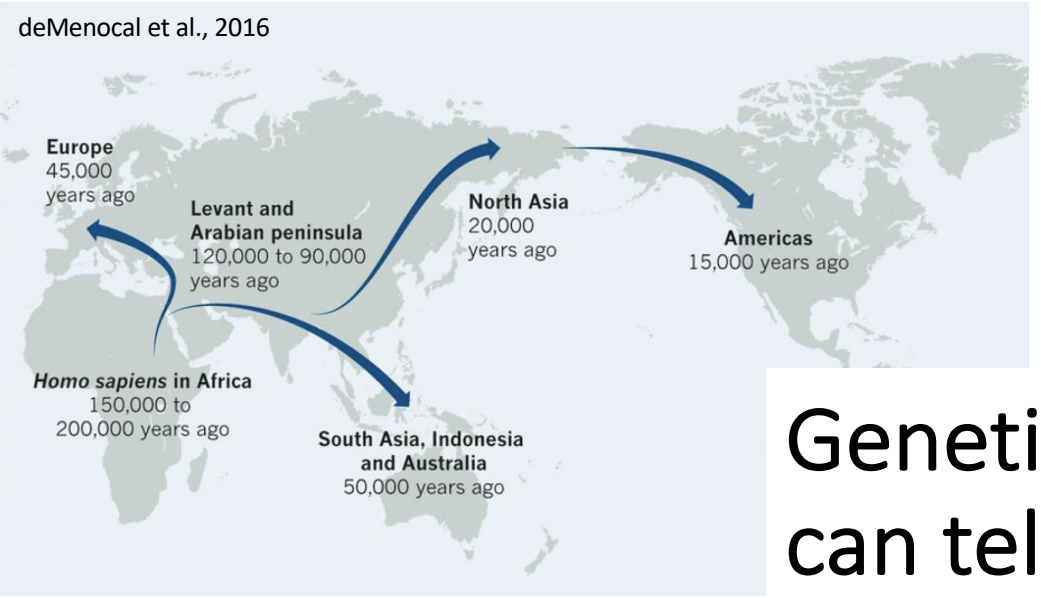
Learning about evolution by building coalescent trees

Simon Myers, 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

deMenocal et al., 2016



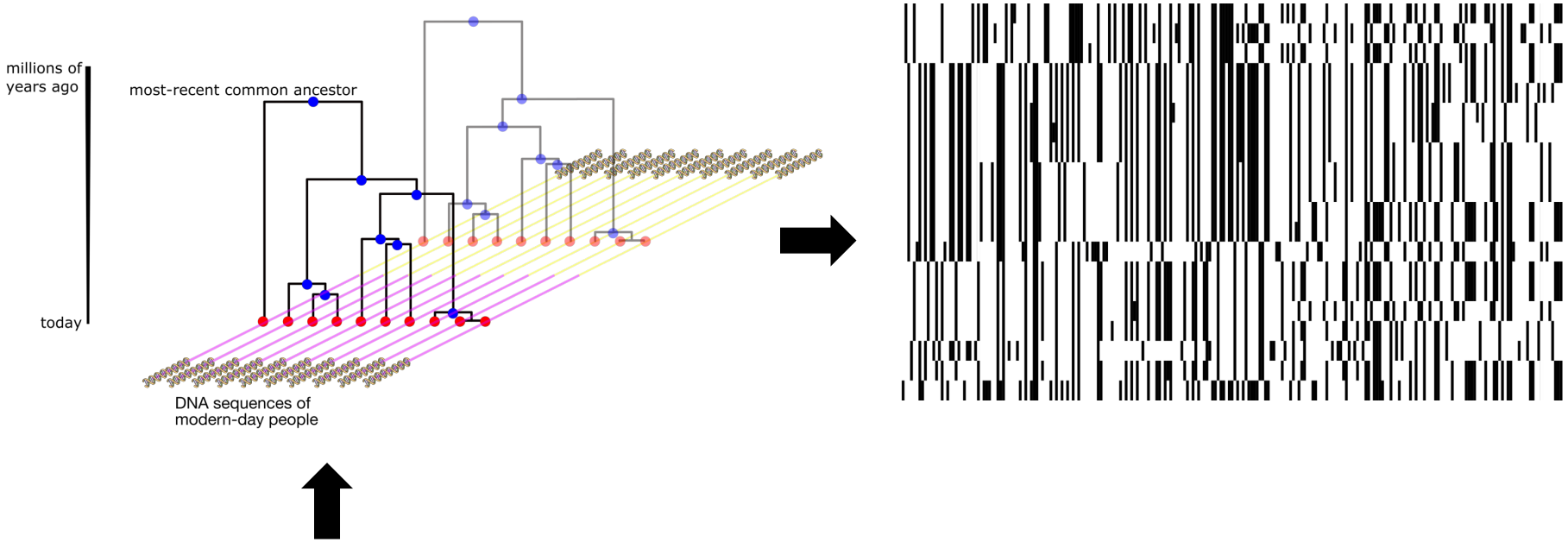
Genetic variation data can tell us about:

- Structure and migrations
- Population bottlenecks
- Admixture

- Mutation
- Recombination
- Selection
- etc

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

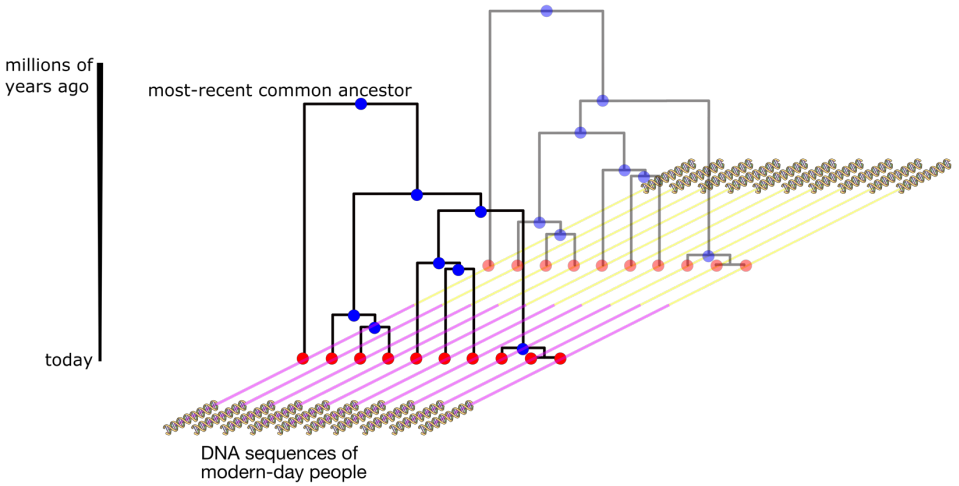
These might themselves evolve through time!

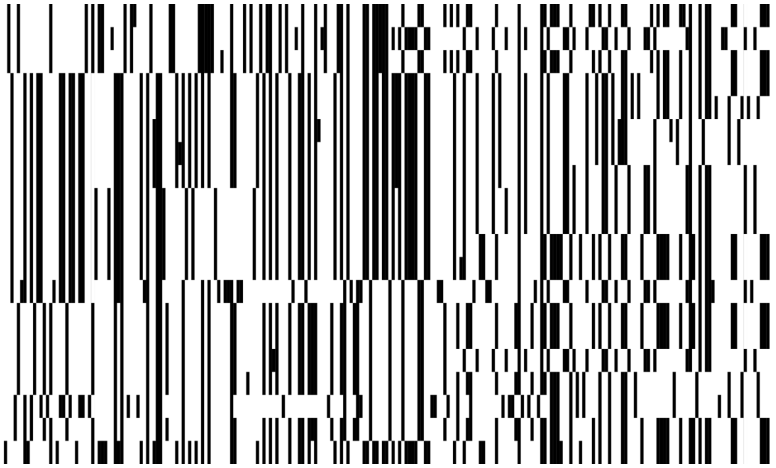


Demographic history
Genetic structure
Mutation, recombination,
etc.

Fundamental forces impact data (only) through underlying genealogies

Many canonical approaches

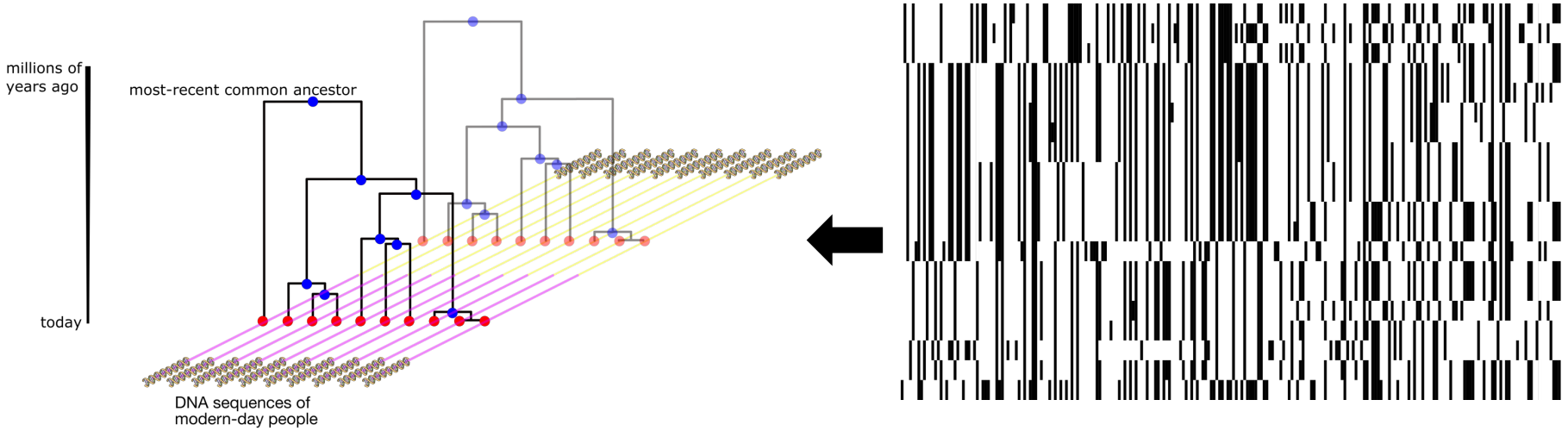




Demographic history
Genetic structure
Mutation, recombination,
etc.

Invent informative statistics, simplify,
"integrate out all possible histories"

Today's approach



Demographic history
Genetic structure
Mutation, recombination,
etc.

In **principle**, trees capture all the information available from the data about these processes

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

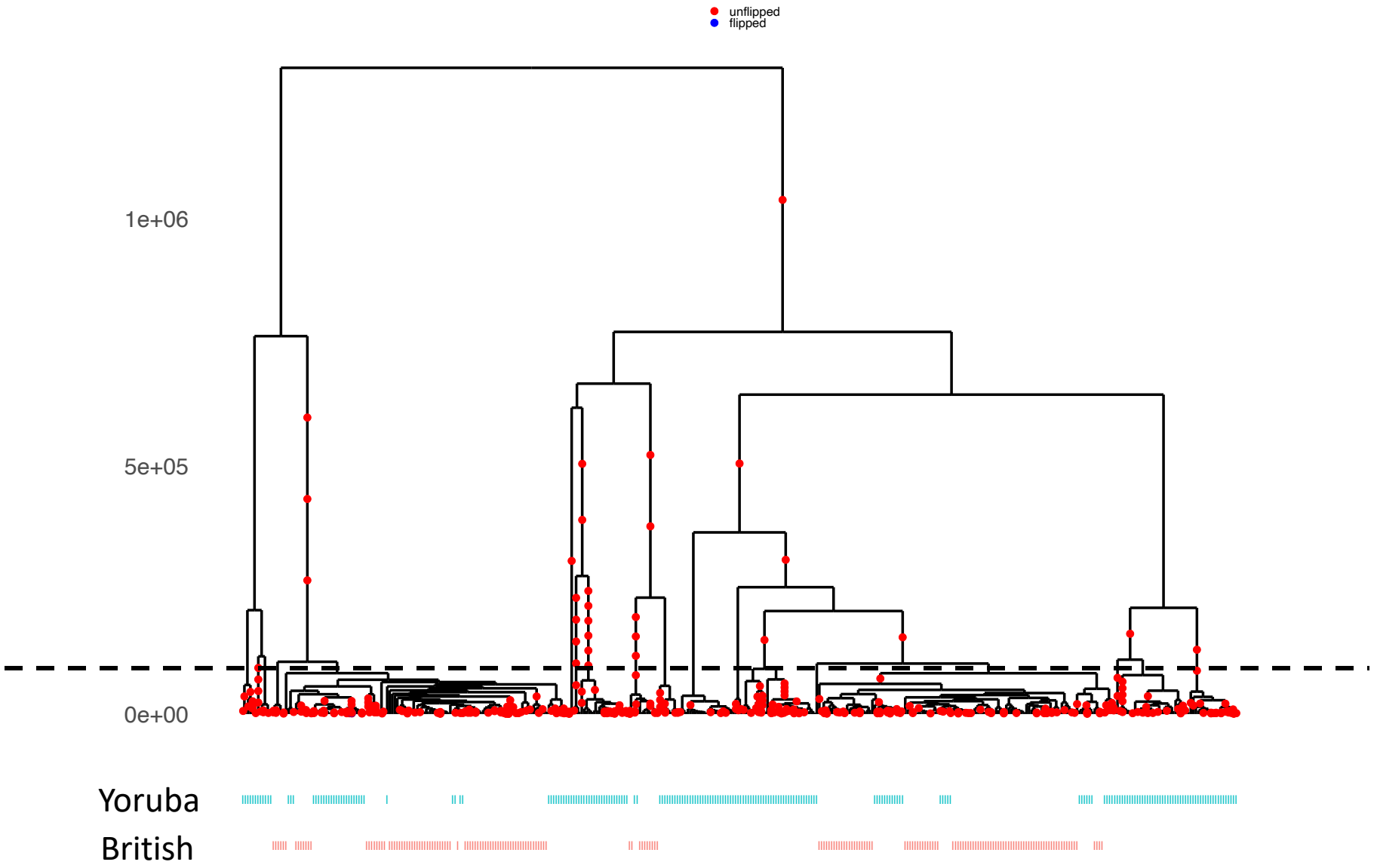
Inferring genealogies

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

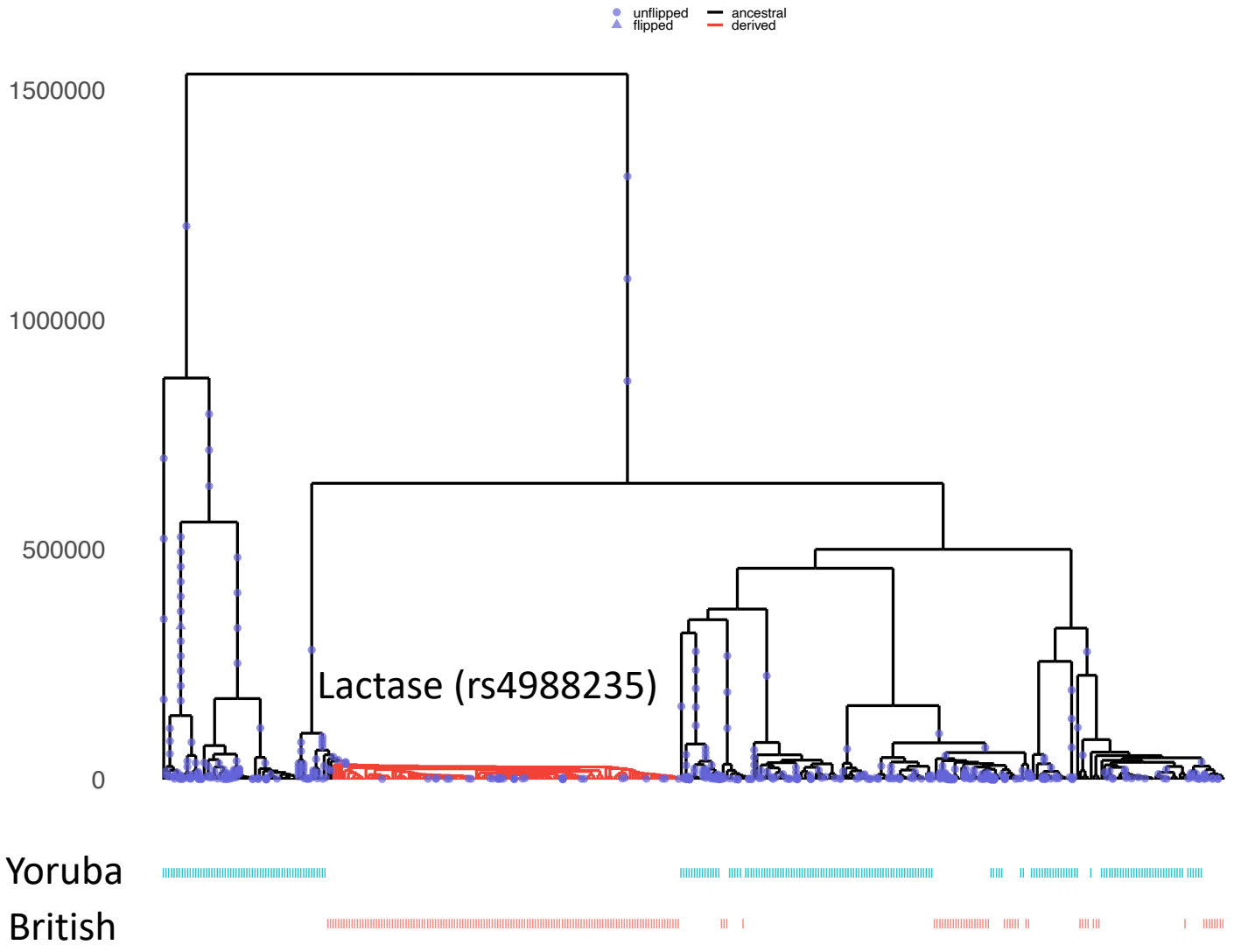
- ARGweaver } Infers Ancestral Recombination Graphs
- Rent+
- Tsinfer } Published in 2019, scale to large sample sizes
- Relate

We will talk about Relate, but principles of tree-based inference applies more generally!

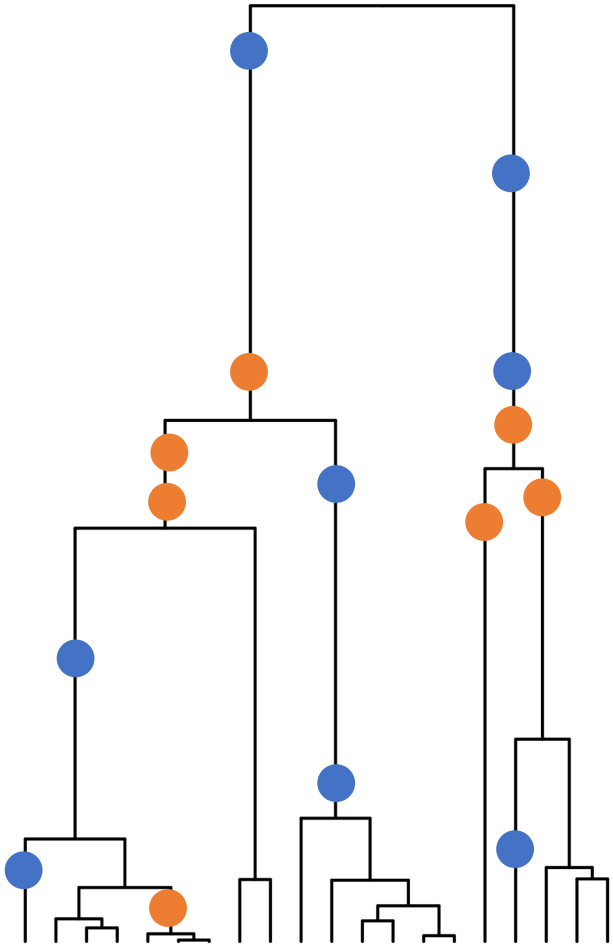
Identifying and dating population splits

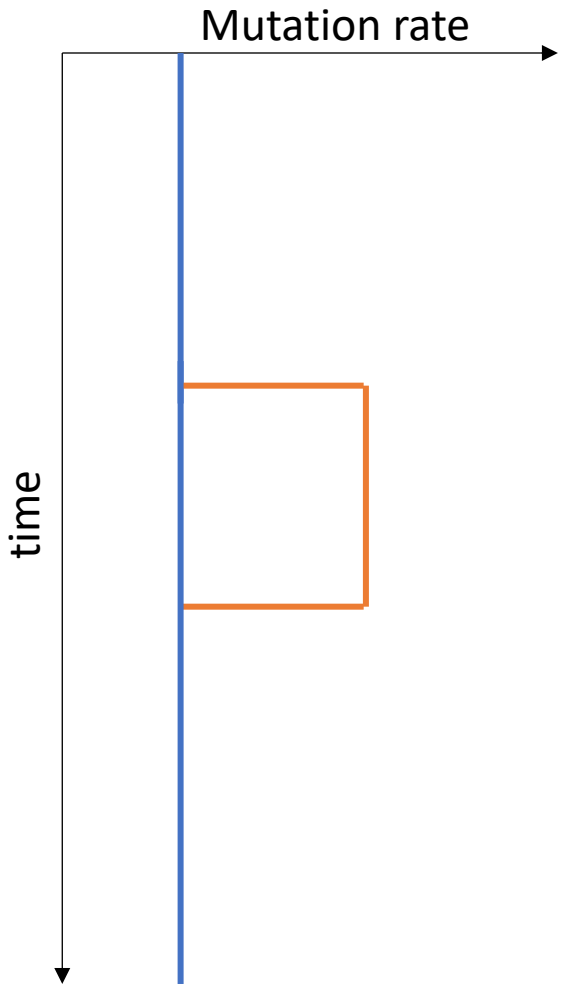


Positive selection: rapidly spreading lineage



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:

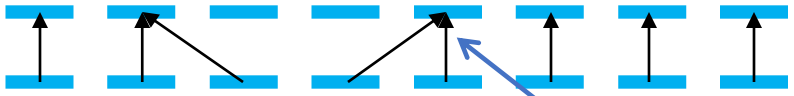
1) Create a method to build trees under a coalescent model, with varying population size, and allowing for recombination
2) Construct statistics to capture information from trees and either
   (i) Interpret parameters in the coalescent, e.g. coalescence rates
   (ii) Reject a null model, e.g. testing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

Revision of coalescent

The **Wright-Fisher model** is able to approximate more realistic models of populations

Each member of the current generation randomly chooses one of *M* parents 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

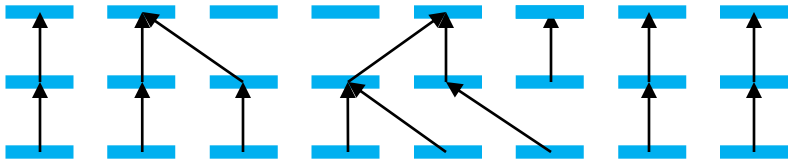
If haplotypes **share** a parent back in time, this is called a **coalescence event**

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 $1/M$



Each haplotype chooses parent in previous generation totally at random

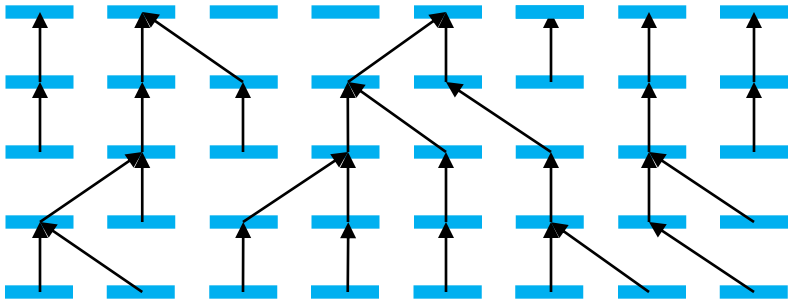
If haplotypes **share** a parent back in time, this is called a **coalescence event**

