doi:10.1111/j.1558-5646.2008.00481.x

WHY DOES A METHOD THAT FAILS CONTINUE TO BE USED?

L. Lacey Knowles[1,2]

[1]*Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109*

[2]*E-mail: knowlesl@umich.edu*

Received July 17, 2008
Accepted July 18, 2008

As a critical framework for addressing a diversity of evolutionary and ecological questions, any method that provides accurate and detailed phylogeographic inference would be embraced. What is difficult to understand is the continued use of a method that not only fails, but also has never been shown to work—nested clade analysis is applied widely even though the conditions under which the method will provide reliable results have not yet been demonstrated. This contradiction between performance and popularity is even more perplexing given the recent methodological and computational advances for making historical inferences, which include estimating population genetic parameters and testing different biogeographic scenarios. Here I briefly review the history of criticisms and rebuttals that focus specifically on the high rate of incorrect phylogeographic inference of nested-clade analysis, with the goal of understanding what drives its unfettered popularity. In this case, the appeal of what nested-clade analysis claims to do—not what the method actually achieves—appears to explain its paradoxical status as a favorite method that fails. What a method promises, as opposed to how it performs, must be considered separately when evaluating whether the method represents a valuable tool for historical inference.

KEY WORDS: **Biogeography, demographic history, historical inference, nested-clade analysis, phylogeography.**

From its humble introduction in 1995, where the authors cautioned a need for thorough investigation, nested clade phylogeographical analysis (NCPA, formerly known as NCA) has had an amazing rise, especially given that this cautionary advice was never heeded. With over 1700 citations to date (see Petit 2008), there is no debate that this approach has had an enormous impact on the field of phylogeography—at issue, is what will be NCPA's legacy?

Two recent commentaries (Garrick et al. 2008; Petit 2008) deliver two very different verdicts. Both commentaries were motivated by the same simulation study that documents a disturbingly high error rate of NCPA: incorrect historical processes were inferred in over 75% of the simulated datasets (Panchal and Beaumont 2007). Neither commentary disputes this finding. Yet, comparison of these two commentaries reveals that NCPA is apparently either (1) a potentially flawed method that should be abandoned until further testing, or (2) an unique approach with no substitute that should inspire "practical" approaches to validate or strengthen its inferences.

Such paradox is not limited to the diametrically opposed interpretations (i.e., Petit 2008 versus Garrick et al. 2008) of the extremely high false inference rate documented by Panchal and Beaumont (2007). Such contradiction is also mirrored by the communities' response to an earlier simulation study (Knowles and Maddison 2002). This study also found that NCPA made incorrect inferences about the population history in over 75% of the datasets. The high failure rate certainly drew the attention of NCPA users; Knowles and Maddison (2002) now stands at over 210 citations. However, examination of the studies citing Knowles and Maddison ironically reveals the unfaltering popularity of NCPA—the majority of empirical studies that referenced Knowles and Maddison, nevertheless still used NCPA (Fig. 1).

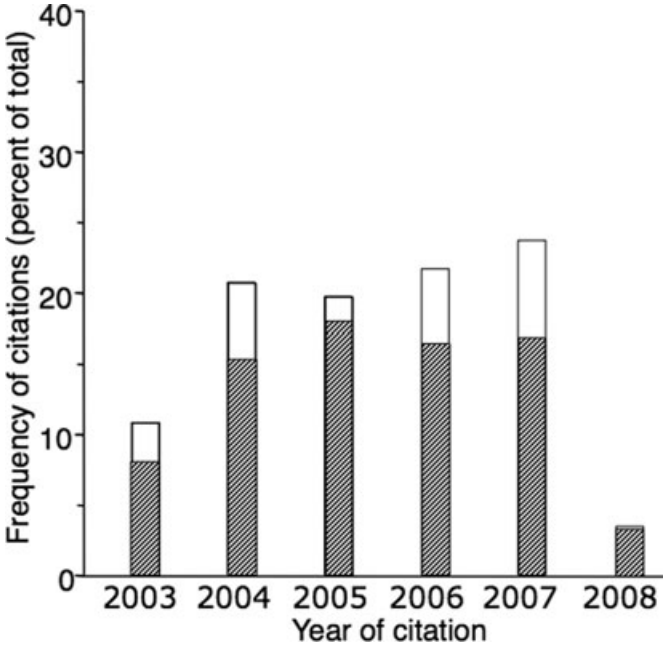

Figure 1. Comparison of the proportion of citations for a computer simulation study that identified a high rate of incorrect inference with NCPA (i.e., Knowles and Maddison 2002, shown in white) that nevertheless still used NCPA to analyze their empirical data (as shown by the hatched bars).

