

Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation

ALAN R. TEMPLETON

Department of Biology, Washington University, St. Louis, MO 63130-4899, USA

Abstract

Nested clade phylogeographical analysis (NCPA) and approximate Bayesian computation (ABC) have been used to test phylogeographical hypotheses. Multilocus NCPA tests null hypotheses, whereas ABC discriminates among a finite set of alternatives. The interpretive criteria of NCPA are explicit and allow complex models to be built from simple components. The interpretive criteria of ABC are ad hoc and require the specification of a complete phylogeographical model. The conclusions from ABC are often influenced by implicit assumptions arising from the many parameters needed to specify a complex model. These complex models confound many assumptions so that biological interpretations are difficult. Sampling error is accounted for in NCPA, but ABC ignores important sources of sampling error that creates pseudo-statistical power. NCPA generates the full sampling distribution of its statistics, but ABC only yields local probabilities, which in turn make it impossible to distinguish between a good fitting model, a non-informative model, and an over-determined model. Both NCPA and ABC use approximations, but convergences of the approximations used in NCPA are well defined whereas those in ABC are not. NCPA can analyse a large number of locations, but ABC cannot. Finally, the dimensionality of tested hypothesis is known in NCPA, but not for ABC. As a consequence, the 'probabilities' generated by ABC are not true probabilities and are statistically non-interpretable. Accordingly, ABC should not be used for hypothesis testing, but simulation approaches are valuable when used in conjunction with NCPA or other methods that do not rely on highly parameterized models.

Keywords: Bayesian analysis, computer simulation, nested clade analysis, phylogeography, statistics

Received 23 May 2008; revision received 28 October 2008; accepted 5 November 2008

Intraspecific phylogeography is the investigation of the evolutionary history of populations within a species over space and time. This field entered its modern era with the pioneering work of *Avise et al.* (1979, 1987) who used haplotype trees as their primary analytical tool. A haplotype tree is the evolutionary tree of the various haplotypes found in a DNA region of little to no recombination. The early phylogeographical studies made qualitative inference from a visual overlay of the haplotype tree upon a map of the sampling locations. As the field matured, there was a general recognition that the inferences being made were subject to various sources

of error, and the next phase in the development of intraspecific phylogeography was to integrate statistics into the inference structure.

One of the first statistical phylogeographical methods was nested clade phylogeographical analysis (NCPA) (*Templeton et al.* 1995). This method incorporates the error in estimating the haplotype tree (*Templeton et al.* 1992; *Templeton & Sing* 1993) and the sampling error due to the number of locations and the number of individuals sampled (*Templeton et al.* 1995). Since 2002, nested clade analysis has been radically modified and extended to take into account the randomness associated with the coalescent and mutational processes; to minimize false-positives and type I errors through cross-validation, and to test specific phylogeographical hypotheses through a likelihood ratio

Correspondence: Alan R. Templeton, Fax: 1 314 935 4432; E-mail, temple_a@wustl.edu

testing framework (Templeton 2002b, 2004a, b; Gifford & Larson 2008). Statistical approaches to phylogeography have also been developed through simulations of specific phylogeographical models for both hypothesis testing and parameter estimation (Knowles 2004).

Beaumont & Panchal (2008) have questioned the validity of NCPA and held up the approximate Bayesian computation (ABC) method (Pritchard *et al.* 1999; Beaumont *et al.* 2002) as an alternative. They specifically cite the application of ABC by Fagundes *et al.* (2007) as an example of how tests of phylogeographical hypotheses should be performed. Therefore, I will compare the statistical properties of the ABC method to those of NCPA with a focus upon the work of Fagundes *et al.* (2007).

Strong vs. weak inference

The two main types of statistical hypothesis testing are (i) testing a null hypothesis, and (ii) assessing the relative fit of alternative hypotheses. All inferences in NCPA start with the rejection of the null hypothesis of no association between haplotype clades with geography. When this null hypothesis is rejected, a biological interpretation of the rejection is sometimes possible through the application of an inference key (discussed in the next section). Knowles & Maddison (2002) pointed out that these interpretations from the inference key were not phrased as testable hypotheses. This deficiency of single-locus NCPA has been corrected in multilocus NCPA (Templeton 2002a, 2004a, b). All inferences in multilocus NCPA are phrased as testable null hypotheses, and the only inferences retained are those that are confirmed by explicit statistical tests. The multilocus likelihood framework also allows many other specific phylogeographical models to be phrased as testable null hypotheses. For example, the out-of-Africa replacement hypothesis that human populations expanded out of Africa around 100,000 years ago and drove to complete genetic extinction all Eurasian human populations can be phrased as a testable null hypothesis within multilocus NCPA. This null hypothesis is decisively rejected with a probability less than 10^{-17} with data from 25 human gene regions using a likelihood-ratio test (Templeton 2005, 2007b).

Nested clade analysis can also test null hypotheses about a variety of other data types. Indeed, nested clade analysis was first developed for testing the null hypothesis of no association between phenotypic variation with the haplotype variation at a candidate locus (Templeton *et al.* 1987). This robustness allows one to integrate hypothesis testing on morphological, behavioural, physiological, and ecological data with the phylogeographical analysis (e.g. Templeton *et al.* 2000b; Templeton 2001), despite erroneous claims to the contrary (Lemmon & Lemmon 2008). Moreover, the same multilocus statistical framework can test null hypothesis about correlated phylogeographical

events in different species, as illustrated by the likelihood-ratio test that humans and their malarial parasite shared common range expansions (Templeton 2004a, 2007b). The criticism that single-locus interpretations are not testable hypotheses is true but increasingly irrelevant as the entire field of phylogeography moves towards multilocus data sets. Given that the Fagundes *et al.* (2007) is a multilocus analysis, the legitimate comparison is multilocus ABC vs. multilocus NCPA and not with the pre-2002 single-locus version of NCPA criticized by Beaumont & Panchal (2008).

In contrast to multilocus NCPA, the ABC method posits two or more alternative hypotheses and tests their relative fits to some observed statistics. For example, Fagundes *et al.* (2007) used ABC to test the relative merits of the out-of-Africa replacement model of human evolution and two other models of human evolution (assimilation and multiregional). Of these three models of human evolution, the out-of-Africa replacement model had the highest relative posterior probability of 0.781.

Karl Popper (1959) argues that the scientific method cannot prove something to be true, but it can prove something to be false. In Popper's scheme, falsifying a hypothesis is strong scientific inference. The falsification of the replacement hypothesis by NCPA is an example of strong inference. Contrasting alternative hypotheses can also be strong inference when the alternatives exhaustively cover the hypothesis space. When the hypothesis space is not exhaustively covered, testing the relative merits among a set of hypotheses results in weak inference. A serious deficiency of weak inference occurs when all of the hypotheses being compared are false. It is still possible that one of them fits the data much better than the alternatives and as a result could have a high *relative* probability. The three basic models of human evolution considered by Fagundes *et al.* (2007) do *not* exhaustively cover the hypothesis space. For example, the model of human evolution that emerges from multilocus NCPA (Templeton 2005, 2007b) has elements from the out-of-Africa, assimilation, and multiregional models as well as completely new elements, such as an Acheulean out-of-Africa expansion around 650 000 years ago that is strongly corroborated by fossil, archaeological and palaeo-climatic data. Indeed, all the models given by Fagundes *et al.* (2007) have already been falsified by multilocus NCPA, so having a high *relative* probability does not mean that a hypothesis is true or supported by the data. The fact that the out-of-Africa replacement hypothesis is rejected with a probability of less than 10^{-17} is compatible with the out-of-Africa replacement hypothesis having a probability of 0.781 *relative* to two alternatives that have also been falsified. In the Popperian framework, the strong falsification of the out-of-Africa replacement has precedent over the weak *relative* support for the out-of-Africa replacement model against other falsified alternatives.

