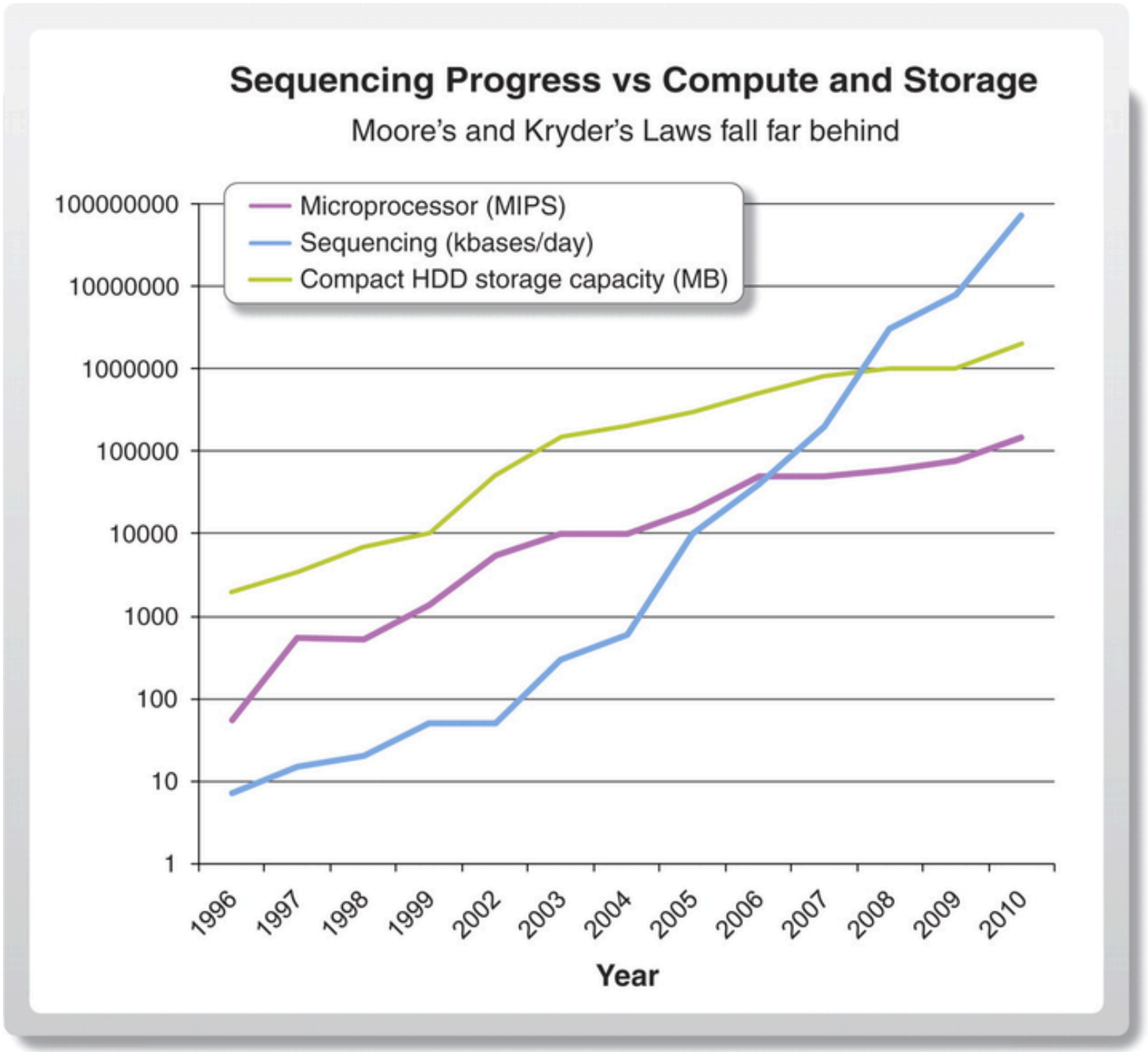
Genealogical analysis

Trees, tree sequences, and tskit

Genomic data is big

and tedious to analyse



Kahn, S: 2011 (!)
doi: 10.1126/science.119789

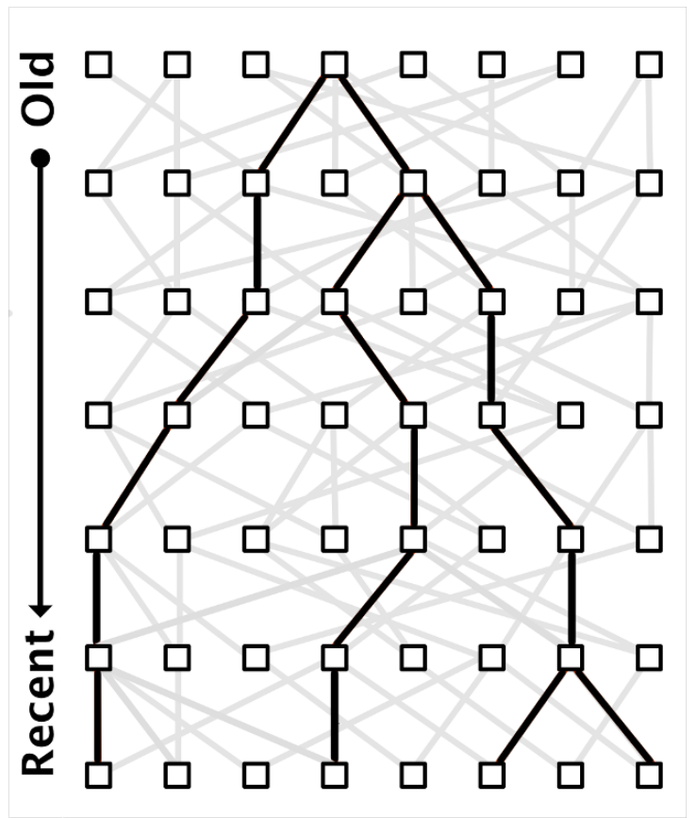
How can the right data structure help?

(1) Succinct

- stores each bit of information once …
- … but still allows efficient query operations

