Efficient simulation of introgression, admixture and local ancestry

Georgia Tsambos

University of Melbourne, Australia

Quantitative Genomics, 10 June 2019

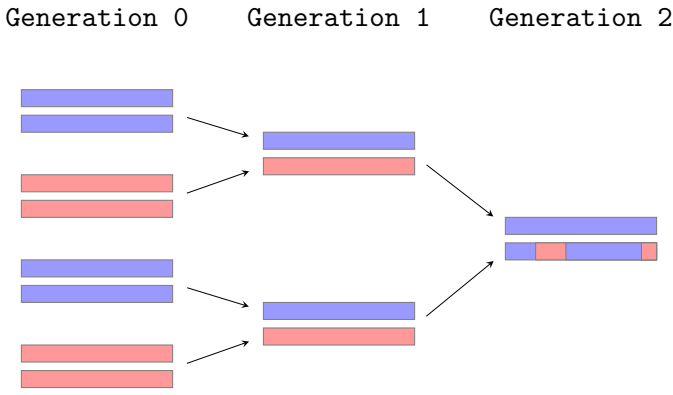
Talk outline

1. Introduction to admixture and local ancestry
2. Tree sequences
3. Simulating local ancestry with tree sequences: existing methods
4. Simulating local ancestry with tree sequences: new method

Intro to admixture and local ancestry

What is admixture?

- An organism has ancestry with a given population if they have inherited some genetic material from ancestors who belonged to that population.
- Any organism with > 1 ancestry is admixed.
- Introgression is admixture between different species.


Generation 0
Generation 1
Generation 2

Reporting ancestry

Global ancestry		Local ancestry
60%	40%	
10%	90%	
70%	30%	

My PhD work is about simulating and inferring local ancestry.

Being able to simulate genetic ancestry is useful!

- **Benchmarking and evaluating method performance**
  To assess the accuracy of methods that infer ancestry, we need test datasets in which we know the true local ancestry.

- **Model training**
  Some methods for ancestry inference are trained on simulated data.

- **Exploration**
  Simulations allow us to explore the influence of various historical scenarios on observed patterns of genetic variation and inheritance.

