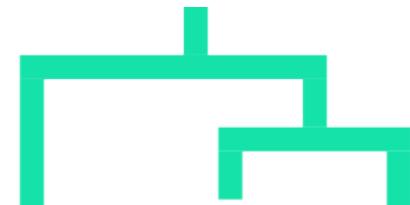
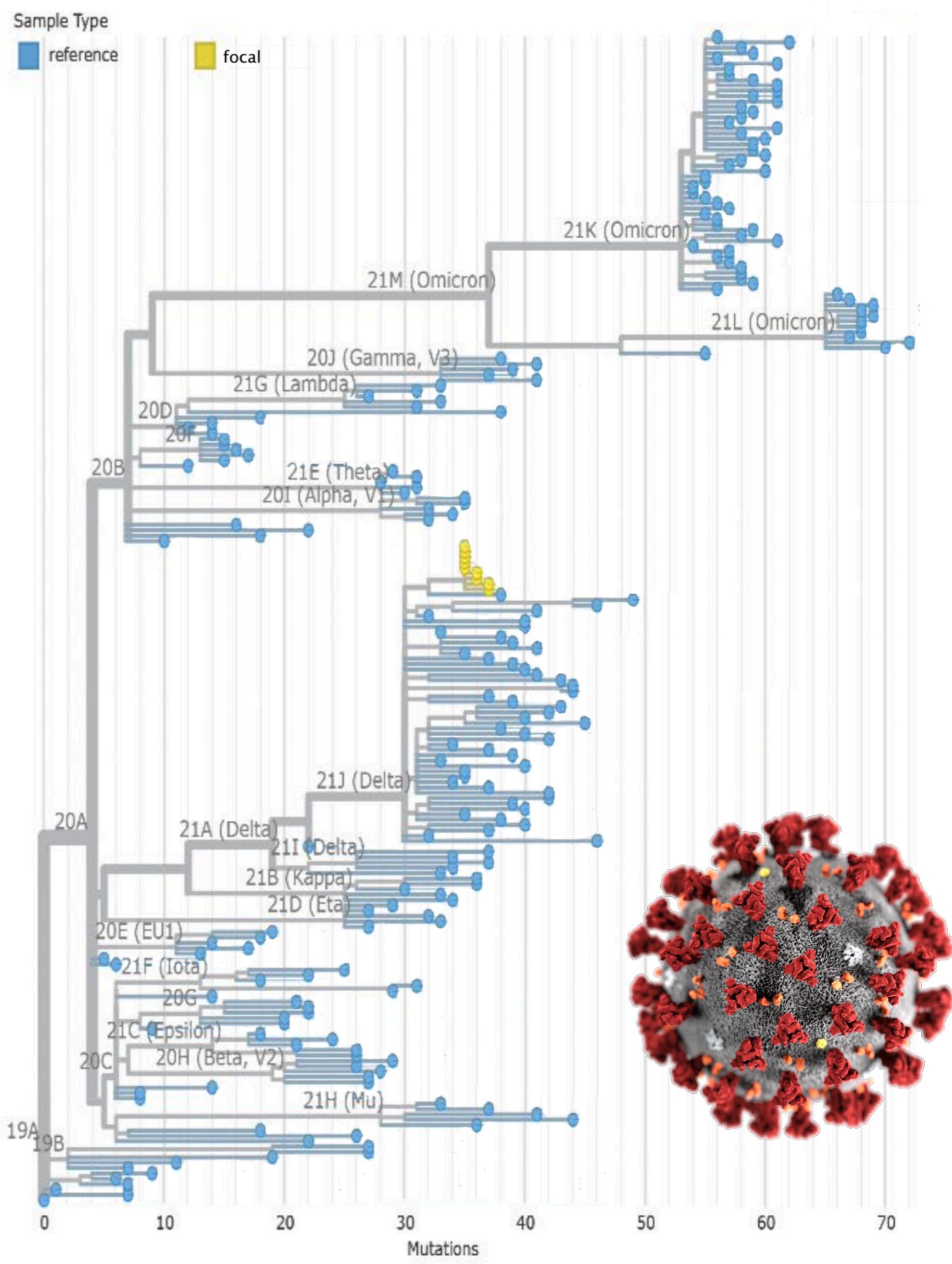


# Genealogical analysis

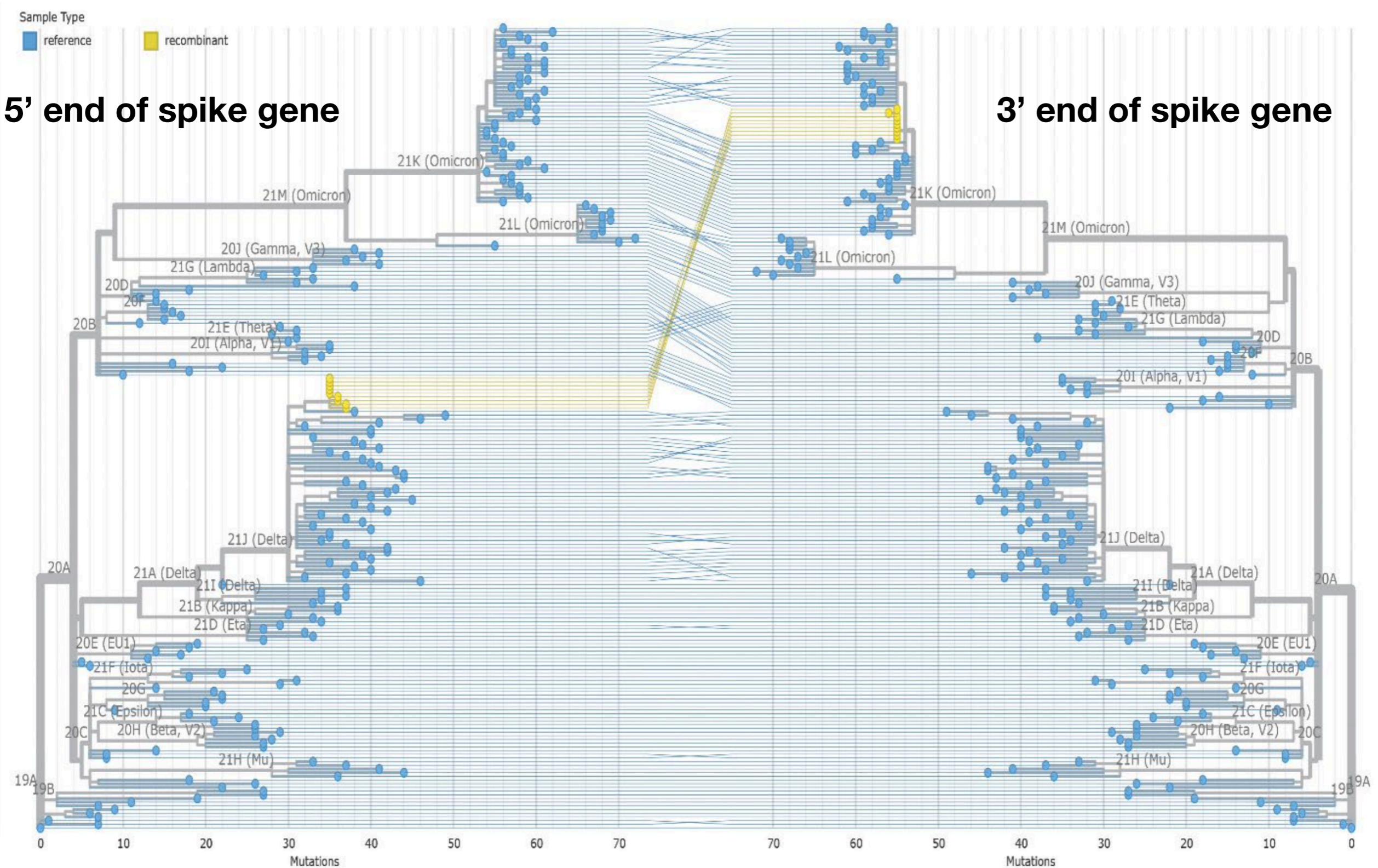
Trees, tree sequences, and  **ts kit**

# Trees are fundamental in comparing & analysing genetic sequences...

from: SARS-CoV-2 Delta–Omicron Recombinant Viruses, United States (Lacek *et al.*, *Emerg Infect Dis*. 2022: <https://doi.org/10.3201/eid2807.220526>)