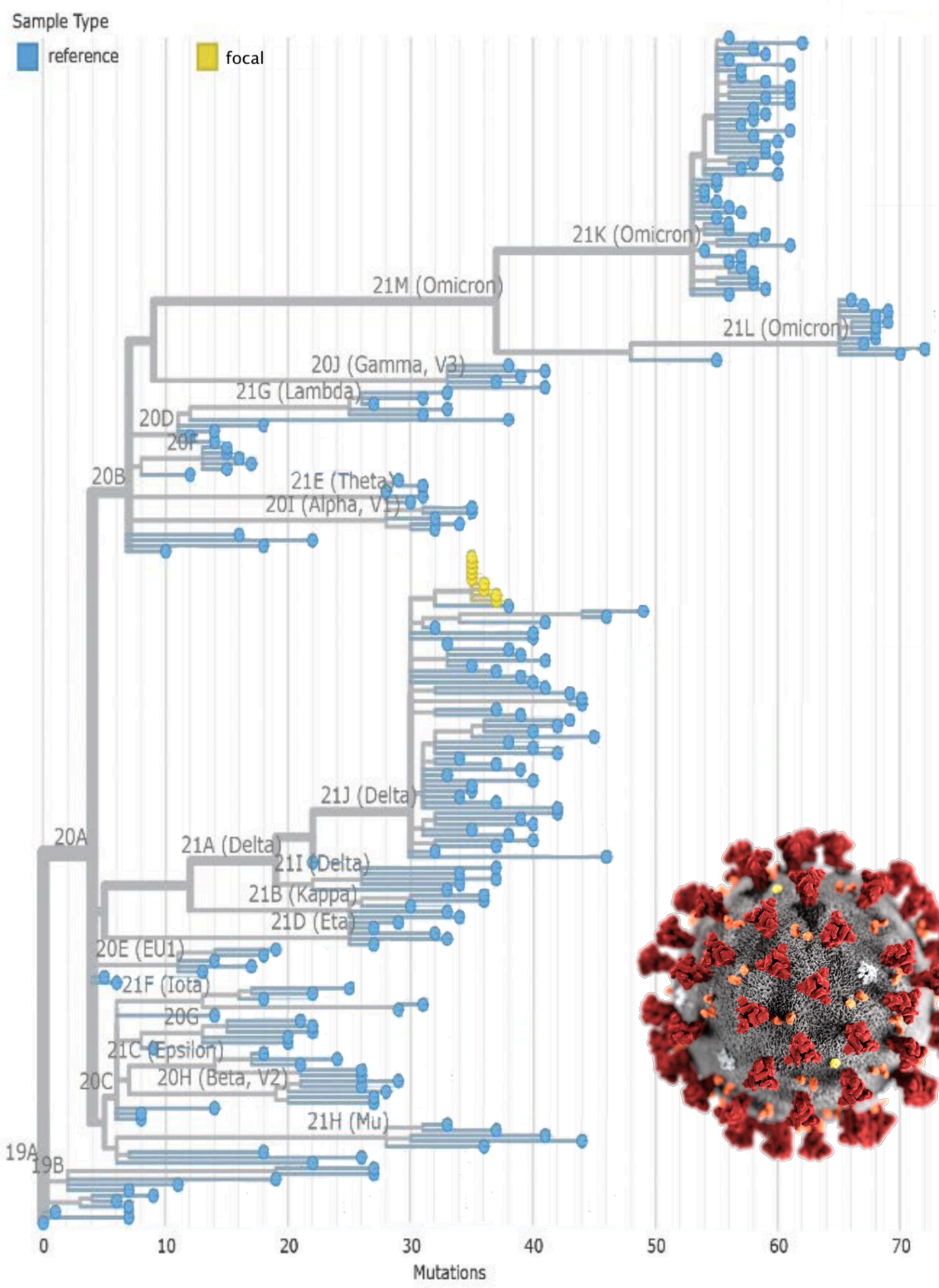
(2) Descriptive

- reflects underlying generative process

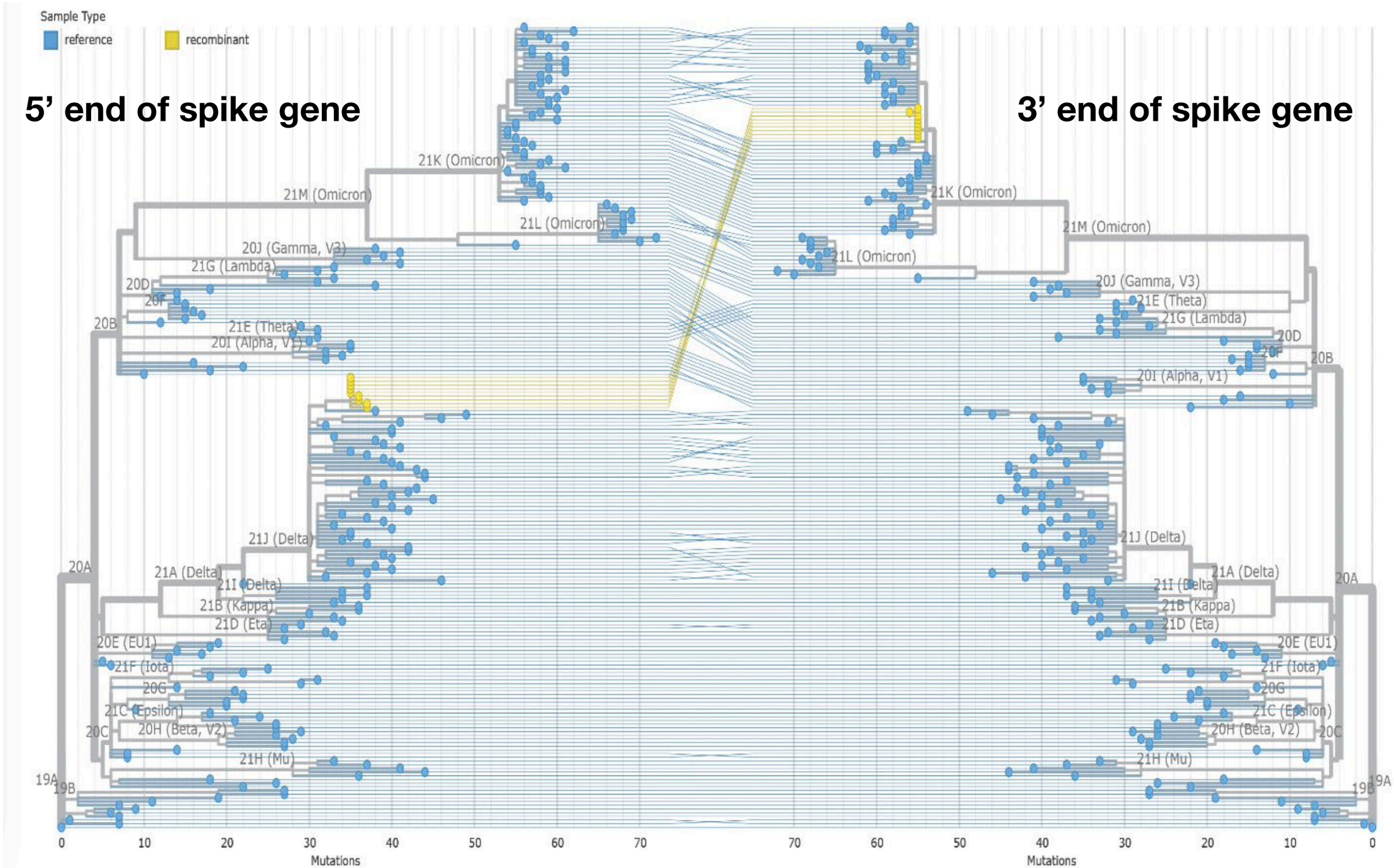


Trees are fundamental in comparing & analysing genetic sequences…

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

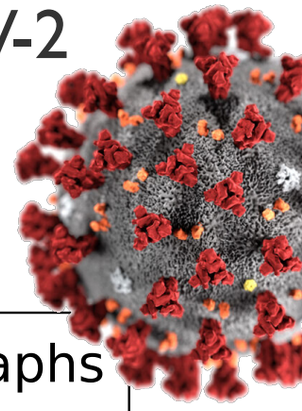


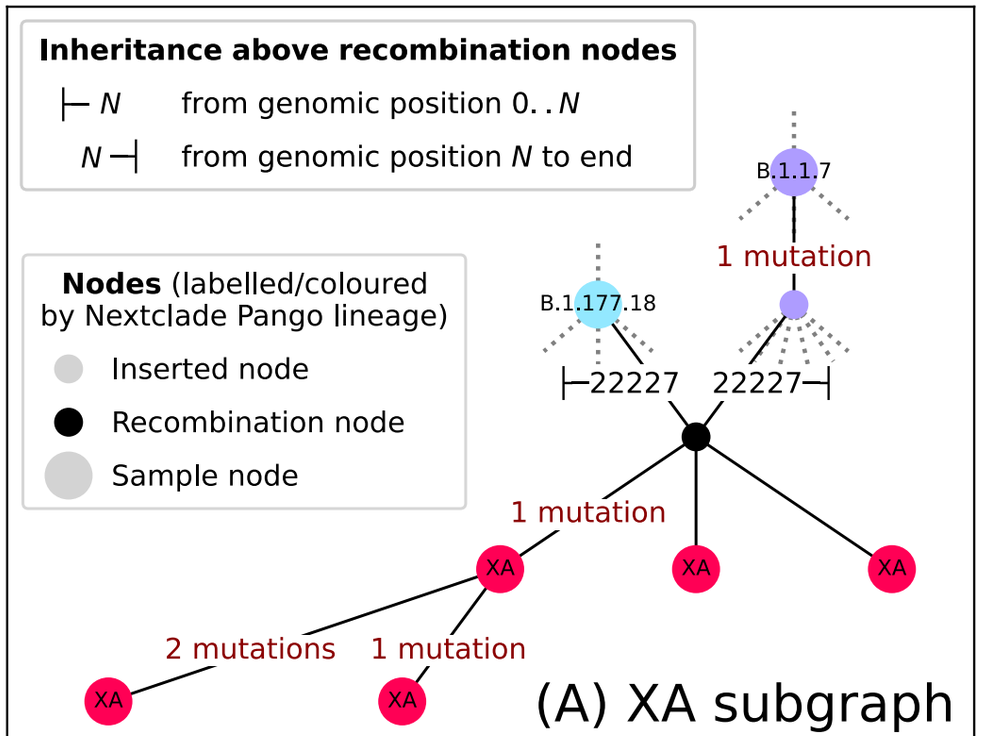
… but recombination causes different genetic regions to have different trees

