doi:10.1111/j.1558-5646.2007.00124.x

THE AUTOMATION AND EVALUATION OF NESTED CLADE PHYLOGEOGRAPHIC ANALYSIS

Mahesh Panchal[1] and Mark A. Beaumont[2]

School of Biological Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading RG6 6AJ, United Kingdom

[1]*E-mail: m.panchal@rdg.ac.uk*

[2]*E-mail: m.a.beaumont@rdg.ac.uk*

Received August 10, 2006
Accepted February 6, 2007

Nested clade phylogeographic analysis (NCPA) is a popular method for reconstructing the demographic history of spatially distributed populations from genetic data. Although some parts of the analysis are automated, there is no unique and widely followed algorithm for doing this in its entirety, beginning with the data, and ending with the inferences drawn from the data. This article describes a method that automates NCPA, thereby providing a framework for replicating analyses in an objective way. To do so, a number of decisions need to be made so that the automated implementation is representative of previous analyses. We review how the NCPA procedure has evolved since its inception and conclude that there is scope for some variability in the manual application of NCPA. We apply the automated software to three published datasets previously analyzed manually and replicate many details of the manual analyses, suggesting that the current algorithm is representative of how a typical user will perform NCPA. We simulate a large number of replicate datasets for geographically distributed, but entirely random-mating, populations. These are then analyzed using the automated NCPA algorithm. Results indicate that NCPA tends to give a high frequency of false positives. In our simulations we observe that 14% of the clades give a conclusive inference that a demographic event has occurred, and that 75% of the datasets have at least one clade that gives such an inference. This is mainly due to the generation of multiple statistics per clade, of which only one is required to be significant to apply the inference key. We survey the inferences that have been made in recent publications and show that the most commonly inferred processes (restricted gene flow with isolation by distance and contiguous range expansion) are those that are commonly inferred in our simulations. However, published datasets typically yield a richer set of inferences with NCPA than obtained in our random-mating simulations, and further testing of NCPA with models of structured populations is necessary to examine its accuracy.

KEY WORDS: **Coalescent, nested clade phylogeographic analysis, phylogeography, sequence analysis, simulation.**

The task of trying to disentangle the historical processes that shape the evolution of a species and its geographical distribution is challenging. As pointed out by Hey and Machado (2003), the problem has typically been addressed in two rather different ways. The traditional approach is parametric, based on (typically) relatively simple models, whose parameters are inferred via estimators based on summary statistics or through calculation of likelihoods. This type of analysis is necessarily somewhat abstract, and once it became possible to routinely construct phylogenetic trees from DNA sequences (Cann et al. 1987), methods were developed to infer demographic history from tree topology (Avise et al. 1987). This approach invites us to "read" history directly through visualization of the trees—a much more concrete and intuitively appealing procedure. A widely used example of this latter technique is nested clade phylogeographic analysis (NCPA, Templeton et al. 1995), which is designed to distinguish between a wide array of processes

and events that shape a species' history, such as allopatric fragmentation, contiguous range expansion, and restricted gene flow due to isolation by distance. This breadth of scope is far in excess of what can currently be addressed using summary statistics or likelihood, and NCPA has grown in popularity and has been applied to a wide variety of species (over 400 citations are recorded for Templeton et al. 1995, and over 340 citations are recorded for Posada et al. 2000, in the Web of Science citation index).

The evidence for the reliability of NCPA stems primarily from concordance between conclusions drawn by the method and existing prior knowledge of a species' history (Templeton 1998, 2004). In some cases discrepancies have been found, and these have prompted suggestions for improvements (Paulo et al. 2002; Petit and Grivet 2002; Masta et al. 2003), some of which have been incorporated (Templeton 2004). Concerns have been raised on certain technical aspects (e.g., Petit and Grivet 2002; although see Templeton 2002a). Other criticisms have been made on more general grounds, for example, that it offers little statistical support for the inferences made, and does it provide information regarding the relative likelihood of alternative inferences (Knowles and Maddison 2002). As pointed out by Hey and Machado (2003), a problem with methods that are based on detailed interpretation of gene trees is that the same demographic history can lead to very different gene genealogies, and also it is very unlikely that the true tree will be recovered from sequence data.

The NCPA approach originates from the earlier nested clade analysis (NCA) proposed by Templeton and colleagues (Templeton et al. 1988), which was designed as a method for studying association between phenotype and genotype. The basic idea behind NCA still has significant application, and has inspired a number of statistical approaches to disease mapping (e.g., TREESCAN; Templeton et al. 2005; Posada et al. 2005; CLADHC; Durrant et al. 2004; Evolutionary Tree TDT; Seltman et al. 2001; Cladistic Test; Bardel et al. 2005). What makes the NCPA method unusual, however, and contentious (see e.g., Knowles and Maddison 2002 and Templeton 2004) has been the introduction of the inference key to interpret the summary statistics yielded by the method (Templeton et al. 1995).

There has been little detailed statistical testing of this widely used technique, particularly in comparison to that involved in the evaluation of parametric methods (Wang and Whitlock 2003; Choisy et al. 2004; Beerli 2006). The performance of NCPA has mainly been tested under models of fragmentation and range expansion, using a large number of empirical datasets in which there is strong prior independent evidence for the event (Templeton 1998, 2004). The performance of NCPA under a wider range of demographic models has not been explored in any detail. Furthermore, as noted below, a number of variations in the application of the method have been described, and again, have not been tested. These include, for example, the use of rooted versus unrooted gene trees, different ways to take into account uncertainty in the reconstruction of gene-trees, and the performance of the method when using multilocus data, introduced by Templeton (2002b).

Another commonly used approach to investigate a statistical procedure is through simulation of many replicate test datasets, on which its performance is then examined. Although in principle this could also be applied to NCPA, in practice it is difficult to do so because no fully automated package has yet been developed. Thus, evaluation of NCPA has to be carried out "by hand," and only a limited number of simulated datasets have thus far been analyzed (Knowles and Maddison 2002). The results suggest that NCPA may produce a number of false positives under a scenario of fragmentation, although certain assumptions were criticized by Templeton (2004). These criticisms highlight potential sensitivity to various factors in the NCPA method and thus further investigation seems desirable.

Described in this article is a fully automated procedure for performing NCPA, for single loci such as mitochondrial sequence data. This is implemented in the software ANeCA (Panchal 2007). Certain steps in NCPA have involved subjective choice and judgment on the part of the user, which makes the method difficult to automate, and we include as supplementary material how these have evolved since the method's inception, and show how the steps in the automated procedure are representative of those made in earlier publications. A survey of authors who have published using the method illustrates the (relatively small) degree of variation in interpretation, and supports the interpretations made for the automated method (provided as supplementary material). We apply the method to published datasets, previously analyzed manually by NCPA. We show that the automated method very often leads to identical or similar conclusions, and discuss in more detail those cases in which it does not. We examine the false-positive rate by applying NCPA to simulated datasets in which there is no genetic substructuring, and compare the frequencies of different types of inferences with those obtained in published papers from empirical datasets.

Implementation of Automated NCPA

Nested clade phylogeographic analysis is a method that involves several steps to obtain inferences from the locus being studied. Beginning typically with DNA sequences and the geographical coordinates of each individual, the NCPA procedure can be applied as follows:

(1) *Creation of a haplotype network.* A graph is constructed, with vertices consisting of observed and inferred haplotypes and edges representing single mutations. The package TCS version 1.18 (Clement et al. 2000) is used for the automated procedure described here.

(2) *Nesting of clades.* The haplotypes, both observed and inferred, are clustered hierarchically by progressively grouping vertices in the graph according to certain rules. These clusters are termed "clades." Lower order clades are nested within higher order clades, creating a nested design. Until now it has been necessary to perform this by hand.

(3) *Calculation of summary statistics and tests of significance.* Summary statistics are calculated that measure the geographical spread of members of a clade relative to their mean location, D_c, and the geographical spread of members of a clade relative to the mean location of all members of the nesting clade, D_n. These summary statistics are used to test the null hypothesis of no geographic association (Templeton et al. 1995). Tests of significance are made through a permutation procedure. The calculation of statistics and tests of significance are carried out by the program GeoDis version 2.2 (Posada et al. 2000).

(4) *Interpretation of results through an inference key.* Templeton and coworkers (Templeton et al. 1995; Templeton 1998, 2004) have developed and refined an inference key with which to interpret the summary statistics. From the inference key it is possible to conclude, for example, that there has been long-distance colonization, isolation by distance, or other demographic processes. Consultation of the inference key has until now been performed by hand.

We have developed a fully automated procedure that will take haplotype information in Nexus sequential format and lead to conclusions drawn from the inference key. Important issues that arose during the development of the process are discussed below.

