread of this subspecies in the Western Mediterranean area, we previously investigated hundreds of populations representative of the *aspersa* range (western Mediterranean and European coastline) for conchological, anatomical

* Corresponding author. Tel: +33 223234819. Fax: +33 22323 4540. e-mail: annie.guiller@univ-rennes1.fr

and molecular characters (Guiller, 1994; Guiller *et al*., 1994, 1998, 2001; Madec *et al*., 1996; Madec & Guiller, 1994). However, all these studies have been conducted on a partial set of populations. Moreover, the populations sampled have not all been investigated for the different markers we have successively developed. Thus, only populations coming from the suspected centre of origin from which *H. aspersa* dispersed (i.e. North Africa) have been analysed using part of the mitochondrial large subunit (16S) gene. Whatever the set of populations and/or the markers used, the combination of all different types of data led to a distinct pattern in geographical structure. Within North Africa, results showed a clear West versus East pattern of variation with Lesser Kabylia at the intersection. In this area, data indicate the occurrence of a suture zone due to secondary contact between peripheral populations of the two regions (Guiller *et al*., 1996). This East–West distinction becomes still more apparent when almost all European populations are considered as well (Guiller *et al*., 1994). Indeed, nearly all these European populations clustered with those of western North Africa, with smaller genetic distances than those between western and eastern North Africa. However, the set of European populations that has been analysed jointly with North African samples was restricted since it did not include the 42 French populations still studied (see Section 2 for reasons). Because these populations appeared evolutionarily less interesting since they would have been introduced from the Mediterranean area, they have not yet been integrated in a previous search for knowledge on the history of *H. aspersa*. A comparison of the geographical structure between these populations and the North African ones, meaning between two contrasted zones of the distribution area that form discrete historical entities, has, however, been done to assess the processes that produced the observed spatial distribution (Madec *et al*., 1996). Results showed a real pattern of regional differentiation for African samples while the genetic relationships between French samples, probably influenced by man's activity, are not associated with the spatial position of sampling localities.

The present work provides the first broad-scale screening of allozyme variation in *H. aspersa*. Indeed, all populations from the Western Mediterranean to Northern France that have previously been investigated are treated together, i.e. a total of 2861 individuals from 102 sampling sites. Analysing the whole set of overall populations simultaneously instead of separate historical entities should improve our understanding of *H. aspersa*'s expansion in the Western Mediterranean and provide information on colonization routes used from the centre of origin of the species to the northern part of its distribution area.

In the expectation of an investigation that would yield enough information to provide the geographical and temporal details required for a full phylogeographical analysis, it was obviously tempting to combine European *sensu stricto* and North African allozyme data to investigate more thoroughly the description of the spatial structure of populations. For that, we used clustering methods with spatial contiguity constraint (Ray & Berry, 1966; Legendre & Fortin, 1989). Producing compact geographical groups made up of adjacent sampling sites that should be genetically homogeneous is based on two arguments. First is the hypothesis that dispersal close to a stepwise model has played a significant part in shaping the present-day distribution of the genetic variation. Nevertheless, if genetic structuring were a function only of such kinds of migratory movements, we should expect strong correlation between genetics and geography across the entire range of the species. The use of spatial constraint would then be unjustified. In fact, imposing a geographic constraint is helpful in identifying regions where colonization may be facilitated by human actions. Discrepancies between genetic versus genetic plus geographic patterns may therefore function as indicators of the invasion process. In a historical framework, we would be able to outline different modes of invasion such as a stepwise colonization process in North Africa or a long-distance colonization model in Europe (Müller, 2001). Second, clustering methods free of spatial constraints have been effectively used to analyse partial or overall population versus individual datasets. However, there was no clear evidence of spatial structuring of genetic information whatever the hierarchical level used and whatever the distance-based methods (e.g. FST, Nei's and χ^2 distances) or the model-based methods (e.g. Bayesian clustering approach implemented in STRUCTURE; Pritchard *et al*., 2000) applied. As partial results have outlined, only the clear East–West distinction between lineages together with some local differentiations came out clearly when clusterings are based on genetic component only (Nei's distance). In addition to the fact that the genetic structuring is not always reflected in the geographical proximity of populations or individuals, the great number of individuals/populations investigated tends to complicate analyses and make results ambiguous.

This spatial approach, which tends to discard relationships between geographically distant populations that are close genetically, cannot compensate for the lack of information on temporal variation even if in some cases, as documented in *H. aspersa* by Arnaud & Laval (2004), time is tightly correlated with the spatial dimension. However, we hope that the most meaningful geographical partition obtained will shed more light on the scenario we previously considered to account for the disjunctive distribution of

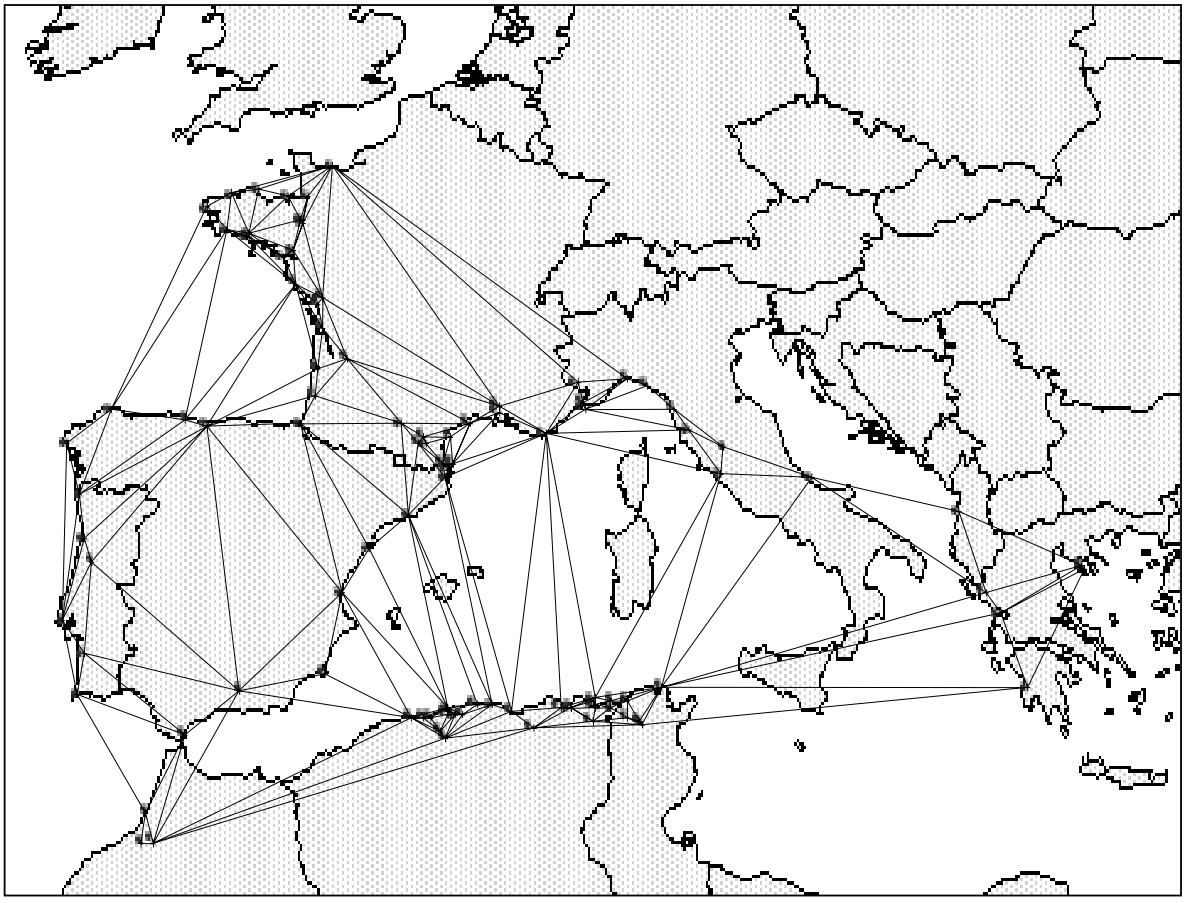

Fig. 1. Location map of the 102 North African and European populations of *Helix aspersa* sampled, with a Delaunay graph (lines) superimposed.

H. aspersa (Guiller *et al.*, 1994). In the process of exploring these scenarios which are based on historical factors (Pliocene geological changes, Quaternary cold periods), we will address substantial questions from the spatial structuring observed: (i) To what extent is the partition retained consistent with the previous scenario inferred from a partial set of data (North Africa and Southern European samples)? (ii) Is genetic variation within or between groups approximately constant from one group to another? If not, how far does extreme divergence fit with previous explanations? (iii) Can the integration of French populations revise the direction of the dispersal? The occurrence of a higher diversity amongst southern regions than northern ones (Guiller, 1994; Guiller *et al.*, 1994) supports observations reported in the literature (Germain, 1908; Taylor, 1913; Chevallier, 1977) that North Africa is the centre of origin of *H. aspersa*. However, according to groups defined and previous haplotype diversity estimated (Guiller *et al.*, 2001), one might imagine an opposite scheme where European immigrants may undergo a strong differentiation after colonizing the North African region.

2. Materials and methods

(i) *Populations and enzymes*

Assuming that the range expansion of introduced populations of *Helix aspersa* occurred from coastline to inland, the 102 populations studied were sampled from 1989 to 1994 on the periphery of the western Mediterranean and the European Atlantic coastline (Fig. 1). North African samples were collected along an East–West transect between Bizerte and Rabat: 5 from Tunisia, 24 from Algeria and 3 from Morocco. European collections were sampled from Portugal (6), Spain (12), France (38), Italy (8), Greece (5) and Albania (1). Each sampling area was less than 5000 m² in area or no longer than 70 m (Madec, 1989). Except for a few collection sites where snails were rare, the sample size varied from 15 to 50 individuals.

