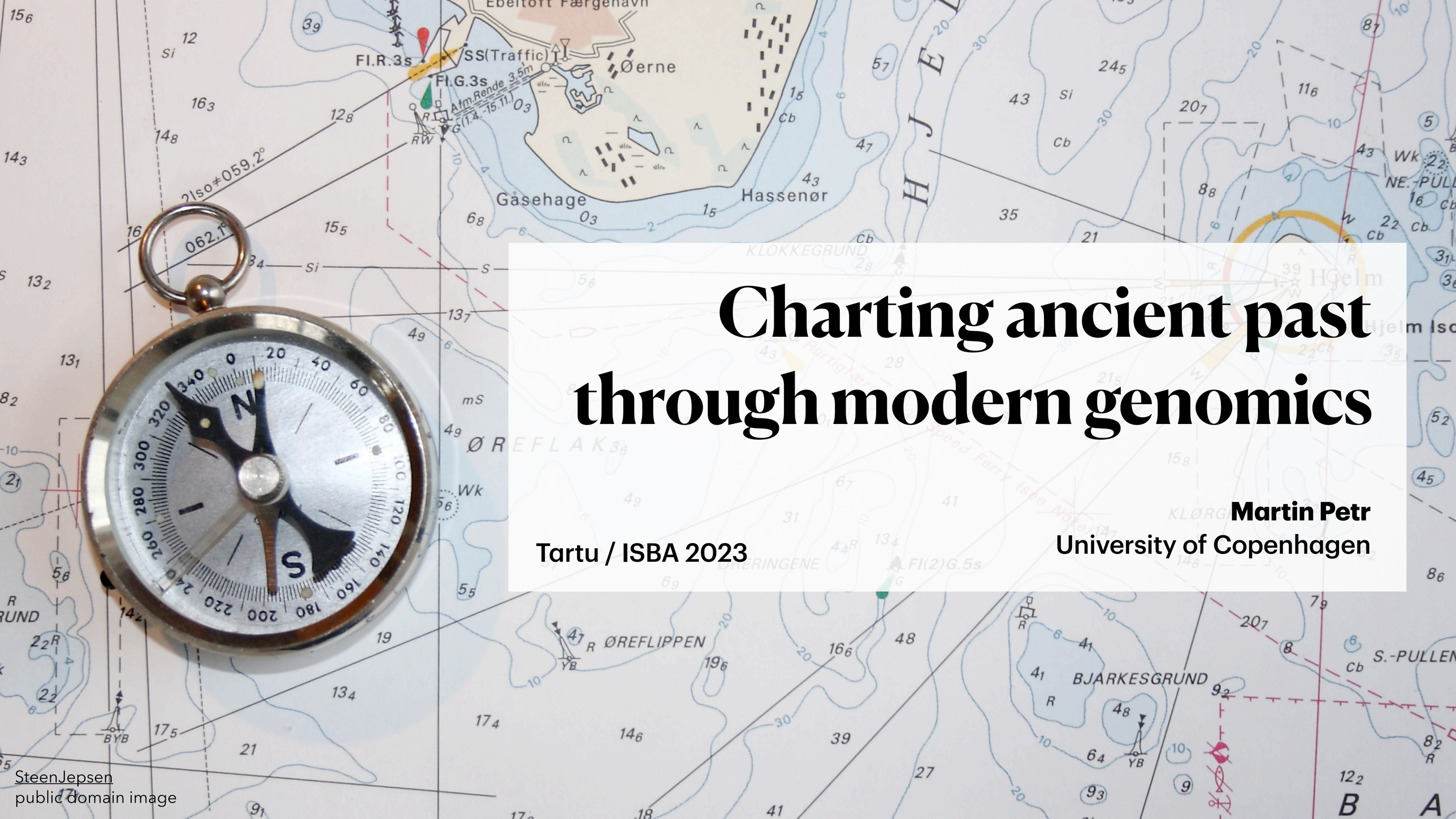
Charting ancient past through modern genomics

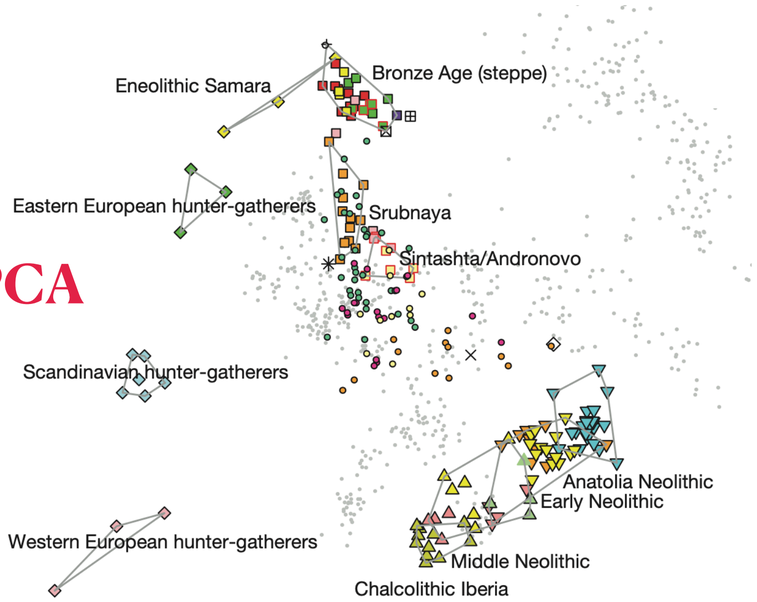
Martin Petr
University of Copenhagen

Tartu / ISBA 2023



SteenJepsen
public domain image

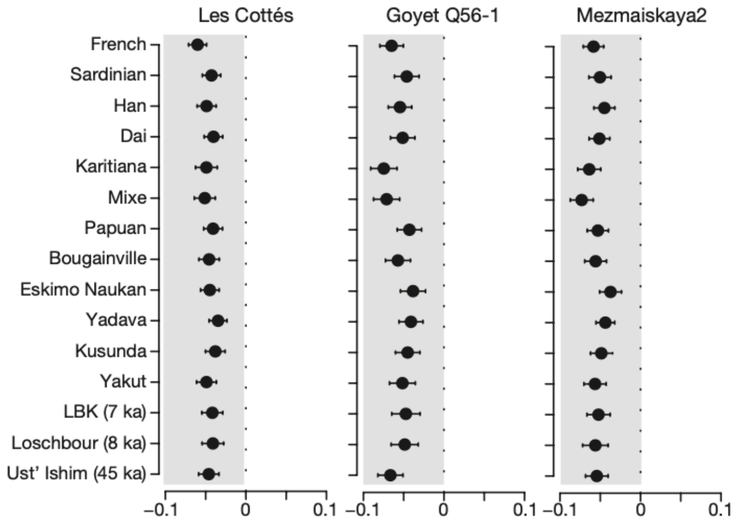
The all-stars of ancient population genomics



PCA

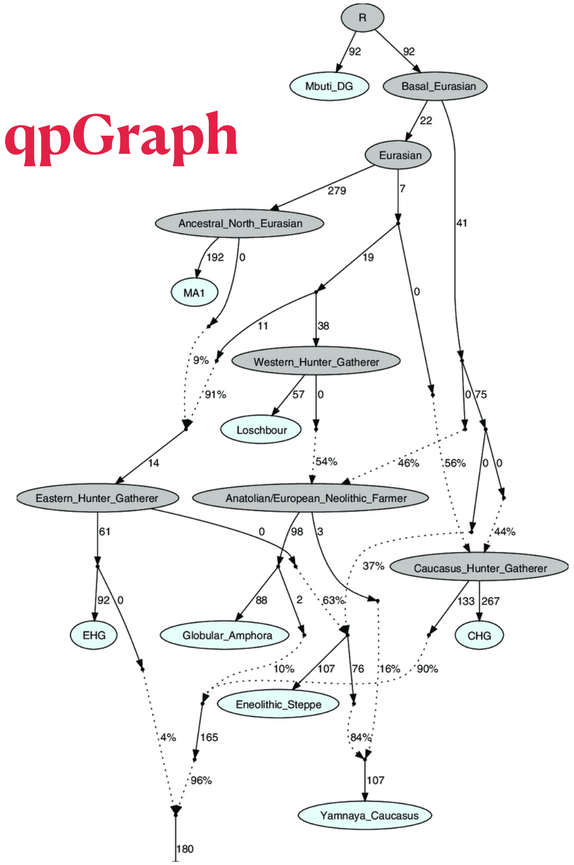
Mathieson *et al*. 2015

f-statistics

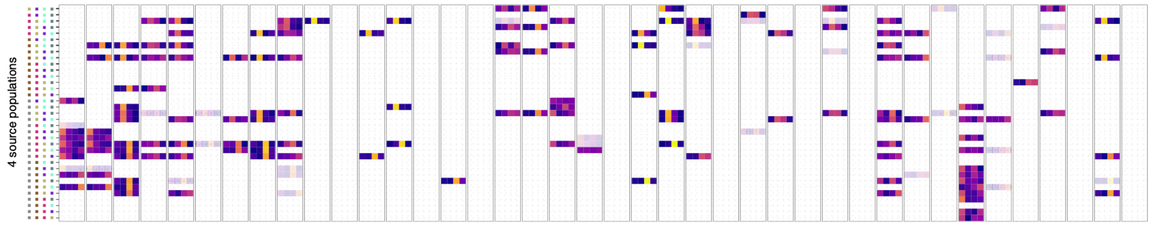


Hajdinjak *et al*. 2018

qpGraph



Wang *et al*. 2019

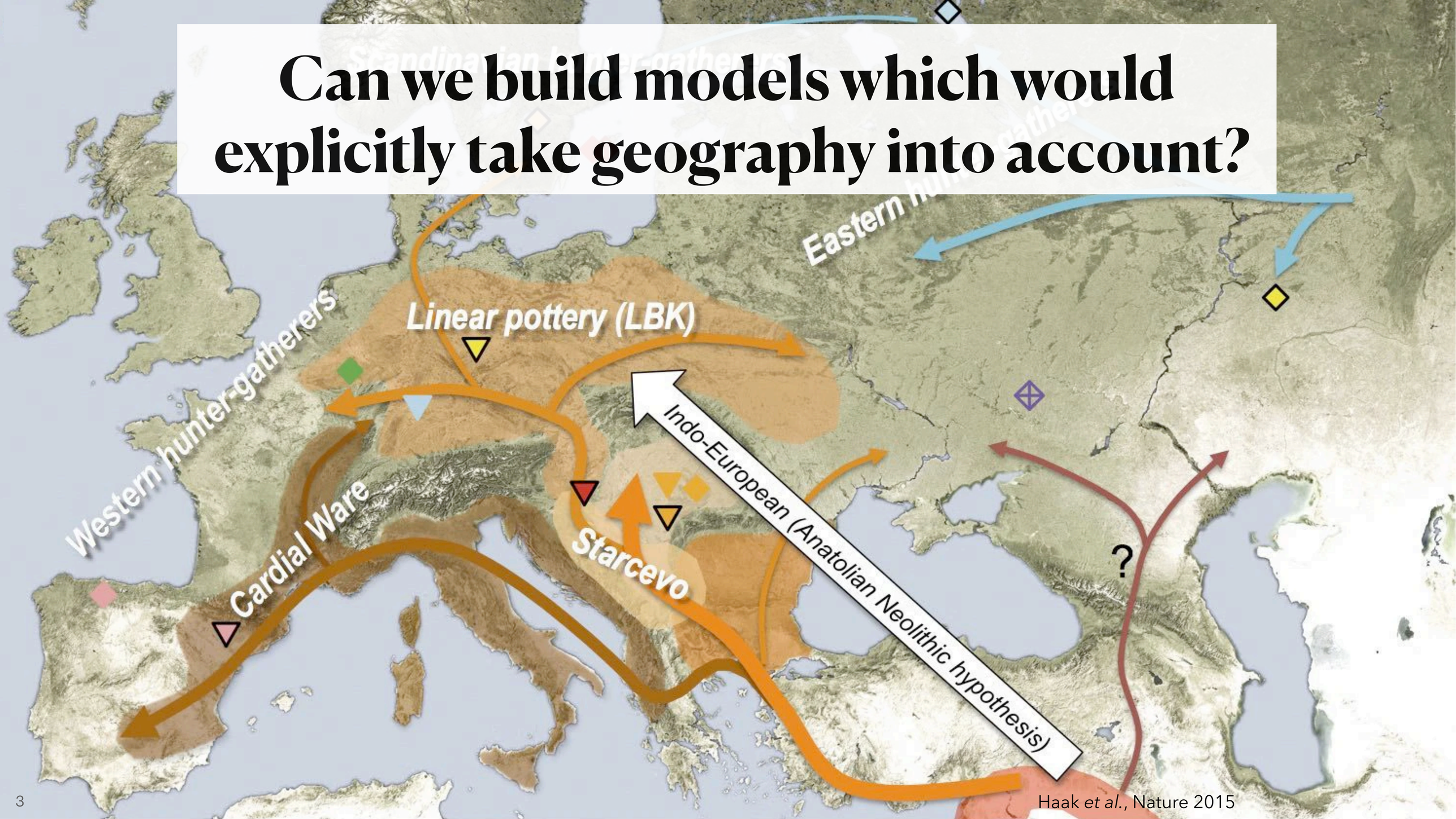


qpAdm

Sikora *et al*. 2019

Haak *et al.*, Nature 2015

Can we build models which would explicitly take geography into account?



Haak *et al.*, Nature 2015

PNAS

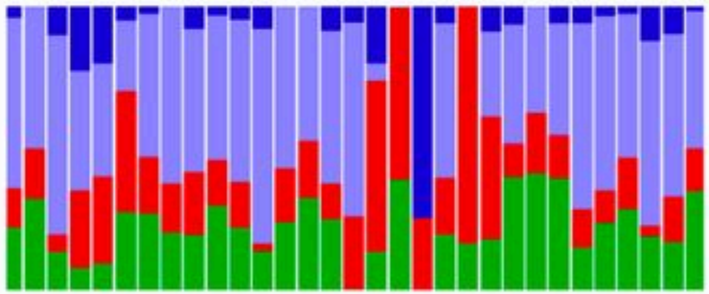
The spatiotemporal spread of human migrations during the European Holocene

Fernando Racimo[a,1], Jessie Woodbridge[b], Ralph M. Fyfe[b], Martin Sikora[a], Karl-Göran Sjögren[c], Kristian Kristiansen[c], and Marc Vander Linden[d]

[a]Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark; [b]School of Geography, Earth, and Environmental Sciences, University of Plymouth, Plymouth PL4 8AA, United Kingdom; [c]Department of Historical Studies, University of Gothenburg, 405 30 Gothenburg, Sweden; and [d]Department of Archaeology, University of Cambridge, Cambridge CB2 1TN, United Kingdom

Spatio-temporal mapping of genetic ancestry

ADMIXTURE analysis



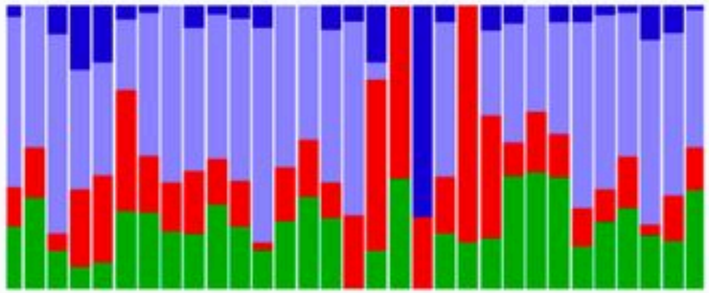
farmer (NEOL) ancestry
Yamnaya (YAM) ancestry
hunter-gatherer ancestry
in each sequenced individual

Racimo *et al.*, PNAS 2020

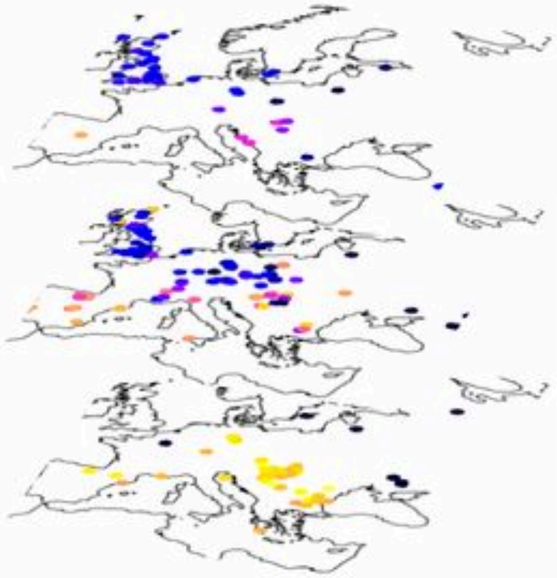
Spatio-temporal mapping of genetic ancestry

position of each individual
(and their ancestry) on a map

ADMIXTURE analysis



farmer (NEOL) ancestry
Yamnaya (YAM) ancestry
hunter-gatherer ancestry
in each sequenced individual



