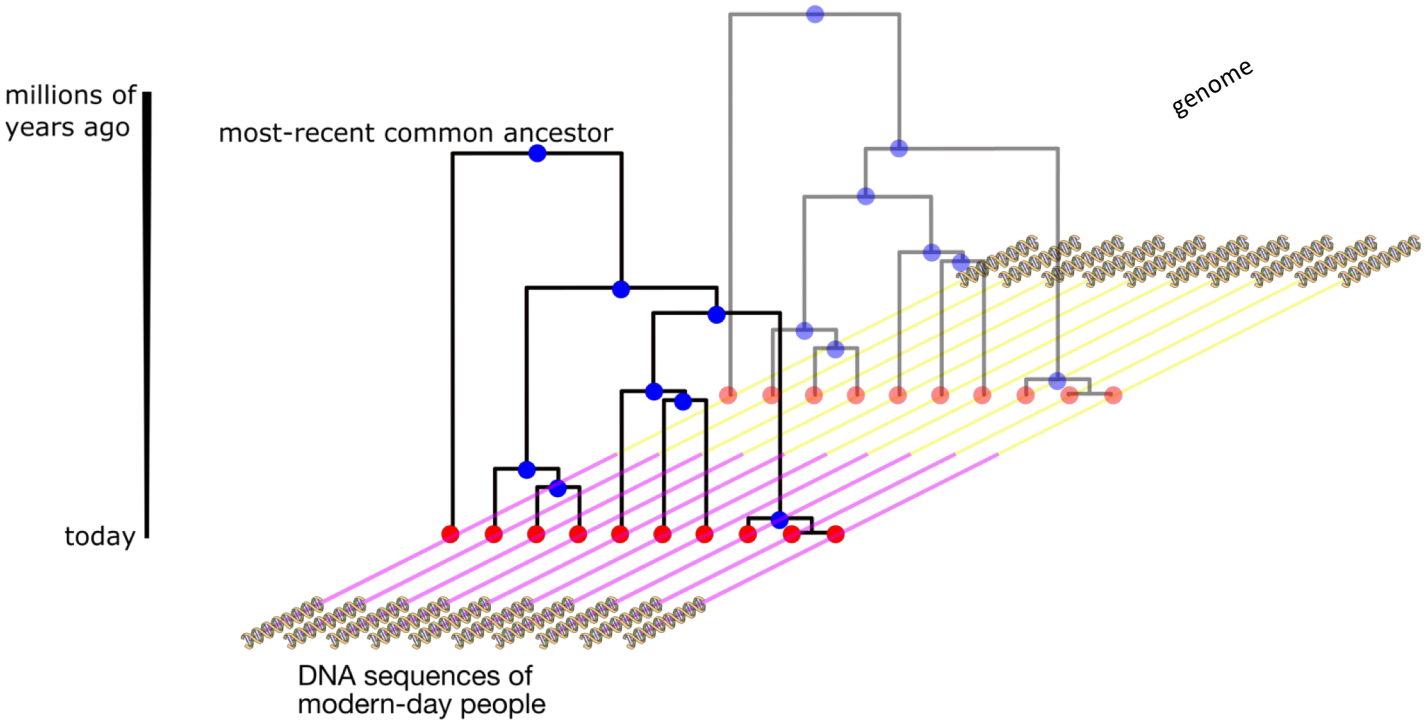

UNIVERSITY OF
OXFORD
DEPARTMENT OF
STATISTICS
wellcome
Seeing the forest from the trees
Simon Myers
Photo: Yunli Song

Leveraging genealogies to understand evolutionary processes

We are genetically related through a sequence of genealogical trees, varying along the genome



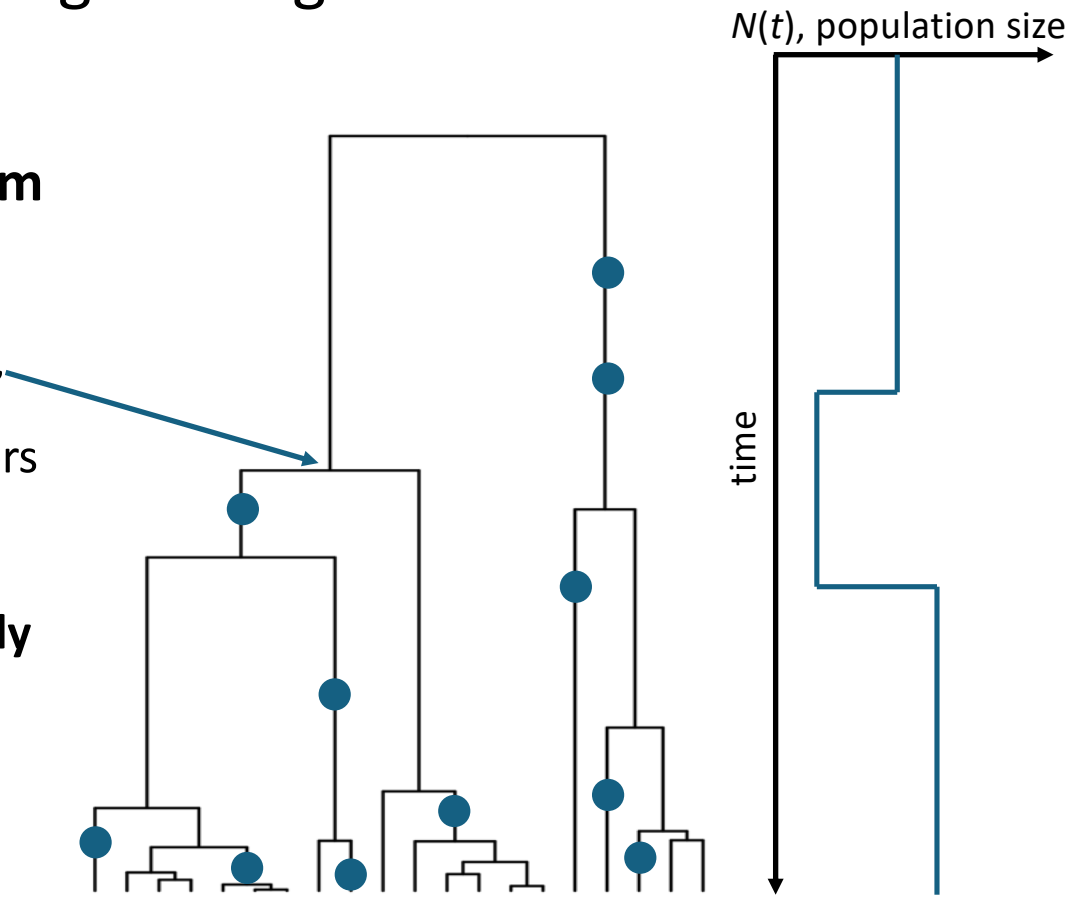
For humans, we can learn about events from present day to >3M years ago

The coalescent (Kingman 1982) is perhaps the simplest genealogical model

- Consider a single position in the genome
- The simplest model: parents in previous generations are chosen **uniformly at random**
- Ancestral lineages can share ancestors further back in time, so **coalesce**
  - Population size $N(t)$ chromosomes a time t ago, coalescence probability $1/N(t)$ per pair
  - Provided $N(t)$ is large, coalesce while j ancestors

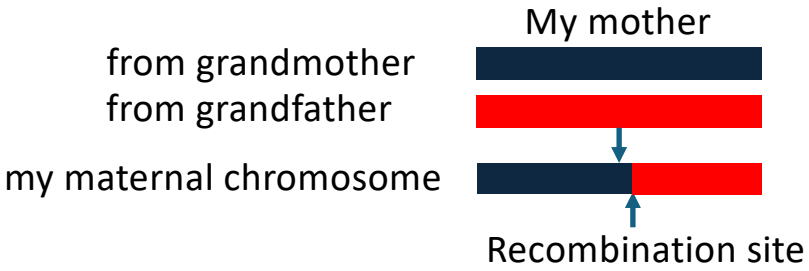
    rate $\binom{j}{2}/N(t)$

    (after suitable scalings)
  - NB: by simple symmetry, **all lineages are equally likely to split, forward in time** (discussed later)
  - Mutations on tree edges (rate $\theta/2$)
- Best inference methods still **use this model**
  - $N(t)$ estimated from the data itself (Relate)
- Genealogies then **vary along the genome**



Why do trees change along the genome?

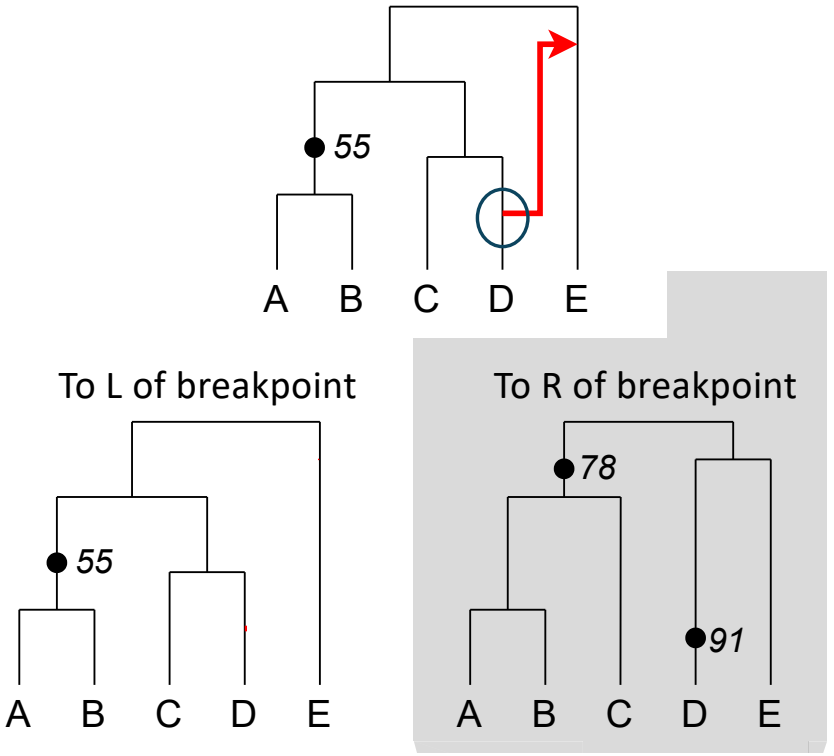
- In us, and our ancestors, a process of recombination occurs (we will discuss this a bit more later)



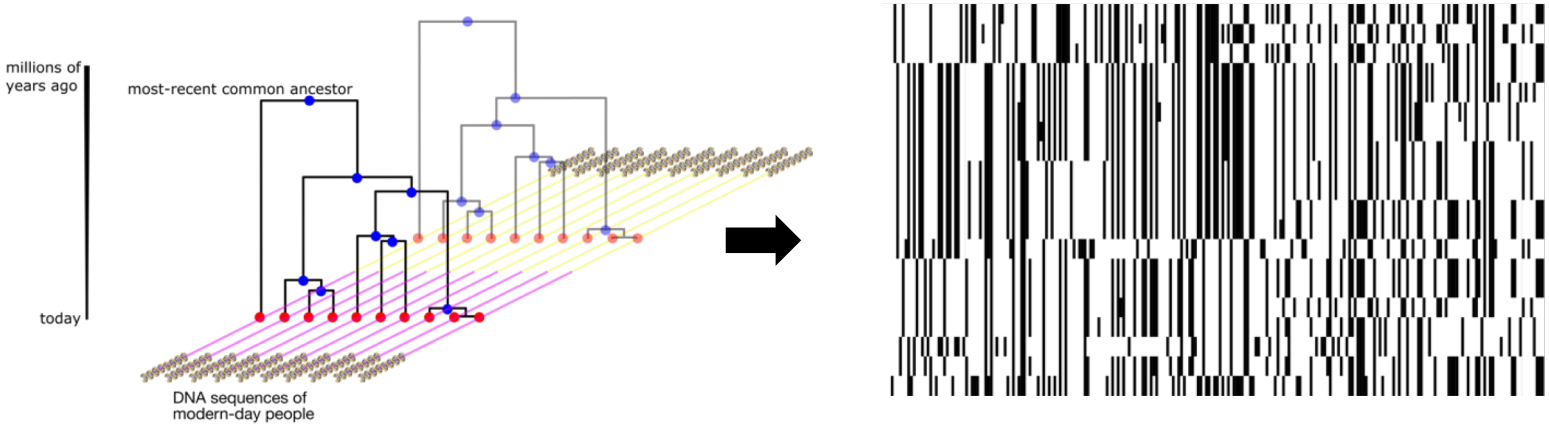
Modelling recombination in the coalescent

- Back in time, in a particular region each of our ancestors has some probability of being formed via recombination

- History becomes a graph (ARG): other events occur as before
- Generates a **sequence of correlated trees** along the genome
- Recombination added to model: **constant** rate $\rho/2$ through time in all ancestors (see later)
- The coalescent with recombination (Hudson et al. 1983, Griffiths and Marjoram 1997)



Many canonical approaches only consider trees <u>indirectly</u>



Demographic history
Splits and speciation
Genetic structure
Mutation, recombination,
Natural selection, etc.

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

In **principle**, trees capture **all the information** available, allowing **self-consistent inference**

Challenges: computationally <u>very</u> challenging to sample trees from the data.
Modern datasets can contain >50,000 individuals and >100,000,000 mutations

Inferring genealogies

Old problem, lots of methods……

……but only recently have methods been developed that can scale

- ARGweaver (Rasmussen et al. 2014)
- Rent+ (Mirzaei et al. 2017)

} Infers Ancestral Recombination Graphs

- Tsinfer+tsdate (Kelleher et al. 2019)
- Relate (Speidel et al. 2019)
- ARG-Needle (Zhang et al. 2023)

} Published since 2019, scale to large sample sizes of thousands+

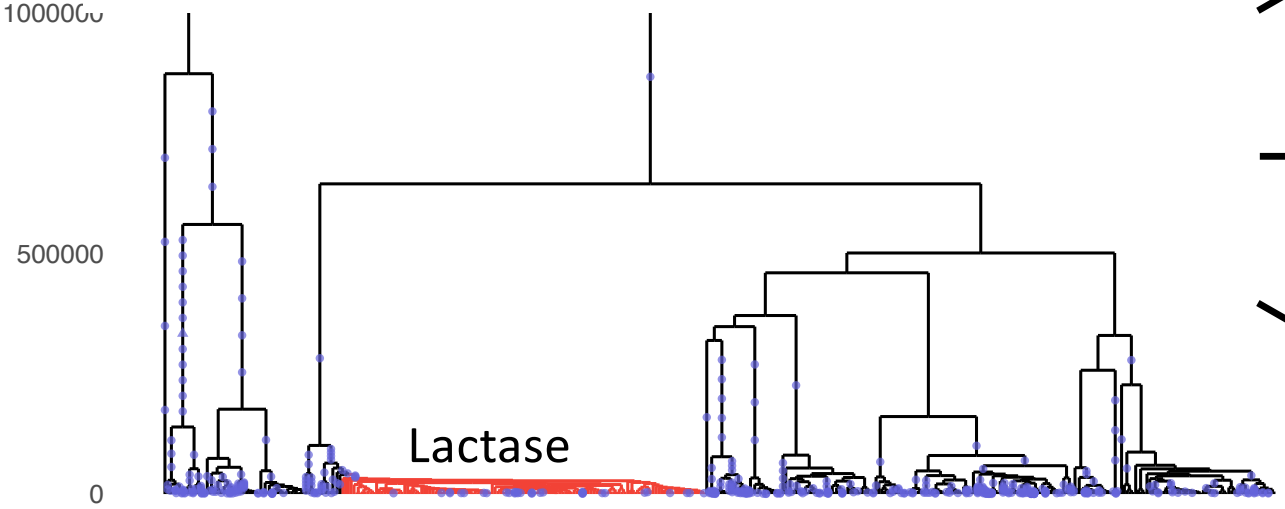
We will focus on Relate today….key ideas largely method-agnostic

Relate

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

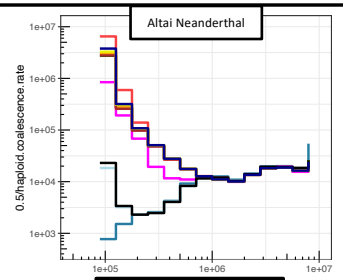
Key features:

- Patterns of mutation sharing ➡ trees genome-wide
- Mutations mapped to edge; branch lengths, $N(t)$ estimated
- Fast & accurate, robust to errors and repeat mutations
- Modern samples and ancient DNA
- Add-on tools for various types of analyses exist

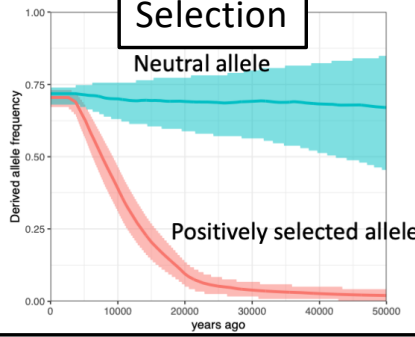


Applied to hundreds of studies on
humans, apes, dogs, birds, fish, plants, bacteria etc!

