MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 436–446

NEWS AND VIEWS

REPLY

In defence of model-based inference in phylogeography

MARK A. BEAUMONT,[1] RASMUS NIELSEN,[2] CHRISTIAN ROBERT,[3] JODY HEY,[4] OSCAR GAGGIOTTI,[5] LACEY KNOWLES,[6] ARNAUD ESTOUP,[7] MAHESH PANCHAL,[8] JUKKA CORANDER,[9] MIKE HICKERSON,[10] SCOTT A. SISSON,[11] NELSON FAGUNDES,[12] LOUNÈS CHIKHI,[13] PETER BEERLI,[14] RENAUD VITALIS,[15] JEAN-MARIE CORNUET,[7] JOHN HUELSENBECK,[2] MATTHIEU FOLL,[16,17] ZIHENG YANG,[18] FRANCOIS ROUSSET,[19] DAVID BALDING[20] and LAURENT EXCOFFIER[16,17]

[1]*School of Animal and Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading, RG6 6AJ, UK;* [2]*Integrative Biology, UC Berkeley, 3060 Valley Life Sciences Bldg #3140, Berkeley, CA 94720-3140, USA;* [3]*CEREMADE, Université Paris Dauphine, Paris, France;* [4]*Department of Genetics, Rutgers University, 604 Allison Road, Piscataway, NJ 08854, USA;* [5]*Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP 53, 38041 GRENOBLE, France;* [6]*Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079, USA;* [7]*INRA UMR Centre de Biologie et de Gestion des Populations (INRA/IRD/Cirad/Montpellier SupAgro), Campus international de Baillarguet, Montferrier-sur-Lez, France;* [8]*Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany;* [9]*Department of Mathematics and statistics, University of Helsinki, Finland;* [10]*Biology Department, Queens College, City University of New York, 65-30 Kissena Boulevard, Flushing, NY 11367-1597, USA;* [11]*School of Mathematics and Statistics, University of New South Wales, Sydney, Australia;* [12]*Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;* [13]*Université Paul Sabatier-UMR EDB 5174 118, 31062 Toulouse Cedex 09, France;* [14]*Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA;* [15]*CNRS–INRA, CBGP, Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez, France;* [16]*CMPG, Institute of Ecology and Evolution, University of Berne, 3012 Berne, Switzerland;* [17]*Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;* [18]*Department of Biology, University College London, Gower Street, London WC1E 6BT, UK;* [19]*Institut des Sciences de l'Évolution, Universté Montpellier 2, CNRS, Place Eugène Bataillon, CC065, Montpellier, Cedex 5, France;* [20]*Institute of Genetics, University College London, 2nd Floor, Kathleen Lonsdale Building, 5 Gower Place, London WC1E 6BT, UK*

Correspondence: Laurent Excoffier;
e-mail: Laurent.excoffier@iee.unibe.ch

Recent papers have promoted the view that model-based methods in general, and those based on Approximate Bayesian Computation (ABC) in particular, are flawed in a number of ways, and are therefore inappropriate for the analysis of phylogeographic data. These papers further argue that Nested Clade Phylogeographic Analysis (NCPA) offers the best approach in statistical phylogeography. In order to remove the confusion and misconceptions introduced by these papers, we justify and explain the reasoning behind model-based inference. We argue that ABC is a statistically valid approach, alongside other computational statistical techniques that have been successfully used to infer parameters and compare models in population genetics. We also examine the NCPA method and highlight numerous deficiencies, either when used with single or multiple loci. We further show that the ages of clades are carelessly used to infer ages of demographic events, that these ages are estimated under a simple model of panmixia and population stationarity but are then used under different and unspecified models to test hypotheses, a usage the invalidates these testing procedures. We conclude by encouraging researchers to study and use model-based inference in population genetics.

Keywords: molecular evolution, phylogeography, population genetics-empirical, population genetics-theoretical

Received 30 October 2009; revision received 4 December 2009; accepted 7 December 2009

Introduction

How is it possible to use genetic data from related populations or species to figure out their recent evolutionary history? Each data set is open to various interpretations, yet in any particular case some interpretations might be better justified than others. The challenge is to develop a genetical and evolutionary theory that is general enough to include real histories, and yet simple but detailed enough that it can be used in a statistical framework to infer details of a specific history, including (importantly) measures of uncertainty.

The idea of a genealogy, or gene-tree, to represent the history of a sample of homologous gene copies is one of biology's most successful models thanks to its generality and flexibility. However, statistical inference under the gene-tree model is difficult. For many years investigators,

often using mitochondrial sequences, struggled to interpret trees generated from their data in terms of demographic processes, such as population separation or gene exchange. In the early days, this field of phylogeography relied on heuristic and descriptive analyses, and it was essentially not statistical.

The situation changed with the introduction of Nested Clade Phylogeographical Analysis (NCPA) (Templeton 1998; Templeton *et al.* 1995). In combining an analysis of estimated gene-tree structure with an inference key to make conclusions about the demographic causes of the shape of the gene-tree, the method served a generation of evolutionary biologists eager to make sense of their data. To address the concern that gene-tree estimates can be wrong, the method accommodates a network of connections based on which haplotypes are likely to be connected in the true genealogy (Crandall 1996; Templeton *et al.* 1992). To address the concern that different unlinked genes can have widely different histories, even when sampled from the same organisms, 'cross-validation' of multiple loci was proposed (Templeton 2002, 2004a). Notwithstanding the apparent flexibility and generality of NCPA, or its popularity, the method has been subject to a number of criticisms (Knowles & Maddison 2002; Petit & Grivet 2002; Hey & Machado 2003; Panchal & Beaumont 2007; Knowles 2008; Manolopoulou 2008), and has been vigorously defended (Templeton 2004b, 2008, 2009b).

