Molecular Ecology (2009) **18**, 1034–1047 doi: 10.1111/j.1365-294X.2008.04059.x

INVITED REVIEW

Statistical inferences in phylogeography

RASMUS NIELSEN*† and MARK A. BEAUMONT‡

*Departments of Integrative Biology and Statistics, University of California, Berkeley, 4096 VLSB, Berkeley, CA 94720, USA, †Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Kbh O, Denmark, ‡School of Biological Sciences, University of Reading, PO Box 68, Whiteknights, Reading RG6 6BX, UK

Abstract

In conventional phylogeographic studies, historical demographic processes are elucidated from the geographical distribution of individuals represented on an inferred gene tree. However, the interpretation of gene trees in this context can be difficult as the same demographic/geographical process can randomly lead to multiple different genealogies. Likewise, the same gene trees can arise under different demographic models. This problem has led to the emergence of many statistical methods for making phylogeographic inferences. A popular phylogeographic approach based on nested clade analysis is challenged by the fact that a certain amount of the interpretation of the data is left to the subjective choices of the user, and it has been argued that the method performs poorly in simulation studies. More rigorous statistical methods based on coalescence theory have been developed. However, these methods may also be challenged by computational problems or poor model choice. In this review, we will describe the development of statistical methods in phylogeographic analysis, and discuss some of the challenges facing these methods.

Keywords: Coalescence theory, likelihood based inference, phylogeography

Received 17 June 2008; revision revised 22 October 2008; accepted 30 October 2008

Introduction

The objective of phylogeographic studies is to use phylogenetic methods for elucidating historical and ancestral processes in a geographical context (e.g. Avise *et al*. 1987; Avise 1989). The field rose to prominence in the late 1980s and early 1990s when estimation of gene trees from human mitochondrial DNA (mtDNA) placed the root of the human gene tree in Africa, supporting the out-of-Africa hypothesis (e.g. Vigilant *et al*. 1991; Fig. 1). Since then, phylogeographic studies have become one of the central pillars of population genetic analysis. Basic phylogeographic analysis typically consists of the estimation of a tree using phylogenetic methods, or possibly the estimation of a network. The branches of the tree are then related to historical events in a geographical context. For example, the emergence of a clade only existing in a particular area may be interpreted as evidence of a historical event by which a population, or group of individuals, separated themselves from other individuals. Likewise, if the individuals in two geographical areas form reciprocal monophyletic clades, with the exception of one or a few individuals, the existence of these individuals may be interpreted as evidence for migration between the two geographical areas. In general, a basic premise for much phylogeographic work is that the branches on the tree can be interpreted as evidence for the occurrence of specific historical demographic events in a geographical context.

The use of phylogeographic methods has bloomed during a period where population genetic theory also has been increasingly dominated by tree-thinking. The seminal work by Kingman (1982) and Hudson (1983) laid the foundations for modern coalescent theory. Coalescent theory provides a mathematical framework which describes the distribution of gene trees in populations. It can be used as a mathematical tool for deriving theoretical population genetic results, and also more directly, to connect demographic models with gene trees. Coalescent theory has helped merge the areas of population genetics and phylogenetics, making the gene tree the focus of study in both areas. However, while the tradition from phylogenetics is to estimate a tree and use the estimated tree to deduce evolutionary relationships, the population genetic tradition

Correspondence: Rasmus Nielsen, Fax: (510) 643-6264, E-mail: rasmus_nielsen@berkeley.edu

Fig. 1 The human mtDNA tree has its root in African populations (A) and not in populations outside Africa (O). This has been interpreted as independent evidence supporting the out-of-Africa theory of an origin of modern humans on the African continent.

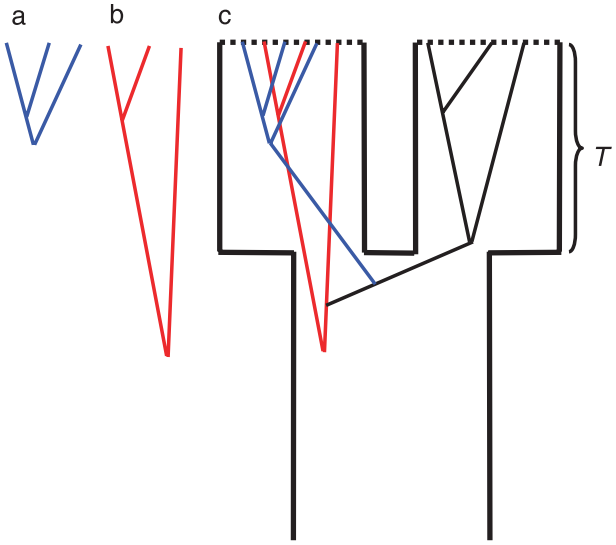

Fig. 2 Gene trees sampled within populations have a strong random component and the tree in (a) and in (b) could easily have been obtained under the same demographic histories. The fact that gene trees are random also implies that there is a strong random component associated with the presence or absence of reciprocal monophyly between populations and the placement and time of the most recent common ancestor (c).

sees the tree as a random outcome of a population genetic process. Therein lies the fundamental difference between phylogeographic and theoretical population genetic thinking: phylogeographic studies traditionally assume that ancestral history can be directly deduced from estimated gene trees, whereas population genetic theory asserts that gene trees are random outcomes of stochastic population-level processes.

It was recognized quite early on that phylogeographic studies must be followed by statistical inferences (e.g. Templeton *et al*. 1995). One line of research for doing this is based on coalescent theory and tools from computational statistics (e.g. Slatkin 1987; Griffiths & Tavare 1994a, b; Kuhner *et al*. 1995; Wakeley & Hey 1997; Beerli & Felsenstein 1999; Nielsen & Wakeley 2001; Beaumont *et al*. 2002). Another approach is based on analysing estimated gene trees, or networks, in a cladistic framework (e.g. Templeton *et al*. 1987, 1995, 1998, 2004; Posada *et al*. 2000, 2005, 2006). In this review, we will discuss and compare these two different methodological frameworks. The aim is to give an overview for the nonmathematical practitioner that introduces current methods of statistical inference in phylogeographic studies and provides some guidelines for their use. We will not provide a detailed description of individual computer programs for performing data analyses, but instead refer to the recent review by Excoffier & Heckel (2006). We will start this review by briefly discussing some of the challenges facing phylogeographic analysis, mainly from the perspective of theoretical population genetics.

Why are trees random?

A basic insight from population genetic theory is that gene trees sampled from different individuals in a population are random realizations of a stochastic process. To realize this, it may help to think of the gene trees describing the ancestry of a genetic marker, for example mtDNA, in three individuals in a population (Fig. 2). Individuals who are very closely related (through the maternal line) share a common ancestor very recently, and individuals who are distantly related share a more ancient most recent common ancestor (MRCA). The gene trees relating the individuals then depend on which individuals we have sampled. Three individuals which are very closely related will have gene trees which tend to be very shallow with late MRCAs (Fig. 2a), but if one of the individuals happens to be more distantly related, the gene trees will be deeper with a more ancient MRCA (Fig. 2b).

