Letter to the Editor

Optimal Randomization Strategies When Testing the Existence of a Phylogeographic Structure

Rémy J. Petit[1] **and Delphine Grivet**

Institut National de la Recherche Agronomique (INRA), Recherches Forestières, F-33611 Cestas, France

Manuscript received October 17, 2001
Accepted for publication February 11, 2002

TO allow quantitative analysis of phylogeographic data, Templeton *et al.* (1995) have developed a statistical framework called the nested cladistic analysis (NCA) of geographical distance, which can be used to infer past and ongoing processes using the genealogical and geographical information available in population genetic surveys. It is the only available technique that incorporates the explicit testing of the geographic spread of a haplotype relative to its center or to the center of other haplotypes or higher-order lineages. Recently, a user-friendly software has been developed for NCA (Posada *et al.* 2000) and there is a growing literature that makes use of this approach using human, animal, and plant data sets (*e.g.*, Templeton 1998; Cruzan and Templeton 2000 and references therein). We discuss briefly the rationale for the randomization strategy used in NCA. This work was motivated by surprising results obtained when applying this method to our own data on forest trees. A data set of chloroplast (cp) DNA variation in oaks is used for illustration purposes.

Among the parameters used in NCA to quantify geographic structure, $D_c(X)$ measures the geographic spread of the individuals that bear haplotype X (mean distance in kilometers between each individual bearing haplotype X and the geographic center of the haplotype). The permutation method proposed in the NCA to test if this parameter (among others) differs from expectations on the basis of a null hypothesis makes use of the algorithm proposed by Roff and Bentzen (1989). This consists of a random permutation of haplotypes (individuals in the case of haploid genetic data) or clades across the geographical space covered by the nesting clade category, keeping all marginal values constant (total number of each haplotype and population sample sizes). As a consequence, two factors can contribute to the outcome of the test: the uneven frequency of haplotypes across populations (regardless of their relative distribution in space) and the existence of a spatial pattern *sensu stricto* (due to a spatial dependence of haplotype frequencies in a population to those of the neighboring populations).

We consider that both sources of genetic structure should be disentangled and tested successively. To test the spatial distribution of haplotypes, it is important to keep other aspects of the genetic structure constant, such as the level of fixation within populations. A similar point, within another field, was recently made by Hosworth *et al.* (2001). The inference keys used by Templeton imply some knowledge of the (relative) spatial distribution of haplotypes, such as whether newly derived haplotypes are more widely distributed or more restricted than those from which they derive by mutation. For such purposes, which are primarily spatial in scope, we suggest that the null hypothesis should be different, such that a sample site approach is adopted, using a permutation of populations against geographic locations. As a consequence, only the assumption of independence of populations is being tested, and not the assumption of independence of individuals within populations. The existence of genetic differentiation (F_{ST}) can be tested separately for each haplotype using one of several approaches proposed so far (*e.g.*, Hudson *et al.* 1992; Pons and Petit 1995; Raymond and Rousset 1995). Actually, the algorithm implemented in NCA has been used to test the null hypothesis of absence of differentiation between two or more populations (Roff and Bentzen 1992). The use of separate tests for genetic fixation within population and spatial structure can help distinguish among different processes that occur at different spatial scales. Processes responsible for genetic fixation within a population include local dispersal, drift in small populations, or even clonal reproduction in some organisms, whereas a broad-scale geographic pattern can be due to population expansion, including through rare long-distance dispersal events, or isolation by distance. Below, we show that the choice of the permutation method makes a large difference when testing phylogeographical hypotheses.

[1]*Corresponding author:* Institut National de la Recherche Agronomique (INRA), Recherches Forestières, BP45, F-33611 Gazinet cedex, France. E-mail: petit@pierroton.inra.fr

TABLE 1

Comparison of the two methods of randomization for three haplotypes belonging to the same cpDNA lineage in oaks

Haplotypes	10	11	12
D_c (kilometers)	581	423	279
P value (populations)	0.219	0.661	0.124
P value (individuals)	0.072	0.046	<0.001

A total of 1000 permutations were used in each case. The P values correspond to the probability that the observed D_c value is significantly smaller than the expected one.

A WORKING EXAMPLE

In European oaks, cpDNA variation has been the subject of intensive studies (*e.g.*, Dumolin-Lapègue *et al.* 1997). Recently, cpDNA variation was studied in >2600 populations across Europe by a consortium of 16 laboratories (Petit *et al.* 2002). The sampling strategy was to maximize the number of populations with fewer individuals per population (in general about four or five trees only), because fixation index is high in this species complex (G_{ST} = 0.83) and there is therefore little reward in increasing sample sizes per population (Pons and Petit 1995). Such a large data set is especially well suited to test hypotheses on the past history of the species.