Today, in contrast to the years when NCPA first came on the scene, there are other approaches available for developing complex demographic inferences. The origins of these methods actually predate NCPA, going back to the first likelihood-based models for demographic and phylogenetic inference (Cavalli-Sforza & Edwards 1967; Thompson 1973; Felsenstein 1981) and the development of coalescent theory (Kingman 1982; Hudson 1983; Tajima 1983). Although they vary considerably in details, these methods differ sharply from NCPA in two fundamental ways. First, they are explicitly based on demographic models that include parameters such as population size and migration rates. Second they use the genealogy as an unobserved variable that connects data to model parameters but need not be explicitly inferred (Hey & Nielsen 2007). These model-based approaches share the goal of computing a likelihood function (i.e. the probability of the data as a function of the parameters within a given model). Being likelihood-based, these methods open doors for population geneticists and phylogeographers to the repertoire of likelihood-based analyses, including maximum likelihood estimation of model parameters and likelihood-ratio hypothesis tests (e.g. Griffiths & Tavaré 1994; Kuhner *et al.* 1995; Beerli & Felsenstein 1999), as well as Bayesian analyses (Wilson & Balding 1998), including Approximate Bayesian Computation (ABC) (Tavare *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002).

Templeton (2010), in response to Nielsen & Beaumont (2009), heavily promotes NCPA for analysing phylogeographic data, incorrectly asserting that it uses 'a likelihood function that explicitly incorporates the randomness associated with the coalescent and mutational processes'. He also repeats many claims from Templeton (2009a) where he strongly criticizes the use of ABC methods for analysing phylogeographic data in general, and their application to discriminate between various human evolutionary scenarios in particular (Fagundes *et al.* 2007). He concluded that 'because of its multiple flaws, ABC should not be used for hypothesis testing'. Yet ABC is simply a Monte Carlo method that can be used to approximate posterior distributions or likelihood surfaces from a model (see e.g. Tavare *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002, for more details on ABC approaches). It is a numerical tool for solving problems within a statistical framework. Thus the majority of criticisms that Templeton (2009a, 2010) aims at ABC are also aimed more generally against model-based inference in population genetics. We feel compelled to react against this broadly unsupported attack on model-based inference, and to point out important misconceptions underlying Templeton's critique.

First, we highlight Templeton misconceptions of model-based inference, of Bayesian methods in general and of ABC in particular. Next, we underline major deficiencies of NCPA when inferring past demographic scenarios, and errors or misleading statements in Templeton's promotion of the method.

Misconceptions about model-based methods

Model specification

In population biology, as in many other scientific areas, there has been a longstanding tension between proponents and opponents of model-based inferences. The most familiar example is the debate between cladists and likelihoodists in phylogenetics. Although Templeton (2009a) claims to accept both hypothesis testing and models, including likelihood and Bayesian methods, many of his criticisms echo old arguments against the use of model-based inferences in phylogenetics. He argues that it is a flaw of ABC, and of model-based methods that they do not cover the entire 'hypothesis space' (Templeton 2009a, p. 320), but instead compare only a small number of potentially misspecified and subjectively chosen models (Templeton 2010). However, for realistic problems, exhaustive coverage of all hypotheses is impossible. Moreover, the situation that 'all hypotheses being compared are false' (Templeton 2009a, p. 320) is in fact the norm in science, since models at best only approximate reality, as recognized in the widely cited words attributed to George Box: 'all models are wrong, but some are useful' (Box & Draper 1987, p. 424). As an aside, the distinction between '(i) testing a null hypothesis and (ii) assessing the relative fit of alternative hypotheses' (Templeton 2009a, p. 320) is reminiscent of the 1930s debate between Fisherian and Neyman-Pearson hypothesis testing; the Neyman–Pearson approach of choosing among a limited set of competing models came to dominate statistical practice (Gigerenzer *et al.* 1990).

Strong vs. weak inference

Invoking Popper (1959), Templeton (2007) contends that by relying on successive dichotomous tests NCPA can make 'strong' phylogeographic inferences, which is not possible with model-based methods. However, 'strong scientific inference' (cf. Platt 1964) arises when the influence of unknown factors on the final result is minimized by randomization (Macneil 2008), which also underlies Fisher's (1925) null hypothesis testing. That is, without a properly randomized experiment, causal explanations are necessarily weak because they are potentially confounded with unobserved effects. Since they are based on observational data, phylogeographic studies are not amenable to randomized interventions and therefore *all* phylogeographic inference methods, including NCPA, lead to 'weak scientific inference' in the sense that it does not arise from planned scientific experiments. Popper was fiercely opposed to inductivism, whereby facts are gathered and then general laws identified. In this regard, rather than being a Popperian falsification method, NCPA can in fact be viewed as an anti-Popperian inductivist approach (Beaumont & Panchal 2008), since a story is built out of the patterns in the data.

Unspecified models cannot be tested

Templeton (2009a) argues that since NCPA tests null hypotheses without reference to an explicit alternative, it does not rely on a restricted set of alternative models. However, except for testing the null hypothesis of no correlation between geographic and genetic distances, we show below that NCPA's inferences about specific phylogeographic hypotheses are invalid. Moreover, since no alternative model is specified, there can be no measure of the relative support for the different hypotheses entertained by NCPA. The specification of alternative models is necessary to correctly assess the support of data for a complex demographic model. This inevitably incurs additional possibilities of model misspecification, but there are many statistical techniques for assessing the fit of a model. The use of explicit models expose their authors to critiques, but it is the price to pay for science to make progress, as other researchers may propose better models that can be tested against the data, leading to an increasing refinement of the models, and in our understanding of the demographic patterns that they reflect.

Simulations under specific models

Templeton continuously rejects the use of simulations to validate models and to infer parameters. As evidence for 'the extreme ambiguity of inference via computer simulations', Templeton (2010) mentions two studies on human evolution (Eswaran *et al.* 2005; Fagundes *et al.* 2007) which simulate different evolutionary scenarios using different data sets and arrive at different conclusions. Two studies leading to different conclusions of course do not invalidate the common tools that are used. As previously stated, the use of simulations in the ABC inference procedure criticized by Templeton is just a means to evaluate or approximate the likelihood function. Templeton also argues against the use of simulations for evaluating the relative merits of different inference methods, because this requires the full specification of the parameter space to be explored, and implies that choices need to be made concerning which models are used and contrasted. A related criticism by Templeton (2009b) is that the models that have been used to test NCPA are unlikely and therefore the high false-positive rate attributed to NCPA is also unlikely. However, an explicit model specification procedure, which is the rule in physics and most other sciences, involves no hidden assumption, and the impact of alternative parameterizations can be conveniently studied. Because it is transparent, it is open to criticism and the use of alternative specifications. By varying the conditions of the simulations it is possible to determine when methods fail and when they perform well. Indeed, without such objective testing, it is impossible to have any assessment of the performance of a statistical procedure. If a method consistently leads to wrong inferences under all or most conditions explored, as we later argue is the case with NCPA, it should be discarded.