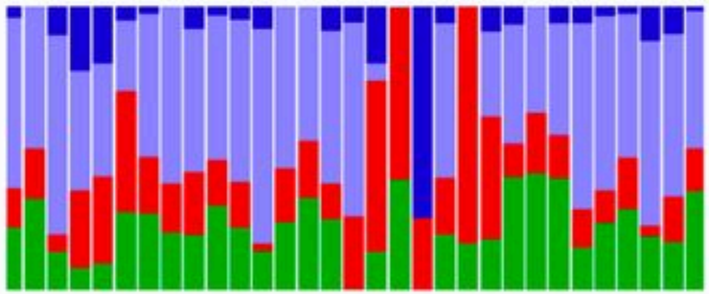
Racimo *et al.*, PNAS 2020

Spatio-temporal mapping of genetic ancestry

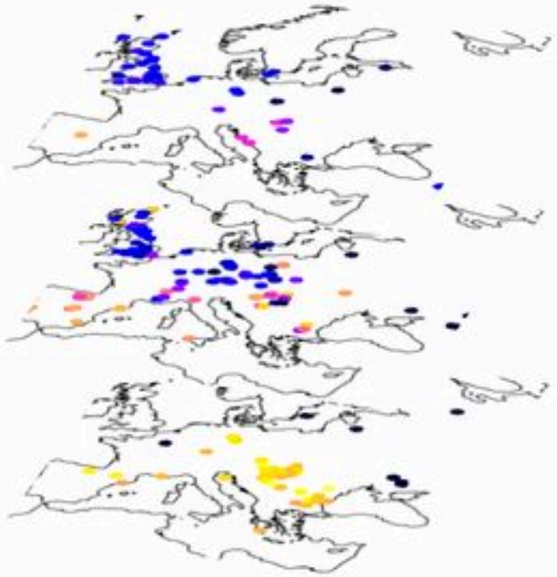
ADMIXTURE analysis

position of each individual (and their ancestry) on a map

interpolated ancestry maps



farmer (NEOL) ancestry
Yamnaya (YAM) ancestry
hunter-gatherer ancestry
in each sequenced individual



Racimo *et al.*, PNAS 2020

Interpolated ancestry maps

"NEOL" ancestry

-10800

"YAM" ancestry

-10800

2020

Interpolated ancestry maps

"NEOL" ancestry

-10800

"YAM" ancestry

-10800

2020

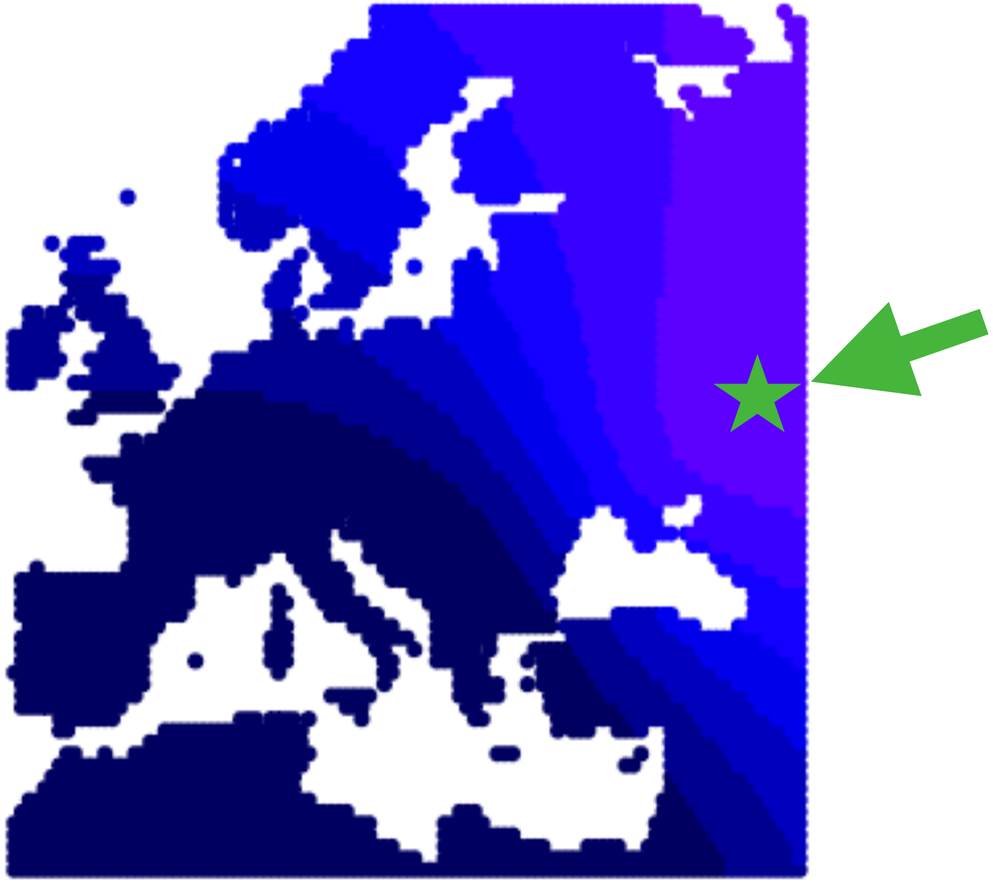
How fast did an "ancestry wave" move from its origin?

"NEOL" ancestry

-10800 BP

"YAM" ancestry

-10800 BP



Racimo *et al.*, PNAS 2020

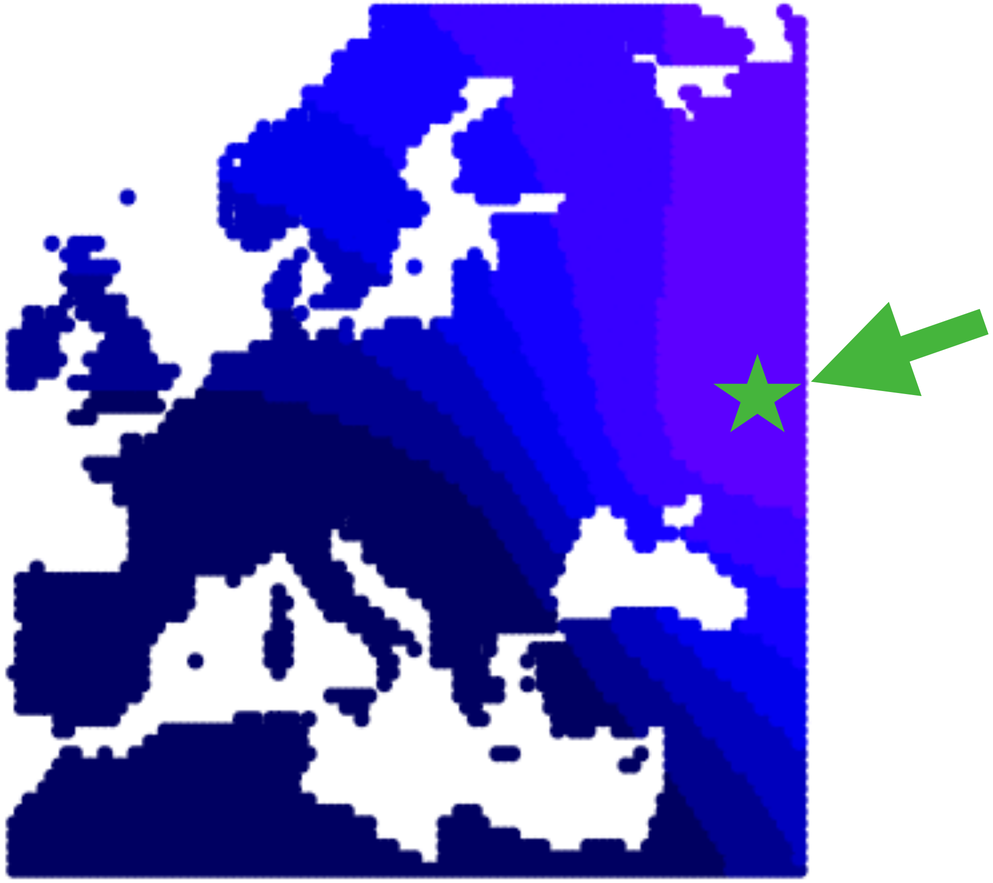
How fast did an "ancestry wave" move from its origin?

"NEOL" ancestry

-10800 BP

"YAM" ancestry

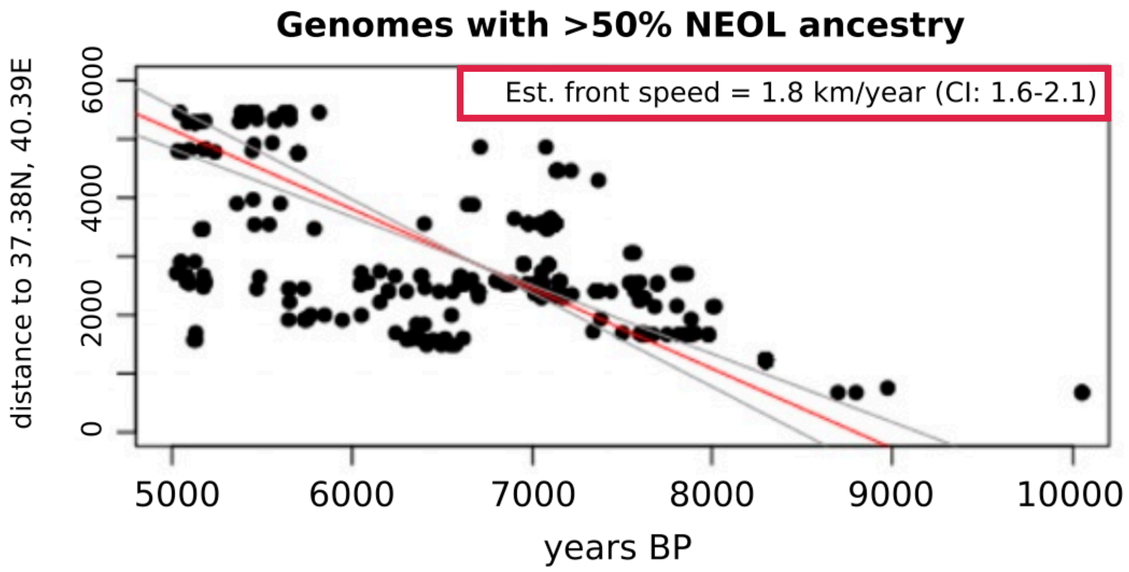
-10800 BP



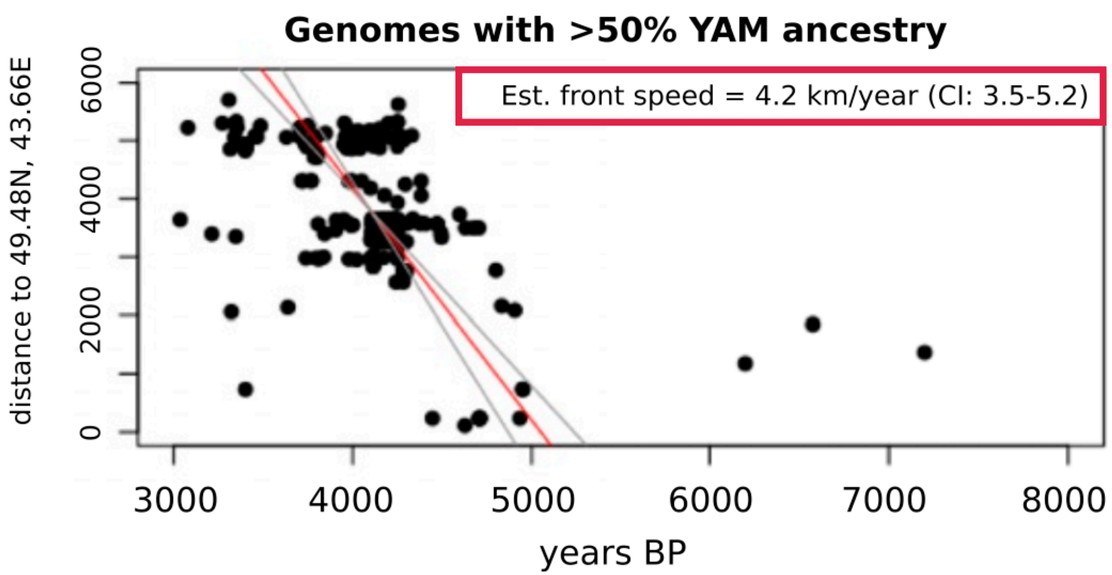
Racimo *et al.*, PNAS 2020

How fast did an "ancestry wave" move from its origin?

"NEOL" ancestry



"YAM" ancestry



Racimo *et al.*, PNAS 2020

PNAS

RESEARCH ARTICLE | GENETICS
ANTHROPOLOGY

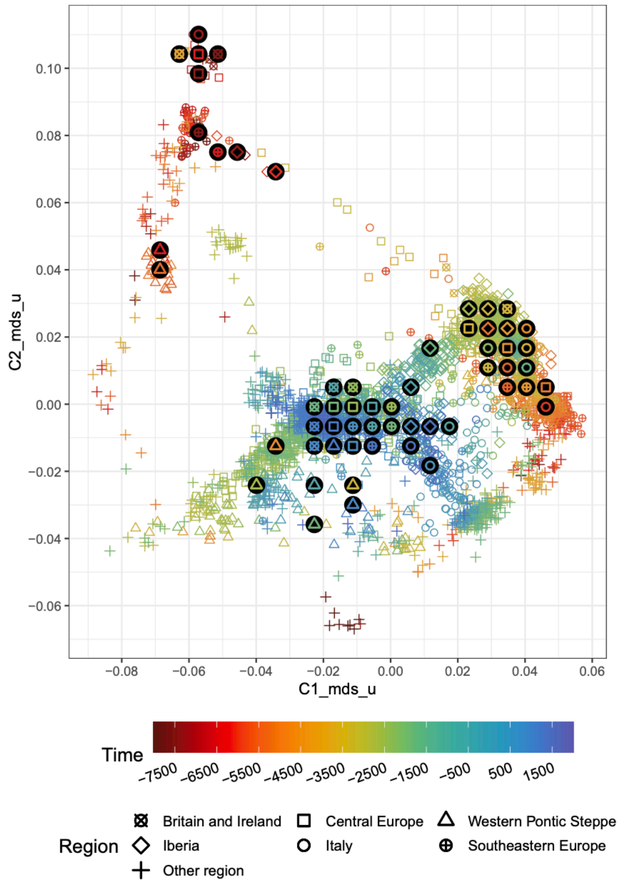
Estimating human mobility in Holocene Western Eurasia with large-scale ancient genomic data

Clemens Schmid[a,b] and Stephan Schiffels[a,1]

Edited by Liisa Loog, University of Oxford, Cambridge, UK; received November 1, 2022; accepted December 5, 2022, by Editorial Board Member Richard G. Klein

Where does a sample trace its ancestral origin?

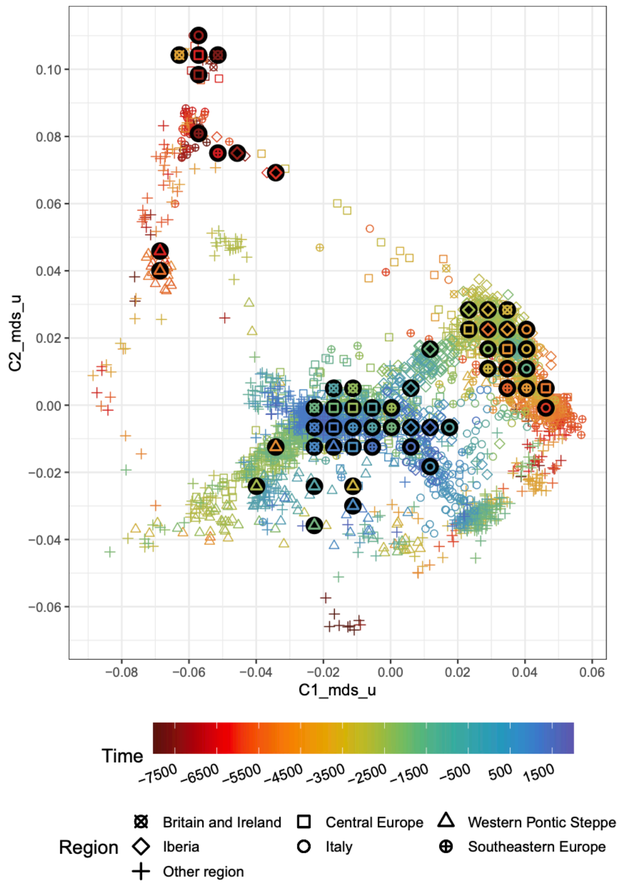
MDS / PCA



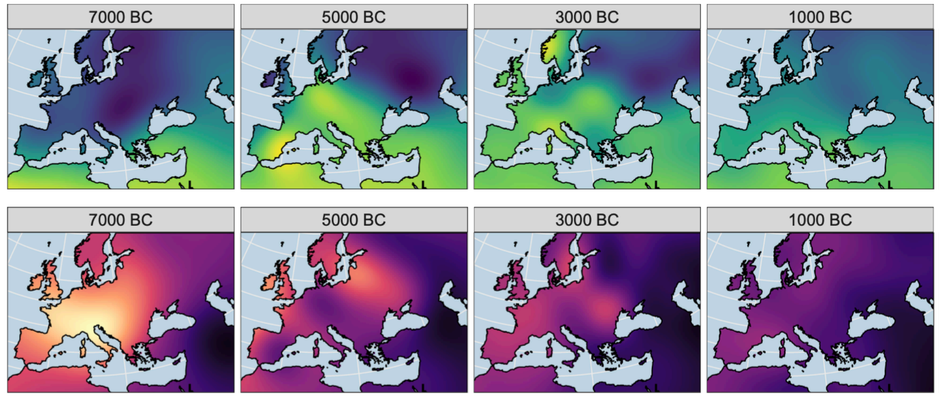
Schmid and Schiffels, PNAS 2023

Where does a sample trace its ancestral origin?