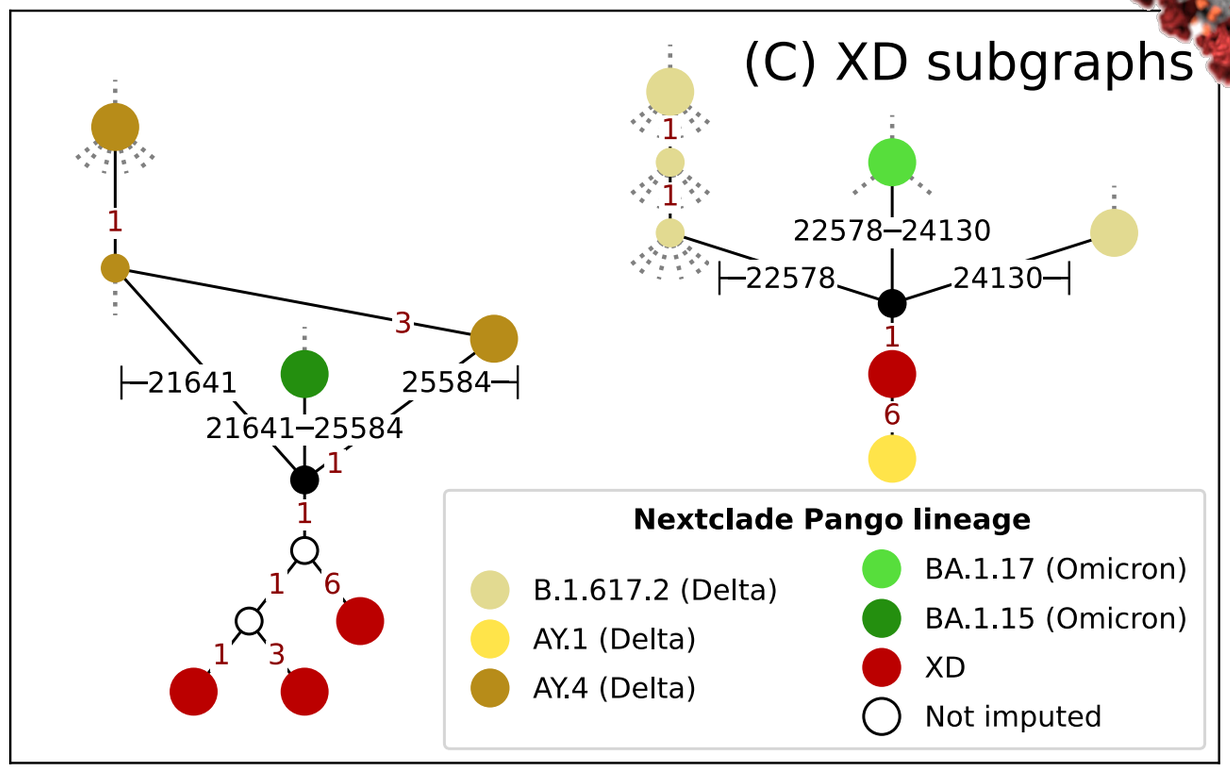


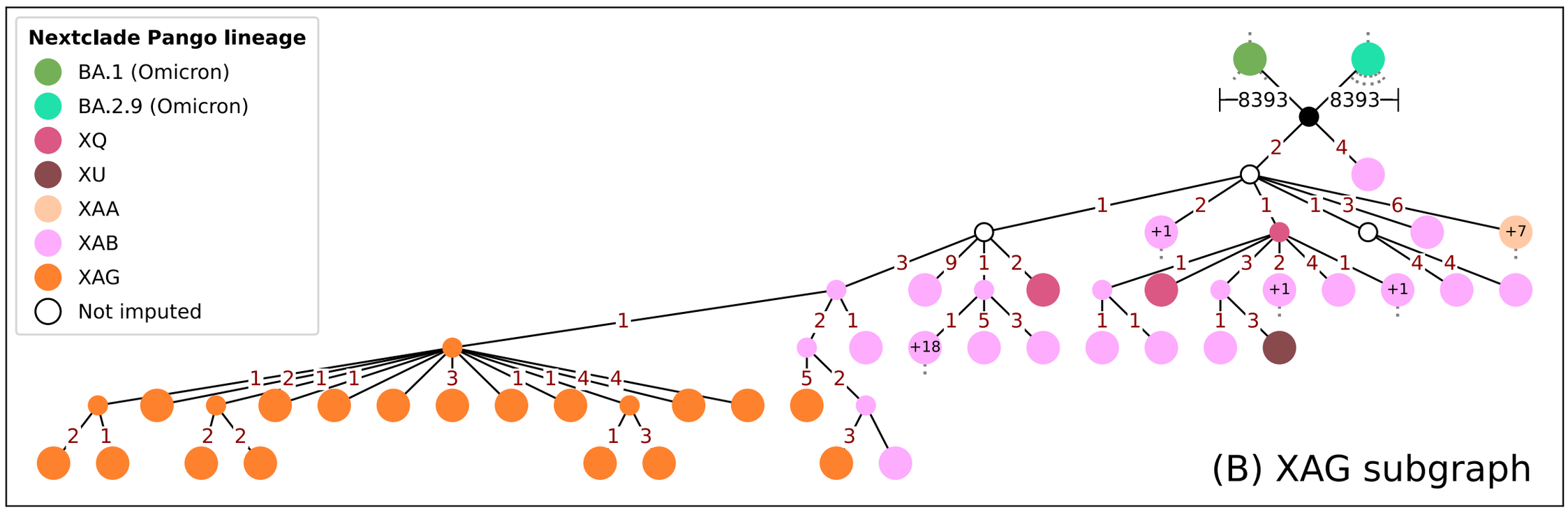
Towards Pandemic-Scale Ancestral Recombination Graphs of SARS-CoV-2

Zhan *et al.* (2023): https://doi.org/10.1101/2023.06.08.544212



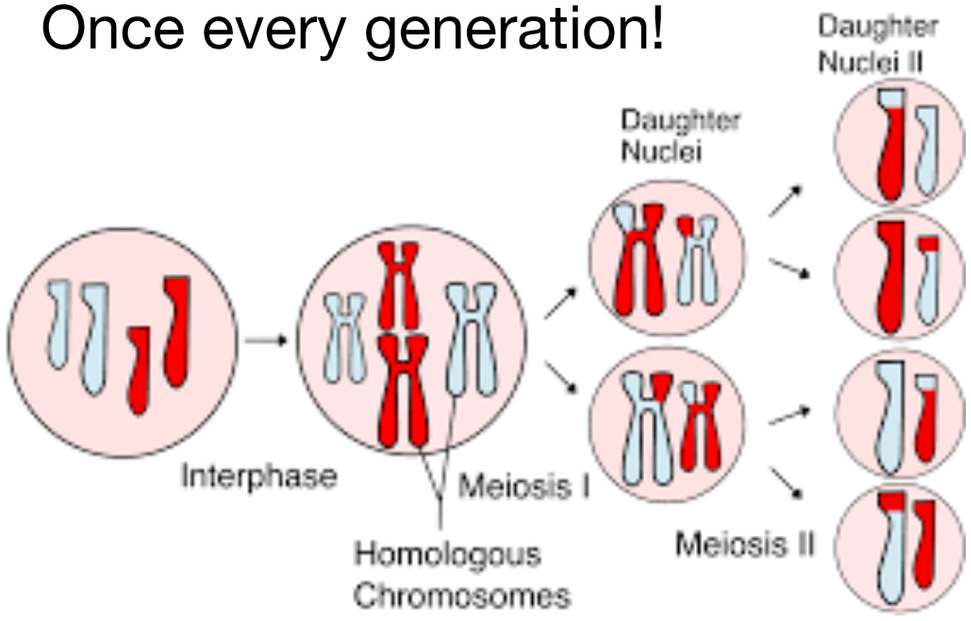






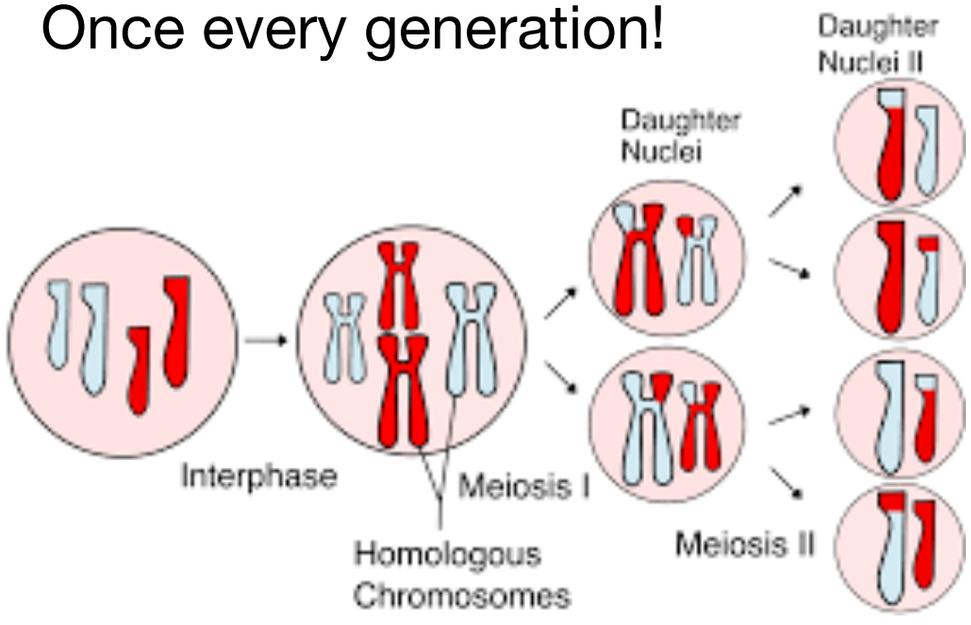
Problem: most organisms (including ourselves) recombine much *more* that coronaviruses