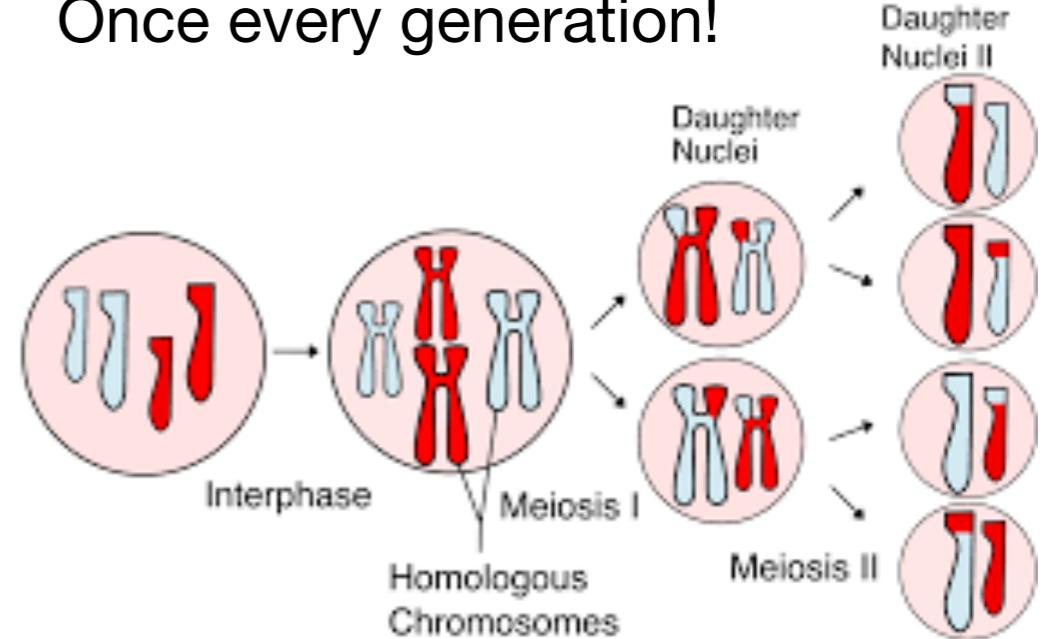


# ... but recombination causes different genetic regions to have different trees



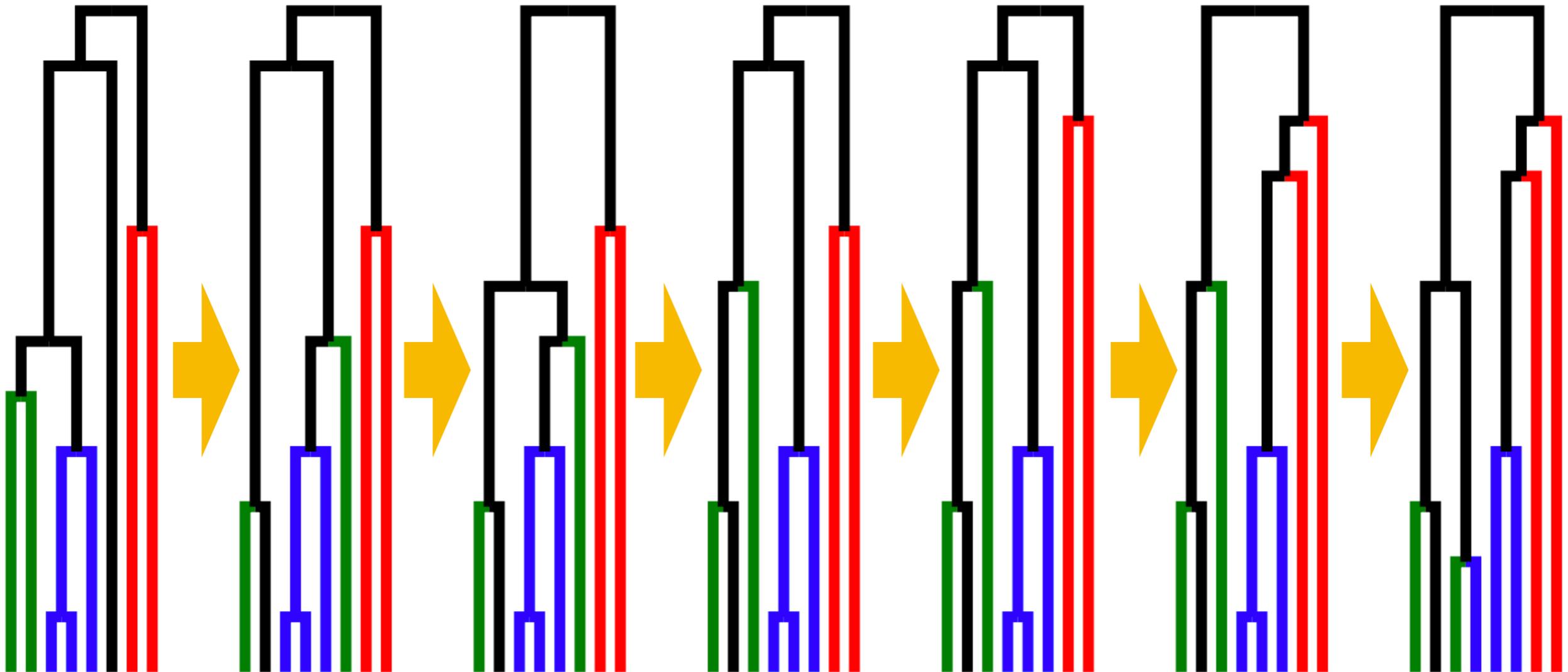
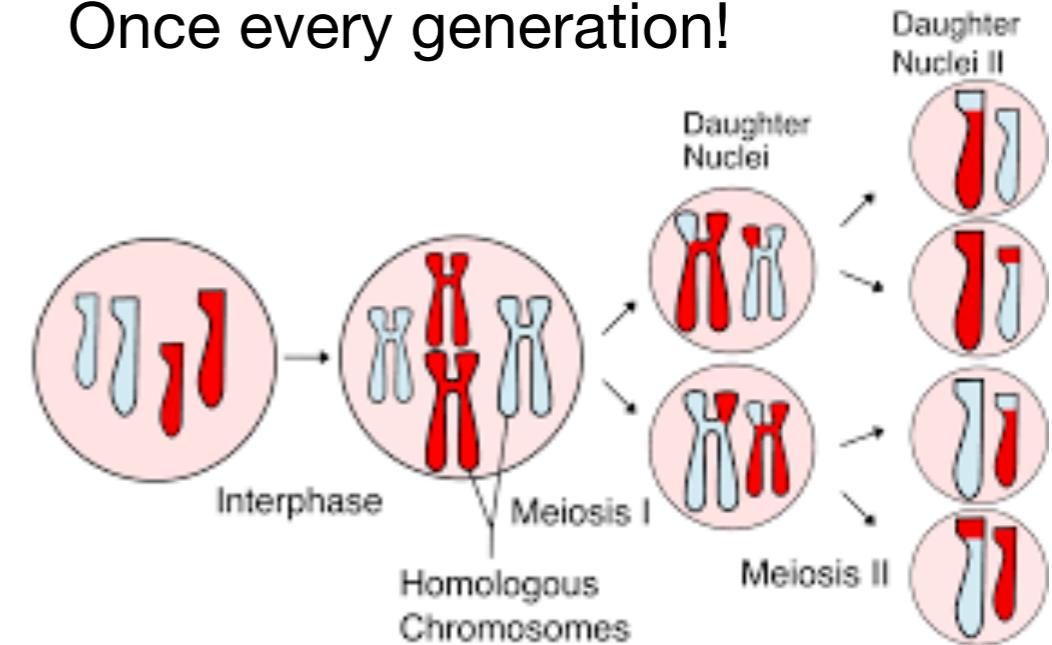
**Problem:** most organisms (including ourselves) recombine much *more* than coronaviruses

Once every generation!