Clearly, the (mtDNA) gene trees relating individuals within a population will differ depending on which individuals we have sampled.

However, it is even more important to realize that the distribution of gene trees may be radically different among populations that have experienced the exact same demographic history, i.e. same population sizes and same geographical distributions, through time, and among loci within the same population. The randomness of gene trees arises because some individuals leave many offspring and others only a few, and because of the random segregation of alleles in diploid organisms. If all females in a population always have exactly one female offspring in each generation, there would be no mtDNA gene tree and the

mitochondrial evolutionary lineages would extend back to the dawn of time. However, in real populations, there is some variance in the number of copies of a particular allele transmitted to the next generation. In each generation, some lineages will die out because of individual(s) not transmitting the alleles they carry to the next generation, while other lineages will emerge as some individuals pass the same allele on to multiple offspring. The structure of the gene tree then depends on which particular individuals happened to leave descendents in the next generation(s) and, in nuclear loci of diploid organisms, which of the two alleles in a locus that was transmitted during reproduction. Thus, even if we consider all of the individuals in the population, the time to the MRCA and other properties of the tree will have a strong random component. It will depend on those ancestral random events occurring while offspring replaced parents in past generations. For example, in a large diploid population of constant effective size N_e, the expected time, and the variance in the time to the MRCA of all individuals is $4N_e$ and $16/3N_e^2(\pi^2-9)$, respectively. Notice the very large variance (it would be smaller if the population had experienced population growth), which tells us that the time to the MRCA of a single locus may in itself not be very informative about the population history — even if we could estimate it with absolute certainty. The age of the 'mitochondrial Eve' in humans, may provide a bit of information about the effective population size of humans but, without making a series of further assumptions, it tells us very little about the origin and demographic history of modern humans. When sampling mtDNA or Y chromosome data, we are only observing one possible genetic history out of many possible genetic histories for the same population, and it may be misleading to interpret the structure of the gene tree, and properties such as the time to the MRCA too strongly. For nuclear data, the problem can be circumvented by considering many loci at the same time.

The fact that gene trees from individuals within a population are random also implies that they are similarly random when using individuals from different populations. For example, if we have sampled individuals from two populations which diverged from each other T generations in the past, the gene tree of the sampled individuals for a particular marker may or may not show reciprocal monophyly. If T is on the same order as the effective population size for one of the populations, the individuals in the population may (Fig. 2c, blue), or may not (Fig. 2c, red), have found an MRCA at T generations in the past when the populations split. If they have not found an MRCA, individuals may be more closely related to each other between populations rather than within populations. Such events, known as lineage sorting (e.g. Hudson 1983; Tajima 1983; Neigel & Avise 1986; Pamilo & Nei 1988), may seriously confuse studies based on the assumption that gene genealogies directly reflect population histories. The probability that lineage sorting occurs depends on a number of factors, including divergence times and effective and historical population sizes.

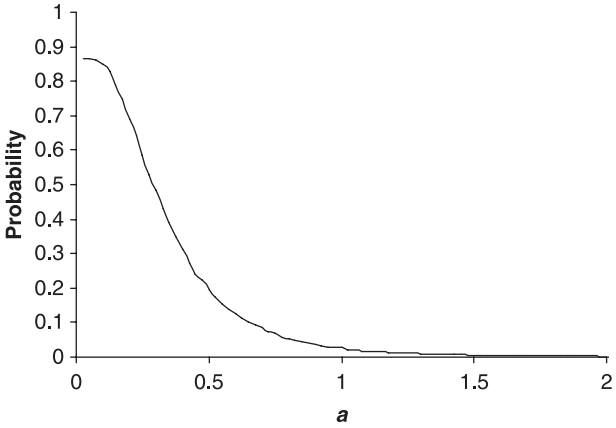

Fig. 3 The probability of monophyly of one population and the root in the tree falling inside the other population, as a function of the relative population size (a) of the two populations.

Phylogeographic uncertainty

The randomness of genealogies has profound implications for the interpretation of estimated gene trees. Any particular structure of the tree may have arisen, not because of a particular ancestral demographic, such as movement of people, but because of the random processes by which some individuals leave many descendents and some leave only few. Armed with this knowledge, it may be worth considering the old problem of inferences regarding the location of the root in a gene tree. For example, the placement of the root in humans within the African population has been interpreted as evidence for an African origin of humans. However, the placement of the root is strongly affected by random events and the relative population sizes. To illustrate this, we used a very simple model in which two populations, of size N and aN, diverged from each other $N/2$ generations ago, and have experienced no gene flow between them. The ancestral population size is also equal to N and we assume 30 gene copies have been sampled from each population. We then find the probability that the first population is monophyletic, while the second is not, so that the root is unambiguously placed in the second population (Fig. 3).

As expected, the relative effective population sizes determine the chance that the root falls within the first population. Placement of the root may not, without further assumptions, identify which population is ancestral. Only if we make the additional assumption that the derived population always has a smaller effective population size can we interpret the placement of the root as unambiguously informative regarding which population is ancestral.

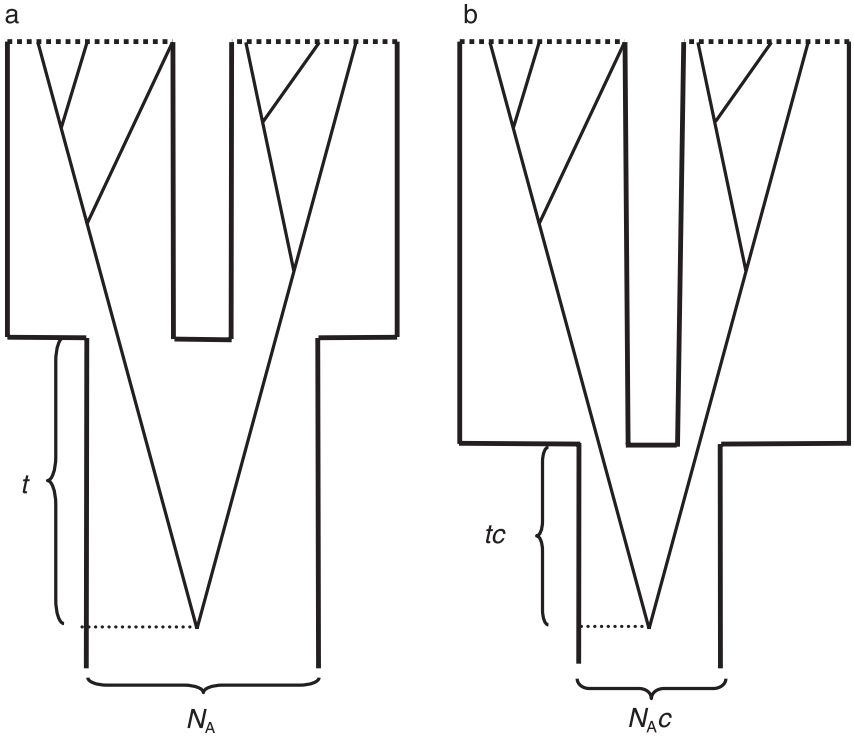

Fig. 4 A model with recent divergence between populations (a) and large ancestral population size (N_A) may, in the presence of reciprocal monophyly, provide just as good a fit to the data as a model with old divergence and small ancestral population (b). If the time the two ancestral lineages diverge from each other in the ancestral population (t) is changed by a factor, c, a model with ancestral population size N_Ac will provide exactly as good a fit as the original model.

