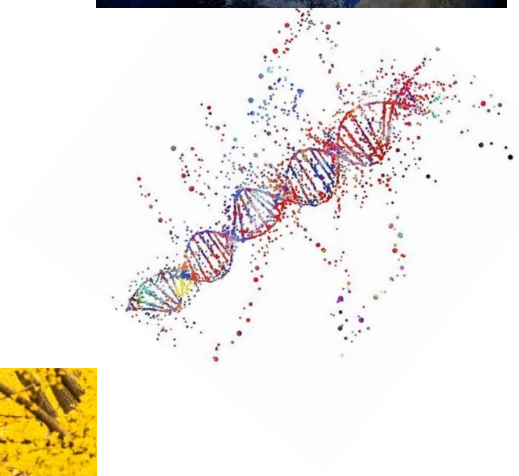
2021 EMBO Course in Population Genomics

Learning about evolution by building coalescent trees

Leo Speidel

How are we all related?



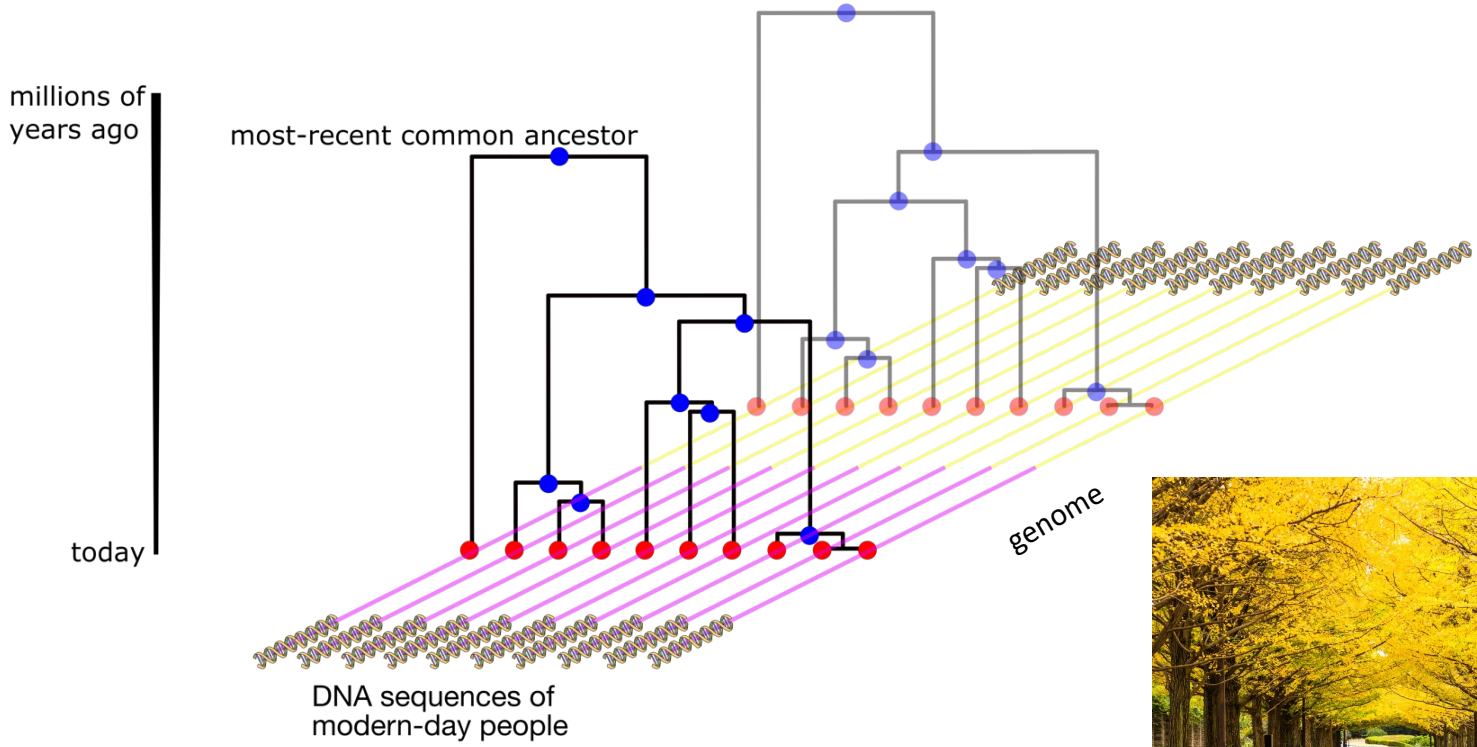
Everyone is different

- Yet people share lots of characteristics
  (hair colour, blood group, disease susceptibility)
- Relatives tend to be more similar

How does genetic variation arise and how is it maintained?

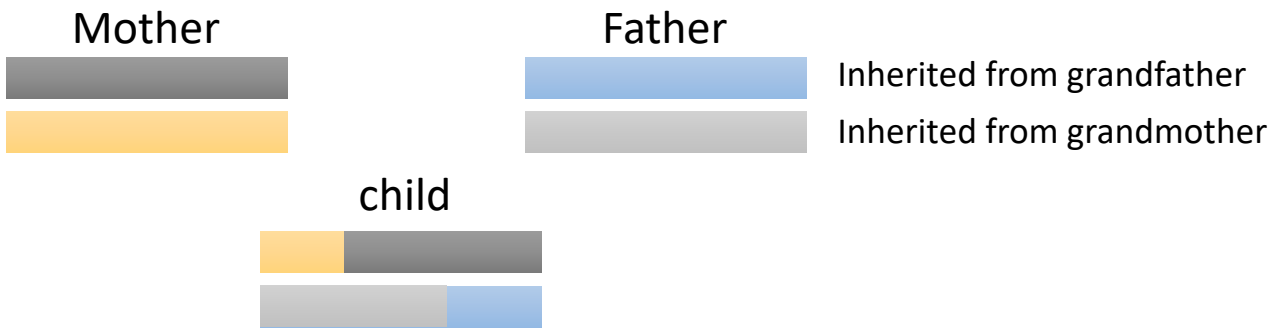
Big family tree?

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



Reason for why trees change along the genome

Recombination:



Processes impacting tree shapes and genetic diversity

Harris et al, eLife 2017

Genetic variation is shaped by

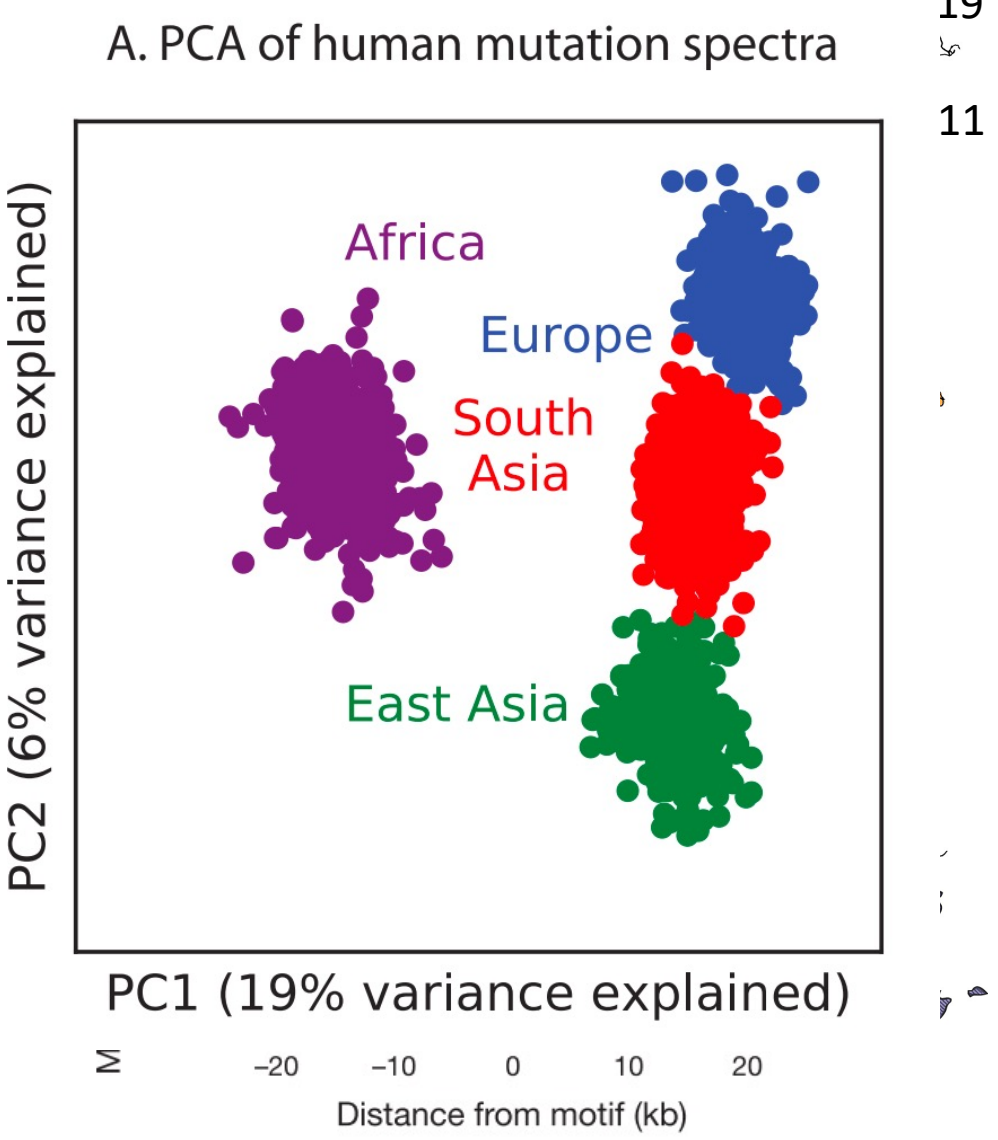
- Structure and migrations
- Population bottlenecks
- Admixture

- Mutation
- Recombination

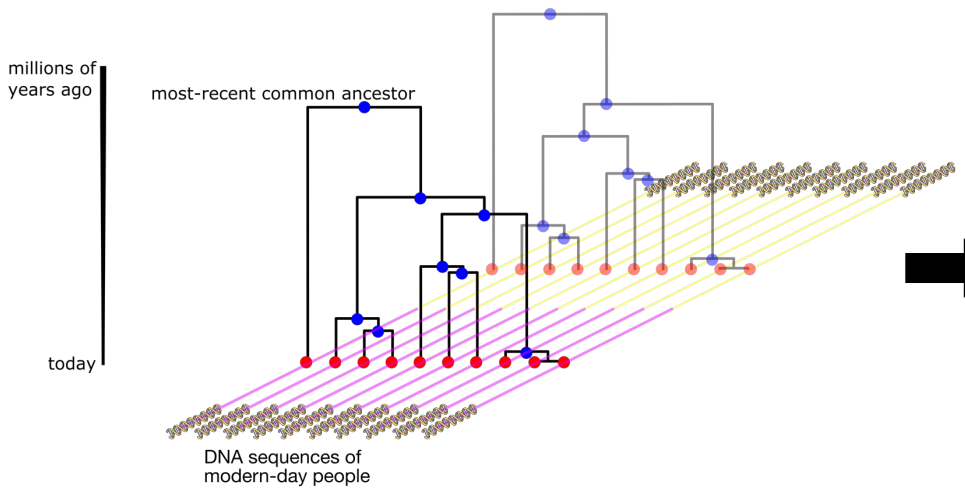
- Selection
- Etc

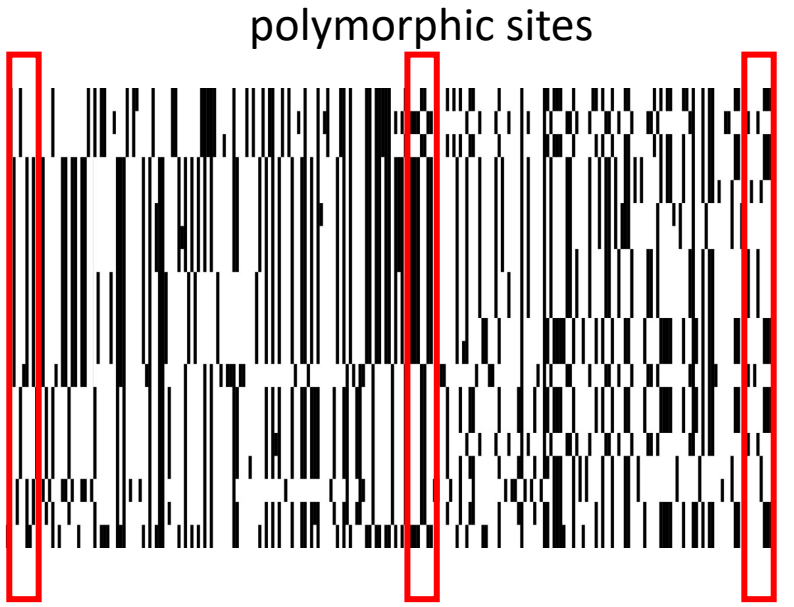
These evolve over time!