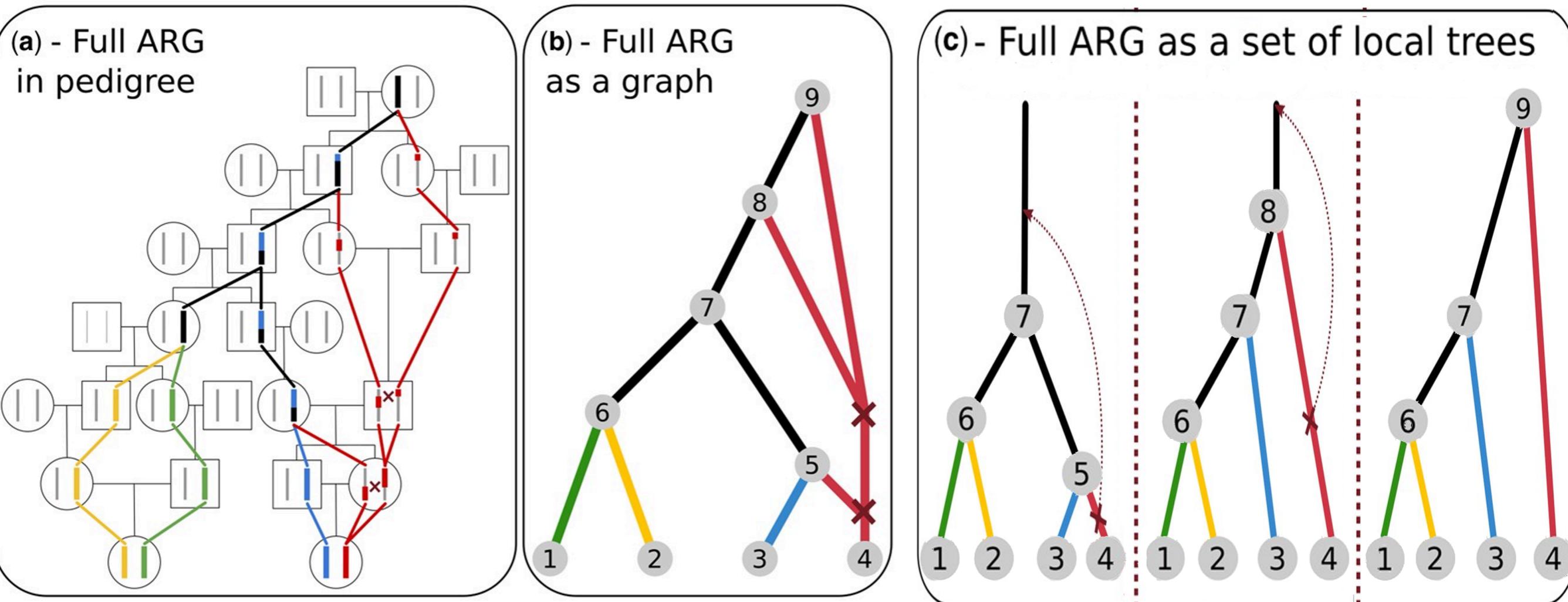


**Problem: most organisms (including ourselves) recombine much *more* than coronaviruses**

Once every generation!

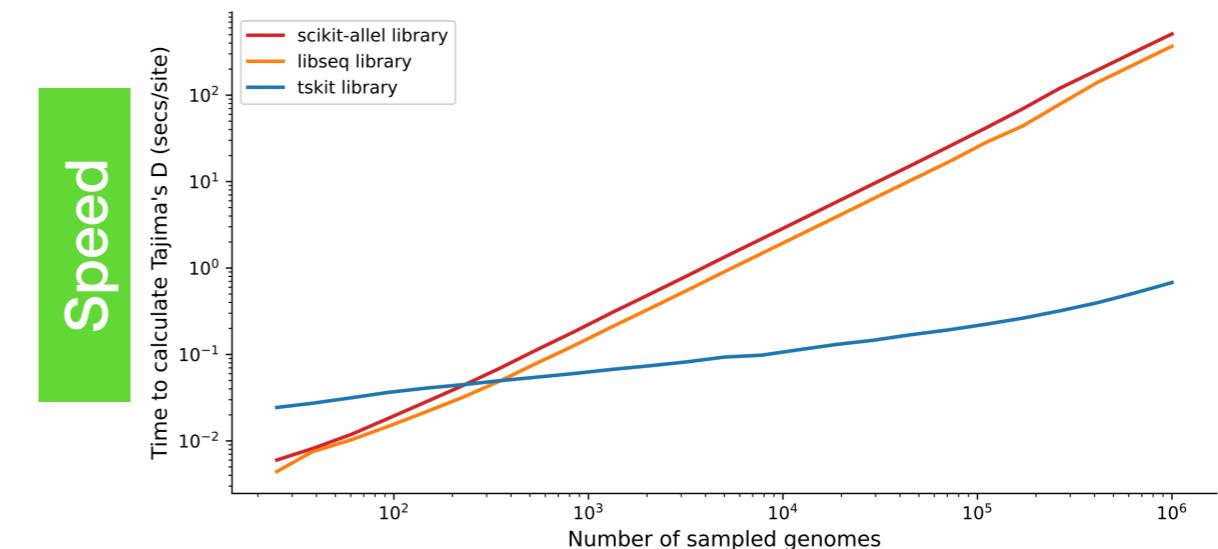
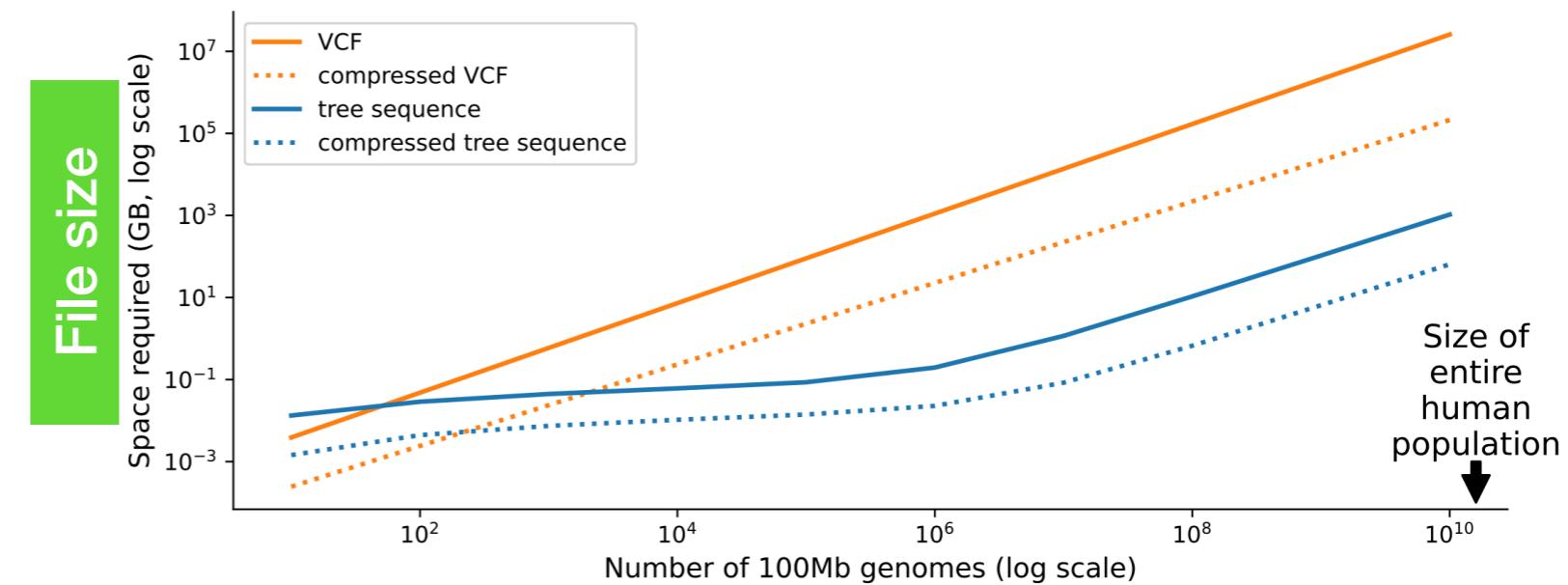