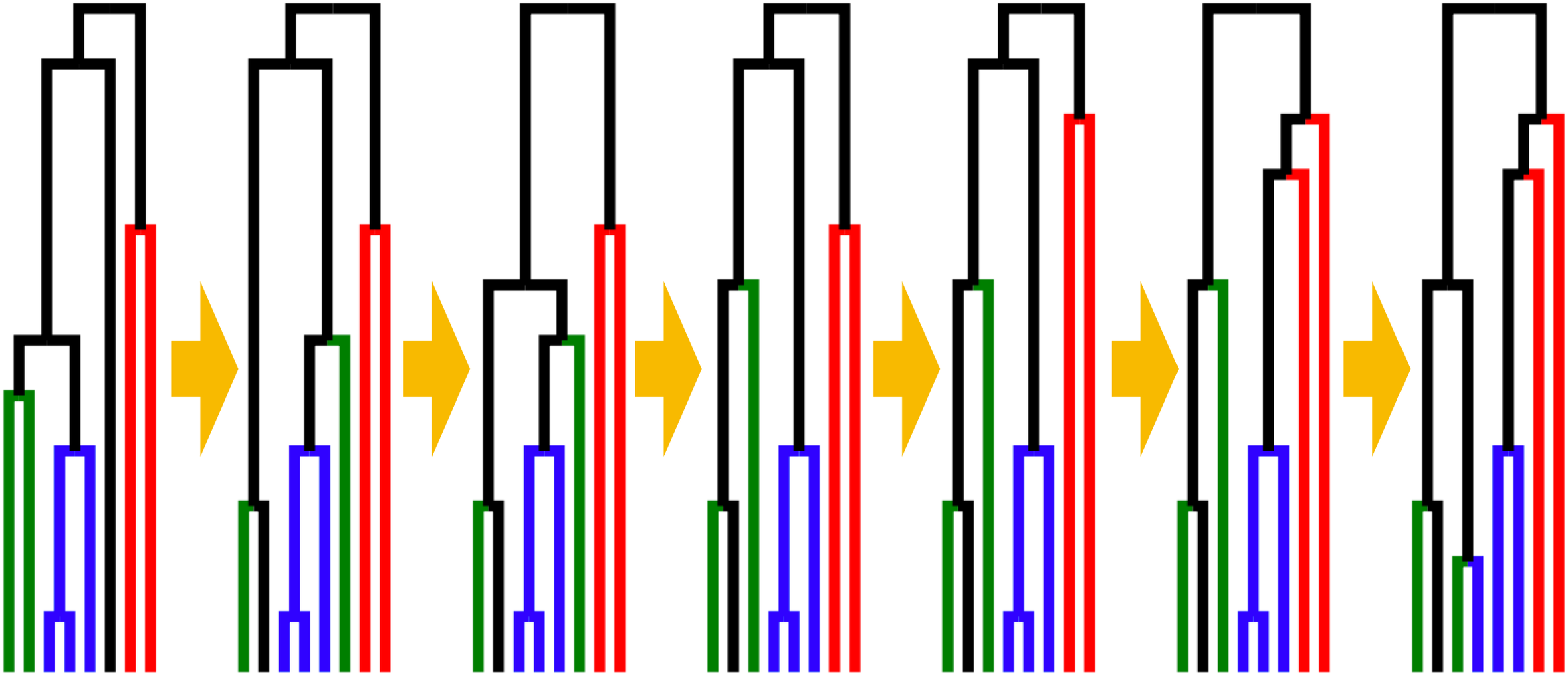
Once every generation!



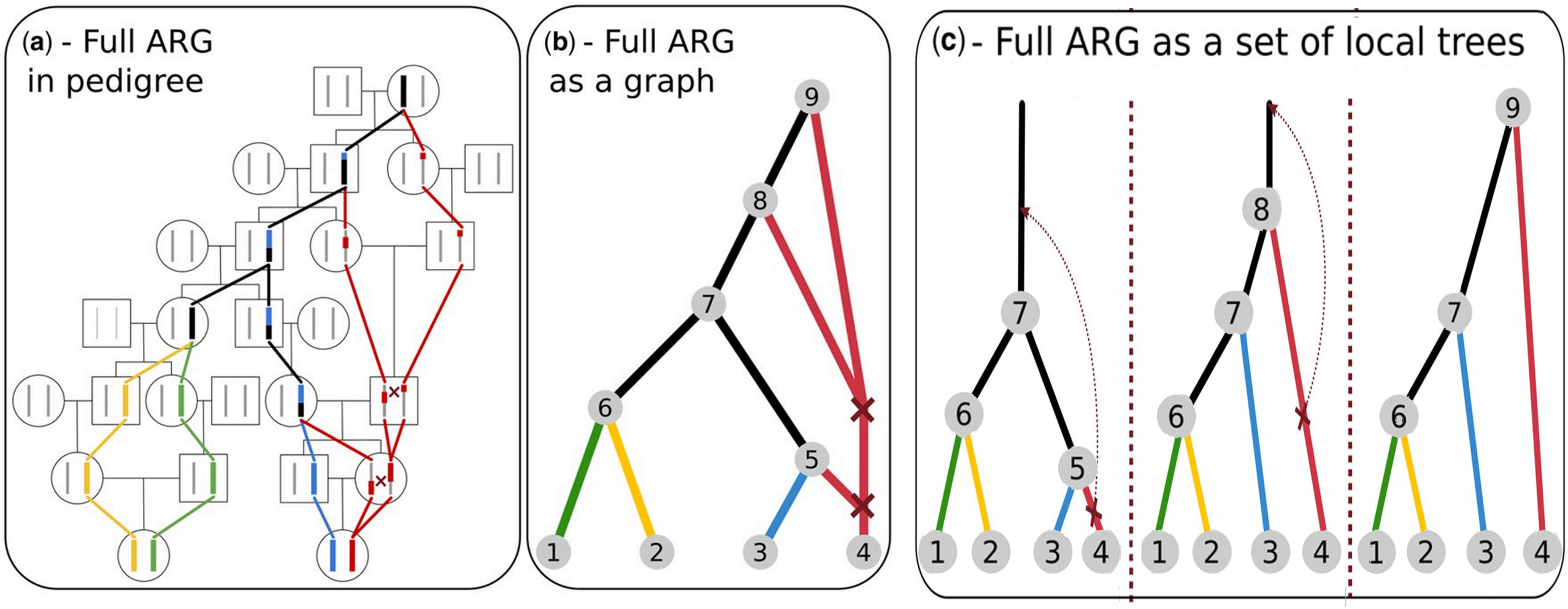
Problem: most organisms (including ourselves) recombine much *more* that coronaviruses

Once every generation!





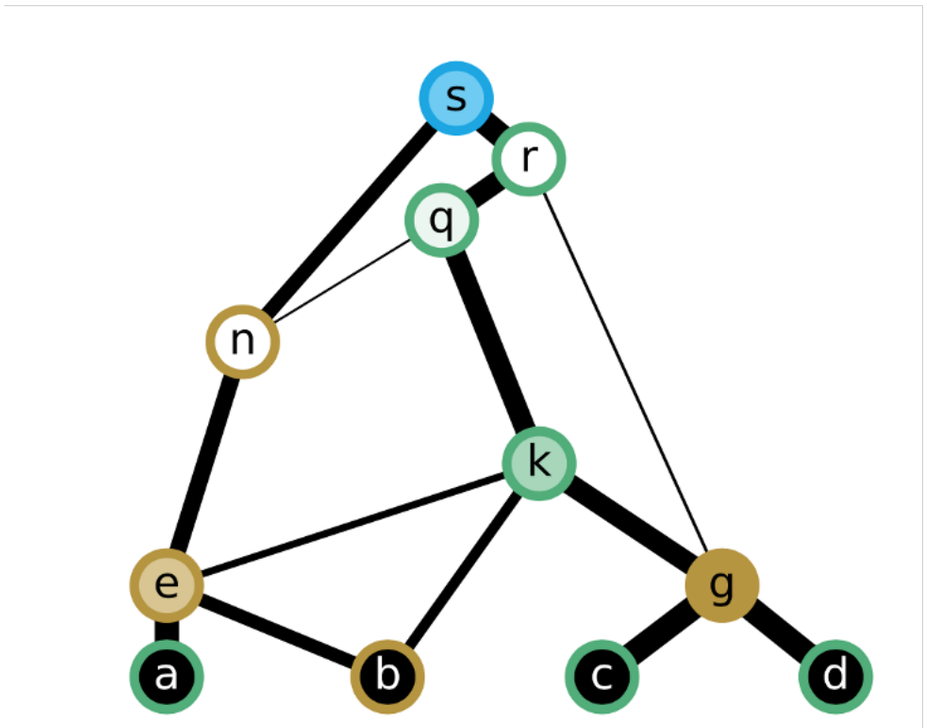
A “genetic genealogy” tracks genome-wide inheritance paths