It is important to note that the difference between strong and weak inference is separate from the difference between likelihood and Bayesian statistics. Both likelihood and Bayesian methods can and have been used for both strong and weak inference. Indeed, NCPA uses both Bayesian and likelihood approaches. The haplotype network that defines the nested design is estimated (with explicit error) by the procedure now known as statistical parsimony, which quantifies the probability of deviations from parsimony by a Bayesian procedure (Templeton *et al.* 1992).

In general, there are always many possible combinations of fragmentation events, gene flow processes, expansions, and their times so as to preclude an exhaustive set of alternative hypotheses. Accordingly, the first distinction between multilocus NCPA and ABC is that multilocus NCPA can yield strong inference whereas ABC yields weak inference. Weak inference also means that ABC can give strong *relative* support to a hypothesis that is false, and there is no way within ABC of identifying or correcting for this source of false-positives.

Interpretative criteria

A statistical test returns a probability value, but rarely is the probability value per se the reason for an investigator performing the test. In such circumstances, statistical tests need to be interpreted. Such interpretation is part of the inference process for both NCPA and ABC.

NCPA provides an explicit interpretative key that is based upon predictions from neutral coalescent theory coupled with considerations from the sampling design. Because of explicit, a priori criteria, biological interpretations in NCPA are based on the same criteria for all species. The explicit nature of the interpretations has also allowed many users to comment upon ways of improving the interpretation of the statistical results and to allow the interpretations to be extensively validated by a set of 150 positive controls (Templeton 2004b, 2008). As a result, the interpretative key is dynamic and open sourced, constantly being improved and subject to re-evaluation, just as the ABC method has been revised over the years. The interpretative key focuses upon the simple types of processes and events that operate in evolutionary history. A statistical analysis of the 150 positive controls also showed that there is no statistical interference among the components (Templeton 2004b). The overall phylogeography of the sampled organism emerges as these simple individual processes and events are put together. As a consequence, NCPA allows the discovery of unexpected events and processes and the inference of complex scenarios from simple events.

The use of positive controls to validate the inference key also revealed that the false-positive rate of single-locus NCPA exceeded the nominal 5% level (Templeton 1998,

2004b). One of the prime motivations for the development of multilocus NCPA was to eliminate the false-positives from single-locus inferences through another series of statistical tests that cross-validate the initial set of single-locus inferences. The false-positive rates for single locus NCPA are therefore irrelevant to multilocus NCPA. These additional tests also mean that *every* inference in multilocus NCPA has been treated as a testable null hypothesis within a maximum likelihood framework.

The biological interpretations in ABC are defined by the finite set of models that are simulated: no biological interpretations outside of this set are allowed. This interpretative set is defined in an ad hoc, case-by-case basis. The interpretative set in ABC consists of models that specify an entire phylogeographical history rather than just phylogeographical components, as in NCPA. Because the interpretations are strictly limited to a handful of fully pre-specified phylogeographical models, no novelty or discovery is allowed in ABC.

An illustration of the sensitivity of the simulation approaches to their interpretative set is provided by a contrast of Ray *et al.* (2005) vs. Eswaran *et al.* (2005). Both papers used computer simulations to measure the goodness-of-fit of several models of human evolution, including the out-of-Africa replacement model. However, Eswaran *et al.* (2005) included in their interpretative set a model of isolation by distance with a small number of genes under natural selection, a model not included in the interpretative set of Ray *et al.* (2005). Whereas Ray *et al.* (2005) concluded that the out-of-Africa replacement model was by far the best-fitting model, Eswaran *et al.* (2005) concluded that the isolation-by-distance/selection model was so superior that the replacement model was refuted. Indeed, Eswaran *et al.* (2005) estimated that 80% of the loci in the human genome were influenced by admixture vs. the 0% of the replacement model. Such dramatically different conclusions are not incompatible with one another because the interpretative sets were different in these two studies. This heterogeneity in *relative* fit as a function of different interpretative sets is an unavoidable feature of weak inference.

Another disadvantage of ABC is that the interpretative set cannot be validated, unlike the interpretative key of NCPA. First, positive controls cannot be used to validate ABC. The interpretations in NCPA are highly amenable to validation via positive controls because the units of inference are individual processes or events and not the entire phylogeographical history of the species under study. For example, suppose a species today occupies in part a region that was under a glacier 20 000 years ago. In such a case, one can be confident that the range of the species expanded into this formerly glaciated region sometime in the last 20 000 years. Consequently, this species can be used as a positive control for the inference of range expansion. There may be many other aspects of the species'

phylogeographical history, but a complete knowledge of the species phylogeography is not needed for it to be used as a positive control in NCPA. In contrast, ABC specifies the entire phylogeographical history of the sample under consideration. Although prior knowledge of specific events is commonplace, prior knowledge of the complete phylogeography of a species is rare, making ABC less amenable to positive controls than NCPA. There is also a logical difficulty to using positive controls with ABC. The only way for ABC to yield the correct phylogeographical model is to have the correct phylogeographical model in the interpretative set. Since by definition, one knows the correct model in a positive control, this is easy to achieve, but it circumvents the primary cause of false-positives in ABC: failing to include the correct model in the interpretative set. This same logical problem also means that computer simulations cannot determine the false-positive rate of ABC. Because NCPA does not pre-specify its inference, false-positive rates are easily determined (Templeton 1998, 2004b) and therefore corrections can and have been implemented: modifications of the original inference key, the development of multiple test corrections for single locus NCPA for both inferences across nesting clades and (if desired) within nesting clades (Templeton 2008), and the development of multilocus NCPA that eliminates false-positives through cross-validation. This ability to quantify and deal with false-positives is a large statistical advantage of NCPA over any method of weak inference in which the false-positive rate cannot be determined even in principle.

The impact of implicit assumptions

Another advantage of NCPA is its transparency. Single-locus NCPA is based on testing simple null hypotheses using well-defined and long established permutational procedures (Edgington 1986). The inference key is explicit. Multilocus NCPA takes the inferences emerging from multiple single locus studies on the same species and subjects them to an explicit cross-validation and hypothesis testing procedure. The multilocus tests are based on explicit probability distributions and likelihood ratios, a standard and well-established statistical procedure. The complexity of the final multilocus NCPA model, such as that for human evolution (Templeton 2005, 2007a, b), is built up from simple inferences, with each individual inference being tested as a null hypothesis in the cross-validation procedure, making it obvious how the final model was achieved and its statistical support.