MDS / PCA



spatio-temporal interpolated "PCA field"



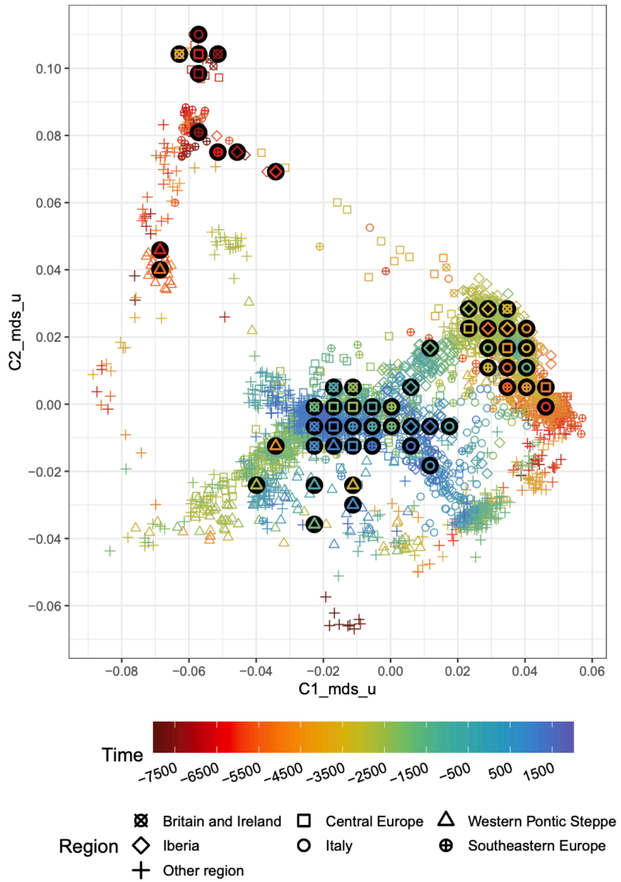
"farmer" component

"Yamnaya" component

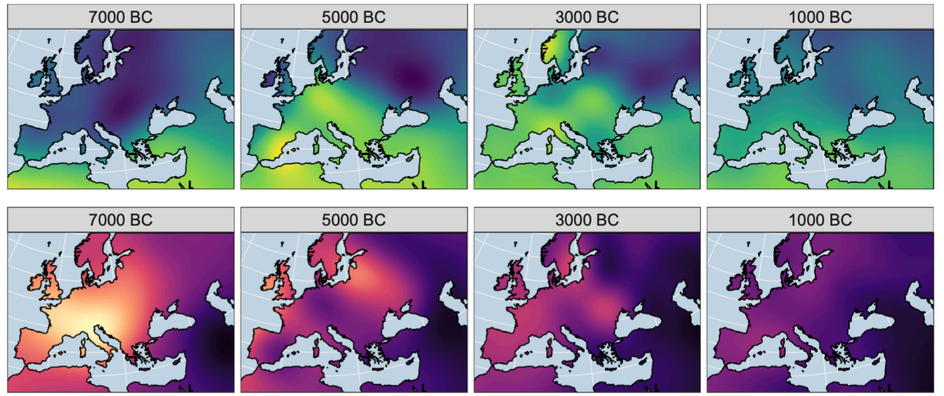
Schmid and Schiffels, PNAS 2023

Where does a sample trace its ancestral origin?

MDS / PCA

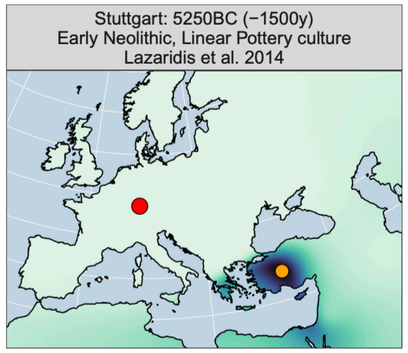


spatio-temporal interpolated "PCA field"



"farmer" component

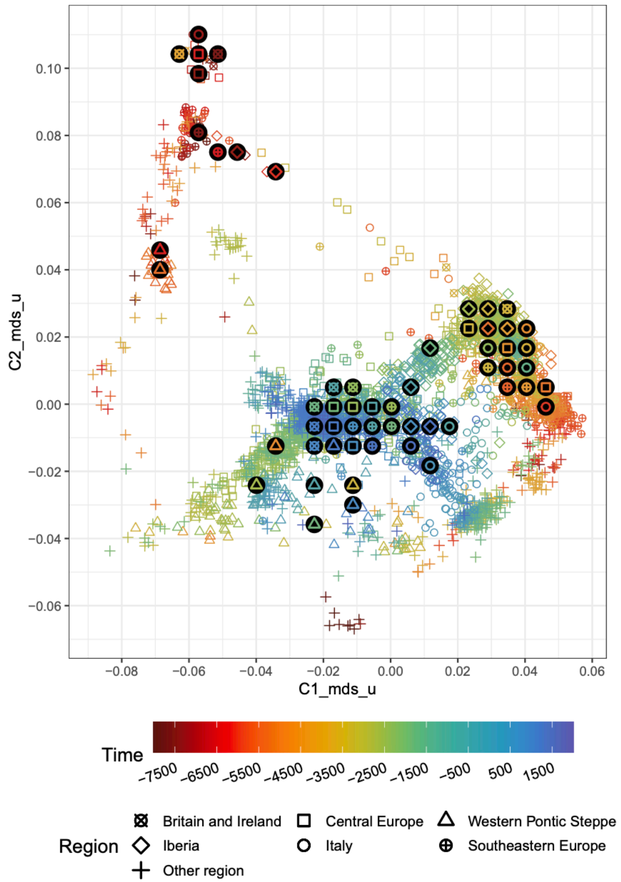
"Yamnaya" component



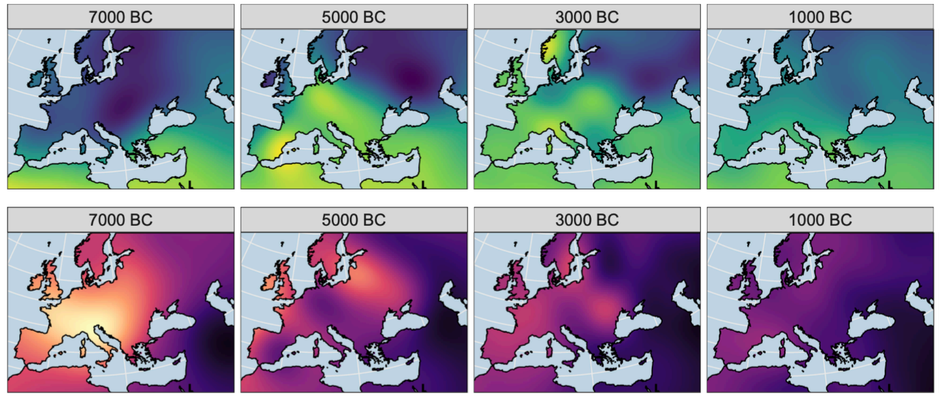
Schmid and Schiffels, PNAS 2023

Where does a sample trace its ancestral origin?

MDS / PCA



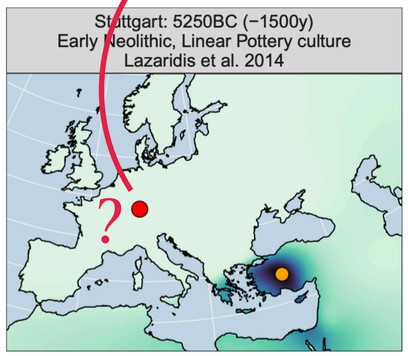
spatio-temporal interpolated "PCA field"



"farmer" component

"Yamnaya" component

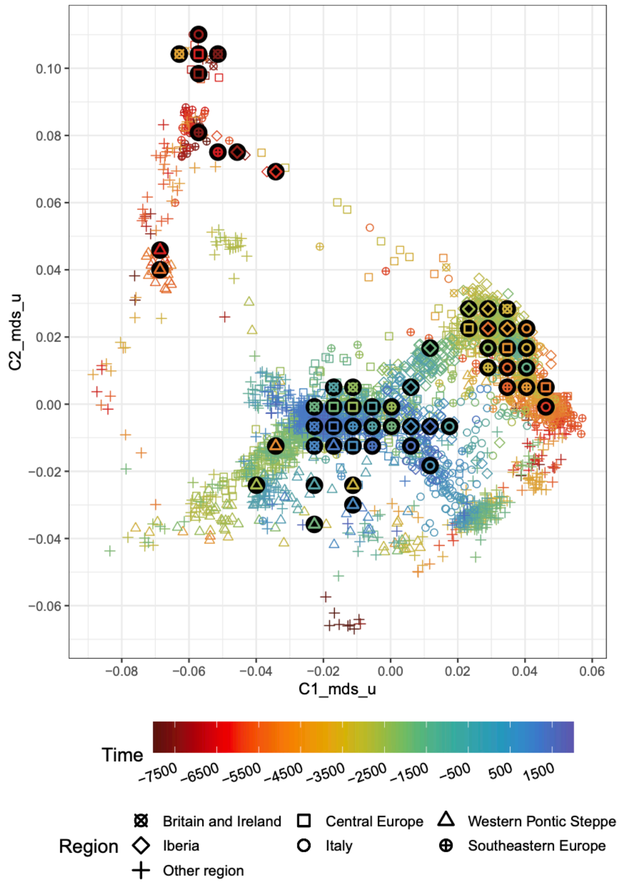
where in the PCA field are the closest component values?



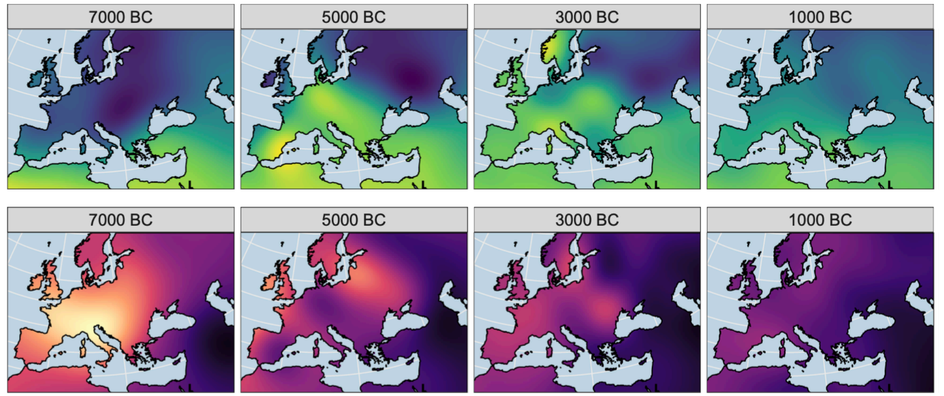
Schmid and Schiffels, PNAS 2023

Where does a sample trace its ancestral origin?

MDS / PCA



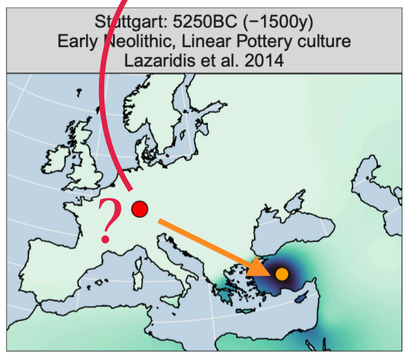
spatio-temporal interpolated "PCA field"



"farmer" component

"Yamnaya" component

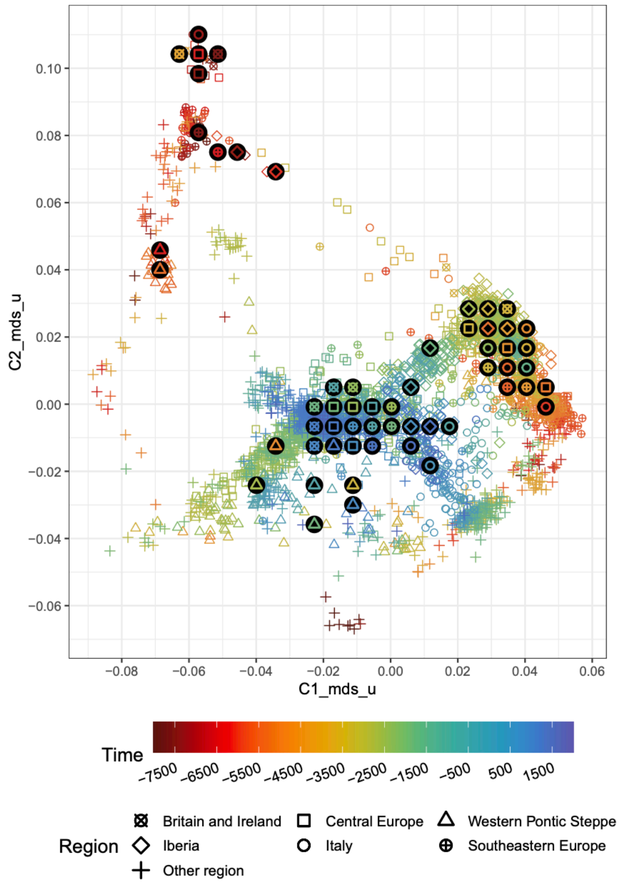
where in the PCA field are the closest component values?



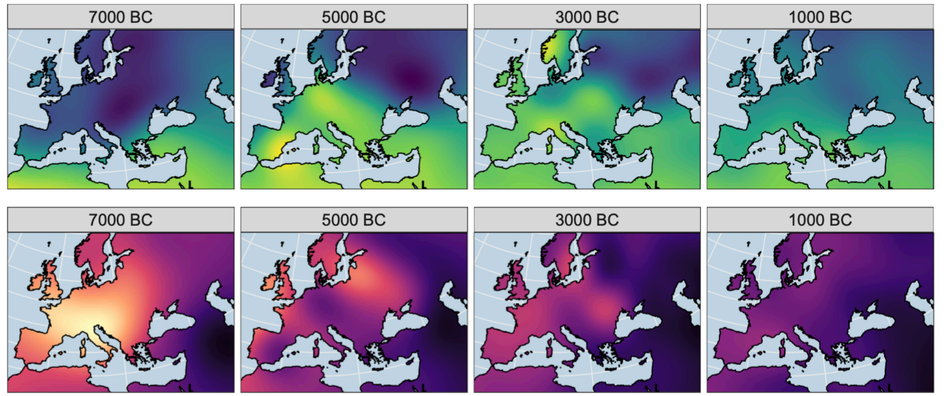
Schmid and Schiffels, PNAS 2023

Where does a sample trace its ancestral origin?

MDS / PCA



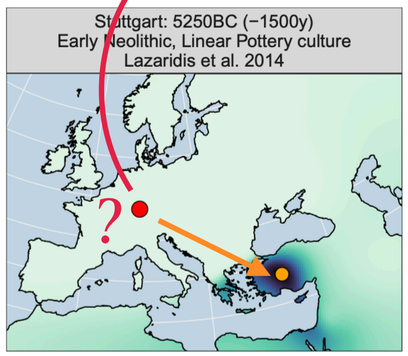
spatio-temporal interpolated "PCA field"

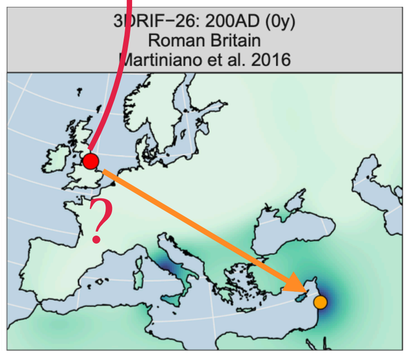


"farmer" component

"Yamnaya" component

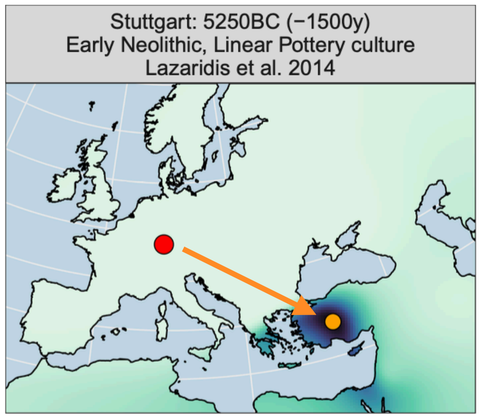
where in the PCA field are the closest component values?