So the question is: will the failings of a favorite method continue to go unheeded? Second, why would error rates of over 75%, now documented in two separate studies conducted by independent laboratories and for two different population histories (i.e., allopatric population divergence and an unstructured population; Knowles and Maddison 2002; Panchal and Beaumont 2007, respectively), as well as other concerns over the validity of NCPA's inferences (e.g., Knowles and Maddison 2002; Petit and Grivet 2002; Felsenstein 2004; Panchal and Beaumont 2007) be dismissed? Here I review the evidence used to support the contrasting opinions about NCPA to gain insight into the possible answers to such questions.

Will an Error Rate of 75% Be the Coup de Grâce for NCPA?

Perhaps the appeal of what NCPA claims to accomplish is simply too much of an allure for empiricists to abandon the approach, especially because there is no single methodological substitute (Garrick et al. 2008; except, researchers have a huge variety of methodologies to choose from, for example see Excoffier and Heckel 2006). Without a way to evaluate how much of the method's popularity is driven by what it purports to accomplish, let us consider the rebuttals put forth by the creator of NCPA, and how persuasive they might be in convincing a would-be user (Templeton 2004, 2008): (1) NCPA is an extensively validated method for statistical phylogeographic inference, and (2) NCPA cannot be adequately tested with simple evolutionary scenarios.

How can a method be both extensively validated (e.g., Templeton 2008) and have error rates of over 75% (e.g., Knowles and Maddison 2002; Panchal and Beaumont 2007)? To accept the first argument, you would have to agree with the second point as well, which is at the crux of Templeton's dismissal of the use of simulation to rigorously test NCPA. Below the validity for these claims are discussed in detail.

ACCURATE AND EXTENSIVELY VALIDATED?

The support for this claim comes from a compilation of empirical datasets representing "positive controls," in which the results from NCPA were compared to each author's a priori expectations about what processes might have generated the data (i.e., range expansion or fragmentation events). Based on this validation procedure, NCPA appears to have done a fair job of inferring the process that matched the original author's expectations (Templeton 2004), identifying population fragmentation in 30 of 34 cases, and range expansions in 34 of the 55 cases, where it was predicted (i.e., a "true" positive rate of 88% and about 62%, respectively).

What was not included in this tabulation/validation procedure was how many times processes other than those that were expected were also inferred, which is the most salient result of the simulation studies—NCPA repeatedly infers processes when no such events have occurred (Knowles and Maddison 2002; Panchal and Beaumont 2007). The argument that NCPA has been extensively tested and shown to be accurate (Templeton 2004, 2008) is based on blatantly confusing type I and type II errors (Sokal and Rohlf 1995). The simulation studies both clearly show that NCPA incorrectly identifies significant geographic associations at a disturbingly high rate, which leads to inferences about process that never occurred. This finding cannot be rebuffed by the argument that NCPA has a high rate of detecting an expected fragmentation or range expansion (i.e., a high rate of true positives) and a low rate of failing to detect an expected fragmentation or range expansion (i.e., a low rate of false negatives)(see Templeton 2004). This logic is fundamentally flawed. A method can have a high false positive rate (as detected with the simulations and remains untested by reference to empirical data) even if it has a high true positive rate and low rate of false negatives. In fact, in almost every one of the compiled empirical datasets used to validate NCPA's accuracy (Templeton 2004), a process other than the expected range expansion or fragmentation was also inferred (which is consistent with the high false positive rate documented with simulated data; Knowles and Maddison 2002; Panchal and Beaumont 2007). It is possible that these were also true events. But it is just as likely that they are indicative of an incredibly high false positive rate. Moreover, there is evidence to suggest that the empirical datasets

Templeton (2004) analyzed with NCPA do suffer from an high false positive rate. First, Panchal and Beaumont (2007) simulation results not only indicated a high rate of incorrect phylogeographic inference with NCPA, but also that the errors were biased toward detecting isolation by distance (IBD) when there was no such structure. IBD is the very process repeatedly inferred, but was not expected a priori, in Templeton's own analysis of the compiled empirical datasets intended to validate NCPA's accuracy (see appendix 1 of Templeton 2004). Moreover, IBD was not just the most frequently inferred process in studies published between 2000 and 2004 that used NCPA, but the frequency that IBD was inferred in analyses of empirical data actually matched the rate at which IBD was incorrectly inferred in simulated datasets with no actual IBD (i.e., significant and strong correlations; see table 2, Panchal and Beaumont 2007).