In contrast, the interpretive set for ABC and other simulation approaches start with the final, complex phylogeographical models that by necessity contain many assumptions and parameters in a confounded manner. For example, Fagundes *et al.* (2007) estimated the 95% highest posterior density (a Bayesian analogue of a 95% confidence interval)

of the admixture parameter between archaic Africans with archaic Eurasians to be 6.3×10^{-5} to 0.023 under their assimilation model. This statistical claim is based on only eight East Asian individuals to represent all of Eurasia (Eurasia is the potentially admixed population) and 50 loci. Such small admixture values are difficult to estimate accurately, and current admixture studies generally recommend samples in the several hundreds with thousands of loci (Bercovici *et al.* 2008). The sample sizes are so small and the geographical sampling so sparse that the data set given by Fagundes *et al.* (2007) does not achieve the minimums given in Templeton (2002b) for testing the out-of-Africa hypothesis with NCPA. I was only able to replicate their claims about the admixture parameter if I assumed that archaic Eurasians and archaic Africans were highly genetically differentiated and treated the eight East Asians as the population of inference rather than being a sample from the Eurasian human population (more on sampling in the next section).

This assumption of extreme genetic differentiation between archaic Africans and Asians was confirmed by one of the co-authors (L. Excoffier, personal communication), who explained that it arose from information in Fig. 1 in the published paper coupled with information about parameter values in table 7 from the online supplementary material. Figure 1 in Fagundes *et al.* (2007) shows that they modelled archaic Africans and archaic Eurasians as being completely isolated populations for a long period of time *before* the admixture event. Supplementary table 7 shows that they assumed that this period of isolation began between 32 000 to 40 000 generations ago from the present and lasted until the out-of-Africa expansion that they modelled as occurring between 1600 to 4000 generations ago. Assuming a generation time of 20 years for these archaic populations (Takahata *et al.* 2001), this translates into a period of genetic isolation lasting between 560 000 to 768 000 years during which genetic differences could accumulate. Extreme genetic differentiation is ensured by their assumptions of small population size (an average of 5050) in both archaic populations. These assumed population sizes would result in an average coalescence time for an autosomal locus ($4N$) of 20 100 generations (402 000 years). This average coalescence time is shorter than the interval of isolation, which is essential to create high levels of genetic divergence. Moreover, they assumed that the initial colonization of Eurasia by *Homo erectus* involved between 2 to 5000 individuals, and this bottleneck would further reduce coalescence times in the archaic Eurasian population and enhance genetic differentiation between archaic Eurasians and archaic Africans. These choices of parameter values produce the extreme genetic differentiation that is necessary to obtain their published confidence interval on the admixture rate from only 8 individuals and 50 loci. However, these same parameter

values lead to discrepancies with what is known about coalescence times of autosomal loci in humans. For example, fig. 3 in Templeton (2007a) shows the estimated coalescence times of 11 autosomal loci, all of which are greater than 402 000 years, and indeed 10 of the 11 have coalescence times greater than 1 million years. Similarly, Takahata *et al.* (2001) estimated the coalescence times of four autosomal loci, all of which were greater than 402 000 years, and two of which were greater than a million years. Takahata *et al.* (2001) used a 5-million-year-ago calibration date for the human/chimpanzee split, whereas this split is now commonly put at 6 million years ago because of better fossil data. Using the newer calibration point, three out the four autosomal loci in Takahata's analysis now have coalescence times greater than 1 million years ago. In either case, it is patent that the parameter values chosen by Fagundes *et al.* (2007) are strongly discrepant with the empirical data on autosomal coalescence times.

There are two basic reasons for rejecting a model in ABC. One, the model is wrong (or at least more wrong than the alternatives to which it is compared); and two, the biological model is correct but the simulated parameter values are wrong. It is not clear if the rejection of the assimilation model by Fagundes *et al.* (2007) is due to the assimilation model being wrong or is due to the unrealistic parameter values they chose for this model. These two causes of poor fit are confounded in ABC, thereby preventing meaningful biological interpretations of the rejection.

The rejection of the assimilation model by Fagundes *et al.* (2007) reveals another fundamental weakness of the ABC inference structure. Although Fagundes *et al.* (2007) interpreted their rejection of the assimilation model in terms of its admixture component, their assimilation model also includes the long period of prior isolation between archaic Eurasians and Africans. There is no necessity to couple admixture with a long period of complete genetic isolation. Indeed, in the model of human evolution that emerges from multilocus NCPA, there is both admixture and gene flow with no extended period of isolation. ABC simulates the entire, complex phylogeographical scenario as a single entity. Thus, the rejection of the assimilation model could be due to the assumed extended period of isolation in the model being wrong and *not* be due to the admixture component. The small confidence interval they obtained for the admixture rate does not discriminate between these two biological interpretations because this small confidence interval arises directly from their assumption of a period of prior isolation. Thus, these two components of their assimilation model are confounded in their simulations and no clear biological interpretation for the reason for rejecting the assimilation model is possible.

In contrast, multilocus NCPA can test individual features of complex models. For example, with 95% confidence, archaic Eurasians and archaic Africans have been exchange-

ing genes under an isolation-by-distance model with no significant interruption for the past 1.46 million years (Templeton 2005). Phrased as a null hypothesis, this means that we can reject the null hypothesis of isolation between these archaic human populations over the past 1.46 million years at the 5% level of significance. The statistical framework of multilocus NCPA is easily extended to test the null hypothesis of no gene flow (isolation) between two geographical areas in an arbitrary interval of time, say l to u . Let t_i be the random variable describing the possible times of gene flow inferred from locus i between two geographical areas of interest, T_i the estimated mean time of an NCPA inference of gene flow from locus i between the two areas, and k_i the average pairwise mutational divergence between haplotypes that arose since T_i at locus i . Then, using the gamma distribution described in Templeton (for more details see Templeton 2002b, 2004a), the probability of no gene flow from locus i in the interval l to u is:

Pr(no gene flow in the interval $[l, u]) =$

$$1 - \int_l^u \frac{t_i^{k_i} e^{-t_i(1+k_i)/T_i}}{\left(\frac{T_i}{1+k_i}\right)^{1+k_i} \Gamma(1+k_i)} dt_i \quad (\text{eqn 1})$$

Under the null hypothesis of isolation of the two areas in the time interval l to u , the probability of no gene flow is 1. Hence, if j is the number of loci that yield an inference of gene flow between the two areas of interest, the likelihood ratio test (LRT) of the null hypothesis of isolation between the two areas in the time interval l to u is:

LRT(isolation in $[l, u]) =$

$$-2 \sum_{i=1}^j \ln \left[1 - \int_l^u \frac{t_i^{k_i} e^{-t_i(1+k_i)/T_i}}{\left(\frac{T_i}{1+k_i}\right)^{1+k_i} \Gamma(1+k_i)} dt_i \right] \quad (\text{eqn 2})$$

with the degrees of freedom being j because the null hypothesis that all loci yield a probability of 1 is of zero dimension.

The minimum interval of isolation in the assimilation model of Fagundes *et al.* (2007) is between 4000 generations (80 000 years ago) to 32 000 generations (640 000 years ago) from their supplementary table 7. As this minimum interval is fully contained within their broader interval, rejection of the null hypothesis for the minimum interval automatically implies rejection of isolation in the broader time interval, so multiple testing is not necessary. With the data given in Templeton (2005, 2007b) on 18 cross-validated inferences of gene flow between African and Eurasian populations during the Pleistocene, the null hypothesis of complete genetic isolation between Africans

and Eurasians during this time interval is rejected with a likelihood-ratio test value of 30.02 and 18 degrees of freedom, yielding a probability level of 0.0094. Hence, the isolation component of their assimilation model has been strongly falsified by testing it as a null hypothesis.