Electrophoresis was conducted as previously described (Guiller *et al.*, 1994). Seventeen loci were scored in each individual in all samples except French ones. Among the 17 loci, 14 were polymorphic (*Lap*-1, *Lap*-2, *Aat*-1, *Mdh*-1, *Me*-2, *Pgm*-1, *Pgm*-2, *Mpi*-1, *Pgi*-2, *Sod*-1, *6Pgd*-2, *Idh*-1, *Est*-3, *Est*-6); the other 3

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

loci were *Eno*-1, *Eno*-2 and *Gpd*-2. Because of difficulties with the preservation of protein extracts, only 12 loci (3 of which were monomorphic) could be scored for French samples (8 polymorphic loci: *Lap*-2, *Aat*-1, *Mdh*-1, *Pgm*-2, *Pgi*-2, *Sod*-1, *Est*-3, *Est*-5, *Est*-6; 3 monomorphic ones: *Eno*-1, *Eno*-2 and *Gpd*-2). Consequently, only loci scored in all the 102 populations were considered here. Fifty-five alleles were distinguished over all these 11 loci and populations.

(ii) *Data analysis*

The aim of the analysis was to cluster the 102 samples of *H. aspersa* into groups following rules of geographical proximity and genetic similarity between populations. Allele frequencies computed at various loci were treated as follows.

(a) *Clustering with spatial contiguity constraint*

Allozyme frequencies were converted into measures of genetic distance among populations. According to the partitioning method we used to establish geographical groups that were homogeneous with respect to genetic variation (see below), pairwise genetic distances were converted into euclidean distances by means of a non-metric multidimensional scaling (NMDS) using the most commonly used genetic distance, i.e. Nei's unbiased genetic distance (Nei, 1978). The choice of the Nei/NMDS combination stems from the good performance it provided in summarizing relationships between populations (see Guiller *et al*., 1998 for the iterative procedure of this ordination method). Note that the goodness of fit between original and final distances was obtained here in a 10-dimensional space (stress $S = 0{\cdot}054$).

Homogeneous groups were produced by using a *k*-means algorithm (MacQueen, 1967; Anderberg, 1973) computed using the 'R' package (Ihaka & Gentleman, 1996). This algorithm is based on an iterative procedure of sample reallocation in order to minimize the sum of within-group dispersion, meaning the sum of squares of the euclidean distances to the centroid of the objects of their respective group (for full details of the *k*-means procedure, see Faber, 1994; Moore, 2001; Mackay, 2003; for applications in ecological and spatial analyses, see Legendre & Fortin, 1989; Fortin & Drapeau, 1995; Jacquez *et al*., 2000). The initial configuration of spatially constrained populations we provided to the program was furnished by using the 'Stony Brook method' (see the R package for details). A statistic *D* (which is the sum of within-group sums of squares) is computed for each iteration and the solution that minimizes *D* is kept as the final solution (Späth, 1980).

The spatial contiguity constraint used to perform clustering was represented by a list of connections among close neighbours. Of the published graphical methods for connecting nearest neighbours, we used the Delaunay triangulation method (Green & Sibson, 1977; Ripley, 1981) because it has the advantage of achieving a great number of connections between neighbouring sample locations. Then, the presence of a connection between two samples tells the *k*-means program that these localities are located close to one another and thus may eventually be included in the same group according to their genetic similarity. Without information on the partitioning of the populations, we tested several values of group number for the *k*-means procedure (*k* ranged from 2 to 9), meaning that for each *k* value, steps 6 to 11 were performed (see Fig. 2). A *k*-means algorithm with 1000 iterations was computed for each *k* value.

In order to achieve a 'consensus' of the 1000 optimal configurations obtained, we proposed first to estimate the percentage of association (or probability of co-occurrence) of each pair of populations within the same group (e.g. the percentage of co-occurrence between A01 and A02 was 98·8% since A01 et A02 were found in the same group in 988 of the 1000 final configurations). From the matrix of associations between the $n(n-1)$ pairs of populations, we computed a NMDS in a *p*-dimensional space, with *p* fixed to 3 since pairwise matrix correlations (Mantel tests) between initial and euclidean distances obtained after NMDS were higher with three than with 10 dimensions whatever the number of groups value tested (e.g. for $k = 8$; $r = 0{\cdot}68$ and $r = 0{\cdot}50$ for three- and 10-dimensional space, respectively). Groups were then identified from ordination diagrams using Ward's method (Ward, 1963), which is a hierarchical clustering analysis based on minimizing the loss of between-inertia. To highlight the structure revealed by this clustering approach, successive partitions with group number $g = 1$ to 8 were then represented by convex hulls superimposed on the map of sample locations (Thioulouse *et al*., 1997).

(b) *Reliability of the partition*

Because additional step analyses were required to provide a final group classification using the *k*-means algorithm, the statistic *D* might not be optimal. A test based on the minimization of *D* was then performed to examine the stability of *D* and the robustness of the partition. For each group obtained, we first searched for populations that were geographically connected to within-group populations but that were not included in the corresponding group. Then, we reassigned these 'outgroup' populations within the group tested and estimated *D*. Populations which tended to minimize *D* were then re-allocated in their new group.

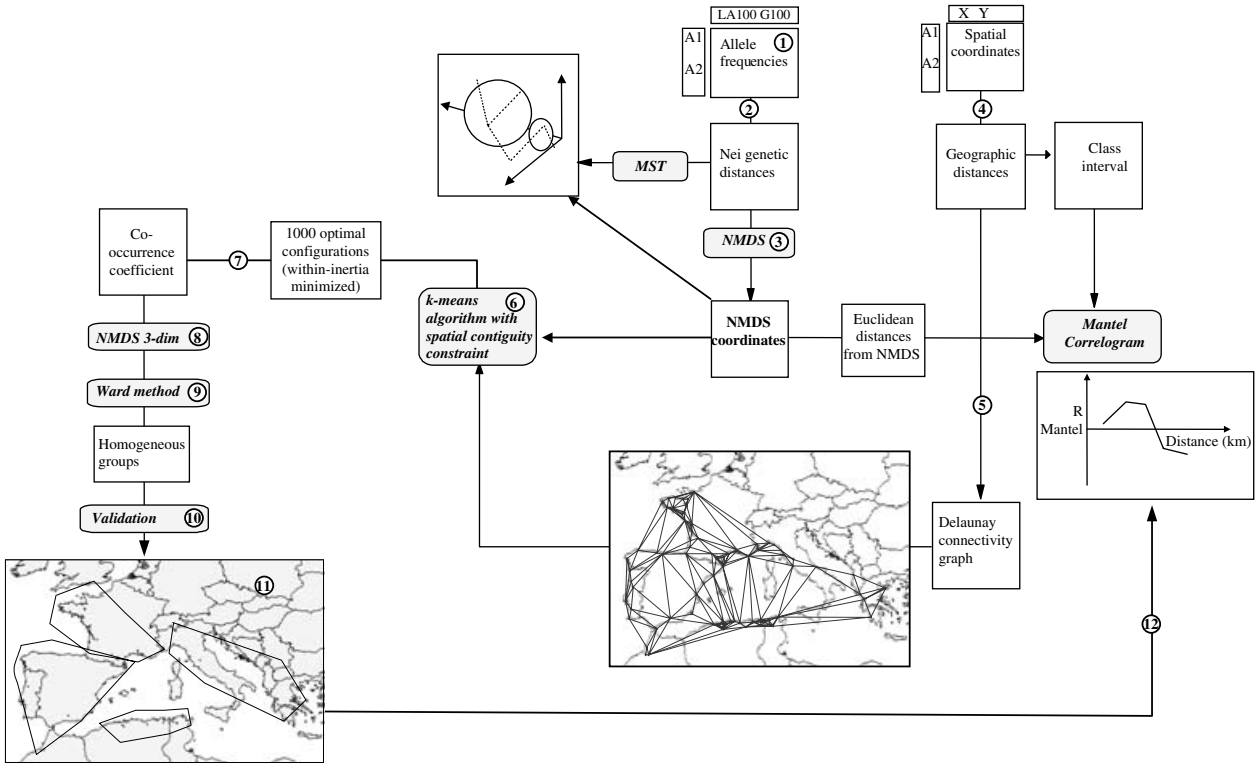

Fig. 2. Data analysis steps from allele frequencies to optimal groups defined: (1) the computation of allele frequencies at various loci; (2) the computation of a matrix of Nei's genetic distances between populations; (3) NMDS on the matrix of Nei's distances in a 10-dimensional space; (4) the computation of a matrix of geographic distances between populations; (5) the computation of a matrix of Delaunay connectivity graph; (6) the clustering of populations using the *k*-means algorithm with spatial contiguity constraint ($k = 2$ to 9, $N = 1000$ iterations); (7) the computation of the matrix of co-occurrence; (8) NMDS on the matrix of co-occurrence in a three-dimensional space; (9) hierarchical clustering of the 102 populations using Ward's method; (10) re-assignation tests to validate the spatial structuring; (11) the description of the final homogeneous groups using inertia criteria; (12) the search for the presence of spatial arrangement in the genetic data by computing correlations (or autocorrelations) between genetic and geographic distances (or classes of geographic distances).

(c) *Description of the final* g*-partition*

The final groups modified according to the re-allocation test were characterized using two statistics based upon inertia. One statistic accounts for the genetic variation within a group (within-inertia) whereas the other accounts for the degree of between-inertia. Observed values of inertia obtained using coordinates of populations obtained by ordination (NMDS on Nei's distance) were then compared with expected values obtained under the null-hypothesis of no genetic relationships between populations spatially constrained. To perform these randomization tests, a group of spatially connected populations of size P was drawn 1000 times from the dataset of 102 populations. For each of these 1000 replicated groups, we estimated its within-inertia versus its between-inertia. Then, the probability of obtaining a random group of size P having a within-inertia (between-inertia) smaller (greater) than or equal to the observed group of the same size was recorded.

