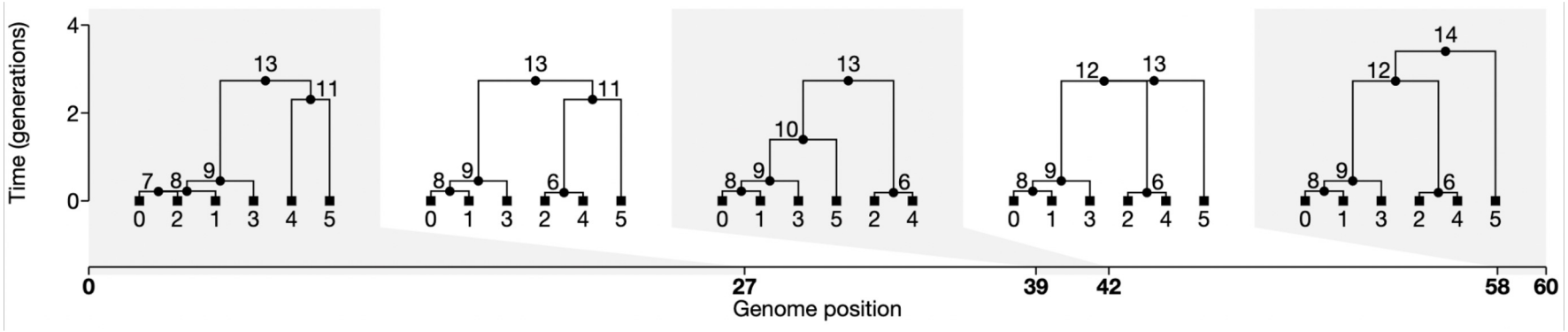
# Module 3: Simulation with msprime

AGAR Workshop | July 28, 2022




**Colin M. Brand** (he/him/his)
Bakar Computational Health Sciences Institute
Department of Epidemiology and Biostatistics
University of California San Francisco

UCSF

# Module Outline

1. **Lecture**: A very brief introduction to evolutionary forces, the Wright-Fisher model, and coalescent theory.
2. **Exercise**: Ancestry simulation in msprime.
3. **Mini Lecture**: Adding mutations to ancestry simulations.
4. **Exercise**: Mutation simulation in msprime.
5. **Mini Lecture**: Demographic models.
6. **Exercise**: Demographic modelling in msprime.

AGAR Workshop 2022
Module 3: Introduction to msprime
Schedule

First Hour (0900 - 1000)
Introduction and Announcements: 5 mins
Breakout Room: 10 mins
Wright-Fisher Models and Coalescent Lecture: 25 mins
Ancestry Simulations Breakout Room: 15 mins
Ancestry Simulations Group Discussion: 5 mins

Second Hour (1000 - 1100)
Mutation Simulations Lecture: 10 mins
Mutation Simulations Breakout Room: 15 mins
Mutation Simulations Group Discussion: 10 mins
Break: 10 mins
Demographic Models Lecture: 15 min

Third Hour (1100 - 1200)
Demographic Models Breakout Room: 25 mins
Demographic Models Group Discussion: 20 mins
Wrap Up: 15 mins

# Module Objectives

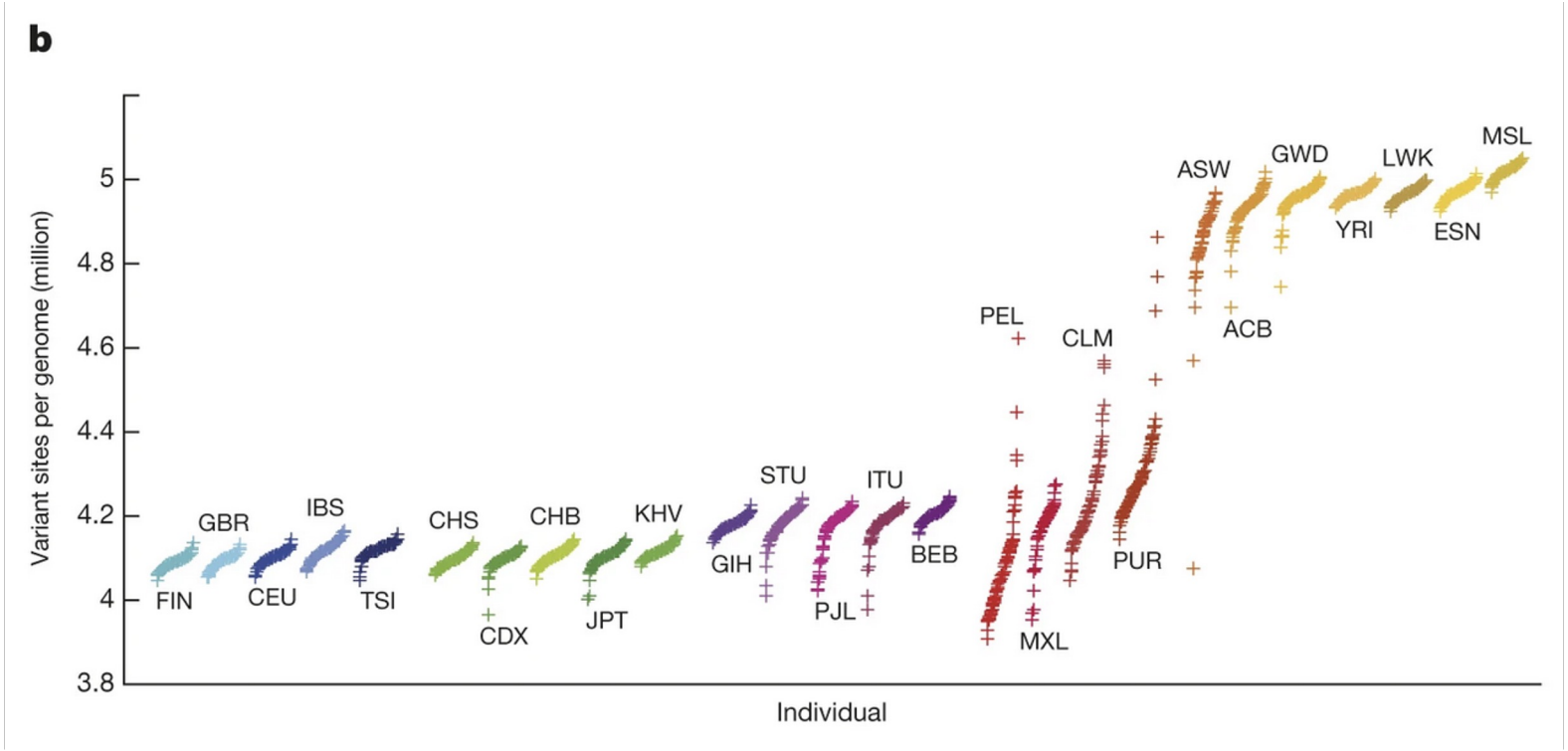
1. Understand the **coalescent** and its relevance to simulating ancestry.
2. Identify the **key parameters** underlying the coalescent and many demographic models.
3. Implement basic **simulations** using msprime.

# Quick Breakout Session

Discussion Points:

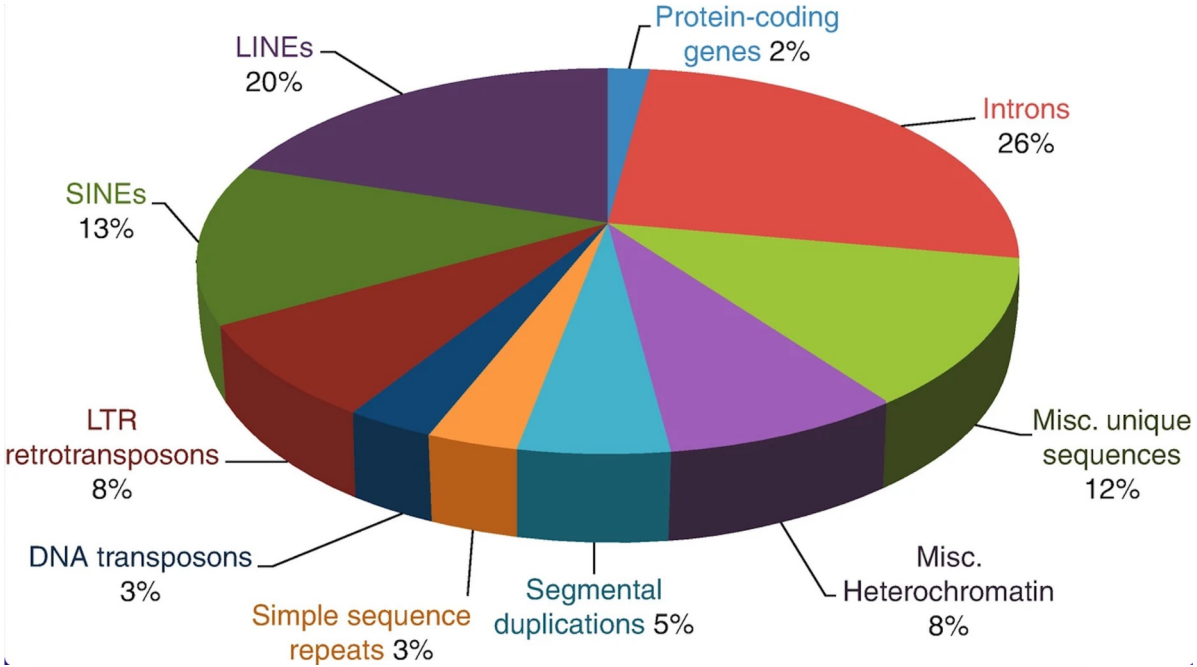
- Introduce yourself.
- Describe something new that you learned yesterday from either module.
- Think of a scenario related to your own research or an interesting question where simulating data would be useful.