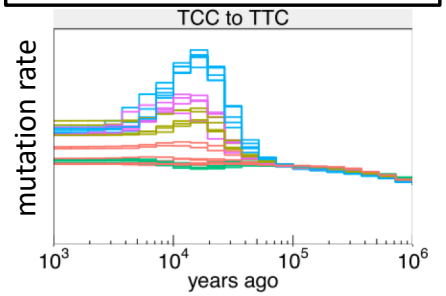
Demographic inference



Selection



Mutation rate evolution



Example: 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)

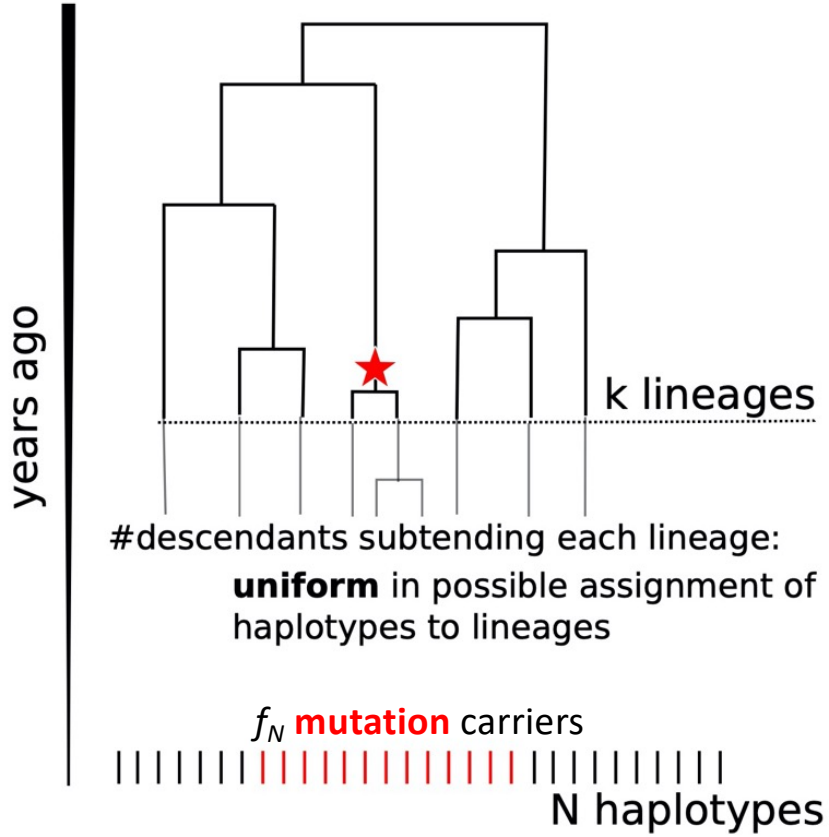
Run time: ~4 days on 300 cores

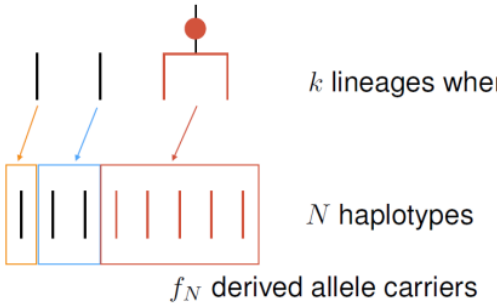
Key ideas

1) We can now analyse many (and in principle, almost all) phenomena relating to population histories by studying real-world trees genome-wide, for thousands of individuals
2) Modelling appropriate **departures** from expectations under the simple coalescent model used to build the trees is key, for many of the most interesting questions

Simple example: **P-values for evidence of positive selection**

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





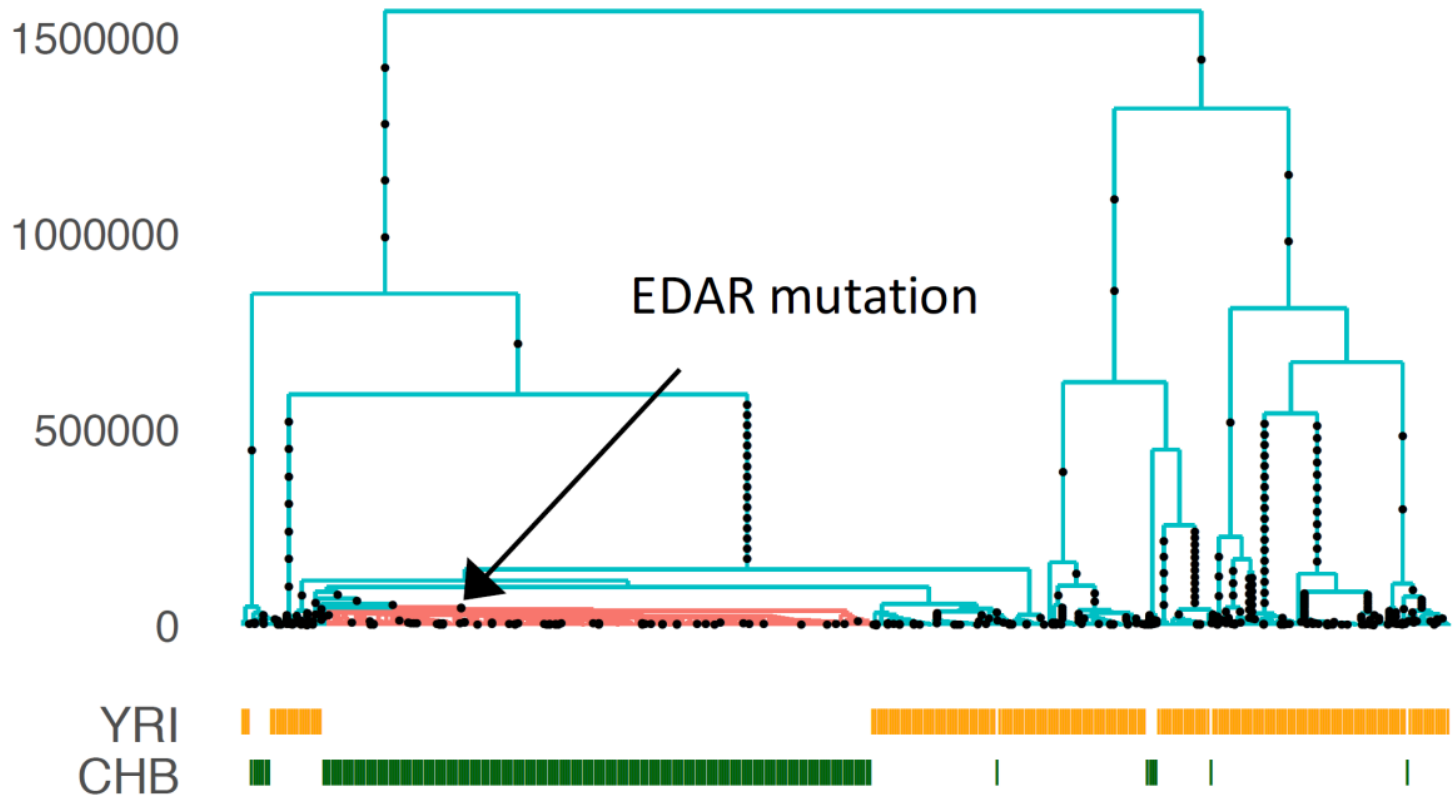
$$P(f_N|k) = \frac{\binom{f_N-1}{2-1}\binom{N-f_N-1}{(k-2)-1}}{\binom{N-1}{k-1}}$$

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

- Very well calibrated in simulated data, and more powerful than **all current non-genealogical approaches**

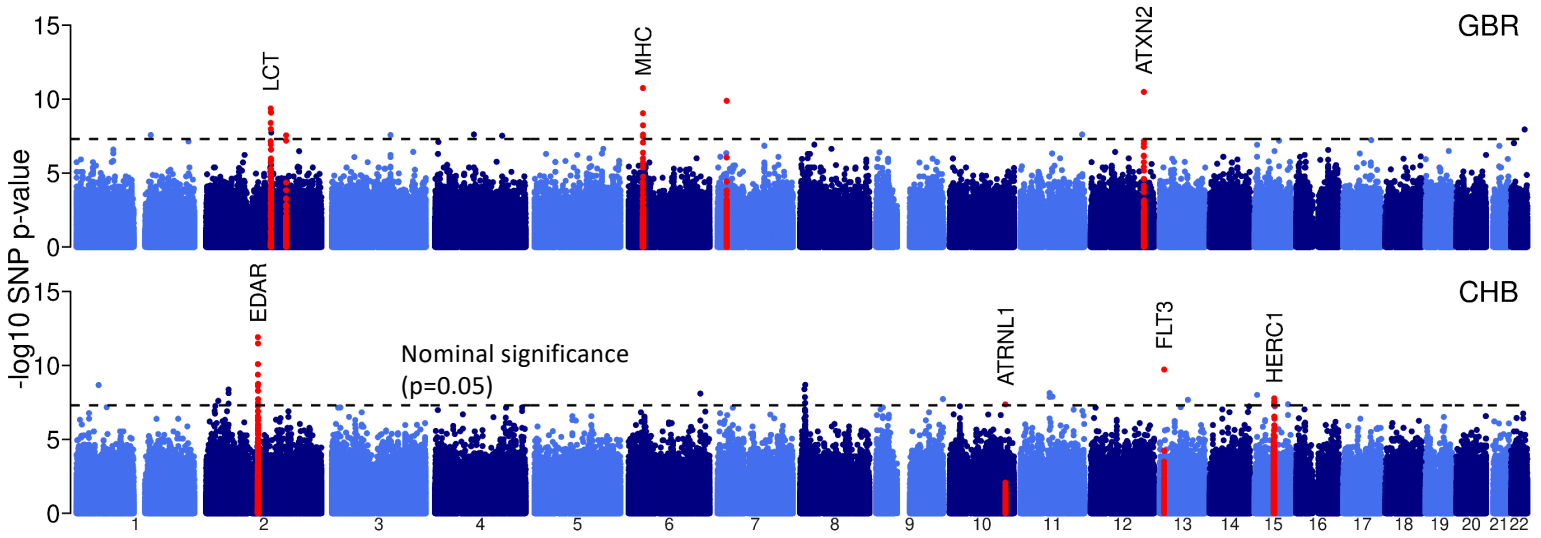
- CLUES: improved sophistication (though slower), Stern *et al.* PLOS Genetics, 2019
- SIA (Hejase *et al.* MBE 2022: Machine Learning)

An example of a mutation showing tree-based evidence of positive selection



Non-synonymous change associated in GWAS with thicker, straighter scalp hair, along with other traits such as shovelling of incisors, altered ear and chin shape, and increased fingertip sweat gland density

Genome-wide selection p-values



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

Evidence for selection on many **traits**, e.g. Lighter hair colour in European populations, Body composition in African ancestry groups, …..

Modelling of departures 2: identifying suppression of recombination using DoLoReS (**Ana Ignatieva**)

Our trees are built under a model that all lineages have undergone recombination at the **same** rate

But might different lineages have **different** recombination rates?

A variety of processes might cause this to happen, including e.g. variable expression during sexual reproduction, **presence of inversions**

Inversions likely suppress recombination in heterozygous individuals

In heterozygotes for an inversion, any recombination in the inverted region produces **non-viable gametes**

In the ARG, if inversion has a unique origin, inversion carriers form a clade

Recombination is suppressed between this clade and others

"Inversion" clade has **longer-than-expected genomic span**, compared with other clades

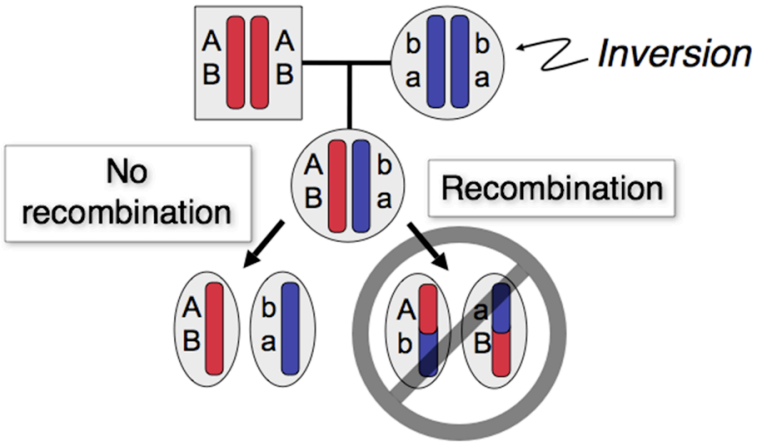
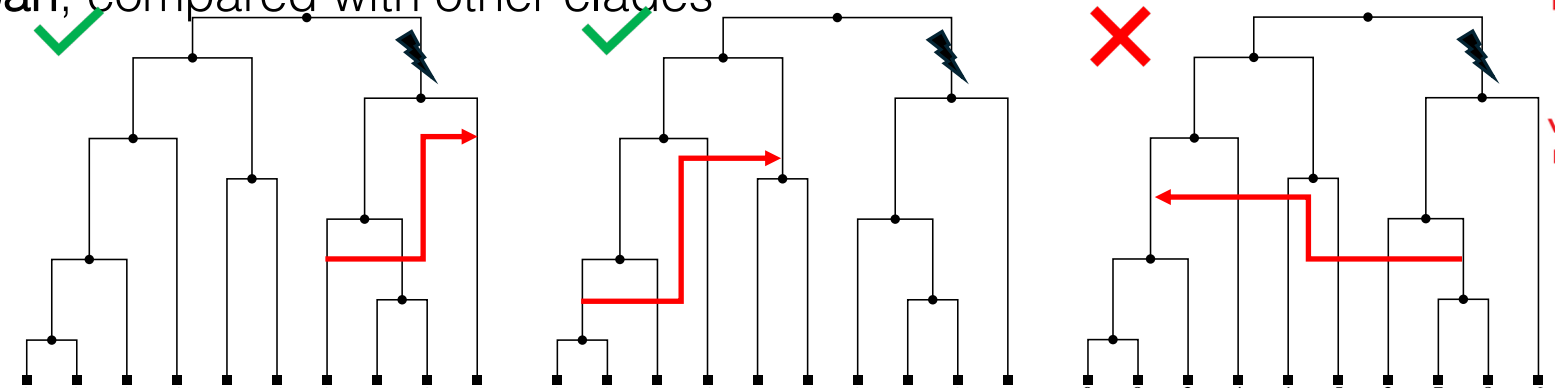
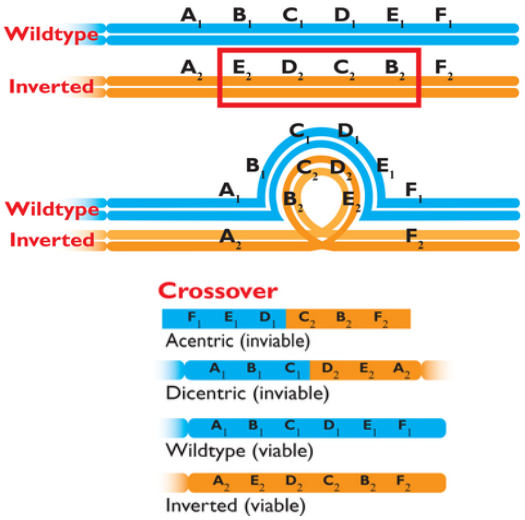


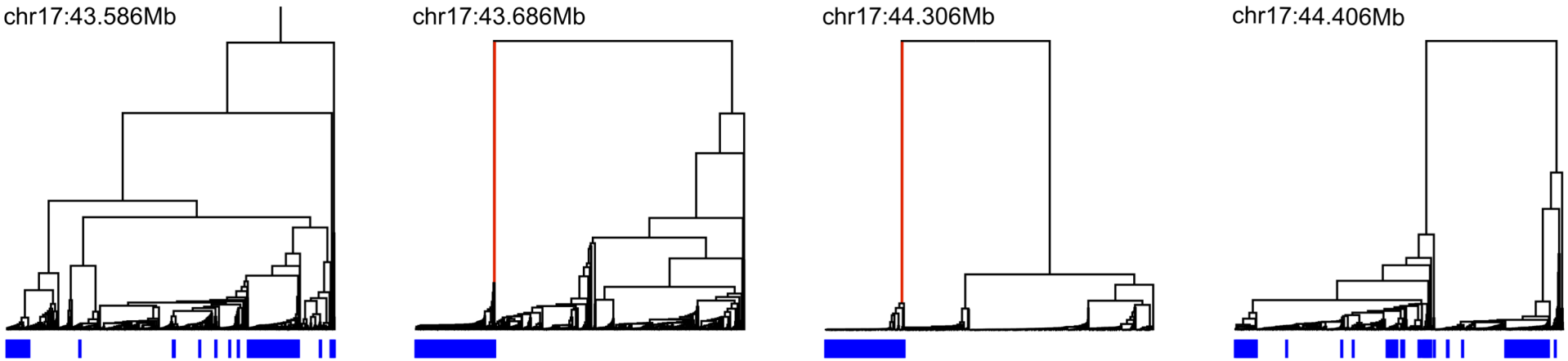
Figure from Kirkpatrick, PLOS Biology, 2010





In reconstructed ARGs, do we see this phenomenon?