(d) *Spatial structuring of genetic variation*

Within-inertia values provide no information on the spatial organization of the genetic variation. To test for the presence of spatial arrangement in the genetic data, we computed a Mantel correlogram for the whole set of populations and for each group defined (Legendre & Fortin, 1989). The multivariate correlograms were drawn by calculating rz between the Nei's genetic distances and binary matrices built for each class of geographical distances (Oden & Sokal, 1986). Two different measures of rz were computed: (i) rz_{gen} based on euclidean distances obtained from NMDS on Nei distances), (ii) rz_{geo} based on genetic distances between populations geographically related (euclidean distances obtained from NMDS on the matrix of co-occurrence in a three-dimensional space). In other words, the correlogram based on rz_{gen} mirrored the real geographical pattern of genetic variation between populations, whilst the rz_{geo} profile expressed the spatial organization of populations expected

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

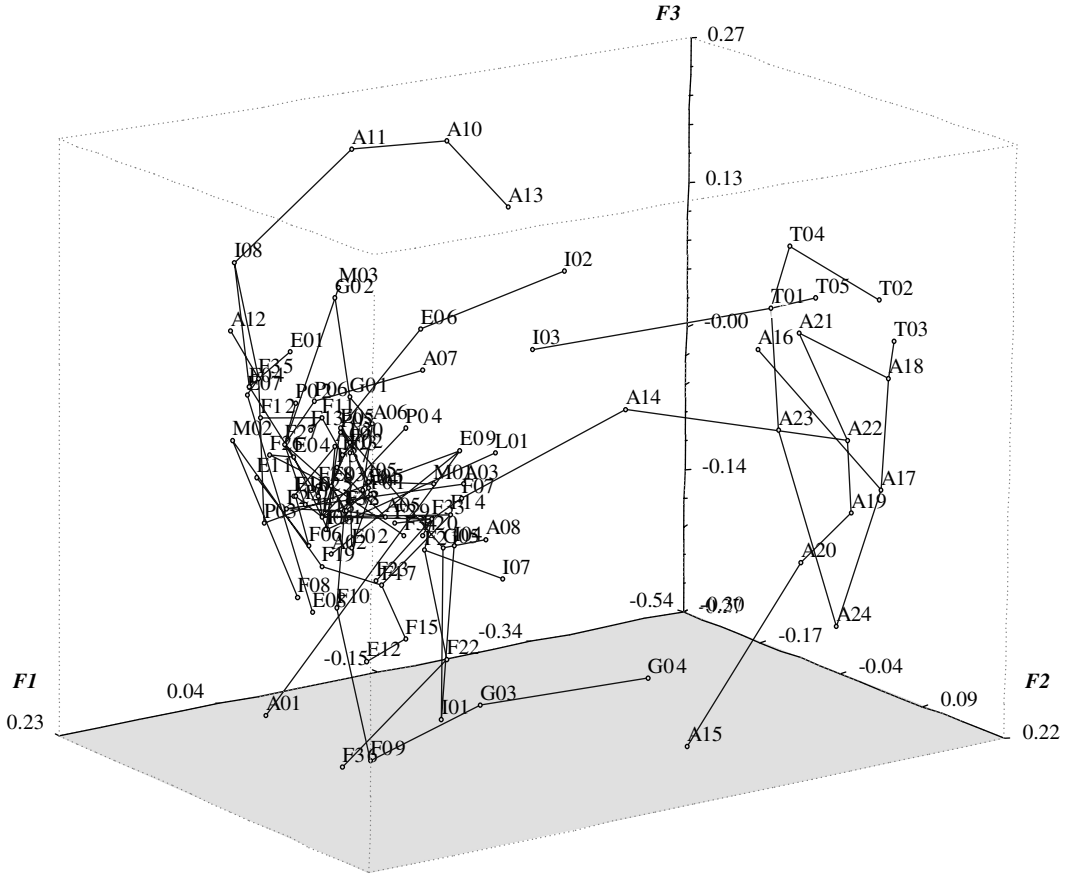

Fig. 3. Three-dimensional ordering graphic of NMDS on Nei's genetic distances between the 102 samples free of spatial constraint. Edges between populations represent a minimum spanning tree (MST) computed from Nei's distances.

under a model optimizing strong association between pairwise genetic and geographical distances. The overall significance of correlograms was assessed using the Bonferroni technique (Oden, 1984).

A summary of the different steps of the whole data analysis is given in Fig. 2. (A table of allele frequency distributions per locus and matrix of pairwise F_{ST} values are available on the journal's web site.)

3. Results

(i) *Clustering without spatial contiguity constraint*

A three-dimensional ordering graphic with NMDS on Nei's genetic distances between the 102 samples of *H. aspersa* is presented Fig. 3. To help detect local distortions due to ordination (i.e. two objects initially close can be distant after ordination), a minimum spanning tree (MST) computed from Nei's distances was superimposed upon the ordination scatter. Populations separated either side of A14 (the Kabylian population of Djemila in Algeria) into two unequal-sized groups that correspond to two well-differentiated geographical regions, namely Eastern North Africa (ENAf, from Lesser Kabylia to Eastern Tunisia) and Europe–Western North Africa (WNAf-Eu). The smaller number of samples in the ENAf group ($n=17$) than in the WNAf-Eu group ($n=85$) made the separation of populations clearer in the eastern block than in the western one. However, the length of edges between points implied a greater differentiation between ENAf populations than between WNAf-Eu ones that are more distant geographically. Whereas it was possible to distinguish a Tunisian subcluster in the eastern group, in which the Italian sample I03 was included, the superposition of points and connections between distant populations that are close genetically prevented the detection of potential patterns of regional structuring in the western group.

(ii) k-*means clustering with spatial contiguity constraint*

The emergence of homogeneous groups from $g=1$ to 8 is illustrated by hierarchical clustering and maps with convex hulls in Fig. 4*a* and *b* respectively (results shown only for $k=8$). The content of the hierarchical decomposition from 2 to 10 groups is given in Table 1. Partitions into one to four groups ($g=1$ to 4) were quite similar whatever the k-value tested. A first group separated early with a high between-inertia

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

value; it clustered the East-North African populations and one trans-Mediterranean sample – I02 and not I03 as previously found. The decrease in inertia between groups was much lower for the following levels of clustering. The partition into three and four groups cleared populations from the Western part of the sampling area, the third group (WNAf-WI) including Western North African and some Iberian populations, and the fourth (NWAt-NI) clustering samples from the North-West Atlantic coastline (i.e. French and North Iberian populations). In contrast, partitions into five groups and more ($g \geqslant 5$) could differ when k ranges from 5 to 9. However, three groups remained unmodified whatever the k-value, especially those including North African samples, i.e. WNAf-EI (West North African and East Iberian samples), WNAf-WI and ENAf. For the two other unstable groups from the five-group partition, configurations depended on the clustering of a set of populations sampled either in the French Mediterranean area (FrMed), or on the Iberian and French coastline (NWAt-NI). The next dichotomy ($g=7$) of the hierarchy tended to homogenize the discrepancies observed between all previous k-configurations since the seven groups produced were quite similar whatever the value of k. That is, only the partition into eight groups showed for the first time a split within North African groups. Indeed, WNAf-WI split up into NAW, which included North African samples together with a Spanish one (E06), and IW mixing North African and Iberian populations. Whilst it was isolated early, ENAf remained unmodified, demonstrating once more the genetic particularity of populations from this geographic zone. The first dichotomy within this group arose only at the seventh node of the hierarchy with the emergence of a subgroup geographically well defined since all populations of ENAf1 are located in the Bônois, whilst samples in ENAf2 form a group more difficult to define spatially since they originated from Algeria, East Tunisia and Italia.

(iii) *Optimal* k- *and* g-*values*

The k-means procedure implies that we initially mapped the 102 populations into k groups. Except for the two large and heterogeneous groups described above when no spatial connections are introduced into the analysis, we had no reliable argument on which to *a priori* impose a value for k, which is why the classification was computed for several k-values ranging from 2 to 9. We considered that k-values >9 would lead to groups too small for further analysis. Optimized groups produced by each k-means classification followed by the Ward hierarchical clustering analysis were then compared by recording the populations that were assigned in the same group whatever the k-value tested. For $g=6$ and $k<8$, 23% of all populations were unstable, and when $k \geqslant 8$, the group composition was nearly constant since only two populations (E06 and F23) moved from one group to another. For $g=7$ and $k>6$, more than 96% of populations belonged to the same groups. Even though clustering in seven groups provided better results, the sizes of some groups were too small to perform Mantel tests and correlograms. Consequently, the following description of population classification has been made using an 8-mean algorithm and a final partition with $g=6$ homogeneous groups (Table 2).

(iv) *Optimization of the statistic* D

Amongst populations not included in their corresponding group, seven (A09, E10, F24, F33, I03, P02 and P06) tended to minimize the sum of inertia values estimated within each of the six groups formed. However, the decrease in D was very weak (5·58 instead of 5·59), demonstrating once again the robustness of the clustering performed. Due to the contiguity constraint, A09 and F33 could not be reassigned in a new group, whereas E10, P02 and P06 joined WNAf-WI, F24 joined NWAt-NI and I03 was reassigned to ItGr (Fig. 5, Table 2).

(v) *Genetic variation among and within groups*

Results of the hierarchical structure implied a heterogeneous distribution of the genetic variation with stronger divergence observed in the North African territory. Because the statistic D alone gave no information on the partitioning of this variation, we computed variances within and between the six optimized groups retained. With randomization tests performed to account for the reliability of the partition, we were able to assign a probability that topology and group composition resulted from a genetic component (Table 2). Then, WNAf-EI and NWAt-NI seemed to be genetically homogeneous since the genetic variance observed for each is much lower than the inertia expected under the null hypothesis ($P=0{\cdot}028$ and 0·017, respectively). For the differentiation between groups, NWAt-NI and especially ENAf dissociated themselves from the other groups with higher between-inertia values than the expected ones ($P=0{\cdot}073$ and 0·001 respectively) (Fig. 6). More than 55% of the total variance accounted for the separation of ENAf.