What about other possible explanations for the apparent consistency between the NCPA analyses of empirical data (Templeton 2004; Panchal and Beaumont 2007) and simulated data (Knowles and Maddison 2002; Panchal and Beaumont 2007), which show an apparently high rate of inferring processes that did not occur (or for which there was no a priori expectation, as with the analyses of the empirical data)? Perhaps it is just a coincidence. Maybe the most commonly inferred processes inferred in analyses of empirical data with NCPA just happen to also be the same incorrectly inferred processes when data simulated under panmixia are analyzed with NCPA. Similarly, just by chance, analyses of data simulated under two different historical scenarios (i.e., population panmixia and allopatric divergence) both found that NCPA had a high rate of incorrect phylogeographic inference. How should we decide?

The automation of NCPA (Panchal 2007) provides a framework in which the accuracy of each inference from NCPA could be tested in principle. There is no reason why simulations should not be used to demonstrate that these historical inferences should be trusted (and hence, acceptable for publication). Simulation approaches are not only accepted, but they have also been used to explore the accuracy of many proposed methods in phylogeography, genetics, and systematics. Moreover, even if someone still wishes to believe in the claim that NCPA provides a framework for accurate phylogeographic inferences, it is worth noting how accuracy is being defined in defense of NCPA. According to Templeton's analyses of the compiled empirical datasets with a "known" event (Templeton 2004), the error rate may be just a paltry 38%, in the case of failures to detect predicted range expansions, or a range expansion may be inferred when no such event was expected in just 23% of the datasets (Templeton 2008). I suppose you might find some solace in error rates of 38% and 23%, after all they are indeed lower than an error rate of 75%. However, these rates certainly should not inspire a lot of confidence in any inference derived from NCPA (or any method). Claims that criticisms of NCPA really are becoming "increasingly irrelevant"(Templeton 2008) are simply counterproductive.

ERRONEOUS ERROR RATES OF 75%?

What is the basis for claims that the simulation studies incorrectly identified high rates of incorrect phylogeographic inference? This assertion rests upon two basic tenets: (1) because NCPA is designed to infer multiple processes, NCPA cannot be tested using simulation of simple historical scenarios, and (2) unrealistic assumptions in the simulations and errors with the automated implementation of NCPA, not any failures of NCPA, is the cause of incorrect phylogeographic inference.

Yes, the scenarios considered in the simulations (to date) are simple and the histories of species can be complicated. However, if NCPA cannot correctly infer a population history for a simple evolutionary scenario, why should anyone have faith in its ability to infer a complicated evolutionary history (Knowles and Maddison 2002)? This is a straightforward and legitimate question that has yet to be answered. It is not complicated—there is no need to obfuscate the issue. For example, the question of why any method should be trusted for inferring complicated histories when it fails with simple evolutionary scenarios is not addressed by a diatribe on the relative merits of model fitting as opposed to hypothesis testing, issues of a priori versus a posterior interpretation, or debates over which method is best or better. Such discourse (see Templeton 2004, 2008) does not assuage concerns over the accuracy of NCPA as a method for phylogeographic inference (e.g., Knowles and Maddison 2002; Petit and Grivet 2002; Felsenstein 2004; Panchal and Beaumont 2007; Beaumont and Panchal 2008).

To demonstrate that the high rate of incorrect phylogeographic inference is not due to any failures of NCPA, rebukes of the simulation studies' findings have also been accompanied by either comparisons of the NCPA error rates attained for the simulated data to previously published data (Templeton 2008), or a reanalysis of the simulated datasets themselves (Templeton 2004). As with the incorrect assertions about the validation of NCPA (i.e., confusing type I and type II errors), there is no legitimacy to the argument that there must be a mistake with the analyses of the simulated data (see Templeton 2004, 2008).