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 $1/M$



Each haplotype chooses parent in previous generation totally at random

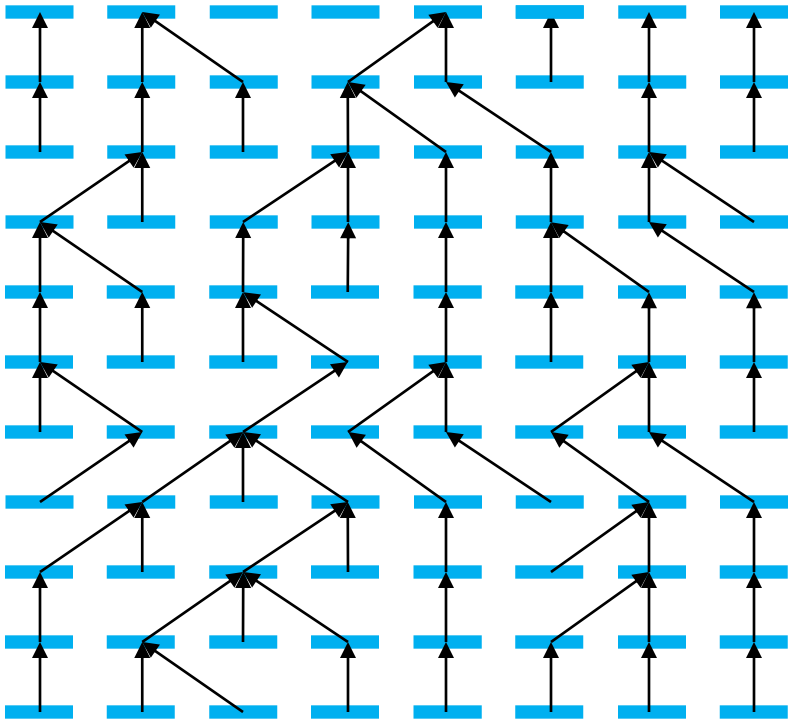
If haplotypes **share** a parent back in time, this is called a **coalescence event**

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 $1/M$



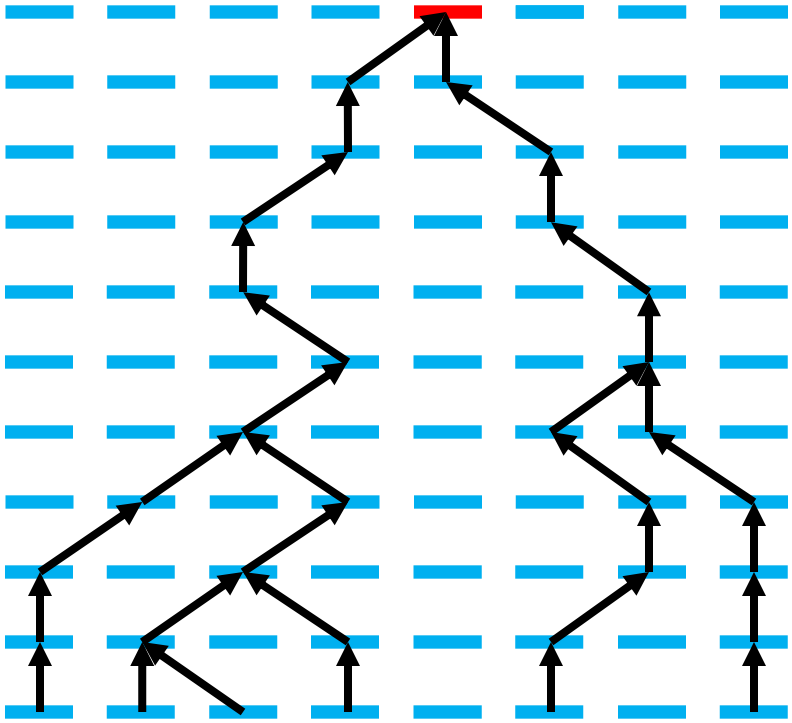
Each haplotype chooses parent in previous generation totally at random

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/*M* coalescences per generation, so 1 coalescence per *M* generations



Sample of size *n*=6

M ~10-50,000 for all human

populations, highest in Africa

So a typical pair of human chromosomes share an ancestor on average around 2x20,000x28=1 million years ago

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

25,000,000 for *E.coli*

2,000,000 for fruit fly

D. Melanogaster

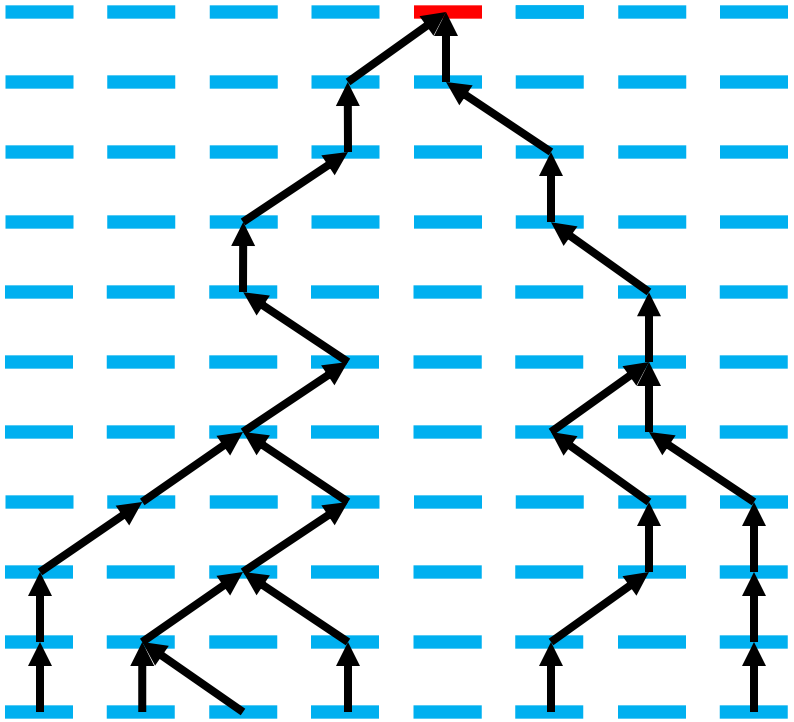
<100 for Salamanders
(Funk et al. 1999)

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

Any pair of lineages coalesces at rate $1/M$

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 $\binom{j}{2}/M$

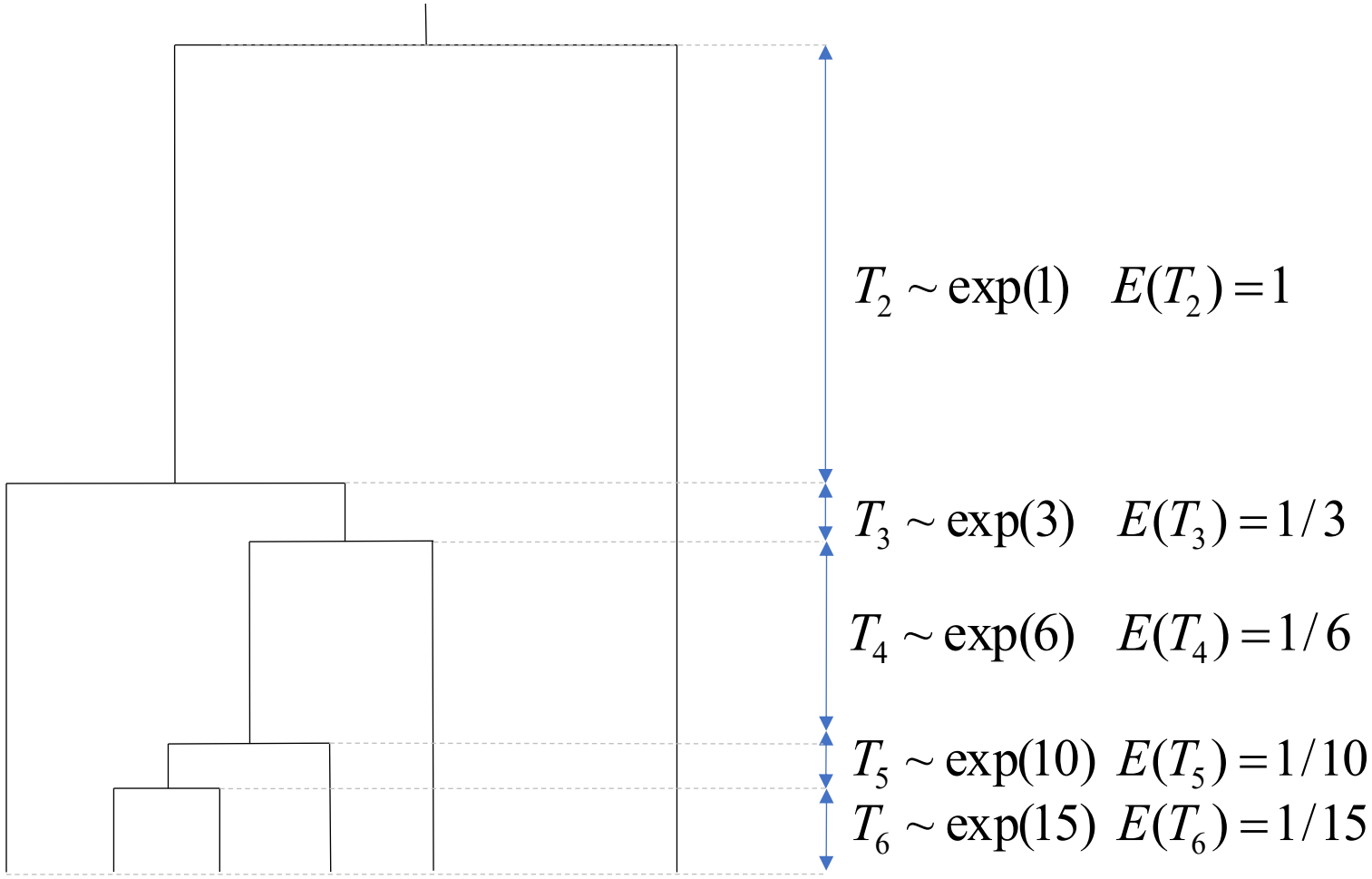
[this leads to an exponential distribution of time until coalescence, with rate $\binom{j}{2}/M$]



j=2, 1 pair

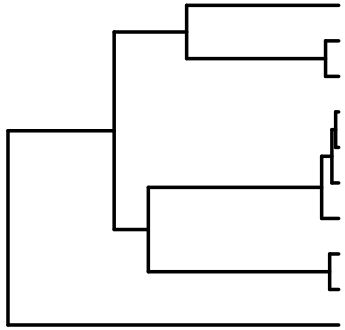
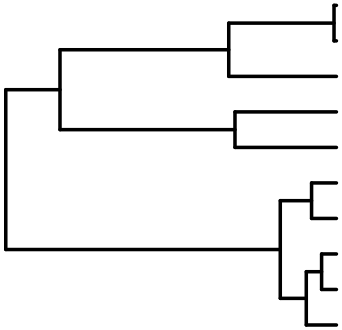
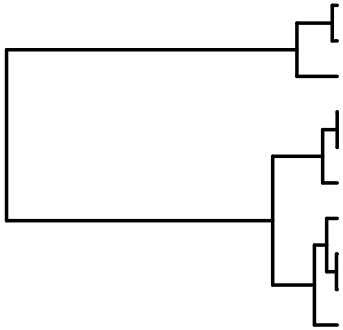
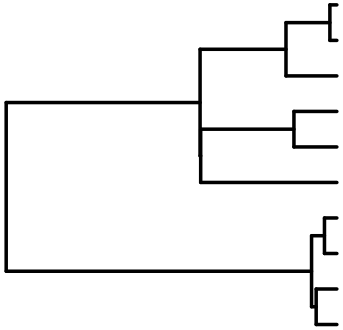
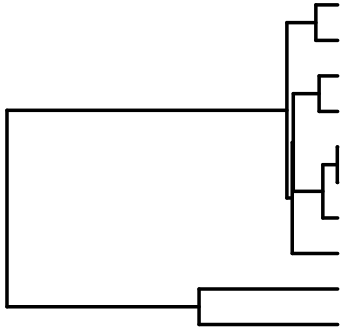
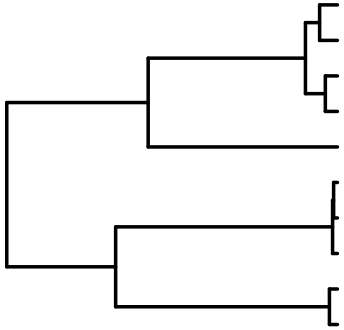
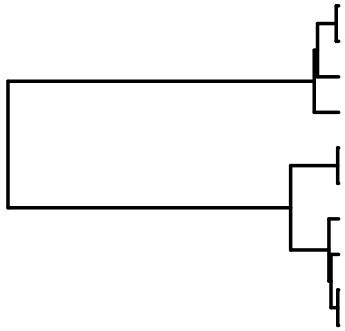
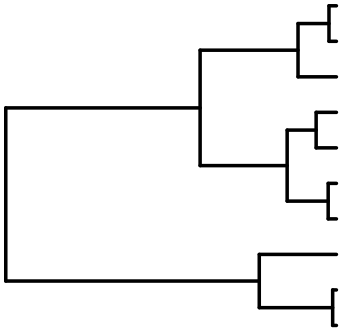
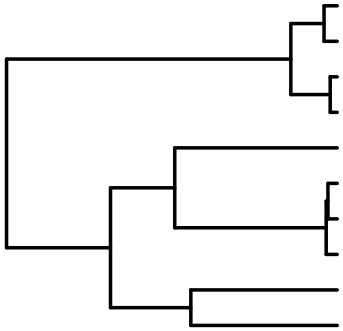
j=4, 6 pairs

Sample of size n=6



n=6

(after scaling time by *M* generations)

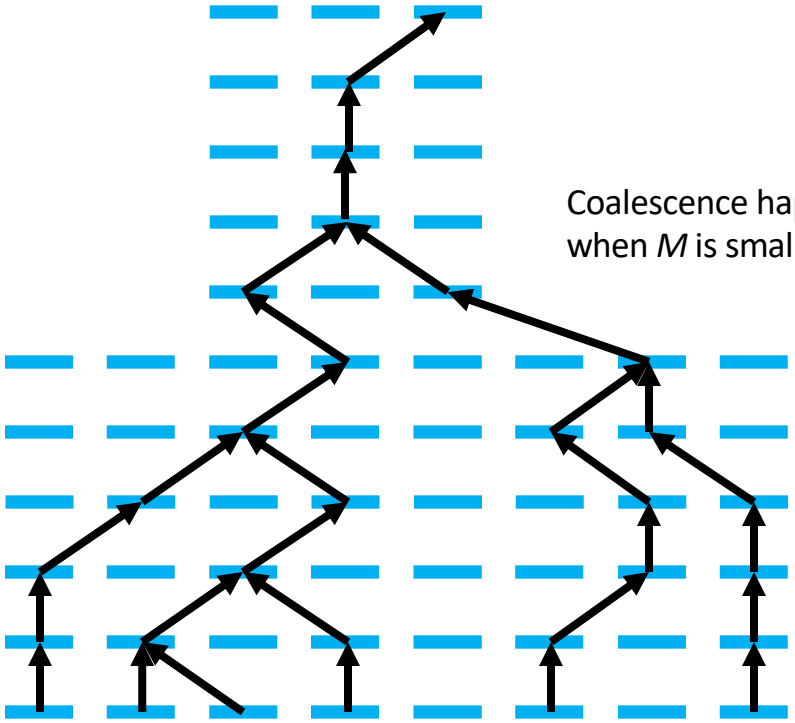










Varying population size

If M changes, so does the chance of coalescing

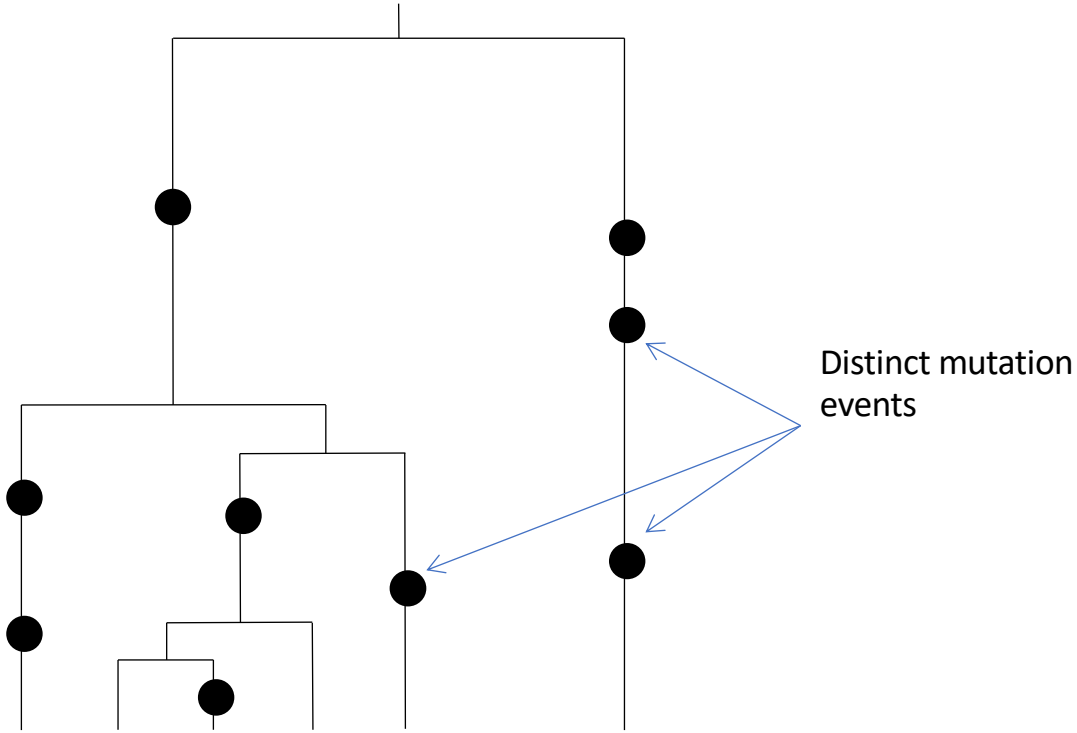
While there are *j* lineages, the rate at which a coalescence happens is just $\binom{j}{2}/M(t)$ a time *t* ago

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



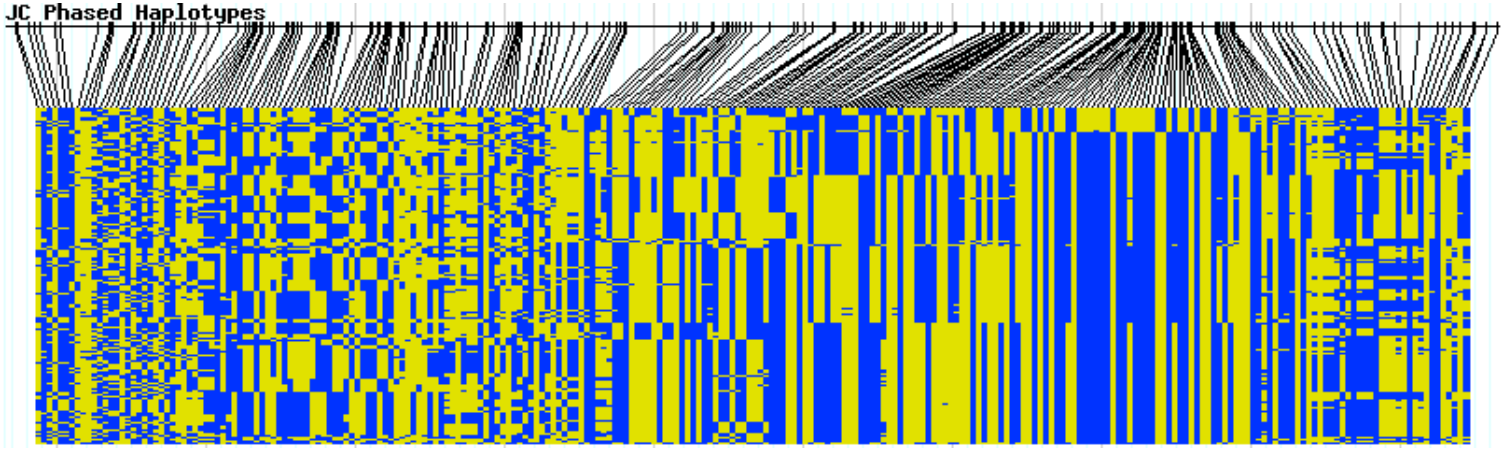
Sample of size *n*=6

Adding mutation to the mix

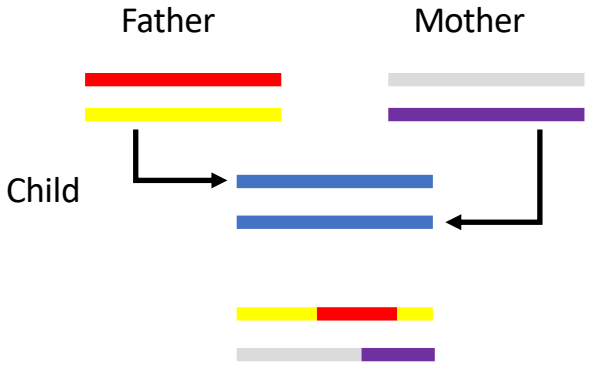


- Mutations are dropped randomly on the edges of the tree (e.g. in many simulation software packages)
- They are seen in descendants of this edge, so this totally specifies diversity patterns
- We will talk about some theory results about spread of mutations in the coalescent later

What about recombination?