Tree sequences: the data format

Context: genetic data is BIG and REPETITIVE

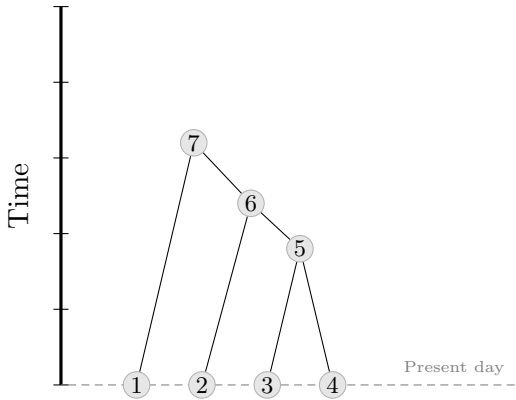
```
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTA...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

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

Repeated haplotypes are often just a consequence of shared history.

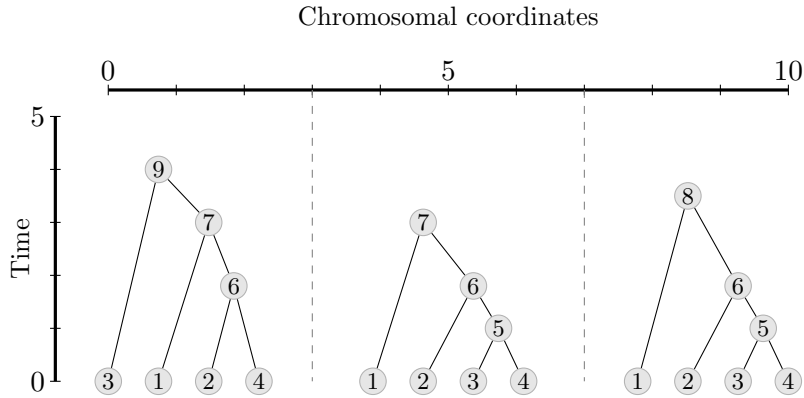
Q. Can we use this history to represent DNA sequences more compactly?

Trees show genealogy at a single locus



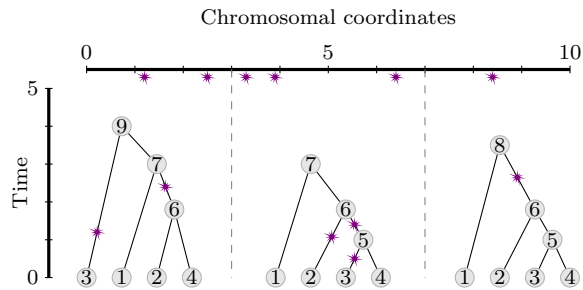
- Nodes = haplotypes
- Edge heights = time

Tree sequences show genealogy over an interval of loci

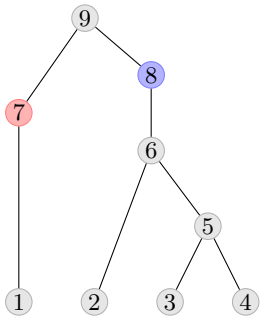


Tree sequences can encode haplotypes

Node 1	0	0	0	0	0	0
Node 2	1	0	1	0	0	1
Node 3	0	1	0	1	1	1
Node 4	1	0	0	0	1	1



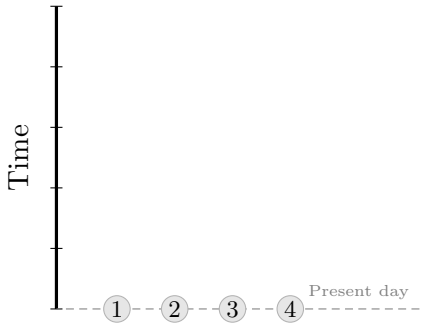
Tree sequences can hold info on local ancestry



Nodes may be assigned to populations.
The branch joining a sample node to an ancestral node shows the sample's ancestry.

