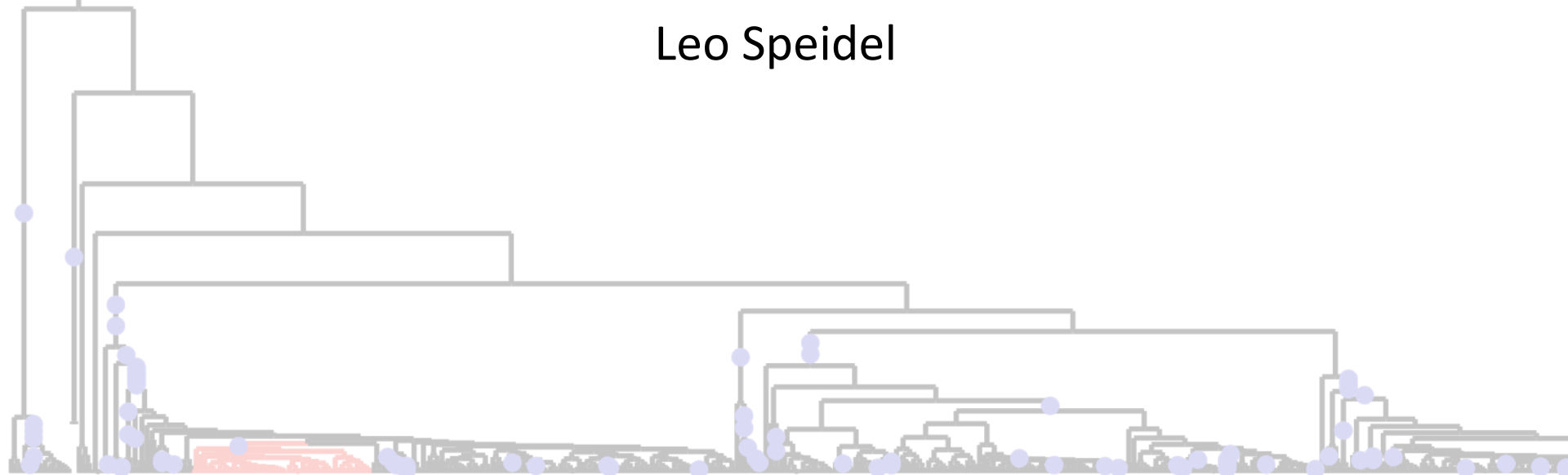
Learning about evolution by building coalescent trees

Leo Speidel



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

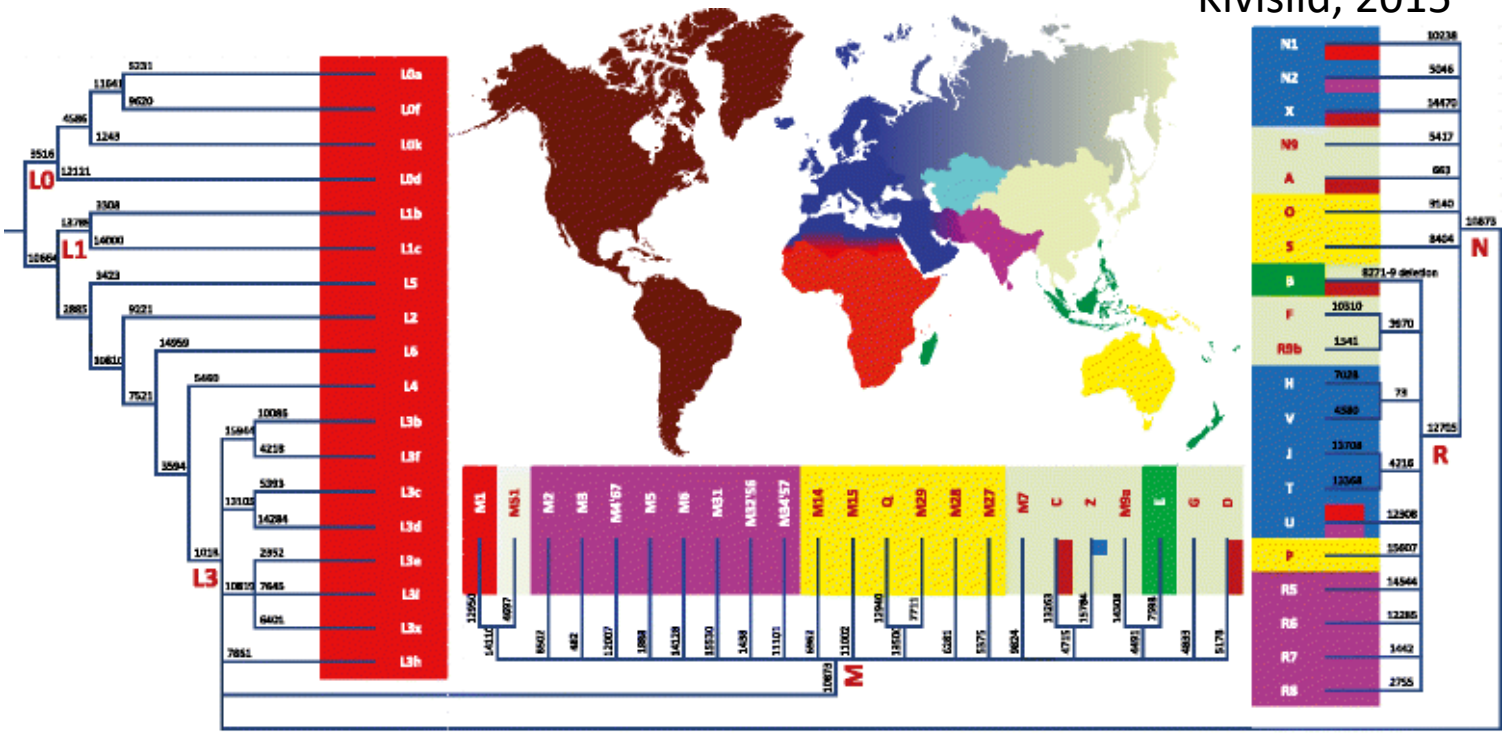
Reconstructing >100,000s years of evolution!

- 22 autosomal chromosomes
- 2 sex chromosomes
  - X chromosome
  - Y chromosome
- Mitrochondrial genome

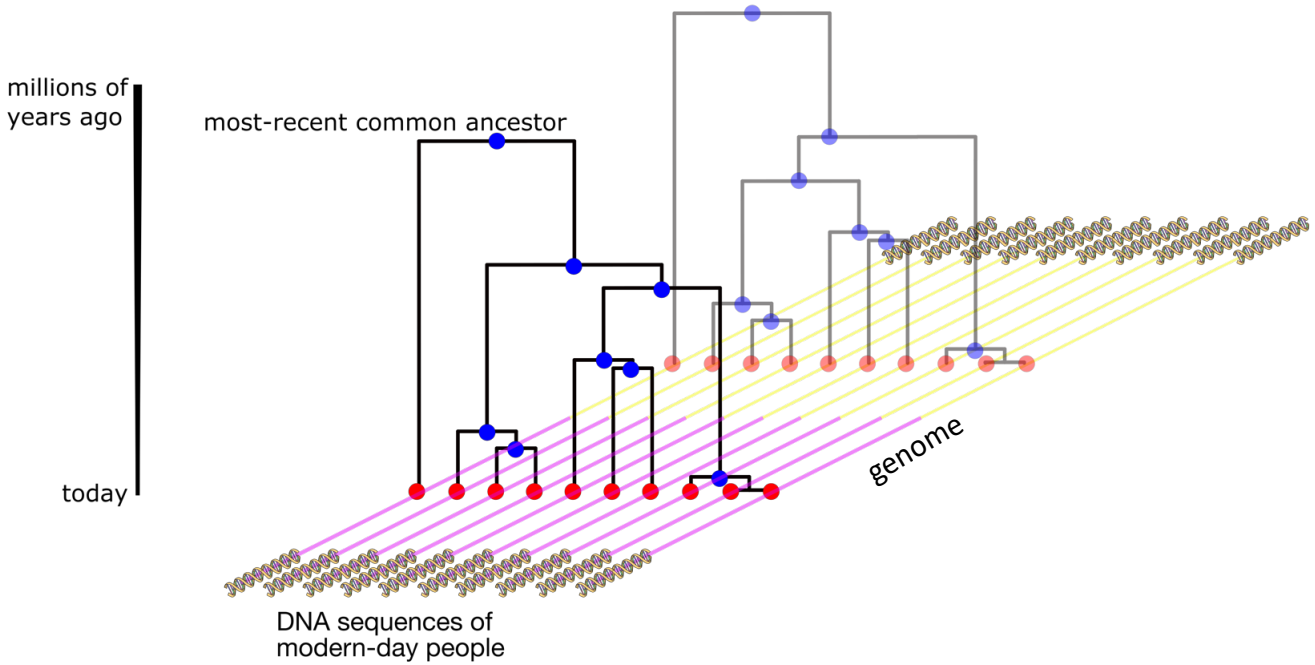
Different trees in different parts of the genome, due to recombination

1 tree each (maternal/paternal)

Kivisild, 2015

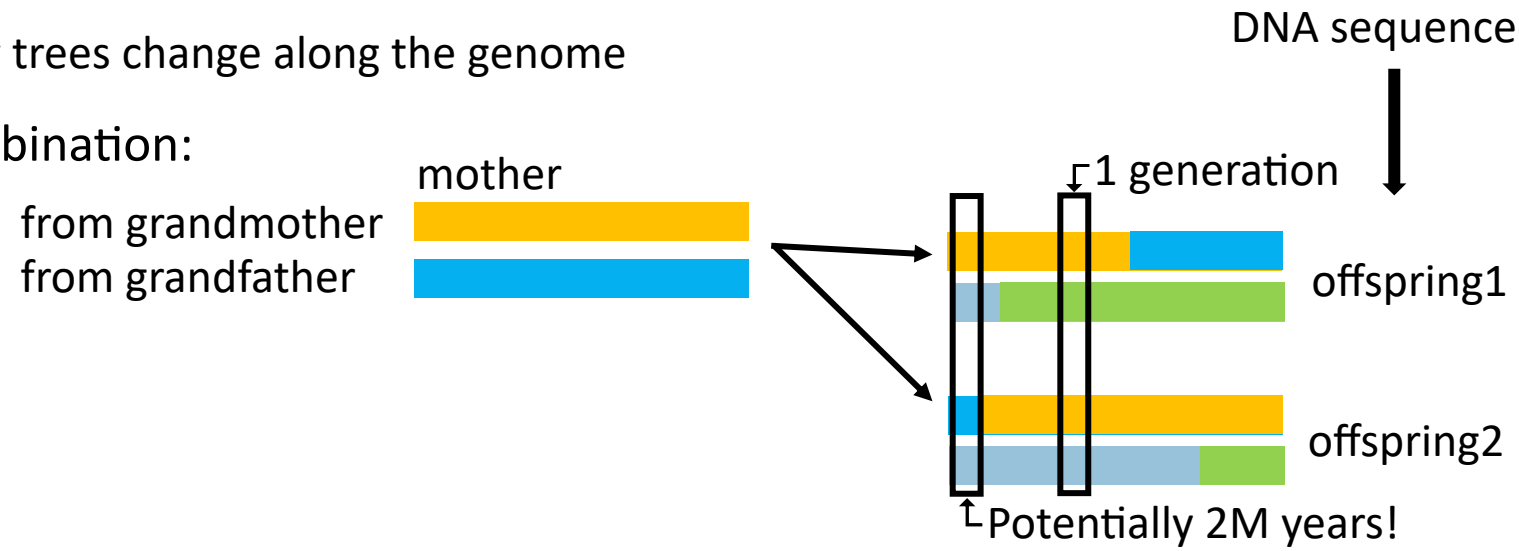


Key concept: Genealogies
We are genetically related through a sequence of trees

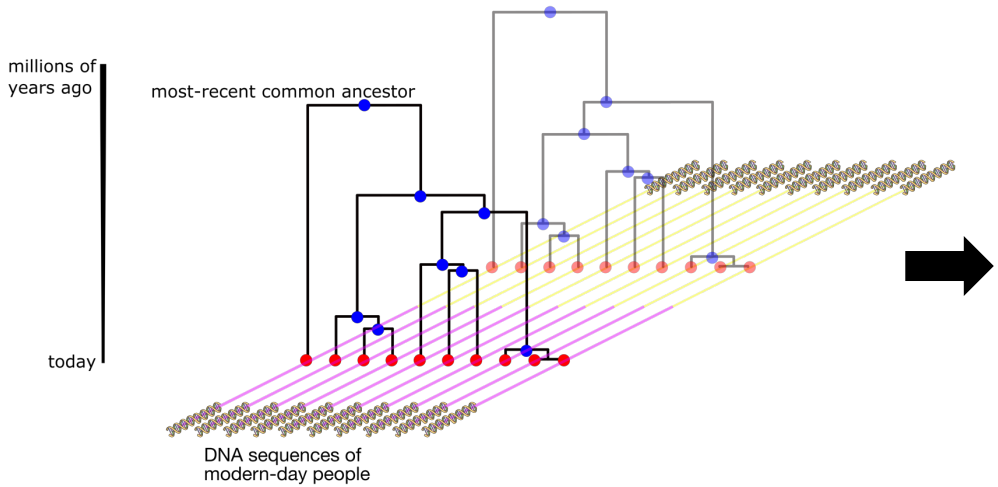


Reason for why trees change along the genome

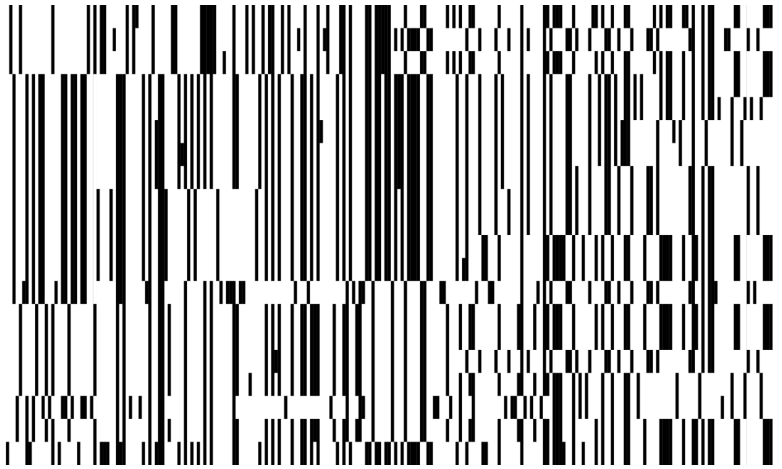
Recombination:



Fundamental forces impact data (only) through underlying genealogies

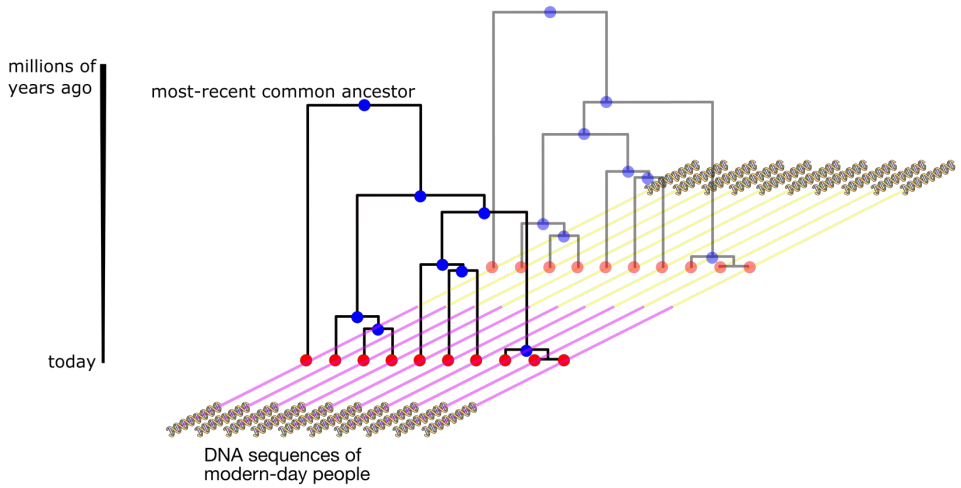


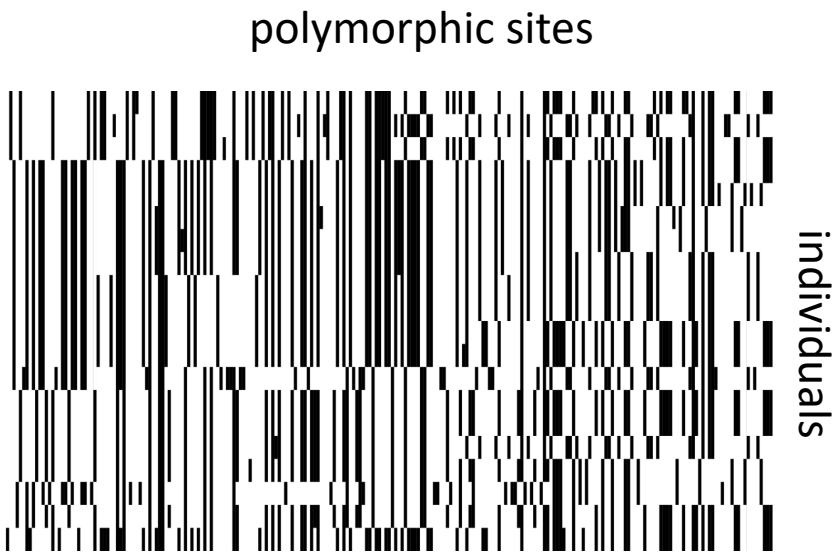
polymorphic sites



Demographic history
Genetic structure
Admixture
Selection, etc.

Many canonical approaches



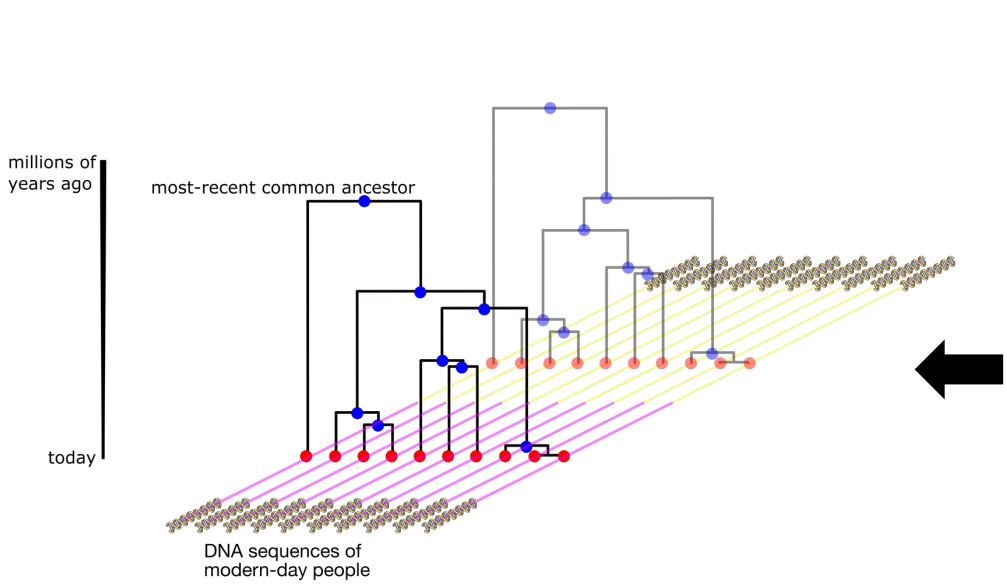


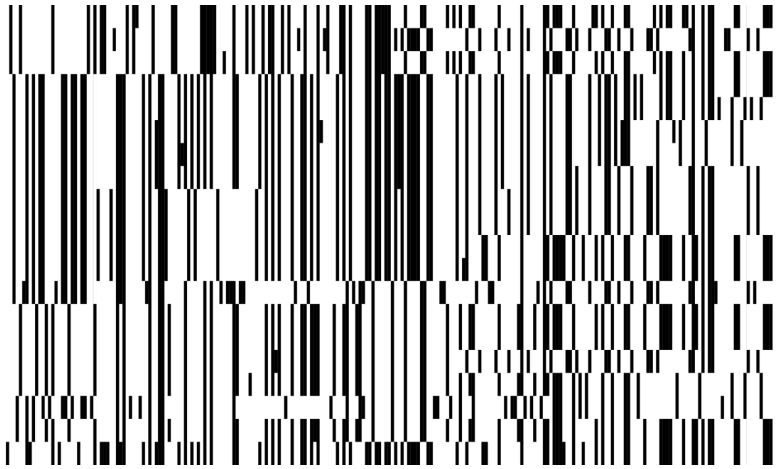
Demographic history
Genetic structure
Admixture
Selection, etc.

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

- Allele frequency-based methods
- Local haplotype sharing
- Mutation sharing (e.g., f statistics)

Genealogies are the “unobserved link” between evolutionary processes and genetic variation





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

- Allele frequency-based methods
- Local haplotype sharing
- Mutation sharing (e.g., f statistics)

Demographic history
Genetic structure
Admixture
Selection, etc.

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

We can now build these trees for many thousands of individuals and add a time dimension onto our genetic data

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

- ARGweaver
- Rent+

} Infers Ancestral Recombination Graphs

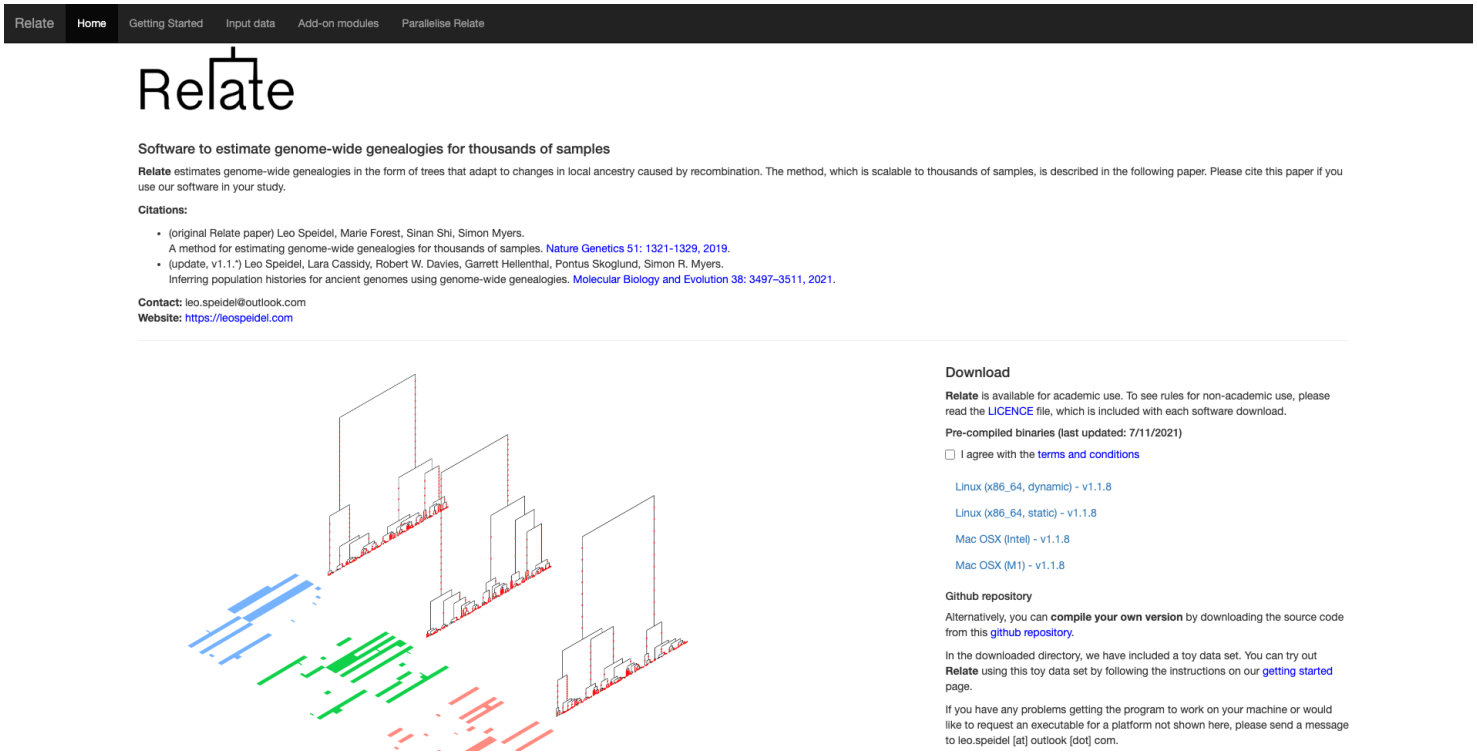
- Tsinfer + tsdate
- ARG-Needle
- Relate

} Published in since 2019,
scale to large sample sizes

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

Relate

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

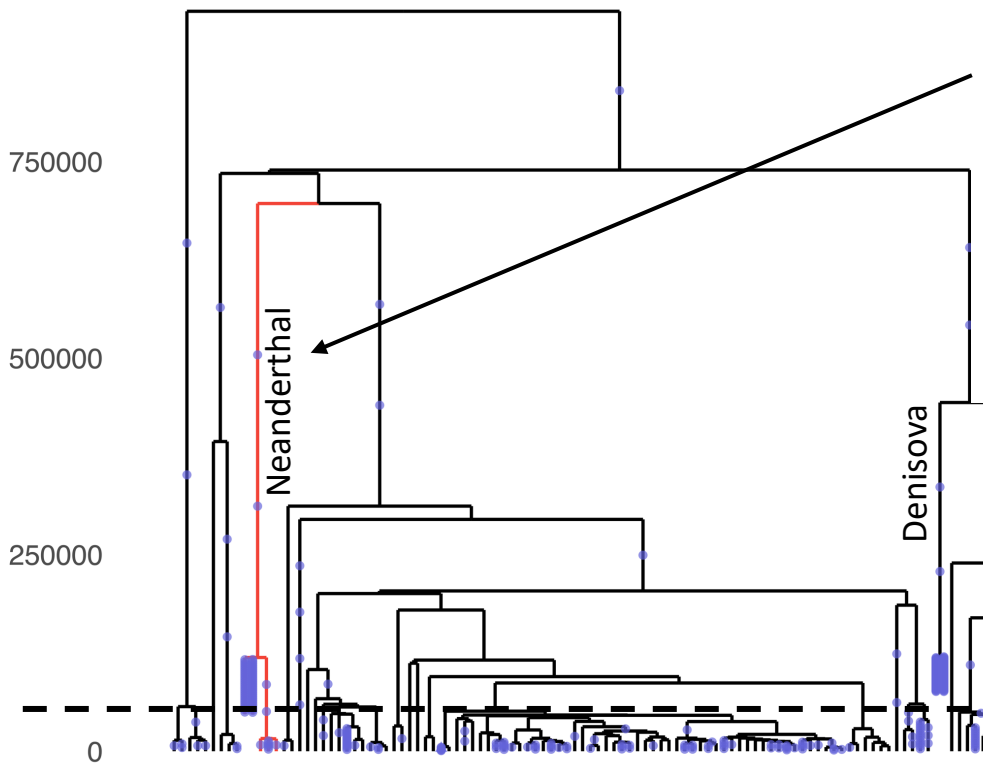


https://myersgroup.github.io/relate/

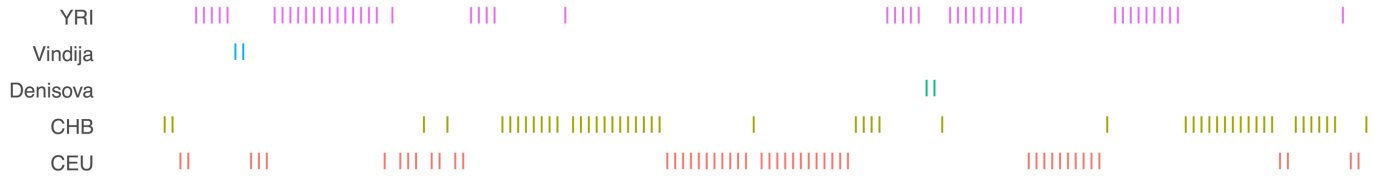
Key features:

- 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



~60k years ago



Article

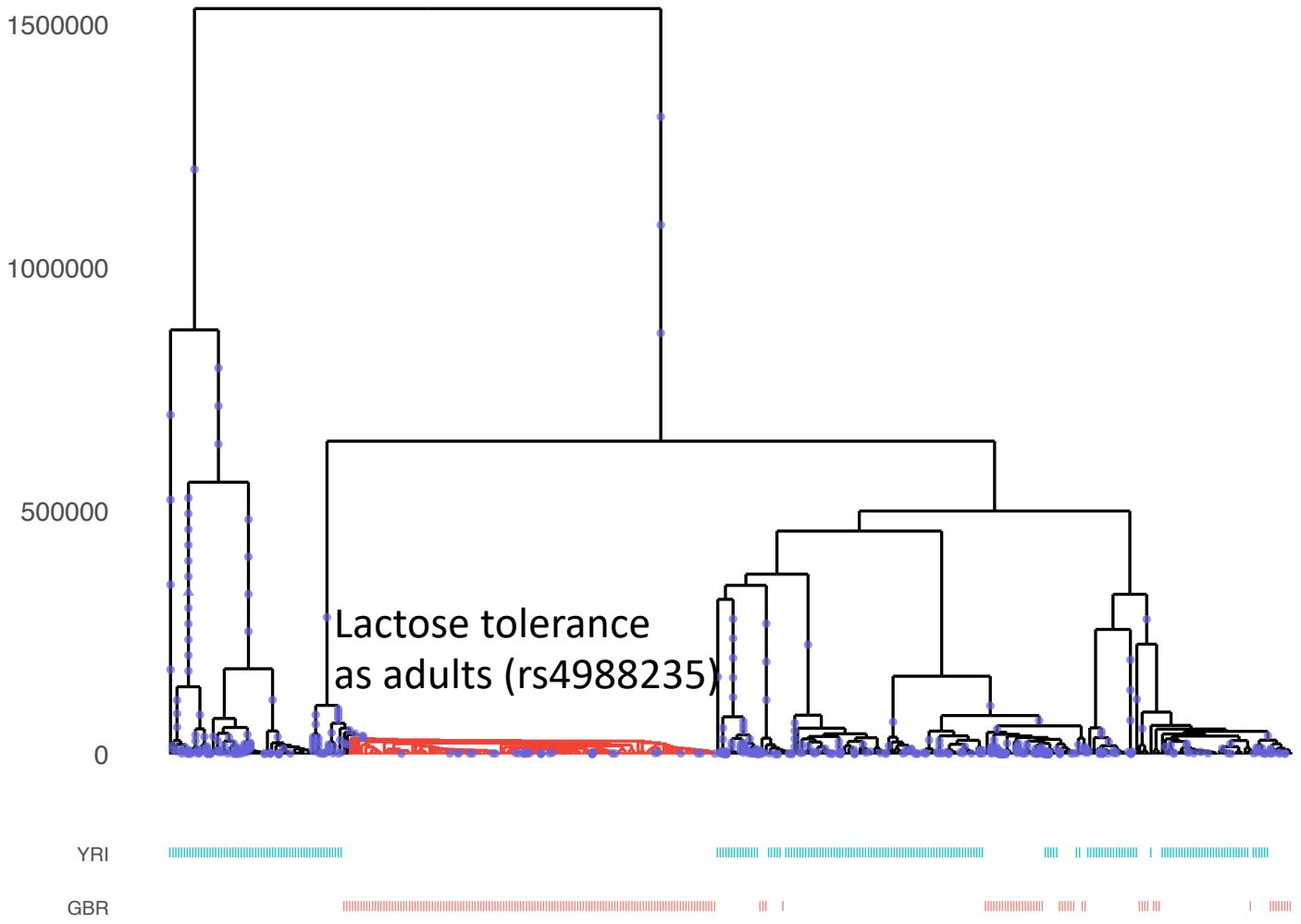
The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

https://doi.org/10.1038/s41586-020-2818-3
Received: 3 July 2020
Accepted: 22 September 2020
Published online: 30 September 2020
Check for updates

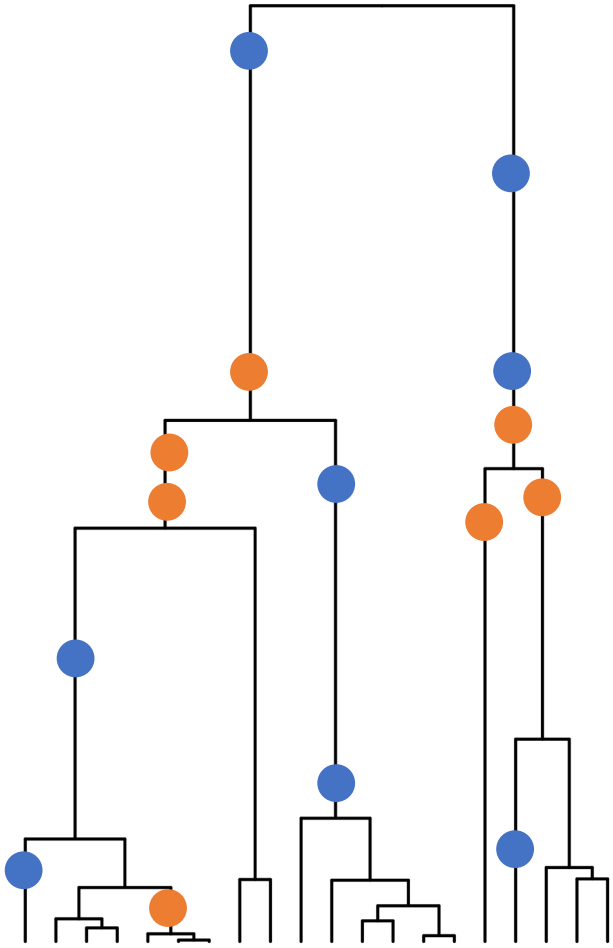
Hugo Zeberg[1,2] & Svante Pääbo[1,3]

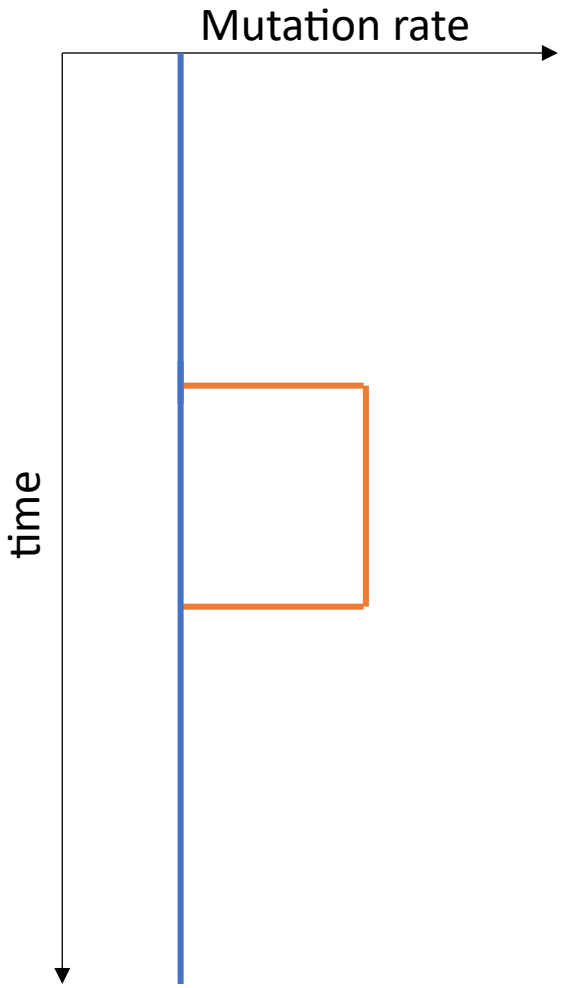
A recent genetic association study[1] identified a gene cluster on chromosome 3 as a risk locus for respiratory failure after infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A separate study (COVID-19 Host Genetics Initiative)[2] comprising 3,199 hospitalized patients with coronavirus disease 2019 (COVID-19) and control individuals showed that this cluster is the major genetic risk factor for severe symptoms after SARS-CoV-2 infection and hospitalization. Here we show that the risk is conferred by a genomic segment of around 50 kilobases in size that is inherited from Neanderthals and is carried by around 50% of people in south Asia and around 16% of people in Europe.

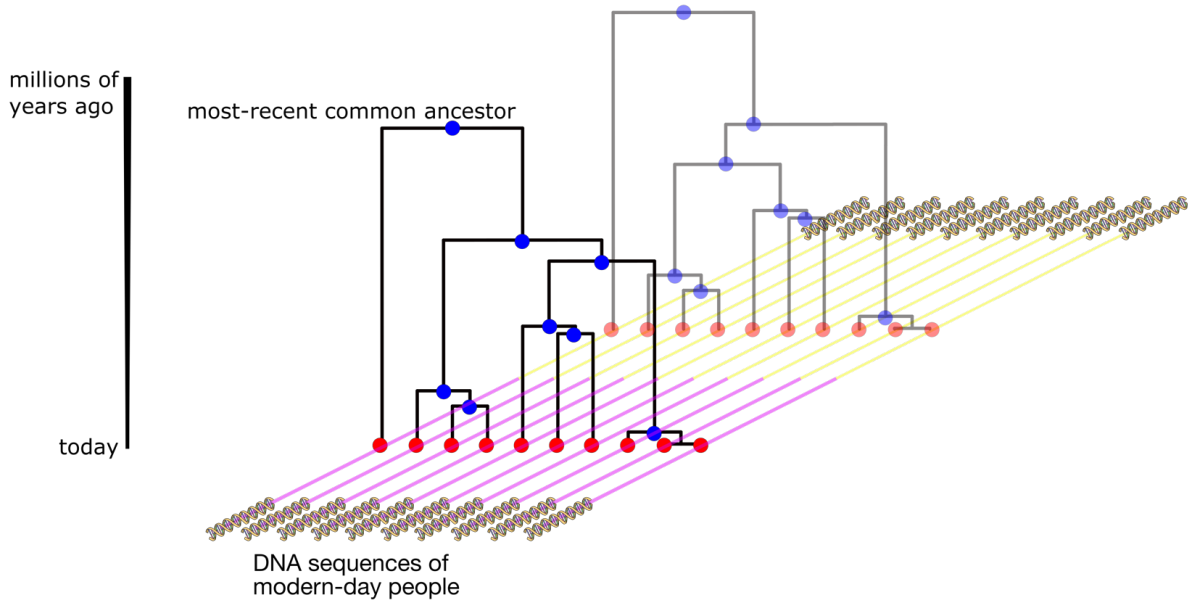
Positive selection: rapidly spreading lineage



Clusters of mutations in time can capture changes in mutation rate

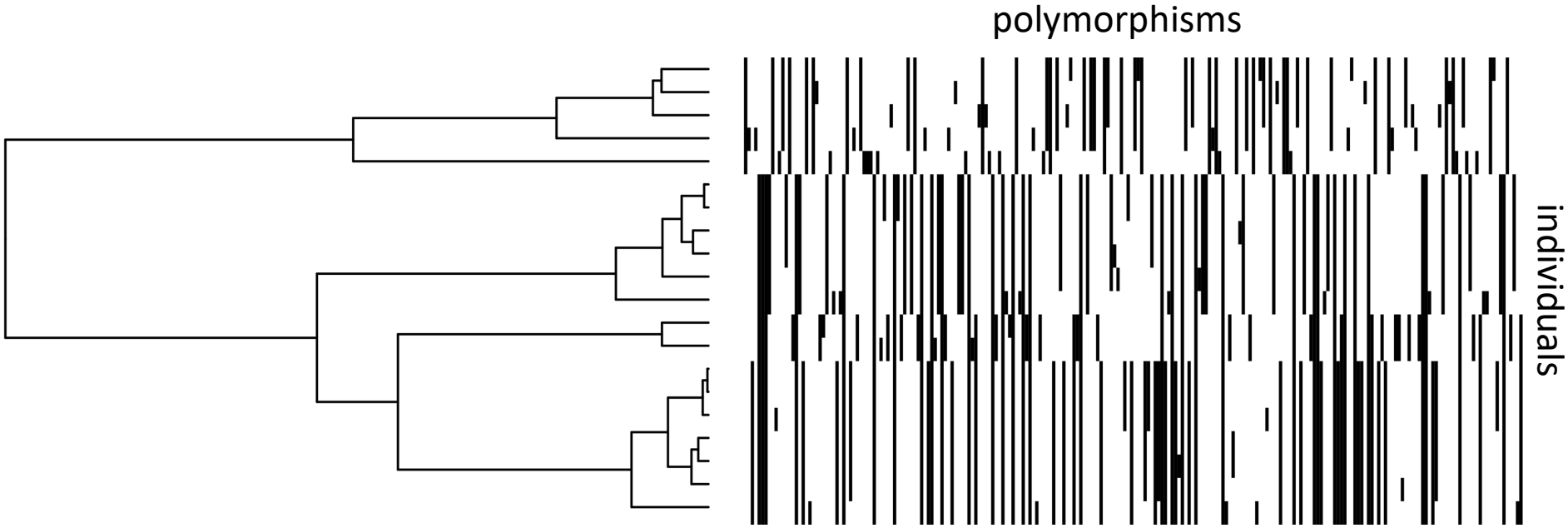






Inferring genealogies from genetic variation

Data and the underlying tree structure

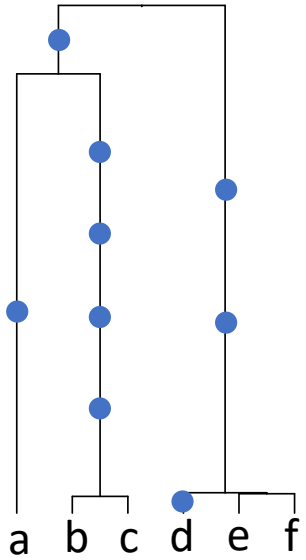


- **Every mutation shows the existence of a branch**
- Mutations are “ordered by inclusion”
- No two branches (mutations) ever show only partial overlap

Count derived mutations to reconstruct tree topology

For tree topology, we want to quantify the order in which we are related to others

1. Count number of "derived mutations" to get order of coalescences
   - E.g., sequence a has 1 derived mutation to (b,c)
     2 derived mutation to (d,e,f)
2. Coalesce mutually closest lineages