To compare both permutation procedures, we used a smaller data set of cpDNA variation where oaks had been sampled in 29 forests throughout Europe, with higher sample sizes per population (10–30). D_c was measured for the three cpDNA haplotypes belonging to the westernmost lineage (two "tip" haplotypes: nos. 11 and 12, and one "interior" haplotype, no. 10, from which they derive). The values of this statistic were then compared with their expectations obtained using either permutation method (Table 1). The values of D_c for haplotypes 11 and 12 were significantly small ($P < 0.05$ and $P < 0.001$, respectively) when permutation of individuals was carried out. On the other hand, after permutation of the populations, no test was significant (Table 1). When using the inference key proposed in NCA, a different path is therefore followed at step 2, depending on the permutation method adopted. In particular, with significantly small D_c values for the tip haplotypes (*i.e.*, a restricted distribution of these haplotypes compared to the other haplotypes of the lineage), possible inferences include restricted gene flow with isolation by distance or range fragmentation. Instead, if D_c values for tip haplotypes are significantly large or not significant, hypotheses of range expansion are preferred. Results from the more comprehensive genetic survey indicated that haplotypes 11 and 12 had a large distribution, comparable to that of haplotype 10, although their distribution was patchy at the local scale. These data and pollen evidence both suggest that a major expansion occurred from a glacial refugium located in the Iberian peninsula, involving long-distance seed dispersal events (Dumolin-Lapègue *et al.* 1997; Petit *et al.* 1997), a scenario compatible with the population permutation test.

We also carried out subsampling of varying numbers of oak populations (5–50, taken at random among the 600 that were characterized by haplotypes 10–12) to examine what type of results GEODIS would return. This led to a range of heterogeneous results (D_c values alternatively significantly large, nonsignificant, or significantly small, for a given haplotype) and to diametrically opposite inferences, including past fragmentation, restricted gene flow, or range expansion of some type. With more (*e.g.*, 200) populations, the results converged toward inferences of range expansion.

This suggests that the statistical tests implemented by GEODIS were insufficiently conservative, at least as far as spatial inferences are concerned. Templeton (1998) has also identified a similar situation (involving long distance colonization followed by a lack of genetic variation) where the NCA method is not appropriate, among several others where the method seemed to perform well. The example of European oak may represent another such example and so may not have much generality *per se* in demonstrating the limits of the NCA method.

The situation that is most likely to be problematic occurs when the level of fixation is high, the number of populations low, and the number of individuals per population high. In such a case, if a lineage consists of haplotypes distributed in only three populations, parameters (such as D_c) may be found to be significantly small, when using a random permutation of haplotypes. On the other hand, random permutation of populations would never detect a spatial structure in this example (because each configuration has one chance out of six to appear). Clearly, with such a limited sampling of populations, variation in haplotype frequencies across populations can be studied but the question of the relative distribution of haplotypes in space should not be addressed. For spatial inferences, our previous experience points to the need for sufficient and homogeneous sampling of populations, even at the expense of sample sizes per population (Pons and Petit 1995). The necessity for a homogeneous sampling has also been underlined by Cruzan and Templeton (2000).

The oak example shows that the choice of the permutation method substantially affects the outcome of the tests. If inferences about spatial patterns are to apply generally (*i.e.*, to other unsampled populations), we argue that the populations should constitute the unit of permutation, not the haplotypes. Given that several of the studies that had used the NCA method included few populations (<20), we consider that the conclusions reached (*i.e.*, the processes inferred) may need reappraisal. Also, alternative methods are needed that will not lead investigators to underestimate the number of populations for inferring spatial processes.

We thank David Posada and Alan Templeton for discussing some concepts involved in the nested clade analysis of geographical distance method as well as for responding in depth to first drafts of this letter. The comments of Martin Lascoux, Frédéric Austerlitz, and Kate Porter on the manuscript are also gratefully acknowledged. The study has been carried out with financial support from the Commission of the European Communities, Agriculture and Fisheries (FAIR) specific RTD programme, CT97-3795, "CYTOFOR."

LITERATURE CITED

Cruzan, M. B., and A. R. Templeton, 2000 Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record. Trends Ecol. Evol. **15:** 491–496.

Dumolin-Lapègue, S., B. Demesure, V. Le Corre, S. Fineschi and R. J. Petit, 1997 Phylogeographic structure of white oaks throughout the European continent. Genetics **146:** 1475–1487.

Housworth, E., J. G. Mezey, J. M. Cheverud and G. P. Wagner, 2001 The test distribution of modularity statistics: a correction and a clarification. Genetics **158:** 1381.

Hudson, R. R., D. D. Boos and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Petit, R. J., E. Pineau, B. Demesure, R. Bacilieri, A. Ducousso *et al.*, 1997 Chloroplast DNA footprints of postglacial recolonisation by oaks. Proc. Natl. Acad. Sci. USA **94:** 9996–10001.

Petit, R. J., U. M. Csaikl, S. Bordács, K. Burg, E. Coart *et al.*, 2002 Chloroplast DNA variation in European white oaks: phylogeography and patterns of diversity based on data from over 2600 populations. For. Ecol. Manage. **156:** 5–26.

Pons, O., and R. J. Petit, 1995 Estimation, variance and optimal sampling of gene diversity. I. Haploid locus. Theor. Appl. Genet. **90:** 462–470.

Posada, D., K. A. Crandall and A. R. Templeton, 2000 A program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Mol. Ecol. **9:** 487.

Raymond, M., and F. Rousset, 1995 An exact test of population differentiation. Evolution **49:** 1280–1283.

Roff, D. A., and P. Bentzen, 1989 The statistical analysis of mitochondrial DNA polymorphisms: chi-square and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

Roff, D. A., and P. Bentzen, 1992 Detecting geographic subdivision: a comment on a paper by Hudson *et al.* (1992). Mol. Biol. Evol. **9:** 968.

Templeton, A. R., 1998 Nested clade analysis of phylogeographical data: testing hypotheses about gene flow and population history. Mol. Ecol. **7:** 381–397.

Templeton, A. R., E. Routman and C. A. Phillips, 1995 Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics **140:** 767–782.

Communicating editor: A. H. D. Brown