Revealing the demographic histories of species using DNA sequences

Brent C. Emerson, Emmanuel Paradis and Christophe Thébaud

Various methodological approaches using molecular sequence data have been developed and applied across several fields, including phylogeography, conservation biology, virology and human evolution. The aim of these approaches is to obtain predictive estimates of population history from DNA sequence data that can then be used for hypothesis testing with empirical data. This recent work provides opportunities to evaluate hypotheses of constant population size through time, of population growth or decline, of the rate of growth or decline, and of migration and growth in subdivided populations. At the core of many of these approaches is the extraction of information from the structure of phylogenetic trees to infer the demographic history of a population, and underlying nearly all methods is coalescent theory. With the increasing availability of DNA sequence data, it is important to review the different ways in which information can be extracted from DNA sequence data to estimate demographic parameters.

Brent C. Emerson*
School of Biological Sciences, University of East Anglia, Norwich, UK NR4 7TJ.
*e-mail: b.emerson@uea.ac.uk

Emmanuel Paradis
Institut des Sciences de l'Evolution, UMR 5554 CNRS, Case courrier 64, Université Montpellier II, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France.

Christophe Thébaud
Centre d'Ecologie de Toulouse, UMR 5552 CNRS, Université Paul Sabatier, 13 Avenue du Colonel Roche, BP 4072, F-31029 Toulouse Cedex 4, France.

Over the past two decades, dramatic progress has been made in our ability to obtain DNA sequence data. With the advent of automated sequencing technology and an ever-increasing database of comparative sequence data, evolutionary geneticists are probably now faced with more obstacles to obtaining biological material, than to extracting DNA sequence data from such material once obtained. The relative ease with which we are now able to obtain DNA sequence data has produced a concomitant shift from typically higher level phylogenetic studies of taxa to studies that also address within-species variability, most notably within the field of phylogeography[1]. Similarly, population genetics has also undergone a reformation of sorts, in particular with the application of coalescent theory (Box 1) to sequence data to estimate population parameters, such as EFFECTIVE POPULATION SIZE (see Glossary). These recent advances within the fields of phylogenetics and population genetics are providing biologists with new analytical tools to investigate the demographic histories of populations (Box 2).

A fundamental result of coalescent theory in population genetics is the finding of a relationship between COALESCENT TIME and population size. For any two sequences drawn from a population, the probability that they coalesce at a given point in history is a function of population size[2]. Two DNA sequences drawn at random from a small population have a higher probability of having a more recent coalescence (i.e. fewer substitutional differences separating them) than do two DNA sequences drawn at random from a large population. Thus, a change in population size over time will leave a signature in the pattern of DNA substitutions among individuals within a population that will depend on the direction (growth or decline) and tempo (ancient or recent) of this change (Fig. 1).

Summary statistics and pairwise difference distributions

Early studies focused on how changes in population size affect the distribution of PAIRWISE DIFFERENCES between the DNA sequences of individuals, and the number of SEGREGATING SITES within a population[3–7]. For a sudden population expansion, theoretical expectations of a Poisson distribution of pairwise nucleotide differences have been verified by simulation and empirical DNA sequence data[6]. Such a relationship allows the assessment of the sequence and approximate timing of population expansion events, such as using mitochondrial DNA (mtDNA) sequence data[8] to map the expansion of modern humans across Europe. Recently, simulation analysis has been used to discriminate between different growth patterns from distributions of pairwise differences[9]. With extremely large samples, distinguishing between patterns of stepwise, exponential and logistic population growth seems possible and, when applied to human mtDNA data, a growth pattern not dissimilar to logistic population growth is suggested[9]. Summary statistics from the distribution of pairwise differences and simulation analyses have been used to compare the rate of spread of different subtypes of human immunodeficiency virus type 1 (HIV-1)[10].

Although the utility of pairwise difference distributions for revealing population growth has been demonstrated, the method fails to characterize clearly a history of constant population size[6,7]. This shortcoming has been rightly attributed to the signal in the data being swamped by nonindependence[11], because methods that seek to estimate population parameters from the distribution of pairwise differences, or the number of segregating sites, do not make full use of the data. By incorporating information from the genealogical tree structure of DNA sequences, the problem of nonindependence can be circumvented[11]. Recently, several conceptual approaches, which make use of the genealogical tree structure, but each with its own set of assumptions (Box 3), have been developed to obtain better estimates of the demographic histories of populations.

http://tree.trends.com 0169-5347/01/$ – see front matter PII: S0169-5347(01)02305-9

Box 1. Coalescent theory

Since the mid-1980s, population geneticists have been studying the effects of genetic drift and mutation using the model called 'the coalescent'[a,b]. The roots of coalescent theory can be traced back to the work of Ronald Fisher[c] and Sewall Wright[d] but it was the development of the theory of neutral evolution of nucleotide sequences[e] that led to fuller development of coalescent theory.

The descent of any set of individuals (DNA sequences) can be traced back, with common ancestry denoting coalescent events. Under the assumption of neutral evolution, the probability in any generation that a DNA lineage will give rise to two distinct daughter lineages is the same for all lineages. Thus, if two DNA sequences are sampled from a population, the probability (h) that they share a common ancestor (i.e. that they coalesce) in the previous generation is the same for any two sequences chosen.

Consider the following simple example (Fig. I). A population consists of ten individuals ('a' to 'j') each with a different DNA sequence. What is the probability (h) that 'd' and 'f' coalesce (i.e. share a common ancestor) in the previous generation, under the assumptions of: (1) constant population size; (2) random mating; and (3) discrete nonoverlapping generations? There were ten ancestors in the previous generation, and only one ('D') can be the ancestor of 'd'. The question is, what is the probability that 'D' is also the ancestor of 'f'?

Each individual in the previous generation (ten genes) is equally likely to be the parent of 'f', so $h = 1/10$. Thus in any population satisfying the above assumptions, the coalescence of any two lineages in the preceding generation will occur with a probability equal to the reciprocal of the number of genes (G) in the population consisting of n individuals, where $G = n/2$ for mitochondrial genes (because there is one maternally inherited gene per individual) and $G = 2n$ for autosomal genes (because there are two biparentally inherited genes per individual).