(vi) *Spatial pattern of genetic variation*

Whether the genetic distance is constrained or not by the geography, there is overall significance in both correlation coefficients and correlograms performed for the whole set of populations (for $rz_{\text{gen}}=0{\cdot}278$, $P<0{\cdot}001$; Bonferroni test $P<0{\cdot}01$; for $rz_{\text{geo}}=0{\cdot}545$,

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

$P<0{\cdot}001$, Bonferroni test $P<0{\cdot}01$; Table 3). Both profiles showed a progressive decline in genetic similarities with increasing geographical distances in classes up to 800 km, there was a transition zone of fluctuating autocorrelation values in classes from 800 km to 1100 km, and a progressive increase in genetic similarities with geographical isolation in classes up to 1700 km (Fig. 7*a*). The overall shape of both types of correlogram per group was suggestive of four different responses (Fig. 7*b–g*): (i) a progressive decrease in rz_{gen} and rz_{geo} underlying a significant genetic differentiation and geographically structured populations for ENAf (Fig. 7*d*), (ii) random fluctuations in rz_{gen} but a quasi-continuous decline in rz_{geo} for WNAf-WI and NWAt-NI (Fig. 7*b*, *e*), (iii) random fluctuations in rz_{gen} and a discontinuous decline in rz_{geo} for WNAf-EI (Fig. 7*c*; a decline up to 100 km and from 650 to 750 km) and for FrMed (Fig. 7*f*; a decline up to 180 km and from 475 to 900 km), (iv) random fluctuations in rz_{geo} and a discontinuous decline in rz_{gen} for ItGr (Fig. 7*g*). The Bonferroni method used to test the Mantel correlograms showed differences among the six groups (Table 3). Only one significant rz_{gen} profile (ENAf, $p<0{\cdot}01$) is observed, meaning that genetic distances are correlated with geography in ENaf only. Notwithstanding the spatial constraint, only four of six rz_{geo} correlograms are significant (WNAf-WI, WNAf-EI, ENAf and NWAt-NI, $p<0{\cdot}01$). It seems that the spatial organization of populations in the FrMed and ItGr groups does not fit at all the population structure and migration model tested.

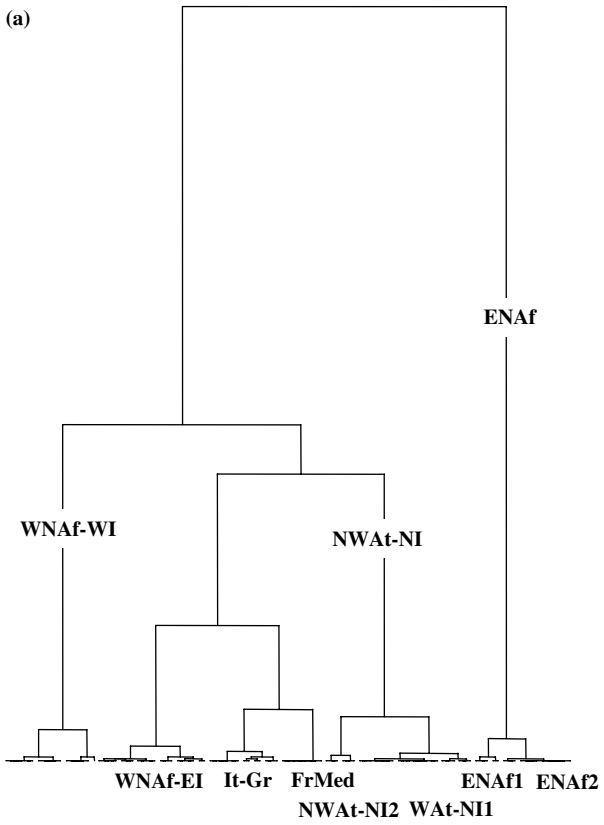

Fig. 4. Partitions from one to eight groups deduced from Ward's hierarchical method after 8-means clustering with spatial contiguity constraint. (*a*) Dendrogram; (*b*) location maps with one to eight convex hulls (ENAf1 and ENAf2 abbreviations correspond to subgroups of ENAf, NWAt-NI1 and NWAt-NI2 abbreviations correspond to subgroups of NWAt-NI; see footnote to Table 1 for other group abbreviations).

4. Discussion

(i) *Description of the optimal partition: group contents and population structure*

Several criteria based on intra-cluster versus inter-cluster measures have been developed for determining cluster validity and allowing k to be fixed automatically (Davies & Bouldin, 1979; Pal & Bezdek, 1995; Ray & Turi, 1999). However, we have preferred to use the stability of populations within their reference group to judge the reliability of k and subsequently to fix the optimal level of g for cutting the hierarchical clustering. The partition we used to adapt and explore the scenario previously proposed is the one corresponding to k and g equal to 8 and 6, respectively.

Of the six groups, only the two easternmost ones, ENAf and ItGr, are totally separated. By contrast, the remaining western groups partially overlap, especially the French ones. The differing level of group isolation could be explained (i) by the various strengths of the spatial constraint according to the genetic background of the group concerned, (ii) by the lack of extensive sampling of neighbouring populations in some areas. That is especially true for the ItGr group, which includes a very small set of populations relative to the great distance covered. Moreover, the genetic differentiation within this group is the highest found despite its reduced size (estimate of Wright's $F_{ST}=0{\cdot}338\pm0{\cdot}140$, *sensu* Weir & Cockerham, 1984), demonstrating that populations of this group have probably been strongly constrained to be clustered together. On the other hand, this peculiar group defined by the Italian and Adriatic populations is sufficiently homogeneous to emerge and be separated from ENAf. For this latter group, its clear and unequivocal separation is not surprising since all previous works based on various markers have already stressed the distinctive genetic feature of snails inhabiting this restricted part of the range. There is also nothing new in the grouping of the Italian and eastern North African populations since a strong affinity between populations located around the Sicilian Canal has already been detected even without applying spatial constraints. What is new is the origin of the Italian population that makes a

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

(b)

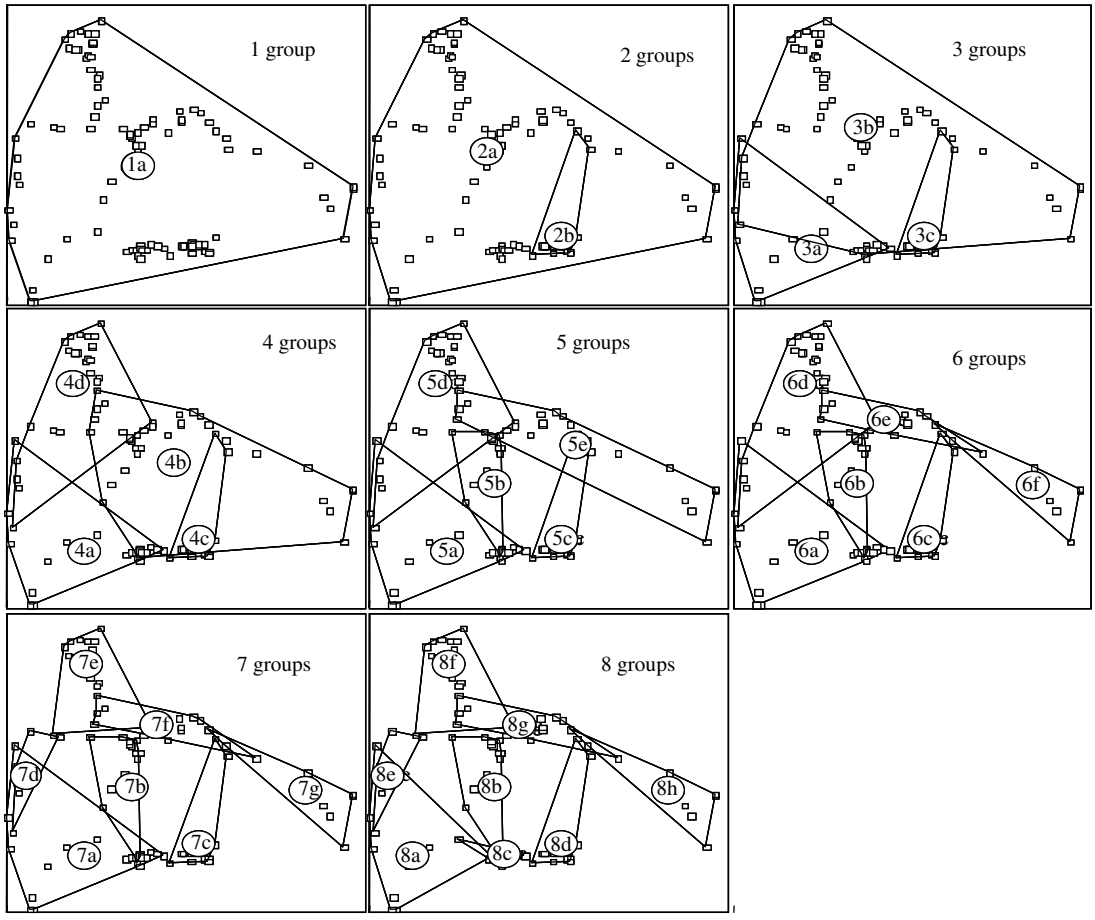

Fig. 4. (Cont.)

connection possible between Italy and Africa. Indeed, I02 substitutes for I03 in joining ENAf when spatial contiguity constraints are taken into account and, consequently, I03 moves to ItGr. The weakness of the relationships between the two trans-Mediterranean areas is due to the higher genetic similarities between I03 and the Tunisian samples. Whilst both I03 and I02 are equally distant from ENAf ($D_{Nei}=0{\cdot}138$ and $0{\cdot}129$, respectively), I03 is closer than I02 to Tunisian populations ($D_{Nei}=0{\cdot}083$ and $0{\cdot}127$, respectively). Whatever the origin of the Italian population involved in the Tyrrhenian connection, it is tempting to think that those populations may represent turning points between distinct eastern routes taken by the species (see below).