The second component of their confounded assimilation model is admixture with the expanding out-of-Africa population. The null hypothesis of no admixture (replacement) is rejected with a probability level of less than 10^{-17} (Templeton 2005). Other recent studies (Eswaran *et al.* 2005; Garrigan *et al.* 2005; Plagnol & Wall 2006; Garrigan & Kingan 2007; Cox *et al.* 2008) also report evidence of admixture. Hence, the part of the assimilation model of Fagundes *et al.* (2007) that is wrong (and that is also discrepant with observed autosomal coalescence times) is the part concerning prior isolation of small populations of archaic Africans and Europeans and *not* the admixture component. The ability of the robust hypothesis-testing framework of NCPA to separate out different phylogeographical components is a great advantage over ABC that produces inferences that have no clear biological interpretations when complex phylogeographical hypotheses are being simulated.

Even assumptions not directly related to the phylogeographical model can influence phylogeographical conclusions. Programs such as SimCoal, used by Fagundes *et al.* (2007), specifically exclude mutation models that depend on the nucleotide composition of the sequence. However, mutation in humans is highly nonrandom, and *all* of this nonrandomness depends upon the nucleotide composition (Templeton *et al.* 2000a and references therein). Hence, the mutational models used in SimCoal are known to be unrealistic for human nuclear DNA. Unrealistic mutation models in turn can influence phylogeographical inference. For example, Palsbøll *et al.* (2004) surveyed mitochondrial DNA (mtDNA) in fin whales (*Balaenoptera physalus*) from the Atlantic coast off Spain and the Mediterranean Sea in order to test two alternative hypotheses: recent divergence with no gene flow vs. recurrent gene flow. They discovered that if they used a finite sites mutation model, they inferred recurrent gene flow, but when they used an unrealistic infinite sites mutation model, they inferred the recent divergence model. Until there is a more general assessment of the problems documented by Palsbøll *et al.* (2004), it is unwise to accept any inferences from simulation programs such as SimCoal that use unrealistic models of mutation. In contrast, NCPA does not have to simulate a mutational process but simply uses the mutations that are inferred from statistical parsimony. Statistical parsimony, the first step in the cladistic analysis of haplotype trees, has proven to be a powerful analytical tool in revealing these nonrandom patterns of mutation (Templeton *et al.* 2000a) precisely because it is robust to complex models of mutation. The phylogenetic ambiguities that nonrandom mutagenesis

creates can then be incorporated into tree association tests (Templeton & Sing 1993; Brisson *et al.* 2005; Templeton *et al.* 2005).

Sampling error

The field of statistics focuses upon the error induced by sampling from a population of inference. In NCPA, the sampling distributions of the primary statistics are estimated under the null hypothesis by a random permutation procedure. The statistical properties of permutational distributions have been long established in statistics (Edgington 1986), and NCPA implements the permutational procedure in such a manner as to capture both sources of sampling error; a finite number of individuals being sampled, and a finite number of locations being sampled. Other suggested permutational procedures (Petit & Grivet 2002; Petit 2008) do not capture all of the sampling error under the null hypothesis, and such flawed procedures can create misleading artefacts (Templeton 2002a). Hence, sampling error is fully and appropriately taken into account by NCPA.

Another source of error in NCPA is the randomness associated with the coalescent process itself and the associated accumulation of mutations, which will be called evolutionary stochasticity in this paper. The multilocus NCPA explicitly takes evolutionary stochasticity into account through the cross-validation procedure, and variances induced by both the coalescent and mutational processes are explicitly incorporated into the likelihood-ratio tests used for cross-validation and for testing specific phylogeographical hypotheses.

There is a misconception that the single-locus NCPA ignores evolutionary stochasticity and equates haplotype trees to population trees (Knowles & Maddison 2002). However, the nested design of NCPA insures that most inferences focus upon just one branch and haplotypes near that branch. This local inference structure gives NCPA great robustness to evolutionary stochasticity and ensures that inferences are not based upon the overall topology of the haplotype tree. Figure 1 illustrates this robustness with respect to five haplotype trees (mtDNA, Y-DNA, and three X-linked regions) from African savanna, African forest, and Asian elephants (Roca *et al.* 2005). NCPA was performed on the five African haplotype trees, using the Asian haplotypes as the outgroup. In all five cases, a single fragmentation event was inferred at the locations shown by the arrows in each haplotype tree in Fig. 1. In all five cases, this fragmentation event primarily separated the savanna from the forest populations. The null hypothesis that all five genes are detecting a single fragmentation event is accepted (the log-likelihood ratio test is 1.497 with 4 degrees of freedom, yielding a probability tail value of 0.8272). Hence, the same inference is cross-validated from

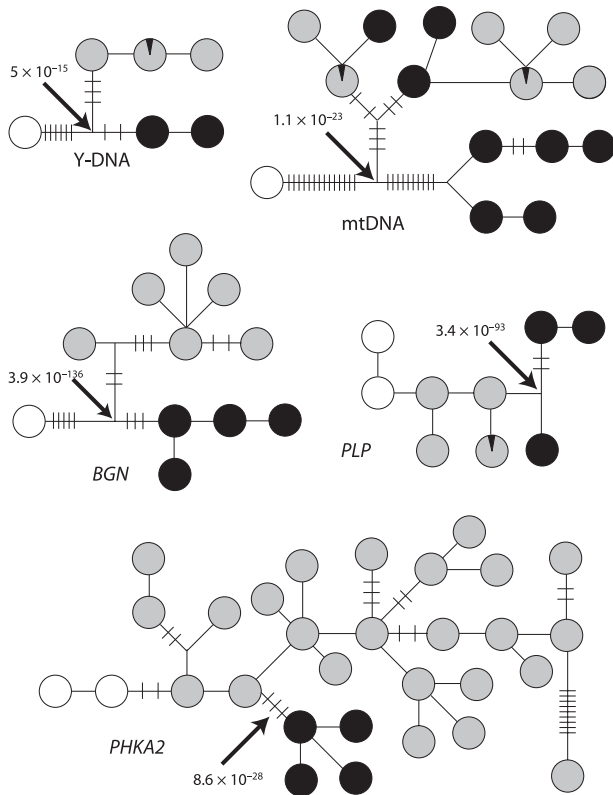


Fig. 1 Five haplotype trees estimated from samples of African savanna elephants (black), African forest elephants (grey), and Asian elephants as an outgroup (white circles) by Roca *et al.* (2005). A black arrow shows the points in the haplotype trees at which a significant fragmentation event was inferred by NCPA that primarily separates the forested and savanna areas of Africa.

five different haplotype trees. Note from Fig. 1 that each of the five trees differs in branch lengths and in the topological positions of the three taxa (Fig. 1). Only BGN corresponds to a clean species tree. Hence, four of the five trees were influenced by lineage sorting and/or introgression, yet the inferences from single-locus NCPA were completely robust to these potential complications, and the multilocus cross-validation procedure confirms this robustness through a likelihood-ratio test. This example also illustrates that NCPA inference does not depend upon equating a haplotype tree to a population tree as all five haplotype trees in this case would yield different population trees under such an equation. Indeed, NCPA does not even assume that a population tree exists at all, as turned out to be the case in human evolution that is characterized by a trellis of genetic interchanges rather than a population tree (Templeton 2005).