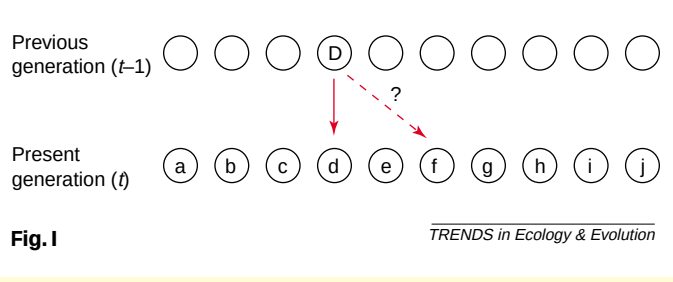

Fig. I

Now consider the case where the ten sequences are a sample from a population of 4000 sequences. The number of possible pairs of genes is $10(10-1)/2$, but this will decrease with each coalescent event such that when there are i genes in the population the number of possible pairs of genes is $i(i-1)/2$. From this, and the probability of coalescence ($h = 1/4000$) for a given pair of genes, it is possible to calculate the probability of a coalescent event in the ith generation containing G_i genes (Eqn I):

$$p = \frac{1}{G} \times \frac{i(i-1)}{2} = \frac{i(i-1)}{2G} \qquad [I]$$

The expected length of the coalescent interval can be expressed as (Eqn II):

$$length = \frac{2G}{i(i-1)} \qquad [II]$$

In the hypothetical example of ten sequences, under the assumptions of constant population size, random mating, and discrete nonoverlapping generations, it is possible to predict the lengths of the nine coalescent intervals:

First interval ($I = 10$) $= (2 \times 4000) / (10 \times 9) = 89$

Second interval ($i = 9$) $= (2 \times 4000) / (9 \times 8) = 111$

…

Eighth interval ($i = 3$) $= (2 \times 4000) / (3 \times 2) = 1333$

Ninth interval ($i = 2$) $= (2 \times 4000) / (2 \times 1) = 4000$

Thus, from coalescent theory, we can test the hypothesis that a population has been of constant size through time, using a sample of DNA sequences.

References

a Kingman, J.F.C. (1982) The coalescent. *Stoch. Process. Appl.* 13, 235–248
b Kingman, J.F.C. (1982) On the genealogy of large populations. *J. Appl. Prob.* 19A, 27–43
c Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*, Clarendon
d Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16, 139–156
e Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press

Coalescent-based graphical methods

Lineage through time plots

Coalescent theory studies have shown that homologous DNA sequences sampled from a population contain information about the demographic history of that population. When these sequences are summarized as a genealogical tree, the information, which is now contained within the relative positions of the internal nodes within the genealogy (Fig. 2), can be exploited further. Because each internal node corresponds by definition to a coalescence event, historical changes in population size can be inferred from the temporal distribution of nodes in a DNA sequence genealogy[12]. If the cumulative distribution of coalescence events is plotted against time, the resultant curve shape, or lineage through time (LTT) plot, can be used to test specific hypotheses about the population history.

These hypotheses, derived from theoretical expectation from coalescent theory and simulation studies, are summarized as LTT plots with various transformations of either the lineage or time axis. Choosing between several competing hypotheses (represented as distinct LTT curves) is achieved by statistical testing of the fit of the different theoretical LTT curves to the observed LTT plot. This methodology has recently been used to analyse the population dynamic histories of coconut crabs *Birgus latro*[13], HIV-1 (Ref. 14), the hepatitis C virus[14,15] and human papillomavirus[15,16]. Existing software (Box 2) requires an ULTRAMETRIC TREE as

Box 2. Software availability

- END-EPI performs data transformations and constructs lineage through time (LTT) plots. Data input is limited to ultrametric tree topologies. A Macintosh executable is available at (http://evolve.zoo.ox.ac.uk/software/End-Epi/End-Epi.html). It was written by A. Rambaut *et al.* (Dept of Zoology, University of Oxford, UK).
- GENIE is a C++ program that constructs skyline plots and performs likelihood analysis of demographic models. Data input is limited to ultrametric tree topologies. A Macintosh PPC executable is available, as well as the code for compiling on Unix/Linux, at (http://evolve.zoo.ox.ac.uk/software/Genie/main.html). It was written by O. Pybus and A. Rambaut (Dept of Zoology, University of Oxford, UK).
- DIVERSI is a Fortran program that performs likelihood analysis of survival models. Incomplete and uncertain phylogenetic data can be analysed, and the statistical power of the tests can be assessed. A Windows executable is available, as well as the code for compiling on Unix/Linux, at (http://www.isem.univ-montp2.fr/~paradis/index.html). It was written by E. Paradis (Institut des Sciences de l'Evolution, CNRS, France).
- LAMARC is a package of four C programs that implement the coalescent theory to estimate population parameters (including θ) with a Metropolis–Hastings Monte Carlo sampling algorithm. COALESCE estimates the effective population size of a single constant population using nonrecombining sequences. FLUCTUATE estimates the effective population size and growth rate of a single exponentially growing or declining population using nonrecombining sequences. MIGRATE estimates the effective population sizes and migration rates of *n* constant populations using nonrecombining sequences, microsatellite data or enzyme electrophoretic data. RECOMBINE estimates the effective population size and per site recombination rate of a single constant size population. Executables for Windows, Macintosh PPC, and some Unix systems are available, as well as the code for compiling, at (http://evolution.genetics.washington.edu/lamarc.html). These programs were written by P. Beerli and J. Felsenstein (MIGRATE) M. Kuhner, J. Yamato, and J. Felsenstein (COALESCE, FLUCTUATE and RECOMBINE) (Dept of Genetics, University of Washington, USA).
- TCS is a Java program that implements a statistical parsimony algorithm for constructing haplotype networks for DNA sequence data. Both Macintosh and Windows executables are available (http://bioag.byu.edu/zoology/crandall_lab/tcs.htm). It was written by M. Clement and D. Posada (Dept of Zoology, Brigham Young University, USA).
- GEODIS is a Java program that implements statistical tests of hypotheses of population structure and population history. Both Macintosh and Windows executables are available (http://bioag.byu.edu/zoology/crandall_lab/geodis.htm). It was written by D. Posada (Dept of Zoology, Brigham Young University, USA) and A. Templeton (Dept of Biology, University of Washington, USA).
- EVE is a program that simultaneously estimates $\theta = 2N\mu$ and the population growth rate using a distance matrix in PAUP* format obtained from nucleotide sequence data. It is available for Unix and Macintosh PCC at (http://bioag.byu.edu/zoology/crandall_lab/Vasco/eve.htm). It was written by D. Vasco (Dept of Zoology, Brigham Young University, USA).
- BATWING is a program implementing Metropolis-Hastings algorithms to investigate models of constant population size, population growth, population subdivision, and population subdivision with growth. It is available for Unix, Windows 95/NT and Macintosh (OS 8.0 and above) at (http://www.maths.abdn.ac.uk/~ijw/). It was written by I. Wilson (Dept of Mathematical Sciences, University of Aberdeen, UK), M. Weale (University College, London, UK) and D. Balding (University of Reading, UK).
- GENETREE is a program using Markov chain simulation to generate likelihood surfaces for θ (genetic diversity), mutation rate, migration and population growth with or without population subdivision. A Windows executable is available, as well as the code for compiling on Unix/Linux, at (http://ftp.monash.edu.au/pub/gtree). It was written by M. Bahlo and R. Griffiths (Mathematics Dept, Monash University, Australia).

input data. As the raw data for analysis are essentially the relative times of the coalescence events within a tree topology, new methods[17–19] can be used to estimate the ages of coalescence events from a nonultrametric tree. The scaling of the axes can then be applied to these nodal ages, circumventing the restrictions imposed by an ultrametric tree. Care must be taken when the assumption of a molecular clock is not met but is enforced during tree reconstruction. Resulting genealogies are susceptible to misleading demographic inferences being made[20].