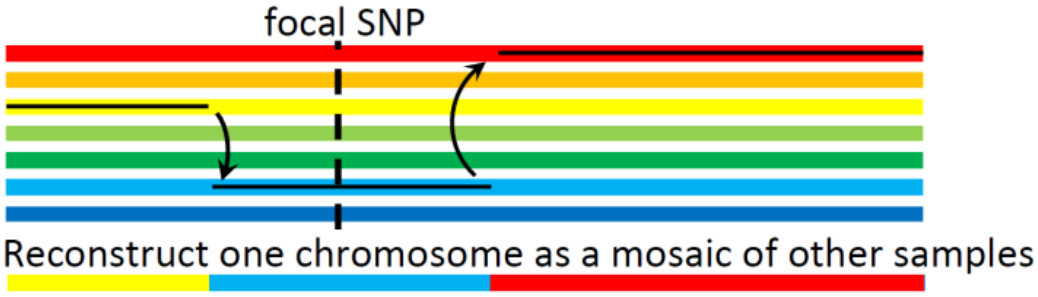


- No recombination: Guaranteed to build tree consistent with data
- This is not the case if we use "pairwise differences" (UPGMA)
  - a and (d,e,f) are closer than a and (b,c)

- Use **chromosome painting** to count derived mutations accounting for recombination

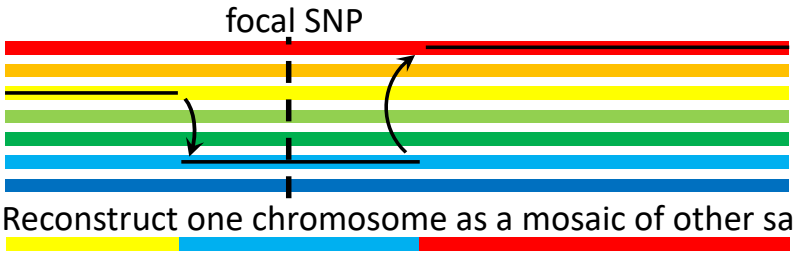
Hidden Markov model (HMM)

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

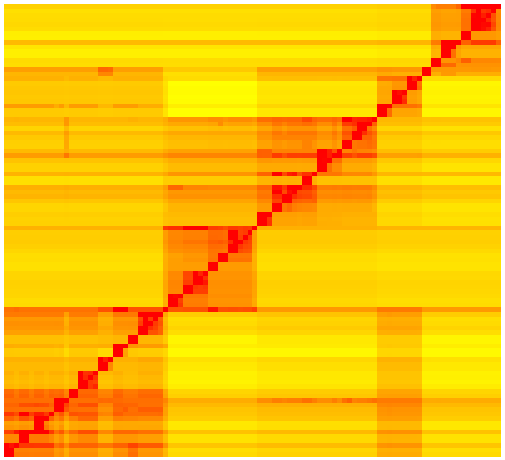


Summary of Relate

Hidden Markov model (HMM)

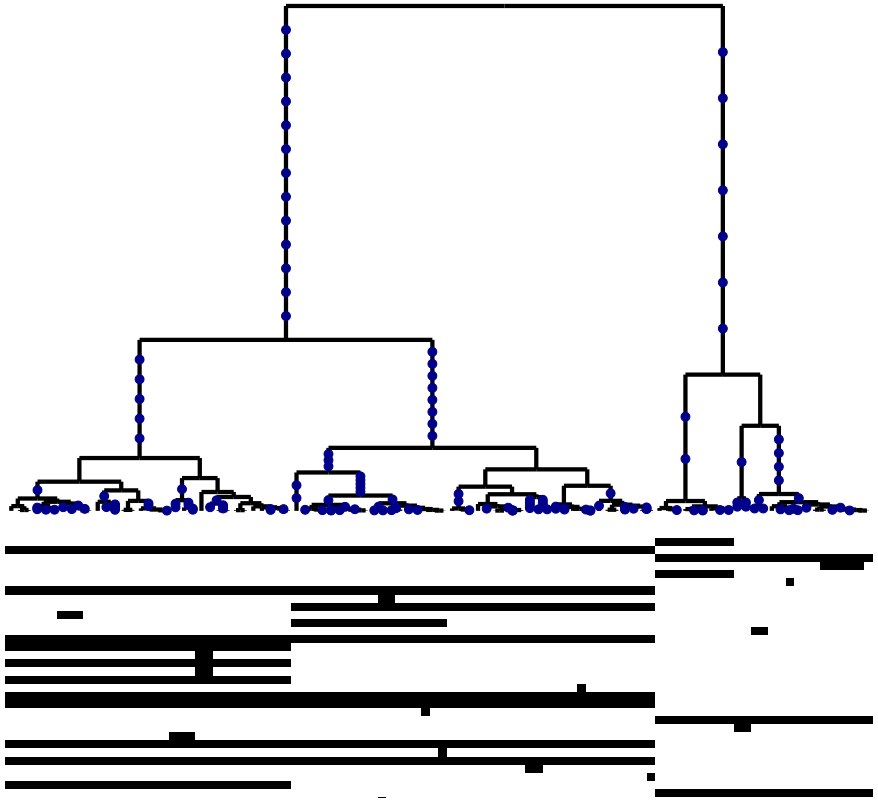


Distance matrix for focal SNP



"number of derived mutations"

Hierarchical clustering
&
MCMC for branch lengths

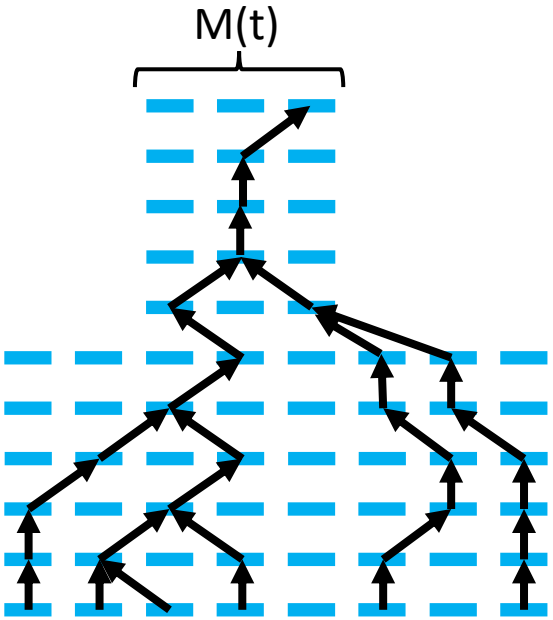


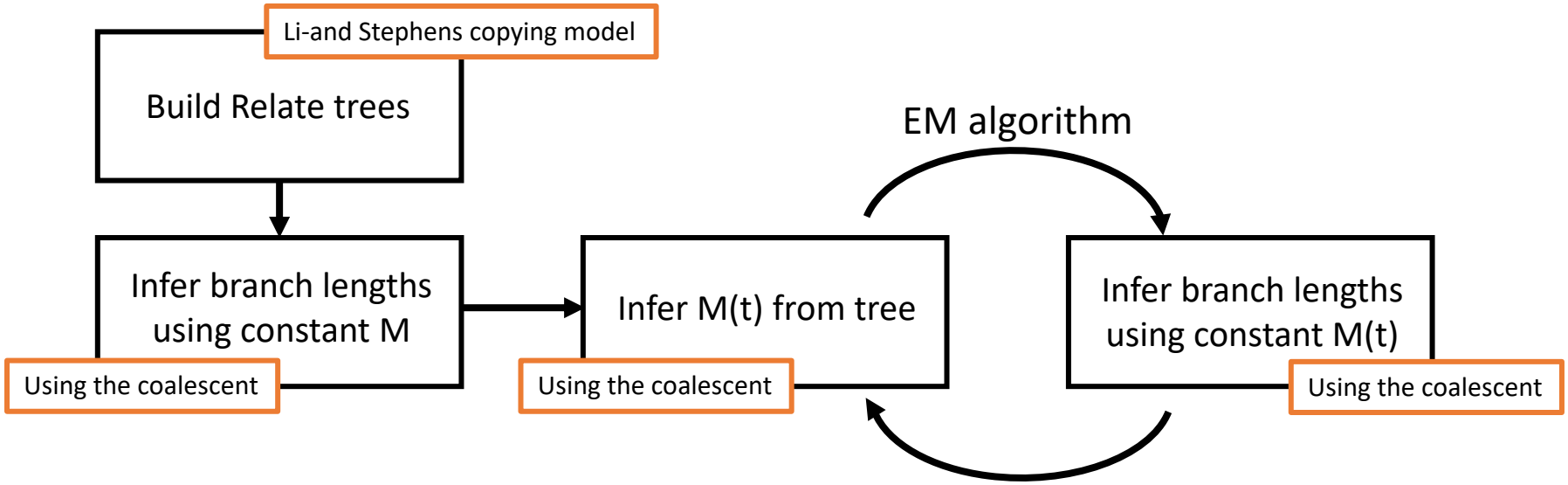
haplotype data sorted using constructed tree

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

- Expected branch lengths depend on population size M(t)

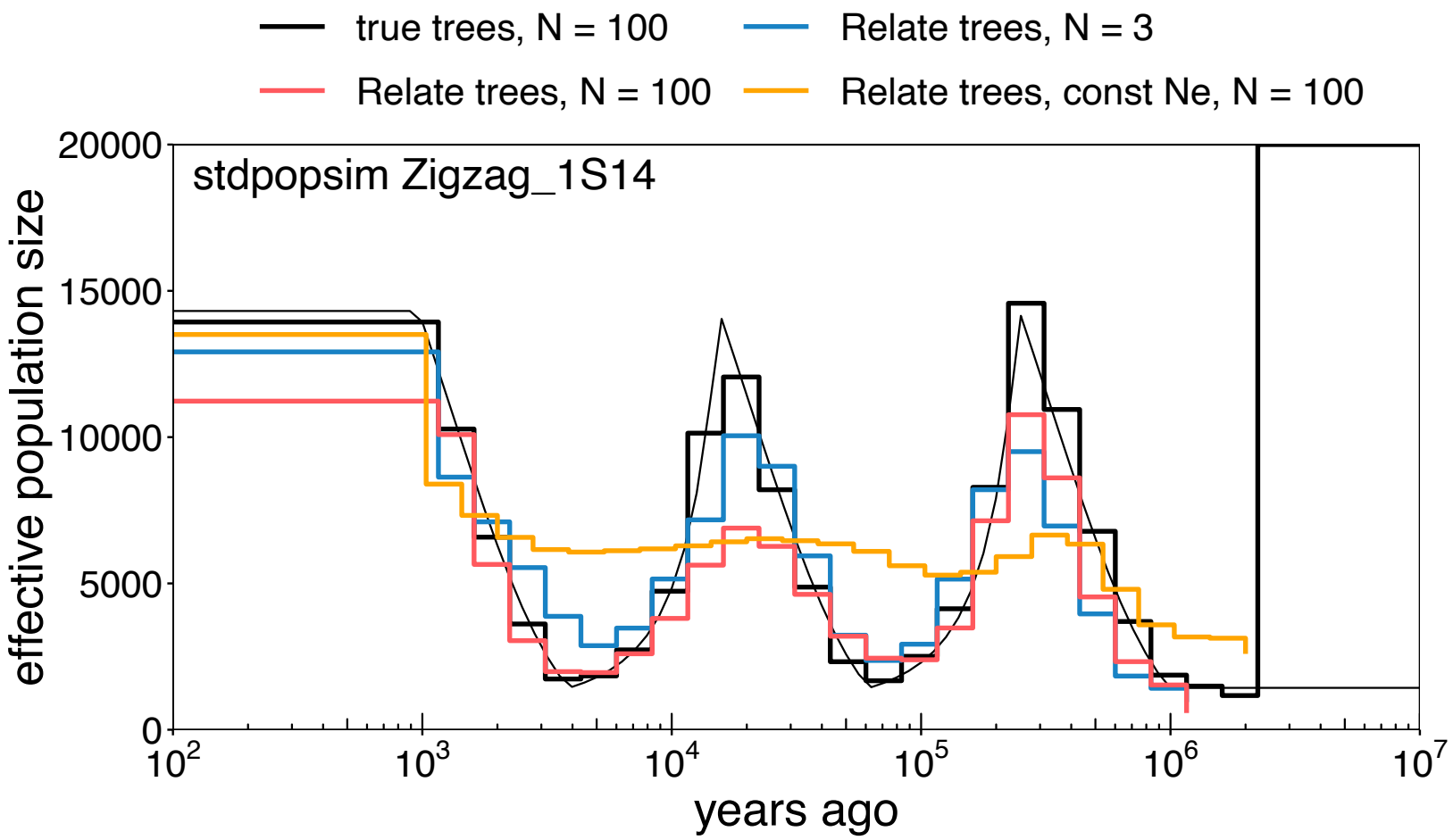
  (or coalescence rates 1/M(t))
- While there are *j* lineages, the rate at which a coalescence happens is $\binom{j}{2}/M(t)$ a time *t* ago
- Demography is shared genome-wide, so we average across trees
- So within a time interval, scaled fraction of trees where coalescence occurs is inversely proportional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





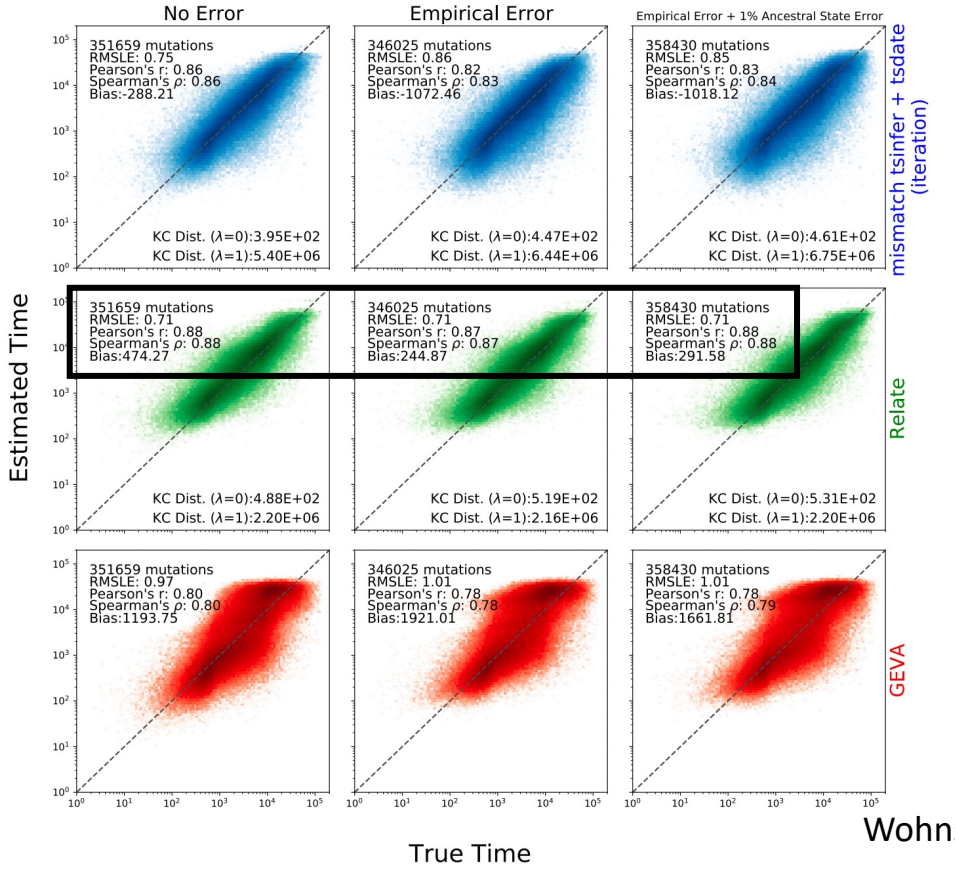
Population size changes through time are jointly inferred in Relate