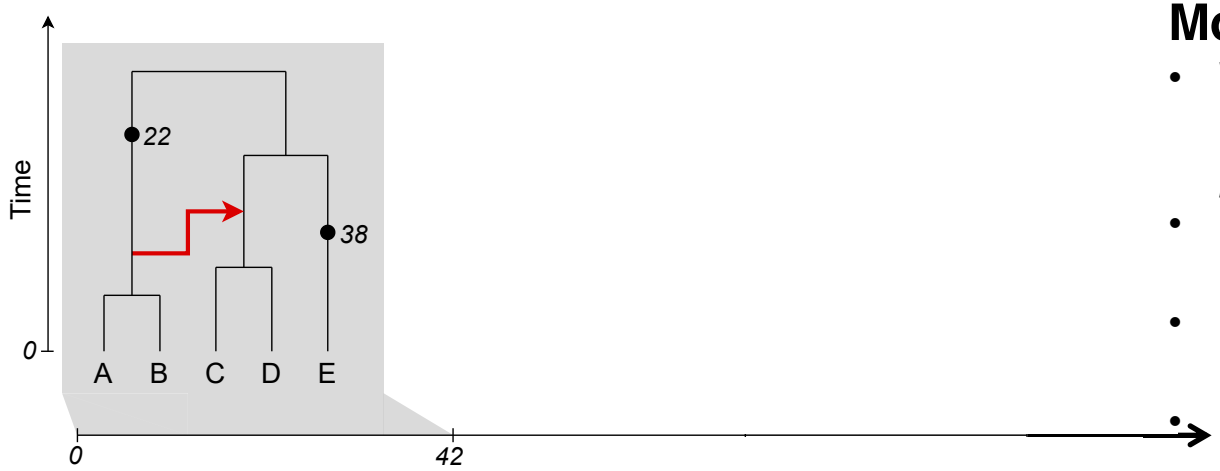
- These are local trees at four positions on human chromosome 17 (1000GP EUR ARG)



- Clade of 242 samples (blue) has a genomic span of around 700kb before it is eventually broken up by recombination
- Perhaps this is longer than expected?
- Obtain survival times of clades under the SMC'

Local trees change along the genome due to recombination

- We leverage SMC': widely used approximation to the coalescent with recombination (accurate, and more tractable for calculations)
- **Markov** changes in local trees along the genome
- Define
  - ρ: population-scaled recombination rate (varies along genome)
  - $\mathcal{L}(T)$: total branch length of tree T



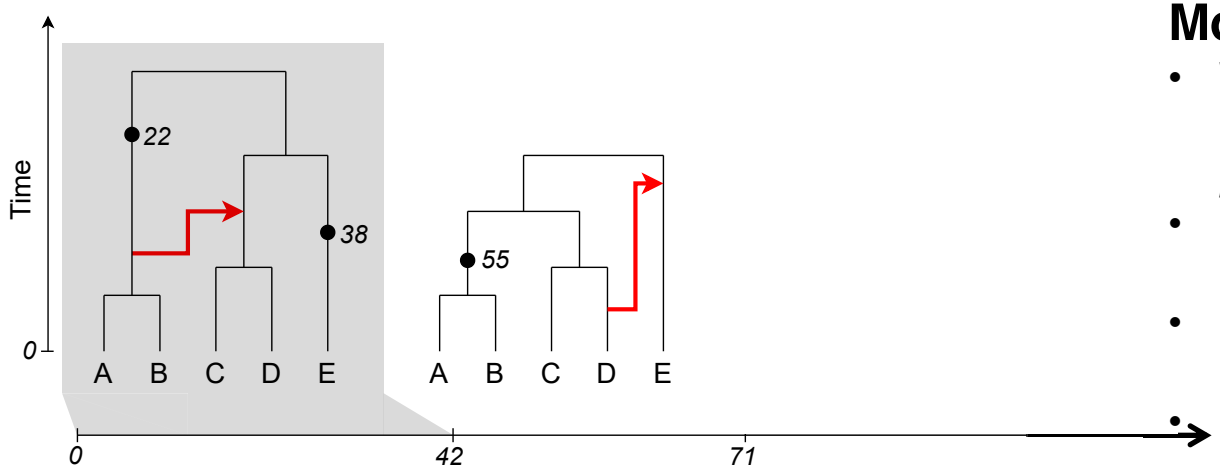
Model

- Waiting distance along genome to next recombination event: exponential, rate $\rho\mathcal{L}(T)/2$
- Recombining ancestor then chosen uniformly at random among tree branches
- Re-coalesce the recombinant branch to generate the next tree along; redefine $\mathcal{L}(T)$
- Repeat

SMC': Marjoram and Wall (BMC Genetics, 2006)

Local trees change along the genome due to recombination

- We leverage SMC': widely used approximation to the coalescent with recombination (accurate, and more tractable for calculations)
- **Markov** changes in local trees along the genome
- Define
  - ρ: population-scaled recombination rate (varies along genome)
  - $\mathcal{L}(T)$: total branch length of tree T



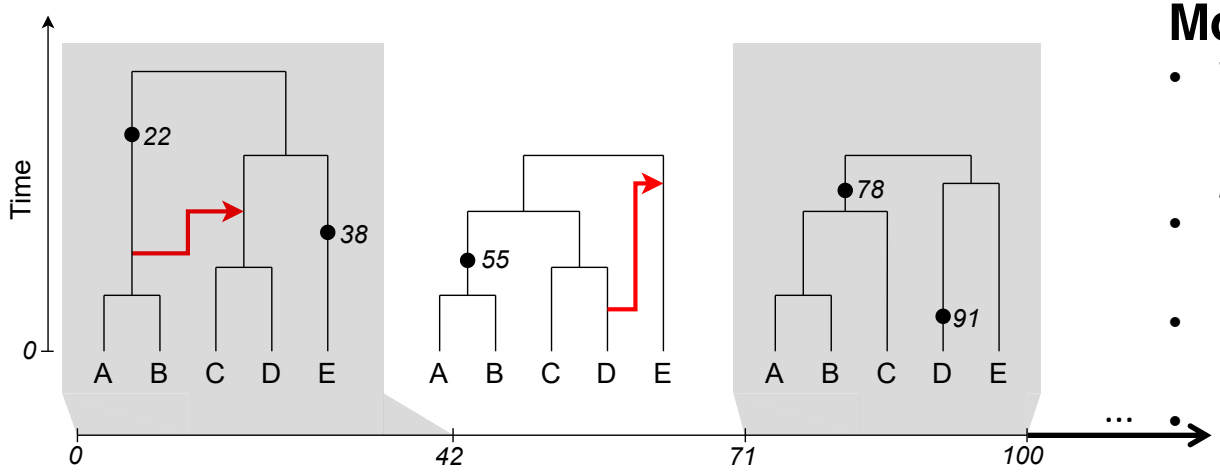
Model

- Waiting distance along genome to next recombination event: exponential, rate $\rho\mathcal{L}(T)/2$
- Recombining ancestor then chosen uniformly at random among tree branches
- Re-coalesce the recombinant branch to generate the next tree along; redefine $\mathcal{L}(T)$
- Repeat

SMC': Marjoram and Wall (BMC Genetics, 2006)

Local trees change along the genome due to recombination

- We leverage SMC': widely used approximation to the coalescent with recombination (accurate, and more tractable for calculations)
- **Markov** changes in local trees along the genome
- Define
  - ρ: population-scaled recombination rate (varies along genome)
  - $\mathcal{L}(T)$: total branch length of tree T



Model

- Waiting distance along genome to next recombination event: exponential, rate $\rho\mathcal{L}(T)/2$
- Recombining ancestor then chosen uniformly at random among tree branches
- Re-coalesce the recombinant branch to generate the next tree along; redefine $\mathcal{L}(T)$
- Repeat

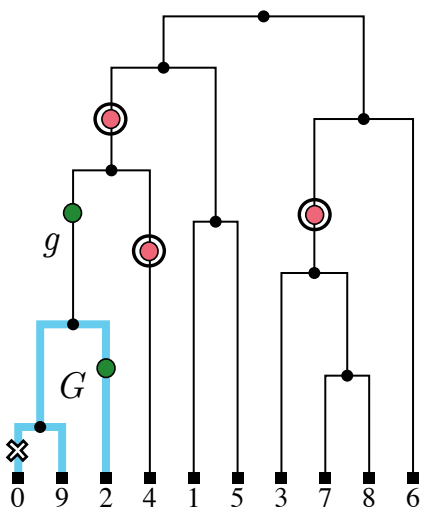
SMC': Marjoram and Wall (BMC Genetics, 2006)

Probability that a clade is broken up by next recombination event

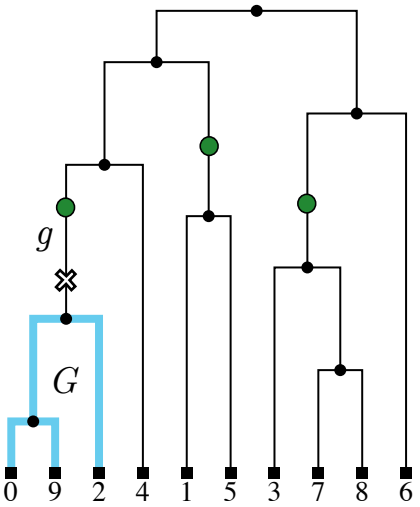
- Clade G contains samples 0, 2, 9
- Subtended by branch g

- x = recombination point, R
- o = coalescence point, C

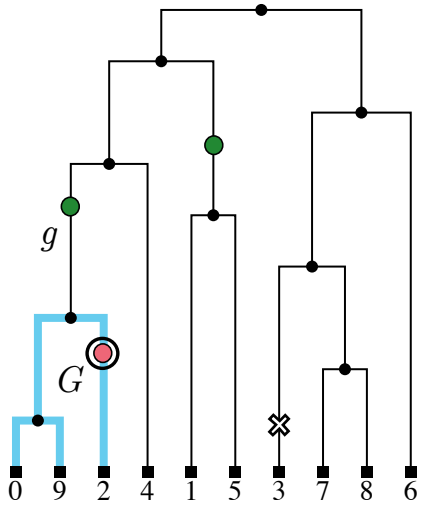
- green o = G <u>not</u> broken up
- red ⊙ = G <u>is</u> broken up



Recombination within clade



Recombination on subtending branch



Recombination elsewhere

$$P(G \text{ broken} \mid T) = P(R \in G) \cdot [P(C \notin G \cup g | R \in G)] + P(R \notin G \cup g) \cdot [P(C \in G | R \notin G \cup g)]$$

Probability that a clade is broken up by next recombination event

$$P(G \text{ broken} \mid T) = P(R \in G) \cdot [P(C \notin G \cup g | R \in G)] + P(R \notin G \cup g) \cdot [P(C \in G | R \notin G \cup g)]$$

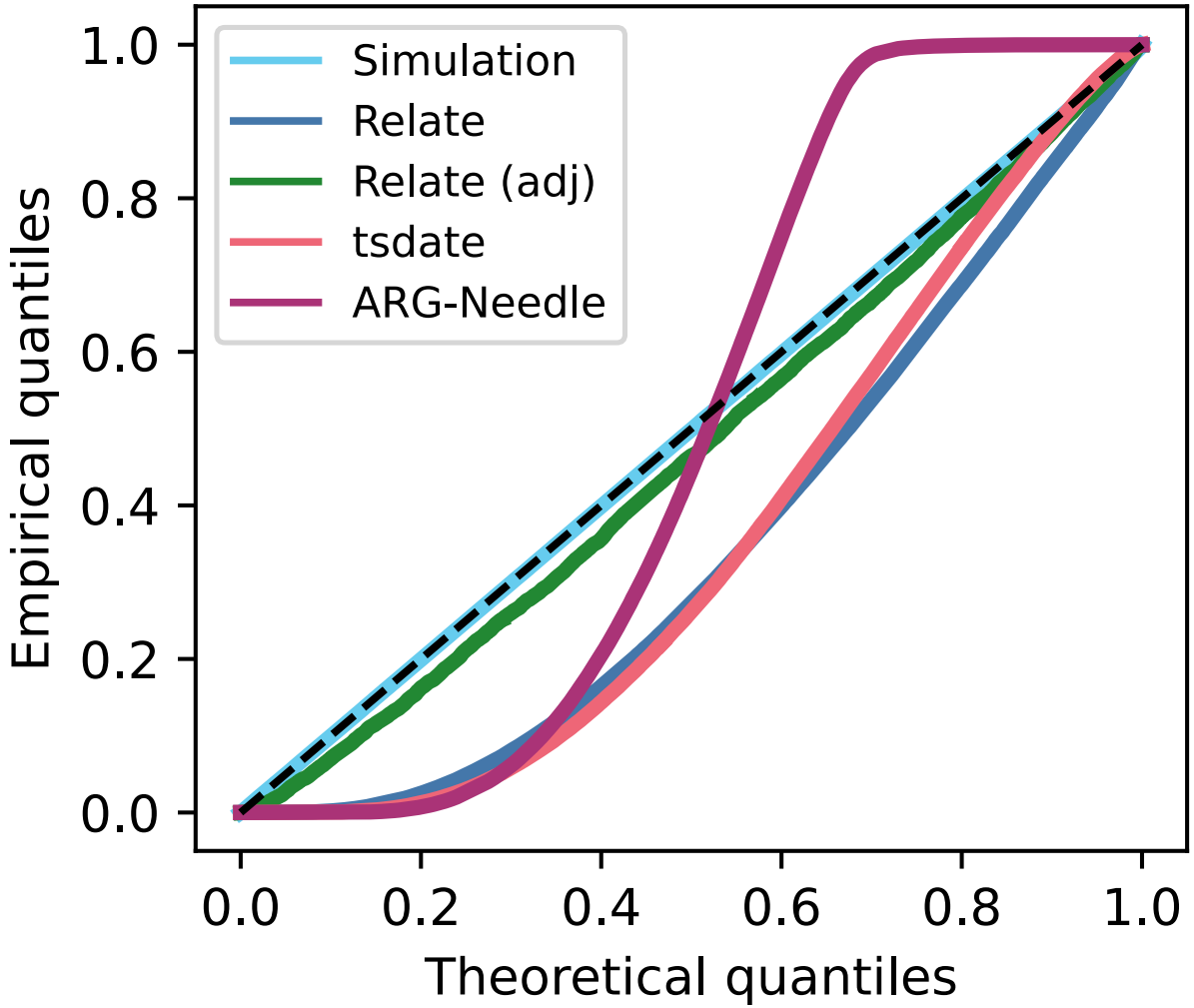
- Can calculate each term
- The expected instantaneous rate at which G is broken up at position τ is $P(G \text{ broken} \mid T) \cdot \rho \mathcal{L}(T)/2$
- Genomic span of G is then waiting time in an inhomogeneous Poisson process whose rate changes along the genome with the local trees T
- For a tree T , let d be span of clade G in recombination (ρ) units

