# 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

Genetic sequence data is BIG and REPETITIVE:

```
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTA...
```

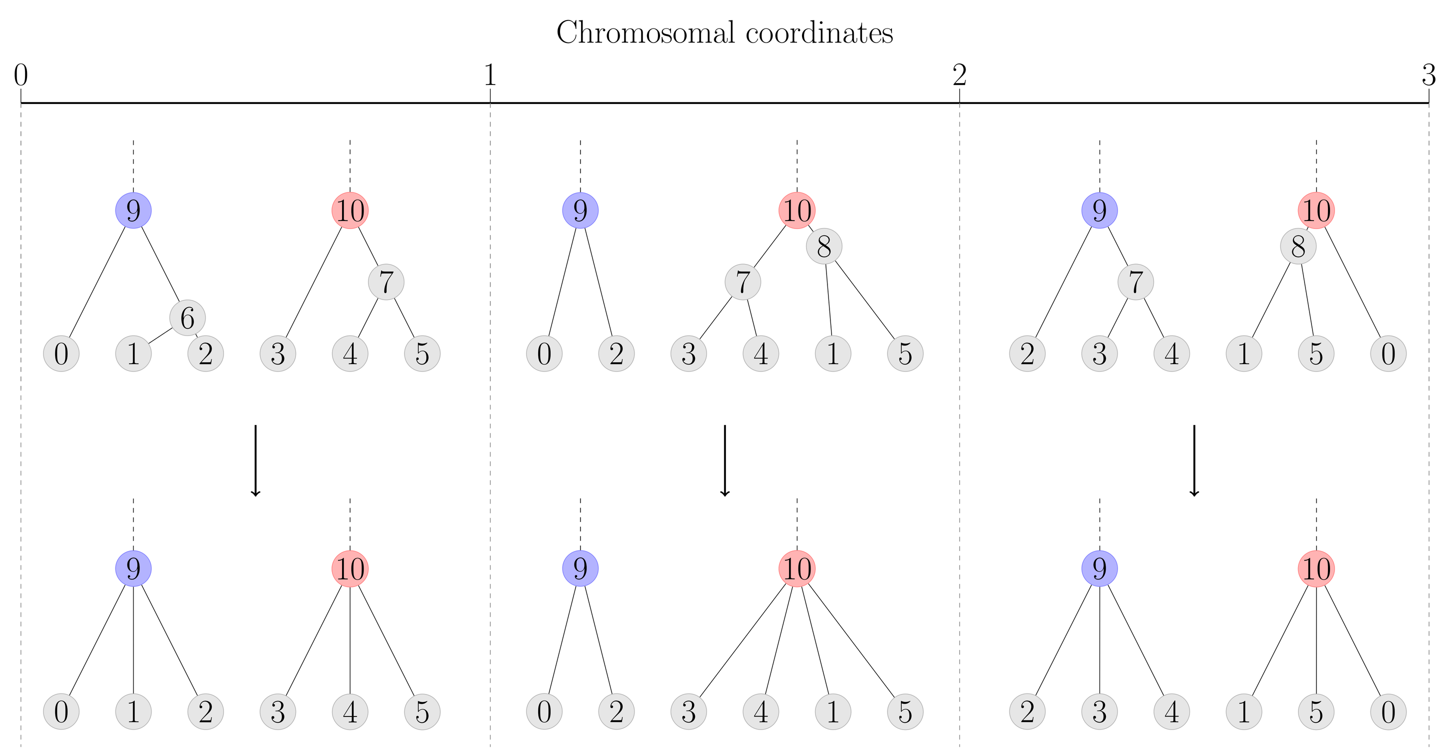
$\leftarrow 5 \times 10^7$ bases for small human chromosome $\rightarrow$

However, common haplotypes in a sample are often simply a consequence of some common history. So if we know this history (as we always do in simulations!), storing it directly is often more convenient and efficient than storing the raw haplotypes. This is the key idea behind the tree sequence data structure [1,2], which encodes a complete genealogy for a sample of chromosomes in a succinct set of tables.

Tree sequences offer a few benefits to population geneticists compared with traditional sequence-based file formats:

- They can store large simulated datasets extremely compactly.
- As they hold rich detail about the history of the sample, many important processes can be observed directly from the tree structure.
- They can be queried and modified extremely quickly.