Skyline plots

A more recent approach, also using graphical representations of demographic information, is the use of skyline plots[21]. These display estimates of effective population size (N_e) along a time axis and are therefore easier to interpret than are the LTT plots, which require the axes to be transformed[12]. Information about the size of internode intervals (measured in mutational events corresponding to time periods) and their sequential order from a reconstructed genealogy allows estimation of the HARMONIC MEAN of the effective population size for each internode interval (denoted M_i)[21]. Plotting values of M_i against time gives a graphical representation of population demographic history. Specific demographic models can then be fitted to the data, and the corresponding parameter(s) estimated by maximum likelihood. When applied to HIV-1 data, this approach indicates a constant-rate exponential increase in the population size of subtype A and logistic growth for

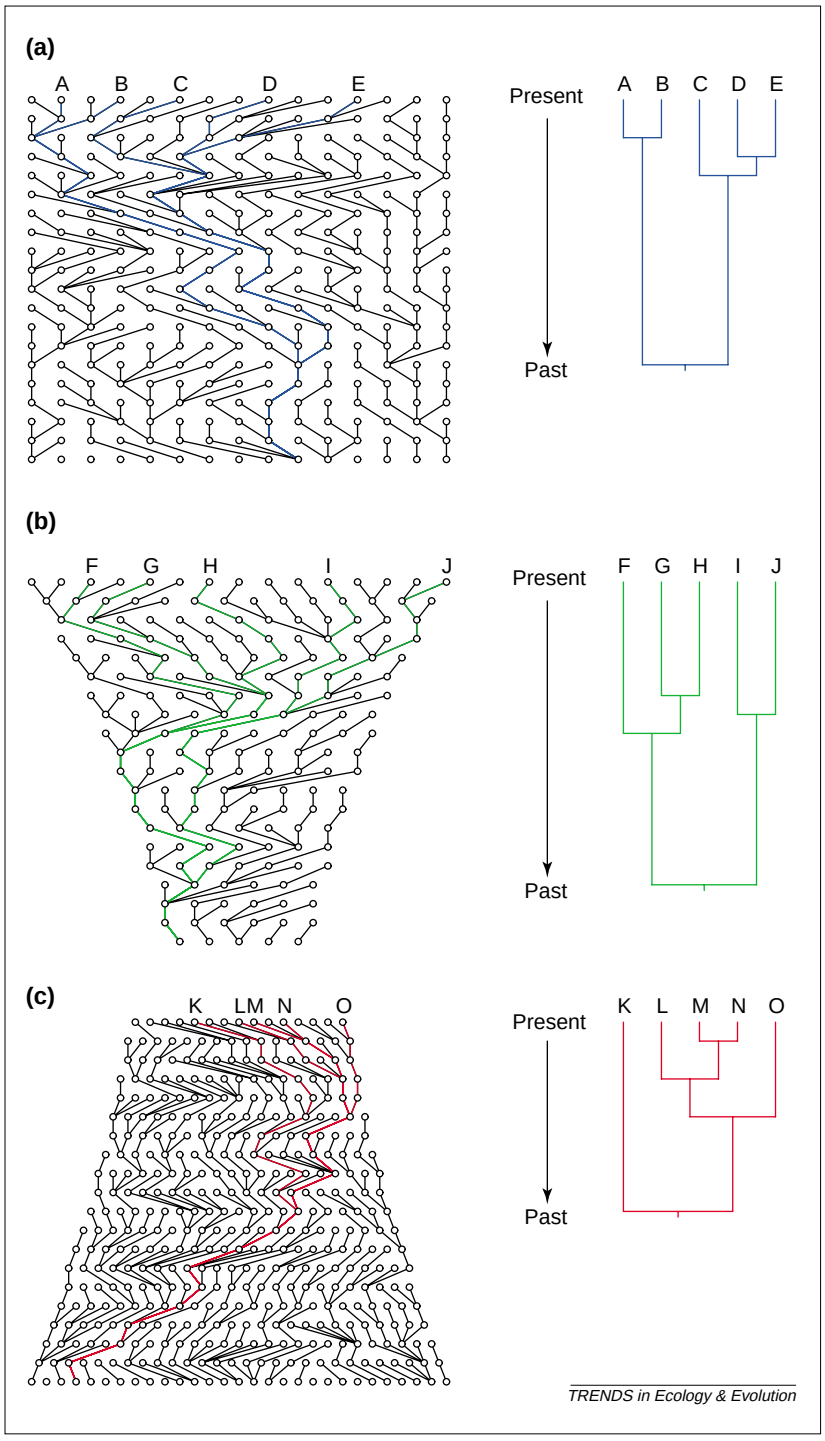

Fig. 1. Hypothetical genealogies and corresponding phylogenetic trees from (a) a population of constant size, (b) an exponentially growing population, and (c) an exponentially declining population. In all cases, current populations are of equal size ($n = 15$) and five sequences (A–E, F–J, and K–O) were sampled from each population. Differences in tree topologies are due only to the effects of sampling sequences from populations with different dynamics. The coalescent provides a robust theoretical framework that relates the expected time between coalescent intervals (node heights in the tree) and past population dynamics. Thus, using coalescent-based methods, it is possible to infer historical changes in population size from samples of gene sequences obtained from extant populations.

subtype B (although exponential growth cannot be ruled out). This suggests that, for HIV-1, skyline plots have greater discriminatory power than do LTT plots, which indicate only that both subtypes have increased exponentially[22].

Both LTT and skyline plots are appropriate for exploratory analyses of population change because they do not make assumptions about the form of this change. Thus, they help in the subsequent choice of a demographic model for parameter estimation; however, they do not provide formal tests of hypotheses or model selection procedures.

