Systematic Biology Advance Access published May 24, 2010

© The Author(s) 2010. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syq022

Evaluating Nested Clade Phylogeographic Analysis under Models of Restricted Gene Flow

MAHESH PANCHAL[1,2,*] AND MARK A. BEAUMONT[1]

[1]*School of Biological Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK; and* [2]*Present address: Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön;*
[*]*Correspondence to be sent to: Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön; E-mail: panchal@evolbio.mpg.de.*

Received 11 November 2008; reviews returned 31 December 2008; accepted 4 January 2010
Associate Editor: L. Lacey Knowles

Abstract.—Nested clade phylogeographic analysis (NCPA) is a widely used method that aims to identify past demographic events that have shaped the history of a population. In an earlier study, NCPA has been fully automated, allowing it to be tested with simulated data sets generated under a null model in which samples simulated from a panmictic population are geographically distributed. It was noted that NCPA was prone to inferring false positives, corroborating earlier findings. The present study aims to evaluate both single-locus and multilocus NCPA under the scenario of restricted gene flow among spatially distributed populations. We have developed a new program, ANeCA-ML, which implements multilocus NCPA. Data were simulated under 3 models of gene flow: a stepping stone model, an island model, and a stepping stone model with some long-distance dispersal. Results indicate that single-locus NCPA tends to give a high frequency of false positives, but, unlike the random-mating scenario presented previously, inferences are not limited to restricted gene flow with isolation by distance or contiguous range expansion. The proportion of single-locus data sets that contained false inferences was 76% for the panmictic case, 87% for the stepping stone model, 79% for the stepping stone model with long-distance dispersal, and more than 99% for the island model. The frequency of inferences is inversely related to the amount of gene flow between demes. We performed multilocus NCPA by grouping the simulated loci into data sets of 5 loci. The false-positive rate was reduced in multilocus NCPA for some inferences but remained high for others. The proportion of multilocus data sets that contained false inferences was 17% for the panmictic case, 30% for the stepping stone model, 4% for the stepping stone model with long-distance dispersal, and 54% for the island model. Multilocus NCPA reduces the false-positive rate by restricting the sensitivity of the method but does not appear to increase the accuracy of the approach. Three classical tests—the analysis of molecular variance method, Fu's *F*s, and the Mantel test—show that there is information in the data that gives rise to explicable results using these standard approaches. In conclusion, for the scenarios that we have examined, our simulation study suggests that the NCPA method is unreliable and its inferences may be misleading. We suggest that the NCPA method should not be used without objective simulation-based testing by independent researchers. [Nested clade analysis; phylogeography; restricted gene flow; simulation.]

Populations of organisms typically have a complex history of dispersal, with fluctuations in density (Woods and Davis 1989). These processes are captured imperfectly in the genealogy of contemporary organisms (Wiuf 2003), which we try to decipher through genetic analysis. The causal links between demographic history and patterns of haplotypic variation can be tenuous, but nonetheless, a variety of approaches have been developed that attempt to infer history from these patterns (Knowles 2009). These can be grouped into methods that infer parameters in specific demographic models (Williamson et al. 2005; Fagundes et al. 2007), and nonparametric techniques more closely based on the shape of inferred gene genealogies (Templeton et al. 1995; Bandelt et al. 1999). The term "nonparametric" is used here to describe methods in which the model structure and parameters are not specified beforehand but are derived from the data. The focus of the present study is to extend the work of Panchal and Beaumont (2007) and examine further the performance of one of these latter approaches (nested clade phylogeographic analysis [NCPA]; Templeton et al. 1995) using data simulated under a number of known scenarios.

NCPA is a method, based on the analysis of single or multiple haplotype networks, that aims to identify past demographic events that have shaped the history of a population. A clustering method based on the rules presented in Templeton et al. (1992) is used to create groups referred to as clades from which summary statistics are calculated. These summary statistics are computed using the equations in Posada et al. (2006) implemented in GeoDis (Posada et al. 2000). Multiple summary statistics are calculated within each clade. It is assumed that the presence of a significant summary statistic indicates that there is some association between the geography and the haplotypes. Panchal and Beaumont (2007) found that many summary statistics can be significant when there is no association between haplotypes and geography, leading to many false inferences (i.e., significant associations between haplotypes and geography were inferred even though no such structure existed because the data were simulated under a condition of population panmixia). Furthermore, when significant associations are detected, the inferences about the underlying processes by NCPA do not appear to be accurate when there is geographic structuring. Knowles and Maddison (2002) found that NCPA yielded many inferences that did not correspond to the history of fragmentation simulated. However, this was criticized by Templeton (2004b), who suggests NCPA can accurately infer processes and discriminate between them (for further discussion, see Knowles 2008 and Templeton 2009b). This final stage of single-locus NCPA consists of a series of questions that interpret patterns of summary statistics and geographical distribution known as the inference key. It is the inference

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

key that finally identifies which process or historical event occurred that caused the particular association and geographic distribution of haplotypes within a clade.

Recently, there has been much discussion of the NCPA procedure. In particular, responding to criticism by Petit (2008), Templeton (2008) has argued against the findings in Panchal and Beaumont (2007), who have in turn defended their results (Beaumont and Panchal 2008). Further criticisms of the NCPA procedure have been made by Knowles (2008) and defended by Templeton (2009b). Additional comments are to be found in Templeton (2009a) and Nielsen and Beaumont (2009). A key feature of recent defences of NCPA by Templeton has been to acknowledge the tendency of the method to produce false positives, and the need to correct for multiple testing (Templeton 2008, 2009b), but then to emphasize the utility of multilocus extensions to the NCPA method, even though it has never been tested.

Multilocus NCPA is designed to combat the problem of false positives and false negatives from single-locus NCPA (Templeton 2004b) and was first used in Templeton (2002) to describe a model of human evolution. In population genetics, generally, there has been a recognition of the need to base inferences on multiple loci (Jennings and Edwards 2005; Carstens and Knowles 2007), and the switch in emphasis from single-locus to multilocus NCPA has been claimed to be part of this trend (Templeton 2004b). However, there is a fundamental difference between the arguments favoring multiple loci in model-based statistical tests, which hinge on the increased power (Rannala and Yang 2003), in comparison with the arguments in favor of multiple-locus NCPA that acknowledge the need to reduce false positives (Templeton 2009a, 2009b). In the first case, inferences were simply never made because the tests were underpowered. In the second case, many inferences were made, yet these may well have been wrong.

It should be noted that the vast majority of published studies have used single-locus NCPA. We have found fewer than 20 articles that have used multiple loci, yet more than 265 articles (Posada et al. 2006) have reported inferences using single-locus NCPA. It is thus important to establish the validity of claims in previously published single-locus studies and that is the main aim of the current paper.

With multiple loci, NCPA is applied to each locus as before. However, each inference must be corroborated by similar inferences in another locus. For an inference to be "cross-validated," it must follow certain criteria, which are 1) any inference must be supported by clades in at least 2 loci, 2) the inferences must be the same qualitative type of inference, 3) the clades must be geographically concordant, and 4) inferences of historical events must be temporally concordant (Templeton 2004b). This extension of NCPA, although used in a number of more recent studies (Templeton 2002; Templeton 2005; Brisson et al. 2005), and described as the solution to the false positives for single-locus NCPA (Templeton 2008), has not been tested and validated in any published study. We have developed a new program (ANeCA-ML) that allows the user to take the single-locus output from the program ANeCA (Panchal 2007) and perform multilocus NCPA.

Because of the wide scope of inferences claimed for NCPA, there is much that needs to be validated and tested. No single simulated scenario can examine all the possibilities. Templeton (2004b) tested a subset of NCPA's capability using real data sets that had strong prior evidence of fragmentation and range expansion. Templeton (2004b) noted, however, that simulations were also needed to test NCPA: to determine the true type 1 error rate under a panmictic model (these simulations were performed by Panchal and Beaumont 2007) and to test recovery of inferences under various patterns of gene flow. It is this last set of simulations that we aim to investigate here. In addition to a study of single-locus NCPA, we also provide data for multilocus NCPA, treating data sets from the single-locus study as independent loci.

Templeton et al. (1995) provide an overview of the expected phylogeographic distribution under the 3 scenarios of restricted gene flow, range expansion, and fragmentation. Templeton (1998, 2004b) tests NCPA using empirical data sets in which the populations have strong a priori evidence of range expansion and/or fragmentation. We chose to add to this information by investigating 3 models of gene flow: a stepping stone model, an island model, and a stepping stone model with some long-distance dispersal. A stepping stone model (Kimura and Weiss 1964) represents the extreme of short-distance movements, whereas an Island model (Wright 1931) represents the extreme of long-distance dispersal. The stepping stone model with some long-distance dispersal represents a more realistic compromise between these 2 models, with mostly short-distance movements, but with some long-distance dispersal events. These models are described in more detail below. More complex models of gene flow would incorporate at least one of these models.

Simulation Model Parameters and Methods

Each gene flow model (stepping stone, island, and stepping stone with long-distance dispersal) was simulated on a flat $x \times x$ lattice of demes, where x was either 3, 7, or 10. The total population size was 90,000 for a 3×3 grid (10,000 individuals per deme), 98,000 for a 7×7 grid (2000 individuals per deme), and 100,000 for a 10×10 grid (1000 individuals per deme). Two sampling schemes were also used. The first sampling scheme sampled all demes on the grid. In the second sampling scheme, referred to as the half-sampled scheme, demes were sampled beginning with the first deme, sampling only diagonally adjacent demes to those that were sampled. Three groups of sample sizes were also used. In the first group, when all demes were sampled, the total sample size was 900 for the 3×3 grid, 980 for the

