BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 21 2009, pages 2845–2847
doi:10.1093/bioinformatics/btp534

Genetics and population analysis

COMPASS: a program for generating serial samples under an infinite sites model

Mattias Jakobsson

Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

Received on July 1, 2009; revised on August 14, 2009; accepted on September 4, 2009

Advance Access publication September 17, 2009

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: The program *COMPASS* can generate samples that have been collected at various points in time from a population that is evolving according to a Wright–Fisher model. The samples are generated using coalescence simulations permitting various demographic scenarios and the program uses an infinite sites model to generate polymorphism data for the samples. By generating serially sampled population-genetic data, *COMPASS* allows investigating properties of polymorphism data that has been collected at different time points, and aid in making inference from ancient polymorphism data.

Availability: The program and the manual are available at: http://www.egs.uu.se/evbiol/Research/JakobssonLab/compass.html

Contact: mattias.jakobsson@ebc.uu.se

1 INTRODUCTION

Ancient DNA (aDNA) can be useful to detect and date demographic events in the history of populations and species (see e.g. Hofreiter, 2008; Willerslev and Cooper, 2005). The number of aDNA studies have increased over the last few years and spectacular data have been generated from Neanderthals, cave bear and Mammoth (Miller *et al.*, 2008; Noonan *et al.*, 2005, 2006).

Rodrigo and Felsenstein (1999) extend the standard coalescent model by considering serially sampled gene copies. The idea of serial samples has been exploited in the *BEAST* software (Drummond *et al.*, 2002) to estimate demographic parameters of populations or species using data from multiple time points. A software that simulates data from serial samples is *Serial SimCoal* by Anderson *et al.* (2005). These two softwares have different primary aims—estimation of demographic parameters and simulating data—and both programs have different limitations and strengths (Anderson *et al.*, 2005). However, neither of these two programs use an infinite sites model (Kimura, 1969), and in many circumstances the infinite sites model may be appropriate and/or more straightforward to use. For example, the infinite sites model is appropriate when simulating or analyzing population–genetic SNP data, and aDNA studies focusing on SNPs have increased in popularity in the last few years (Burger *et al.*, 2007; Ludwig *et al.*, 2009; Svensson *et al.*, 2007). Regardless of model details, simulations of serially sampled data can provide means for exploring properties of population–genetic data sampled at multiple time points. Simulations may also be used as part of analysis frameworks, such as an approximate Bayesian computation approach (see e.g. Beaumont *et al.*, 2002).

2 METHODS

The program *COMPASS* generates samples and polymorphism data assuming an infinite sites model, under a coalescent model allowing serially sampled gene copies and permitting various demographic scenarios. *COMPASS* can generate many independent replicate samples under various assumptions about sample times, population sizes and population size changes. The samples are generated using standard coalescent approaches where the random genealogy of the sample is first generated, followed by randomly 'dropping' mutations to the genealogy (Hudson, 1990; Kingman, 1982; Nordborg, 2001). An infinite sites model (Kimura, 1969) is assumed so that every mutation gives rise to a new variable site. To allow serial samples, the basic genealogy-generating algorithm has been modified to allow lineages sampled in the past to be incorporated when the sample times are passed (backwards in time). To simulate five replicate samples from two time points (six chromosomes from the present and four chromosomes $4N$ generations in the past) and for one SNP, we would type:

```
COMPASS 10 5 -s 1 -h 0.0 6 -h 1.0 4.
```

The output from *COMPASS* is very similar to the output from the program *ms* by Hudson (2002), which will minimize the need to transform output from *COMPASS* to fit programs and scripts written to handle output from *ms*, such as the program *seq-gen* (Rambaut and Grassly, 1997) that can generate sequence data under different mutation models. Briefly, the output of *COMPASS* consists of the command-line arguments, the random number generator's seed and the simulated data (one chromosome per line and one site per column) for each replicate sample. *COMPASS* is a command-line program, easily run using batch scripts, and parameter values can be specified for each replicate run using the 'tbs' option. Other demographic scenarios and output options are available, for which additional command-line arguments are needed. The *COMPASS* manual describes these options. The program is written in C++, and precompiled executables for UNIX/Linux, Windows and Mac are available for download.

