Efficient simulation of introgression, admixture and local ancestry

Georgia Tsambos

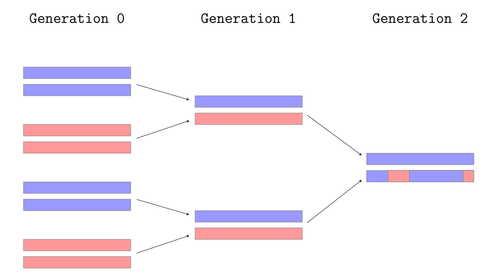
University of Melbourne, Australia

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 ancestral populations that particular genomic segments have been inherited from, often called the local ancestry of the sample. In all pictures here, we represent ancestries with colours.



Section 1 introduces the tree sequence [1], a data structure that is capable of encoding a complete genealogy for a sample of chromosomes at each chromosomal location. Section 2 shows how local ancestry information can be stored and extracted from tree sequences. Section 3 outlines a method for simulating such information using recent advances implemented in the tree sequence simulation software msprime [1] and SLiM [2]. Section 4 demonstrates the performance of this method. Section 5 provides further information for the interested viewer.

1. Tree sequences

Genetic sequence data is BIG and REPETITIVE:

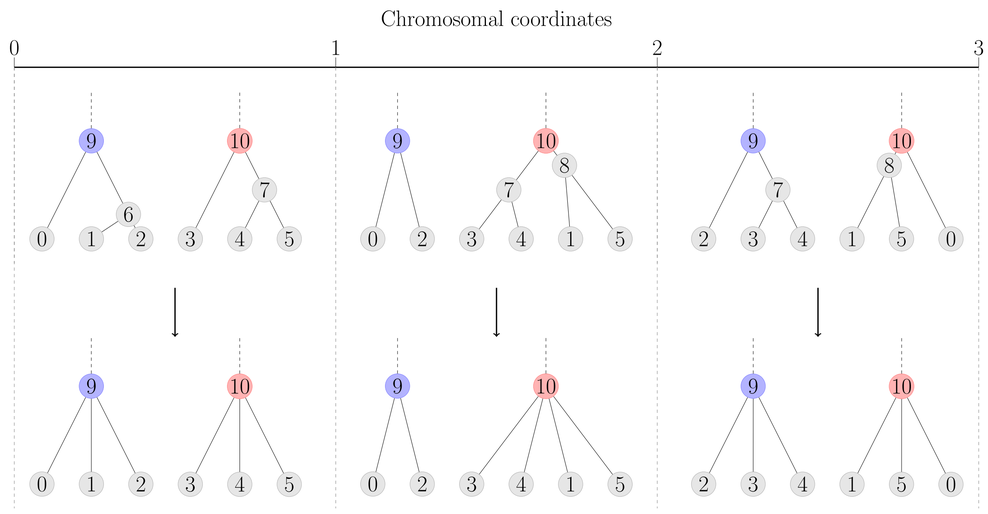
```
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTA...
...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], which encodes a complete genealogy for a sample of chromosomes at each chromosomal location. Tree sequences offer a few benefits to population geneticists compared with traditional sequence-based file formats:

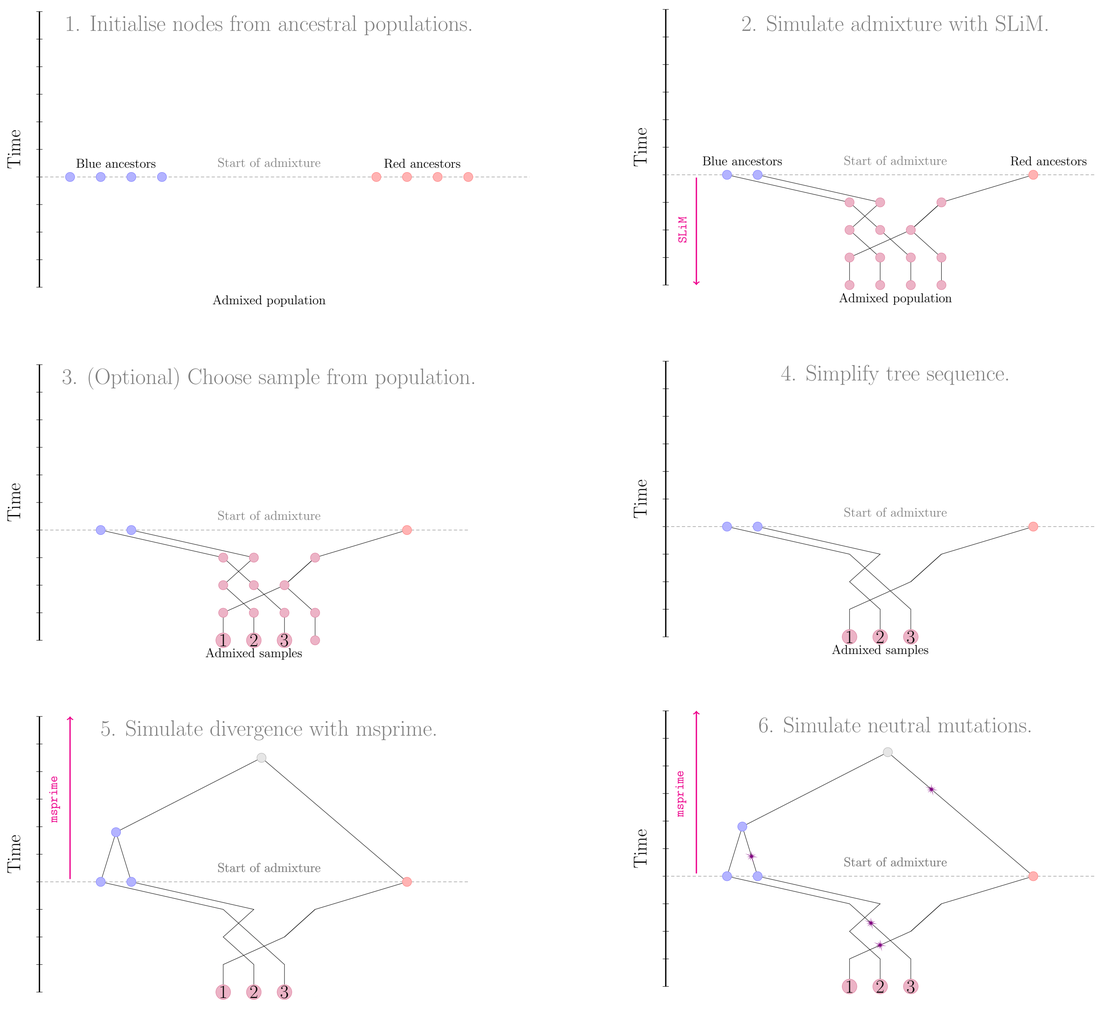
- They can store large simulated datasets extremely compactly.
- As they hold rich detail about the history of the sample, many important processes can be observed directly from the tree structure.
- They can be queried and modified extremely quickly.

2. Local ancestry in tree sequences