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

7×7 grid, and 1000 for the 10×10 grid, and when using the half-sampled scheme, the total sample size was 500 for the 3×3 grid, 500 for the 7×7 grid, and 500 for the 10×10 grid. In the second group, sampling all demes, the total sample size was 450 for the 3×3 grid, 490 for the 7×7 grid, and 500 for the 10×10 grid. Under the half-sampled scheme, the total sample sizes for the second group were 250 for all 3 grid sizes. In the third group, when all demes were sampled, the total sample sizes were 180 for the 3×3 grid, 196 for the 7×7 grid, and 200 for the 10×10 grid, and under the half-sampled scheme, 100 under all 3 grid sizes. For each model of gene flow, we chose migration matrices that had the same probability of sharing a common ancestor within a deme without intervening migration. In an island model with low mutation rate, this becomes the same as the parameter F_{ST}, defined in, for example, Rousset (2003), and we will refer to it here as F_{ST} (as in Beaumont and Nichols 1996 and Beaumont and Balding 2004). It should be noted that standard estimators of F_{ST} will then have expectations close to the value of F_{ST} used in the simulations only in the case of the island model and will be larger than expected for the stepping-stone model (Crow and Aoki 1984). Genealogies were simulated with F_{ST} of either 0.03, 0.05, 0.1, or 0.2, where, from coalescent theory, the probability that 2 lineages in a randomly chosen deme coalesce before migrating (backward in time) is $F_{ST} = (1/k)\sum_{i=1}^{k} 1/(1+2N_i m_i)$, following Beaumont and Nichols (1996). Here, N_i is the number of individuals in deme i, k is the number of demes, $m_i = \sum_{j=1}^{k} m_{ij}$ is the total immigration rate into deme i, and m_{ij} is the migration rate from deme j to deme i. In the stepping stone model, migration rates were identical between neighboring demes or 0 otherwise. In the island model, migration rates were identical between all demes (migration between a given deme and itself being zero). In the stepping stone distribution with some long-distance dispersal, migration rates were based on a Cauchy distribution, given by $f(x\,;x_0,\gamma) = \gamma/(\pi((x-x_0)^2+\gamma^2))$, where $x_0 = 0$ is the location parameter and γ is the shape parameter, given in Table 1. We chose a Cauchy because it is quite widely used in dispersal models (e.g., Skarpaas et al. 2005) for modeling fat-tailed distributions. For each combination of factors (72 combinations of parameters for each model), 100 genealogies were generated using the software Simcoal v1.0 (Excoffier et al. 2000) and analyzed using the software ANeCA v1.2 (Panchal 2007), which implements the November 2005 inference key and includes the software TCS 1.21 (Clement et al. 2000) and GeoDis 2.5 (Posada et al. 2000), using both the longitude–latitude and geographic distance matrix method of input to GeoDis. The program TCS is used to construct haplotype networks, and the program GeoDis computes the statistics D_c and D_n and performs permutation tests to generate P values. The statistic D_c measures the geographical spread of members of a clade relative to their mean location and D_n measures the geographical spread of members of a clade relative to the mean location of all members of the nesting clade. DNA sequence data were simulated 500 bp 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 4 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. The intention was to simulate data with a $\theta = 2Dn_d\mu \approx 2$, where D is the number of demes and n_d is the deme size. The results obtained by Panchal and Beaumont (2007) used ANeCA v1.0. The present version includes a minor bug fix and updates to TCS, GeoDis, and the inference key described in the software distribution. We have reanalyzed the data sets generated under panmixia in Panchal and Beaumont (2007) using ANeCA v1.2, and the results are indistinguishable within sampling error to those described in that publication.

TABLE 1. The shape parameters used to determine the Cauchy distribution that characterize the migration rates for each grid size to obtain an F_{ST} as close as possible to the value given

Number of demes	F_{ST}	Shape parameter
100	0.03	225
100	0.05	95
100	0.1	40
100	0.2	17
49	0.03	105
49	0.05	55
49	0.1	25
49	0.2	11
9	0.03	35
9	0.05	20
9	0.1	9
9	0.2	4

This study also uses classical summary statistics to examine if there is information in the data that could be used by NCPA to make inferences. For each parameter set, the analysis of molecular variance (AMOVA) method within Arlequin (Excoffier et al. 2005) was used to calculate F_{ST} and detect significant structure. We also chose to look at Fu's Fs (Fu 1997), which is sensitive not only to selection but also to demographic expansion when loci are neutrally evolving. Both selection and demographic expansion can result in an excess of rare haplotypes, and Fu (1997) states that θ estimated by the average pairwise difference, $\hat{\theta_\pi}$, is likely to be smaller than θ based on the number of alleles when there is an excess of recent mutations. Fu's Fs assumes a single closed population. However, its behavior in structured populations has been previously studied (e.g., Ray et al. 2003), and it has been shown to be able to detect evidence of population expansion during spatial expansions in structured populations that also include an expansion of metapopulation size. The Mantel test was used to look for the presence of isolation by distance. The Mantel test is a test of correlation between 2 matrices, in this case the geographic distances between demes and Slatkin's linear F_{ST} (given by $F_{ST}/(1-F_{ST})$).

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

If the null hypothesis of no correlation between the matrices is true, then permuting the rows and columns of the matrix should be equally likely to produce a larger or smaller coefficient. If the null hypothesis of no correlation is rejected, it implies that there is isolation by distance (increasing differentiation between demes as geographic distance increases).

Implementation of Multilocus NCPA

To test multilocus NCPA, an application (ANeCA-ML) was written in the Java language and uses the output files from ANeCA to perform the multilocus analysis. A full description of the software is given in online Appendix 5 (available from http://www.sysbio.oxfordjournals.org/) with a brief overview given here. The methods described below follow those given by Templeton as closely as possible. It should be noted that we do so in order to test the method. We do not endorse any of these procedures as valid. In particular, the methods for dating historical demographic events have been severely criticized (Tavare et al. 1997; Beaumont et al. 2010).

The software collects together a list of all the clades with inferences that are of the type restricted gene flow with isolation by distance (RGF_IBD), restricted gene flow with long-distance dispersal, contiguous range expansion (CRE), allopatric fragmentation, and long-distance colonization. Each clade is then clustered if it shares greater than 75% overlap in geographic area with at least one other clade in the cluster. The choice of percentage overlap is somewhat arbitrary. Templeton has stated that the inferences found at different loci should be "qualitatively" similar (Templeton 2004b), which involves judgement on the part of the user. In the study of human demographic history (Templeton 2002), for example, some loci were sampled at up to 4 geographic sites and others at 35 sites, requiring some arbitrary grouping of geographic areas when judging overlap. Inferences of RGF_IBD and restricted gene flow with some long-distance dispersal are not subjected to further "cross-validation" because there is no stipulation that the inferences should be concordant across time. However, if a cluster of geographically concordant inferences (in different clades) are all inferred at the same locus, they are also discarded (because they are not cross-validated).

For historical events (allopatric fragmentation, CRE, and long-distance colonization), it is necessary to compute the times associated with the events and to determine whether the inferred times are consistent across loci. Templeton (2002, 2004a) uses theory from Takahata et al. (2001) to estimate the Time to Most Recent Common Ancestor (T_{MRCA}) associated with each clade, which is taken to date the event. Essentially (for the haploid case), the expected pairwise difference $E[\hat{\pi}] = 2N\mu$, where μ is the mutation rate and N is the population size. Under the standard coalescent, $E[T_{MRCA}] = 2N$, and therefore, an estimate of T_{MRCA} is given by $\hat{\pi}/\mu$ using the empirical mean pairwise difference. In this calculation, we used the same mutation rate of 0.00002 per 500 bp sequence that was used to generate the data. Templeton refers to this as the TLS (Takahata-Lee-Satta) estimator.

Templeton (2002) devised a test for homogeneity of estimated coalescence times, based on the assumption that they follow a gamma distribution, computed by means of a *G*-statistic (equations 16 and 19 in Templeton 2004a). The *G*-statistic is calculated for the geographically concordant group. If the P value is < 0.95 (unable to reject the null hypothesis of a single event), the cluster is regarded as temporally concordant. This test is used to identify temporally concordant groups of clades. However, as noted in Templeton (2002), a number of temporally distinct events, "cross-validated" at different loci, may be identified in this way. For example, in Templeton (2002, fig. 3), the *G*-statistic was used to reject the null hypothesis of a single date of population range expansion, inferred at 5 loci in humans. These were then grouped (apparently, visually) into 2 different temporal groups within which the *G*-statistic was not significant: 3 (autosomal) loci supporting an older expansion and 2 (mitochondrial DNA and Y) loci supporting a younger expansion. A general algorithm therefore requires that subgroups of temporally concordant clades are identified in an automatic way.

To achieve this, the program proceeds as follows. Starting with the geographically concordant clades with the same type of inference, we enumerate the powerset. Each subgroup must satisfy the following tests:

1. The *G*-statistic is nonsignificant.
2. Within a geographically concordant group, all clades must not be from the same locus.
3. For allopatric fragmentation, all the isolates that are identified using the inference key must be concordant too. If the intersection of the set of populations is greater than 75% of the union of populations for 2 isolates, then they are considered as geographically concordant. If there are 2 or more isolates identified for a locus, then they all must be geographically concordant with the 2 or more isolates identified on the other loci. If just a single isolate is identified on a locus, it must be geographically concordant with at least one of the isolates identified on another locus.
4. For CRE and long-distance colonization, the intersection of the populations for tip clades must be greater than 75% of the union of populations for the tip clades and similarly for the interior clades to be geographically concordant.

The program then retains the largest subsets that satisfy these criteria. Thus, in the example given in figure 3 of Templeton (2002), this algorithm would result in the 2 identified groups of clades.

To summarize, inferences of restricted gene flow are cross-validated only if they are concordant in space, and historical events are only cross-validated if they are concordant in time and also identify the same isolates (in the case of allopatric fragmentation) or identify the

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

same origin and destination (in the case of CRE and long-distance colonization).

In order to test the multilocus NCPA procedure, we grouped the simulated single-locus data sets, described above, into sets of 5 loci. Thus, each replicate of 100 loci was reduced to 20 data sets. We then applied ANeCA-ML and recorded the inferences obtained in the same way as for the single-locus data.

Results

Prior Expectations of NCPA

For each gene flow model, different results are expected from NCPA. Under the stepping stone model, only inferences that include isolation by distance are expected, such as "RGF_IBD." Under the island model, only inferences that include long-distance dispersal are expected, such as "restricted gene flow/dispersal but with some long-distance dispersal." In the stepping stone model with some long-distance dispersal, inferences of both isolation by distance and long-distance dispersal are expected, although a greater proportion are expected to be of isolation by distance.

Distribution of NCPA Inferences—Single Locus

The distribution of mean frequencies for each inference in each model is shown in Figures 1–3. Following Panchal and Beaumont (2007), we distinguish between inferences made from clades individually (clade-level inferences) and those made at least once within an entire data set (data set level). The distribution of mean frequencies for each inference at the clade level are the proportions of clades within a parameter group (a particular combination of grid size, sample size, and sampling scheme) that gave rise to that particular inference regardless of clade level or data set/genealogy. The distribution of mean frequencies for each inference at the data set level are the proportions of data sets within a parameter group that gave rise to at least one inference of that type. Although it is interesting to see how frequently an inference is made within a clade, generally only one clade needs to contain an inference within a data set to potentially form the basis of a publication and is why the distinction is made.

The results show that under all 3 models, the null hypothesis of no geographic association is rejected with greater frequency compared with a panmictic model. This also increases with F_{ST}, showing that the results of NCPA are sensitive to the amount of geographic structuring. The results also show, however, that NCPA is not able to discriminate effectively between demographic scenarios, shown by the increase in frequency of most other inferences, in particular at the data set level. Under the panmictic model, false positives were only commonly observed for the inferences of RGF_IBD and CRE (Panchal and Beaumont 2007). By contrast, under these 3 models of gene flow, a much wider range of inferences are obtained. In this case, inferences other than those that include restricted gene flow are false positives and their frequency increases with increasing population structure. Table 2 shows which inferences are considered true and which inferences are considered false for each model.

For the stepping stone model, we found that 24% of all clades resulted in inferences of RGF_IBD, marginal to the parameters. We also found that 22% of all clades were false conclusive inferences. The remainder of the clades were either unable to reject the null hypothesis of no geographic association, inconclusive outcome, or one of the various inferences of insufficient sampling. On the data set level, only 13% of all data sets inferred at least one RGF_IBD and no other concrete inferences (inferences of inconclusive outcome or sampling inadequacy were ignored). However, 8% of data sets did infer at least one other concrete inference that was not RGF_IBD in which there were no inferences of RGF_IBD.

In the stepping stone model with some long-distance dispersal, we found that 21% of all clades inferred the above inferences, marginal to the parameters, and 15% of clades inferred a concrete inference other than that of restricted gene flow. At the data set level, only 21% of all data sets inferred only restricted gene flow and no other concrete inference (again ignoring inferences of inconclusive outcomes or sampling inadequacy) and 10% of all data sets inferred at least one concrete event other than restricted gene flow and did not infer restricted gene flow.