Assuming a stepping-stone model of diffusion, we have tried to determine how drift and historical gene flow have influenced the population structure of each group detected (Hutchison & Templeton, 1999). Evaluations of the association between genetic (F_{ST}) and geographical distances estimated between pairwise populations reveal equilibrium in one set of populations only, i.e. ENaf (when I02 is removed). The changing relative influences of gene flow and drift as populations in this area become more spatially distant, produce effectively a pattern indicative of isolation by distance (IBD). In all remaining groups, the patterns implied that drift has been of relatively greater importance in the western than in the eastern Mediterranean area. The low dispersal capacity of the species could explain this pattern, in which allele frequencies fluctuate independently from geography. However, besides the large variance in estimates of divergence, one might also expect quite a high level of divergence between populations. That is the case for the western group WNAf-WI (mean pairwise $F_{ST}=0{\cdot}273\pm0{\cdot}122$) but much smaller divergence occurred between populations of the three central groups of the Mediterranean area, i.e. WNAf-EI, NWAt-NI and FrMed (mean pairwise F_{ST}, is $0{\cdot}123$, $0{\cdot}145$ and $0{\cdot}154$, respectively). The lack of significant association between genetical variation and geography may also indicate that snails have not occupied these regions long enough to have approached any pattern resembling an IBD. Since the correlograms show IBD for three of these four groups (WNAf-WI, WNAf-EI and NWAt-NI), it is possible that populations of this part of the Mediterranean

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

Table 1. *Geographical subdivisions for various cutoff levels of Ward hierarchy for* k = *8* (g *ranges from 4 to 10*)

WNAf-WI		WNAf-EI + FrMed + It-Gr				NWAt-NI		ENAf	
WNAf-WI		WNAf-EI		FrMed + It-Gr		NWAt-NI		ENAf	
WNAf-WI		**WNAf-EI**		**FrMed**	**It-Gr**	**NWAt-NI**		**ENAf**	
WNAf-WI-1	WNAf-WI-2	WNAf-EI-S	WNAf-EI-N	FrMed	It-Gr	NWAt-NI1	NWAt-NI2	ENAf1	ENAf2
G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
A09	A01	A03	E02	F20	G01	E11	E12	A14	A15
A10	A02	A04	E04	F21	G02	F01	P04	I02	A16
A11	E07	A05	F26	F22	G03	F02	P05	T04	A17
A12	E08	A06	F28	F23	G04	F03		T05	A18
A13	E09	A07	F29	F31	G05	F04			A19
E06	E10	A08	F30	F33	I01	F05			A20
	M01	E01	F37	F34	I03	F06			A21
	M02	E03	F38	F35	I04	F07			A22
	M03	E05		F36	I08	F08			A23
	P01	F25		I05	L01	F09			A24
	P03	F27		I06		F10			T01
	P02			I07		F11			T02
	P06					F12			T03
						F13			
						F14			
						F15			
						F16			
						F17			
						F18			
						F19			
						F24			
						F32			

Abbreviations (in bold) correspond to the six homogeneous groups defined: *WNAf-WI*, West-North Africa + West Iberia; *WNAf-EI*, West-North Africa + East Iberia; *FrMed*, French Mediterranean; *It-Gr*, Italia + Greece; *NWAt-NI*, North-West Atlantic + North Iberia; *ENAf*, East-North Africa.

Table 2. *Characteristics of the six final homogeneous groups defined: group size, sample composition, inertia*

Group name	Group size (P)	Group composition	Within-inertia		Between-inertia	
			Obs.	*P*	Obs.	*P*
WNAf-WI	19	A01 A02 A09 A10 A11 A12 A13 E06 E07 E08 E09 E10 M01 M02 M03 P01 P02 P03 P06	1·507	0·597	0·582	0·137
WNAf-EI	19	A03 A04 A05 A06 A07 A08 E01 E02 E03 E04 E05 F25 F26 F27 F28 F29 F30 F37 F38	0·626	0·028	0·319	0·331
ENAf	17	A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 I02 T01 T02 T03 T04 T05	1·297	0·575	2·564	0·001
NWAt-NI	25	E11 E12 F01 F02 F03 F04 F05 F06 F07 F08 F09 F10 F11 F12 F13 F14 F15 F16 F17 F18 F19 F23 F24 F32 F33 P04 P05	0·870	0·017	0·776	0·073
FrMed	12	F20 F21 F22 F23 F31 F33 F34 F35 F36 I05 I06 I07	0·398	0·133	0·194	0·520
It-Gr	10	G01 G02 G03 G04 G05 I01 I03 I04 I08 L01	0·880	0·727	0·186	0·474
Total inertia			5·578		4·621	

Obs., within- or between-inertia observed; *P*, probability (1000 permutations) of obtaining a random group of size P having a within-inertia (between-inertia) smaller (greater) or equal to the observed group of the same size.
See footnote to Table 1 for group abbreviations.

basin belong to different waves of dispersion. The populations that first colonized these areas would be spatially structured but equilibrium cannot be detected because of more recent invaders from different source populations. In the light of pairwise F_{ST} values averaged over groups WNAf-EI,

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

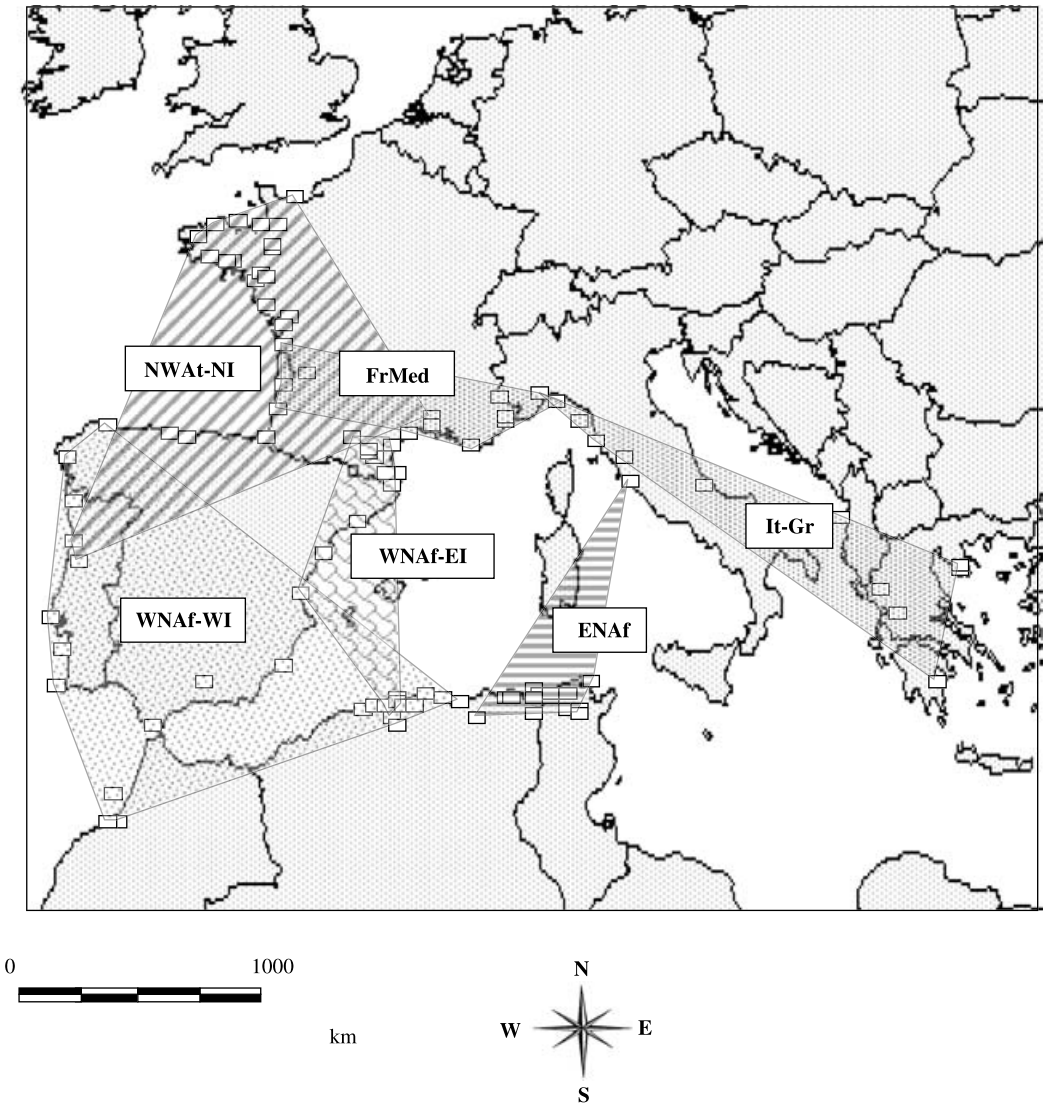

Fig. 5. Partition into six homogeneous areas after minimization of within-inertia of groups obtained from the *k*-means procedure (see footnote to Table 1 for group abbreviations).

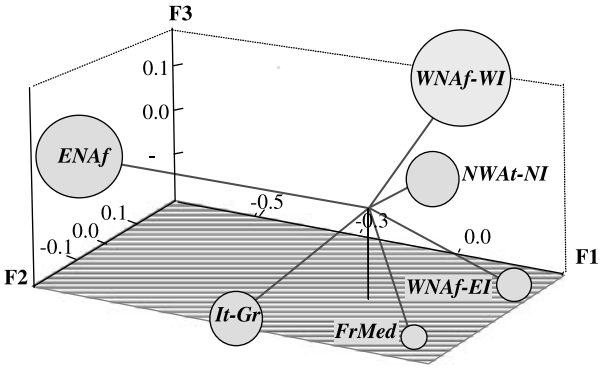

Fig. 6. Graphical representation of within- and between- inertia associated with each homogeneous group. Each sphere, the diameter of which is proportional to within-inertia, is centred on the group barycentre. The length of the unbroken line joining each sphere to the general barycentre is proportional to between-inertia. (See footnote to Table 1 for group abbreviations.)

NWAt-NI and FrMed (F_{ST} is only 0·156 $\pm$0·079), native populations of recurrent invaders would have originated from the same region in this area, whereas snails that dispersed around the Strait of Gibraltar (WNAf-WI) would have originated from more highly differentiated sources, probably Europe and North Africa. Moreover, this kind of within-group genetic mixture would surely play a major role in explaining the strong spatial constraint in defining European groups. Indeed, the inferred six-group structure obtained showed that a large proportion of populations (31%) were not assigned correctly. However, the number of misclassified samples is more than twice as high in Europe (38·5%) as in North Africa (15·6%).