Test 1: p-value: the probability of observing a clade span greater than d, $p = e^{-q}$

Test 2: p-value for number of clade changes (i.e. recombination events) elsewhere on the tree that each clade "survives" (differing false positives in real-world data)

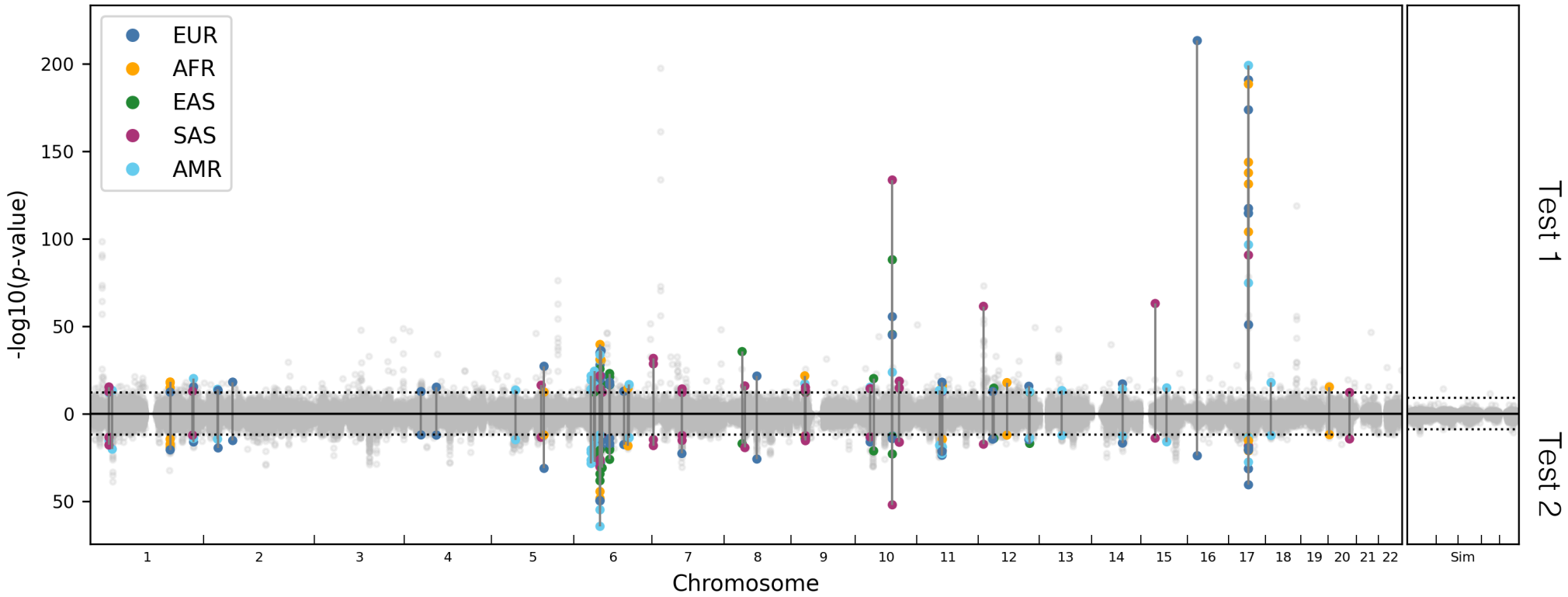
Adjustments to allow for: homogeneous process approximation, varying recombination rates, ARG reconstruction artefacts (use SNP span and “smooth” clades), phasing errors, population structure

Simulation results show imperfect reconstructed genealogies, but for Relate, adjustments fix!



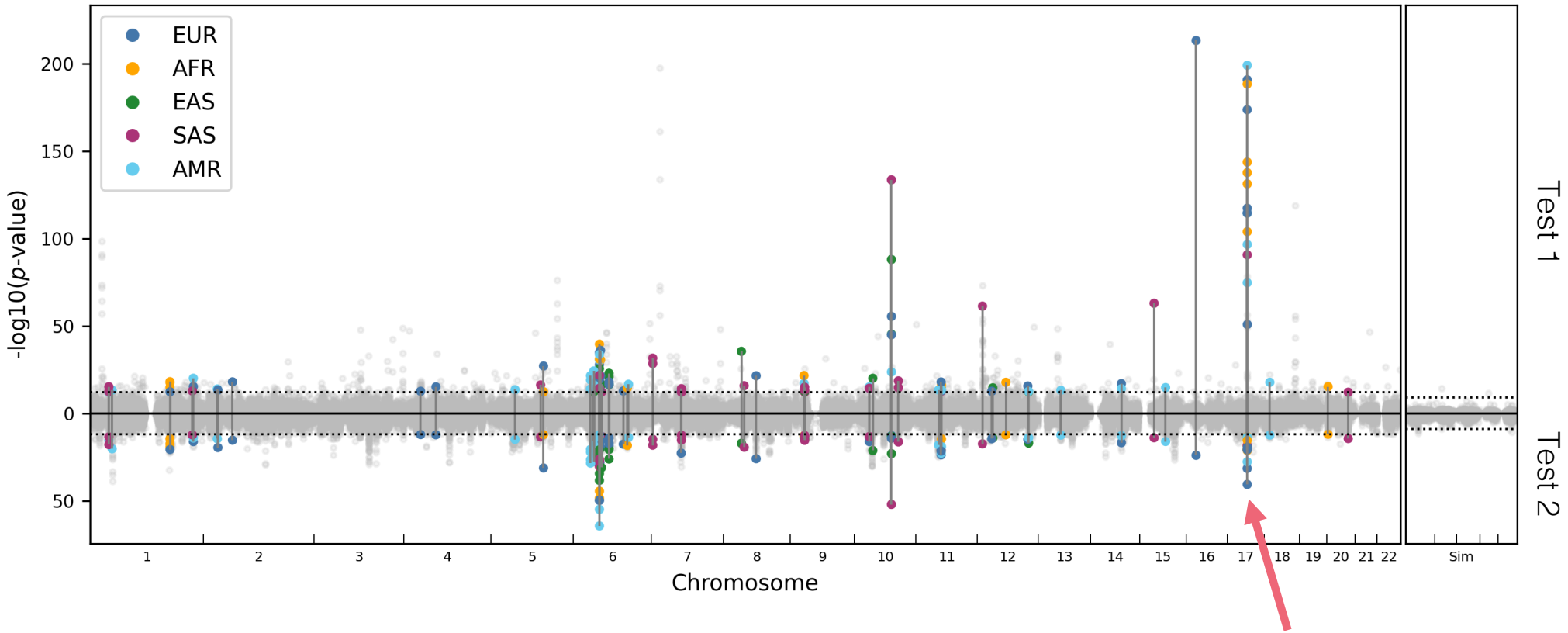
- Simulated ARGs under the null of homogenous recombination rates with stdpopsim (n=100, chromosome 21 recombination map, constant population size: variable size is similar)
- *p*-values using true ARGs are uniformly distributed (light blue points lie on the diagonal), so **SMC' approximation works well**
- Distribution of genomic spans of clades in **reconstructed** ARGs depart strongly from null expectations (all methods)
- However, for Relate, **adjustments** largely fix this; green line

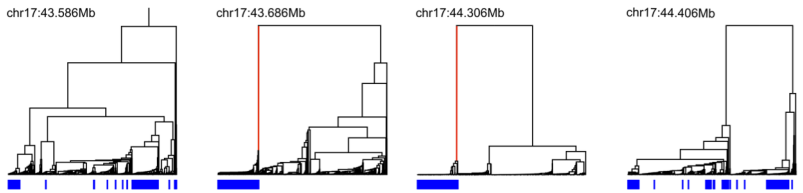
Application to the 1000 Genomes Project ARGs



- Testing all common clades, for all trees in the genome – each point = one test, using Test 1 (above 0 line) and Test 2 (below 0 line)
- Dotted black line = Bonferroni-corrected significance threshold
- Points significant using both tests: coloured by corresponding population

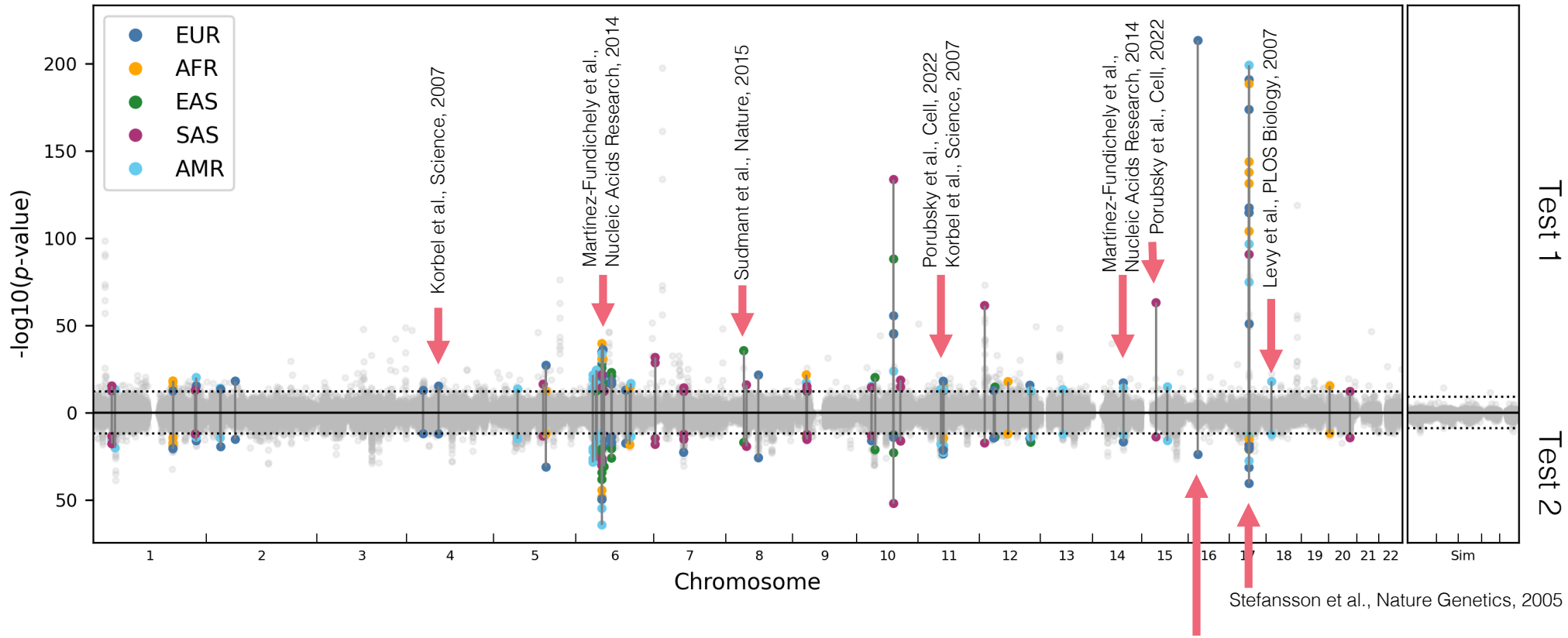
The top hit in the genome is a well-known inversion



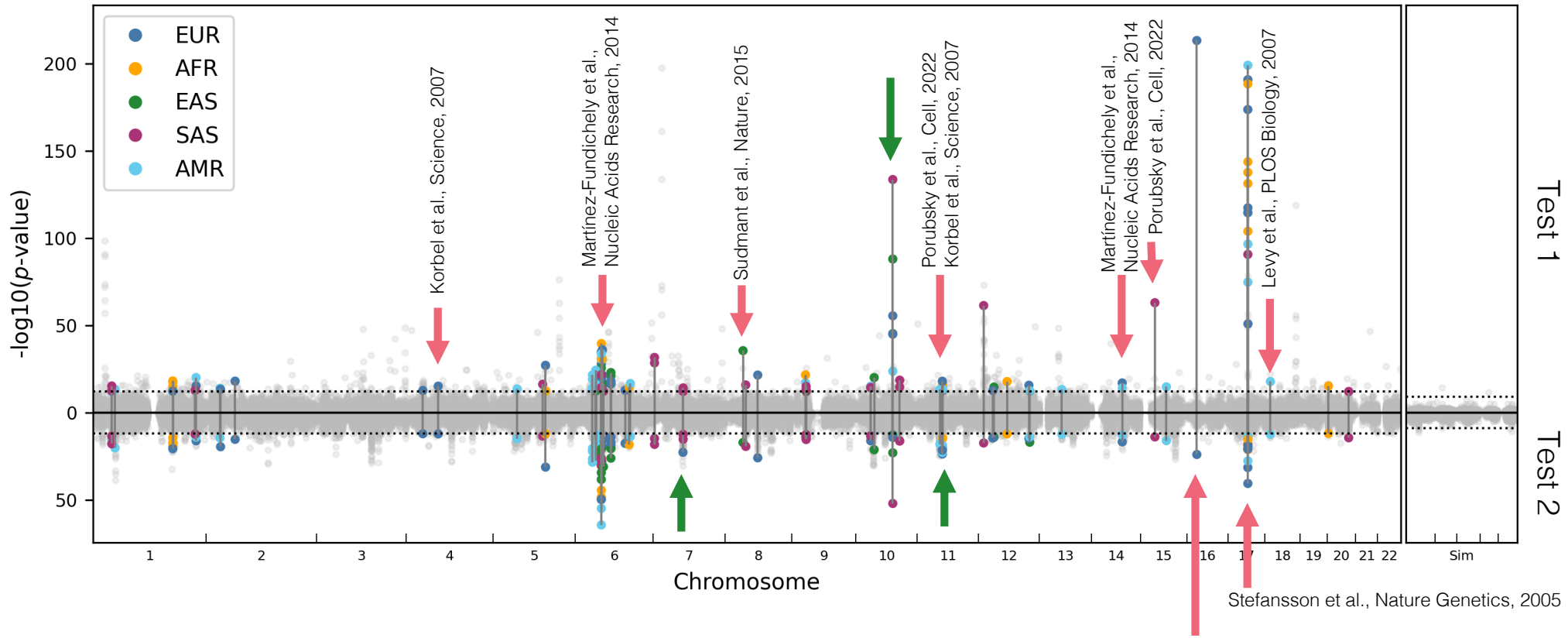


This is a known (large) chromosomal inversion on 17q21.31
(Stefansson et al., Nature Genetics, 2005)
We can estimate the **age** of the inversion, at >4MY old!

Some of the other hits correspond to known inversions or SVs….