But in that case, it might be easier simply to estimate the effective population sizes to determine which population is ancestral.

The relationship between demographic models and gene trees can at times be highly complex. There may be multiple demographic models that fit a gene tree equally well. For example, ancestral population sizes and divergence times can sometimes be confounded. With data from a single locus (e.g. mtDNA), models with short divergence times and large ancestral population sizes may sometimes provide just as good an explanation of a gene tree as a model with long divergence times and small ancestral population sizes (Fig. 4). Long divergence times in a gene tree may be caused by either large ancestral population sizes or long divergence times in the population tree. Phylogeographic studies face the challenge that there may be many equally parsimonious demographic explanations for the same data. As the complexity of proposed demographic scenarios increase, the chance that there are multiple equally parsimonious demographic explanations of the data also increase. Only by considering explicit demographic models will it be possible to determine which population histories are compatible with the data and which are not.

Assumptions and inferences

Given the preceding discussion, one might get the impression that it is near impossible to make meaningful statements about ancestral population history based on gene trees. This contrasts with the fact that phylogeographic analyses have been a tremendously powerful tool in the analysis of population genetic data. The reason might be that there is a set of realistic assumptions under which the gene tree directly reflects population history: if population migration events always are associated with population bottlenecks, new populations are more likely to be monophyletic because many lineages will be forced to coalesce during the bottleneck. It is possible that the bottleneck assumption is justified in a number of species, where new geographic territory is colonized by only few individuals. For example, it could be assumed that only very few individuals were involved in the first migrations of humans out of the African continent. Simple phylogeographic analyses may lead to the right results, even without the use of any statistical methods or model fitting, if population movements always are associated with very strong bottlenecks in the population size, and if there has been a subsequent absence of gene flow among populations. However, if the assumption of population movements being associated with strong bottlenecks is not true, naive inferences from gene trees may be misleading.

Nested clade phylogeographic analysis

Nested clade phylogeographic analysis (NCPA) is a method for turning phylogenetic information into inferences about the demographic history of populations. The procedure was first introduced in Templeton *et al*. (1995), based on an earlier cladistic approach (nested clade analysis, NCA)

designed to study association between genotype and phenotype (Templeton *et al*. 1987, 1988, 2005; Posada *et al*. 2000, 2005, 2006). NCA is based on the idea that we cannot study the associations between phenotype and genotype as if observations are independent, because there is an underlying correlation structure induced by the genealogy (Felsenstein 1985). In the original NCA procedure it was noted that there is a similarity between the natural structure of a gene tree and a hierarchically nested analysis of variance: lineages can be grouped together into clades, which can then be nested in an agglomerative way. Thus, by performing a nested analysis of variance on phenotypic traits it is possible to test whether there is significant variation in the trait along different branches of the tree. In this regard NCA is related to the comparative method (Felsenstein 1985; Harvey & Pagel 1991; Pagel 1997) in which the association between two traits is analysed, taking into account phylogenetic history. Whereas the comparative method has been concerned with the evolution of traits along phylogenetic trees, NCA has focused on the evolution of traits within populations. The NCPA procedure naturally arises from the early NCA by regarding the geographical location as a phenotypic trait evolving along the genealogy.

The first step in standard NCPA is to construct a haplotype network. The network is then used to group haplotypes into nested clades. Given the clade structure and either the geographical coordinates of sampling locations, or a matrix of geographical distances between locations, it is then possible to compute statistics that describe the geographical spread of clades. Associated with each clade are a number of statistics, principal among them D_c, which measures the geographical spread of members of a clade relative to their mean location, and D_n, which measures the geographical spread of members of a clade relative to the mean location of all members of the nesting clade. By permuting the geographical locations of samples, the probability of observing as extreme or more extreme values of these statistics can be computed. If any clade has at least one significant statistic (typically at the 0.05 level), then an 'inference key' is consulted, providing a qualitative interpretation of the result. The key leads to statements such as:

12 Are the D_n and/or I–T D_n values significantly reversed from the D_c values?

- No – contiguous range expansion.
- Yes – go to step 13.

These statements are then used directly in published analyses. Since there may be many clades with significant statistics, the key is often consulted many times, and this may lead to a number of different inferences. Those conclusions that are derived from higher-level clades in the nesting hierarchy are assumed to pertain to events that occurred earlier in time.

The key was first enunciated in Templeton *et al*. (1995), based on detailed reasoning outlined in that study. The justification for particular inferences is plausible. For example, it is intuitively plausible that under isolation by distance, the D_c statistic should tend to increase with increasing clade level, as noted in question 4 of the inference key of Templeton (2004).

The procedure has been extended to deal with multiple loci (Templeton *et al*. 2002). In this case, inferences are regarded as concordant in space if two or more loci infer the same process at the same location.

Ambiguity in the NCPA

The multiple steps of the NCPA procedure each rely on a series of specific methodological choices. For example, most sequence data will not give rise to a uniquely parsimonious network due to the presence of homoplasy, caused by recurrent mutation and recombination. Some additional assumptions must, therefore, be introduced to accommodate homoplasy. The TCS package, which implements a method called statistical parsimony (Clement *et al*. 2000), is often used to construct the haplotype network. TCS accommodates homoplasy by inserting additional edges into the network representing alternative evolutionary paths, producing loops in the network. However, there are many other algorithms available, and it appears that these may often lead to rather different networks (Cassens *et al*. 2005).

The second step in the NCPA procedure, the construction of nested clades, also relies on specific methodological choices. Most publications aim to reproduce the procedures suggested in the studies of Templeton and colleagues (Templeton *et al*. 1987; Templeton & Sing 1993), although, as noted in Panchal & Beaumont (2007), there are some uncertainties in the interpretation of the published suggestions for nesting.

However, arguably the biggest level of ambiguity is introduced by the interpretation of the inference key. Given the complexity of information that could be covered by the key, one might speculate whether different inferences could also have been obtained equally plausibly by focusing on other patterns in the data. To put it another way, each question in the inference key requires the user to look at a particular pattern in their data (such as in question 12, above). The decision to focus on this pattern, among the myriad patterns that are possible in the data, has been made by the author of the key, and it is not clear whether different inferences would be made had a different, similarly plausible, question been asked.