CREATING THE HAPLOTYPE NETWORK

There is a choice of methods for constructing haplotype networks (for a review, see Posada and Crandall 2001). However, most commonly used for NCPA is the TCS package (Clement et al. 2000), which implements statistical parsimony (SP) (Templeton et al. 1992). Statistical parsimony was originally developed for restriction fragment length polymorphism (RFLP) data; however, it has been updated to estimate haplotype networks from DNA sequence data (Crandall 1996). Version 1.18 of TCS is implemented for the automated version of NCPA described here. We use TCS as the default haplotype network estimation method because it is the most commonly cited in the NCPA literature, however, both median-joining networks (Bandelt et al. 1999), and the union of maximum parsimony trees (Cassens et al. 2005) appear to provide better estimates of the true genealogy from DNA sequence data (Cassens et al. 2005).

The aim at this stage is to construct a rooted or unrooted tree. This is a graph in which each vertex (node) has, conceptually, only one ancestor and any number of descendants, although in an unrooted tree the ancestral and descendent edges are undefined. The current version of the automated algorithm assumes an unrooted tree, although it would be relatively straightforward to allow for trees to be rooted in future developments. Trees are a type of graph that has no cycles; that is, there is only exactly one path between any two nodes in the graph. However, it is possible to have a number of different and equally parsimonious gene-trees for particular haplotypes, caused by, for example, recombination or recurrent mutation. Rather than representing this uncertainty by different trees, some methods, such as SP used here, represent the uncertainty by adding extra edges in the tree, forming cycles (loops). Uncertainty in ancestry, as depicted by the loops, leads to uncertainty in the nesting of clades.

Historically, three general approaches have been taken to address the consequences of loops in the network: 1 choose the best resolution of the haplotype network according to certain criteria; 2 include within the process of nesting clades certain rules for dealing with loops; 3 explore all possible trees within the network noting which inferences are robust to uncertainty in the tree topology. In fact, generally, procedure 1 above does not always result in a tree (e.g., Pfenninger and Posada 2002), but results in a network with relatively simple loops that can be analyzed with method 2 or 3. We have implemented the original method for resolving loops given in Templeton et al (1995), based on Templeton and Sing (1993), and explicitly recommended in Templeton (2005a). In future it may be possible to include the recent alternative techniques that involve 1 and 3 above (e.g., Brisson et al. 2005).

For completeness, we summarize here examples of NCPA that have taken the first approach, approach 1, to address loops in the network. Methods based on genealogical reasoning (loosely, coalescent theory) are the most commonly used for simplifying networks that have loops, and are discussed in Crandall and Templeton (1993) and Pfenninger and Posada (2002). These procedures break loops in the network by removing edges. For example, Crandall and Templeton (1993) have suggested three criteria. The first is the geographic criterion, which states that the haplotypes are more likely to be connected to haplotypes from the same geographic area than to haplotypes from a distant geographic area. The second is the frequency criterion, which states that haplotypes are more likely to be connected to haplotypes with a higher frequency. The third is the topological criterion. This states that haplotypes are more likely to be connected to interior haplotypes (haplotypes with more than one connection to another clade) than to tip haplotypes (a haplotype with a single connection to another clade). It may be possible to incorporate these criteria in future implementations of the automated algorithm, once the current version, based on the original description in Templeton et al. (1995), has been fully explored, so that there is a baseline against which to test them. For example, a testable hypothesis is

that use of the geographic criterion will create a tendency to infer short-distance movements, because haplotypes that are geographically closer will tend to be united in the same clade.

As noted, for example, by Hey and Machado (2003), it is unlikely that gene-trees from multiple independent loci will agree with each other on the inferred demographic history. To address this, Templeton (2002b) proposed the use of a form of cross-validation in which inferences are regarded as concordant in space if more than one locus infers the same historical process involving the same locations. Temporal concordance of these events is then assessed using a maximum likelihood framework (Templeton 2005a), and this has allowed multiple, temporally distinct events to be identified (Templeton 2002b, 2005a). This extension of the original NCPA method, including the temporal analysis, appears not to have been widely used (examples include Templeton 2002b; Templeton 2005b [as cited in Templeton 2005a]). Spatial cross-validation (as defined above) has been used for two loci in Brisson et al. (2005), but without a temporal analysis. A number of other analyses have used NCPA with more than one locus (e.g., Zhang et al. 2005), but not in the formal framework suggested by Templeton (2002b). Ideally, a future goal would be to include the use of multiple loci in the automated procedure, but in the meantime we note that the current automated method will make it considerably easier to perform NCPA with cross-validation on such datasets.

AUTOMATING THE NESTING ALGORITHM

The objective of nesting is to hierarchically cluster haplotypes into related groups called clades. The process of nesting has until now been conducted manually using the rules given by Templeton et al. (1987), Templeton and Sing (1993), Crandall (1996), and Templeton (2002b). Some justification of the rules that have been suggested for nesting clades is given in Templeton (1999) and Templeton (2005a). For a network that is a tree the algorithm is straightforward.

(1) Identify the tip clades (haplotypes with only one connection to rest of the network).
(2) The haplotypes that are connected to a tip are then grouped with the tips to form 1-step clades, which are simply clusters of two or more haplotypes.
(3) Then identify the next set of "tip" clades among the ungrouped haplotypes. These are nodes of the graph with only 1 edge leading to an ungrouped haplotype.
(4) Steps 2 and 3 are applied repeatedly to create a complete set of 1-step clades.
(5) Once all the haplotypes have been grouped the 1-step clades are then treated as the haplotypes described in step 1 and steps 2 to 4 are applied to create 2-step clades.
(6) Step 5 is applied repeatedly until all the n-step clades would eventually be grouped into a single $n + 1$ step clade.

During grouping, ungrouped haplotypes may no longer be directly connected to other ungrouped haplotypes. These haplotypes are termed symmetrically stranded clades (Templeton and Sing 1993). In this situation a symmetrically stranded clade is grouped with the connecting $n + 1$-step clade that has the smallest sample size, choosing randomly between them if there are more than one with lowest sample size.

It is also possible to apply the nesting algorithm to networks that contain loops. As discussed above, the representation of uncertainty in the gene tree by means of loops is simply a device, used by, for example, TCS, to avoid returning multiple, equally parsimonious, trees. In this case the algorithm above has to be modified. An outline algorithm is as follows (commented code with the detailed algorithm available on request):

(1) Use the standard nesting rules until a loop is encountered.
(2) Two possibilities arise. The first is that a node within the loop is grouped into an $n + 1$ step clade (Fig. 1A), in which case follow step 3. The second possibility is that a node within the loop is not nested within an $n + 1$ step clade, but is connected to a node that is (Fig. 1B). In this case, follow step 4.
(3) Continue nesting using the standard nesting rules.

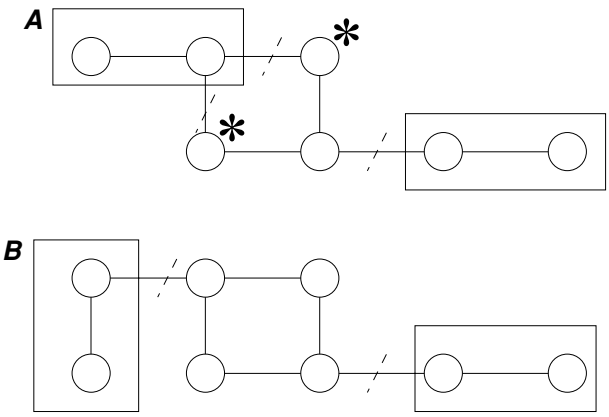

Figure 1. (A) It shows how a loop is broken by the standard nesting rules. We begin by nesting the clades one mutational step from the tip clades. This means that a node within the loop is grouped into an $n + 1$ step clade. By pruning off the $n + 1$ step clades the remainder of the haplotype network is no longer ambiguously connected, allowing us to continue using the standard nesting rules for trees. (B) It shows how haplotype networks can be left with no tip haplotypes to nest after the initial round of nesting and pruning, leaving ungrouped haplotypes and no way using the standard nesting rules for trees to continue. In this case the remainder of the haplotype network needs to be dealt with using additional rules, described in the main text. The dotted lines indicate which edges would be cut by pruning, and circles (haplotypes) with an asterisk indicate which haplotype would be next identified as a "tip" clade.

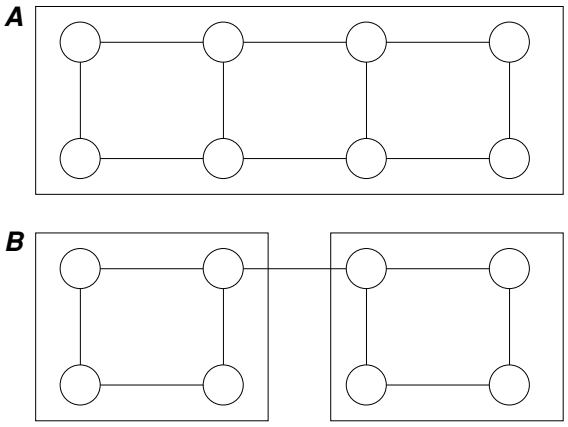

Figure 2. Both (A) and (B) show how loops are nested in the automated software. Haplotypes that are connected with edges that are part of a loop are grouped together into a single $n + 1$ step clade. (A) It shows a network where all the nodes are connected by edges within a loop. (B) It shows a network in which one edge is not part of a loop and so separates the two loops into different $n + 1$ step clades.