"evolution of evolution"



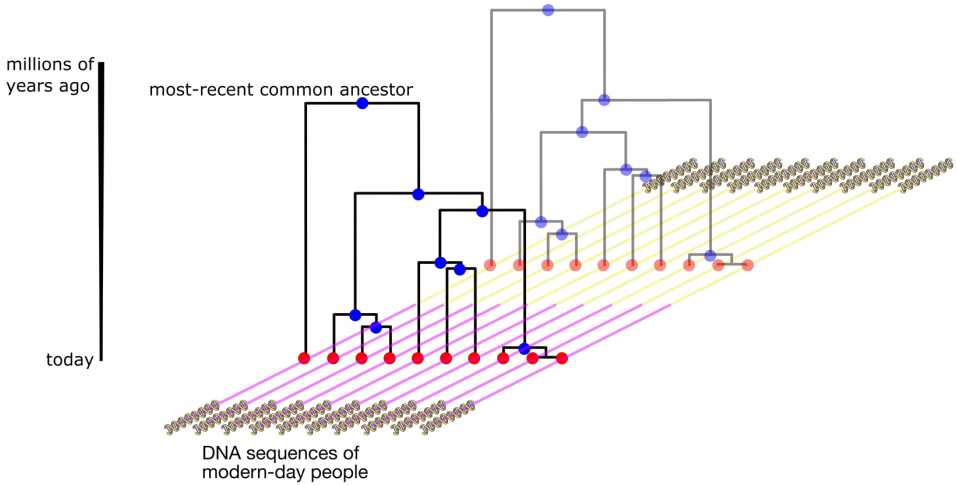
Fundamental forces impact data (only) through underlying genealogies



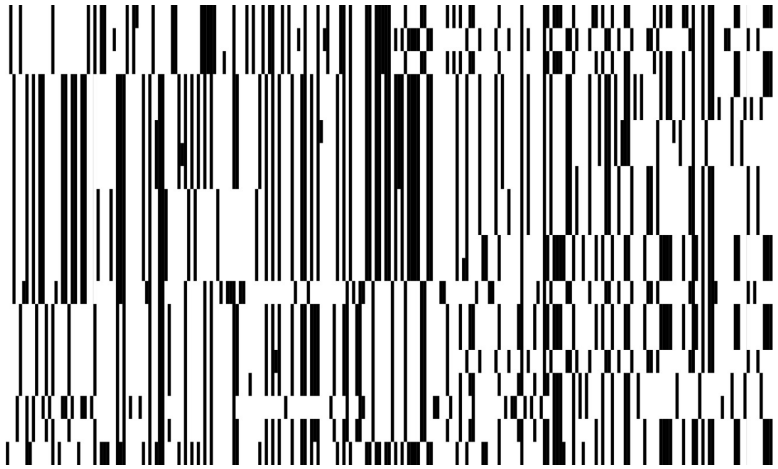


Demographic history
Genetic structure
Mutation, recombination, etc.

Many canonical approaches



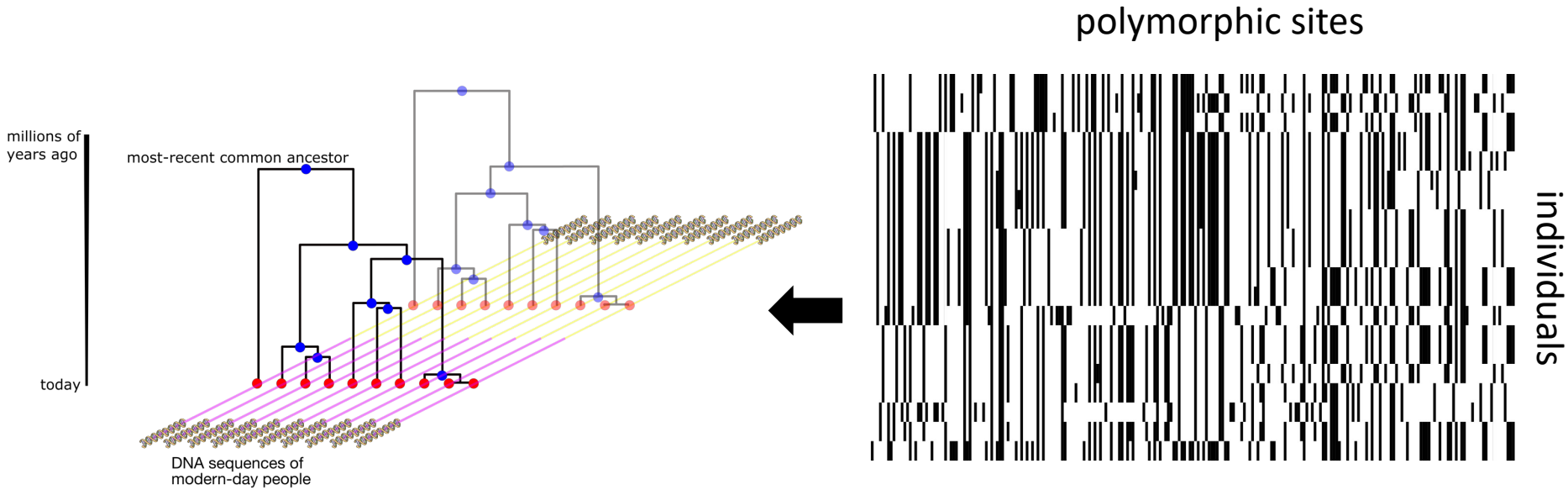
polymorphic sites



Demographic history
Genetic structure
Mutation, recombination,
etc.

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

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



Demographic history
Genetic structure
Mutation, recombination, etc.

Inferred trees reorganise all the information available from the data about these processes in a very accessible & powerful way

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
- Rent+

} Infers Ancestral Recombination Graphs

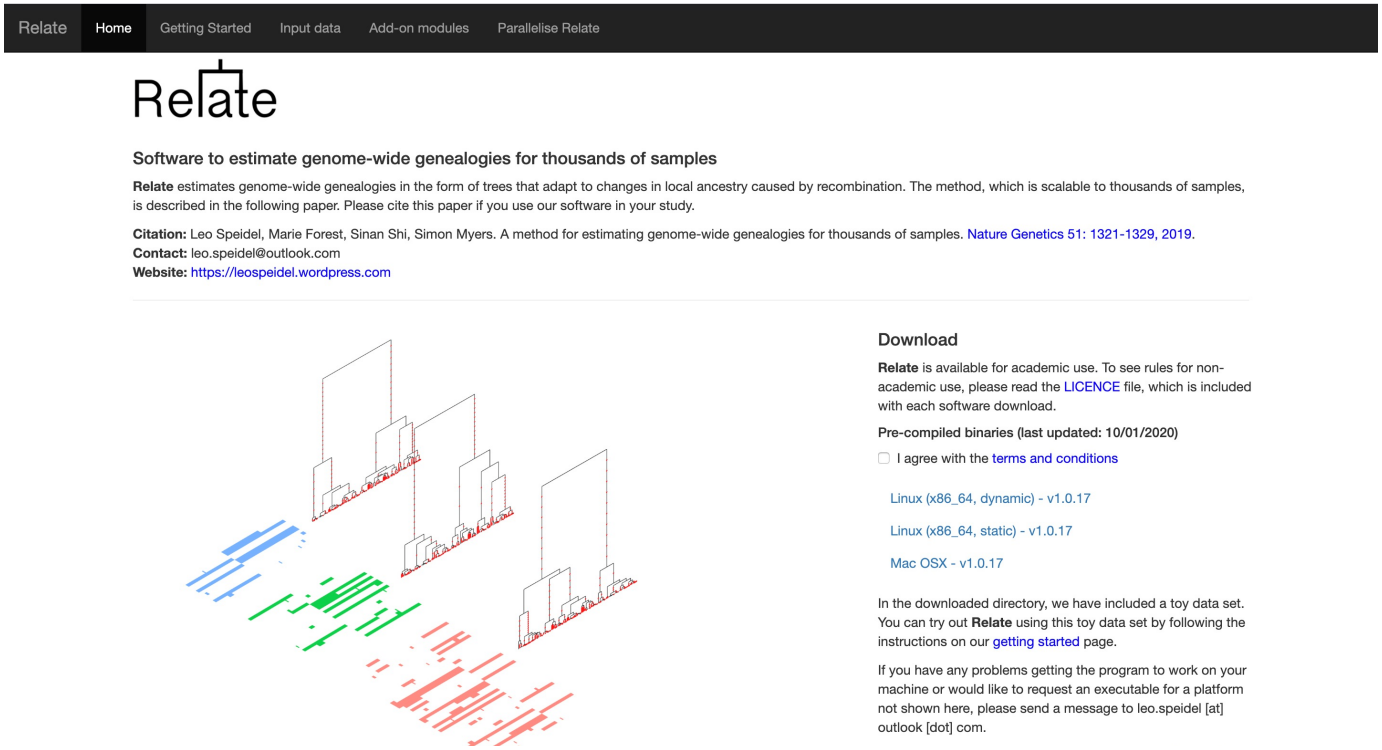
- Tsinfer + tsdate
- Relate

} Published in 2019/2021, scale to large sample sizes

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

Relate

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

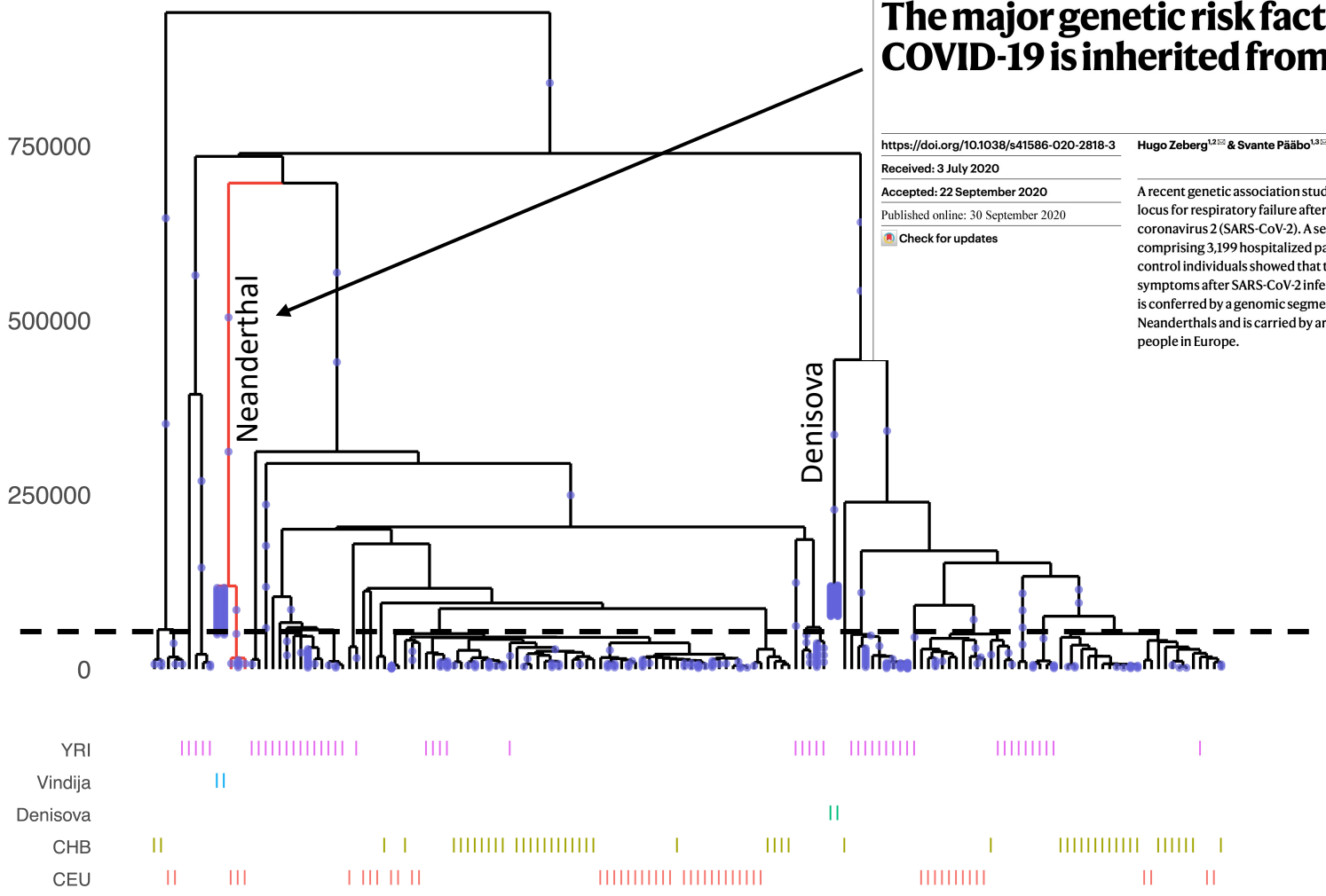


https://myersgroup.github.io/relate/

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



Article

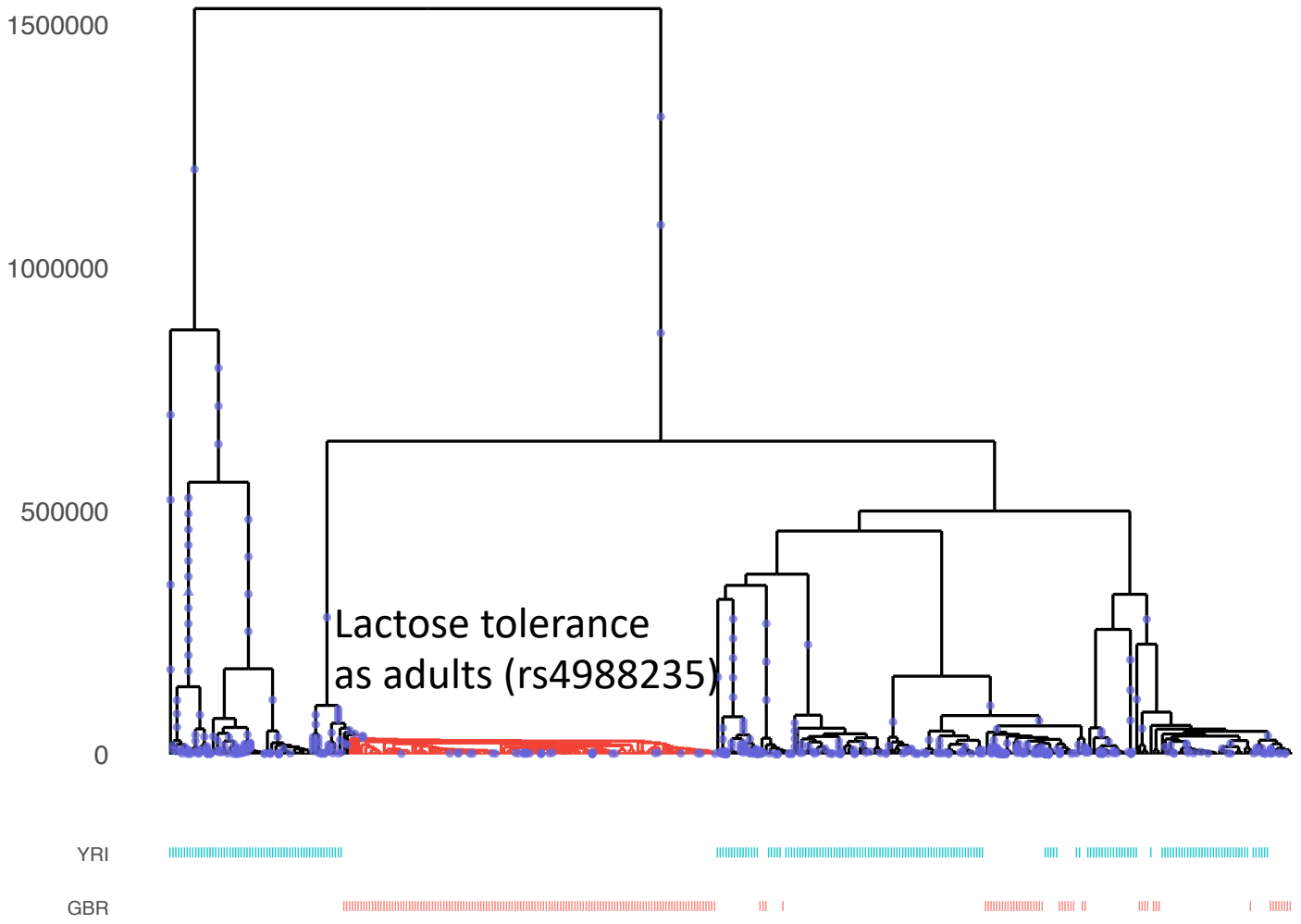
The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

