Bioinformatics Advance Access published July 13, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw365
Advance Access Publication Date: 17 June 2016
Applications note

Genetics and population analysis

SELAM: simulation of epistasis and local adaptation during admixture with mate choice

Russell Corbett-Detig[1,*] **and Matt Jones**[1]

[1]Department of Integrative Biology, University of California Berkeley, Berkeley, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on February 22, 2016; revised on May 16, 2016; accepted on June 6, 2016

Abstract

Summary: SELAM is a forward time population genetic simulation program that provides a flexible framework for simulating admixture between any number of ancestral populations. The program can be used to simulate complex demographic and selection models, including dioecious or monoecious populations, autosomal or sex chromosomes, local adaptation, dominance, epistasis, and mate choice.

Availability and Implementation: The SELAM package (C++ source code, examples and manuals) is available via github at https://github.com/russcd/SELAM. This package is distributed under version 3 of the GNU general public license.

Contact: russcd@gmail.com

1 Introduction

There is a growing appreciation for the importance of admixture in a variety of evolutionary processes. Admixture has played a critical role in the evolutionary histories of non-African modern humans (Sankararaman *et al.* 2014), *Drosophila melanogaster* (Pool *et al.*, 2012), and numerous other species. Admixture has the potential to influence a wide variety of evolutionary processes including shaping patterns of neutral genetic variation, introducing adaptive (Pardo-Diaz *et al.*, 2012) and deleterious alleles (Sankararaman *et al.*, 2014), and producing novel ecologically and reproductively isolated lineages (Rieseberg, 2003). In order to study patterns of local ancestry across the genome and to develop novel theoretical predictions, an efficient and a flexible simulation framework that can track complete ancestry information and incorporate an array of selective and demographic functions is essential.

Previous studies have used coalescent approaches to study the distributions of haplotype structure resulting from relatively ancient admixture events (Sankararaman *et al.*, 2014). Though efficient, the coalescent has limited utility for studies that focus on natural selection, which is of central interest for many admixed populations (although see *e.g.* Ewing and Hermisson 2010). For this reason, forward in time simulations are appealing, and recently a number of efficient forward-time simulation programs have been released (Thornton, 2014; Messer, 2013; Hernandez, 2008). The majority of these software packages are designed to simulate the evolution and inheritance of mutations distributed along a chromosome. Although this approach is capable of simulating admixture (see Cui *et al.* 2015 for one implementation), by design these programs do not retain full ancestral information and many require substantive burn-in periods to achieve genetic equilibrium.

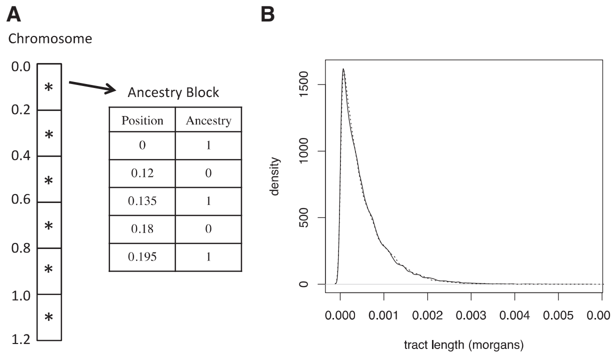
If the process and consequences of admixture are the primary interest, and not the evolutionary factors that generate the genetic diversity present in ancestral populations, an efficient simulation framework can be achieved by tracking local ancestry, rather than specific mutations, along chromosomes. Similar methods have been used previously in the context of studying relatively short evolutionary processes such as artificial selection (Kessner and Novembre, 2014; Haiminen *et al.*, 2013), and more generally for long term population genetic inference (Aberer and Stamatakis, 2013)). As the number of generations increases, this approach often performs poorly due to the rapid accumulation of recombination events that must be tracked in all descendent individuals. Here, we propose mitigating these issues by tracking local ancestry in genomic segments that can be accessed and inherited as a group. We implement this solution in SELAM, a forward-time population genetic simulator that is designed to efficiently handle both short and long-term admixture simulations while retaining complete local ancestry information. Furthermore, this program incorporates a broad spectrum of biologically realistic aspects of admixture which are, to our knowledge, not available within any existing simulation framework; among other things, SELAM can simulate local adaptation, separate sexes (with distinct migration rates, selection coefficients and sex chromosomes), pairwise epistasis, and mate choice.

2 Methods and features

SELAM is based on the extended Wright-Fisher model. Each generation consists of migration between subpopulations, mating based on individuals' fitness and mate choice loci, and offspring production. Once all the offspring are created, the parent generation dies. In the beginning of the simulation, all subpopulations are composed of proportions of individuals from any number of ancestral populations. Each subsequent generation consists of migration between subpopulations, as well as migration from non-admixed ancestral populations into admixed subpopulations. All demographic parameters can be modified during simulation. SELAM therefore supports simulations of complex evolutionary processes.