(4) Follow step 3 until only ungrouped nodes within a loop remain.

(5) Ungrouped nodes within a loop are grouped into an $n + 1$ step clade, as shown in Figure 2.

(6) Continue nesting again with the standard nesting rules.

Another possibility is that the network estimation method, in this case SP, may not connect all haplotypes into a single network, because the number of mutations connecting two observed haplotypes exceeds a preset threshold (Templeton and Sing 1993). The nesting rules are applied simultaneously to each disjoint network. When a disjoint network has been grouped into a single n-step clade, it is nested by itself until all the disjoint networks are single n-step clades, at which stage they are all grouped together terminating the algorithm. The treatment of disjoint networks described here involves a modification of their treatment in Templeton and Sing (1993).

During automation of the nesting procedure, some ambiguities in the earlier literature became apparent, and decisions needed to be taken so that the algorithm was consistent in its application. These are detailed in the supplementary material.

SUMMARY STATISTICS AND TESTS OF SIGNIFICANCE

The summary statistics, D_c and D_n, are calculated using GeoDis version 2.2 (Posada et al. 2000), which includes a refinement to the original method given in Templeton et al. (1995) to include relative abundance into the equations for D_c and D_n. The details, and equations used are given in Posada et al. (2006). In a number of clades these statistics are undefined and ignored by Geodis: for example, clades with only a single haplotype, or clades only consisting of inferred rather than observed haplotypes, or when individuals within a clade come from a single location. The equations are based on the use of great circle distances (the shortest distance as measured on the surface of a sphere) between sampled locations.

Sometimes the use of great circle distances may be inappropriate, such as with riparian species (Templeton et al. 1995; Posada et al. 2000; Turner et al. 2000, Fetzner and Crandall 2003). In this case a matrix of pairwise distances between sample locations may be better, and so an alternative method for calculating D_c and D_n is provided.

Two further statistics are calculated for each clade if possible. They are D_c and D_n for the interior-tip contrast of clades, $I - T$. Although these are given the same symbols, the definition of these statistics is slightly changed in that it is the difference between the average of the D_cs (or D_ns) calculated for the interior clade(s) and the average of the D_cs (or D_ns) calculated for the tip clade(s) (see Posada et al. 2006). These $I - T$ statistics indicate whether the average spread of the tip clades (presumed to be the younger haplotypes; Castelloe and Templeton 1994), and the average spread of the interior clades (presumed to be older) are significantly different from each other. Thus temporal information is incorporated into the analysis (Templeton et al. 1995).

A P-value is attached to each of these statistics by carrying out random permutations of the geographic locations of individuals and finding the proportion of repetitions in which values of the statistic more extreme than observed are generated. The procedure preserves haplotype frequencies and sample sizes (Roff and Bentzen 1989). In the automated software the default setting is to use 10,000 permutations to reduce the variance associated with the estimated P-values (the default for the standard GeoDis software is 1000).

AUTOMATION OF THE INFERENCE KEY

Outline of the approach

Each clade that is analyzed (1-step clades and above) will have associated with it a number of D_c and D_n statistics, depending on the number of clades nested within it. The final stage in NCPA is the interpretation of these statistics by means of an inference key. The inference key guides the user to see if the observed patterns emerging from the nested tests correspond to a priori expectations from certain models (Templeton et al. 1995; Templeton 2002a, 2004). However, it has not been shown that specific geographic patterns (either described by clade boundaries, or combinations of D_c and D_n) arise only from these models (and not others). The inference key also suggests where sampling inadequacies might be, and clades in which additional tests could reveal further information. Details of the difficulties in automation of the key are provided in the online supplementary material.

Application of the Automated NCPA to Published Datasets

We applied the automated software to three published datasets to investigate to what extent the software could recover original conclusions drawn from these data in manual analyses. These analyses are uncensored (i.e., no other datasets have been analyzed, and excluded). The choice of datasets here is largely arbitrary: five authors were contacted on the basis of recent publications that came to our attention while developing the method, and three provided sufficient information to enable a comparison to be made.

Leaf beetle, Gonioctena pallida

The first dataset, from Mardulyn (2001), consists of 363 nucleotide long portion of the mtDNA control region from 242 individuals. The beetles were sampled primarily from the Voges Mountains in France, with a couple of localities in the Black Forest in Germany. The minimum number of individuals from any sample location was 13.

We applied the automated software to the data. The initial network provided by TCS contained many loops, and was very similar to the published network. The difference in networks may be explained by changes to the TCS software (ver. 1.01 alpha vs. ver. 1.18) because there was no manual manipulation of the sequence data. Due to the high number of loops within the network, Mardulyn (2001) chose to resolve the ambiguities using the methods presented in Crandall and Templeton (1993). As a result, the haplotype network obtained using TCS version 1.18 was modified by removing the branches and missing intermediate haplotypes to obtain the tree analyzed by Mardulyn (2001). Nesting was then performed on the modified haplotype tree using the automated software, after which GeoDis version 2.2 was applied using a setting of 1000 permutations to obtain the distance statistics and probabilities of each clade. In both analyses there were 37 clades, out of which 24 yielded distance statistics. The statistics obtained matched those that were published (obtained with GeoDis ver. 2.0), and the *P*-values for the significance tests were similar. The automated key was then applied, but in this case information regarding unsampled locations was omitted because it was unavailable. There was a close similarity between the results obtained from the software and those published, as shown in Table S2 provided in the online supplementary material. Although different inference keys are used in the evaluation, they are similar. Mardulyn (2001) uses the appendix to Templeton et al. (1995), while the automated inference key is based on the version dated 14 July 2004. The difference in inference in Clade 1–1 is due to the statistic being very close to the 0.05 significance level, and so due to variance the statistic will occasionally be nonsignificant. The difference in the chain of inference for Clade 1–8 is due the use of different versions of the inference key, although the final inference is the same. The difference in Clade 1–16 appears to be due to a difference in interpretation of Question 2b in the July 2004 version of the inference key, regarding the phrase ". . . and the D_c's for some but *not* all of the interiors are significantly small." Clade 1–16 contains a single interior clade with a significantly small D_c. The software interprets the phrase as: there is at least one interior clade that has a significantly small D_c, but also that there exists another interior clade that does not have a significantly small D_c. Clade 3–1 yields the same inference, however, the difference in the chain of inference is again because different versions of the inference key are used.

We also ran the automated software on the unmodified haplotype network. We found that nesting produced 36 more clades in the entire nesting design. As a result there were 31 clades out of 64 that could potentially contain inferences—that is, clades for which the summary statistics can be calculated, and are therefore processed by Geodis. This compares with 24 clades out of the 37 for the modified haplotype network. Out of these 31 clades only 6 rejected the null hypothesis compared to 13 out of 24 clades in the modified network. This appears to be because of the reduced sample sizes, and the increased number of clades. The inferences made were of restricted gene flow with isolation by distance (three occurrences), inconclusive outcome (two occurrences), and contiguous range expansion (one occurrence).

Iberian Lizard, Lacerta schreiberi

The next dataset is from Paulo et al. (2002), consisting of 83 individuals sequenced at the Cytochrome b region in the mitochondrial DNA. Sequences were 663 base pairs long. Individuals were sampled from 18 locations in the Iberian Peninsula, with sample sizes ranging from one to 10 individuals per site.

Applying the software, we obtained the same haplotype network that was published and an almost identical nesting design. The only difference to the nesting design was the placement of haplotype 3 in figure 3 of Paulo et al. (2002). As a result, Clade 2–1 in Paulo et al. (2002) became potentially informative. We applied GeoDis with a setting of 1000 permutations and used user-defined distances, followed by the application of the automated inference key. The results are presented in Table S3 in the online supplementary material. Again information about unsampled or potentially habitable areas was unavailable and so it was omitted from our data. The geographic data provided by Paulo et al. (2002) was provided as user-defined distances; however, the automated software requires the longitude and latitude of locations. Relative coordinates were used in place of longitude and latitude, and were obtained by placing a grid over the map, figure 2 in Paulo et al. (2002).