Survival models

New work has approached the problem of estimating demographic history from gene sequence data using statistical models that were originally designed for the analysis of survival data[23–25]. The data used are divergence times among a group of sequences as estimated from a phylogenetic tree. The number of lineages within a reconstructed phylogeny increases with time but, if the time axis is reversed, it provides a representation of survival data with one lineage dying at each dichotomous node. The ages of the divergences measured on the phylogeny from present to past can thus be considered as failure times. Some failure times are often not precisely known in survival data, and this is called CENSORING[26]. Consequently, branching times that are not exactly known can be treated as censored or interval-censored data[23,24]. This method assumes that, for each lineage, there is an instantaneous diversification rate $\delta(t)$, which is estimated by maximum likelihood analysis and has two components, an instantaneous birth rate $\sigma(t)$ and an instantaneous death rate $\varepsilon(t)$. It is assumed that $\sigma(t)$ and $\varepsilon(t)$ cannot be estimated separately, which can be justified because the information about death events can only be found in the extinct lineages[24,25]. An advantage of this survival method is that it allows for MULTIFURCATIONS in tree topology.

In survival analyses, the change (or constancy) through time of $\delta(t)$ is specified by a simple curve, called a hazard function. LIKELIHOOD RATIO TESTS (LRT) and AKAIKE INFORMATION CRITERIA (AIC) are then calculated to compare models with different assumptions with respect to $\delta(t)$, and thus test hypotheses on the tempo of growth in lineage number. Specifically defined models of diversification can be constructed for analysing differences in demographic histories between clades (or groups of lineages). With several clades under analysis, models can range from the null model, where all clades have the same δ, to a model where all clades have a different δ. Although developed primarily to examine phylogenies of recent species, the use of survival models is equally suited to intraspecific gene genealogies, and has been used to compare population demographic histories of island populations of beetle species on the Canary Islands[27,28].

Box 3. Assumptions of methods for estimating demographic history

Different methods for estimating demographic parameters each involve a particular set of assumptions (Table I).

Table I. Methods for demographic inference and their associated assumptions

Method	Assumptions						
	Infinite sites model	Genealogy known	No recombination	Molecular clock	No population structure or subdivision	Constant population size	Neutral evolution
Summary statistics	×		×	×		×	
Graphical methods	×		×	×		×	
Survival models	×		×	×		×	
Likelihood methods		×[a]	×	×[a]	×[a]	×	
Mid-depth method		×	×	×	×		×
Least squares estimators	×	×		×	×		×
Nested cladistic analysis	×						×

[a]One of these assumptions can be relaxed to estimate the relevant parameters.

Coalescent-based demographic parameter estimation

The integration of coalescent theory into a statistical framework has led to the growing development of coalescent-based methods that analyse genetic diversity among a sample of DNA sequences to infer population demographic history. The fundamental relationship exploited by these coalescent-based methods is between the distribution of divergence times among individuals and effective population size. This relationship is embodied by the genetic diversity parameter $\theta = 4 N_e \mu$, where N_e is the effective population size ($\theta = 2 N_e \mu$ for haploids) and μ is the mutation rate. Early developments in the estimation of θ were based on summary statistics, such as the number of segregating sites among DNA sequences[3] or the mean number of nucleotide

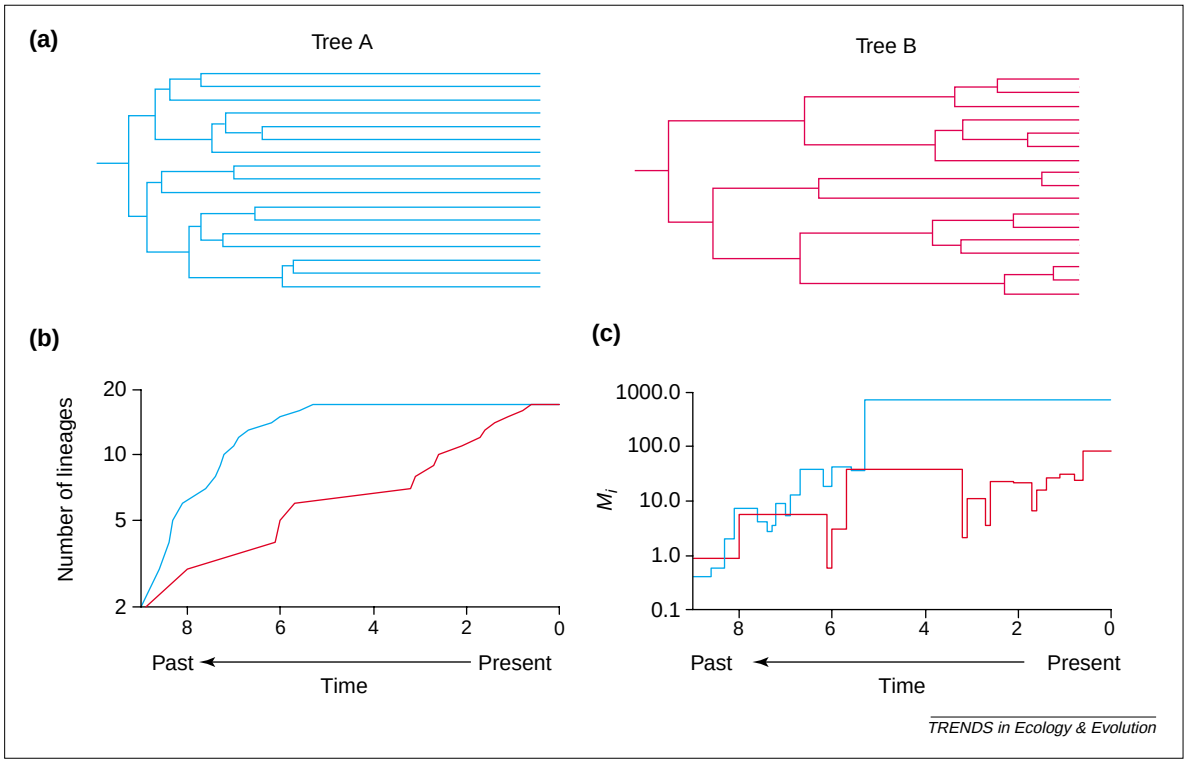

Fig. 2. Lineage through time (LTT) plots and skyline plots are two graphical methods that aim to explore patterns of population histories to eventually select an appropriate parametric model for population parameter estimation. (a) Two phylogenetic trees with the same number of tips and the same distance from root to tips. The branch lengths differ between both trees and this might reveal different population histories. (b) The LTT plot shows, on a logarithmic scale, a curvilinear increase in the number of lineages through time for tree A, indicative of a population increase, whereas the roughly linear increase of the same number for tree B suggests a constant population through time. (c) The skyline plots suggest a more subtle picture. The population numbers corresponding to tree A seem to have increased at a constant exponential rate (thus the increase appears linear on the logarithmic scale), and stabilized after some time. Conversely, for tree B, population numbers seem to have remained at a fairly stable level, perhaps slightly increasing.

differences in all pairwise comparisons[4]. However, this approach is limited because it does not take into account the genealogical structure of the data[11]. Recent developments have tried to remedy this using: (1) maximum likelihood estimates incorporating METROPOLIS–HASTINGS SAMPLING and integration across several phylogenies[29]; (2) calculating likelihood surfaces for a tree by recursive evaluation of trees with progressively fewer coalescence events[30]; (3) maximum likelihood estimation from a reconstructed tree that is assumed to be correct[20,21]; and (4) information from a tree topology and correcting the estimate using simulation[31].