Misconceptions about Bayesian methods

We recognize that there are alternative ways to perform statistical inference. This is well reflected in this paper authorship, and arises from different epistemological traditions lying deep in the history of statistics. Our aim in this section is not to argue for the relative merits of one approach over another, but simply to correct factual errors concerning Bayesian inference that are to be found in Templeton (2009a, 2010), and to present the main arguments that underpin it.

Statistical validity of ABC

Templeton (2009a) presents an extensive critique of the ABC method, which is simply a way to perform model-based inference in a Bayesian setting when model likelihoods are intractable and thus need to be approximated by simulations. For example Templeton questions 'the statistical validity of all inferences made by the ABC method' (p. 325) and argues that 'the 'posterior probabilities' that emerge from ABC [are] mathematically impossible … to be probabilities' (p 329). However, when the summary statistics used in ABC are statistically sufficient and parameter estimation uses only the simulations that exactly match the observed data, ABC is exact Bayesian inference (Marjoram & Tavare 2006). Thus Templeton is in effect claiming that standard Bayesian inferences are invalid, and that Bayesian posterior probabilities are mathematically incapable of being probabilities.

Comparison of alternative models

Bayesian analysis is fundamentally a decision-making approach, in which the goal is to evaluate the relative sup-

port for different models under comparison. In contrast, Fisherian testing of a point null hypothesis using *P*-values only rejects models that inadequately explain the data. There is a large literature on the problems that arise when taking null hypothesis testing out of its original context in the analysis of designed experiments (see e.g. Berger & Sellke 1987). Templeton's claim that in 'ABC there is no null hypothesis, which complicates the computation of sampling error' (2009a, p.325) is incorrect: sampling error is evaluated in each model under consideration, and is not dependent on the specification of a null hypothesis.

Priors

Templeton's criticisms that in ABC a model can be rejected because 'the simulated parameter values are wrong' (Templeton 2009a, p. 323), and that 'parameter ranges and distributions are only guessed based upon the subjective opinion of the investigators' (Templeton 2010), are classical objections made against Bayesian approaches, which need the specification of a prior distribution for all the parameters of a model. Priors might be mis-specified and their choice may indeed carry some subjectivity, but their impact on posterior distributions, parameter inference, and model choice can be quantified (Berger 1990; Gelman *et al.* 1996).

Global parameter inference

Templeton's comment that NCPA 'separate[s] out different phylogeographical components is a great advantage over ABC' (Templeton 2009a, p. 324) ignores the fact that testing subsets of the data separately precludes any assessment of uncertainty in the overall conclusions. The fact that a method, like ABC, permits this assessment is a clear advantage over NCPA. A sound statistical approach should work with all data and parameters at once, and thus incorporate dependencies among the parameters and avoid multiple uses of the data. In particular, unlike NCPA, Bayesian methods avoid the problem of using an estimate as if it were the true value. Uncertainty in parameter values is explicitly modelled, at odds with NCPA, where for instance very little or no uncertainty in the topology of the gene-tree is assumed for the analysis.

Sampling error

Templeton's argument that simulated statistics and observed statistics cannot be compared because the observed statistic (s) is 'current generation' while a simulated statistic (s') is 'long-term' (see fig. 2 in Templeton 2009a) is wrong. The error in the argument can be made explicit by replacing 'statistics' with 'data'. The aim of model-based methods is to examine the relative probability of obtaining the data for different combinations of parameter values. It is acknowledged that the observations are influenced by both sampling error and evolutionary stochasticity in the model, and this is explicitly accounted for by ABC which simulates data sets with sample sizes and number of loci matching exactly those observed. As mentioned before, ABC is then simply a way of using simulations to make inferences.

Simulation weighting

Templeton's claim of an artefactual increase in statistical power by computing a distance between observed and simulated summary statistics, $||s - s'||$, is incorrect. In ABC, $||s - s'||$ is not 'a generalized goodness of fit statistic' (Templeton 2009a, p. 328), but is used to determine if a simulation is retained for parameter estimation. For retained simulations, $||s - s'||$ is also used as a weight allocated to the simulated parameter values in approximating the posterior distribution. Note that the ABC method is exact when simulations are retained if $||s - s'|| = 0$ and s is sufficient, since the fraction of retained simulations provide a direct estimate of the likelihood. If the retention interval increases then, typically, the posterior distributions become wider, and the posterior tends to the prior with increasing retention intervals. Thus the ABC approach is inherently conservative. How the approximated density converges to the true distribution (conditional on the summary statistics) as $||s - s'||$ tends to zero is an area of active research (e.g. Ratmann *et al.* 2009; Blum & François 2010).

Posterior densities and Bayesian model choice

The section in Templeton (2009a, p. 326–327) that discusses full distributions and local probabilities contains a number of erroneous statements, as explained below. Templeton's Figure 3 is used to suggest that conditioning inferences on observed statistics may lead to wrong decisions in Bayesian model choice. The interpretation of the figure is actually problematic in itself. The graph plots the posterior density against the value of a summary statistic. Bayesian inference typically aims to compute the posterior distribution of parameter values, not statistics. Conceivably what is meant is the posterior predictive distribution of the values of a summary statistic, conditional on the observed summary statistic. The posterior predictive distribution is typically used in Bayesian model checking (Gelman *et al.* 1996). Central to Templeton's argument are (i) the assumption that observed statistics may often lie in the tails of this distribution, and (ii) that ABC (and by extension, Bayesian) model choice procedures are based on an examination of this distribution around the observed statistics, while the center of mass of the distribution can be further away from the observed statistics, and thus lead to wrong inferences. These premises are incorrect, because, if the model fits well, the observed summary statistic does not necessarily lie within the tails of the posterior predictive distribution. Furthermore, as discussed in more detail below, Bayesian model choice is not based on the posterior predictive distribution at all, as implied in the discussion in Templeton (2009a, p. 326–327). An alternative interpretation of Templeton's Figure 3 is that it is, in fact, the prior predictive

