BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1733–1734
doi:10.1093/bioinformatics/bti154

Genetics and population analysis

Serial SimCoal: A population genetics model for data from multiple populations and points in time

Christian N. K. Anderson, Uma Ramakrishnan, Yvonne L. Chan and Elizabeth A. Hadly*

Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA

Received on September 17, 2004; revised on November 9, 2004; accepted on November 12, 2004
Advance Access publication November 25, 2004

ABSTRACT

Summary: We present Serial SimCoal, a program that models population genetic data from multiple time points, as with ancient DNA data. An extension of SIMCOAL, it also allows simultaneous modeling of complex demographic histories, and migration between multiple populations. Further, we incorporate a statistical package to calculate relevant summary statistics, which, for the first time allows users to investigate the statistical power provided by, conduct hypothesis-testing with, and explore sample size limitations of ancient DNA data.

Availability: Source code and Windows/Mac executables at http://www.stanford.edu/group/hadlylab/ssc.html

contact: senka@stanford.edu

1 INTRODUCTION

Biological sampling from ancient sources, including the sequencing of ancient DNA, is becoming increasingly relevant to the fields of biosystematics, population genetics and molecular evolution. Recent progress has focused on the development and standardization of molecular methods (e.g. Hofreiter *et al.*, 2001); however, the statistical analysis of ancient DNA data still remains challenging.

The original coalescent framework (Hudson, 1990; Kingman, 1982) allowed modeling of population genetic data from a single point in time from a single population, and significantly improved the efficiency of population genetics simulations. To model data from two points in time, serial coalescent approaches have been developed (Rodrigo and Felsenstien, 1999). Serial coalescent models are used in a Bayesian framework (e.g. BEAST) to estimate mutation rate and effective population size (e.g. Drummond *et al.*, 2002; Lambert *et al.*, 2002) from data at multiple time points. However, BEAST is limited by assumptions of (1) a single population, (2) simple demographic processes like exponential growth and (3) a time-span long enough relative to mutation rate to make sequence-based genetic data useful.

In practice, ancient DNA data often violate these three assumptions. First, while single population models may provide inference regarding species history, multiple population models are necessary for answering questions about realistic population history. Second, in refining species-based questions to population-level analyses, we often investigate complex demographic hypotheses; for example, whether observed genetic data (ancient and modern) are consistent with a discrete, linear, or arbitrary climate-mediated change in population size (Hadly *et al.*, 2004). Finally, in studies where temporal samples are fairly recent (e.g. 3000 year-old human samples are 150 generations old), mutation rates for sequence-based data are too low to provide useful information. In such cases, fast mutating short tandem STR or microsatellite polymorphisms can increase power. For the three situations above (multiple populations, population-level questions and short time-scales), existing approaches for analysis of historic genetic data have proved inadequate.

*To whom correspondence should be addressed.

2 METHODS

Here we develop a model that simulates ancient and modern genetic data for user-specified population histories. We modified the publicly available SimCoal (Excoffier *et al.*, 2000) to include serial sampling. Our program (Serial SimCoal) allows multiple sampling time points from several populations. As with SimCoal, it is capable of simulating migration between populations and 'historical events' (user-defined changes in migration patterns, population size and/or population growth rate). Further, Serial SimCoal simulates either STR polymorphisms or DNA sequences.

We integrated a statistical analysis package into the program to facilitate hypothesis-testing. This allows users to output simulated values of chosen statistics based on user-specified input. As with SIMCOAL2, multiple coalescences per generation are allowed. In addition, we also make available a small shell program that allows multiple input files to be generated quickly and run in a batch. A continuous time method was added, which increases computation efficiency without sacrificing accuracy. The program is backwards-compatible with SIMCOAL (any input files previously created for SIMCOAL can also be used in Serial SimCoal). The program is written in C++; source code and executable versions for Windows 95/98/NT/XP, and MacOS X are available. The program tested successfully for simple models against theoretical expectations.

3 RESULTS

An example of simulations from our program reveals the power provided by ancient DNA data to understand population history. We modeled the following population history: two populations ($N_e = 25{,}000$ each) diverged 10,000 generations ago. We assumed both populations were sampled 5000 generations ago (ancient sample) and in the present (modern sample). Sample size was assumed to be 20 in all cases. Given this basic scenario, we allowed (1) no migration during the entire 10,000 generations, (2) recent migration during the last 5000 generations and (3) ancient migration during the first 5000 generations following population divergence. Migration was assumed to be symmetric, at 10 migrants per generation. Figure 1 shows the proportion of private haplotypes (haplotypes