In the island model less than 1% of clades inferred restricted gene flow with long-distance dispersal and 32% of clades were some other concrete inference. Less than 1% of data sets inferred at least one restricted gene flow with some long-distance dispersal and no other concrete inference. We found that 85% of data sets inferred at least one inference that was not restricted gene flow with long-distance dispersal and did not infer at least one restricted gene flow with long-distance dispersal.

It can also be seen that the proportions of inferences are similar for each inference between the 3 models. A closer comparison between the 3 models also indicates that NCPA finds more signals of long-distance dispersal in the stepping stone model compared with the island model; however, this may be due to slightly greater geographic structuring within the stepping stone models in general (see below, and online Appendices 2–4—http://www.sysbio.oxfordjournals.org). In fact, there is little inference of long-distance dispersal at all in the island model and of those inferences that do include long-distance movements the majority also indicate past fragmentation (RGF_LDD_PF and LDC_PF). CRE also continues to be inferred in high proportions under all 3 models.

There are wide intervals on the mean frequencies of inferences because these are marginal to the factors investigated, which are sample size, grid size, and sampling scheme. Additionally, the difference in frequencies of inferences between F_{ST} can also be explained by the greater structuring as F_{ST} increases, leading to larger haplotype networks. Larger haplotype networks will

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

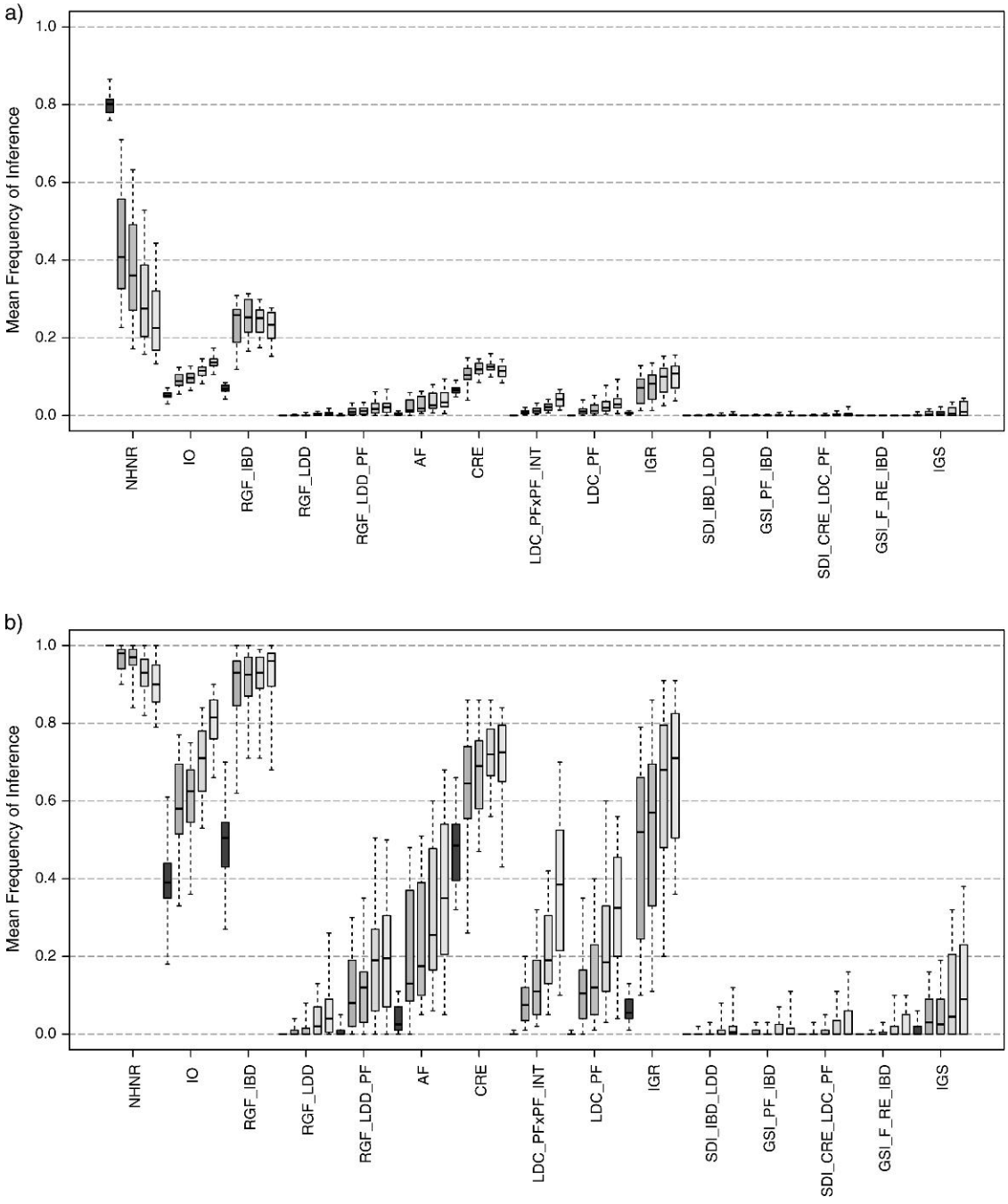

FIGURE 1. Stepping stone model: a) the distribution of means of all inferences at the clade level from each parameter group and b) the distribution of means of all inferences at the data set level from each parameter group. Distributions of means are separated by F_{ST} (panmixia = $F_{ST}=0$, $F_{ST}=0.03$, $F_{ST}=0.05$, $F_{ST}=0.1$, and $F_{ST}=0.2$ from darkest to lightest). The key to the inference codes are as follows and are given in online Appendix 1 that also includes the questions that lead to the given inference. NHNR, the null hypothesis of no geographic association is not rejected; IO, inconclusive outcome; RGF_IBD, restricted gene flow with isolation by distance (restricted dispersal by distance in nonsexual species); RGF_LDD, restricted gene flow/dispersal but with some long-distance dispersal; RGF_LDD_PF, restricted gene flow/dispersal but 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_PFxPF_INT, long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion; LDC_PF, long-distance colonization and/or past fragmentation; 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_PF_IBD, geographical sampling scheme inadequate to discriminate between fragmentation and isolation by distance; SDI_CRE_LDC_PF, sampling design inadequate to discriminate between CRE, long-distance colonization, and past fragmentation; GSI_F_RE_IBD, geographical sampling scheme inadequate to discriminate between fragmentation and isolation by distance; IGS, inadequate geographical sampling. The whiskers of each box plot indicate the maximum and minimum mean frequency. The box range is from the upper quartile to lower quartile of the distribution of mean frequencies, and the line through each box is the median.

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

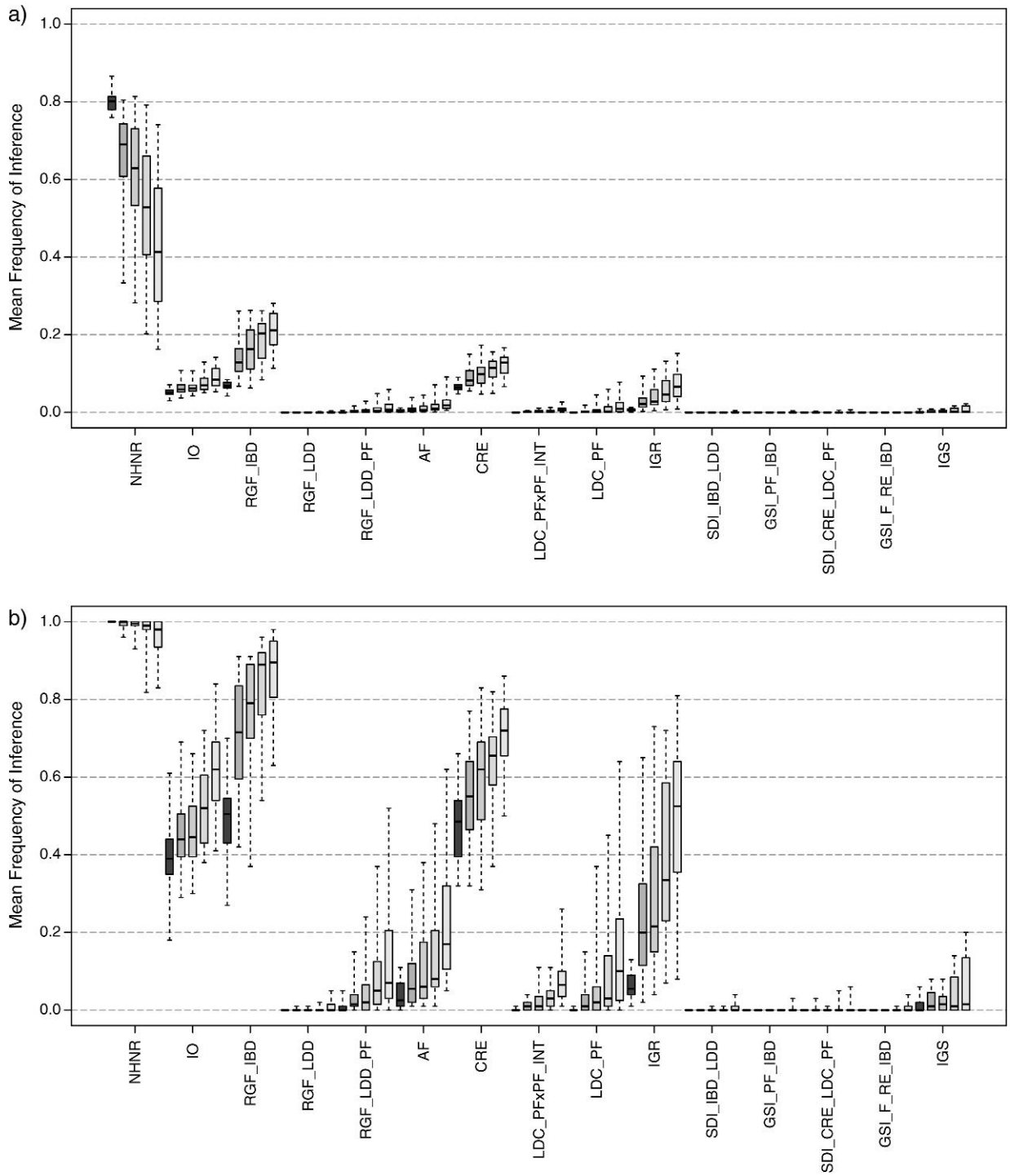

FIGURE 2. Island model: a) the distribution of means of all inferences at the clade level from each parameter group and b) the distribution of means of all inferences at the data set level from each parameter group. Distributions of means are separated by F_{ST} (panmixia $=F_{ST}=0$, $F_{ST}=0.03$, $F_{ST}=0.05$, $F_{ST}=0.1$, and $F_{ST}=0.2$ from darkest to lightest). The key to the inference codes along the axis are found in 1 and in online Appendix 1. The whiskers of each box plot indicate the maximum and minimum mean frequency. The box range is from the upper quartile to lower quartile of the distribution of mean frequencies, and the line through each box is the median.

mean either the creation of more clades, potentially more statistics per clade, or perhaps both, potentially leading to more inferences. Table 3 shows that under all 3 scenarios, the average number of clades per data set increases with F_{ST}; however, the average number of statistics per clade decreases slightly.