In ABC there is no null hypothesis, which complicates the computation of sampling error since there is no single statistical model under which to evaluate sampling error. Fortunately, Ewens (1983) clarified the sampling issues in

situations such as this. He pointed out that there are two basic categories of statistics used in population genetics: 'long-term' statistics whose sampling error includes both the randomness of the evolutionary process that created the current population (evolutionary stochasticity) and the error associated with a finite sample from the current population; and 'current generation' statistics whose sampling error includes only that associated with a finite sample from the current generation. The errors associated with these two types of statistics are shown pictorially in Fig. 2. This clarification of sources of sampling error led to a whole new generation of statistics in population genetics, such as the Tajima's *D* statistic (Tajima 1989). Tajima's *D* statistic, and many subsequent ones, can be expressed generically as:

$$\| S_{cg} - S_{lt} \| \quad (\text{eqn 3})$$

where S_{cg} is the current generation statistic, S_{lt} is the long-term statistic, and $\|\bullet\|$ designates some sort of normalization operation. This is the same class of statistics used in the ABC method (Beaumont *et al.* 2002), which is based on $\|s' - s\|$ where s is the observed value from the sample from the current generation and s' is the value obtained from a long-term coalescent simulation that yields the current generation followed by a sampling scheme identical to that used to obtain s . Tajima (1989), like Ewens (1983), noted that to test the evolutionary model being invoked for the S_{lt} statistic, the goodness-of-fit statistic in equation 3 *must* take into account all sources of sampling error in *both* the S_{cg} and S_{lt} statistics (Fig. 2). Hence, Tajima (1989) calculated the evolutionary stochasticity and current generation sampling error of S_{lt} and the current generation sampling error of S_{cg} . It is at this point that a major discrepancy appears between ABC and Tajima's use of equation 3. In ABC, s' is the long-term statistic, and the contribution of evolutionary stochasticity and current generation sampling error is taken into account via the computer simulation. However, s is the current generation statistic, and it is treated as a fixed constant in the ABC methodology, thereby violating the known sources of error made explicit by Ewens (1983) and that are incorporated into population genetic statistics such as Tajima's *D* statistic (Tajima 1989). The impact of treating s as a fixed constant is to increase statistical power as an artefact. This increase in pseudo-power can be quite substantial for small sample sizes (see equation 29 in Tajima 1989), and explains why great statistical resolution is claimed for the ABC method based on miniscule sample sizes (Fagundes *et al.* 2007). Ignoring the sampling error of s undermines the statistical validity of all inferences made by the ABC method.

The pseudo-power achieved by ignoring sampling error in s also undermines the primary method of validation of the ABC: the analysis of simulated populations. Validating

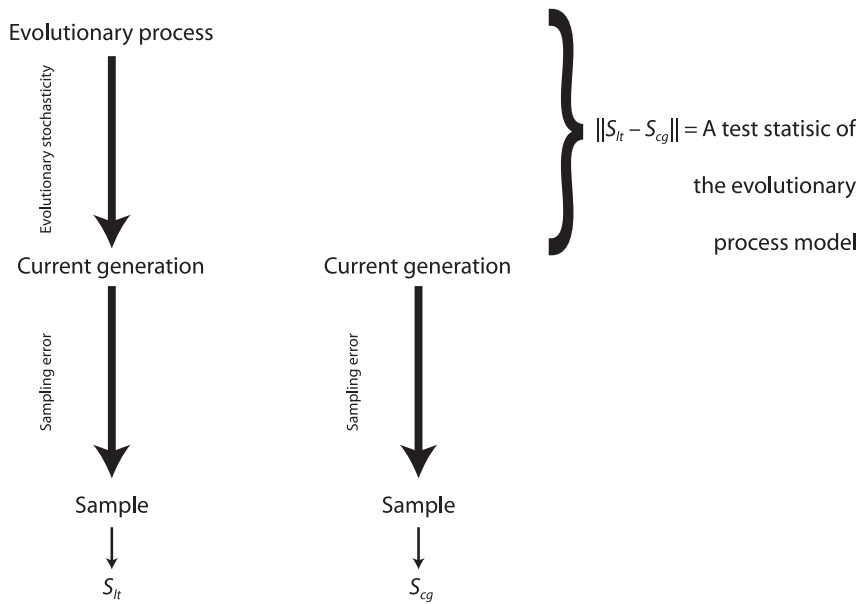


Fig. 2 A diagram of the sampling considerations made explicit by Ewens (1983) for long-term statistics (S_{lt}) and current generation statistics (S_{cg}). A contrast of these two types of statistics will focus its power on the evolutionary model used to generate the long-term statistic *only* if both types of statistics include the impact of sampling error from the current generation.

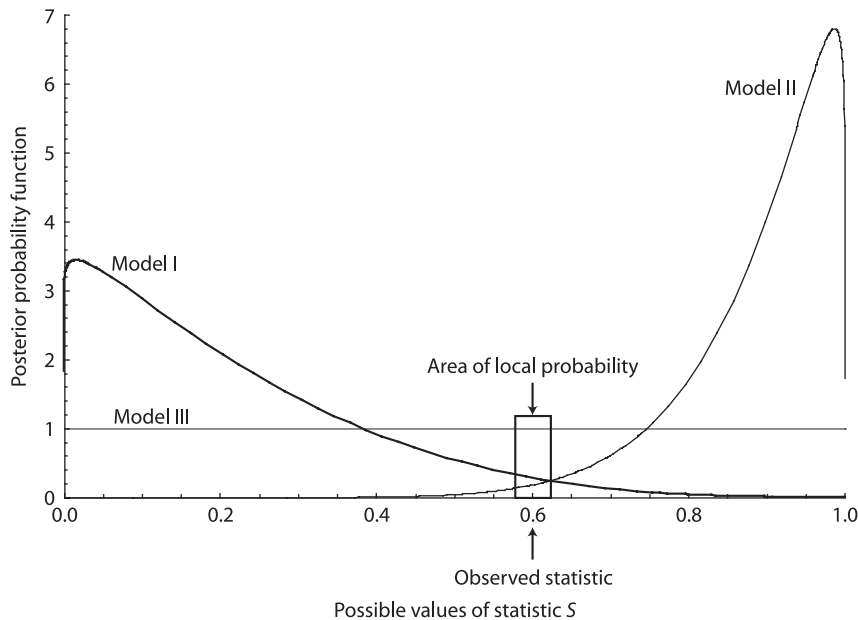


Fig. 3 Hypothetical posterior distributions for three models for a univariate statistic and an area around the observed value of statistic where local probabilities are evaluated.

a simulation method of inference with simulations hides the problem of unrealistic assumptions, as these assumptions are typically shared by both the analytical method and by the simulated data sets. The problems of weak inference and of ad hoc interpretative sets are ignored because the true model is typically included in the interpretative set. By including the true model in the interpretative set, the pseudo-power created by ignoring sampling error will cause the ABC method to appear to perform well with these simulated data sets. Unfortunately, when the truth is not fully known (i.e. real data), this pseudo-power can generate many false-positives and misleading inferences.