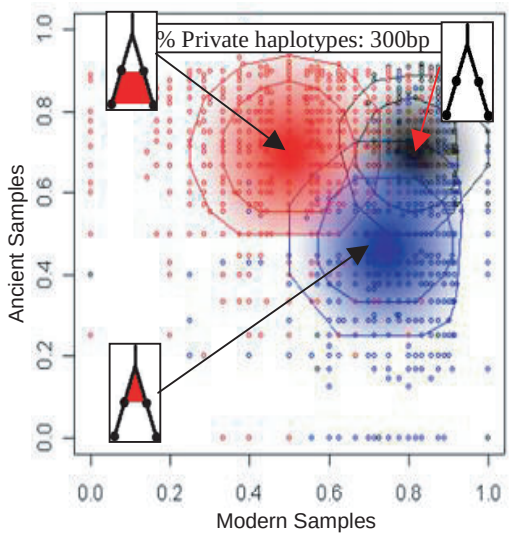

Fig. 1. Sample output for a simulation. Two populations of 25,000 split 10,000 generations ago; 20 samples were taken from each deme at 5000 generations ago and at present. The percentage of private haplotypes is determined for the ancient and modern samples. After 1000 runs, there is a clear difference between models with ancient (blue), recent (red), and no (black) migration between demes for 300bp of mtDNA. The outer polygons contain 90% of the simulated results, the inner polygons contain 70%.

unique to either population, calculated by our program) for all three scenarios when 300 bp of sequence data are modeled (mutation rate = 3%/bp/million years, typical of cyt b).

There is a clear difference between the three models in the proportion of private haplotypes in the ancient and modern samples for the sequence data (Fig. 1a). Although the 90% confidence polygons overlap somewhat, there is almost no overlap between the 70% polygons.

4 DISCUSSION

We foresee many potential applications for this program. Discrimination between times of migratory events is often difficult with only modern data. As we demonstrate, both our program and ancient DNA data illuminate this problem. Additionally, Serial SimCoal can be used to generate thousands of combinations of ancient and modern genetic diversity for any population history. These results may be used to statistically reject the hypothesis that, for example, the drop in genetic diversity between two sampled time points occurs by chance with a stable population size. Furthermore, source materials (for example, bones and teeth) for ancient DNA data are rare and extraction difficult, so sampling will always be an issue in ancient DNA studies. Serial SimCoal allows users to investigate the impact of small sample sizes and quantify the gains in power available by further sequencing. Finally, Serial SimCoal can be used in an approximate Bayesian computational framework (Beaumont *et al.*, 2002) to estimate distributions for demographic and genetic parameters in the evolutionary past of species and populations. A version of the program that does just this will be available shortly. In summary, Serial SimCoal will prove valuable to the field of ancient DNA in the following situations: (1) theoretical investigation of power provided by ancient DNA approaches, (2) hypothesis-testing of particular evolutionary histories using bootstrap analysis, (3) investigation of sampling limitations and (4) parameter estimation.

ACKNOWLEDGEMENTS

Funding for development of this program and website was received from NSF (grant DEB# 0108541 to E.A.H). We thank M. van Tuinen, L. Excoffier and two anonymous reviewers for comments on earlier versions of this manuscript.

REFERENCES

Beaumont,M.A., Zhang,W. and Balding,D.J. (2002) Approximate Bayesian computation in population genetics. *Genet.*, **162**, 2025–2035.

Drummond,A.J., Nicholls,G.K., Rodrigo,A.G. and Solomon,W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genet.*, **161**, 1307–1320.

Excoffier,L., Novembre,J. and Schneider,S. (2000) SimCoal: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered.*, **91**, 506–509.

Hadly,E.A., Ramakrishnan,U., Chan,Y.L., van Tuinen,M, Conroy,C., Spaeth,P.E. and O'Keefe,K. (2004) Genetic response to climatic change: insights from ancient DNA and phylochronology. *PLoS Biol.*, **2**, e290.

Hudson,R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 1–44.

Hofreiter,M., Serre,D., Poinar,H.N., Kuch,M. and Paabo,S. (2001) Ancient DNA. *Nat. Rev. Genet.*, **2**, 353–359.

Kingman,J.F.C. (1982) The coalescent. *Stoc. Proc. Ap.*, **13**, 235–248.

Lambert,D.M., Ritchie,P.A., Millar,C.D., Holland,B., Drummond,A.J. and Baroni,C. (2002) Rates of evolution in ancient DNA from Adelie penguins. *Science*, **295**, 2270–2273.

Rodrigo,A.G. and Felsenstein,J. (1999) Coalescent approaches to HIV-1 population genetics. In Crandall,K.A., ed. *The Evolution of HIV*. Johns Hopkins University Press.