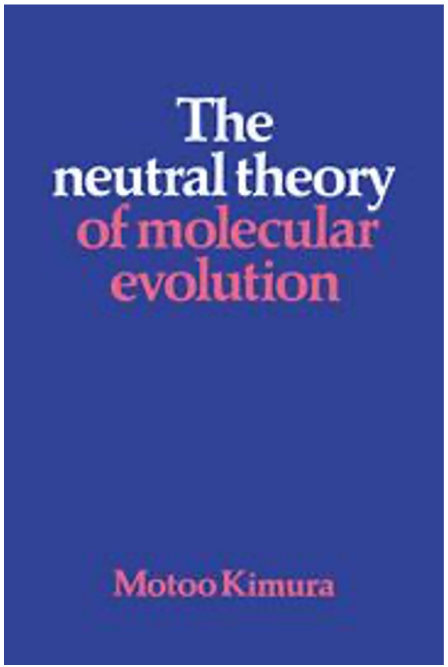
# Genetic variation exists within and among populations/species.




The 1000 Genomes Consortium (2015)

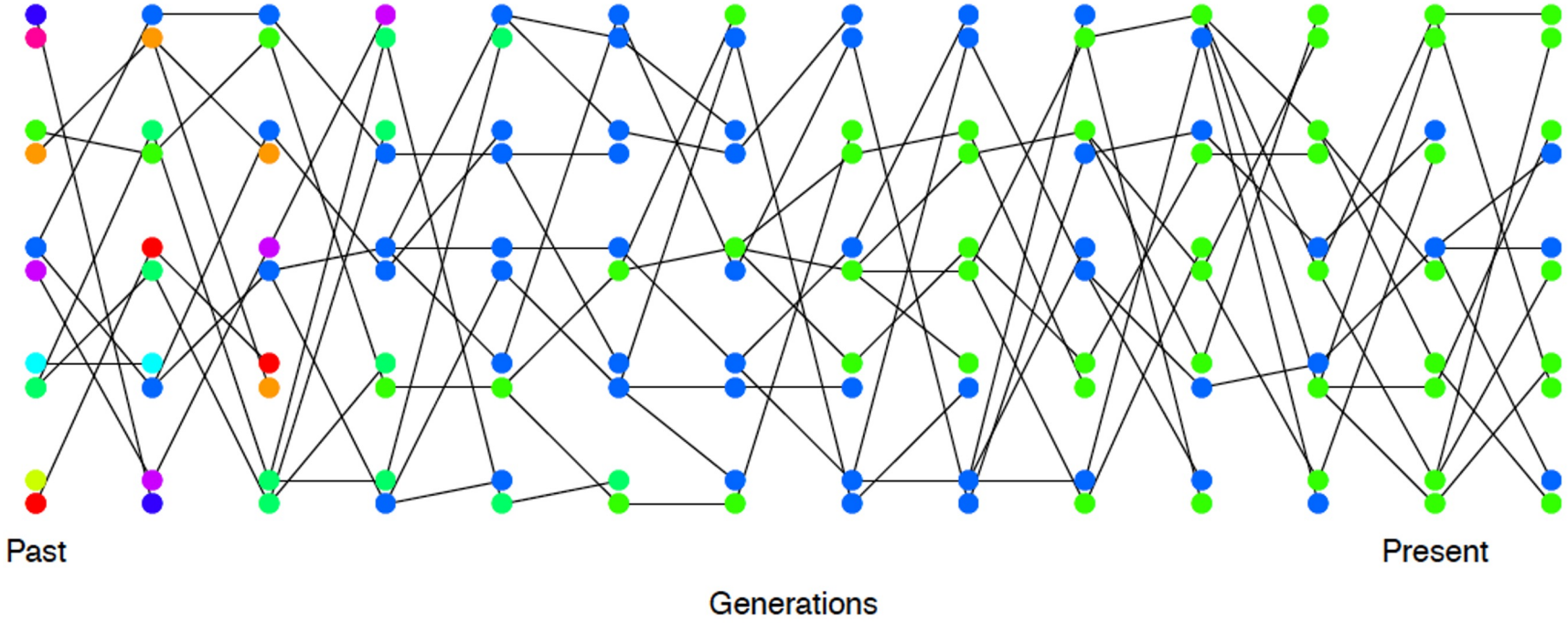
# The vast majority of mutations are minimally consequential.



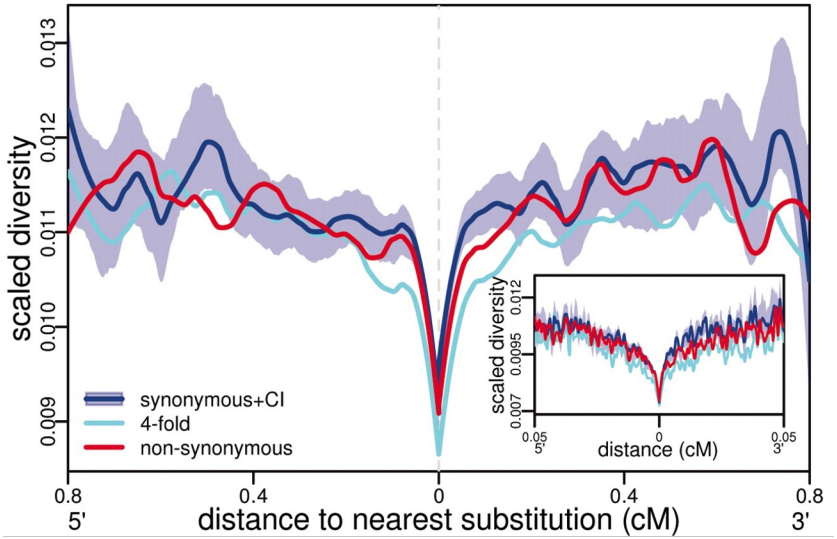

Kimura 1968, Pena 2021 (Figure)

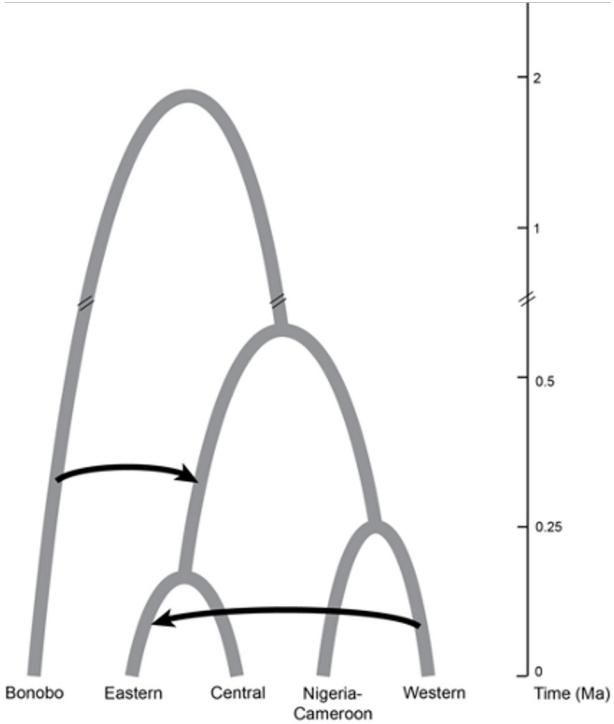
# Genetic drift reduces heterozygosity in the absence of mutation.




Coop 2020

In addition to mutation and genetic drift, other forces shape genetic diversity in populations.




Brand et al. 2022, Hernandez et al. 2011

Much of population genetics is based on Wright-Fisher models.