- 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



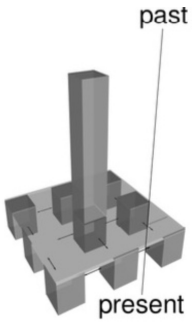
Wohns et al. Science 2022

Genealogy-based inference of human evolutionary history

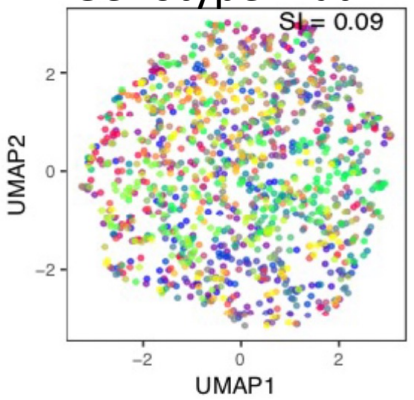
One reconstruction of history, many applications that are self consistent

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

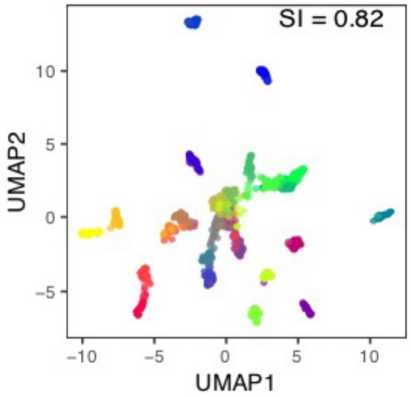
Fan, Mancuso, Chiang, AJHG, 2022



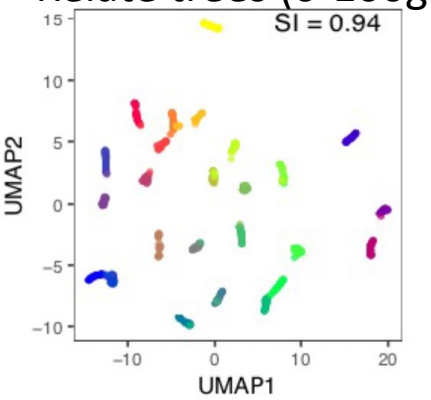
Genotype matrix

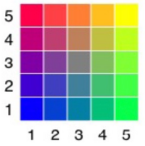


Relate trees

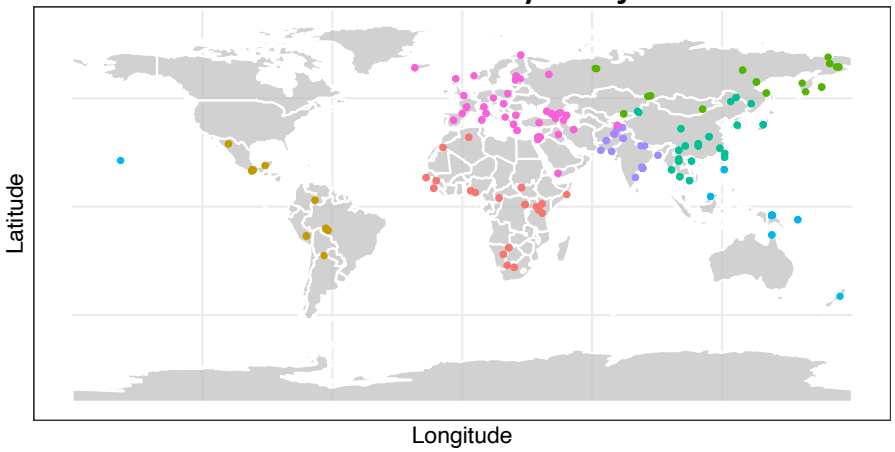


Relate trees (0-100gen)

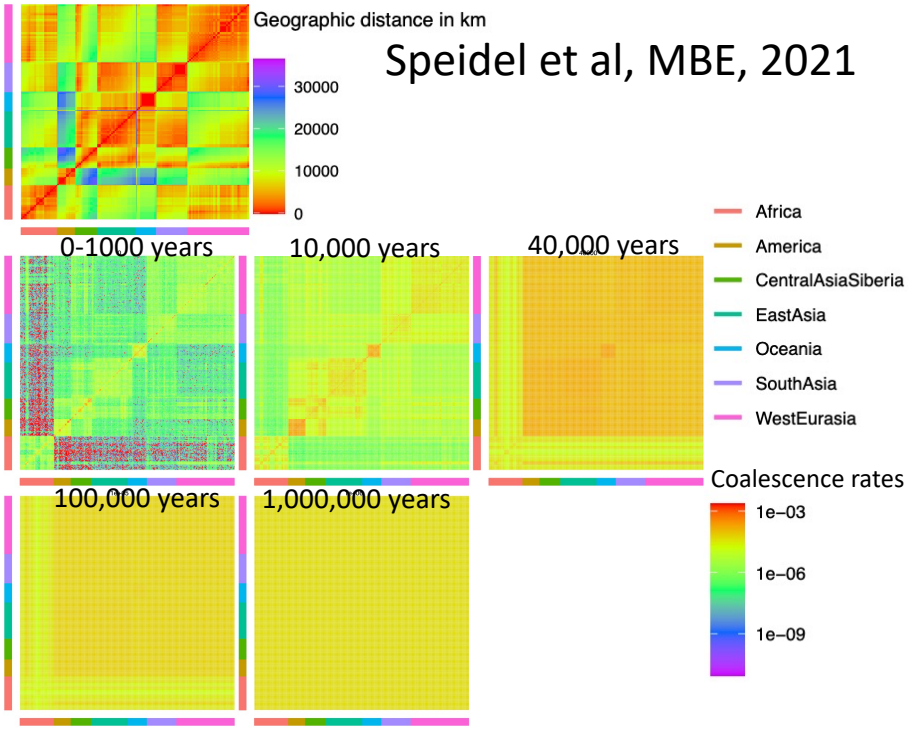




Simons Genome Diversity Project

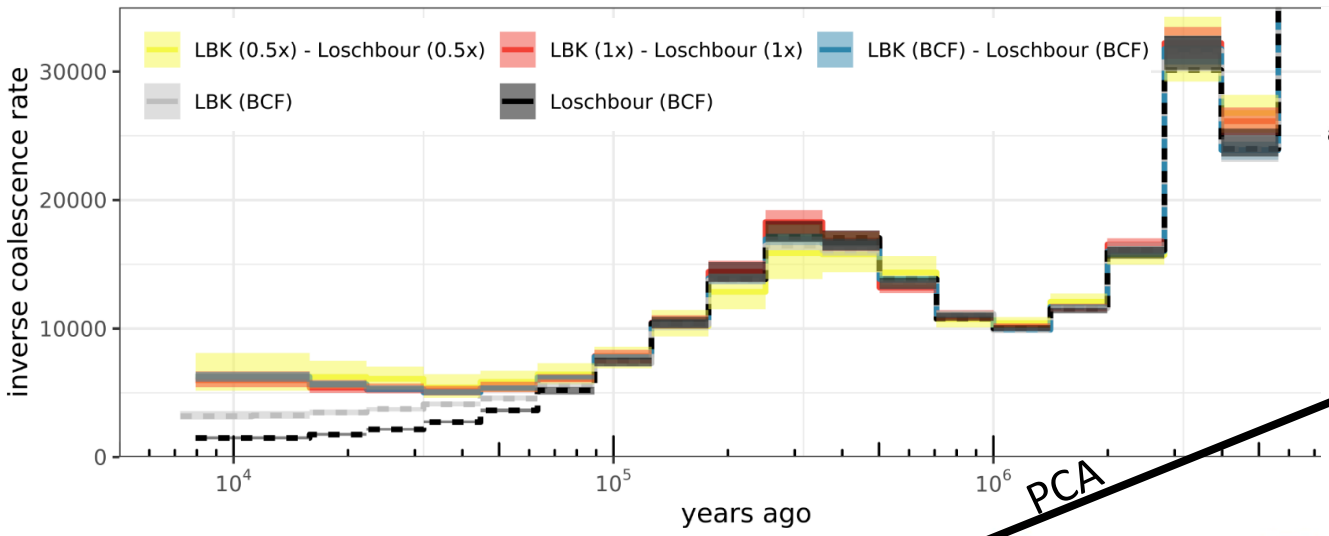


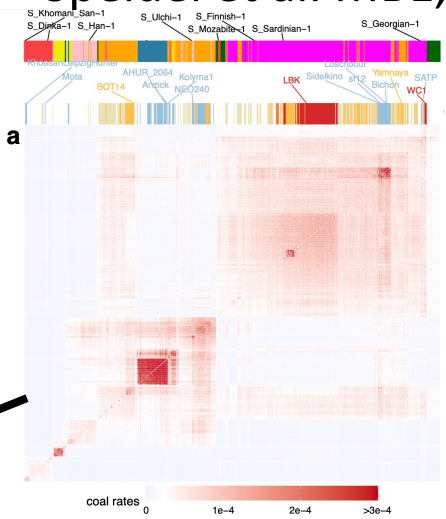
Speidel et al, MBE, 2021



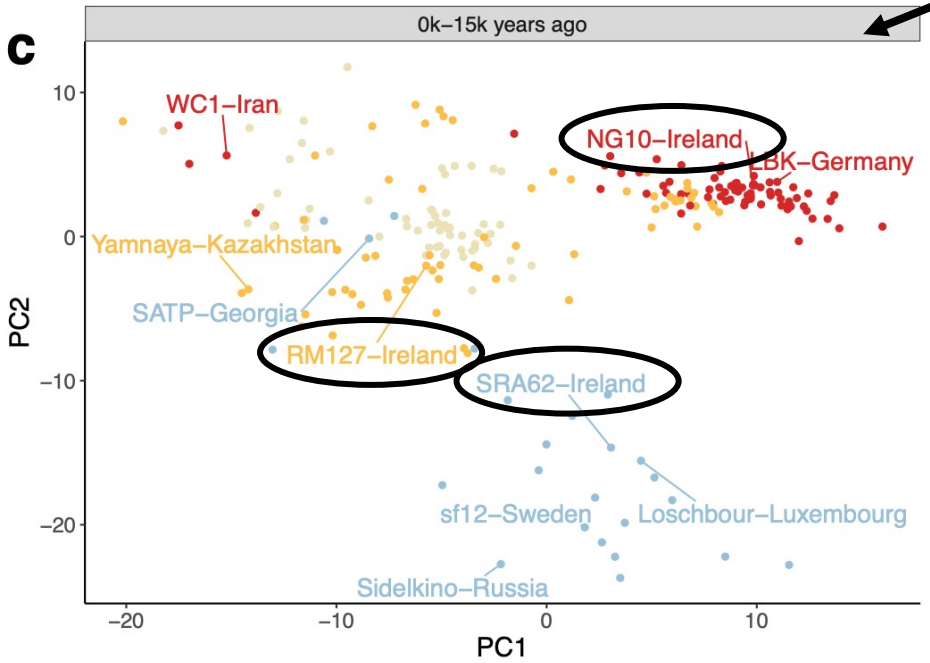
Colate: Inferring coalescence rates for low-coverage, unphased (ancient) genomes