Again we find explainable differences in the inferences obtained. At the first level all inferences match. At the second level of nesting is where discrepancies begin. Clade 2–1 was made potentially informative due the mismatch in nesting. The differences in Clades 2–2 and 2–4 are due to there being less than three clades

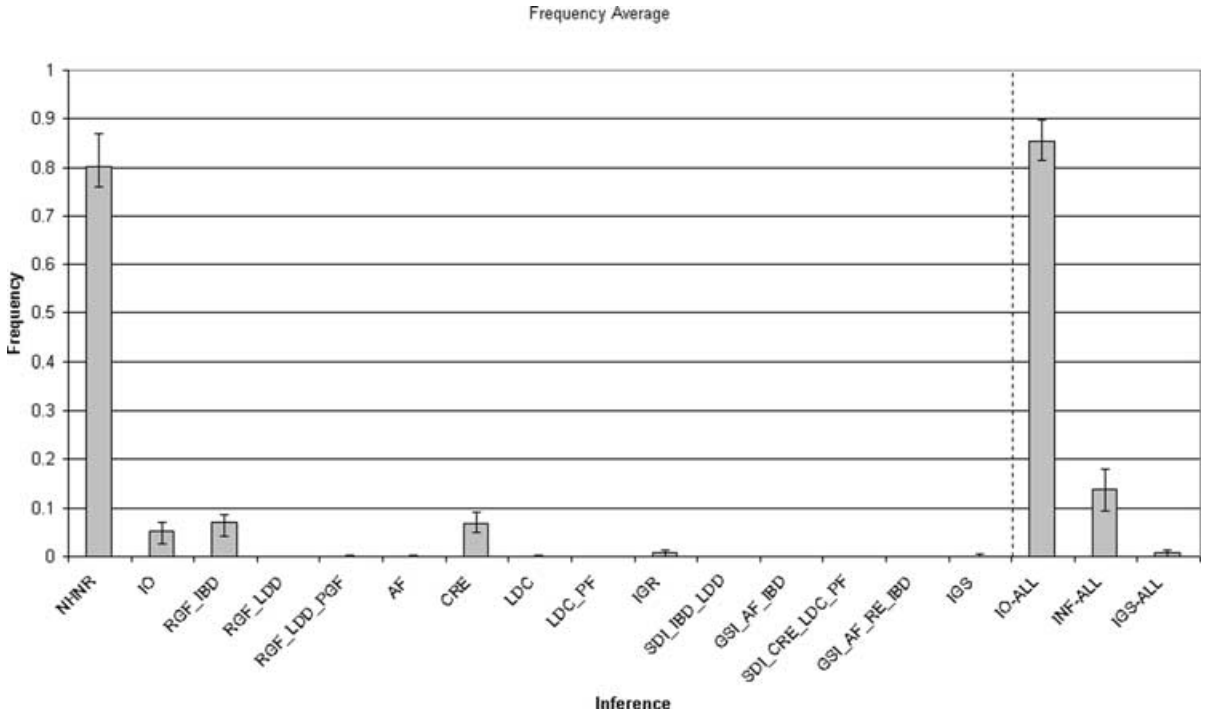

Figure 3. This figure shows the frequency at which an inference was made by NCPA among all replicates. The level at which the inference was made was ignored. The error bars show the range of frequencies obtained from the different parameter sets. Each column refers to a particular inference made by the inference key as follows. NHNR, null hypothesis of no geographic association was not rejected; IO, inconclusive outcome; RGF_IBD, restricted gene glow with isolation by distance; RGF_LDD, restricted gene glow with some long-distance dispersal; RGF_LDD_PGF, restricted gene flow with some long-distance dispersal over intermediate areas not occupied by the species, or past gene flow followed by extinction of intermediate populations; AF, allopatric fragmentation; CRE, contiguous range expansion; LDC, long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion; LDC_PF, long-distance colonization and/or past fragmentation (not necessarily mutually exclusive); IGR, insufficient genetic resolution to discriminate between range expansion/colonization and restricted dispersal/gene flow; SDI_IBD_LDD, sampling design inadequate to discriminate between isolation by distance (short-distance movements) versus long-distance dispersal; GSI_AF_IBD, geographic sampling inadequate to discriminate between fragmentation and isolation by distance; SDI_CRE_LDC_PF, sampling design inadequate to discriminate between contiguous range expansion, long-distance colonization, and past fragmentation; GSI_AF_RE_IBD, geographic sampling inadequate to discriminate between fragmentation, range expansion, and isolation by distance; IGS, inadequate geographical sampling; IO_ALL, inference is either NHNR or IO; INF_ALL, inference is RGF_IBD, RGF_LDD, RGF_LDD_PGF, AF, CRE, LDC, or LDC_PF; IGS_ALL, inference is IGR, SDI_IBD_LDD, GSI_AF_IBD, SDI_CRE_LDC_PF, GSI_AF_RE_IBD, or IGS.

with significant reversals. This is because of a difference of interpretation of the inference key. The difference in 2–3 is due to rooting by Paulo et al. (2002) between the two disjoint networks, where haplotype 17 is treated as an interior. Rooting also affected inferences in Clades 3–1, 3–2 and the Total Cladogram, as did the missing geographic information.

Leaf beetle, Timarcha goettingenis *complex*

Gómez-Zurita and Vogler (2003) sequenced 167 individuals from 37 localities in the Iberian Peninsula and other parts of Europe. Sequences were 639 base pairs long from the nuclear marker, the Internal Transcribed Spacer Region 2. Sample sizes ranged from one to 14 individuals per location.

Application of TCS to this dataset resulted in four disjoint networks. Gómez-Zurita and Vogler (2003) applied NCA to these networks, which were rooted according to Gómez-Zurita et al. (2000a,b), and this information was used in the NCPA analysis. It should be noted that this will result in some differences with the automated analysis, because the latter does not include the additional information that arises from the choice of root. Our automated nesting algorithm gave a design that was identical in the early levels to that in Gómez-Zurita and Vogler (2003) (exceptions arising in Clades W2-4, W2-5, and W2-6 upward). Geodis was run with the options to use 1000 permutations and decimal degrees. The automated inference key was applied, and unsampled locations were omitted from the data. The results are in found in Table 4 of the supplementary material.

The results are discordant with those published by Gómez-Zurita and Vogler (2003). The difference in clades W1-8 is due to a difference in interpretation: the software looks for at least

two clades that have significantly small D_c values, whereas clade W1-8 contains only one. The differences in clades W2-8, E2-2, W3-6, E3-1, and W4-4 are due to missing geographic information regarding the unsampled areas. The difference in clade W5-1 is due to a difference in the nesting design. Rooting the network has changed the Tip/Interior status of several subclades in clades W3-5, W4-3, E4-1, and the Total Cladogram, which changes the inferences. We have no explanations for the discrepant results obtained for Clades W1-2, W1-3, W1-4, W1-6, W-17, E1-1, W2-1, W2-4, W3-2, and W3-3. The D_c and D_n values provided in the supplementary material to Gómez-Zurita and Vogler (2003) match those that we obtain in these clades. However, the levels of significance for these statistics given in Gómez-Zurita and Vogler (2003) do not match with those that we have obtained with GeoDis. This cannot be explained by sampling variance in the permutation procedure. Although we use GeoDis version 2.2, we ran GeoDis version 2.0 on the GeoDis input to verify that this was not due to any update of GeoDis.

Application to Simulated Data Under a Panmictic Model

We aimed to investigate the false-positive rate in our automated version of NCPA by applying it to spatially distributed data simulated to have no genetic population structure. Under this scenario, it is expected that NCPA is unable to reject the null hypothesis of no geographic association. A more moderate expectation would be that even if some distance statistics are found to be significant, the inference key would be unable to reach a conclusion that infers structural or historical processes.

SIMULATION OF SAMPLES

SIMCOAL version 1.0 (Excoffier et al. 2000) was used to generate 100 genealogies from which DNA sequence data were sampled. We then used the automated NCPA software, which includes TCS version 1.18 and GeoDis version 2.2, to build the haplotype network, perform the nesting and permutation tests, and apply the inference key.

The effect of several factors was investigated. The first was the effect of the topology of the haplotype network. A comparison was made between the panmictic simulation and another simulation where the geographic locations of individuals were randomly permuted, effectively removing geographic associations in a general model. For this latter case, five additional genealogies were simulated, using the same parameters as for the 100 genealogies discussed above. For each of these genealogies, 100 random permutations were performed.

Another factor that was tested was how the number of demes affected results. Demes were assigned positions in a grid under three different sizes, 3×3, 7×7, and 10×10 prior to running the simulations (although SIMCOAL itself does not incorporate spatial information). Spacing between adjacent demes was (notionally) set to 1.53 decimal degrees (about 170 km) and each deme has a radius of 70.0 km. The units themselves are unimportant, but a scale needs to be set so that the radius can be specified for the inference key, as discussed below.

The proportion of demes sampled was another factor that was investigated. Simulations were conducted where all the demes were sampled and half the demes were sampled. In the case when half the demes were sampled, demes were selected that were diagonal to each other from the first deme in the grid. This was applied to all three sizes of grids.

The final factor that was investigated was the total sample size. Three ranges for sample sizes were used, the first from 900 to 1000, and the second from 450 to 500, and the third from 180 to 200. It should be noted that the total sample sizes are generally greater than sample sizes obtained in real data, to avoid the problem of insufficient sample size. In the 10×10 grids the sample sizes were 10, 5, and 2 for each deme from each sample range respectively. For the 7×7 grids the sample sizes were 20, 10, and 4, and for the 3×3 grid the sample sizes were, 100, 50, and 20. The methods of calculating D_c and D_n, were also investigated by specifying the geographic information (sample locations) in both decimal degrees, and as a distance matrix, where distances were calculated as great circle distances.