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

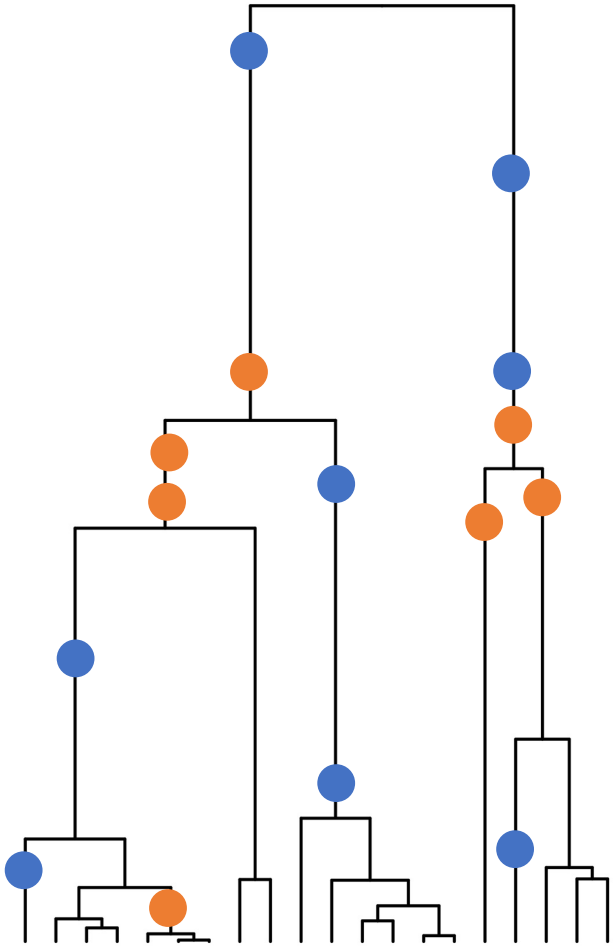
Hugo Zeberg[1,2] & Svante Pääbo[1,3]

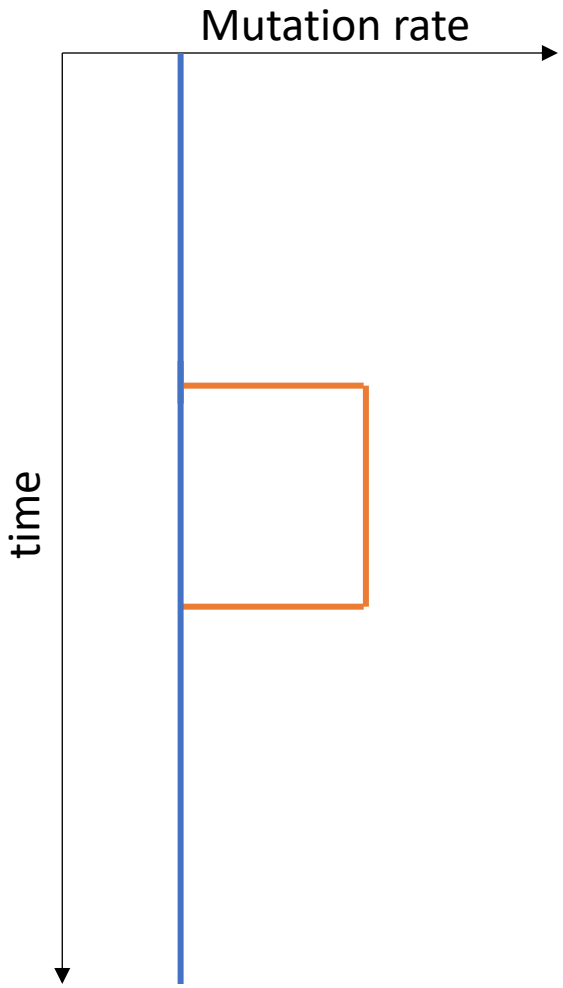
A recent genetic association study[1] identified a gene cluster on chromosome 3 as a risk locus for respiratory failure after infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A separate study (COVID-19 Host Genetics Initiative)[2] comprising 3,199 hospitalized patients with 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.

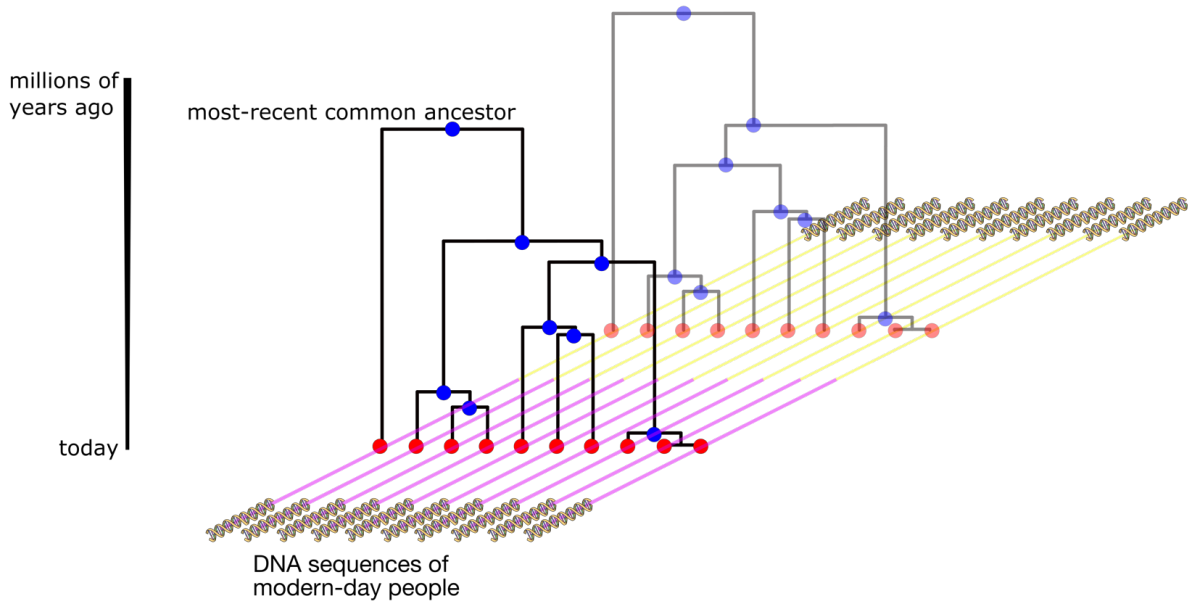
Positive selection: rapidly spreading lineage



Clusters of mutations in time can capture changes in mutation rate







Inferring genealogies

Old problem

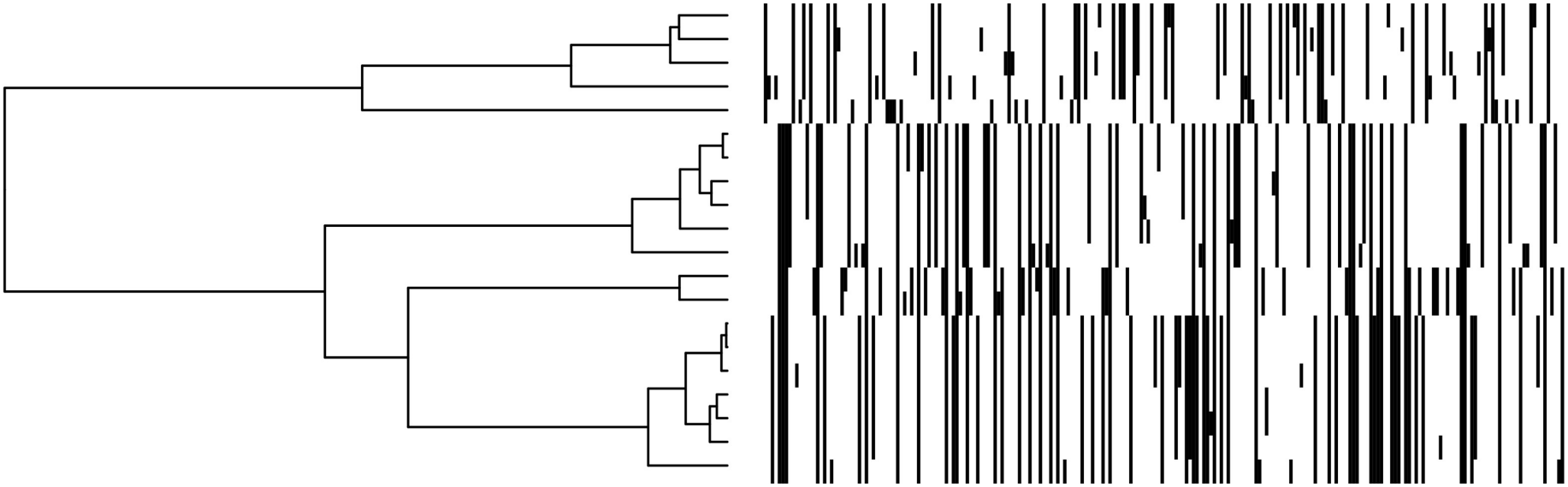
- 90s – 2000s: Use the coalescent with recombination to infer genealogies, e.g. MCMC

$$P(\text{tree} \mid \text{Data}) \propto P(\text{Data} \mid \text{tree}) * \underbrace{P(\text{tree})}_{\text{coalescent}}$$

- Relate and tsinfer (2019):
  - Separate inference of tree topology and branch lengths
  - Use fast but approximate approach for tree topology
  - Use coalescent for branch lengths

Scalable to many thousands of samples

Data and the underlying tree structure

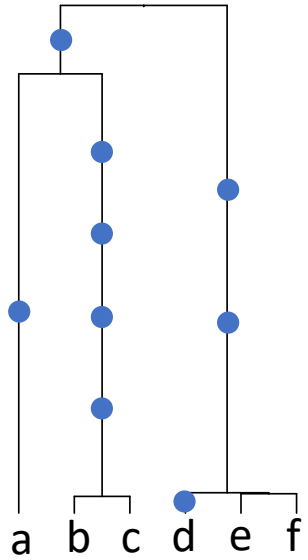


- **Every mutation shows the existence of a branch**
- Mutations are “ordered by inclusion”
- No two branches (mutations) ever show only partial overlap

Use hidden Markov model to identify stretches where relationships to other are unchanged

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

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

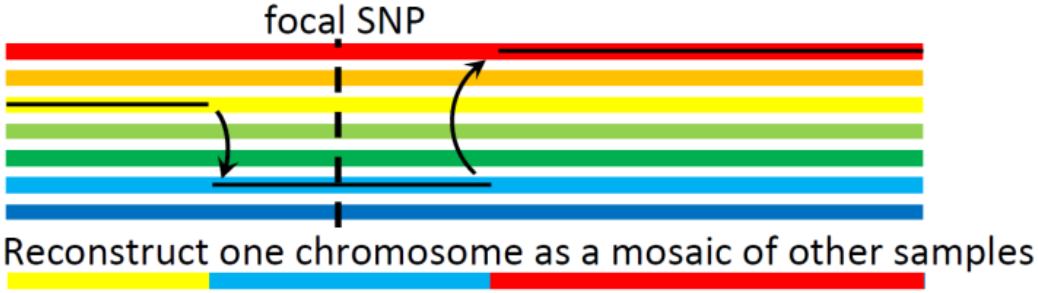


- No recombination: Guaranteed to build tree consistent with data
- This is not the case if we use "pairwise differences" (UPGMA)
  - a and (d,e,f) are closer than a and (b,c)

- Use HMM to count derived mutations accounting for recombination

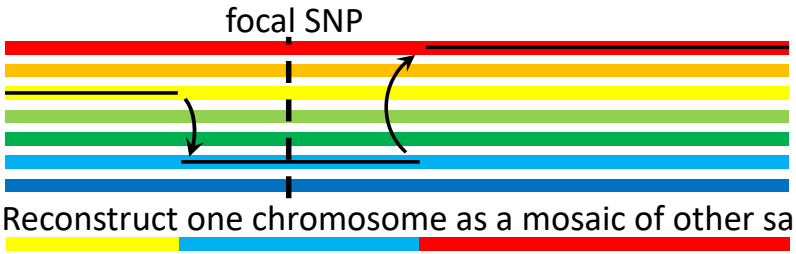
Hidden Markov model (HMM)