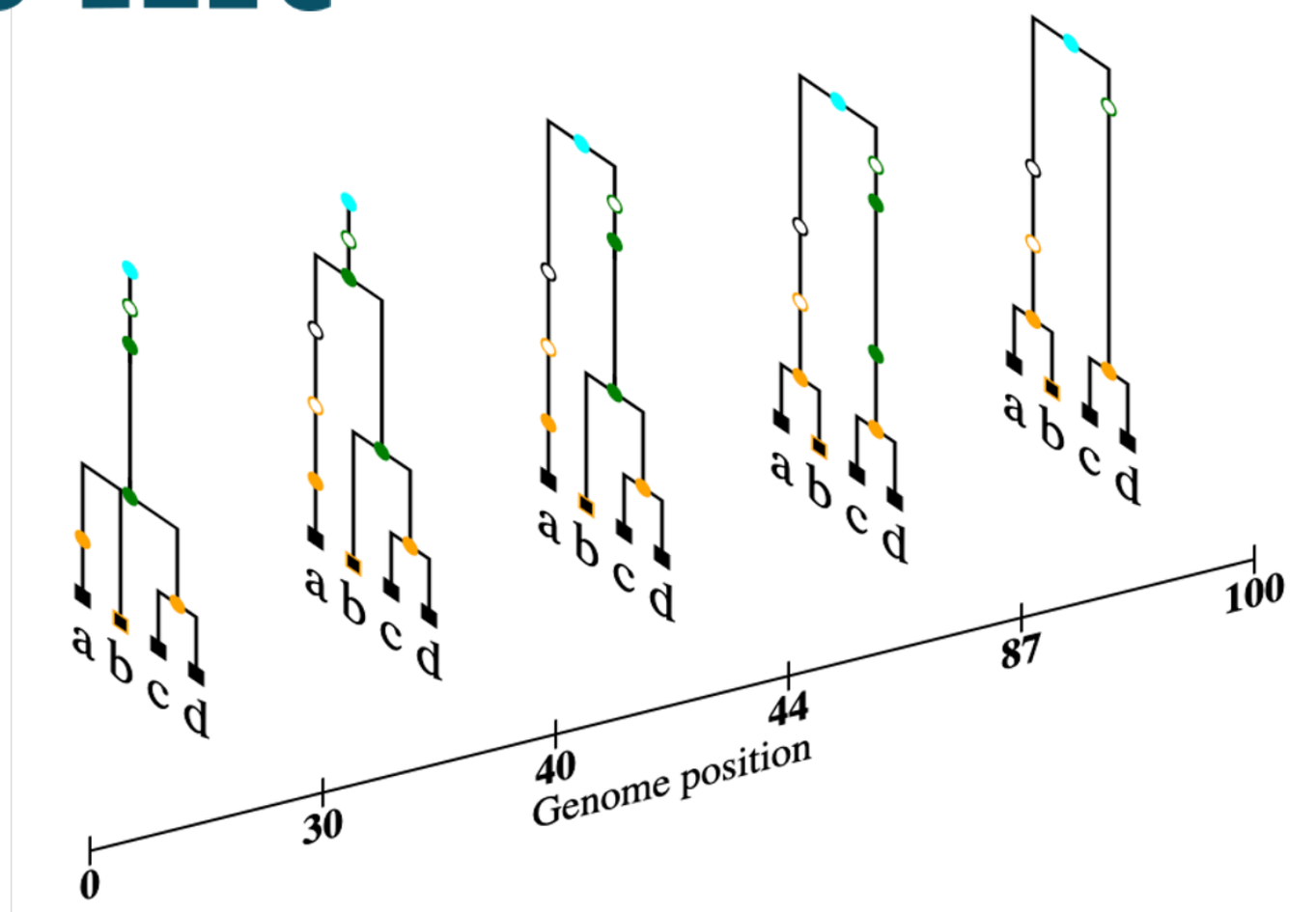


“succinct tree sequence format”

from: Evaluation of methods for estimating coalescence times using ancestral recombination graphs (Brandt *et al.*, **Genetics.** 2022: https://doi.org/10.1093/genetics/iyac044)

ts kit





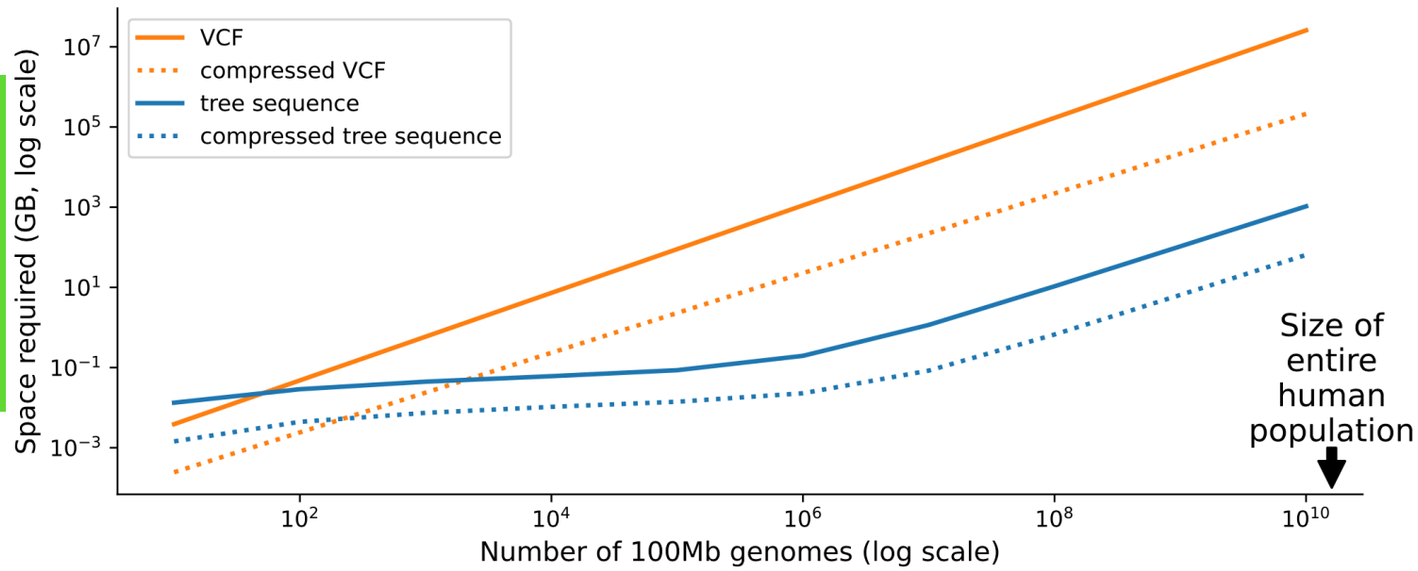
Graph representation

Local tree representation

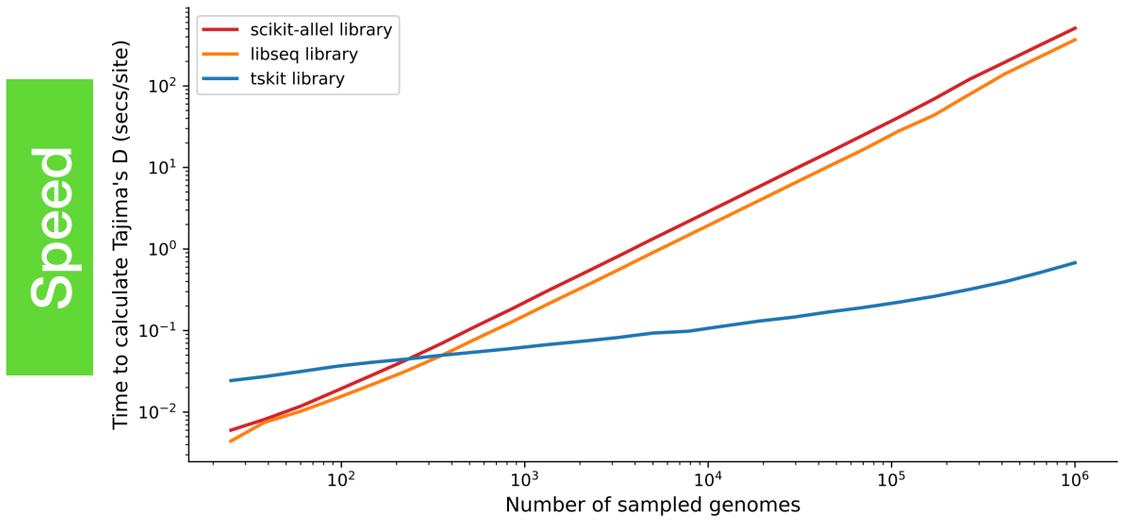
Storing a genetic genealogy in the (succinct) tree sequence format has many benefits

- **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)