Performance of NCPA in simulation studies

The performance of NCPA has been investigated either by applying the method to simulated data, in which the true

demographic history is known (Knowles & Maddison 2002; Panchal & Beaumont 2007), or to data from natural populations in which the history is assumed to be known (Templeton 2004). A problem with NCPA is that it was not developed as an automated procedure. Two key places — the nesting of clades and the consultation of the inference key — require some judgment by the user. This has meant that its performance on simulated data sets has been difficult to examine. Knowles & Maddison (2002) performed simulations in which they simulated a population tree for three populations (i.e. a common ancestral population that is then subdivided by two sequential vicariance events). They then applied the NCPA procedure 'by hand'. They examined 10 cases, and the results indicated that the NCPA procedure was unable to recover the true scenario of allopatric fragmentation. In addition, since demographic processes were inferred that were not actually simulated in the data, they concluded that NCPA had a high false-positive rate. Knowles and Maddison noted that for the particular parameter settings they used, it might have been difficult for any method to infer the true population history, but emphasized that standard approaches typically had 'honest' false-positive rates whereas NCPA typically appeared to result in an inference that was not correct.

It was uncertain, however, whether these false-positives arose because the method detected some pattern or structure in the data but misattributed the cause, or whether there is an intrinsic feature of the method that generates false positives. To examine this further, Panchal & Beaumont (2007) automated NCPA by devising algorithms to perform the nesting of clades and the consultation of the key, and then pipelined these together with the application of TCS and GeoDis to provide a unified process that takes DNA sequences and their geographical location and outputs inferences from the key. Panchal & Beaumont (2007) simulated sequence data from a panmictic population, but allocated the sequences randomly to geographical locations on a lattice. A very high false-positive rate was the main observation from this investigation, agreeing with the earlier observation of Knowles & Maddison (2002). For a given data set, there is often a much greater than 50% chance that the inference key is consulted at least once even if there is no geographical structuring in the data. Of the positive inferences generated by NCPA on these simulated data sets, the two that were most common were 'restricted gene flow with isolation by distance' and 'contiguous range expansion'. Panchal & Beaumont (2007) surveyed 68 publications published between 2000 and 2004 that used NCPA, and showed that these inferences were also the two most commonly observed in the natural data. Furthermore, there was a significant association between the rank ordering of the frequency of different inferences in the simulations and in the natural data.

The main reason for the high false-positive rate appears to be that the method does not account for the multiple testing problem that arises because there are many statistics associated with each clade. In the use of NCA for testing phenotype/genotype associations based on the program TreeScan (Posada *et al*. 2005; Templeton *et al*. 2005), a statistical procedure known as free step-down resampling (Westfall & Young 1993) is used to control for multiple testing. However, in the case of NCPA, the structure of tests is more complicated, and it is unclear how to rectify the multiple testing problem.

NCPA and the 'Forer effect'

Empirical studies often find that the outcomes of NCPA are consistent with known historical information, and are supported by other approaches, such as the model-based methods discussed below. For example Pfenninger & Posada (2002), studying a species of snail, and Sunnucks *et al*. (2006), studying two species of flatworm, compared results from NCPA with those of, respectively, Migrate (Beerli & Felsenstein 1999) and Fluctuate (Kuhner *et al*. 1995, 1998), both model-based coalescent methods, and concluded that there is good concordance between methods.

There are two possible explanations for the discrepancy between the conclusions of empirical studies and simulation studies. One possibility is that real data, somehow, are more suitable for NCPA than simulated data, as noted above. Chikhi & Beaumont (2005) discuss the possibility that demographic histories consisting of sequential bottlenecks could give a strong signal in the haplotype network that is easier to detect using NCPA. Alternatively, the inherently subjective nature of the NCPA procedure seems to enable researchers to find answers that coincide with those from other approaches. There is a tendency for NCPA to give a number of different 'answers' for a data set, particularly if there are a large number of clades. To some degree it is, therefore, up to the user to choose the most suitable among several possible answers. For example, in the study of Sunnucks *et al*. (2006), two inferences of long distance dispersal are considered unlikely and discounted, whereas inferences of fragmentation, contiguous range expansion, and restricted dispersal by distance appear to be consistent with other tests, while an inference of past fragmentation followed by range expansion, while not supported by any standard test, was considered reasonable. It is a well-known phenomenon in psychology that when predictions are sufficiently ambiguous, there is a tendency for subjects to find that their experiences are consistent with these predictions — an example is the 'Forer effect' ('Barnum effect') in personality assessment (Forer 1949). Forer gives an example whereby a group of people were subjected to a personality test but then provided with identical assessments (a list of 13 statements, for instance: 'you have a tendency to be

critical of yourself'; 'you have found it unwise to be too frank in revealing yourself to others'; 'disciplined and self-controlled outside, you tend to be worrisome and insecure inside'). Virtually all respondents agreed that the personality test was effective, and that their own assessments were accurate. Such a tendency, of course, underpins the widespread interest in newspaper horoscopes (Fichten & Sunerton 1983).

However, we should emphasize that the possibility for over-interpretation of results based on estimation of trees and networks is not unique to the NCPA method. There has been a long tradition, particularly in human genetics, to interpret estimated trees or network very strongly in a geographical context without properly accounting for the stochasticity introduced by the coalescent and/or applying explicit statistical methods. A more detailed discussion of this problem can be found in Goldstein & Chiki 2002).

Coalescent-based methods

Model-based inference in population genetics comes in a number of different flavours, discussed in more detail in, for example, Hey & Machado (2003) and Beaumont & Rannala (2004). In this study, we will concentrate on methods based on likelihood, which includes Bayesian inference. The likelihood function is simply the probability of obtaining the data (or any function proportional to this probability). By 'data', we mean the types of different genetic variants and their frequencies in a sample. In mathematical notation, the likelihood is given by $p(X \mid \Theta)$, where X is the data, Θ is a vector containing all the parameters of interest, and '$\mid$' is read as 'given' or 'conditional on', indicating here that the probability is calculated for a particular value of the parameters. Statisticians often prefer to base inferences on the likelihood function, because all the information in the data regarding the parameters is captured by this function. The parameters can, for example, be migration rates, effective population sizes, and/or population growth rates. Estimates of parameters can then be obtained, for example, by maximum likelihood. The maximum-likelihood estimate of a parameter is the value of the parameter that maximizes the likelihood, i.e. which gives the highest value of $p(X \mid \Theta)$.