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

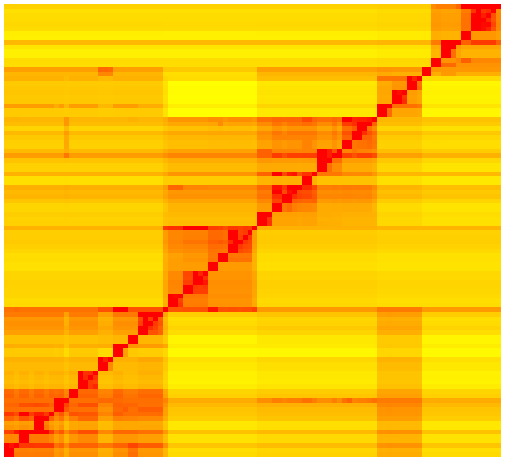


Summary of Relate

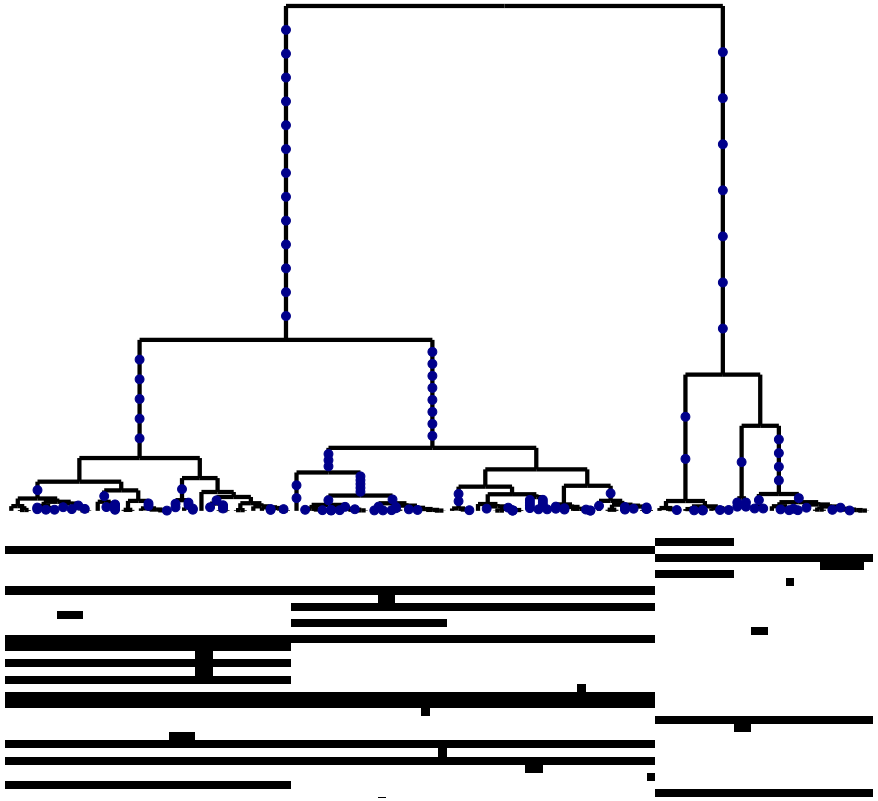
Hidden Markov model (HMM)



Distance matrix for focal SNP



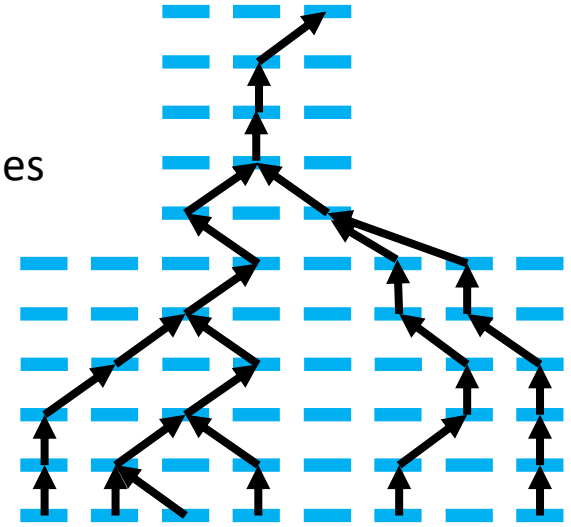
Hierarchical clustering
&
MCMC for branch lengths

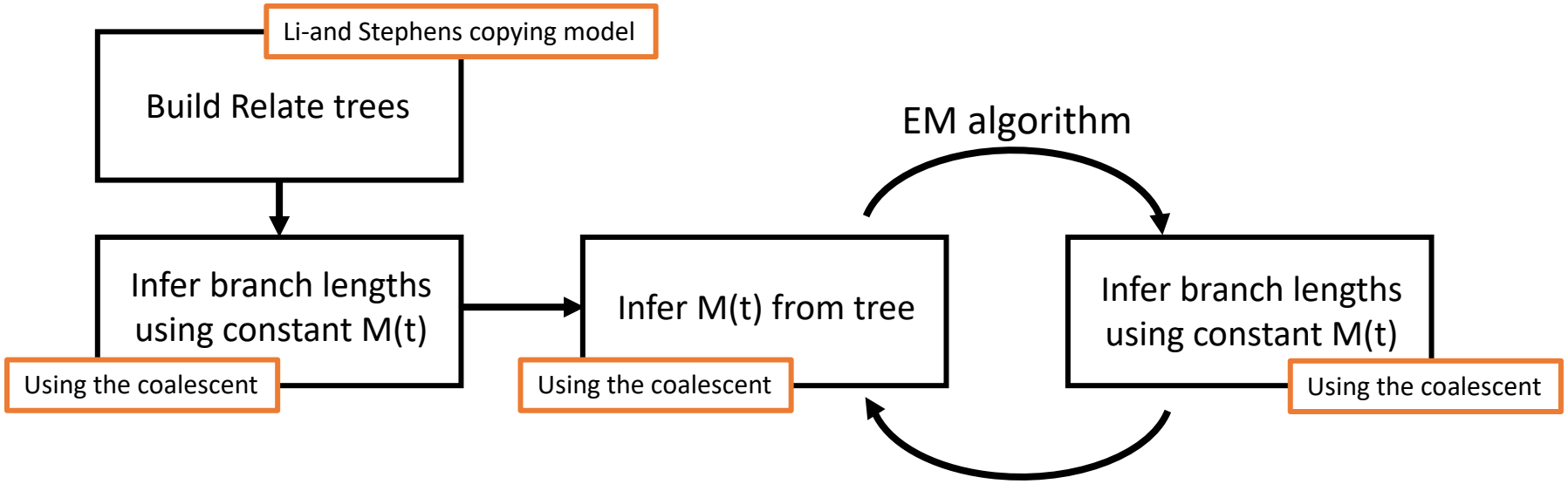


haplotype data sorted using constructed tree

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

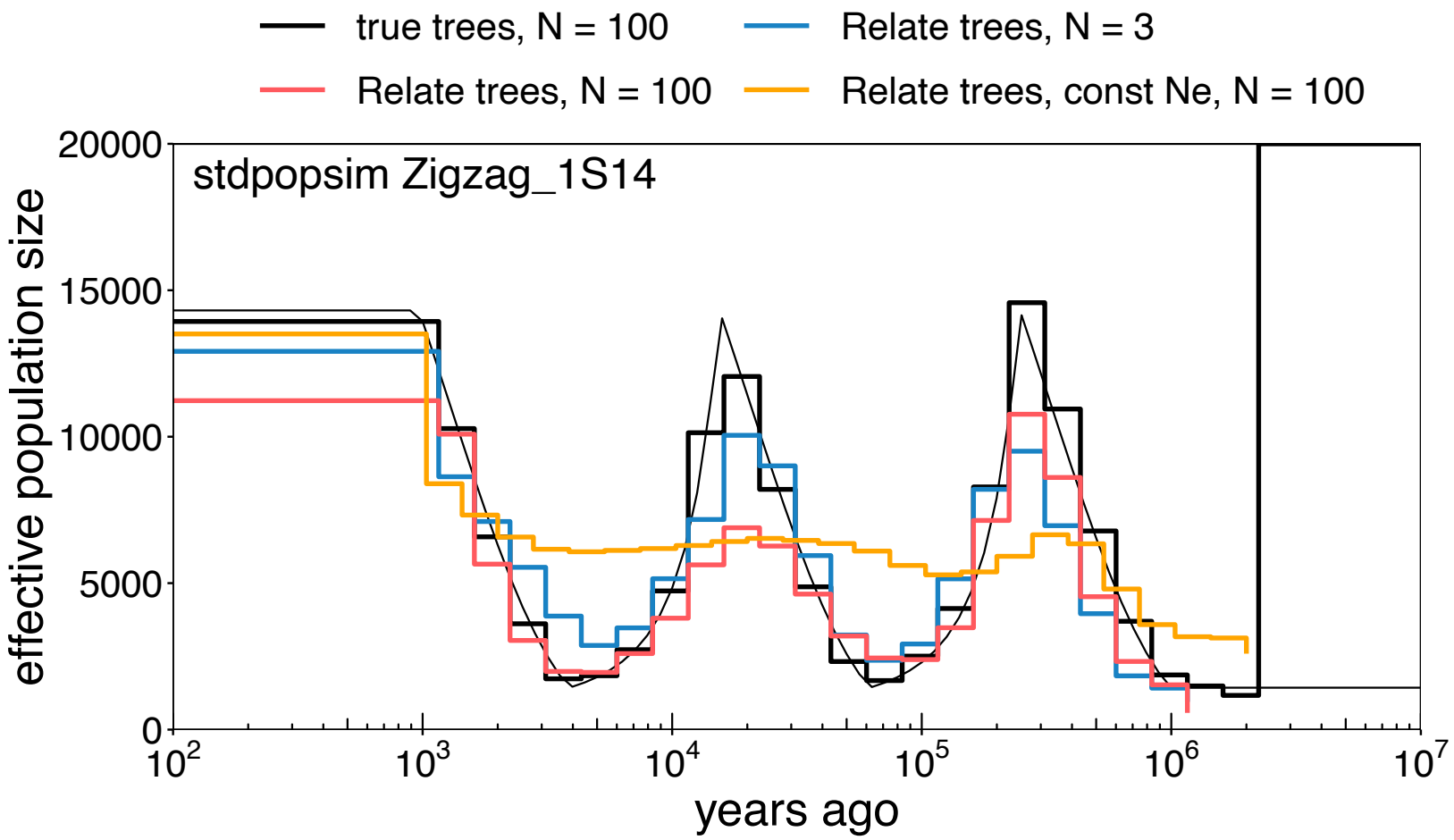
- While there are *j* lineages, the rate at which a coalescence happens is $\binom{j}{2}/M(t)$ a time *t* ago
- Demography is shared genome-wide, so we average across trees
- So within a time interval, scaled fraction of trees where coalescence occurs is inversely proportional to M(t)



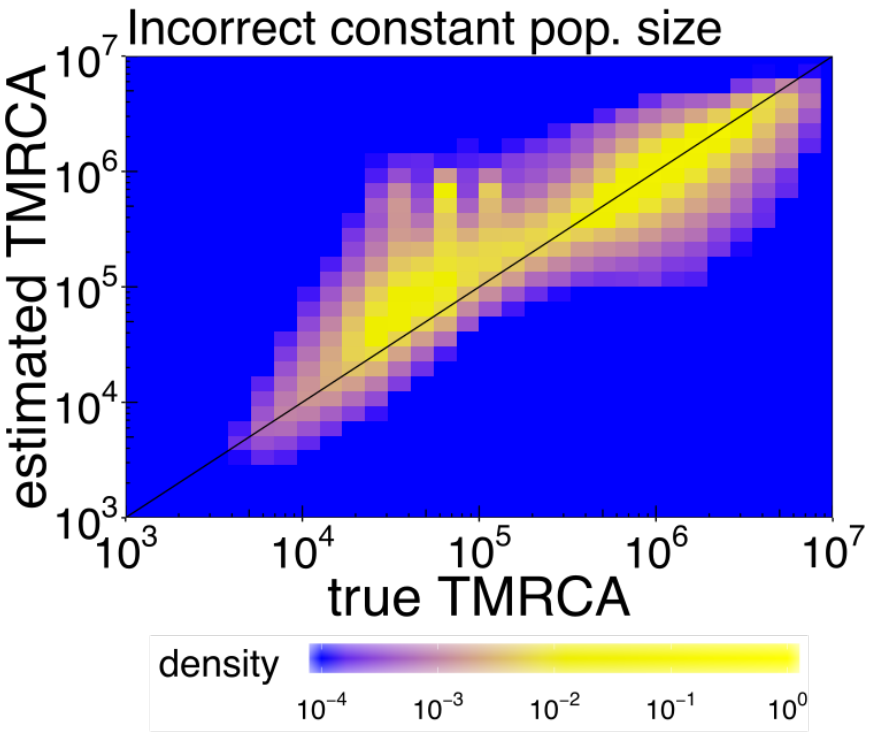


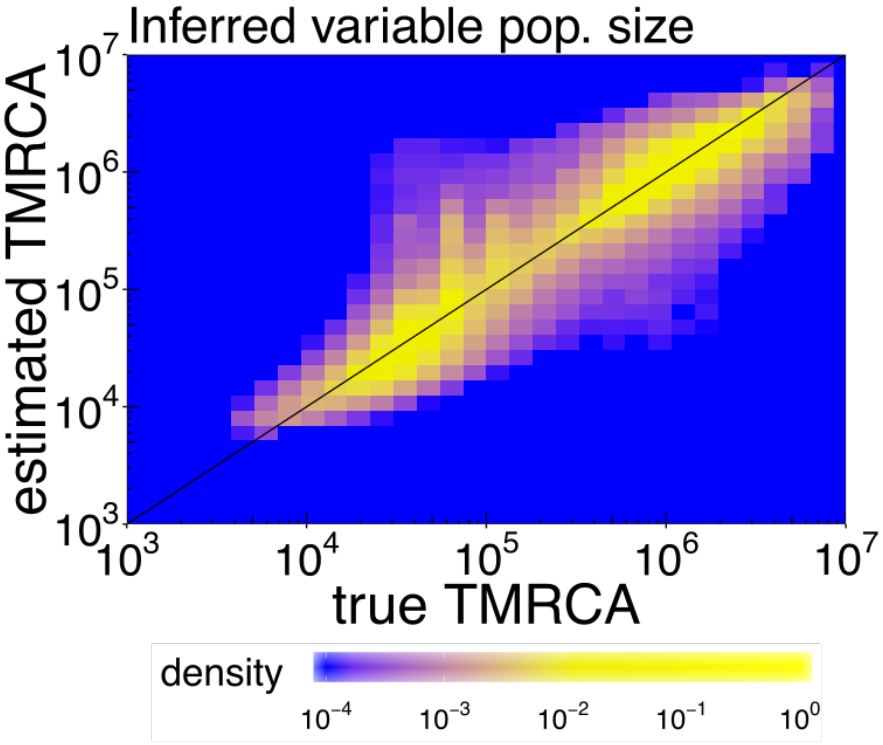
Population size changes through time are jointly inferred in Relate