# A “genetic genealogy” tracks genome-wide inheritance paths



# Storing a genetic genealogy in the (succinct) tree sequence format has many benefits

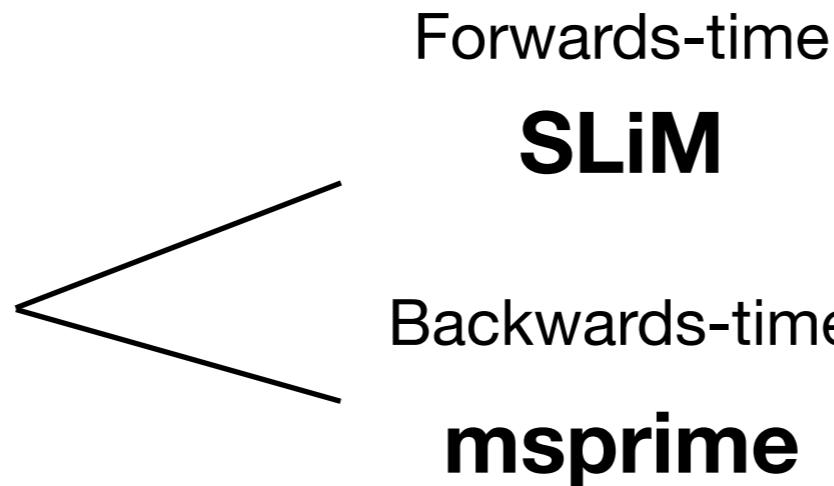
- **Compact storage**  
("domain specific compression")
- **Fast, efficient analysis**  
(a "succinct" structure)
- **Well tested, open source**  
(active dev community)
- **Built-in functionality**  
(well documented:  
<http://tskit.dev>)



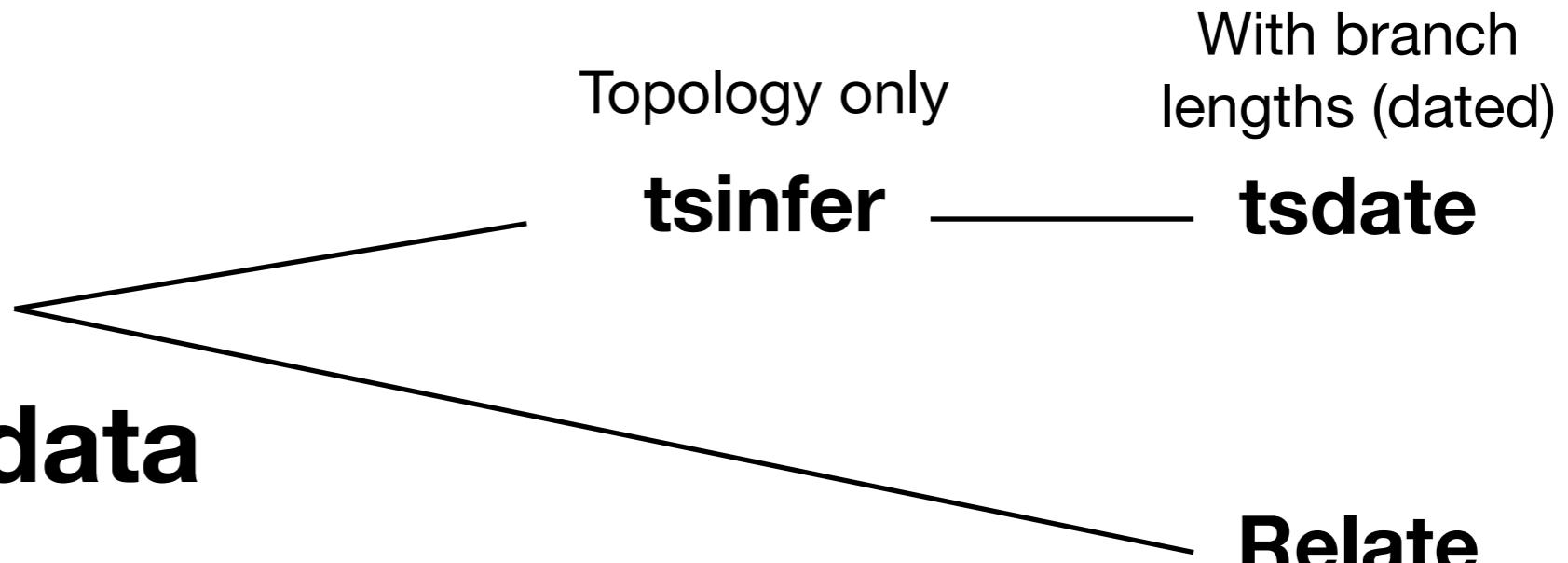
...but limited support for major genomic rearrangements (e.g. inversions, large indels): genomes should be (reasonably) aligned => current primary focus = **population genetics**

# How do we get hold of a tree sequence?

- **Simulation**



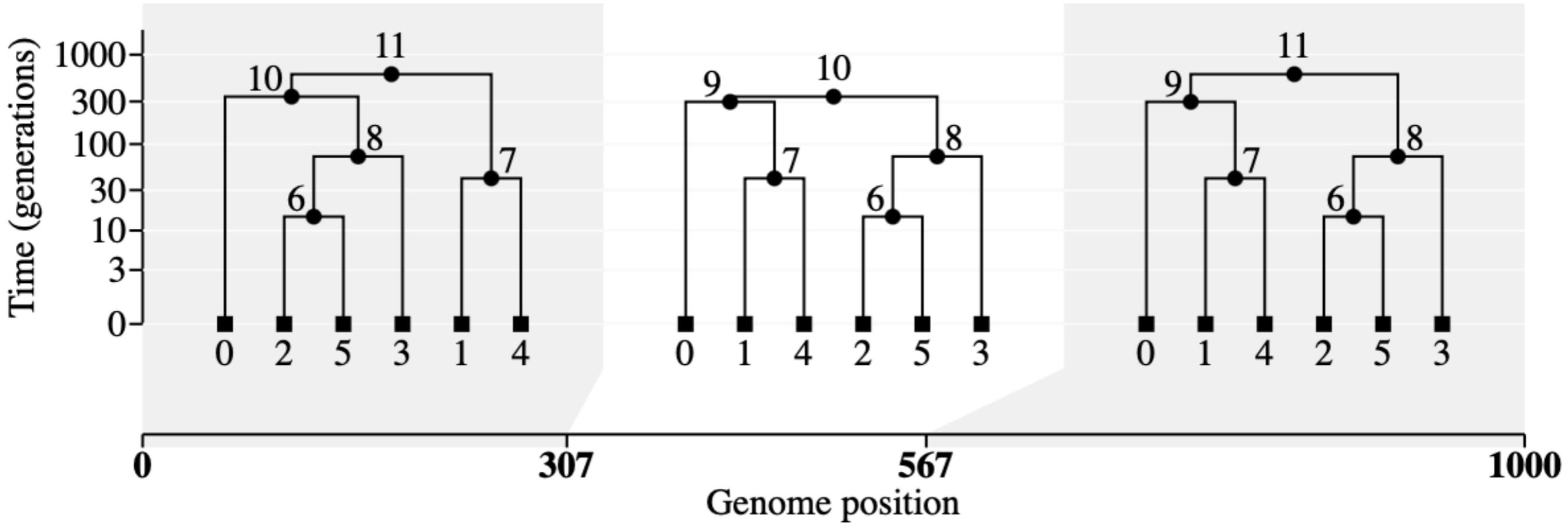
- **Inference  
from real data**



Also other inference programs that don't yet output in  
the tree sequence format (ARGweaver, Argneedle)