This shows that the increase in frequency of inferences can be due to an increase in the number of clades,

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

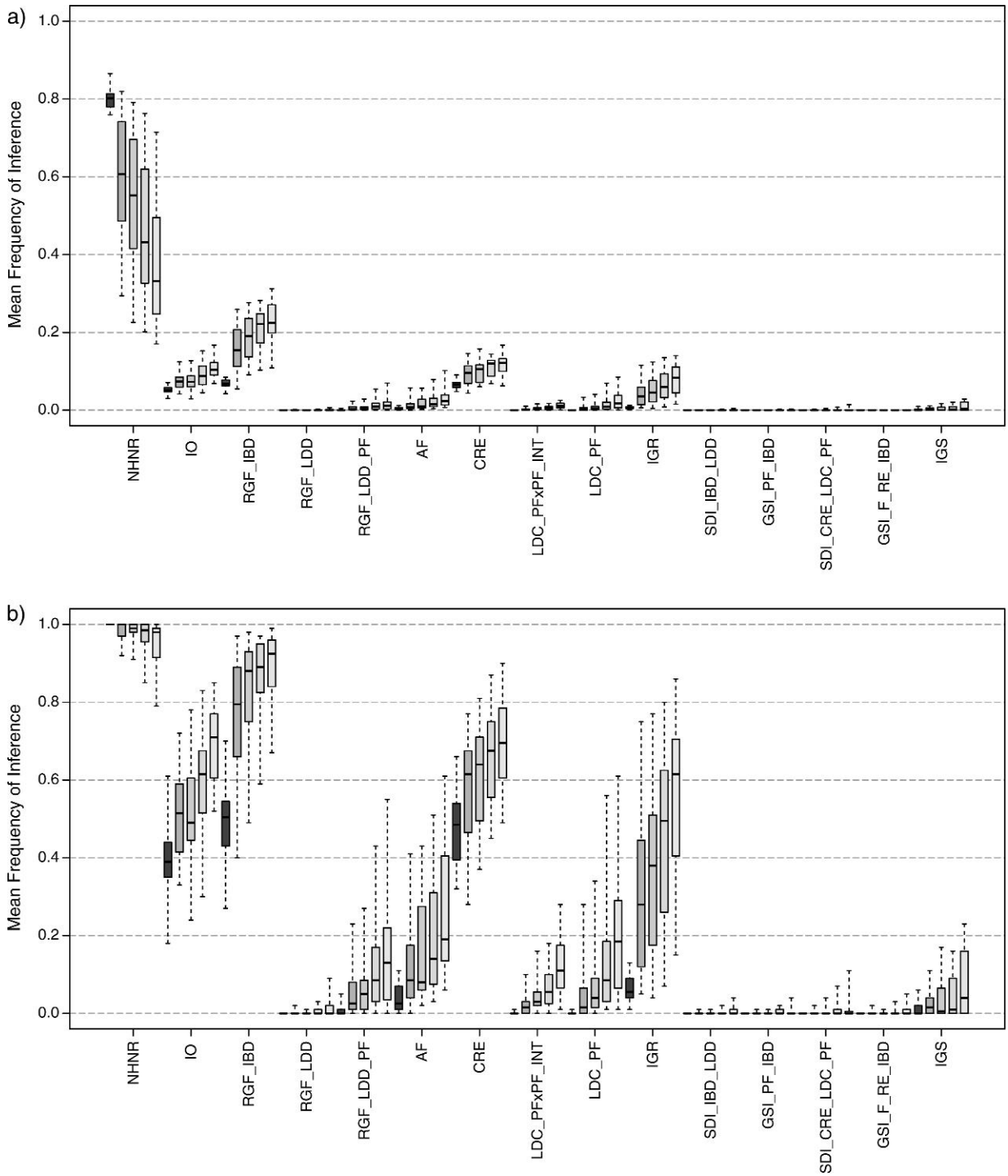

FIGURE 3. Stepping stone model with some long-distance dispersal: a) the distribution of means of all inferences at the clade level from each parameter group and b) the distribution of means of all inferences at the data set level from each parameter group. Distributions of means are separated by F_{ST} (panmixia = $F_{ST} = 0$, $F_{ST} = 0.03$, $F_{ST} = 0.05$, $F_{ST} = 0.1$, and $F_{ST} = 0.2$ from darkest to lightest). The key to the inference codes along the axis are found in 1 and in online Appendix 1. The whiskers of each box plot indicate the maximum and minimum mean frequency. The box range is from the upper quartile to lower quartile of the distribution of mean frequencies, and the line through each box is the median.

but not to an increase in the number of statistics per clade. This does not tell us whether the underlying geographic structure is affecting the chance of having at least one significant statistic per clade. To determine the factors that affect the probability of having at least one significant statistic per clade, we performed a step-down regression, including the number of statistics (numerical), the sample size (factor), the sampling scheme

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

TABLE 2. Which inferences are defined as true and which are defined as false under the given model

Model	True inferences	False inferences
Panmixia	The null hypothesis of no geographic association is not rejected (NHNR)	Restricted gene flow with isolation by distance (restricted dispersal by distance in nonsexual species) (RGF_IBD), restricted gene flow/dispersal but with some long-distance dispersal (RGF_LDD), restricted gene flow/dispersal but with some long-distance dispersal over intermediate areas not occupied by the species; or past gene flow followed by extinction of intermediate populations (RGF_LDD_PF), contiguous range expansion (CRE), allopatric fragmentation (AF), long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion (LDC_PFxPF_INT), long-distance colonization and/or past fragmentation (LDC_PF)
Stepping stone model	Restricted gene flow with isolation by distance (restricted dispersal by distance in nonsexual species)	Restricted gene flow/dispersal but with some long-distance dispersal, restricted gene flow/dispersal but with some long-distance dispersal over intermediate areas not occupied by the species; or past gene flow followed by extinction of intermediate populations, contiguous range expansion, allopatric fragmentation, long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion, long-distance colonization and/or past fragmentation
Stepping stone model with some long-distance dispersal	Restricted gene flow with isolation by distance (restricted dispersal by distance in nonsexual species), restricted gene flow/dispersal but with some long-distance dispersal, restricted gene flow/dispersal but with some long-distance dispersal over intermediate areas not occupied by the species; or past gene flow followed by extinction of intermediate populations	Contiguous range expansion, allopatric fragmentation, long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion, long-distance colonization and/or past fragmentation
Island model	Restricted gene flow/dispersal but with some long-distance dispersal, restricted gene flow/dispersal but with some long-distance dispersal over intermediate areas not occupied by the species; or past gene flow followed by extinction of intermediate populations	Restricted gene flow with isolation by distance (restricted dispersal by distance in nonsexual species), contiguous range expansion, allopatric fragmentation, long-distance colonization possibly coupled with subsequent fragmentation or past fragmentation followed by range expansion, long-distance colonization and/or past fragmentation

(factor), the grid size (factor), the F_{ST} (factor), and the method of calculating the distance statistics (factor) as independent variables. The dependent variable was whether there was at least one significant statistic within a clade, and we used a binomial generalized model with a logit link function. Interactions were not included in the model. Under all 3 scenarios, excluding F_{ST} caused a significant ($P \leq 0.0001$, χ^2_3) change in deviance, which shows geographic structure, independent of the other factors, has an effect on the chance of finding at least one significant statistic within a clade. The number of statistics per clade was the only variable that caused a greater change in deviance in all 3 models. Under the stepping stone model, and stepping stone model with some long-distance dispersal, the exclusion of sampling scheme as an explanatory variable was nonsignificant; it does not effect the chance of finding at least one significant statistic within a clade.

TABLE 3. How the average number of haplotypes per data set, average number of clades per data set, and the average number of statistics per clade varies with F_{ST} within each model

F_{ST}	Average number of haplotypes	Average number of clades	Average number of statistics
Model—isolation by distance			
0.03	17.96	9.75	7.15
0.05	18.13	9.90	7.13
0.1	19.00	10.52	7.08
0.2	19.99	11.42	6.96
Model—isolation by distance with some long-distance dispersal			
0.03	17.75	9.59	7.16
0.05	18.03	9.76	7.17
0.1	18.26	10.00	7.11
0.2	18.98	10.54	7.05
Model–long-distance dispersal			
0.03	17.65	9.28	7.32
0.05	17.79	9.34	7.32
0.1	18.16	9.62	7.29
0.2	19.06	10.36	7.18

Overall we see that NCPA has performed poorly, leading to a similar pattern of frequency of inferences made, regardless of the model of gene flow (i.e., Figs. 1–3 look very similar). It is also shown that a greater frequency of inferences of all types are likely to be made under a scenario of restricted gene flow unlike under a model of panmixia. Inferences of RGF_IBD and CRE are the most common outcome, as previously noted for the panmictic scenario (Panchal and Beaumont 2007). The results also show that as geographic structure increases, so do the number of clades per data set. However, even though the number of clades increases per data set as geographic structure increases, the number of statistics per clade do not. The results also show that there is an increasing chance of finding at least one significant statistic within a clade as geographic structure increases,

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

TABLE 4. The proportion of data sets with at least one "cross-validated" inference in a data set under the 4 models studied

Model	Proportion of data sets with at least one cross-validated inference					
	RGF_IBD	RGF_LDD	CRE	AF	LDC	False
Panmictic	0.1639	0.0042	0.0028	0	0	0.1694
Stepping stone	0.8076	0.2792	0.0278	0.0028	0.0031	0.3038
Stepping stone with LDD	0.6319	0.1559	0.0406	0.0021	0.0021	0.0444
Island	0.5306	0.1118	0.0368	0.0007	0.0014	0.5403

Notes: Each data set consists of 5 loci. The proportions shown here are each based on 720 data sets for the panmictic case, and 2880 data sets for each of the other models. RGF_IBD is the chance of getting at least one cross-validated inference of restricted gene flow with isolation by distance. Similarly, RGF_LDD is restricted gene flow with some long-distance dispersal, CRE is contiguous range expansion, AF is allopatric fragmentation, and LDC is long-distance colonization. The column "False" is the proportion of data sets that falsely cross-validate at least one inference not simulated. These are all inferences for the panmictic model, the set of RGF_LDD, CRE, AF, and LDC for the stepping stone model; the set of CRE, AF, and LDC for the stepping stone model with some long-distance dispersal; and the set of RGF_IBD, CRE, AF, and LDC for the island model.

which is independent of the number of statistics within that clade.

Distribution of NCPA Inferences—Multiple Loci

The results that we have obtained using multilocus NCPA with GeoDis are presented in Table 4. As with the single-locus results described above, for ease of exposition, we have chosen to present the results for the 5-locus data sets marginal to the parameters that were varied (sample size, number of demes, proportion of demes sampled, F_{ST}, and whether coordinates or distance matrices are used as input to GeoDis). Thus, each proportion is based on either 2880 samples or, in the panmictic case where F_{ST} is not varied, 720 samples.