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



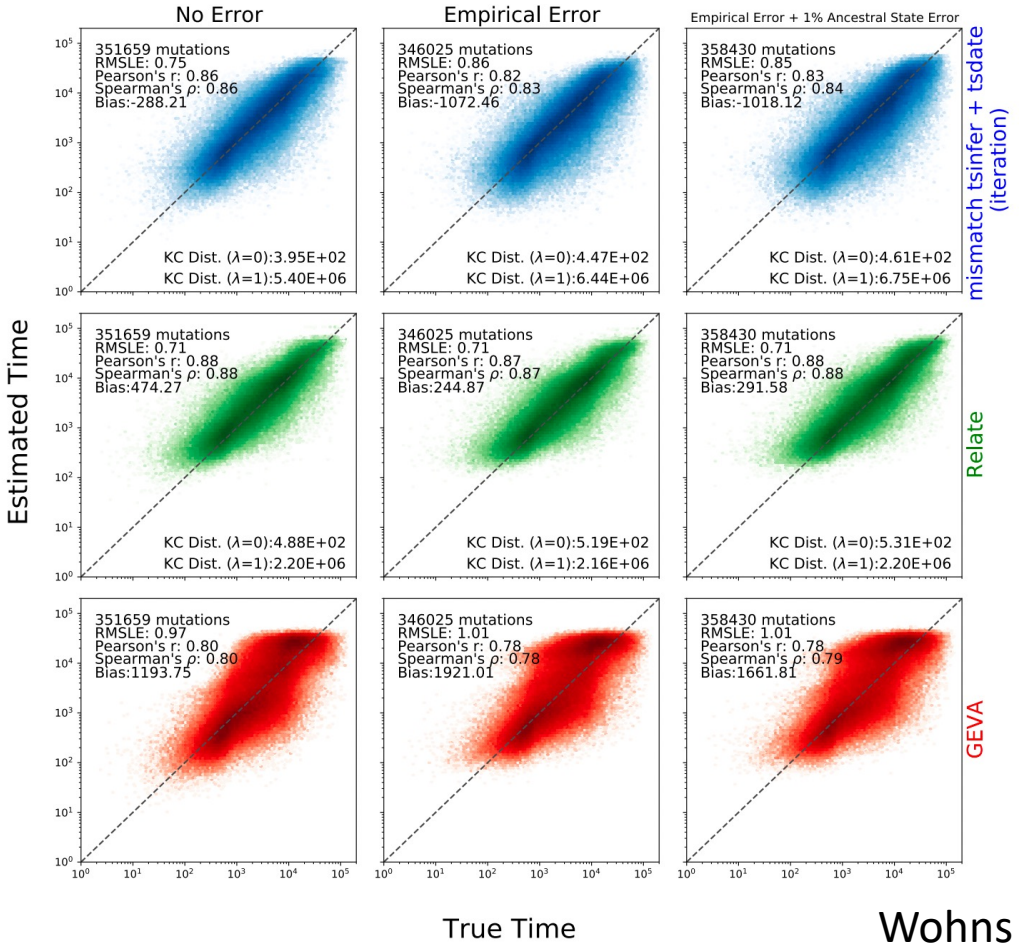
Simulated data with variable population size (European-like demographic history)





Speed and accuracy of Relate

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



Wohns et al, bioRxiv 2021

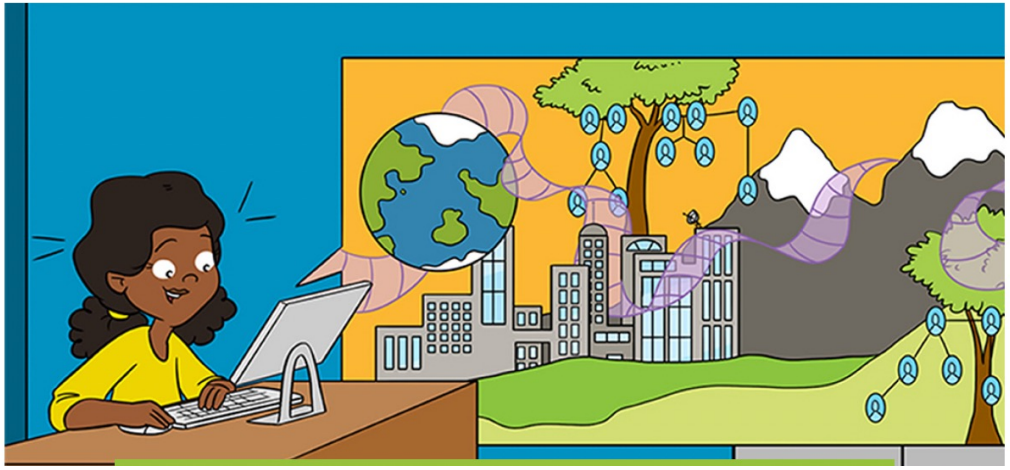
What Our DNA Can Tell Us About the History of Humans

Authors

Leo Speidel Clare Bycroft

Young Reviewers

Mariana Anna-Marie Zara



Luckily not our reviews…

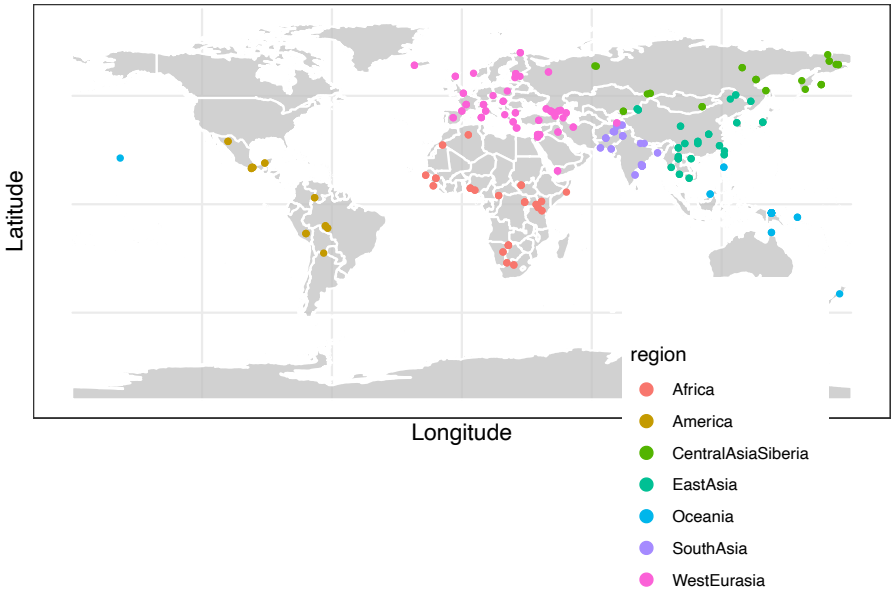
> This seems important, but the way it is written is so boring I can't even get to the end. Could the authors maybe sound excited about what they are doing?
>
> Reviewer, Age 12

> The writers of the article did not make it clear why such an expensive and involved research project was done to begin with … It seemed like a fruitless task.
>
> Reviewer, Age 14

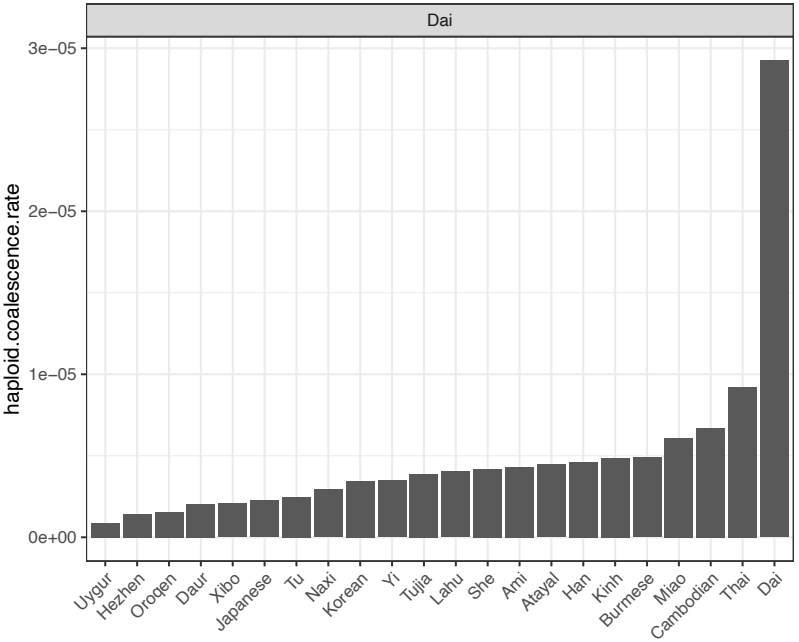
Genealogy-based inference of human evolutionary history

Coalescence rates track relatedness through time

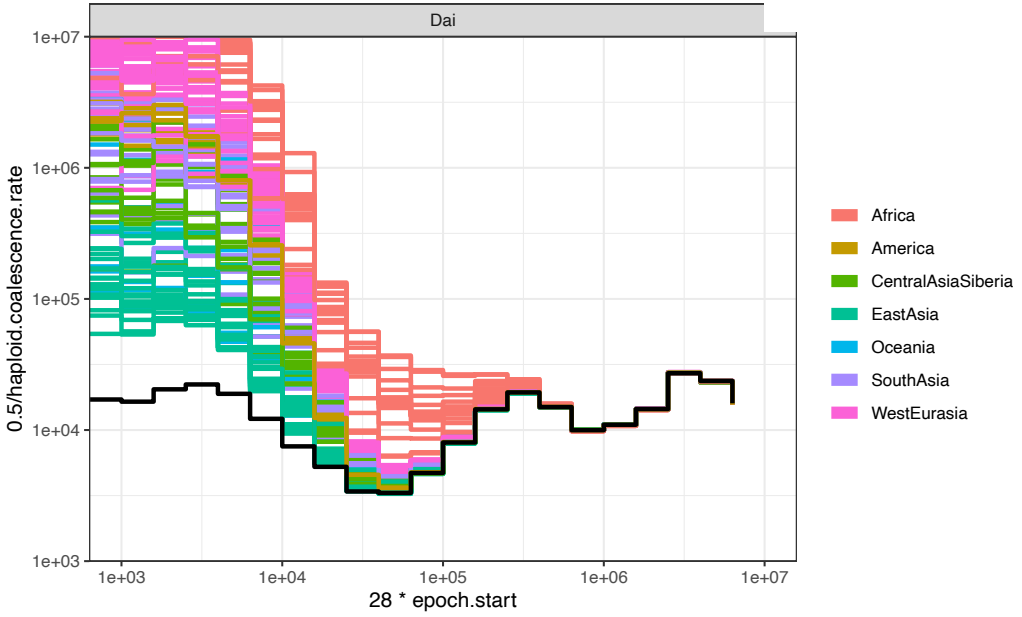
Coalescence rates = 1/M(t)
M(t): effective population size