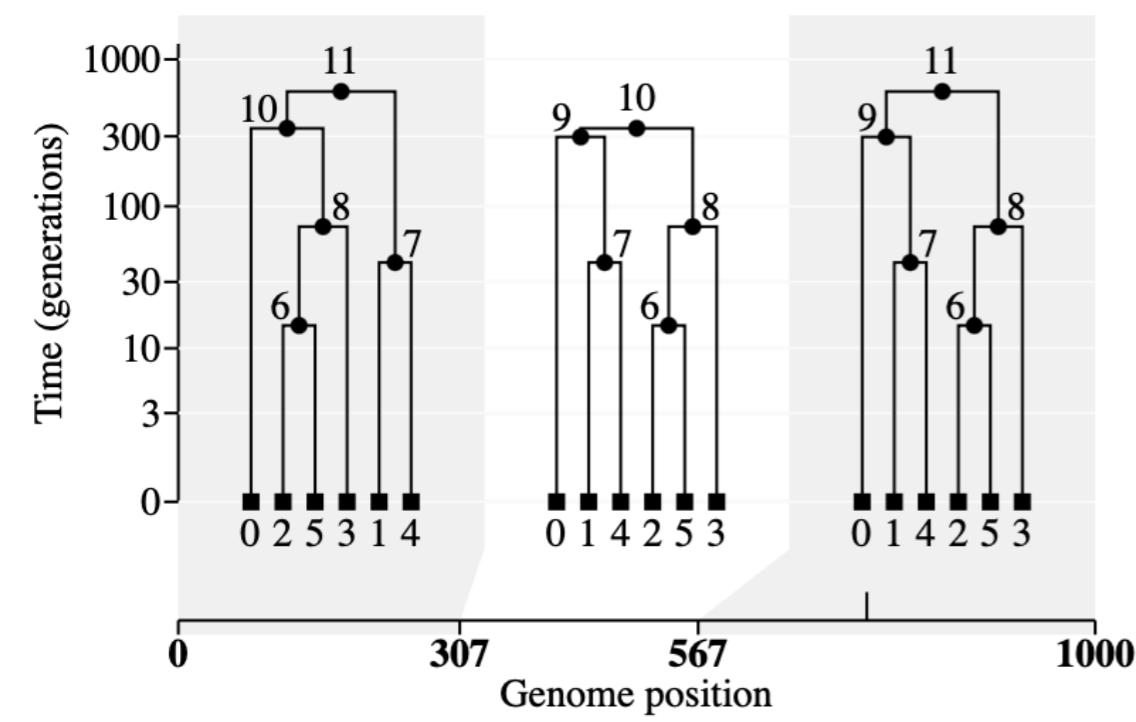
(but this tree seq is more like a  
sequence of separate trees)

# Tskit terminology: the basics

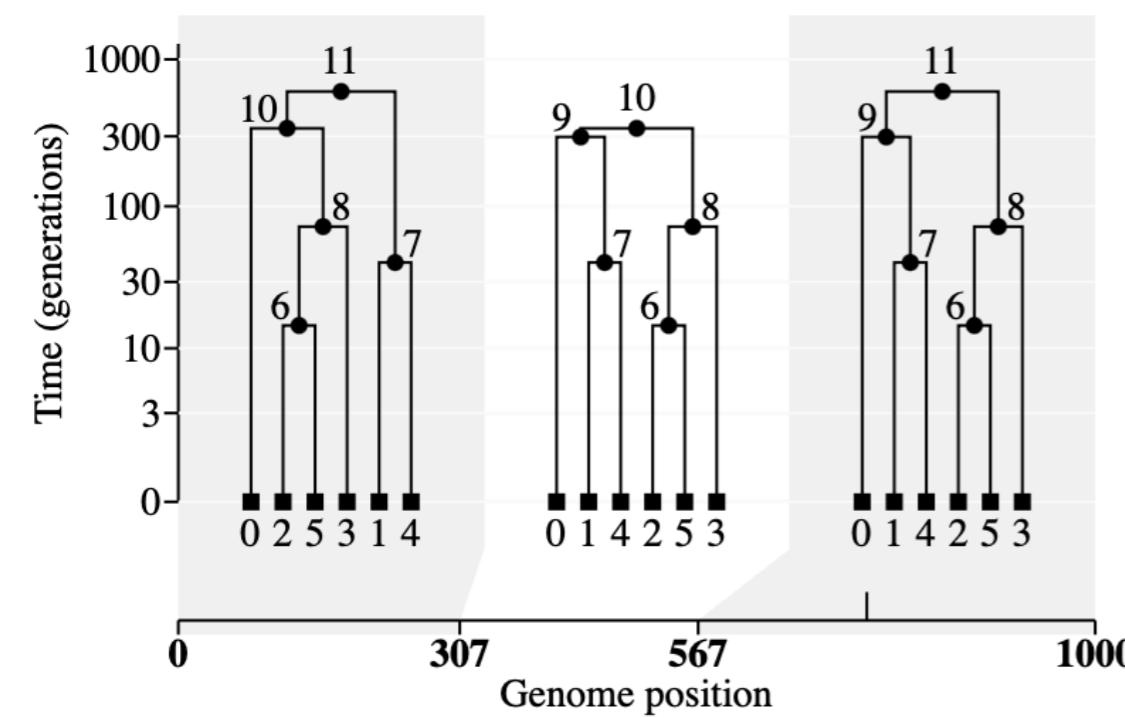


- Multiple **local trees** exist along a genome of fixed length (by convention measured in base pairs)
- Genomes exist at specific times, and are represented by **nodes** (the same node can persist across many local trees)
- Some nodes are most recent common ancestors (MRCAAs) of other nodes
- Entities are zero-based: the first node has id 0, the second id 1, ...

# Tskit terminology: nodes and edges



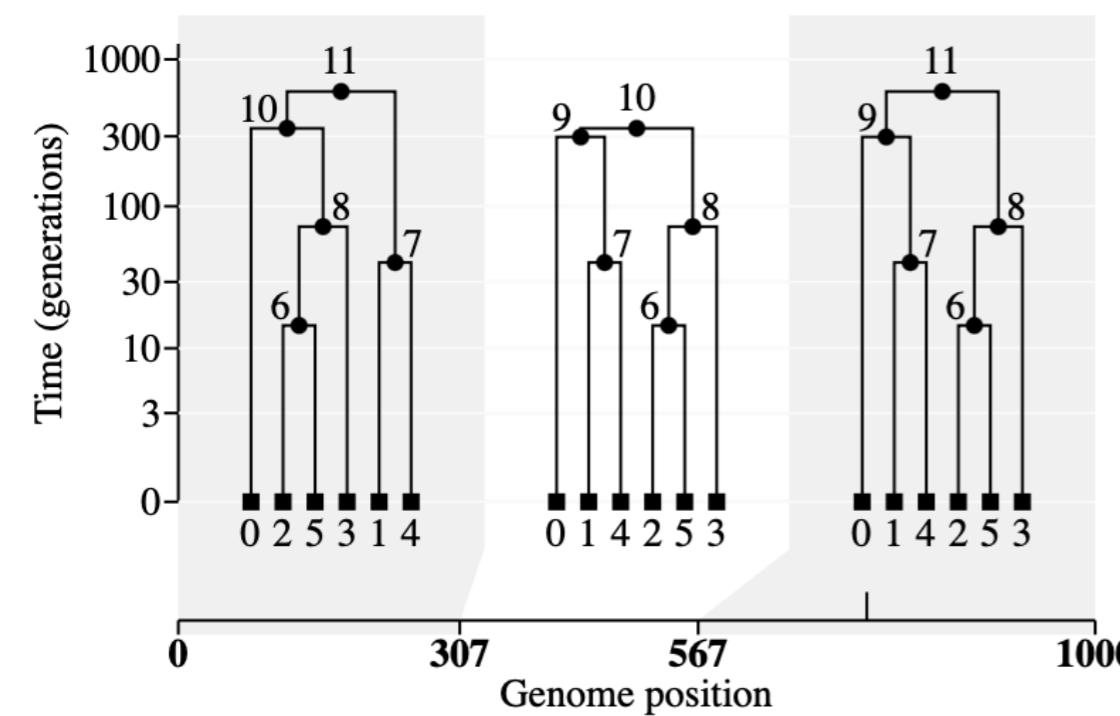
# Tskit terminology: nodes and edges



## Nodes (=genomes)

- exist at a specific *time*
- can be flagged as “samples”
- can belong to “*individuals*” (e.g. 2 nodes per individual in humans) and, if useful, “*populations*”

id	flags	population	individual	time
0	1	0	0	0.000000000
1	1	0	0	0.000000000
2	1	0	1	0.000000000
3	1	0	1	0.000000000
4	1	0	2	0.000000000
5	1	0	2	0.000000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866