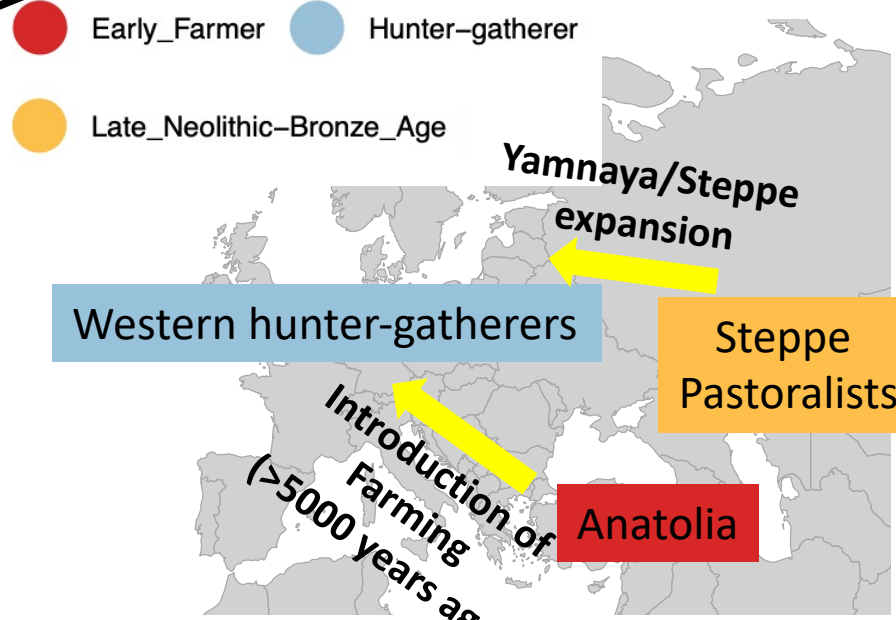
Speidel et al. MBE, 2021



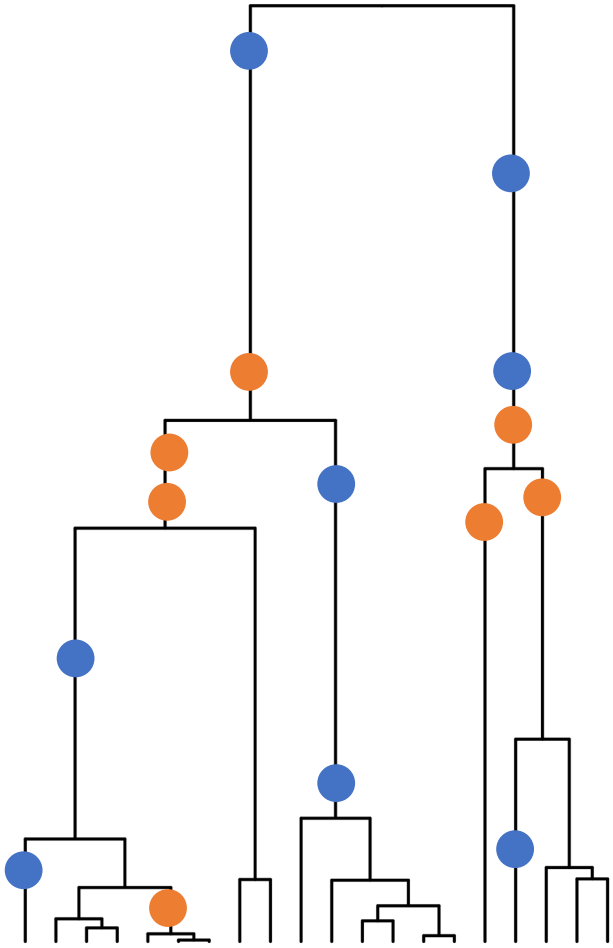


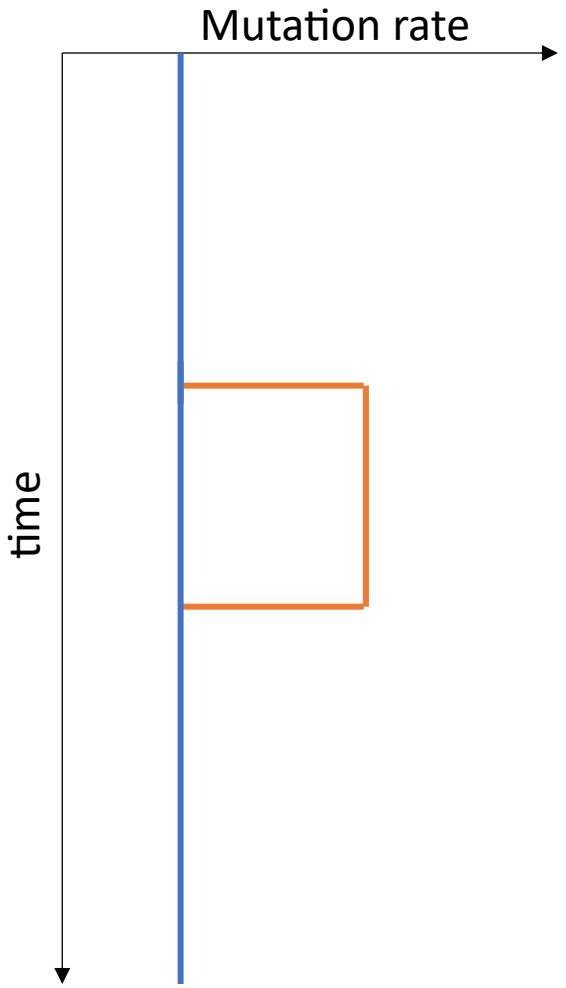
PCA





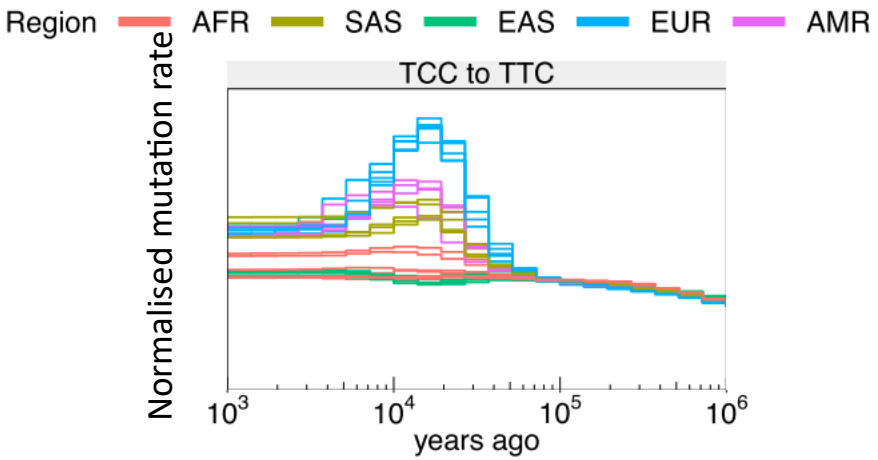
Clusters of mutations in time can capture changes in mutation rate

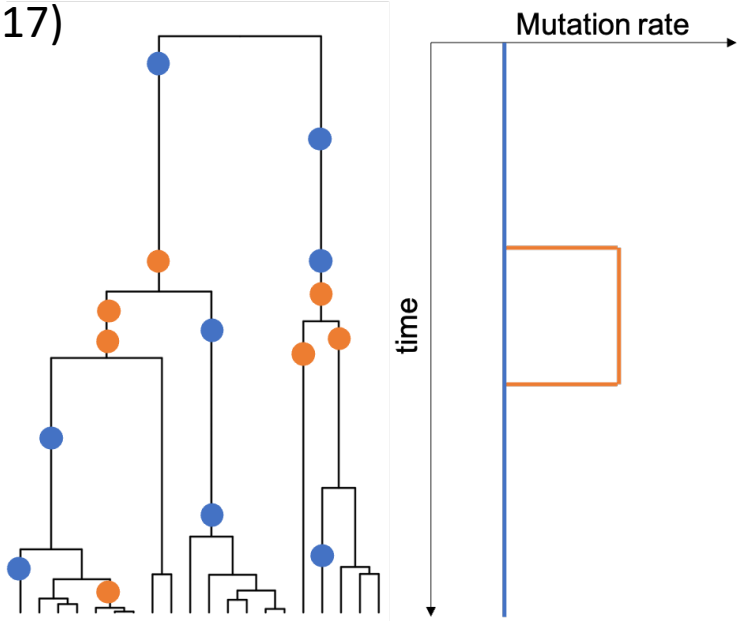




TCC/TTC mutation rates have experienced a strong pulse in the Upper Paleolithic

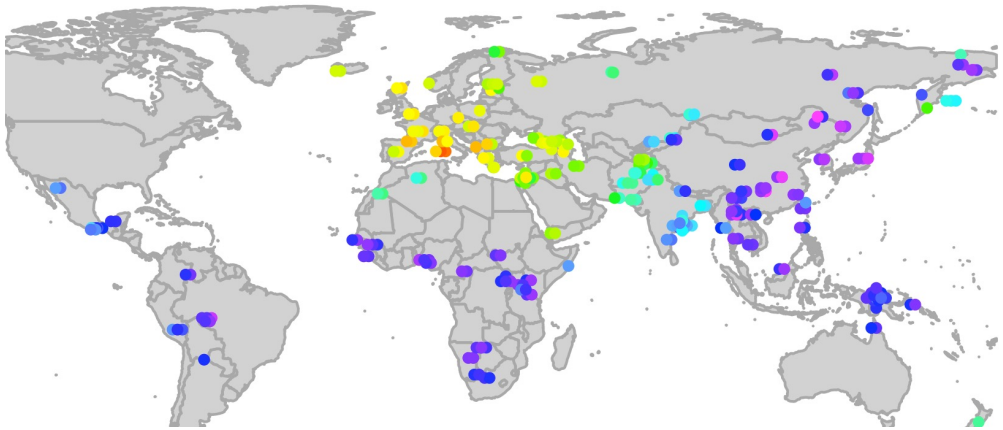
- First reported by Kelley Harris (PNAS 2015, eLife 2017)
- Unknown cause (genetic?, environmental?)
- Previously mainly studied in modern groups



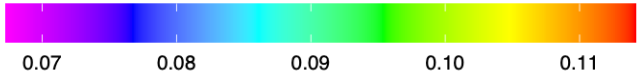


Speidel, Nature Genetics, 2019

How did this spread to all West Eurasians today?

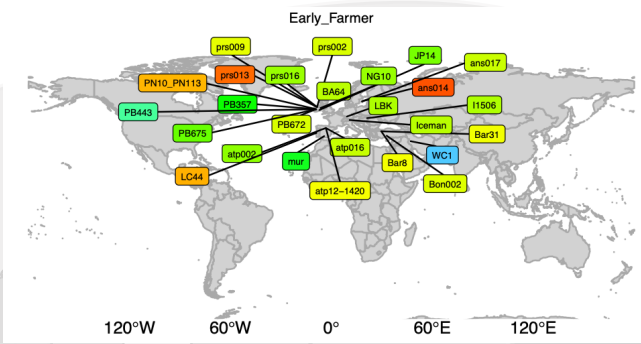


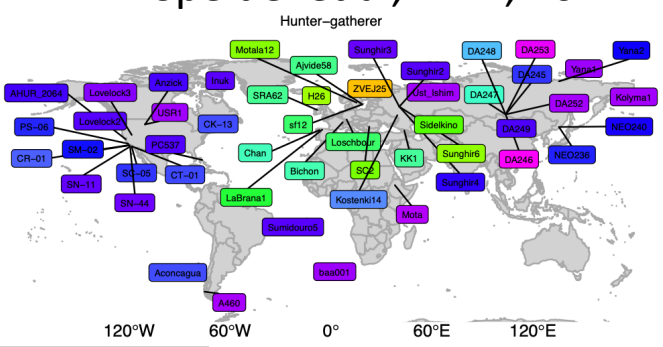
Colour shows strength of elevation in TCC/TTC mutation rates



TCC/TTC mutation rate increase happened and spread before farming, >15k years ago!

Speidel et al, MBE, 2021





Ice Age Europe

>20,000 years ago

ICE

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

earlier HG groups

e.g., Sidelkino (**11k year old**)

Eastern HG

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

Western HG

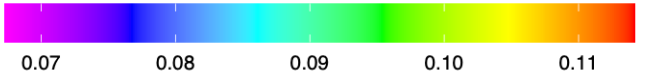
Caucasus HG

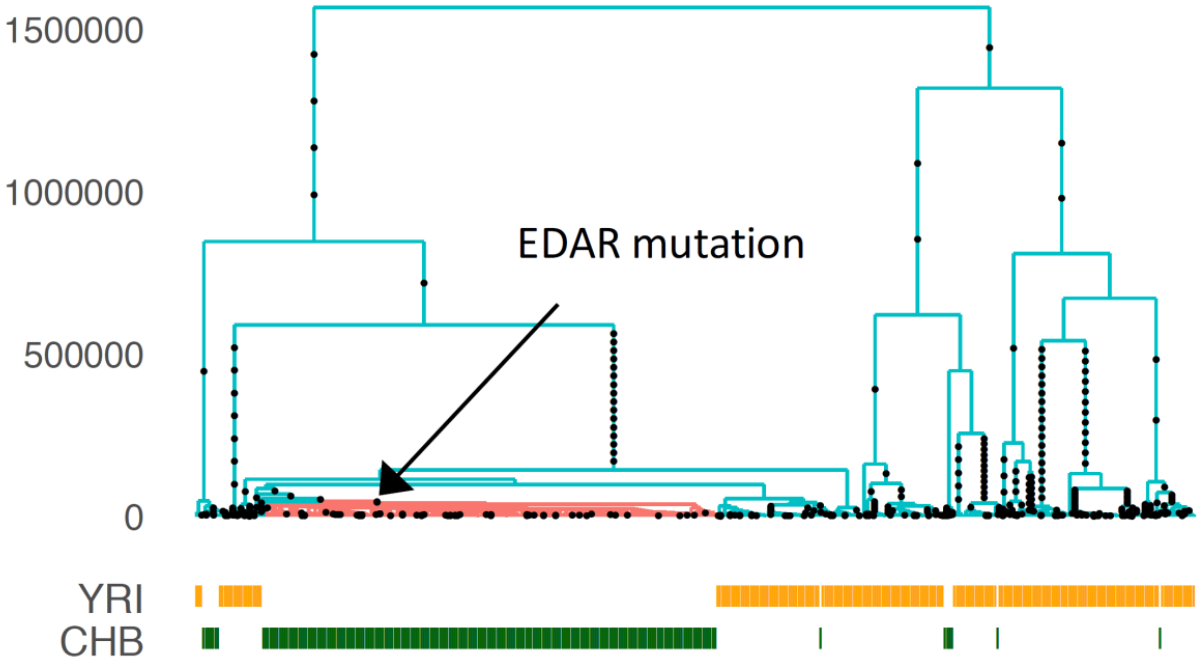
e.g., KK1 (**9k year old**)

Anatolia

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

Colour shows strength of elevation in TCC/TTC mutation rates

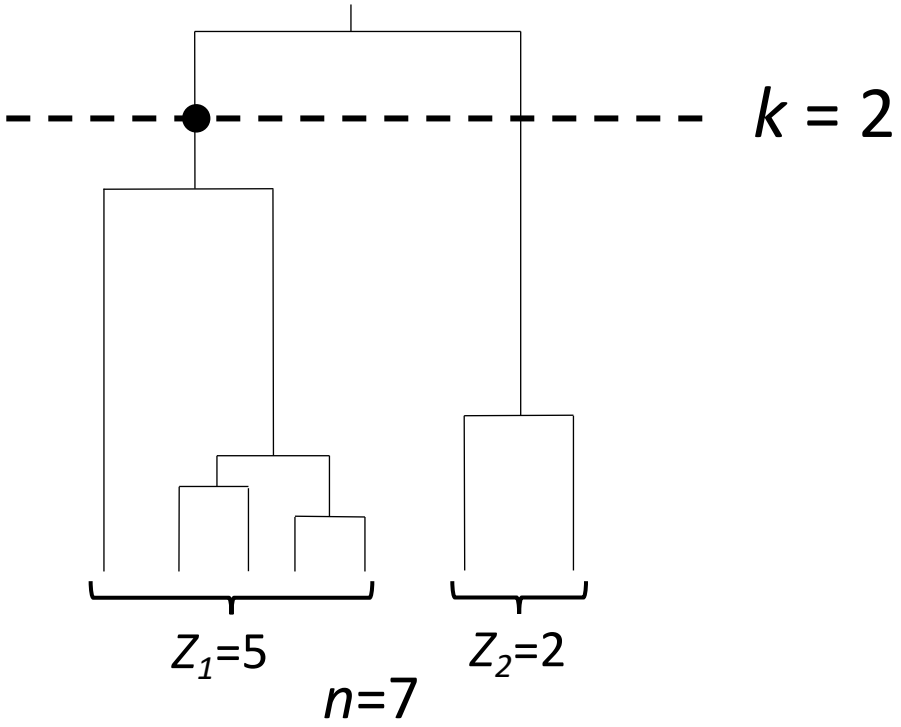




Quantifying positive natural selection on a single mutation

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

How quickly does a mutation spread in the neutral case?