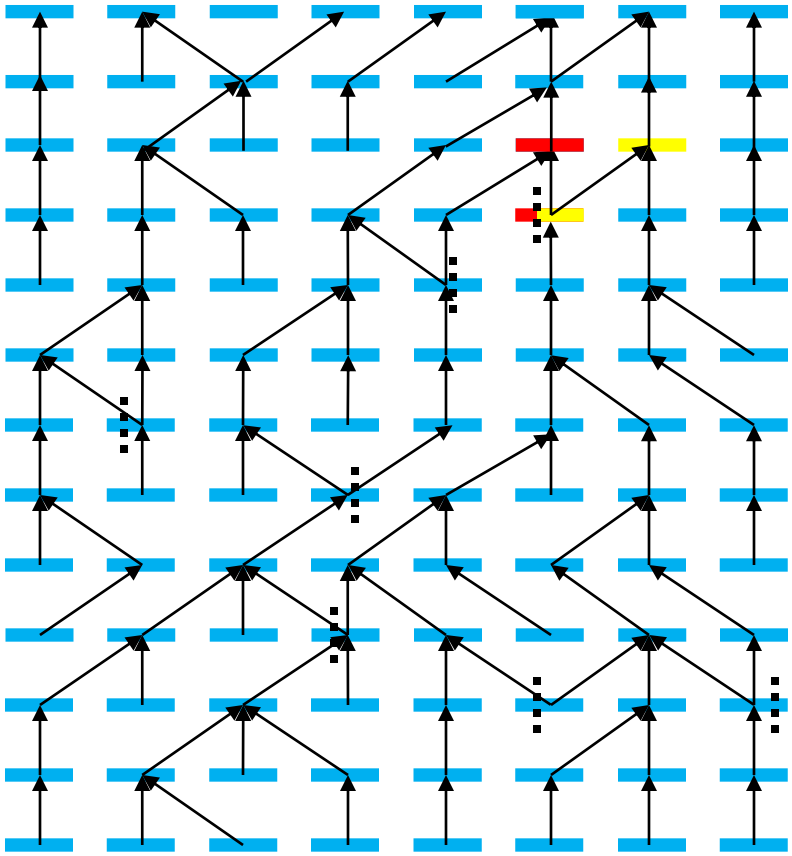
- Unlike mutation, recombination events actually change the trees



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

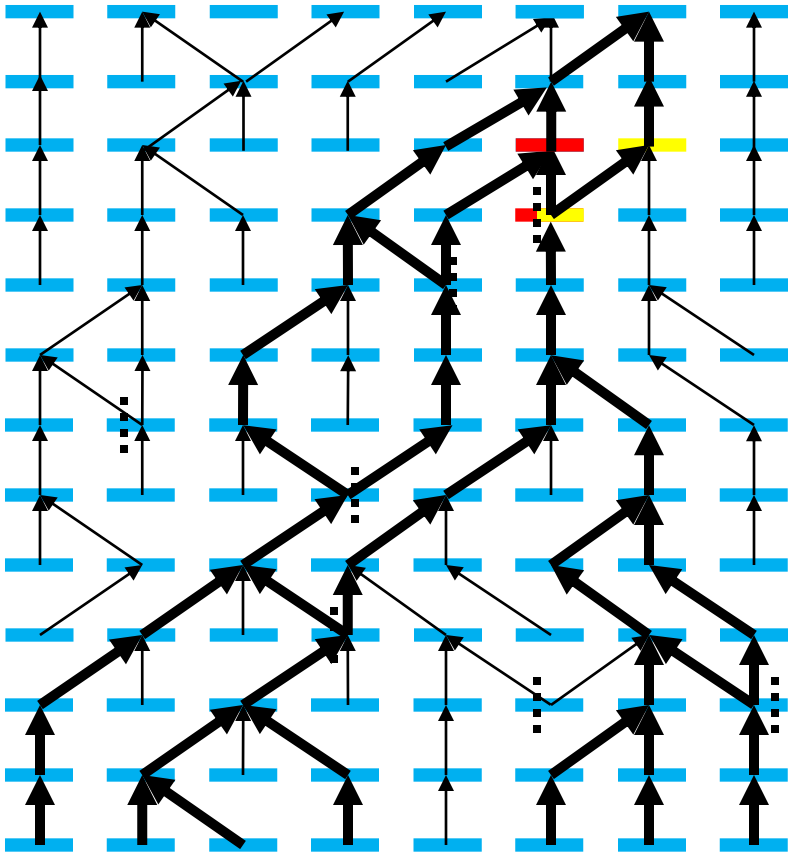
Principles of adding recombination to the coalescent

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

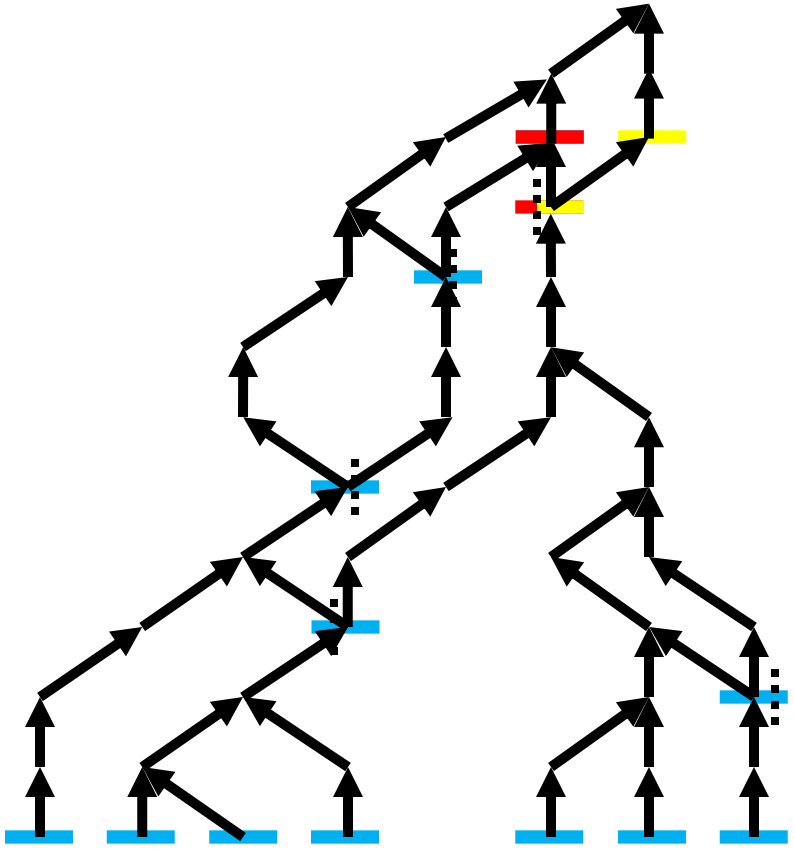


Principles of adding recombination to the coalescent

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

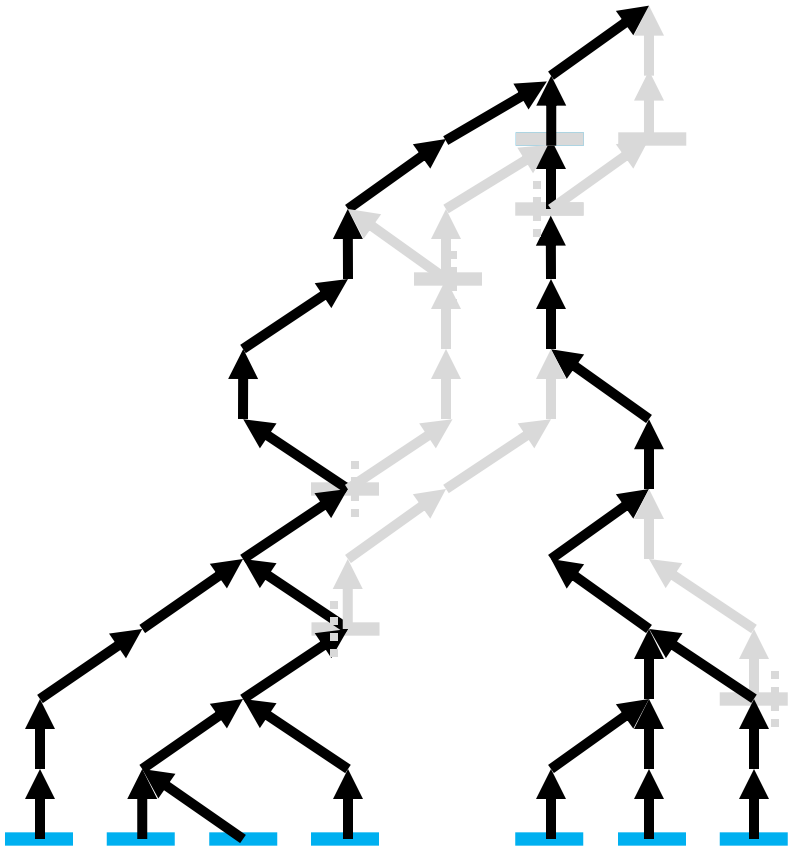


The ancestral recombination graph



The tree at the left-most position in the region

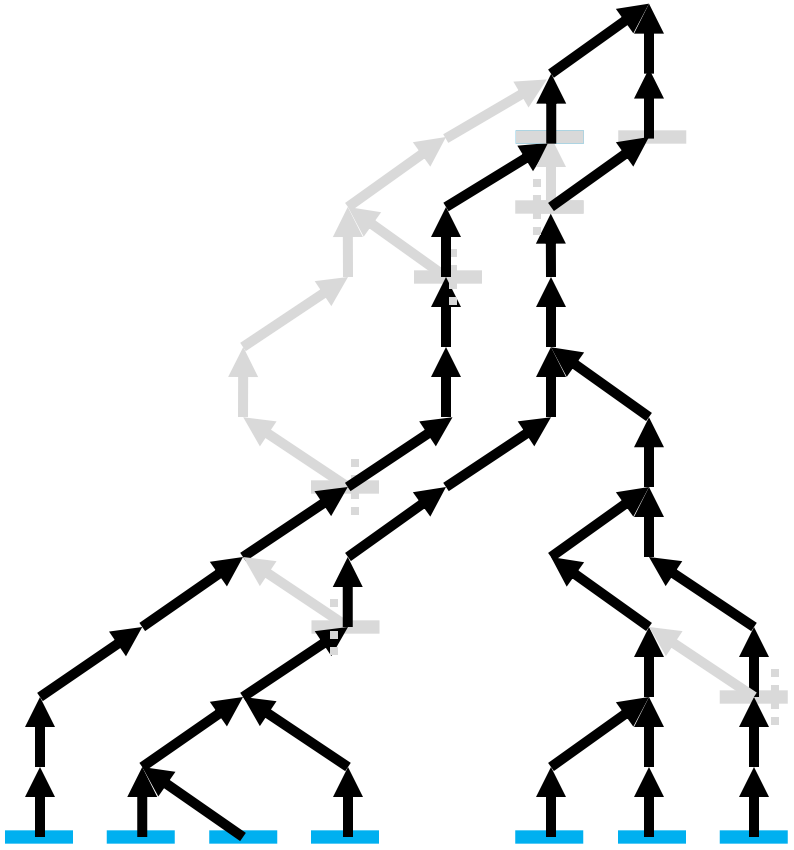
A small piece of DNA is not impacted by recombination, so the coalescent model still applies



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



Relate

L. Speidel, M. Forest, S. Shi, S. Myers. Nature Genetics 2019

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

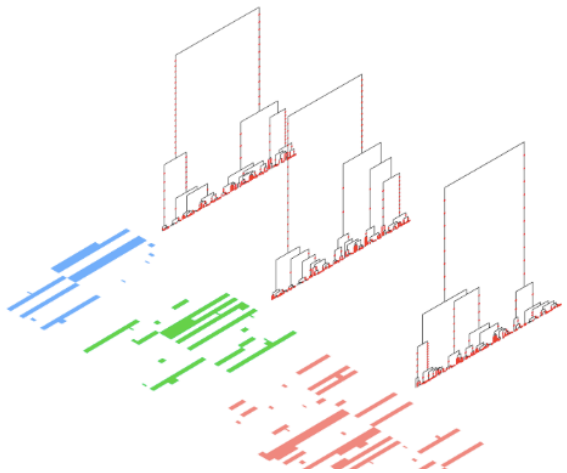
Relate | Home | Getting Started | Input data | Add-on modules | Parallelise Relate

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

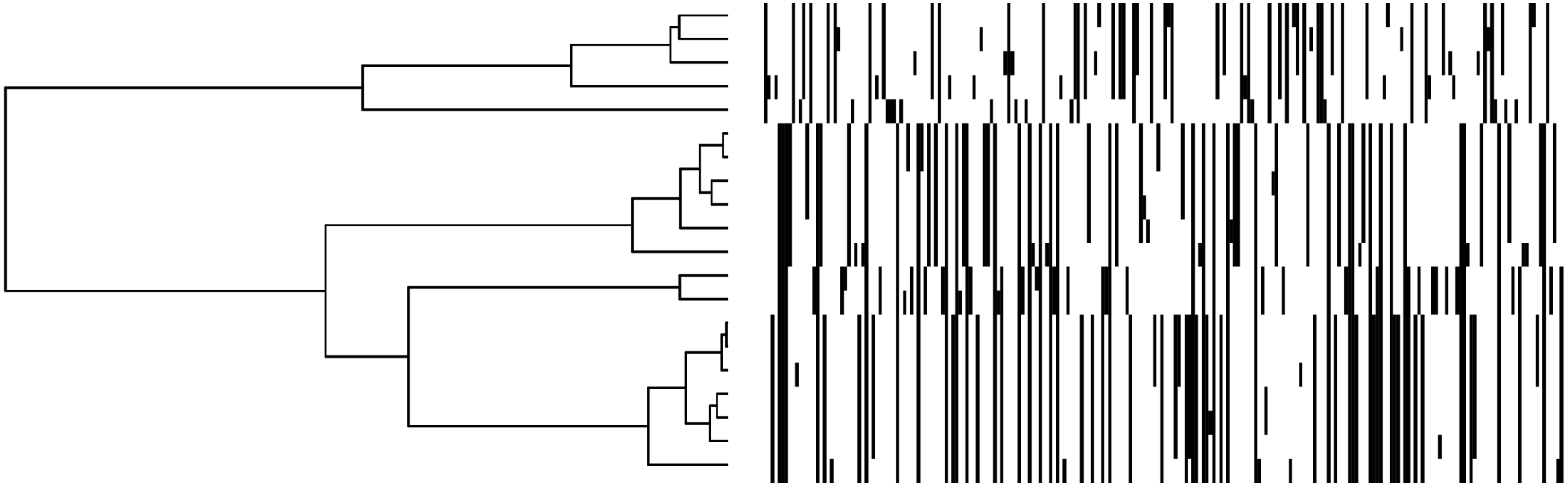
Mac OSX - v1.0.16

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.

We document changes to previous versions in a change-log.

Data, and the underlying tree structure



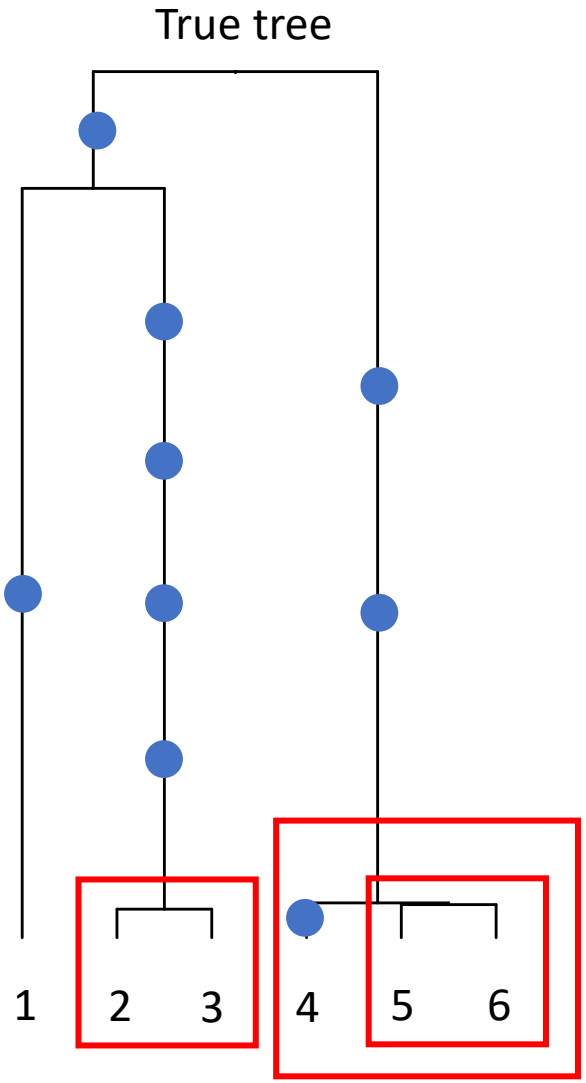
- **Every mutation shows the existence of a branch**
- Mutations are “ordered by inclusion”
- No two branches (mutations) ever show only partial overlap

A basic tree-builder: UPGMA

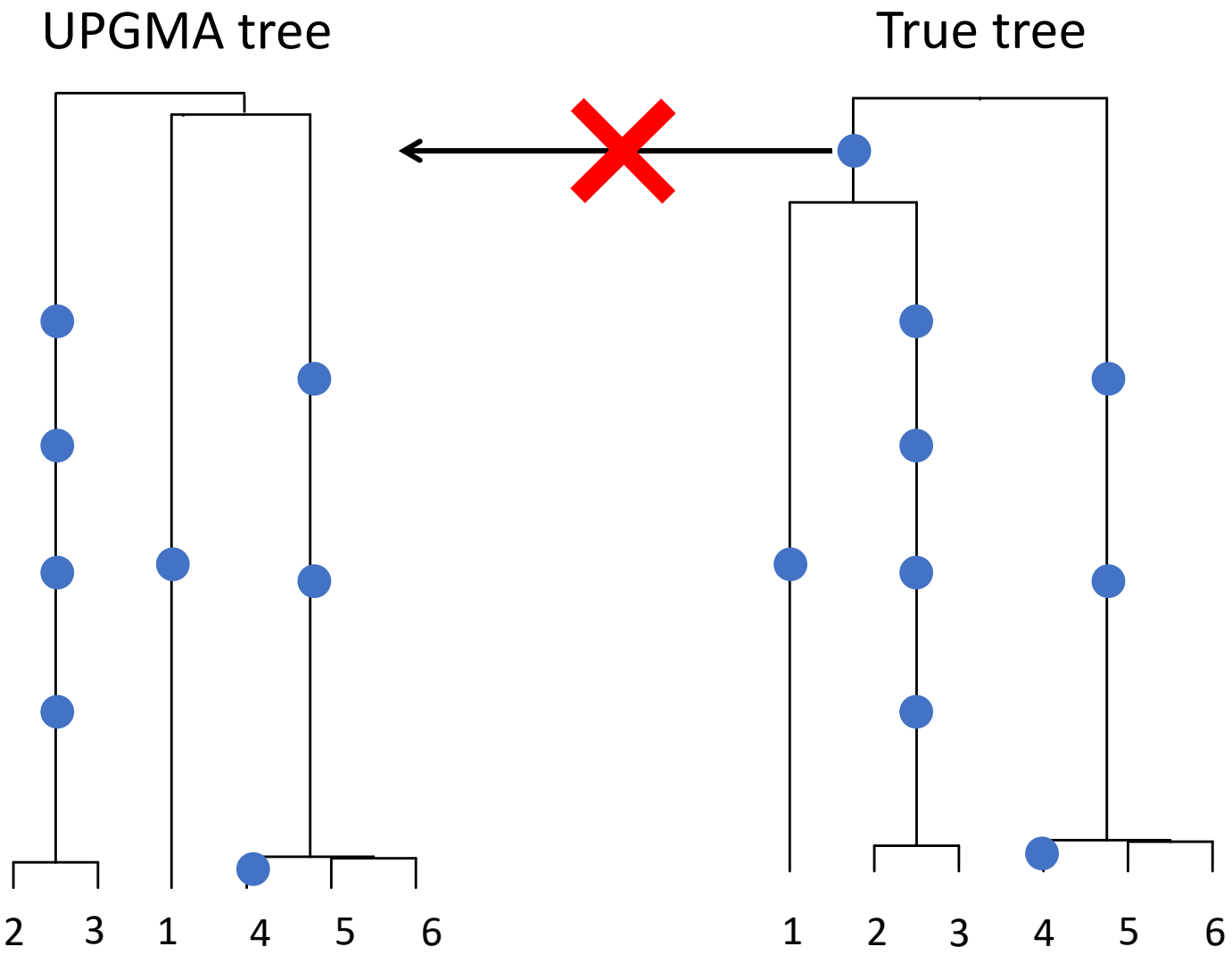
- UPGMA coalesces lineages with smallest number of pairwise differences
- (2,3) and (5,6) are coalesced first
- (5,6) and 4 are coalesced

- 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



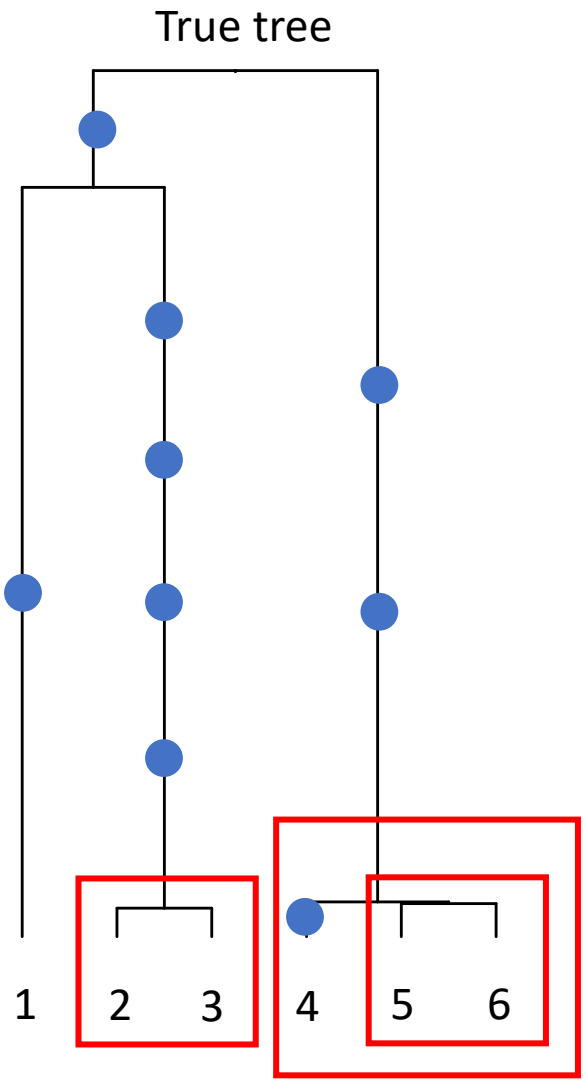
The UPGMA tree cannot be correct, because it does not include any branch whose descendants are sequences 1,2,3. How can we fix this?

Counting **derived** mutations to build the correct tree

relative to

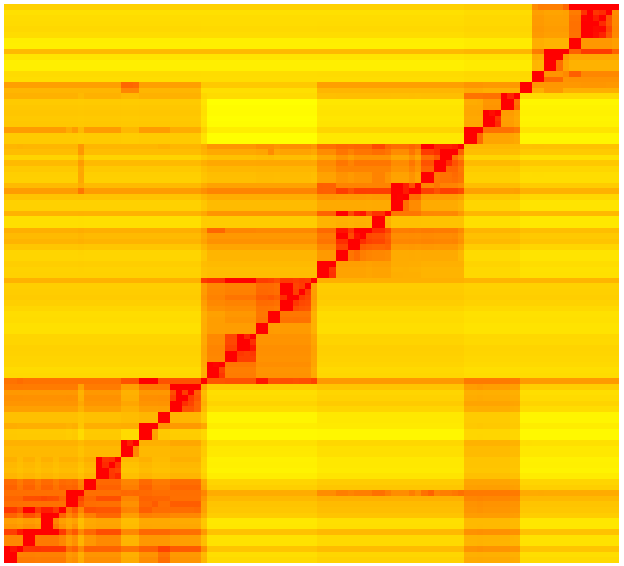
Mutation carrier	1	(2,3)	(4,5,6)
1	0	1	2
(2,3)	4	0	5
(4,5,6)	2	2	0

Avoids combining information of two branches!



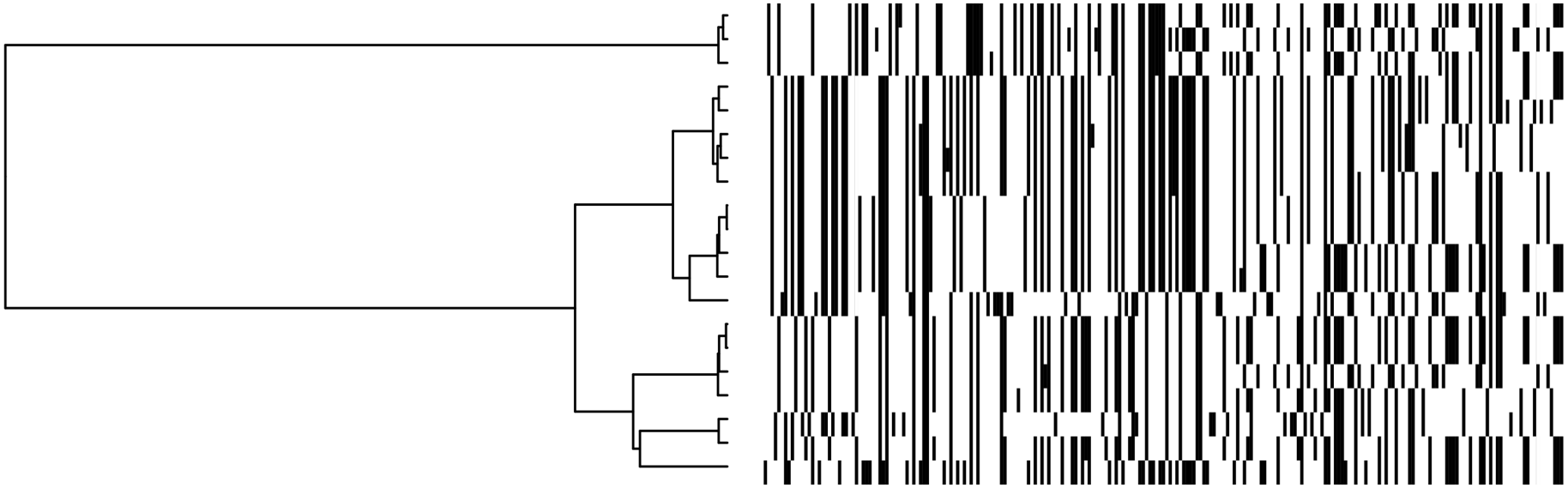
How does Relate work?

- In the **no-recombination** case, it first counts numbers of derived mutations for each pair and builds a tree structure/topology (times are deferred for now)



- Performs coalescences between mutually most similar lineages
- Guaranteed to produce a tree matching the data!