The calculation of the likelihood function in population genetics is very challenging. Only in very simple cases (e.g. Ewens 1972) can we explicitly write down a formula that gives the probability of the data, given values for a parameter. However, such formulas are not available in closed form, even for the simplest demographic models, for DNA sequence data, microsatellite markers, etc. A breakthrough in demographic inference came in the late 1980s and early 1990s with the observation that the probability of the data in population genetics could be calculated by combining computational methods from phylogenetics with coalescence models (Felsenstein 1988, 1992). While the field of phylogenetics had developed methods for connecting data with a tree, coalescence theory now provided mathematical methods for connecting demographic or ecological models with a tree. The likelihood function could then be calculated by considering all possible trees and multiplying the probability of the data given the tree with the probability of the tree given the demographic parameters. In mathematical notation, we write (Felsenstein 1988):

$$p(x \mid \Theta) = \int_{\Omega} p(x \mid T) p(T \mid \Theta) dT \qquad \text{(eqn 1)}$$

The integral is really a sum over all possible trees (T), and a multiple integral over all the possible branch lengths. Ω indicates the set of all possible trees. The first expression inside the integral is the probability of the data given the tree, which can be calculated using the methods from phylogenetics. The second term inside the integral is the probability of the tree given the parameters of the demographic model, which can be calculated using coalescence theory. This expression takes advantage of the fact that when the tree is known, the probability of the tree can be calculated without knowledge of the demographic model. The only problem here is that the integral, in general, cannot be solved directly by any known method. Instead, a number of different numerical and simulation-based methods have been developed.

Methods based on the full likelihood

The two main methods used to evaluate equation 1 using simulations are based on Markov chain Monte Carlo (MCMC, e.g. Kuhner *et al*. 1995, 1998; Wilson & Balding 1998; Beaumont 1999; Beerli & Felsenstein 1999; Nielsen 2000; Nielsen & Wakeley 2001; Hey & Nielsen 2004) and importance sampling (IS, e.g. Griffiths & Tavare 1994a, b; Nielsen 1997; Stephens & Donnelly 2000; Fearnhead & Donnelly 2001). Both methods are based on simulation of a large number of trees. If the methods are constructed correctly, calculations based on these samples of trees can provide a very close approximation to the likelihood function, and inferences proceed by finding maximum-likelihood estimates, or by the use of Bayesian methods (for a review, see Stephens 2007). In Bayesian methods, a so-called prior distribution, $p(\Theta)$, of the parameters is assumed, which enables the calculation of a posterior distribution, $p(\Theta \mid x)$. The prior and posterior distributions then summarize the researcher's knowledge about the parameter before and after observing the data.

These methods have been widely applied to a number of different scenarios: the standard model with constant population size with different mutation models (Griffiths & Tavaré 1994a, b; Kuhner *et al*. 1995; Wilson & Balding

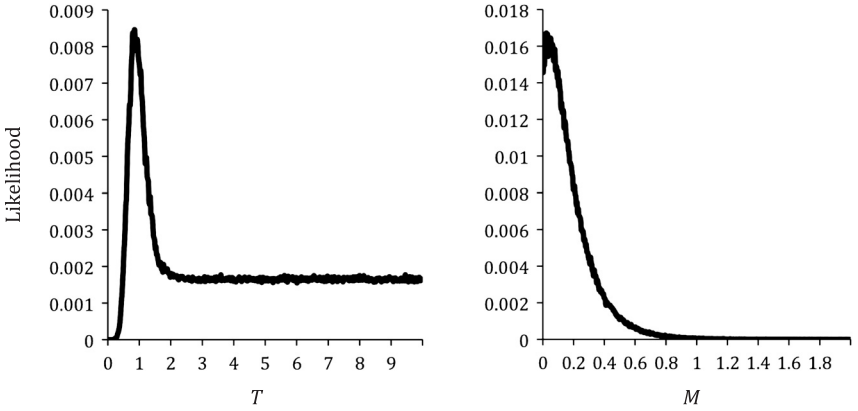

Fig. 5 Likelihood surface for divergence time (T) and migration rate (M) between western and central Eurasian wood lemmings (*Mycopus schisticolor*), estimated using the Mdiv program (Nielsen & Wakeley 2001) based on data from Fedorov *et al*. (2008).

1998; Nielsen 1997); models of varying population size (Griffiths & Tavaré 1994b; Kuhner *et al*. 1998; Beaumont 1999; models with migration (Nath & Griffiths 1996; Beerli & Felsenstein 1999, 2001; De Iorio *et al*. 2005); models with vicariance and migration (Nielsen & Wakeley 2001; Hey & Nielsen 2004); and even models with vicariance, migration, and population growth (Hey 2005). Two features are immediately apparent: the development of these methods occurred at a similar time to, and in parallel with, the work on network-based approaches such as NCPA; and, while not yet able to deliver the full panoply of inferences claimed for network methods, it is still possible to address a much wider set of questions than hitherto. A difficulty with these techniques is that they are complicated to program, which slows development, and it is difficult to obtain a sufficiently representative sample of genealogies over the parameter space, which manifests itself as poor convergence of the MCMC. To solve the latter problem typically requires extensive computational resources, and some ingenuity on the part of the user in tweaking the details of how the MCMC is run. These issues, which are discussed in more detail in the next section, have delayed the introduction of more model-based population genetic analysis, but have also encouraged the development of approximations.

An example of a likelihood-based inference is shown in Fig. 5. Fedorov *et al*. (2008) analysed cytochrome *b* sequences from the Eurasian wood lemming (*Myopus schisticolor*). They used the IM program (Hey & Nielsen 2004; Hey 2005; Won & Hey 2005), in addition to a number of other phylogeographic analyses, to elucidate the demographic history of this species. We have reproduced part of their analyses of DNA sequences from the western and eastern populations using the program Mdiv (Nielsen & Wakeley 2001). The results are presented in terms of the likelihood function of two of the parameters: M, the migration rate per generation multiplied by the effective population size, and T, the divergence time in generations between the populations, also multiplied by the population size.

Notice that the highest likelihood for M is found very close to zero. This implies that there is little or no evidence for ongoing migration between these two populations. In contrast, the estimate of T, given by the peak of the likelihood function, is about $T = 1$. The likelihood function for T is sharply peaked and the likelihood is close to zero for values of $T < 0.25$. As is often the case in analyses of a single locus, large values of T are more difficult to exclude when allowing for migration. The likelihood converges to a positive value for large values of T, corresponding to the likelihood obtained under a pure migration model without a vicariance event. However, the likelihood is larger for $T = 1$ than for high values of T, showing that the data do not support the hypothesis of equilibrium migration between the two populations ($T = \infty$).

The conclusions from the analysis of the Fedorov *et al*. (2008) data are not surprising, considering that there is reciprocal monophyly between the two populations. Reciprocal monophyly is more likely to arise when there is no ongoing gene flow. However, the likelihood analyses allowed us to quantify what the absence of reciprocal monophyly, and the other genealogical information in the data, implies about the demography of the species.