Likelihood estimation by Metropolis–Hastings sampling

Rather than using a single estimate of the genealogy for a population sample of DNA sequences, Metropolis–Hastings sampling allows for genealogical uncertainty when estimating θ (Ref. 29). This is done by not only analysing the phylogenetic tree that best describes the DNA sequence data, but also a sample of those trees that are less probable. However, compared with other methods for inferring population demographic history, this approach is computationally demanding, because algorithms are based on MONTE CARLO resampling. An algorithm for the estimation of population growth rate under the condition that growth (or decline) is exponential has been implemented within this framework[32]. These algorithms have been used in studies of mosquitoes *Anopheles dirus*[33], a skink *Chalcides viridus*[34] and two bird species, the California gnatcatcher *Polioptila californica* and the yellow warbler *Dendroica petechia*[35,36], to estimate the rate of growth within populations of these taxa. Comparisons of genealogical[29] and nongenealogical[4] estimates of θ have also provided insights into recent population histories of the grey wolf *Canis lupus* and coyote *C. latrans*[37], indicating that both were much more numerous in the recent past. Comparisons with contemporary census data can delimit population changes that are too recent to be recorded within the diversity of DNA sequences[37]. Demographic analysis of DNA sequence data from coconut crab populations[13] also provides an example of the different outcomes that can occur between genetic and population censusing[38].

Likelihood estimation by recurrence equations

An alternative method also calculates likelihood curves for the parameter θ using Monte Carlo integration, but with a different method to the Metropolis–Hastings algorithm. Recurrence equations are used to express the probabilities of mutational and coalescent events for a given value of θ. These recurrence equations are used to construct a MARKOV CHAIN: this provides a simple Monte Carlo sampling method for approximating complex sampling probabilities[39]. The likelihood function for a given value of θ is computed by independently, repeatedly, simulating the Markov chain and taking the mean of the simulated values. Likelihood curves are then constructed for a range of θ values. Recently, this computational technique has been extended to subdivided populations for the estimation of migration rate, detection of population growth, determining in which population the most recent common ancestor of all the sequences occurred, and determining the age and location of subpopulation ancestors[30]. The method performs well on relatively monomorphic data that approximate an INFINITE SITES MUTATIONAL MODEL of evolution (less homoplasy and back mutation), but deviation from this model will probably result in under-performance with regard to the Metropolis–Hastings sampling method. A more detailed comparison of the two methods is given in Ref. 40.

Mid-depth method

The mid-depth method[20] has been used to test the hypothesis of constant population size from a reconstructed phylogeny. A tree statistic σ (the number of coalescence events between the root and the mid-depth point of a genealogy) is used to test this hypothesis. This is done using an expectation from coalescent theory that, for sample sizes up to 200 sequences, if $\sigma > 3$, the hypothesis of constant population size can be rejected with a 5% risk. If the hypothesis is rejected, and LTT plots indicate constant exponential growth, the method can then be used for the estimation of exponential growth rate (denoted r). The parameter $\alpha = N_0 r$ (where N_0 is the present population size) can be estimated using the tree statistic σ as a result of the approximately linear relationship between σ and $\log(\alpha)$ for a wide range of α values. Monte Carlo simulations are used to generate a large number of genealogies with a given value of α. For each of these, the probability that $\sigma(\text{simulated}) = \sigma(\text{observed})$ is tested with LRTs. Higher estimates of α and its confidence limits for human HIV-1 subtype B, compared with those for subtype A, have been interpreted as indicating higher exponential growth rate for subtype B (Ref. 20).

Other approaches also make use of maximum likelihood analysis to infer population history by estimating model parameters. Likelihood estimates of exponential growth rate (λ) and current effective population size for HIV-1 subtypes indicate that subtype B is either spreading faster than subtype A, has a smaller current effective population size, or there is a combination of these two effects[10]. Genetic diversity (θ) and the ratio (ρ) of the current to the initial population size have been estimated using a likelihood approach to refine these conclusions for HIV subtypes[21]. Recent work applying likelihood estimates of ρ, θ and the time when a population

Box 4. Nested clade analysis

Evidence for the geographical structuring of haplotypes can arise as a result of contemporary factors, such as restricted gene flow, or historical population events, such as past population fragmentation, range expansions or colonization. Nested clade analysis[a] is a method to distinguish which of these four factors (or combination of them) offers the best explanation when nonrandom geographical association is found. This is achieved by framing each of the four factors as hypotheses that have been developed from several studies using coalescent theory and computer simulation. For example, one such study has modelled the dispersal of mitochondrial DNA (mtDNA) lineages as a random walk process[b]. Because mtDNA is usually maternally inherited, it provides an indication of dispersal from the maternal birth site in one generation to the birth site of the next.

An essential first step is the construction of a network of haplotypes using a methodology that accounts for population level phenomena[c]. A nesting arrangement is subsequently derived. Figure I illustrates such a network with a nested clade arrangement for 14 chromosome haplotypes (labeled a–n) sampled from 2198 males from 60 populations worldwide (reproduced, with permission, from Ref. d). The 14 individual haplotypes comprise zero-step clades. One-step clades are then created by uniting zero-step clades (e.g. the two zero-step clades a and b unite to form the one-step clade 1-1). These six one-step clades are then nested at a higher level into three two-step clades. The third and final nesting level comprises the three two-step clades. This nesting arrangement results in a total of ten hierarchically arranged clades, with higher level nesting correlating with more distant evolutionary time.

The four hypotheses have different expectations regarding the relationship between the genealogical and geographical distances between haplotypes. Because the nested arrangement of the network corresponds to evolutionary time (genealogical distance) it is possible to test the fit of the data to each of the hypotheses. This is achieved using two types of calculated geographical distances: (1) clade distances, which measure how geographically widespread the haplotypes within a clade are (e.g haplotypes i, j and k, in clade 1-5) and (2) nested clade distances, which measure how far the haplotypes of one clade are from the haplotypes of the sister clades in the higher nesting level (e.g. clades 1-5 and 1-6 within 2-3). Statistical comparison of clade and nested clade distances for tip (e.g. 1-6) and interior (e.g. 1-5) subclades inside a given tested clade (e.g. 2-3) is performed to search for patterns characteristic of the four hypotheses[a].

