Efficient simulation of introgression, admixture and local ancestry

Georgia Tsambos

Melbourne Integrative Genomics, Australia

Quantitative Genomics, 10 June 2019

Talk outline

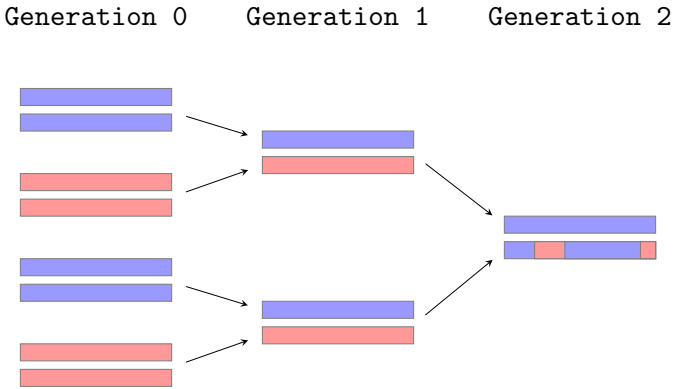
1. Introduction to admixture and local ancestry
2. Tree sequences
3. Simulating local ancestry with tree sequences - existing ways
4. Simulating local ancestry with tree sequences - new way

Intro to admixture and local ancestry

What's admixture?

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

What is admixture?



Reporting ancestry

Global ancestry		Local ancestry
60%	40%	G A T T T G C C A A
10%	90%	A A C C T G T C G A
70%	30%	G A T C T A T T G G

My PhD work is about simulating and inferring local ancestry.

Understanding local ancestry is important in studies of demography and history, medicine and genetic pipelines.

Storing genomes with history using tree sequences

Context - genetic data is BIG and REPETITIVE

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

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

Storing $n = 1 \times 10^5$ chromosomes in a compressed VCF requires ≈ 50 GB.