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

- Discrete, non-overlapping generations

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

- Discrete, non-overlapping generations
- Offspring drawn from previous generation only (i.e., no migration)

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

- Discrete, non-overlapping generations
- Offspring drawn from previous generation only (i.e., no migration)
- Individuals are equally fit

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

- Discrete, non-overlapping generations
- Offspring drawn from previous generation only (i.e., no migration)
- Individuals are equally fit
- Constant population size

# Much of population genetics is based on Wright-Fisher models.

Key assumptions:

- Discrete, non-overlapping generations
- Offspring drawn from previous generation only (i.e., no migration)
- Individuals are equally fit
- Constant population size

**Question**: What complexities in human (or other) populations does this model fail to capture?

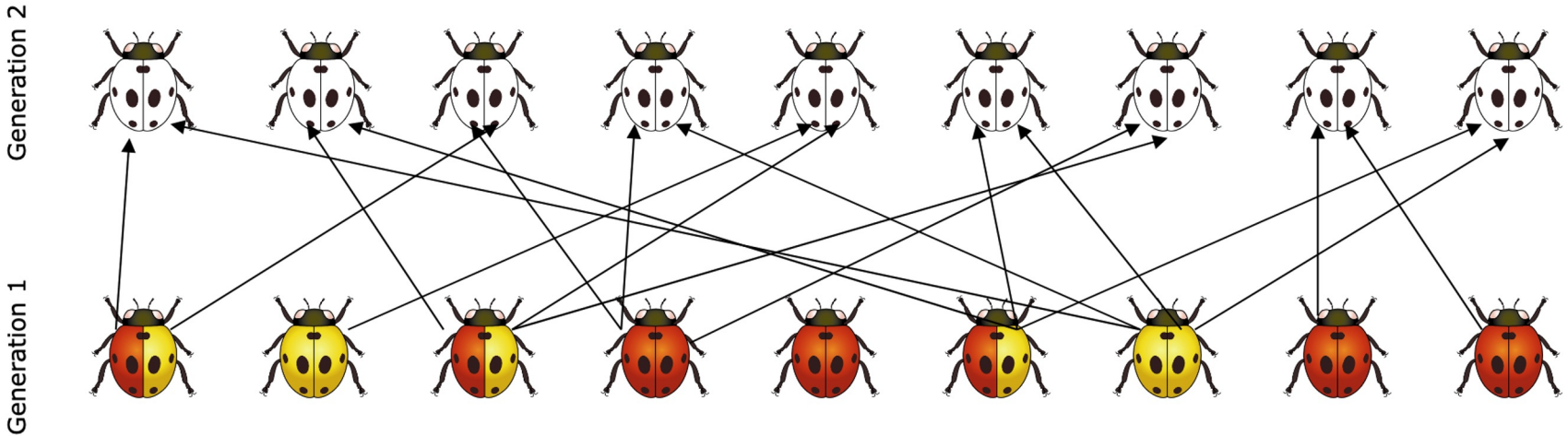
Generation 1

from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

Generation 2

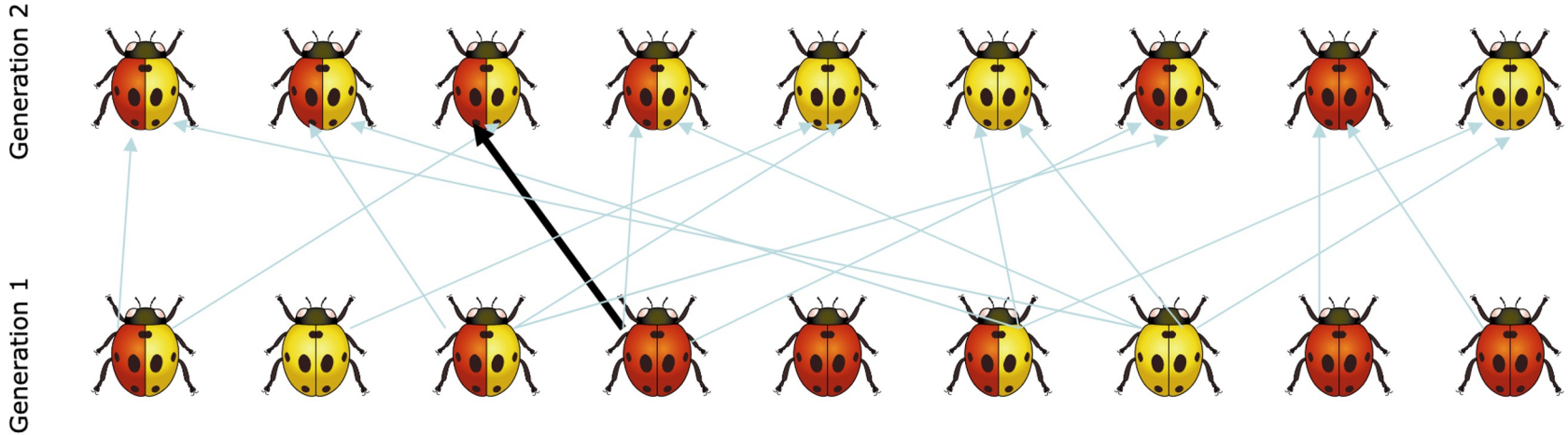
Generation 1

from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture



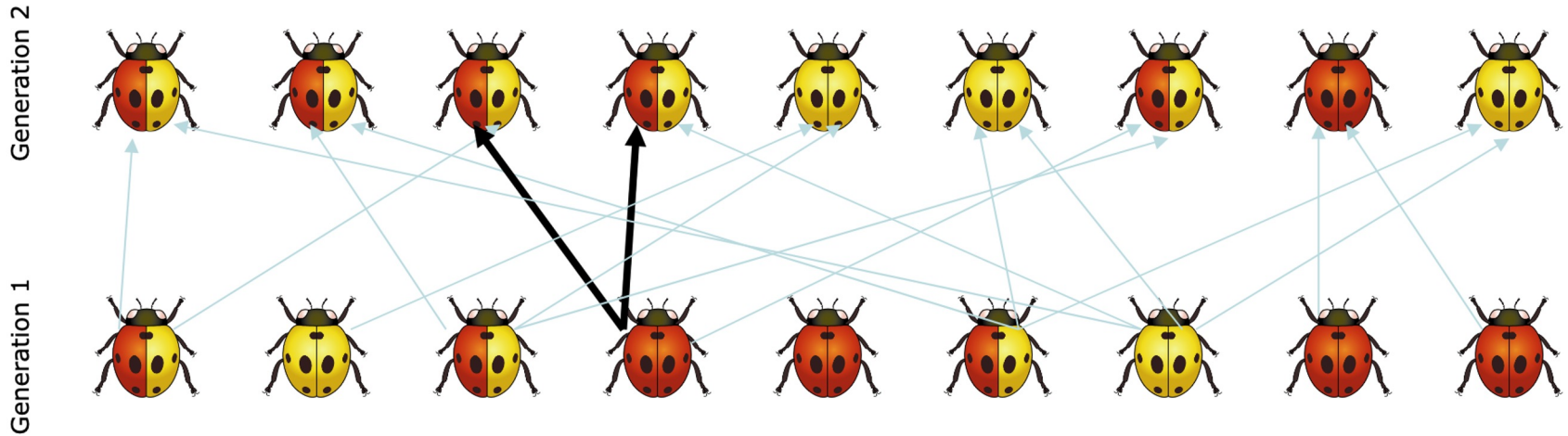

from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

The probability that an allele in generation 2 has a parent in generation 1 is 1.

