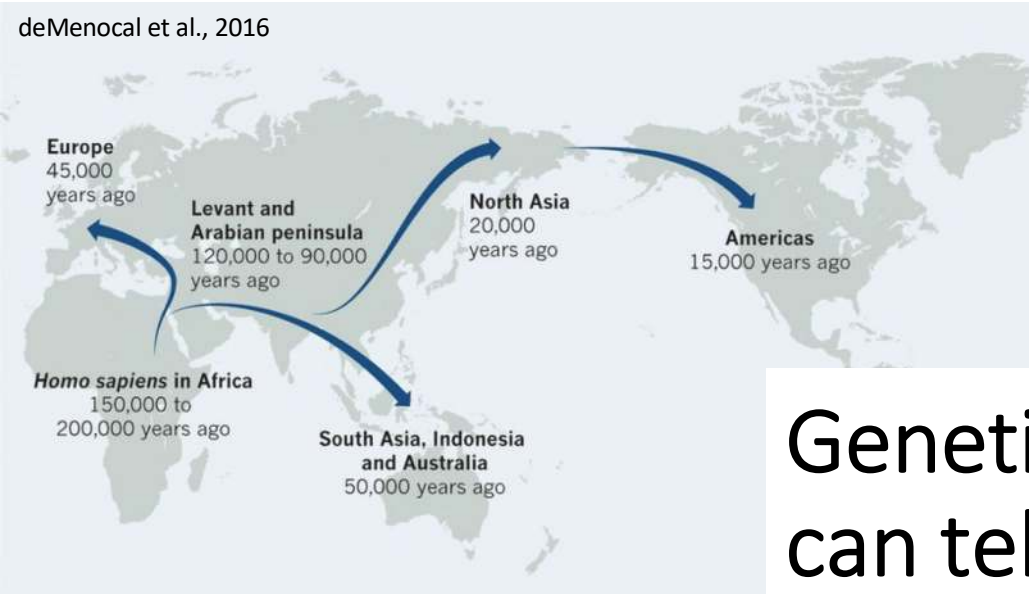


Learning about evolution by building coalescent trees

Simon Myers, Jasmin Rees, Leo Speidel

- This am: introductory lectures
- This pm: "Relate in the Prelate"
 - Running Relate on a human dataset of 130 different populations
 - Population structure and how it changes through time
 - Identifying directional selection





Genetic variation data can tell us about:

- Structure and migrations
- Population bottlenecks
- Admixture
- Mutation
- Recombination
- Selection
- etc



Step 1: Let's model these
Step 2: Inference

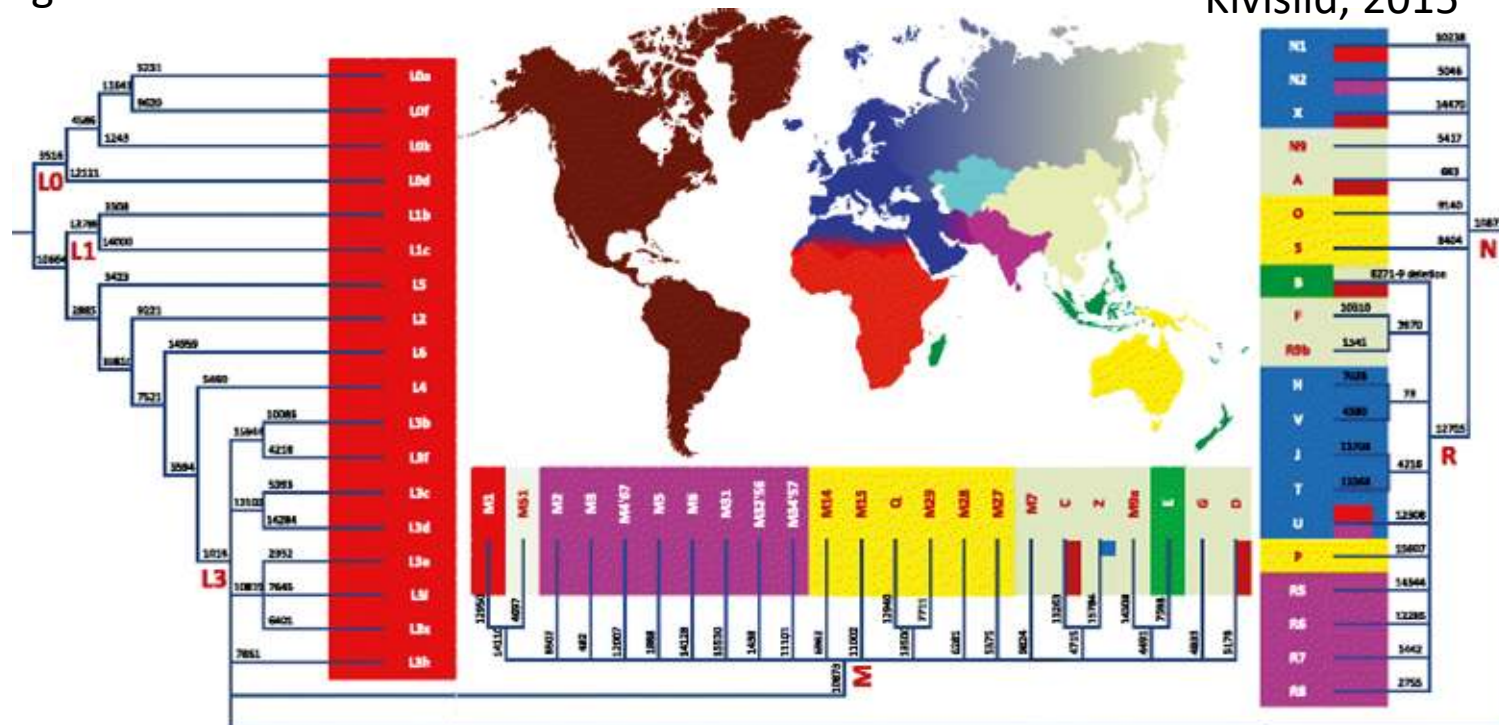
These might themselves evolve through time!

The genetic tree(s) relating humans (or any other species)

Reconstructing >100,000s years of evolution!

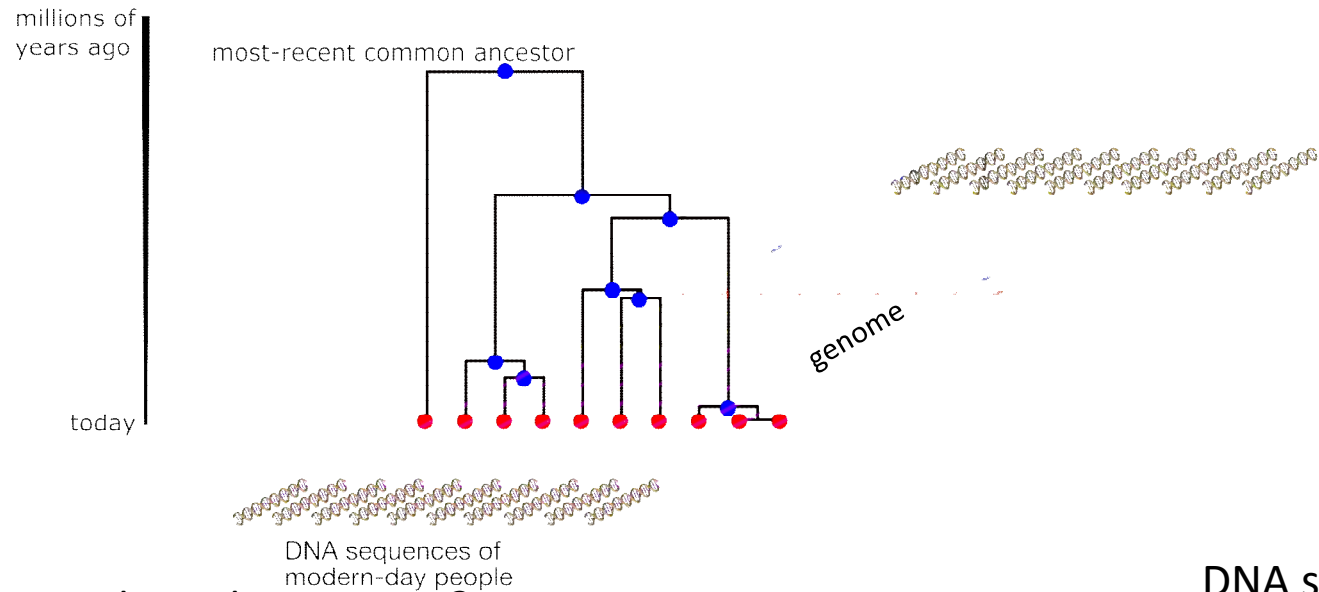
- 22 autosomal chromosomes
 - 2 sex chromosomes
 - X chromosome
 - Y chromosome
 - Mitochondrial genome
- Different trees in different parts of the genome, due to recombination
- 1 tree each (maternal/paternal)

Kivisild, 2015



Key concept: Genealogies

We are genetically related through a sequence of trees

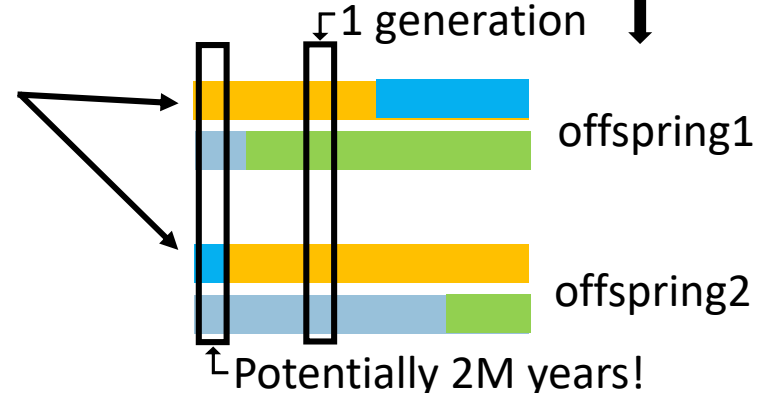


Why do trees change along the genome?

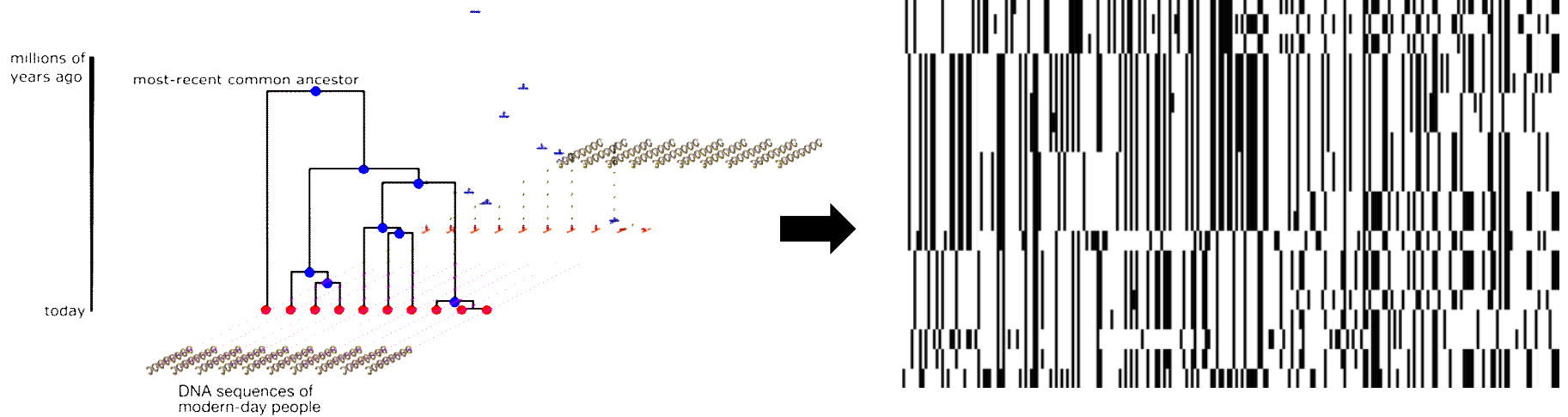
Recombination:

from grandmother
from grandfather

mother



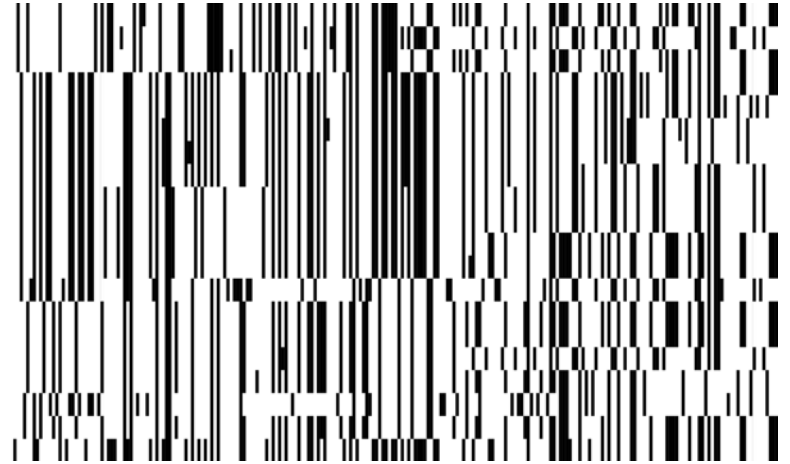
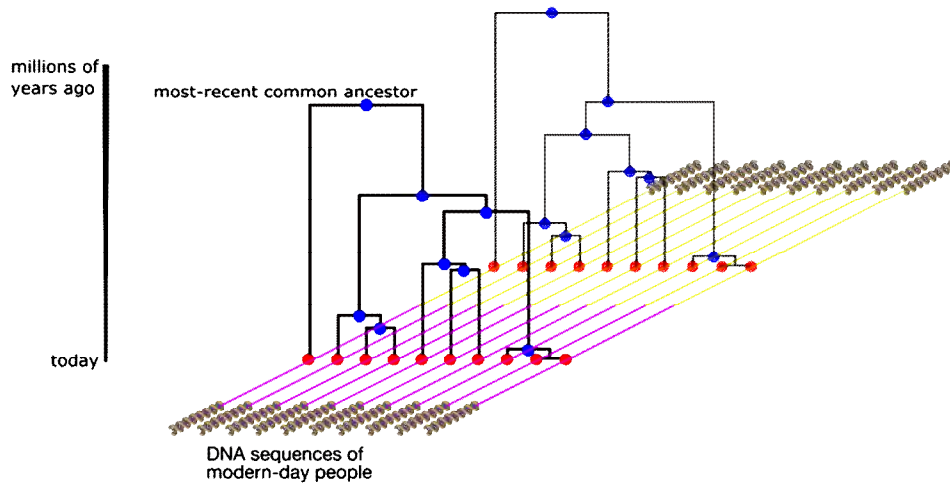
DNA sequenced



Demographic history
Genetic structure
Mutation, recombination,
etc.

Fundamental forces impact data (only) through underlying genealogies

Many canonical approaches

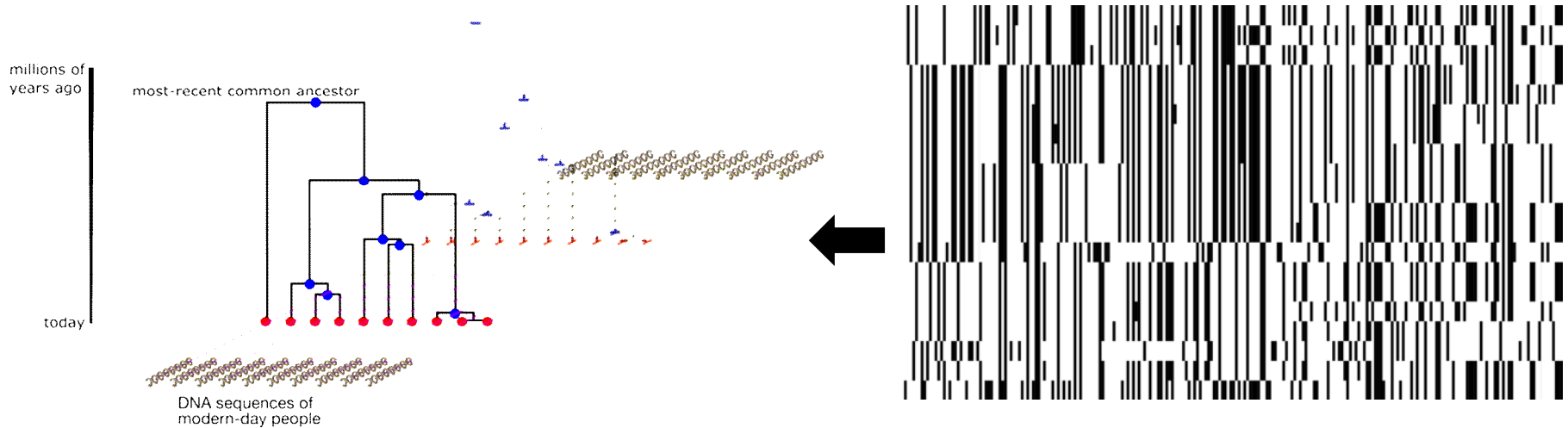


Demographic history
Genetic structure
Mutation, recombination,
etc.



Invent informative statistics, simplify,
"integrate out all possible histories"
(machine learning methods typically
learn these statistics from the data)

Today's approach



Demographic history
Genetic structure
Mutation, recombination,
etc.

Build (simple) statistics directly based on trees, to answer questions in population genomics

In **principle**, trees capture **all the information** available from the data about these processes, allowing **self-consistent inference**

Challenges: computationally very challenging to sample trees from the data, and modern datasets can contain >50,000 individuals and >100,000,000 mutations

Inferring genealogies

Old problem, lots of methods, but few can scale:

- ARGweaver } Infers Ancestral Recombination Graphs
- Rent+
- Tsinfer+tsdate } Published since 2019, scale to large
- Relate } sample sizes
- ARG-Needle }

We will talk about Relate today – Georgia will introduce tsinfer tomorrow.....

but principles of tree-based inference apply more generally!

Relate

Speidel, Forest, Shi, Myers. Nature Genetics 2019
Speidel, Cassidy, Davies, Hellenthal, Skoglund, Myers, MBE, 2021

[Relate](#) [Home](#) [Getting Started](#) [Input data](#) [Add-on modules](#) [Parallelise Relate](#)

Relate

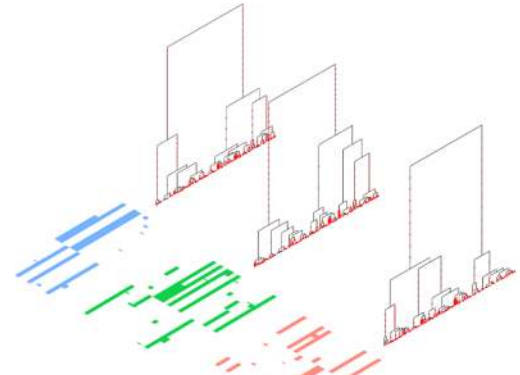
Software to estimate genome-wide genealogies for thousands of samples

Relate estimates genome-wide genealogies in the form of trees that adapt to changes in local ancestry caused by recombination. The method, which is scalable to thousands of samples, is described in the following paper. Please cite this paper if you use our software in your study.

Citations:

- (original Relate paper) Leo Speidel, Marie Forest, Sinan Shi, Simon Myers.
A method for estimating genome-wide genealogies for thousands of samples. [Nature Genetics 51: 1321-1329, 2019.](#)
- (update, v1.1.1) Leo Speidel, Lara Cassidy, Robert W. Davies, Garrett Hellenthal, Pontus Skoglund, Simon R. Myers.
Inferring population histories for ancient genomes using genome-wide genealogies. [Molecular Biology and Evolution 38: 3497-3511, 2021.](#)

Contact: leo.speidel@outlook.com
Website: <https://leospeidel.com>



Download

Relate is available for academic use. To see rules for non-academic use, please read the [LICENCE](#) file, which is included with each software download.

Pre-compiled binaries (last updated: 7/11/2021)

☐ I agree with the [terms and conditions](#)

[Linux \(x86_64, dynamic\) - v1.1.8](#)

[Linux \(x86_64, static\) - v1.1.8](#)

[Mac OSX \(Intel\) - v1.1.8](#)

[Mac OSX \(M1\) - v1.1.8](#)

GitHub repository

Alternatively, you can **compile your own version** by downloading the source code from this [github repository](#).

In the downloaded directory, we have included a toy data set. You can try out **Relate** using this toy data set by following the instructions on our [getting started](#) page.

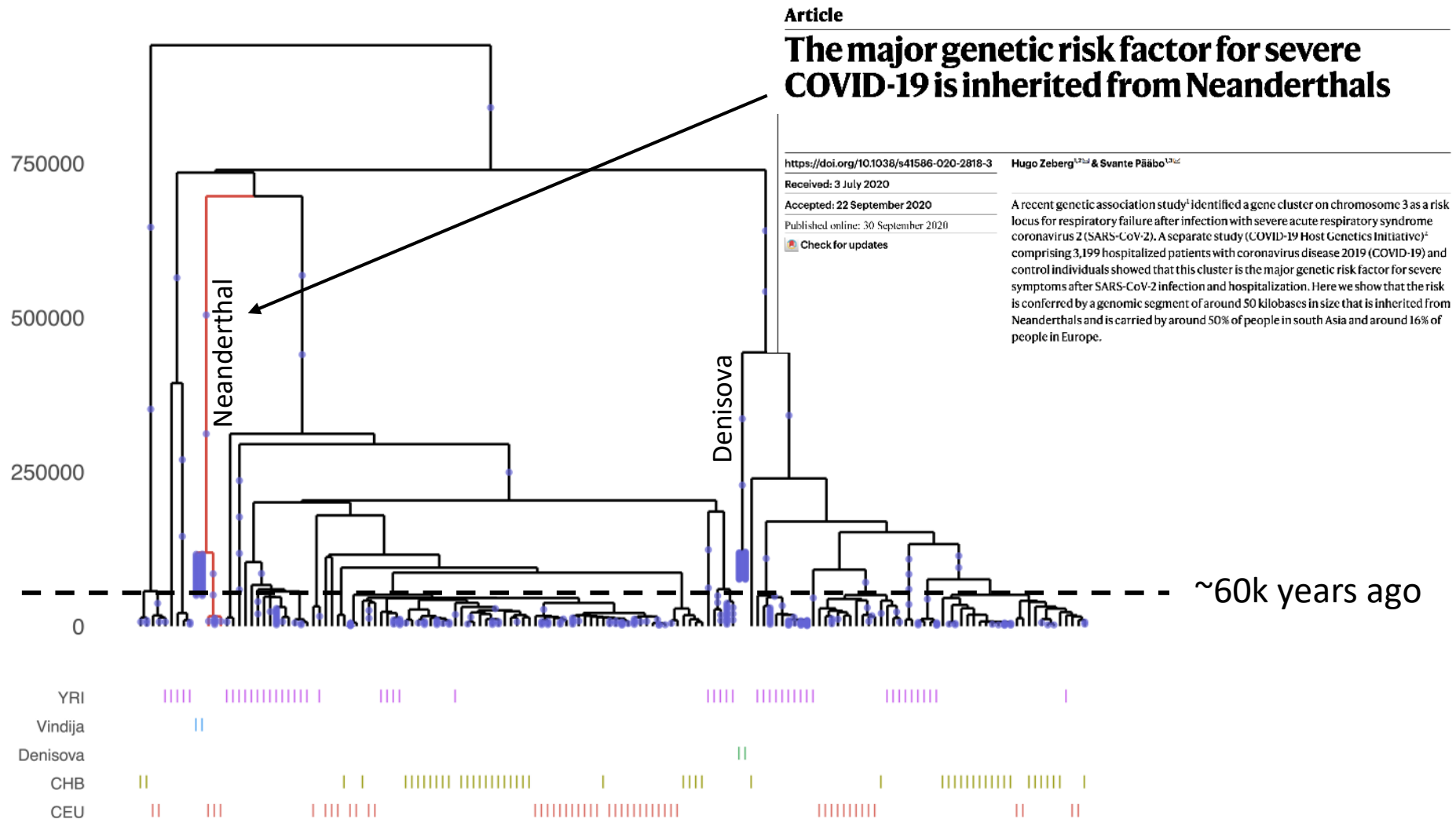
If you have any problems getting the program to work on your machine or would like to request an executable for a platform not shown here, please send a message to [leo.speidel\[at\]outlook\[dot\]com](mailto:leo.speidel[at]outlook[dot]com).