Shared haplotypes are often due to shared history

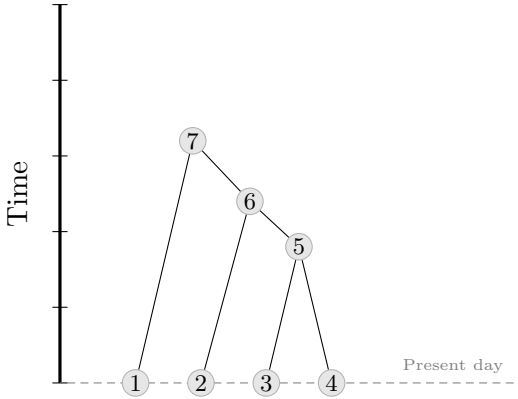
```
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTA...
...GTAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTA...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

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

Tree sequences are the future!

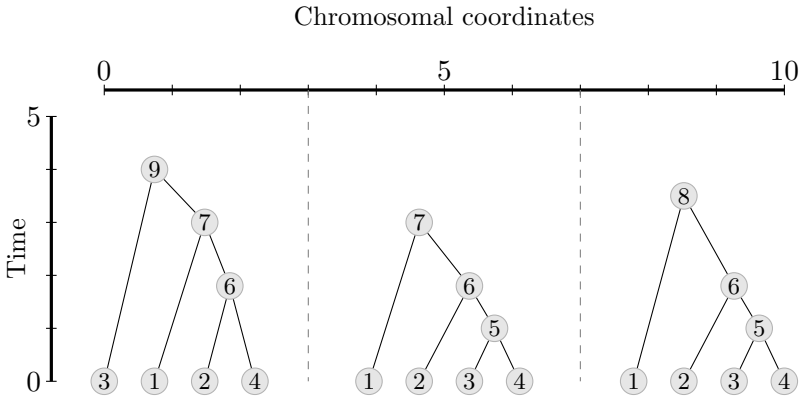
- Tree sequences contain rich information about the history of a sample, not just the genotypes.
- We can simulate this data structure using well-established software (`msprime`, `SLiM`). (With some caveats...)
- We can infer this data structure with some success (`tsinfer`).

Trees show genealogy at a single allele



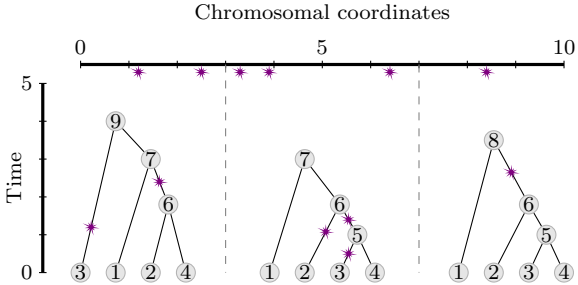
- Nodes = haplotypes
- Edge heights = time

Tree sequences show genealogy over an interval of alleles

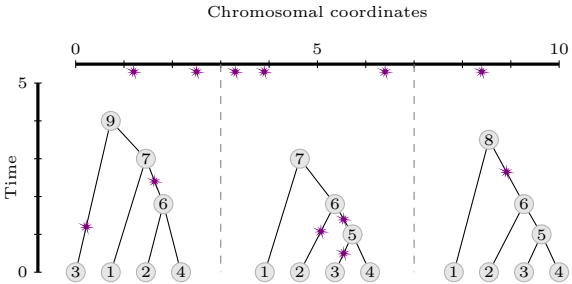


Tree sequences can encode haplotypes