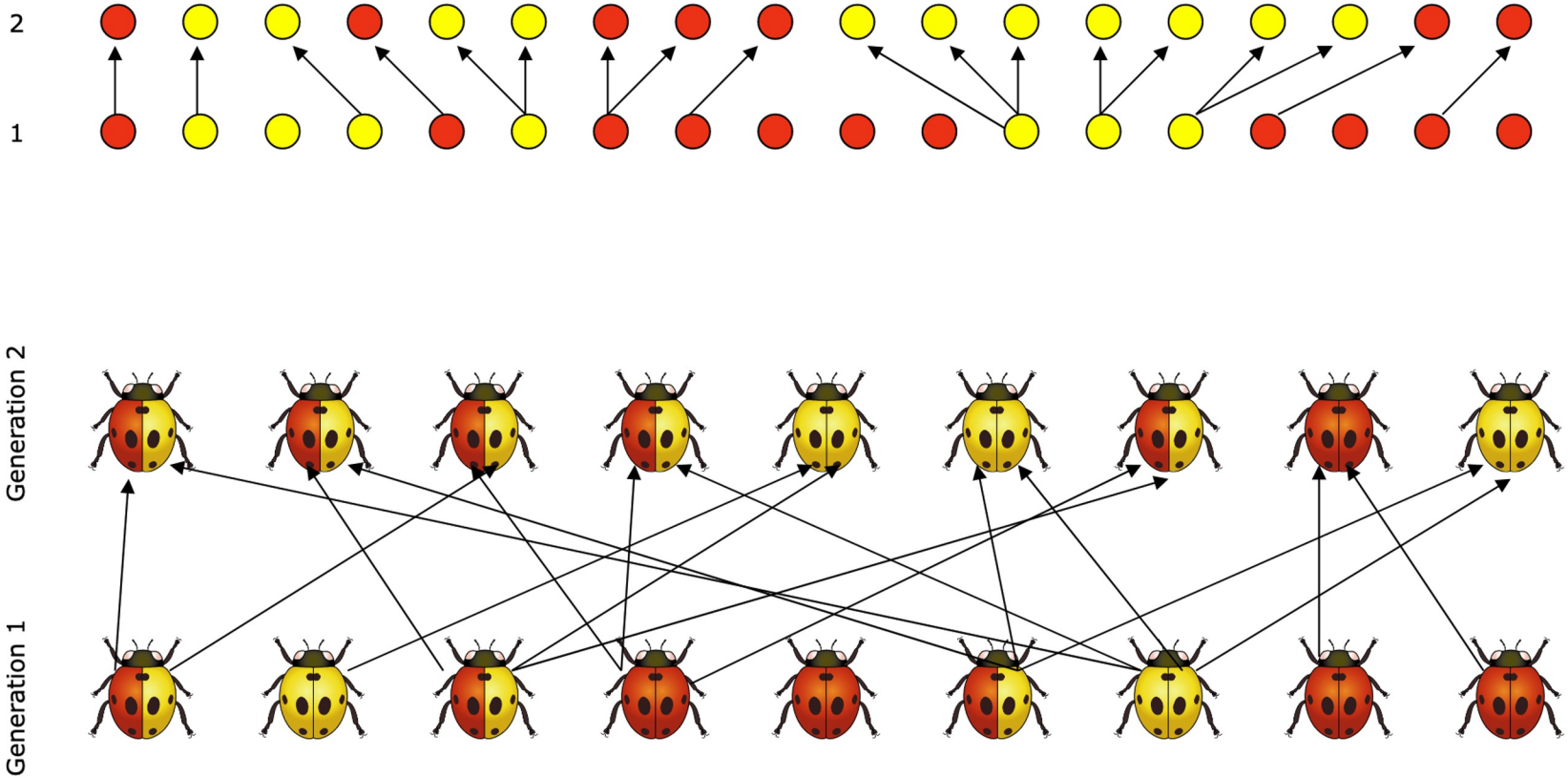



from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

The probability that two alleles in generation 2 share the same parent in generation 1 is 1/(2N).

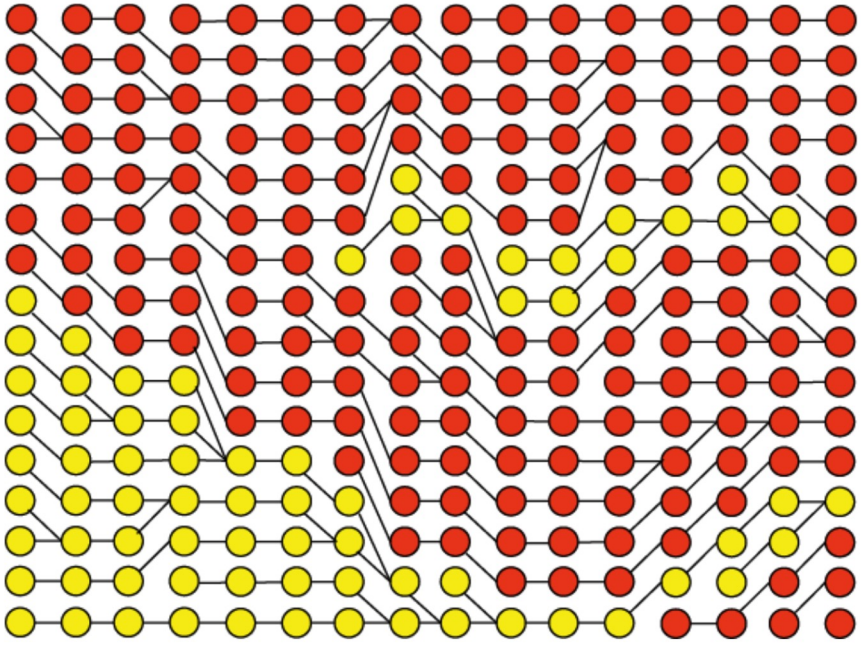



from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture



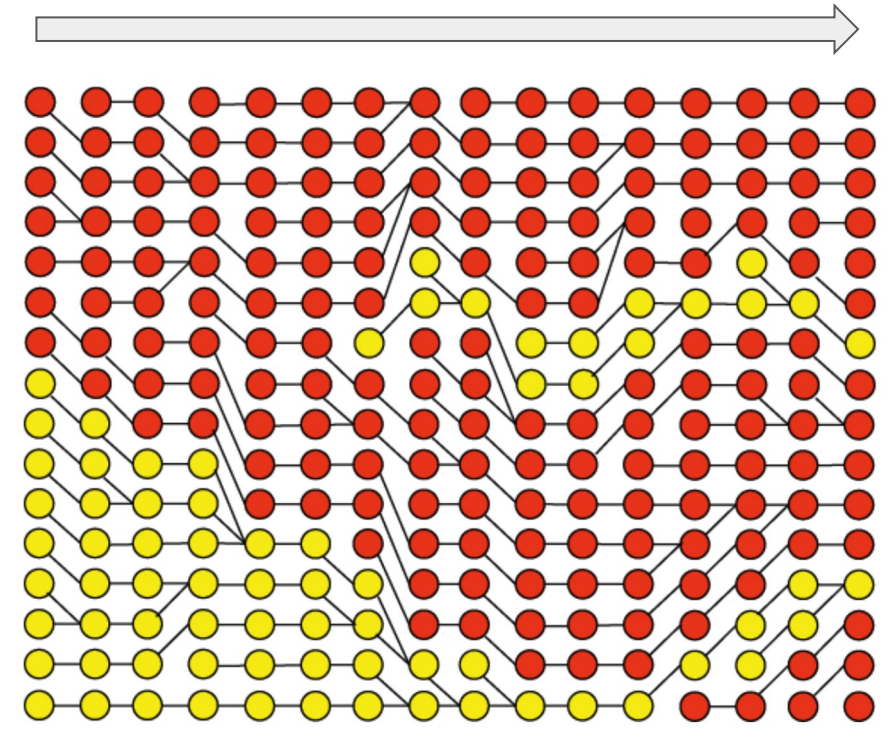

from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

The **coalescent** is a model of the expected relationship among genes in a gene genealogy.



from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

# The coalescent views a population “backwards” in time.



from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

# The coalescence of two genes is a geometric distribution.

| | |
|---|---|
| P (coalesces 1 generation ago) | 1/2N |
| P (coalesces 2 generations ago) | (1-1/2N) * 1/2N |
| P (coalesces 3 generations ago) | (1-1/2N)^2 * 1/2N |
| P (coalesces t generations ago) | (1-1/2N)^t-1 * 1/2N |

The **average** pair of genes last shared a common ancestor 2N generations ago.

This observation establishes a link between the gene genealogy and population size.

# What about a sample of K genes?

Everytime we move “back” one generation, our sample is reduced: K - 1

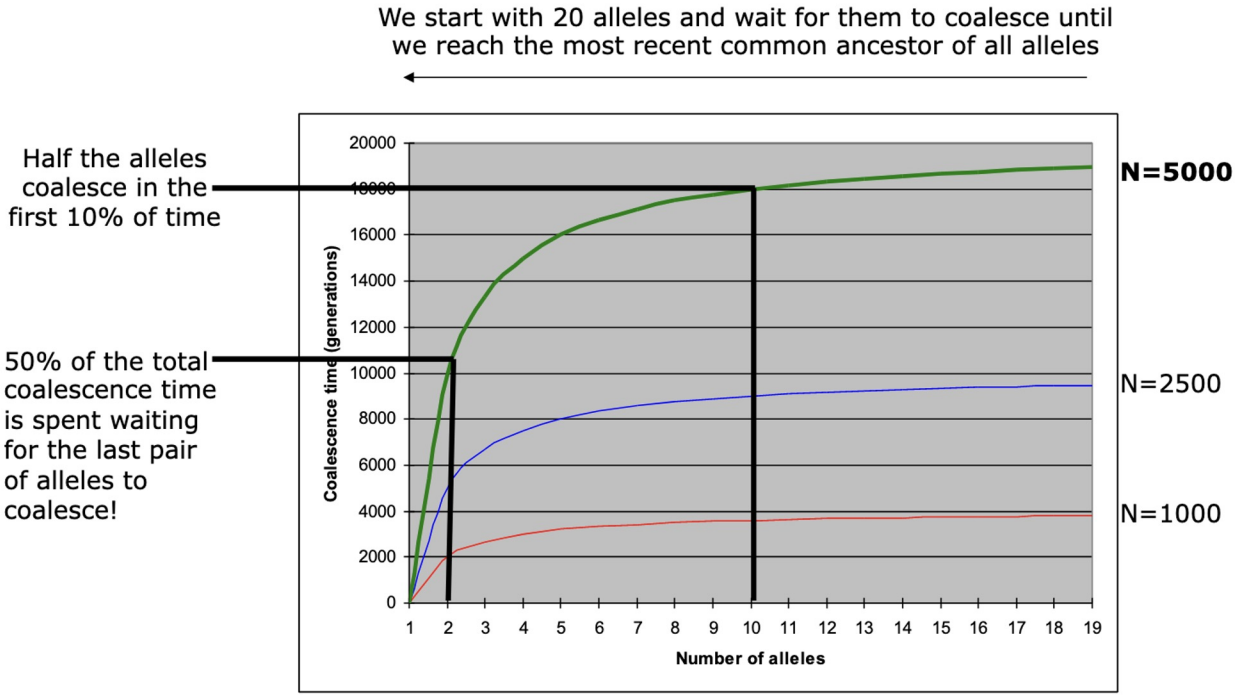
Therefore, there are K - 1 intervals to consider.