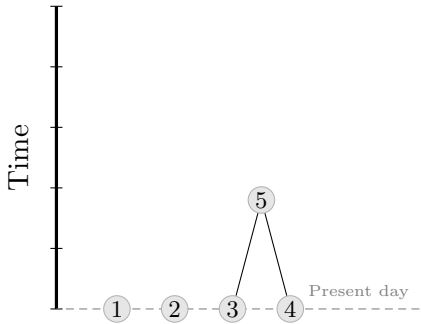
Simulating admixture: existing methods

`msprime` simulates a sample backwards-in-time



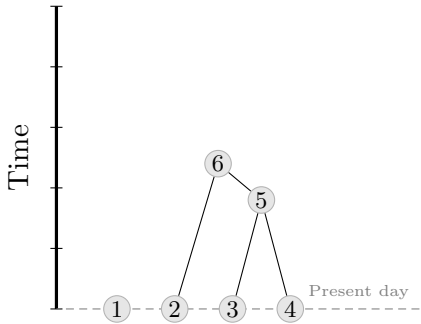
Simulates tree sequences under the coalescent model.

`msprime` simulates a sample backwards-in-time



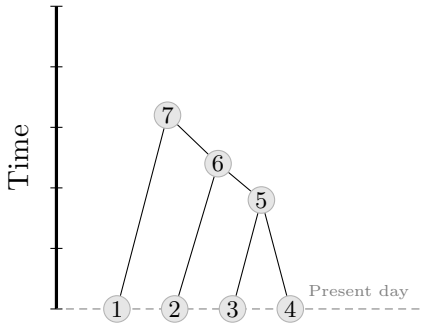
Simulates tree sequences under the coalescent model.

`msprime` simulates a sample backwards-in-time



Simulates tree sequences under the coalescent model.

`msprime` simulates a sample backwards-in-time



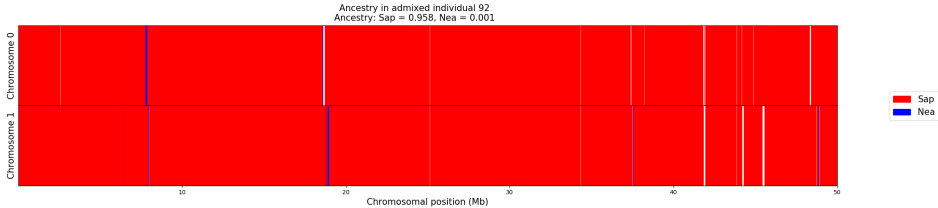
Simulates tree sequences under the coalescent model.

A simplified toy example: Neanderthal introgression

Generations	**Event**
$\approx 240\,000$	Common ancestor of all modern Eurasians and Neanderthals at all loci.
$20\,000$	Divergence of Eurasians and Neanderthals.
$2\,500$	2% introgression of Neanderthals into Eurasians.
0	Samples from 100 Eurasian individuals obtained.