Full distributions vs. local probabilities

The random permutation procedure of NCPA generates the full sampling distribution under the null hypothesis of no geographical associations. In contrast, the ABC method only generates local probabilities in the vicinity of the observed statistics ($||s' - s|| \leq \delta$ where δ is a number that determines the level of acceptable goodness of fit). Bayesian inference is based upon having the posterior distribution, not just local probabilities. The local behaviour of the posterior distributions can sometimes be very misleading, as shown in Fig. 3. This figure shows a case in

which there is a single, one-dimensional statistic, and ABC concentrates its inference in a small region around the observed value. The complete posterior distributions for three models are shown in Fig. 3. Consider first models I and II. The observed statistic lies in the tails of both posterior distributions. This is probably a common situation. For example, Ray *et al.* (2005) simulated the out-of-Africa replacement model and other models of human evolution, but they found that even their best-fitting model only explained about 10% of the variation. Hence, the observed statistic falling on the tail of the posterior is not unlikely. Note that in Fig. 3, model I has a higher relative probability (the area under the curve in the local area indicated around the observed statistic) than model II. Yet, the observed statistic is actually closer to the higher values where model II places most of its probability mass than to the lower values where model I places most of its probability mass. Hence, if one had the entire posterior distributions, the conclusion about which of these two models was the better fitting might well be reversed.

However, it is model III in Fig. 3 that is the clear winner, having much more local probability mass than either models I or II. Note that model III has a flat posterior. This can occur for two reasons. First, it may be that model III is completely uninformative relative to the calculated statistic; that is, the data are irrelevant to this model. Alternatively, model III may be over-determined such that it has so many parameters relative to the sample size that it can fit equally well to any observed outcome. As Fig. 3 illustrates, the ABC method cannot distinguish between a truly good fitting model, an uninformative model, or an over-determined model. This is the danger of using only local relative probabilities rather than true posterior distributions.

Sample size

NCPA is computationally efficient, so even large sample sizes can be handled on a desktop computer. For example, the multilocus NCPA of human evolution included data sets with up to 42 locations and sample sizes of 2320 individuals (Templeton 2005), and this is not the limit for NCPA.

Because ABC must simulate every population in a complex model, ABC has severe constraints on sample size. For example, Fagundes *et al.* (2007) had the genetic data on all 50 loci for five populations: 10 sub-Saharan Africans, 10 Europeans, 2 East Indians, 8 East Asians, and 12 American Indians. However, the data from the European and East Indian Eurasian populations were 'excluded from the analysis to avoid incorporating additional parameters in our scenarios' (SI text of Fagundes *et al.* 2007). The inability of ABC to deal with just five locations and a total sample size of 42 individuals in this case is a serious defect for any phylogeographical tool. As shown by the inference

key for NCPA, many types of phylogeographical inference cannot be made if few sites are sampled, such as distinguishing fragmentation from isolation by distance.

There is no obvious statistical rationale for the populations actually analysed by Fagundes *et al.* (2007). Given that the American portion of their simulations is identical in all three models, the 12 American Indians are irrelevant to discriminating among three models of human evolution that they tested. The three models differ only in their relationship between African and Eurasian populations, yet the majority of their Eurasian sample was excluded from the analysis. The 40% of their data that is irrelevant to the tested models flattens the posterior. The informative subset of the data consists of just two geographical regions and 18 individuals. Because of the geographical hierarchy in sampling, the degrees of freedom in the informative data subset is bounded between 2 and 18, making over-determination of their models a real possibility as the number of parameters varied from 10 to 18 after excluding the parameters related to the Americas.

Approximate probabilities in NCPA vs. ABC

NCPA approximates the null distribution via a well-characterized random permutation procedure (Edgington 1986). The convergence of the approximation depends upon the number of random permutations performed. The default in NCPA is 1000 random permutations, which is adequate for most statistical inference. However, if any user wants more precision in the approximation, it is easily manipulated simply by increasing the number of permutations.

ABC approximates its local posterior probabilities via random computer simulations of complex scenarios. The convergence of this approximation depends upon the parameter δ (Beaumont *et al.* 2002). However, Beaumont *et al.* (2002) do not examine the convergence properties and only give ad hoc, heuristic guidelines in choosing δ . They also do not investigate the impact of sample size on the convergence of these probability measures. A Euclidian space is defined for the observed statistics even though many of the statistics used do not naturally fall into a Euclidian space, such as the number of segregating sites. As shown by Billingsley (1968), convergence of probability measures depends in part on the type of space being used, and it is not clear what mixing statistics that naturally fall into different types of spaces would do to the convergence properties. Finally, a spherical Euclidian space is assumed. Spherical spaces should be invariant to rigid rotations and reflections of their axes, yet the highly correlated nature of the statistics being used patently violates the assumption of sphericity. The impact of all of these factors upon the convergence properties of these probability measures is not addressed by Beaumont *et al.* (2002).

Table 1 A hypothetical two-by-two contingency table

	A1	A2
B1	52	20
B2	1	3

Approximations are common in statistics, but they should only be used when the limits of the approximation are known. For example, Table 1 gives a hypothetical two-by-two contingency table. One can test the null hypothesis of homogeneity in such a table with a chi-squared goodness-of-fit, which approximates the null distribution under some conditions. Doing so yields a chi-squared of 4 with 1 degree of freedom, which is significant at the 0.05 level. However, the bottom row has too few observations in this case for the chi-squared approximation to be valid. Instead, Fisher's exact test should be used in this situation, and with this test, there is no significant departure from homogeneity at the 0.05 level. The ABC method, like any approximation in statistics, should never be used unless the user is confident that their choices for δ , sample size, and statistics allow the approximation to be valid. This was not done by Fagundes *et al.* (2007).

Dimensionality and co-measurability

How well a model fits a given data set depends not only upon the extent to which the model captures reality, but also upon the dimensionality of the model relative to the size and structure of the sample. For example, a Hardy–Weinberg model can always fit exactly the phenotypic data from a one-locus, two allele, dominant allele model regardless of whether or not the population is in Hardy–Weinberg. Dominance means that there is only one degree of freedom available from the data, and one degree of freedom is used to estimate the allele frequency under Hardy–Weinberg. The resulting goodness-of-fit statistic indicates a perfect fit, but the statistic has zero degrees of freedom. Hence, the perfect fit of the Hardy–Weinberg model to a dominant allele model is meaningless statistically. As this example illustrates, the goodness-of-fit of a model and its statistical support can be very different.

NCPA builds up its inferences from simple, one-dimensional tests at the single locus level. The cross-validation statistics and the tests of specific phylogeographical hypotheses can be of higher dimension in NCPA. All of these higher dimension tests are nested in that the null hypothesis can be regarded as a lower-dimension special case relative to alternative hypotheses. This is important because the statistical theory for testing nested hypotheses is well developed and straightforward whereas it is far

more difficult to compare hypotheses that are not nested. Moreover, the parameters in the alternative models all play a similar role with respect to their respective probability distributions. As a consequence of nesting and comparable parameterization, the degrees of freedom in the NCPA multilocus tests are easy to calculate, so over-determined models can be avoided. For example, in testing the out-of-Africa replacement model with likelihood-ratio test, the degrees of freedom were 17, indicating that much information was available in the multilocus data set to test this hypothesis.