Nine of the ten nesting clades for the 14 Y chromosome haplotypes in Fig. I are significant for the nonrandom geographical association of haplotypes. The analysis of structure within each of the ten nested clades means that it is possible to have a combination of population structure and population history factors offering the greatest explanatory power across the network. Within the human Y chromosome data, there are three episodes of restricted gene flow, one instance of long distance dispersal, six instances of range expansion, but no instances of allopatric fragmentation.

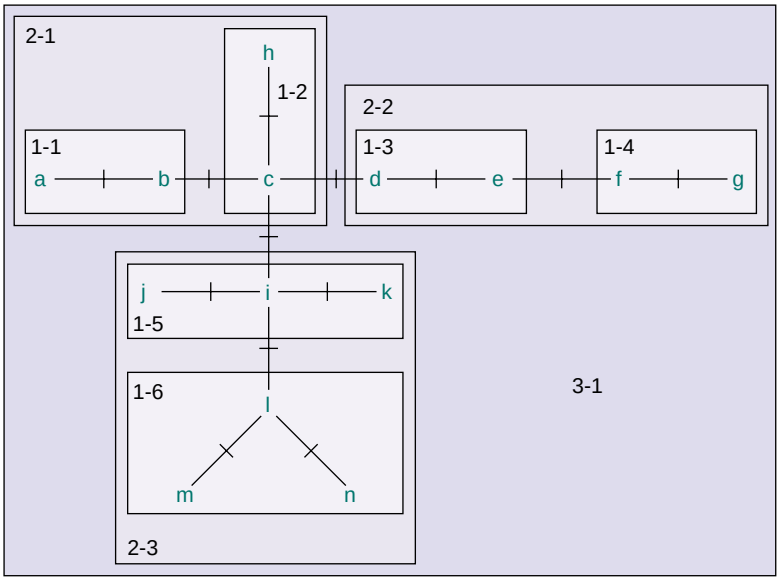

Fig. I TRENDS in Ecology & Evolution

References

a Templeton, A.R. (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* 7, 381–397

b Neigel, J.E. *et al.* (1991) Estimation of single generation migration distances from geographical variation in animal mitochondrial DNA. *Evolution* 45, 423–432

c Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45

d Karafet, T.M. *et al.* (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* 64, 817–831

starts to change (τ), to sequences from the hypervariable region I of mtDNA of humans from different geographical areas, suggest that the Basque people have expanded, Biaka pygmies have decreased, and the native American Nuu-Chah-Nulth have probably remained constant in population size[41].

Least squares estimators

Although the statistical properties of maximum likelihood estimators make them desirable, they are also computationally intensive. As an alternative, computationally less intensive LEAST SQUARES estimators, such as the best linear unbiased estimator[31], have been applied to various population parameters[42]. For a given tree topology, θ is estimated from the partition of the number of mutations by the branch on which the mutations occurred. This is done using a particular least squares algorithm (recursive least squares) that takes into account the dependence structure of the data resulting from the genealogy. These estimates

Glossary

Akaike information criterion (AIC): a likelihood-based selection criterion that can compare any hypotheses (nested or otherwise): the hypothesis with the smallest AIC value is considered to be the most probable. The AIC is a tradeoff between the fit of the model to the data (measured by its likelihood value) and the number of parameters in the model (because models with more parameters will fit the data better).

Censoring: this occurs when the true value of a variable is unknown. When the true value of the variable is known to be greater than an observed value, the observation is said to be censored. When the true value of the variable is known to be within an observed interval, the observation is said to be interval-censored. This is typical of lifetime data but can be applied to other kinds of observations (e.g. geographical distances).

Coalescent time: the time at which two alleles share their most recent common ancestor. Any two alleles can be followed down the phylogenetic tree (towards the root) to the point at which the two genetic lineages coalesce (i.e. they unite at their most recent common ancestor). This union of the two alleles is called a coalescent event.

Effective population size: the effective size of a population is the size of the pool from which genes could be drawn at random to construct the next generation resulting in the same rate of genetic drift (i.e. loss of genetic diversity) as the actual population under consideration. This can be expressed in three different ways, leading to three concepts of effective population size: the probability of homozygosity owing to common ancestry (inbreeding effective size), the expected variance of gene frequencies at the next generation (variance effective size), or the rate of decay of segregating loci (eigenvalue effective size).

Harmonic mean: the reciprocal of the arithmetic mean of the reciprocals. The harmonic mean h of a sample x^1, x^2,..., x^n is given by: $1/h$ = the sum (from $i = 1$ to $i = n$) of the values $1/x^i$. Because it is dominated by smaller terms, it provides the best means for summarizing fluctuations in population size when events characterized by a small value (e.g. bottlenecks) are of great biological significance.

Infinite sites mutational model: under this model, sequences are said to evolve such that any new mutational event will occur at a nucleotide site that has not previously experienced one.

Least squares: least squares methods fit a model to data by minimizing the squared differences between the observed values and those predicted by the model.

Likelihood ratio test (LRT): a likelihood value measures the support for a hypothesis (in this case a demographic model) given by the available data (in this case a phylogenetic tree). Likelihood values for different hypotheses can be compared only for the same data (they are not probabilities, and only describe relative support). One way to do this is to compute the ratio of the likelihood values of two hypotheses, providing they are nested (i.e. that one is a particular case of the other). This test usually follows a χ^2 distribution: the null hypothesis is that the most parsimonious hypothesis (with the least number of parameters) is true.

Markov chains: used to model discrete stochastic processes with a finite number of states where the state in the next unit of time depends only on the current state of the system, and is in no way influenced by the state one or more steps before (this is usually referred to as the Markovian hypothesis). Whether the system changes its state or not is determined by transition probabilities.

Metropolis–Hastings sampling: a widely used Monte Carlo method based on a modification of the original Monte Carlo scheme[a,b]. It uses Markov chains to sample through the different states of the stochastic process under study to sample the most probable states with a higher probability, thus avoiding sampling all states (whose number can be extremely large).

Monte Carlo (methods): this generic term covers a wide range of methods whose basic principle is to obtain by stochastic simulation the distribution of random variables that would be impossible to obtain through analytical calculus, typically because this involves the integration of an extremely large number of components or dimensions.

Multifurcations: when one genealogical lineage instantaneously gives rise to more than two descendant lineages (cf. bifurcation, in which one lineage gives rise to two descendant lineages, typical of species level phylogenies).

Nesting algorithm: an algorithm that identifies clades grouped by mutational changes, step by step, until the final level of nesting comprises the entire network. The end result is a phylogenetic network of DNA haplotypes, grouped hierarchically, with higher level nestings corresponding to earlier coalescent events.

Pairwise difference: the number of nucleotide sites at which two alleles differ in their nucleotide state. The average pairwise difference for more than two sequences is the average difference across all possible pairwise comparisons.