File size



Speed



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

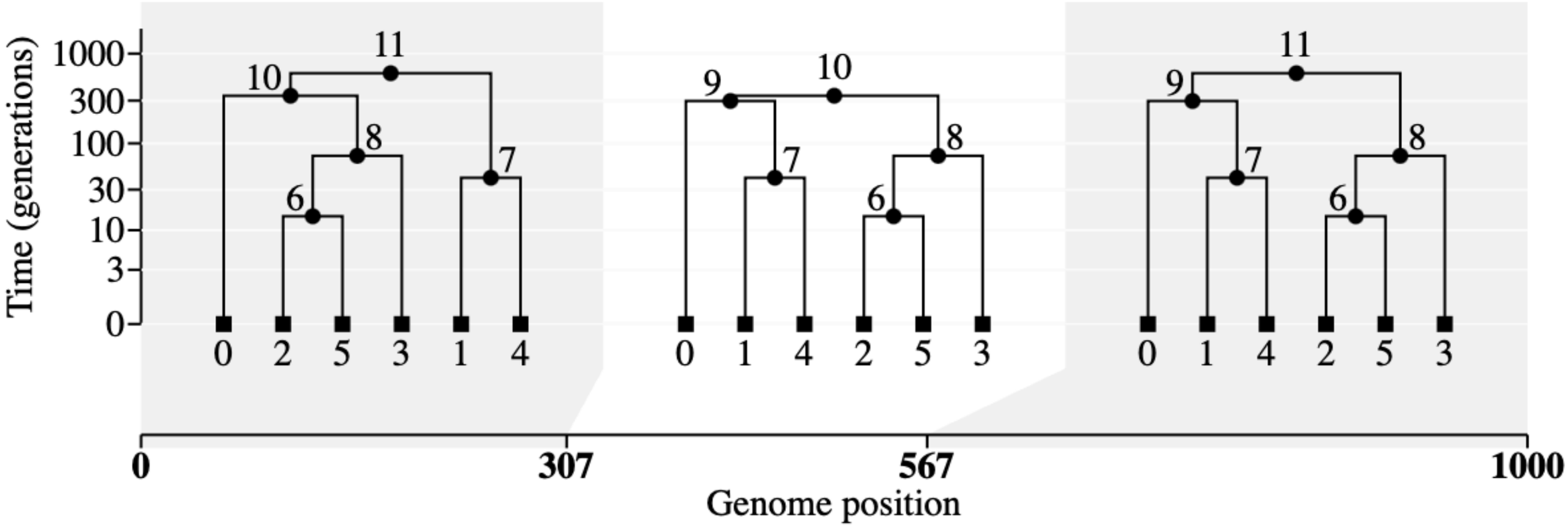
Images from online tutorials: "What is a tree sequence": https://tskit.dev/tutorials/what_is.html

How do we get hold of a tree sequence?

- **Simulation**
  - Forwards-time: **SLiM**
  - Backwards-time: **msprime**
- **Inference from real data**
  - Topology only: **tsinfer** — With branch lengths (dated): **tsdate**
  - **Relate** (but this tree seq is more like a sequence of separate trees)

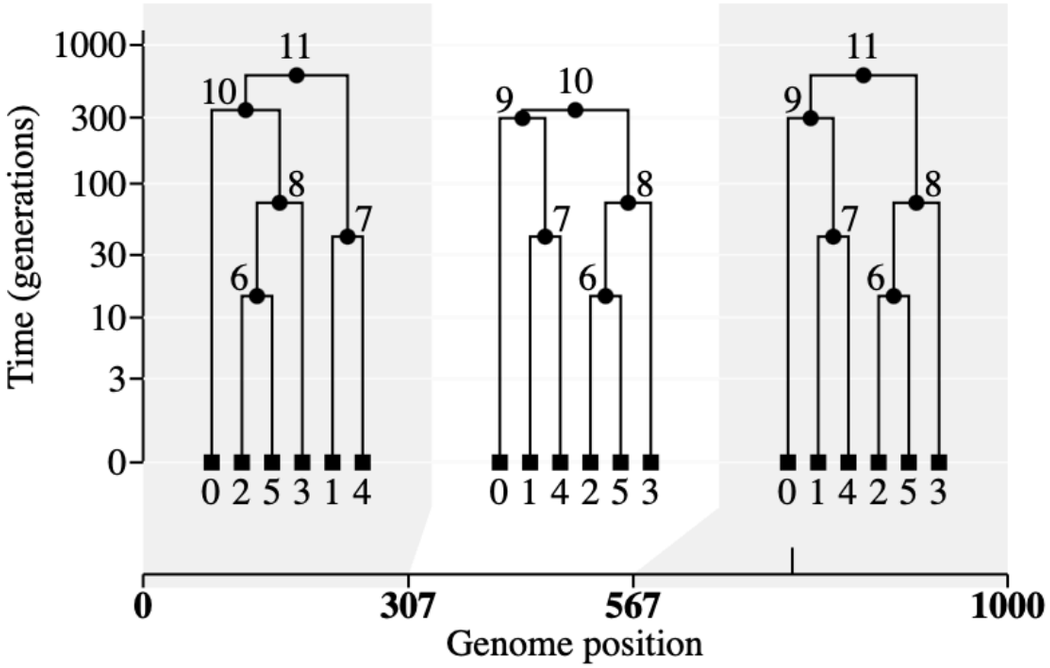
Also other inference programs that don't output in the tree sequence format by default (ARGweaver, Argneedle)

Tskit terminology: the basics

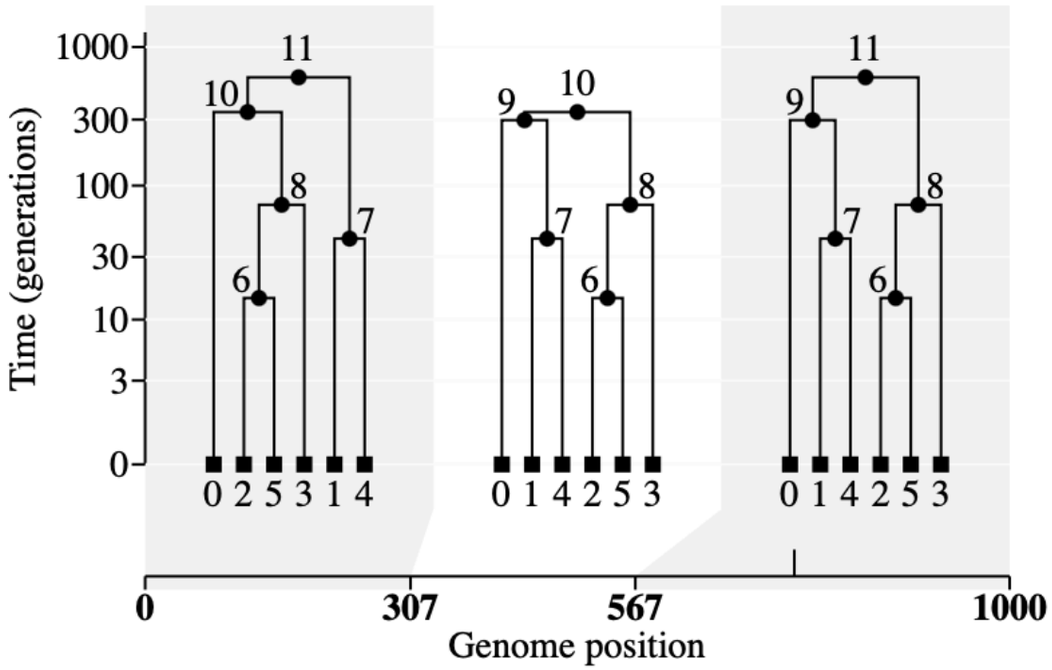


- Multiple **local trees** exist along a genome of fixed length
  (by convention measured in base pairs)
- Genomes exist at specific times, and are represented by **nodes**
  (the same node can persist across many local trees
- Some nodes are most recent common ancestors (MRCAs) of other nodes
- Entities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

Images from online tutorials: “Terminology & concepts”:
https://tskit.dev/tutorials/terminology_and_concepts.html



Tskit terminology: nodes and edges



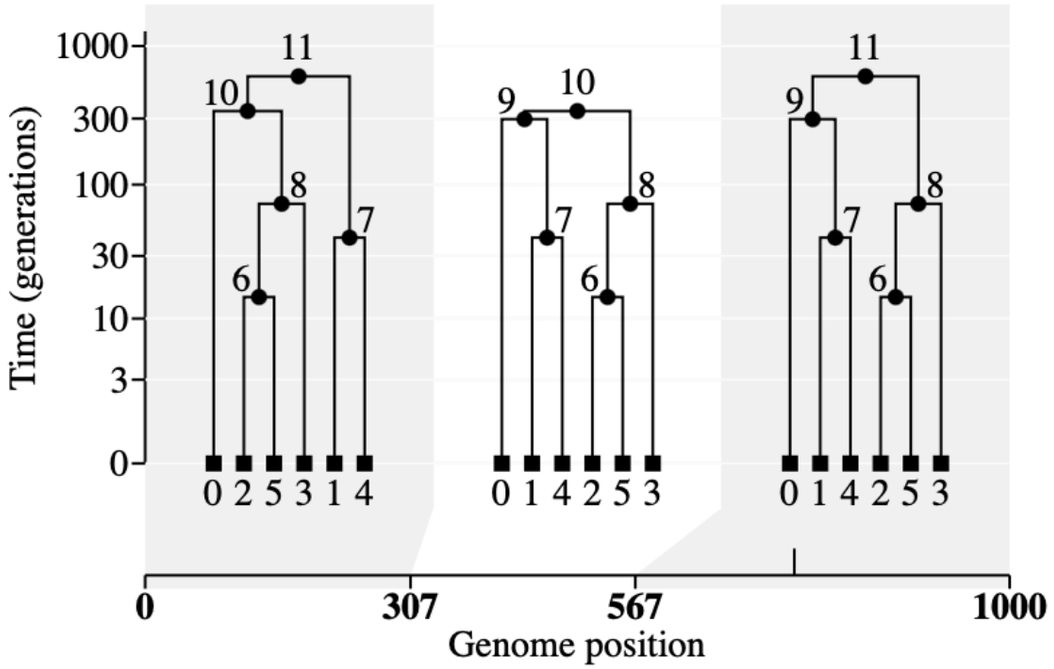
Tskit terminology: nodes and edges

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	0.00000000
1	1	0	0	0.00000000
2	1	0	1	0.00000000
3	1	0	1	0.00000000
4	1	0	2	0.00000000
5	1	0	2	0.00000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866