Preliminary tests we concurrently performed to assign all 2861 individuals to *k* virtual populations using the Bayesian algorithm implemented in the

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

Table 3. *Mantel tests performed on initial and constrained distances*

		Initial Nei distances			Constrained distances		
Group	Pairs	Global *rz*	*P*	Bonferroni test	Global *rz*	*P*	Bonferroni test
WNAf-WI	171	0·009	0·435	NS	0·550	<0·001	*P*<0·01
WNAf-EI	171	0·199	0·019	NS	0·442	<0·001	*P*<0·01
ENAf	136	0·387	0·058	*P*<0·01	0·766	<0·001	*P*<0·01
NWAt-NI	300	0·020	0·435	NS	0·651	<0·001	*P*<0·01
FrMed	66	−0·044	0·399	NS	0·533	0·001	NS
It-Gr	45	0·270	0·057	NS	0·267	0·029	NS
102 pops.	5151	0·278	<0·001	*P*<0·01	0·545	<0·001	*P*<0·01

Initial Nei distances, euclidean distances after NMDS with 10 dimensions on Nei distances matrix; Constrained distances, euclidean distances after NMDS with three dimensions on pairwise association coefficients calculated on 1000 *k*-means iterations; global *rz*, Mantel test on genetic distances and geographical distances via Delaunay graph; Bonferroni test performed on Mantel correlograms.
See footnote to Table 1 for group abbreviations.

software STRUCTURE (Pritchard *et al.*, 2000; Evanno *et al.*, 2005) confirmed the greatest admixture of European populations. Indeed, only 20% of European samples, compared with 56·3% of North African ones, have more than 75% of individuals probabilistically assigned in the same cluster. Considering a lower hierarchical level than population to identify homogeneous groups shows that individuals of the same population might originate from different source populations. Such an assumption, strengthened by the extremely high levels of mitochondrial diversity so far recorded in some European populations, supports a North African rather than an European origin of the species (see above for details).

(ii) *The choice of a* k*-means clustering approach*

Numerous univariate and multivariate analysis methods are available to depict discontinuities in the spatial distribution of gene frequencies (Barbujani *et al.*, 1989; Legendre & Fortin, 1989; Barbujani & Sokal, 1990; see Manel *et al.*, 2003 for a review). However, few of these techniques integrate directly the contribution of the geographical component in defining genetically homogeneous groups of populations. Genetic variation is more often considered subsequent to geography (i) to test for a spatial structure (autocorrelation, Mantel test and Mantel correlogram), (ii) to detect zones of sharp genetic discontinuities or genetic barriers such as wombling (Womble, 1951) and tessellation used for example in Monmonier's algorithm (Monmonier, 1973), or (iii) to give insight into a specific spatial pattern by means of interpolated maps (interpolation such as kriging mapping, trend surface analysis, ordination). Clustering methods with spatial constraint such as *k*-means may describe spatial structure but, as a by-product, they are also a way of producing maps of multivariate data. Unlike other techniques mentioned above, the *k*-means clustering includes space as 'predictor' and the only approach that could compare with ours is the SAMOVA method recently developed by Dupanloup *et al.* (2002) (see below for comparison). To be efficient and find good-quality partitions, the use of our clustering approach should, however, imply beforehand that populations may be arranged in a clustered fashion. All previous studies we have done on the evolutionary history of *H. aspersa* verify such an assumption of population grouping, in showing for instance the existence of two main geographically differentiated entities (Guiller *et al.*, 1994, 1996, 1998, 2001). That is the reason we initially opted with confidence for clustering methods.

Another point that can be questioned is the choice of a *k*-means procedure rather than other clustering methods. Whilst this algorithm is popular in different fields requiring data exploration and data compression (e.g. genome biology to analyse microarray data), it has several *ad hoc* features and can behave badly (Falkenauer & Marchand, 2003; Merz, 2003). The main reason behind the use of this centre-based clustering algorithm was the possibility of spatially constraining the algorithm to produce groups.

(iii) *The standard* k*-means algorithm: popular but open to criticism*

Before discussing the performance of the *k*-means procedure and comparing the inferred genetic structure obtained with other methods for defining homogeneous groups, i.e. the Monmonier and SAMOVA algorithms, we will report its limitations and cases where it might be viewed as failing.

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

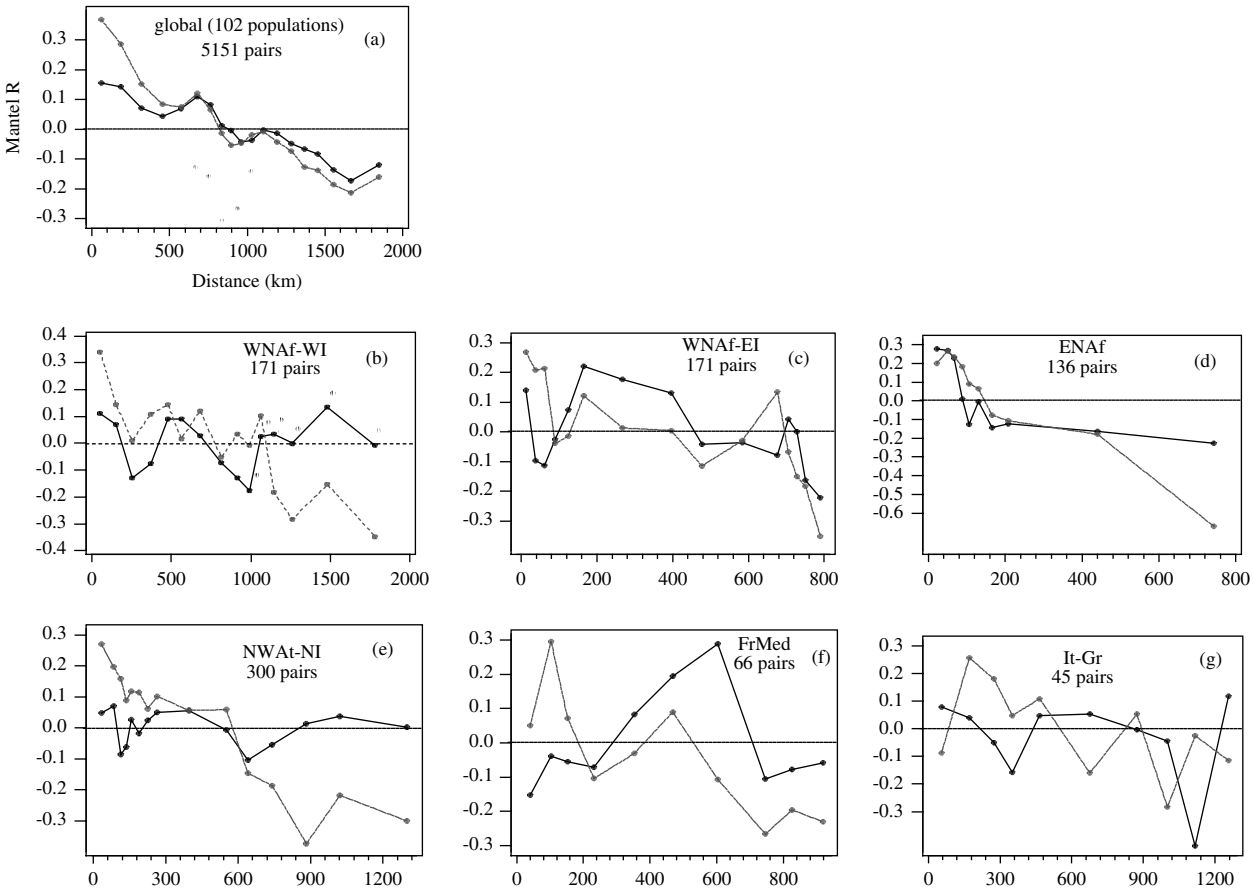

Fig. 7. The global (*a*) and intra-group (*b–g*) spatial pattern of genetic variation illustrated by Mantel correlograms: rz_{gen} (unbroken line) is based on genetic distance between populations (euclidean distances obtained from NMDS on Nei distances); rz_{geo} (dotted line) is based on genetic distance between populations geographically related (euclidean distances obtained from NMDS on the matrix of co-occurrence in a three-dimensional space). (See footnote to Table 1 for group abbreviations.)

As for other clustering methods of the same family, the goal of the *k*-means algorithm is to minimize its objective function, i.e. the squared distance between each centre and its assigned points. The computation starts with an initialization of the centre positions and is followed by iterative refinement of these positions until a local optimum corresponding to convergence (stable assignments) is reached. The main problem of the standard *k*-means algorithm, or to be precise 'NP-hard problems', is that it can easily converge to a local minimum that is far from the global minimum (*D*), then providing solutions that are only locally optimal. One of the reasons for convergence to poor solutions is its hard membership function, which defines the proportion of points *xi* that belong to centre *cj* with constraint *m* ($0 \leqslant m \leqslant 1$). Because *m* is strictly 0 or 1 for *k*-means, all points belonging to the same cluster have an equal vote (vote 1) and no vote (vote 0) in any other clusters, even ambiguous points located near the border between two or more clusters. This algorithm takes account only of the distance between the means and the data points and has no way of representing the weight, size or shape of each cluster. Alternative methods softer than the *k*-means algorithm also exist and make assignments using more information about the data in the refitting step of the algorithm. Comparison of the performance functions of several of these variants methods (e.g. GEM: Gaussian expectation-maximization; FKM: fuzzy *k*-means; KHM: *k*-harmonic means and KHM-derived algorithms) showed the superiority of KHM and FKM for finding clusterings of high quality (Zhang *et al.*, 1999; Hamerly & Elkan, 2002). These clusterings would be effectively less sensitive to the initialization step. In the absence of experience with both these methods, the strategy we used to prevent *k*-means getting stuck at local minima was (i) to run the algorithm several times from different starting points (1000 iterations), (ii) not to simply select the best solution (configuration with smaller *D*) to define groups, but take into account pairwise relationships between populations over all configurations. Whilst the 'R' package makes the former strategy feasible, the latter is part of the procedure we proposed.