Schmid and Schiffels, PNAS 2023

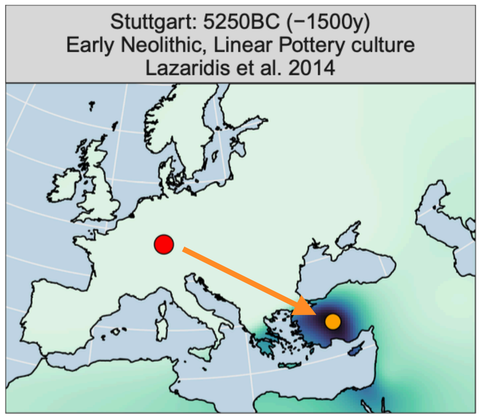
A "global" summary of individual mobilities across time



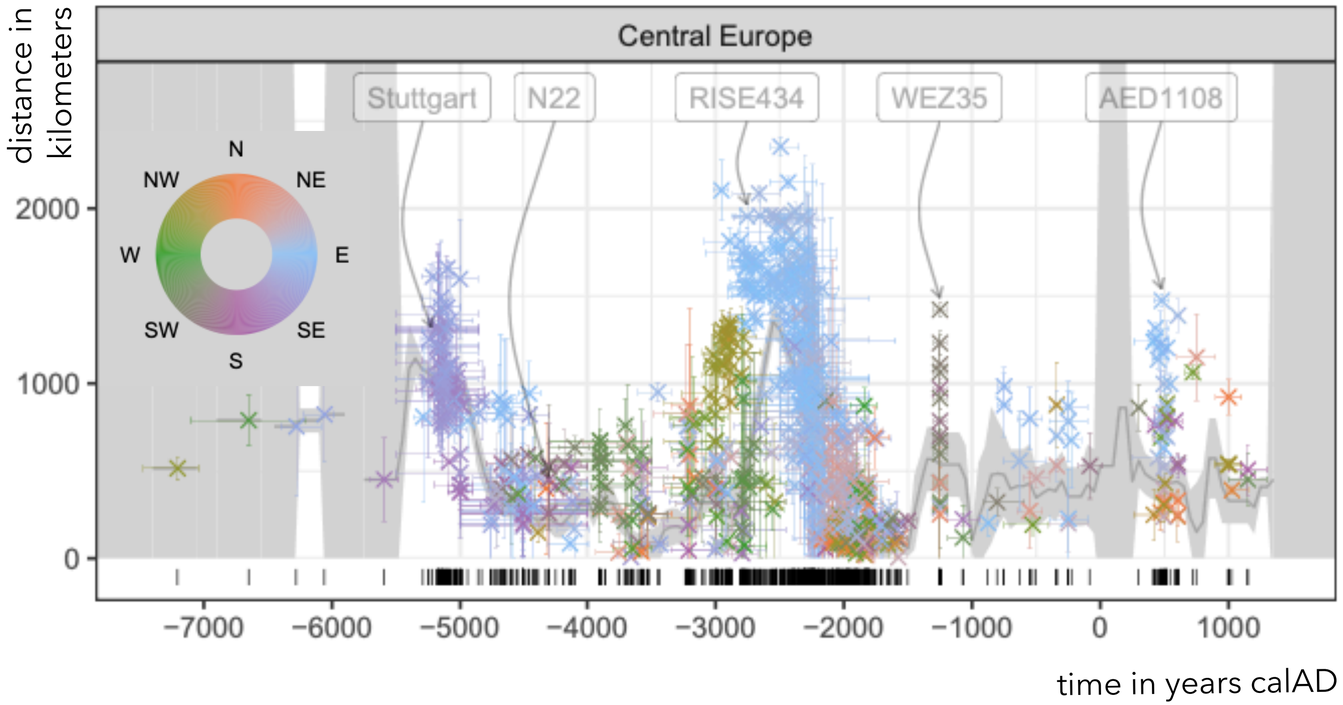
aggregation of **lengths** and **directions** of many mobility vectors

Schmid and Schiffels, PNAS 2023

A "global" summary of individual mobilities across time



aggregation of **lengths** and **directions** of many mobility vectors



Schmid and Schiffels, PNAS 2023

www.slendr.net

slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes

Martin Petr, Benjamin C. Haller, Peter L. Ralph, Fernando Racimo

(2023), bioRxiv, ver.5, peer-reviewed and recommended by PCI Evol Biol

https://doi.org/10.1101/2022.03.20.485041

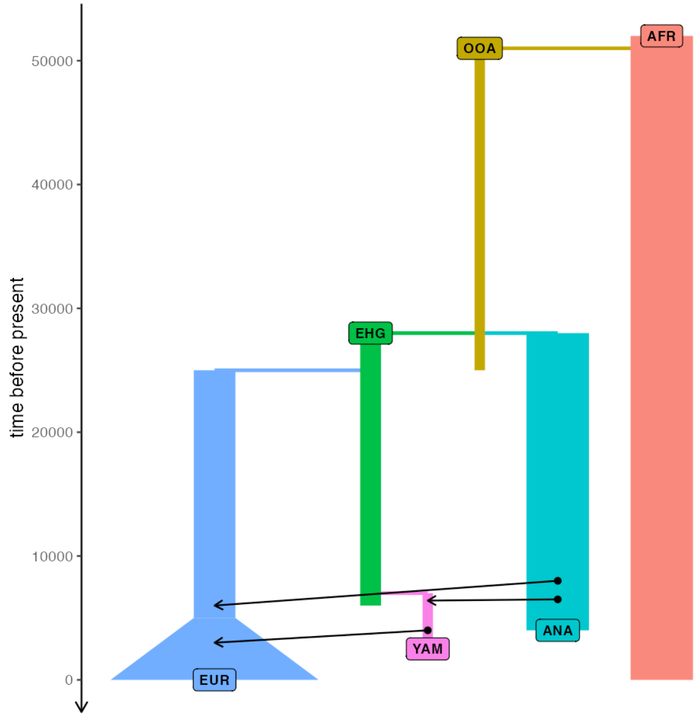
PCI Evol Biol

https://peercommunityin.org/

slendr can simulate genomic data from...

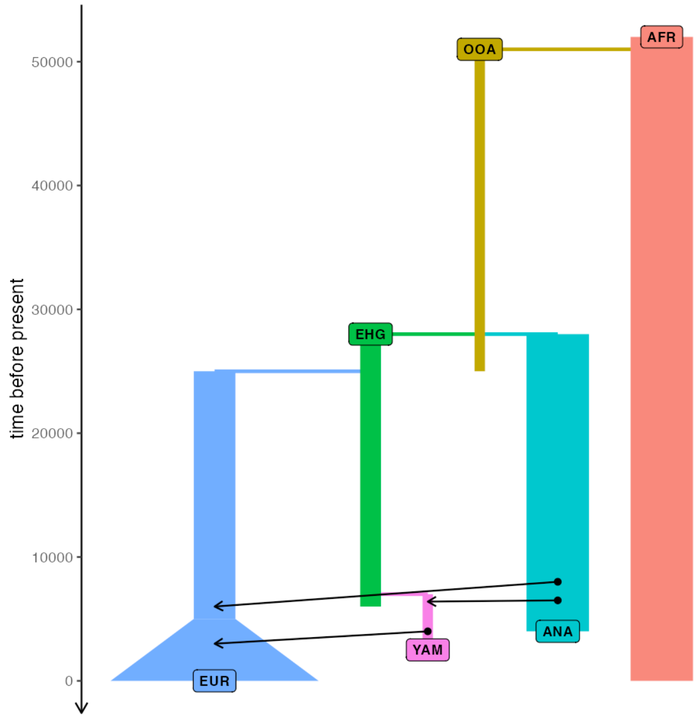
slendr can simulate genomic data from...

... traditional demographic models...



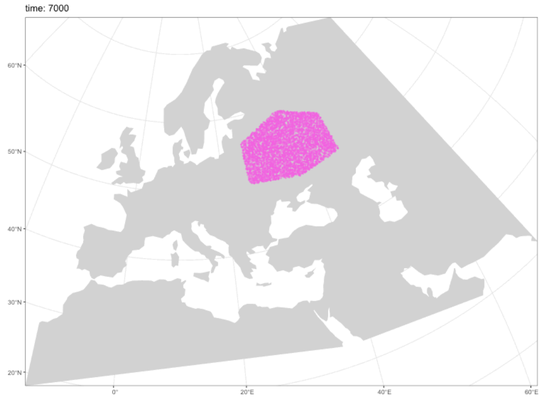
slendr can simulate genomic data from...

... traditional demographic models...

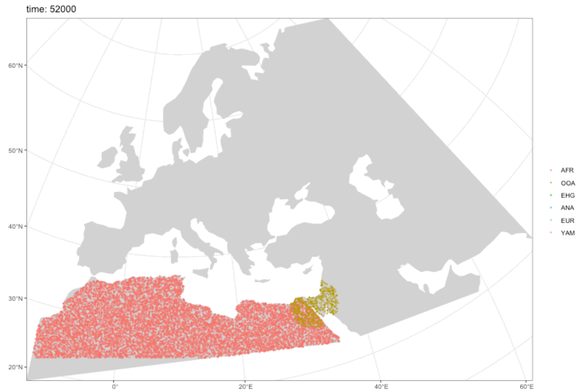


... but also explicitly spatial models!

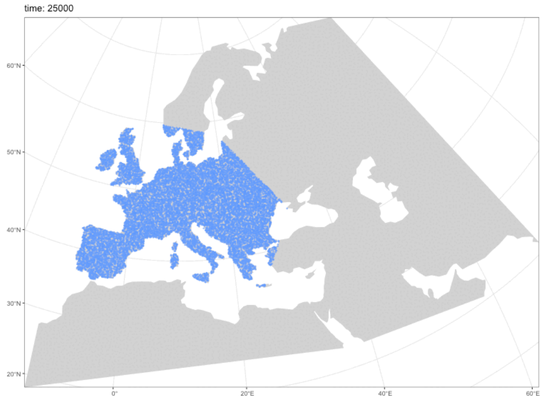
"farmer expansion"



"Yamnaya migration"



"Out-of-Africa migration"



"Europeans"

Tree sequence / Ancestral Recombination Graph (ARG)

tsinfer

Inferring whole-genome histories in large population datasets

Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers and Gil McVean

2019

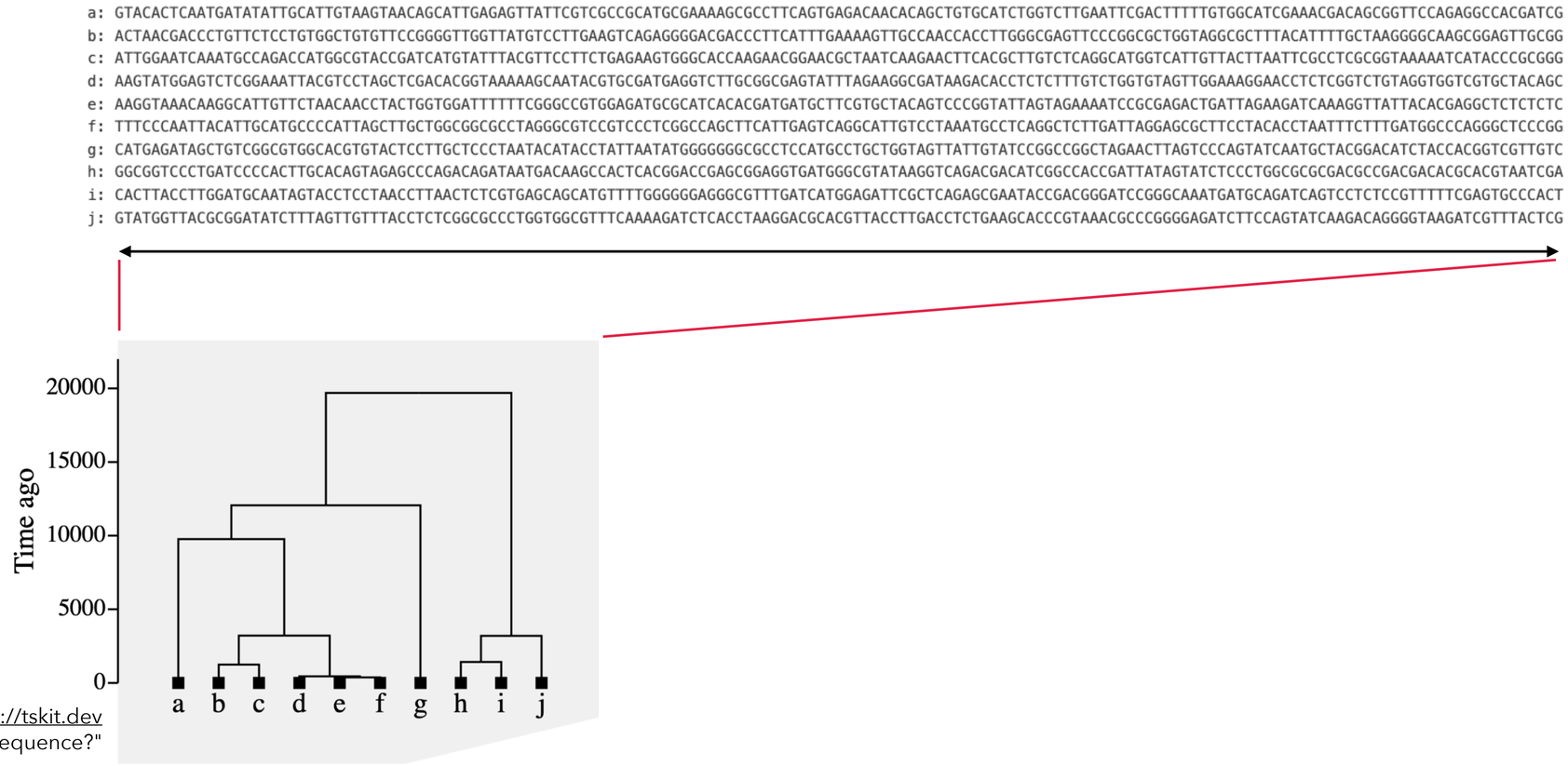
Relate

A method for genome-wide genealogy estimation for thousands of samples

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myers[1,3]*

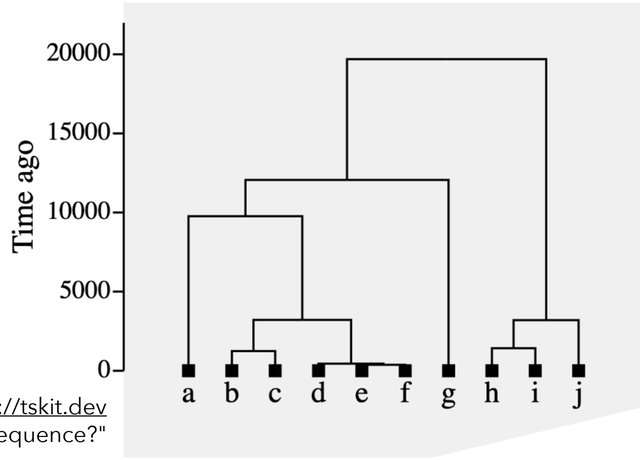
2019

Tree sequence / Ancestral Recombination Graph (ARG)



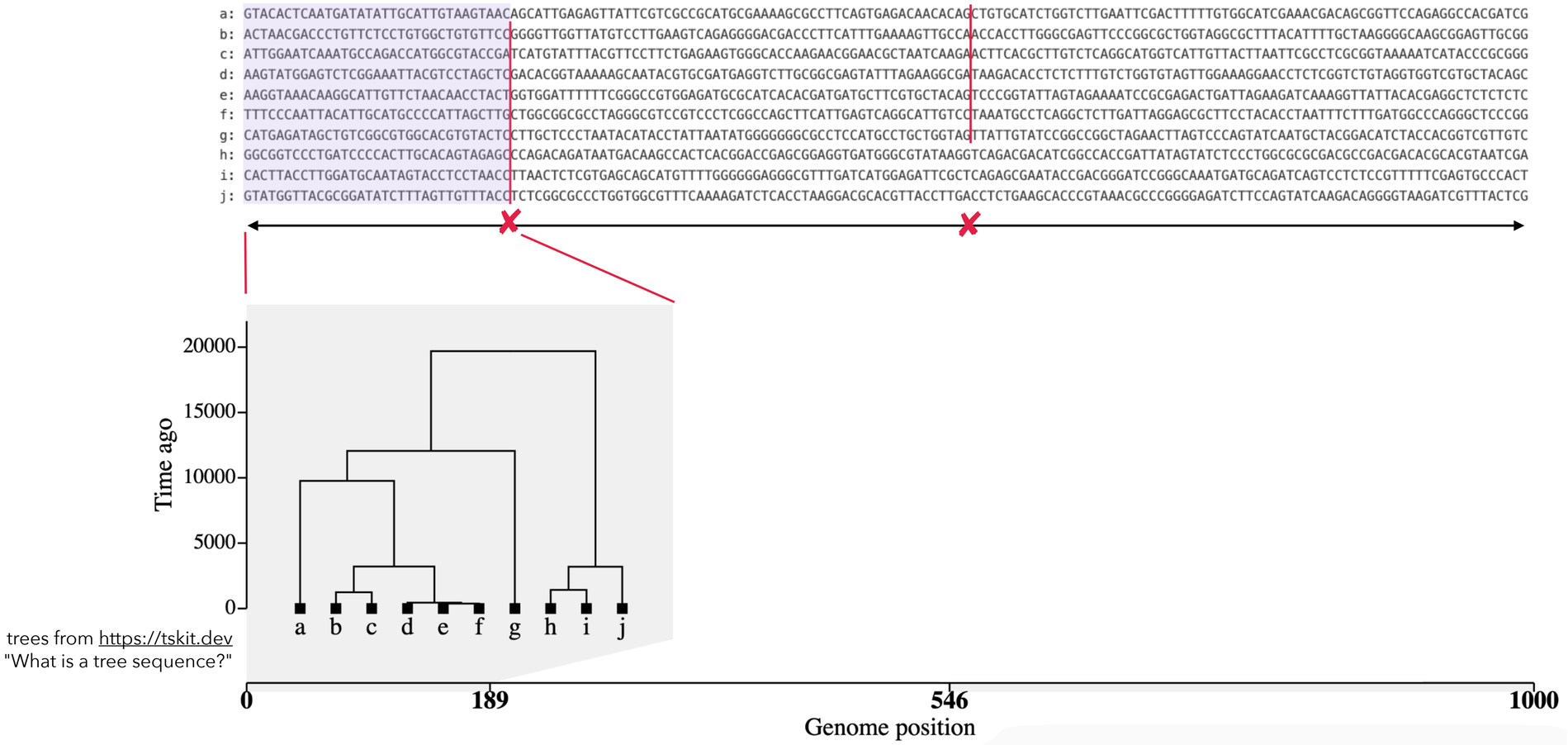
trees from https://tskit.dev
"What is a tree sequence?"