Chromosome of $50\,000\,000$ base pairs, constant effective population sizes of 5000 individuals, uniform recombination rate 1×10^{-8} bp/generation, uniform mutation rate 1×10^{-8} bp/generation, all variation is neutral.

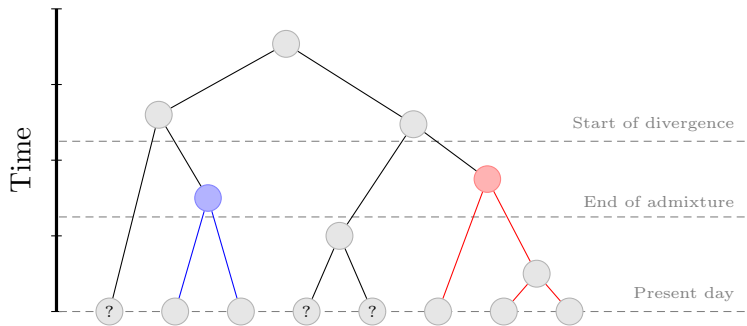
`msprime` results: missing ancestry data



Global ancestry averaged over all of the simulated samples was

- 96.0% Sapiens
- $< 0.05\%$ Neanderthal
- 4.0% unassigned.

Missing ancestry is due to incomplete lineage sorting



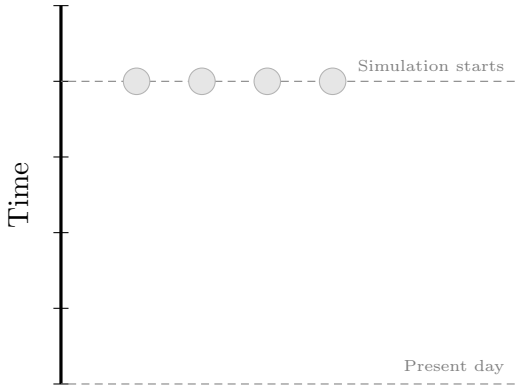
Some samples do not have simulated ancestors in the given populations.

`msprime` performance

	Missing data	Run time	File size	Realism
default	4.0%	6 sec	9 Mb	✗
+ all ancestors	0.0%	53 sec	1700 Mb	✗

Restricted by the limitations of the coalescent model: no selection, random mating, small sample sizes, no more than 1 mutation at any location ...

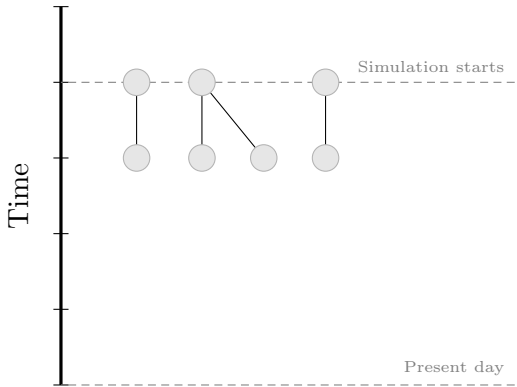
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

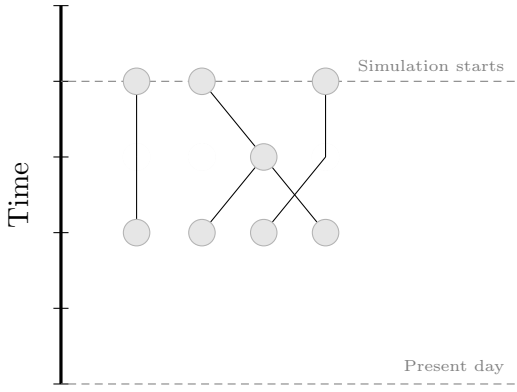
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

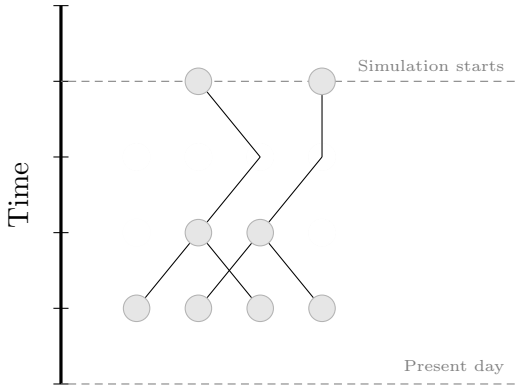
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