It can be seen that the cross-validation procedure does indeed reduce the level of false positives. In particular, in the stepping stone model with long-distance dispersal, the false-positive rate is reduced from 79% of data sets that have at least one false inference in the single-locus case to less than 5% that falsely cross-validate at least one inference that was not simulated in the multilocus case. (We will refer to the former quantity as the false-positive rate for the single-locus case and the latter as the false-positive rate for the multilocus case.) The degree to which the false positives are reduced is somewhat patchy. For example, in the island model, the majority of cross-validated inferences are of RGF_IBD, as indeed is the case for all the scenarios that we looked at (including the panmictic case). Thus, for the island model, the false-positive rate is 54%. In the case of the stepping stone model, 81% of cross-validated inferences are indeed of isolation by distance, but, in addition, 28% are of long-distance dispersal, and overall, there is a 30% false-positive rate. In the panmictic case, 16% of the inferences are of RGF_IBD, which is by far the commonest inference, and the overall false-positive rate is 17%. In the panmictic scenario, only 13% of data sets have no significant D_c or D_n, and 24% have no significant outcome (either the null hypothesis could not be rejected or resulted in an "inconclusive outcome" in the inference key) (Panchal and Beaumont 2007). Whereas for the multilocus case, 83% have no cross-validated inference. The inference of restricted gene flow with long-distance dispersal is rare (0.4%) in the panmictic case and more common in all the other scenarios. In this sense, it appears more sensitive to the scenario modeled than the inference of isolation by distance, which is generally by far the commonest inference in the table, even with panmictic data. The inferences of CRE, allopatric fragmentation, and long-distance colonization are much rarer, generally. This probably reflects the additional requirements for cross-validation of the timing of events and the additional geographic requirements. For example, in the single-locus case, CRE is observed in approximately 50% of data sets simulated under panmixia, whereas in the cross-validated data sets, it is reduced to less than 1%.

Results of Classical Summary Statistics

Although the results show NCPA performs poorly, one possible explanation is that perhaps the simulated data show signals other than those expected. Figures 4–6 show the proportions of significant results under each model, separated by F_{ST}. We found the measured F_{ST}s to be very close to the simulated F_{ST}s expected from the simple theory in the case of the island model (online Appendix 4). However, for the stepping stone models, there was greater differentiation than under the island model, for the same F_{ST}, as expected (online Appendices 2 and 3).

Under the stepping stone model, there is generally a high chance of detecting structure using AMOVA when it is present, increasing as geographic structuring increases (Fig. 4a), compared with the 5% false-positive rate when there is no structure (Panchal and Beaumont 2007). Results are similar under the island model (Fig. 5a), and stepping stone model with some long-distance dispersal (Fig. 6a), although it appears that there is greater difficulty in detecting structure when there are more long-distance movements.

Under the stepping stone model, the Mantel test detects isolation by distance on average in less than half of the data sets, even with stronger geographic structuring (Fig. 4b). Under the island model, we do not expect

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

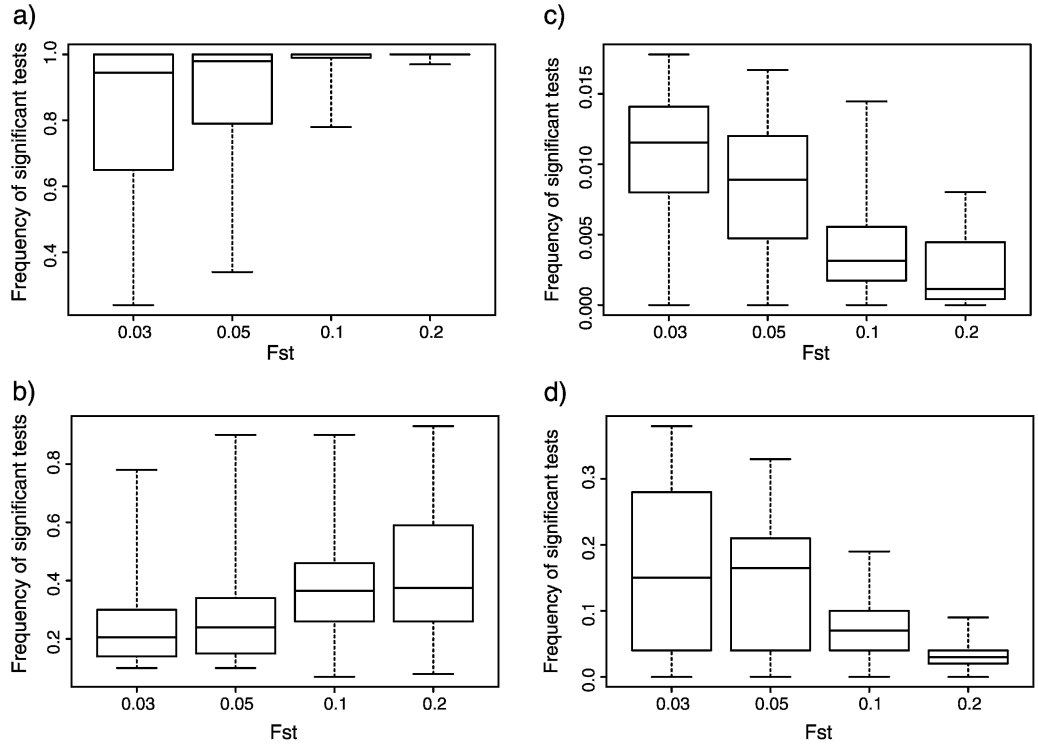

FIGURE 4. These charts show the frequency of significant AMOVA *P* values (a), significant Mantel tests (b), and significant Fu's *F*s at both the deme level (c) and the data set level (d) (no multiple test correction), broken down by F_{ST} under the stepping stone model. The whiskers of each box plot indicate the maximum and minimum frequency of a significant result. The box range is from the upper quartile to lower quartile of the distribution of frequencies of significant results, and the line through each box is the median.

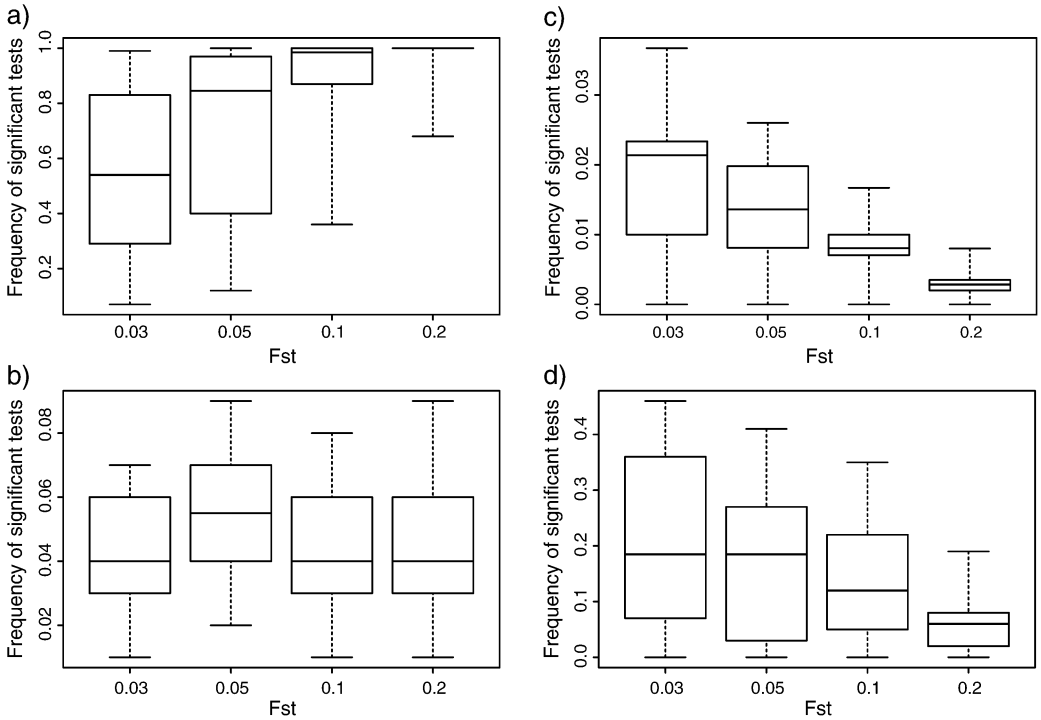

FIGURE 5. These charts show the frequency of significant AMOVA *P* values (a), significant Mantel tests (b), and significant Fu's *F*s at both the deme level (c) and the data set level (d) (no multiple test correction), broken down by F_{ST} under the island model. The whiskers of each box plot indicate the maximum and minimum frequency of a significant result. The box range is from the upper quartile to lower quartile of the distribution of frequencies of significant results, and the line through each box is the median.

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

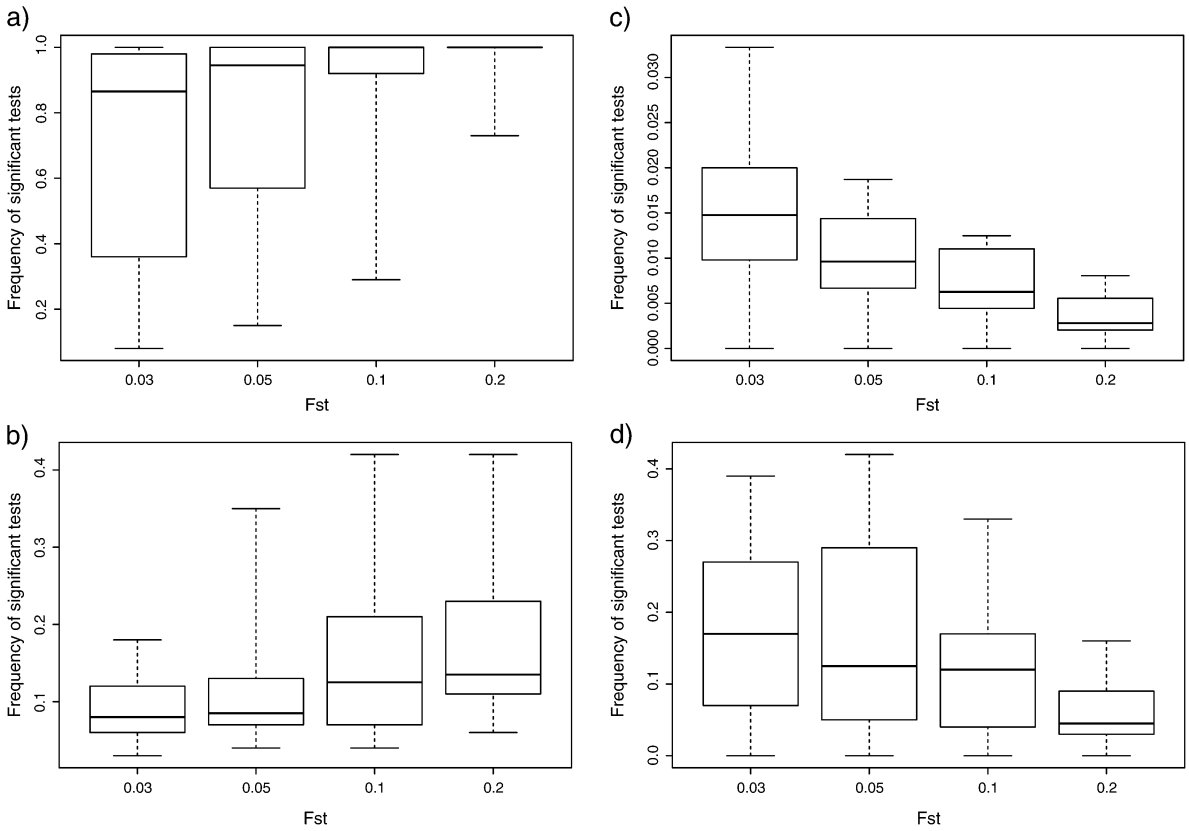

FIGURE 6. These charts show the frequency of significant AMOVA P values (a), significant Mantel tests (b), and significant Fu's Fs at both the deme level (c) and the data set level (d) (no multiple test correction), broken down by F_{ST} under the stepping stone model with some long-distance dispersal. The whiskers of each box plot indicate the maximum and minimum frequency of a significant result. The box range is from the upper quartile to lower quartile of the distribution of frequencies of significant results, and the line through each box is the median.

signals of isolation by distance, and this is confirmed with a false-positive rate of about 5% regardless of the level of structuring (Fig. 5b). Under the stepping stone model with some long-distance dispersal, we do expect some signal of isolation by distance, but not as strong, and this is exactly what is shown. The proportion of data sets in which isolation by distance is detected is generally small (between 10% and 15%; Fig. 6b). Under these 3 models, the Mantel test is accurate to the extent that it is not prone to false positives; however, it does not have much power under these models.