Tree sequence / Ancestral Recombination Graph (ARG)



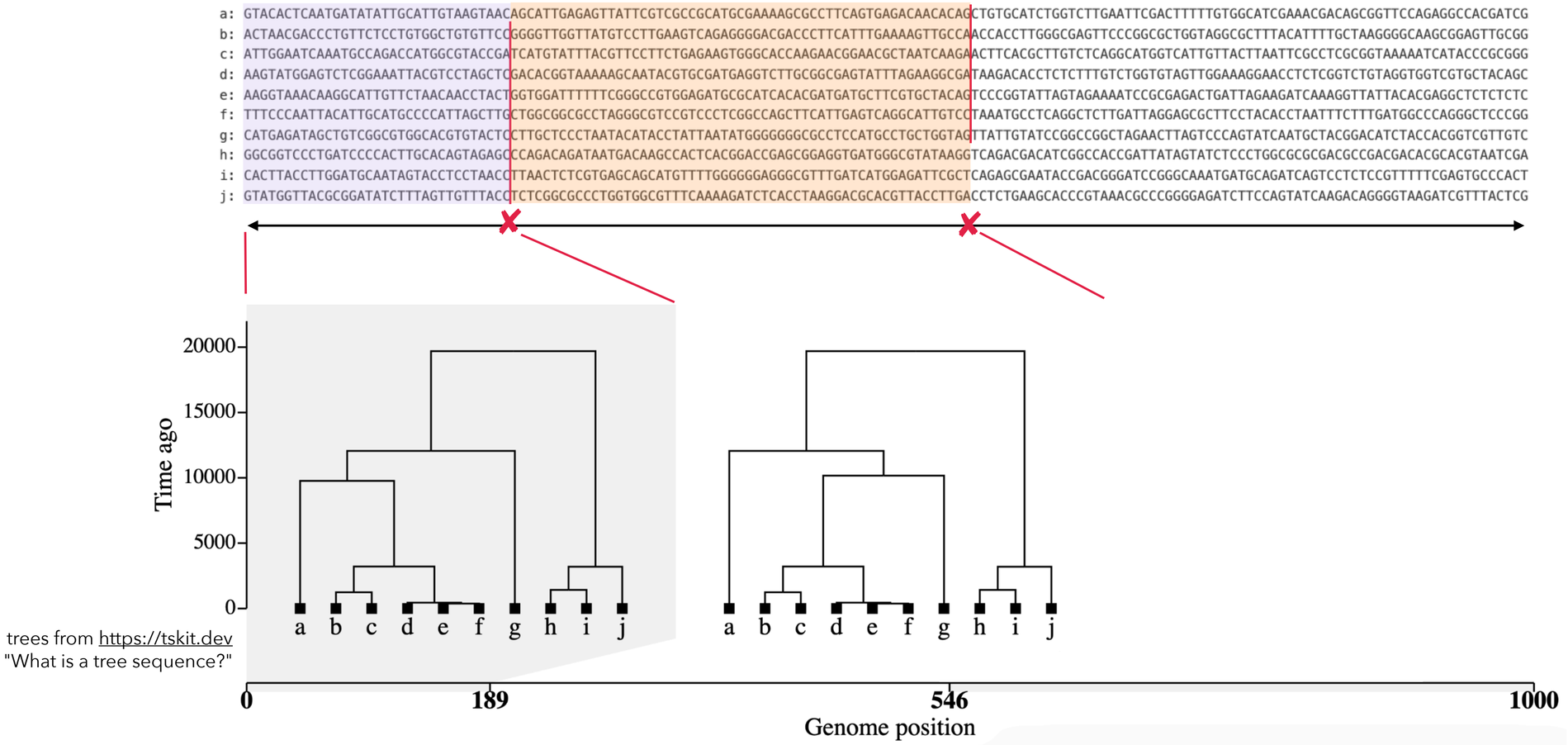
trees from https://tskit.dev
"What is a tree sequence?"

Tree sequence / Ancestral Recombination Graph (ARG)



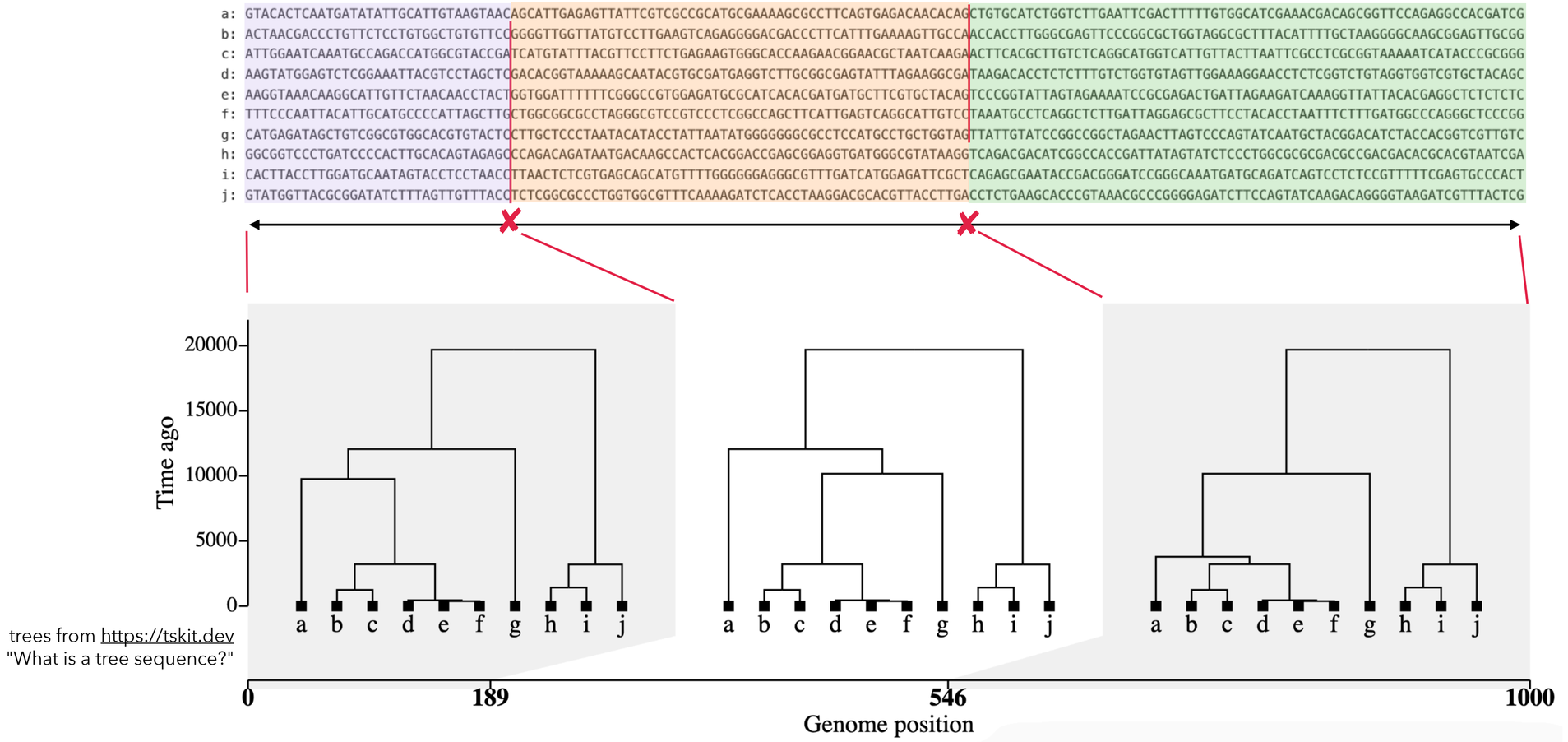
trees from https://tskit.dev
"What is a tree sequence?"

Tree sequence / Ancestral Recombination Graph (ARG)



trees from https://tskit.dev
"What is a tree sequence?"

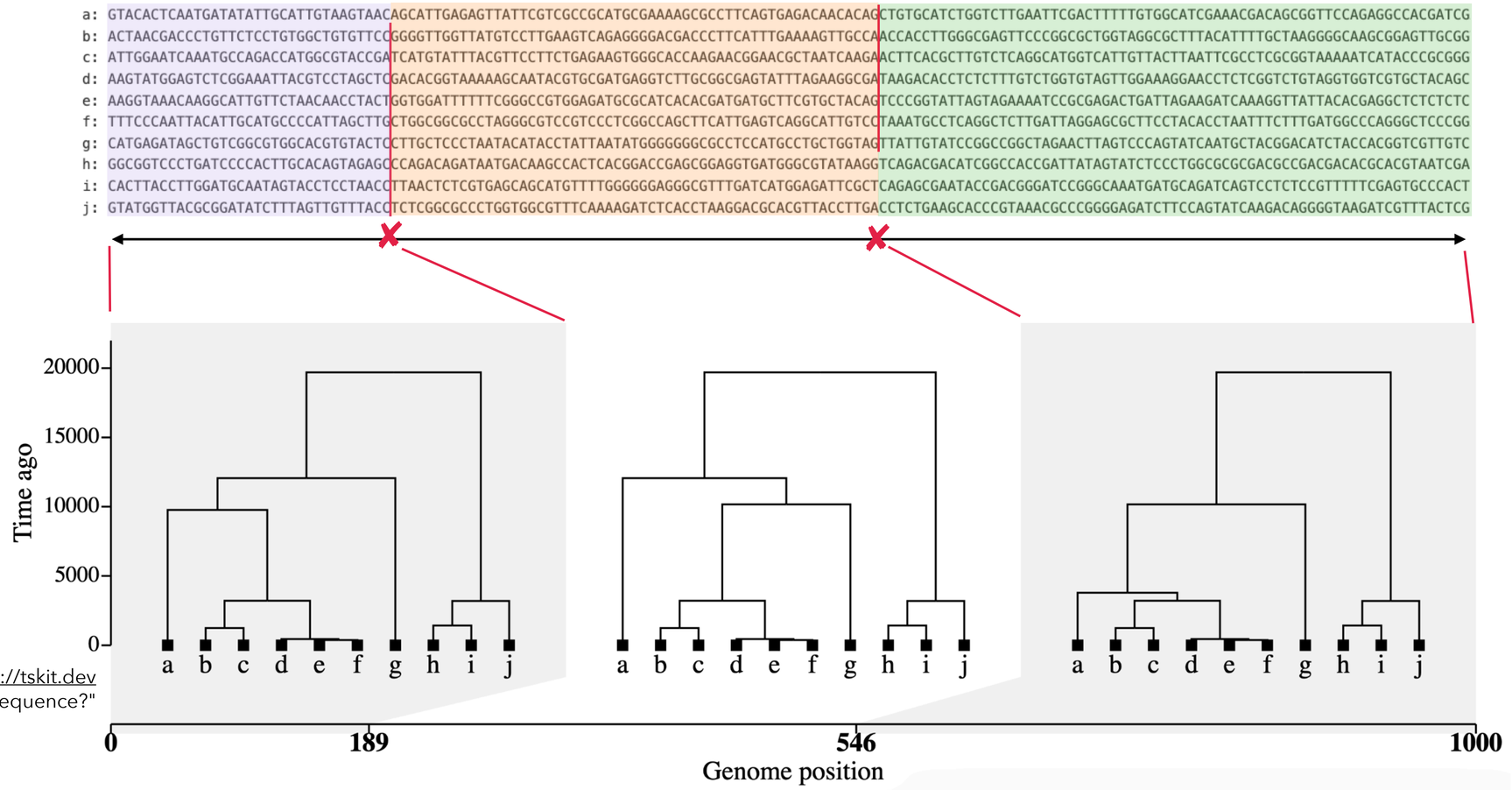
Tree sequence / Ancestral Recombination Graph (ARG)



trees from https://tskit.dev
"What is a tree sequence?"

Recombination produces a sequence of (correlated) trees along the genome

Tree sequence / Ancestral Recombination Graph (ARG)



trees from https://tskit.dev
"What is a tree sequence?"

A tree sequence represents

everything knowable about an evolutionary process.

What we (usually) have

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00099	HG00101
12	60076	.	A	C	100	PASS	.	GT	1|0	0|0	0|0
12	60252	.	A	G	100	PASS	.	GT	0|0	0|0	0|1
12	60317	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	60344	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60383	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	60405	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60474	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	60614	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60628	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60654	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61021	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61107	.	G	T	100	PASS	.	GT	0|0	0|0	0|0
12	61172	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61220	.	G	A	100	PASS	.	GT	0|0	0|0	0|1
12	61258	.	C	T	100	PASS	.	GT	0|0	0|0	0|1
12	61272	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	61329	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61341	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	61368	.	C	T	100	PASS	.	GT	0|0	0|1	0|1
12	61392	.	T	A	100	PASS	.	GT	0|0	0|0	0|0
12	61405	.	G	C	100	PASS	.	GT	0|0	0|0	0|0
12	61411	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	61416	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61422	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61476	.	C	G	100	PASS	.	GT	0|0	0|0	0|0
12	61510	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61516	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61552	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61604	.	T	G	100	PASS	.	GT	0|0	0|0	0|0
12	61687	.	G	A	100	PASS	.	GT	1|0	0|1	0|1
12	61700	.	C	T	100	PASS	.	GT	0|0	0|0	0|0

What we (usually) have

positions along a chromosome

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00099	HG00101
12	60076	.	A	C	100	PASS	.	GT	1|0	0|0	0|0
12	60252	.	A	G	100	PASS	.	GT	0|0	0|0	0|1
12	60317	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	60344	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60383	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	60405	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60474	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	60614	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60628	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60654	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61021	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61107	.	G	T	100	PASS	.	GT	0|0	0|0	0|0
12	61172	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61220	.	G	A	100	PASS	.	GT	0|0	0|0	0|1
12	61258	.	C	T	100	PASS	.	GT	0|0	0|0	0|1
12	61272	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	61329	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61341	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	61368	.	C	T	100	PASS	.	GT	0|0	0|1	0|1
12	61392	.	T	A	100	PASS	.	GT	0|0	0|0	0|0
12	61405	.	G	C	100	PASS	.	GT	0|0	0|0	0|0
12	61411	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	61416	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61422	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61476	.	C	G	100	PASS	.	GT	0|0	0|0	0|0
12	61510	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61516	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61552	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61604	.	T	G	100	PASS	.	GT	0|0	0|0	0|0
12	61687	.	G	A	100	PASS	.	GT	1|0	0|1	0|1
12	61700	.	C	T	100	PASS	.	GT	0|0	0|0	0|0