SELAM also supports a variety of selection models. Users may specify a single locus selection, pairwise epistatic selection, population-specific selection and loci influencing mate choice. The user must also supply SELAM with an input file, which specifies how many individuals to output from a given subpopulation at a given generation during the simulation. SELAM's ability to output individuals at multiple time points throughout the simulation makes this software applicable for studies where samples are distributed across many generations (e.g. ancient DNA, artificial selection and time series experiments). The SELAM package also includes a basic statistics program that enables users to obtain linkage disequilibria, ancestry frequencies and haplotype structure information across the genome.

Forward time simulations must account for all individuals within a simulation, which can pose a substantial computational burden, especially for longer simulations. SELAM's primary data structure, the *ancestry block*, stores all relevant genetic information for a genomic region. A chromosome is then represented as a vector of pointers to ancestry blocks (Fig. 1). Instead of copying all genetic information to the offspring chromosome, an ancestry block is retrieved only when it is needed (*i.e.* during recombination with another ancestry block), otherwise a pointer to the ancestry block is passed to the child chromosome without modification. The data structure within each ancestry block, a list of positions where ancestry switches, resembles that of other haplotype-based simulators (Haiminen *et al.*, 2013; Kessner and Novembre, 2014; Aberer and Stamatakis, 2013, Fig. 1), but also contains a list of selected and mate choice alleles that are present within the ancestry block. By varying the frequency of extinct ancestry block deletion as well as ancestry block length, both run time and memory usage can be optimized for a given set of simulation parameters and available computational resources (described in detail in the SELAM manual). These features enable SELAM to perform long simulations efficiently.



Position	Ancestry
0	1
0.12	0
0.135	1
0.18	0
0.195	1

Fig. 1. In SELAM, a chromosome is a vector of pointers to ancestry blocks, which contain lists of pairs of start positions and ancestry types in the subsequent tract (A). In comparisons between SELAM (solid) and a coalescent simulation (dotted), SELAM produced an indistinguishable and correct distribution of ancestry tracts ($P = 0.78$, Kolmogorov-Smirnov Test, B)

As with other haplotype-based simulations (Haiminen *et al.*, 2013; Kessner and Novembre, 2014; Aberer and Stamatakis, 2013), it is straightforward to place neutral mutations onto ancestral haplotypes once the simulation has finished. If the sample is small relative to the size of the simulated population, it is sufficient to record the local population ancestry because there is little risk of sampling individuals who share segments of their genome that are identical be descent. However, if the simulated population is relatively small, or the simulation time is very long, the probability of sampling individuals who share genetic ancestors is non-negligible. In this case, it may be desirable to track each unique ancestral haplotype individually. SELAM can support this as well—see, *e.g.*, example 6 in the SELAM package.

We have found that SELAM reproduces the expected ancestry tract length distribution via comparisons to coalescent-based simulations (Fig. 1, additional validations are reported in the SELAM manual). Furthermore, simulating the full human genome with a demographic history approximating that of modern Europeans in the time since admixture with Neanderthal populations—*i.e.* 2000 generations and including a maximum population size of 20,000 diploid individuals—took 51 minutes and had a peak memory usage of approximately 3.4 gigabytes on a personal computer using SELAM's default parameters (see example 3 of the SELAM package). SELAM is therefore capable of performing complex and biologically realistic simulations of long-term admixture using modest computational resources.

Acknowledgements

The authors thank Shelbi Russell, Tyler Linderoth, Rasmus Nielsen, and two anonymous reviewers for helpful discussion and comments on the manuscript.

Funding

RBC-D is supported by a UCB Chancellor's Postdoctoral Fellowship. This work was supported in part by a National Science Foundation Dissertation Defense Improvement Grant (Number 1311260) to RBC-D.

Conflict of Interest: none declared.

References

Aberer,A.J. and Stamatakis,A. (2013) Rapid forward-in-time simulation at the chromosome and genome level. *BMC Bioinformatics*, **14**, 216.

Cui,R. *et al.* (2015) Admix'em: a flexible framework for forward-time simulations of hybrid populations with selection and mate choice. *Bioinformatics*, **32**, 1103–1105.

Ewing,G. and Hermisson,J. (2010) MSMS: a coalescent simulation program including in recombintion, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.

Haiminen,N. *et al.* (2013) Efficient in silico Chromosomal Representation of Populations via Indexing Ancestral Genomes. *Algorithms*, **6**, 430–441.

Hernandez,R.D. (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

Kessner,D. and Novembre,J. (2014) forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics*, **30**, 576–577.

Messer,P.W. (2013) SLiM: simulating evolution with selection and linkage. *Genetics*, **194**, 1037–1039.

Pardo-Diaz,C. *et al*. (2012) Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *PLoS Genet*, **8**, e1002752–e1002713.

Pool,J.E. *et al*. (2012) Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLoS Genet*, **8**, e1003080–e1003024.

Rieseberg,L.H. (2003) Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science*, **301**, 1211–1216.

Sankararaman, S. *et al*. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**, 354–357.

Thornton,K.R. (2014) A C++ template library for efficient forward-time population genetic simulation of large populations. *Genetics*, **198**, 157–166.