.....but many others novel, and several appear to be new inversions



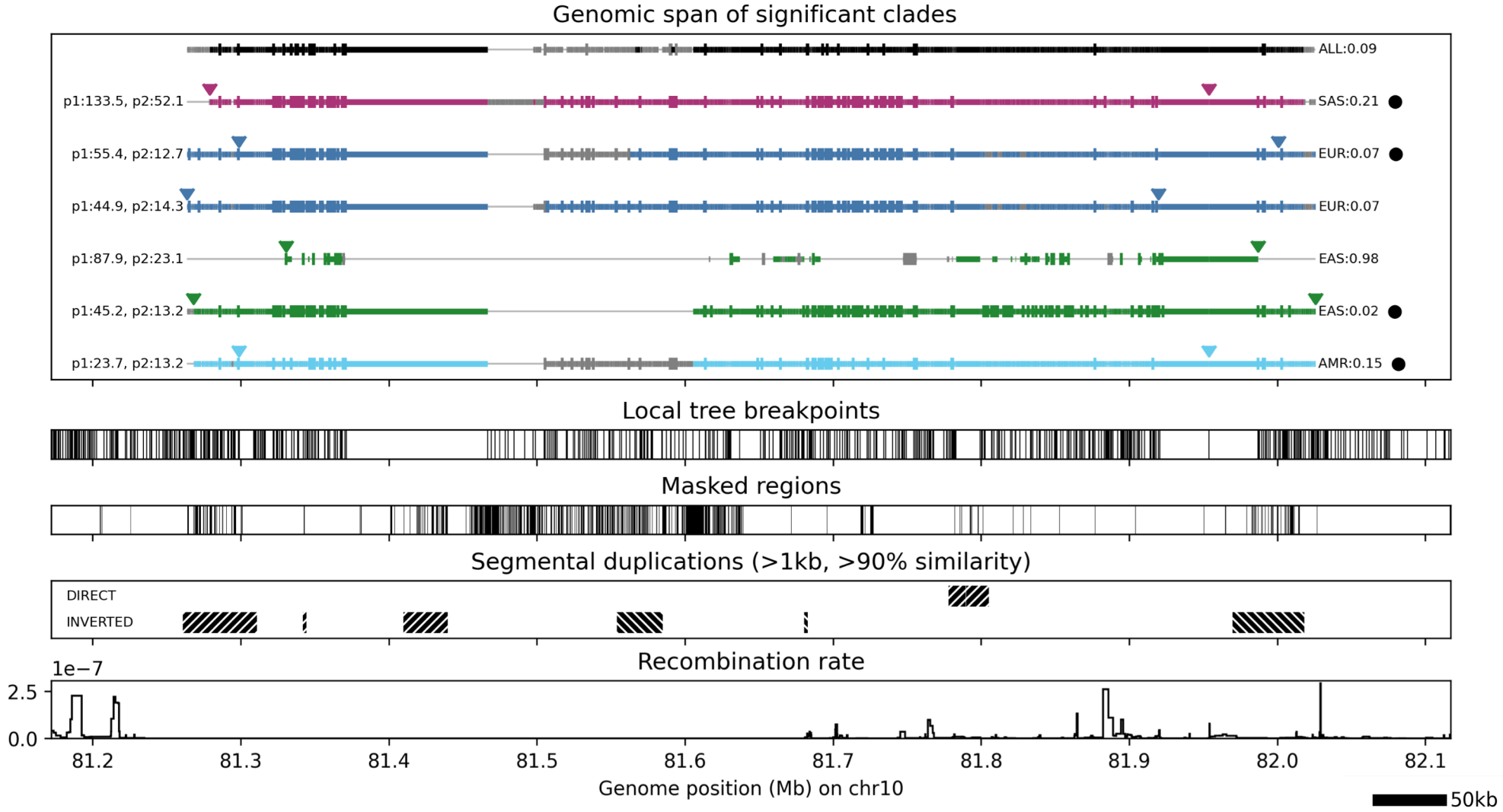
new

known

Known region of complex structural variation
(Antonacci et al., Nature Genetics, 2010)

(still investigating many hits)

A novel signal on 10q22.3 is likely explained by an inversion



Departures part 3: Ghostbuster identifies admixture (Hrushi Loya, Leo Speidel)

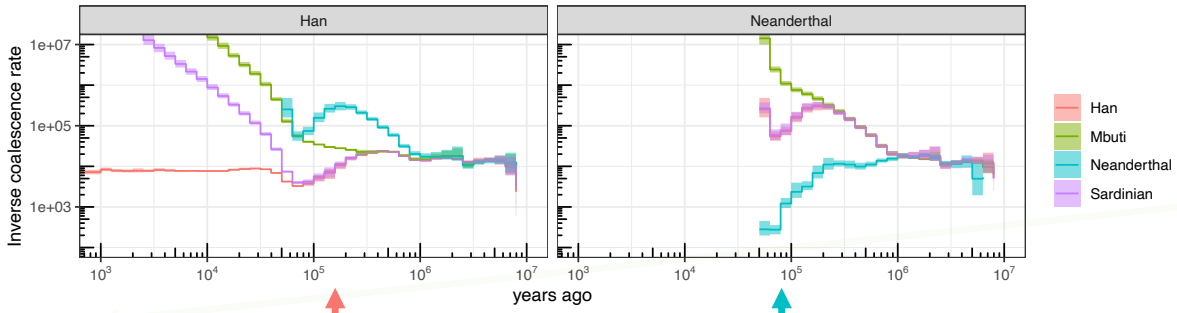
Our trees are built under a model where all pairs of lineages coalesce at the **same** rate, the inverse population size

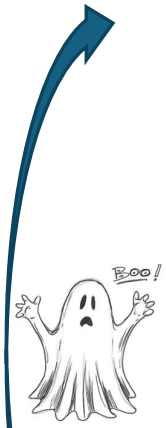
Can we detect different coalescence rates between more closely related groups, or groups with smaller population sizes, within the trees?

GHOSTBUSTER aims to find e.g. Hunter-gatherer or Neanderthal segments within in our genomes due to **admixture**, as well as ghost populations never directly sampled, “hidden” in the trees

GHOSTBUSTER identifies ancestry segments by detecting varying coalescence rates along the genome

After building trees, for any pair of labelled samples we can infer their (inverse) rate of coalescence through time, averaged genome-wide





Neanderthal segments (hidden)

Simulation with Neanderthal-like admixture:

- 4 populations “Han”, “Sardinian”, “Neanderthal”, “Mbuti”
- “Han” & Sardinian have 3% Neanderthal DNA
- “Mbuti” has 0%
- Ran Relate **on all samples together**

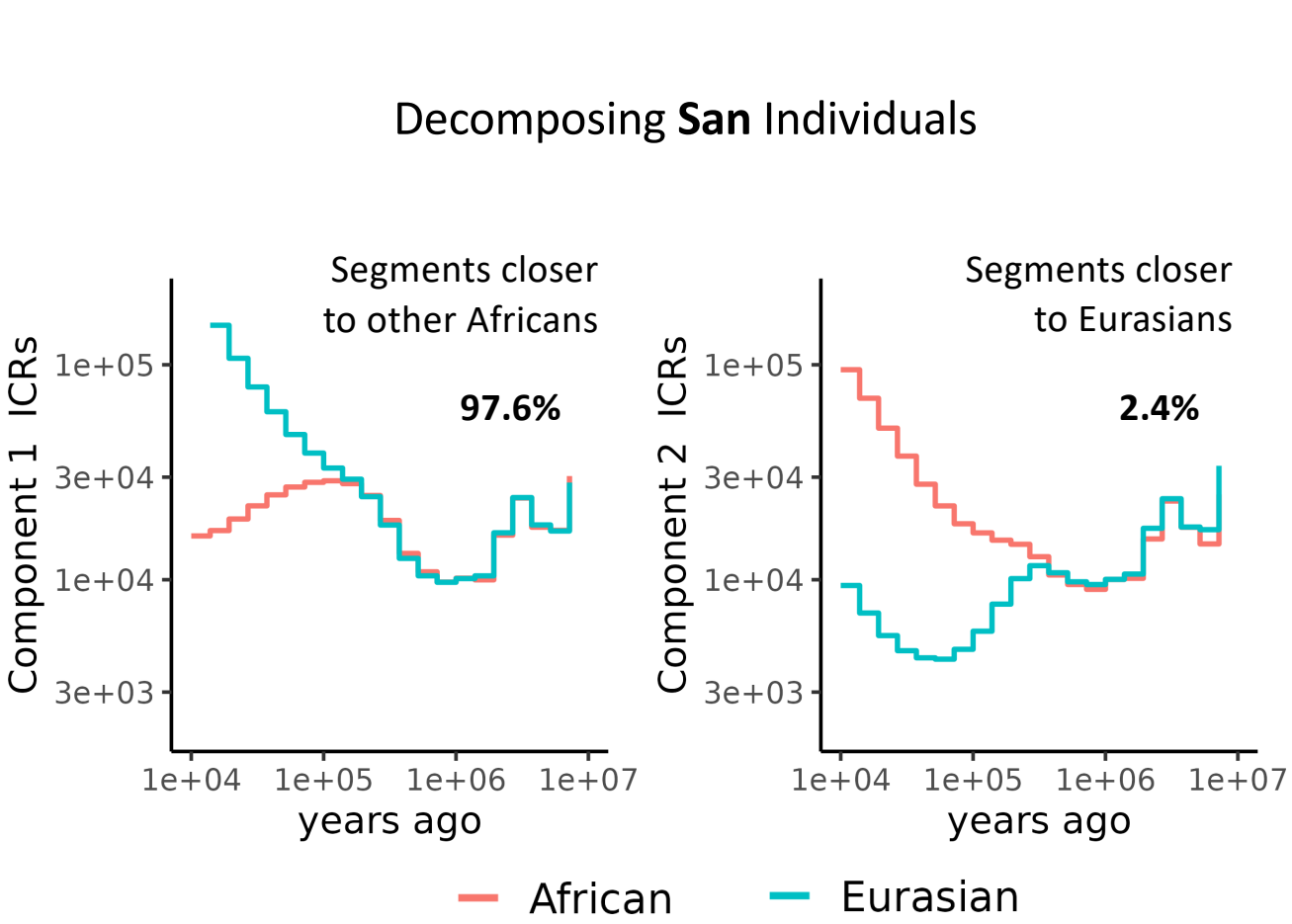
Given a genealogy: GHOSTBUSTER calculates the likelihood for any lineage it comes from the Neanderthal vs. Han population…….infer the Neanderthal segments within a Han individual’s genome!

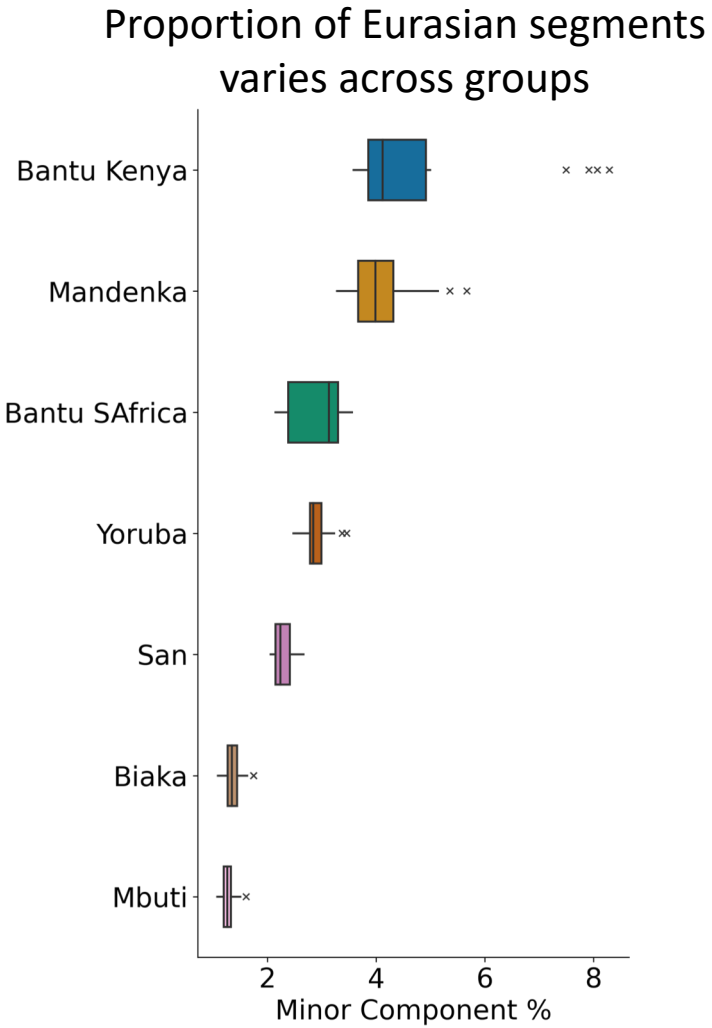
Old mixture events in Africa

i) Old Africa-wide back-migration

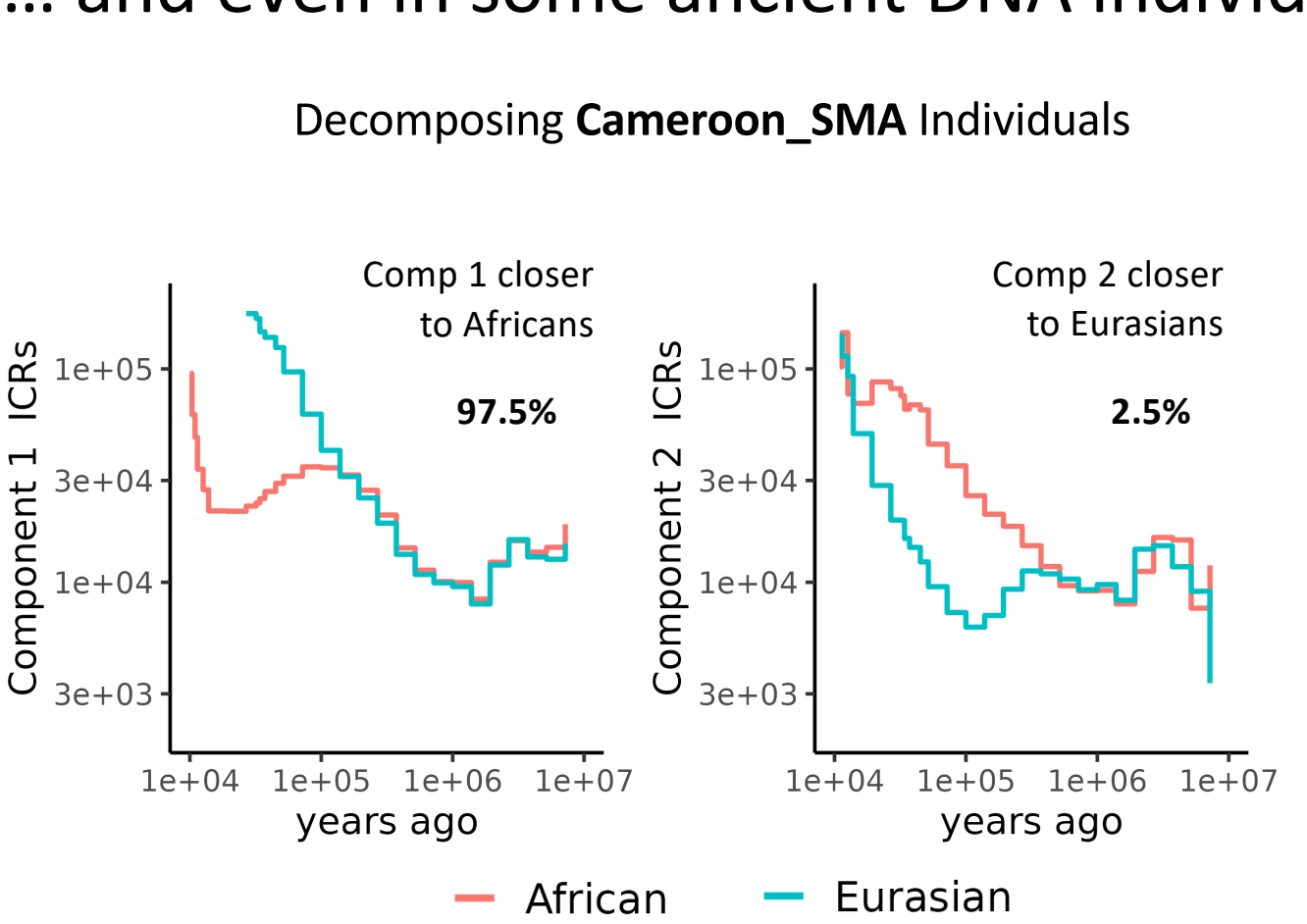
ii) Deep structure within and outside Africa