Relatedness < 1000 years in East Asia

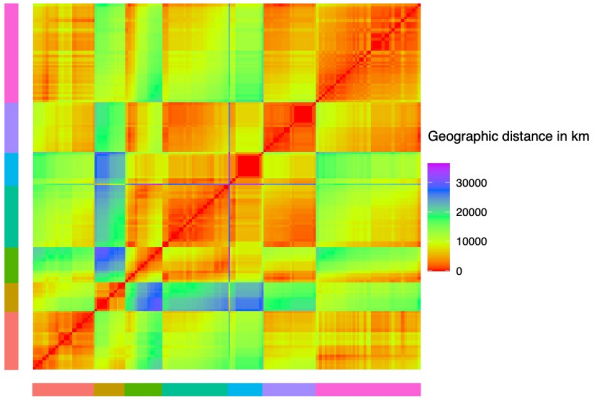


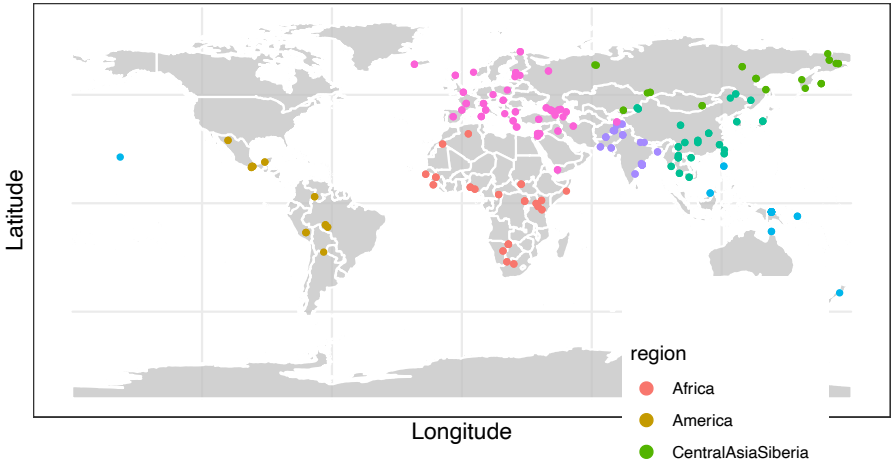
Relatedness through time globally



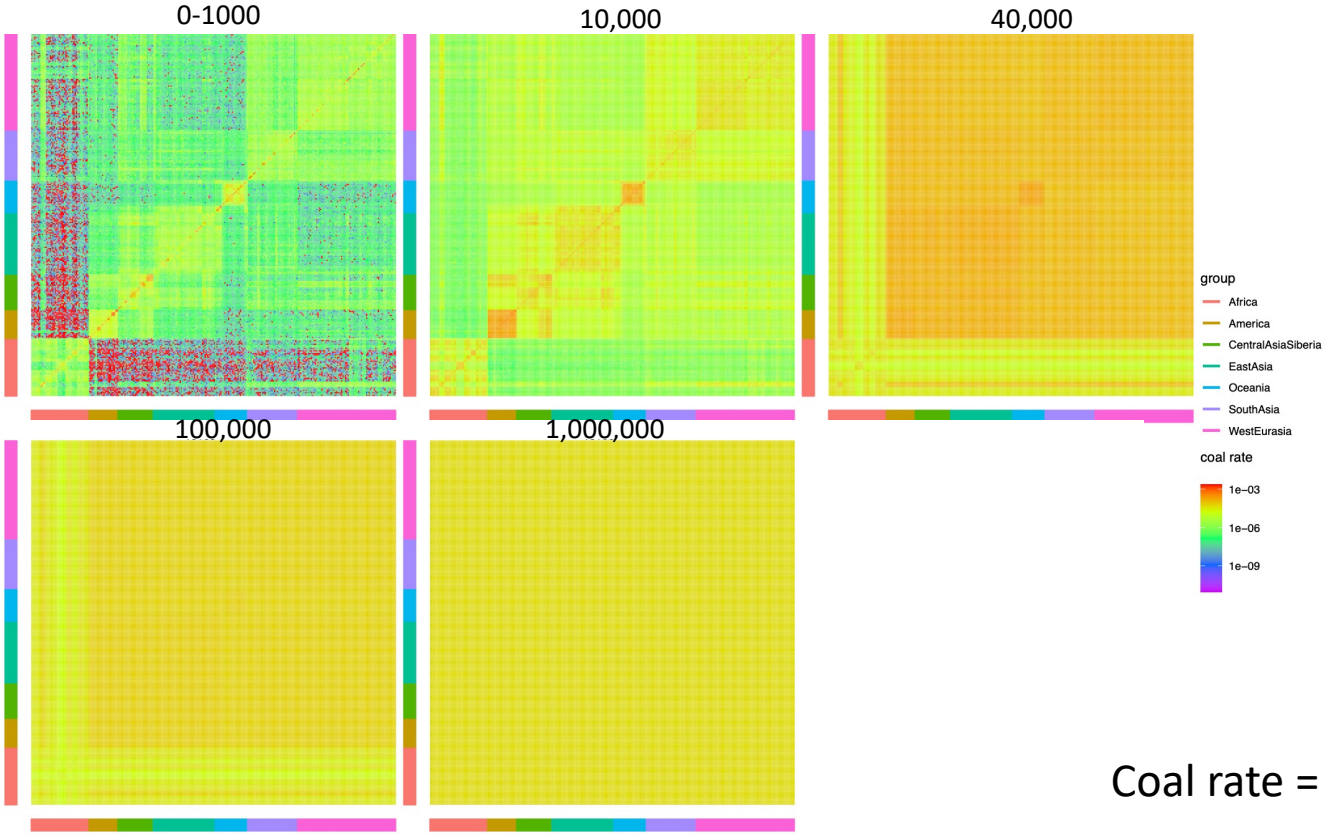
Structure through time

Geographic distance



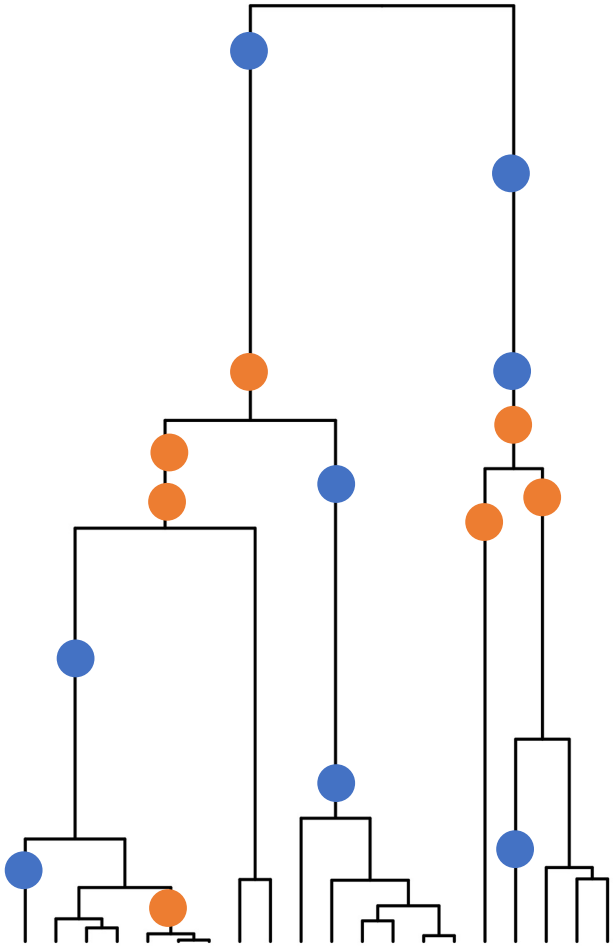


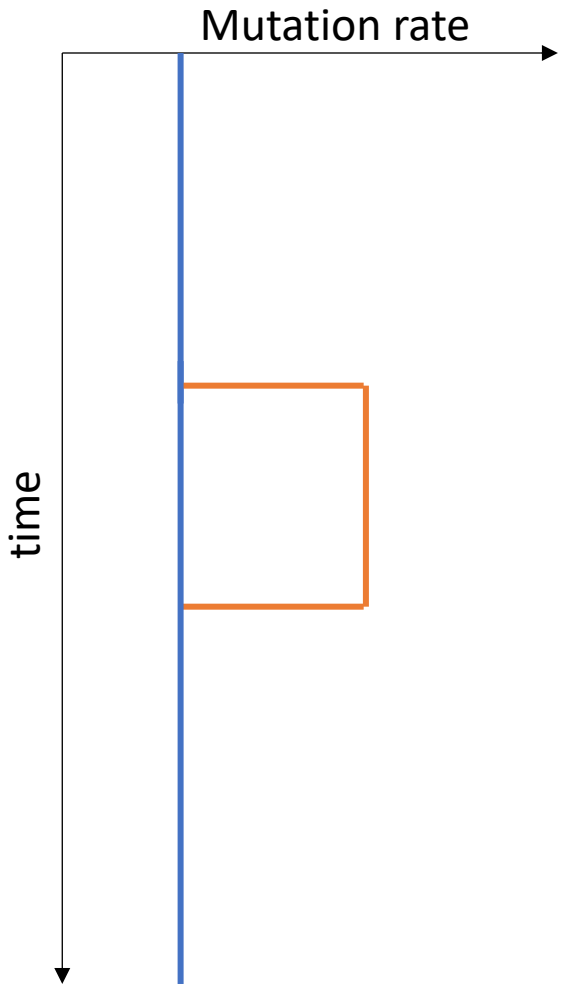
Coalescence rates

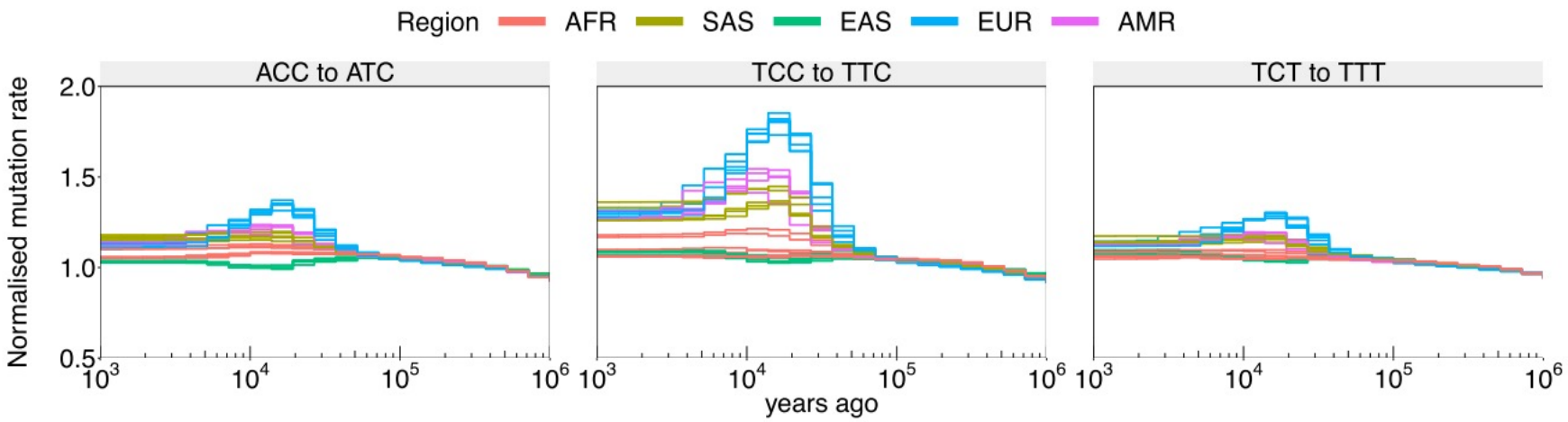


Coal rate = 1/M(t)

Clusters of mutations in time can capture changes in mutation rate







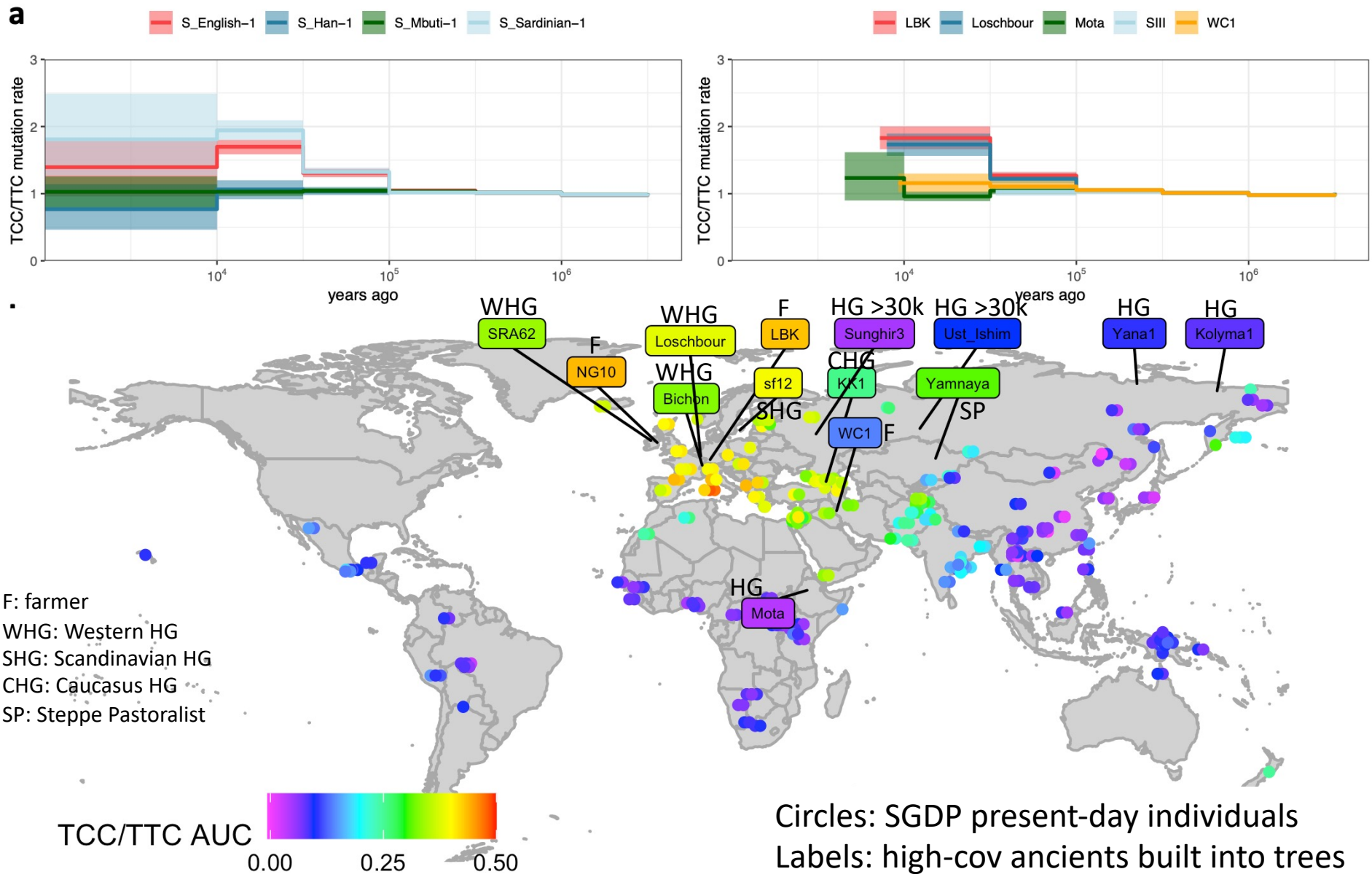
"Mysterious" TCC/TTC mutation rate in West Eurasia