Fu's Fs can be used to test for evidence of population expansion in single closed populations. Ray et al. (2003) have shown that Fu's Fs does detect signals of expansion in various models of spatial expansion that include an expansion of metapopulation size. In our case, the metapopulation size is constant and so we do not expect to find signals of expansion. However, in addition, it is also known that models with restricted gene flow will show signals of population contraction (De and Durrett 2007; Nielsen and Beaumont 2009). Thus, overall, we expect to find no evidence of population expansion through Fu's Fs, but some signals of contraction that increase with increasing F_{ST}. Fu's Fs is calculated on each deme, and so the frequency of significant Fu's Fs and the proportion of data sets with at least one significant Fu's Fs (one tailed) are presented. The reason we present a one-tailed Fu's Fs test is so that we then distinguish between signals of expansion and contraction. At the deme level, we see a false-positive rate of around 3% in all 3 models (Figs. 4c, 5c, and 6c). At the data set level though, the proportion of data sets that have at least one significant Fu's Fs is about 18% when structuring is low, decreasing as structuring increases (Figs. 4d, 5d, and 6d). However, this is without correcting for multiple testing. Because each test within a data set is independent, a Bonferroni correction can be easily applied. After application of the Bonferroni correction, no significant Fu's Fs were found. Figures 4–6 also support the prediction in Nielsen and Beaumont (2009) that as geographic structuring increases, so does the chance of finding individual populations with signals of contraction (Fig. 7).

Overall the summary statistics have shown that the island model has estimates of F_{ST} close to the expected values. The stepping stone model has higher estimates of F_{ST}. Unfortunately, we cannot compare the simulated results with analytical theory because there are no results for lattice models without periodic boundaries (Crow and Aoki 1984). In all cases, there is good evidence of population structure but no evidence of expansion. Furthermore, using the Mantel test, the island

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

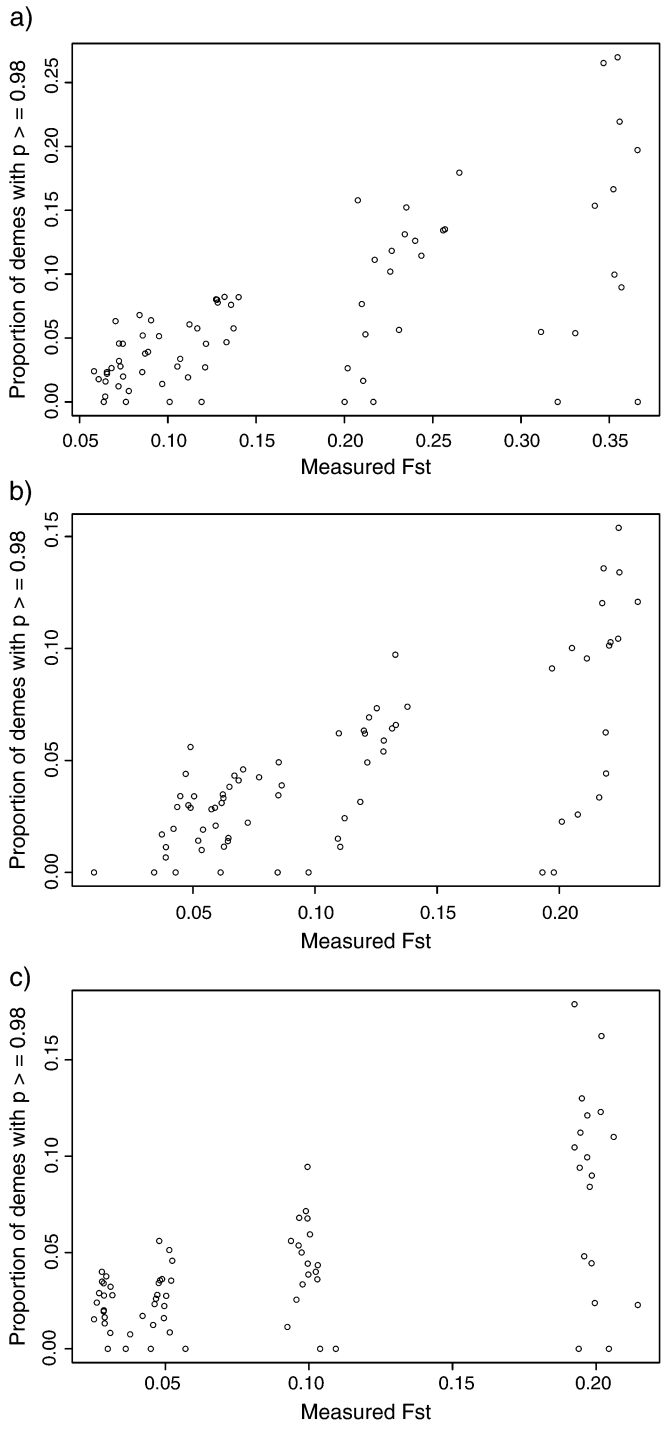

FIGURE 7. These charts show the measured F_{ST} against the proportion of demes with a significant positive Fu's Fs ($P \geq 0.98$) under a) the stepping stone model, b) the stepping stone model with some long-distance dispersal, and c) island model.

model shows no signals of isolation by distance, whereas an appreciable proportion of simulations of the stepping stone model show this pattern.

The Distribution of Statistics

So where do the inaccuracies in NCPA arise? By examining the distributions of statistical significance given by the NCPA statistics, and comparing them with the patterns described by Templeton et al. (1995), we can attempt to determine at which stage in the NCPA analysis errors occur. According to Templeton et al. (1995), the following patterns will be found in a clade if restricted gene flow is the cause of geographic association (the D_c and D_n statistics were originally defined in Templeton et al. 1995, but currently use the equations described by Posada et al. 2006 in the GeoDis software).

1. Significantly small D_cs primarily for tip clades. Some interior clades with significantly large D_cs.
2. $\bar{D}_c(I) - \bar{D}_c(T)$ significantly large.
3. Average D_cs should increase (and occasionally level off) with increasing clade level in a nested series of clades. If the distances level off, the null hypothesis of no geographical should no longer be rejected even though rejected at lower clade levels.
4. The above patterns also hold for the D_ns unless some gene flow is due to long-distance dispersal events, then significant reversals are of the above pattern can occur with the D_ns.

In general, although the patterns described by Templeton et al. (1995) can be found, other patterns also occur, indicating that it is not due to the inference key alone that false positives are made. Under all 3 scenarios, we see that for tip clades, the D_cs are primarily small, but for interior clades, the D_cs are also primarily small (Tables 5–7), which indirectly contrasts with the pattern given by Templeton et al. (1995). However, it is true that there are a proportion of interior clades with significantly large D_cs (coupled with significantly large D_ns), and so Pattern 1 is supported by the distribution of statistics, although it indicates that other patterns are present too. Pattern 2 is also generally supported by the statistics under all 3 models, but again other patterns are present. Pattern 3, which requires a more detailed look at the results, is interestingly only used to provide additional confirmation of restricted gene flow. Under Pattern 4, we expect to see the same patterns for D_n as we do for D_c unless long-distance dispersal occurs. Under the stepping stone model, we see that a greater proportion of D_n for tip clades are significantly large; however, when D_cs are significantly small, a greater proportion of tip clades have significantly small D_ns as well (which is why we see a large number of inferences of RGF_IBD). This also holds under the other 2 models, but with the proportions of significantly large D_ns and significantly small D_ns becoming more similar as more long-distance dispersal is involved (which is perhaps why we see less inferences of long-distance dispersal when more are occurring in the model). Under the stepping stone model, we also see that the proportions of significantly small $\bar{D}_n(I) - \bar{D}_n(T)$ and significantly large

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

TABLE 5. The proportions of significantly small (Dcs, Dns), nonsignificant (Dc, Dn), and significantly large (Dcl, Dnl) D_c and D_n statistics under the stepping stone model

		Tip			Interior			I-T		
		Dns	Dn	Dnl	Dns	Dn	Dnl	Dns	Dn	Dnl
0.03	Dcs	0.0903	0.1725	0.0410	0.1128	0.0901	0.0096	0.0434	0.0217	0.0030
	Dc	0.0174	0.5856	0.0592	0.0077	0.6539	0.0514	0.0575	0.5484	0.0225
	Dcl	0.0000	0.0000	0.0318	0.0000	0.0026	0.0720	0.0336	0.1749	0.0951
0.05	Dcs	0.1132	0.1963	0.0531	0.1308	0.1119	0.0182	0.0460	0.0273	0.0049
	Dc	0.0163	0.5228	0.0655	0.0064	0.5905	0.0600	0.0604	0.4845	0.0262
	Dcl	0.0000	0.0015	0.0312	0.0000	0.0024	0.0799	0.0423	0.1965	0.1119
0.1	Dcs	0.1447	0.2229	0.0873	0.1674	0.1368	0.0355	0.0516	0.0313	0.0098
	Dc	0.0153	0.4246	0.0726	0.0062	0.4869	0.0774	0.0730	0.3929	0.0343
	Dcl	0.0000	0.0008	0.0318	0.0000	0.0025	0.0873	0.0694	0.2020	0.1357
0.2	Dcs	0.1773	0.2316	0.1203	0.1930	0.1463	0.0632	0.0488	0.0322	0.0196
	Dc	0.0125	0.3540	0.0735	0.0057	0.4098	0.0870	0.0799	0.3396	0.0421
	Dcl	0.0000	0.0004	0.0303	0.0000	0.0019	0.0931	0.0906	0.1939	0.1532

Notes: The table shows the distribution of statistics for the tip clades, interior clades, and the I-T clades, as well as over each level of F_{ST} (0.03, 0.05, 0.1, 0.2) under the stepping stone model. The equations for the D_c and D_n statistics are defined in Posada et al. (2006).