$$P(\text{coalescence}) = K(K-1)/2 * (1/2N)$$

or

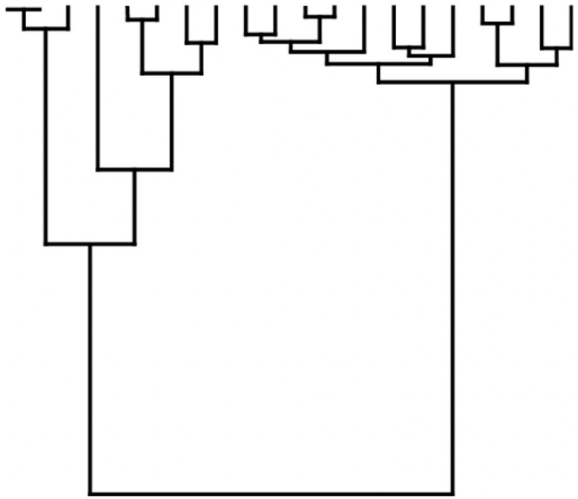
$$P(\text{coalescence}) = 4N(1-1/K)$$

# “Interval” times get larger with each subsequent generation.

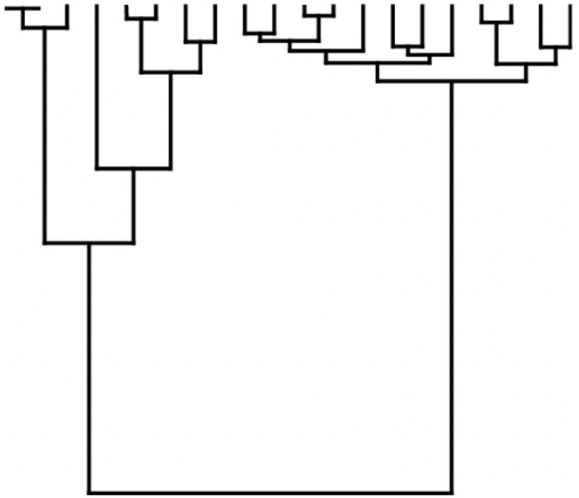



from Mick Elliot and Arne Mooers “Introduction to Coalescent Theory” lecture

These trees are “top heavy”. But also exhibit variation so replicates are needed.

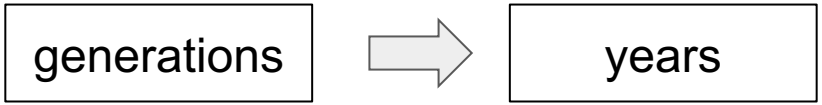


# These trees are “top heavy”. But also exhibit variation so replicates are needed.



The fact that most branches are near the top of the tree underscore that relatively small samples can infer deeper nodes.

The coalescent yields estimates in coalescent units. Additional data are needed to estimate time in years.



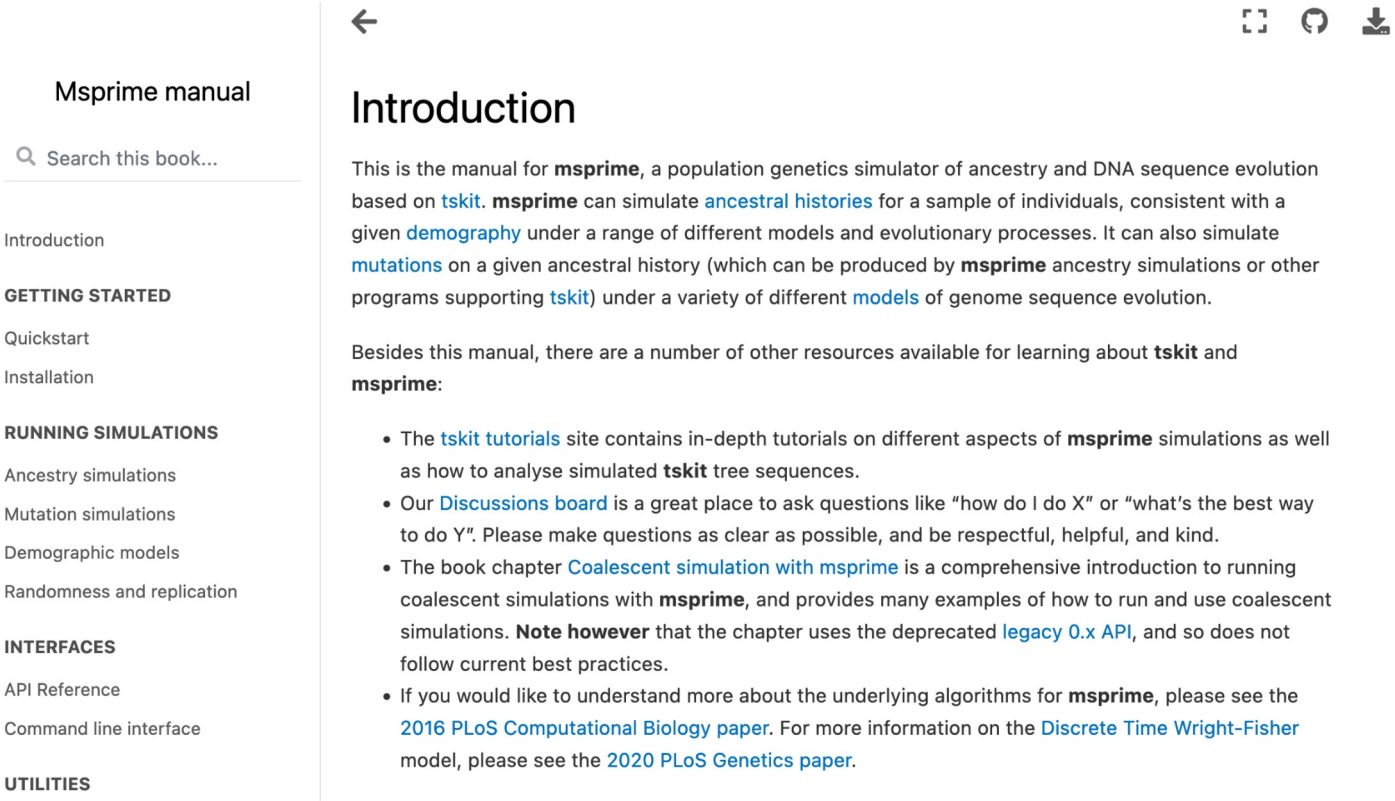

**Question**: What data inform this conversion above?

Now that we’ve got a handle on neutral evolution and the coalescent, let’s learn how to put this into practice by simulating the ancestry of a sample of genomes.

# You have many options when choosing a coalescent simulator.

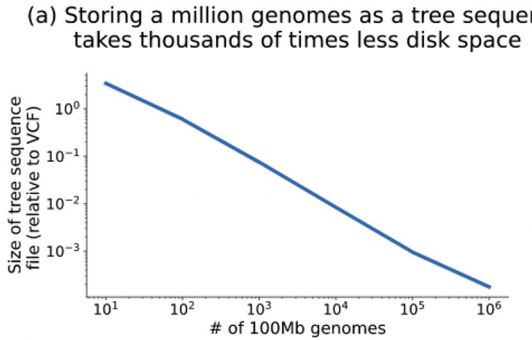
- ms (hence msprime)
- msHOT
- MaCS
- fastsimcoal
- scrm
- msms
- cosi2
- coala
- discoal

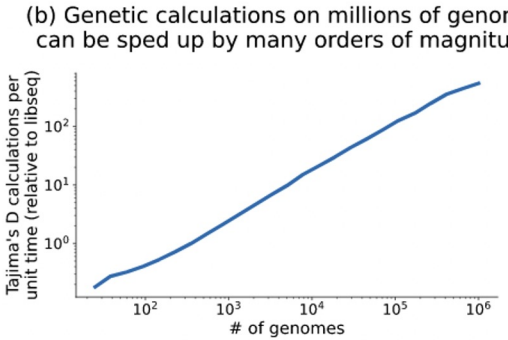
# Today we will focus on msprime.