By assigning population labels to the nodes that correspond to ancestors of the sample, tree sequences can store the sample's local ancestry. The branch joining a sample node to an ancestral node shows its ancestry. Extracting this information efficiently is challenging due to correlations in genealogical structure between samples, and across chromosomes; in an upcoming paper, we will describe an algorithm that enables us to do this.

3. Method outline



4. Method performance

	Missing data	Run time	File size	Selection
`msprime`	4.0%	6 sec	9 Mb	No
`msprime` + full ARG	0.0%	53 sec	1700 Mb	No
`SLiM`	0.0%	$>$ 1 hr	41 Mb	Yes
`slime`	0.0%	86 sec	39 Mb	Yes

To illustrate the power of our method, `slime` [3], we simulated a toy demographic scenario inspired by the history of Neanderthal introgression into the Eurasian population. We simulated 200 chromosomes of length 50 Mb from 100 present-day Eurasian individuals, assuming a 2% introgression of Neanderthals into Eurasians 2500 generations ago. For simplicity, we assumed a constant effective population sizes of 5000 individuals, a uniform recombination rate of 1×10^{-8} bp per generation, a uniform mutation rate of 1×10^{-8} bp per generation and neutral variation. However, note that all of these methods can deal with more complexity than this. In particular, both `slime` and `SLiM` are capable of simulating under selection.

Although still under development, `slime` appears to outperform existing tree sequence simulation softwares on various metrics by orders of magnitude. As a principled fusion of `SLiM` and `msprime`, `slime` will allow users to track local ancestry in large simulations under realistically complex demographic scenarios, and with minimal computational overhead.

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

[2] Galloway, J., et al. (2018). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Molecular Ecology Resources, (November 2018), 552–566.

[3] [Link to public `slime` page]

More info

Come say hi!