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

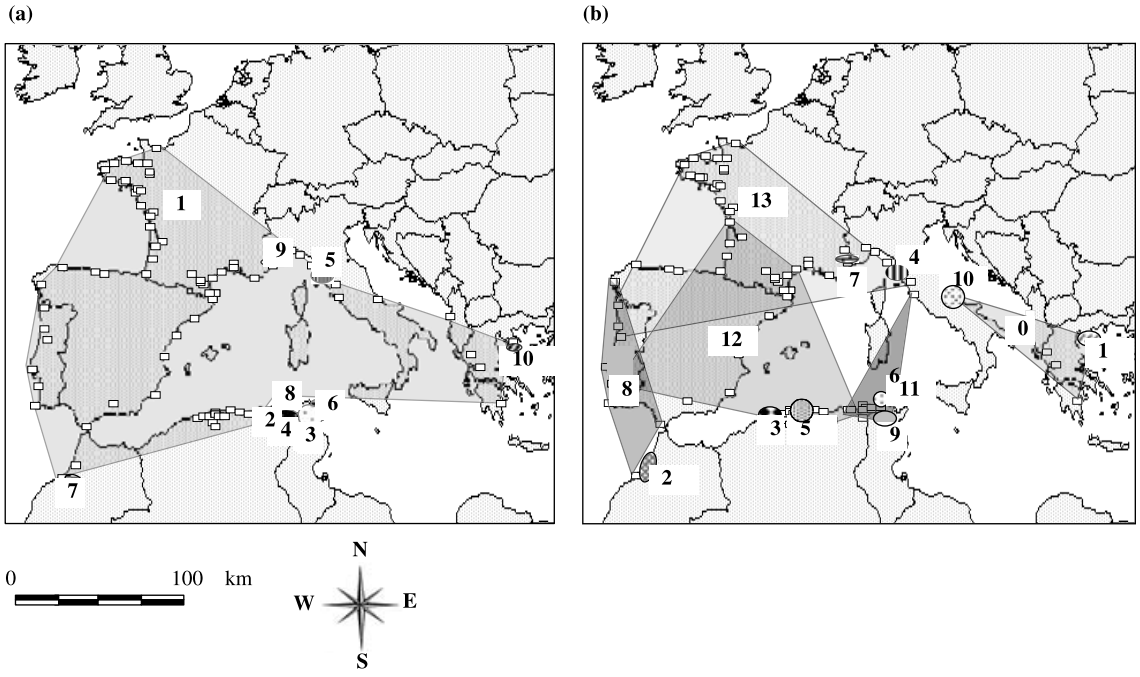

Fig. 8. Partitions into homogeneous areas defined directly by SAMOVA (*a*) or subsequent to genetic barriers by the Monmonier algorithm (*b*) (for optimal comparison, *k* is 10 and 13 for SAMOVA and the Monmonier algorithm, respectively).

(iv) *Inferences from the SAMOVA and Monmonier algorithms*

The Monmonier algorithm (Monmonier, 1973) is designed for visualizing potential genetic barriers between groups, whereas the aim of spatial analysis of molecular variance (SAMOVA), developed by Dupanloup *et al.* (2002), is to maximize the proportion of total genetic variance due to differences between groups (F_{CT} index). The spatial partitioning inferred from these methods leads to quite different groups of populations (Fig. 8). A major difference concerns the group size. For an equivalent six-group structure, the number of populations per group fluctuates between 10 and 25 for *k*-means, but ranges from 1 up to 87 for SAMOVA and 1 up to 91 for the Monmonier algorithm. The occurrence of 'singletons' or a reduced group size with SAMOVA and the Monmonier algorithm showed that even if *k*-means partitioning is one of the so-called NP-hard problems, it seems to be less sensitive to local minima than are the other two methods. All small groups defined might not imply false minima. As a result of differences in group size, the variance components calculated at different levels of population hierarchy differed between methods but, as expected, the variance component among groups and F_{CT} for SAMOVA were found to be nearly 2 and 3 as high as for *k*-means and the Monmonier algorithm, respectively (Table 4). Despite large differences in group size between SAMOVA and *k*-means, 'within-group variation' is equivalent for the two methods (18%). The two first genetic barriers that appear with the Monmonier algorithm isolate six samples that are part of the ItGr group, whilst the eastern North African samples were separated rather later (barrier 6). Apart from singletons or small group size that reflects reduced zones of sharp genetic changes, the partition of populations in the western part of Mediterranean mirrors that by *k*-means. In contrast, SAMOVA gives a quite different configuration. It confirms the genetic peculiarities of the eastern North African region, since for a six-group structure, four groups emerge from this small area. However, partitioning of this region still continues when *k* increases. This continuous fragmentation in the most differentiated group is not surprising since it shows quite a high level of genetic variation. However, neither the Monmonier nor *k*-means procedures lead to this finding. Such discrepancies could be attributed to population structure and pattern of migration with two arguments: (i) the sensitivity of SAMOVA to pattern of isolation by distance (IBD) (Dupanloup *et al.*, 2002), (ii) the presence of IBD only detected in the ENAf group. The effective number of migrants ($N_m < 1$) would not be sufficiently high for SAMOVA to perform well. Simulations showed that the undesirable effect of IBD was reduced only by increasing the amount of N_m within the group ($N_m >> 1$) (Dupanloup *et al.*, 2002).

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

Table 4. *Variance components and F-statistics analogues at three levels of hierarchical analyses (AMOVA) on a six-group structure defined by the k-means procedure (K), SAMOVA (S) or the Monmonier algorithm (M)*

Source of variation	d.f.		Variance components	Percentage of variation	Fixation indices		*P*
Among groups (s^2a)	5	S	0·550	27·5	FCT	0·275	<0·0001
		M	0·157	9·2		0·092	<0·0001
		K	0·238	14·7		0·147	<0·0001
Among populations within groups (s^2b)	96	S	0·357	17·9	FSC	0·246	<0·0001
		M	0·456	26·7		0·294	<0·0001
		K	0·292	18·0		0·211	<0·0001
Within populations (s^2c)	5620	S	1·093	54·7	FST	0·453	<0·0001
		M	–	64·1		0·327	<0·0001
		K	–	67·3		0·359	<0·0001

P, the probability of having a higher variance component and F-statistic than the observed values by chance alone; test based on permutations.

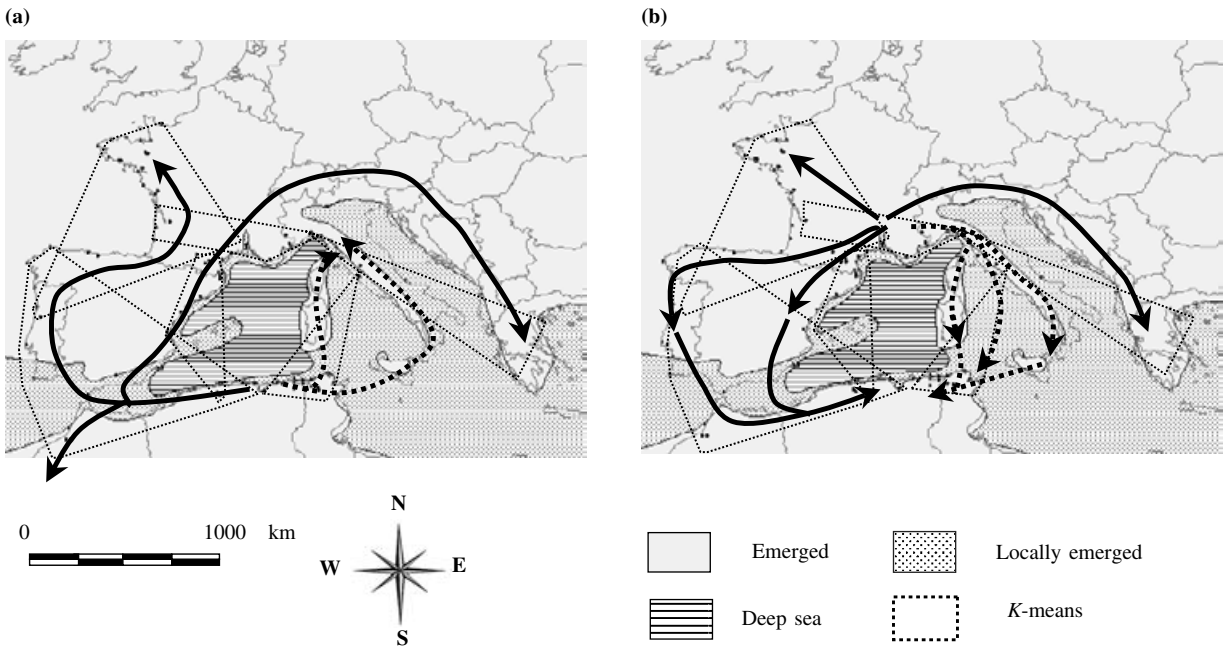

Fig. 9. Colonization routes following two scenarios involving (*a*) migrations from North Africa to Europe, (*b*) secondary contacts in the Kabylian area of populations originating from Western Europe (western and eastern lineages of *H. aspersa* are represented by unbroken and dotted arrows, respectively).

(v) *Inferring the evolutionary history of the species: new insights from present results*

Previous inferences based on allozyme, genital and shell affinities between Italian (stations in Latium, Tuscany, Calabria and Sicily) and East African populations presumed that Italy would have become secondarily a contact zone between the southern Italian–Tunisian populations and northern ones. On the plate tectonics assumption we proposed, such contact would arise after several colonization steps from North Africa to Europe (Guiller *et al*., 1994; Madec & Guiller, 1994; Madec *et al*., 2003). After the disjunction and drift of continental microplates in the western Mediterranean since the Lower Miocene, snails would have first progressed from the Kabylia plates to Tunisia via the Tellian Atlas Chain (a massif arising from the collision of the Kabylia blocks with north-west Africa) (Giusti & Manganelli, 1984; Dercourt *et al*., 1985, 1993). The eastern edge of the Tellian Atlas Chain (Babor Mountains, Constantine Mountains and Numidian Mountains) could have prevented and still continue to prevent gene flow between eastern and western populations. Then, from this intermediate area, a second later dispersal of snails would have allowed colonization of Sicily and Calabria. This would have occurred before the split of the Tellian massif at the position of the Sicilian

https://doi.org/10.1017/S0016672306008305 Published online by Cambridge University Press

channel during the Pliocene (Sacchi, 1958). According to this scenario implying that the species would originate from North Africa, the geographic and genetic isolation of eastern African populations would be the signature of geological events and consequently historical events that occurred during the Pliocene and later. During the Pleistocene glacial/interglacial cycles, populations would also have experienced large founder effects (due to the low dispersal capability of the species) at any time when more favourable climatic conditions allowed animals to spread from refugia located in the Kabylias and in the High Steppe area of Tunisia (Flint, 1971 and Brown & Gibson, 1983 in Blondel, 1986).