<https://myersgroup.github.io/relate/>

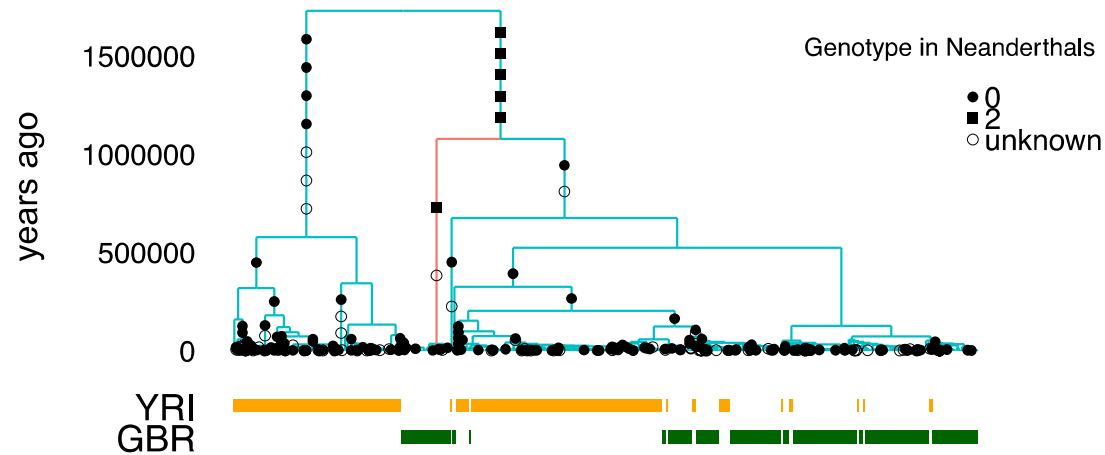
Key features:

- Fast & accurate
- Robust to errors!
- Jointly infers branch lengths and demographic history
- Moderns and ancients
- Lots of add-on tools for various types of analyses

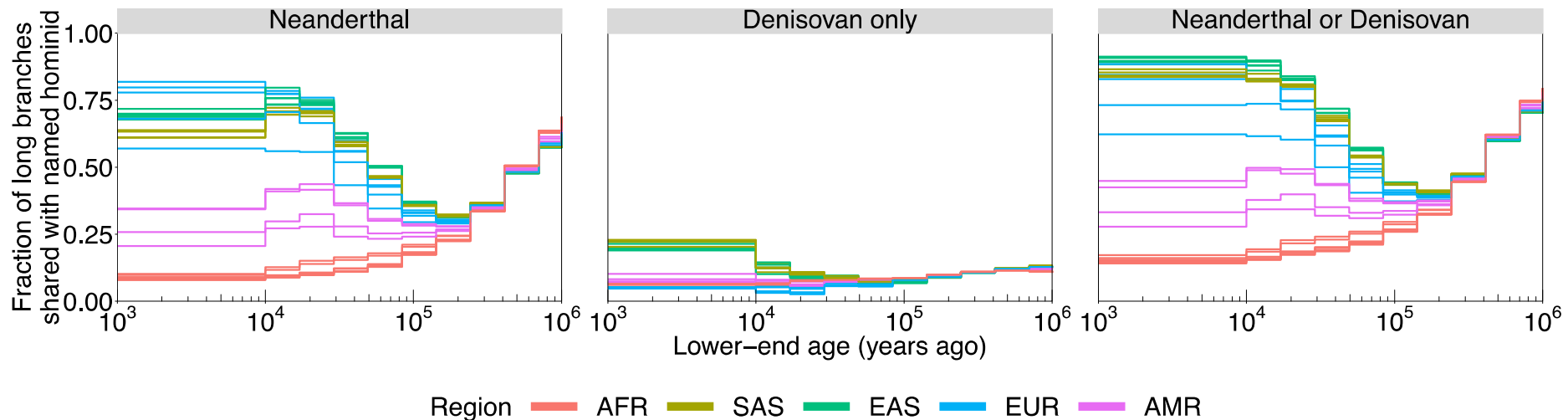
One locus can already tell us a lot about our history



....combining across loci tells us much more (e.g. ancient introgression in non-Africans mainly from Neanderthal/Denisovan relatives)

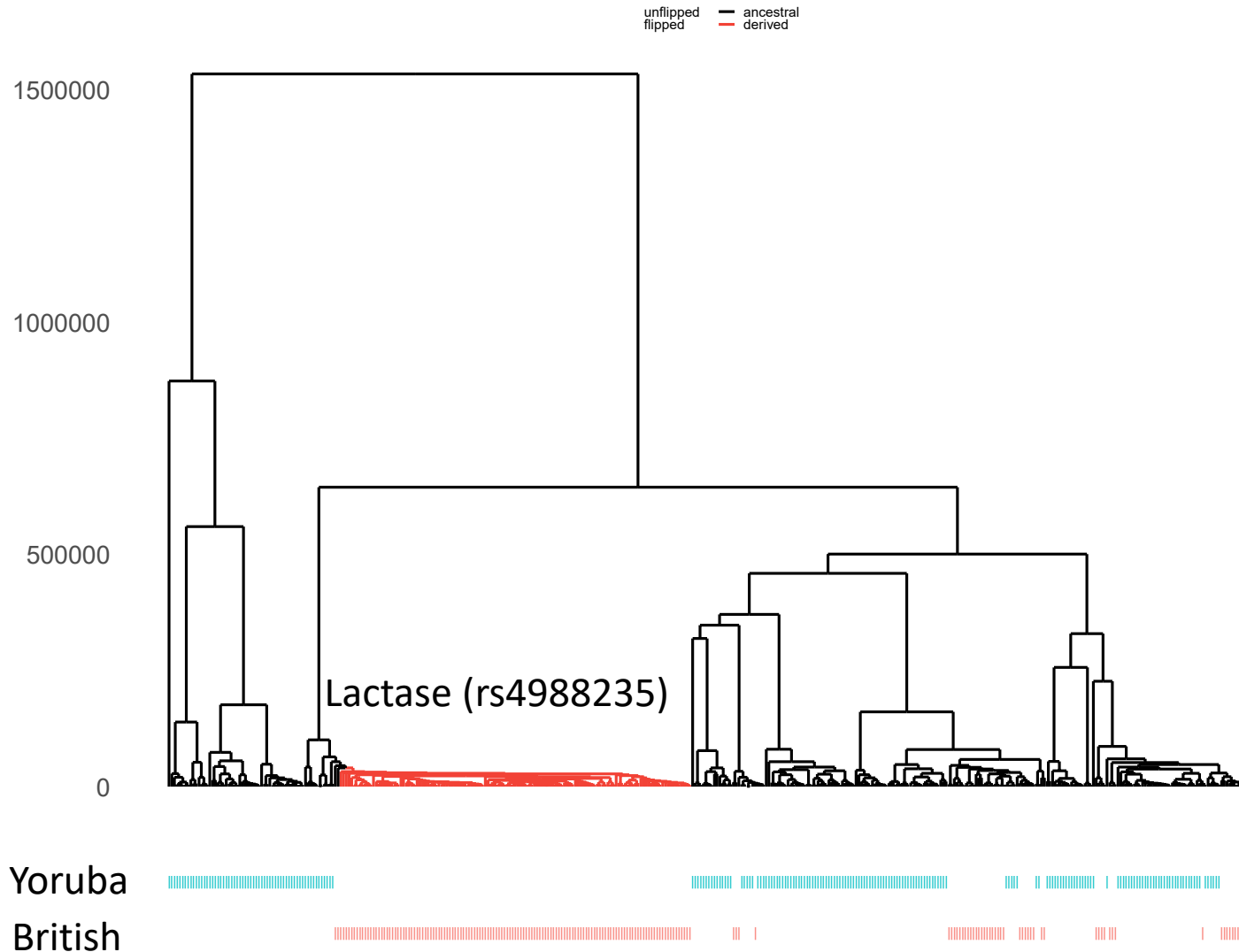


Fraction of **deep branches** shared with named hominid:
(Upper end > 1M years)

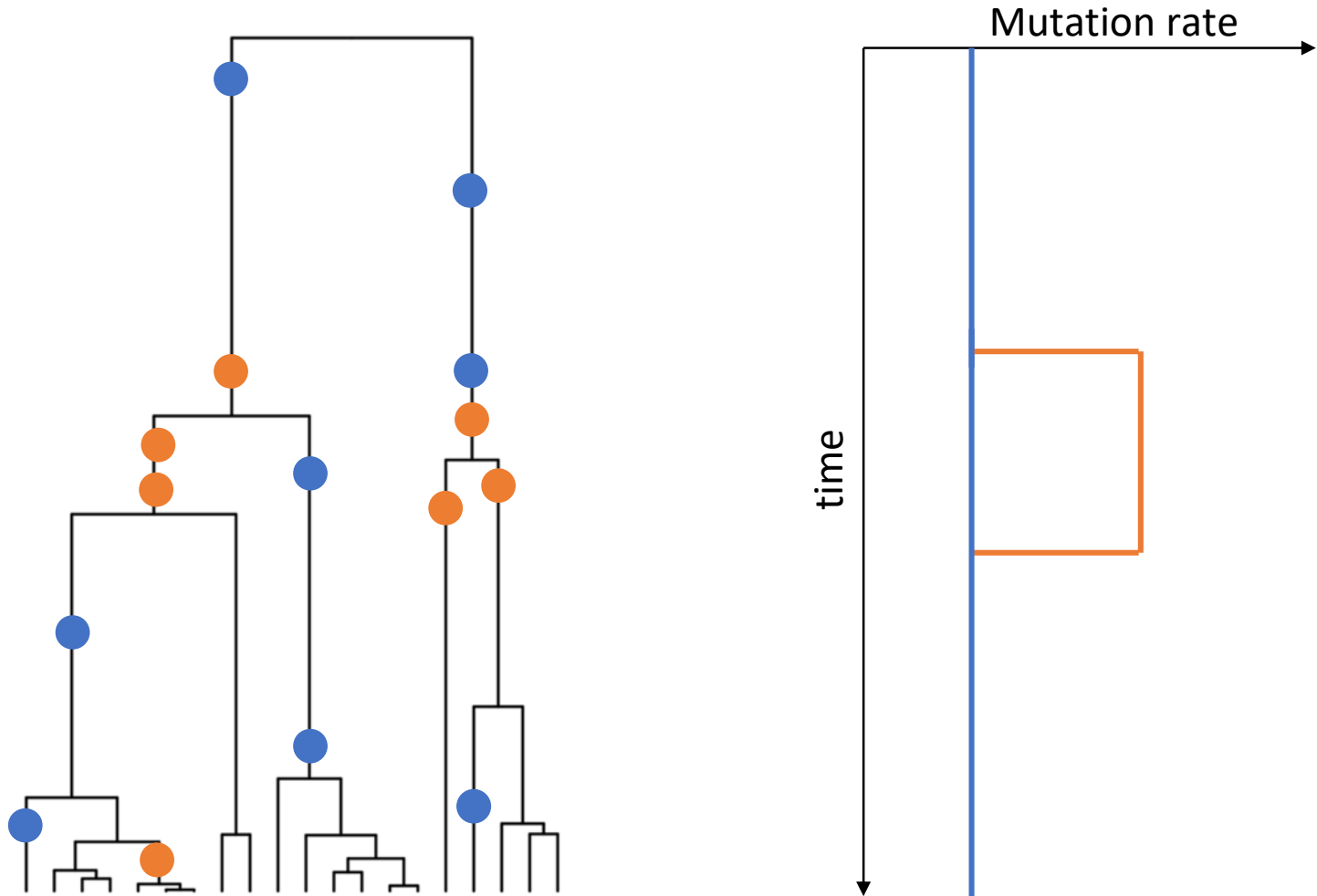


NB: Africa shows little Neanderthal or Denisovan introgression, but actually has a **huge excess** of long (unexplained) branches shared only within Africa

Example: Positive selection, rapidly spreading lineage



Example: clusters of mutations in time can capture changes in mutation rate



To actually do inference, we need to (re)visit the coalescent model to help:

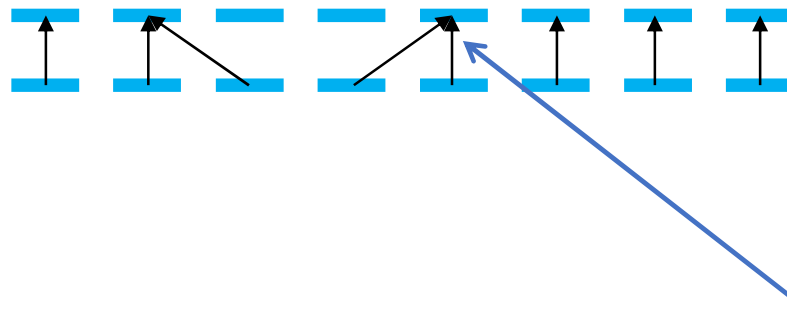
- 1) Create a method to build trees under a coalescent model, with varying population size, and allowing for recombination
- 2) Construct statistics to capture information from trees and either
 - (i) Interpret parameters in the coalescent, e.g. coalescence rates
 - (ii) Reject a null model, e.g. testing for selection

Revision of coalescent

The **Wright-Fisher model** is able to approximate more realistic models of populations, but is itself the “simplest possible model incorporating inheritance”

N individuals; each member of the current generation randomly chooses one of $2N$ parent chromosomes and inherits their DNA

Some population members have 0 children, others more than 1 child:



Each haplotype chooses parent in previous generation totally at random

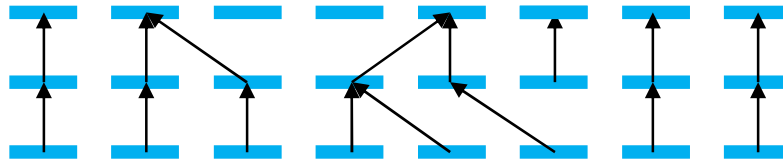
If haplotypes **share** a parent back in time, this is called a **coalescence event**

Revision of coalescent

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time

In a single generation, chance two haplotypes choose the same parent
is $1/2N$



Each haplotype chooses parent in previous generation totally at random

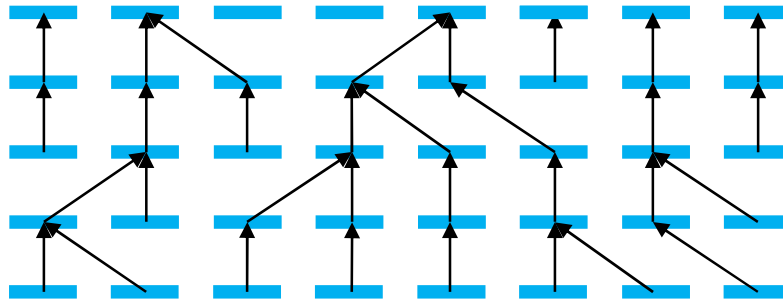
If haplotypes **share** a parent back in time, this is called a **coalescence event**

Revision of coalescent

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time

In a single generation, chance two haplotypes choose the same parent is $1/2N$



Each haplotype chooses parent in previous generation totally at random

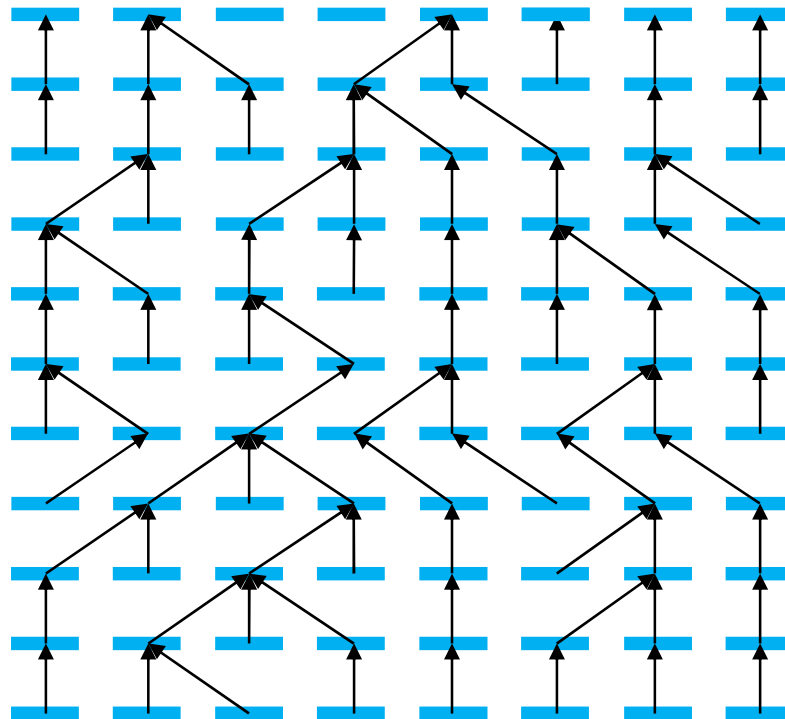
If haplotypes **share** a parent back in time, this is called a **coalescence event**

Revision of coalescent

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time

In a single generation, chance two haplotypes choose the same parent
is $1/2N$



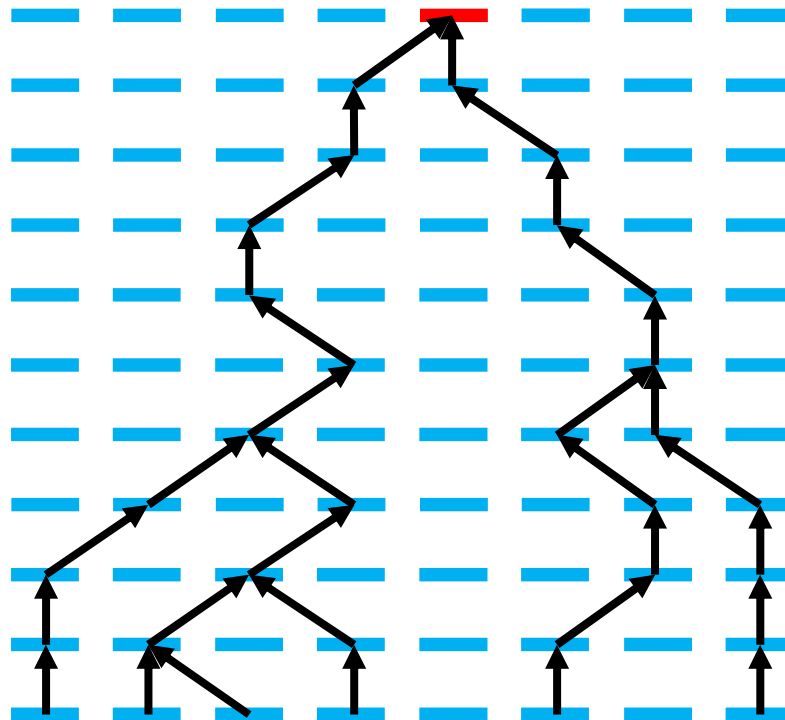
Each haplotype chooses
parent in previous
generation totally at
random

If haplotypes **share** a
parent back in time, this
is called a **coalescence**
event

If we take a sample from the population, we can trace their ancestry: a random tree

In this tree, the number of ancestors decreases back in time from n to 1

Each pair of lineages has $1/2N$ coalescences per generation, so 1 coalescence per $2N$ generations



Sample of size $n=6$

$N \sim 10\text{-}50,000$ for all human populations, highest in Africa



So a typical pair of human chromosomes share an ancestor on average around $2 \times 20,000 \times 28 = 1$ million years ago

N varies dramatically across species
(Charlesworth, Nature Reviews Genetics 2009):

25,000,000 for *E.coli*

2,000,000 for fruit fly

D. Melanogaster



<100 for Salamanders
(Funk et al. 1999)

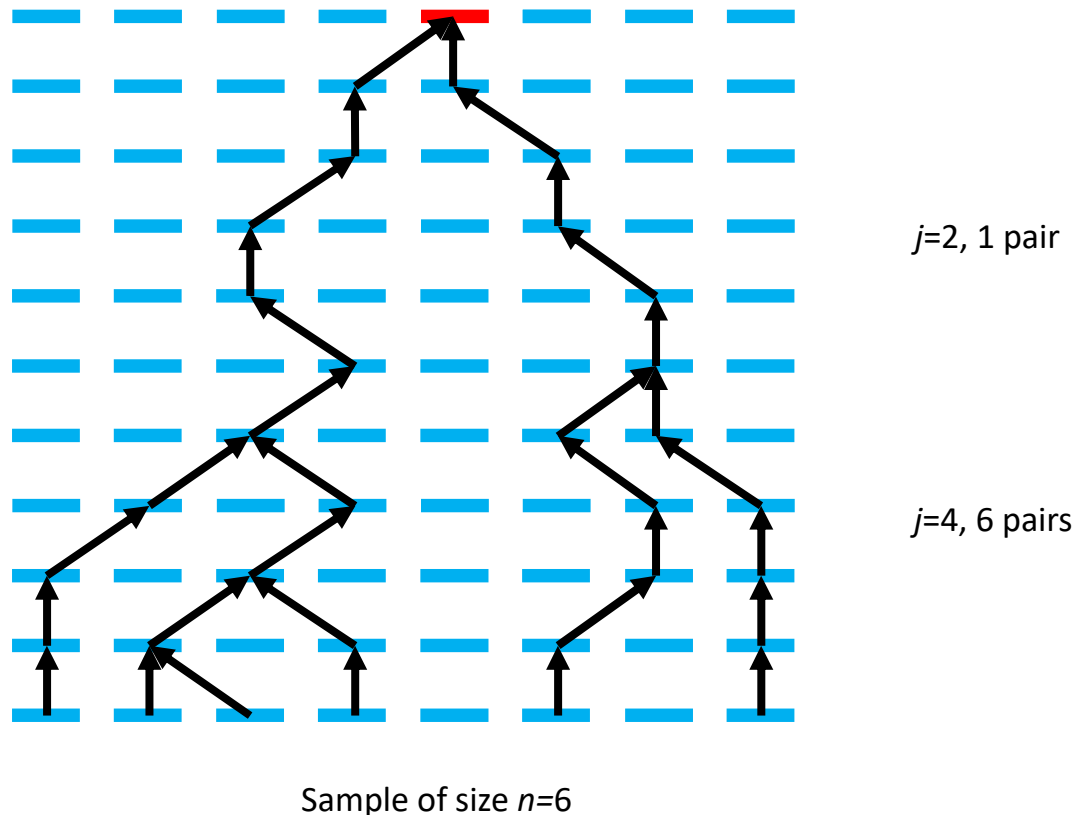
Let's not even talk about plankton!