For each simulation the total number of individuals simulated ranged between 90,000 and 100,000. To simulate panmixia in the SIMCOAL framework, each deme coalesced into a single deme at time zero going backward in the simulation. There was no deme growth or further migration. (An exactly equivalent alternative would have been to simulate the samples from a single population, and then distribute them among different demes. However, it was convenient to follow the former method because the study reported here was part of a wider study that included a number of complex demographic scenarios, the results of which will be published in a later report.) We simulated DNA sequences 500bp long with a mutation rate per generation of 0.00002 and a transition bias of 0.66666666. The mutation rate per generation was derived by assuming four years per generation, and that mitochondrial DNA has an overall mutation rate of 1×10^{-8}per site (Pesole et al. 1999). We also assumed that the mutation rate was gamma distributed with a shape parameter of 4, using 10 rate classes.

Results of Simulations with Panmictic Data

The results from NCPA can be analyzed on a clade-by-clade basis and also for each dataset. Thus, a researcher would be tempted to publish an inference from a dataset if any of the clades gave rise to an inference. This would typically happen if the *P*-value of a statistic from Geodis in any clade is significant, because then

the researcher would work through the inference key, although in some cases this might result in an "inconclusive outcome."

In each panmictic simulation we found at least one significant statistic in 20% of the clades in each parameter set. This is shown in Figure 3, where the null hypothesis of no geographic association is not rejected on average 80% of the time. The results show that restricted gene flow with isolation by distance and contiguous range expansion are the most common conclusive inferences if an inference is made. The results also show that other conclusive inferences (e.g., allopatric fragmentation, long-distance colonization) are very rare under panmixia (Fig. 4). Per dataset, these results show that there is on average a 50% chance of obtaining restricted gene flow with isolation by distance, and a 50% chance of inferring contiguous range expansion (Fig. 4). It also shows there is $< 1\%$ of finding one of the other conclusive inferences at the dataset level. The randomized models generated for each simulation also showed similar results (provided in the online Supplementary Tables S5–S10).

On a dataset basis, 13% of all the datasets analyzed have no significant D_c or D_n statistic. However, in some of the cases where a statistic is significant, the inference key would lead to "inconclusive outcome." If we also treat these as "nonsignificant," and treat all other outcomes as "significant" (including those cases that lead to some demographic inference but also highlight sampling inadequacy) we find that 24% of all datasets have no "significant" outcome.

In these simulations the relationship between the probability, P, of obtaining at least one significant statistic in any clade tested and the number of summary statistics in the clade, n, is complex. In comparison to the naïve expectation of $p = 1 - (1 - \alpha)^n$, where α is the significance level used, we find that P is too low at small n, and too high for large n (results not shown). This is unsurprising because the various correlations among summary statistics in the nested clade design mean that the tests are not independent. Permutation methods have been developed to control the Familywise error rate (i.e., the probability of making at least one error in the family of tests) when there are dependencies, and a particular algorithm—Free Step-Down resampling (Algorithm 2.8; Westfall and Young 1993)—has been applied in a related program, TREESCAN (Templeton et al. 2005; Posada et al. 2005). Our results suggest that future development of NCPA requires the application of similar methods to control the false-positive rate, but it would be fair to say that none of the published analyses so far have done so, and our results suggest that the false-positive rate may have had some effect on the outcome of previous analyses that have used NCPA.

We also used binomial logistic regression and Akaike's An Information Criterion (AIC) to determine which other variables affected if a clade had at least one significant distance statistic. The analysis revealed that along with the number of statistics, the level of the clade, the number of locations, and the proportion of locations sampled, affected if at least one significant statistic is

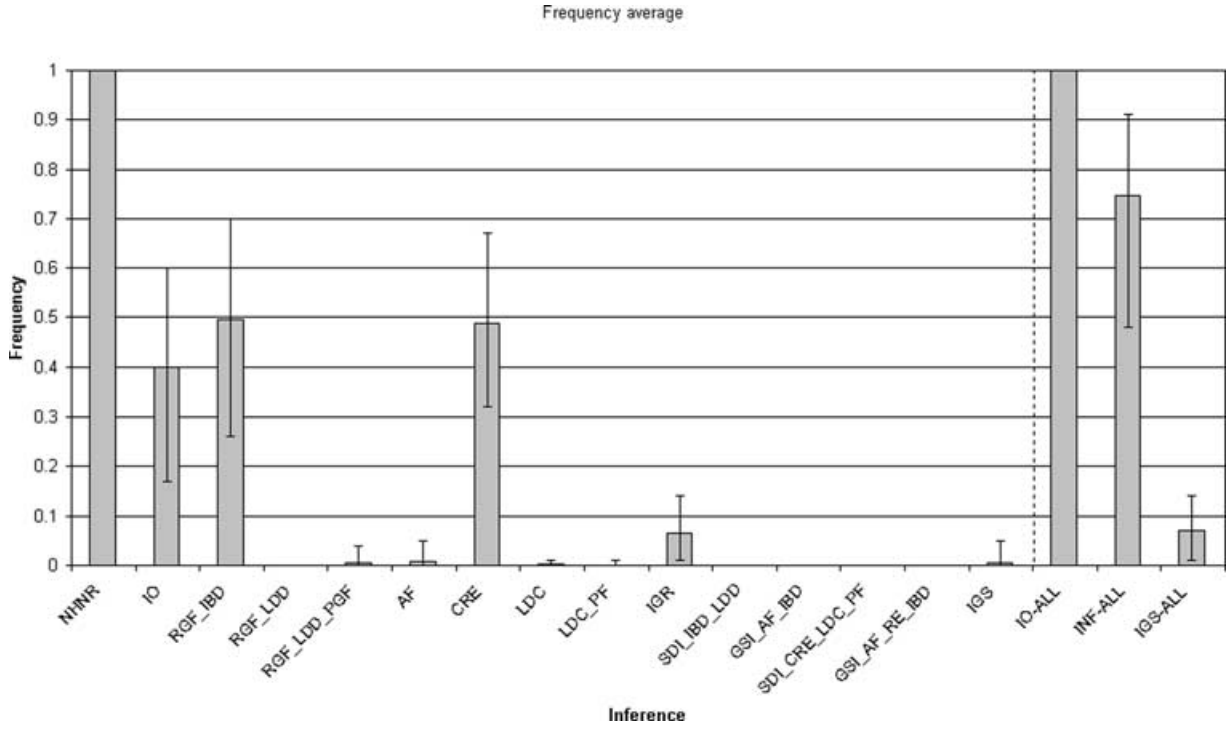

Figure 4. This figure shows the probabilities of obtaining an inference under panmixia in a dataset. The inference codes are the same as those in Figure 3.

Table 1. Model comparisons showing both the AIC and χ^2 values. The models are given in the format independent variable ~ dependent variables. S, a binary variable indicating if a clade has at least one significant statistic; *N*, the number of distance statistics; CL, the level (0 if at the total cladogram level); SL, the number of locations; PS, the proportion sampled; SS, the sample size; M, the method of calculating the distance statistics. Each χ^2 result is a comparison to the model in the first row.

Model	AIC	χ^2 *P*-value
S~N+CL+SL+PS	31912	
S~N+CL+SL+PS+SS	31914	0.438
S~N+CL+SL+PS+M	31914	1.000
S~CL+SL+PS	34435	<0.001
S~N+SL+PS	32011	<0.001
S~N+CL+PS	31985	<0.001
S~N+CL+SL	31916	0.019

found. The sample size and the method of calculating the distance statistics were found not to be important (Table 1).

We also applied to each dataset the randomization test in the AMOVA method of Arlequin version 3.01 (Excoffier et al. 2005). Across all simulated datasets (1800), marginal to the parameters, we found a false-positive rate of 5.34% in the direction of having too high an F_{ST} value at the (one tailed) 5% level. The mean of the estimates of F_{ST} was -0.000123 and not significantly different from zero. Of the 5.34% of datasets that had a significant AMOVA *P*-value, we found that 92.78% of these had at least one inference made by NCPA. Of the datasets that did not have a significant AMOVA *P*-value, 86.44% had at least one inference made by NCPA. The proportions reported are from NCPA inferences using geographic coordinates as input. Similar values are found when using a geographic distance matrix as input. A χ^2-test on the difference in proportions given above is not significant ($P = 0.101$). However, if we model the dataset using a binomial logistic regression we find a significant relationship between whether an NCPA inference was made in a dataset and the AMOVA *P*-values ($P = 0.00654$ and $P = 0.00223$ using the geographic coordinates and geographic distance matrix, respectively). The probability of at least one inference in NCPA appears to vary weakly from around 0.85 at AMOVA *P*-values close to 1 to around 0.93 at AMOVA *P*-values close to 0, which is why the difference is not detected in the χ^2-test.

