BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 15 2004, pages 2485–2487
doi:10.1093/bioinformatics/bth264

SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history

Guillaume Laval* and Laurent Excoffier

Computational and Molecular Population Genetics, Zoological Institute, University of Bern, Baltzerstrasse 6, Bern CH-3012, Switzerland

Received on October 25, 2003; revised on February 19, 2004; accepted on April 3, 2004
Advance Access publication April 29, 2004

ABSTRACT

Summary: We present an extension of the program SIMCOAL, which allows for simulation of the genomic diversity of samples drawn from a set of populations with arbitrary patterns of migrations and complex demographic histories, including bottlenecks and various modes of demographic expansion. The main additions to the previous version include the possibility of arbitrary and heterogeneous recombination rates between adjacent loci and multiple coalescent events per generation, allowing for the simulation of very large samples and recombining genomic regions, together with the simulation of single nucleotide polymorphism data with frequency ascertainment bias.

Availability: http://cmpg.unibe.ch/software/simcoal2/

Contact: guillaume.laval@zoo.unibe.ch

Supplementary information: http://cmpg.unibe.ch/software/simcoal2/

Recent studies have shown the importance of past demographic events in explaining current patterns of linkage disequilibrium (LD) (Slatkin, 1994; Nordborg and Tavare, 2002; Stumpf and Goldstein, 2003). It seems therefore important to be able to simulate large genomic regions with complex patterns of recombination under realistic demographic models to get better insights about expected patterns of LD.

Here, we introduce a new version of a coalescent-based simulation software program (Excoffier *et al.*, 2000), SIMCOAL2, which includes a number of new features, such as the possibility to simulate datasets with arbitrary recombination rates between partially linked loci. Unlike fast algorithms based on continuous-time approximations (Hudson, 2002; Posada and Wiuf, 2003), SIMCOAL2 is based on a slower discrete-generation coalescent approach, which has several advantages: the possibility of multiple coalescent events per generation, as well as complex migration patterns and demographic scenarios (Leblois *et al.*, 2003). Note that the additional time needed for discrete-generation simulation is unlikely to be critical, since the analysis of the simulated data often takes much longer than their generation (e.g. Beaumont *et al.*, 2002). Moreover, the simulation of discrete generations while allowing multiple defining events (such as coalescences, migrations or recombinations) per generation is identical to a reversed Wright–Fisher model, which may be more realistic than a standard coalescent model in some situations. For instance, Wakeley and Takahashi (2003) have shown that the distribution of the expected number of segregating sites under the reversed Wright–Fisher model in a large sample (relative to the effective size) can markedly differ from the prediction of the standard coalescent. On the other hand, the description of how to define subdivided populations, migration matrices and complex demographic histories incorporating historical events, such as admixture events, population splits or fusions, or sudden changes in deme size can be found at http://cmpg.unibe.ch/software/simcoal/, and has not changed since the last version.

Simulation or arbitrary patterns of recombination: The main innovation of the current version is to allow the simulation of arbitrary patterns of recombination between loci, and thus the modelling of chromosomal segments with recombination rate heterogeneity or hot spots of recombination surveyed with single nucleotide polymorphisms (SNPs) or short tandem repeat (STR) (microsatellite) data, or a combination of different markers in arbitrary order. An example of an input file for the simulation of the genetic diversity of a chromosomal segment consisting of three blocks separated by two recombination hot spots, with 25 SNPs, 50 microsatellites and 25 SNPs in the three blocks respectively, can be downloaded at http://cmpg.unibe.ch/software/simcoal2/

Multiple coalescent, migration or recombination events per generation: The possibility of multiple migrations per generation was implemented in the previous version of the program, but multiple coalescent events and multiple recombinations per generation have now been introduced. Allowing multiple coalescent events is necessary under the discrete generation framework because the number of lineages to follow,

*To whom correspondence should be addressed.

say n, present at any generation in a given deme can increase dramatically under the effect of recombination. When the number of partially linked loci to simulate is large, this number n can indeed reach the same order of magnitude (or may be larger) than the deme size (N), which violates the underlying assumption of the coalescent that $n \ll N$. For instance, when we simulate 100 SNPs over 10 centimorgans (cM), the number of recombination events along a genealogy of 200 sampled genes can exceed 6000 in a diploid population of 1000 individuals, implying that the number of ancestral lineages segregating at a given past generation can be of the same magnitude than that of or even exceed the population size. Implementing multiple coalescent events per generation can also be useful in the absence of recombination, so as to simulate cases where $n \sim N$, or even $n > N$, where N is considered as an effective size (Wakeley and Takahashi, 2003), and to allow the simulation of multiple bifurcations or trifurcations or additional multifurcations of ancestral lineages.