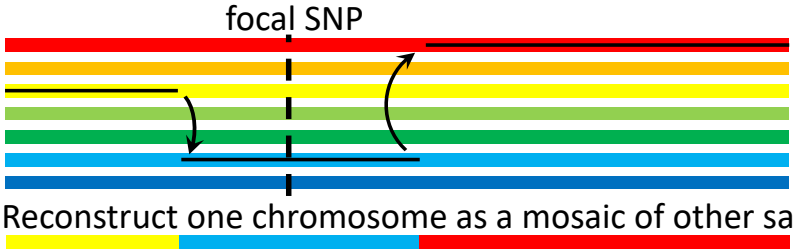
Accounting for recombination



- 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

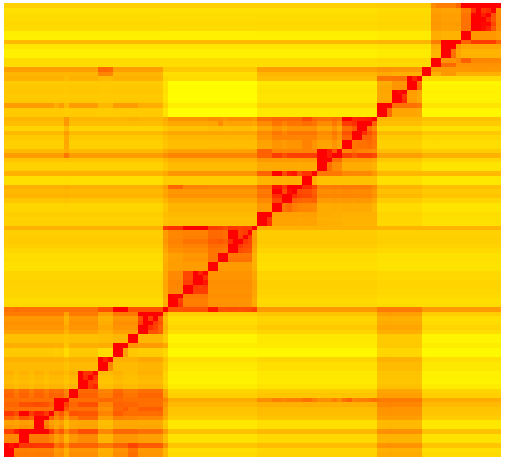
Summary of Relate pipeline

Hidden Markov model (HMM)

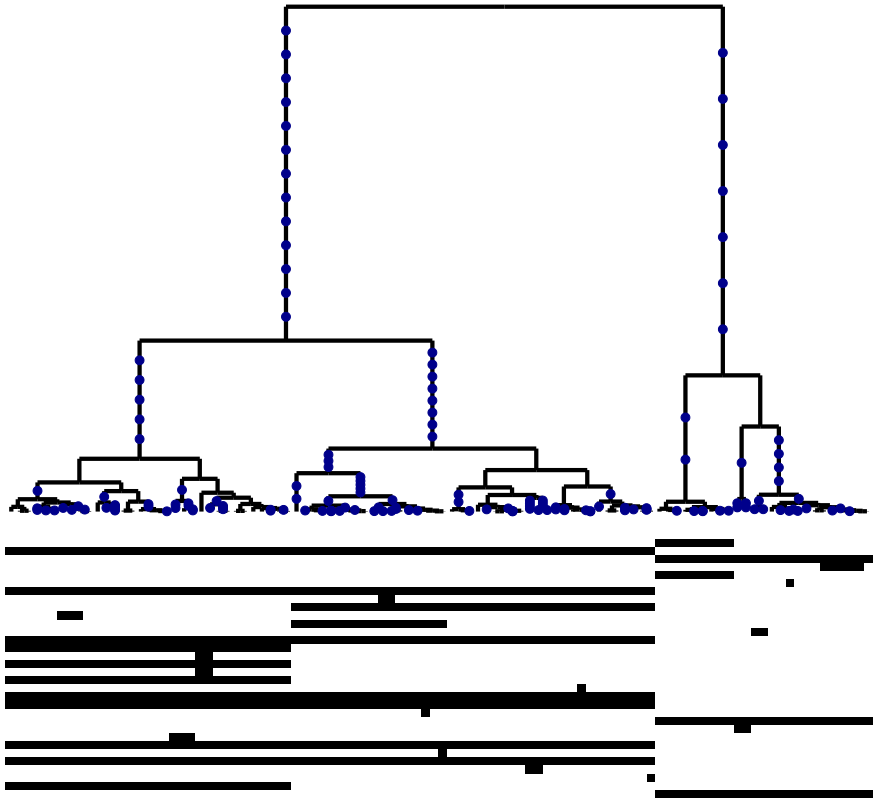


Reconstruct one chromosome as a mosaic of other samples

Distance matrix for focal SNP



Hierarchical clustering
&
MCMC for branch lengths

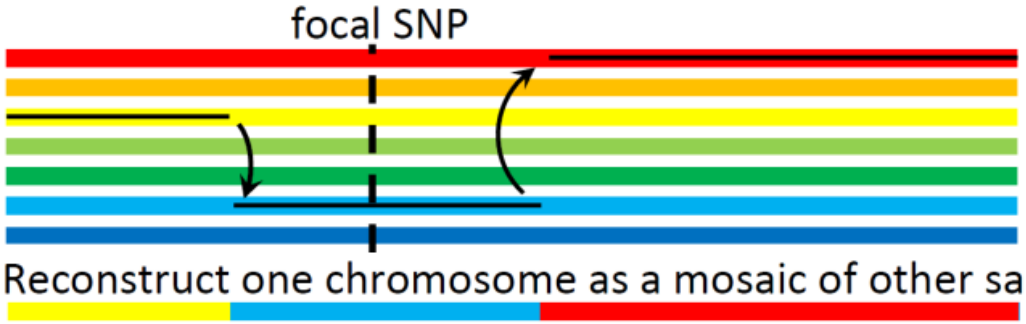


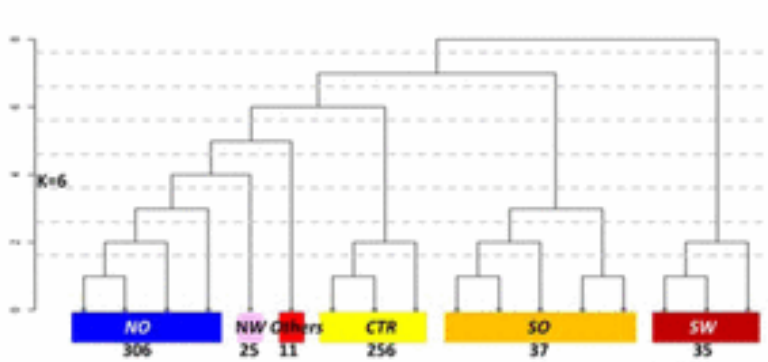
haplotype data sorted using constructed tree

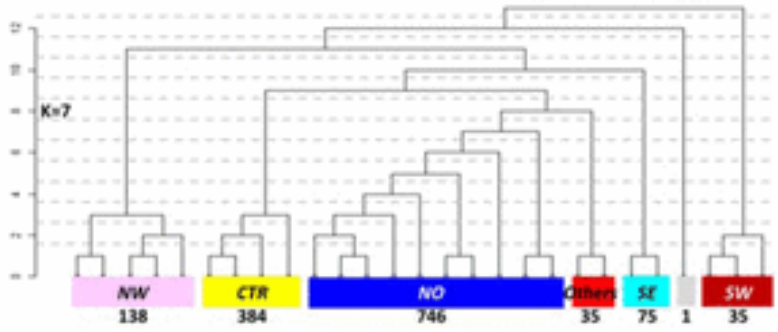
Model used by Relate is very similar to that used in fineSTRUCTURE (Monday)

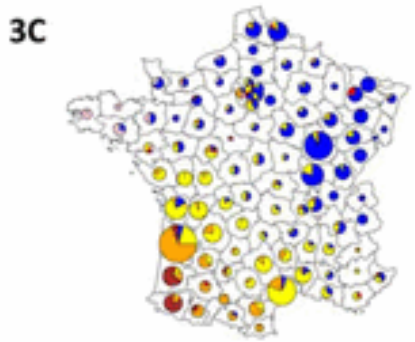
Hidden Markov model (HMM)

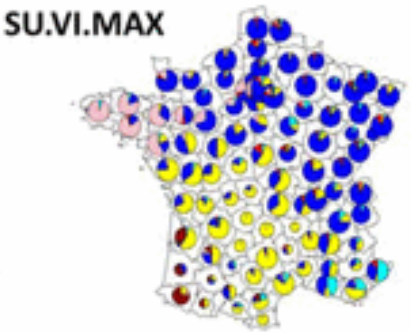
Li and Stephens, Genetics, 2003; Lawson et al., PLOS Genetics, 2012









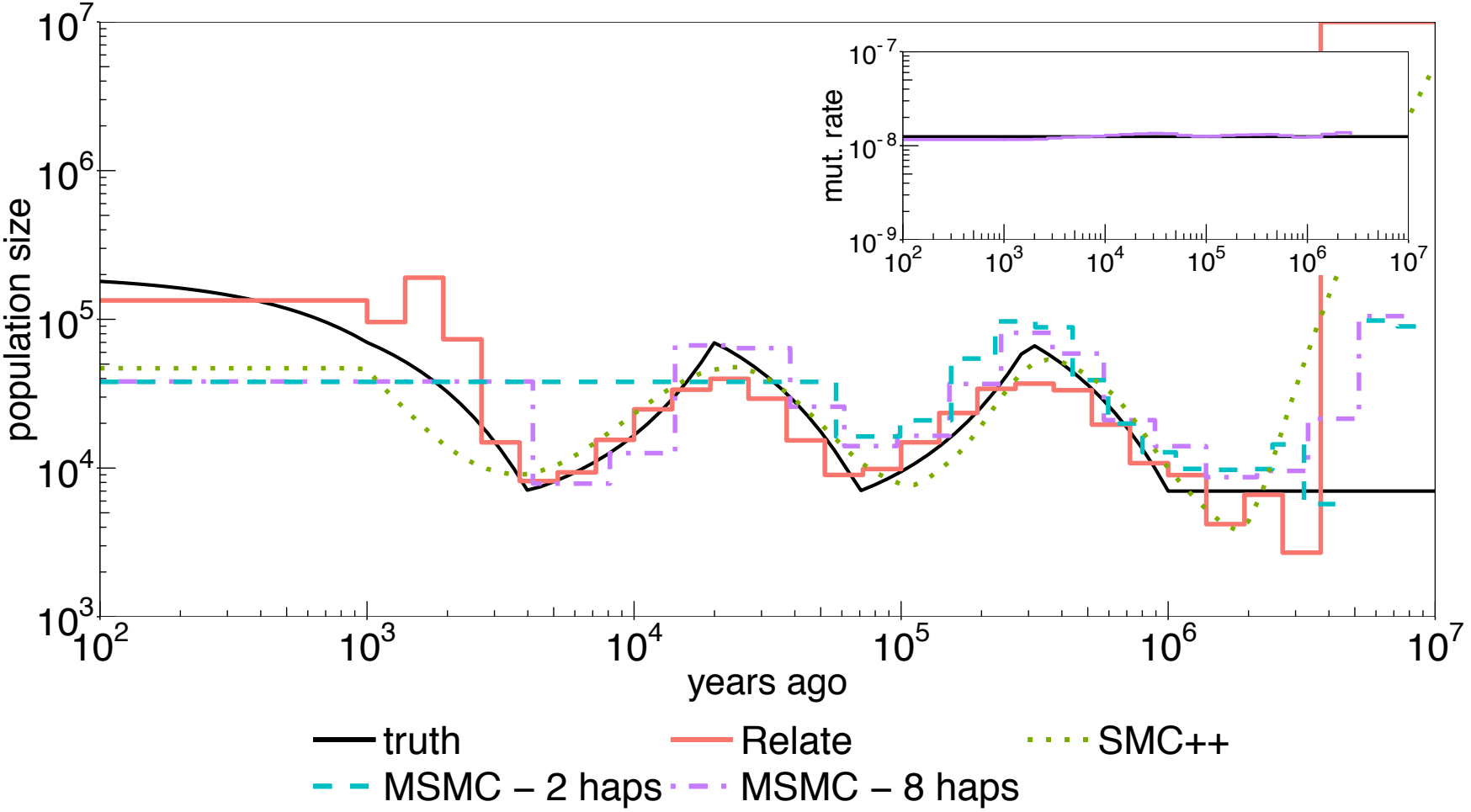


fineSTRUCTURE analysis of France:
Saint-Pierre et al. 2019

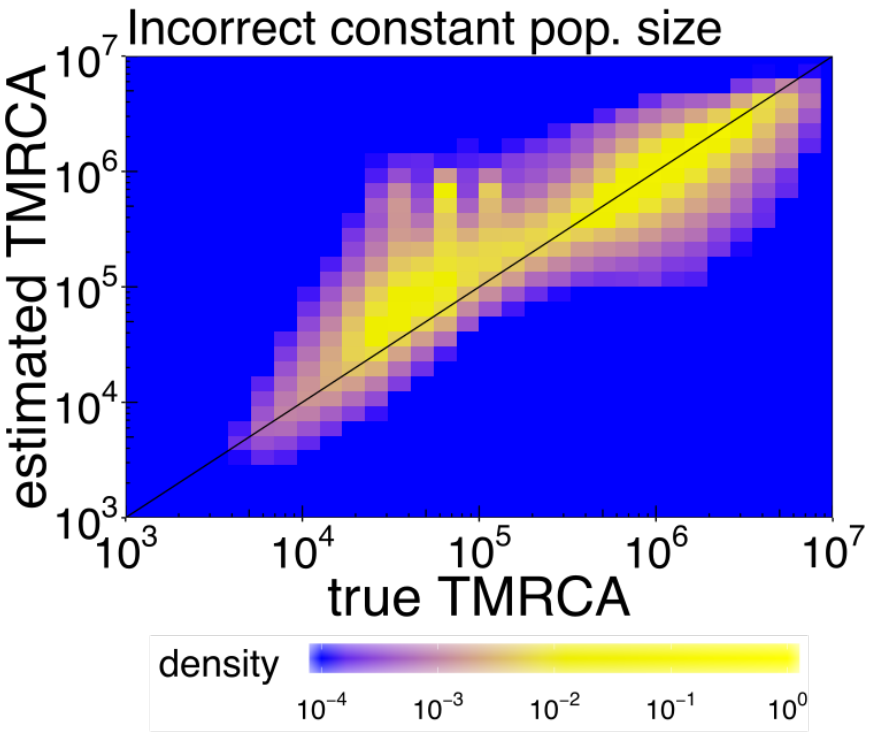
Variable population sizes

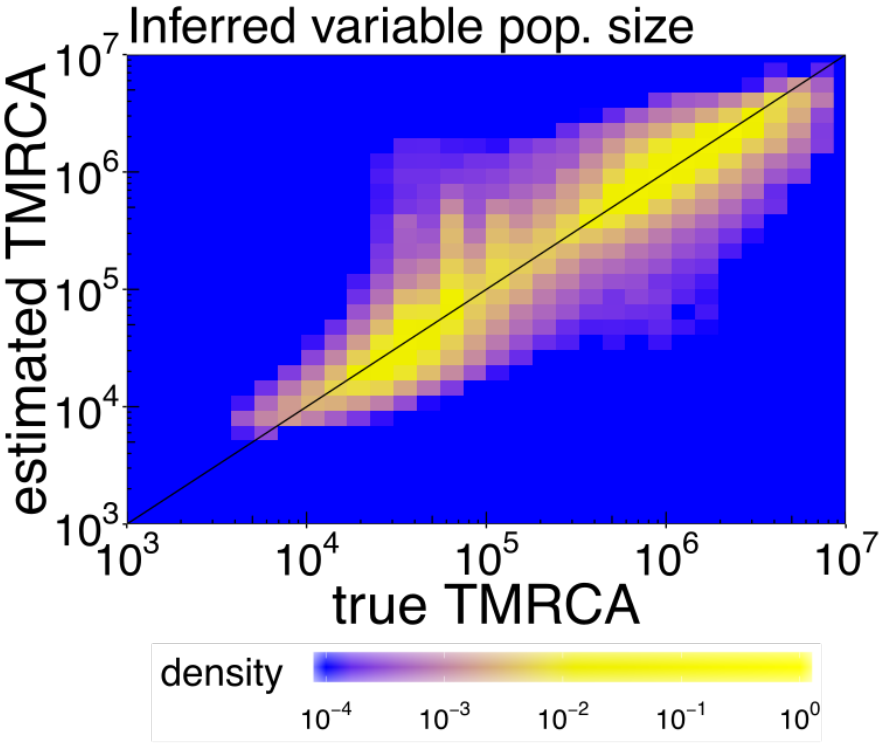
- Population size is varying; estimated from the trees, within Relate
- Recall: while there are *j* lineages, the rate at which a coalescence happens is $\binom{j}{2}/M(t)$ a time *t* ago
- Demography is shared genome-wide, so we average across trees
- So within a time interval, scaled fraction of trees where coalescence occurs is inversely proportional to M(t)
- Iterate and rebuild trees…

Simulation:
population size changes through time



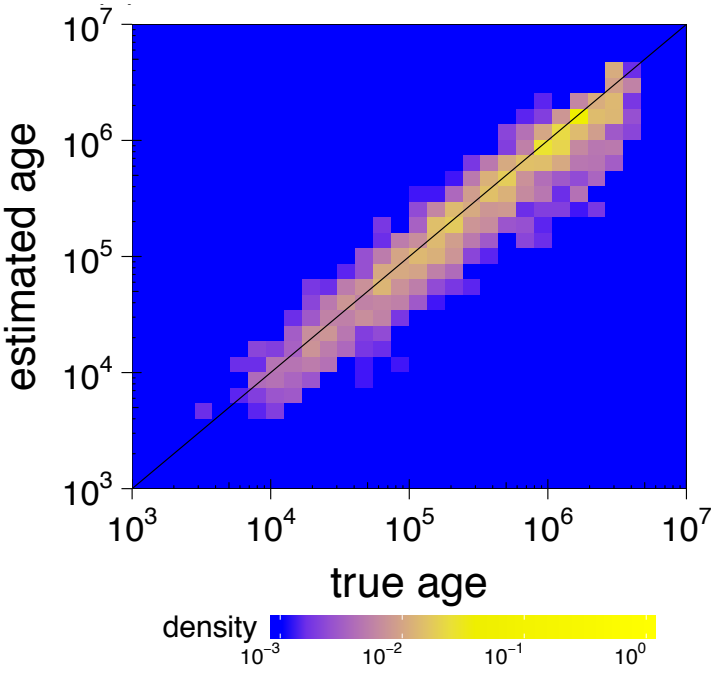
Simulated data with variable population size (European-like demographic history)

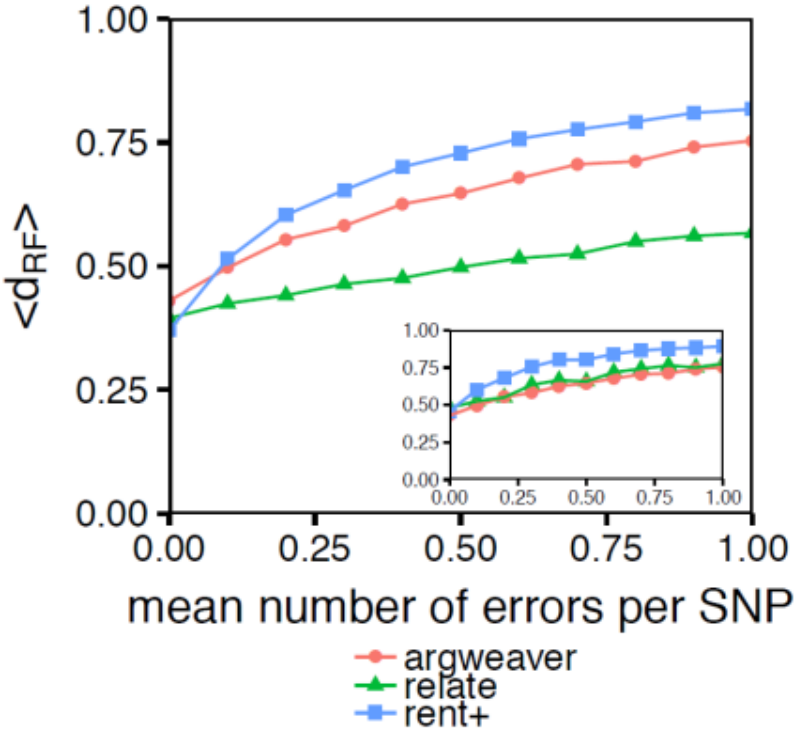




Speed and accuracy of Relate

- About 14,000 times faster than previous method, ARGWEAVER (1 min. vs. 200 hours)
- Builds “correct” tree if no recombination
- Accurate, robust to data errors





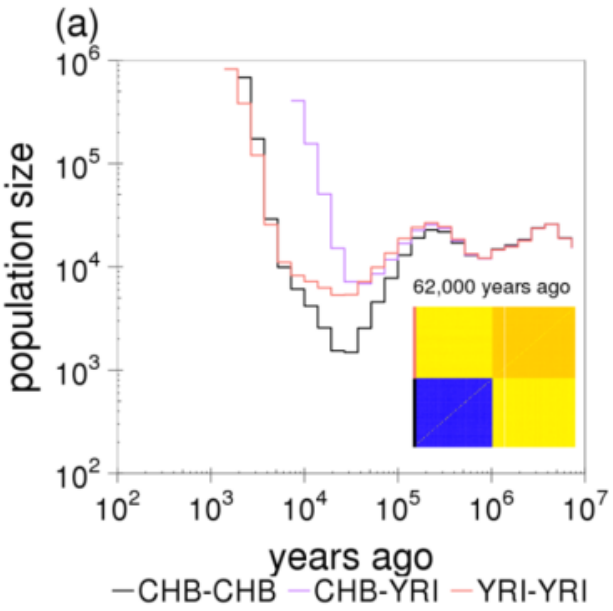
200 haplotypes; chr1
recom. rates; $\mu=1.25\times10^{-8}$

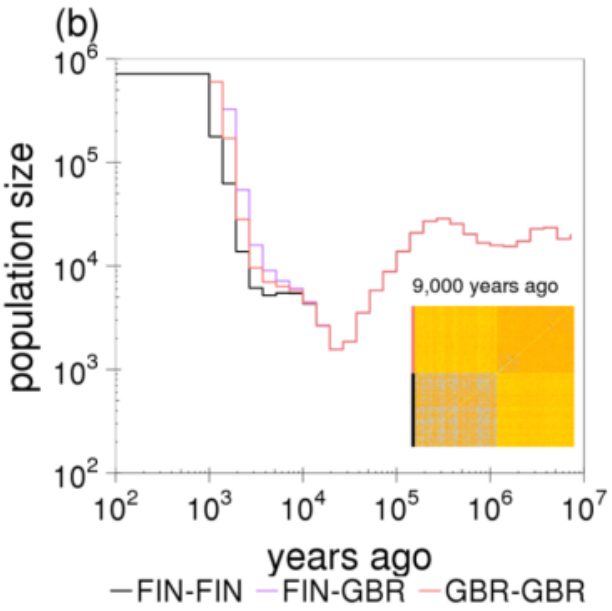
1000 Genomes Project data:

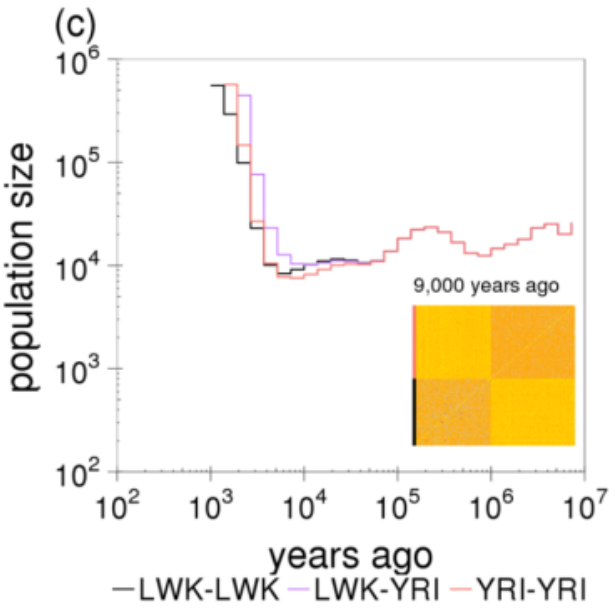
- 4956 haplotypes from 26 populations
- ~71,000,000 biallelic SNPs
- ~93% of SNPs map uniquely to a tree (80% for CpG mutations due to repeat mutations)