3 AN EXAMPLE OF THREE SAMPLING TIMES

An example will illustrate the generation of population–genetic data from a demographic model where samples have been taken at three different time points. We simulate data (using *COMPASS*) from a scenario of a population that instantaneously decreased to $1/2$ of the ancestral population size (N_2) 100 000 years ago, and that started growing 60 000 years ago to reach four times the ancestral population size at present (Fig. 1). Assuming that the population size at present is $N_0 = 40\,000$ and that the generation time is 25 years, the population size at the start of the growth is $N_1 = 1/8 \times N_0$ and the ancestral population size is $N_2 = 1/4 \times N_0$. The sampling times T_0, T_1 and T_2 corresponds to $T_0 = 0$ (the present), $T_1 = 0.015 \times 4N_0$ and $T_2 = 0.025 \times 4N_0$ generations ago. The growth parameter α is

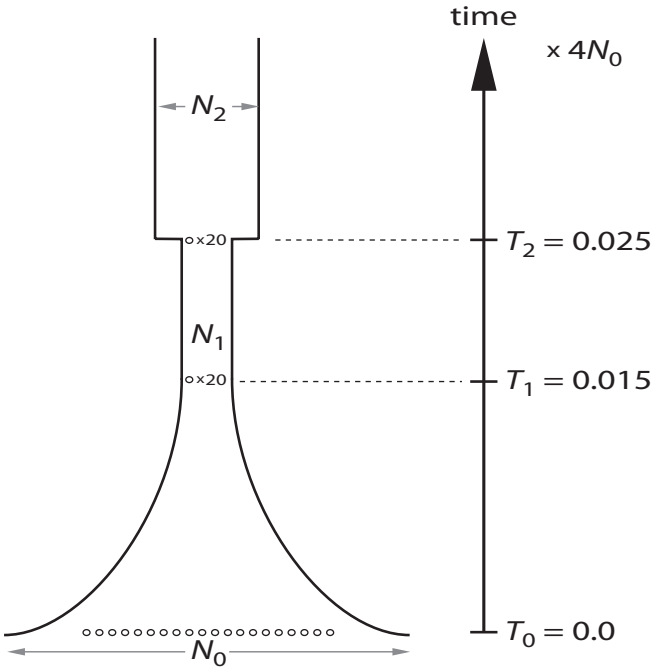

Fig. 1. A demographic model of a population that at time T_2 decreased in size to a half of its ancestral population size (from N_2 to N_1), and at time T_1 started growing exponentially reaching a size at present four times as large as the ancestral size (N_0). Three samples were taken from the population, one sample (S_0) of 20 gene copies at present (T_0), one sample (S_1) of 20 gene copies at time T_1 and one sample (S_2) of 20 gene copies at time T_2.

given by solving the following equality for α; $N = N_0 e^{-\alpha T}$, where N is the population size after T time units of growth, and T is time scaled in units of $4N_0$. We have $\alpha \approx 138.6$. We replicate the simulation 1000 times, where each simulation contains 100 unlinked biallelic markers (e.g. SNPs) for 3×20 haploid gene-copies (corresponding to 10 diploid individuals sampled at each time point).

The mean (standard error of the mean, s.e.) heterozygosities in the samples S_0, S_1 and S_2 across 1000 replicates were 0.1692 (0.00059), 0.1631 (0.00060) and 0.1745 (0.00057), respectively, and the mean (s.e.) heterozygosity across all individuals (and across the replicates) was 0.1788 (0.00054). Note that we could have simulated these three scenarios without using a program that generates serial samples, but we would not have been able to compute a meaningful mean heterozygosity for all 60 gene copies (or the F_{ST}-values in the next paragraph). The diversities of the samples appear reasonable with respect to the past population sizes, i.e. we expect that the 'effective population size' would be largest for sample S_2 followed by sample S_0 and sample S_1.