Typically, as N is large we just model time as continuous

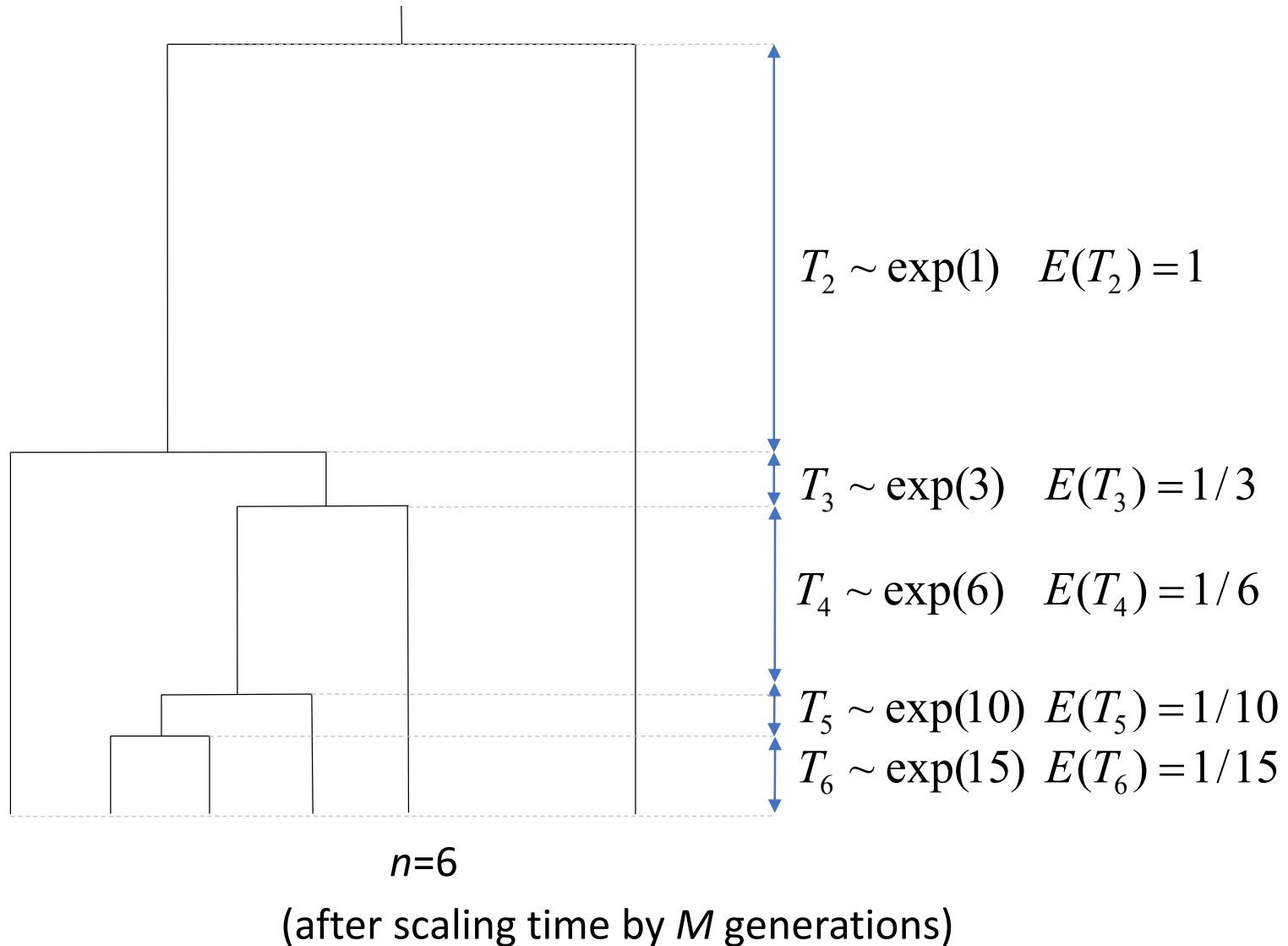
Any pair of lineages coalesces at rate $1/2N$

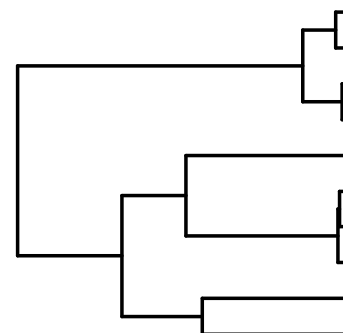
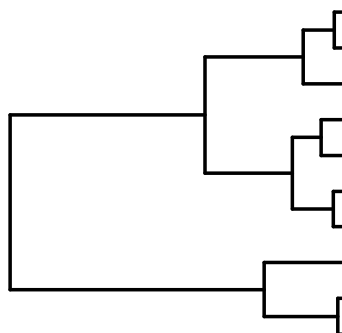
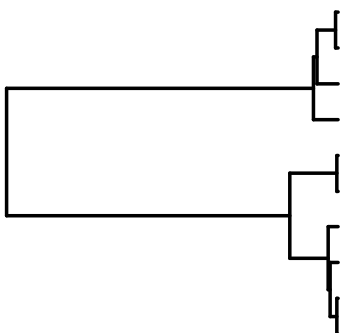
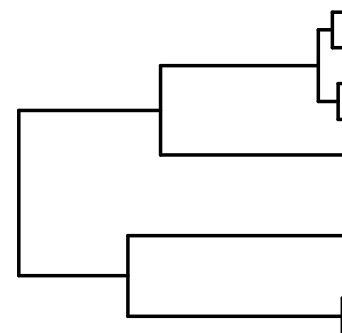
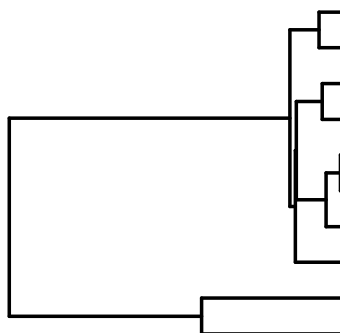
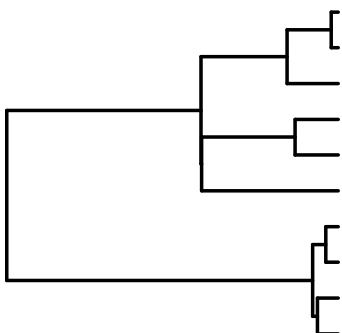
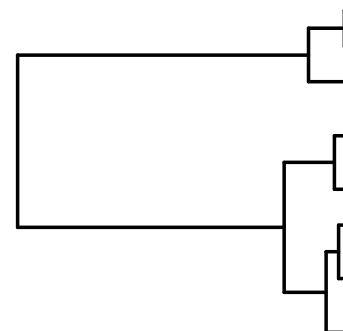
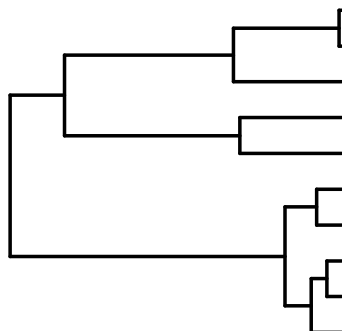
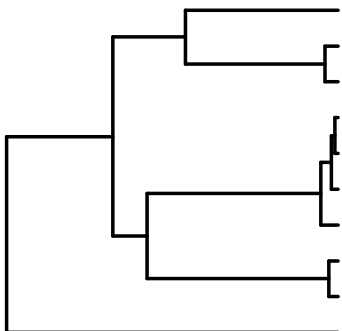
Then while there are j lineages, there are $\binom{j}{2}$ pairs that can coalesce - so the rate at which a coalescence happens is just $\binom{j}{2}/2N$

[this leads to an exponential distribution of time until coalescence, with rate $\binom{j}{2}/2N$]



We have come to a model – the Coalescent



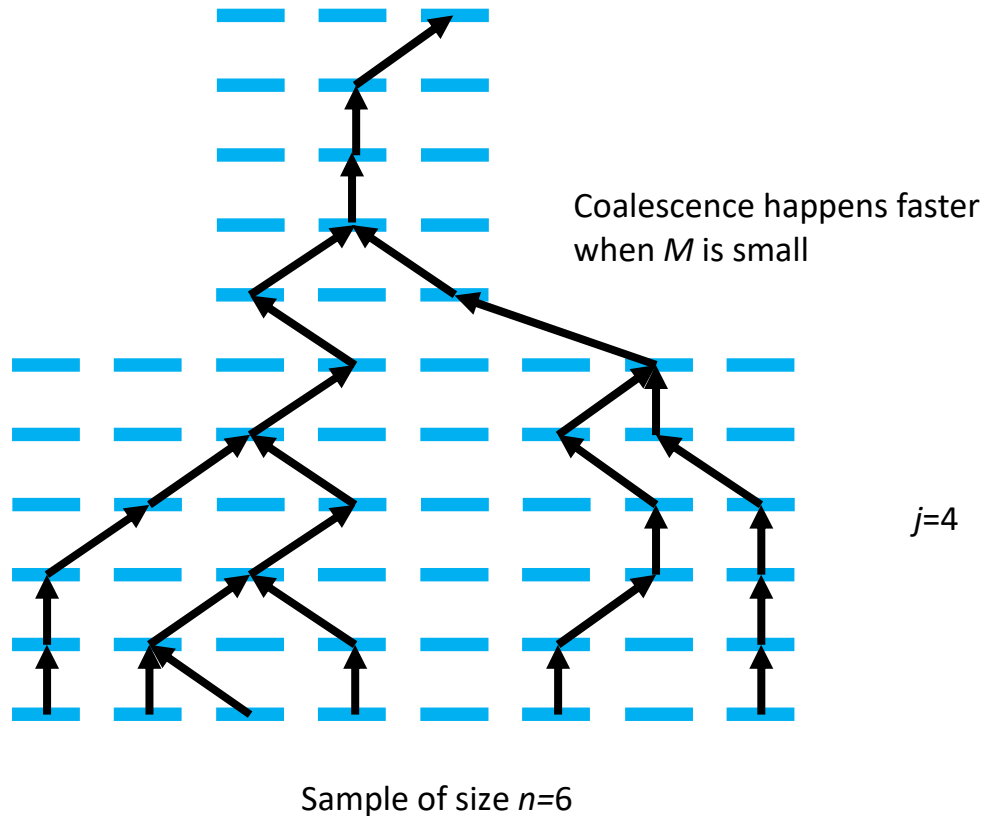


Varying population size

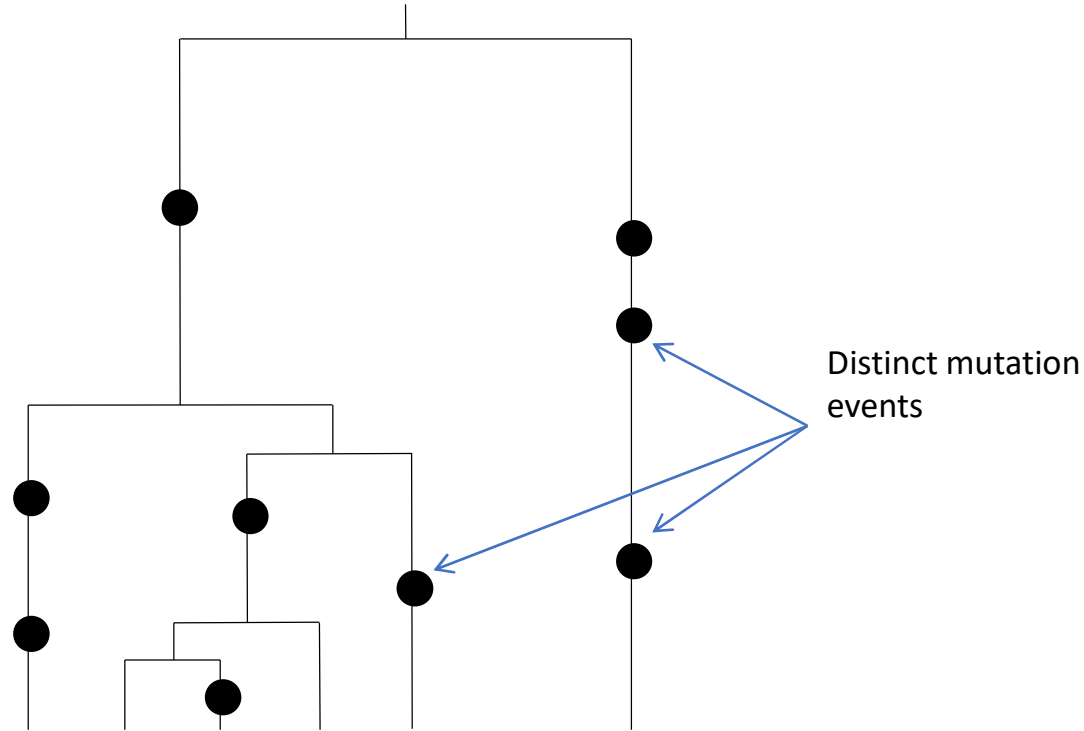
If N changes, so does the chance of coalescing

While there are j lineages, the rate at which a coalescence happens is just $\binom{j}{2}/2N(t)$ a time t ago

Shapes of trees can tell us about $N(t)$

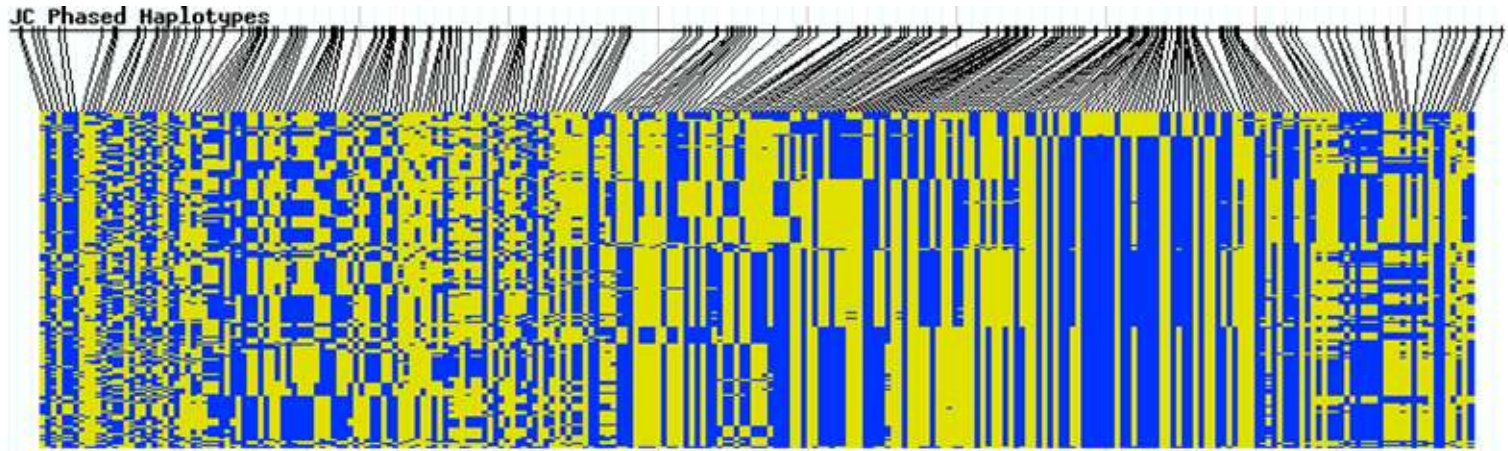


Adding mutation to the mix

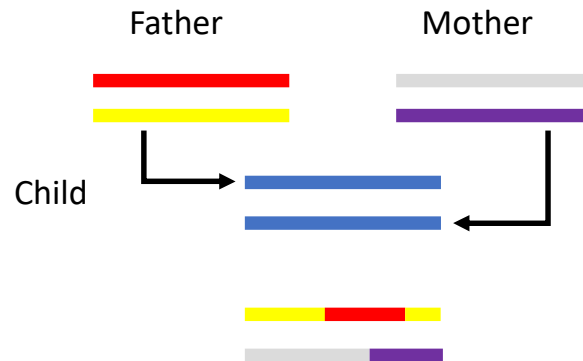


- Mutations are dropped randomly on the edges of the tree (e.g. in many simulation software packages)
- They are seen in descendants of this edge, so this totally specifies diversity patterns
- We will talk about some theory results about spread of mutations in the coalescent later

What about recombination?



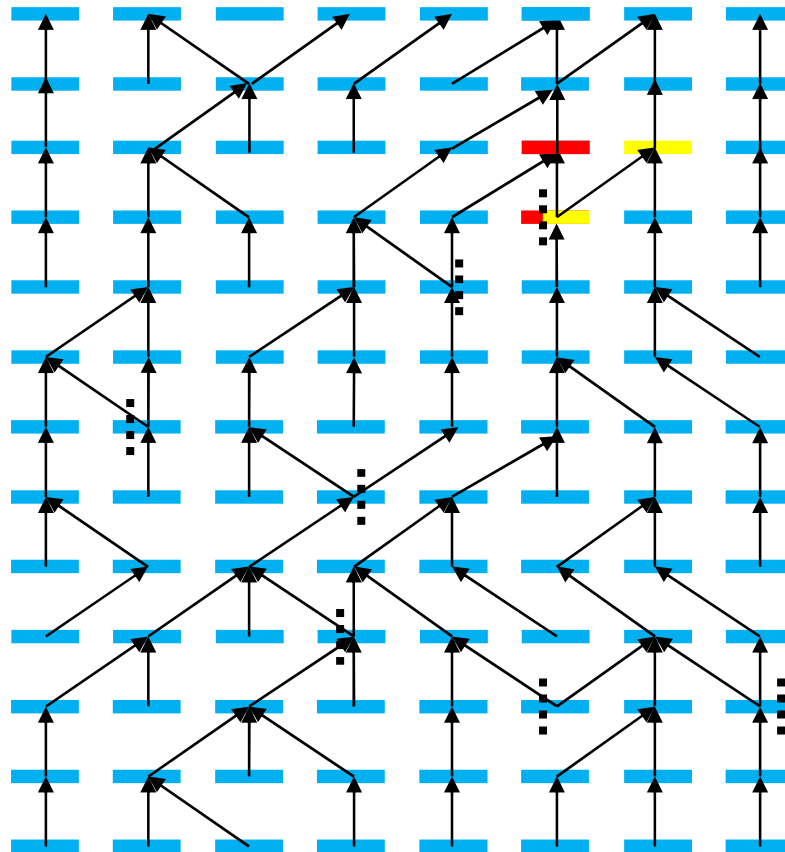
- Unlike mutation, recombination events actually **change the trees**



- One piece of DNA can be inherited from **two** different parental chromosomes, as a mosaic

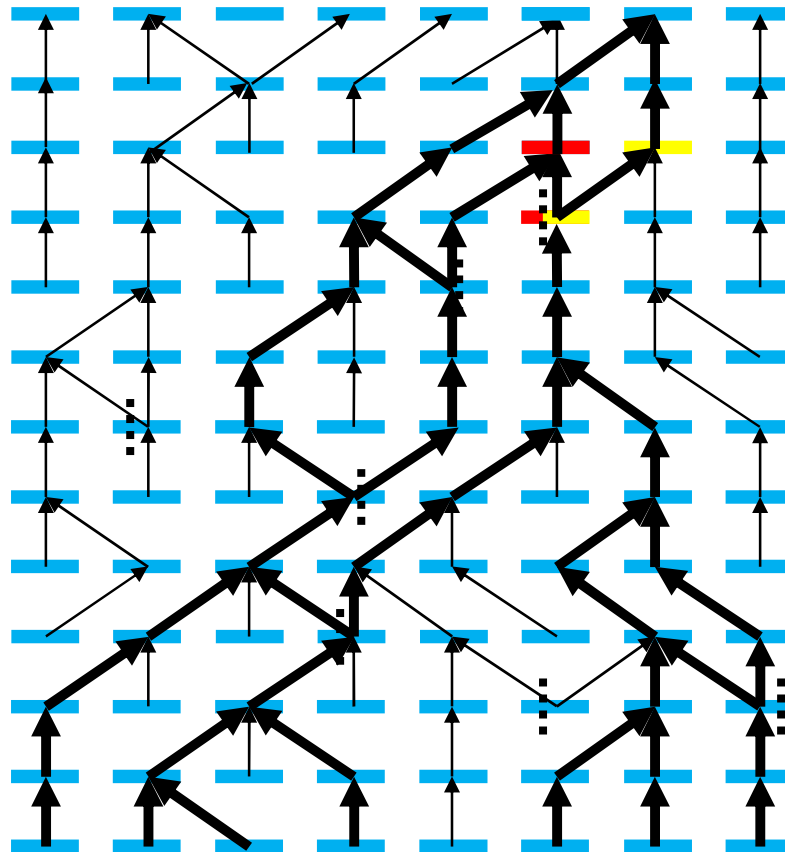
Principles of adding recombination to the coalescent

In the Wright-Fisher model, if recombination occurs then a chromosomal segment has two parents

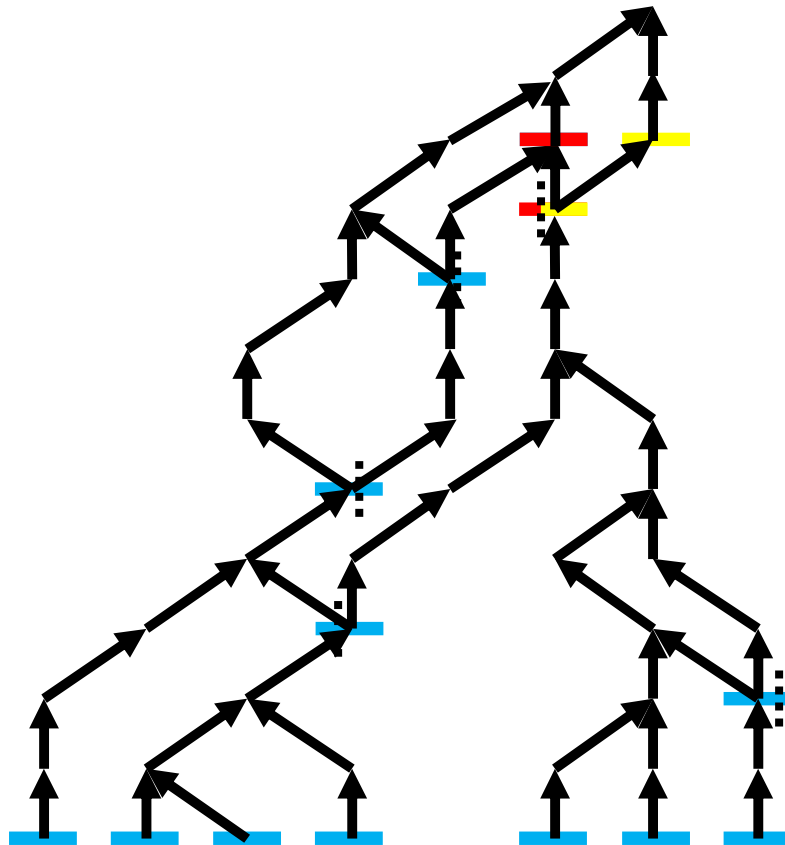


Principles of adding recombination to the coalescent

In the Wright-Fisher model, if recombination occurs then a chromosomal segment has two parents