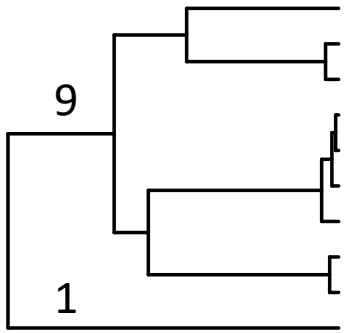
We can write down the analytical distribution for the number of descendants of a mutation arising while k lineages remain

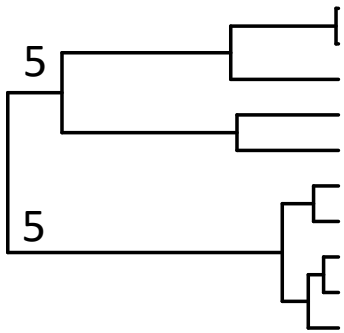
Example: if *k*=2, this is just a **uniform distribution**

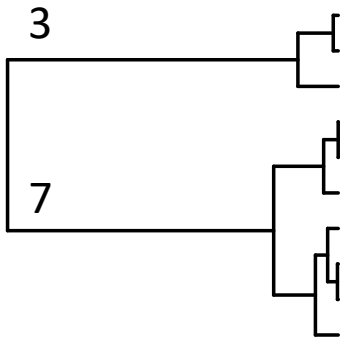
P(5 descendants)=1/6

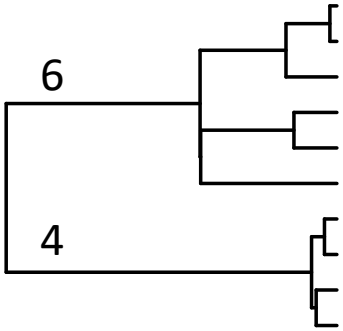
The k = 2 case

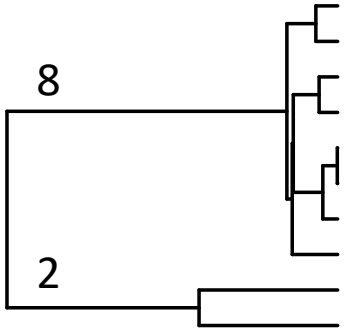
We expect “unbalanced” tree shapes!

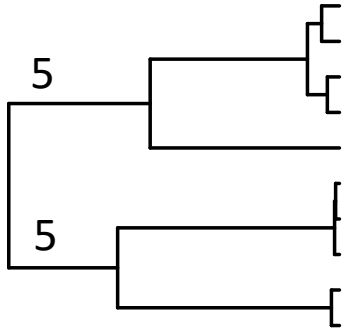


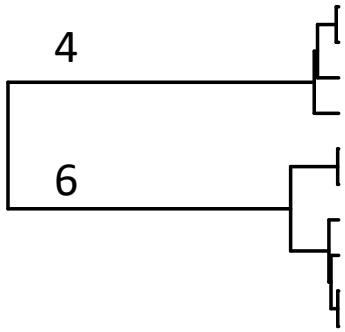


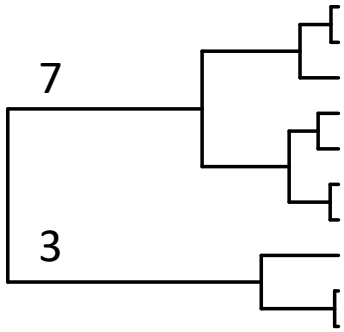


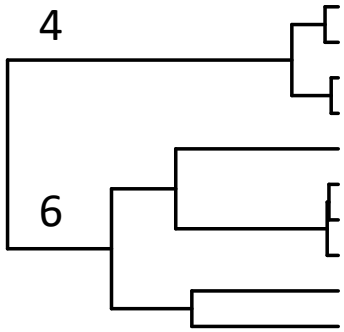










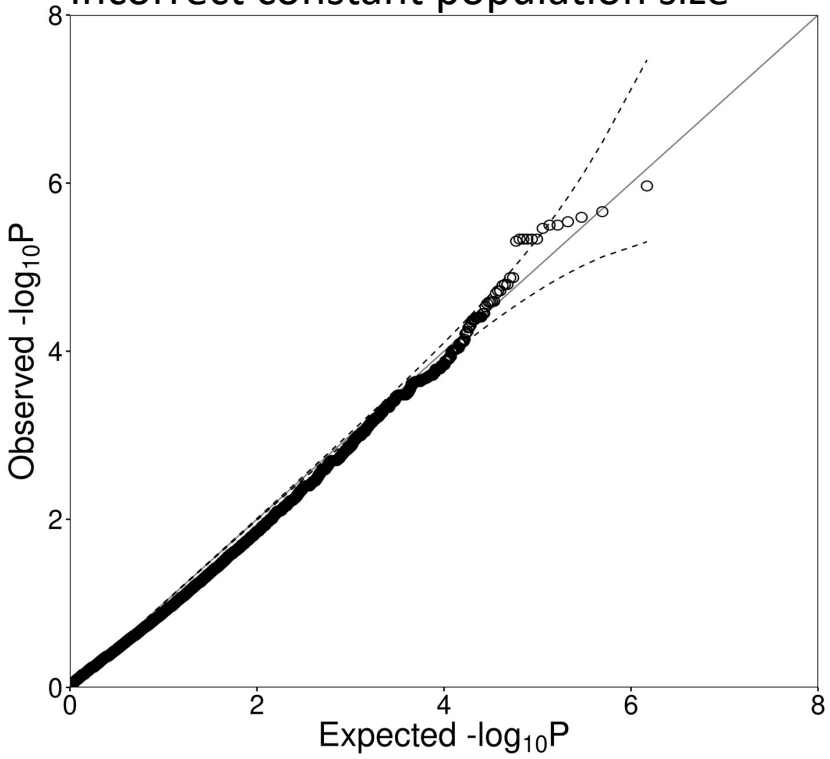


P-values: very well calibrated under null simulations of no selection

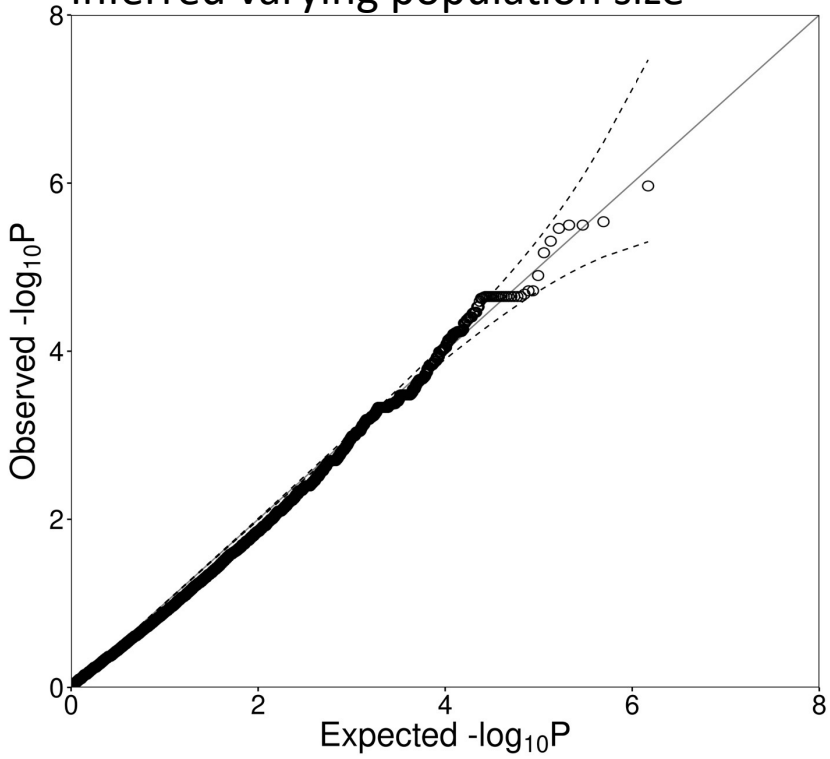
N=1000, 250Mb
Bottleneck population size

Quantile-quantile plot:

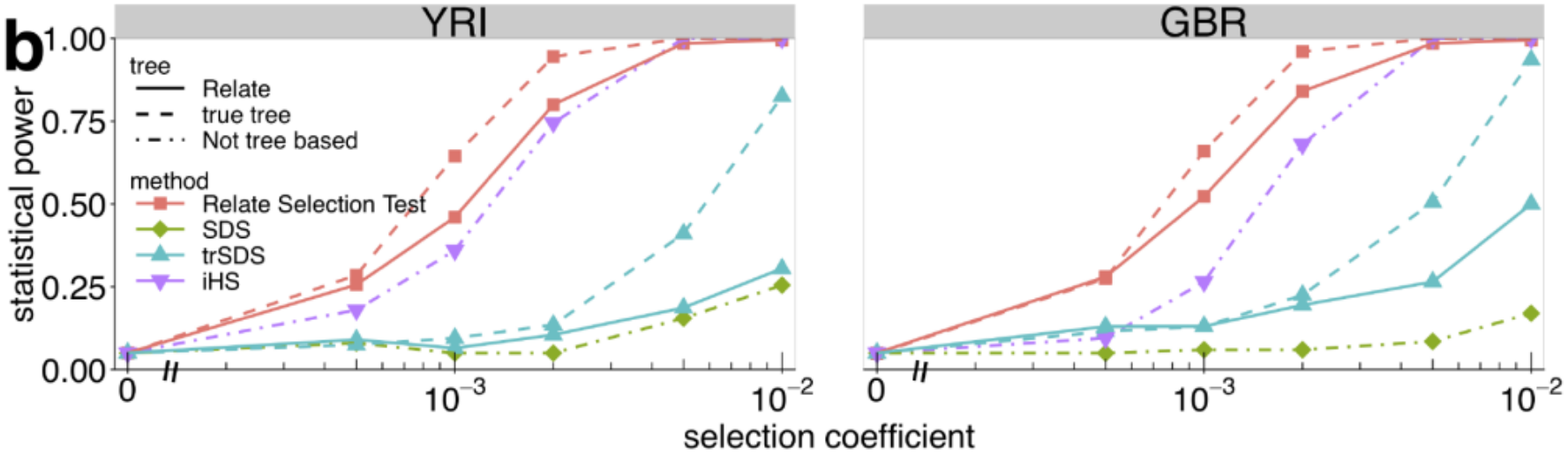
Incorrect constant population size



Inferred varying population size

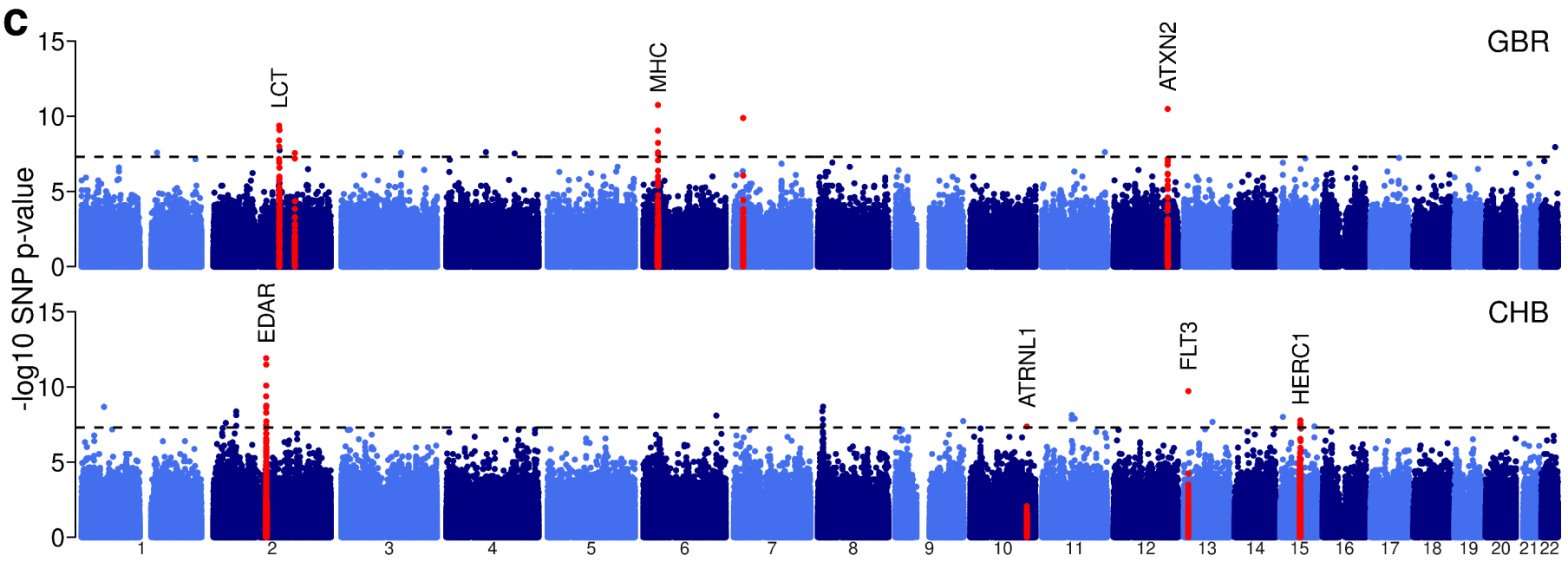


Improved power to see weak selection



Genome-wide selection p-values

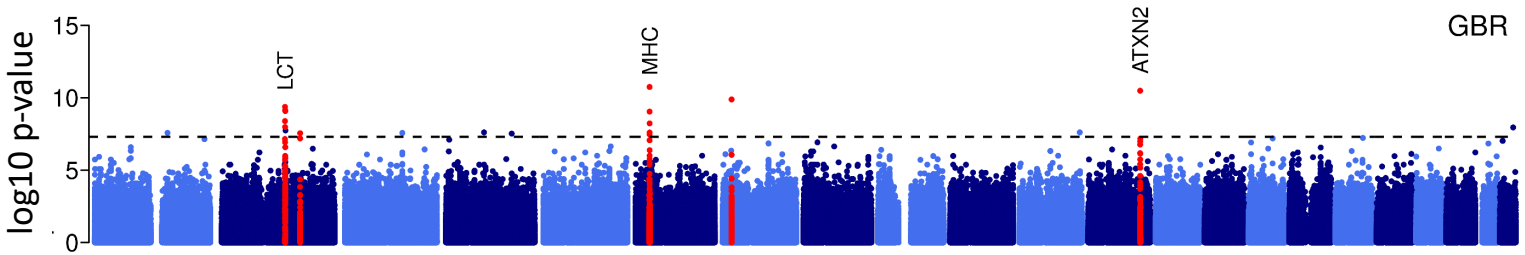
Given most traits are highly polygenic, expect mainly weak, polygenic selection