The level of differentiation for pairs of samples, measured as mean (s.e.) F_{ST} across replicates [computed using Equation (5.3) in Weir (1996)], were 0.0469 (0.00056), 0.0736 (0.00071) and 0.1120 (0.00086) for the sample pairs $S_0 - S_1$, $S_1 - S_2$, and $S_0 - S_2$, respectively. Mean (s.e.) F_{ST} among all three samples was 0.0788 (0.00054). The level of differentiation among time samples is comparable with moderate spacial differentiation generated by, e.g. an island model. Intuitively, it makes sense that time samples show differentiation—the more scaled time that passes between two sample points, the more lineages from the younger sample will coalesce before joining the older sample—and the sharing of variation between samples decrease with increasing scaled time between the samples.

4 CONCLUSIONS

Several studies point to the importance of using time-serial samples to answer biological questions, including studies focusing on viruses (Reid *et al.*, 2000; Rodrigo and Felsenstein, 1999) and studies of aDNA (Hofreiter, 2008; Miller *et al.*, 2008; Noonan *et al.*, 2006; Willerslev and Cooper, 2005). Using simulations, Depaulis *et al.* (2009) explored how commonly used population–genetic metrics were affected by serially sampled non-recombining sequence data (mimicking mitochondrial DNA). Depaulis *et al.* (2009) concluded that serially sampled data can have significant effect on commonly used metrics, potentially leading to erroneous conclusions if one ignores the time dimension of the data.

COMPASS is a flexible simulation tool that can be used to understand properties of time-serial data. The program can also be used for making inferences from time-serial data, for example, approaches that rely on simulations can use *COMPASS* to generate simulated time-serial population–genetic data. As the technical procedures for extracting and genotyping aDNA are steadily improving (Burger *et al.*, 2007; Ludwig *et al.*, 2009; Miller *et al.*, 2008; Svensson *et al.*, 2007), the challenges, and potential rewards, for analyzing these data will only increase, and *COMPASS* can be a helpful tool for analyzing these serially sampled data.

ACKNOWLEDGEMENTS

I thank A. Götherström for comments on the manuscript, and E. Svensson and P. Båtelsson for testing the program.

Funding: Swedish Research Council Formas and the Magn Bergvall foundation.

Conflict of Interest: none declared.

REFERENCES

Anderson,C.N.K. *et al.* (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, **21**, 1733–1734.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Burger,J. *et al.* (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA*, **104**, 3736–3741.

Depaulis,F. *et al.* (2009) Using classical population genetics tools with heterochroneous data: time matters! *PLoS ONE*, **4**, e5541.

Drummond,A.J. *et al.* (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Hofreiter,M. (2008) Long DNA sequences and large data sets: investigating the quaternary via ancient dna. *Quat. Sci. Rev.*, **27**, 2586–2592.

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

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Kimura,M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

Kingman,J.F.C. (1982) On the genealogy of large populations. *J. Appl. Prob.*, **19A**, 27–43.

Ludwig,A. *et al.* (2009) Coat color variation at the beginning of horse domestication. *Science*, **324**, 485.

Miller,W. *et al*. (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–390.

Noonan,J.P. *et al*. (2005) Paleontology: Genomic sequencing of pleistocene cave bears. *Science*, **309**, 597–600.

Noonan,J.P. *et al*. (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**, 1113–1118.

Nordborg,M. (2001) Coalescent theory, Ch. 7. In Balding,D.J. *et al*. (eds), *Handbook of Statistical Genetics*, Wiley, Chichester, pp. 179–212.

Rambaut,A. and Grassly,N.C. (1997) Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appli. Biosci.*, **13**, 235–238.

Reid,A. *et al*. (2000) Characterization of the 1918 "Spanish" influenza virus neuraminidase gene. *Proc. Natl Acad. Sci. USA*, **97**, 6785–6790.

Rodrigo, A. G. and Felsenstein,J. (1999) Coalescent approaches to HIV population genetics. In Crandall,K.A. (ed.) *The Evolution of HIV*. Johns Hopkins University Press, Baltimore, pp. 233–272.

Svensson,E.M. *et al*. (2007) Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Animal Genet.*, **38**, 378–383.

Weir,B.S. (1996) *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Willerslev,E. and Cooper,A. (2005) Ancient dna. *Proc. R. Soc. B Biol. Sci.*, **272**, 3–16.