



## GENIE: estimating demographic history from molecular phylogenies

O. G. Pybus\* and A. Rambaut

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received on February 15, 2002; revised on April 12, 2002; accepted on April 17, 2002

### ABSTRACT

**Summary:** GENIE implements a statistical framework for inferring the demographic history of a population from phylogenies that have been reconstructed from sampled DNA sequences. The methods are based on population genetic models known collectively as coalescent theory.

**Availability:** GENIE is available from <http://evolve.zoo.ox.ac.uk>. All popular operating systems are supported.

**Contact:** [oliver.pybus@zoo.ox.ac.uk](mailto:oliver.pybus@zoo.ox.ac.uk)

Coalescent theory refers to a group of mathematical models that describe the statistical properties of intra-population phylogenies (Kingman, 1982). The theory is used to estimate population genetic parameters (e.g. effective population size, recombination and migration rates) from sampled gene sequences, and has been applied in many fields, including anthropology, conservation biology and epidemiology (e.g. Pybus *et al.*, 2001). The shared history of the sampled sequences creates a *genealogy*, the lineages of which ‘coalesce’ as time progresses into the past, until the most recent common ancestor of the sample is reached.

One particular coalescent model, the variable population size model, describes how the shape of the genealogy depends on the demographic history of the sampled population (Griffiths and Tavaré, 1994). It provides a probability distribution for the waiting times between coalescent events (when two lineages merge) in a genealogy. This distribution depends on the *demographic function*, denoted  $N_e(t)$ , which represents the effective population size at time  $t$  before the present. GENIE (GENealogy Interval Explorer) implements a statistical framework for inferring the demographic function from reconstructed phylogenies (Pybus *et al.*, 2000).

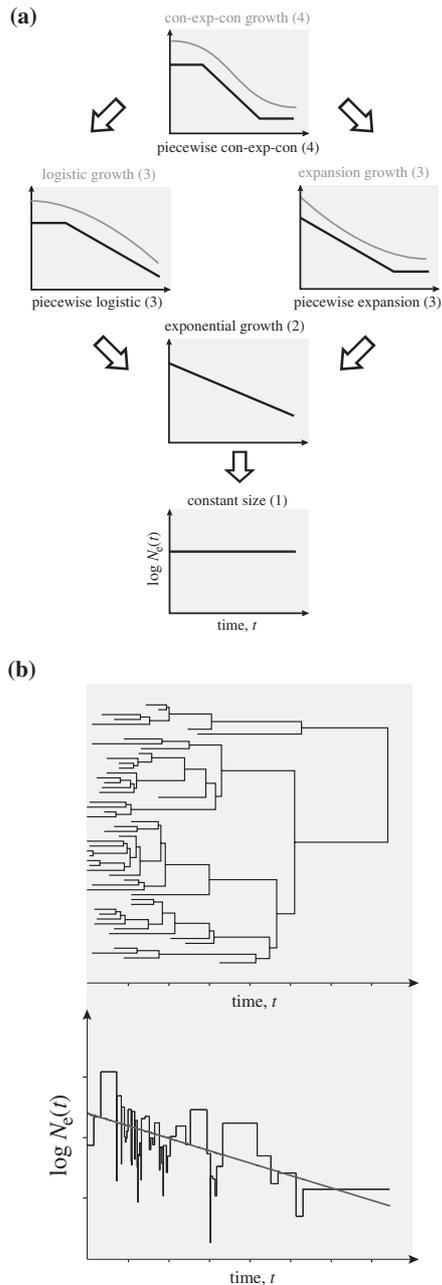
A fully-specified demographic function  $N_e(t)$  constitutes a *demographic hypothesis*, whose likelihood is given by the coalescent probability distribution. GENIE calculates this likelihood and implements two approaches for estimating  $N_e(t)$ . The first approach represents  $N_e(t)$  using *demographic models*, simple mathematical functions that characterize biologically plausible population

dynamic histories, such as exponential or logistic growth. The demographic models have 1 to 4 demographic parameters and form two families, one containing continuous models, the other containing piecewise models (Figure 1a). Parameters are estimated numerically using maximum likelihood, and approximate 95% confidence intervals are obtained using the likelihood ratio test statistic (Pybus *et al.*, 2000). This statistic can also be used to test the goodness-of-fit of any two nested models. If the models are not nested then they can be compared using the corrected Akaike Information Criterion (AICC). Further details about the demographic models are available in the GENIE manual.