distribution—that is the distribution of summary statistics under the model when the parameters are drawn from the prior. With this interpretation, the prior predictive distribution at the observed summary statistic is also the marginal likelihood. In the context of ABC, ratios of marginal likelihoods (Bayes factors) can be approximated as the ratio of the number of simulations made under alternative models that are arbitrarily close to the observed data. Within the Bayesian framework this procedure is correct and is not based on a notion of 'local probability', and Templeton's criticisms of a specific deficiency in ABC are therefore also unfounded. Templeton further argues against the use of ABC (and hence Bayesian) methods for model comparison because they cannot take dimensionality into account, and he implies that they will always choose over-determined models. Indeed he appears to criticize ABC approaches for not using the correction of Schwarz (1978) in his Bayesian information criterion (BIC). However, from a Bayesian perspective there is no need to correct for dimension, nor to call for Schwarz (1978), since the marginal likelihood naturally allows for differences in model dimensionality (see e.g. MacKay 2002, chapter 28.1 about Occam's razor). In fact, the penalty in Schwarz's (1978) BIC stems from a Taylor expansion of a standard Bayes Factor (see also Schervish 1995), which illustrates the automatic penalty for dimension and over-parameterization when using Bayes factors.

Sample size

In the section on 'Sample size', Templeton (2009a, p. 327) claims that 'ABC has severe constraints on sample size'. This is a misleading statement. Indeed one of the main motivations behind the approach is that it can potentially deal with larger data sets than can currently be handled with other model-based procedures. There are constraints set by computation time for very large data sets, but with efficient simulation methods implemented on computer clusters sample size is not a major limitation of the approach for most practical applications. Further, Templeton argues that the samples sizes (8–12 individuals per continent) used in Fagundes *et al.* (2007) are too small to lead to reliable estimates, arguing that such size do not meet NCPA requirements. However, as noted above, the ABC framework, by simulating exactly the observed sample sizes, handles any sample sizes correctly. Small sample sizes simply lead to wider credible intervals than large sample sizes. ABC methods are not markedly constrained by the use of multiple loci, and, as is to be expected, the precision of estimates tends to increase when summary statistics are based on many loci (e.g. Excoffier *et al.* 2005).

Recent developments in ABC methods

In order to put the comments of Templeton (2009a) in context it is perhaps helpful to provide a brief overview of the current status of ABC, which is now quite widely used in statistical inference. For example, it has been applied to infectious disease epidemiology (Tanaka *et al.* 2006; Luciani *et al.* 2009; McKinley *et al.* 2009) and systems biology (Ratmann *et al.* 2009; Toni *et al.* 2009). Whereas several studies have now shown that parameter posterior distributions inferred by ABC are very similar to those provided by full-likelihood approaches (see e.g. Marjoram *et al.* 2003; Bortot *et al.* 2007; Beaumont *et al.* 2009; Leuenberger & Wegmann 2010), the approach is still in its infancy and continues to evolve, and to be improved. For instance, Marjoram *et al.* (2003) developed a Markov chain Monte Carlo (MCMC) ABC approach, improving the sampling efficiency of conventional ABC, which must otherwise explore sometimes very wide priors while posterior distributions may only occupy a narrow region of parameter space. This MCMC-ABC has some problems (Sisson *et al.* 2007), which are addressed in variants of the original approach (see e.g. Becquet *et al.* 2007; Bortot *et al.* 2007; Wegmann *et al.* 2009). Recently, sequential Monte Carlo (SMC) techniques have been adapted to ABC in order to further improve its efficiency (see e.g. Sisson *et al.* 2007; Beaumont *et al.* 2009; Del Moral *et al.* 2009). As noted by Beaumont *et al.* (2002) efficient conditional density estimation is a key aspect of ABC, and this has been developed further in Blum & Francois (2009). Further related developments involve the choice of statistics to summarize datasets (Joyce & Marjoram 2008; Sousa *et al.* 2009) and how they can be combined (Hamilton *et al.* 2005; Wegmann *et al.* 2009). A number of software packages now allow an easy implementation of ABC models, such as DIY-ABC (Cornuet *et al.* 2008) or popABC (Lopes *et al.* 2009), which can accommodate a wide range of evolutionary models, and be used for both model choice and parameter estimation.

Theoretical and statistical problems in NCPA

NCPA inferences are typically not tested by users

Templeton (2009a,b) claims that NCPA is embedded into a strong statistical framework, as it is based on the rejection of null models and hypothesis testing based on likelihood ratios contrasting NCPA inferences. It is interesting to examine what aspects of the NCPA procedure actually involve hypothesis testing and the rejection of null models. In the hundreds of published empirical studies based on this method, the only statistical procedure of NCPA is a simple permutation test of the null hypothesis of no association between clades and geographic location (see e.g. Knowles 2008; Petit 2008). However, the processes inferred by NCPA have never been tested as null models to see if they can actually give rise to data sets similar to those observed. Therefore NCPA inferences are typically presented without further attempt at model checking or validation. There is thus no measure of confidence that can be assigned to the inferences being made, nor any indication of support in the data for alternative processes. Moreover, almost all published NCPA inferences are based on the analysis of a single locus and NCPA internal cross-validation is not used.

Lack of NCPA expectations under different scenarios

When (i) there is a lack of strong prior knowledge of the universe of biological possibilities, or (ii) because of the possibility of multiple processes leading to the same output, it has been claimed that the 'broader coverage' of processes makes NCPA the method of choice (Templeton 2004b). However, as emphasized above, because the interpretation of the patterns of genetic variation is not associated with a defined model, there is no basis for evaluation of the inferences made with the dichotomous inference key of NCPA. In other words there is no explicit description of the patterns of variation in NCPA outcome expected under one historical scenario relative to another. There is no study verifying that the interpretations of the distance statistics used in NCPA (i.e. D_C and D_N values) actually correspond to what is expected under the processes NCPA claims to be able to distinguish. This does not mean that model-based inference is not without its challenges, especially with regards to issues surrounding model choice (as reviewed in Hey & Machado 2003; Knowles 2004, 2009; Nielsen & Beaumont 2009), but these difficulties should not be used as a justification for resorting to a method with undefined statistical properties (Knowles 2008). Any sound statistical method needs to provide an assessment of its error or uncertainty. Even if NCPA was not flawed in the many other ways described in this paper, the inference of phylogeographic processes based on pure verbal logic with no alternate models and no statistical support should be enough to relegate it to be regarded as an exploratory tool at best.