## 2. IBD and ancestry in tree sequences

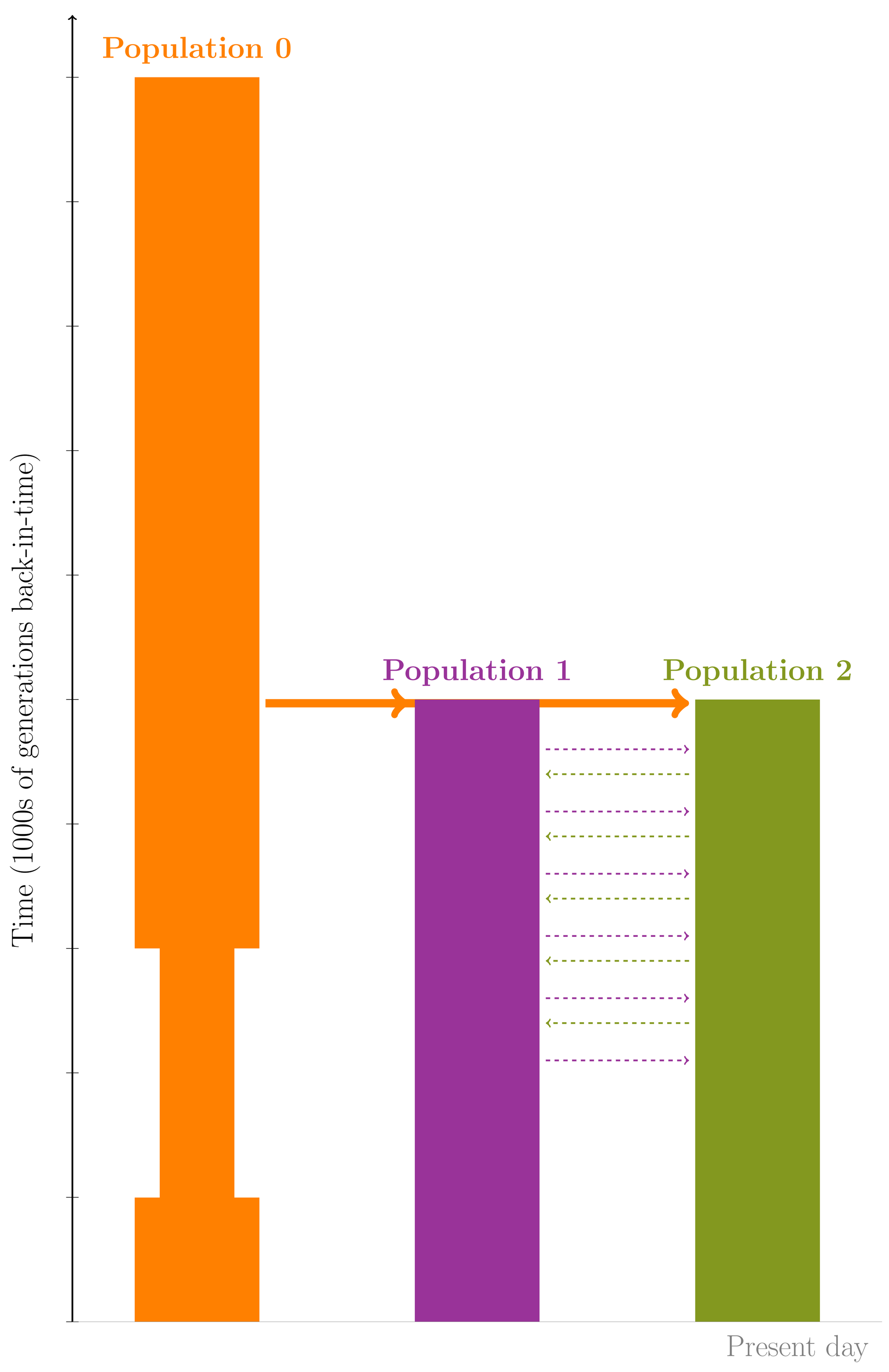


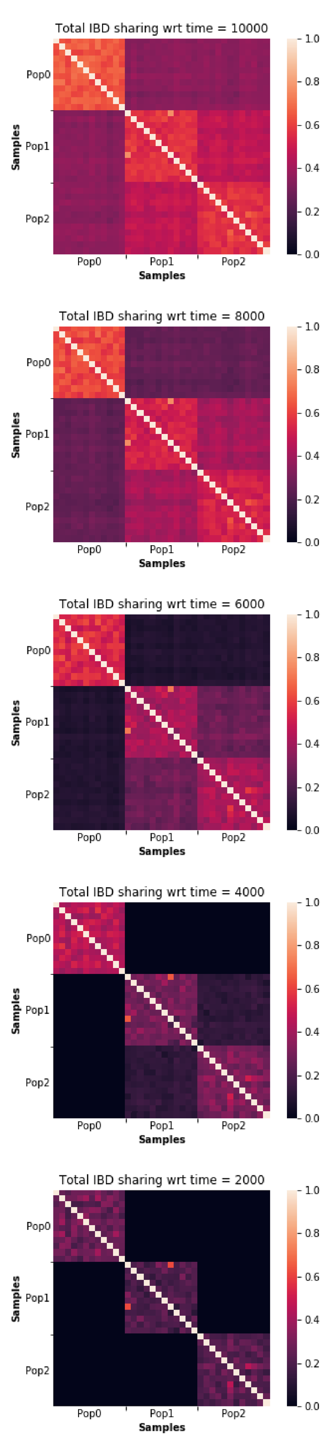

Questions about inheritance and ancestry can be reframed as questions about the underlying tree sequence that represent our datasets:

- *Identity-by-descent: which samples share a common ancestor where?*
- *Local ancestry: which ancestors have what population labels, and which samples descend from them?*

Extracting this information efficiently is challenging due to correlations in genealogical structure between samples, and across chromosomes; in an upcoming paper, we will describe algorithms that allow us to do this.

## 3. Example: total IBD sharing over time




## 4. Application: demography inference

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 50 000 simulations of various demographic scenarios. 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.

### Results: inference of demographic scenario

| Posterior probability | Votes for correct scenario | Prior error rate |
|---|---|---|
| $93.36\% \pm 4.69\%$ | $0.9633 \pm 0.0268$ | $0.0993 \pm 0.0008$ |

### Results: inference of divergence time

| Mode | True value | Our 95% HPDI | 95% HPDI in [2] |
|---|---|---|---|
| 13.59 | 14.93 | (11.22, 19.90) | (2.00, 29.00) |

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!