The distinct cleavage between ENAf and ItGr and the unstable relationships of some Italian populations could, however, argue against this scenario. Indeed, they could suggest that populations would have progressed not from South to North but in an opposite direction from Western Europe to south-eastern areas either to North Africa via a Tyrrhenian route or to Adriatic/Ionian regions via an Aegean route. For the western margin of the Mediterranean, dispersal would have occurred via the Strait of Gibraltar and around before the Neogene extension of the Alboran basin (Calvert *et al*., 2000) (Fig. 9*a*). Part of the phylogeographic results related to North African samples is consistent with this opposite direction of dispersal (Guiller *et al*., 2001). The mtDNA lineage corresponding to ENAf (without Italian samples) is effectively characterized by a nucleotide diversity twice as low as in the western lineage, and more unresolved relationships between populations. Whilst we ascribed the lower diversity in the East to the occurrence of transitory and prolonged bottlenecks inherent to successive migrations that populations experienced from Kabylia to Tunisia, this may also simply indicate a more recent divergence within the eastern region. However, in the expectation of future arguments that would strengthen the idea of an European origin of the species, it is rather tempting to favour a North African origin since all other findings would strongly support this view (Fig. 9*b*). First is the occurrence of a greater genetic diversity in North Africa than in Europe, illustrated by the emergence of three distinct groups in Africa, and contrasted with smaller genetic divergence between populations of the three central groups of the Mediterranean area. Substantial differences in population structure with the lack of drift–migration equilibrium in all groups except ENAf are also suggestive of a more recent establishment of European populations without ruling out the possibility of multiple introductions due to man's activity as early as the Neolithic revolution. The chronology of diversification events within the species estimated by assessing divergence times between lineages also fits better with a South–North colonization. Assuming that the eastern *aspersa* form in North Africa would have started to diverge at the end of the Tertiary (Guiller *et al*., 2001), European populations would have occurred earlier in the case of a dispersion from Europe. Pliocene shell fossils reported in the literature show that *H. aspersa* was already quite widespread in the Tertiary but only in circum-Mediterranean sites (Oran, Sicily, Nice; Taylor, 1913), not in northern parts of Europe. Preliminary phylogeographical inferences based on relationships between European and North African haplotypes would also support an African range expansion of the species (unpublished data).

We thank G. Evanno and P. Jarne for their advice and valuable criticisms on earlier versions of the manuscript, and P. Legendre, A. Bar-Hen and three anonymous referees for critically reviewing the manuscript.

References

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Arnaud, J.-F. & Laval, G. (2004). Stability of genetic structure and effective population size inferred from temporal changes of microsatellite DNA polymorphisms in the land snail *Helix aspersa* (Gastropoda: Helicidae). *Biological Journal of the Linnean Society* **82**, 89–102.

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

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

Blondel, J. (1986). *Biogéographie évolutive*. Paris: Masson.

Calvert, A., Sandvol, E., Seber, D., *et al*. (2000). Geodynamic evolution of the lithosphere and upper mantle beneath the Alboran region of the Western Mediterranean: constraints from travel time tomography. *Journal of Geophysical Research* **105**, 10871–10898.

Chevallier, H. (1977). La variabilité de l'escargot petit-gris *Helix aspersa* Müller. *Bulletin du Muséum National d'Histoire Naturelle, Paris*, 3rd Series **448**, 425–442.

Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**, 224–227.

Dercourt, J., Zonenshain, L. P., Ricou, L. E., *et al*. (1985). Geological evolution of the Tethys belt from the Atlantics to the Pamirs since the Lias. *Tectonophysics* **123**, 241–315.

Dercourt, J., Ricou, L. E. & Vrielinck, B. (1993). *Atlas Tethys: Paleoenvironnemental Maps*. Paris: Gauthier-Villars.

Dupanloup, I., Schneider, S. & Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**, 2571–2581.

Evanno, G., Regnaut, S. & Goudet, J. (2005). Detecting the number of clusters of individuals using the software Structure: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Faber, V. (1994). Clustering and the continuous *k*-means algorithm. *Los Alamos Sciences* **22**, 138–144.

Falkenauer, E. & Marchand, A. (2003). Clustering microarray data with evolutionary algorithms. In *Evolutionary Computation in Bioinformatics* (ed. G. B. Fogel &

D. W. Crone), pp. 219–230. San Francisco: Morgan Kauffman.

Fortin, M.-J. & Drapeau, P. (1995). Delineation of ecological boundaries: comparison of approaches and significance tests. *Oikos* **72**, 323–332.

Germain, L. (1908). Etude sur les Mollusques recueillis par M. Henri Gadeau de Kerville pendant son voyage en Khroumirie (Tunisie). In *Voyage zoologique en Khroumirie* (ed. Gadeau de Kerville), pp. 129–206. Paris: Baillière.

Giusti, F. & Manganelli, G. (1984). Relationships between geological land evolution and present distribution of terrestrial gastropods in the western Mediterranean area. In *World-wide Snails* (ed. A. Solem & A. C. Van Bruggen), pp. 70–92. Leiden: E. J. Brill.

Green, P. J. & Sibson, R. (1977). Computing Dirichlet tessellations in the plane. *The Computer Journal* **21**, 70–92.

Guiller, A. (1994). Aspect géographique de la différenciation génétique des populations de l'escargot terrestre *Helix aspersa* Müller (Gastéropode, Pulmoné). PhD thesis, University of Rennes, France.

Guiller, A., Madec, L. & Daguzan, J. (1994). Geographical patterns of genetic differentiation in the landsnail *Helix aspersa* Müller (Gastropoda: Pulmonata). *Journal of Molluscan Studies* **60**, 205–221.

Guiller, A., Coutellec-Vreto, M.-A. & Madec, L. (1996). Genetic relationships among suspected contact zone populations of *Helix aspersa* (Gastropoda: Pulmonata) in Algeria. *Heredity* **74**, 113–129.

Guiller, A., Bellido, A. & Madec, L. (1998). Ordination and genetic distances: the land snail *Helix aspersa* in North Africa as a test case. *Systematic Biology* **47**, 208–227.

Guiller, A., Coutellec, M.-A., Madec, L. & Deunff, J. (2001). Evolutionary history of the land snail *Helix aspersa* in Western Mediterranean: preliminary results inferred from mitochondrial DNA sequences. *Molecular Ecology* **10**, 81–87.

Hamerly, G. & Elkan, C. (2002). Alternatives to the *k*-means algorithm that find better clusterings. In *Proceedings of the ACM Conference on Information and Knowledge Management* (*CIKM*) (McLean, Virginia, USA), pp. 600–607. New York: ACM Press.

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography or seeing genes in space and time. *Molecular Ecology* **10**, 537–549.

Hutchison, D. W. & Templeton, A. R. (1999). Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution* **53**, 1898–1914.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Jacquez, G. M., Maruca, S. & Fortin, M.-J. (2000). From fields to objects: a review of geographic boundary analysis. *Journal of Geographical Systems* **2**, 221–241.

Legendre, P. & Fortin, M.-J. (1989). Spatial pattern and ecological analysis. *Vegetatio* **80**, 107–138.

Mackay, D. J. C. (2003). An example inference task: clustering. In *Information Theory, Inference, and Learning Algorithms* (ed. D. J. Mackay), pp. 284–292. Cambridge: Cambridge University Press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (ed. L. M. Le Cam & J. Neyman), pp. 281–297. Berkeley: University of California.

Madec, L. (1989). Etude de la différenciation de quelques populations géographiquement isolées de l'espèce *Helix aspersa* Müller (Mollusque, Gastéropode, Pulmoné): aspects morphologiques, écophysiologiques et biochimiques. PhD thesis, University of Rennes, France.

Madec, L. & Guiller, A. (1994). Geographic variation of distal genitalia in the landsnail *Helix aspersa* (Mollusca: Gastropoda). *Journal of Zoology*, **233** 215–231.

Madec, L., Bellido, A. & Guiller, A. (1996). Statistical and biogeographical significances of patterns of morphological and biochemical variation in the land snail *Helix aspersa*. *Comptes Rendus de l'Académie des Sciences Paris,* Series III **319**, 225–229.

Madec, L., Bellido, A. & Guiller, A. (2003). Shell shape of the land snail *Cornu aspersum* in North Africa: unexpected evidence of a phylogeographical splitting. *Heredity* **91**, 224–231.

Manel, S., Schwartz, M., Luikart, G. & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution* **18**, 189–197.

Merz, P. (2003). Analysis of gene expression profiles: an application of memetic algorithms to the minimum sum-of-squares clustering problem. *Biosystems* **72**, 99–109.

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

Moore, A. (2001). K-means and hierarchical clustering. http://www.autonlab.org/tutorials/kmeans.html

Müller, J. (2001). Invasion history and genetic population structure of riverine macroinvertebrates. *Zoology* **104**, 346–355.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.

Oden, N. L. (1984). Assessing the significance of a spatial correlogram. *Geographical Annals* **16**, 1–16.

Oden, N. L. & Sokal, R. R. (1986). Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Systematic Zoology* **35**, 608–617.

Pal, N. R. & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *Transactions on Fuzzy Systems* **3**, 370–379.

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

Ray, D. M. & Berry, B. J. L. (1966). Multivariate socio-economic regionalization: a pilot study in central Canada. In *Papers on Regional Statistical Studies* (ed. S. Ostry & T. Rymes), pp. 75–130. Toronto: University of Toronto Press.

Ray, S. & Turi, R. H. (1999). Determination of number of clusters in *k*-means clustering and application in colour image segmentation. In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques* (ed. N. R. Pal, A. K. De & J. Das), pp. 137–143. Calcutta, India.

Ripley, B. D. (1981). *Spatial Statistics*. New York: Wiley.

Sacchi, C. F. (1958). Les Mollusques terrestres dans le cadre des relations biogéographiques entre l'Afrique du Nord et l'Italie. *Vie et Milieu* **9**, 11–52.

Späth, H. (1980). *Cluster Analysis Algorithms*. Chichester: Ellis Horwood.

Taylor, J. W. (1913). *Monograph of the Land and Freshwater Mollusca of the British Isles*, pp. 236–273. Leeds: Taylor Brothers.

Templeton, A. R. (1998). Nested clade analyses of phylogeographic data: testing hypotheses about gene

flow and population history. *Molecular Ecology* **7**, 381–397.
Thioulouse, J., Chessel, D., Dolédec, S. & Olivier, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing* **7**, 75–83.
Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 238–244.
Weir, B. S. & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.
Womble, W. (1951). Differential systematics. *Science* **28**, 315–322.
Zhang, B., Hsu, M. & Dayal, U. (1999). *K-Harmonic Means: A Data Clustering Algorithm*. Palo Alto, CA: HP Laboratories.