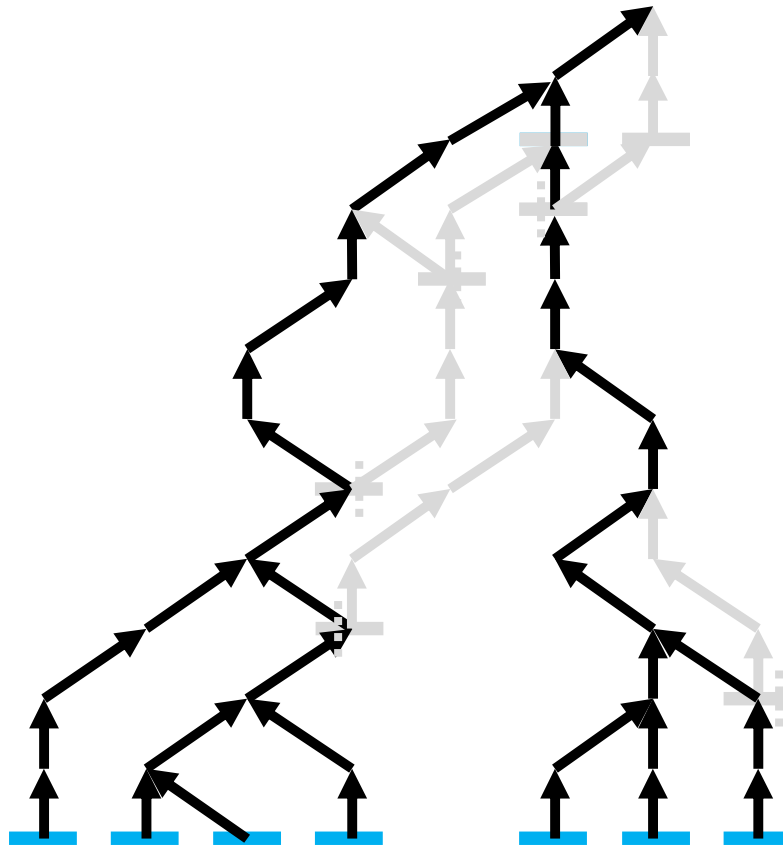


The ancestral recombination graph



The tree at the left-most position in the region

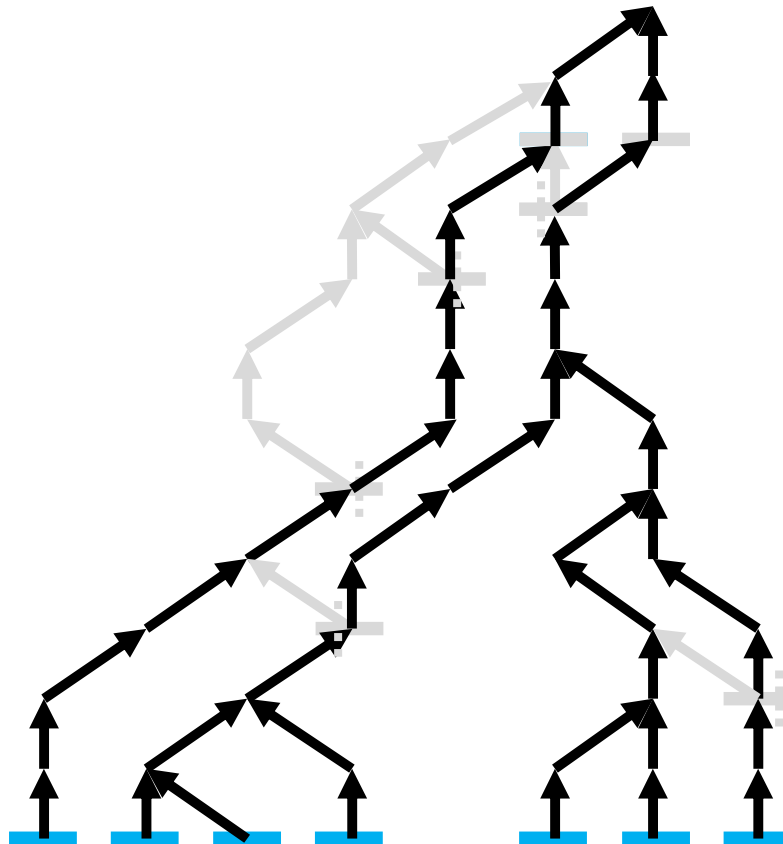
A small piece of DNA is not impacted by recombination, so the coalescent model still applies – with the same rates as usual



The tree at the right-most position in the region

A small piece of DNA is not impacted by recombination, so the coalescent model still applies

The bases of the trees have less recombination so are more similar than the tops



Building trees under this model – approximately – with Relate

<https://myersgroup.github.io/relate/>

Relate Home Getting Started Input data Add-on modules Parallelise Relate

Relate

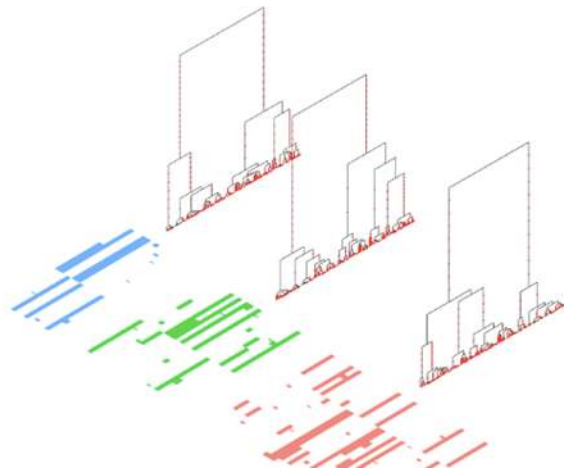
Software to estimate genome-wide genealogies for thousands of samples

Relate estimates genome-wide genealogies in the form of trees that adapt to changes in local ancestry caused by recombination. The method, which is scalable to thousands of samples, is described in the following paper. Please cite this paper if you use our software in your study.

Citation: Leo Speidel, Marie Forest, Sinan Shi, Simon Myers. A method for estimating genome-wide genealogies for thousands of samples. [Nature Genetics 51: 1321-1329, 2019.](#)

Contact: leo.speidel@outlook.com

Website: <https://leospeidel.wordpress.com>



Download

Relate is available for academic use. To see rules for non-academic use, please read the [LICENCE](#) file, which is included with each software download.

Pre-compiled binaries (last updated: 02/09/2019)

☐ I agree with the [terms and conditions](#)

[Linux \(x86_64, dynamic\) - v1.0.16](#)

[Linux \(x86_64, static\) - v1.0.16](#)

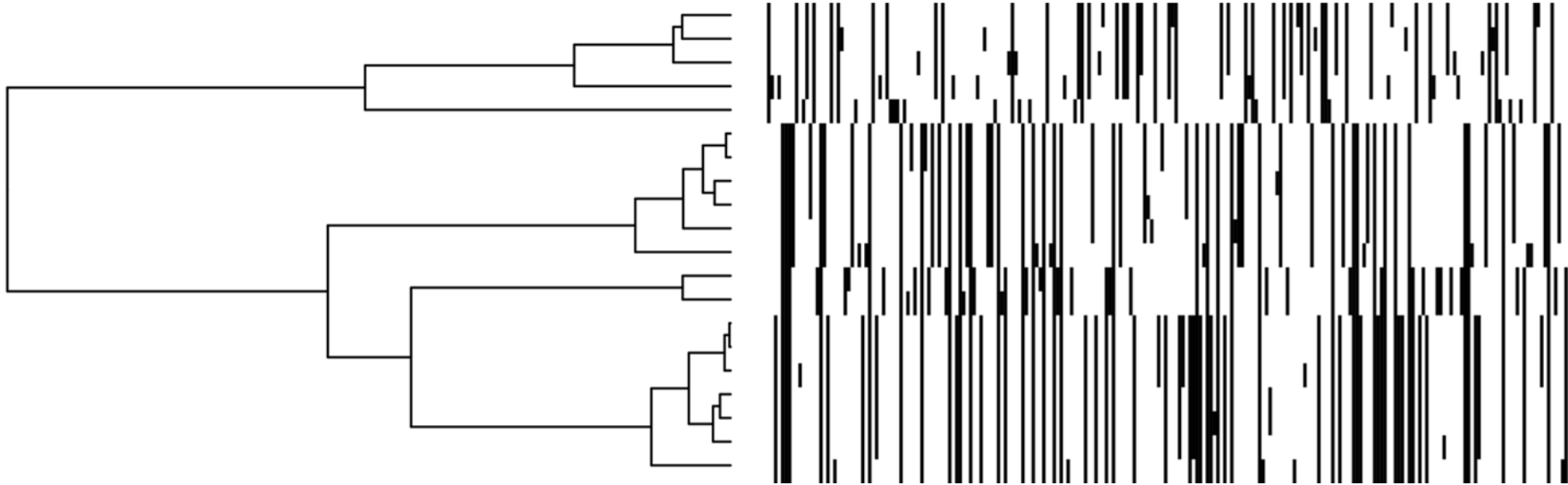
[Mac OSX - v1.0.16](#)

In the downloaded directory, we have included a toy data set. You can try out **Relate** using this toy data set by following the instructions on our [getting started](#) page.

If you have any problems getting the program to work on your machine or would like to request an executable for a platform not shown here, please send a message to [leo.speidel \[at\] outlook \[dot\] com](mailto:leo.speidel[at]outlook[dot]com).

We document changes to previous versions in a [change-log](#).

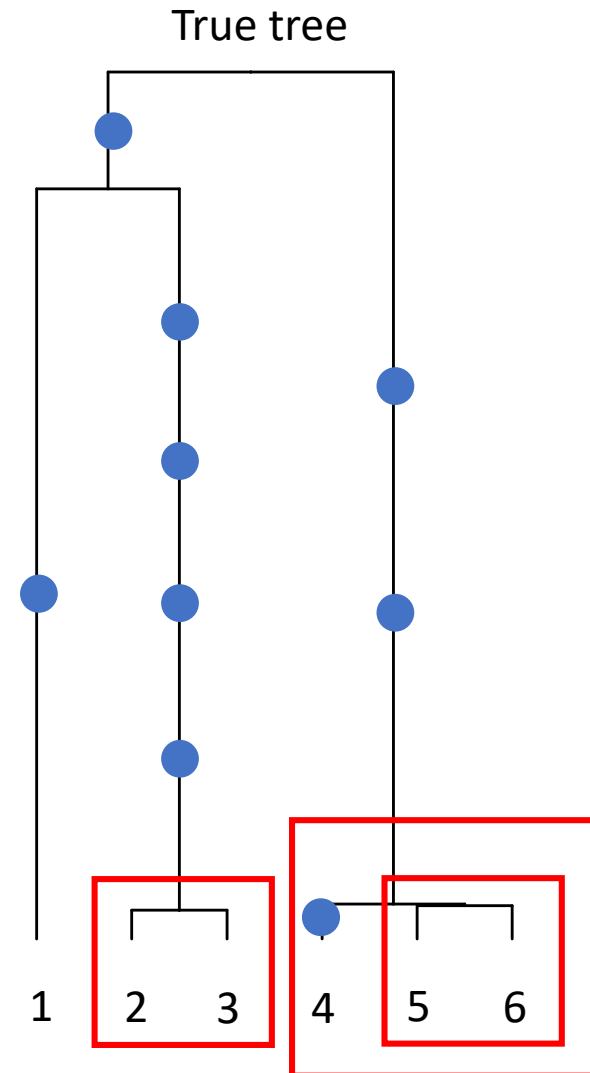
Data, and the underlying tree structure



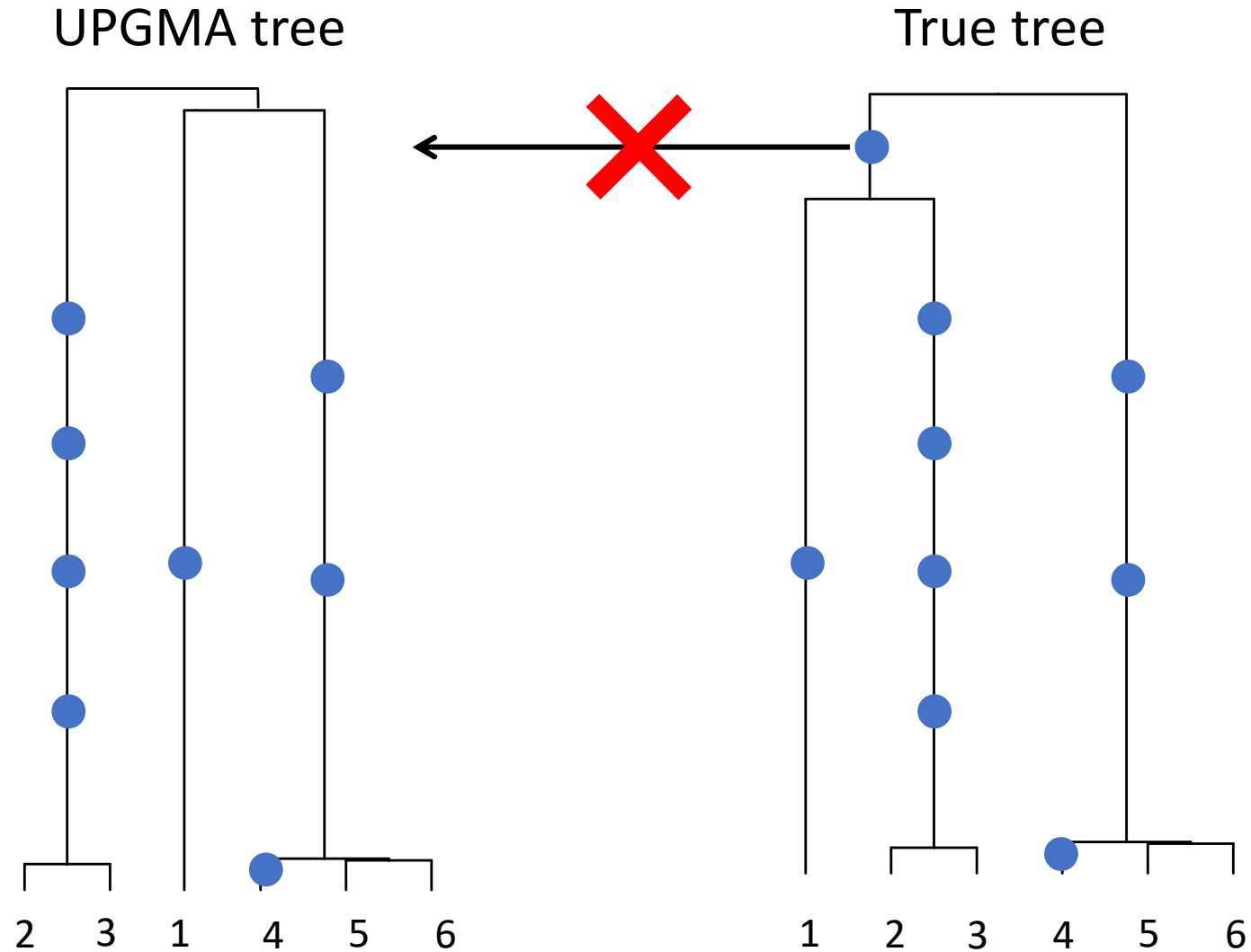
- This is a simulation without (for now) recombination
- Every mutation occurs only once on the tree....so
- **Every mutation shows the existence of a branch, suggesting we can build the tree given enough mutations**
- **Given the tree, we can uniquely map each mutation to its branch**

A basic tree-builder: UPGMA

- We are confronted with variation data – how do we build a tree?
 - UPGMA coalesces lineages with smallest number of (averaged) pairwise differences
 - (2,3) and (5,6) are coalesced first
 - (5,6) and 4 are coalesced
 - Now, pairwise difference of
1 and (2,3) is 5
1 and (5,6) is 4
1 and 4 is 5
- 1 and (4,5,6) are coalesced next!



UPGMA tree cannot be correct, given the data

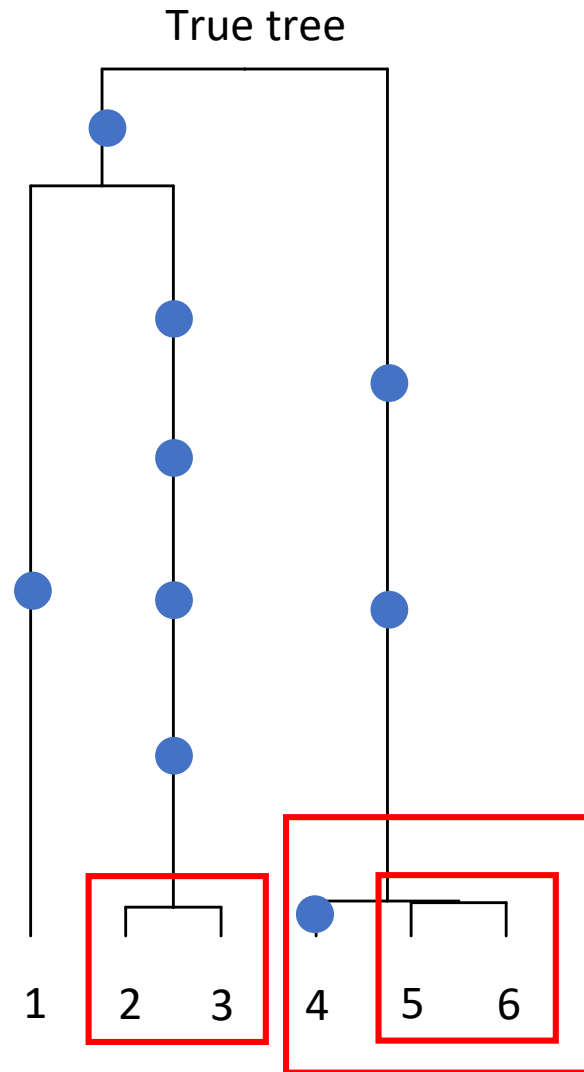


The UPGMA tree cannot be correct, because it does not include any branch whose descendants are sequences 1,2,3. How can we fix this?

Towards Relate: counting **derived** mutations to build the correct tree

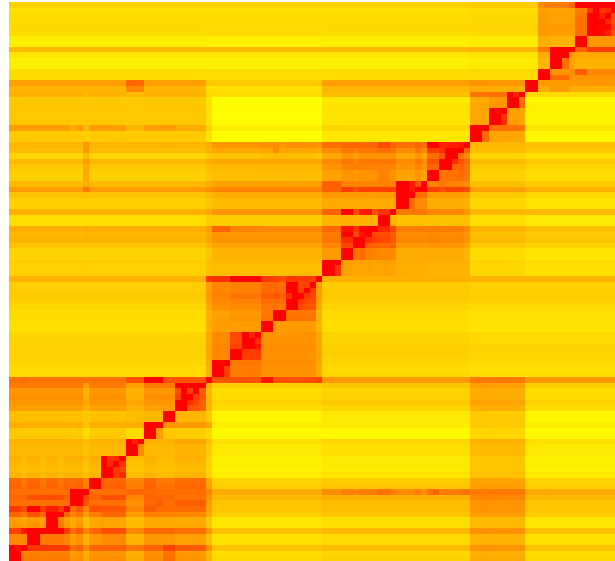
		relative to		
		1	(2,3)	(4,5,6)
Mutation carrier	1	0	1	2
	(2,3)	4	0	5
	(4,5,6)	2	2	0

Avoids adding information of two branches!



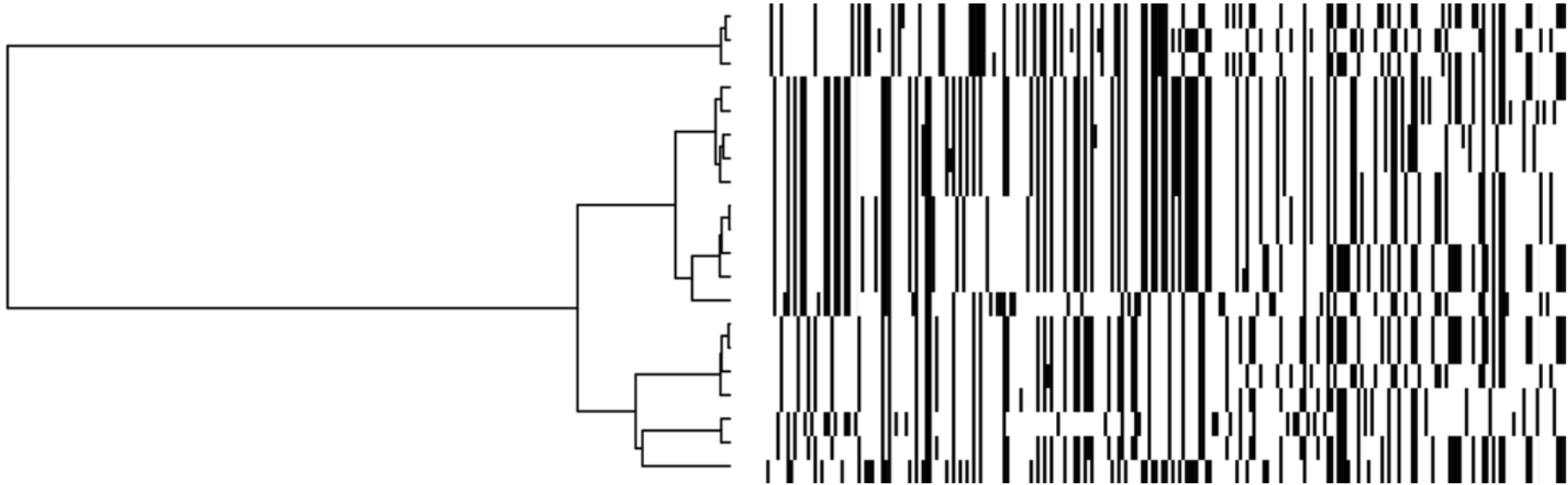
How does Relate work?