## Nodes (=genomes)

- exist at a specific *time*
- can be flagged as “samples”
- can belong to “*individuals*” (e.g. 2 nodes per individual in humans)  
and, if useful, “*populations*”

id	flags	population	individual	time
0	1	0	0	0.000000000
1	1	0	0	0.000000000
2	1	0	1	0.000000000
3	1	0	1	0.000000000
4	1	0	2	0.000000000
5	1	0	2	0.000000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866

# Tskit terminology: nodes and edges

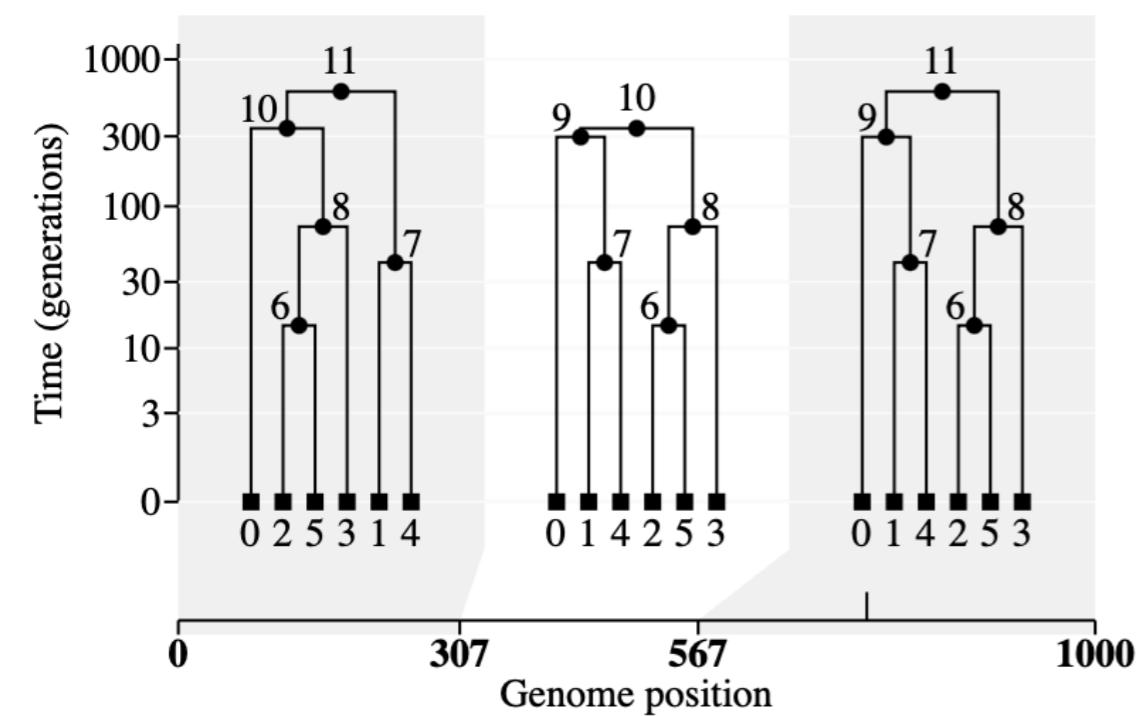
## Edges

Connect a *parent & child*

Have a *left & right* genomic coordinate

And usually *span* multiple trees,  
(e.g. edges connecting nodes 1+7 and 2+7)

id	left	right	parent	child
0	0	1000	6	2
1	0	1000	6	5
2	0	1000	7	1
3	0	1000	7	4
4	0	1000	8	3
5	0	1000	8	6
6	307	1000	9	0
7	307	1000	9	7
8	0	307	10	0
9	0	567	10	8
10	307	567	10	9



## Nodes (=genomes)

- exist at a specific *time*
- can be flagged as “samples”
- can belong to “*individuals*” (e.g. 2 nodes per individual in humans) and, if useful, “*populations*”

<b>id</b>	<b>flags</b>	<b>population</b>	<b>individual</b>	<b>time</b>
0	1	0	0	0.000000000
1	1	0	0	0.000000000
2	1	0	1	0.000000000
3	1	0	1	0.000000000
4	1	0	2	0.000000000
5	1	0	2	0.000000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866

# Tskit terminology: nodes and edges

## Edges

Connect a *parent & child*

Have a *left & right* genomic coordinate

And usually *span* multiple trees,  
(e.g. edges connecting nodes 1+7 and 2+7)

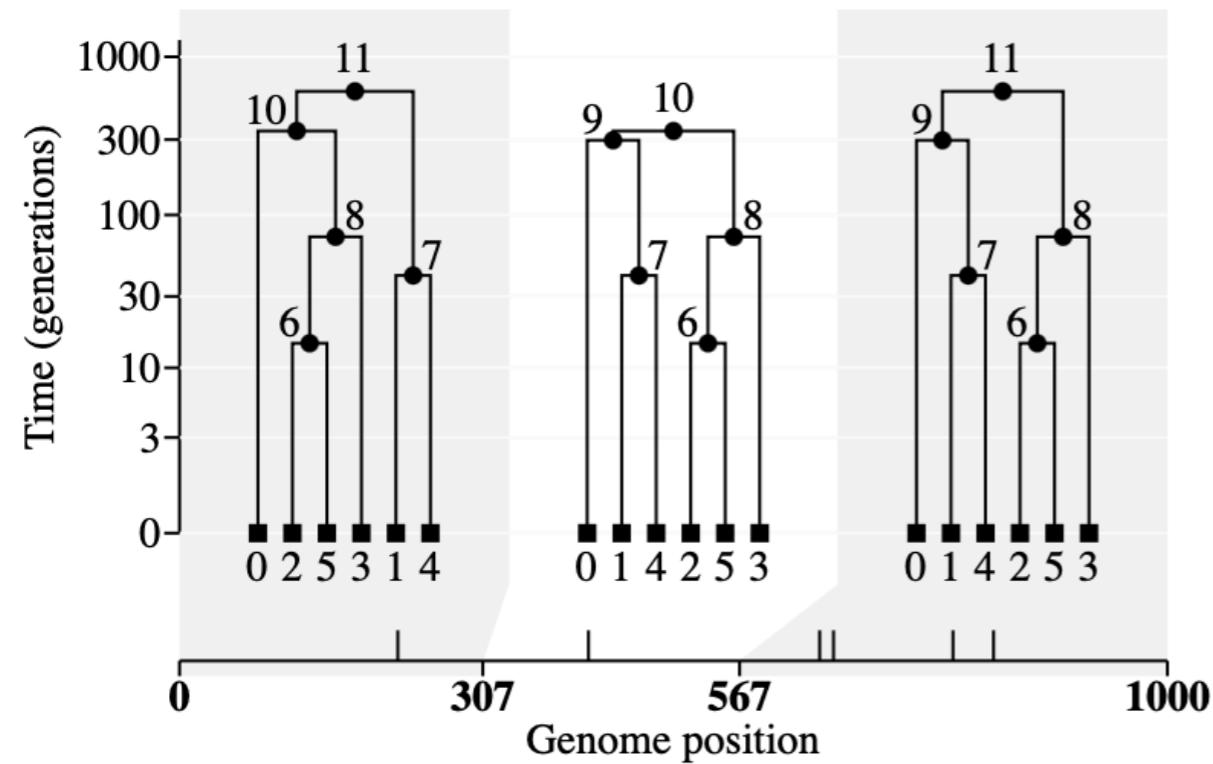
<b>id</b>	<b>left</b>	<b>right</b>	<b>parent</b>	<b>child</b>
0	0	1000	6	2
1	0	1000	6	5
2	0	1000	7	1
3	0	1000	7	4
4	0	1000	8	3
5	0	1000	8	6
6	307	1000	9	0
7	307	1000	9	7
8	0	307	10	0
9	0	567	10	8
10	307	567	10	9