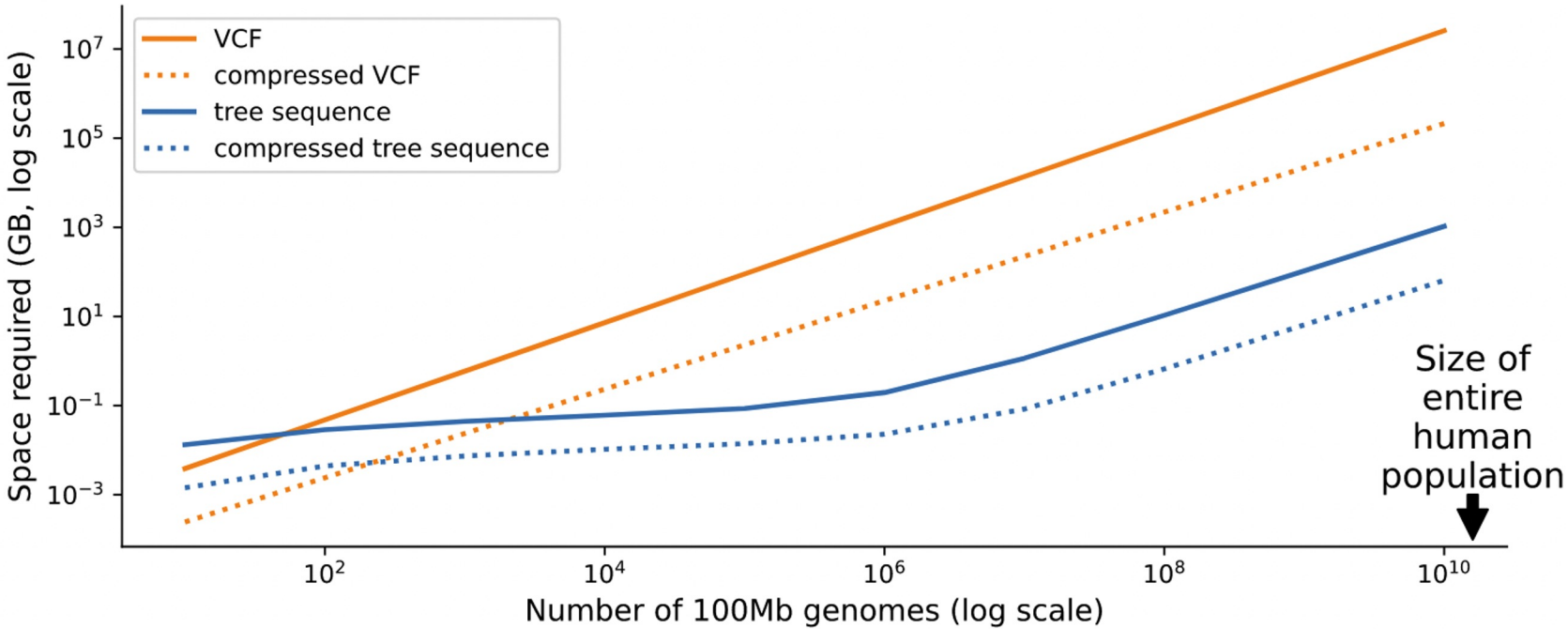
# Msprime uses **tskit** to process simulated data.

- Tskit stores data as a tree sequence
- A tree sequence represents the evolutionary relationships between a set of DNA sequences.
- They take up less space than other sequence formats and enable faster processing.

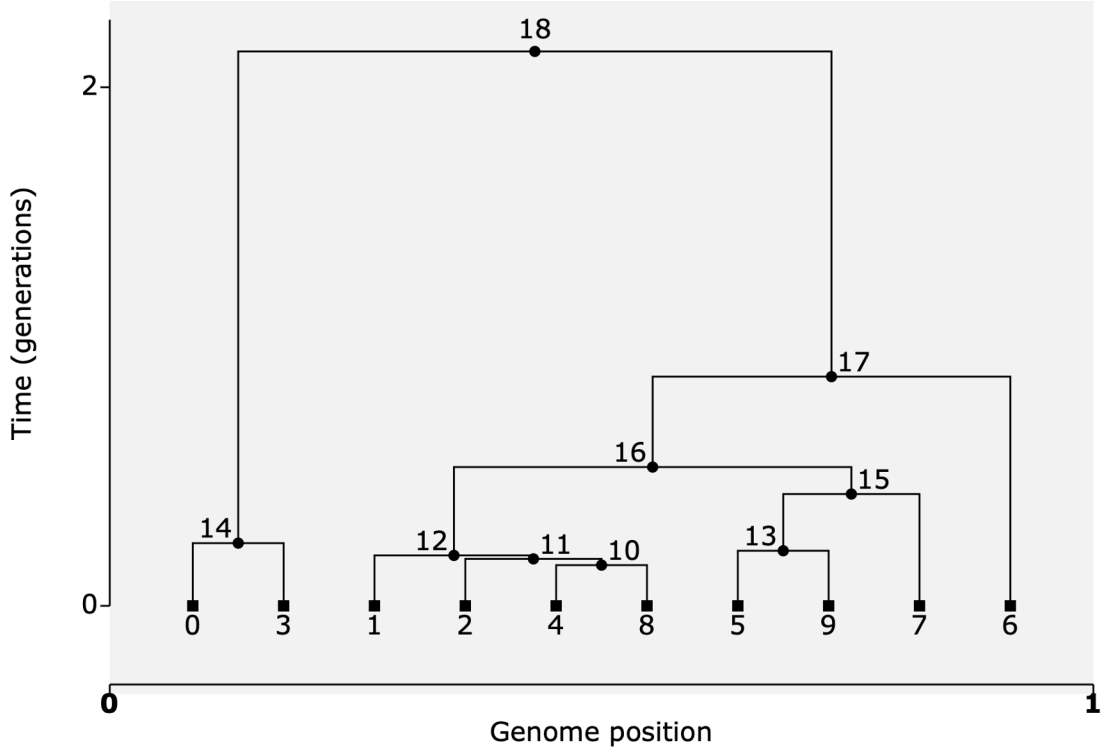


# Tree sequences efficiently store data at large sample sizes.



This tree looks familiar! Msprime allows for easy visualization of simulated ancestry.



Tree sequence objects also contain lots of information about the simulation.

ts kit **Tree Sequence**

| | |
|---|---|
| Trees | 1 |
| Sequence Length | 1.0 |
| Time Units | generations |
| Sample Nodes | 10 |
| Total Size | 2.7 KiB |
| Metadata | No Metadata |

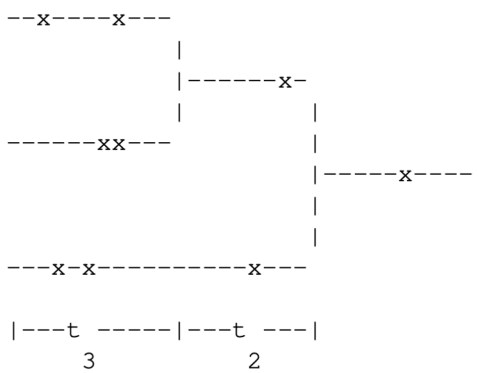
| Table | Rows | Size | Has Metadata |
|---|---|---|---|
| Edges | 18 | 584 Bytes | |
| Individuals | 5 | 164 Bytes | |
| Migrations | 0 | 8 Bytes | |
| Mutations | 0 | 16 Bytes | |
| Nodes | 19 | 540 Bytes | |
| Populations | 1 | 224 Bytes | ✅ |
| Provenances | 1 | 1015 Bytes | |
| Sites | 0 | 16 Bytes | |

# EXERCISE

Ancestral simulations are useful and provide of a summary of population history. But they do not speak to the actual sequence. This is where mutation comes in.

The central question is how many mutations do we expect to see?