- In the **no-recombination** case, it first counts numbers of derived mutations for each pair:



- Performs coalescences between mutually most similar lineages
- Guaranteed to produce a tree matching the data!
- First builds a tree structure/topology (times are deferred for now)

Accounting for recombination

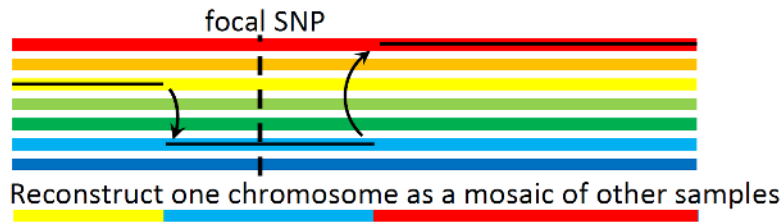


- Recombination means pairs of sequences are most similar for only **stretches of DNA**
- Use a HMM to (intuitively) identify these stretches, count derived mutations only within them, then proceed as for no-recombination case
- More formally, we use a modified version of Li+Stephens (Genetics, 2003) where we count differences

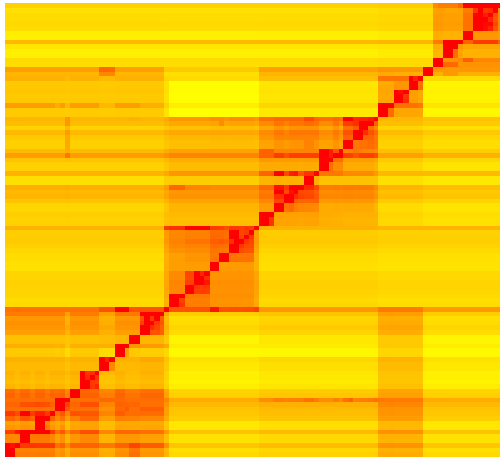
Summary of Relate pipeline

Hidden Markov model (HMM)

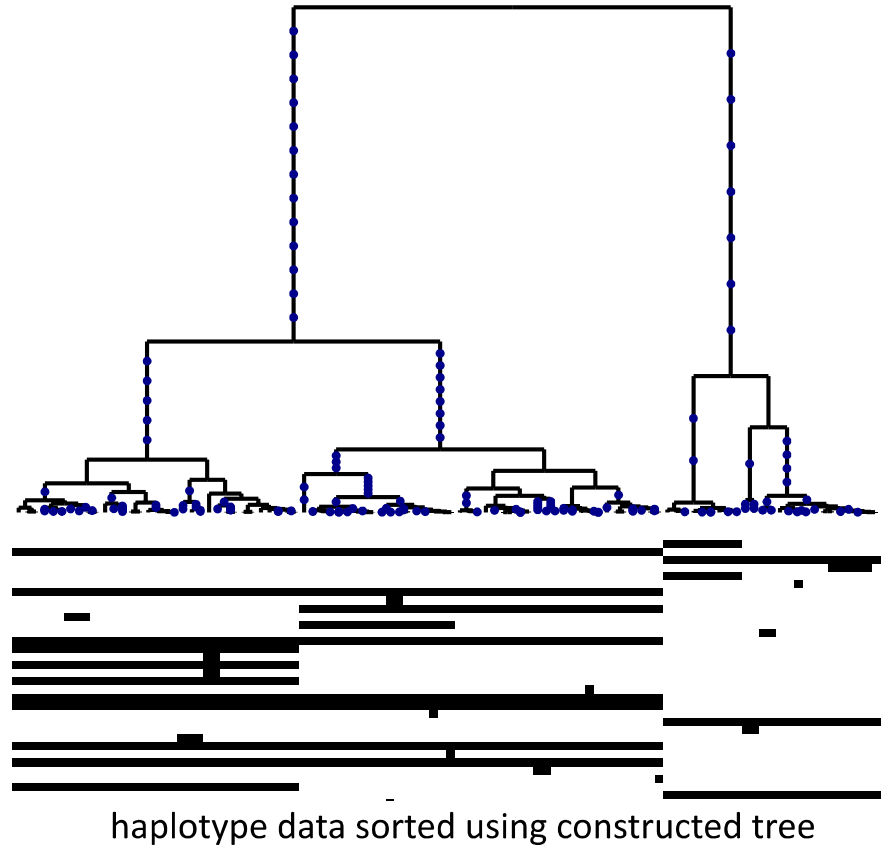
Li and Stephens, Genetics, 2003; Lawson et al., PLOS Genetics, 2012



Distance matrix for focal SNP



Hierarchical clustering
&
MCMC for branch lengths

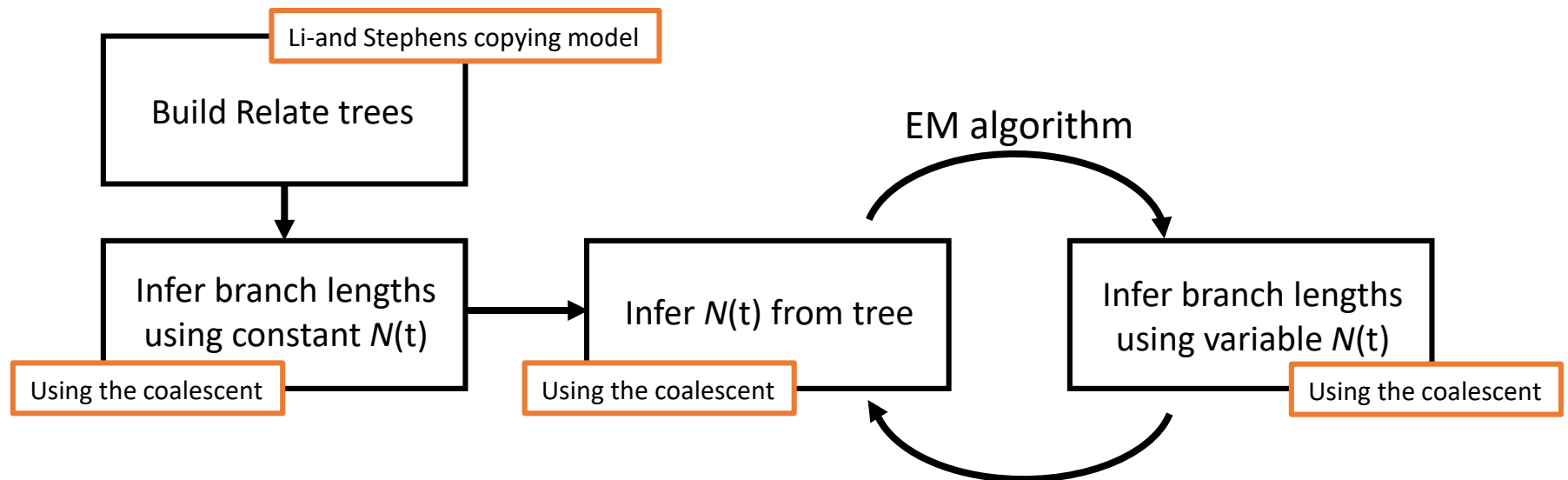
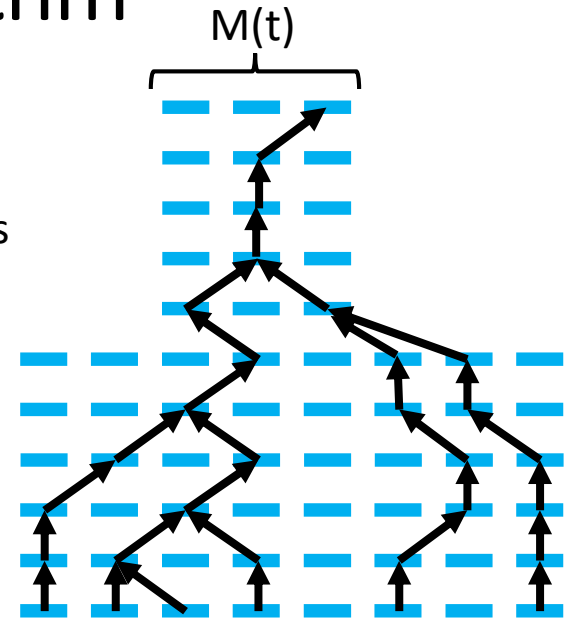


Coffee break!

After the break:
branch lengths,
variable population
sizes, and
applications of
Relate

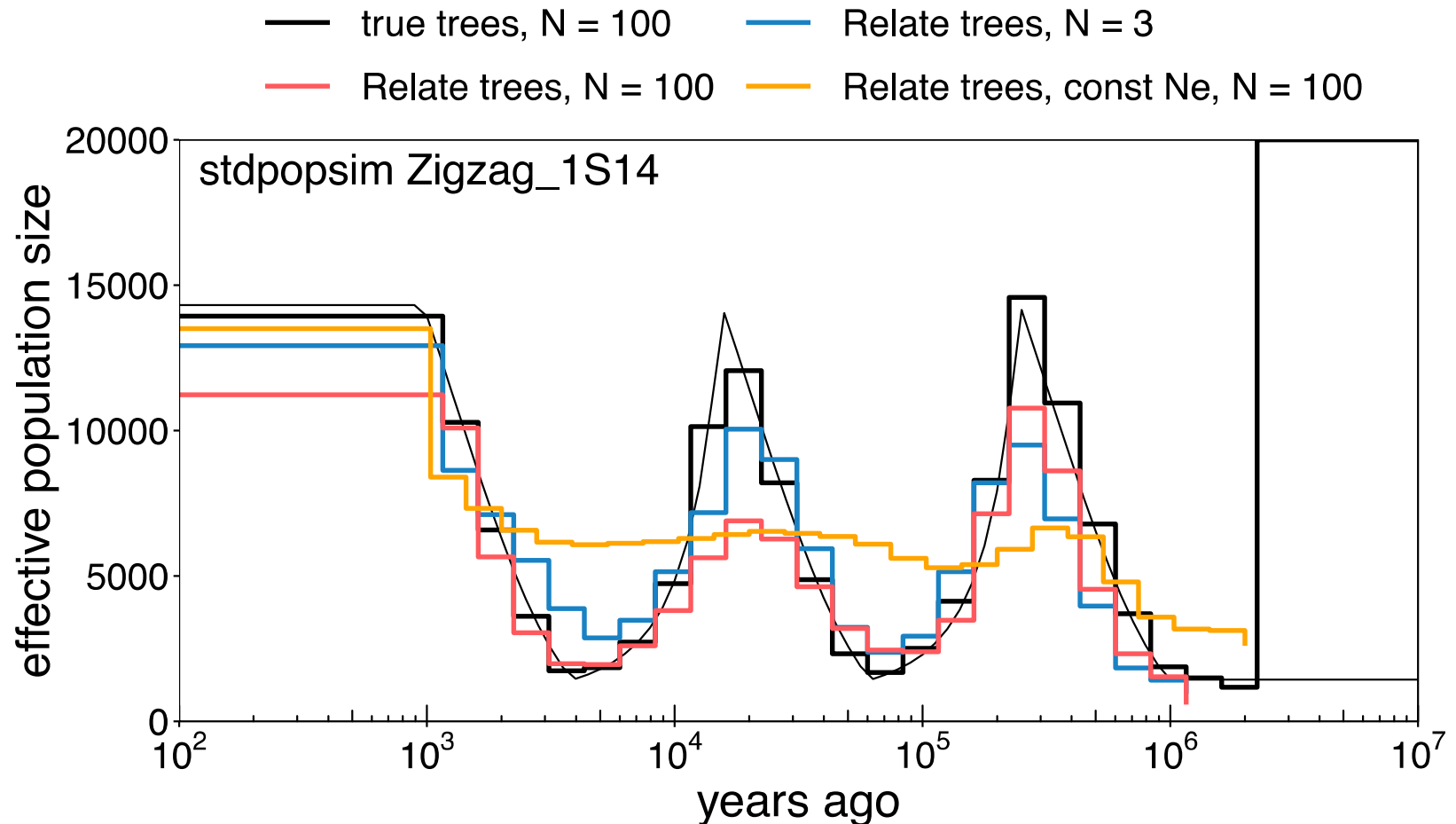
Branch lengths and population size are estimated jointly in an EM algorithm

- Expected branch lengths depend on population size $N(t)$
(or coalescence rates $1/2N(t)$)
- While there are j lineages, the rate at which a coalescence happens is $\binom{j}{2}/2N(t)$ a time t ago
- Demography is shared genome-wide, so we average across trees
- So within a time interval, scaled fraction of trees where coalescence occurs is inversely proportional to $N(t)$



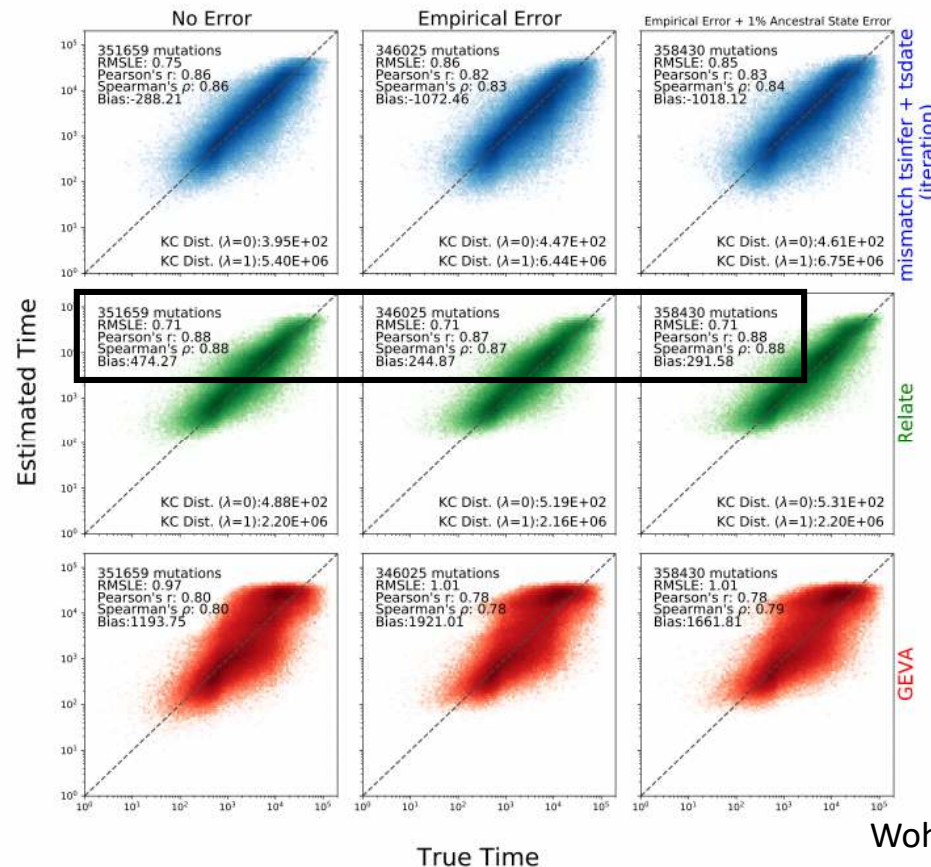
Simulation: population size changes through time

- Effective population size = inverse coalescence rate
- N: number of diploid samples



Speed and accuracy of Relate

- About 14,000 times faster than previous method, ARGWEAVER (1 min. vs. 200 hours), slower than tsinfer + tsdate
- Builds “correct” tree if no recombination
- Accurate, robust to data errors
- Can sample posterior branch lengths





Example I: 1000 Genomes Project data:

- 4956 haplotypes from 26 populations
- ~71,000,000 biallelic SNPs
- ~93% of SNPs map uniquely to a tree (80% for CpG mutations due to repeat mutations)

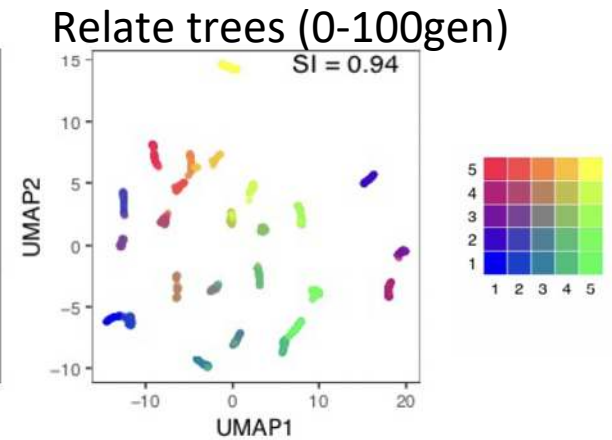
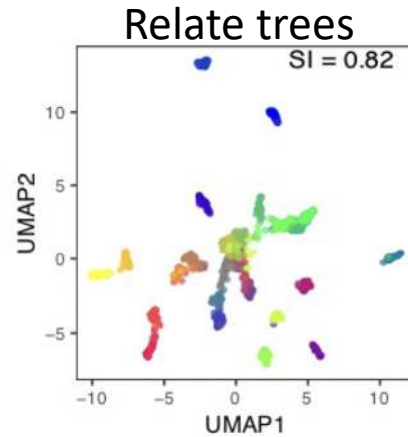
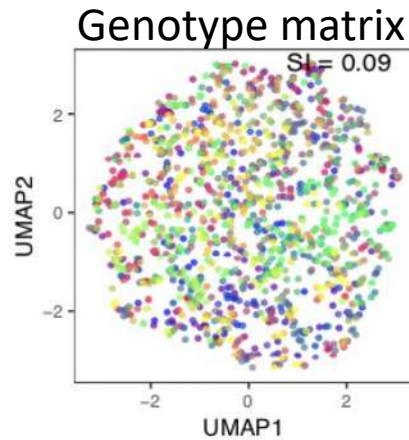
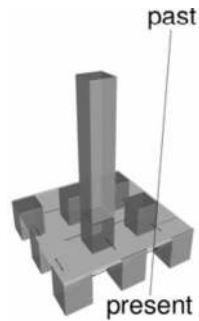
Run time: ~4 days on 300 cores

Genealogy-based inference of human evolutionary history

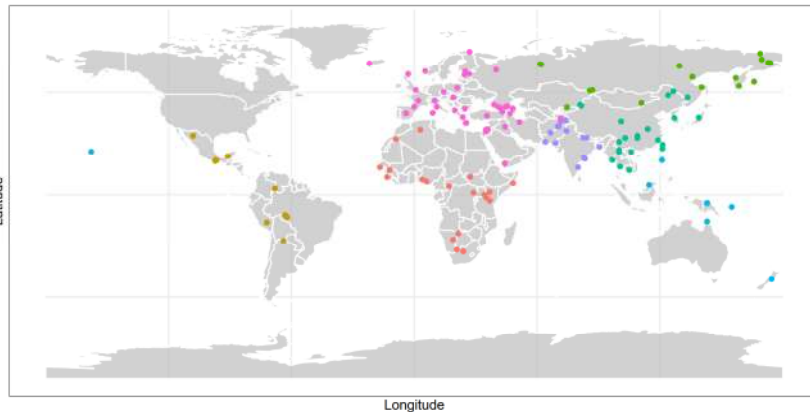
One reconstruction of history, many applications that are **self consistent**

Inferring fine-scale population structure and how it changed through time

Fan, Mancuso, Chiang, bioRxiv, 2021

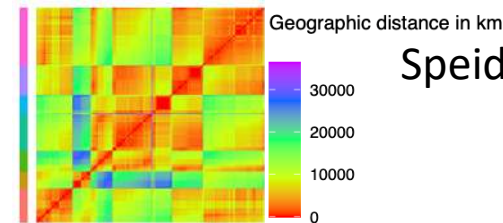


Simons Genome Diversity Project

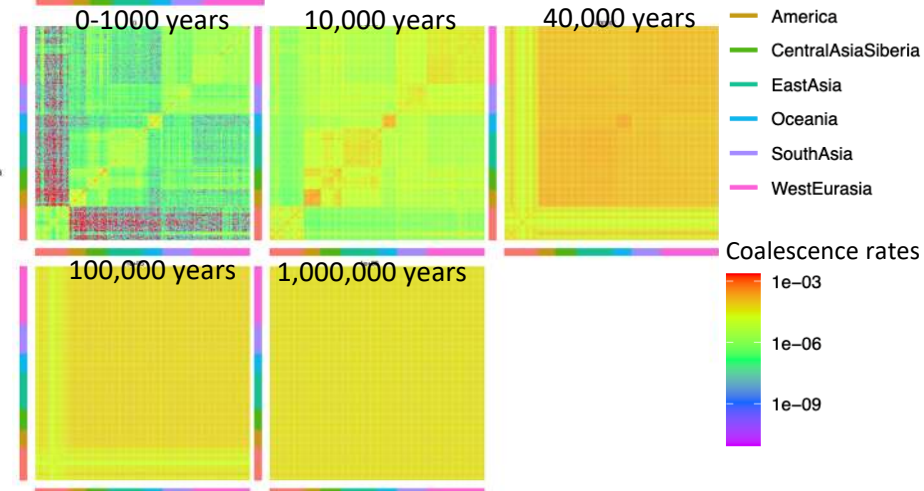


region

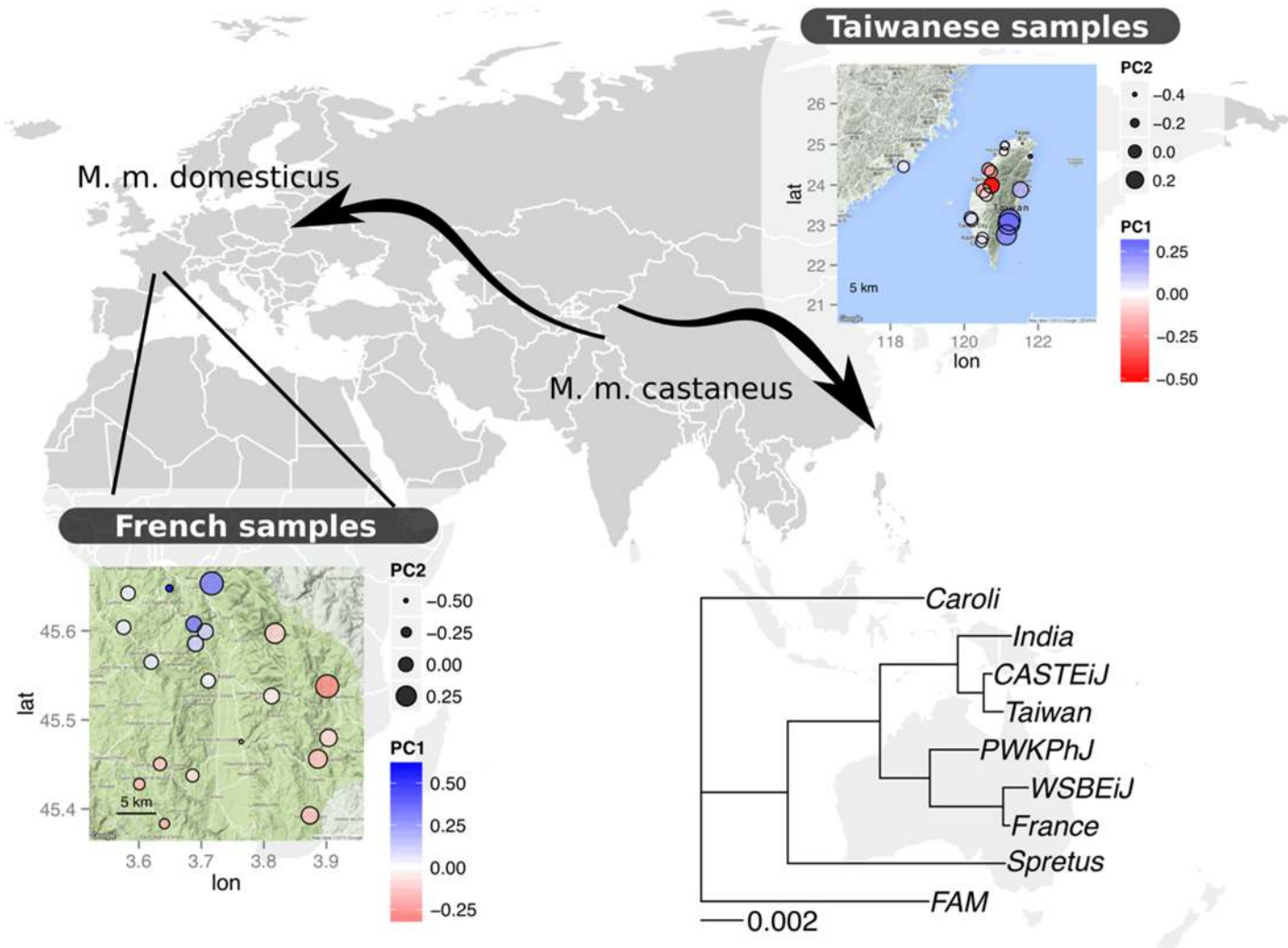
- Africa
- America
- CentralAsiaSiberia
- EastAsia
- Oceania
- SouthAsia
- WestEurasia