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



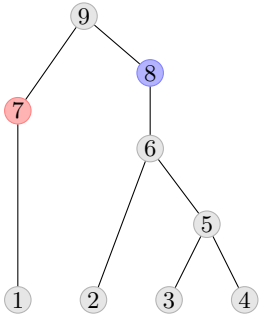
Tree sequences can be stored in tables



Nodes		Edges			
id	time	left	right	parent	child
1	0.0	0.0	7.0	5	3
2	0.0	0.0	7.0	5	4
3	0.0	0.0	10.0	6	2
4	0.0	0.0	3.0	6	4
5	1.1	3.0	10.0	6	5
6	1.8	0.0	7.0	7	1
7	3.0	0.0	7.0	7	6
8	3.5	1.0	10.0	8	1
9	4.0	7.0	10.0	8	6
		0.0	3.0	9	3
		0.0	3.0	9	7

Mutations		
location	time	nearest node
1.2	2.5	6
2.5	1.2	3
3.3	1.3	2
3.9	0.4	3
6.4	1.4	5
8.4	2.7	6

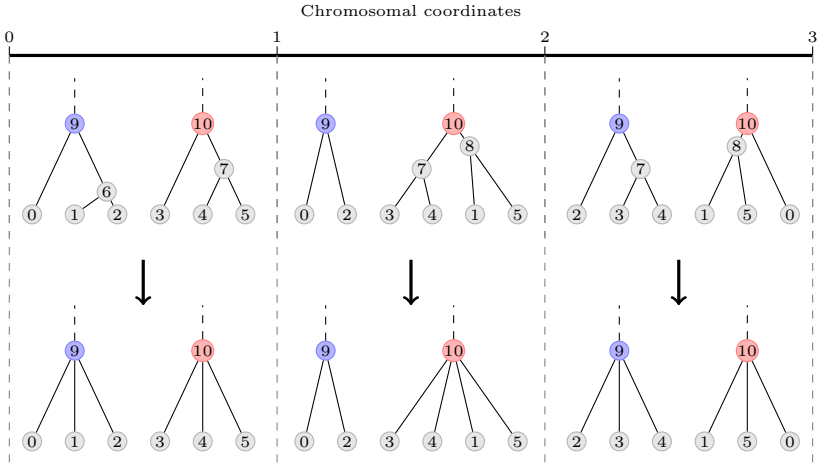
Tree sequences can hold info on population structure



Nodes may be assigned to populations.

The branch joining a sample node to an ancestral node shows the sample's ancestry.

Local ancestry can be extracted from tree sequences



Simulating tree sequences - old methods

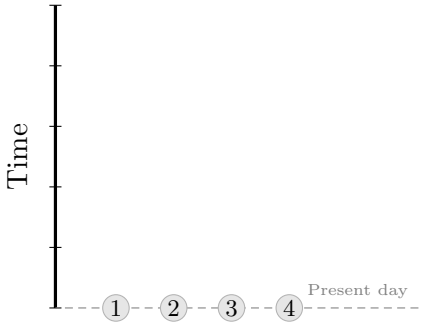
Realistic simulations are important

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

- **Benchmarking and evaluating method performance**