Approximate methods without mutation

There are a number of approximations which have been used when full likelihood methods have not been feasible, computationally tractable, or flexible enough. One useful class involves the assumption that the genetic variation we see at several locations in space or time can be regarded as being derived from a common ancestral stock of variation that is then distributed either spatially or temporally purely through the agency of random genetic drift. The advantage here is that the models deal with gene or haplotype frequencies, and do not take mutations into account. This simplifies the calculations considerably. Examples include the estimation of effective population

size from temporally spaced samples (Wang & Whitlock 2003; Anderson 2005), the estimation of immigration rates in island models (Balding & Nichols 1997; Foll & Gaggiotti 2006), the estimation of divergence times/effective population sizes in allopatric populations (Nielsen *et al*. 1998), and the analysis of admixture (Chikhi *et al*. 2001; Wang 2003). The use of these models is restricted to situations in which the mutation rate is typically much lower than the migration rate, or the reciprocal of the divergence time.

ABC methods

Recently, a group of techniques, variously called likelihood-free inference, or approximate Bayesian computation (ABC), have been quite widely applied in population genetics. These methods typically require the data to be compressed into summary statistics. A large number of simulations are then performed. The main idea is that an approximation of the likelihood — in this case, the probability of obtaining the observed summary statistics measured from the data — is proportional to the number of simulated data sets yielding summary statistics that lie within some small distance of the summary statistic computed from the observed data. With this approximation, it is then possible to apply all standard likelihood-based techniques for inference, both frequentist (Weiss *et al*. 1997) and Bayesian (Pritchard *et al*. 1999). The key to successful application of these methods is how well the summary statistics capture the relevant properties of the data, and how similar the approximated likelihood surface is to the true likelihood surface at a given distance from the target (observed) summary statistics. Typically, the likelihood surface varies rapidly with increasing distance from the target (Beaumont *et al*. 2002), and a variety of techniques have been developed to try to correct for this (Beaumont *et al*. 2002; Marjoram *et al*. 2003; Sisson *et al*. 2007).

The statistical properties of likelihood-free methods can be examined in precisely the same way as for likelihood-based techniques. Thus, it is possible to simulate data sets with known parameter values and examine performance. Indeed, it is particularly easy for ABC methods because the simulations that are used in inference need only be performed once, and then applied to many different data sets. Most published studies that have used ABC have therefore been able to examine in some detail the statistical properties of their implementation. In those cases where it has been possible to make comparisons with full-likelihood approaches, it has been shown that the estimates are generally concordant, although the power of the approximate method tends to be rather lower (e.g. Beaumont *et al*. 2002; Tallmon *et al*. 2004; Excoffier *et al*. 2005). Unlike NCPA, the coverage (the adherence to some expected false-positive rate) is generally good, although often rather conservative (e.g. Tallmon *et al*. 2004; Hamilton *et al*. 2005; Excoffier *et al*. 2005). However, in contrast to standard likelihood methods, there is no underlying theory, and thus, the properties of likelihood-free approaches have to be examined on a case-by-case basis.

An advantage of these methods is that it is possible to address highly parameterized models, potentially providing inferences of similar detail to those that are claimed for network-based techniques. A further advantage is that the relative likelihoods, or posterior probabilities of different models can be compared, for instance, the various scenarios that have been proposed for recent human demography (Fagundes *et al*. 2007). Additionally, they have been used to compare different invasion scenarios for economically important pests (Estoup *et al*. 2004; Miller *et al*. 2005), and to infer the demographic history of populations from ancient DNA (Chan *et al*. 2006). With the potential to have very complex models, there is always the question whether the data can support such complexity. Certainly, there are reasons for believing that a number of different demographic histories will give rise to genealogies of a similar structure (Wakeley 2004), as discussed in more detail in the next section. However, it will always be possible to use the simpler models as a baseline against which to compare more complex models.

PAC methods

Another, recent and less well-used approximation has been called the product of approximating conditionals (PAC) approach (Li & Stephens 2003). This can be explained by reference to the Ewens sampling formula discussed above. It is possible to derive this formula by considering the following procedure. A gene is chosen at random, and we note its type. We then choose another gene at random, and ask what is the probability of getting a gene that is the same or a different type, conditional on the type we have at the moment. This is repeated until the full sample size is achieved, at each stage asking what is the probability of getting a gene that corresponds to a particular type in the sample, or is of another type, conditional on the data already in hand. The product of all these probabilities, multiplied by a constant that is equal to the number of different ways of getting the same sample by this sequential approach, is equal to the Ewens sampling formula (Ewens 1972; see Durrett 2008; also http://www.math.leidenuniv.nl/~verduyn/djb_notes.ps for a useful review). A similar procedure can be used to obtain the finite-allele equivalent (commonly used for modelling population structure: Balding & Nichols 1995 Beaumont & Balding 2004). Interestingly, this probability does not depend on the order that the genes are chosen. However the formula can only be applied exactly for these two special kinds of model. Li & Stephens (2003) proposed to apply a result in Stephens &

Donnelly (2000), who used coalescent theory to derive an approximate formula for the probability of obtaining a gene of a particular type, given the data in hand. Originally, this formula was employed to provide an efficient way of sampling genealogies given the data. Because it is an approximation, when it is used to provide the likelihood directly — that is, to get an equivalent of the Ewens sampling formula — it no longer has the property of giving a probability that is independent of the order in which the genes are chosen (at least for sample sizes greater than two, where it is exact for all mutation models). What can be done in practice is to average the probability over a number of different sequences of draws. For microsatellites under a model of constant population size, evolving according to the stepwise mutation model, the computed approximate probability is indistinguishable from estimates obtained by importance sampling, and much quicker to obtain (Cornuet & Beaumont 2007). The original application was for recombining markers (Li & Stephens 2003), but there is scope for the PAC method to be widely used for different genetic markers and complex demographic models.

Composite likelihood methods

Composite likelihood methods are often used in population genetics when data, typically SNP genotyping or DNA sequencing data, have been obtained from many loci. A so-called composite likelihood function is formed by calculating the likelihood in individual nucleotide sites, or pairs of sites, and then taking the product of these likelihood functions (e.g. Nielsen 2000; Hudson 2001; Wooding & Rogers 2002; Polanski & Kimmel 2003; Adams & Hudson 2004; Marth *et al*. 2004; Williamson *et al*. 2005). The resulting function is not a likelihood function because the data are not independent due to linkage disequilibrium. However, estimation of demographic parameters is still possible, and there are some theoretical results, which suggest that these methods may have desirable statistical properties such as consistency (Fearnhead 2003; Wiuf 2006). Computationally, these methods are useful because they can be very fast and the computational time does not rapidly increase with the number of nucleotide sites sequenced. They form the background for one of the highly used methods for estimating recombination rates (Hudson 2001; McVean *et al*. 2002) and have also been employed to estimate divergence times, migration rates, and population growth rates (e.g. Nielsen 2000; Wooding & Rogers 2002; Polanski & Kimmel 2003; Adams & Hudson 2004; Marth *et al*. 2004; Williamson *et al*. 2005). The disadvantage of these methods is that they do not take advantage of haplotype structure, leading to a loss of information. Also, as the composite likelihood function is not a true likelihood function, confidence intervals and hypothesis testing must be performed using simulations.