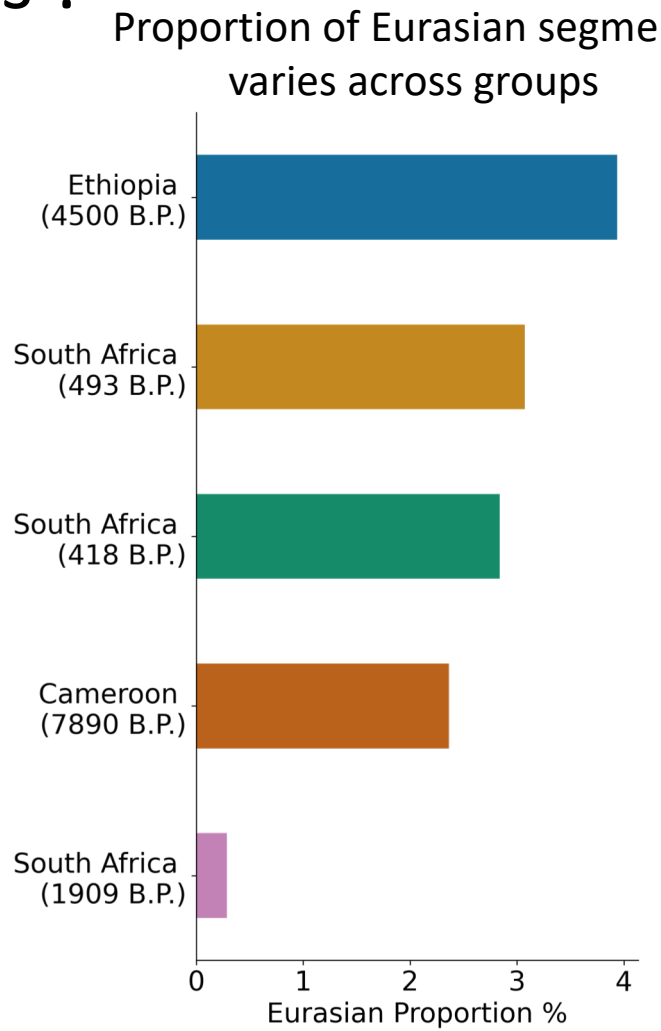
We identify 1-8% Eurasian ancestry in **all** African groups from HGDP – other analyses imply 8-20KY old





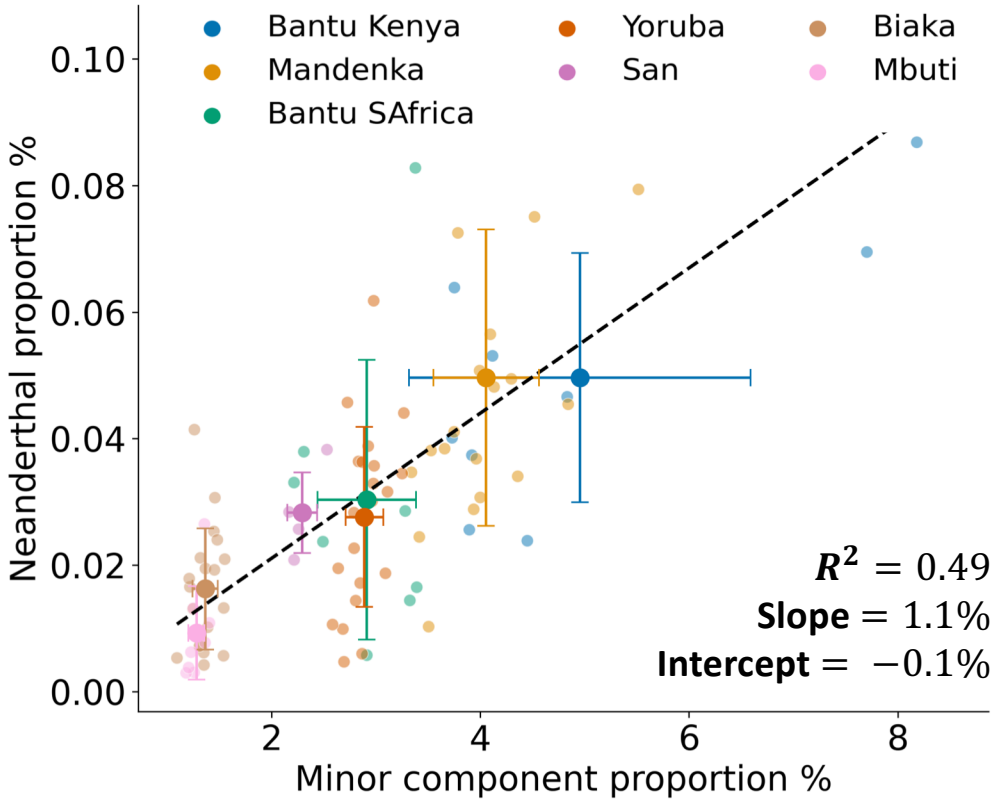
… and even in some ancient DNA individuals !



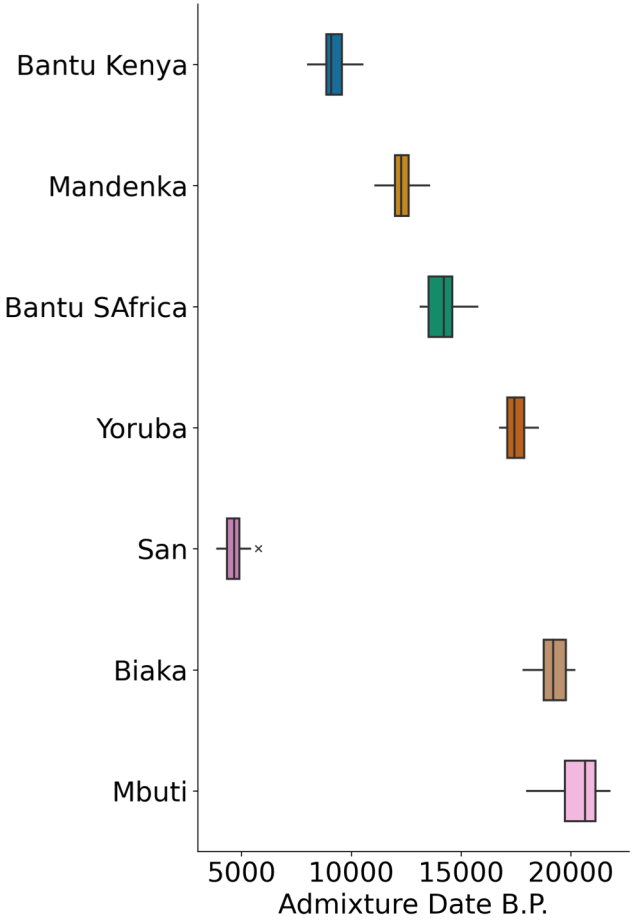


Llorente et al. 2015, Schlebusch et al. 2017, Lipson et al. 2020

The Eurasian ancestry explains the Neanderthal ancestry in Africa & is 5000-20000 years old



Inferred dates for various African groups

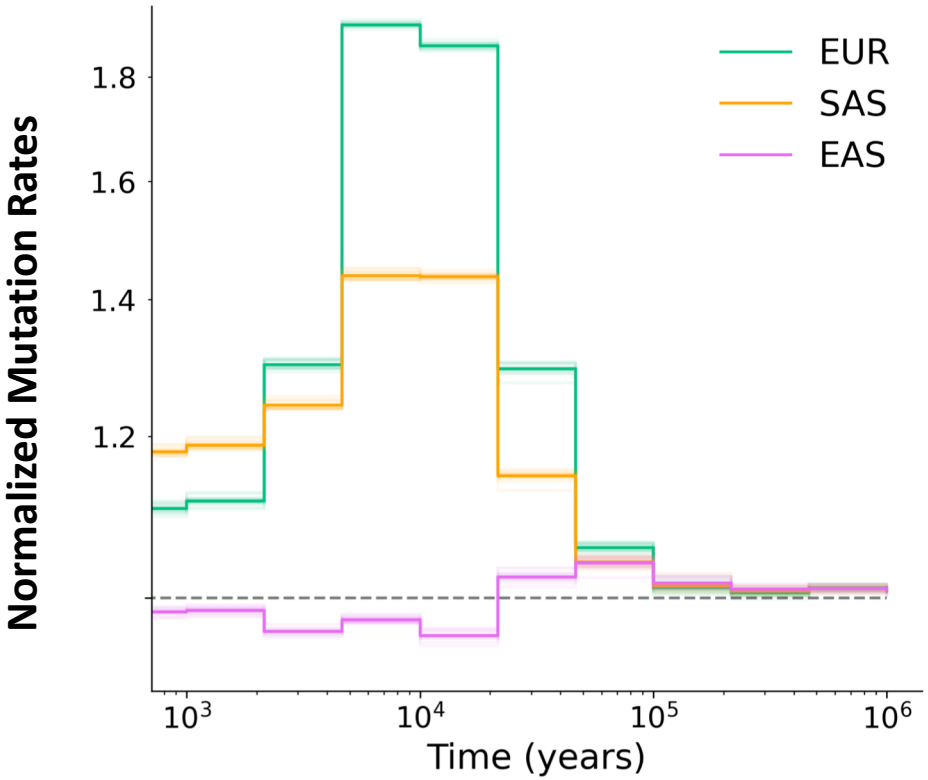


Older than previously suggested?
Pickrell+ PNAS 2014, Chen+ Cell 2020

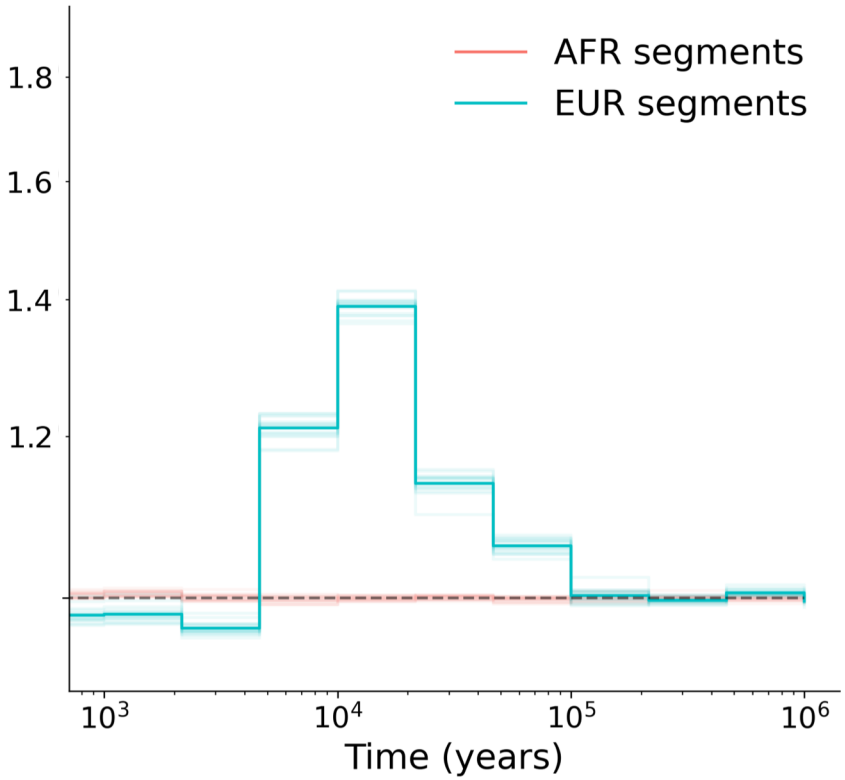
- Comparing the local ancestry, we found ~**9x** enrichment for Neanderthal in Eurasian segments in Africa

Enrichment for TCC>TTC mutations in Eurasian segments

TCC>TTC mutations are enriched in EUR and SAS



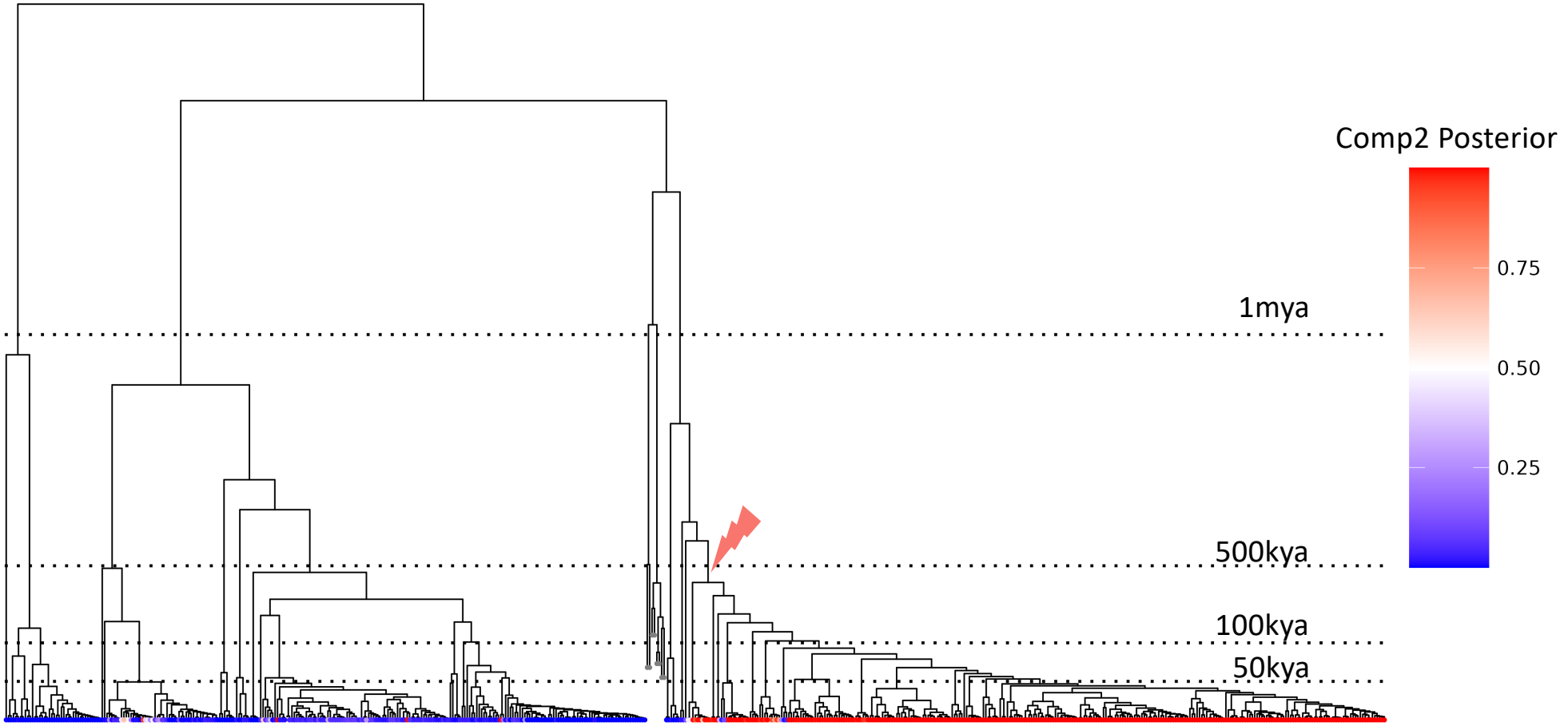
TCC>TTC mutations also enriched in Eurasian segments in Africans



Kelley Harris 2015, Harris & Pritchard 2017, Speidel et al. 2019

Removing the Eurasian segments and looking at deeper times reveals a potential older event, between “ghost” populations diverging ~500,000 years ago

… and still exploring



Summary

- We can now infer joint genealogical trees for modern and ancient people, and other species too
- These are built under (approximations to) the simplest coalescent models
- Departures from these models are hidden within the trees, can themselves be modelled, and offer insights into many phenomena, including
  - **Natural selection**
  - Evolution of mutation rates
  - **Evolution of recombination rates and recombination suppression**
  - **Fine-scale population history, and modern or ancient admixture events**
  - Ghost populations we have not observed
  - Evolution of traits/diseases
  - …..
- Lots of open questions on the formation of present-day genetic structure, the deeper human past, trait evolution & similar stories around the world

Credit: Aina Colomer

Thank you!