  To assess the accuracy of inferential methods, we need test datasets for which the true values of important parameters are known.

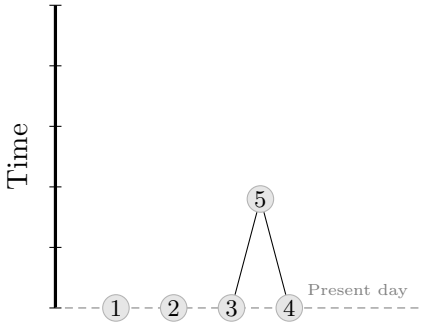
- **Model training**
  Some methods for ancestry inference are trained on simulated data.

msprime simulates a sample backwards-in-time



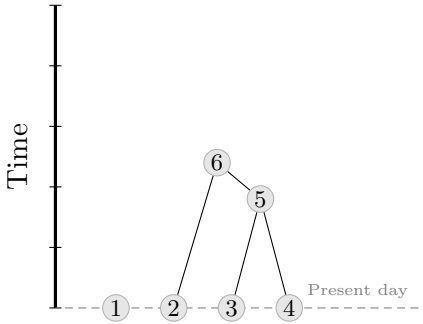
- Simulates tree sequences under an implementation of the coalescent model.

msprime simulates a sample backwards-in-time



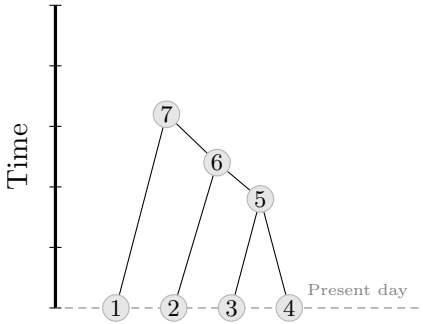
- Simulates tree sequences under an implementation of the coalescent model.

msprime simulates a sample backwards-in-time



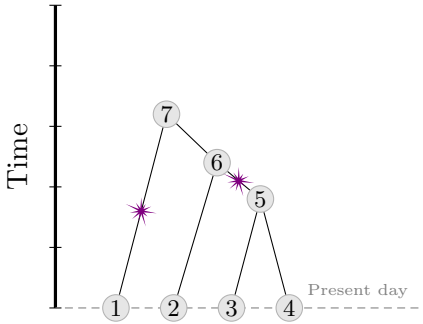
- Simulates tree sequences under an implementation of the coalescent model.

msprime simulates a sample backwards-in-time



- Simulates tree sequences under an implementation of the coalescent model.

msprime simulates a sample backwards-in-time