Unlike NCPA, ABC invokes complex phylogeographical models at the onset. ABC then uses a goodness-of-fit criteria upon these models. Indeed, $\|s' - s\|$ is a generalized goodness-of-fit statistic that includes the standard chi-squared statistic of goodness-of-fit. The goodness-of-fit nature of the ABC is further accentuated by the application of the Epanechnikov kernel and local regression (Beaumont *et al.* 2002). How close a simulated model will approximate the observed statistics will depend in part upon the number and nature of the parameters in the model and upon the size and structure of the data set. This information is needed in order to interpret the goodness-of-fit statistic. This is not only true of likelihood-ratio tests, but of Bayesian procedures as well (Schwarz 1978). The dimensionality of the models used in ABC has not been determined in any application that I have seen, and certainly not in Fagundes *et al.* (2007). Because the models are complex, determining dimensionality is not just a simple task of counting up the number of parameters, as different parameters influence the models in qualitatively different fashions and interact with one another in the simulation. Further complicating dimensionality of test statistics is the fact that the models in ABC are often not nested, and one model may contain parameters that do not have analogues in the other models and vice versa. Finally, the data are often sampled in a hierarchical fashion in phylogeographical studies, making the calculation of the available degrees of freedom difficult.

Knowing hypothesis dimensionality is critical for valid statistical inference. For example, suppose two models were being evaluated with a traditional chi-squared goodness-of-fit statistic, and model I yields a chi-squared statistic of 5 and model II yields a chi-squared statistic of 10. Which is the better fitting model? Since the chi-squared statistic decreases in value as the model fits better and better, it would be tempting to say that model I is the better fitting model as its chi-squared goodness-of-fit statistic is smaller. However, suppose that model I has 1 degree of freedom and model II has 5. Using this information, we can transform the chi-squared statistics into tail probabilities, yielding a probability of 0.025 for model I and 0.075 for model II. Hence, model II actually fits the data better than model I.

This hypothetical example illustrates the importance of a property called co-measurability. Co-measurability requires

Table 2 Properties of multilocus nested clade analysis vs. approximate Bayesian computation

Property	NCA	ABC
Genetic data used	Haplotype trees	Broad array of genetic data
Data analysable	Geographical, phenotypic, ecological, interspecific, etc.	Geographical, ecological, interspecific, etc.
Nature of inference	Strong	Weak
Interpretative criteria	Explicit, a priori, universal	Explicit, ad hoc, case specific
False-positives	False-positive rate estimable; false-positives reduced via cross-validation	False-positive rate not estimable; no mechanism for correcting for false-positives
A priori phylogeographical model	No: allows discovery of unexpected	Yes: inference limited to finite set of a priori alternatives
Nature of inferred phylogeographical model	Built-up from simple components, each with explicit statistical support	Full models specified a priori, resulting in confoundment when model has several components
Mechanics of inference	Interpretive key followed by phrasing as null hypotheses tested with likelihood ratios	Simulation requiring multiple parameters: model and parameter values confounded
Sampling error	Incorporates errors due to tree inference, number of locations, number of individuals, and evolutionary stochasticity	Incorporates sampling error and evolutionary stochasticity in simulations; ignores error in current generation statistics
Probability distributions and probabilities	Simulates full sampling distribution; likelihood ratios based on full distributions	Local probabilities, obscuring good-fitting, irrelevant, and over-determined models
Sample size	Handles large numbers of locations and individuals	Severely restricted on the number of locations analysable
Convergence	Well defined	Unknown
Dimensionality of tests	Well defined	Ignored
Final test product	Tail probability of null hypothesis	A non-co-measurable fit metric with no probabilistic meaning

an absolute ability to say if $A > B$, $A = B$, or $A < B$. Goodness-of-fit statistics from the class $\|s' - s\|$ do not have the property of co-measurability, as illustrated by the chi-squared example above. For statistical inference, it is critical to have a metric that is co-measurable, and a probability measure is one such entity. That is why the chi-squared goodness-of-fit statistics had to be transformed into tail probabilities in order to compare the fits of models I and II. Indeed, much of statistical theory is devoted to transforming statistics that are not inherently co-measurable (chi-squares, t -tests, likelihood ratios, mean squares, least-squares, etc.) into co-measurable probability statements.

The posterior probabilities in ABC are constructed by using a numerator that is a function of the goodness-of-fit measure $\|s' - s\|$ for a particular model. The numerator is then divided by a denominator that ensures that the 'probabilities' sum to one across the finite set of hypotheses (see equation 9 in Beaumont *et al.* 2002). Beaumont *et al.* (2002) make no adjustments for the dimensionality of these hypotheses, violating the Schwarz (1978) proposition, nor do they distinguish between nested and non-nested

alternative models. As a result, the numerators are not co-measurable across hypotheses, and the denominators are sums of non-co-measurable entities. Hence, the 'posterior probabilities' that emerge from ABC are not co-measurable. This means that it is mathematically impossible for them to be probabilities.

Fagundes *et al.* (2007) assign an ABC 'posterior probability' of 0.781 to the out-of-Africa replacement model. Because it is mathematically impossible for this number to be a probability, the out-of-Africa replacement model is *not* necessarily the most probable model out of the three that they considered. Indeed, the number 0.781 has no meaningful statistical interpretation. Thus, the final product of the ABC analysis are numbers that are devoid of statistical meaning. The ABC method is not capable of even weak statistical inference.

Discussion

Table 2 summarizes the points made in this paper about the statistical properties of multilocus NCPA and ABC. As

can be seen, ABC has multiple statistical flaws and does not yield true probabilities. In contrast, multilocus NCPA provides a framework for hard inference based upon well-established statistical procedures such as permutation testing and likelihood ratios. There has been much criticism of NCPA starting in 2002. Some of this criticism is based on factual errors; such as the misrepresentation that NCPA equates haplotype trees to population trees, or that nested clade analysis cannot analyse data other than geographical data. Other criticisms had validity; such as the interpretations of single-locus NCPA were not phrased as testable null hypothesis or that the false-positive rate was high. These flaws of single locus NCPA were addressed and solved by multilocus NCPA. Consequently, the criticisms of Knowles & Maddison (2002), Panchal & Beaumont (2007), Petit (2008), and Beaumont & Panchal (2008) are irrelevant to multilocus NCPA. As shown in this paper, multilocus NCPA is a robust and powerful method of making hard phylogeographical inference and does not suffer from the limitations attributed to single-locus NCPA.

Because of its multiple flaws, ABC should not be used for hypothesis testing. This does not mean that simulation approaches have no role in statistical phylogeography. The strength of NCPA is to falsify hypotheses and to build up complex phylogeographical models from simple components without any need for prior knowledge. NCPA achieves its robustness in testing hypotheses by taking a nonparametric approach, but this also means that NCPA will not yield much insight into the details of the emerging phylogeographical models. For example, isolation by distance may be inferred, but there is no estimation of the parameters of an isolation-by-distance model. A range expansion might be inferred, but there is no insight into the demographic details of that expansion through NCPA. Moreover, NCPA is limited to haplotype data from DNA regions with little to no recombination, but other types of data may have considerable phylogeographical information as well. Hence, the best phylogeographical analyses are those that use NCPA or some other statistically valid procedure to outline the basic phylogeographical model, followed by the use of simulation techniques to estimate phylogeographical and gene flow parameters and to incorporate additional information from other types of data (for some examples, see Garrick *et al.* 2007, 2008; Strasburg *et al.* 2007; Brown & Stepien 2008). Of particular relevance is the work of Gifford & Larson (2008) who used multilocus NCPA as their primary hypothesis testing tool and coalescent-based simulations for some parameter estimation. Hence, NCPA and simulation approaches are not so much alternative techniques as they are complementary, and potentially synergistic, techniques. Both add to the statistical toolkit of intraspecific phylogeographers, and both should be used when appropriate.