Francis Crick Institute, UCL
Leo Speidel

University of Warwick
Jere Koskela
Martina Favero

Glasgow University
Ana Ignatieva

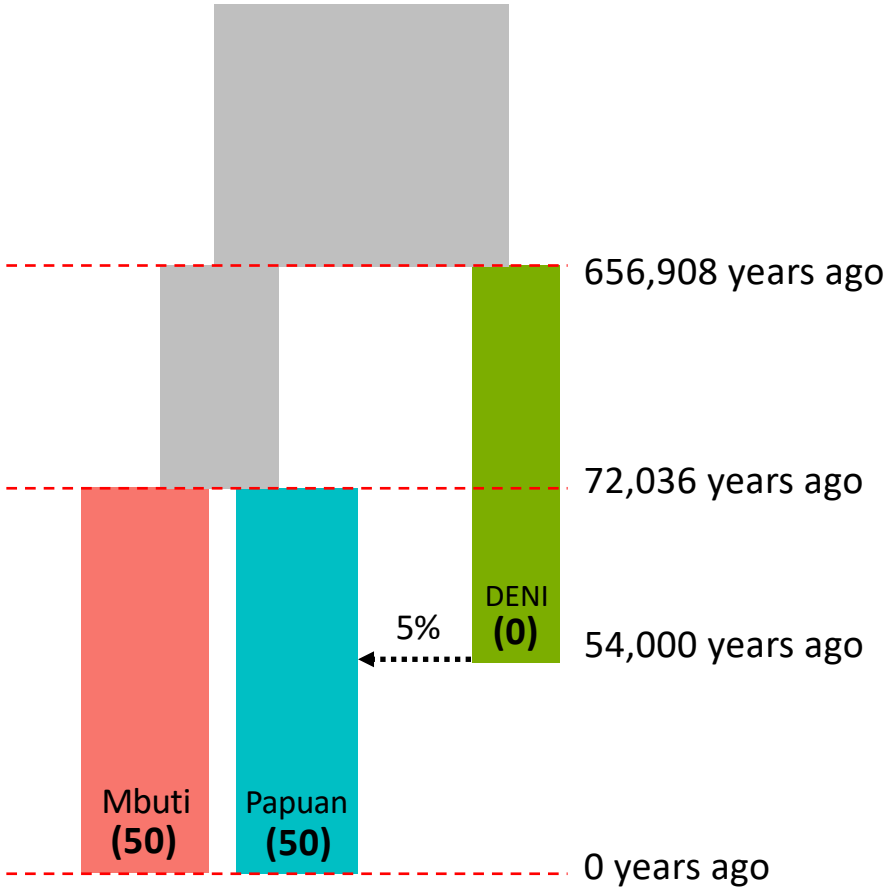
University of Oxford
Ana Ignatieva
Hrushikesh Loya
Jaromir Sant

University of Newcastle
Jere Koskela

Stockholm
Martina Favero

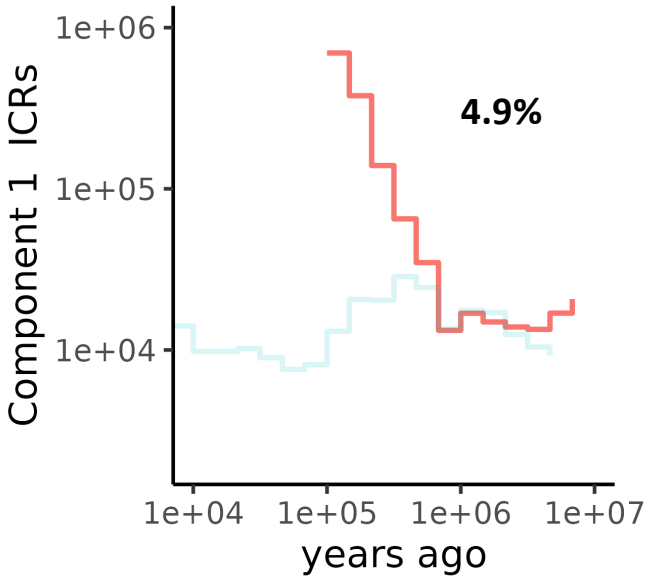
THE FRANCIS CRICK INSTITUTE
UCL
UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS
wellcome
him HAUSDORFF RESEARCH INSTITUTE FOR MATHEMATICS
Knut and Alice Wallenberg Foundation
erc
EPSRC Engineering and Physical Sciences Research Council

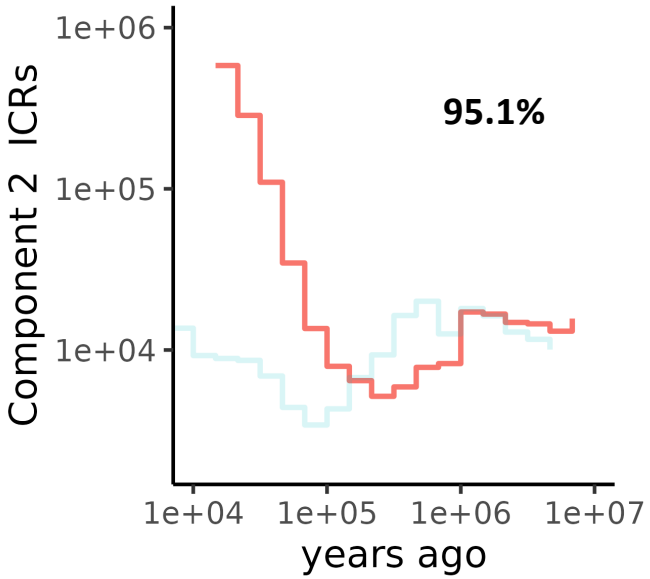
Simulation: Finding Denisova-like ancestry in Papuans without any Denisovans

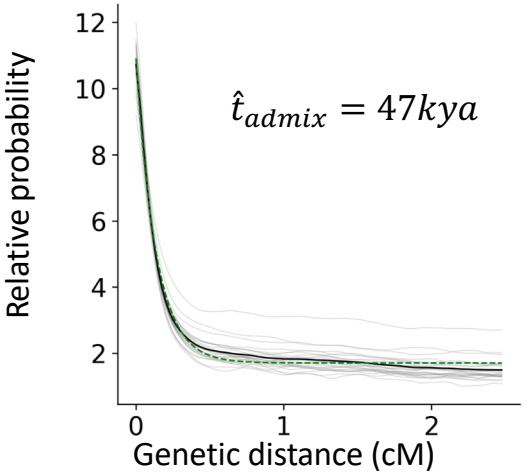


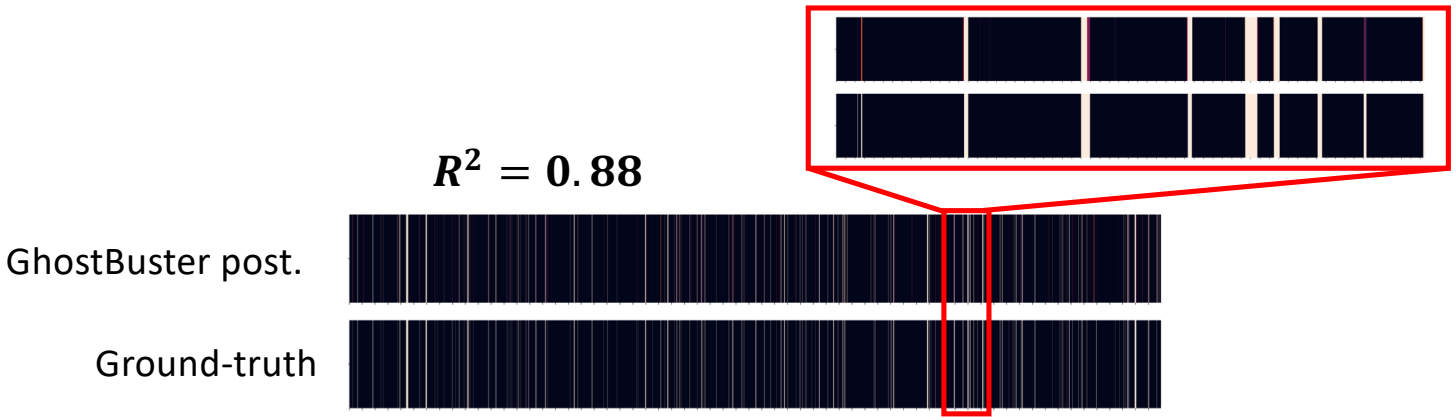
Simulations: Skov et al. 2018

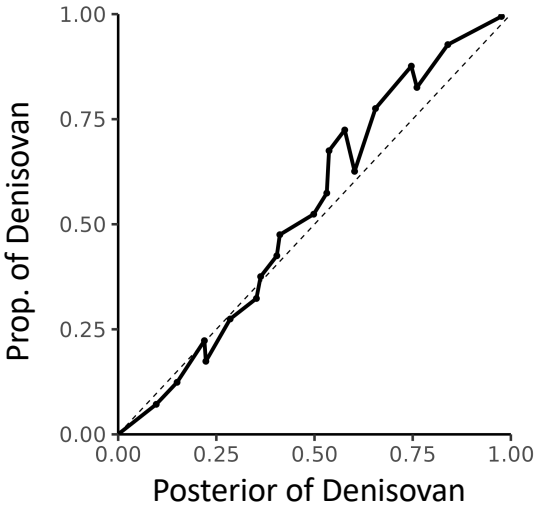
Simulation: Method finds Denisova-like segments without using a Denisovan reference genome









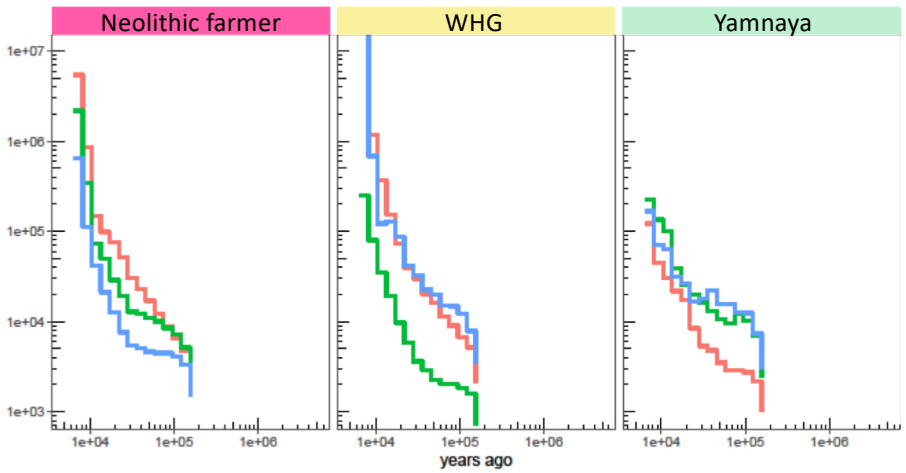


Reconstruct Steppe genome even w/o Yamnaya reference population

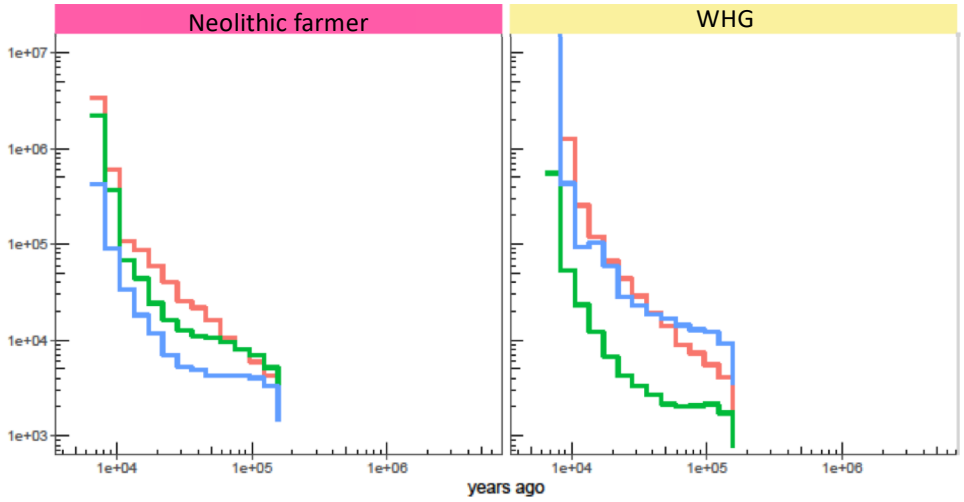
Haak et al, Nature 2015

Target: Ireland_BA (Cassidy et al, PNAS 2016)

Farmers, WHG and Yamnaya reference populations



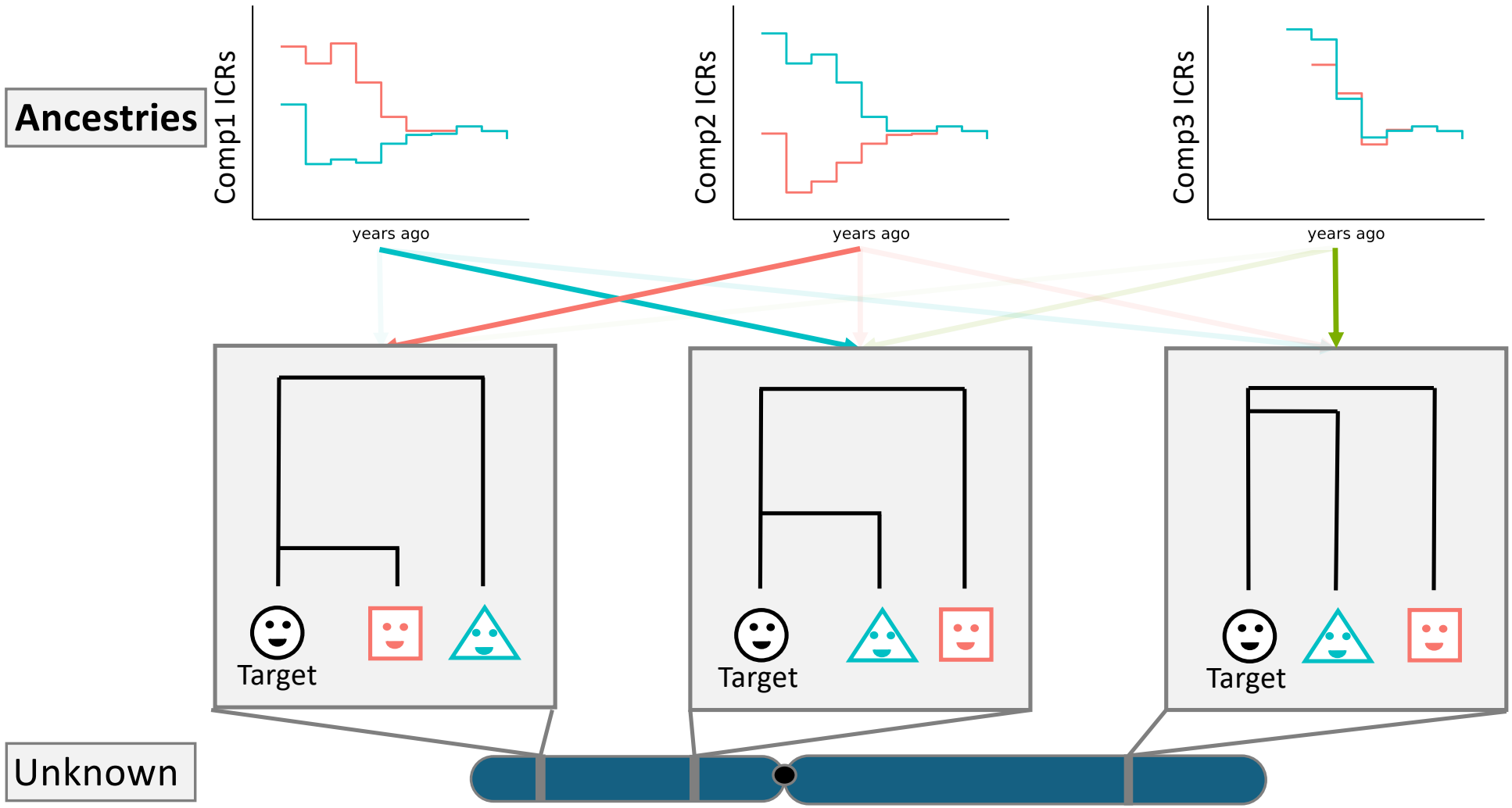
Farmers and WHG reference populations only