- Mutations are assumed to be neutral and can be generated independently of the tree topologies and edge lengths.

A simplified toy example: Neanderthal introgression

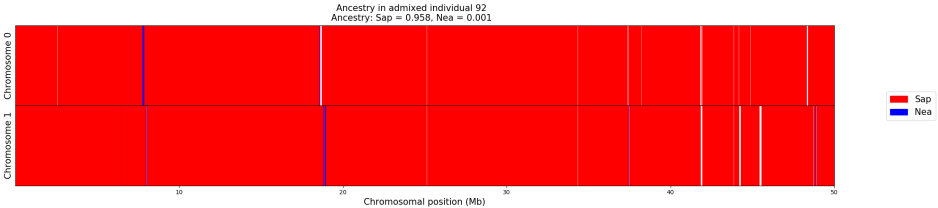
Generations	**Event**
$\approx 240\,000$	Common ancestor of all modern Eurasians and Neanderthals at all loci.
$20\,000$	Divergence of Eurasians and Neanderthals.
$2\,500$	2% introgression of Neanderthals into Eurasians.
0	Samples from 100 Eurasian individuals obtained.

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` performance

	Missing data	Run time	File size	Realism
default	4.0%			✗
+ full ARG	0.0%			✗
+ migration records	0.0%			✗

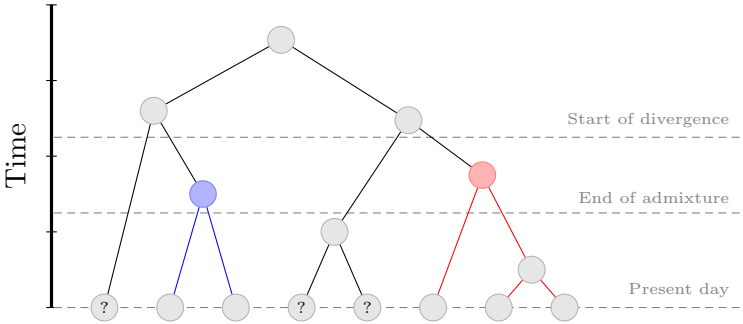
msprime results - missing ancestry data



Global ancestry averaged over all of the simulated samples was

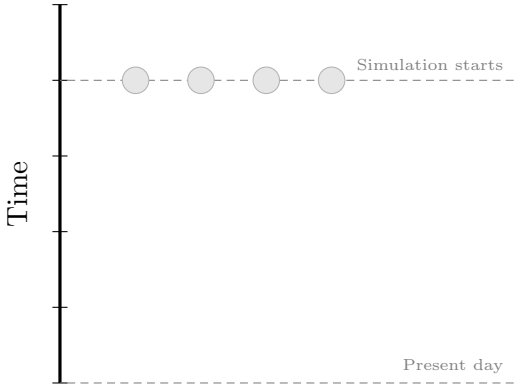
- 96.0% Sapiens
- $< 0.05\%$ Neanderthal
- 4.0% unassigned.

Missing ancestry is due to incomplete lineage sorting



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