Cross-validation using multi-locus data lacks rigor

The suggestion that the new multilocus NCPA somehow overcomes these problems is likewise indefensible, and the statistical test on which it relies is flawed (see details below). Additionally, the claim that when NCPA analyses of two or more loci lead to the same inference, this constitutes a rigorous 'cross-validation', is not based on any statistical concept of validation. Any concordance in observed patterns across two loci depends on the evolutionary variance of the process itself, which is not evaluated in NCPA, and which may vary extensively among different evolutionary processes. For instance, patterns of molecular diversity after a range expansion can be highly correlated among unlinked loci, and the observation of similar patterns at two loci is expected (e.g. Di Rienzo *et al.* 1998), whereas a population bottleneck often induces a much larger evolutionary variance across loci (e.g. Bonneuil 1998; Teshima *et al.* 2006). Thus, the probability for a given number of loci to show congruent patterns can only be evaluated under a given evolutionary model. The fact that the number of false inferences drops with additional loci is expected, but there is no control over the resulting type II error.

NCPA inference key has still not been properly tested

The NCPA procedure consists of four main tasks: (i) the construction of cladograms; (ii) the computation of summary statistics based on geographic patterns associated with these cladograms; (iii) permutation tests to assess their statistical significance; (iv) biological interpretation of the 'significant' summary statistics. Task (iv) is carried out via an 'inference key', which is consulted each time a statistically significant summary statistic is identified. The concomitant problem of multiple testing has been previously highlighted (Knowles & Maddison 2002; Panchal & Beaumont 2007) and acknowledged by Templeton (2008, 2009b). The inference key was originally provided in Templeton *et al.* (1995), and leads to a conclusion either that there are insufficient data to make an inference, or that some specified demographic event has occurred in the history of the population. Examples and discussion of the high rate of false positives generated by use of the inference key are given in Nielsen and Beaumont (2009) and in Panchal and Beaumont (2007), as well as in a later section of this article (see Table 1). An important point to note, however, is that the procedure is superficially similar in scope to the decision tree, or classification tree, used in machine learning and statistics (Breiman *et al.* 1984). The aim of the classification tree is to model a categorical dependent variable (the classification) as a function of independent variables. A *sine qua non* of such a procedure is that it must be validated on a training set to measure classification error and compare its performance against different algorithms, before it is applied to real classification problems. There is no evidence that the rules encapsulated in the key of Templeton *et al.* (1995), including its later revisions (Templeton 2004b) have been generated through a training set, as required for a valid statistical procedure. It would appear that the rules are based solely on reasoned opinions (Templeton *et al.*, 1995). A *post hoc* justification of this inference tree, which appears to uphold the purely verbal reasoning by which it was originally constructed, has been made through analysis of empirical data sets, but the demographic history in these empirical data sets is not known for certain. In the following section, further grounds for doubt about the validity of these conclusions are raised.

Positive controls do not rigorously test the validity of NCPA

The repeated claim that the inferences from NCPA have been 'extensively validated' refers to two studies in which, respectively, 13 and 150 empirical data sets with 'strong a priori expectations' were analysed (Templeton 1998, 2004b). Vigorous defence of this approach as a rigorous test of NCPA performance (and hence, its validation) has been made (e.g. Templeton 2009b), including claims that any former criticisms are 'outdated' or based on 'factual errors'. However we emphasize that NCPA has never been successfully verified by researchers independent of its author.

Evaluations of NCPA based on simulated data (Knowles & Maddison 2002; Panchal & Beaumont 2007) and empirical data (Templeton 1998, 2004b) consistently inferred multiple processes other than those expected (in case of the empirical datasets) or other than the actual processes (in

Table 1 NCPA false positive rates of inference. Results are broken down by F_{ST}, lattice size, and model. The false positive rate, as defined in Panchal & Beaumont (2010), is shown for both single and multiple loci

		Lattice size					
		3 × 3		7 × 7		10 × 10	
Model	F_{ST}	Single locus	Multiple loci	Single locus	Multiple loci	Single locus	Multiple loci
Panmictic	0	0.665	0.088	0.788	0.204	0.783	0.217
Stepping Stone	0.03	0.753	0.300	0.747	0.150	0.781	0.129
	0.05	0.791	0.333	0.834	0.175	0.851	0.200
	0.10	0.879	0.386	0.903	0.300	0.903	0.358
	0.20	0.895	0.429	0.942	0.396	0.969	0.479
Lattice with Cauchy Disperal Kernel	0.03	0.723	0.188	0.618	0.000	0.576	0.000
	0.05	0.783	0.179	0.654	0.000	0.649	0.000
	0.10	0.818	0.125	0.753	0.000	0.694	0.000
	0.20	0.888	0.042	0.831	0.000	0.802	0.000
Island Model	0.03	0.914	0.613	0.883	0.358	0.823	0.291
	0.05	0.950	0.704	0.886	0.467	0.878	0.321
	0.10	0.975	0.717	0.949	0.567	0.904	0.454
	0.20	0.980	0.708	0.982	0.667	0.941	0.617

The proportions shown here are (for single loci) each based on 1200 data sets. In the case of multiple loci, the single locus simulations were grouped into sets of five, and hence the proportions are based on 240 data sets.

case of the simulated data with known history). However, as mentioned above, Templeton has never conducted any validating simulation study. When applied to empirical data he has even suggested that these additional inferences may not be false positives, but rather unexpected discoveries. When these 'unexpected discoveries' were found by other authors in simulated datasets, they were of course classified as false positives (Knowles & Maddison 2002; Panchal & Beaumont 2007), but again, not by Templeton (2009a,b), who strongly argues that the simulated data and/or their interpretation must be flawed in one way or the other. It is also worth noting that while Templeton's 'extensive validation' relies almost exclusively on 'positive controls' based on single-locus studies, he charges that any critique of NCPA that is applied to single-locus data is outdated and unfair, given the more recent multilocus NCPA (Templeton 2009a,b). It should not be ignored that in doing so he is implicitly suggesting that all preceding papers that have used NCPA may have led to wrong inferences.