First consider the flawed logic for discrediting the results of Panchal and Beaumont's simulation study. Projecting the error rate characterized by Panchal and Beaumont's analysis for a panmictic population onto human demographic empirical data, Templeton (2008) showed that there was a numerical disagreement in the error rates of NCPA when applied to the two different sets of data. What does this demonstrate? A numerical disagreement in the error rates of NCPA when applied to data simulated under panmixia (Panchal and Beaumont 2007) versus empirical data on the demographic history of humans out of Africa (Templeton 2008) is just that—a difference in error rates. Any

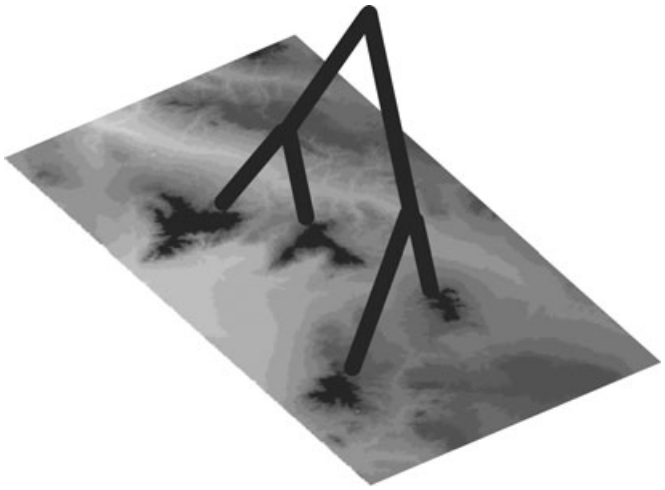

Figure 2. Graphical depiction of the history of allopatric divergence among the four populations used to simulate data, where the populations are restricted to specific habitats (as shown here in this example of sky islands from the desert southwest of North America, where suitable habitat is identified in black).

claims to the contrary (see Templeton 2008) are based on the implicit assumption that methods are expected to perform equally well (or poorly, as in this case) under all historical scenarios and across all parameter space (e.g., for different times of divergence, rates of migration, changes in population size, histories of vicariance). This rationale is obviously untenable for any method (including ones with failure rates as high as NCPA), even if the conditions under which NCPA will perform well are yet to be identified.

Similarly, Templeton's challenge to the conclusions from Knowles and Maddison's simulation study is also built upon faulty reasoning (Knowles and Maddison 2002). Emboldened by claims that the history used to simulate the data was too simple and artificial, Templeton reanalyzed the datasets, but assumed a new evolutionary history that differed from the one used to simulate the data. The data were simulated under a history of allopatric divergence with four populations (Fig. 2). For example, consider four montane populations of grasshoppers (just as a random example), which were founded from two different ancestral source populations (e.g., two different glacial refuges). Such populations are restricted to mountain-tops and are surrounded by inhospitable intervening habitat, as with other sky island plants and animals (McCormack et al. 2009). However, in Templeton's reanalysis of the data, he decided that there were intervening populations. In this reanalysis he showed that instead of incorrectly inferring phylogeographic processes, in most cases in which significant population structure was identified, the inference key indicated that there was inadequate sampling. Again, what does this really demonstrate? Showing that you can get different results if you assume a new population history is just that—analyzing data with NCPA under a history that is inconsistent with the model used to generate the data will not result in the same phylogeographic inferences. Any claims that this somehow invalidates the high rate of false inference documented by Knowles and Maddison (see Templeton 2004) is indefensible. Such an argument implicitly assumes that the results of a methodological analysis are robust to varying the conditions under which the data might have evolved. This may certainly be a desirable attribute of a method under certain conditions. For example, a phylogenetic method may be insensitive to using a model that departs from the actual model of nucleotide evolution—that is, the estimate of the phylogeographic relationships may be accurate despite a mismatch between the actual and assumed model of nucleotide evolution used to infer the species relationships. However, demonstrating that the method is sensitive (i.e., gives different estimates of species relationships) when the model used to infer the phylogeny does not actually match the one used to generate the data certainly is not evidence that the species relationships inferred under a model matching the actual history of species divergence is untrustworthy. Likewise, it would be unjustified to conclude that the high rate of incorrect phylogeographic inference documented by Knowles and Maddison (2002) was mistaken because a reanalysis of the data showed that conclusions drawn from NCPA are sensitive to departures from the model of evolution under which the data are simulated.

Is An Error Rate of 75% a Premature Obituary?

Despite the appeal of a method that promises to do what is impossible with any single alternative method (i.e., identify jointly the relevant processes of IBD, past fragmentation, population expansions . . . that characterize a species' history), there is no evidence to support the contention that NCPA can deliver on this lofty goal. With (1) a rate of 75% of false inference by NCPA, and (2) the alarming correlation between the most frequently inferred processes in analyses of empirical data and those incorrectly inferred in data simulated under panmixia (Panchal and Beaumont 2007), all the available data seem to indicate that any conclusions based on NCPA should be met with significant skepticism. So why would anyone use NCPA or advocate its use (Garrick et al. 2008)? Why not follow the cautionary advise that any would-be-users should await further evaluation of NCPA (Petit 2008)?