The second approach for estimating  $N_e(t)$  is the skyline plot, which represents  $N_e(t)$  using a very general stepwise function, such that  $N_e(t)$  is constant within each step, but changes between steps (Figure 1b). The number of steps equals the number of demographic parameters. The classic skyline plot allows  $N_e(t)$  to change every time there is a coalescent event in the genealogy (Pybus *et al.*, 2000). It has one parameter for each data point and thus has the highest likelihood of any demographic hypothesis. The *generalized skyline plot* allows adjacent steps to be grouped, thereby reducing both the number of parameters and the likelihood. The optimal amount of grouping can be found by maximizing the AICC of the plot (Strimmer and Pybus, 2001). The skyline plot is quick to compute and is useful as a model selection tool, suggesting which simple demographic model will fit the data. Similarly, skyline plot and demographic model estimates of  $N_e(t)$  should correspond closely when the correct model is chosen.

The coalescent models GENIE uses require that the branch lengths of inputted genealogies are proportional to time. Such trees can be obtained using a maximum likelihood approach under the assumption of a molecular clock. Alternatively, there are several methods for estimating genealogies when the rate of evolution varies among branches in the tree (Sanderson, 2002 and references therein). GENIE also implements the serial-sample coalescent model (Rodrigo and Felsenstein, 1999) and can therefore estimate  $N_e(t)$  when sequences have been sampled at different points in time (Figure 1b). This is

\*To whom correspondence should be addressed.



**Fig. 1.** (a) The demographic models. Grey curves belong to the continuous family, black curves belong to the piecewise family. The constant size and exponential models belong to both families. The number of demographic parameters are shown in brackets. Arrows indicate nested models that differ by one parameter; the simpler model can be rejected in favour of the more complex model using a likelihood ratio test with one degree of freedom. (b) An example analysis. A genealogy with sequences sampled at different time points (top) is shown on the same time scale as the classic skyline plot for this tree (bottom; jagged line). The maximum likelihood estimate of  $N^e(t)$  using the exponential model is also shown (bottom; straight line).

especially useful for data sets containing rapidly evolving viral genes, or ancient DNA. The time of sampling must be accounted for when estimating branch lengths, so specialized software must be used (Rambaut, 2000; Drummond and Rodrigo, 2000). If multiple genealogies are imported into GENIE then they can be treated either as independent data sets, or as multiple unlinked loci from the same population sample.

The analysis results are written to the screen and an optional log file. Results can be opened directly in spreadsheet applications. GENIE has a command-line user interface that can also be automated by appending a list of commands to the input file.

GENIE infers demographic history from estimated phylogenies, not sequence data. Complex analyses can therefore be performed rapidly, but the results do not incorporate error arising from phylogenetic uncertainty. Thus GENIE is best suited to data sets containing much phylogenetic information and is complementary to other packages that do incorporate phylogenetic error (LAMARC; <http://evolution.genetics.washington.edu>, and GENETREE; <http://www.stats.ox.ac.uk>).

## ACKNOWLEDGEMENTS

Many thanks to our collaborators and the referees for their ideas and feedback. Funded by the Wellcome Trust (grant 50275) and The Royal Society.

## REFERENCES

- Drummond, A. and Rodrigo, A.G. (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial sample UPGMA. *Mol. Biol. Evol.*, **17**, 1807–1815.
- Griffiths, R.C. and Tavaré, S. (1994) Sampling theory for neutral alleles in a varying environment. *Phil. Trans. Roy. Soc. Lond. B*, **344**, 403–410.
- Kingman, J.F.C. (1982) On the genealogy of large populations. *J. Appl. Prob.*, **A19**, 27–43.
- Pybus, O.G., Rambaut, A. and Harvey, P.H. (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**, 1429–1437.
- Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. and Harvey, P.H. (2001) The epidemic behaviour of the Hepatitis C virus. *Science*, **292**, 2323–2325.
- Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood estimates. *Bioinformatics*, **16**, 395–399.
- Rodrigo, A.G. and Felsenstein, J. (1999) Coalescent approaches to HIV population genetics. In Crandall, K. (ed.), *The Evolution of HIV*. John Hopkins University Press, Baltimore.
- Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalised likelihood approach. *Mol. Biol. Evol.*, **19**, 101–109.
- Strimmer, K. and Pybus, O.G. (2001) Exploring the demographic history of DNA sequences using the generalised skyline plot. *Mol. Biol. Evol.*, **18**, 2298–2305.