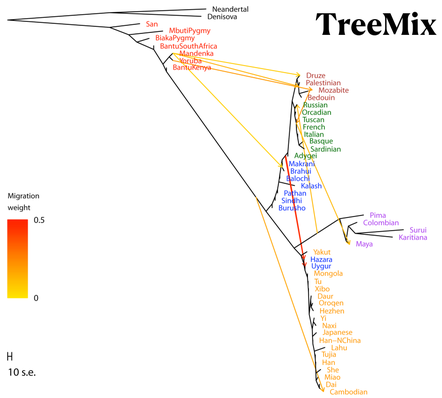
What we (usually) have

positions along a chromosome

									ind. #1	ind. #2	ind. #3	...
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00099	HG00101	
12	60076	.	A	C	100	PASS	.	GT	1|0	0|0	0|0	
12	60252	.	A	G	100	PASS	.	GT	0|0	0|0	0|1	
12	60317	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	60344	.	C	A	100	PASS	.	GT	0|0	0|0	0|0	
12	60383	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	
12	60405	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	
12	60474	.	G	A	100	PASS	.	GT	0|0	0|1	0|1	
12	60614	.	C	A	100	PASS	.	GT	0|0	0|0	0|0	
12	60628	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	
12	60654	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61021	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61107	.	G	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61172	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61220	.	G	A	100	PASS	.	GT	0|0	0|0	0|1	
12	61258	.	C	T	100	PASS	.	GT	0|0	0|0	0|1	
12	61272	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	
12	61329	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61341	.	G	A	100	PASS	.	GT	0|0	0|1	0|1	
12	61368	.	C	T	100	PASS	.	GT	0|0	0|1	0|1	
12	61392	.	T	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61405	.	G	C	100	PASS	.	GT	0|0	0|0	0|0	
12	61411	.	C	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61416	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61422	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61476	.	C	G	100	PASS	.	GT	0|0	0|0	0|0	
12	61510	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	
12	61516	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61552	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	
12	61604	.	T	G	100	PASS	.	GT	0|0	0|0	0|0	
12	61687	.	G	A	100	PASS	.	GT	1|0	0|1	0|1	
12	61700	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	

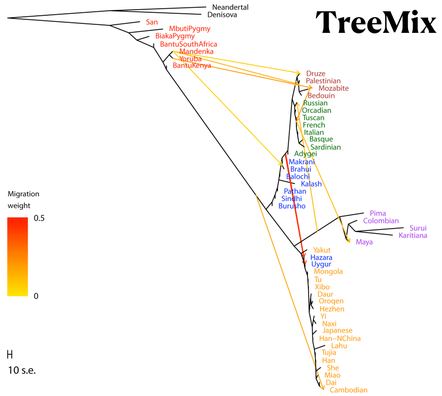
What we (usually) *want*

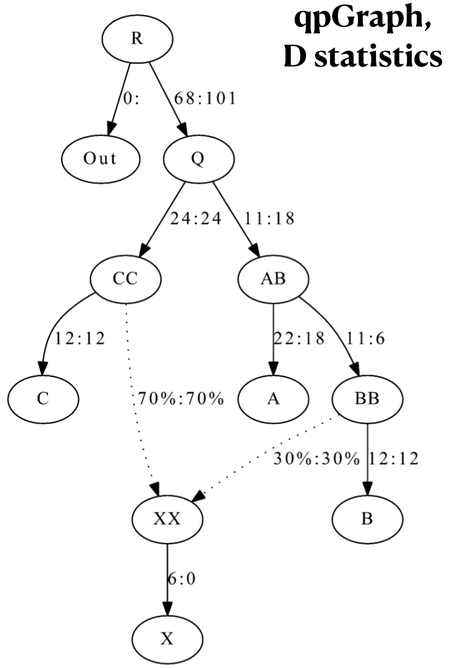
- estimate split times between populations



What we (usually) *want*

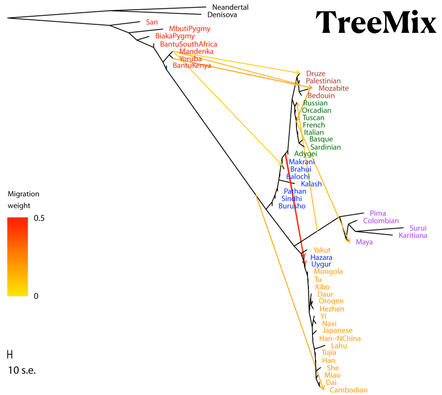
- estimate split times between populations
- investigate admixture hypotheses

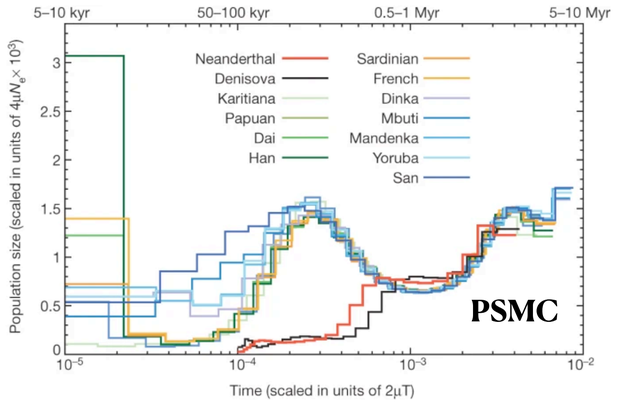


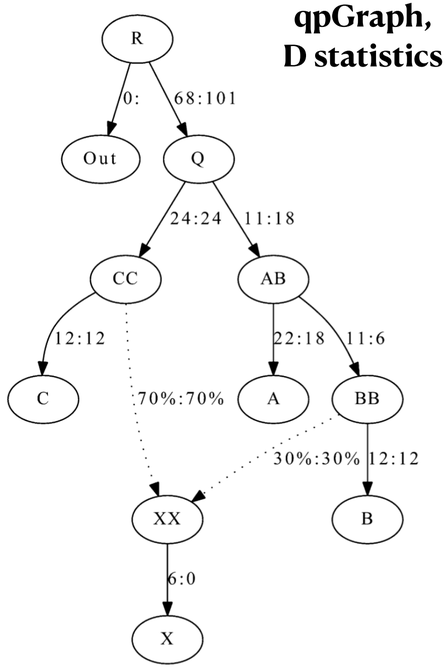


What we (usually) *want*

- estimate split times between populations
- investigate admixture hypotheses
- infer population sizes changes over time



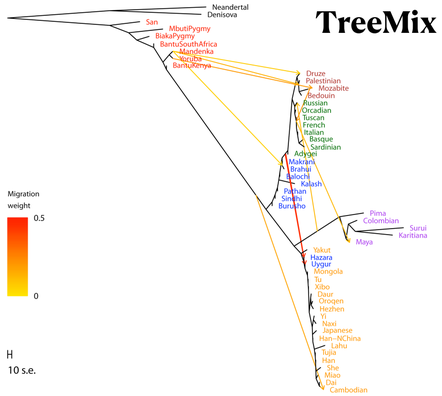


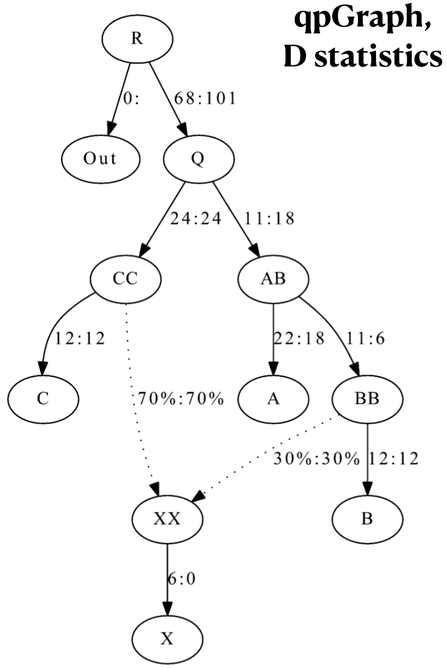


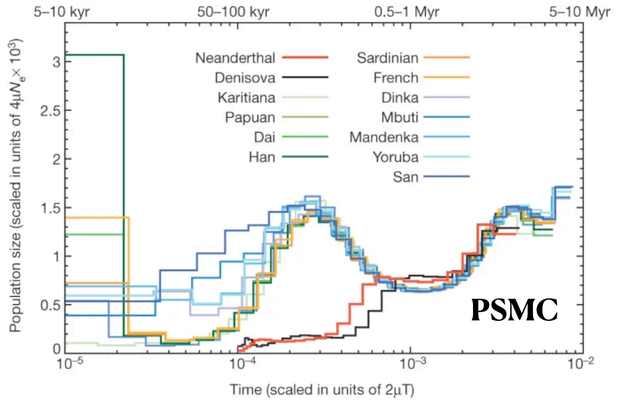
What we (usually) *want*

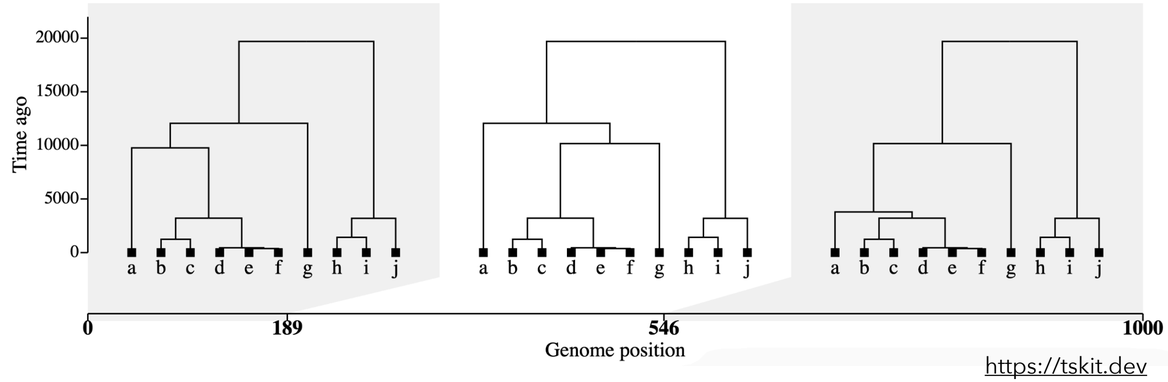
- estimate split times between populations
- investigate admixture hypotheses
- infer population sizes changes over time

These are all features of a real (but hidden) genealogical process.









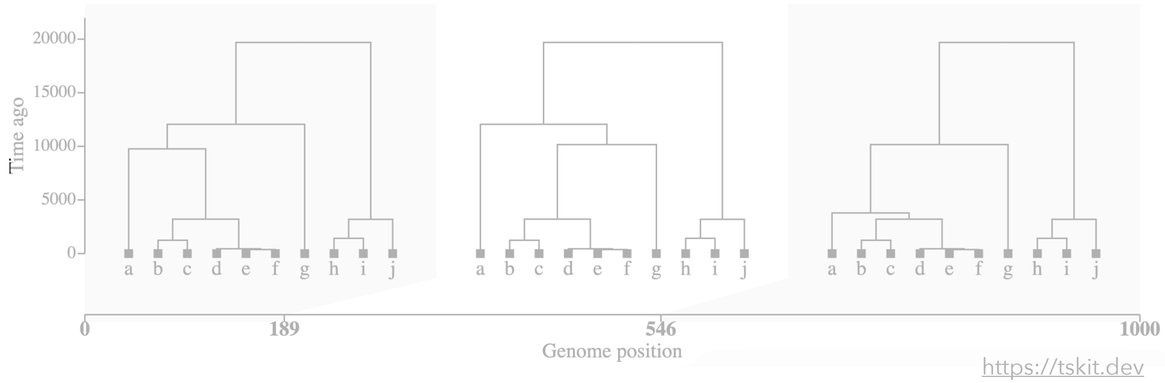
https://tskit.dev
"What is a tree sequence?"

What we (usually) *want*

- estimate split times between populations
- investigate admixture hypotheses
- infer population sizes changes over time

These are all features of a real (but hidden) genealogical process.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00099	HG00101
12	60076	.	A	C	100	PASS	.	GT	1|0	0|0	0|0
12	60252	.	A	G	100	PASS	.	GT	0|0	0|0	0|1
12	60317	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	60344	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60383	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	60405	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60474	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	60614	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60628	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60654	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61021	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61107	.	G	T	100	PASS	.	GT	0|0	0|0	0|0
12	61172	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61220	.	G	A	100	PASS	.	GT	0|0	0|0	0|1
12	61258	.	C	T	100	PASS	.	GT	0|0	0|0	0|1
12	61272	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	61329	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61341	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	61368	.	C	T	100	PASS	.	GT	0|0	0|1	0|1
12	61392	.	T	A	100	PASS	.	GT	0|0	0|0	0|0
12	61405	.	G	C	100	PASS	.	GT	0|0	0|0	0|0
12	61411	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	61416	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61422	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61476	.	C	G	100	PASS	.	GT	0|0	0|0	0|0
12	61510	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61516	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61552	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61604	.	T	G	100	PASS	.	GT	0|0	0|0	0|0
12	61687	.	G	A	100	PASS	.	GT	1|0	0|1	0|1
12	61700	.	C	T	100	PASS	.	GT	0|0	0|0	0|0



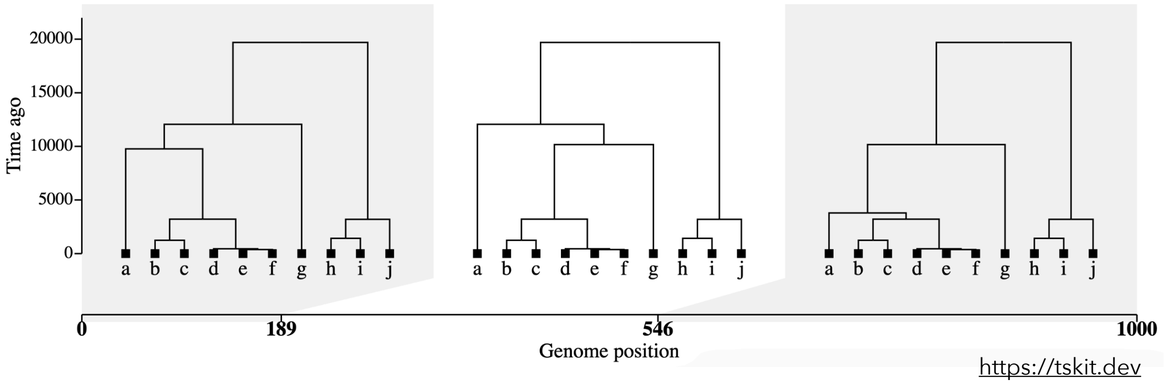
https://tskit.dev
"What is a tree sequence?"

What we (usually) *want*

- estimate split times between populations
- investigate admixture hypotheses
- infer population sizes changes over time

These are all features of a real (but hidden) genealogical process.

If we have a tree sequence, we get direct answers almost "for free".



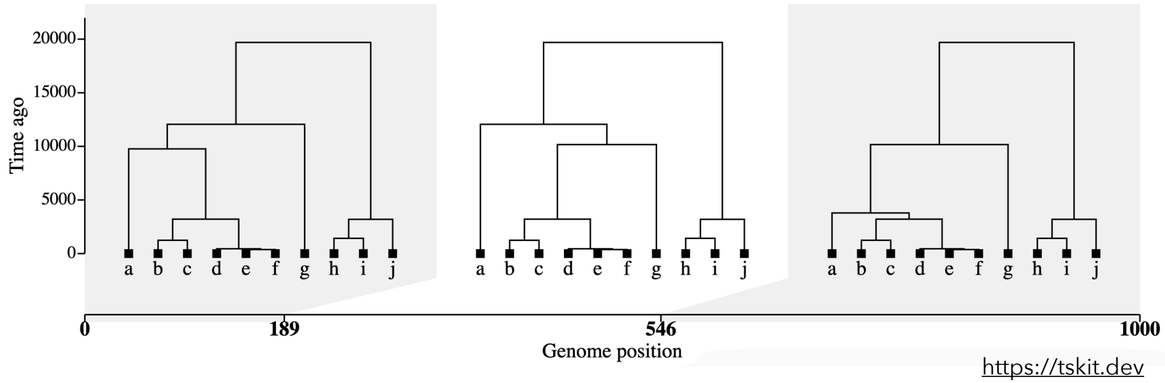
https://tskit.dev
"What is a tree sequence?"

What we (usually) *want*

- estimate split times between populations
- investigate admixture hypotheses
- infer population sizes changes over time

These are all features of a real (but hidden) genealogical process.

If we have a tree sequence, we get direct answers almost "for free".