TABLE 6. The proportions of significantly small (Dcs, Dns), nonsignificant (Dc, Dn), and significantly large (Dcl, Dnl) D_c and D_n statistics under the island model

		Tip			Interior			I-T		
		Dns	Dn	Dnl	Dns	Dn	Dnl	Dns	Dn	Dnl
0.03	Dcs	0.0517	0.0512	0.0075	0.0639	0.0262	0.0010	0.0440	0.0147	0.0005
	Dc	0.0200	0.7875	0.0339	0.0108	0.8096	0.0190	0.0309	0.7689	0.0165
	Dcl	0.0000	0.0000	0.0394	0.0000	0.0052	0.0644	0.0059	0.0585	0.0602
0.05	Dcs	0.0636	0.0656	0.0130	0.0774	0.0360	0.0029	0.0487	0.0156	0.0008
	Dc	0.0205	0.7434	0.0418	0.0096	0.7714	0.0252	0.0386	0.7203	0.0183
	Dcl	0.0000	0.0085	0.0437	0.0000	0.0049	0.0726	0.0097	0.0760	0.0720
0.1	Dcs	0.0828	0.0951	0.0242	0.0992	0.0512	0.0066	0.0565	0.0184	0.0022
	Dc	0.0184	0.6709	0.0523	0.0094	0.7039	0.0412	0.0477	0.6354	0.0192
	Dcl	0.0000	0.0067	0.0497	0.0000	0.0044	0.0841	0.0226	0.1051	0.0928
0.2	Dcs	0.1083	0.1204	0.0417	0.1289	0.0683	0.0137	0.0670	0.0213	0.0041
	Dc	0.0172	0.5800	0.0686	0.0084	0.6223	0.0545	0.0621	0.5490	0.0219
	Dcl	0.0000	0.0062	0.0577	0.0000	0.0031	0.1007	0.0330	0.1266	0.1149

Notes: The table shows the distribution of statistics for the tip clades, interior clades, and the I-T clades, as well as over each level of F_{ST} (0.03, 0.05, 0.1, 0.2) under the island model. The equations for the D_c and D_n statistics are defined in Posada et al. (2006).

TABLE 7. The proportions of significantly small (Dcs, Dns), nonsignificant (Dc, Dn), and significantly large (Dcl, Dnl) D_c and D_n statistics under the stepping stone model with some long-distance dispersal

		Tip			Interior			I-T		
		Dns	Dn	Dnl	Dns	Dn	Dnl	Dns	Dn	Dnl
0.03	Dcs	0.0617	0.0778	0.0161	0.0782	0.0466	0.0034	0.0449	0.0181	0.0009
	Dc	0.0216	0.7332	0.0438	0.0091	0.7614	0.0285	0.0403	0.7044	0.0184
	Dcl	0.0000	0.0000	0.0394	0.0000	0.0049	0.0679	0.0135	0.0894	0.0701
0.05	Dcs	0.0736	0.1011	0.0222	0.0905	0.0557	0.0070	0.0499	0.0182	0.0020
	Dc	0.0195	0.6886	0.0494	0.0094	0.7198	0.0380	0.0434	0.6509	0.0195
	Dcl	0.0000	0.0056	0.0401	0.0000	0.0041	0.0756	0.0173	0.1136	0.0851
0.1	Dcs	0.1003	0.1355	0.0389	0.1163	0.0748	0.0124	0.0571	0.0215	0.0039
	Dc	0.0195	0.5946	0.0614	0.0088	0.6432	0.0514	0.0541	0.5589	0.0233
	Dcl	0.0000	0.0041	0.0457	0.0000	0.0030	0.0901	0.0334	0.1387	0.1089
	Dcs	0.1271	0.1696	0.0627	0.1434	0.0982	0.0249	0.0594	0.0264	0.0077
0.2	Dc	0.0167	0.5024	0.0724	0.0085	0.5565	0.0706	0.0656	0.4668	0.0244
	Dcl	0.0000	0.0026	0.0465	0.0000	0.0026	0.0953	0.0522	0.1676	0.1299

Notes: The table shows the distribution of statistics for the tip clades, interior clades, and the I-T clades, as well as over each level of F_{ST} (0.03, 0.05, 0.1, 0.2) under the stepping stone model with some long-distance dispersal. The equations for the D_c and D_n statistics are defined in Posada et al. (2006).

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

$\bar{D}_n(I) - \bar{D}_n(T)$ are similar, although a slightly greater proportion are significantly small. The largest proportion, however, is when both $\bar{D}_c(I) - \bar{D}_c(T)$ and $\bar{D}_n(I) - \bar{D}_n(T)$ are significantly large supporting the pattern in Pattern 4. When long-distance dispersal is involved, we should see significantly large $\bar{D}_c(I) - \bar{D}_c(T)$ and significantly small $\bar{D}_n(I) - \bar{D}_n(T)$, and although this combination is present, other combinations are present in greater proportions. These tables show that the predicted patterns of Templeton et al. (1995) under restricted gene flow are not the only patterns that can be observed and highlight where part of the inaccuracy lies.

Discussion

NCPA remains a popular method (Garrick et al. 2009). It is a complex procedure, however, and deserves to be thoroughly tested. The studies in which it has been tested have themselves been subject to vigourous criticisms (see Templeton 2008, 2009a, 2009b). Before discussing the results that have been obtained in the current study, we would like to address some of the recent criticism of previous examinations of NCPA.

Recently, Templeton (2009b) has stated "The critics have come to different conclusions because they have focused on the pre-2002 versions of NCPA and failed to take into account the extensive developments in NCPA since 2002." In fact, only the simulation study of Knowles and Maddison (2002) (necessarily) uses an inference key prior to 2002. Before Templeton (2002), the only modifications to NCPA came from Crandall (1996) to the nesting algorithm and to the various updates of the inference key from Templeton (1998). The work of Panchal and Beaumont (2007) takes both these updates into account.

Templeton (2009b) has defended his tests of NCPA that are based on empirical data sets: "This method validates NCPA in the most relevant way: how it behaves with real data and actual historical events." In fact, unlike the case with simulated data, there is no certainty in any of the assumed demographic histories of these data, and therefore, any validation obtained in this way, uncorroborated by simulation-based tests, must be regarded as weak. Furthermore, as noted by Knowles (2008), the tests based on empirical data sets did not show how many times processes were inferred other than those that were expected. Templeton (2009b) has disputed this by emphasizing the point made in Templeton (2004b) that "*all* the inferences from NCPA with respect to historical events" are presented. The word "all" is emphasized in Templeton (2009b), but actually, it is "with respect to historical events" that is most significant: Templeton does not consider the inferences of restricted gene flow, and there are no data on any inferences other than historical events. This means that the false-positive rate is not an upper bound as claimed because there may have been many more inferences that were not considered tested and verified.

Templeton (2009b) has suggested that the figure of a 75% false-positive rate reported in Panchal and Beaumont (2007) is wrong for a number of reasons. First, he states "The first source of error arises from unrealistic simulations. The simulation program they used, SIMCOAL, only allows the use of unrealistic mutation models .... Simulations with unrealistic mutational models are known to generate false positive phylogeographic inferences (Palsboll et al. 2004)." Panchal and Beaumont (2007) use a Kimura 2-parameter model modified to allow for gamma-distributed mutation rates at each site. Although for particular data sets, other mutational models may be appropriate, as noted by Ripplinger and Sullivan (2008) the Kimura 2-parameter model is a widely used default model. In general, we have no idea what would be a "realistic" model, and no such model is implemented in NCPA. It should be noted that AMOVA, which uses haplotype information, showed a correct false-positive rate with these simulated data. Moreover, Templeton (2009b) miscites Palsboll et al. (2004). That paper simply shows that when using a model-based method, if data are simulated with a mutation model that is different from that assumed by the model, poor inferences may result.

Templeton (2009b) states "Moreover, in half of their simulations they assumed exhaustive sampling of every local deme in their species, and in the other half they assumed 50% coverage of all local demes .... These assumptions eliminate or minimize geographical sampling as a source of error, thereby creating artificial power." The false-positive rate for a panmictic model is not affected by sampling scheme because the false positives are generated by GeoDis and not the sorting of significance patterns by the inference key, which is the only place in NCPA in which the sampling scheme is important. Therefore, this cannot be a source of error.

In continuing his critique of Panchal and Beaumont (2007), Templeton (2009b) states "Second, Panchal and Beaumont (2007) do not use the inference key that has been legitimately validated by the criteria given in Knowles (2008); rather they use their own unvalidated inference algorithm. I know from personal experience that the inferences emerging from their algorithm can be discrepant with inferences from the validated inference key, and I know from direct communication that other users of NCPA have also encountered discrepancies." Knowles (2008) does not specify any criteria for validation, and this appears to be a miscitation. Furthermore, we would argue that Templeton's point is hearsay—unless results are published that show a clear discrepancy between the outcome of using ANeCA and an automated, replicable, implementation of NCPA that is deemed "correct" by Templeton, there can be no valid scientific discussion on this point. Panchal (2007) and Panchal and Beaumont (2007) have published an explicit algorithm laying out the precise details of the simulations, the code of which is publicly available. To date, there has been no explicit statement by Templeton of which steps in the algorithm are incorrect.

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

Templeton (2009b) has suggested that Beaumont and Panchal (2008) are now "rejecting the premise of the entire section of Panchal and Beaumont (2007) that compared their simulated results to actual results, now arguing that actual data and simulated data are so different that no valid comparison can be made .... Hence, all agree now that the 75% figure does not reflect the behaviour of NCPA with real data." This is a clever miscitation. The context in which this arises is that in Beaumont and Panchal (2008), we do not accept that the 75% false-positive figure we observed should be the same as the 23% asserted by Templeton (2008). The reason we do not accept this is that our simulations were based on panmixia, whereas the assumed "truth" in the studies by Templeton (2004b) always involved a complex history. We felt it self-evident that these different cases were not comparable in terms of false-positive rate: the false-positive rate will vary, as indeed is demonstrated by the present paper, depending on what is chosen to be the "true" history. All the inferences under panmixia are false in Panchal and Beaumont (2007), whereas only historical events not expected under prior knowledge are considered false in Templeton (2004b). Now, of course, the false premise in the argument of Templeton (2009b) is that because "real" data are generally not panmictic, the 23% figure is correct. However, as demonstrated in the current paper, even with nonpanmictic data, the false-positive rates can be as high as 99%. It is quite possible that the true demography assumed by Templeton is not, in fact, true and is actually unknown, and, as noted above, the 23% figure is falsely calculated because a large class of inferences were ignored. Moreover, we do not accept that our argument invalidates the comparisons that were made in Panchal and Beaumont (2007), where we showed that there was a statistically significant rank order correlation between the inferences made in the panmictic case and those obtained from real data. If NCPA has specific biases to make particular inferences regardless of the true history, this could still be evident regardless of the true history and hence false-positive rate.

Having addressed some of the earlier criticisms by Templeton in order to justify the current study, we now discuss the results in more detail. Earlier work (Knowles and Maddison 2002; Petit and Grivet 2002; Templeton 2004b; Pulquério 2005; Panchal and Beaumont 2007) shows that changes in parameters and assumptions affect inferences and consequently this study varies particular factors: the sample size, the number of demes simulated (corresponding to geographic locations), the proportion of demes sampled, the amount of structuring (measured though an estimator of F_{ST}), and the method used to calculate the summary statistics on the inferences made.