Challenges to coalescent-based methods

Despite the clear benefits in using explicit models and explicit hypotheses in population genetics, model-based analyses also face a number of challenges.

Computational issues

Several of the methods discussed here are challenged by serious computational issues. In many cases, especially for large genomic data sets, or under very complex models, it is not possible to calculate the likelihood function even using simulation techniques. A number of problems are currently not accessible for analysis. There are no methods developed for calculating the likelihood function for most demographic parameter of interest in the presence of recombination. Even in the absence of recombination, some of the methods based on the full likelihood face serious computational challenges. Methods based on MCMC may not always converge well. It may be difficult to determine when convergence has been achieved, and in some cases, if the data sets are sufficiently large, convergence may not be achieved within a practical amount of time.

The use of approximate methods may lead to a loss of power, and how much power is lost depends on the details of the implementation. For example, in the ABC methods, the summary statistics have to capture as much of the relevant information as possible, but it can often be difficult to devise good strategies for selecting statistics. Developing powerful approximate methods applicable to large data sets is currently the focus of much research.

Models may be too simple

A clear limitation of any model-based method is that the model might be wrong. In fact, the real complexity of the demography of natural populations is unlikely to be captured by any simple model we could propose. In some cases, this may not affect inferences much, but in other cases it will. An example arises when trying to infer changes of population size in structured populations. It turns out that population structure very easily can be confused with changes in the population size. Imagine a model in which there are multiple different demes, but all individuals in the sample are obtained from only one deme. Most of the gene copies in the deme will find a common ancestor relatively recently, but a few may be descendents of migrants from other demes. These genes copies will not find a most recent common ancestor with the other gene copies in the deme until the lineages ancestral to the gene copies have migrated into the same deme and then coalesced. Depending on the number of demes, their population size, and the migration rate, this may take a long time. In terms of pairwise differences, most

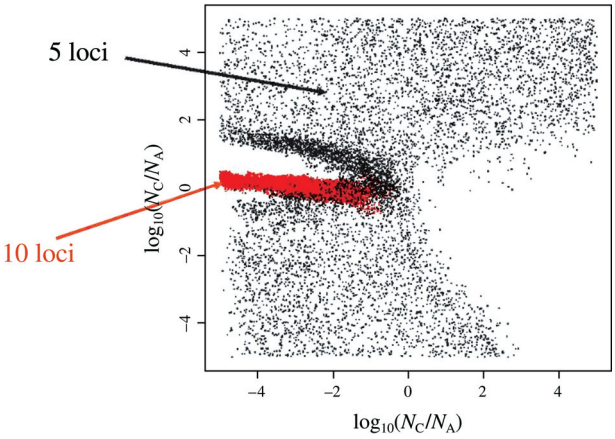

Fig. 6 This figure shows the results of applying the Msvar program to a data set from one deme in a simulation of 100 demes, where each immigrant has an equal probability of having come from any of the other demes. All demes are simulated assuming a constant population size. The probability, F, that two genes sampled from the deme share a common ancestor within the deme, without an intervening migration, is 0.2 (i.e. $F_{ST} = 0.2$, under some definitions). The Msvar program draws samples from the posterior distribution of parameter values in a model of population growth, and these points are plotted in the figure. Here, T is the time over which the population has changed in size, N_A is the ancestral population size, and N_C is the current population size. Ten loci were simulated (results shown in red), and also the results from a subsample of five loci is shown. The areas in the plot with more dots are areas with higher likelihood. It can be seen that by violating the assumption of a closed population, very strong inferences of population decline (or a 'bottleneck') can be made.

will be very small, but a few will be very large. This is exactly the pattern we would expect to observe if there had been a strong reduction in population size, or there had been a bottleneck. Without information about the existence of other demes, we cannot easily separate the hypothesis of a change in population size, from the existence of multiple other unsampled populations.

An example that illustrates this point is presented in Fig. 6, using the program Msvar by Beaumont (1999). Data were simulated assuming an island model with $F_{ST} = 0.2$, and the figure shows the distribution of samples from the posterior distribution of parameters, obtained using MCMC. These posterior distributions are obtained from two data sets — one with five loci, and the other with 10. It can be seen that with five loci, there is a very broad distribution of times and growth rates that are compatible with the data. Strong population growth (the empty 'hole' on the right hand side of the figure), and strong population contraction (empty hole on the left) are ruled out. However, there is a ridge with a higher concentration of points for parameter values giving contraction over timescales (in generations) in the order of 10 to 100 times the current population size. With 10 loci, by contrast, there is stronger evidence in favour of population contraction on a timescale of the order of the current population size. Thus, with 10 loci, one would conclude that there had been a population contraction, or 'bottleneck', although the population has maintained a constant population size.

While this problem is not unique to the method used here, or to explicit model-based methods in general, it does suggest that the most naïve interpretation of inferences based on very simple models should be avoided. Of course, ideally, the models themselves should be improved. In the case of the specific example given here, it would be possible to improve the model to take into account the effect of population structure using the approach of Wakeley (1999), provided that geographically separate samples are available.

No appropriate method available

Several of the methods discussed here are quite laborious to develop and test. As a consequence, there are a number of potentially useful models that have not been implemented. This includes full likelihood-based methods for inference of complex demography in the presence of recombination and models which allow both vicariance events and ongoing gene flow for more many populations at the same time. Obviously, this is just a practical obstacle that does not argue against the general use of these methods. However, it is in many cases a real serious problem for the practical use of these methods. An advantage of approaches that use ABC or composite likelihood, is that they easily allow the construction of inference methods for new scenarios as long as it is possible to simulate data under these models. The challenge is then to construct these methods so they are computationally efficient and maintain reasonable statistical power and accuracy.

Several of the methods discussed in this review, the full likelihood-based methods in particular, will face insurmountable computational challenges in the face of large-scale resequencing data. With the drastically reduced price of DNA sequencing based on next-generation sequencing technologies, such data is likely to become more and more common in population genetic studies. Techniques have been developed for cheaply sequencing just a fraction of a genome (Margulies *et al*. 2005; Bentley 2006), even in organisms without any genomic resources available. It is likely that in the future, most population genetic studies will be based on this type of cheap large-scale sequencing instead of microsatellite, mtDNA, restriction fragment length polymorphism, amplified fragment length polymorphism or allozyme genotyping. Many of the techniques currently available for population genetic analysis do not scale up to the analysis of this type of data. It will be one of the important challenges in the field of molecular ecology over the next 5–10 years to develop new methods of analysis for this type of data.

Conclusion

The field of molecular ecology has benefited greatly by the tree-thinking introduced by the field of phylogeography. However, qualitative interpretation of gene trees typically rely on a series of assumptions that may not always be justified. To move the field forward, it is necessary to relate trees to specific models and to take uncertainty in the estimation of the trees into account. This is the objective of many of the current efforts in statistical population genetics. Arguably, the field of phylogeography has developed tremendously over the past years, allowing more rigorous inference methods to replace simpler approaches based solely on tree estimation and informal interpretation. However, some serious challenges remain in developing techniques that are computationally faster and allow for more realistic demographic models. Also, the emerging availability of large-scale DNA sequencing data sets produced by next-generation sequencing technologies pose new problems in statistical population genetics.