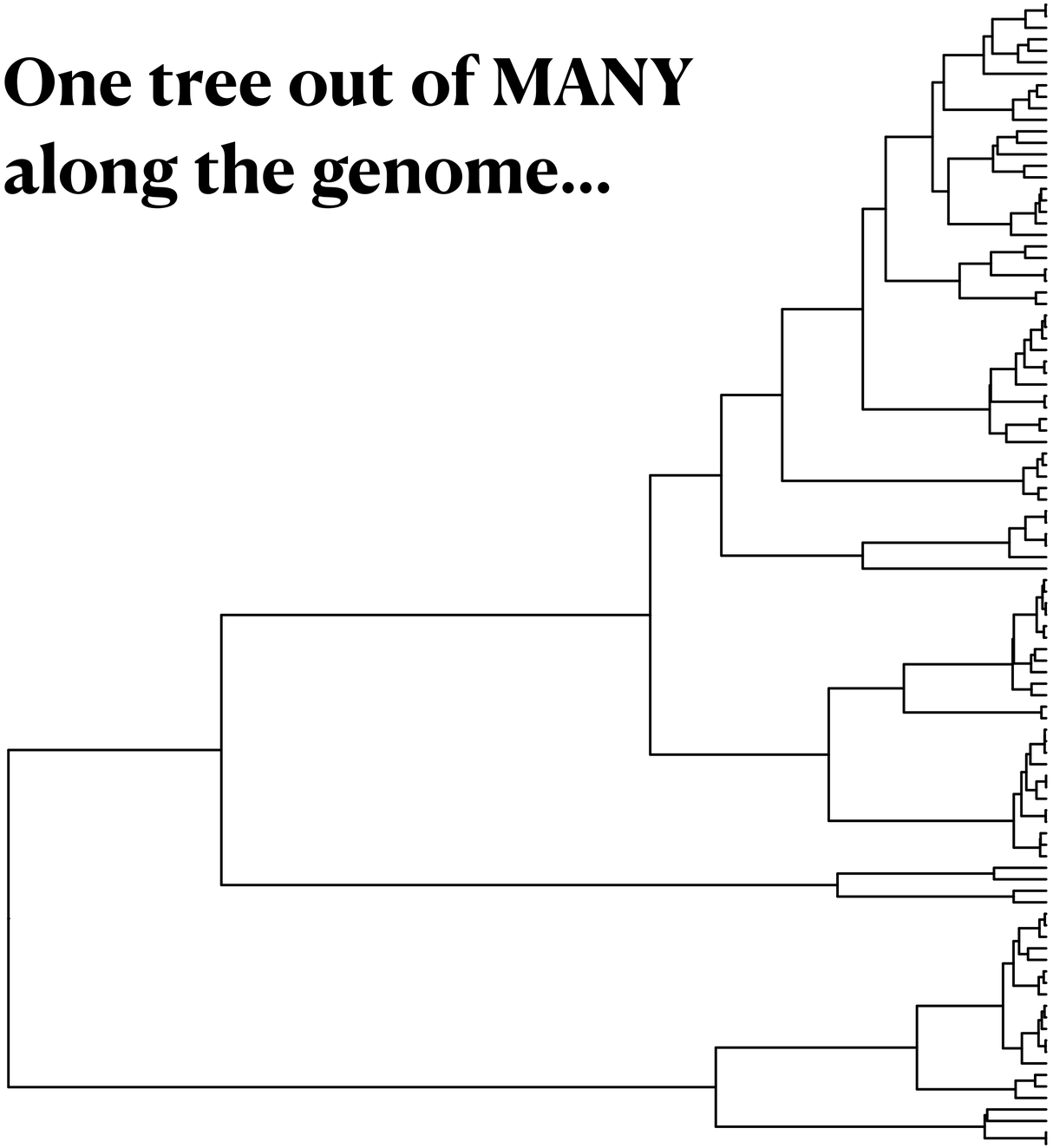


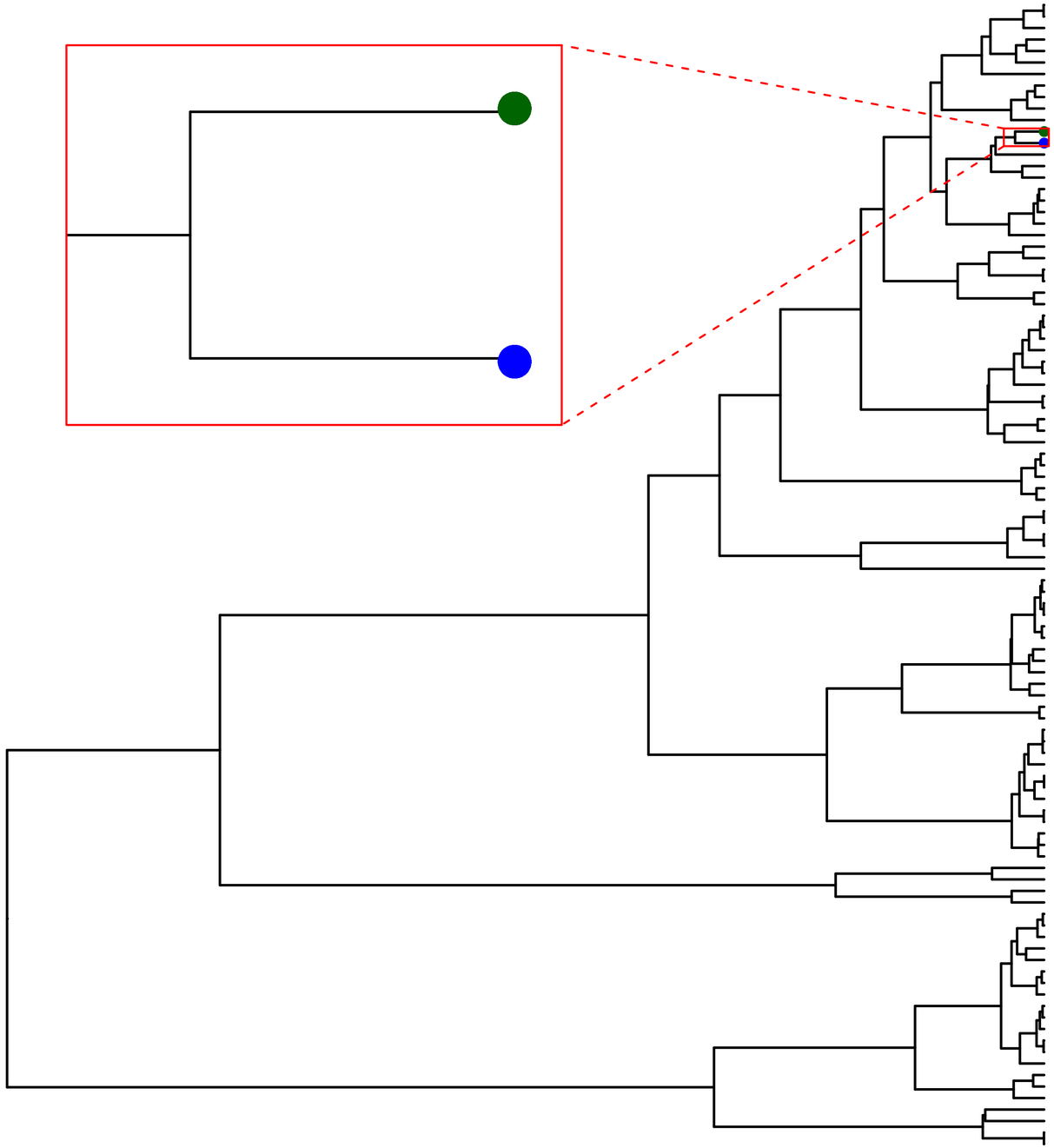
https://tskit.dev
"What is a tree sequence?"

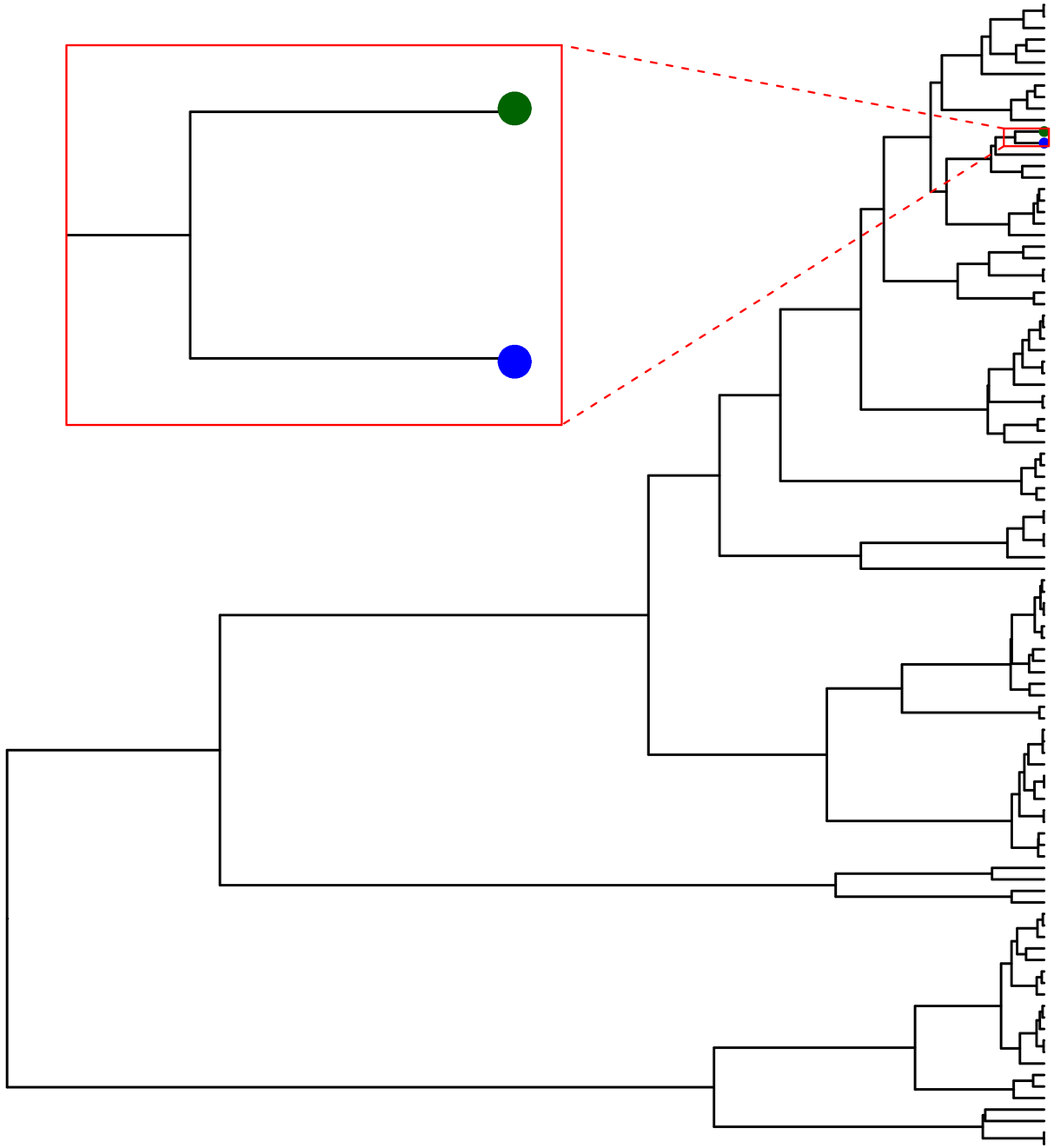
[... although the tree sequence must first be inferred from real data.]

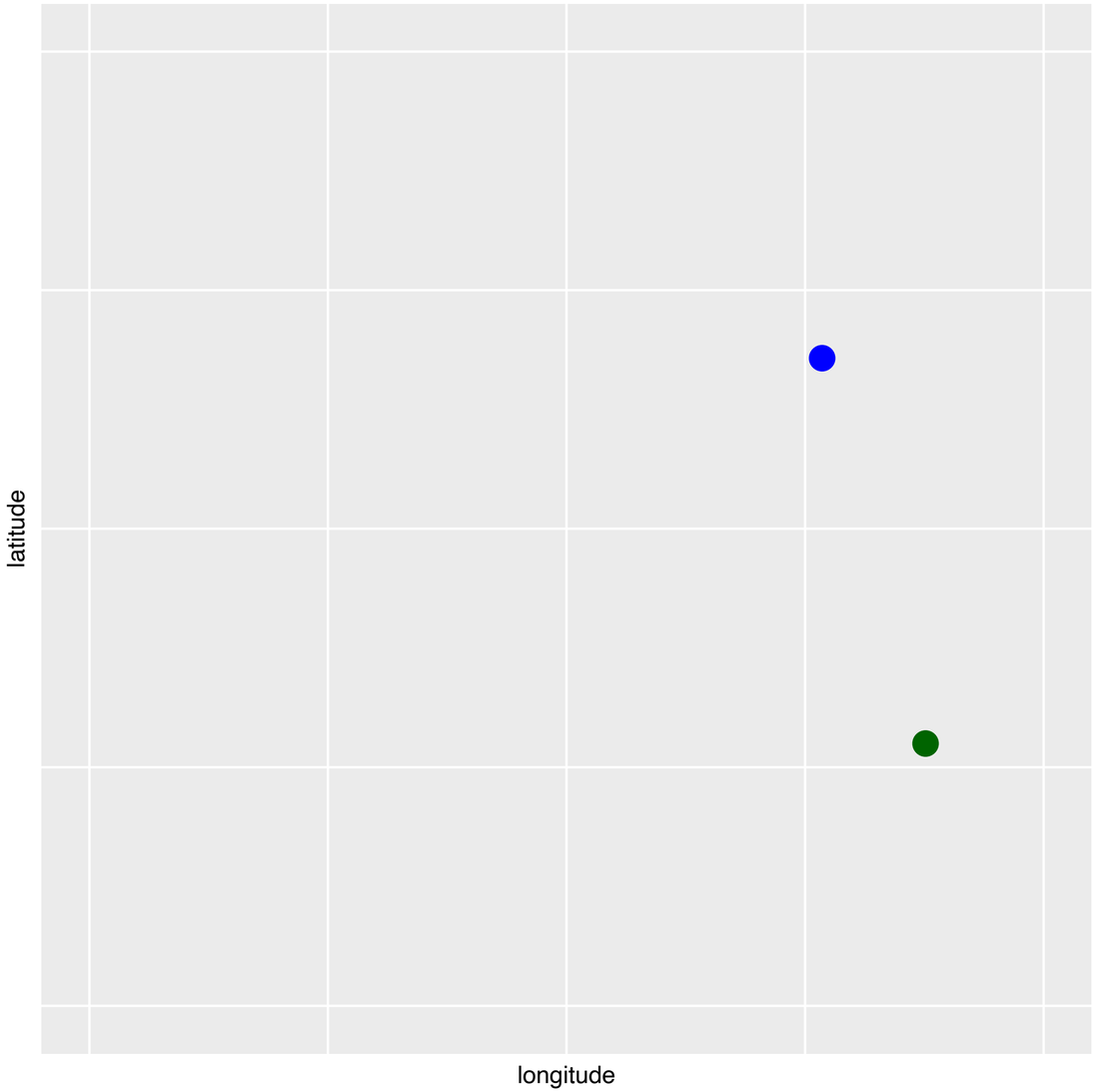

Tree sequences will
push spatial inference
to the next level.

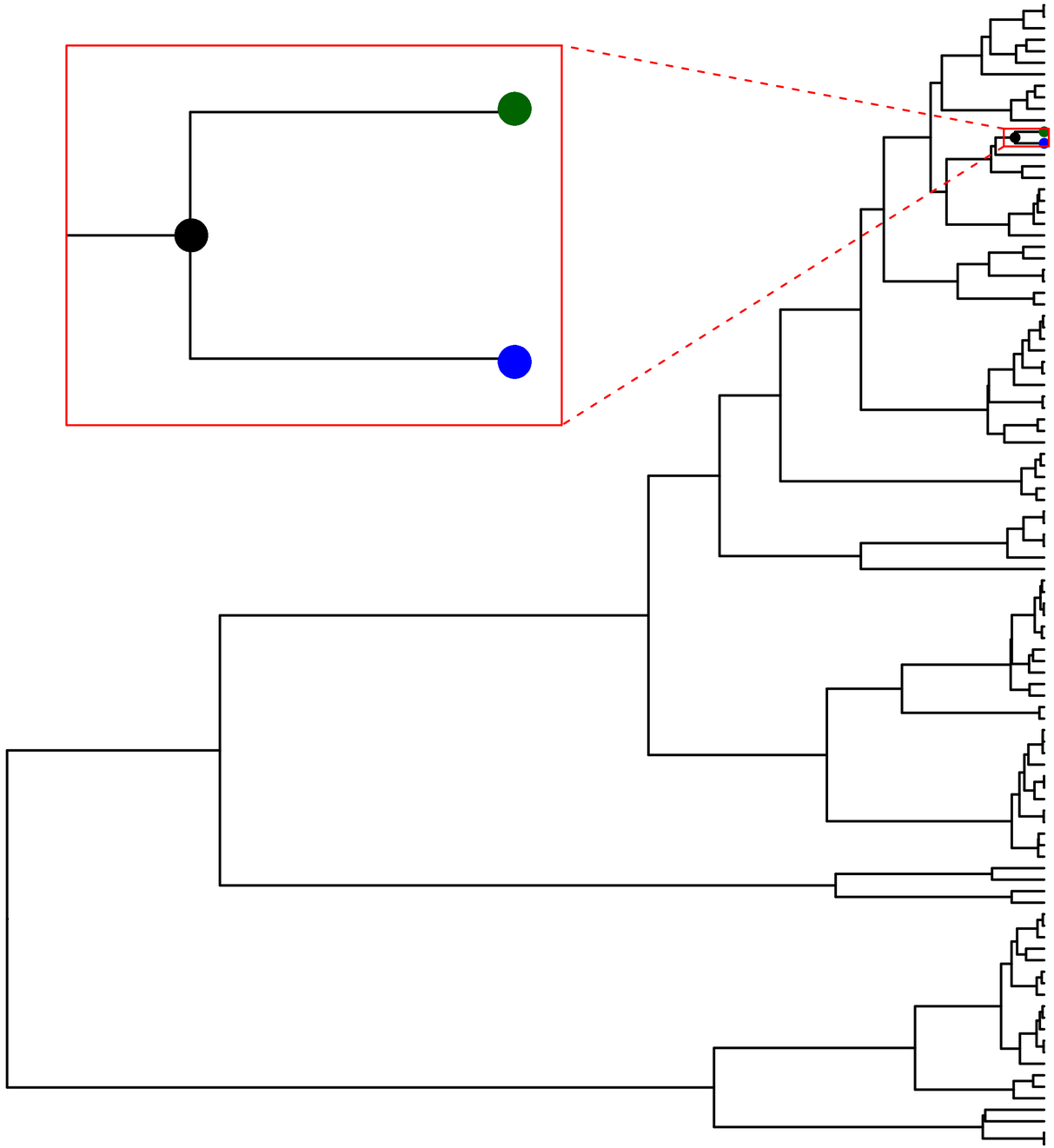
One tree out of MANY along the genome...