Survey of Inferences Obtained in Published Papers

It is possible that the relatively high false-positive rate has had some influence on published results. To test this hypothesis we carried out a survey of the inferences obtained by researchers to determine whether they resembled results obtained under panmixia. We chose articles published between 2000 and 2004 that obtained at least one inference, and supplied the chain of inference (the questions followed to obtain an inference) for each clade. To make the results comparable to those obtained from the simulations, we needed to take into account that most of these papers used a key prior to the 2004 key (although the keys are still quite similar to each other). The chain of inference published in these papers was used to determine which inference would be obtained under the (later) 2004 key. Modifications of the questions were taken into account such that if a modification altered the chain of inference, the chain of inference followed would be that presumed to have been followed in the original chain. For example, Question 14, under the 2001 inference key asks if the species is absent in the nonsampled areas. Under the 2004 inference key, it asks if the species is present in the nonsampled areas. This would result in a change in the chain of inference from ". . . 14 YES" to ". . . 14 NO" and similarly the other way around. This meant that chains of inference under the 2001 key that ended ". . . 14 YES" would still lead to an inference of long-distance colonization under the 2004 key.

The survey comprised of 68 articles in which there were a total of 71 datasets analyzed by NCPA (some articles contained two datasets). The same inference categories as those in the panmictic simulation were used, except for NHNR (null hypothesis of no geographic association was not rejected), which was excluded because these were not always recorded in an article. Both clade level and dataset level frequencies were obtained. The clade level and dataset level frequencies obtained under panmixia were also adjusted to account for the removal of the NHNR inference. Figures 5 and 6 show box plots of the range of frequencies of each inference under the panmictic model and the corresponding frequency found in the survey, at both the clade level and dataset level respectively. Table 2 shows the results of analyses of the correlation between the median frequencies of inferences made in the simulated data and those found in the published data. The Pearson's product-moment correlation is around 0.8 for both clade level and dataset level frequencies, with *P*-values < 0.005. However, two points have a high influence on the correlation, and Spearman's rank correlation is around 0.5 for both levels, with P-values ~ 0.05. These two points correspond to restricted gene flow with isolation-by-distance and contiguous range expansion, which are the most common inferences made both in the simulated and published datasets. Potential explanations for this observation are: (1) the empirical data are drawn from panmictic populations; (2) the empirical data do not come from panmictic populations, but have some other demographic history, and the NCPA procedure has an inherent tendency to infer restricted gene flow with isolation-by-distance and contiguous range expansion; (3) the demographic processes are correctly identified and it is merely coincidence that the two most commonly inferred

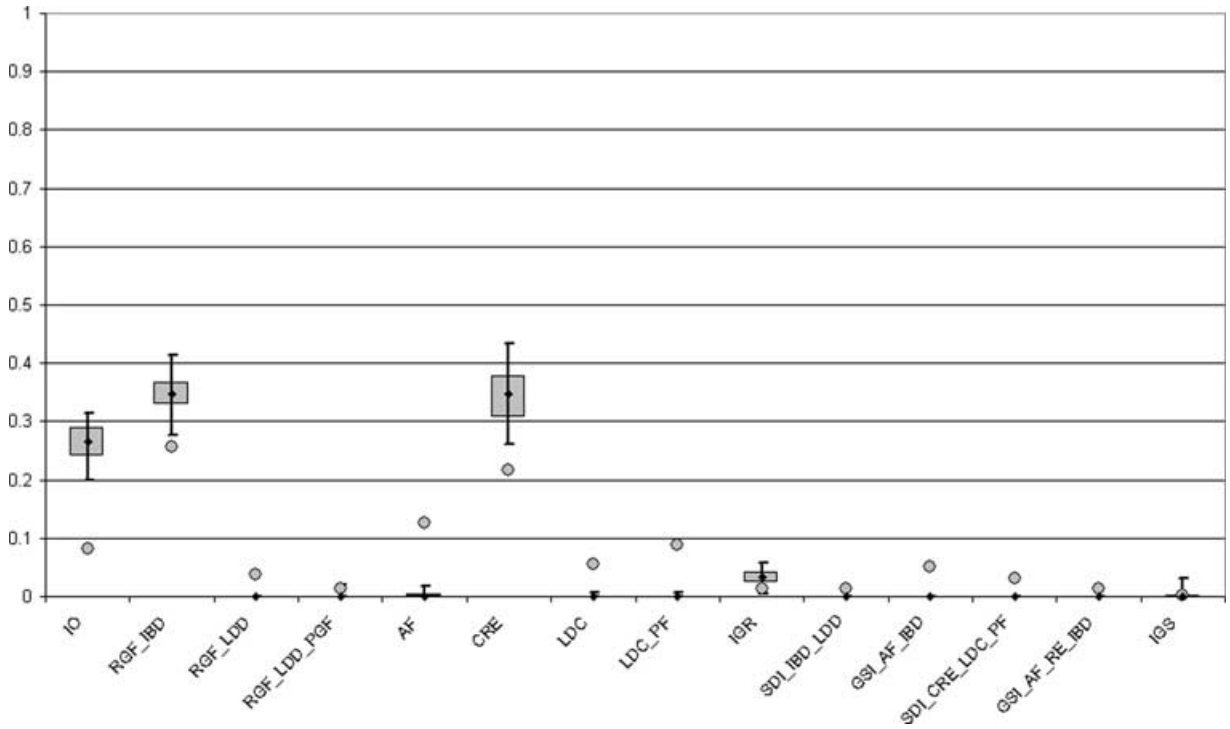

Figure 5. The frequency of inferences of the survey data (gray dots) compared to the frequencies of inferences under the panmictic model at the clade level. The box plots show the range of frequencies obtained under the panmictic model in addition to the median, and the first and third quartiles.

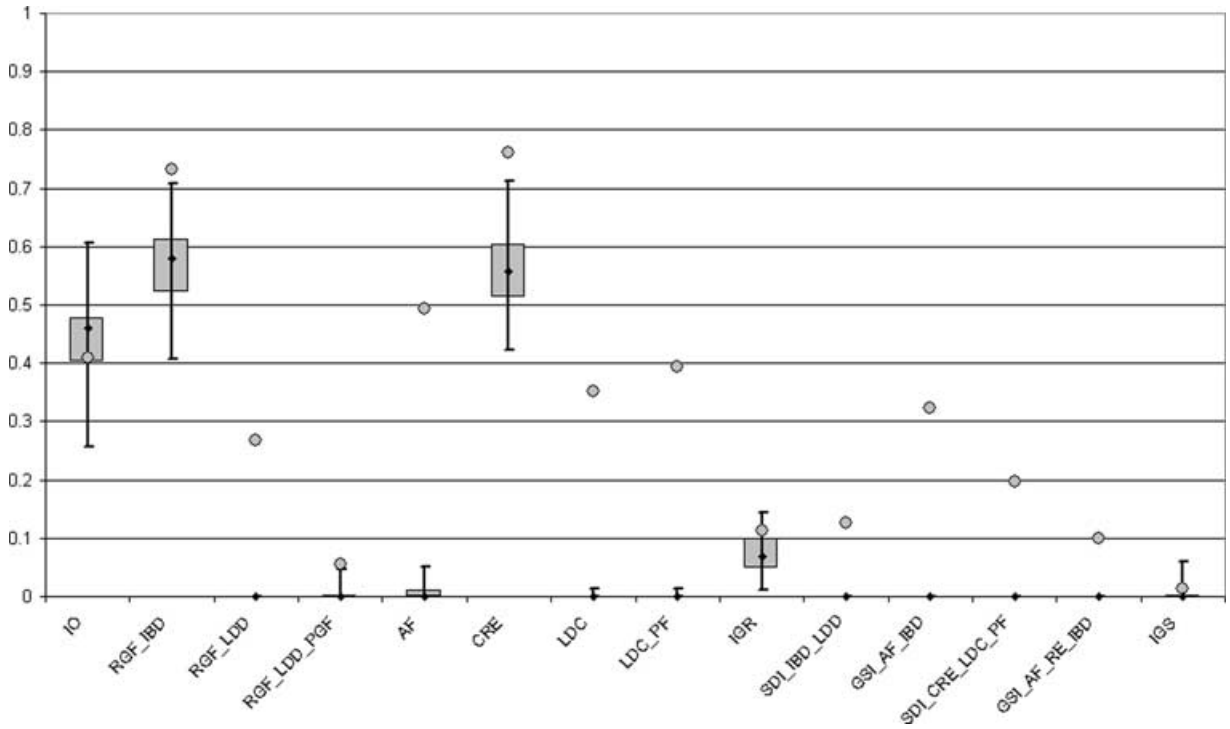

Figure 6. The frequency of inferences of the survey data (gray dots) compared to the frequencies of inferences under the panmictic model at the dataset level. The box plots show the range of frequencies obtained under the panmictic model in addition to the median, and the first and third quartiles.

Table 2. The correlation between the median frequency of the panmictic data and the frequency of the survey data, using both Pearson and Spearman correlation tests.