Results checking: The consistency of the results has been checked against theoretical expectations in relatively simple cases compatible with the coalescent theory ($n \ll N$). In the simulation of the diversity of $n = 200$ genes typed at 11 loci distributed over 2.5 cM, the average T_{MRCA} over 10 000 simulations ranged between 1.97 and 1.98, compared with the continuous time coalescent expectation of $2(1-1/n) = 1.99$. The average number of recombination between two adjacent loci over 10 000 simulated coalescences of two lineages are equal to 0.87 for $R = 4Nr = 1$ and 5.67 for $R = 100$, compared with theoretical expectations [given by $6R/(R+6)$ in equation 31 of Simonsen and Churchill (1997)] of 0.86 and 5.66, respectively.

Simulated data types: By using SIMCOAL2, users can simulate restriction fragment length polymorphism (RFLP), STR, (microsatellites), DNA sequence, SNP markers or any combination of these types of markers over a chromosome segment, with arbitrary recombination rates ($0 \leq r \leq 0.5$) between adjacent loci. For SNP data, one can also define a minimum frequency for the minor allele such as to simulate one common source of ascertainment bias (see, e.g. Nielsen and Signorovitch, 2003). SIMCOAL2 now also outputs genetic data in the form of diploid multilocus genotypes obtained under the assumption of Hardy–Weinberg equilibrium. More information about the mutation models implemented for DNA sequence and STR are described elsewhere (Excoffier *et al*., 2000).

Computation time and output formats: The speed of simulation with recombination is much slower than without recombination and will increase with sample size, effective population size, the number of loci and the total recombination rate. Large amounts of recombinations between markers are easily handled by the program, such as to allow the modelling of large segments (in recombination units) of chromosomes. SIMCOAL2 needs 15 min to simulate 10 000 samples of 100 diploid individuals ($n = 200$ and $N = 1000$) typed at 250 SNPs over a segment of 250 kb (assuming 1 cM = 1 Mb, and thus with total recombination rate $R = 4Nr = 10$ for the whole chromosome segment), and 2.5 h to simulate 10 000 samples of 250 microsatellite loci over 2.5 Mb (total recombination rate $R = 100$), using a Pentium IV (2.8 GHz) under Linux. For every simulation, SIMCOAL2 generates ARLEQUIN (Schneider *et al*., 2000) output files that can be used to estimate several summary statistics from the data. SIMCOAL2 can also produce coalescent trees for each non-recombining segment, or for each locus in a recombining chromosomal segment in the NEXUS format (Swofford, 1999). These outputs can be used to visualize the shape of trees under different evolutionary scenarios, but these trees could also be used to compute some statistics, for instance, enabling the correction of SNP ascertainment bias due to small discovery panels (Nielsen and Signorovitch, 2003).

Note that while SIMCOAL2 currently provides a very flexible way to perform realistic simulations of recombining segments of our genome, it could become part of a powerful estimation procedure for demographic parameters, migration or recombination rates, if coupled with important sampling techniques allowing approximate Bayesian computation (see, e.g. Beaumont *et al*., 2002).

ACKNOWLEDGEMENTS

We are grateful to two anonymous reviewers, whose comments were helpful in better describing our program. This work was supported by a Swiss NSF grant no. 31-56755.99 and by an EU grant no. QLG2-CT-2001-00916.

REFERENCES

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

Excoffier,L., Novembre,J. and Schneider,S. (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Heredity*, **91**, 506–510.

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

Leblois,R., Estoup,A. and Rousset,F. (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a "Continuous" population under isolation by distance. *Mol. Biol. Evol.*, **20**, 491–502.

Nielsen,R. and Signorovitch,J. (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.*, **63**, 245–255.

Nordborg,M. and Tavare,S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet.*, **18**, 83–90.

Posada,D. and Wiuf,C. (2003) Simulating haplotype blocks in the human genome. *Bioinformatics*, **19**, 289–290.

Schneider,S., Roessli,D. and Excoffier,L. (2000) Arlequin: a software for population genetics data analysis. User manual ver. 2.000. Genetics and Biometry Lab, Department of Anthropology, University of Geneva, Geneva.

Simonsen,K.L. and Churchill,G.A. (1997) A Markov chain model of coalescence with recombination. *Theor. Popul. Biol.*, **52**, 43–59.

Slatkin,M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics*, **137**, 331–336.

Stumpf,M.P. and Goldstein,D.B. (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.*, **13**, 1–8.

Swofford,D.L. (1999) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Wakeley,J. and Takahashi,T. (2003) Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, **20**, 208–213.