The ages of inferred events are crudely approximated by gene tree coalescent times

An important outcome of NCPA analysis is the dating of inferred events. Estimated dates are subsequently used (i) to build complex evolutionary scenarios from NCPA (see e.g. Templeton 2002) (ii) to treat estimated dates as if they were the observed ages of inferred events in likelihood-ratio tests (Templeton 2004a), and (iii) to invalidate conflicting results obtained by other authors on other data sets (Templeton 2009a, 2010). It is therefore important to understand the estimation method and its foundations. Templeton (2004a) proposes to estimate the age of a given event inferred by NCPA as the 'age of the youngest monophyletic clade that contributed in a statistically significant fashion to the inference'. The rationale is that 'the age of the youngest clade marking an event or process is expected to be largely coincident with the age of the event itself in most cases' (Templeton 2002), but several authors have underlined the dangers of dating population events from coalescent times on gene trees (see e.g. Pamilo & Nei 1988; Nichols 2001; Degnan & Rosenberg 2009). Therefore, the events whose ages are estimated in NCPA are at best, genealogical events, and not demographic events as claimed. That is not to say that temporal and spatial inferences of genealogical events may not be informative, but by themselves they cannot directly lead to statements about demography.

Coalescent theory is not applied correctly

Templeton (2004a) estimates the time since the most recent common ancestor (TMRCA) of a given clade by applying results of Tajima (1983) on the expected coalescent time of a pair of genes (noted hereafter T_2) conditional on the number of sites at which they differ (say π). There are serious problems attached to this estimation in the NCPA context. First, T_2 is not equal to the TMRCA of a sample of n genes (noted here T_n). In a stationary panmictic population, T_n is roughly twice as large as T_2, but the relation between T_2 and T_n is different for more complex evolutionary scenarios. Second, since Templeton ignores sample sizes and only concentrates on the number of different DNA sequences in a given clade (say k), he is using Tajima's theory as if it could be applied to estimate the average

TMRCA, T_k, among k haplotypes given their average number of pairwise differences $\overline{\pi}_k$, while Tajima's theory can only be used to estimate $\overline{T}_2$ as the average T_2 over all $n(n-1)/2$ pairs of genes in the clade. Third, Tajima's derivations are only strictly valid under a specific evolutionary model, which is that of a panmictic population of constant size, while Templeton applies this theory to haplotypes found in a clade that shows some support for demographic events that depart from stationarity (e.g. short or long range migrations in a subdivided population, population spatial expansion, or vicariance events). Fourth, as noted by Rannala & Bertorelle (2001) subclades within a genealogy do not follow the standard coalescent, but are conditional on the other parts of the genealogy and not independent, contrary to the assumption of Templeton's method. Thus, NCPA age inferences are not model-free, but are in fact based on a simple evolutionary model (isolated, random-mating and constant-size population) that is used precisely to establish that a different model applies! This weakness seems to have previously been overlooked, and suggests that evolutionary scenarios inferred by NCPA are not only based on unreliably-inferred demographic events, but also on a wrong timing of these events.

Likelihood ratio tests are not based on valid likelihoods

Multi-locus hypothesis testing in NCPA is based on the age distribution of inferred events, and basically evaluates the probability of a given number of loci showing NCPA-inferred events within a given time period. We now reexamine the theoretical foundations of this approach.

Templeton (2004a) proposed to take into account the stochasticity of the coalescent process by (incorrectly) assuming that T_k has a Gamma distribution with the same mean and variance as T_2 as derived by Tajima (1983). He obtained its distribution conditional on its mean ($\hat{T}_k$) and on $\overline{\pi}_k$ (defined above) as

$$f(T_k|\hat{T}_k\overline{\pi}_k) = \frac{\hat{T}_k^{\overline{\pi}_k} e^{-T_k(1+\overline{\pi}_k)/\hat{T}_k}}{\left(\frac{\hat{T}_k}{1+\overline{\pi}_k}\right)^{1+\overline{\pi}_k}\Gamma(1+\overline{\pi}_k)}. \quad (1)$$

Note that $\hat{T}_k$ is an estimate but is used here as if it were known without error, but that is a minor point compared to the use of this theory in an evolutionary context where it does not apply. Templeton (2004a) then uses eqn (1) to infer the probability that a given NCPA-inferred event E occurs before a given time T as

$$\Pr(T_E \le T) = \int_{t=0}^{T} f(t|\hat{T}_k, \overline{\pi}_k)\mathrm{d}t. \quad (2)$$

However, $\Pr(T_E \le T)$ is at best the probability that the TMRCA occurred before time T in a panmictic and stationary population. The use of eqn (2) as the probability of a given demographic event within a given time interval thus goes beyond the already doubtful assumption that the TMRCA of a clade can be used to date an inferred event. Indeed, it further assumes that the timings of these events are distributed as if they were coalescent times, which is a very strong assumption. This assumption is invalid because phenomena like vicariance events or episodes of intercontinental gene flow (or any other NCPA-inferred event) will alter the distribution of coalescent time between two DNA sequences, which will therefore not follow eqn (1). Despite these problems, Templeton (2009a) used eqn (1) further to estimate the probability of no gene flow between two continents between times T_l and T_u where an episode of gene flow has been dated by NCPA at the i-th locus at $\hat{T}_{k_i}$ as

$$\Pr(T_l \ge T_E \ge T_u) = \int_{t_i=T_l}^{T_u} f(t_i|\hat{T_{k_i}}, \overline{\pi}_{k_i})dt_i. \quad (3)$$

However, this equation merely describes the probability that two genes drawn from a stationary panmictic population and differing at $\overline{\pi}_{ki}$ sites do not coalesce between T_l and T_u, given their expected coalescence time of $\hat{T}_{k_i}$, and it has nothing to do with the probability of an absence of gene flow between continents. It follows that such an equation cannot be used in likelihood ratio tests as proposed for NCPA, and that these likelihood ratio tests are not testing phylogeographic hypotheses. Moreover, these likelihoods cannot be simply fixed as their inapplicability does not stem from mathematical errors, but from a misinterpretation of what they are supposed to describe. Therefore, Templeton's assertions that NCPA 'multilocus tests are based on explicit probability distributions and likelihood ratios' (Templeton 2009a, p. 322), or that NCPA uses 'a likelihood function that explicitly incorporates the randomness associated with the coalescent and mutational processes' (Templeton 2010) are wrong.