Method	Correlation coeff.	*P*-value
Clade level		
Pearson	0.826	0.0003
Spearman	0.531	0.0506
Dataset level		
Pearson	0.764	0.0015
Spearman	0.545	0.0438

processes in the empirical data are the same as those in the datasets simulated under panmixia. Possibility (1) is unlikely given that the datasets were collected from geographically structured populations, and furthermore a richer set of inferences is typically observed than is present in the simulations. More detailed simulations from structured populations will be necessary to distinguish between the other two explanations.

Discussion

As noted by Posada et al. (2006), NCPA has been widely used. Surprisingly, for such a widely used technique, little is known, objectively, of its basic properties. The aim of this study has been to provide an automated method whereby inferences can be drawn using the NCPA technique, based on a transparent algorithm, the details of which can be debated in an objective way.

AREAS OF UNCERTAINTY IN IMPLEMENTING NCPA

A current limitation of the automated procedure is that it does not preprocess haplotype networks prior to construction of the nested cladograms. The necessity for preprocessing reflects the difficulty in recovering a unique haplotype tree from the data. This uncertainty is represented as loops in the network, and can also give rise to subnetworks that are not joined together. It would appear that there is variability in the literature concerning the decisions taken on how to deal with the effects of homoplasy, and this aspect may be very difficult to capture in any computer program. However, the current version of the automated procedure lends itself to one straightforward way of dealing with uncertainty in tree topology: replication over the distribution of topologies consistent with the data. The replication could be over equally parsimonious trees, bootstrapped samples, or posterior distributions. A distribution of inferences could then be obtained that would reflect uncertainty in tree topology. This is similar to the approach suggested by Templeton for multiple loci (Templeton 2002b). The automation also provides a concrete method against which to test improvements: for example, testing the effects of applying the criteria based on coalescent theory, suggested by Crandall and Templeton (1993) and Pfenninger and Posada (2002), and the potential sensitivity of NCPA to variation in the sample size (Petit and Grivet 2002).

In comparison with the scope for variability in the haplotype networks from which nested cladograms are constructed, there appears to be more consistency in the literature in how nesting is performed and in the application of the inference key. Although there are some conflicting suggestions for how to perform the nesting (e.g., Templeton et al. 1987 vs. Templeton 2002b, as discussed in the online supplementary material), and despite the potential for different interpretations of the questions in the inference key, in practice, there appears to be relatively little variation in how NCPA is applied. This is evident in the answers to a survey that we performed in which we elicited responses from authors who had previously used NCPA, although it should be noted that the sample size is relatively low. The details of the survey are given in the supplementary material.

RECOVERY OF INFERENCES FROM PUBLISHED DATASETS

We applied our automated method to published datasets, but could not recover exactly the inferences that were published. In one out of the three datasets (the leaf beetle *G. pallida*) the haplotype network had been preprocessed before construction of the nested cladogram. In this case we used the same haplotype network as that used in the published analysis. For all three datasets we then largely recovered at least the lower levels of the nested cladograms given in the original publications. In two out of the three datasets (*G. pallida*, and Iberian lizards) the results obtained under the inference key either matched, or, if different, the differences could be reasonably explained. A number of the differences arose from changes in the inference key. Only in the study of Gómez-Zurita and Vogler (2003) was it impossible to provide an explanation for the discrepancies in outcome. Even in this case, however, the same Geodis statistics were obtained. Given that no such discrepancies were obtained with two other publications this illustrates how difficult it may be to construct an automated method that will entirely replicate published results.

Small differences in nesting the clades can result in substantial differences in the structure of the entire cladogram, as illustrated by the analysis of the data of Gómez-Zurita and Vogler (2003), in which there is a perfect match in nesting design at the early levels, and then a small discrepancy results in rather different nesting design at higher levels. These datasets also highlight various factors that can affect NCPA, such as the assumption of exhaustive sampling (also highlighted by Templeton 2004), the usage of criteria to resolve loops, and the effect of rooting the tree. Ideally, these factors in NCPA need to be further explored.

SIMULATION STUDY

A primary objective in the automation of the NCPA procedure has been to test it with simulated data. We chose to simulate data with a relatively low level of homoplasy, and therefore it is likely that the automated procedure will give rise to very similar results to those obtained in a manual analysis. The results of the simulations highlight a structural feature of NCPA that requires some attention: the difficulty in attaching a measure of confidence (via, e.g., a frequentist *P*-value) to the inferences obtained (a point first made by Knowles and Maddison 2002).

Although a particular critical *P*-value is used for deciding whether to draw inferences about a clade using the key, this does not appear to correspond to the frequency of false positives under the null model of no geographic association. Figure 3 shows that, using the 5% level of significance in Geodis, in any given clade there is a 14% chance on average to obtain an inference (INF-ALL). On the dataset level however, there is a good chance that at least one of the summary statistics will be significant, and hence a verbal statement can be made about the demographic history of the study organism. This point has been noted in Pulquério (2005) where NCPA was applied to simulated datasets under a number of different demographic scenarios. This study (also described in M.G.F. Pulquério, M. Panchal, O.S. Paulo, and M.A. Beaumont, unpubl. ms.) used an earlier version of the software ANeCA, in which the consultation of the inference key was not automated. Our results suggest that there is on average a 75% chance of obtaining an inference (INF-ALL) in at least one clade (Fig. 4). The most common inference will be either restricted gene flow with isolation by distance, or contiguous range expansion. For data without genetic structuring, it would appear that at both the clade and dataset levels there is very little chance to obtain another inference, such as allopatric fragmentation, or long-distance colonization.

It is interesting to note that the two commonest outcomes from the simulated datasets, restricted gene flow with isolation by distance, and contiguous range expansion, are also the commonest outcomes in the published datasets. It is tempting to suggest there is an inherent tendency for NCPA to reach this conclusion. For example, even in datasets that may have geographical structuring, some clades may have similar distribution patterns to those obtained in the panmictic simulations, and therefore have a tendency to give rise to statistically significant distance statistics, leading to the two most commonly observed outcomes from the simulations. However, other inferences are found at much higher frequencies in the published datasets in comparison with those in the simulations, and published results are likely to contain a mixture of false positives and the effects of genuine genetic substructuring. Further investigation is required to determine how reliable these particular inferences are, and whether we can replicate the frequencies of inferences found in publications through a single scenario, or mixture of scenarios (M.G.F. Pulquério, M. Panchal, O.S. Paulo, and M.A. Beaumont, unpubl. ms.).

On a per-clade basis, the simulations also show that the number of statistics associated with the clade, the level of the clade, the number of locations, and the proportion of locations sampled all had an effect on inferences made. Interestingly sample size (the total across demes) appeared unimportant (Table 1). We have, however, chosen to study quite large sample sizes, and it is possible that smaller sample sizes might have a larger effect. Overall, the simulations suggest that a complicated mix of factors affects the false-positive rate, and therefore this problem may be difficult to fix.

COMPARISON WITH RECENT FINDINGS

The results from this study are directly contradictory to the findings of Templeton (2004), which finds NCPA conservative and not prone to false positives. However, there are many reasons that could cause differences between the results. One such cause could be the difference in user-executed versus automated NCPA. For example, Templeton (2004) applies the coalescent criteria in Crandall and Templeton (1993), commonly used in the literature, and explores the remaining trees, whereas our automated software uses the original nesting rules to deal with ambiguity (Templeton and Sing 1993; Templeton et al. 1995; Templeton 2005a). Exploring the effects of using one or a combination of these criteria in NCPA is an important factor that needs to be investigated. A second possibility is that these discrepancies reflect a qualitative difference between real data and simulated data. When simulating data there are many assumptions that are made, which may be unintentionally restrictive. The results reported here involve a large number of different scenarios: different numbers of populations, different sampling proportions, different sample sizes, different ways of calculating distance statistics, the effect of conditioning on particular genealogies. However, further simulation testing under other scenarios may be desirable, and the automated software will make this straightforward. A third possibility is that the simulation-based study reported here has so far only been concerned with panmixia, whereas those analyzed by Templeton (2004) contained strong evidence of fragmentation and range expansion. With the automated software it is possible to analyze data under a wide range of demographic histories, and this has been carried out (M.G.F. Pulquério, M. Panchal, O.S. Paulo, and M.A. Beaumont, unpubl. ms.).

Conclusions

This paper describes the first attempt at full automation of the NCPA procedure introduced by Templeton et al. (1995). The automated procedure that we have developed can replicate many, but not all, of the results in three published datasets. We have

identified that there is variation in how the method has been applied by earlier authors: in the manipulation of haplotype networks prior to construction of nested clade; in the nesting of clades; and in the interpretation of the inference key. In simulations of geographically distributed, but genetically unstructured, populations we have noted that it is difficult to make an a priori calculation of the false-positive rate, which is affected by many factors. It is quite likely that the method will lead to at least one positive inference for any dataset. The most likely false positives are inferences of restricted gene flow with isolation by distance, and contiguous range expansion. A survey of published data indicates that these are the most frequently found inferences. The software ANeCA also makes it possible for users to use their own real dataset as a basis for simulations under different models (including panmixia) to estimate their empirical false-positive rate.