Speidel et al, MBE, 2021

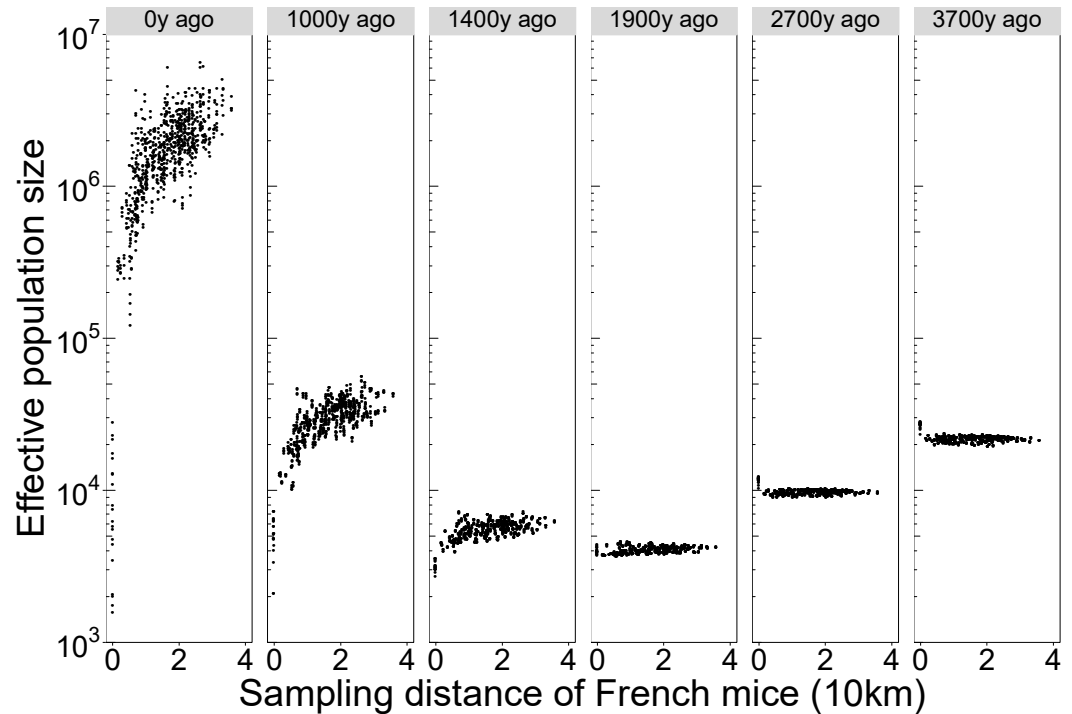
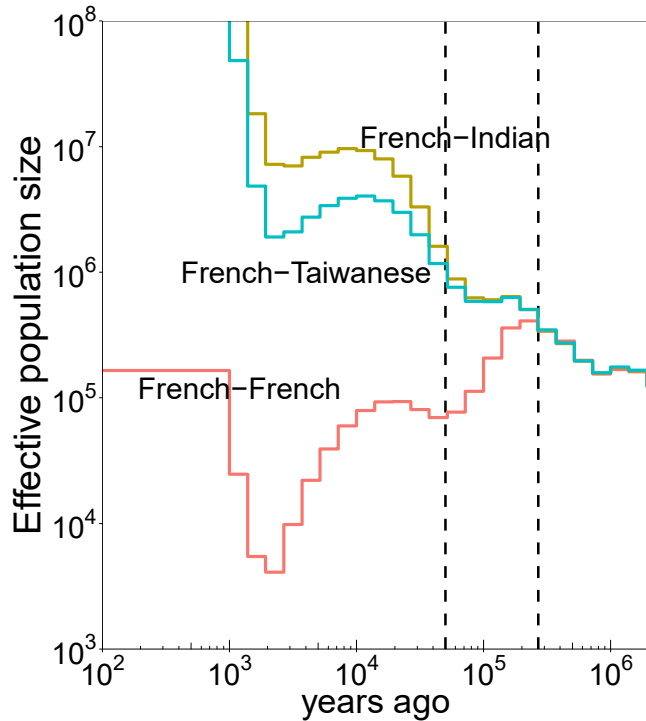


Relate applied to 50 wild mice sampled in India, Taiwan, and France

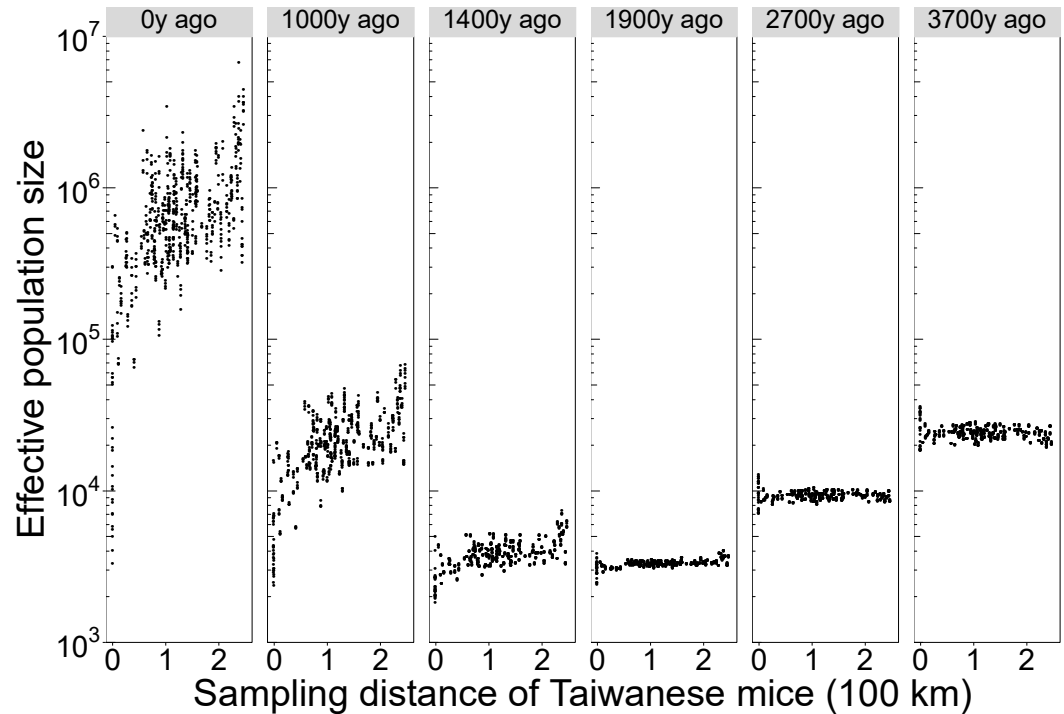
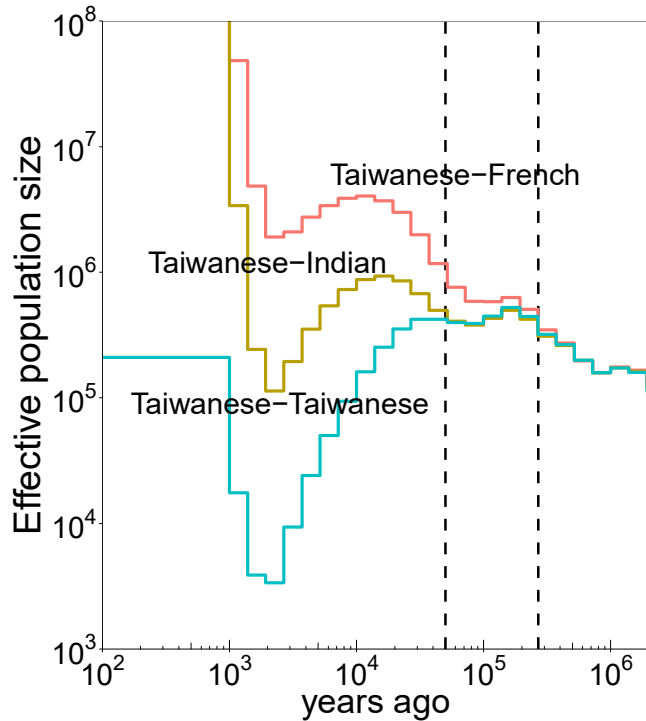


Runtime: 17 CPU hours for 19 chromosomes, Memory usage < 2.5 Gb

Population structure through time in French mice

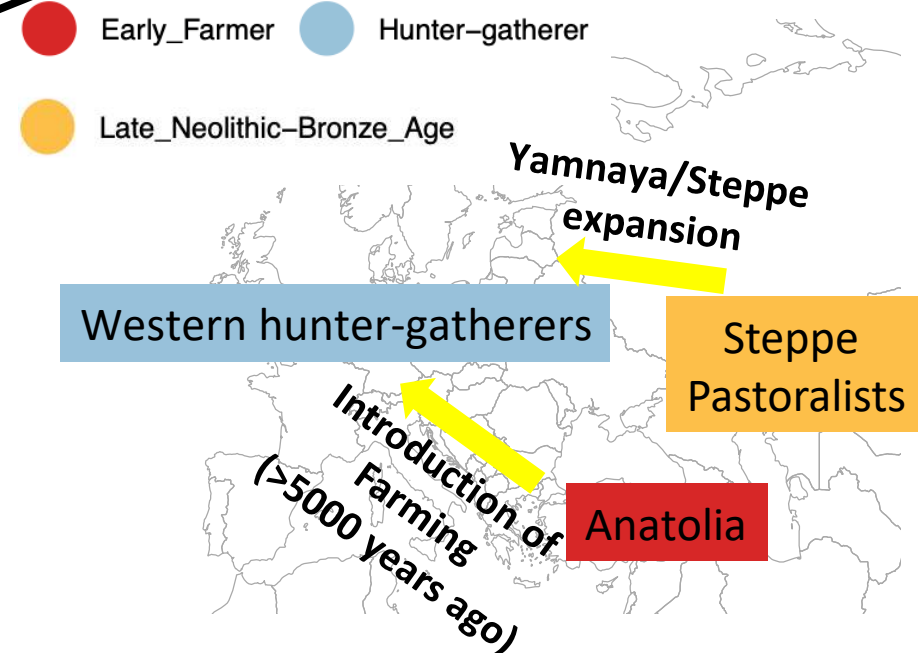
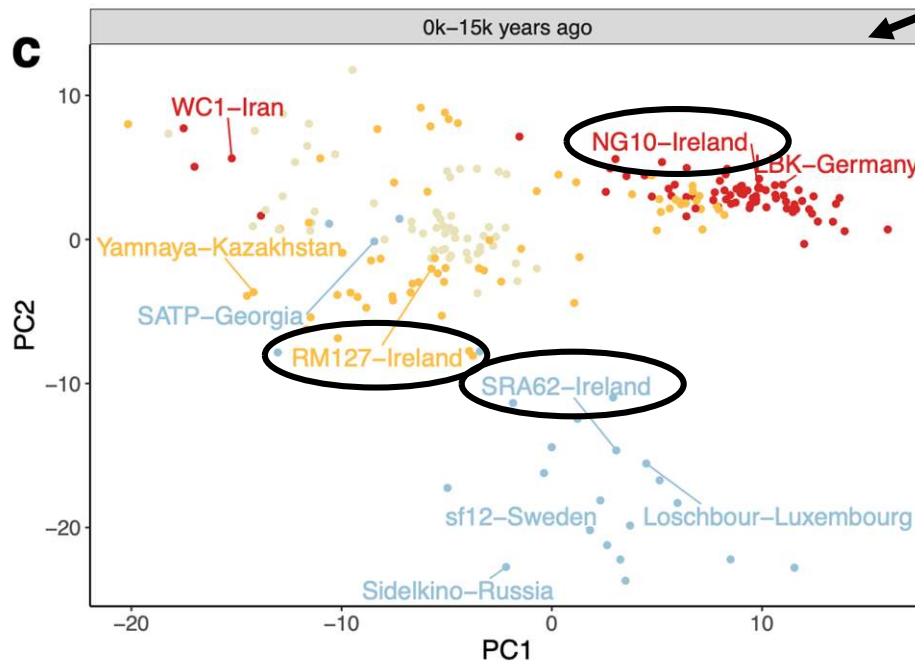
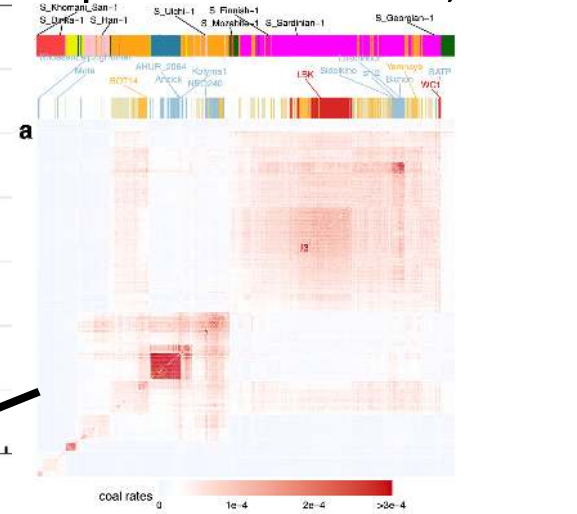
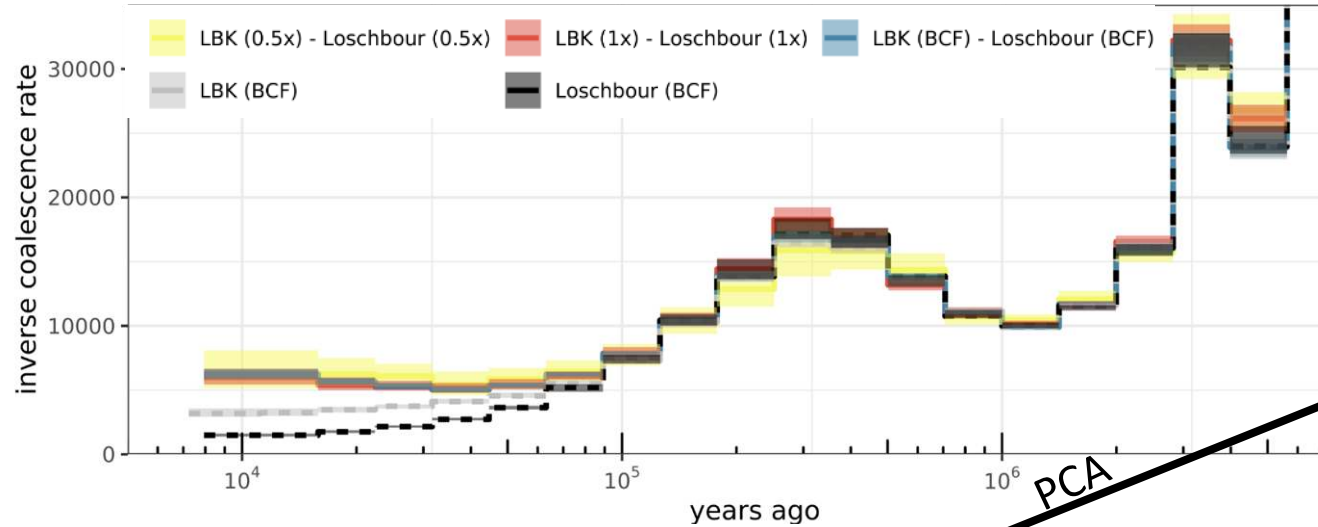


Population structure through time in Taiwanese mice

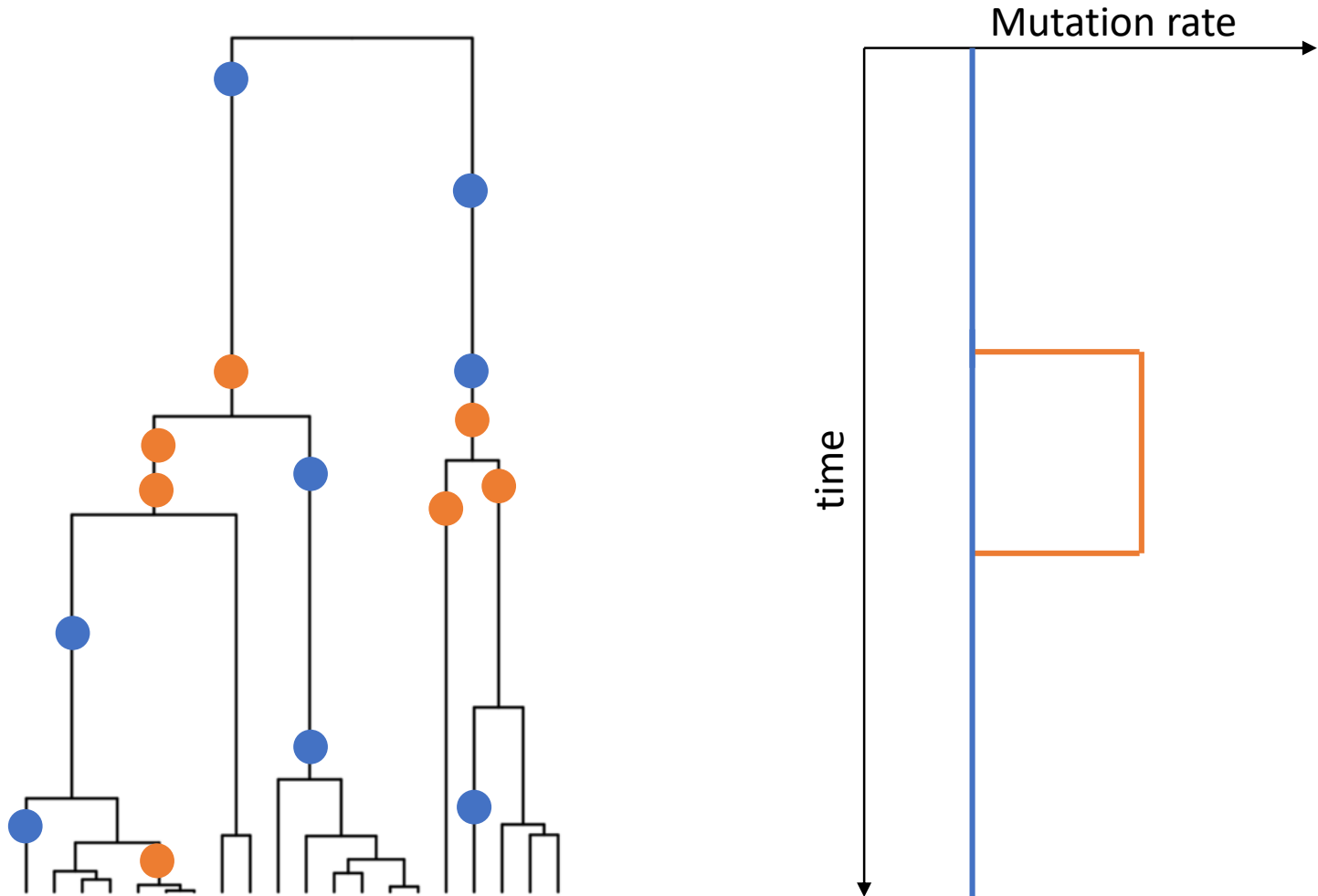


Colate: Inferring coalescence rates for low-coverage, unphased (ancient) genomes

Speidel et al. MBE, 2021

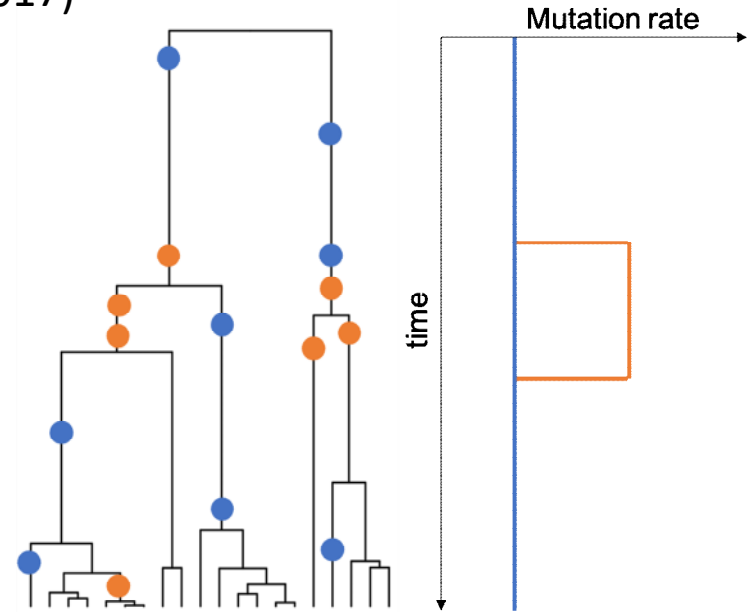
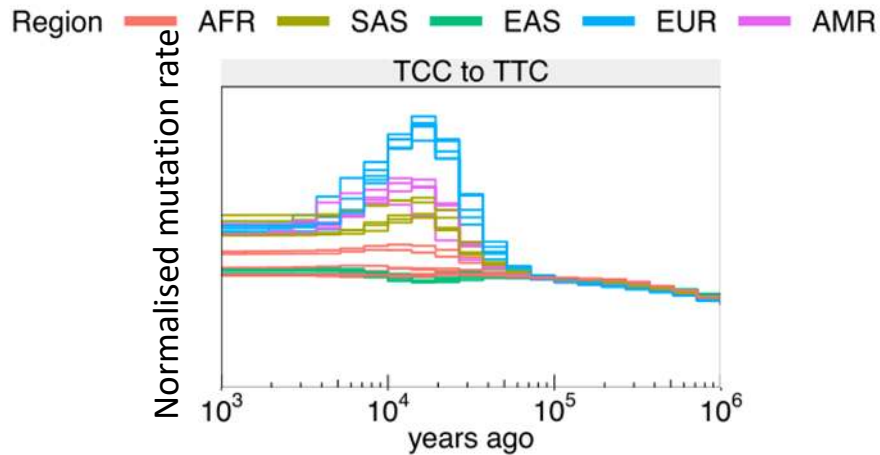


Reminder: Clusters of mutations in time can capture changes in mutation rate



TCC/TTC mutation rates experienced a strong increase in the Upper Paleolithic

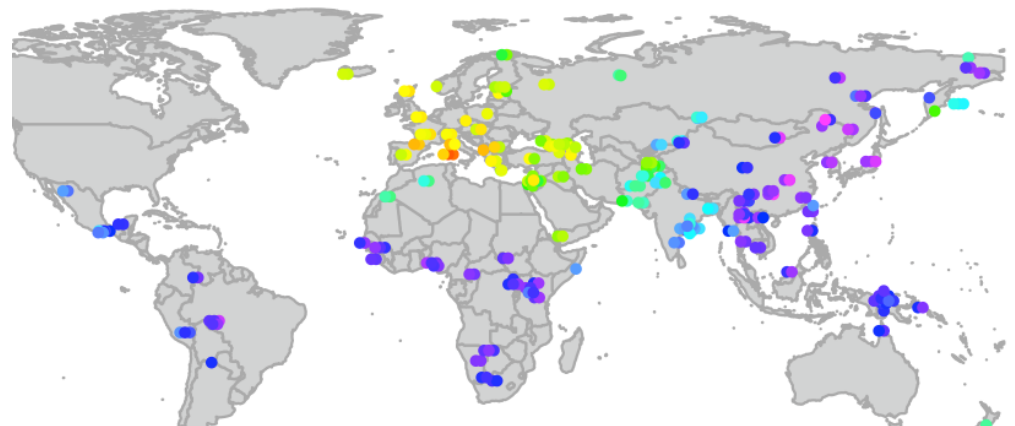
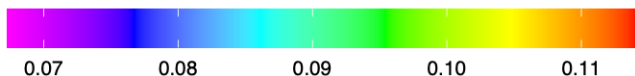
- First reported by Kelley Harris (PNAS 2015, eLife 2017)
- Unknown cause (genetic?, environmental?)
- Previously mainly studied in modern groups



Speidel, Nature Genetics, 2019

How did this spread to all West Eurasians today?