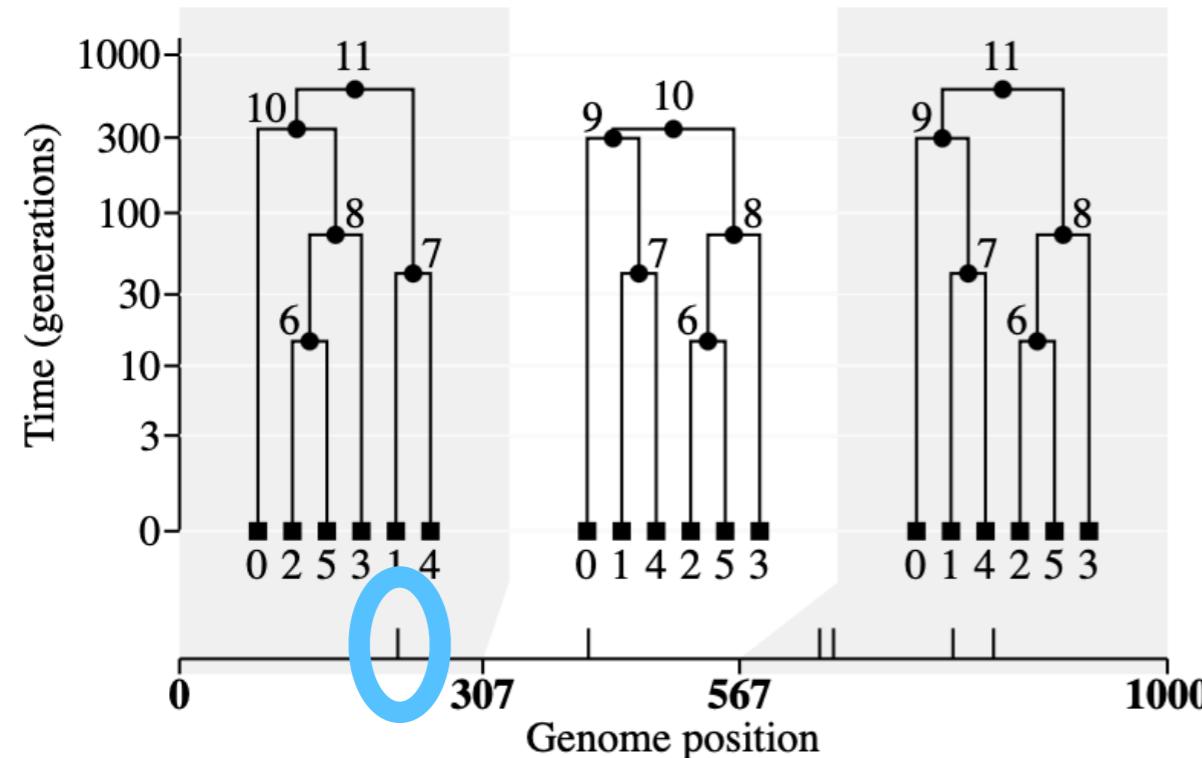
Number of edges  
= primary  
determinant of  
tree sequence  
filesize, and  
processing  
efficiency

(e.g UKBB chr20 ~  
1M sample nodes,  
62 M edges, tree  
iteration ~ 5secs)

# Tskit terminology: sites & mutations

This is how we can encode genetic variation  
Most genomic positions do not vary between  
genomes: usually we don't bother tracking these





# Tskit terminology: sites & mutations

This is how we can encode genetic variation  
Most genomic positions do not vary between  
genomes: usually we don't bother tracking these

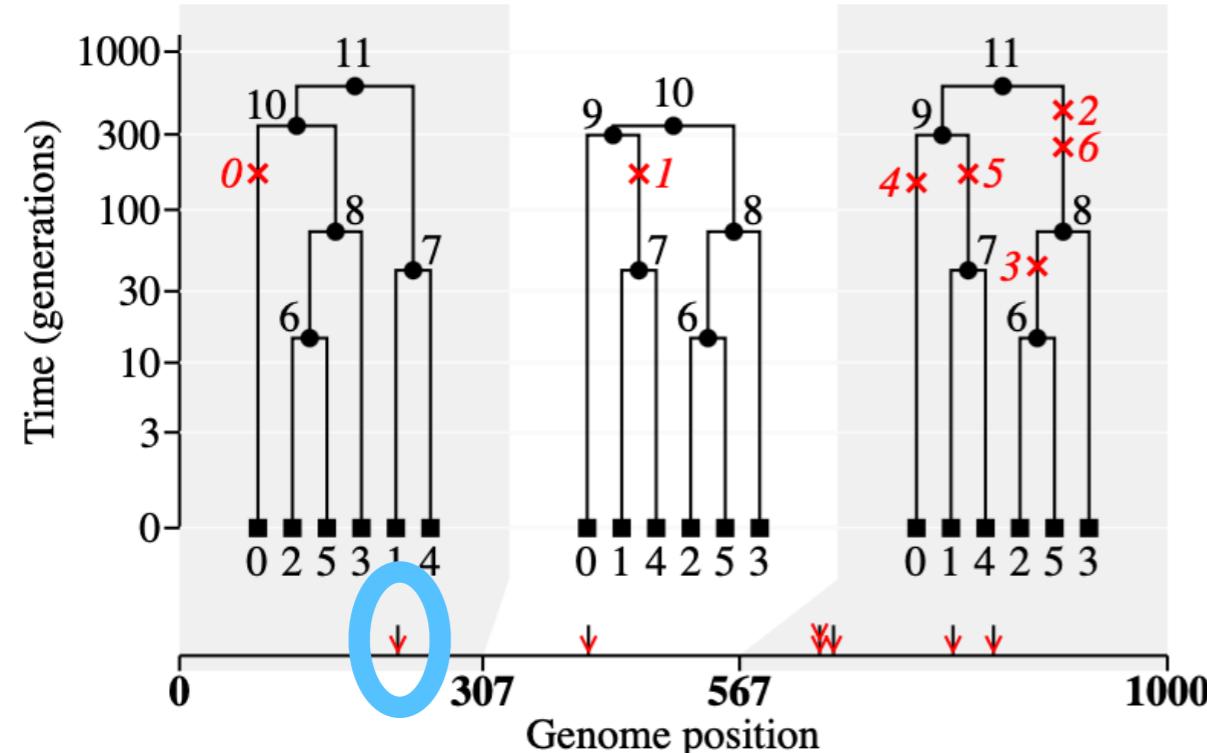
We can create a **site** at a given genomic **position** with a fixed **ancestral state**.

<b>id</b>	<b>position</b>	<b>ancestral_state</b>
0	221	T
1	414	G
2	648	A
3	662	G
4	783	T
5	824	C

**Sites table**

# Tskit terminology: sites & mutations

This is how we can encode genetic variation  
Most genomic positions do not vary between  
genomes: usually we don't bother tracking these



We can create a **site** at a given genomic **position** with a fixed **ancestral state**.

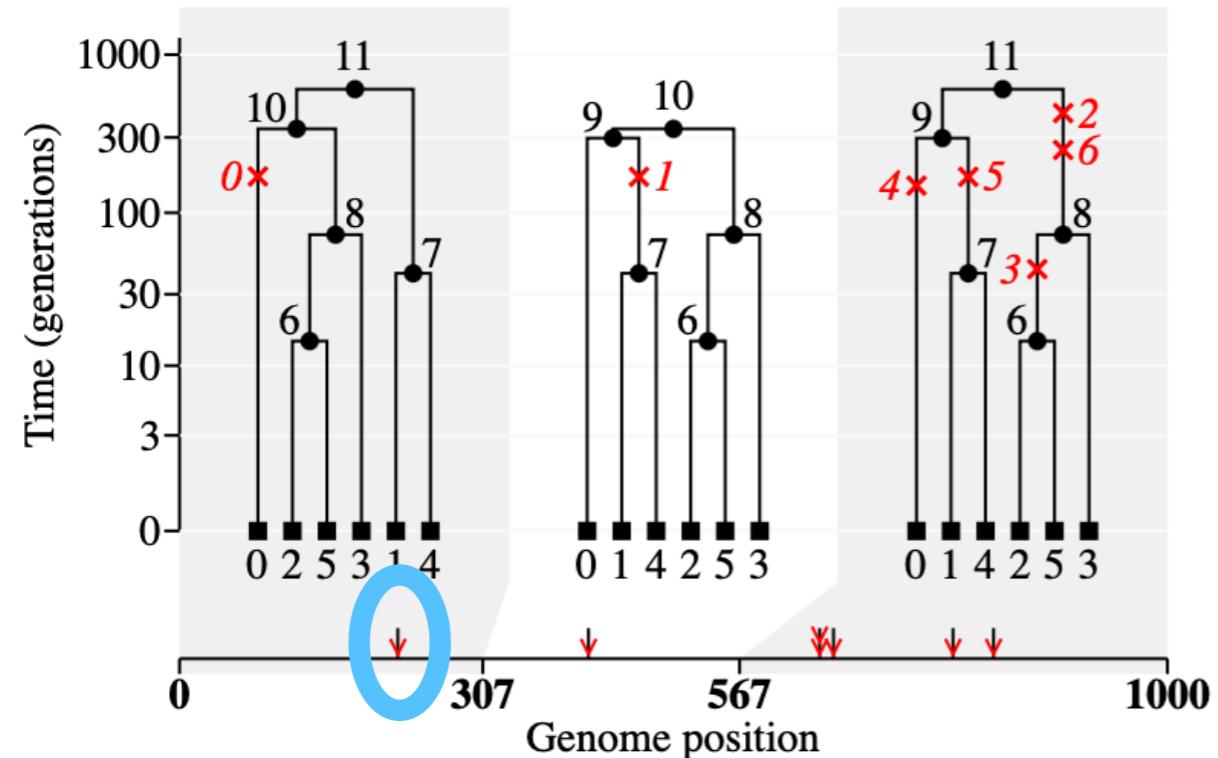
Normally, a site is created in order to place one or more **mutations** at that site