ACKNOWLEDGMENTS

We would like to thank M. Pulquério, O. S. Paulo, A. R. Templeton, K. A. Crandall, D. Posada, J. Gómez-Zurita, S. Creer, I. Carbone, P. Mardulyn, L. L. Knowles, and the anonymous participants in the questionnaire regarding the use of NCPA for their discussions on the methodology and application of NCPA. We would also like to thank M. Pulquério, O. Moya, I. Phillipsen, S. O'Loughlin, J. Gómez-Zurita, and G. Segelbacher, for testing and providing feedback regarding the software, as well as P. Mardulyn, J. Gómez-Zurita and O. Paulo for providing datasets to test the automated NCPA. The Biotechnology and Biological Sciences Research Council is also gratefully acknowledged for funding to MP and the Natural Environment Research Council for funding to MAB. We are also grateful for the comments and suggestions provided by P. Sunnucks, O. Moya, and three anonymous reviewers on earlier versions of the manuscript.

Subsequent to the publication of a program AUTOINFER version 1.0 (Zhang et al. 2006), Ai-bing Zhang has noted that there are some issues in the program that require resolution and it has temporarily been withdrawn. Thus unfortunately we have been unable to consider that program here.

LITERATURE CITED

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography—the mitochondrial-DNA bridge between population-genetics and systematics. Annu. Rev. Ecol. Syst. 18:489–522.

Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37–48.

Bardel, C., V. Danjean, J. P. Hugot, P. Darlu, and E. Genin. 2005. On the use of haplotype phylogeny to detect disease susceptibility loci. BMC Genet. 6:24–36.

Beerli, P. 2006. Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics 22:341–345.

Brisson, J. A., D. C. de Toni, I. Duncan, and A. R. Templeton. 2005. Abdominal pigmentation Variation in *Drosophila* Polymorpha: geographic variation in the trait, and underlying phylogeography. Evolution 59:1046–1059.

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cassens, I., P. Mardulyn, and M. C. Milinkovitch. 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach?. Syst. Biol 54:363–372.

Castelloe, J., and A. R. Templeton. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. 3:102–113.

Choisy, M., P. Franck, and J.-M. Cornuet. 2004. Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. Mol. Ecol 13:955–968.

Clement, M., D. Posada, and K. A. Crandall. 2000. TCS: a computer program to estimate gene genealogies. Mol. Ecol. 9:1657–1659.

Crandall, K. A. 1996. Multiple interspecies transmissions of human and simian T-cell leukaemia/lymphoma virus type I sequences. Mol. Biol. Evol. 13:115–131

Crandall, K. A., and A. R. Templeton. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics 134:959–969

Durrant, C., K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas, and A. P. Morris. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. 75:35–43.

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

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin (ver. 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47–50.

Fetzner, J. W., and K. A. Crandall. 2003. Linear habitats and the nested clade analysis: an empirical evaluation of geographic versus river distances using an Ozark crayfish (Decapoda: Cambaridae). Evolution 57:2101–2118.

Gómez-Zurita, J., and A. P. Vogler. 2003. Incongruent nuclear and mitochondrial phylogeographic patterns in the *Timarcha goettingensis* species complex (Coleoptera, Chrysomelidae). J. Evol. Biol. 16:833–843.

Gómez-Zurita, J., C. Juan, and E. Petitpierre. 2000a. The evolutionary history of the genus *Timarcha* (Coleoptera, Chrysomelidae) inferred from mitochondrial COII gene and partial 16S rDNA sequences. Mol. Phylogenet. Evol. 14:304–317.

———. 2000b. Sequence, secondary structure and phylogenetic analyses of the ribosomal internal transcribed spacer 2 (its2) in the *Timarcha* leaf beetles (Coleoptera: Chrysomelidae). Insect Mol. Biol. 9:591–604.

Hey, J., and C. A. Machado. 2003. The study of structured populations - new hope for a difficult and divided science. Nature Rev. Genet. 4:535–543.

Knowles, L. L., and W. P. Maddison. 2002. Statistical phylogeography. Mol. Ecol 11:2623–2635.

Mardulyn, P. 2001. Phylogeography of the Vosges Mountains populations of *Gonioctena pallida* (Coleoptera: Chrysomelidae): a nested clade analysis of mitochondrial DNA haplotypes. Mol. Ecol. 10:1751–1763.

Masta, S. E., N. M. Laurent, and E. J. Routman. 2003. Population genetic structure of the toad *Bufo woodhousii*: an empirical assessment of the effects of haplotype extinction on nested cladistic analysis. Mol. Ecol. 12:1541–1554.

Panchal, M. 2007. The automation of Nested Clade Phylogeographic Analysis. Bioinformatics 23:509–510.

Paulo, O. S., W. C. Jordan, M. W. Bruford, and R. A. Nichols. 2002. Using nested clade analysis to assess the history of colonization and the persistence of populations of an Iberian lizard. Mol. Ecol. 11:809–819.

Pesole, G., C. Gissi, A. De Chirico, and C. Saccone. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. J. Mol. Evol. 48:427–434.

Petit, R. J., and D. Grivet. 2002. Optimal randomization strategies when testing the existence of a phylogeographic structure. Genetics 161:469–471.

Pfenninger, M., and D. Posada. 2002. Phylogeographic history of the land snail *Candidula unifasciata* (Helicellinae, Stylommatophora): fragmentation, corridor migration, and secondary contact. Evolution 56:1776–1788.

Posada, D., and K. A. Crandall. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. 16:37–45.

Posada, D., K. A. Crandall, and A. R. Templeton. 2000. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Mol. Ecol. 9:487–488.

Posada, D., T. J. Maxwell, and A. R. Templeton. 2005. TreeScan: a bioinformatic application to search for genotype/phenotype associations using haplotype trees. Bioinformatics 21:2130–2132.

Posada, D., K. A. Crandall, and A. R. Templeton. 2006. Nested clade analysis statistics. Mol. Ecol. Notes 6:590–593.

Pulquério, M. J. F. 2005. Evaluation of nested clade phylogeographical analysis using simulated DNA sequence data with different population structures and histories. Master's thesis, Universidade de Lisboa, Lisbon, Portugal.

Roff, D. A., and P. Bentzen. 1989. The statistical-analysis of mitochondrial-DNA polymorphisms - χ^2 and the problem of small samples. Mol. Biol. Evol. 6:539–545.

Seltman, H., K. Roeder, and B. Devlin. 2001. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am. J. Hum. Gen. 68:1250–1263.

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

———. 1999. Uses of evolutionary theory in the human genome project. Annu. Rev. Ecol. Syst. 30:23–49.

———. 2002a. "Optimal" randomization strategies when testing the existence of a phylogeographic structure: a reply to Petit and Grivet. Genetics 161:473–475.

———. 2002b. Out of Africa again and again. Nature 416:45–51.

———. 2004. Statistical phylogeography: methods of evaluating and minimizing inference errors. Mol. Ecol. 13:789–809.

———. 2005a. Haplotype trees and modern human origins. Yrbk. Phys. Anthro 48:33–59.

———. 2005b. Population biology and population genetics of Pleistocene hominins. *in* W. Henke, H. Rothe, and I. Tattersall, eds. Handbook of paleoanthropology. Springer, Berlin.

Templeton, A. R., and C. F. Sing. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping .4. Nested analyses with cladogram uncertainty and recombination. Genetics 134:659–669.

Templeton, A. R., E. Boerwinkle, and C. F. Sing. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping .1. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351.

Templeton, A. R., C. F. Sing, A. Kessling, and S. Humphries. 1988. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping .2. The analysis of natural populations. Genetics 120:1145–1154.

Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data .3. Cladogram estimation. Genetics 132:619–633.

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.

Templeton, A. R., T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle, and C. F. Sing. 2005. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. Genetics 169:441–453.

Turner, T. F., J. C. Trexler, J. L. Harris, and J. L. Haynes. 2000. Nested cladistic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel. Genetics 154:777–785.

Wang, J., and M. C. Whitlock. 2003. Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163:429–446.

Westfall, P. H., and S. S. Young. 1993. Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley series in probability and mathematical statistics. Wiley, New York.

Zhang, A.-B., K. Kubota, Y. Takami, J. L. Kim, J. K. Kim, and T. Sota. 2005. Species status and phylogeography of two closely related *Coptolabrus* species (Coleoptera: Carabidae) in South Korea inferred from mitochondrial and nuclear gene sequences. Mol. Ecol. 14:3823–3841.

Zhang, A.-B., S. Tan, and T. Sota. 2006. AUTOINFER 1.0: a computer program to infer biogeographical events automatically. Mol. Ecol. Notes 6:597–599.

Associate Editor: P. Sunnucks

Supplementary Material

The following supplementary material is available for this article:

Appendix S1, Figures S1-S3, Tables S1-S10.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1558-5646.2007.00124.x
(This link will take you to the article abstract).

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