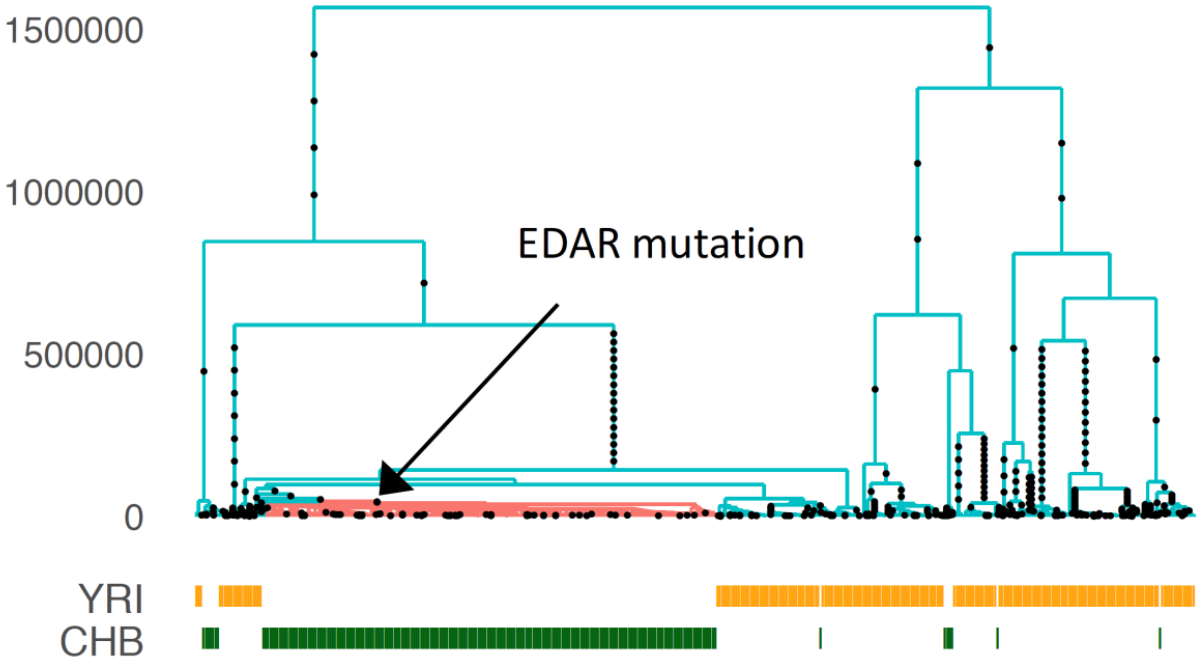
- First reported by Kelley Harris (PNAS 2015, eLife 2017)
- Strongest signal of all triplets and strongest in Europeans among modern-day people
- Unknown cause (genetic?, environmental?)
- Date of onset unclear (dates range from 10k – 80k)
- Studied mainly in moderns, handful of ancients (Mathieson & Reich 2018),

  but detailed geographic & temporal pattern was unexplored

TCC/TTC mutation rate inferred from Relate genealogies

Speidel et al. bioRxiv, 2021



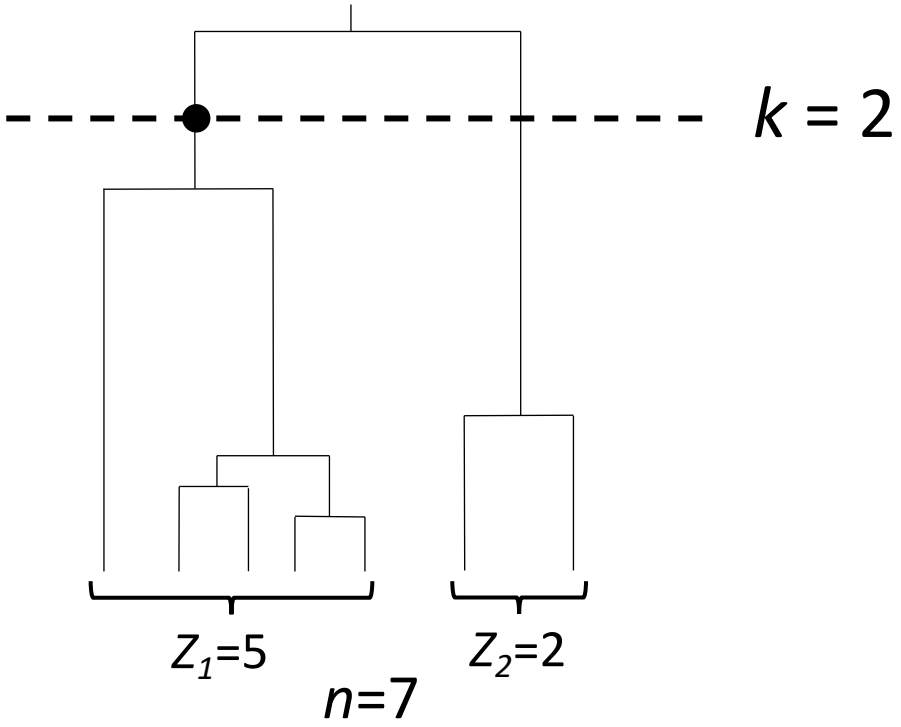


Quantifying positive natural selection on a single mutation

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

Reject a null model

How quickly does a mutation spread in the neutral case?



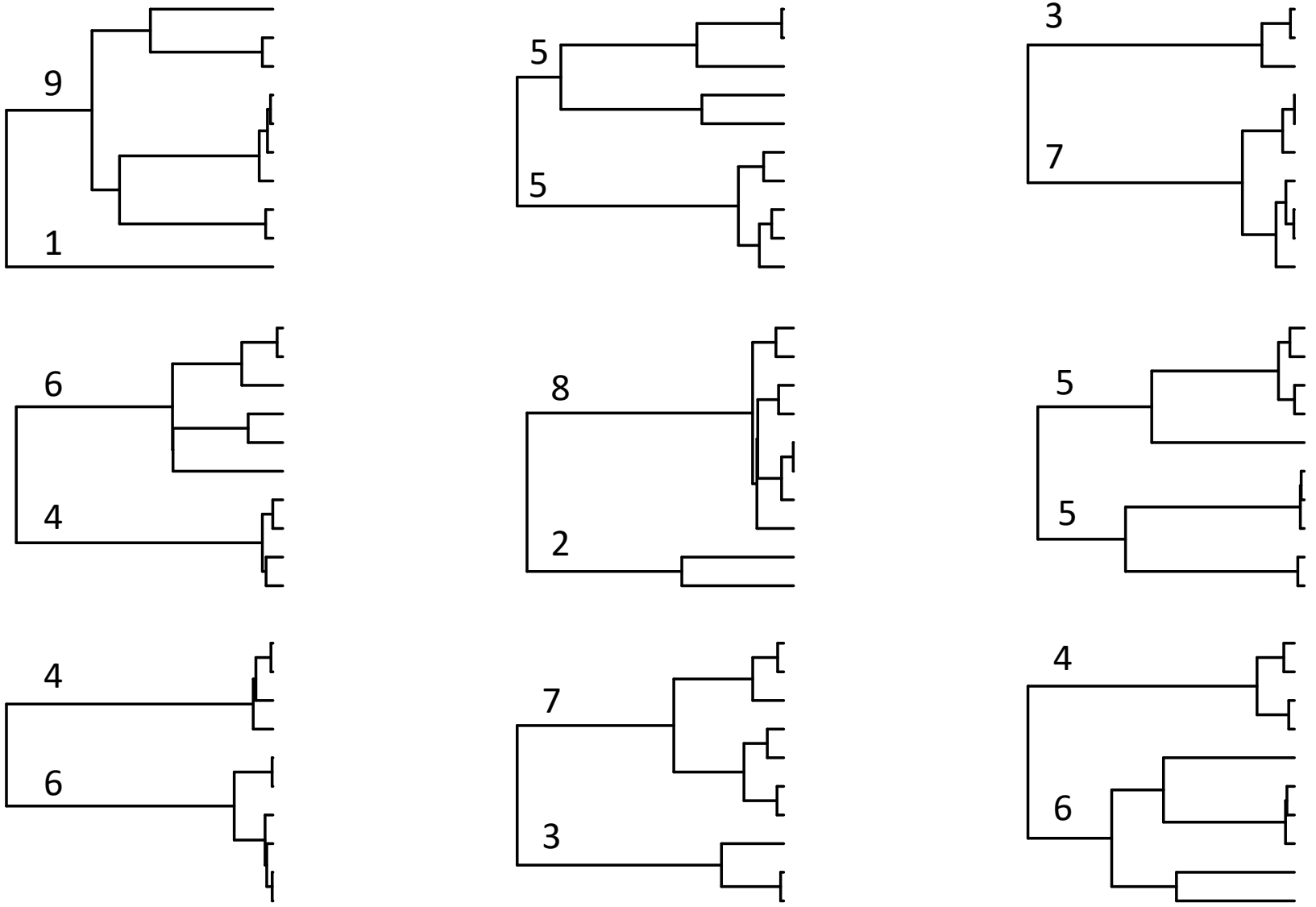
We can write down the analytical distribution for the number of descendants of a mutation arising while k lineages remain

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

P(5 descendants)=1/6

The k = 2 case

We expect “unbalanced” tree shapes!

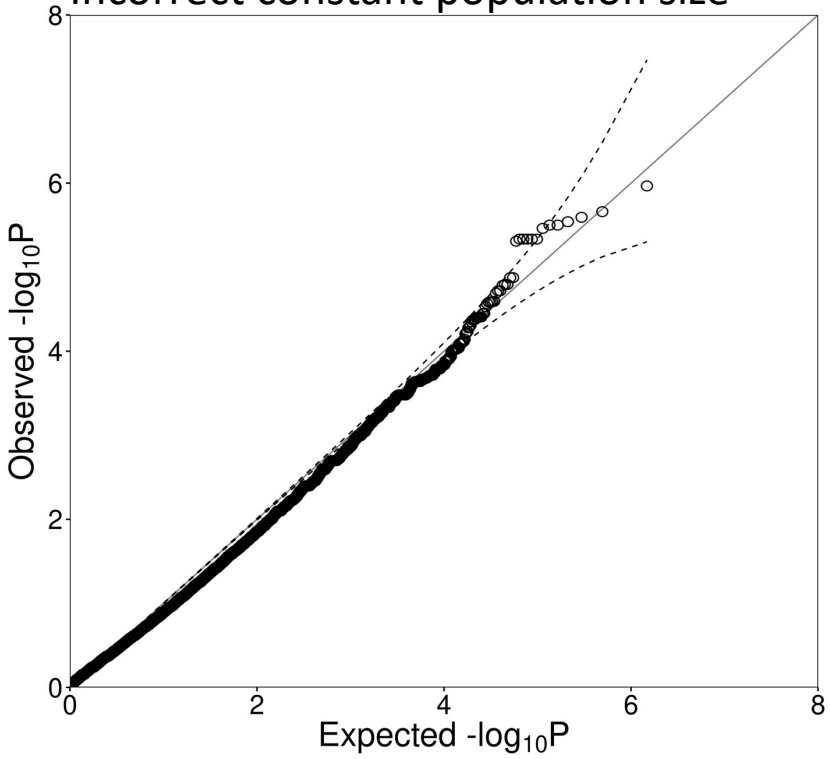


P-values: very well calibrated under null simulations of no selection

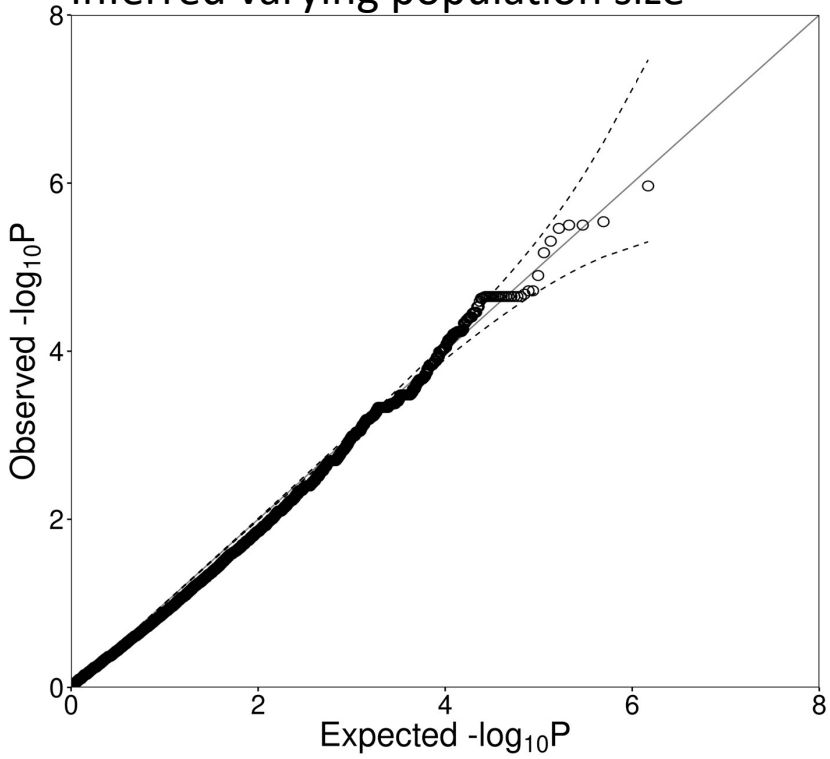
N=1000, 250Mb
Bottleneck population size

Quantile-quantile plot:

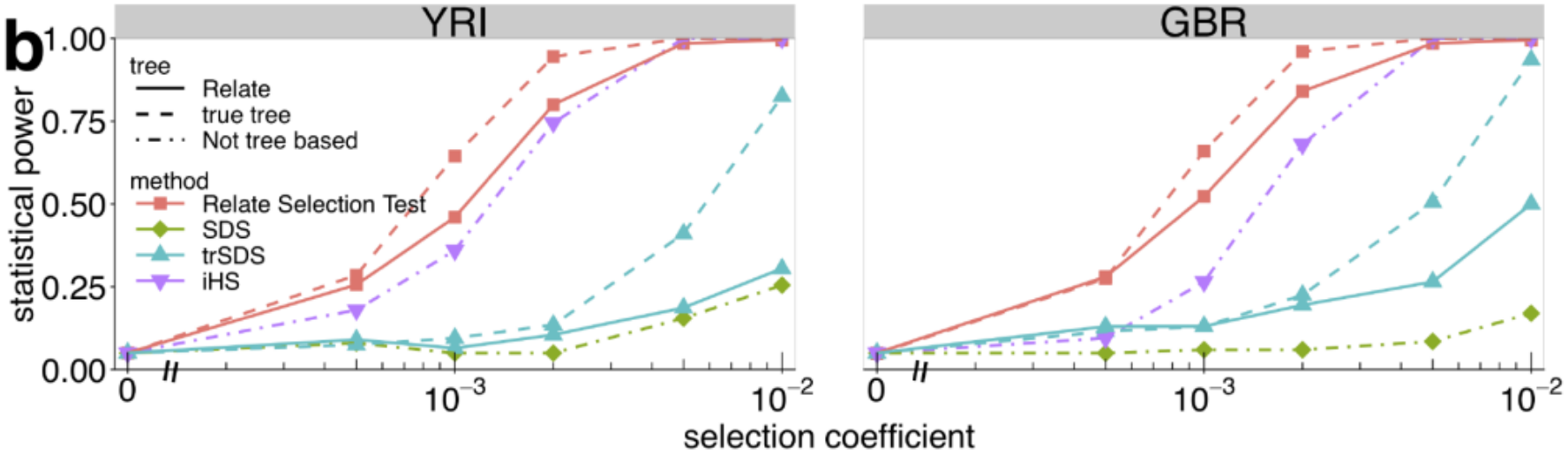
Incorrect constant population size



Inferred varying population size

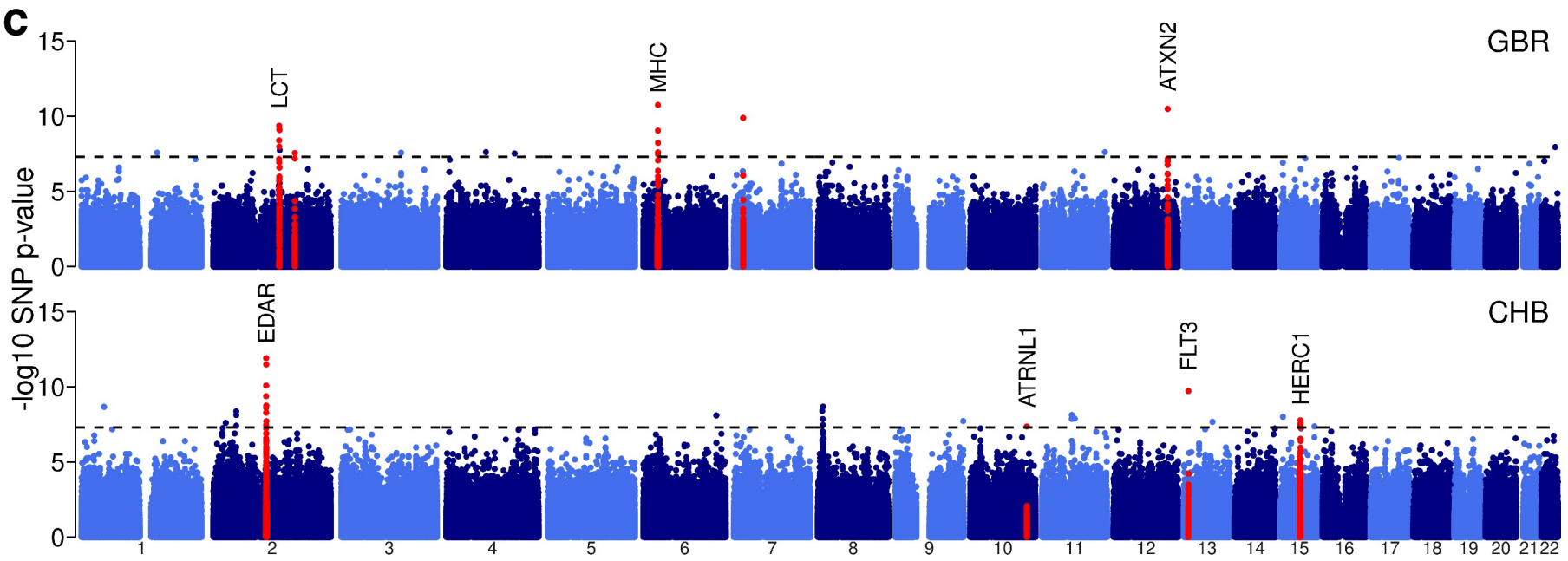


Improved power to see weak selection



Genome-wide selection p-values

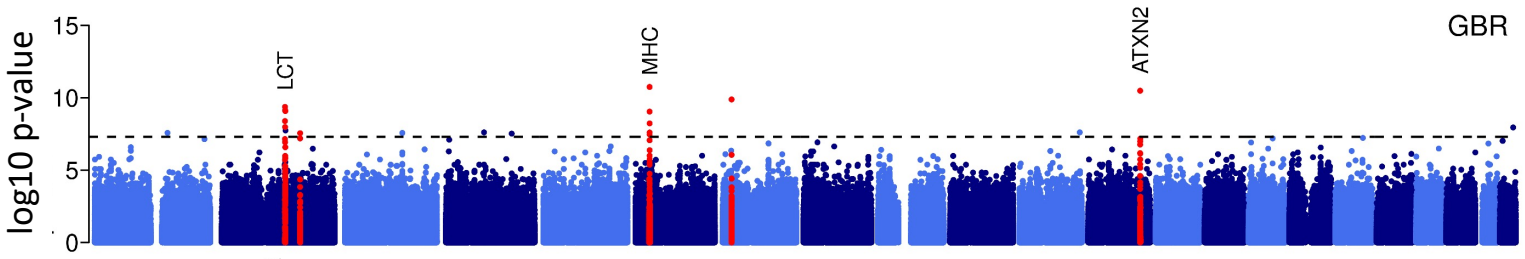
Given most traits are highly polygenic, expect mainly weak, polygenic selection



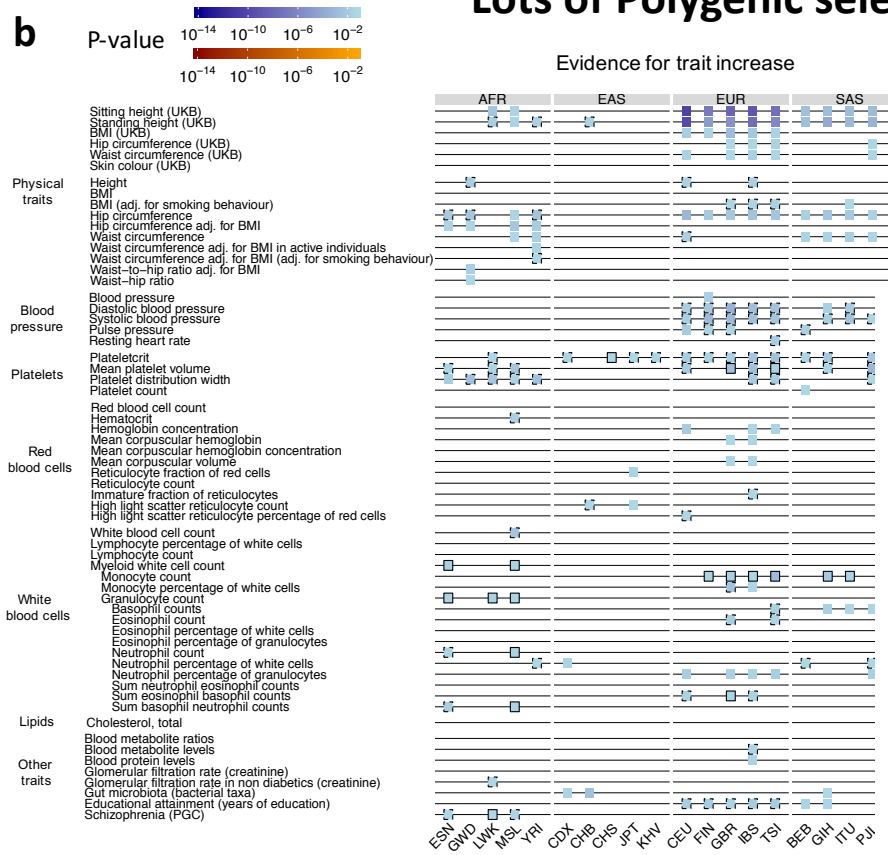
How does weak selection evidence vary by trait?

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

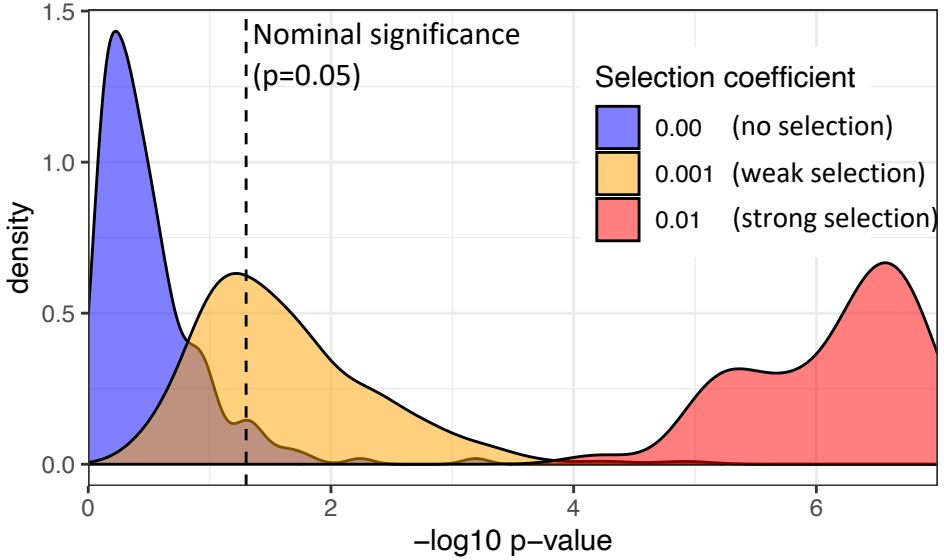
Selection p-values: only a handful of "genome-wide significant" loci



Lots of Polygenic selection signals

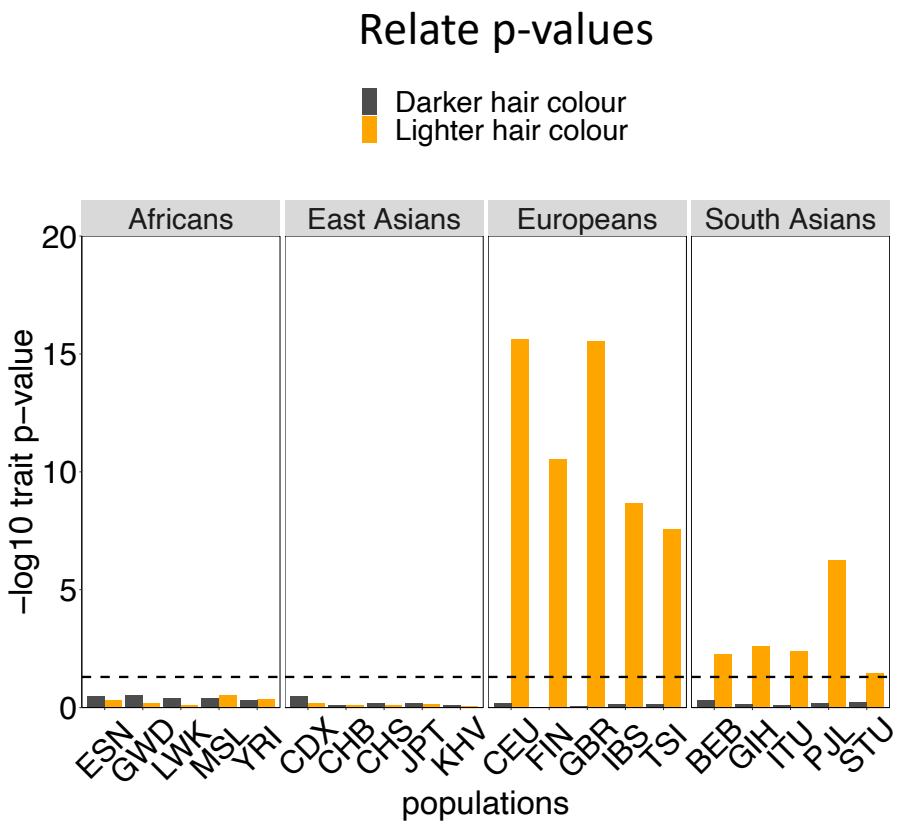


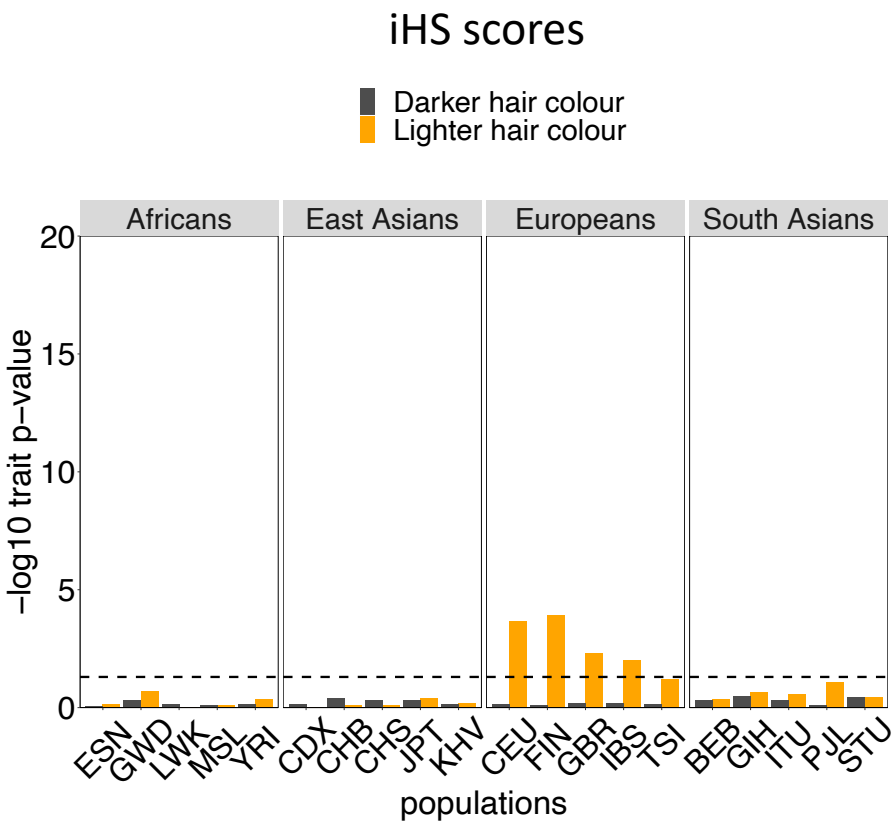
Weak selection on individual mutations is hard to detect, compared to shifts in distributions



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)





Summary & 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
  - Migrations and ancient introgression
  - Mutation rate evolution
  - Trait evolution
  - (and many more things)
- Lots of scope for more methods using inferred genealogies:
  - Many existing approaches are based on idea of coalescent trees
  - These can often be adapted work directly with trees!
  - Deeper analysis of varying types of selection
  - Ghost ancestries, directional migration etc
  - Recombination evolution
  - Etc

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

Thanks!

Simon Myers
Garrett Hellenthal
Pontus Skoglund
Aaron Stern
Lara Cassidy
Robbie Davies
Marie Forest
Sinan Shi
Sile Hu
Anubha Mahajan
Andrew Morris
Mark McCarthy
Daniel Falush
++