The results presented here show that NCPA is likely to infer false positives under models of gene flow. It also fails to discriminate adequately between long-distance and short-distance movements. We showed that there is information in the data that can be used by statistical methods. The results with 3 classical summary statistics suggested that there was readily interpretable information in the data. AMOVA was shown to have a false-positive rate of 5% when no structure was present (Panchal and Beaumont 2007) and increasing power to detect population structure as levels of genetic differentiation increased. The Mantel test had a false-positive rate of approximately 5% under the island model and inferred isolation by distance on average 40% of the time in the stepping stone model. By contrast, using single-locus NCPA, as previously found in the case of a panmictic model, RGF_IBD and CRE, continued to be the commonest inferences. However, in contrast to the panmictic model, all other inferences became more frequent that were otherwise rare when there was no structuring. With increasing levels of F_{ST} increased, there was an increased rate of rejection of the null hypothesis of no geographic association. The randomization procedure in GeoDis yields more positives when there is genetic structure, and our analysis also shows that with increasing genetic differentiation, there is an increase in the number of clades and therefore in the number of statistical tests, leading to an increased number of inferences. However, despite this increased rate of rejection of the null hypothesis of no geographic association, the GeoDis statistics themselves do not convey sufficient information for the inference key to have any discriminative power.

Although Templeton (2004b) claims that NCPA is reasonable under scenarios of fragmentation and range expansion (but see Knowles 2008), this may be due to the creation of fewer clades, especially because in a range expansion setting, haplotype networks will be more star like. However, even if NCPA is more accurate under those 2 scenarios, we have shown that those inferences will also arise under several gene flow models, and so makes NCPA unreliable as a method of discrimination between them. Furthermore, we have shown that although the patterns described by Templeton et al. (1995) do occur under restricted gene flow scenarios, other patterns also arise, and this is not taken into account anywhere in the NCPA methodology. This is also hindered by the lack of proof that the patterns given by Templeton solely occur under a given model. The lack of error estimates, or probability that other scenarios are plausible, makes NCPA much less functional than other model-based analyses (also highlighted by Knowles and Maddison 2002). As pointed out in Beaumont and Panchal (2008) (see also Templeton 2009b), correction for multiple testing is also not simple due to dependence between statistics within a clade and also because clades in different levels are not necessarily independent of each other (data from subclades are included again at higher levels of nesting).

We have shown that the implementation of multilocus NCPA has indeed reduced the false-positive rate, as claimed by Templeton (2004a). However, the false-positive rates are still high and depend on the model. In particular, the method seems to be unable to distinguish between restricted gene flow caused by isolation

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

by distance or caused by long-distance dispersal. Thus, the proportion of inferences are quite similar for the different models of population structure with reduced gene flow. The false-positive rates differ greatly because of how they are defined. Increasing levels of genetic differentiation lead to more cross-validated inferences. There were typically few inferences of CRE, allopatric fragmentation, and long-distance colonization, which would have been false positives under our scheme. However, we note that in these cases, additional levels of cross-validation are required (both temporally and geographically), which will reduce the frequency with which the method makes these inferences, and means the lower false-positive rate cannot be attributed to the methods ability to detect these processes effectively. We also note in passing that there are some problematic aspects to the multilocus procedure. The cross-validation criterion itself is arbitrary. No justification is given for the need for an inference to match at 2 loci. Perhaps it would be better to require a match at half the loci. Presumably, there will be some optimum requirement that depends on the data. The theory for dating the clades is based on the standard coalescent but is then applied to scenarios that explicitly do not correspond to it. Also, as noted by Rannala and Bertorelle (2001), clades within a genealogy do not follow the coalescent. Furthermore, there need be no expectation that the genealogical dates concord with demographic dates (Nielsen and Beaumont 2009).

A working hypothesis of the behavior seen is that NCPA is structurally predisposed to make certain inferences. The rate of making these inferences is dependent on the amount of geographic structuring. This is not surprising given that the method depends on rejection of hypotheses based on randomization tests that relate genetic to geographic distances (much as in the Mantel test). The method, however, appears unable to differentiate accurately between different scenarios (at least in the case considered here of geographically structured models at equilibrium). Cross-validation, although it generally restricts the sensitivity of the method (by requiring the data to jump through additional hoops), and thereby reduces the false-positive rate, does not appear to increase the accuracy of the approach.

In conclusion, having tested the NCPA method under a variety of settings (Panchal and Beaumont 2007, this study), we have demonstrated that the method is at best unreliable. It does appear to be sensitive to population structure, but a standard method such as AMOVA has much better statistical properties. It is, of course, perfectly normal that statistical methods are developed, are evaluated and criticized, and, in response to criticism, modified or superseded. The NCPA procedure continues to be strongly defended (Templeton 2008; Templeton 2009a, 2009b) and is unusual in that it has been particularly difficult to test because it has relied on lengthy manual procedures, as outlined in Panchal and Beaumont (2007). Thus, most of the published, confirmatory, tests have been performed by the originators of NCPA. A large number of studies have used NCPA (over 1600 articles citing the original paper; Petit 2008), and if it continues to be applied in future, it needs to be developed as a unitary, algorithmically defined, procedure that can be straightforwardly tested and compared by independent researchers. Without such development, we would urge extreme caution in its future use.

Supplementary Material

Supplementary appendices can be found at http://www.sysbio.oxfordjournals.org/.

Funding

The Biotechnology and Biological Sciences Research Council is gratefully acknowledged for funding to M.P. and the Natural Environment Research Council for funding to M.A.B.

Acknowledgements

We would like to thank J. Sullivan, L. L. Knowles, L. Excoffier, and 2 anonymous reviewers for their comments and suggestions on this manuscript.

References

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

Beaumont M.A., Balding D. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13:969–980.

Beaumont M.A., Nichols R.A. 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B, Biol. Sci. 263:1619–1626.

Beaumont M.A., Nielsen R., Robert C., Hey J., Gaggiotti O., Knowles L.L., Estoup A., Panchal M., Corander J., Hickerson M., Sisson S.A., Fagundes N., Chikhi L., Beerli P., Vitalis R., Cornuet J.-M., Huelsenbeck J., Foll M., Yang Z., Rousset F., Balding D., Excoffier L. 2010. In defense of model-based inference in phylogeography. Mol. Ecol. 19:436–446.

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

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

Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst. Biol. 56:400–411.

Clement M., Posada D., Crandall K.A. 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 leukemia/lymphoma virus type I sequences. Mol. Biol. Evol. 13:115–131.

Crow J.F., Aoki K. 1984. Group selection for a polygenic behaviourial trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. 81:6073–6077.

De A., Durrett R. 2007. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. Genetics. 176:969–981.

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

Excoffier L., Novembre J., Schneider S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010

interconnected populations with arbitrary demography. J. Hered. 91:506–509.
Fagundes N.J.R., Ray N., Beaumont M.A., Neuenschwander S., Salzano F.M., Bonatto S.L., Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. 104:17614–17619.
Fu Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. Genetics. 147:915–925.
Garrick R.C., Nason J.D., Meadows C.A., Dyer R.J. 2009. Not just vicariance: phylogeography of a Sonoran desert euphorb indicates a major role of range expansion along the Baja peninsula. Mol. Ecol. 18:1916–1931.
Jennings W., Edwards S. 2005. Speciational history of australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution. 59:2033–2047.
Kimura M., Weiss G. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics. 49:561–576.
Knowles L.L. 2008. Why does a method that fails continue to be used? Evolution. 62:2713–2717.
Knowles L.L. 2009. Statistical phylogeography. Annu. Rev. Ecol. Syst. 40:593–612.
Knowles L.L., Maddison W.P. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.
Nielsen R., Beaumont M.A. 2009. Statistical inferences in phylogeography. Mol. Ecol. 18:1034–1047.
Palsboll P.J., Berube M., Aguilar A., Notarbartolo-Di-Sciara G., Nielsen R. 2004. Discerning between recurrent gene flow and recent divergence under a finite-site mutation model applied to north atlantic and mediterranean sea fin whale (balaenoptera physalus) populations. Evolution. 58:670–675.
Panchal M. 2007. The automation of nested clade phylogeographic analysis. Bioinformatics. 23:509–510.
Panchal M., Beaumont M.A. 2007. The automation and evaluation of nested clade phylogeographic analysis. Evolution. 61:1466–1480.
Pesole G., Gissi C., De Chirico A., Saccone C. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. J. Mol. Evol. 48:427–434.
Petit R.J. 2008. The coup de grâce for the nested clade phylogeographic analysis? Mol. Ecol. 17:516–518.
Petit R.J., Grivet D. 2002. Optimal randomization strategies when testing the existence of a phylogeographic structure. Genetics. 161:469–471.
Posada D., Crandall K.A., Templeton A.R. 2000. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Mol. Ecol. 9:487–488.
Posada D., Crandall K.A., Templeton A. R. 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.
Rannala B., Bertorelle G. 2001. Using linked markers to infer the age of a mutation. Hum. Mutat. 18:87–100.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.
Ray N., Currat M., Excoffier L. 2003. Intra-deme molecular diversity in spatially expanding populations. Mol. Biol. Evol. 20:76–86.
Ripplinger J., Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57:76–85.
Rousset F. 2003. Inferences from spatial population genetics. In: Balding D. J., Bishop M., Cannings C., editors. Handbook of statistical genetics. 2nd ed. Vol. 2. Chichester (UK): John Wiley & Sons, Ltd.
Skarpaas O., Shea K., Bullock J.M. 2005. Optimizing dispersal study design by monte carlo simulation. J. Appl. Ecol. 42:731–739.
Takahata N., Lee S.-H., Satta Y. 2001. Testing multiregionality of modern human origins. Mol. Biol. Evol. 18:172–183.
Tavare S., Balding D.J., Griffiths R.C., Donnelly P. 1997. Inferring coalescence times from DNA sequence data. Genetics. 145:505–518.
Templeton A.R. 1998. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. Mol. Ecol. 7:381–397.
Templeton A.R. 2002. Out of Africa again and again. Nature. 416:45–51.
Templeton A.R. 2004a. A maximum likelihood framework for cross validation of phylogeographic hypotheses. In: Wasser S. P., editor. Evolutionary theory and processes: modern horizons. Dordrecht (The Netherlands): Kluwer Academic. p. 209–230.
Templeton A.R. 2004b. Statistical phylogeography: methods of evaluating and minimizing inference errors. Mol. Ecol. 13:789–809.
Templeton A.R. 2005. Haplotype trees and modern human origins. Yearb. Phys. Anthropol. 48:33–59.
Templeton A.R. 2008. Nested clade analysis: an extensively validated method for strong phylogeographic inference. Mol. Ecol. 17:1877–1880.
Templeton A.R. 2009a. Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. Mol. Ecol. 18:319 – 331.
Templeton A.R. 2009b. Why does a method that fails continue to be used: the answer. Evolution. 63:807–812.
Templeton A.R., Crandall K.A., Sing C.F. 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., Routman E., Phillips C.A. 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.
Williamson S., Hernandez R., Fledel-Alon A., Zhu L., Nielsen R., Bustamante C. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. 102:7882.
Wiuf C. 2003. Inferring population history from genealogical trees. Zool. Sci. 46:241–264.
Woods K.S., Davis M.B. 1989. Paleoecology of range limits: beech in the upper peninsula of Michigan. Ecology. 70:681–696.
Wright S. 1931. Evolution in Mendelian populations. Genetics. 16:97–159.

Downloaded from http://sysbio.oxfordjournals.org at MPI fuer Evolutionsbiologie at Ploen on May 25, 2010