BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 4 2003, pages 539–540
DOI: 10.1093/bioinformatics/btg024

PopHist: inferring population history from the spectrum of allele frequencies

Stephen Wooding

Eccles Institute of Human Genetics, University of Utah, 15 North 2030 East, Salt Lake City, UT 84112-5330, USA

Received on February 14, 2002; revised on September 24, 2002; accepted on October 25, 2002

ABSTRACT

Summary: *PopHist* is a computer program that uses the frequency spectrum of alleles to: (a) estimate maximum likelihood parameters describing a population's history; and (b) compare alternative hypotheses about population history using likelihood ratio tests. The program uses the matrix coalescent, a method for calculating theoretical frequency spectra that can be applied to sets of unlinked sites.

Availability: Source code and documentation are available at http://mombasa.anthro.utah.edu/wooding/PopHist

Contact: swooding@genetics.utah.edu

The frequency spectrum of alleles, or frequency spectrum, is the distribution describing the number of derived genetic variants that occur $k = 1, 2, \ldots, n-1$ times in a sample of n homologous chromosomes. This distribution is known to be sensitive to population size changes. Populations that have rapidly increased in size tend to have relative overabundances of rare alleles, for instance, while constant-sized populations do not (Harpending *et al.*, 1998). Thus, the frequency spectrum can be used to make inferences about past population history from present patterns of variation.

Analyses of the frequency spectrum of linked polymorphisms are used as the basis for several well known statistical tests for constant population size and evolutionary neutrality (e.g. Tajima 1989). Recent developments in coalescent theory have provided methods for tests based on the frequency spectrum of unlinked polymorphisms. However, these methods are analytically complicated and have not been adopted widely, even though they achieve substantial statistical power through the use of independent loci. The *PopHist* computer program tries to identify the population history that best explains a frequency spectrum composed of sets of linked and unlinked polymorphisms.

PopHist takes two files as input: a list of segregating polymorphisms and a model of population history. For each polymorphism, the locus, number of chromosomes sampled, ascertainment procedure, and character state polarity are listed in a datafile. Linked sites can be included by specifying that they come from the same locus. The model of population history describes a series of epochs that define population size during different time periods.

PopHist performs two main analyses: parameter estimation and hypothesis testing. Estimates of maximum likelihood parameters are obtained by using a standard Metropolis algorithm (Robert and Casella, 1999, Chap. 5). This algorithm explores the parameter space defined by the user's model of population history as follows. The process begins with a population history of constant size. Then, prospective parameters are chosen by perturbing the current parameters. The likelihood of the new parameters is calculated, and they are adopted with probability $exp((L_p - L_c)/T)$, where L_p is the likelihood of the prospective parameters, L_c is the likelihood of the current parameters, and T is a scaling parameter (Robert and Casella, 1999, p. 199). If L_p is greater than L_c, the prospective parameters are accepted with probability 1. The process of perturbation followed by acceptance or rejection of the new parameters is iterated for a length of time specified by the user. The maximum likelihood parameters returned by this process are not exact, but they closely approximate true values when data with known parameters are used. Pairs of user-specified hypothetical population histories can be compared in a likelihood ratio test that reports the likelihood of each alternative as well as a p-value.

PopHist models the frequency spectrum of mutations using the matrix coalescent method of Wooding and Rogers (2002). The matrix coalescent describes the frequency spectrum of mutations (D) drawn from a sample of homologous chromosomes in a population whose size history (H) is described by a function, $N(t)$. Given $N(t)$, it is possible to calculate the probability of an observed spectrum, yielding the likelihood $L = P(D|H)$. Alternative, hypothetical population histories can be compared using standard likelihood ratio tests (Bain and Engelhart, 1992; Edwards, 1992).

The matrix coalescent can be used to model population histories of arbitrary complexity, but the *PopHist* pro-