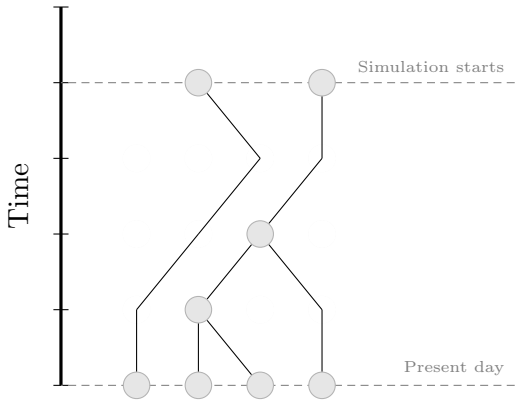
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

SLiM simulates an entire population forward-in-time



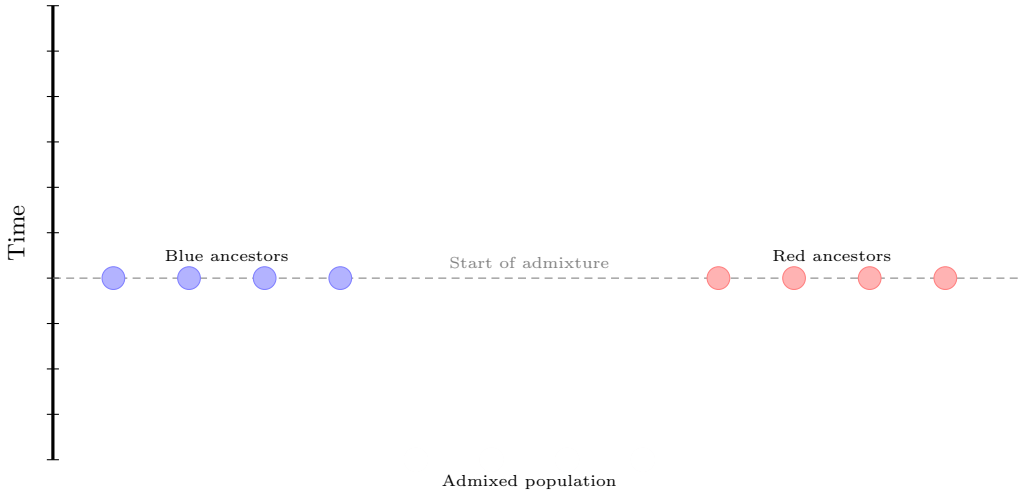
Alternates between

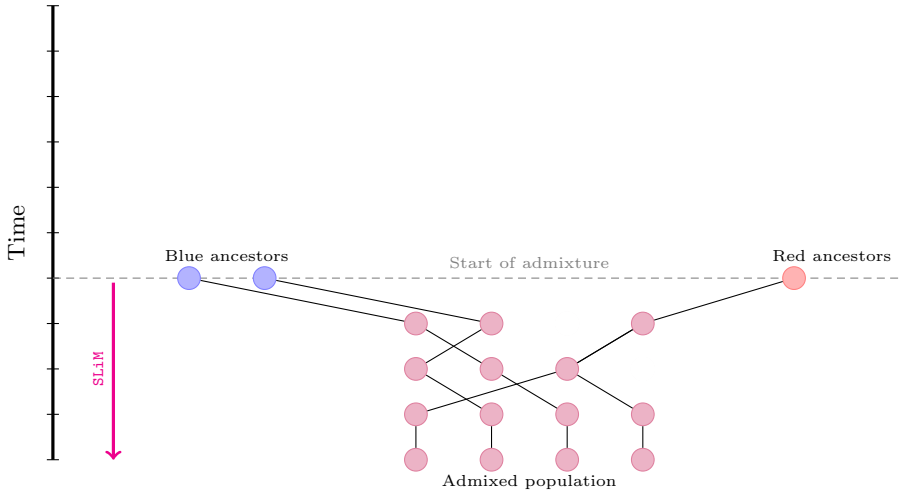
1. forward simulation
2. pruning of irrelevant history.

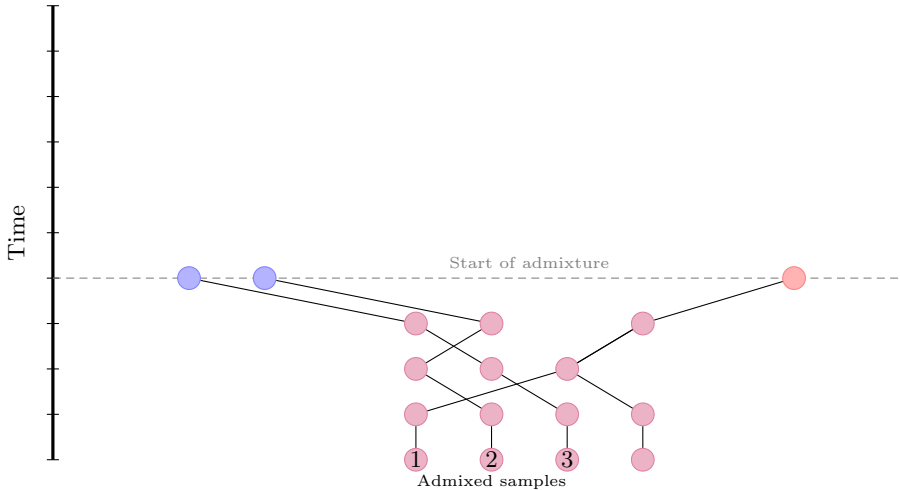
SLiM performance

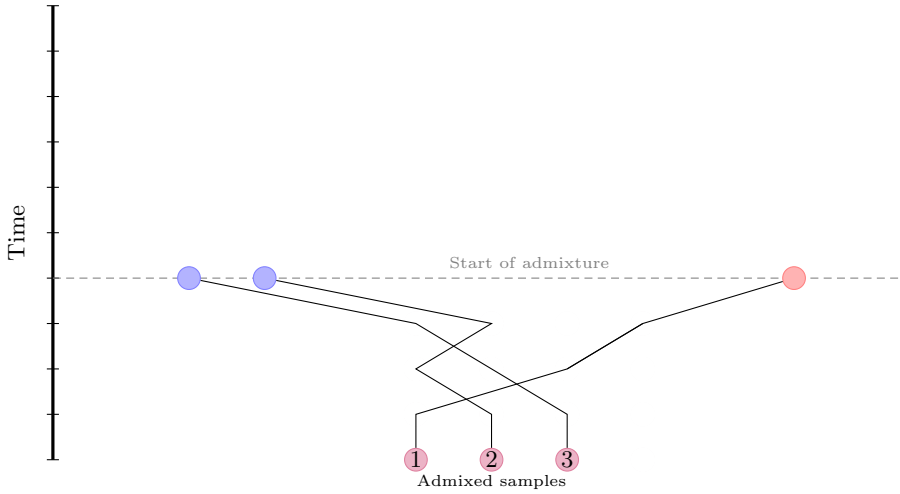
	Missing data	Run time	File size	Realism
msprime	0.0%	53 sec	1700 Mb	✗
SLiM	0.0%	> 1 hr	41 Mb	✓

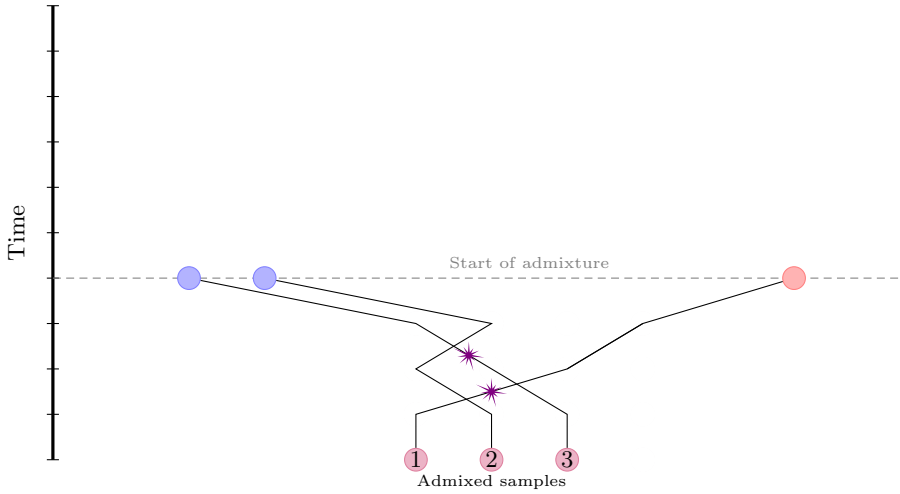
Simulating tree sequences: new method

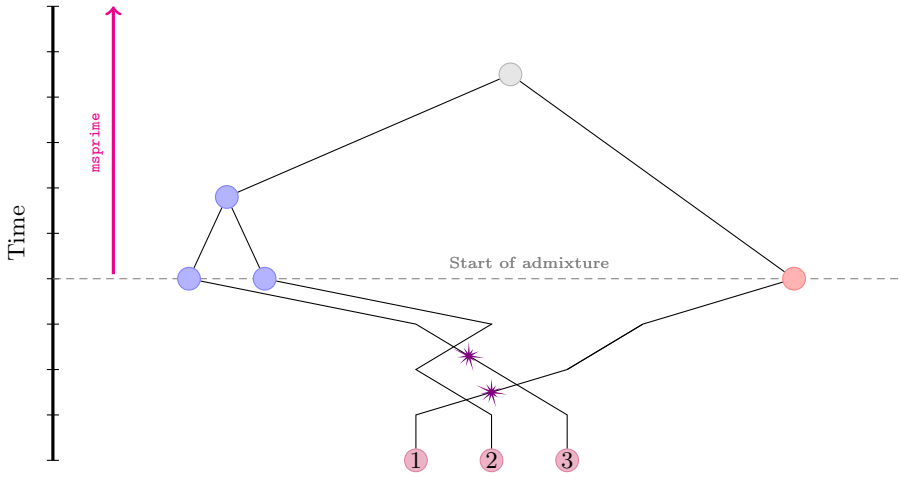

Time
Blue ancestors
Start of admixture
Red ancestors
Admixed population

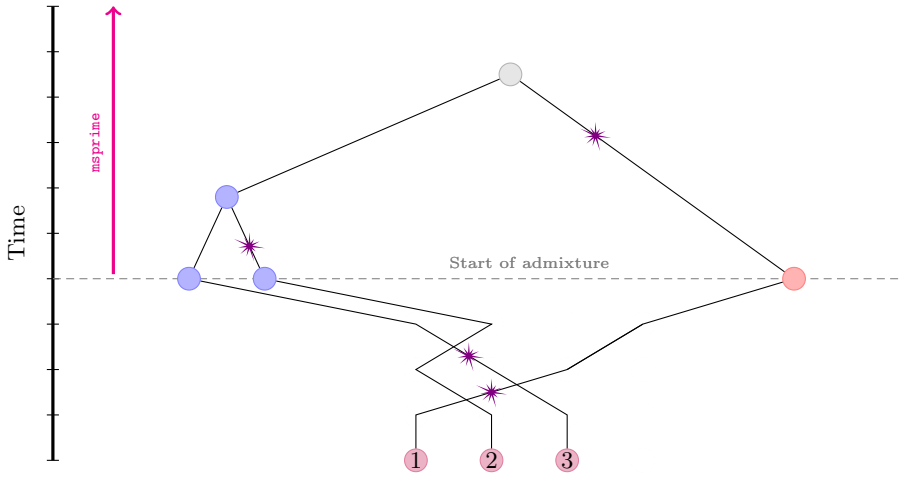

Time
Blue ancestors
Start of admixture
Red ancestors
SLiM
Admixed population


Time
Start of admixture
1
2
3
Admixed samples


Time
Start of admixture
1
2
3
Admixed samples


Time
Start of admixture
1
2
3
Admixed samples


Time
msprime
Start of admixture
1
2
3


Time
msprime
Start of admixture
1
2
3

New method performance

	Missing data	Run time	File size	Realism
`msprime`	0.0%	53 sec	1700 Mb	✗
`SLiM`	0.0%	> 1 hr	41 Mb	✓
Georgia	0.0%	86 sec	39 Mb	✓

Summary

- It is useful to keep track of local ancestry in genetic simulations.
- Standard tree sequence simulators are able to do this with the help of efficient ancestry-extraction algorithms.
- A new simulation method can do this quickly while also allowing the user to model complex admixture scenarios.

Thanks to...

My supervisors	Damjan Vukcevic (University of Melbourne) Stephen Leslie (University of Melbourne)
My collaborators	Peter Ralph (University of Oregon) Jerome Kelleher (BDI, University of Oxford)
Sources of $	Helen Freeman scholarship, UniMelb Maurice Belz Fund, UniMelb Research Training Scheme, Australian Government

Special thanks to Jerome and the University of Oxford for hosting me in the UK this year.

References

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

`SLiM`:
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.