Segregating sites: the number of nucleotide sites that exhibit polymorphism within a sample of DNA sequences.

Ultrametric tree: a phylogenetic tree with branch lengths corresponding to some measure of evolutionary change (genetic distance), where the amount of change from the branch tips to the roots is the same for all lineages. An ultrametric tree enforces the assumption of a molecular clock (that the same rate of nucleotide substitution occurs within each lineage), and enables the ages of branching events to be measured directly from the tree.

References

a Metropolis, N. *et al.* (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092

b Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109

of θ in a population of constant size can be generalized to include parameters for change in population size over time[42]. Although computationally faster, least squares estimators assume an infinite sites mutational model of sequence evolution, which is often less realistic than explicit models of DNA sequence evolution that feature in the calculation of some maximum likelihood estimators[29,32,41] but not all[30]. A future approach might be to combine likelihood and least squares methods, and it has recently been suggested that a hybrid theory of these two approaches would exploit the strengths of both methods and minimize their respective weaknesses[40].

Geography, genealogy and demographic history

Coalescent-based approaches to estimating demographic history from DNA sequence data typically assume no selection, a lack of recombination, random mating within a single population, and random sampling. Although little is known about the effects of selection or recombination on parameter estimates, simulation analyses have shown that the classic unimodal distribution of pairwise differences from a population that has undergone exponential growth[6,7] might be affected by the presence of population substructure leading to a multimodal distribution[43]. New work incorporating predictions from coalescent theory and simulation analysis within a statistical framework have sought to identify the different contributions of population history and population structure to the geographical arrangement of haplotypes within a population[44,45] (Box 4). This strategy can be divided into three steps: (1) a phylogenetic 'network' of haplotype relationships is constructed using statistical parsimony[44,46] to deal with problems encountered when applying traditional methods of phylogenetic reconstruction for intraspecific gene genealogies[44,47]. The problems are lower levels of variability, the persistence of ancestral haplotypes, and multifurcations arising from ancestral haplotypes giving rise to multiple descendant lineages. Networks can also be constructed from nonsequence DNA data such as restriction fragment length polymorphism data[48]; (2) a NESTING ALGORITHM[49,50] is then used to arrange

the haplotype network into a nested series of clades; and (3) the geographical coordinates of each haplotype in the network are recorded. The geographical and phylogenetic correspondence of haplotypes is used to test specific hypotheses of spatial patterns for gene flow and population structure developed from simulation models and coalescent theory. The hypotheses that can be tested with this approach are restricted gene flow with isolation by distance, long distance dispersal, past fragmentation and range expansion[51].

Nested cladistic analysis of phylogeographical data requires broad geographical sampling across the distribution limits of a population. Because of its nested design, this method can partition the combined effects of multiple population structural or historical scenarios that could be acting, or have acted upon, a population across its geographical range. Studies of human evolution using nested cladistic analysis of Y chromosome data[52,53] have revealed patterns of expansion from Asia into North America and also back into Africa. These studies also suggest different sex-specific demographic histories[53] when Y chromosome patterns are compared with mtDNA and β-globulin data (Box 4). Past population fragmentation and contiguous range expansions have been identified from mtDNA as significant factors in the demographic history of the chrysomelid beetle *Timarcha goettingensis* species complex[54]. MtDNA haplotype data suggest isolation by distance, and past population expansions provide an explanation for the phylogeography of the chrysomelid beetle *Gonioctena pallida* in the Vosges mountains of France[55]. Beyond population demography, further extensions of hypothesis testing with nested clade analysis relate to assessing the frequency and direction of cross-species viral transmission[50], and identifying DNA regions responsible for disease risk[44].

Conclusions and future prospects

New approaches using phylogenetics and population genetics can be used to describe evolution below the species level, and the methodological approaches presented here comprise a solid basis for further developments. Although many of the population dynamic history models being tested are restrictive because they assume a single pattern of growth, sequence data from more loci might allow more realistic models, such as exponential growth followed by a steady state period, to be tested[32]. Sampling additional unlinked loci might also overcome inherent upward biases in estimations of growth rate[32] that are a feature of perhaps all approaches to estimating this parameter. These upward biases are caused by a severe lack of information about the most ancient parts of a single gene genealogy, because only a few lineages from that period are still present in the population. Methods for estimating migration rates and population size[56] will also provide more realistic estimations of population demography. It has recently been suggested that future coalescent based methods might be able to generate estimates under a history of extremely complicated evolutionary dynamics, by using computer programs that can operate on complex data types (e.g. trees)[42]. Current work in coalescent-based graphical methods is seeking to incorporate errors in phylogenetic reconstruction for estimates of demographic history.

New approaches are exploring the use of other forms of DNA data that are more typical of population genetics to test hypotheses of population history. Microsatellite data could be incorporated into methods using pairwise distributions to infer population history[9], and a likelihood approach has been assessed using human and hairy-nosed wombat *Lasiorhinus krefftii* DNA microsatellite data[57,58]. Recent work has also shown that single nucleotide polymorphism data can also be used to estimate summary statistics from phylogenies using a maximum likelihood approach[59,60].

The increasing availability of analytical techniques for inferring demographic history (Box 2) is of major importance to population geneticists, molecular systematists, population ecologists and epidemiologists. The methods available and their continued development also offer new perspectives for studying genome evolution as the number of fully sequenced organisms increases. For example, the full complement of sequences for families of transposons, short and long interspersed nuclear elements could potentially be used to estimate diversification rates within each sequence family, allowing researchers to determine which groups are increasing in copy number, which are constant, and which might be declining. With the judicious selection of genetic marker(s), sampling scheme and hypotheses to be tested, biologists are becoming well equipped for the interpretation of the demographic histories of organisms.

Acknowledgements
We thank K.Crandall, O.Pybus, and three anonymous referees for their comments on this review, which was motivated by research projects funded by NERC (GR3/09807 and GR8/03693), BBSRC (83/D09448) and the Royal Society. Box 1 was abstracted with permission from 'Lecture notes on gene genealogies', by Alan Rogers (http://mombasa.anthro.utah.edu/~rogers/ant4221/Lecture/molevol.pdf). This is publication 01-067 of the Institut des Sciences de l'Evolution (Unité Mixte de Recherche 5554 du Centre National de la Recherche Scientifique).

