Efficient simulation of identity-by-descent and ancestry in large datasets

Georgia Tsambos

University of Melbourne, Australia

Georgia Tsambos (1, 2), Peter Ralph (3), Jerome Kelleher (4), Stephen Leslie (1, 2, 5), Damjan Vukcevic (1, 2). (1) School of Mathematics and Statistics, University of Melbourne, Australia (2) Melbourne Integrative Genomics, University of Melbourne, Australia, (3) Department of Mathematics, University of Oregon, United States, (4) Big Data Institute, University of Oxford, United Kingdom, (5) School of Biosciences, University of Melbourne, Australia.

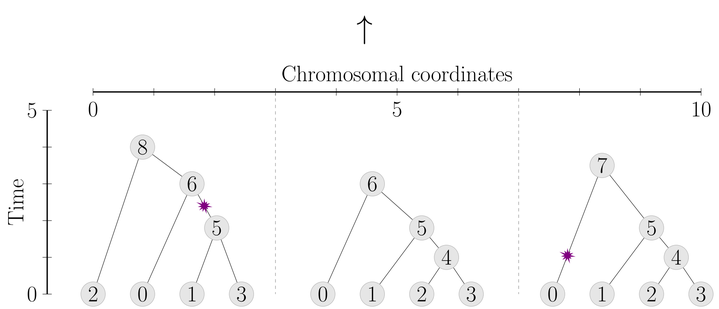
Presenting author: gtsambos (at) student.unimelb.edu.au

0. Introduction

To assess the performance of methods in population genetics, we often wish to simulate realistic genetic datasets while retaining detailed information about the history of the simulated genomes. This poster briefly describes how we can efficiently simulate genetic information with full information about the common ancestry of particular genomic segments, as well as information about the populations that these segments have been inherited from. Although the software is still under development in `tskit` [1], our progress to date suggests that our methods are scaleable and fast enough to be useful in high-powered studies of subtle demographic questions.

1. The data structure: tree sequences

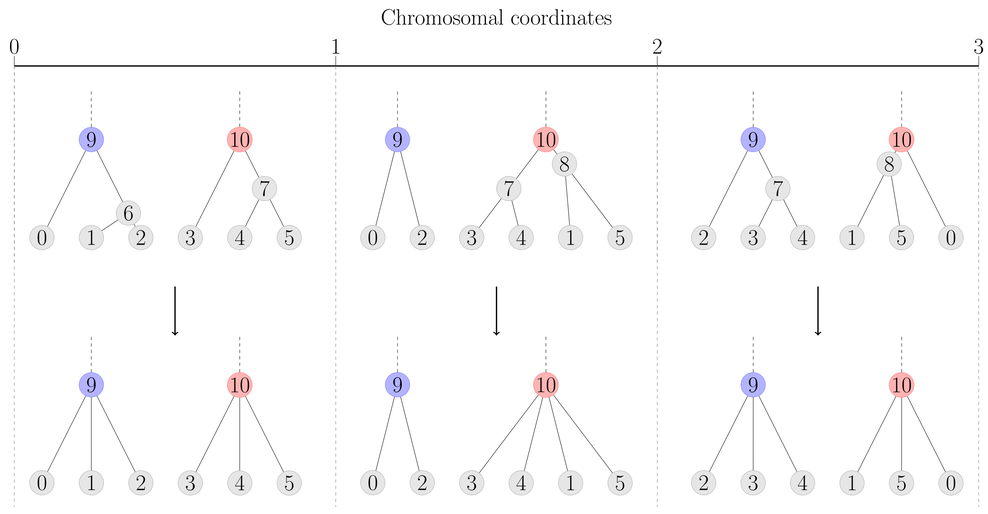
```
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGTAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTA...
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGTAATAGCGTA...
```



Nodes		Edges				Mutations		
ID	time	left	right	parent	child	location	time	node
0	0.0	3.0	10.0	4	2	1.5	2.3	5
1	0.0	3.0	10.0	4	3	8.9	1.2	0
2	0.0	0.0	10.0	5	1			
3	0.0	0.0	3.0	5	3			
4	1.1	3.0	10.0	5	4			
5	1.8	0.0	7.0	6	0			
6	3.0	0.0	7.0	6	5			
7	3.5	7.0	10.0	7	0			
8	4.0	7.0	10.0	7	5			
		0.0	3.0	8	2			
		0.0	3.0	8	6			

The **tree sequence** data structure [1,2] encodes a complete genealogy for a sample of chromosomes in a succinct set of tables. Compared with traditional sequence-based formats, tree sequences are *compact*, *fast* to process and *informative* of the history of the sample.