References

Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single nucleotide polymorphisms. *Genetics*, **168**, 1699–1712.

Anderson EC (2005) An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. *Genetics*, **170**, 955–967.

Avise JC (1989) Gene trees and organismal histories — a phylogenetic approach to population biology. *Evolution*, **43**, 1192–1208.

Avise 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.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Balding DJ, Nichols RA (1997) Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, **78**, 583–589.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251–261.

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

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.

Bentley DR (2006) Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, **16**, 545–552.

Cassens I, Mardulyn P, Milinkovitch MC (2005) Evaluating intraspecific 'Network' construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Systematic Biology*, **54**, 363–372.

Chan YL, Anderson CNK, Hadly EA (2006) Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *Plos Genetics*, **2**, 451–460.

Chikhi L, Beaumont MA (2005) Modelling human genetic history. In: *The Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley, New York.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, **158**, 1347–1362.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1659.

Cornuet JM, Beaumont MA (2007) A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theoretical Population Biology*, **71**, 12–19.

De Iorio M, Griffiths RC, Leblois R, Rousset F (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology*, **68**, 41–53.

Durrett R (2008) *Probability Models for DNA Sequence Evolution*. Springer, New York.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM (2004) Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution*, **58**, 2021–2036.

Ewens W (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

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

Excoffier L, Heckel G (2006) Computer programs for population genetic data analysis: a survival guide. *Nature Reviews Genetics*, **7**, 745–758.

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.

Fearnhead P (2003) Consistency of estimators of the population-scaled recombination rate. *Theoretical Population Biology*, **64**, 67–79.

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Fedorov VB, Goropashnaya AV, Boeskorov GG, Cook JA (2008) Comparative phylogeography and demographic history of the wood lemming (*Myopus schisticolor*): implications for Late Quaternary history of the taiga species in Eurasia. *Molecular Ecology*, **17**, 598–610.

Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.

Felsenstein J (1988) Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, **19**, 445–471.

Felsenstein J (1992) Estimating effective population-size from samples of sequences — inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, **59**, 139–147.

Fichten CS, Sunerton B (1983) Popular horos copes and the 'barnum effect'. *The Journal of Psychology*, **14**, 123–134.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

Forer BR (1949) The fallacy of personal validation — a classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, **44**, 118–123.

Goldstein DB, Chiki L (2002) Human migrations and population structure: what we know and why it matters. *Annals of Review of Genom and Human Genetics*, **3**, 129–152.

Griffiths RC, Tavare S (1994a) Simulating probability-distributions in the coalescent. *Theoretical Population Biology*, **46**, 131–159.

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

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

Harvey PH, Pagel DM (1991). The comparative method in evolutionary biology, Oxford University Press, Oxford, UK.

Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. *Public Library of Science, Biology*, **3**, 965–975.

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 (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

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

Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

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

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.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

Li N, Stephens M (2003) Modeling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, **165**, 2213–2233.

Margulies M, Egholm M, Altman WE *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

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

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**, 351–372.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Miller N, Estoup A, Toepfer S *et al*. (2005) Multiple transatlantic introductions of the western corn rootworm. *Science*, **310**, 992–992.

Nath HB, Griffiths RC (1996) Estimation in an island model using simulation. *Theoretical Population Biology*, **50**, 227–253.

Neigel JE, Avise JC (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: *Speciation and its Consequences* (eds Otte D, Endler JA), pp. 28–59. Sinauer Associates, Sunderland, Massachusetts.

Nielsen R (1997) A likelihood approach to populations samples of microsatellite alleles (Vol. 146, p. 711, 1997). *Genetics* **147**, 349–349.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M (1998) Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*, **52**, 669–677.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Pagel M (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 331–348.

Pamilo P, Nei M (1988) Relationships 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.

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

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, **165**, 427–436.

Posada D, Crandall KA, Templeton AR (2000) GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Molecular Ecology*, **9**, 487–488.

Posada D, Crandall KA, Templeton AR (2006) Nested clade analysis statistics. *Molecular Ecology Notes*, **6**, 590–593.

Posada D, Maxwell TJ, Templeton AR (2005) TreeScan: a bioinformatic application to search for genotype/phenotype associations using haplotype trees. *Bioinformatics*, **21**, 2130–2132.

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

Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **104**, 1760–1765.

Slatkin M (1987) The average number of sites separating dna-sequences drawn from a subdivided population. *Theoretical Population Biology*, **32**, 42–49.

Stephens M (2007) Inference under the coalescent. In: *Handbook of Statistical Genetics* (eds Balding DJ, Bishop M, Cannings C), pp. 878–908. Wiley, Chichester, UK.

Stephens M, Donnelly P (2000) Inference in molecular population genetics. *Journal of the Royal Statistical Society B: Statistical Methodology*, **62**, 605–655.

Sunnucks P, Blacket MJ, Taylor JM *et al*. (2006) A tale of two flatties: different responses of two terrestrial flatworms to past environmental climatic fluctuations at Tallaganda in montane south-eastern Australia. *Molecular Ecology*, **15**, 4513–4531.

Tallmon DA, Luikart G, Beaumont MA (2004) Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics*, **167**, 977–988.

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

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

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

Templeton 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.

Templeton AR, Maxwell T, Posada D *et al*. (2005) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, **169**, 441–453.

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. 4. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Templeton AR, Sing CF, Kessling A, Humphries S (1988) A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. *2. The Analysis of Natural Populations Genetics*, **120**, 1145–1154.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–1507.

Wakeley J (1999) Nonequilibrium migration in human history. *Genetics*, **153**, 1863–1871.

Wakeley J (2004) Recent trends in population genetics: more data! More math! Simple models? *Journal of Heredity*, **95**, 397–405.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Wang JL (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.

Wang JL, Whitlock MC (2003) Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, **163**, 429–446.

Weiss G, von Haeseler A, Paabo S (1997) The myth of bumpy hunter–gatherer mismatch distributions — reply. *American Journal of Human Genetics*, **61**, 983–983.

Westfall PH, Young SS (1993) On adjusting *P*-values for multiplicity. *Biometrics*, **49**, 941–944.

Williamson SH, Hernandez R, Fledel-Alon A *et al*. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences, USA*, **102**, 7882–7887.

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

Wiuf C (2006) Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, **53**, 821–841.

Won YJ, Hey J (2005) Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, **22**, 297–307.

Wooding S, Rogers A (2002) The matrix coalescent and an application to human single nucleotide polymorphisms. *Genetics*, **161**, 1641–1650.

R. Nielsen works on statistical methods in population genetics, medical genetics, and molecular evolution. M. A. Beaumont works on statistical methods and applications in population genetics and molecular ecology.