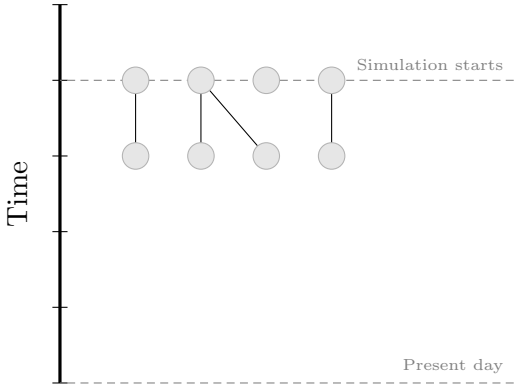
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

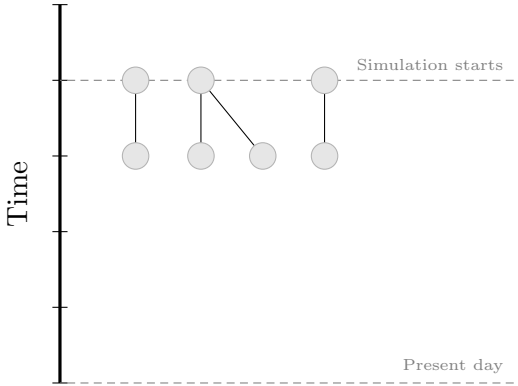
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

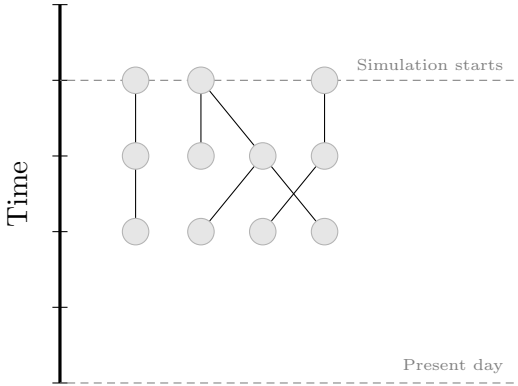
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

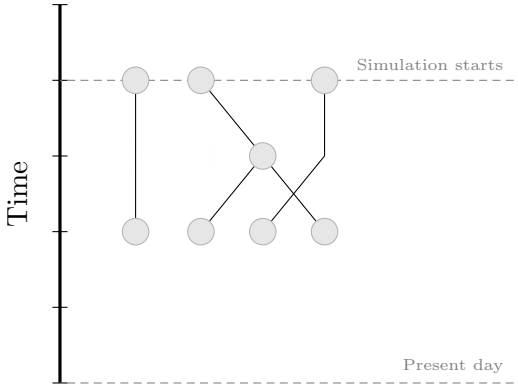
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

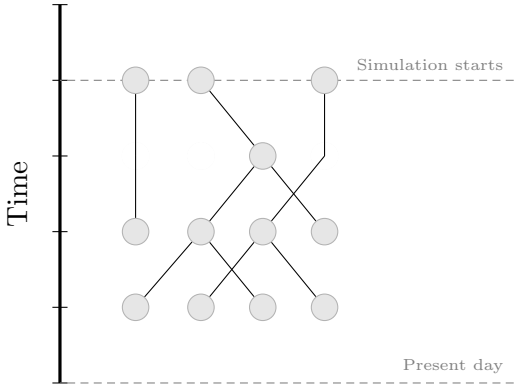
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

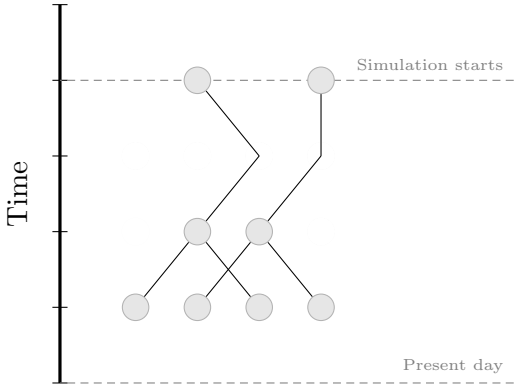
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

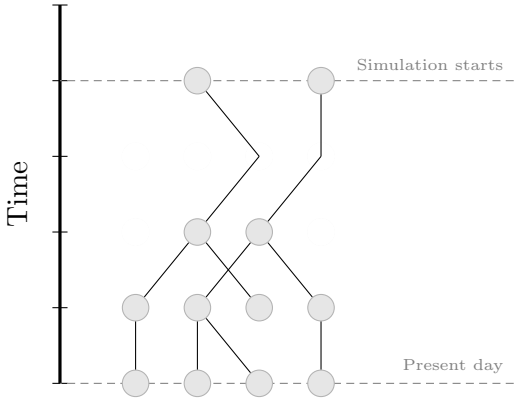
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

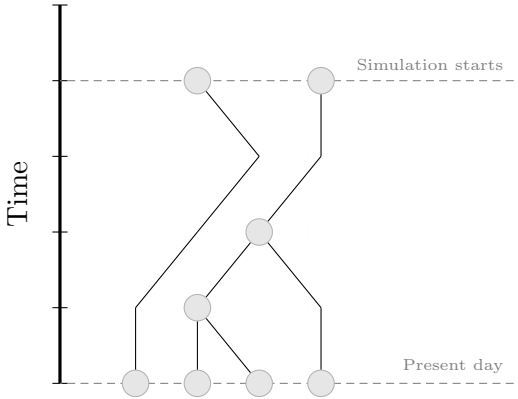
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

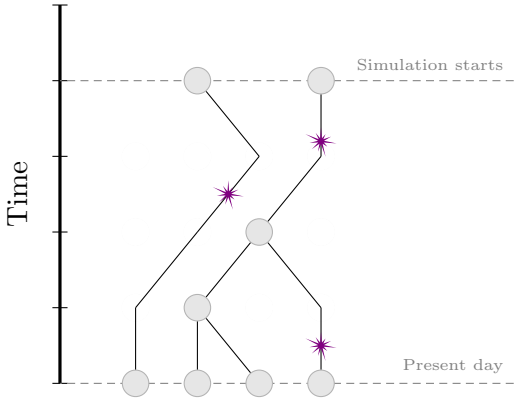
SLiM simulates an entire population forward-in-time



Alternates between

1. forward simulation
2. pruning of irrelevant history.