Run time: ~4 days on 300 cores

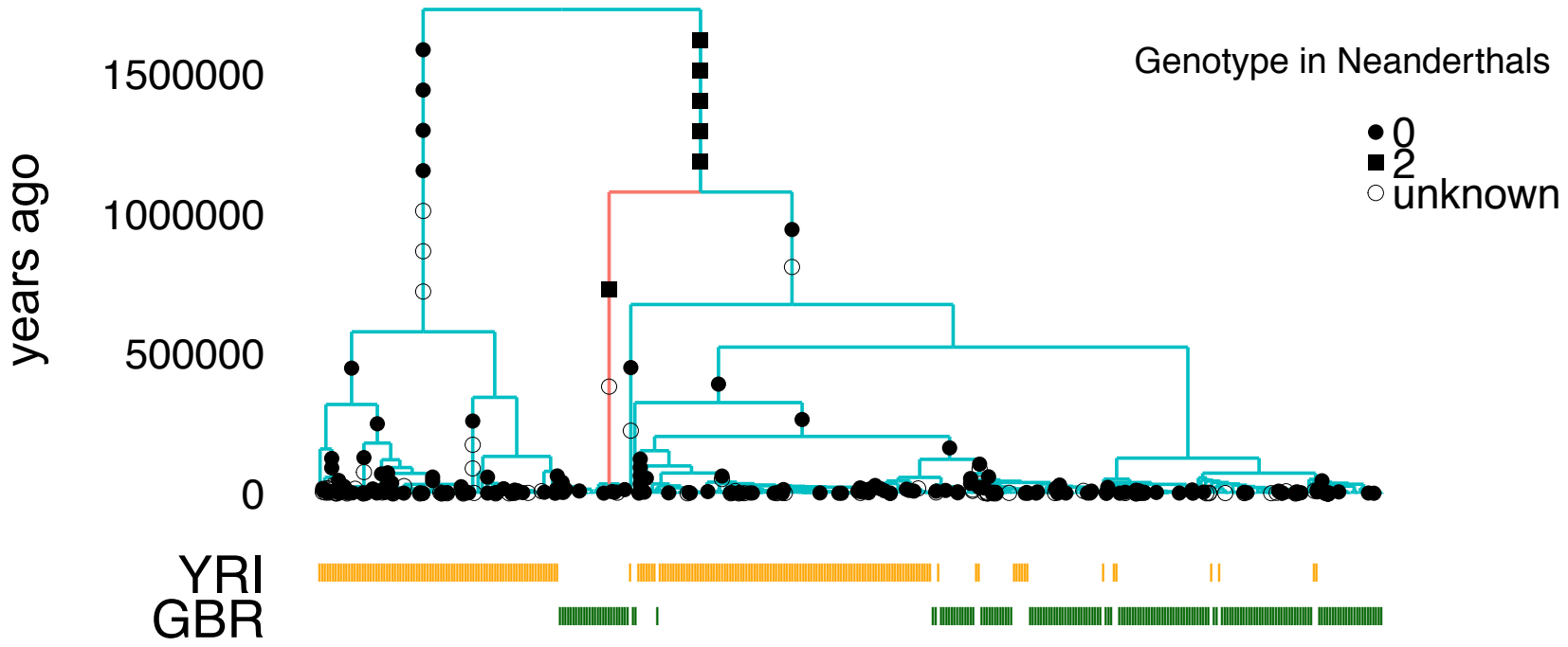
Estimates of human population sizes





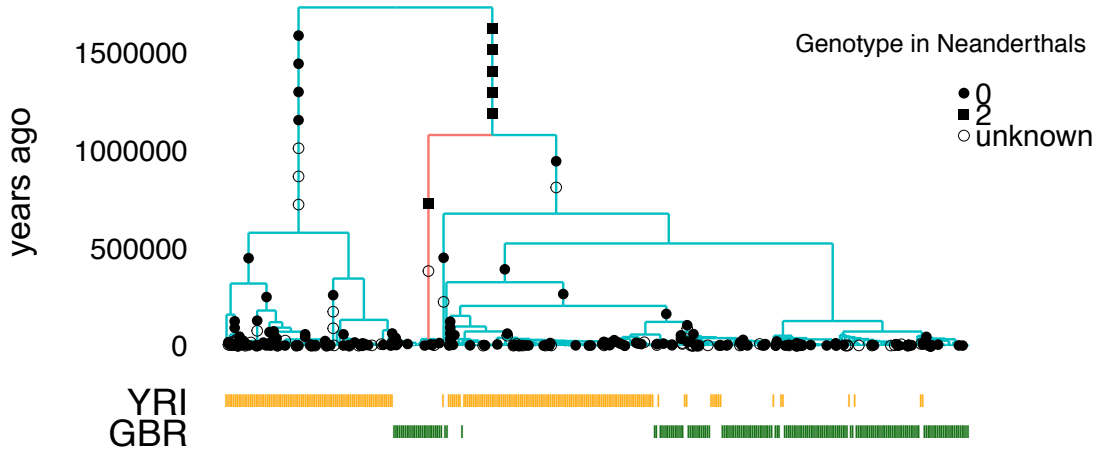


Example tree

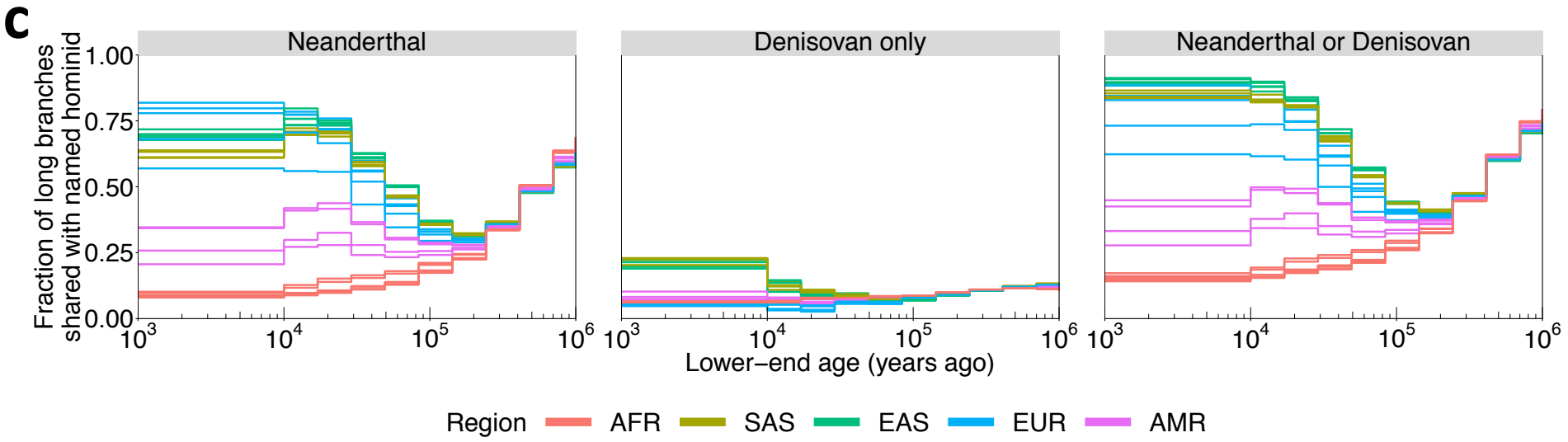


Only 2 populations shown, for clarity;
some GBR individuals carry a Neanderthal haplotype of age <50,000 years

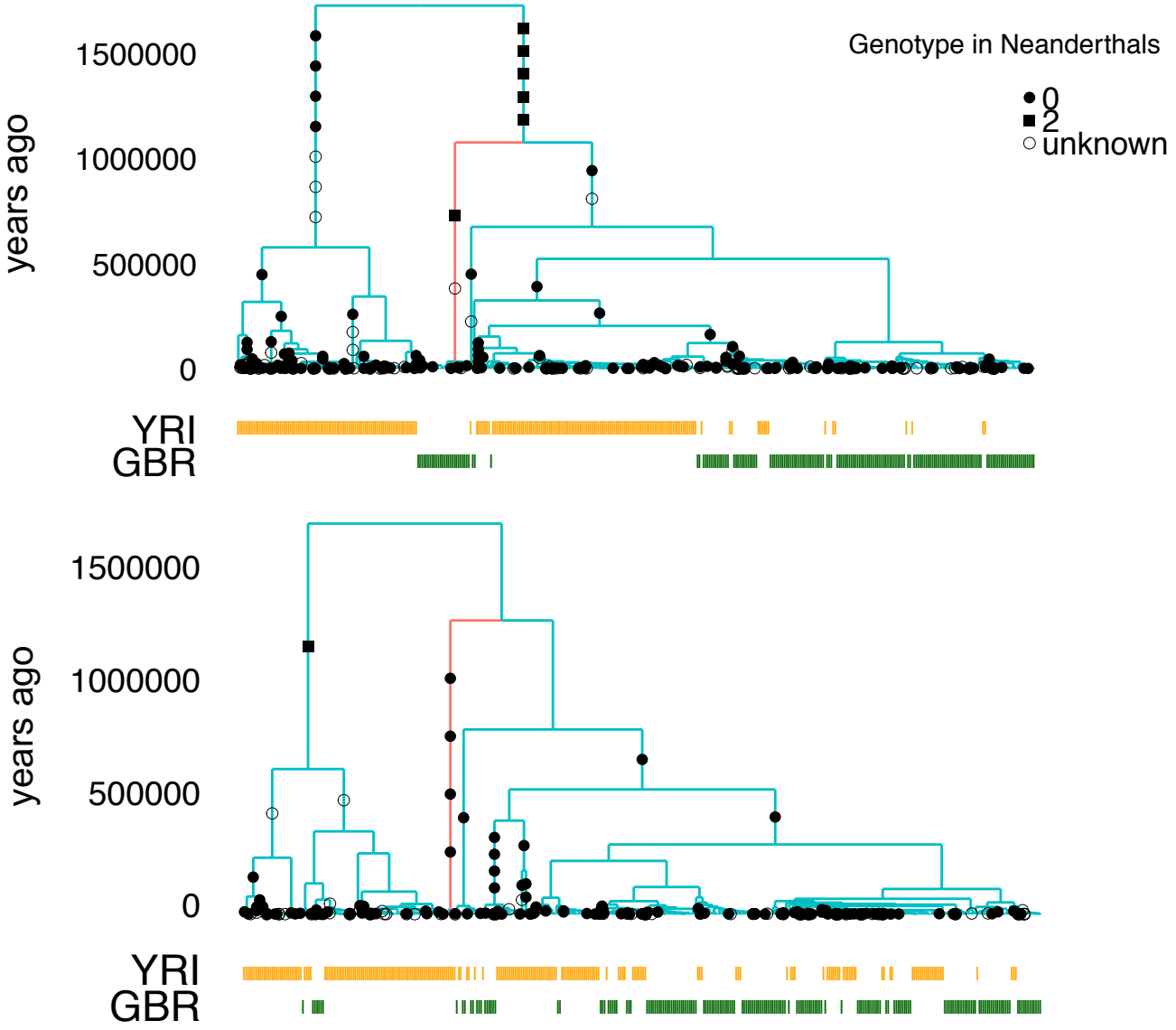
Introgression in humans



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

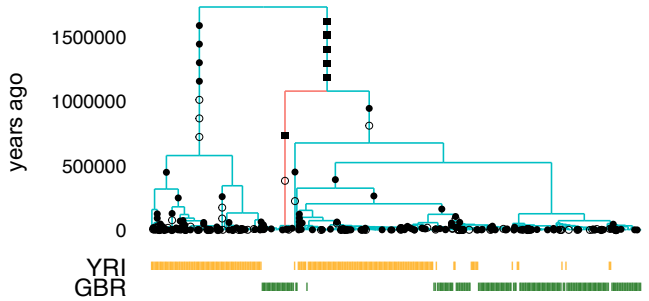


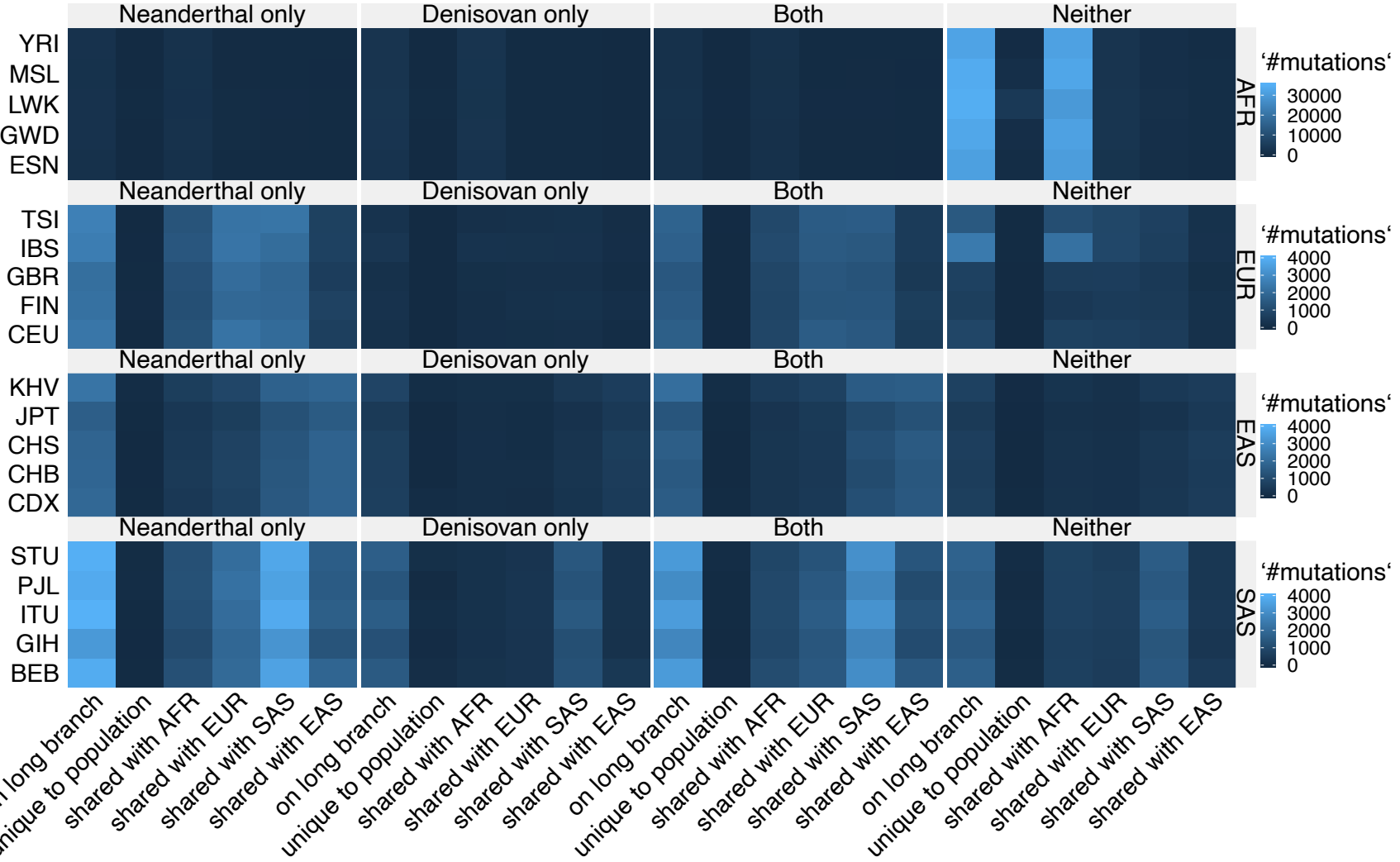
Many long branches in African genealogies



These long branches are private to African populations

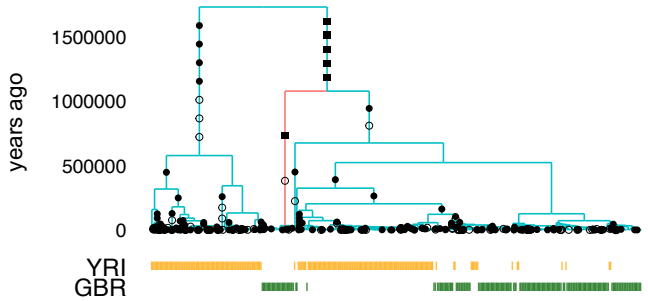
Long branch (upper end >1MY, lower end <30kY)

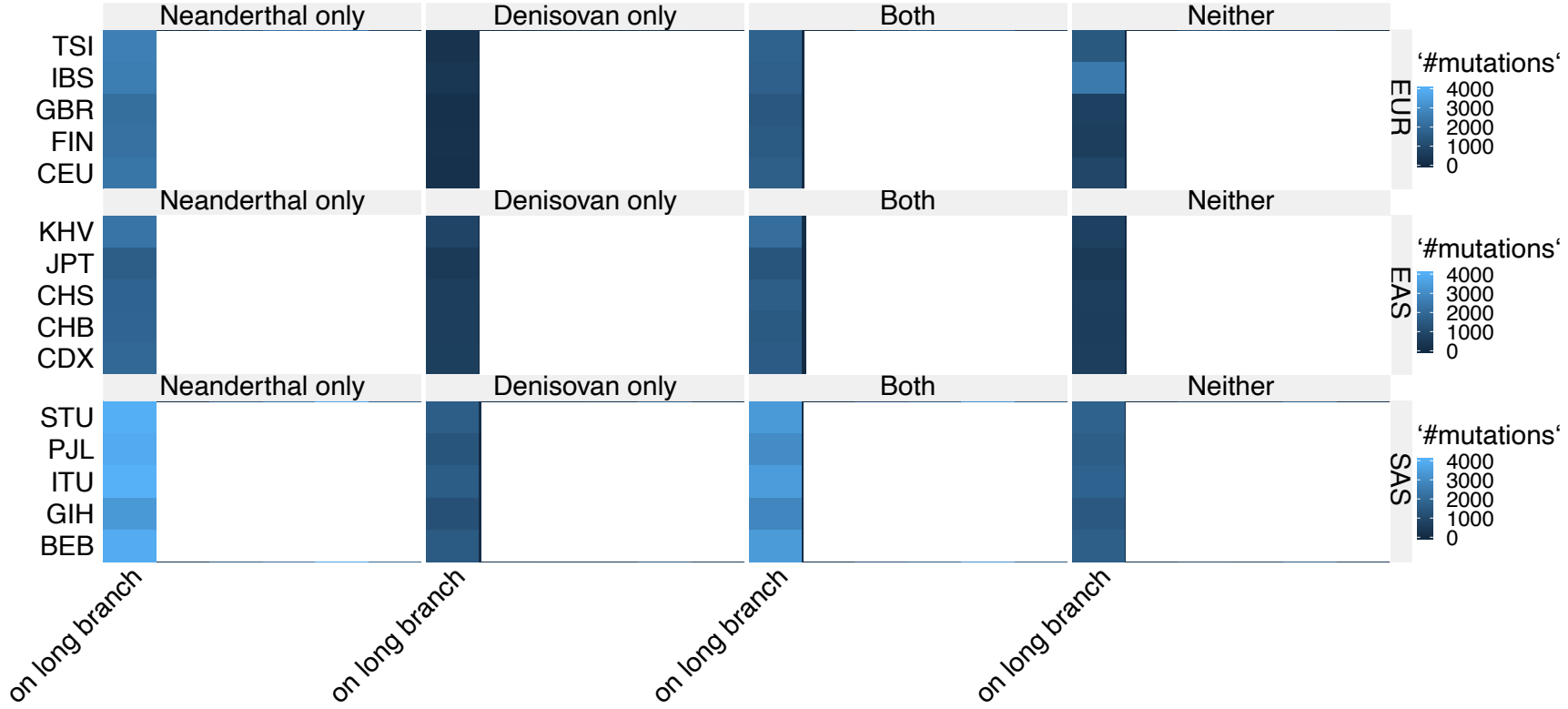




These long branches are private to African populations

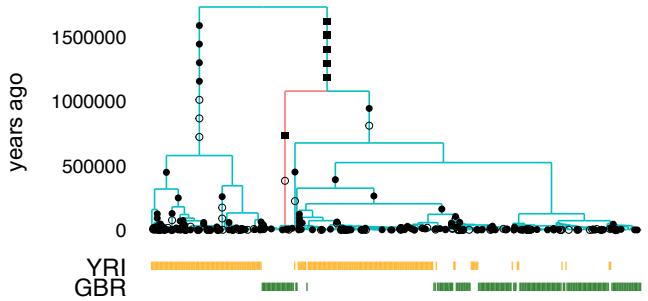
Long branch (upper end >1MY, lower end <30kY)

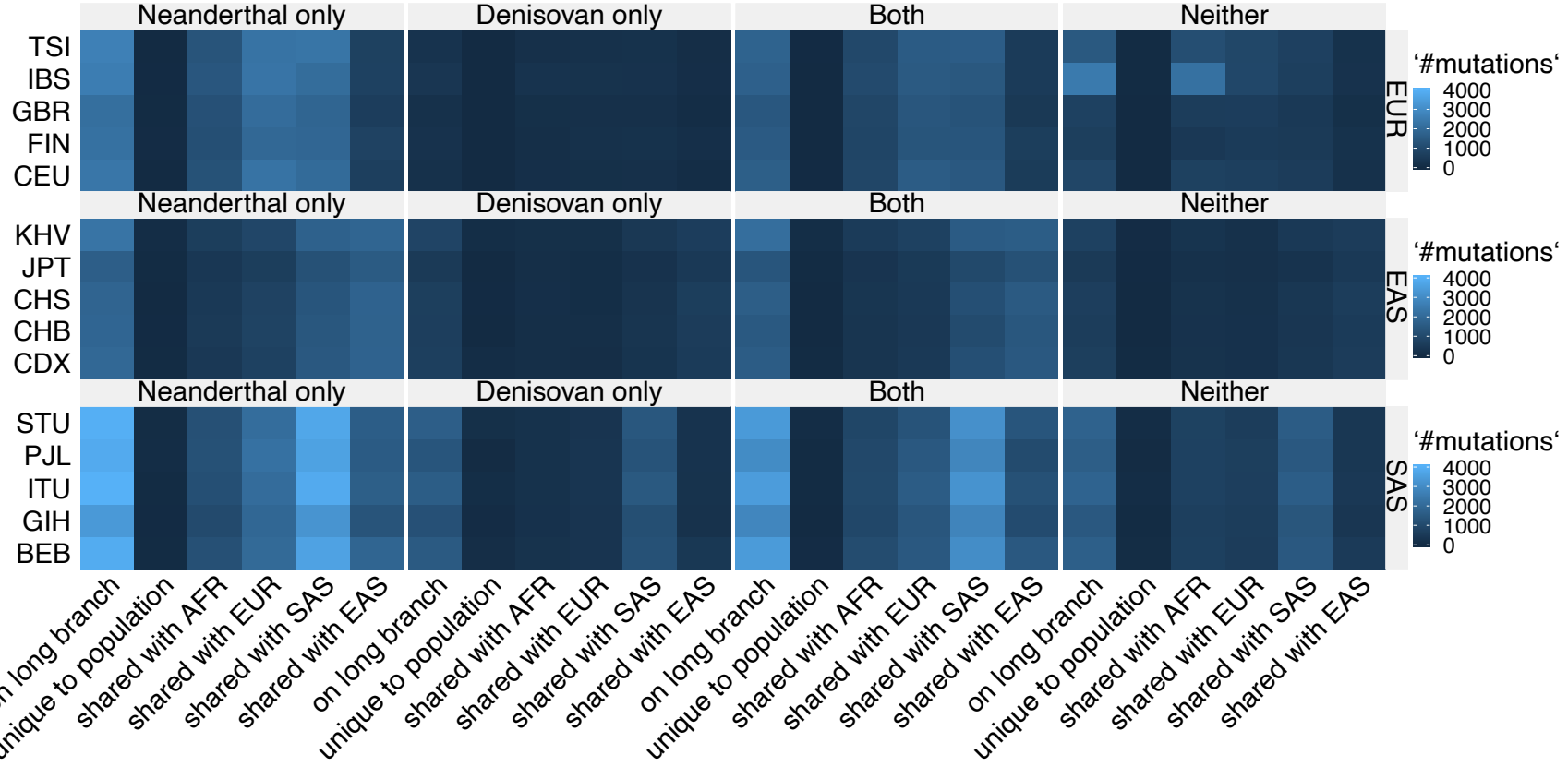




These long branches are private to African populations

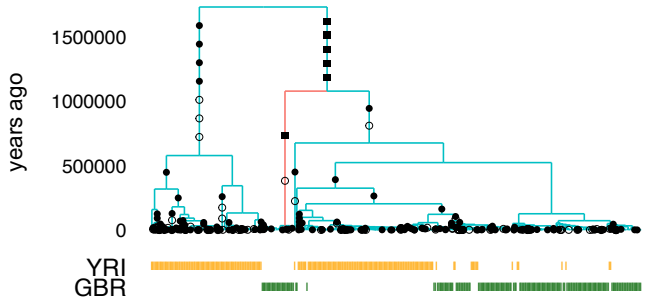
Long branch (upper end >1MY, lower end <30kY)