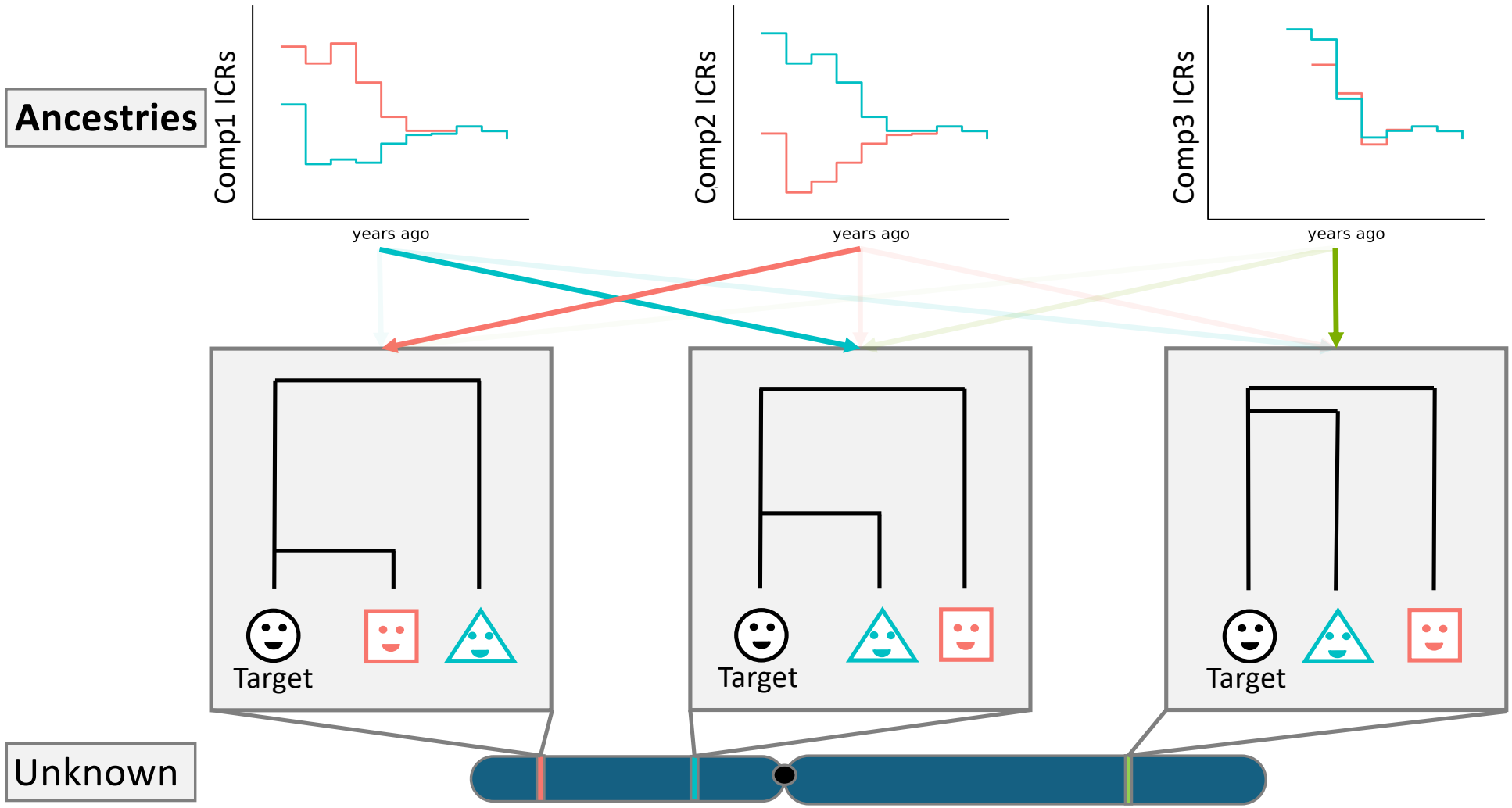
Component 1 is close to Yamnaya
Component 2 is close to WHG
Component 3 is close to farmers

comp1 comp2 comp3

Use trees to assign local ancestry of the target

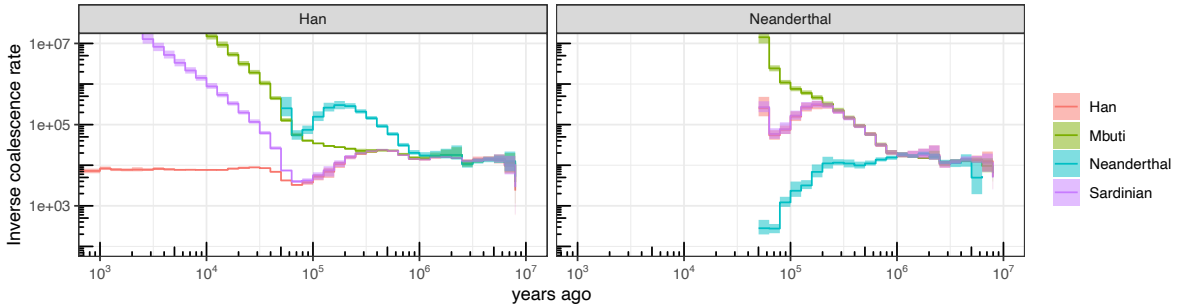


Use trees to assign local ancestry of the target

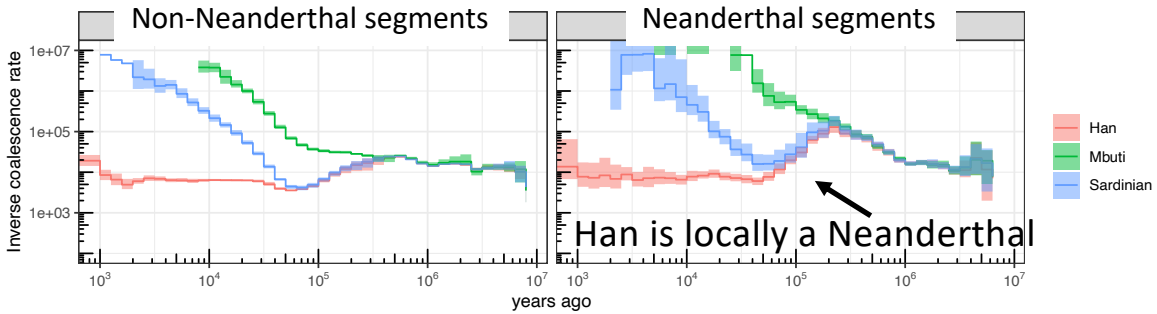


We can also measure genome-wide inverse coalescence rates between groups

Simulation with Neanderthal-like admixture:
- "Han" & Sardinian have 3%
- "Mbuti" has 0%



Given a genealogy, our approach GHOSTBUSTER then can calculate the likelihood for any lineage it comes from the Neanderthal or Han population…….and use this to identify the Neanderthal segments within a Han individual's genome!



After building trees, the population size is estimated as simply the inverse of observed coalescence rate through time

We approximate the distribution of the genomic span of a clade

We find that in practice the rate of clade-breaking events does not vary hugely

$$P(G \text{ broken} \mid T) \cdot \mathcal{L}(T) \approx P(G \text{ broken} \mid T') \cdot \mathcal{L}(T')$$

- So we can calculate once per clade, i.e. approximate as homogeneous process

- For a given ARG, let d_i be span of clade G_i in tree T_i in recombination (ρ) units, then if

$$q_i = P(G_i \text{ broken} \mid T_i) \cdot d_i \mathcal{L}(T_i)/2 \, ,$$

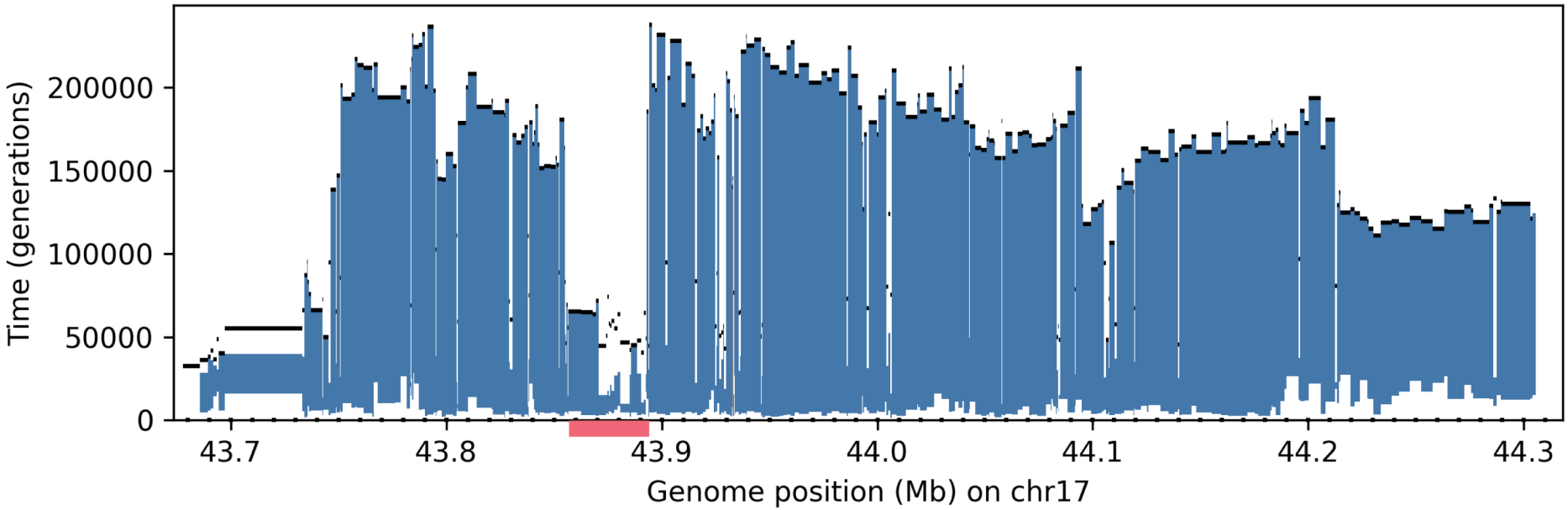
q_i has an approximate Exp(1) distribution

Test 1: p-value: the probability of observing a clade span greater than d_i, $p_i = e_i^{-q_i}$
Test 2: p-values for number of clade changes (i.e. recombination events) elsewhere on the tree that each clade "survives“ (differing false positives in real-world data)

Adjustments to allow for varying recombination rates, ARG reconstruction artefacts (use SNP span and “smooth” clades), phasing errors, population structure

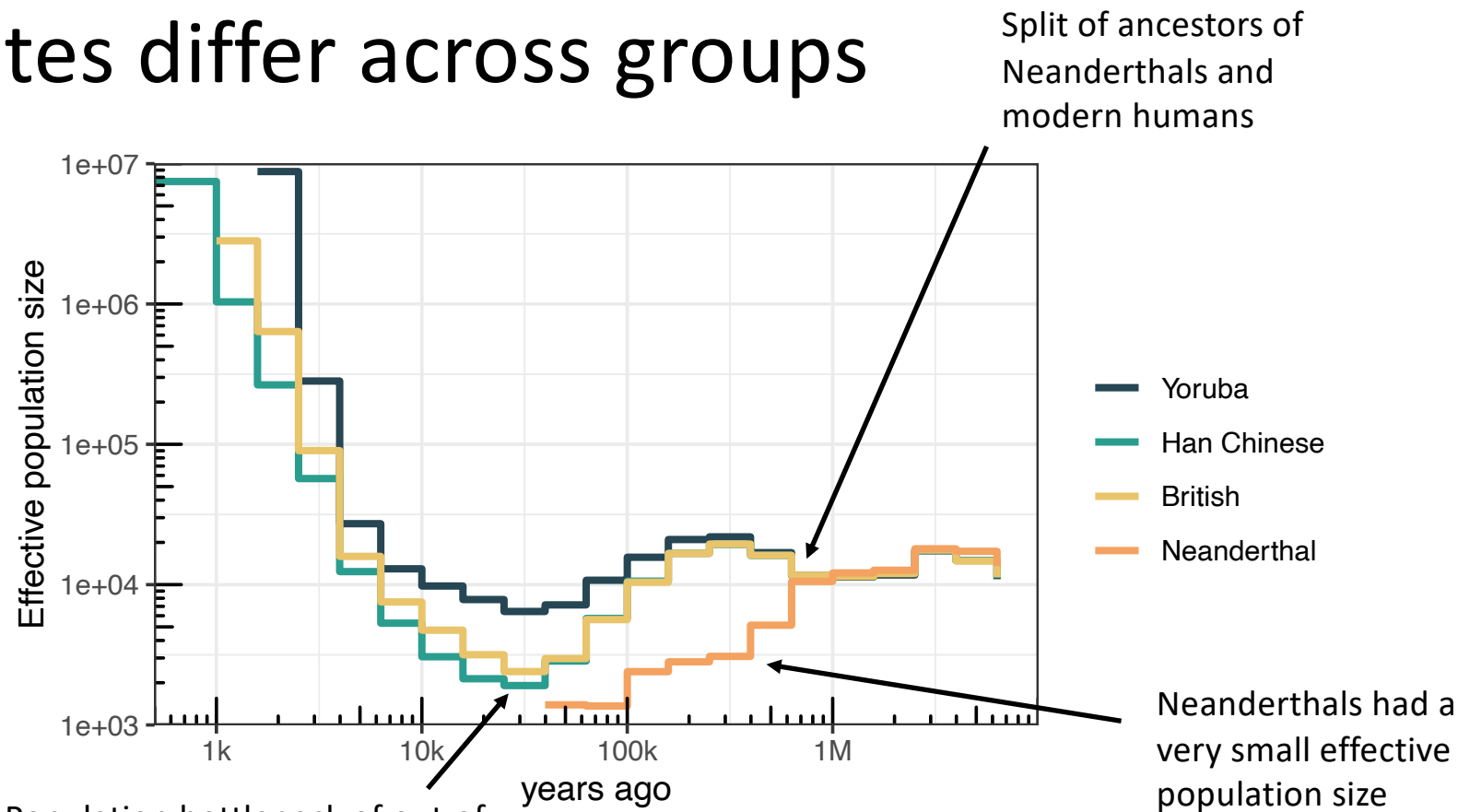
Genealogy-based analysis of 17q21.31 inversion

- Frequency in EUR: 24%, AMR: 15%, SAS: 6%, AFR: 2%, agrees with prior estimates (Donnelly et al., AJHG, 2010)



- Ancient origin; even older than other estimates of 2-3m years (Stefansson et al., Nature Genetics, 2005; Steinberg et al., Nature Genetics, 2012)
- Region where MRCA time is much more recent supportive of a possible historic double crossover event between the inverted and non-inverted haplotypes (Steinberg et al., Nature Genetics, 2012)

Relate-based population size estimates differ across groups



After (jointly) building trees, the population size is estimated as simply the **inverse of mean observed coalescence rate through time**