This depends on two parameters: 1) the size of the locus / the mutation rate and the 2) architecture of the gene genealogy.




from Alan Rogers' "Lecture Notes on Gene Genealogies"

# Msprime allows you to specify the mutation rate model.

**Models**

**JC69 (Nucleotides)**
Jukes & Cantor model (‘69), equal probability of transitions between nucleotides

This is the default

**HKY (Nucleotides)**
Hasegawa, Kishino & Yano model (‘85), different probabilities for transitions and transversions

**F84 (Nucleotides)**
Felsenstein model (‘84), different probabilities for transitions and transversions

**GTR (Nucleotides)**
Generalised Time-Reversible nucleotide mutation model

**BLOSUM62 (Amino acids)**
The BLOSUM62 model of time-reversible amino acid mutation

**PAM (Amino acids)**
The PAM model of time-reversible amino acid mutation

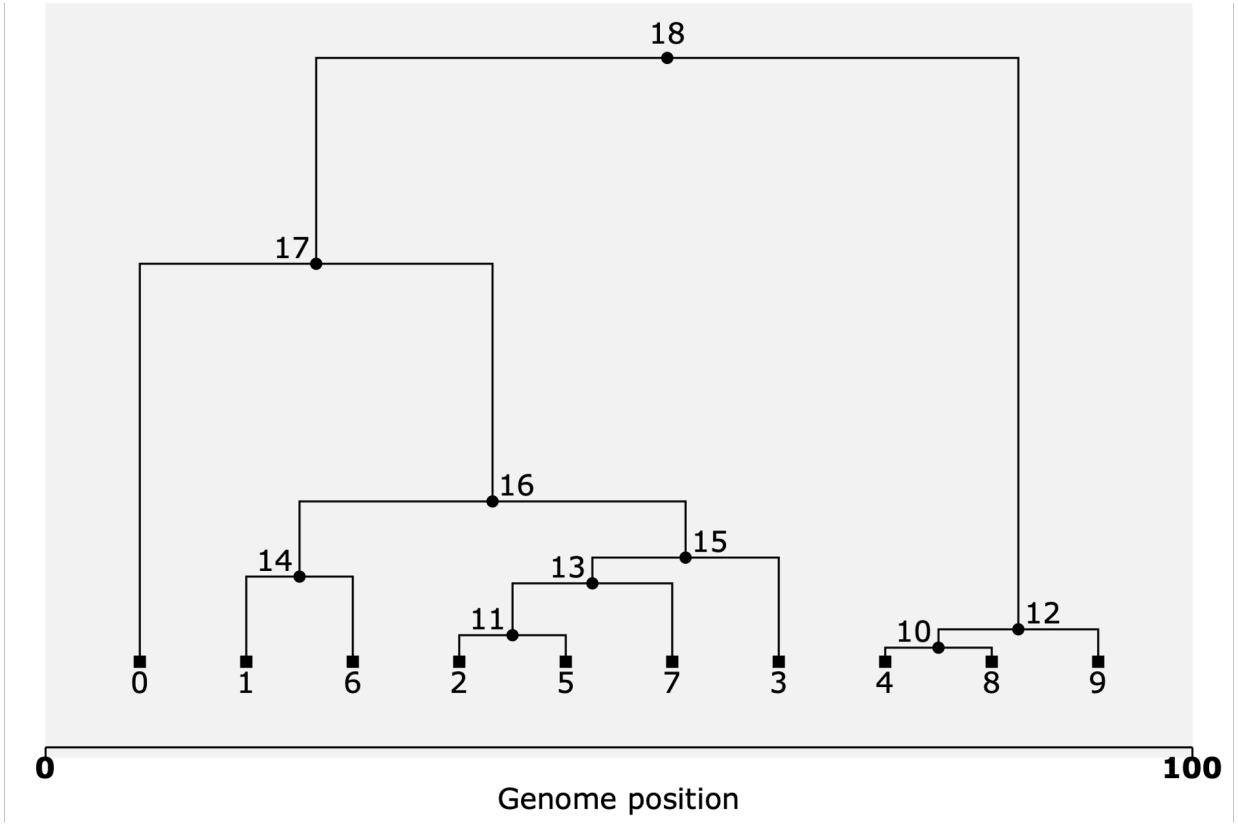
**BinaryMutationModel (Binary ancestral/derived)**
Binary mutation model with two flip-flopping alleles: “0” and “1”.

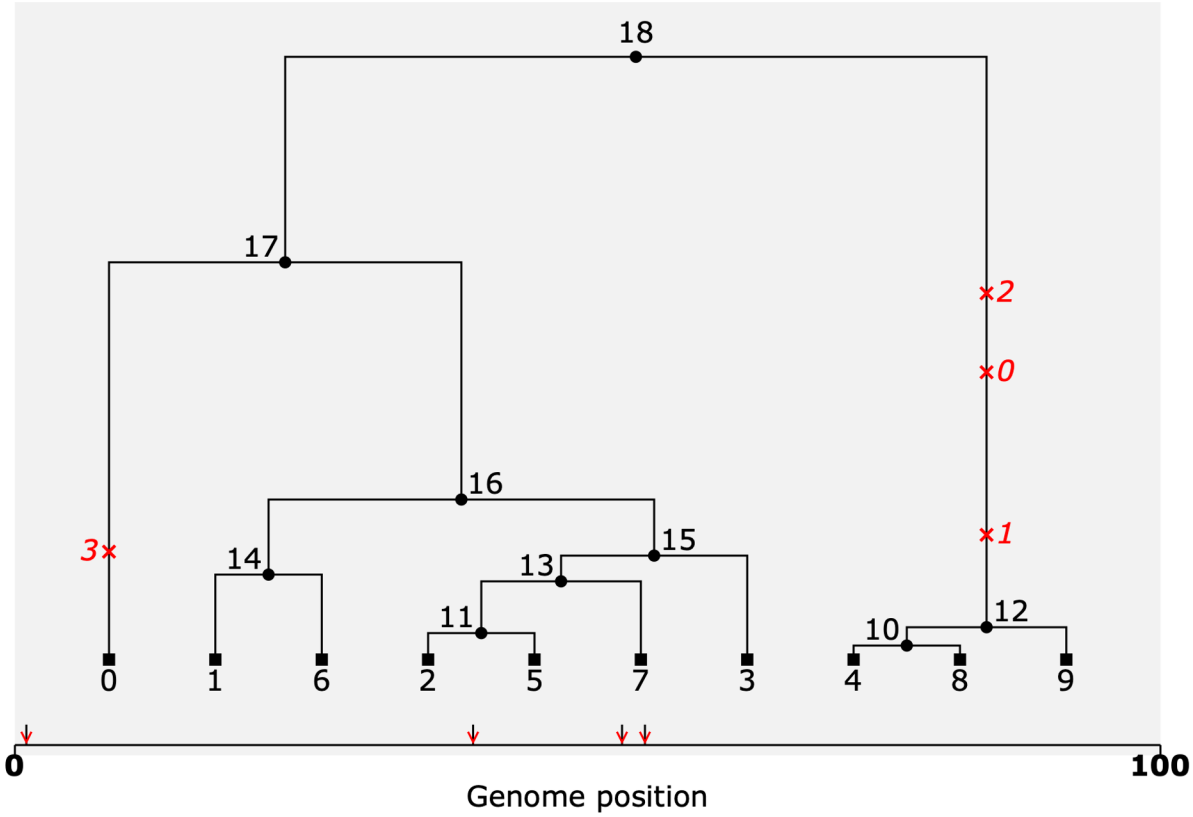
**MatrixMutationModel (General finite state model)**
Superclass of mutation models with a finite set of states

**InfiniteAlleles (Integers)**
A generic infinite-alleles mutation model

**SLiMMutationModel (Integers)**
An infinite-alleles model producing SLiM-style mutations

In most cases, you (probably) don’t need to tinker with the mutation model. But do note this flexibility. You can even customize your own model!