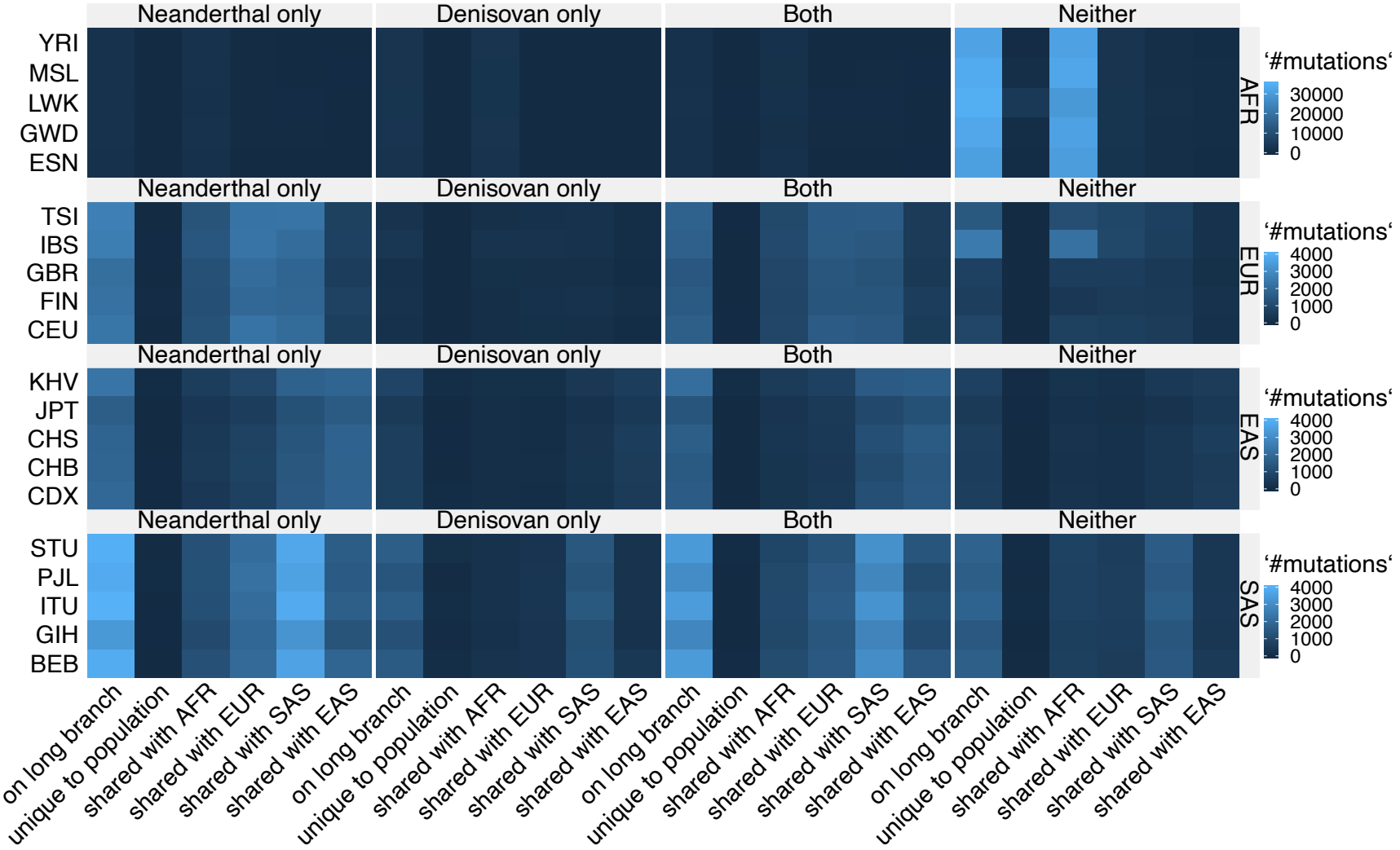




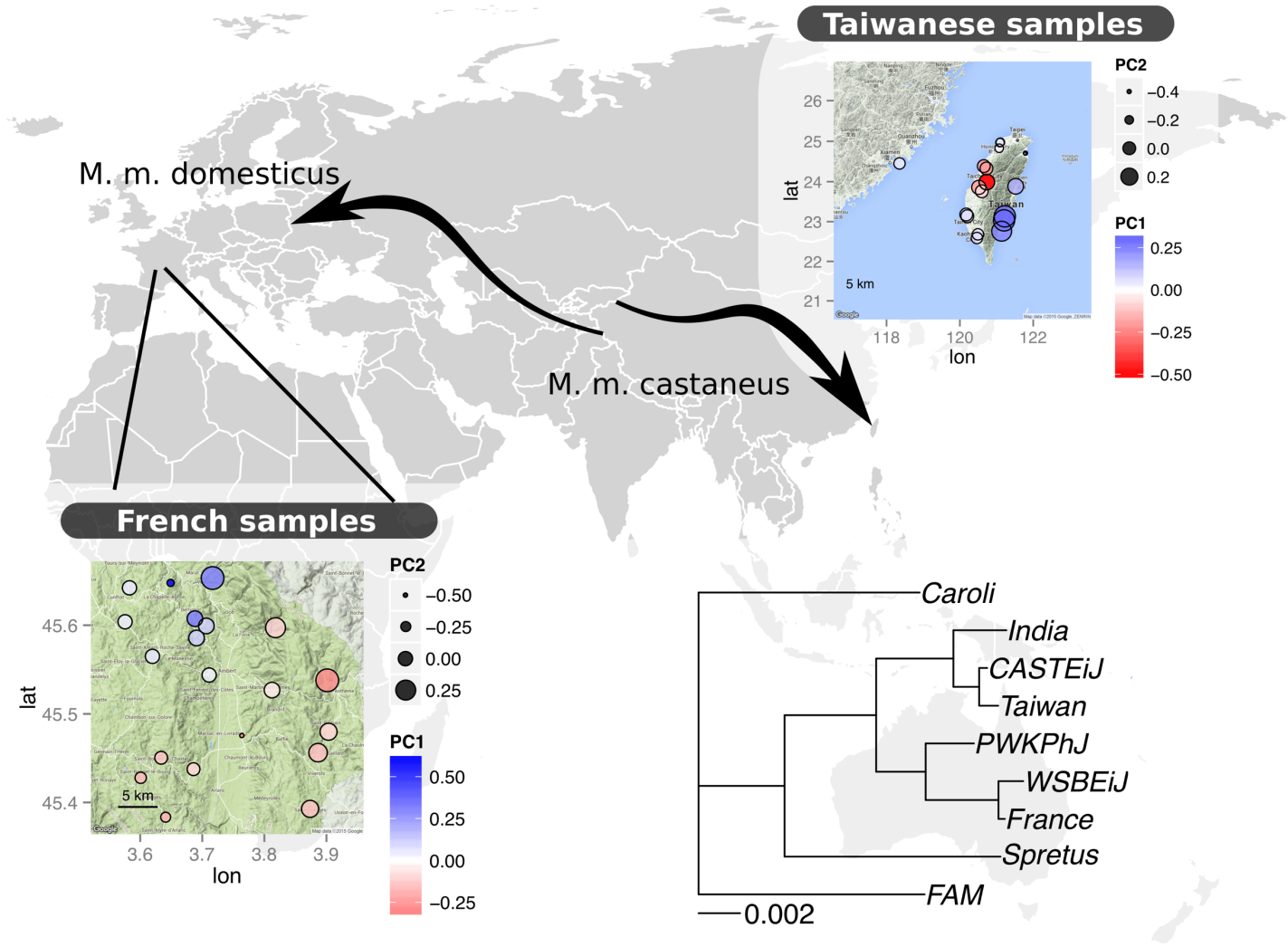
These long branches are private to African populations

Long branch (upper end >1MY, lower end <30kY)



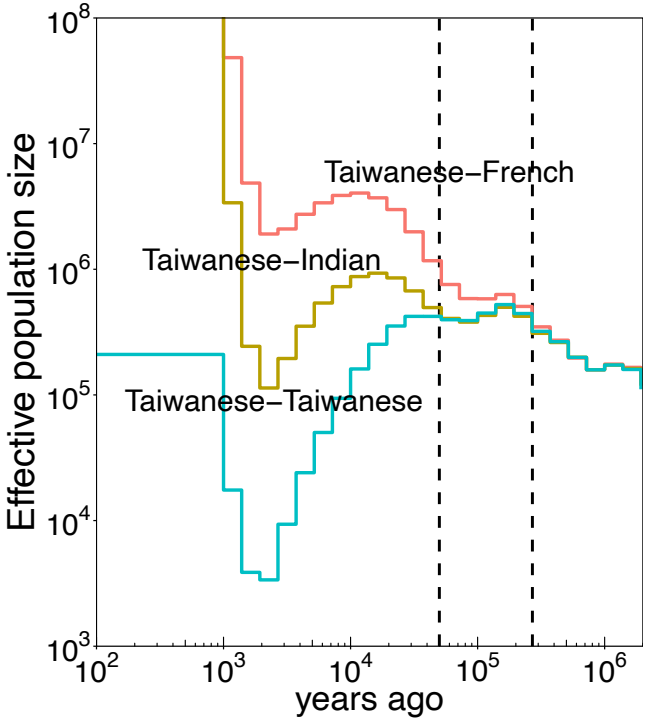


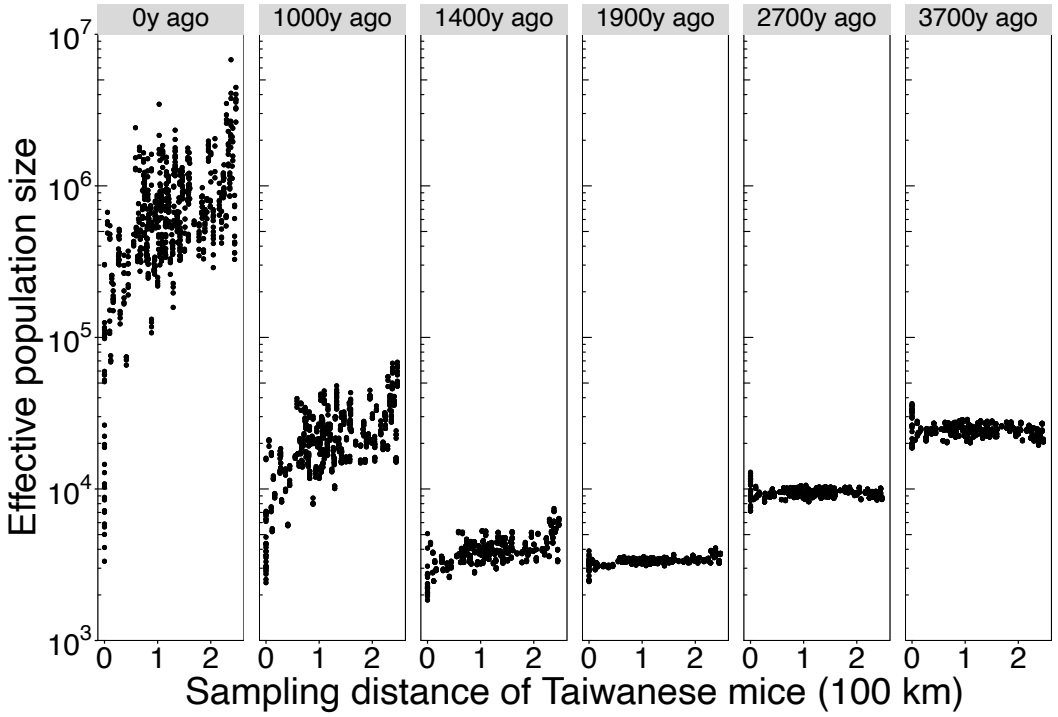
Relate applied to 50 wild mice sampled in India, Taiwan, and France



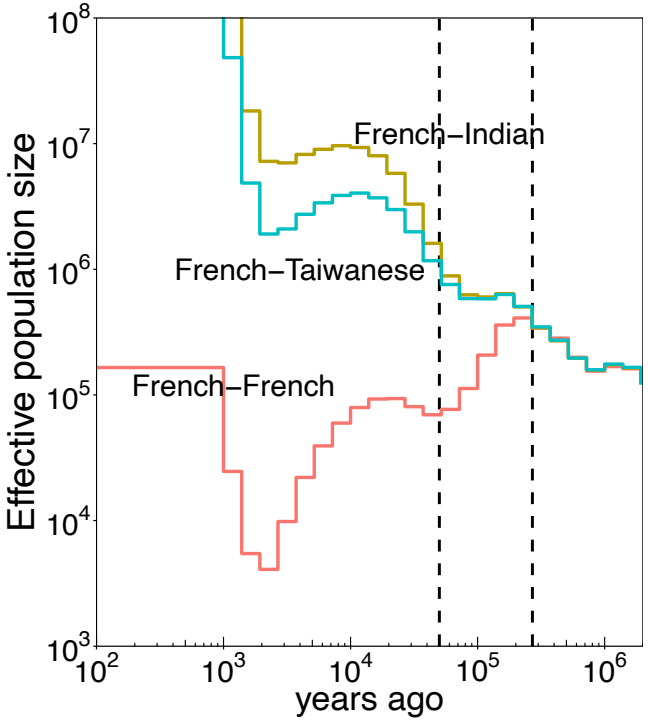
Runtime: 17 CPU hours for 19 chromosomes, Memory usage < 2.5 Gb

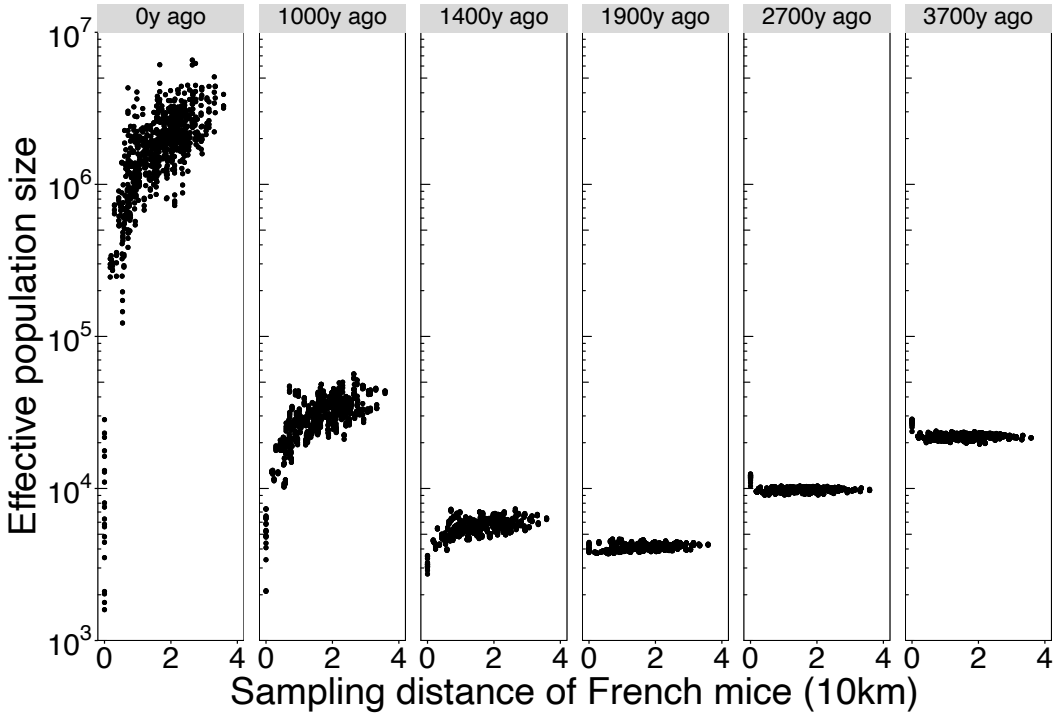
Population structure through time in Taiwanese mice



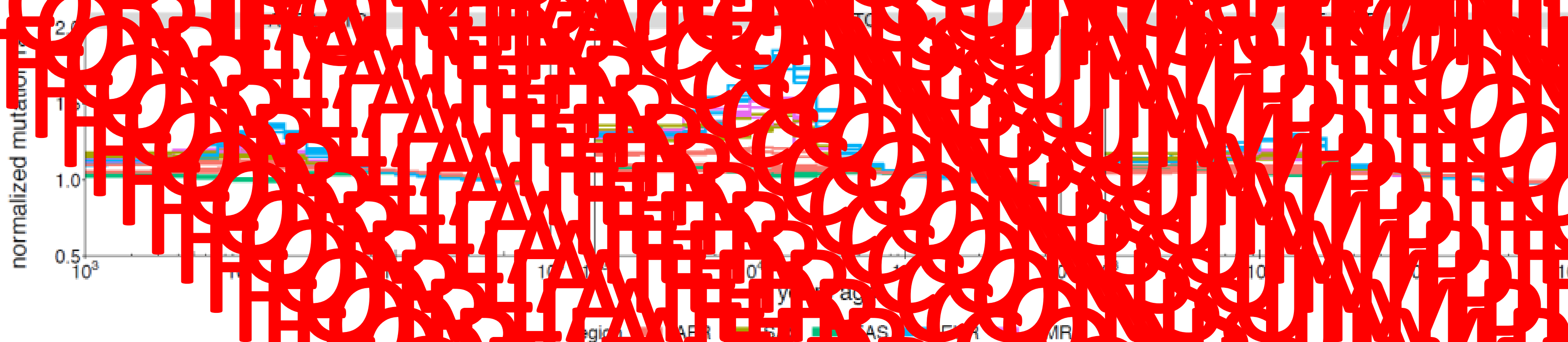


Population structure through time in French mice



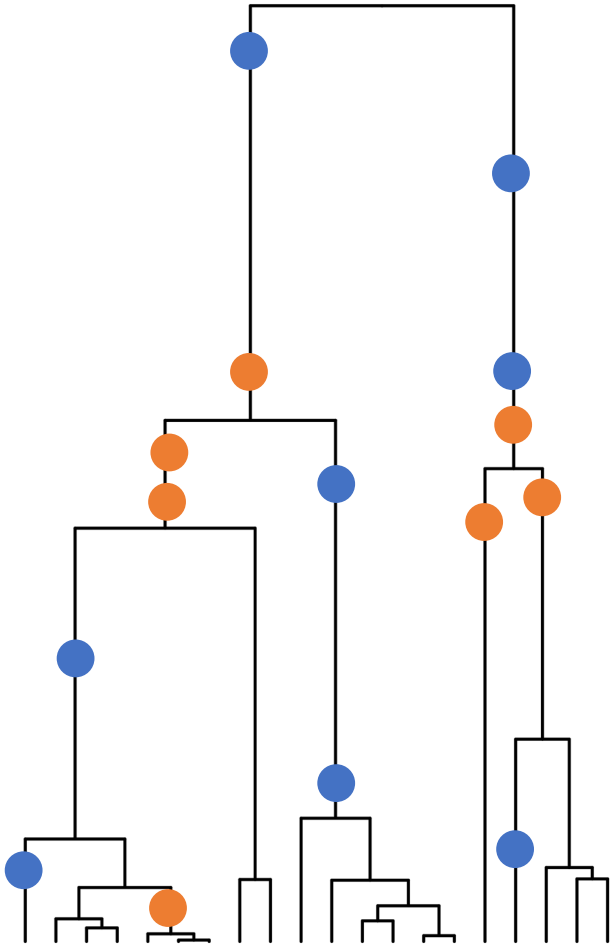


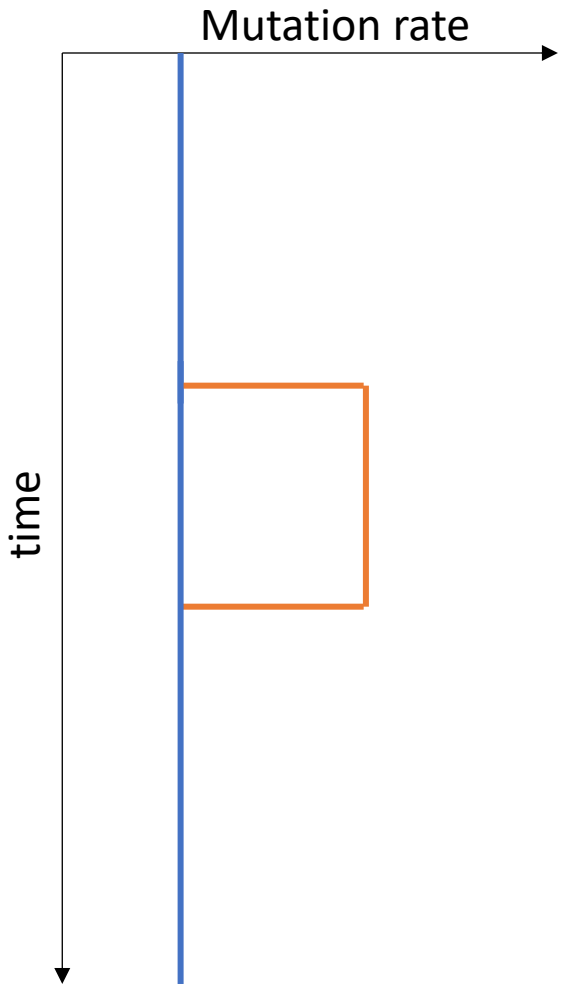
Changing human mutation rates through time



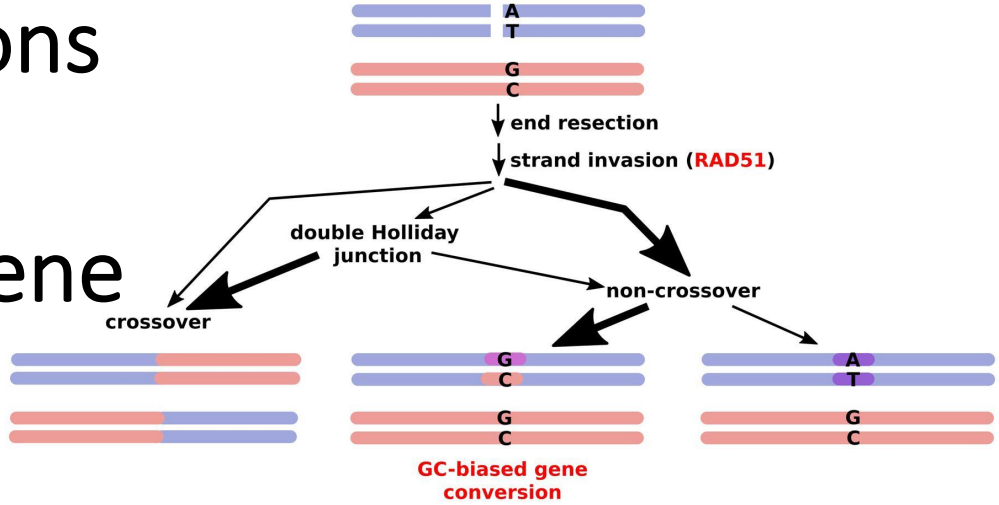
Our results closely mirror those of Kelley Harris (PNAS 2015)