Problems with multi-locus NCPA

In a recent article, Panchal & Beaumont (2010) have evaluated the merit of the multi-locus method promoted by Templeton, using an automated program (ANeCA-ML). They have simulated multi-locus test data sets under a variety of conditions and analysed them under NCPA following closely the descriptions in Templeton (2002, 2004a,b). Four demographic scenarios are considered: panmixia, as described in Panchal & Beaumont (2007); an island model; a strict 4-neighbor stepping stone model; a lattice model with a Cauchy dispersal kernel allowing for long-distance dispersal. All the demes are laid out in a 2-D lattice (of sides 3, 7 and 10 demes) to provide geographical coordinates for NCPA. The data consist of sets of 5 loci, each of 500 bp, evolving under a Kimura 2-parameter model.

The multi-locus analysis reveals a number of problems in addition to those described above for single-locus NCPA:

1 Inferences are identified as cross-validated in NCPA when two different clades provide the same inference for the same geographical region, but the necessary degree of overlap of geographical area covered by clades is unspecified and thus arbitrary. For example in Templeton (2002), because of large variation in sample

size among loci, inferences were deemed cross-validated if the two loci agreed on the same continent.

2 Hypotheses of concordance of temporal events can be rejected for a group of clades, but then subsets of these may be found in which the hypothesis is not rejected. Each of these subsets can then be deemed to support the hypothesis of a particular event. For example, in Templeton (2002, fig. 3) the hypothesis that all 5 loci support the same temporal event is rejected. But then the loci are grouped (apparently by eye from their inferred event times) into a set of two loci (mtDNA and Y-chromosome), and a set of three (autosomal) loci, leading to a claim that forms the basis of the entire paper that there were two colonization events out of Africa. As an aside, no account seems to be taken that these differences are to be expected from the different mode of inheritance of these loci, and therefore the claim may be baseless.

3 There is no relative weight attached to the various inferences that result from NCPA. The outcome is all-or-nothing with no measure of uncertainty, whereas a model-comparison procedure (either Bayesian or frequentist) would allow for the possibility that the data supported an island model but also a lattice model, or it supported a stepping-stone model and a lattice model.

4 The reduction in false-positive rate arising from the use of multiple loci is very patchy, and depends on the (generally unknown) true scenario. This contrasts with the case of model-based inference in which the false positive rate is generally well controlled and the whole motivation for using more loci is to increase power (Rannala & Yang 2003).

Table 1 summarizes results found in Panchal and Beaumont (2010). It can be seen that with more loci the false-positive rate is indeed reduced, but due to the very specific nature of the inferences yielded by NCPA, it is highly variable across simulated scenarios. For example under the stepping stone model only Restricted Gene Flow (RGF) with isolation by distance is regarded as a true positive, and any inference including Long Distance Dispersal (LDD) is regarded as a false-positive. Under the lattice model with LDD, a much larger range of inferences are allowed that include RGF with isolation by distance, and RGF with LDD. In the island model all inferences of RGF are regarded as true positives as long as they do not include isolation by distance. A direct consequence of this is that the false positive rate for the island model remains very high (54%) whereas that for the lattice model with LDD is less than 5%. In the latter case a much wider range of inferences were deemed consistent with the scenario, whereas with the island model any inference with isolation-by-distance was deemed a false positive. The rates decrease with increasing lattice size, and increase with increasing level of population structure. The rates for single loci are typically always quite high. In conclusion, the use of multiple loci tends to reduce the false-positive rate in NCPA. However when there is population structure, it does not lead to improved discrimination among its possible causes because in this case the most frequent inference is restricted gene flow with isolation by distance, irrespective of whether the data comes from an island model or a stepping stone model.

Conclusions

Gleaning useful information about evolutionary processes from population genetic data is hard, and requires appreciation of the mathematical and conceptual underpinnings of population genetics theory. Such requirements are taken for granted by experimentalists in the physical sciences, while in evolutionary biology there remains a tendency to treat statistical procedures uncritically as 'black boxes', and to accept apparently easy solutions, especially those that fit with common-sense nostrums. We argue here that the need for rational, quantitative assessment of population genetics models and estimates is unavoidable.

In this article we have demonstrated that the majority of criticisms by Templeton (2009a, 2010) of ABC are in fact directed at model-based inference more generally, and are unfounded. Other criticisms arise from profound misconceptions of the ABC procedure itself, and are easily rebutted. Templeton promotes the use of NCPA, and we demonstrate that, despite its past popularity among empiricists, there are many problems associated with the method: there is no justification for the adoption of specific alternative hypotheses following the rejection of a simple null hypotheses by a permutation test; there is no measure of confidence in its support for hypotheses or estimates; the inference key of NCPA has not been properly validated, including error rate estimates; the ages of inferred events are estimated from a simple evolutionary model (the standard coalescent) in precisely those situations when it does not apply; the likelihood ratio tests are not based on valid likelihoods. As a result, it maintains a highly inflated false positive rate, even when applied to multi-locus data.

Current model-based statistical methodology does not match in scope the breadth of inference claimed by NCPA, but the latter's claims are not based on real, external validation. ABC has limitations, but like full-likelihood methods, it is based on explicit models, uses all the data simultaneously in inference, and allows an assessment of uncertainty in all inferences. Geographic and genetic information are intimately linked (Novembre *et al.* 2008), and the use of geographic information can certainly bring additional insights on past evolutionary processes such as environmental adaptations, range expansions and migrations. While most inferential approaches integrating geography only use information on allele frequencies (e.g. Guillot *et al.* 2005; Novembre *et al.* 2005; Francois *et al.* 2006; Corander *et al.* 2008), coalescent-based approaches seem in an ideal position to enable us to integrate molecular information into phylogeographic inferences (see e.g. Manolopoulou 2008; Itan *et al.* 2009;). Ongoing advances in computation and methodology will undoubtedly yield increasing flexibility in the range of evolutionary and historical scenarios that can be considered, ensuring a major role for model-based approaches in reconstructing realistic demographic

and evolutionary scenarios from the spatial distribution of genetic data. It should enable us to have a better appreciation of the complex and subtle relationships between demographic history, natural selection, and genomic diversity.

Acknowledgements

We thank the editor and three anonymous reviewers for their helpful comments on an earlier version.