SLiM simulates an entire population forward-in-time

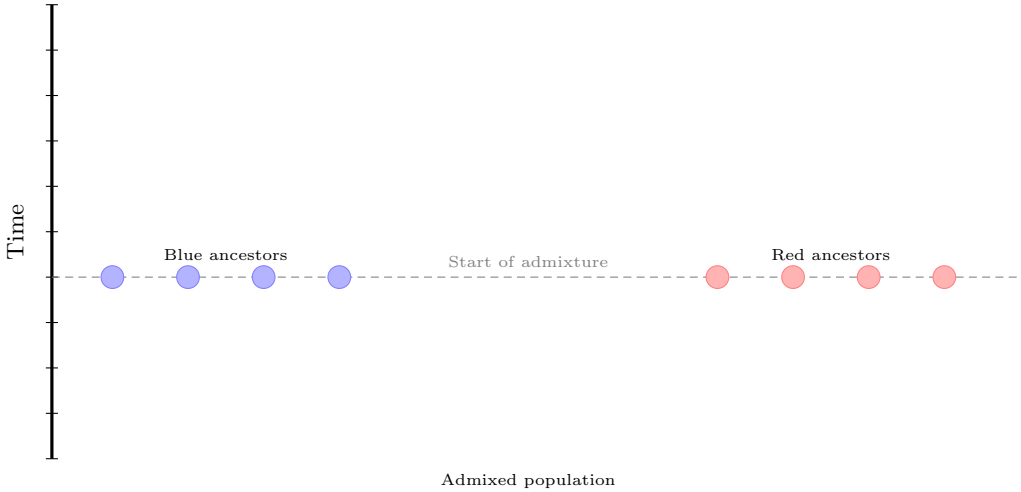


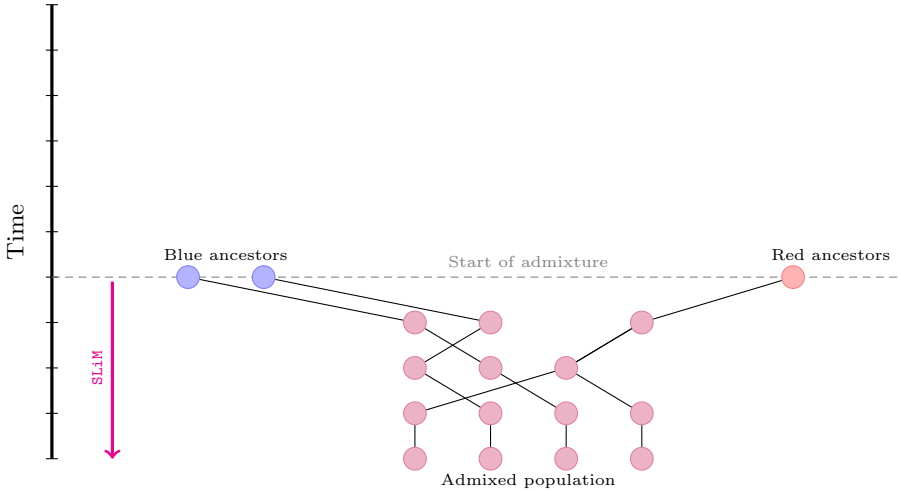
Mutations can be added during the simulation of the tree topologies, or generated independently and added afterwards.

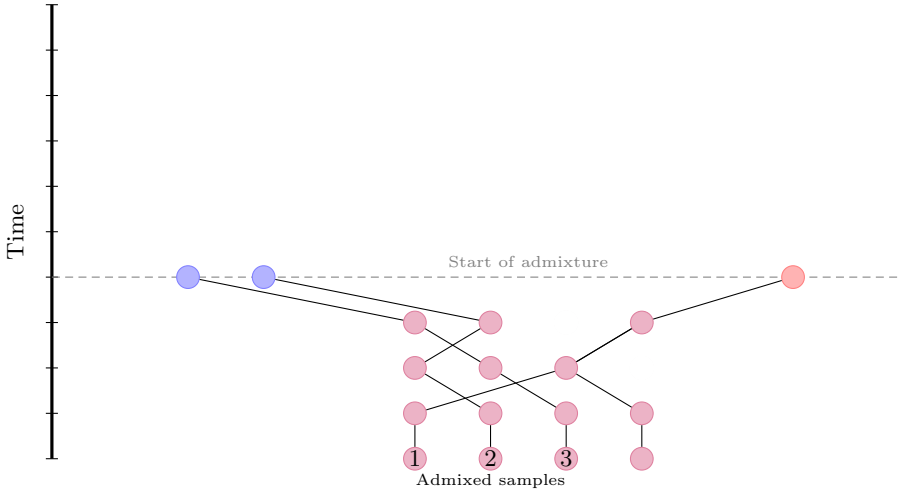
SLiM performance

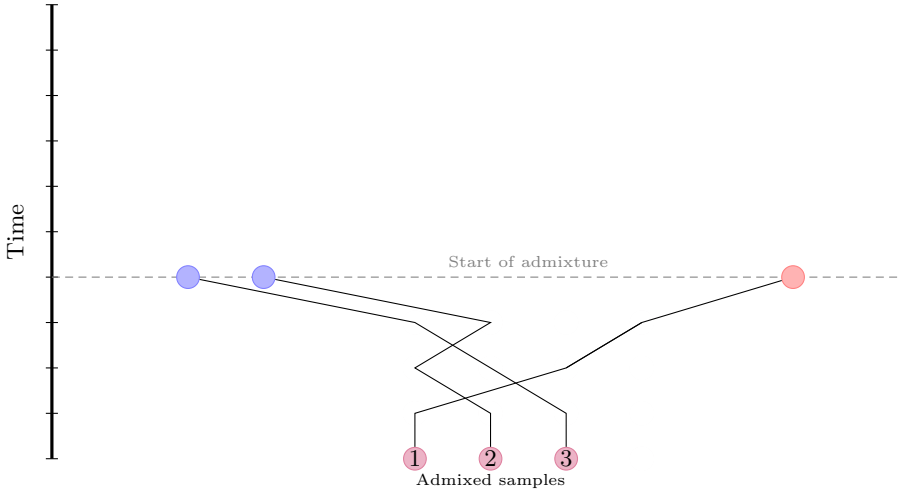
	Missing data	Run time	File size	Realism
default	4.0%			✓
+ unary simplify	0.0%			✓

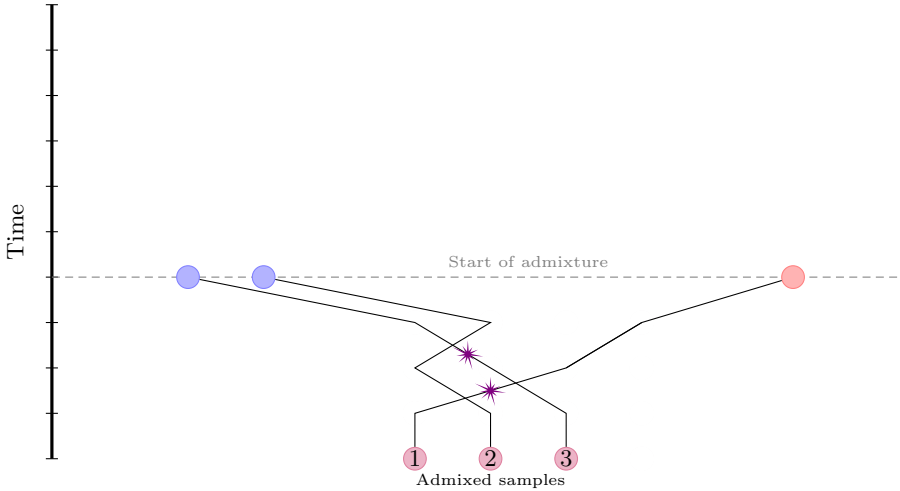
Simulating tree sequences - new method

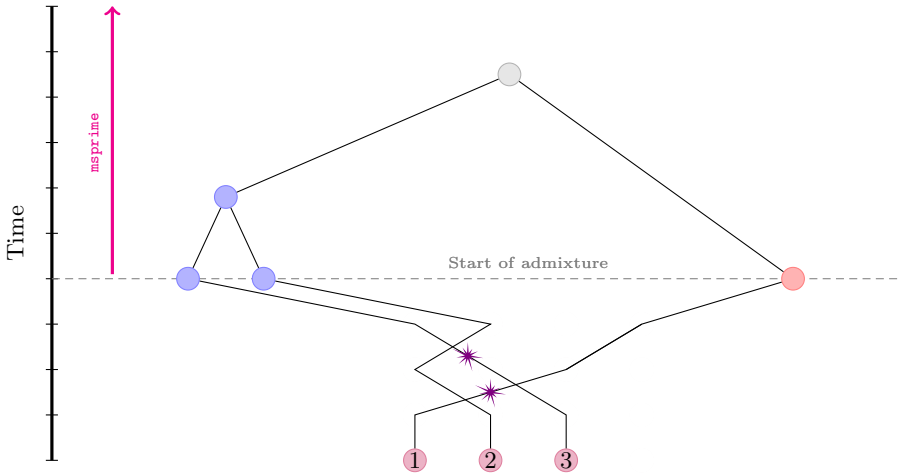

Time
Blue ancestors
Start of admixture
Red ancestors
Admixed population

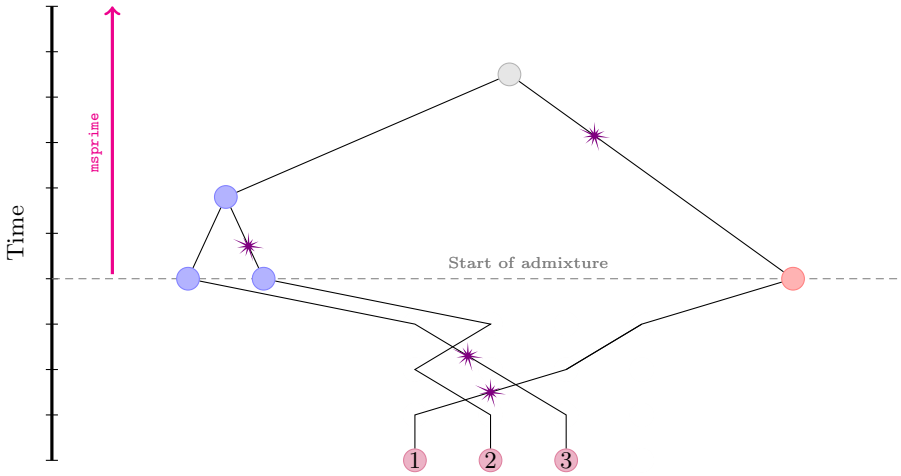

Time
Blue ancestors
Start of admixture
Red ancestors
SLiM
Admixed population


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`	4.0%			✗
+ **full ARG**	0.0%			✗
+ **migration records**	0.0%			✗
`SLiM`	4.0%			✓
+ **unary simplify**	0.0%			✓
Georgia	0.0%			✓

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.
- However, they do not retain the ancestry of all segments due to incomplete lineage sorting.
- A new simulation method overcomes this problem.

Thanks to...

My supervisors	Damjan Vukcevic Stephen Leslie
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

References I

- Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. & Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19, pp. 552 - 566.
- Haller, B. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution*, 36(3), pp. 632 - 637.
- Kelleher, J., and Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, 12(5), pp. 1553 - 7358.
- Kelleher, J., Thornton, K. R., Ashander, J. & Ralph, P. L. (2018) Efficient pedigree recording for fast population genetics simulation. *PLoS Computational Biology* 14(11), pp 1 - 21.

References II

Kelleher, J., Wong, Y., Albers, P. K., Wohns, A. W. & McVean, G. (2018). Inferring the ancestry of everyone. *bioRxiv* (http://dx.doi.org/10.1101/458067)