Acknowledgements

This work was supported in part by NIH grant P50-GM65509. I wish to thank six anonymous reviewers for their useful suggestions and Laurent Excoffier for patiently explaining some of the details of the ABC analysis of human evolution.

References

- Avisé JC, Lansman RA, Shade RO (1979) The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. *Genetics*, **92**, 279–295.
- Avisé JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.
- Beaumont MA, Panchal M (2008) On the validity of nested clade phylogeographical analysis. *Molecular Ecology*, **17**, 2563–2565.
- Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Bercovici S, Geiger D, Shlush L, Skorecki K, Templeton A (2008) Panel construction for mapping in admixed populations via expected mutual information. *Genome Research*, **18**, 661–667.
- Billingsley P (1968) *Convergence of Probability Measures*. John Wiley & Sons, New York.
- Brisson JA, De Toni DC, Duncan I, Templeton AR (2005) Abdominal pigmentation variation in *Drosophila polymorpha*: Geographic variation in the trait, and underlying phylogeography. *Evolution*, **59**, 1046–1059.
- Brown JE, Stepien CA (2008) Ancient divisions, recent expansions: phylogeography and population genetics of the round goby *Apollonia melanostoma*. *Molecular Ecology*, **17**, 2598–2615.
- Cox MP, Mendez FL, Karafet TM *et al.* (2008) Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. *Genetics*, **178**, 427–437.
- Edgington ES (1986) *Randomization Tests*, 2nd edn. Marcel Dekker, New York.
- Eswaran V, Harpending H, Rogers AR (2005) Genomics refutes an exclusively African origin of humans. *Journal of Human Evolution*, **49**, 1–18.
- Ewens WJ (1983) The role of models in the analysis of molecular genetic data, with particular reference to restriction fragment data. In: *Statistical Analysis of DNA Sequence Data* (ed. Weir BS), pp. 45–73. Marcel Dekker, New York.
- Fagundes NJR, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences, USA*, **104**, 17614–17619.
- Garrick RC, Sands CJ, Rowell DM, Hillis DM, Sunnucks P (2007) Catchments catch all: long-term population history of a giant springtail from the southeast Australian highlands—a multigene approach. *Molecular Ecology*, **16**, 1865–1882.
- Garrick RC, Rowell DM, Simmons CS, Hillis DM, Sunnucks P (2008) Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. *Evolution*, **62**, 1103–1118.
- Garrigan D, Kingan SB (2007) Archaic human admixture: a view from the genome. *Current Anthropology*, **48**, 895–902.

- Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF (2005) Evidence for archaic Asian ancestry on the human X chromosome. *Molecular Biology and Evolution*, **22**, 189–192.
- Gifford ME, Larson A (2008) *In situ* genetic differentiation in a Hispaniolan lizard (*Ameiva chrysoleama*): a multilocus perspective. *Molecular Phylogenetics and Evolution*, **49**, 277–291.
- Knowles LL (2001) Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molecular Ecology*, **10**, 691–701.
- Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolution Biology*, **17**, 1–10.
- Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.
- Lemmon AR, Lemmon EM (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic Biology*, **57**, 544–561.
- Palsbøll PJ, 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, Beaumont MA (2007) The automation and evaluation of nested clade phylogeographic analysis. *Evolution*, **61**, 1466–1480.
- Petit RJ (2008) The coup de grace for the nested clade phylogeographic analysis? *Molecular Ecology*, **17**, 516–518.
- Petit RJ, Grivet D (2002) Optimal randomization strategies when testing the existence of a phylogeographic structure. *Genetics*, **161**, 469–471.
- Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. *PLoS Genetics* **2**, e105.
- Popper KR (1959) *The Logic of Scientific Discovery*. Hutchinson, London.
- Pritchard JK, Seielstad MT, Perez-Lezaan A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.
- Ray N, Currat M, Berthier P, Excoffier L (2005) Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Research*, **15**, 1161–1167.
- Roca AL, Georgiadis N, O'Brien SJ (2005) Cytonuclear genomic dissociation in African elephant species. *Nature Genetics*, **37**, 96–100.
- Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Strasburg J, Kearney M, Moritz C, Templeton A (2007) Combining phylogeography with distribution modeling: multiple pleistocene range expansions in a parthenogenetic gecko from the Australian arid zone. *PLoS ONE*, **2**, e760.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585.
- Takahata N, Lee S-H, Satta Y (2001) Testing multiregionality of modern human origins. *Molecular Biology and Evolution*, **18**, 172–183.
- Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.
- Templeton AR (2001) Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology*, **10**, 779–791.
- Templeton AR (2002a) 'Optimal' randomization strategies when testing the existence of a phylogeographic structure: a reply to Petit and Grivet. *Genetics*, **161**, 473–475.
- Templeton AR (2002b) Out of Africa again and again. *Nature*, **416**, 45–51.
- Templeton AR (2004a) A maximum likelihood framework for cross validation of phylogeographic hypotheses. In: *Evolutionary Theory and Processes: Modern Horizons* (ed. Wasser SP), pp. 209–230. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Templeton AR (2004b) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.
- Templeton AR (2005) Haplotype trees and modern human origins. *Yearbook of Physical Anthropology*, **48**, 33–59.
- Templeton AR (2007a) Perspective: genetics and recent human evolution. *Evolution*, **61**, 1507–1519.
- Templeton AR (2007b) Population biology and population genetics of Pleistocene Hominins. In: *Handbook of Palaeoanthropology* (ed. Henke W, Tattersall I), pp. 1825–1859. Springer-Verlag, Berlin, Germany.
- Templeton AR (2008) Nested clade analysis: an extensively validated method for strong phylogeographic inference. *Molecular Ecology*, **17**, 1877–1880.
- Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.
- Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics*, **117**, 343–351.
- Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.
- Templeton AR, Routman E, Phillips C (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.
- Templeton AR, Clark AG, Weiss KM *et al.* (2000a) Recombinational and mutational hotspots within the human *Lipoprotein Lipase* gene. *American Journal of Human Genetics*, **66**, 69–83.
- Templeton AR, Maskas SD, Cruzan MB (2000b) Gene trees: a powerful tool for exploring the evolutionary biology of species and speciation. *Plant Species Biology*, **15**, 211–222.
- Templeton AR, Maxwell T, Posada D *et al.* (2005) Tree scanning: a method for using haplotype trees in genotype/phenotype association studies. *Genetics*, **169**, 441–453.

Alan Templeton has both a master's degree in Statistics and a doctoral degree in Human Genetics. This combination of statistical and genetics training has allowed him to develop many innovative statistical methods in human genetics and evolutionary biology, including the statistical procedure of nested clade analysis.