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	0.00000000
1	1	0	0	0.00000000
2	1	0	1	0.00000000
3	1	0	1	0.00000000
4	1	0	2	0.00000000
5	1	0	2	0.00000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866

Edges

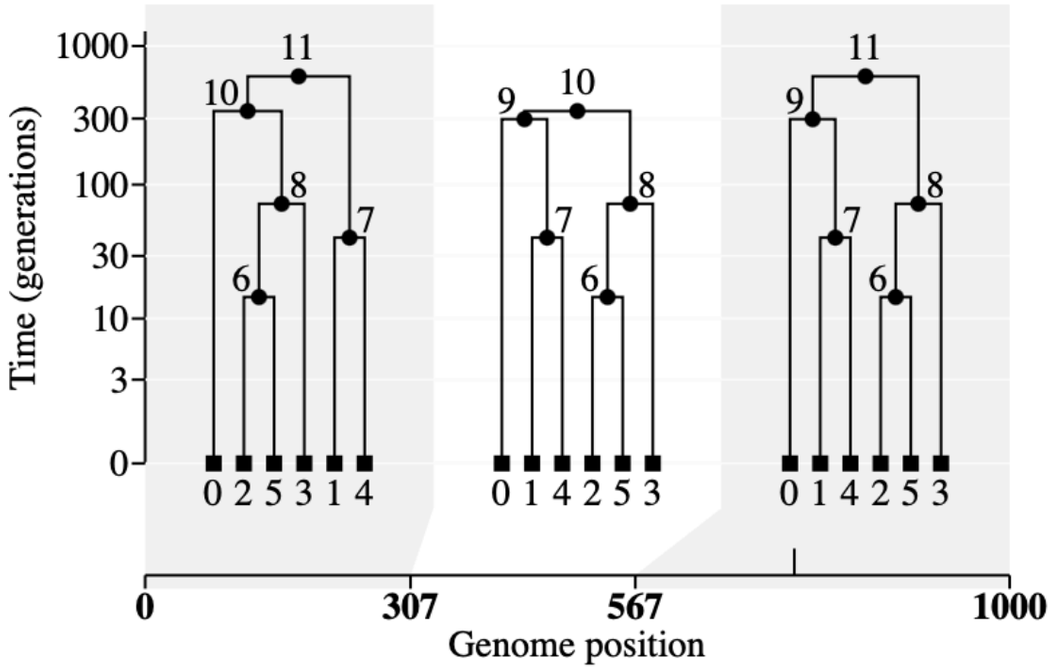
Connect a *parent & child*

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

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	0.00000000
1	1	0	0	0.00000000
2	1	0	1	0.00000000
3	1	0	1	0.00000000
4	1	0	2	0.00000000
5	1	0	2	0.00000000
6	0	0	-1	14.70054184
7	0	0	-1	40.95936939
8	0	0	-1	72.52965866

Edges

Connect a *parent & child*

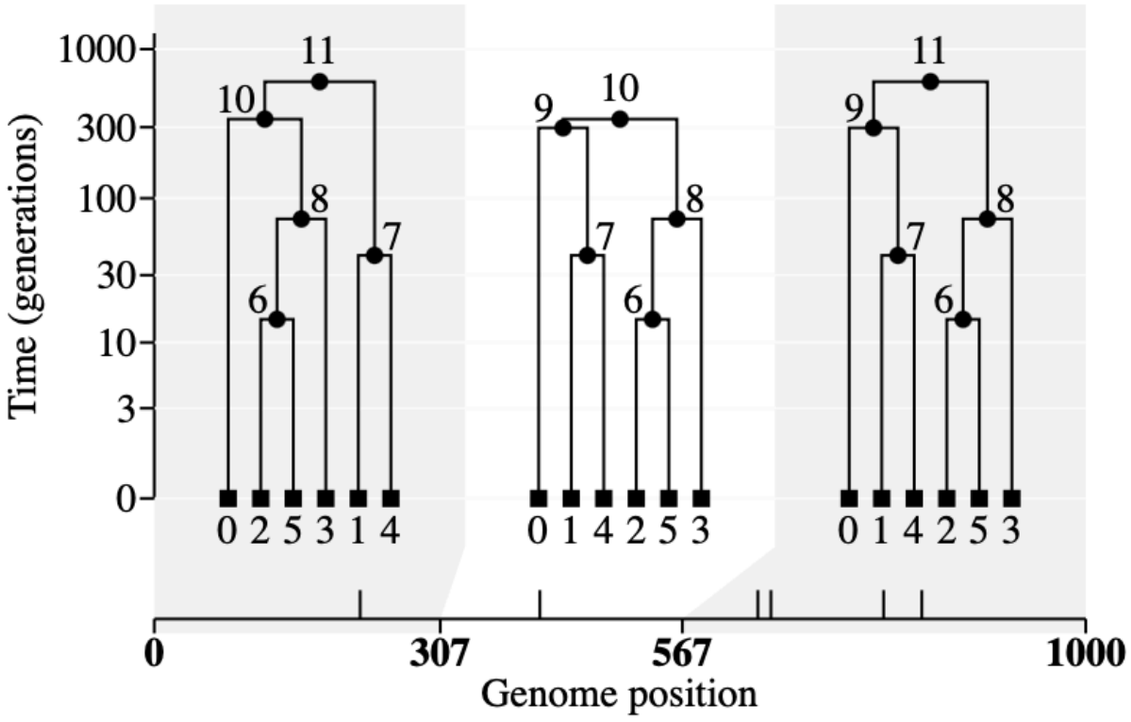
Have a *left & right* genomic coordinate

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

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

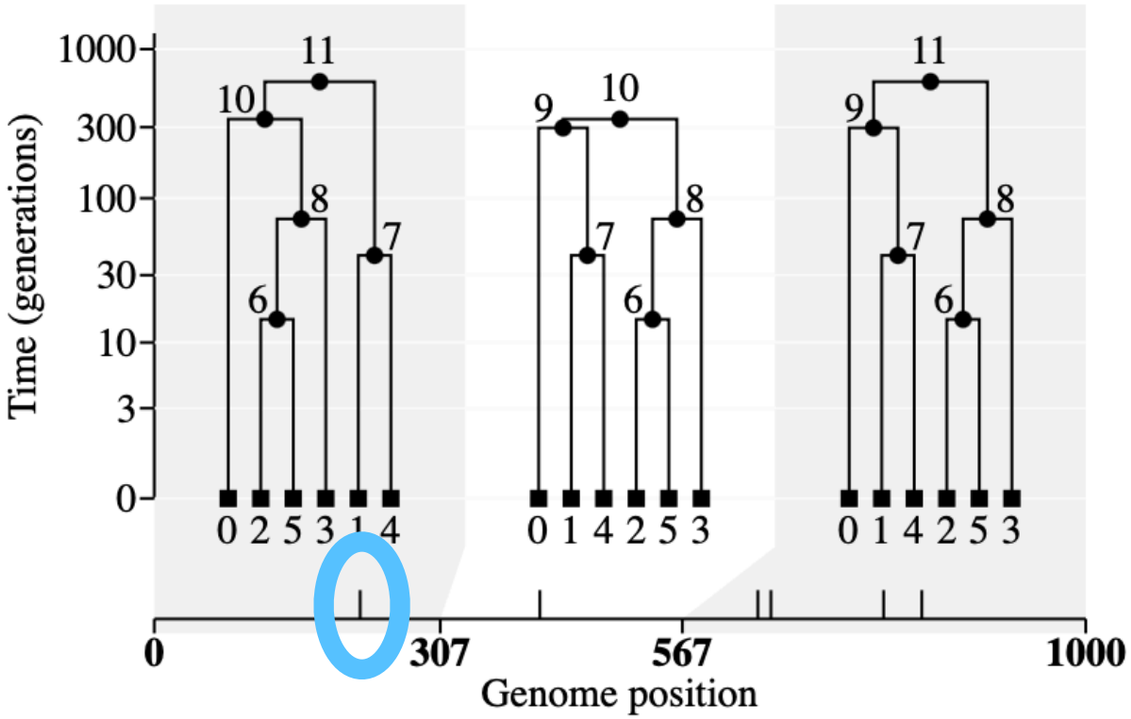
(e.g UKBB chr20 ~ 1M sample nodes, 62 M edges, tree iteration ~ 5secs)