Is the lack of a substitute for NCPA (i.e., a method that can estimate multiple historical and demographic processes) a tenable argument for its continued use (Garrick et al. 2008)? Species do have complicated histories. This is noncontroversial. Clearly we need methods that can accommodate this biological reality, and it represents a laudable ideal that we can strive toward. However, the absence of an alterative approach that attempts to simultaneously

infer all the historical processes attempted by NCPA is not a reason to overlook the method's failures.

What a method promises, as opposed to how it performs, must be considered separately when evaluating whether the method represents a valuable tool to the phylogeographic community. As noted above, the conditions when NCPA is likely to provide reliable results have not yet been demonstrated. It is therefore difficult to understand the rational for using an unreliable method for generating plausible hypotheses, let alone as the sole basis for historical inference. Garrick et al. (2008) imply that any judgment about the potential utility of NCPA based on simulated data might be too harsh because the simulations relied on analysis of single locus data, when the importance of analyzing multiple independent loci is already widely recognized. Is the high error rate really an overestimate because it was unfair to ask NCPA to make historical inferences from single locus datasets? A survey of the literature reveals that 88% of the empirical studies using NCPA relied on analysis of single locus data (based on a review of the 235 studies to date that cited Templeton 2004), and in the few cases in which more than one molecular marker was examined, inferences were not cross-validated by analysis of independent markers with NCPA.

Nested Clade Analysis: An Extensively Invalidated Method for Strong Phylogeographic Inference

How is it that after more than a decade since its introduction and 1700 plus citations later, questions about the performance of NCPA are only now being raised (and perhaps recognized)? More to the point, how is it that these concerns over the validity of NCPA (e.g., Knowles and Maddison 2002; Petit and Grivet 2002; Felsenstein 2004; Panchal and Beaumont 2007) go unheeded? In the end, it is hard to dismiss the possibility that it is perhaps merely the intended goal of NCPA that drives its continued popularity, and not any compelling argument put forth on its behalf. However, no one should overlook the obvious. Achieving a joint estimate of the multiple processes that characterize a species' history is difficult. And no matter how appealing such an inference might be, it is simply untenable (at least at this time).

Irrespective of whether you are, or are not, a believer in NCPA's utility, one undeniable fact remains: the performance of NCPA has yet to be thoroughly investigated. There is not a single analysis, simulation or otherwise, that has shown that NCPA can accomplish what it is purported to do—infer multiple historical processes that may characterize a species' history. The simulation studies documenting high error rates examined only simple histories (e.g., either allopatric population divergence or an unstructured population). Only two specific types of events have been evaluated with the so-called validation approach of comparing the results from NCPA to expectations about historical events (Templeton 2004), namely fragmentation and population expansion. The legacy of NCPA ultimately rests squarely on the entire phylogeography community—how long will a field cling to an ideal rather than requiring objective critical validation?

ACKNOWLEDGMENTS

I would like to thank M. Turelli, L. McBride, D. OFoighil, B. Payne, and members from the Knowles laboratory for helpful suggestions on the manuscript.

LITERATURE CITED

Beaumont, M. A., and M. Panchal. 2008. On the validity of nested clade phylogeographical analysis. Mol. Ecol. 17:2563–2565.

Excoffier, L., and G. Heckel. 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7:745–58.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates. Sunderland, MA.

Garrick, R. C., R. J. Dyer, L. B. Beheregaray, and P. Sunnucks. 2008. Babies and bathwater: a comment on the premature obituary for nested clade phylogeographic analysis. Mol. Ecol. 17:1401–1403.

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

McCormack, J. E., H. Huang, and L. L. Knowles. 2009. Sky islands. *In* R. Gillespie and D. Clague, eds. *Encyclopedia of Islands*. University of California Press, Berkeley, CA. *In press*.

Panchal, M. 2007. The automation of nested clade phylogeographic analysis. Bioinformatics 23:509–510.

Panchal, M., M. A. Beaumont. 2007. The automation and evaluation of nested clade phylogeographic analysis. Evolution 61:1466–1480

Petit, R. J. 2008. The coup de grâce for nested clade phylogeographic analysis? Mol. Ecol. 17:516–518.

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

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. 3rd edn. Freeman and Co., New York.

Templeton, A. R. 2004. Statistical phylogeography: methods of evaluating and minimizing inference errors. Mol. Ecol. 13:789–809.

———. 2008. Nested clade analysis: extensively validated method for strong phylogeographic inference. Mol. Ecol. 17:1877–1880.

Associate Editor: M. Rausher