id	position	ancestral_state
0	221	T
1	414	G
2	648	A
3	662	G
4	783	T
5	824	C

Sites table

id	site	node	time	derived_state	parent
0	0	0	0	nan	-1
1	1	7	nan	T	-1
2	2	8	nan	T	-1
3	2	6	nan	C	2
4	3	0	nan	C	-1
5	4	7	nan	C	-1
6	5	8	nan	T	-1

Mutations table



# Tskit terminology: sites & mutations

This is how we can encode genetic variation  
Most genomic positions do not vary between genomes: usually we don't bother tracking these

We can create a **site** at a given genomic **position** with a fixed **ancestral state**.

Normally, a site is created in order to place one or more **mutations** at that site

<b>id</b>	<b>position</b>	<b>ancestral_state</b>
0	221	T
1	414	G
2	648	A
3	662	G
4	783	T
5	824	C

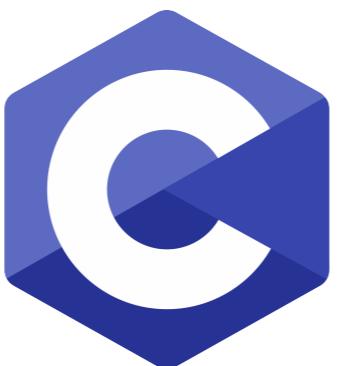
Sites table

<b>id</b>	<b>site</b>	<b>node</b>	<b>time</b>	<b>derived_state</b>	<b>parent</b>
0	0	0	0	nan	A -1
1	1	7	nan	T	-1
2	2	8	nan	T	-1
3	2	6	nan	C	2
4	3	0	nan	C	-1
5	4	7	nan	C	-1
6	5	8	nan	T	-1

Mutations table

But actually you can do lots of analysis even if you don't have sites & mutations:  
[https://tskit.dev/tutorials/no\\_mutations.html](https://tskit.dev/tutorials/no_mutations.html))

# Using tskit



**The Rust  
Programming  
Language**

# Using tskit



The Rust  
Programming  
Language

Docs and tutorials

<https://tskit.dev/tskit/docs>

<https://tskit.dev/tutorials>

# Using tskit

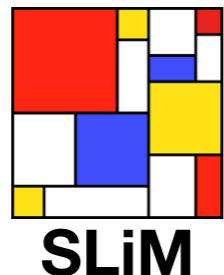
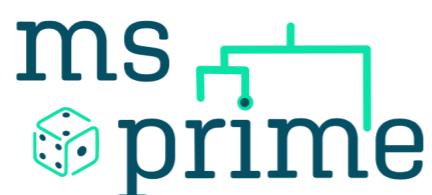


The Rust  
Programming  
Language

Docs and tutorials

<https://tskit.dev/tskit/docs>

<https://tskit.dev/tutorials>



SLiM



tsinfer



# Using tskit

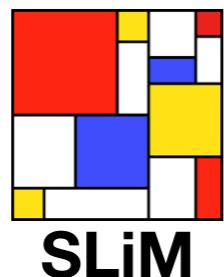
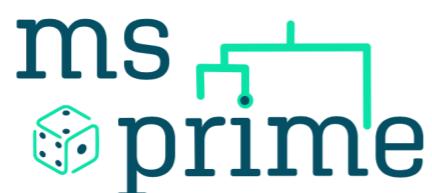


The Rust  
Programming  
Language

Docs and tutorials

<https://tskit.dev/tskit/docs>

<https://tskit.dev/tutorials>



SLiM



tsinfer

New web site soon  
<https://tskit.dev>

# Research-enabling:



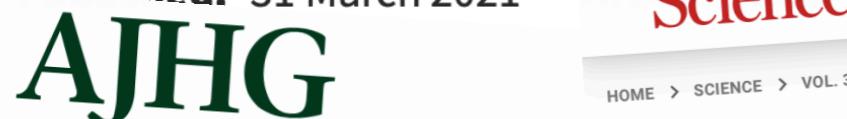
LETTER | Open Access | CC BY  
Mutation load decreases with haplotype age in wild Soay sheep  
Martin A. Stoffel, Susan E. Johnston, Jill G. Pilkington, Josephine M. Pemberton  
DOI: 10.1002/evl3.229 | Citations: 3

## The timing of human adaptation from Neanderthal introgression FREE

Sivan Yair, Kristin M Lee, Graham Coop

Genetics, Volume 218, Issue 1, May 2021, iyab052,  
<https://doi.org/10.1093/genetics/iyab052>

Published: 31 March 2021



Volume 100, Issue 4, 6 April 2017, Pages 635-649

Article

## Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin <sup>1, 2, 3, 4</sup>, Christopher R. Gignoux <sup>4</sup>, Raymond K. Walters <sup>1, 2, 3</sup>,...  
M. Neale <sup>1, 2, 3</sup>, Simon Gravel <sup>5, 6</sup>, Mark J. Daly <sup>1, 2, 3</sup>, Carlos D.

## Analysis of genetic dominance in the UK Biobank

Duncan S. Palmer, Wei Zhou, Liam Abbott, Nikolas Baya, Claire Churchhouse, Cotto Daniel King, Masahiro Kanai, Alex Bloemendaal, Benjamin M. Neale  
doi: <https://doi.org/10.1101/2021.08.15.456387>

**nature**

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open Access | Published: 12 January 2022

**Mutation bias reflects natural selection in *Arabidopsis thaliana***

Gre Monroe, Thavni Srikant, Pablo Carbonell-Bejerano, Claude Becker, Marieke Lensink, Moises Klein, Julia Hildebrandt, Manuela Neumann, Daniel Kliebenstein, Mao-Lun Ågren, Matthew T. Rutter, Charles B. Fenster & Detlef Weigel

Computational and Systems Biology, Genetics and Genomics

**Detecting adaptive introgression using convolutional neural networks**

Matteo Fumagalli, Fernando Racimo, Institute, Faculty of Health and Medical Sciences, University Campus, Imperial College London, United Kingdom

Submit manuscript

Current Issue First release papers Archive About

HOME > SCIENCE > VOL. 375, NO. 6583 > A UNIFIED GENEALOGY OF MODERN AND ANCIENT GENOMES

RESEARCH ARTICLE | HUMAN EVOLUTION

ANTHONY WILDER WOHNS, YAN WONG, BEN JEFFERY, ALI AKBARI, [...] GIL MCVEAN +6 authors Authors Info & Affiliations

eLife

HOME MAGAZINE COMMUNITY INNOVATION

Research Article Epidemiology and Global Health, Genetics and Genomics

**Demographic history mediates the effect of stratification on polygenic scores**

Arslan A Zaidi, Iain Mathieson, Department of Genetics, Perelman School of Medicine, University of Pennsylvania, United States

DOI: <https://doi.org/10.1101/2021.08.15.456387> | CC BY