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



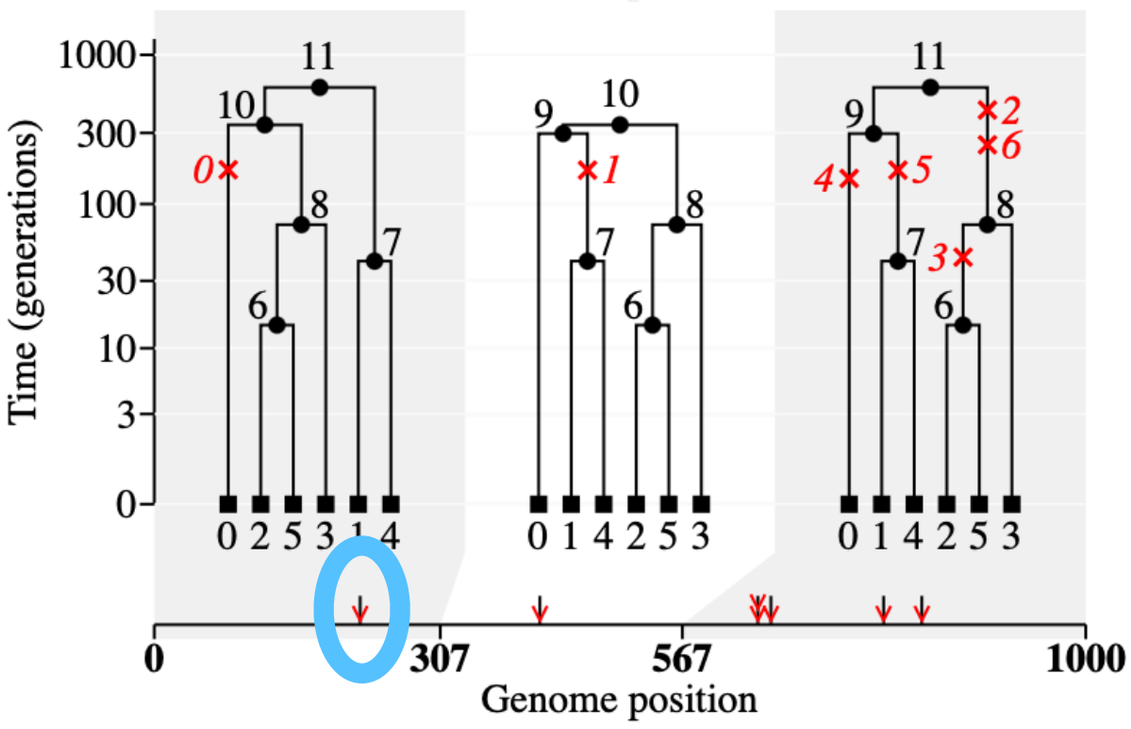
Tskit terminology: sites & mutations

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**.

id	position	ancestral_state
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**.

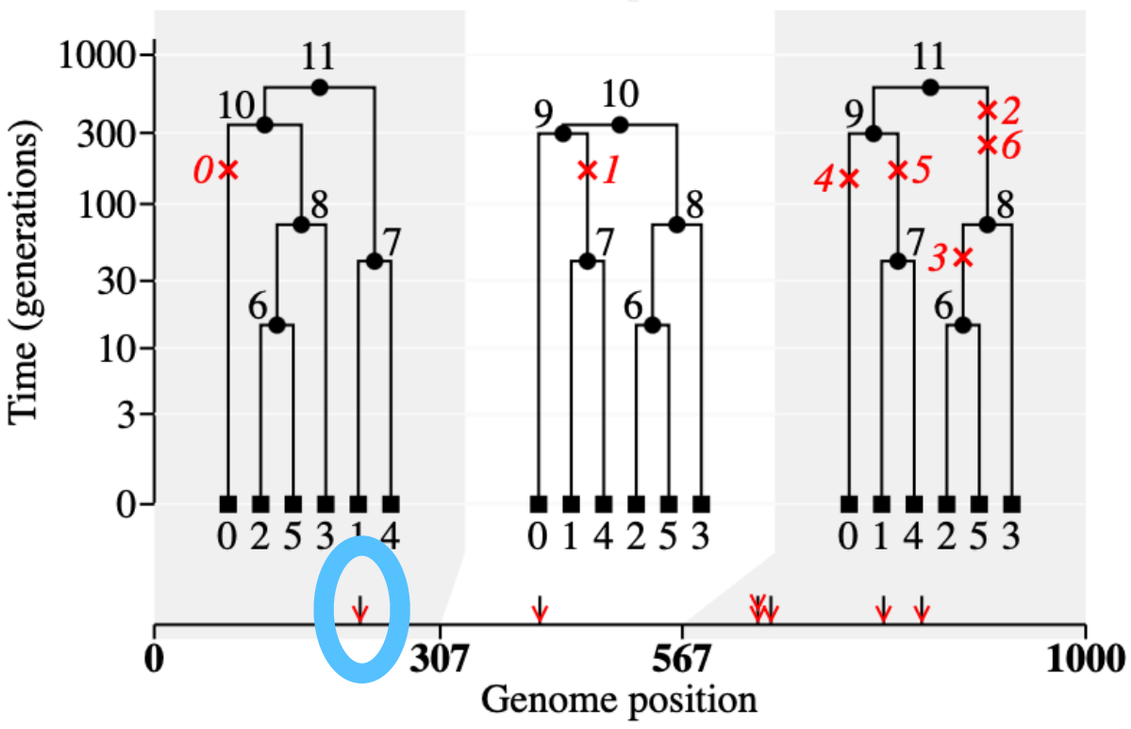
Normally, a site is created in order to place one or more **mutations** at that site

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	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



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	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)

Using tskit

Using tskit

Docs and tutorials

https://tskit.dev/tskit/docs

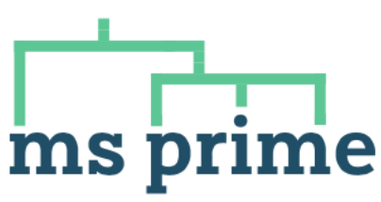
https://tskit.dev/tutorials

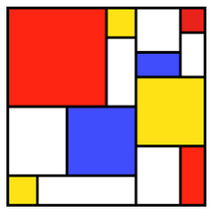
Using tskit

Docs and tutorials

https://tskit.dev/tskit/docs

https://tskit.dev/tutorials





SLiM

Research-enabling:

Evolution Letters

LETTER | Open Access

Mutation load decreases with haplotype age in wild Soay sheep

Martin A. Stoffel, Susan E. Johnston, Jill G. Pilkington, Josephine M. Pemberton

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*

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

Cite this article

The timing of human adaptation from Neanderthal introgression FREE

Sivan Yair, Kristin M Lee, Graham Coop

Genetics, Volume

Computational and Systems Biology, Genetics and Genomics

Detecting adaptive introgression in human evolution using convolutional neural networks

Science

HOME > SCIENCE > VOL. 380, NO. 6647 > ON THE GENES, GENEALOGIES, AND GEOGRAPHIES OF QUEBEC

Current Issue | First release papers | Archive | About | Submit manuscript

On the genes, genealogies, and geographies of Quebec

RESEARCH ARTICLE | HUMAN GENETICS

LUKE ANDERSON-TROCMÉ, DOMINIC NELSON, SHADI ZABAD, ALEX DIAZ-PAPKOVICH, [...] AND SIMON GRAVEL

+7 authors | Authors Info & Affiliations

Archive | About | Submit manuscript

...zo, Matteo Fumagalli, Fernando Racimo

...itute, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark; ...Campus, Imperial College London, United Kingdom

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

Artic...

Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin [1, 2, 3, 4], Christopher R. Gignoux [4], Raymond K. Walters [1, 2, 3], ... M. Neale [1, 2, 3], Simon Gravel [5, 6], Mark J. Daly [1, 2, 3], Carlos ...

eLife

HOME | MAGAZINE | COMMUNITY | INNOVATION | NEWSLETTER | ABOUT

Research Article

Epidemiology and Global Health, Genetics and Genomics

Demographic history mediates the effect of stratification on polygenic scores

Arslan A Zaidi, Iain Mathieson

Nov 17, 2020 · https://doi.org/10.7554/eLife.61548

Analysis of genetic dominance in the UK Biobank

Duncan S. Palmer, Wei Zhou, Liam Abbott, Nikolas Baya, Daniel King, Masahiro Kanai, Alex Bloemendal, Benjam...

doi: https://doi.org/10.1101/2021.08.15.456387

nature

Article | Open Access | Published: 17 May 2023

A weakly structured stem for human origins in Africa

Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn & Simon Gravel