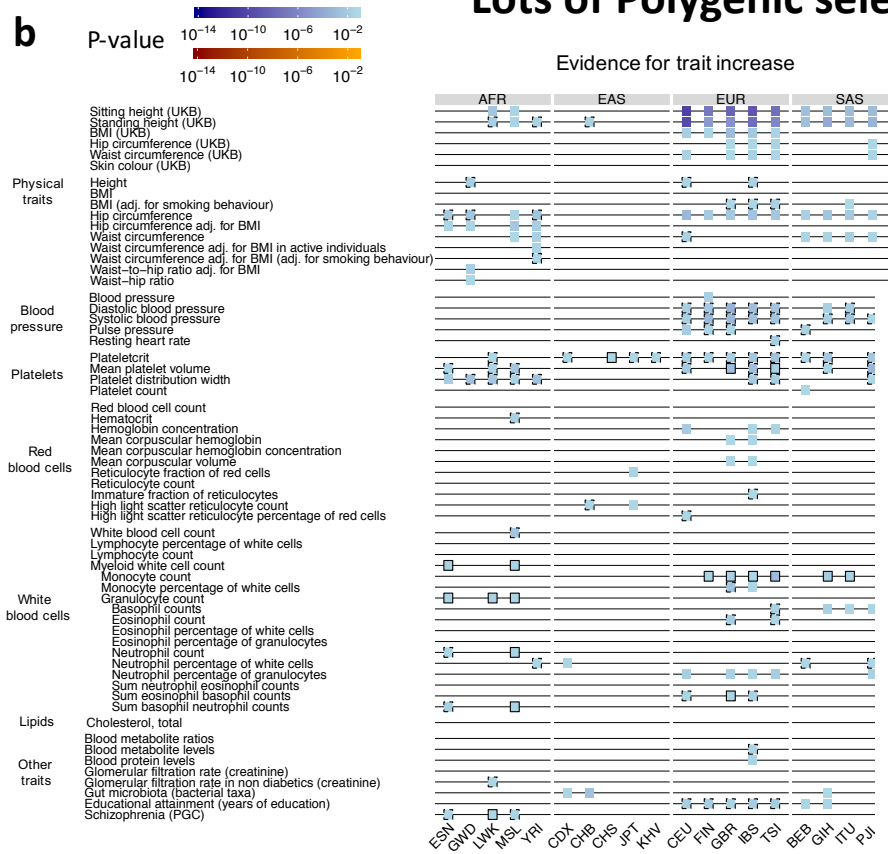
How does weak selection evidence vary by trait?

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

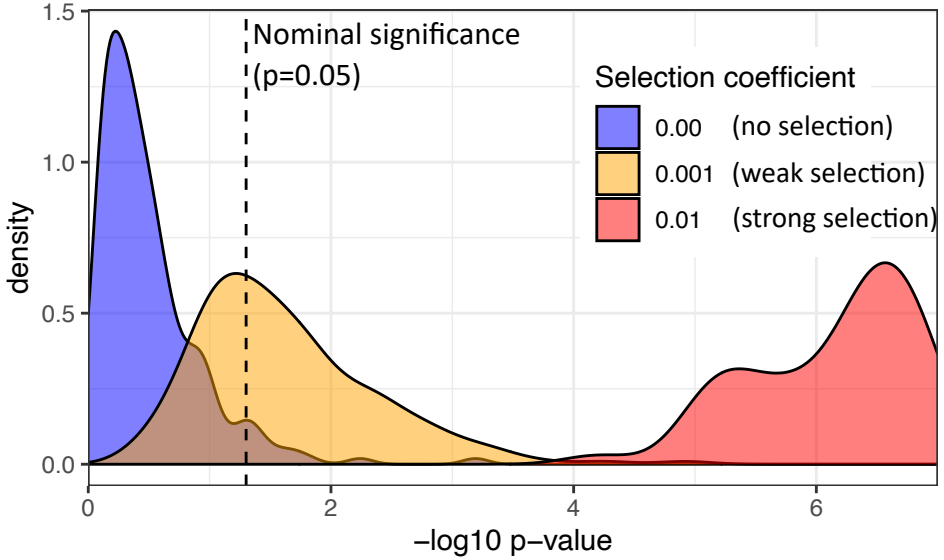
Selection p-values: only a handful of "genome-wide significant" loci



Lots of Polygenic selection signals

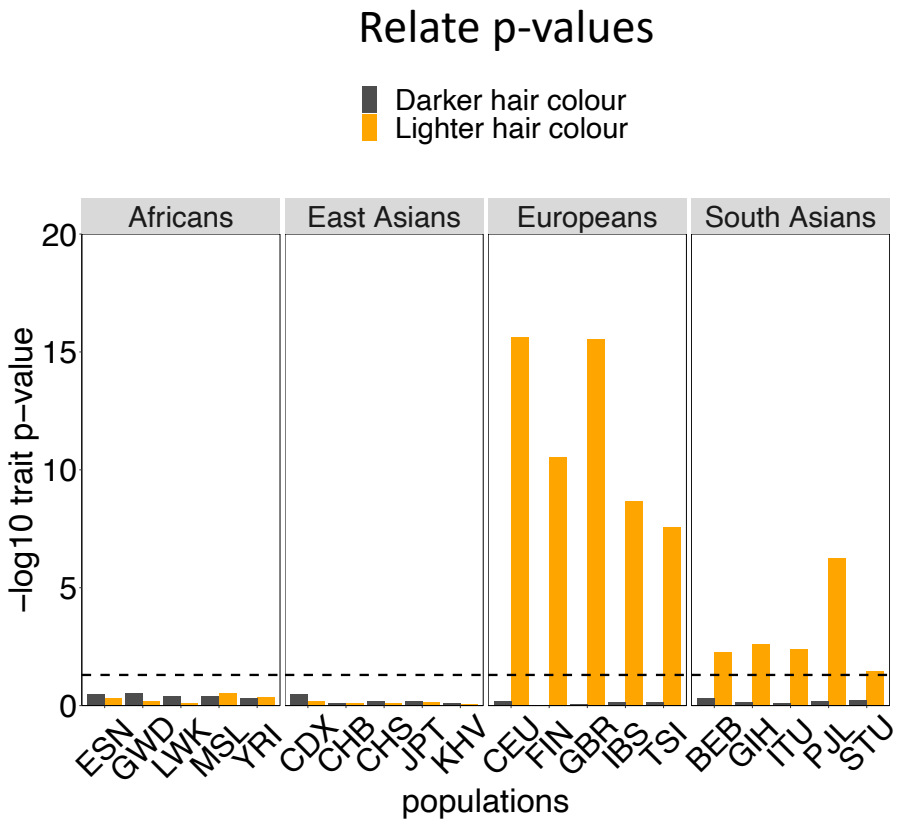


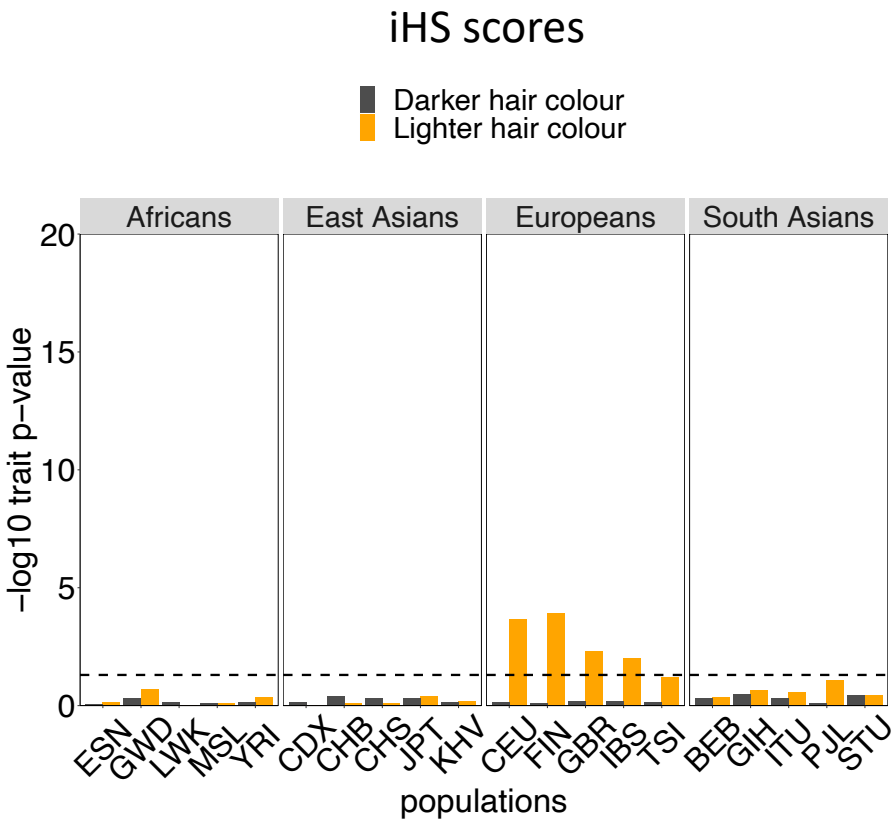
Weak selection on individual mutations is hard to detect, compared to shifts in distributions



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)





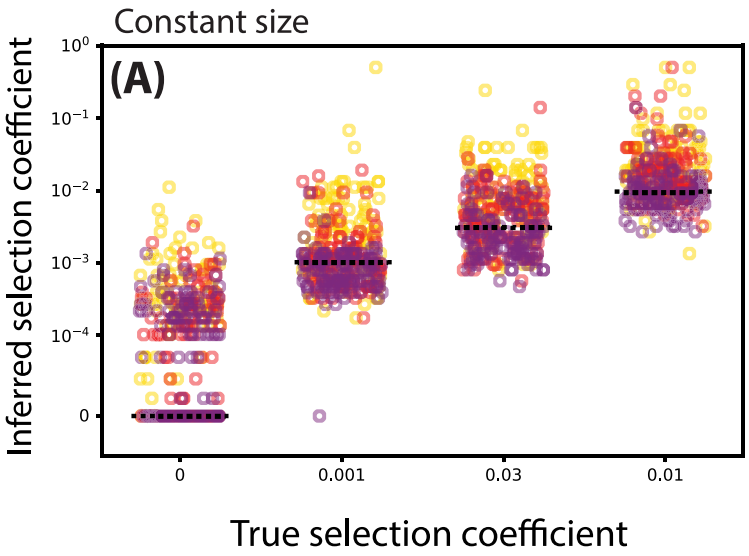
CLUES: Importance-sampling based method for inferring selection coefficients

Aaron Stern

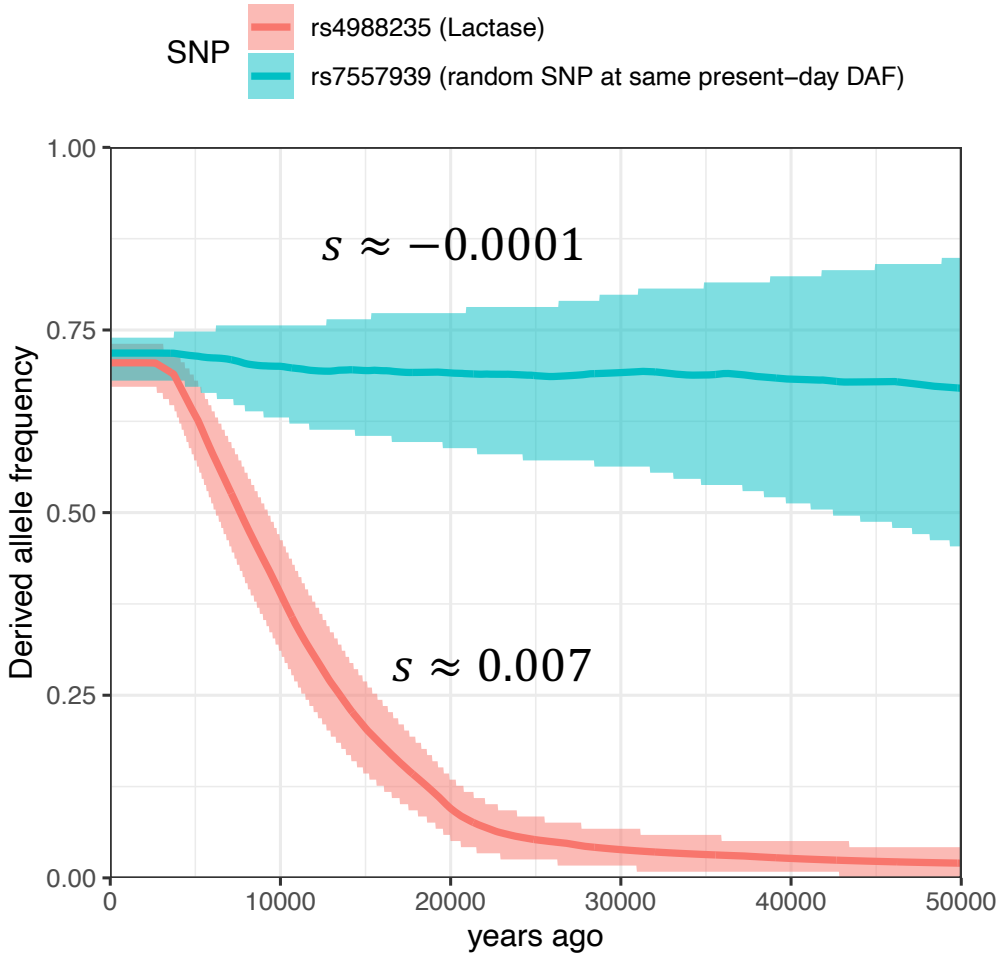
Aaron J. Stern, Peter R. Wilton, Rasmus Nielsen. **PLOS Genetics, 2019**.

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

Simulations:



1000 Genomes Project British:



Summary & outlook

- 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!