Colour shows strength of elevation in TCC/TTC mutation rates



TCC/TTC mutation rate increase happened >15k years ago, and spread among hunter-gatherers before farming

Speidel et al, MBE, 2021

Ice Age Europe

>20,000 years ago



earlier HG groups

e.g., Sunghir3 (30k year old), Russia

e.g., Sidelkino (11k year old),
S Russia

Eastern HG

e.g., Bichon (14k year old),
Switzerland

Western HG

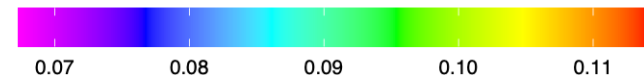
Caucasus HG

e.g., KK1 (9k year old)

Anatolia

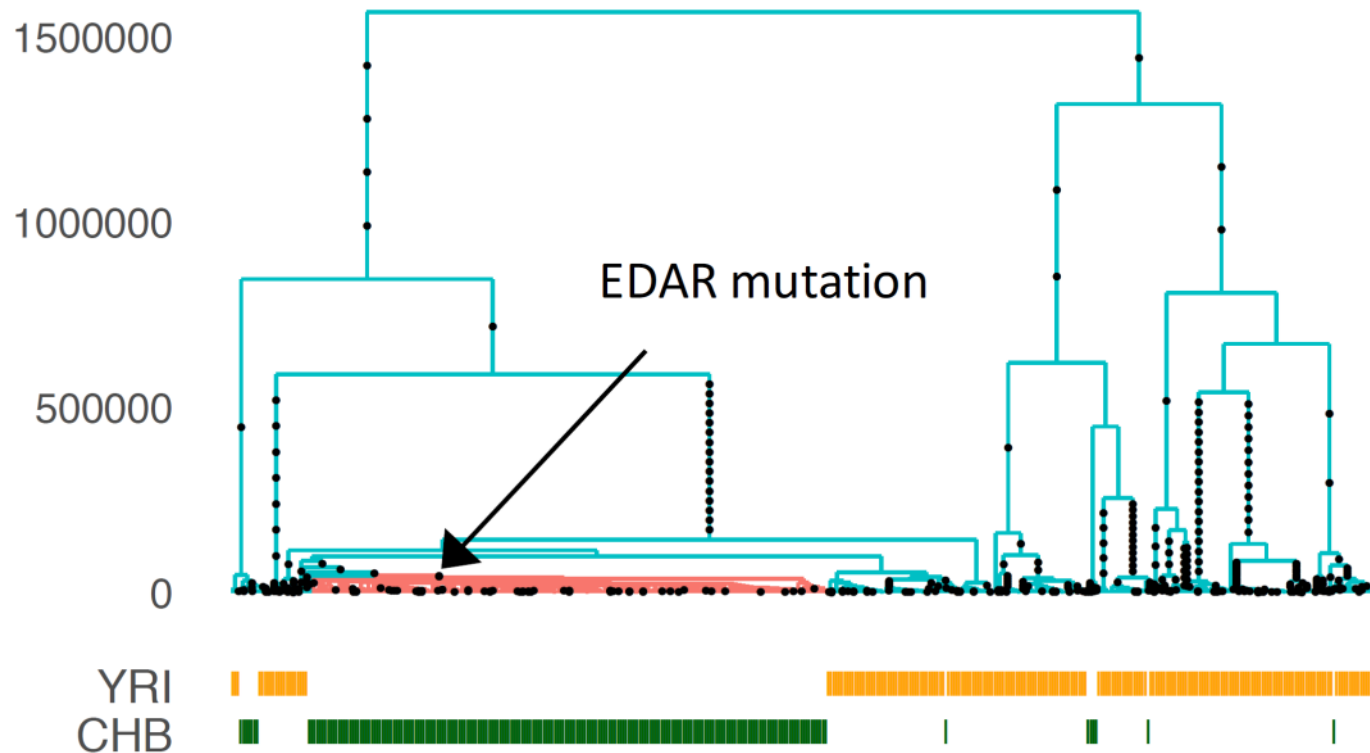
e.g., Bon002 (10k year old)

Colour shows strength of elevation in TCC/TTC mutation rates

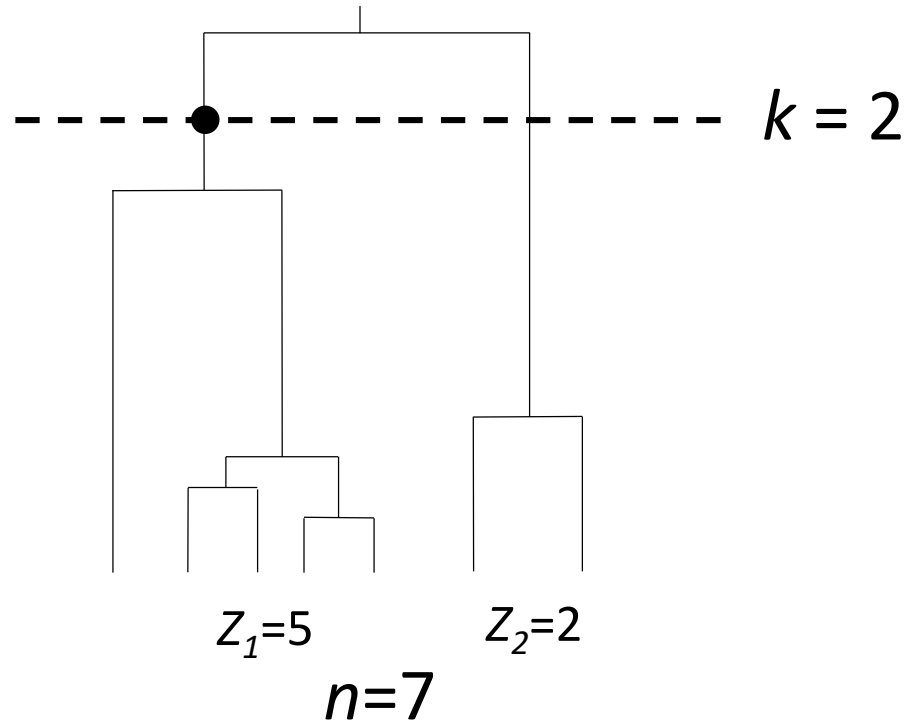


Detecting signals of positive selection

- First, consider a single mutation
- Genetic adaptations to changing environment, diet, lifestyles,...
- Use trees incorporating demographic history:



How quickly does a mutation spread in the neutral case?

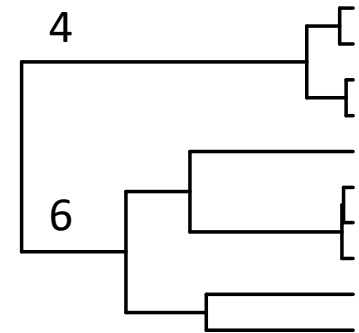
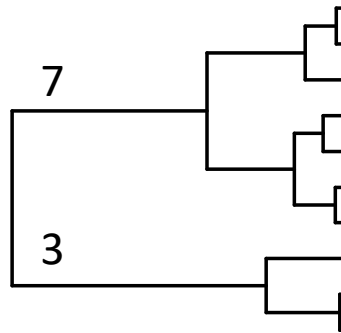
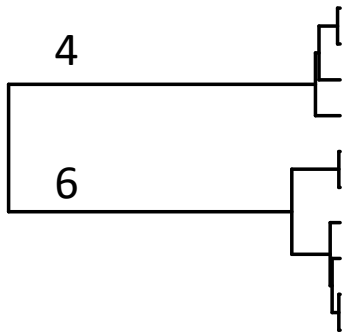
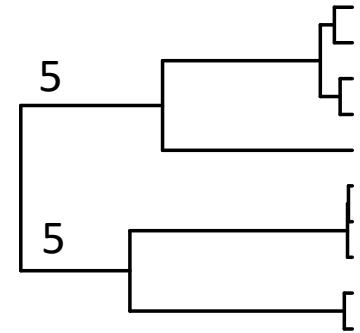
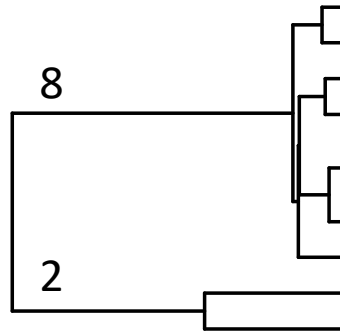
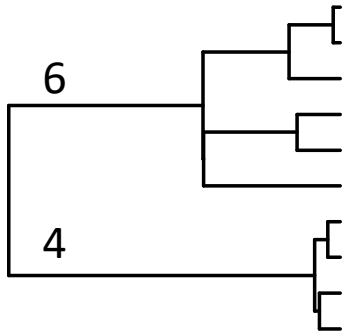
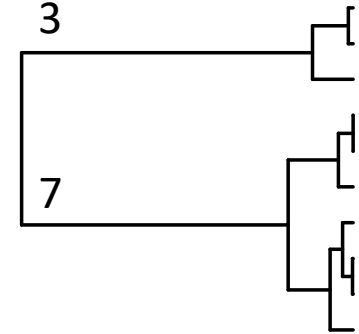
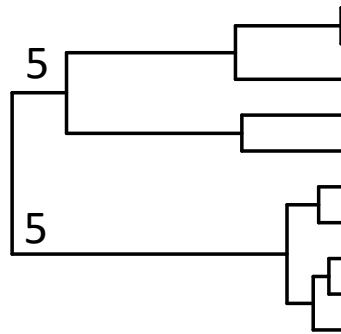
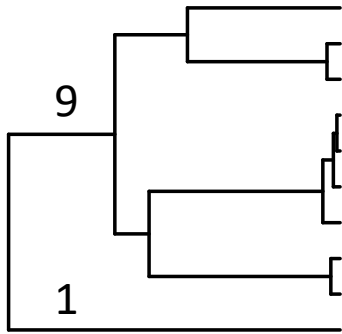


The coalescent is simple so it is possible to analytically write down the probability a mutation arising while k lineages are in the tree has at least some number of descendants: yields a **p-value, testing a null hypothesis of no selection**

Example: if $k=2$, this is just a **uniform distribution**

$P(5 \text{ descendants})=1/6$; $P=P(5 \text{ or more descendants})=1/3$

The $k=2$ case

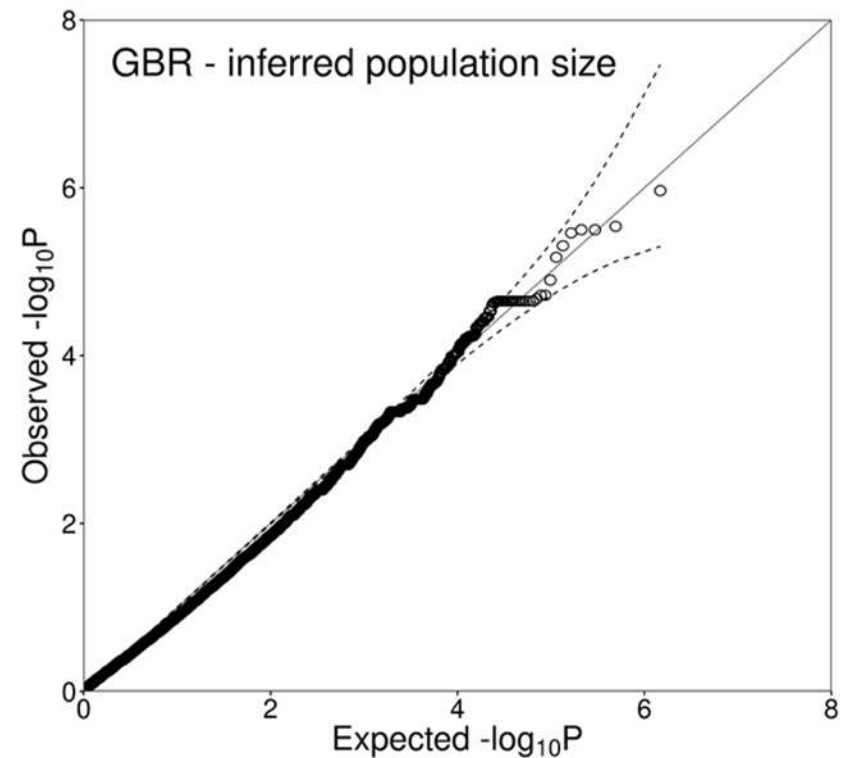
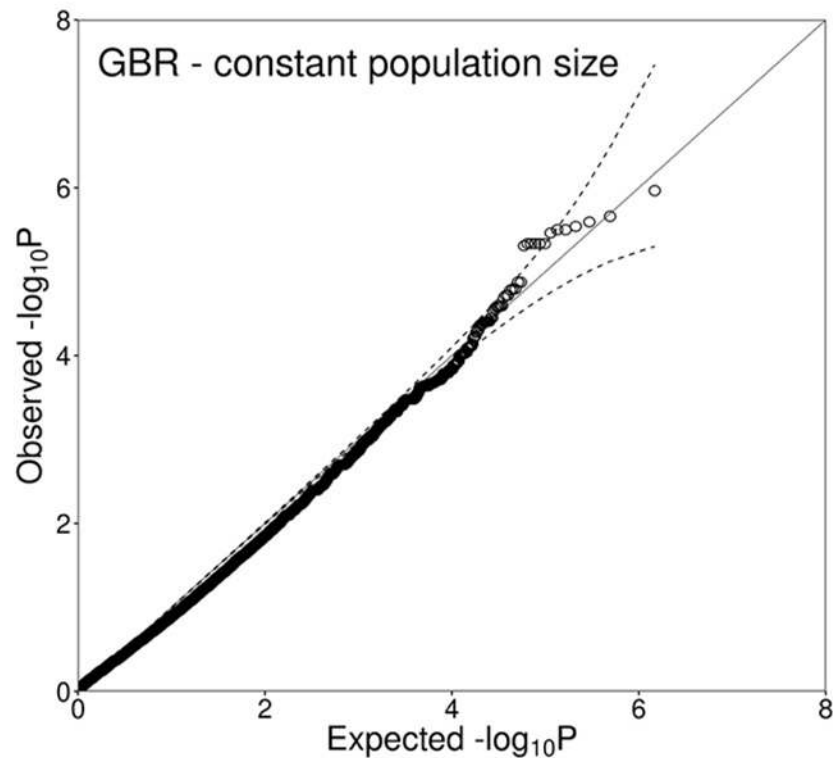


P-values: very well calibrated under null simulations of no selection

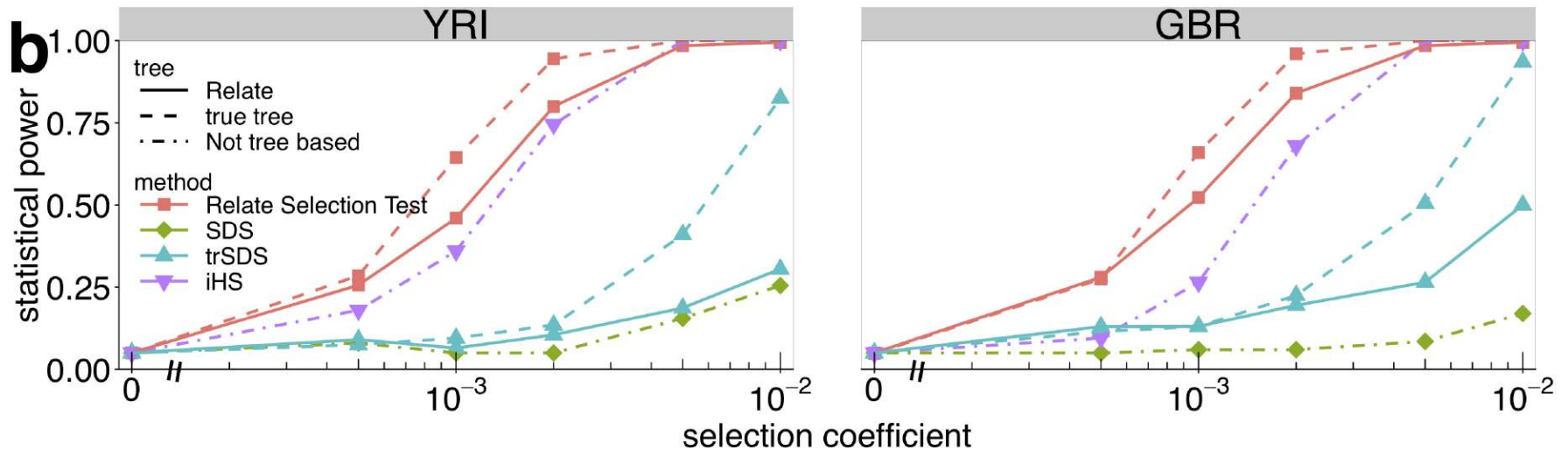
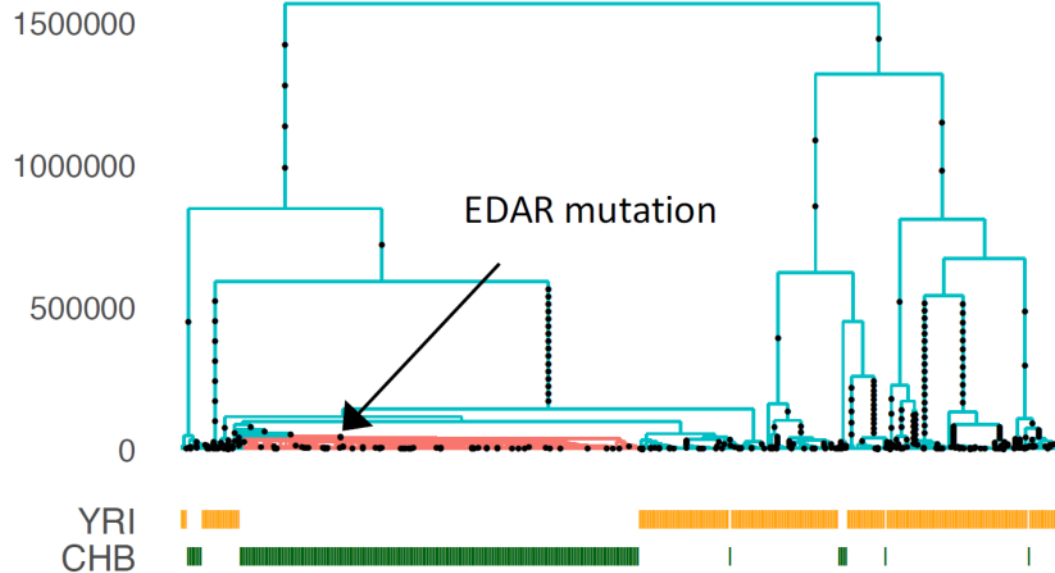
N=1000, 250Mb