References

1 Avise, J.C. (2000) *Phylogeography: The History and Formation of Species*, Harvard University Press
2 Kingman, J.C. (1982) On the genealogy of large populations. *J. Appl. Prob.* 19, 27–43
3 Watterson, G.A. (1975) On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* 7, 256–276
4 Tajima, F. (1983) Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105, 437–460
5 Tajima, F. (1989) The effect of change in population size on DNA polymorphism. *Genetics* 123, 597–601
6 Slatkin, M. and Hudson, R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562
7 Rogers, A.R. and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic divergences. *Mol. Biol. Evol.* 9, 552–569
8 Comas, D. *et al.* (1996) Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol. Biol. Evol.* 13, 1067–1077
9 Polanski, A. *et al.* (1998) Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5456–5461

10 Grassly, N.C. *et al.* (1999) Population dynamics of HIV-1 inferred from gene sequences. *Genetics* 151, 427–438
11 Felsenstein, J. (1992) Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.* 59, 139–147
12 Nee, S. *et al.* (1995) Inferring population history from molecular phylogenies. *Philos. Trans. R. Soc. London Ser. B* 349, 25–31
13 Lavery, S. *et al.* (1996) Genetic patterns suggest exponential population growth in a declining species. *Mol. Biol. Evol.* 13, 1106–1113
14 Holmes, E.C. *et al.* (1995) Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos. Trans. R. Soc. London Ser. B* 349, 33–40
15 Ong, C.K. *et al.* (1996) Inferring the population history of an epidemic from a phylogenetic tree. *J. Theor. Biol.* 182, 173–178
16 Ong, C.K. *et al.* (1997) Elucidating the population histories and transmission dynamics of papillomaviruses using phylogenetic trees. *J. Mol. Evol.* 44, 199–206
17 Sanderson, M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14, 1218–1231
18 Thorne, J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15, 1647–1657
19 Huelsenbeck, J.P. *et al.* (2000) A compound Poisson process for relaxing the molecular clock. *Genetics* 154, 1879–1892
20 Pybus, O.G. *et al.* (1999) The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. *Mol. Biol. Evol.* 16, 953–959
21 Pybus, O.G. *et al.* (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155, 1429–1437
22 Holmes, E.C. *et al.* (1999) The molecular population genetics of HIV-1. In *The Evolution of HIV* (Crandall, K.A., ed.), pp. 177–207, Johns Hopkins University Press
23 Paradis, E. (1997) Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. *Proc. R. Soc. London B Biol. Sci.* 264, 1141–1147
24 Paradis, E. (1998) Detecting shifts in diversification rates without fossils. *Am. Nat.* 152, 176–187
25 Paradis, E. (1998) Testing for constant diversification rates using molecular phylogenies: a general approach based on statistical tests for goodness of fit. *Mol. Biol. Evol.* 15, 476–479
26 Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*, Chapman & Hall
27 Emerson, B.C. *et al.* (2000) Colonisation and diversification of the species *Brachyderes rugatus* (Coleoptera) on the Canary Islands: evidence from mtDNA COII gene sequences. *Evolution* 54, 911–923
28 Emerson, B.C. *et al.* (2000) Tracking colonisation and diversification of insect lineages on islands: mtDNA phylogeography of *Tarphius canariensis* (Coleoptera: Colydiidae) on the Canary Islands. *Proc. R. Soc. London B Biol. Sci.* 267, 2199–2205
29 Kuhner, M.K. *et al.* (1995) Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* 140, 1421–1430
30 Bahlo, M. and Griffiths, R.C. (2000) Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57, 79–95
31 Fu, Y-X. (1994) A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136, 686–692
32 Kuhner, M.K. *et al.* (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149, 429–434
33 Walton, C. *et al.* (2000) Population structure and population history of *Anopheles dirus* mosquitos in southeast Asia. *Mol. Biol. Evol.* 17, 962–974
34 Brown, R.P. *et al.* (2000) Mitochondrial DNA evolution and population history of the Tenerife skink *Chalcides viridanus*. *Mol. Ecol.* 9, 1061–1067
35 Zink, R.M. *et al.* (2000) Genetics, taxonomy, and conservation of the threatened California gnatcatcher. *Conserv. Biol.* 14, 1394–1405
36 Milot, E. *et al.* (2000) Phylogeography and genetic structure of northern populations of the yellow warbler (*Dendroica petechia*). *Mol. Ecol.* 9, 667–681
37 Vilà, C. *et al.* (1999) Mitochondrial DNA phylogeography and population history of the grey wolf *Canis lupus*. *Mol. Ecol.* 8, 2089–2103
38 Crandall, K.A. *et al.* (1999) Effective population sizes: missing measures and missing concepts. *Anim. Conserv.* 2, 317–319
39 Griffiths, R.C. and Tavaré, S. (1996) Monte Carlo inference methods in population genetics. *Math. Comput. Model.* 23, 141–158
40 Felsenstein, J. *et al.* (1999) Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics* (Seiller-Moisewitcsh, F., ed.), pp. 163–185, American Mathematical Society
41 Weiss, G. and von Haesler, A. (1998) Inference of population history using a likelihood approach. *Genetics* 149, 1539–1546
42 Vasco, D.A. *et al.* (2000) Molecular population genetics: coalescent methods based on summary statistics. In *Computational and Evolutionary Analysis of HIV Molecular Sequences* (Rodrigo, A.G. and Learn, G.H., Jr, eds), pp. 173–218, Kluwer
43 Marjoram, P. and Donnelly, P. (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* 136, 673–683
44 Crandall, K.A. and Tempelton, A.R. (1996) Applications of intraspecific phylogenetics. In *New Uses for New Phylogenies* (Harvey, P. *et al.*, eds), pp. 187–202, Oxford University Press
45 Templeton, A.R. (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* 7, 381–397
46 Templeton, A.R. *et al.* (1992) A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132, 619–633
47 Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45
48 Clement, M. *et al.* (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659
49 Templeton, A.R. and Sing, C.F. (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* 134, 659–669
50 Crandall, K.A. (1996) Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. *Mol. Biol. Evol.* 13, 115–131
51 Posada, D. *et al.* (2000) GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* 9, 487–488
52 Karafet, T.M. *et al.* (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* 64, 817–831
53 Hammer, M.F. *et al.* (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15, 427–441
54 Gómez-Zurita, J. *et al.* (2000) Nested cladistic analysis, phylogeography and speciation in the *Timarcha goettengensis* complex (Coleoptera, Chrysomelidae). *Mol. Ecol.* 9, 557–570
55 Mardulyn, P. and Milinkovitch, M.C. Phylogeography of a regional population of the leaf beetle *Gonioctena pallida* (Coleoptera: Chrysomelidae): a nested clade analysis of the geographical distribution of mitochondrial DNA haplotypes within the Vosges Mountains. *Mol. Ecol.* (in press)
56 Beerli, P. and Felsenstein, J. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152, 763–773
57 Beaumont, M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics* 153, 2013–2019
58 Garza, J.C. and Williamson, E.G. (2001) Detection of reduction in population size using data from microsatellite loci. *Mol. Ecol.* 10, 305–318
59 Kuhner, M.K. *et al.* (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156, 439–447
60 Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154, 931–942