References

Beaumont MA, Panchal M (2008) On the validity of nested clade phylogeographical analysis. *Molecular Ecology*, **17**, 2563–2565.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Beaumont MA, Cornuet J-M, Marin J-M, Robert CP (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. *Biometrika*, **96**, 983–990.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. *PLoS Genet*, **3**, e66.

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Berger JO (1990) Robust bayesian-analysis—sensitivity to the prior. *Journal of Statistical Planning and Inference*, **25**, 303–328.

Berger JO, Sellke T (1987) Testing a point null hypothesis—the irreconcilability of P-values and evidence. *Journal of the American Statistical Association*, **82**, 112–122.

Blum MGB, François O (2010) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing* (in press).

Bonneuil N (1998) Population paths implied by the mean number of pairwise nucleotide differences among mitochondrial DNA sequences. *Annals of Human Genetics*, **62**, 61–73.

Bortot P, Coles SG, Sisson SA (2007) Inference for stereological extremes. *Journal of the American Statistical Association*, **102**, 84–92.

Box GEP, Draper NR (1987) *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, Oxford, UK.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees*. Chapman and Hall/CRC Press, New York.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. *Evolution*, **32**, 550–570.

Corander J, Marttinen P, Siren J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Cornuet JM, Santos F, Beaumont MA, et al. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Crandall KA (1996) Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. *Molecular Biology and Evolution*, **13**, 115–131.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.

Del Moral P, Doucet A, Jasra A (2009) An adaptive sequential Monte Carlo method for Approximate Bayesian Computation. *Annals of Applied Statistics*, http://www.cs.ubc.ca/~arnaud/delmoral_doucet_jasra_smcabc.pdf.

Di Rienzo A, Donnelly P, Toomajian C, et al. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, **148**, 1269–1284.

Eswaran V, Harpending H, Rogers AR (2005) Genomics refutes an exclusively African origin of humans. *Journal of Human Evolution*, **49**, 1–18.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Fagundes NJ, Ray N, Beaumont M, et al. (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17614–17619.

Felsenstein J (1981) Evolutionary trees from DNA-sequences—a maximum-likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Fisher RA (1925) *Statistical Methods for Research Workers*. Oliver and Boyd, London.

Francois O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–760.

Gigerenzer G, Swijtink Z, Porter T, et al. (1990) *The Empire of Chance: How Probability changed Science and Everyday Life*. Cambridge University Press, Cambridge.

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. *Statistical Science*, **9**, 307–319.

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Hamilton G, Currat M, Ray N, et al. (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, **170**, 409–417.

Hey J, Machado CA (2003) The study of structured populations—new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535–543.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hudson RR (1983) Testing the constant rate neutral allele model with protein sequence data. *Evolution*, **37**, 203–217.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. *PLoS Computational Biology*, **5**, e1000491.

Joyce P, Marjoram P (2008) Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article26.

Kingman JFC (1982) The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1–10.

Knowles LL (2008) Why does a method that fails continue to be used? *Evolution*, **62**, 2713–2717.

Knowles LL (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics*, **140**, 1421–1430.

Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics* (in press).

Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–2749.

Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM (2009) The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 14711–14715.

MacKay DJC (2002) *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

Macneil MA (2008) Making empirical progress in observational ecology. *Environmental Conservation*, **35**, 193–196.

Manolopoulou I (2008) *A Bayesian approach to Nested Clade Analysis*. University of Cambridge, Cambridge.

Marjoram P, Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, **7**, 759–770.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15324–15328.

McKinley T, Cook AR, Deardon R (2009) Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, **5**, Article 24.

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**, 358–364.

Nielsen R, Beaumont MA (2009) Statistical inferences in phylogeography. *Molecular Ecology*, **18**, 1034–1047.

Novembre J, Glavani A, Slatkin M (2005) The geographic spread of the CCR5 D32 HIV-resistance allele. *PLoS Biology*, **3**, e339.

Novembre J, Johnson T, Bryc K, et al. (2008) Genes mirror geography within Europe. *Nature*, **456**, 98–101.

Pamilo P, Nei M (1988) Relationship between gene trees and species trees. *Molecular Biology and Evolution*, **5**, 568–583.

Panchal M, Beaumont MA (2007) The automation and evaluation of nested clade phylogeographic analysis. *Evolution*, **61**, 1466–1480.

Panchal M, Beaumont MA (2010) Evaluating nested clade phylogeographic analysis under models of restricted gene flow. *Systematic Biology* (in press).

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; author reply 473–465.

Platt JR (1964) Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, **146**, 347–353.

Popper KR (1959) *The Logic of Scientific Discovery*. Hutchinson, London.

Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Rannala B, Bertorelle G (2001) Using linked markers to infer the age of a mutation. *Human Mutation*, **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.

Ratmann O, Andrieu C, Wiuf C, Richardson S (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10576–10581.

Schervish MJ (1995) *Theory of Statistics*. Springer-Verlag, New York.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1760–1765. Errata (2009), 1106, 16889.

Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate bayesian computation without summary statistics: the case of admixture. *Genetics*, **181**, 1507–1519.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, **173**, 1511–1520.

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Templeton AR (2002) 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 Horizon* (ed Wasser SP). pp. 209–230, Kluwer Academic Publishers, Dordrecht.

Templeton AR (2004b) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.

Templeton AR (2007) Genetics and recent human evolution. *Evolution*, **61**, 1507–1519.

Templeton AR (2008) Nested clade analysis: an extensively validated method for strong phylogeographic inference. *Molecular Ecology*, **17**, 1877–1880.

Templeton AR (2009a) Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, **18**, 319–331.

Templeton AR (2009b) Why does a method that fails continue to be used? The answer. *Evolution*, **63**, 807–812.

Templeton AR (2010) Coalescent-based, maximum likelihood inference in phylogeography. *Molecular Ecology*, doi: 10.1111/j.1365-294X.2009.04514.x.

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 CA (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.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Thompson EA (1973) The Icelandic admixture problem. *Annals of Human Genetics*, **37**, 69–80.

Toni T, Welch D, Strelkova N, Ipsen A, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient Approximate Bayesian Computation coupled with Markov Chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.

The authors are all involved in the development of model-based inference methodologies in population genetics, phylogenetics, phylogeography, or statistics.

---

doi: 10.1111/j.1365-294X.2009.04515.x