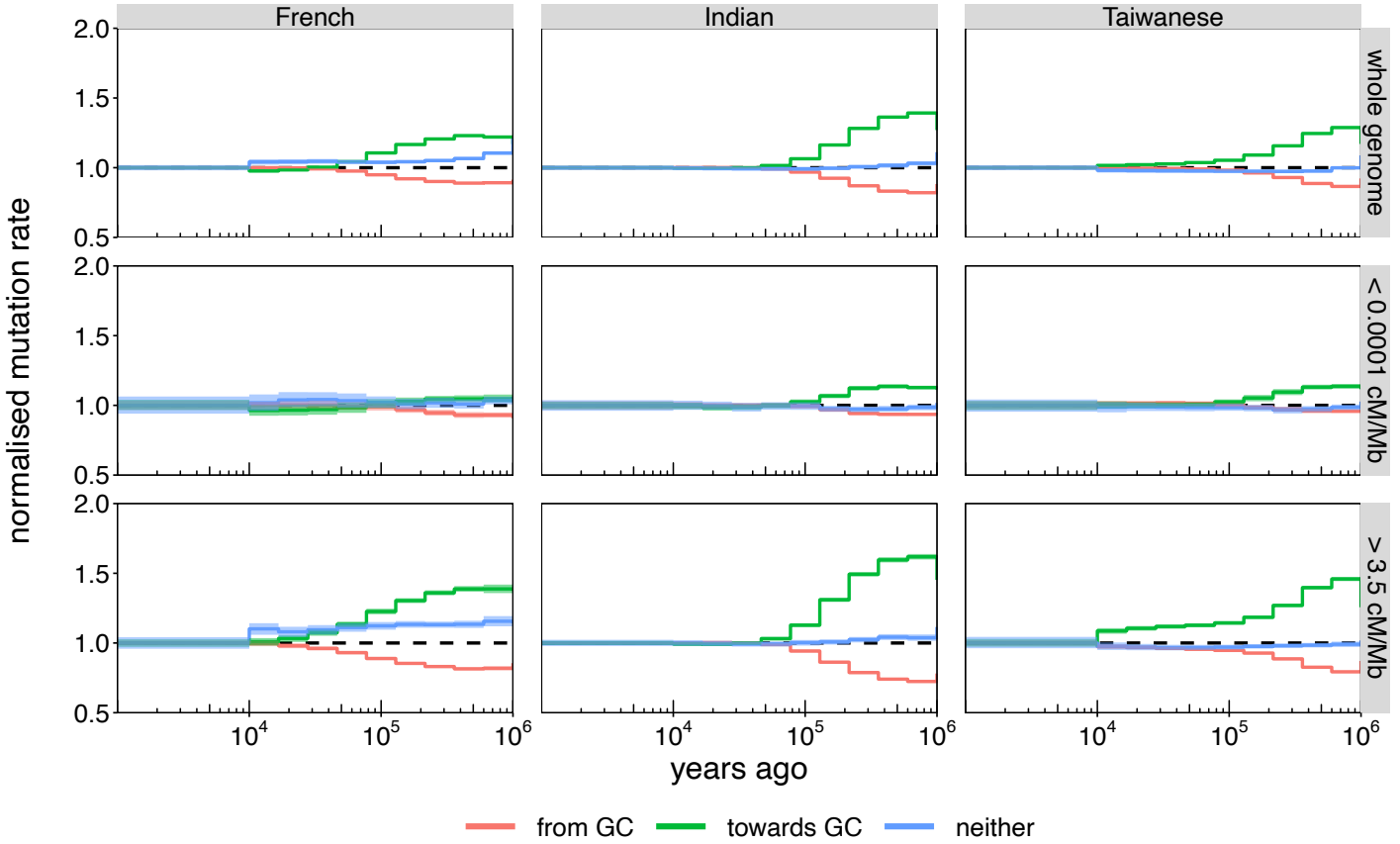
Reminder: Clusters of mutations in time can capture changes in mutation rate





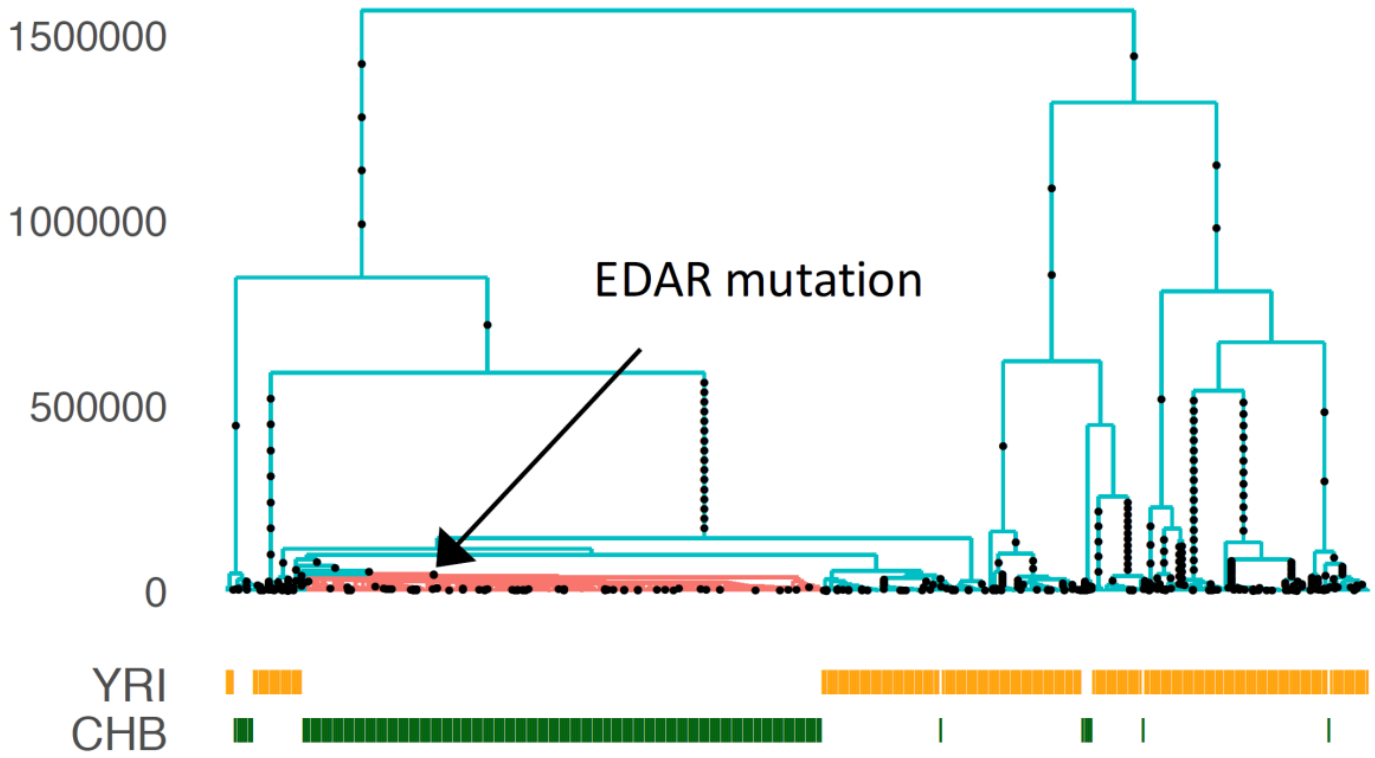
Excess of older mutations in mice towards GC evidences GC-biased gene conversion



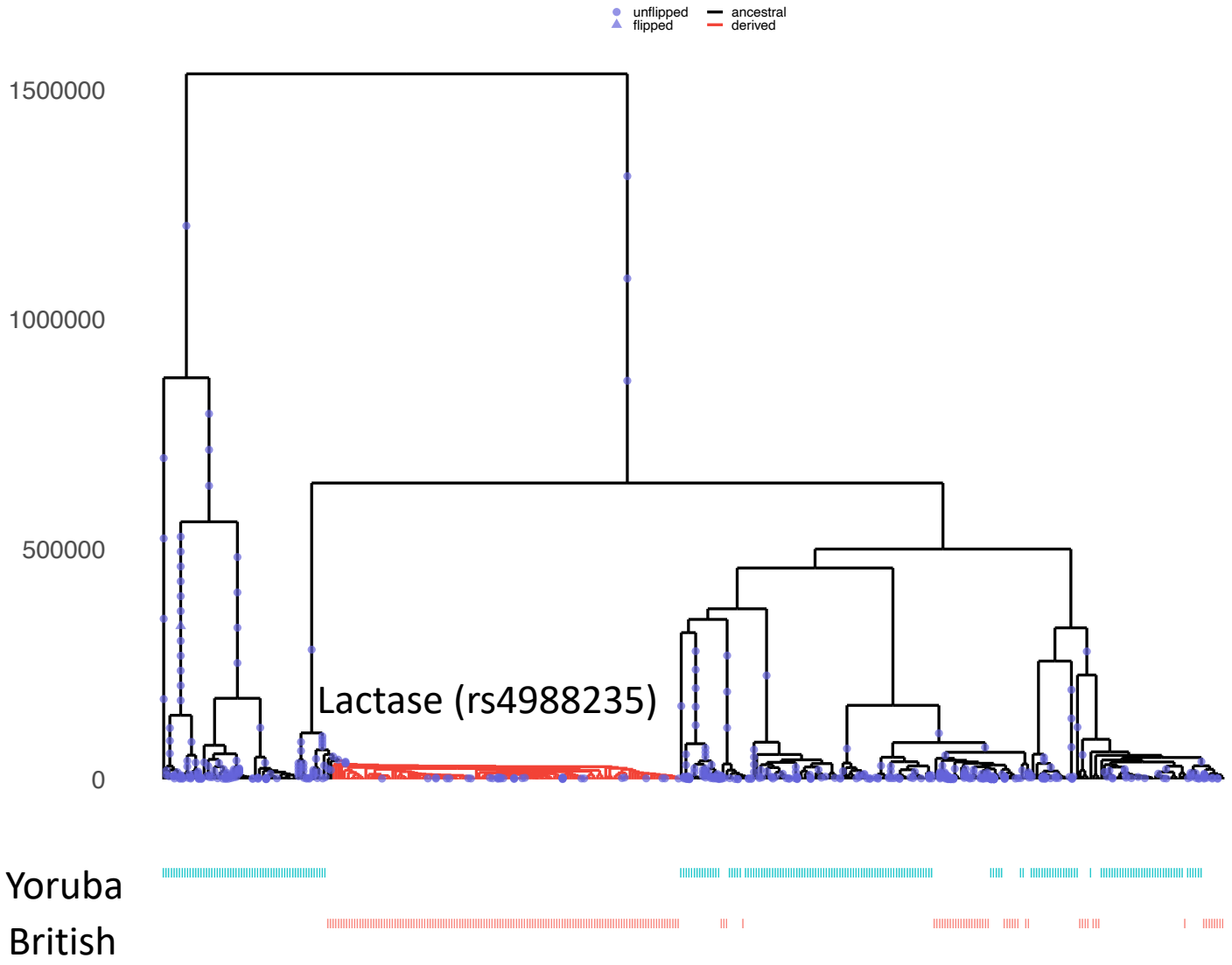


Detecting signals of positive selection

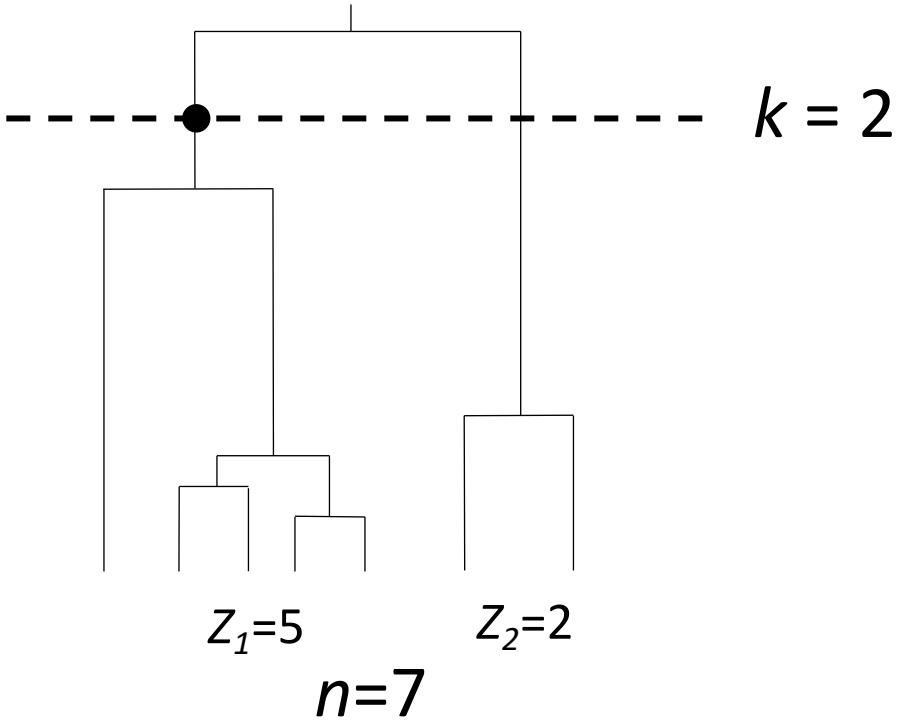
- Genetic adaptations to changing environment, diet, lifestyles,…
- Use trees incorporating demographic history:



Adaptation for increased Lactose tolerance is one of the strongest selection signals genome-wide



How quickly does a mutation spread in the neutral case?

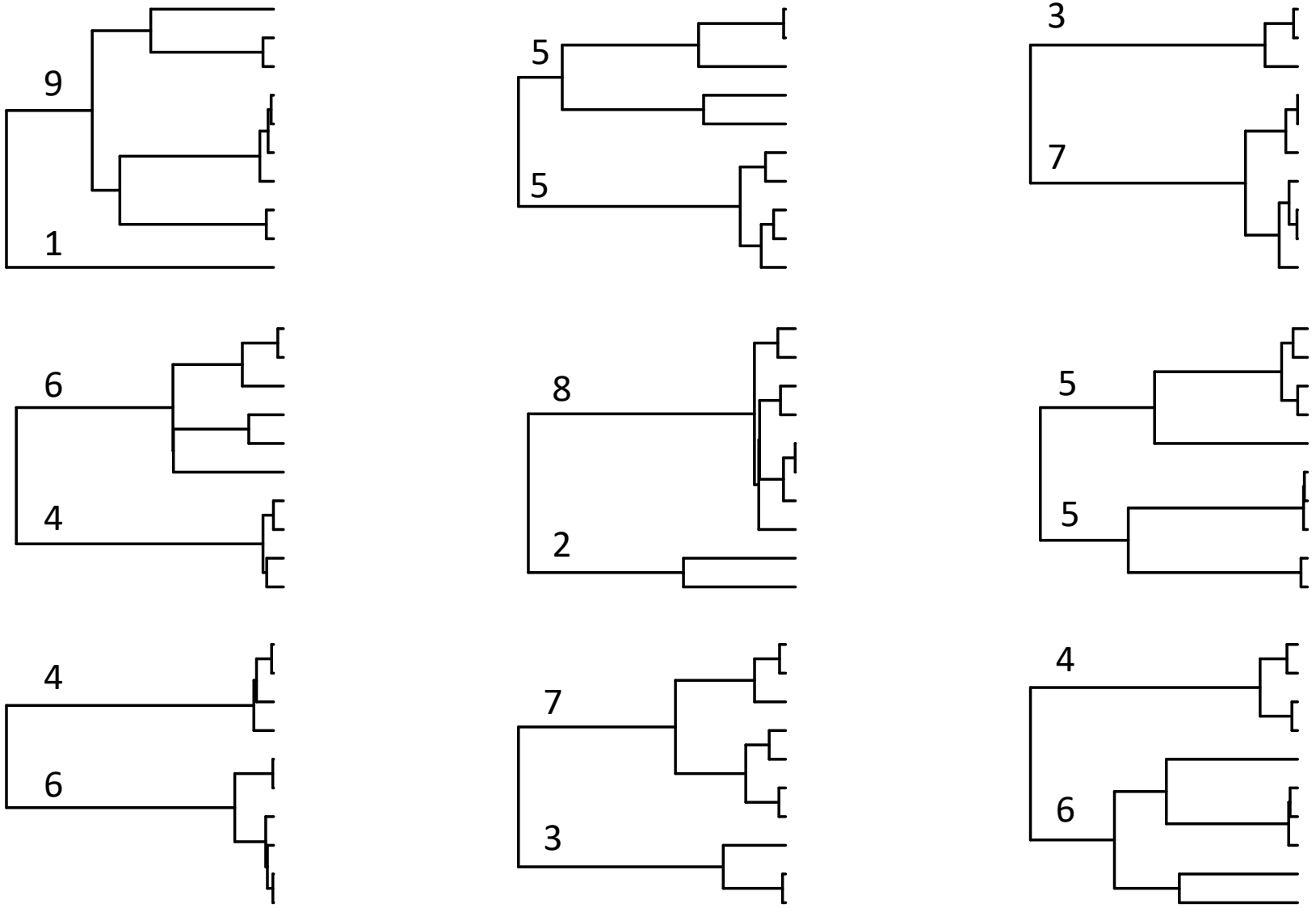


The coalescent is simple so it is possible to analytically write down the probability a mutation arising while *k* lineages are in the tree has some number of descendants

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

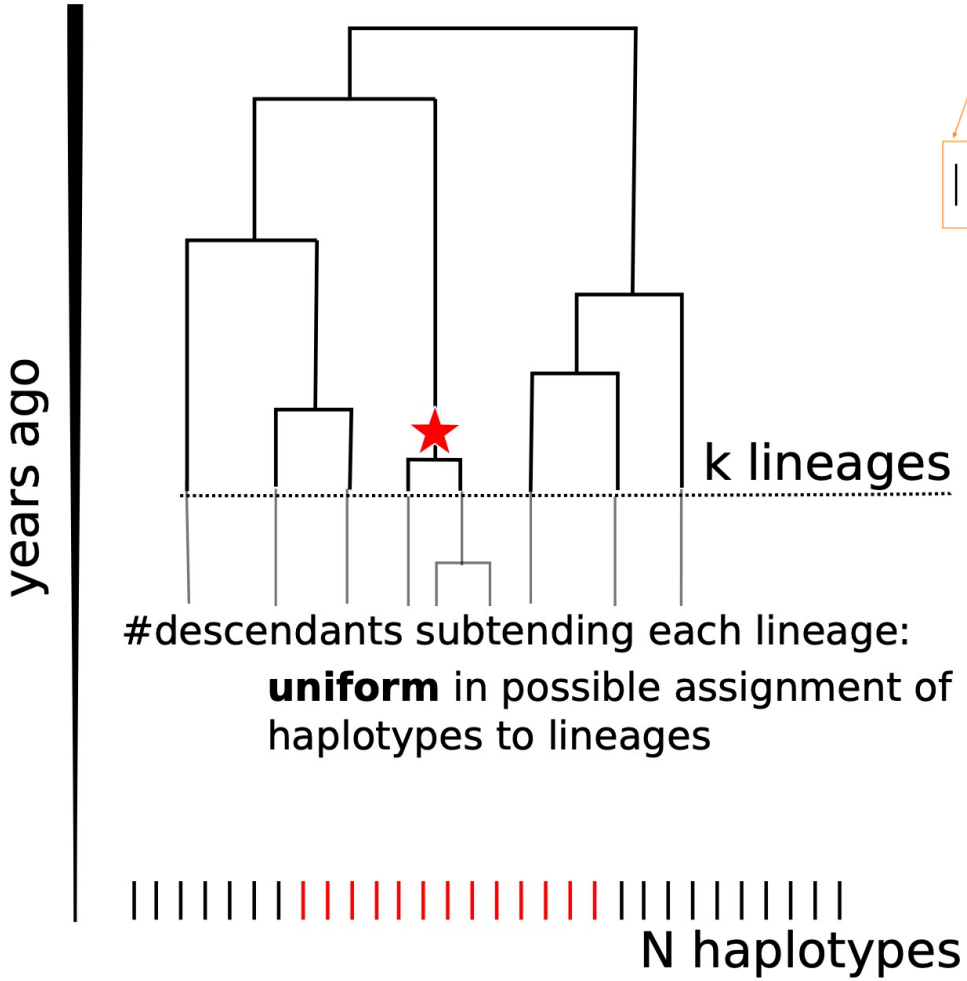
P(5 descendants)=1/6

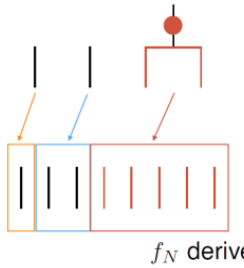
The k=2 case



P-value for evidence of positive selection

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



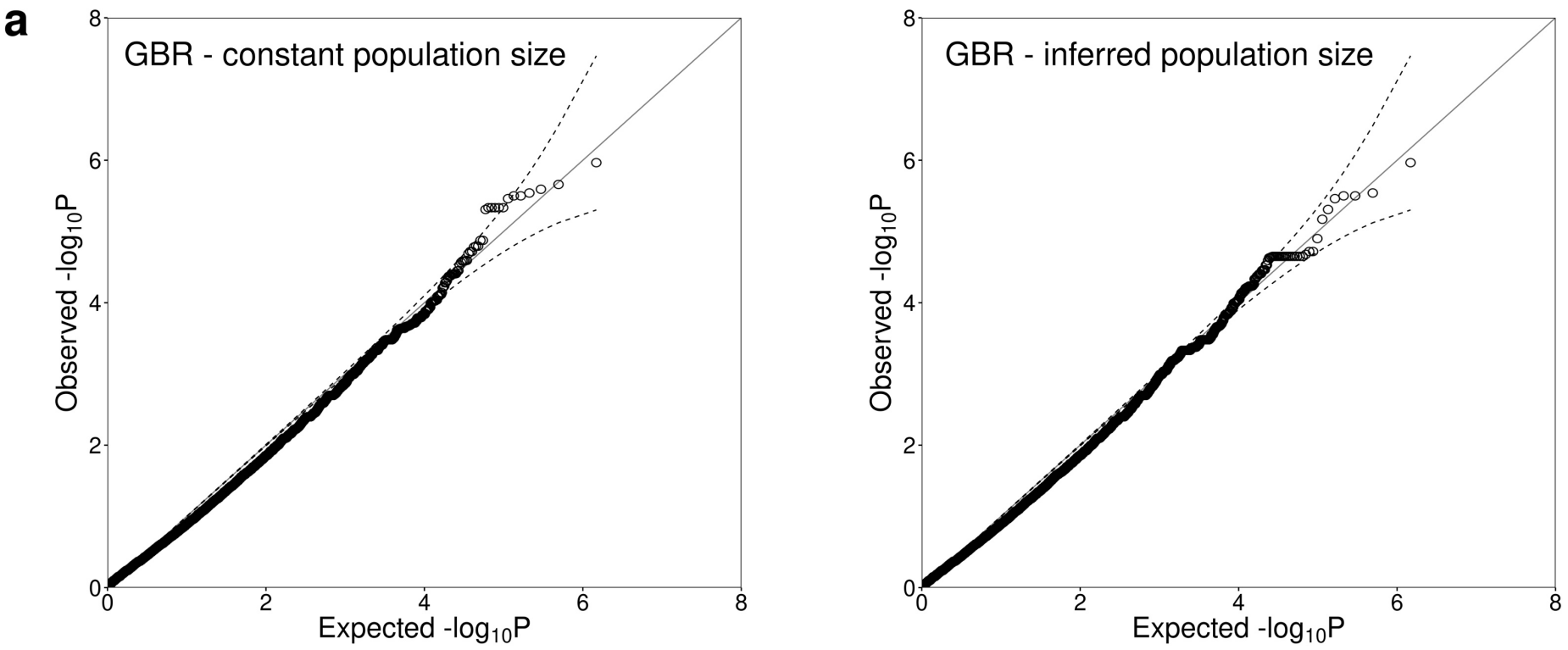


$$P(f_N|k) = \frac{\binom{f_N-1}{2-1}\binom{N-f_N-1}{(k-2)-1}}{\binom{N-1}{k-1}}$$

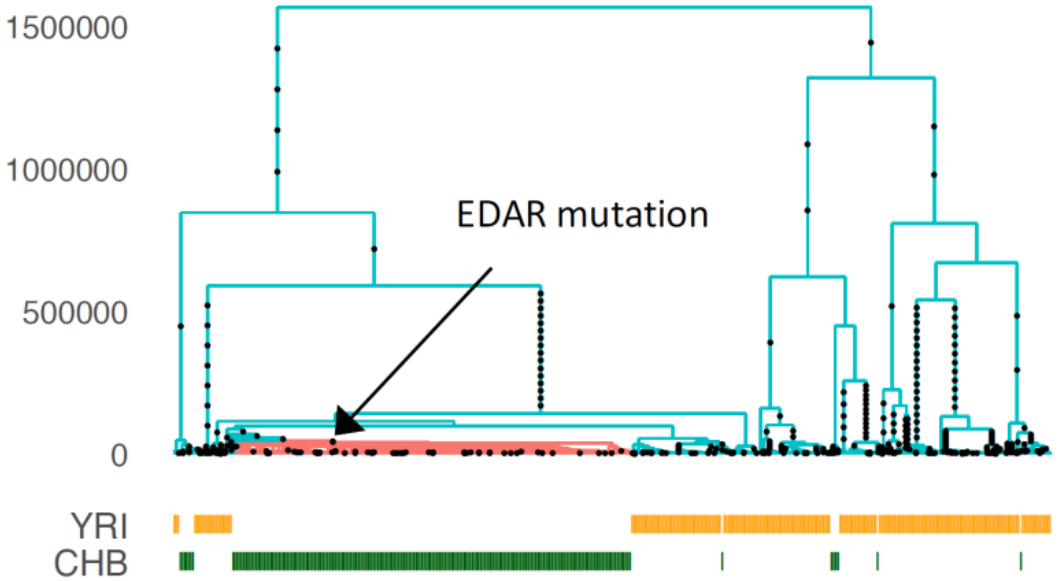
$$\text{p-value} = \sum_{f=f_N}^{N-k+2} P(f|k)$$