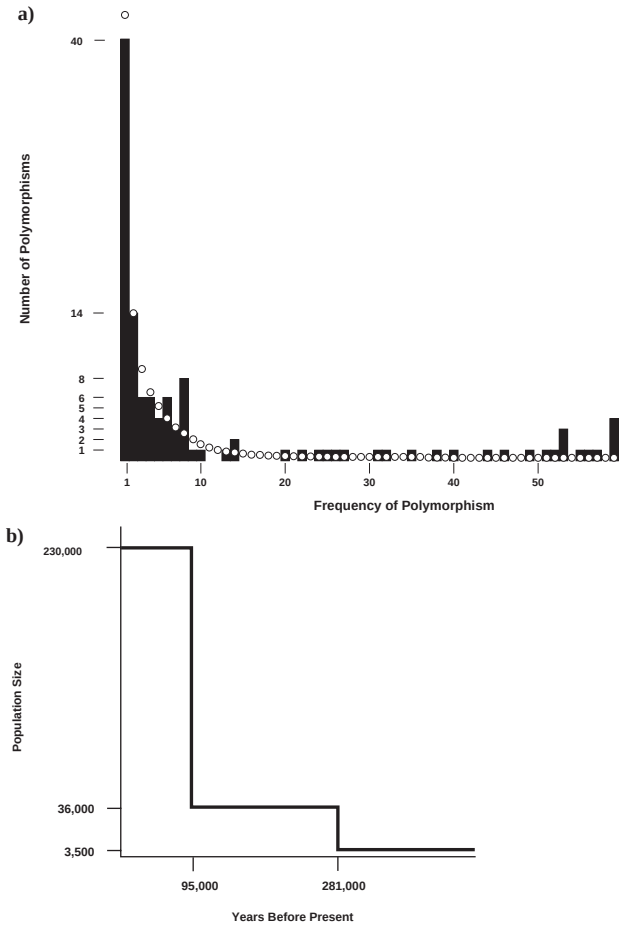

Fig. 1. Example application. (a) Frequency spectrum observed by Yu *et al.* (2002) (bars) and frequency spectrum expected under maximum likelihood parameters estimated by *PopHist* (circles). The data are composed of DNA sequence polymorphisms obtained from a collection of humans spread across Africa, Asia and Europe. (b) Maximum likelihood history inferred by *PopHist*.

gram is limited to histories that are piecewise-constant in order to retain speed and flexibility. Piecewise-constant population histories are composed of a series of epochs of constant population size, each described by two parameters: τ, a value proportional to the length of the epoch, and θ, a value proportional to the population size during the epoch (Fig. 1). Such histories are simple analytically, but are limited by the fact that the number of model parameters grows rapidly with the number of epochs used. The increase in the number of parameters with model complexity lengthens the maximization process rapidly as epochs are added. Complicated histories can sometimes be approximated with a small number of epochs (Rogers, 1997). It may also be possible to use Akaike's information criterion (Akaike, 1974) to group epochs, reducing their number, in a way analogous to that suggested by Strimmer and Pybus (2001) for coalescent intervals.

In spite of the relative simplicity of the maximization algorithm used, *PopHist* performed well in trial runs. With simulated data, *PopHist* converged on known parameters consistently. Known values fell within 95% confidence intervals 48 out of 50 times, for example, when two-epoch models were tested. In applications to real data, *PopHist* was able to distinguish the frequency spectra of different types of single-nucleotide polymorphism (synonymous and non-synonymous), providing evidence for the action of natural selection (Wooding and Rogers, 2002). An example use of *PopHist* is shown in Figure 1.

PopHist is written in C++. Source code is available from the author. A translation of the *matcoal* C++ library from Rogers's *GTree* package (Rogers, 1999) is used by *PopHist* for the calculation of coalescent intervals, which are used in the calculation of the theoretical frequency spectrum.

ACKNOWLEDGEMENTS

The author was supported by a NIH Genome Sciences (Genome Informatics) Training Grant to the University of Utah.

REFERENCES

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **AC-19**, 716–723.

Bain,L.J. and Engelhart,M. (1992) *Introduction to Probability and Mathematical Statistics*, 2nd edition, Duxbury Press, Bolmont, CA.

Edwards,A.W.F. (1992) *Likelihood*. The Johns Hopkins University Press, Baltimore, MD.

Harpending,H.C., Batzer,M.A., Gurven,M., Jorde,L.B. and Rogers,A.R. (1998) Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA*, **95**, 1961–1967.

Robert,C.R. and Casella,G. (1999) *Monte Carlo Statistical Methods*. Springer, New York.

Rogers,A.R. (1997) Population structure and modern human origins. In Donnelly,P.J. and Tavare,S. (eds), *Progress in population Genetics and Human Evolution*. Springer, New York.

Rogers,A.R. (1999) GTree Version 1.0, Computer program distributed by the author. Department of Anthropology, University of Utah.

Strimmer,K. and Pybus,O.G. (2001) Exploring the demographic history of DNA sequences using the generalized skyline. *Plot. Mol. Biol. Evol.*, **18**, 2298–2305.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

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

Yu,N., Chen,F.C., Ota,S., Jorde,L.B., Pamilo,P., Patthy,L., Ramsay,M., Jenkins,T., Shyue,S.-K. and Li,W.-H. (2002) Larger genetic differences within Africans than between Africans and Eurasians. *Genetics*, **161**, 269–274.