18
17
16
14
15
13
11
12
10
0
1
6
2
5
7
3
4
8
9
0
100
Genome position


18
17
16
15
14
13
12
11
10
0
1
6
2
5
7
3
4
8
9
×2
×0
×1
3×
0
100
Genome position

Quite often, the variation data from mutations is what we're after.

```
In [122]: for var in mts.variants():
              print(var.site.position, var.alleles, var.genotypes, sep="\t")
```

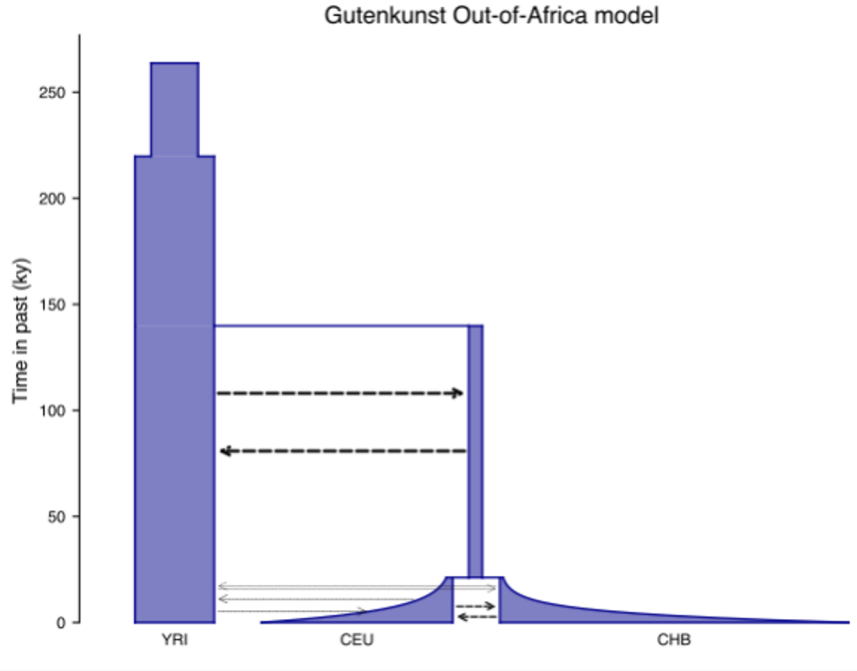
```
1.0     ('G', 'C')      [0 0 0 0 1 0 0 0 1 1]
40.0    ('C', 'G')      [0 0 0 0 1 0 0 0 1 1]
53.0    ('G', 'A')      [0 0 0 0 1 0 0 0 1 1]
55.0    ('C', 'T')      [1 0 0 0 0 0 0 0 0 0]
```

In this example, we ask for the position, the alleles (ancestral vs derived), and genotypes. We can compute lots of different metrics from these data and directly compare them to empirical data.

Plenty of these can actually be computed in msprime!

# EXERCISE

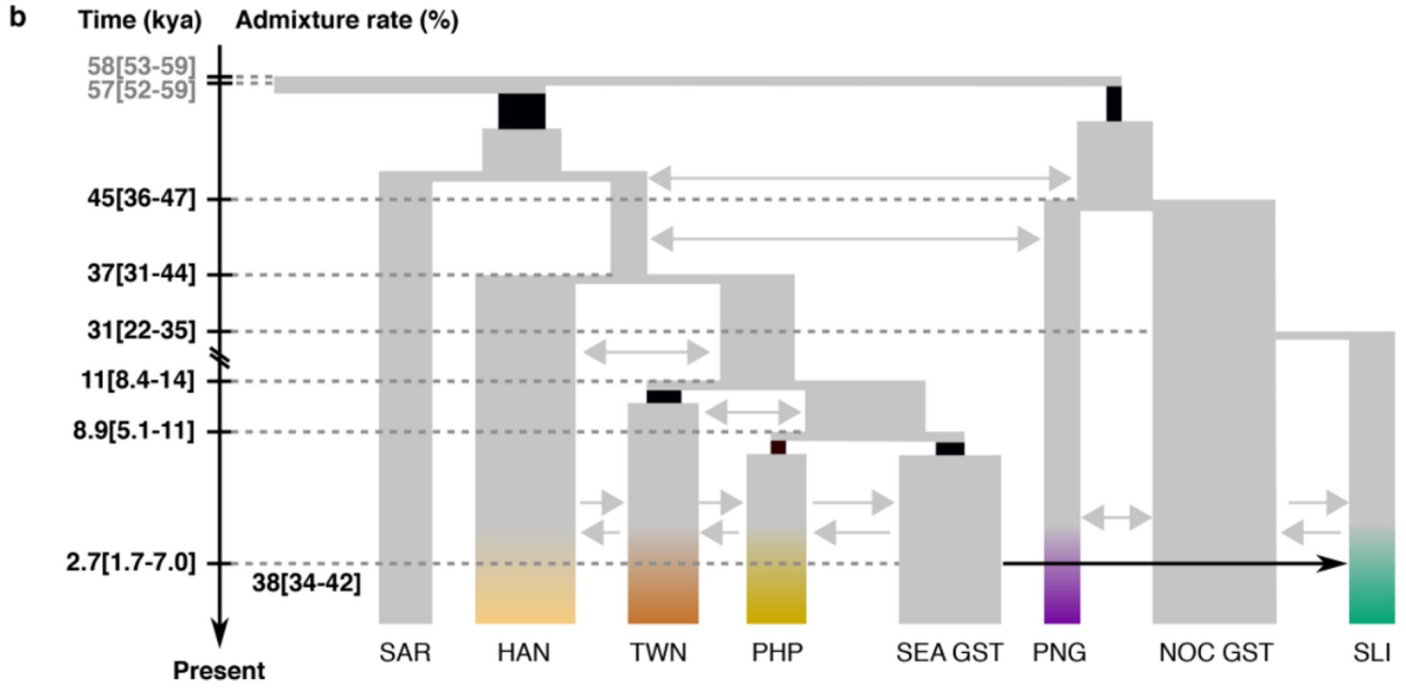
Populations do not exist statically in time nor in isolation. Let's add some complexity to our simulations to generate a better null.



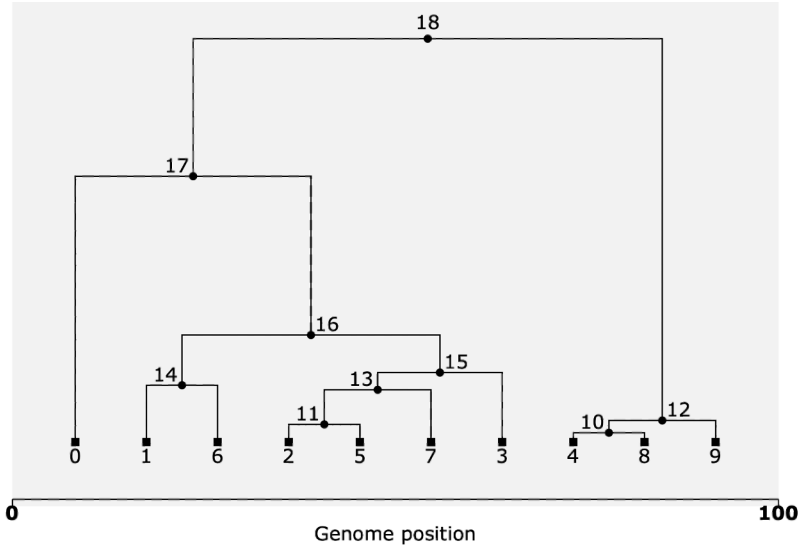
We need to consider a number of parameters when building a demographic model.

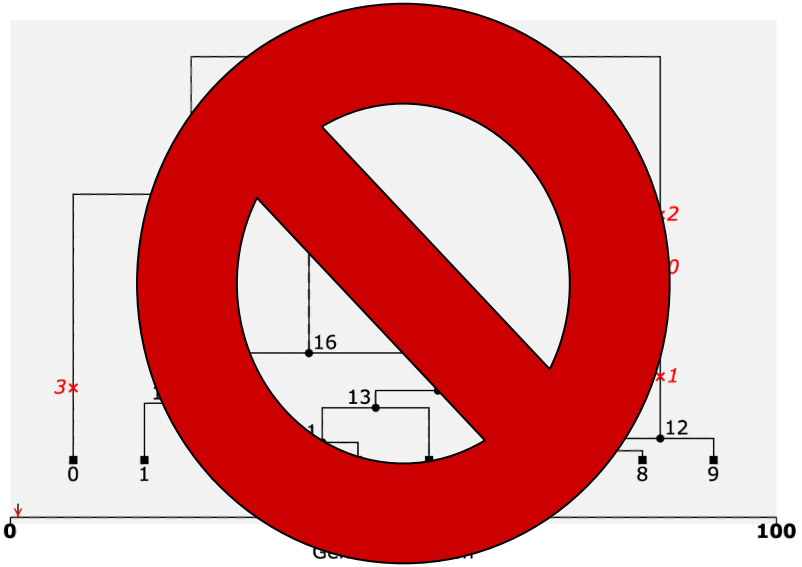
- Sample size
- Population size
- Growth rates
- Timing of events
- Gene flow (we'll cover this in the next module!)

The time and memory needed to run simulations scales with complexity.



When you start to model complex demographics with smallish sample sizes, plot the genealogy only, not the mutations.



# EXERCISE

# Wrap Up

- Msprime is a powerful tool for simulating neutral evolution under a variety of scenarios
- Other coalescent software can handle more specific cases; otherwise forward time simulators such as SLiM may be useful
- Careful thought about the key parameters in your model will pay off later
- Simulations are [usually] fast and cheap. Use a variety of parameter values if needed and don’t hesitate to run computational experiments.
- This must be balanced with the reality that simulations will not cover biological reality. We should strive to get as far as we biologically/computational can.

# Contact

Email: colin.brand@ucsf.edu

Twitter: @colinmbrand

GitHub: brandcm