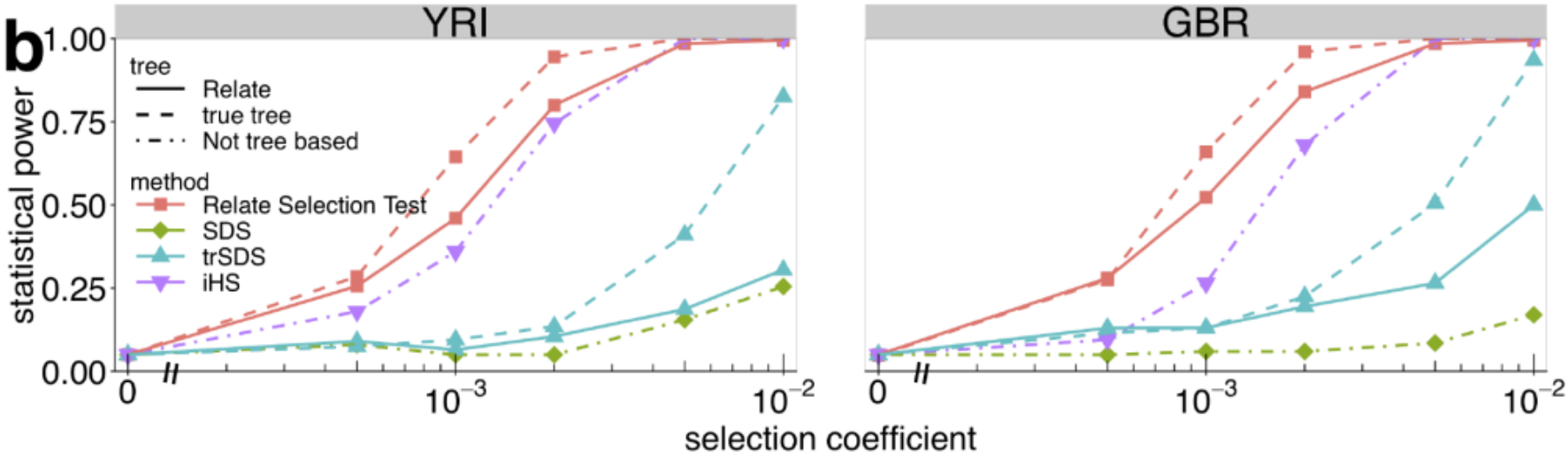
P-values: very well calibrated under null simulations of no selection

N=1000, 250Mb
Bottleneck population size



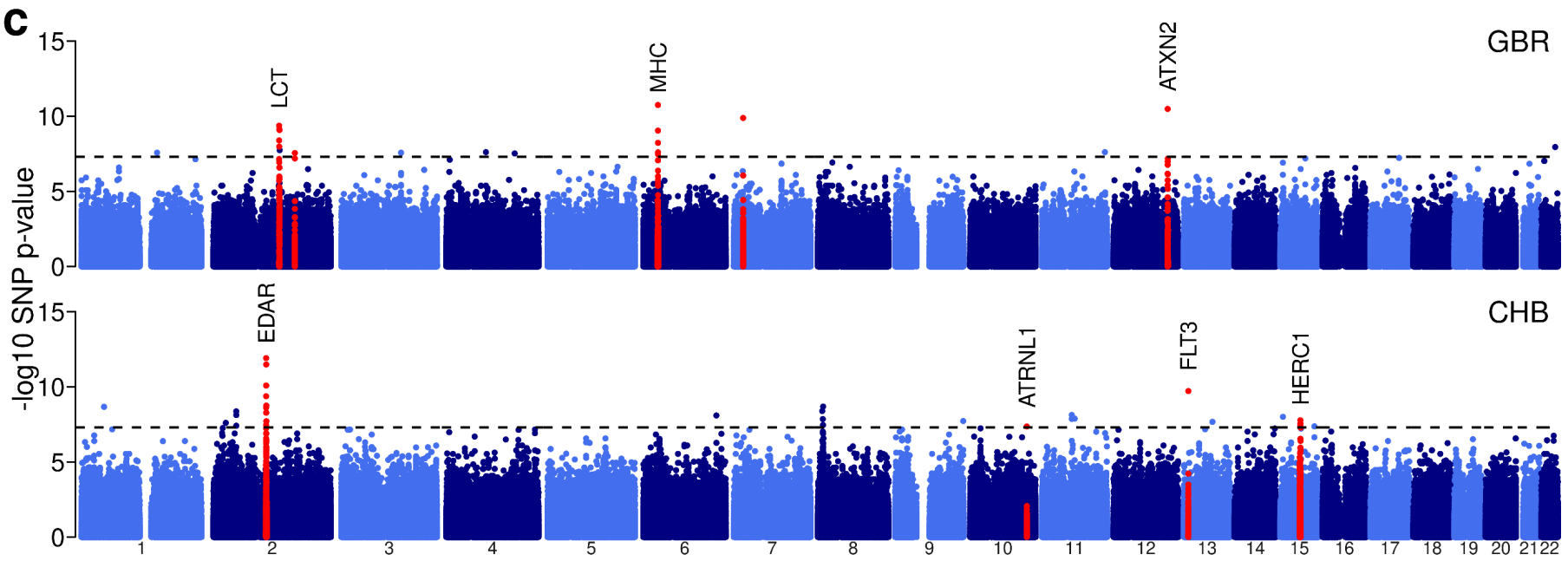
Much better power to see weak selection than existing approaches





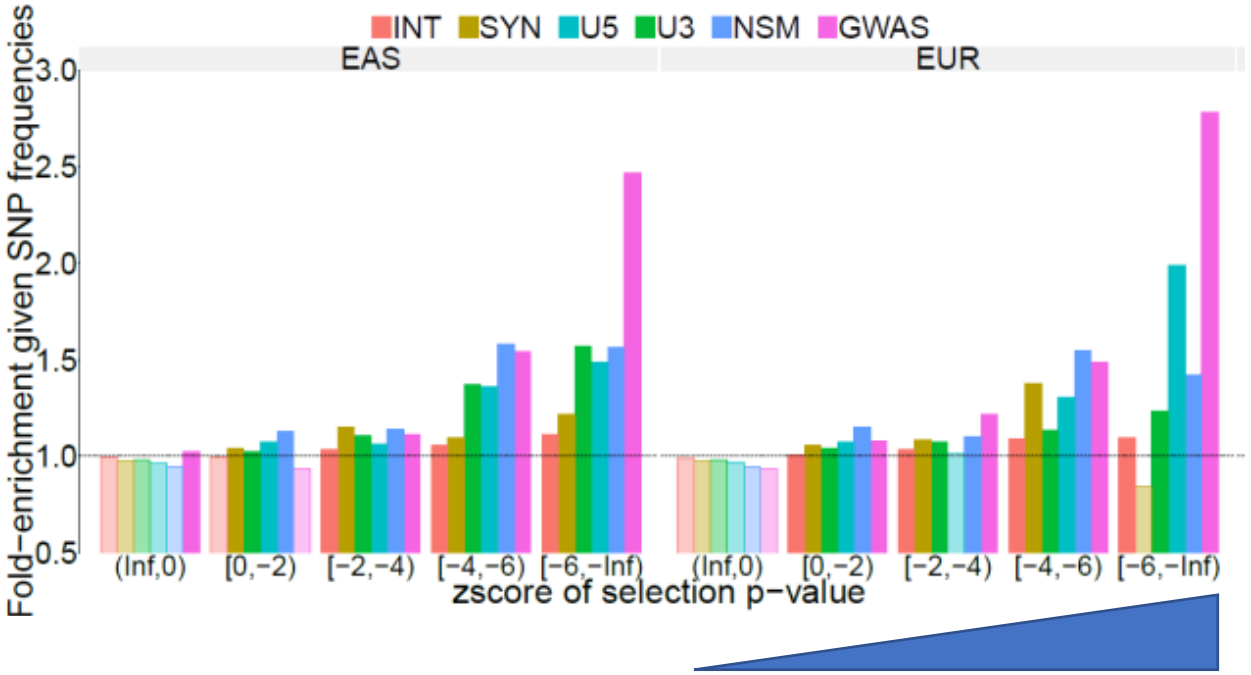
Genome-wide selection p-values

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



How does weak selection evidence vary by functional class of SNP?
By individual trait?

GWAS hits are most enriched, among selection signals we observe

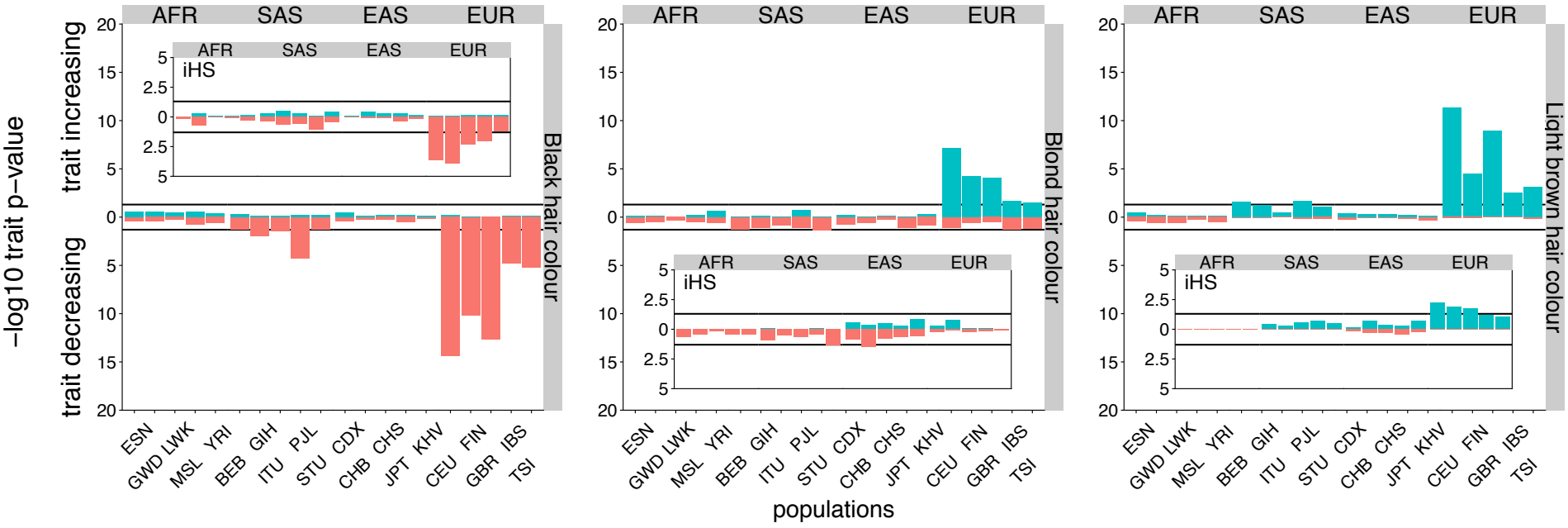


INT	Intronic SNP
SYN	Synonymous coding SNP
U5	5' untranslated region
U3	3' untranslated region
NSM	Non-synonymous coding SNP
GWAS	Genome-wide significant GWAS hits

Evidence of selection on a trait: hair colour

We use

- Effect direction of ”genome-wide significant” associations
- Compare selection p-values to frequency matched random SNPS

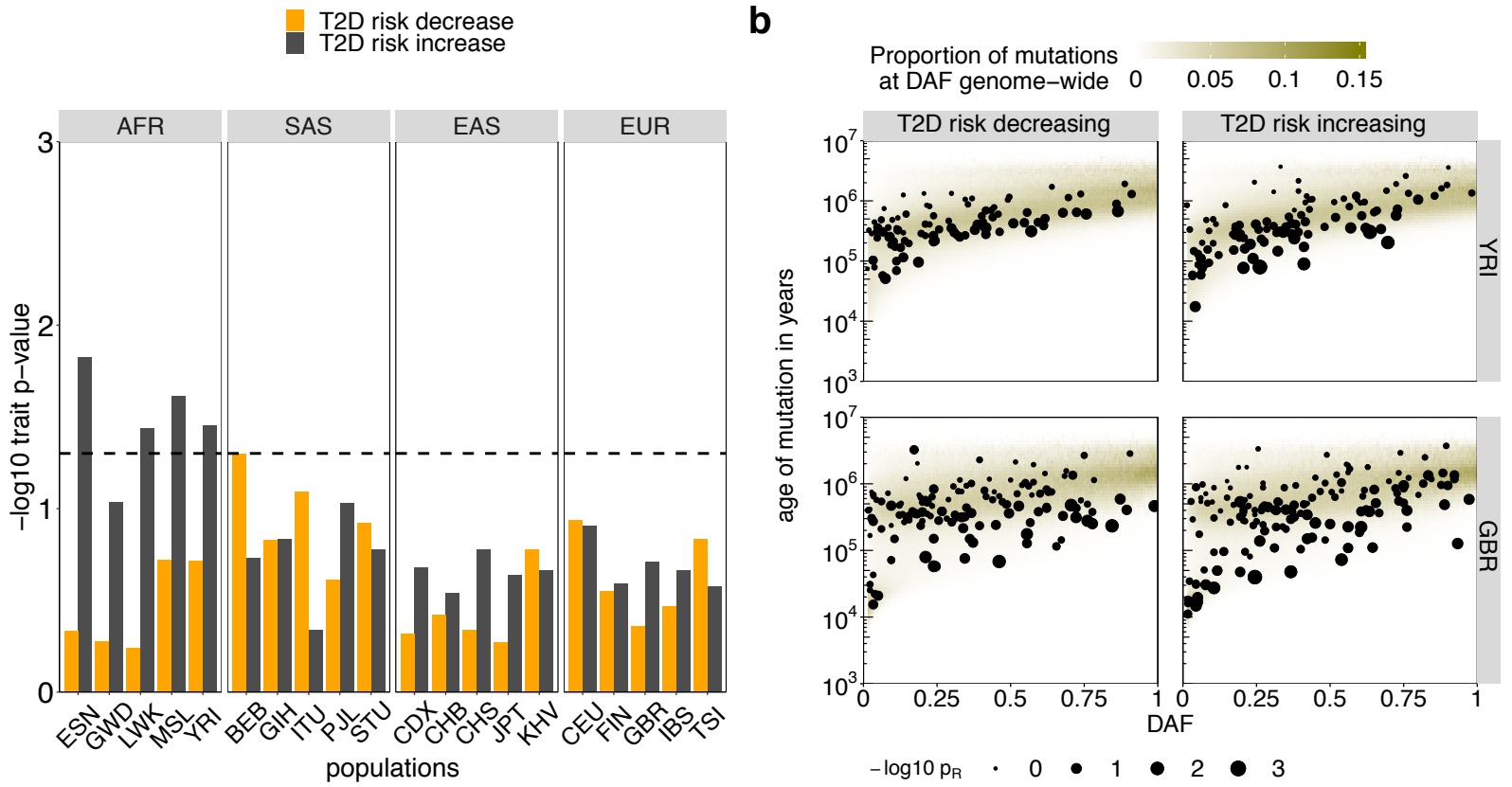


- Correlated phenotypes
- Pleiotropy
- Biased effect sizes (e.g., due to genetic structure)
- Unbalanced power for different ancestries

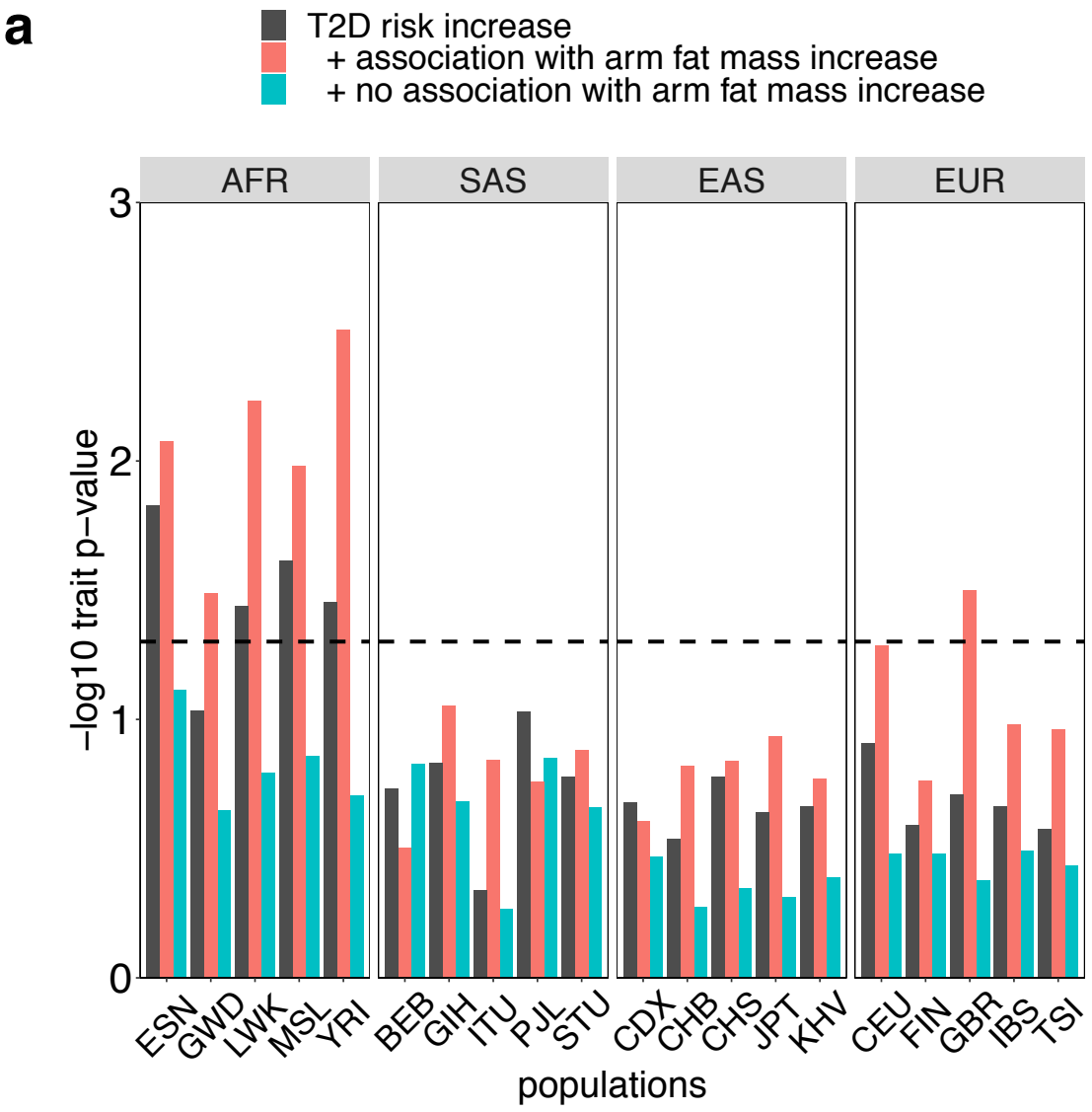
Evidence of selection on a trait: type II diabetes

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

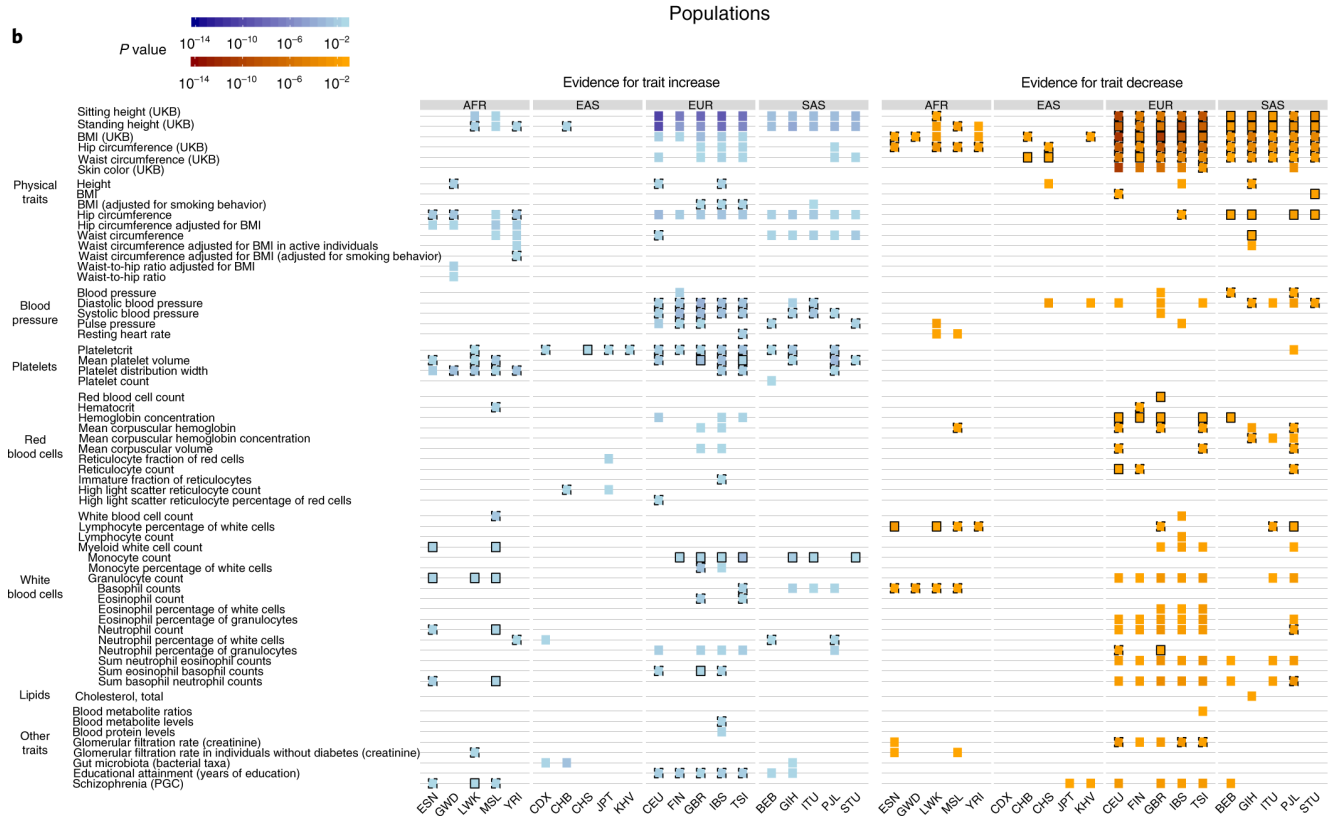


Type II diabetes: selection via fat related traits?



Many GWAS signals of trait selection, clustering geographically

Directional effects (GBR: 36 of 76 traits tested)



Blood pressure
↑ **(EUR, SAS)**

Hip circumference
↓ **(SAS)**

Plateletcrit
↑ **(AFR, EAS, EUR, SAS)**

Hemoglobin
↓ **(EUR, SAS)**

Granulocyte
↑ **(AFR)** ↓ **(EUR)**

Conclusions and future work

- It is now possible to build genealogical trees for huge datasets, in humans and other species (currently 10,000 people or more)
- These trees capture information about many processes including ancient introgression, mutation rate evolution, and trait evolution (and many more things)
- There is evidence of widespread, relatively weak, selection on SNPs impacting human traits
  - Pleiotropy, differences in power among human groups, etc. complicate interpretation
- Future work:
  - deeper analysis of varying types of selection
  - Trait evolution through time and space
  - Recombination evolution, directional migration

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

Thanks!

- Sinan Shi
- Marie Forest
- Anubha Mahajan
- Mark McCarthy
- Thaddeus Aid
- Jonathan Marchini
- Pier Palamara

DEPARTMENT OF
STATISTICS