Bottleneck population size

a

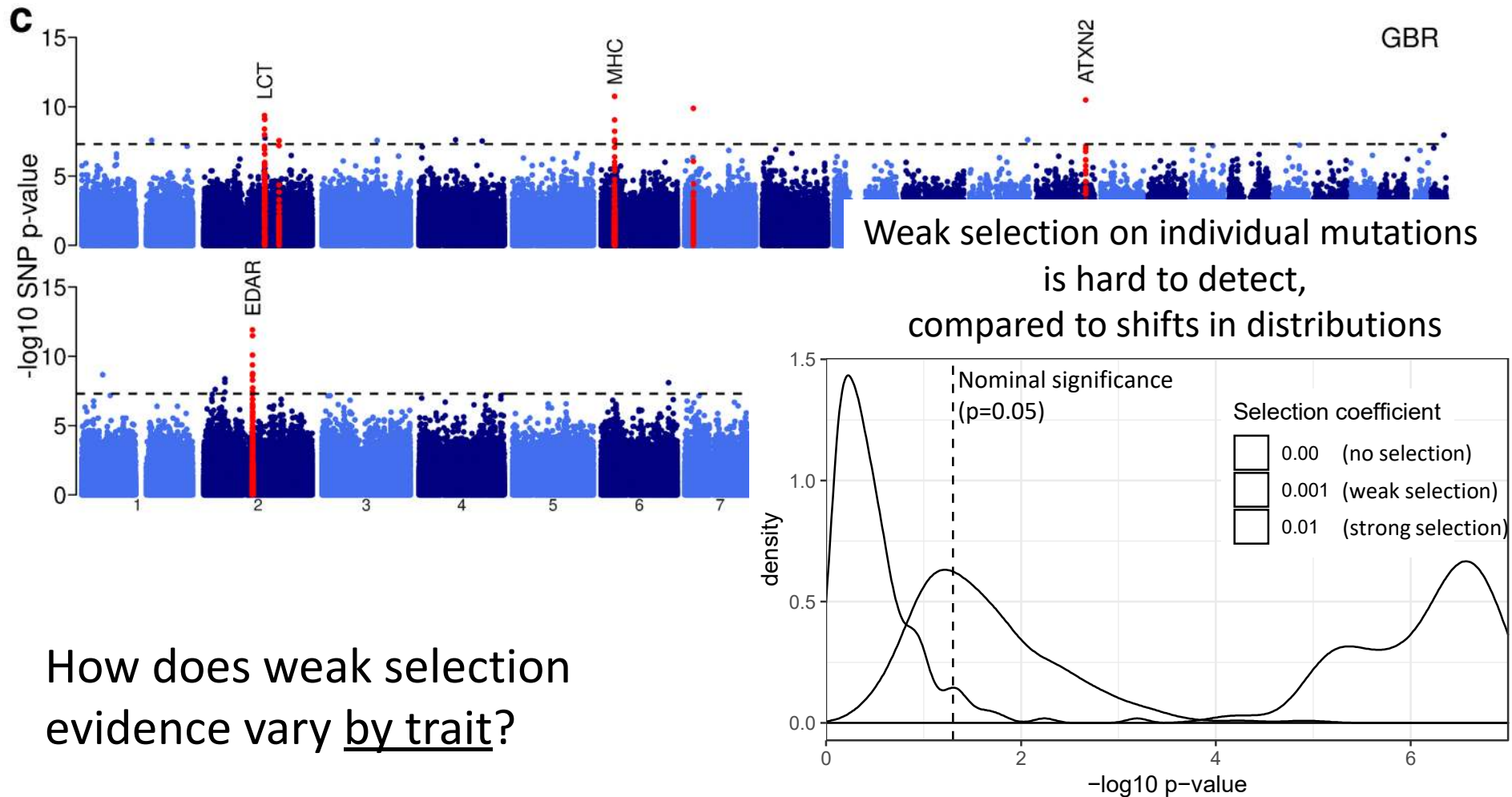


Improved power to see weak selection than existing approaches

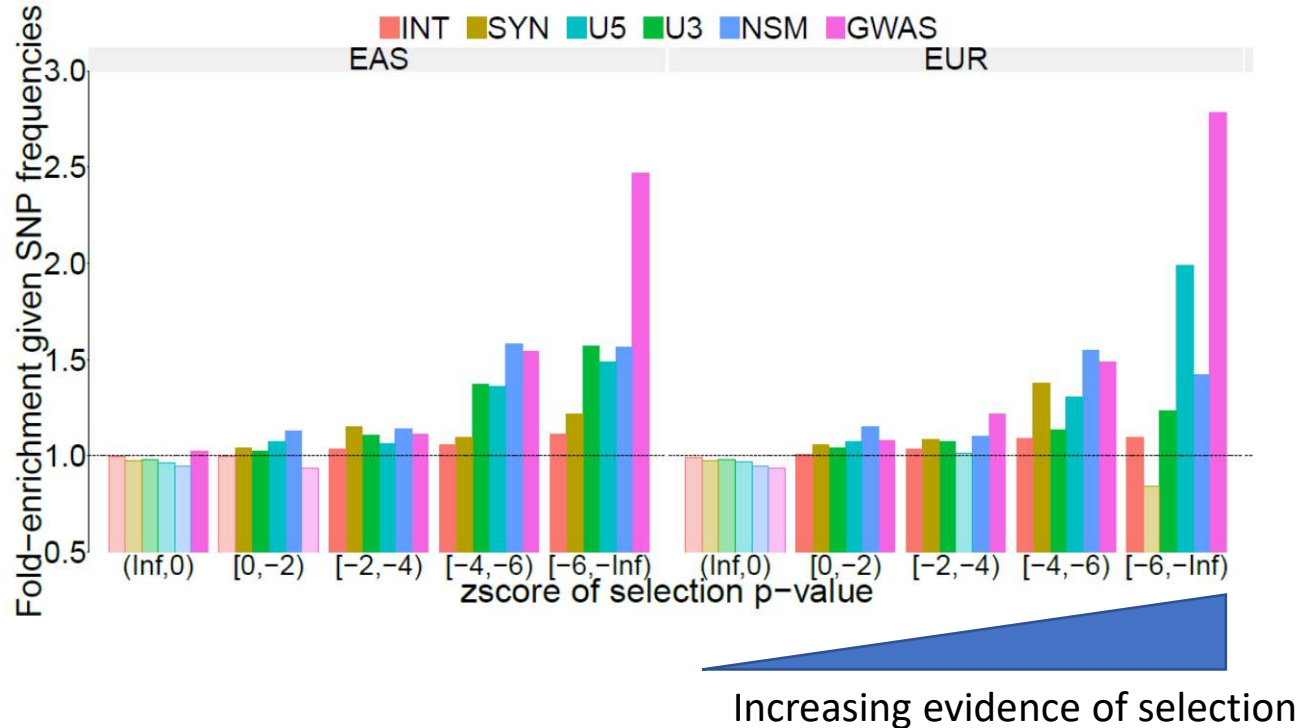


Genome-wide selection p-values

Given most traits are highly polygenic, expect mainly weak, polygenic selection



GWAS hits are most enriched, among selection signals we observe



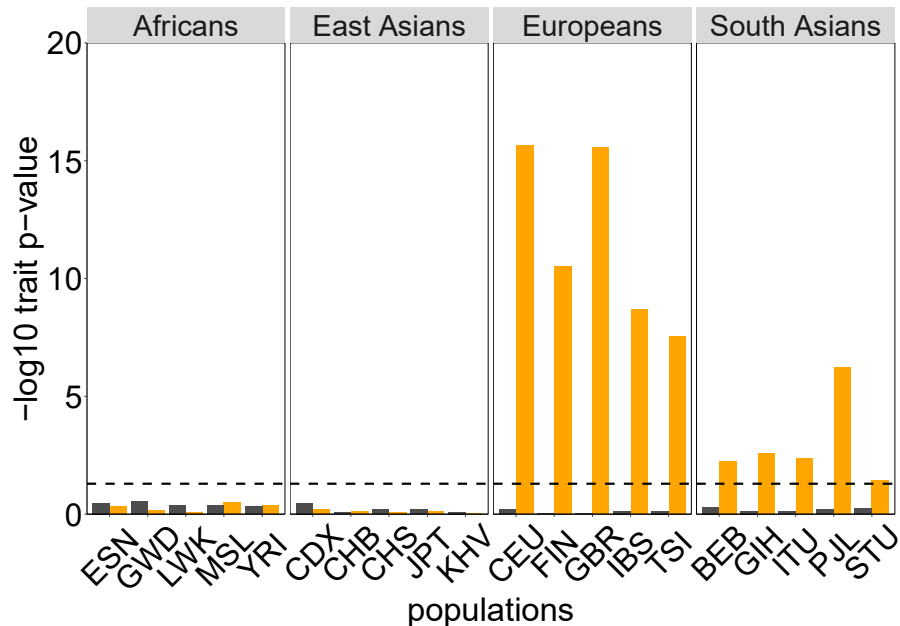
INT	Intronic SNP
SYN	Synonymous coding SNP
U5	5' untranslated region
U3	3' untranslated region
NSM	Non-synonymous coding SNP
GWAS	Genome-wide significant GWAS hits

Evidence of selection on a trait: hair colour

1. Use **effect direction** of "genome-wide significant" associations
2. Compare selection p-values to frequency matched random SNPs (Wilcoxon rank-sum test)

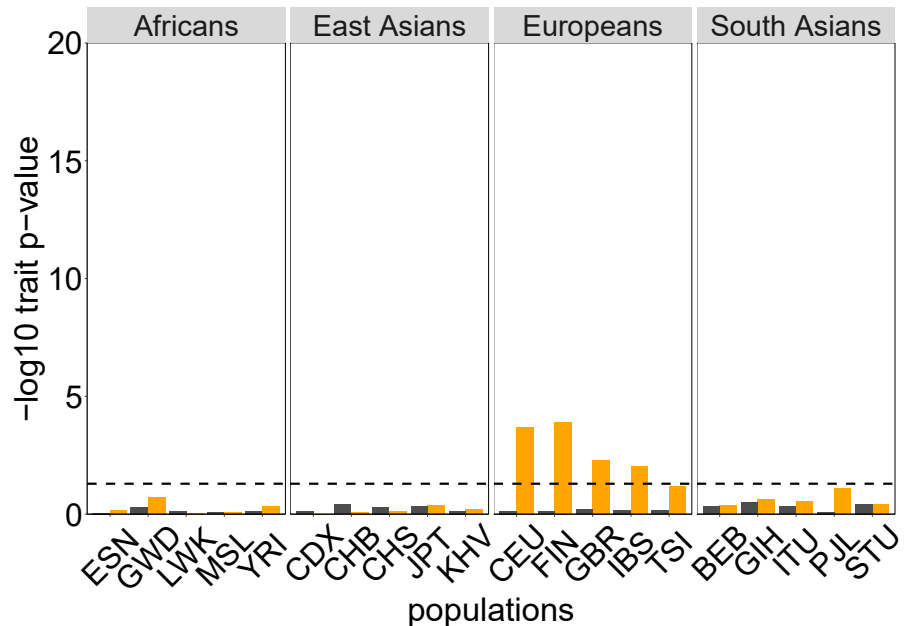
Relate p-values

■ Darker hair colour
■ Lighter hair colour



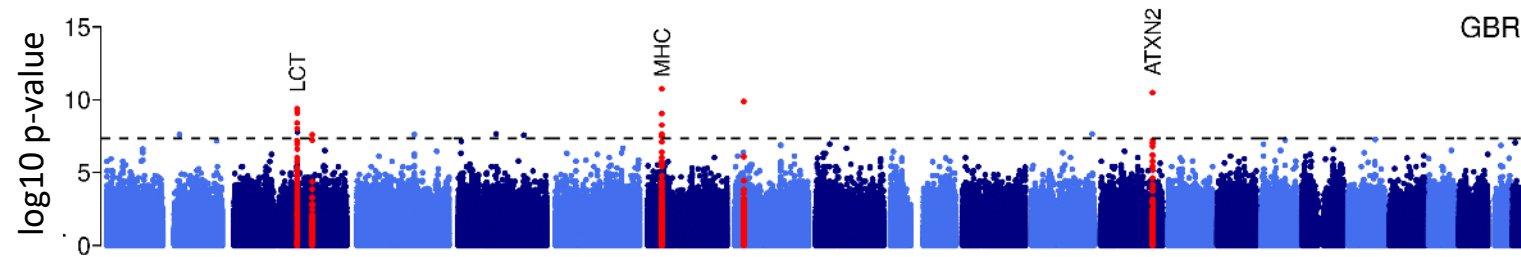
iHS scores

■ Darker hair colour
■ Lighter hair colour



Many key events in our evolutionary history are only implicated as subtle

Selection p-values: only a handful of “genome-wide significant” loci

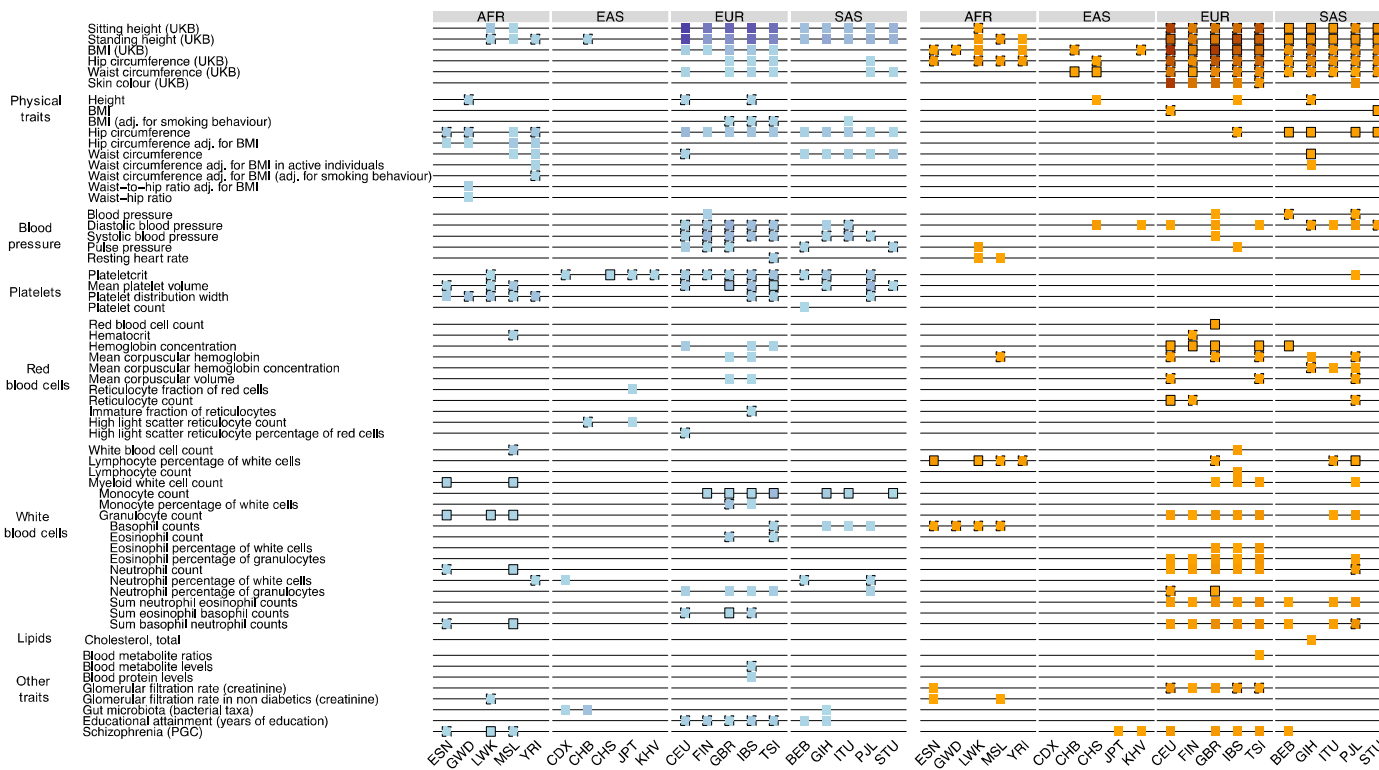


Lots of Polygenic selection signals



Evidence for trait increase

Evidence for trait decrease



Blood pressure

↑ (EUR, SAS)

Hip circumference

↓ (SAS)

Plateletcrit

↑ (AFR, EAS, EUR, SAS)

Hemoglobin

↓ (EUR, SAS)

Granulocyte

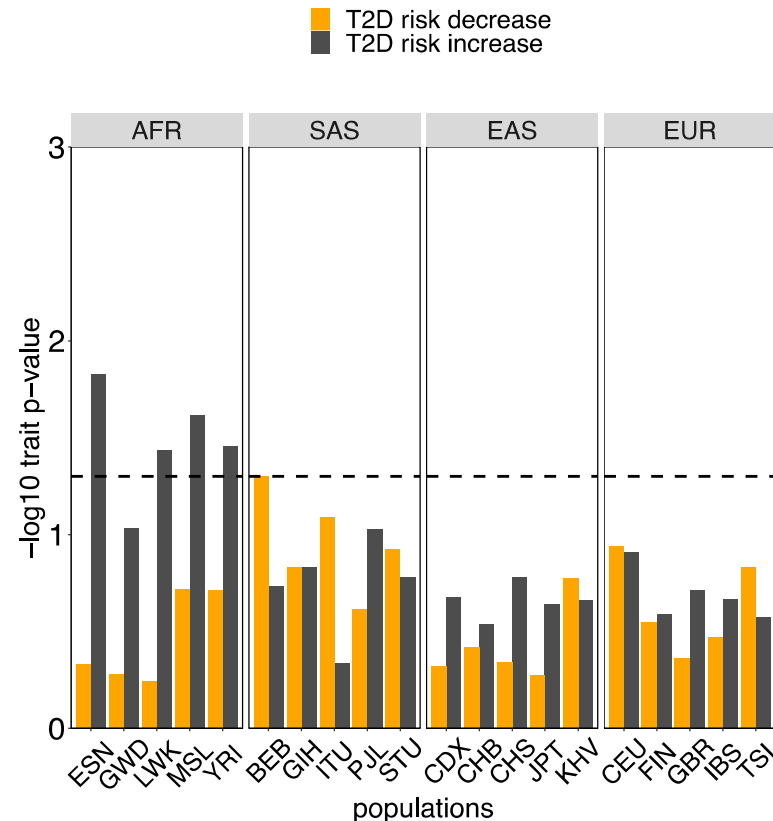
↑ (AFR) ↓ (EUR)

Evidence for selection increasing risk of T2D in African-ancestry people

Nature Genetics (2022), with Anubha Mahajan (Oxford), Mark McCarthy (Genentech)

- 171,262 cases and 1,075,072 controls from diverse ancestries
- 337 independent loci with T2D risk associations
- 209 (MSL) – 297 (FIN) segregating hits per population

a





- Correlated phenotypes
- Pleiotropy (in T2D, risk increase is fully explained by mutations associated with changes in body fat composition)
- Biased effect sizes (e.g., due to genetic structure)
- Unbalanced power for different ancestries

CLUES: Importance-sampling based method for inferring selection coefficients

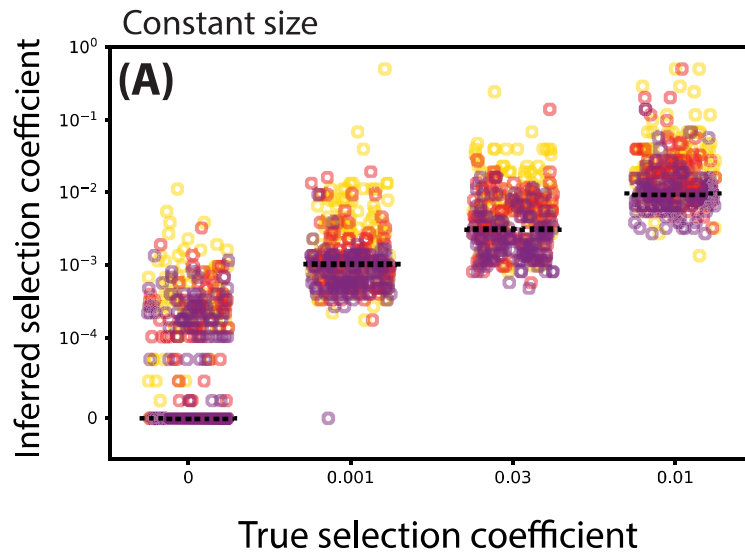
Aaron Stern



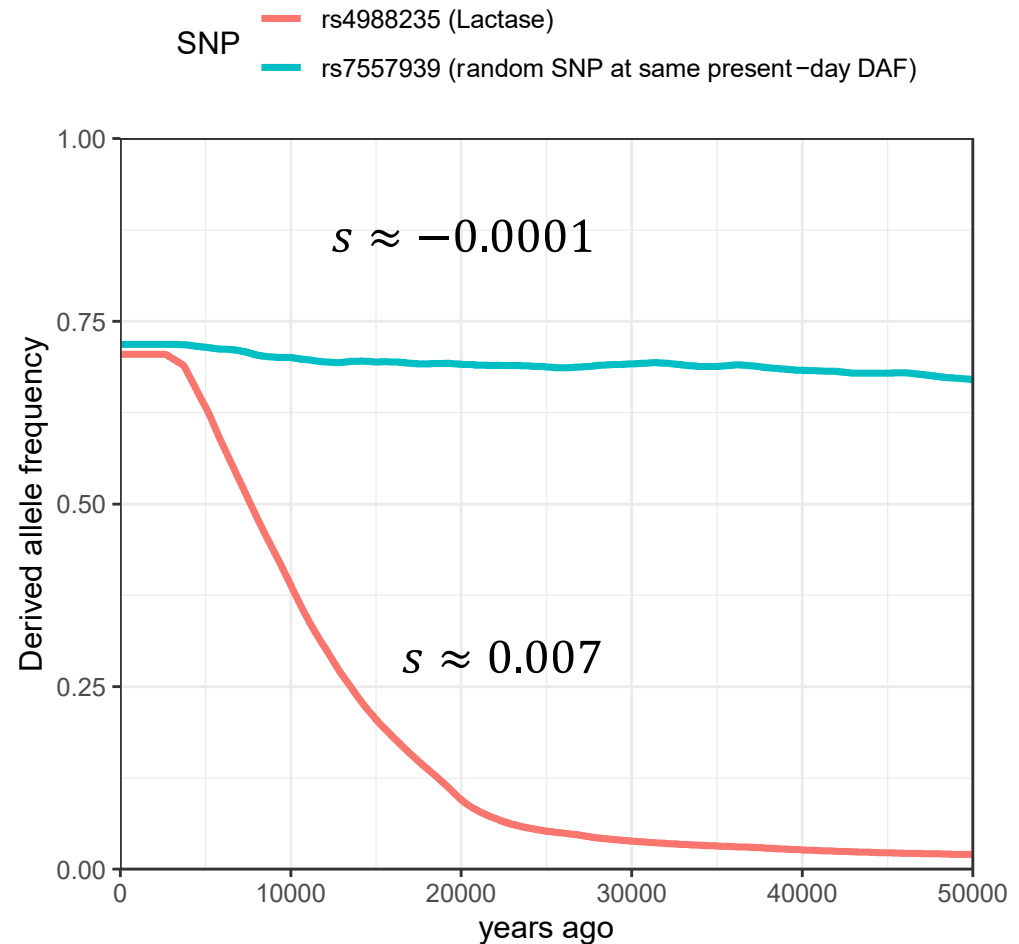
Aaron J. Stern, Peter R. Wilton, Rasmus Nielsen. **PLOS Genetics**, 2019.

Aaron J. Stern, Leo Speidel, Noah A. Zaitlen, Rasmus Nielsen. **AJHG** 2021

Simulations:



1000 Genomes Project British:



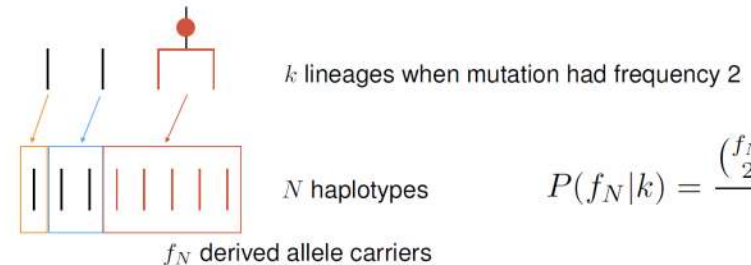
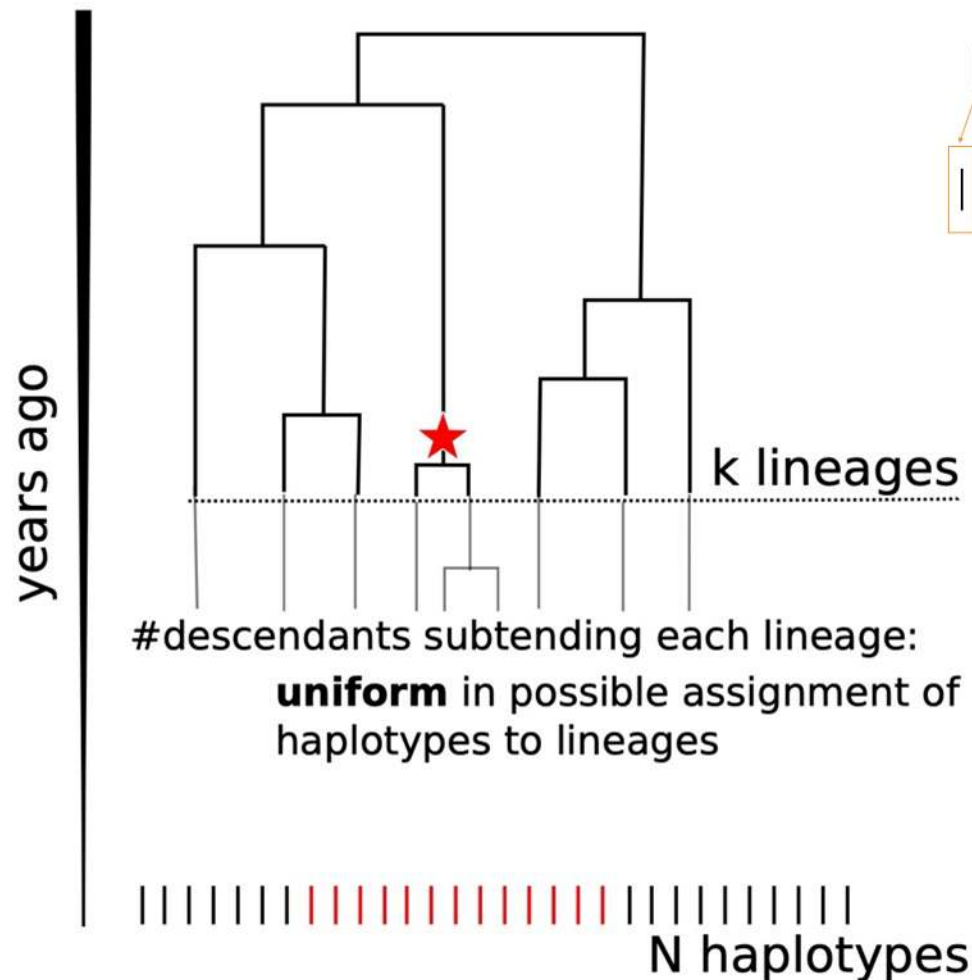
- It is now possible to build genealogical trees for huge datasets, in humans and other species (currently 10,000 individuals or more)
 - Humans (ancient and present)
 - Dogs and wolves
 - Mice
 - Bacteria
 - Atlantic cod, Cichlids
 - Waterhemp, Arabidopsis
- These trees capture information about many processes including
 - Migrations and ancient introgression
 - Mutation rate evolution
 - Trait evolution
 - (and many more things)
- Lots of scope for more methods using inferred genealogies (under development)



....creative approaches to leverage trees to answer biological questions!

P-value for evidence of positive selection

- How much has a mutation out-competed other mutations?
- Robust to population size history



$$P(f_N|k) = \frac{\binom{f_N-1}{2-1} \binom{N-f_N-1}{(k-2)-1}}{\binom{N-1}{k-1}}$$

$$\text{p-value} = \sum_{f=f_N}^{N-k+2} P(f|k)$$