2. IBD and ancestry in tree sequences

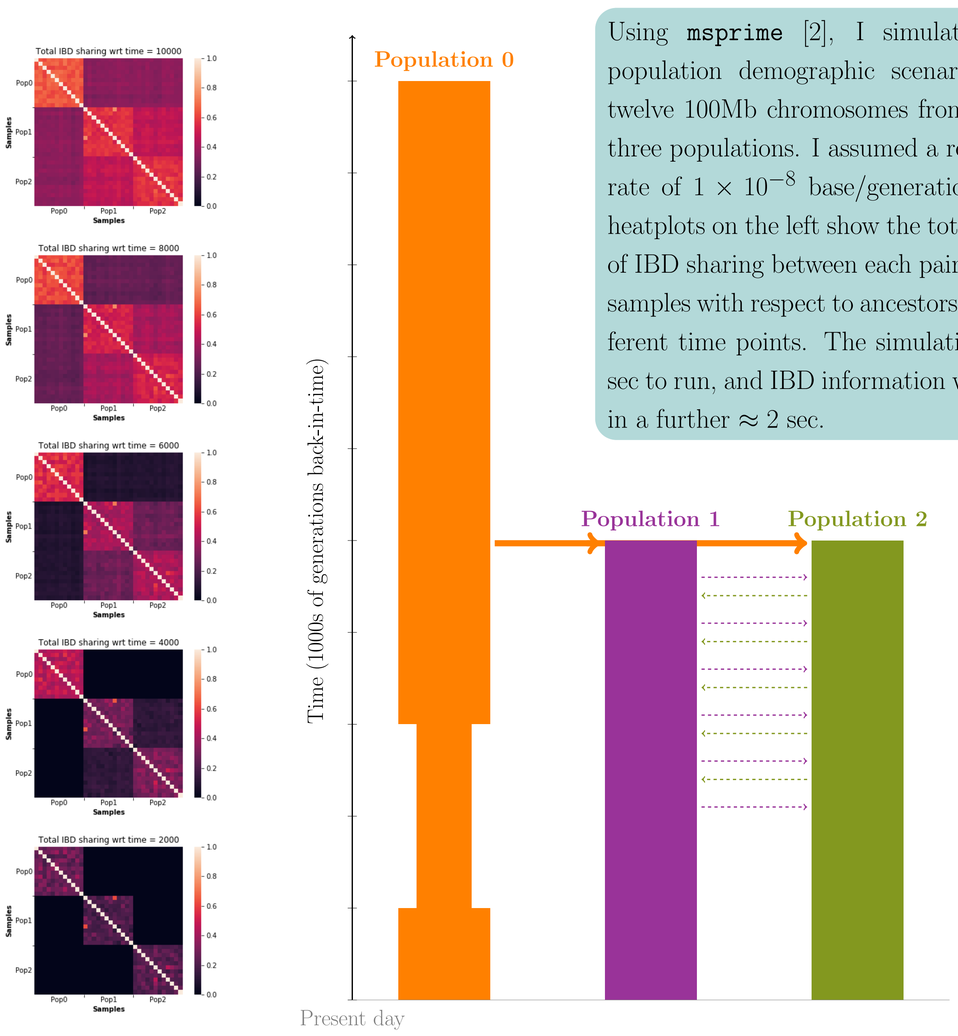


Questions about inheritance and ancestry can be reframed as questions about the underlying tree sequence that represents the data:

- *Identity-by-descent: which samples share a common ancestor? How recently? Over which genomic interval?*
- *Local ancestry: which ancestors have which population labels? Which samples descend from them? Over which genomic interval?*

Extracting this information from large datasets requires efficient algorithms that account for the correlations in genealogical structure between samples, and across chromosomes.

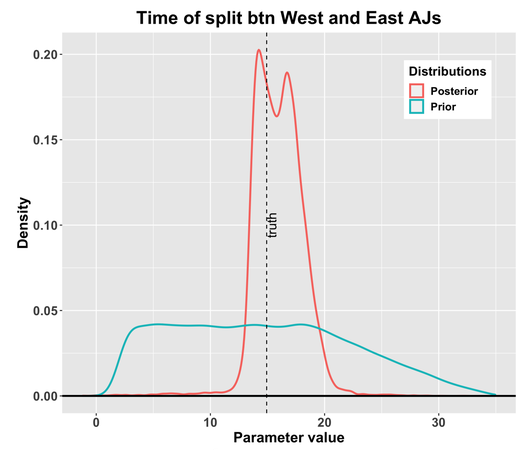
3. Example: total IBD sharing over time



Using `msprime` [2], I simulated a three-population demographic scenario, sampling twelve 100Mb chromosomes from each of the three populations. I assumed a recombination rate of 1×10^{-8} base/generation. The five heatplots on the left show the total proportion of IBD sharing between each pair of simulated samples with respect to ancestors from five different time points. The simulation took ≈ 6 sec to run, and IBD information was extracted in a further ≈ 2 sec.

4. Application: demography inference

Model inference	
Posterior probability for correct model	$96.33\% \pm 2.68\%$
Votes for correct demography	$93.36\% \pm 4.69\%$
Prior error rate	0.0993 ± 0.0008



To explore the power of our method, we attempted to recreate some of the important findings from a recent study of Ashkenazi Jewish (AJ) demographic history [3]. This work provided evidence for a recent divergence event between Eastern and Western communities of AJ people. It was estimated that this happened about 15 generations ago.

Closely following [3], we performed 100 000 simulations of various demographic scenarios, and 20 simulations under the parameter values inferred by [3], which we treated as proxies for the truth. For each simulation, we calculated moments of IBD segment lengths at multiple time points. We used the `abcrf` package [4] to infer the most plausible demographic scenario, and the `abc` package [5] with neural-net regression to estimate parameter values.

All simulations and analyses ran on a standard desktop in < 12 hours.

5. Acknowledgements, references and further information

GT is funded by the Helen Freeman scholarship, the Maurice Belz Fund and the Australian Government's Research Training Scheme.

[1] `https://tskit.readthedocs.io/en/latest/`

[2] Kelleher, J., et al. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology, 12(5).

[3] Gladstein, A and Hammer, M. (2019). Substructured Population Growth in the Ashkenazi Jews Inferred with Approximate Bayesian Computation. Molecular Biology and Evolution, 36(6), 1162-1171.

[4] Raynal, L., et al. (2019). ABC random forests for Bayesian parameter inference. Bioinformatics, 35(10), 1720 - 1728.

[5] Csillery, K., et al. (2012). Abc: An R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution, 3(3), 475 - 479.

Come say hi!