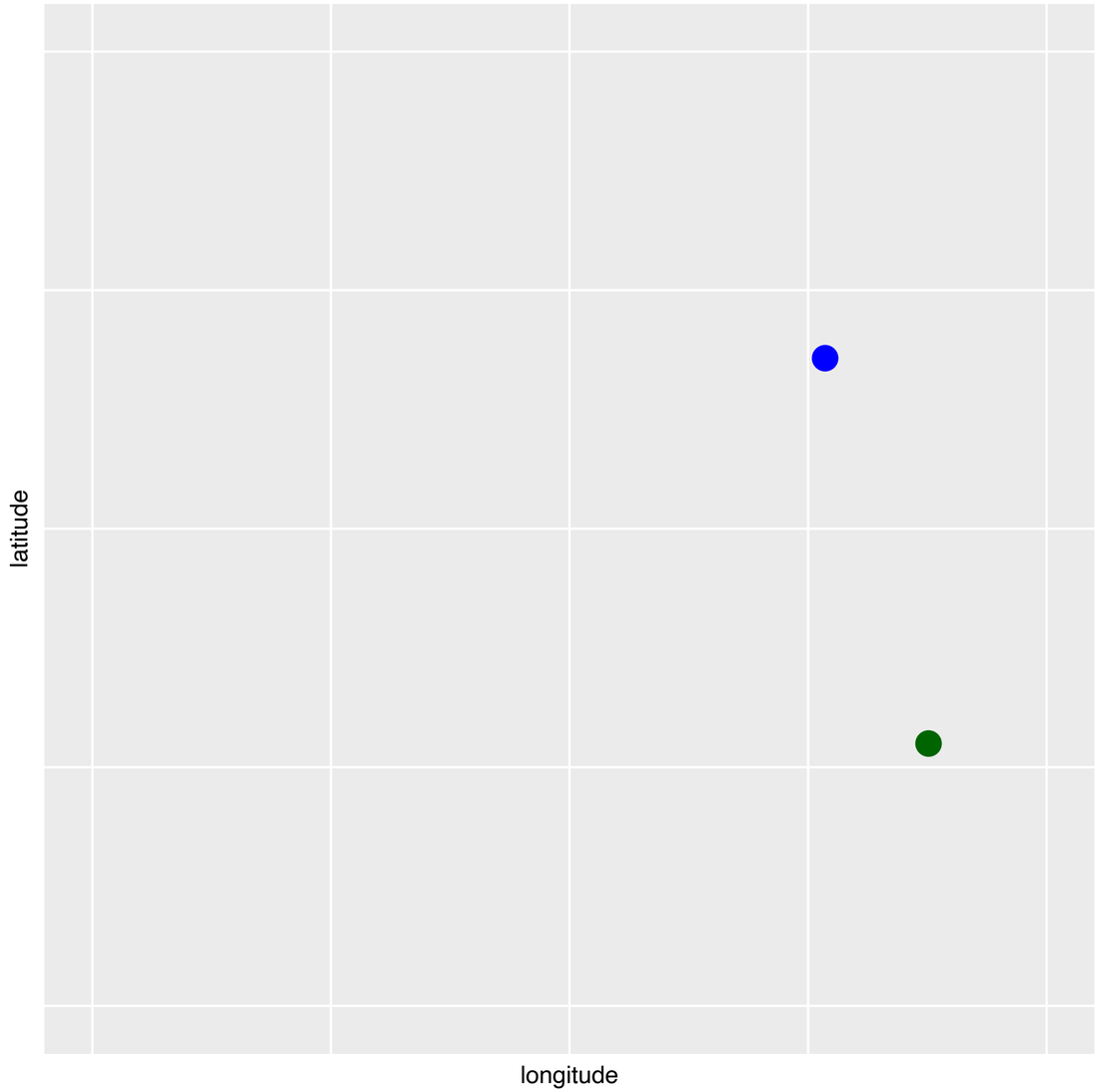


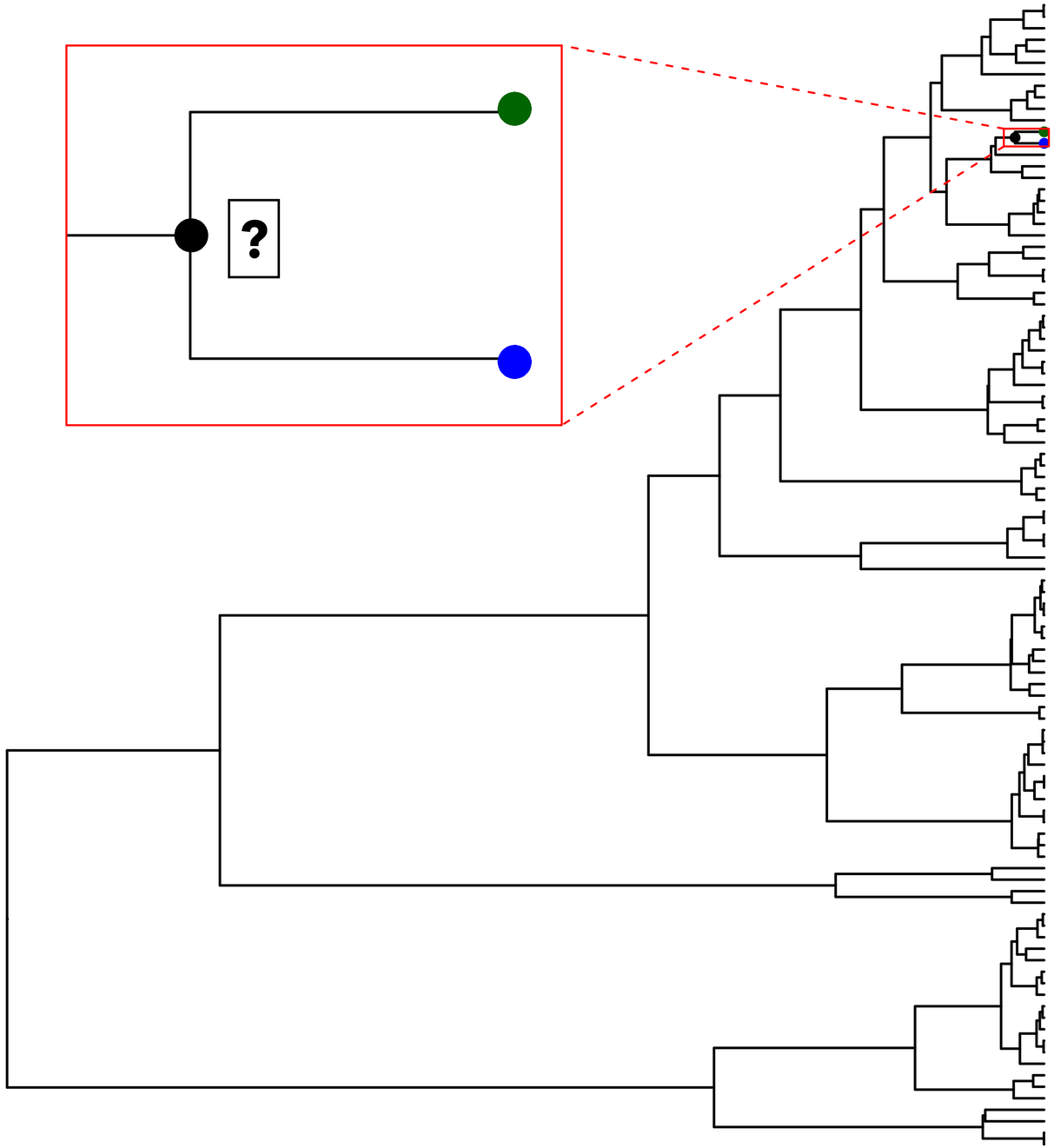


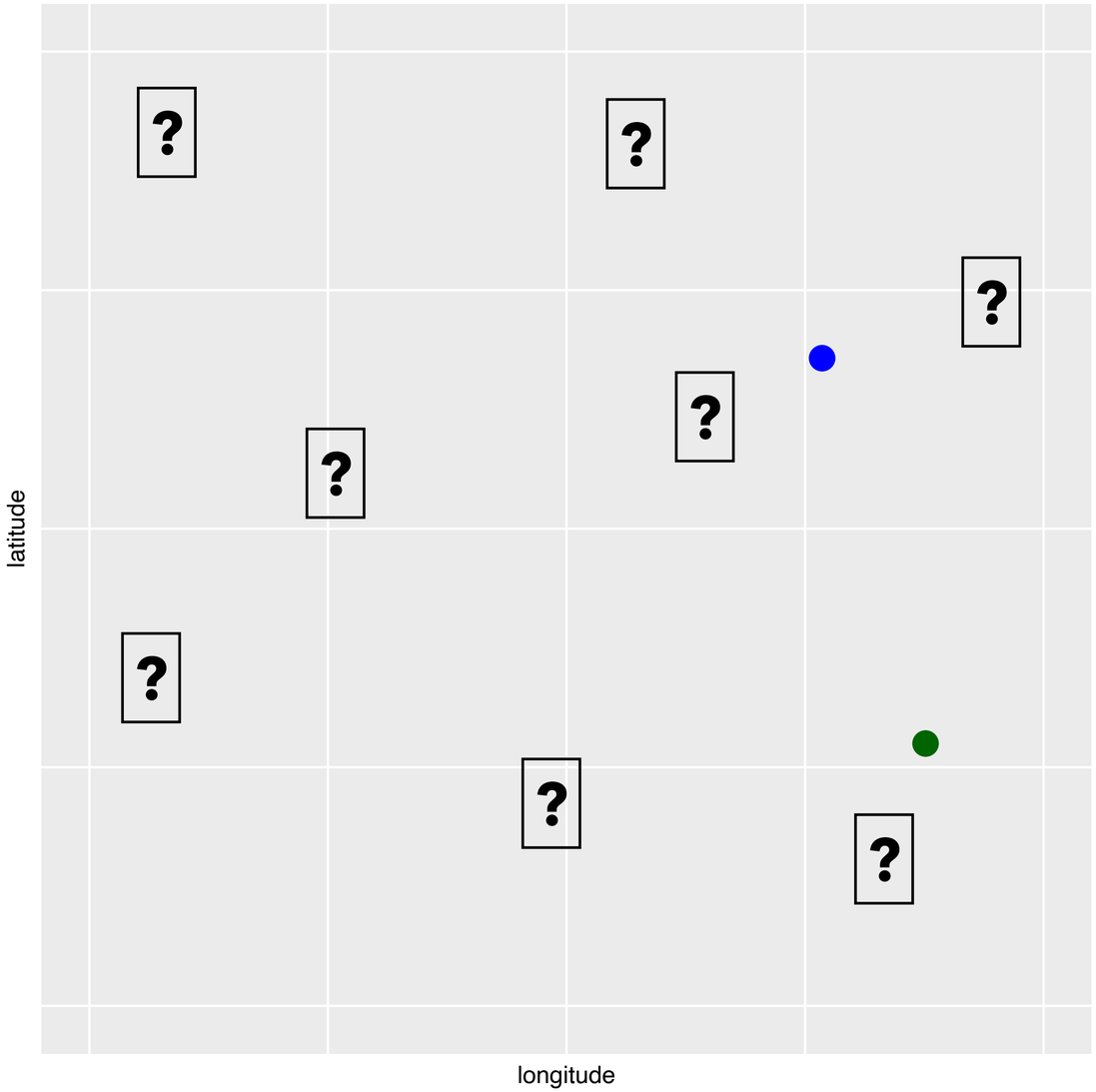


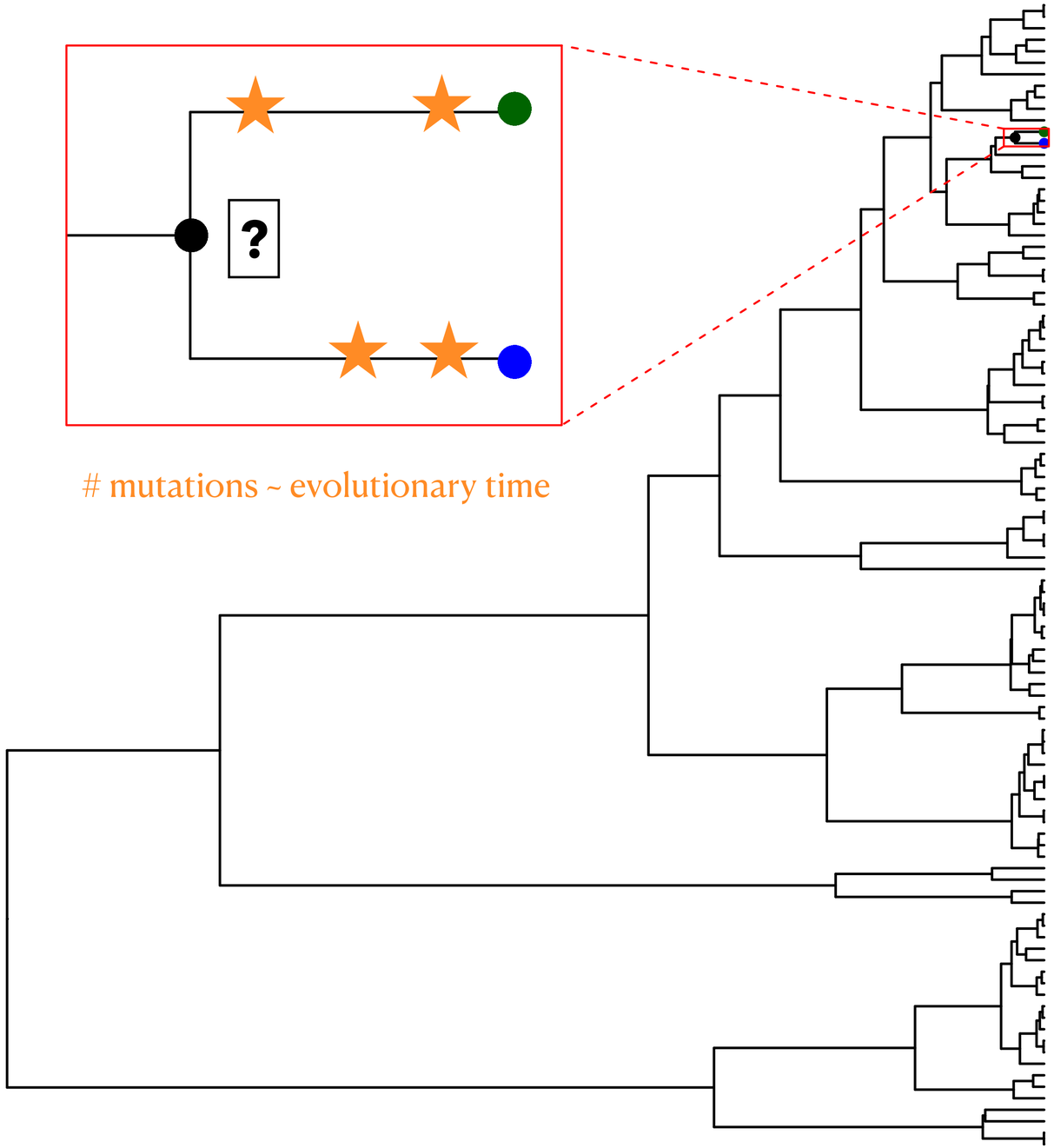

latitude
longitude

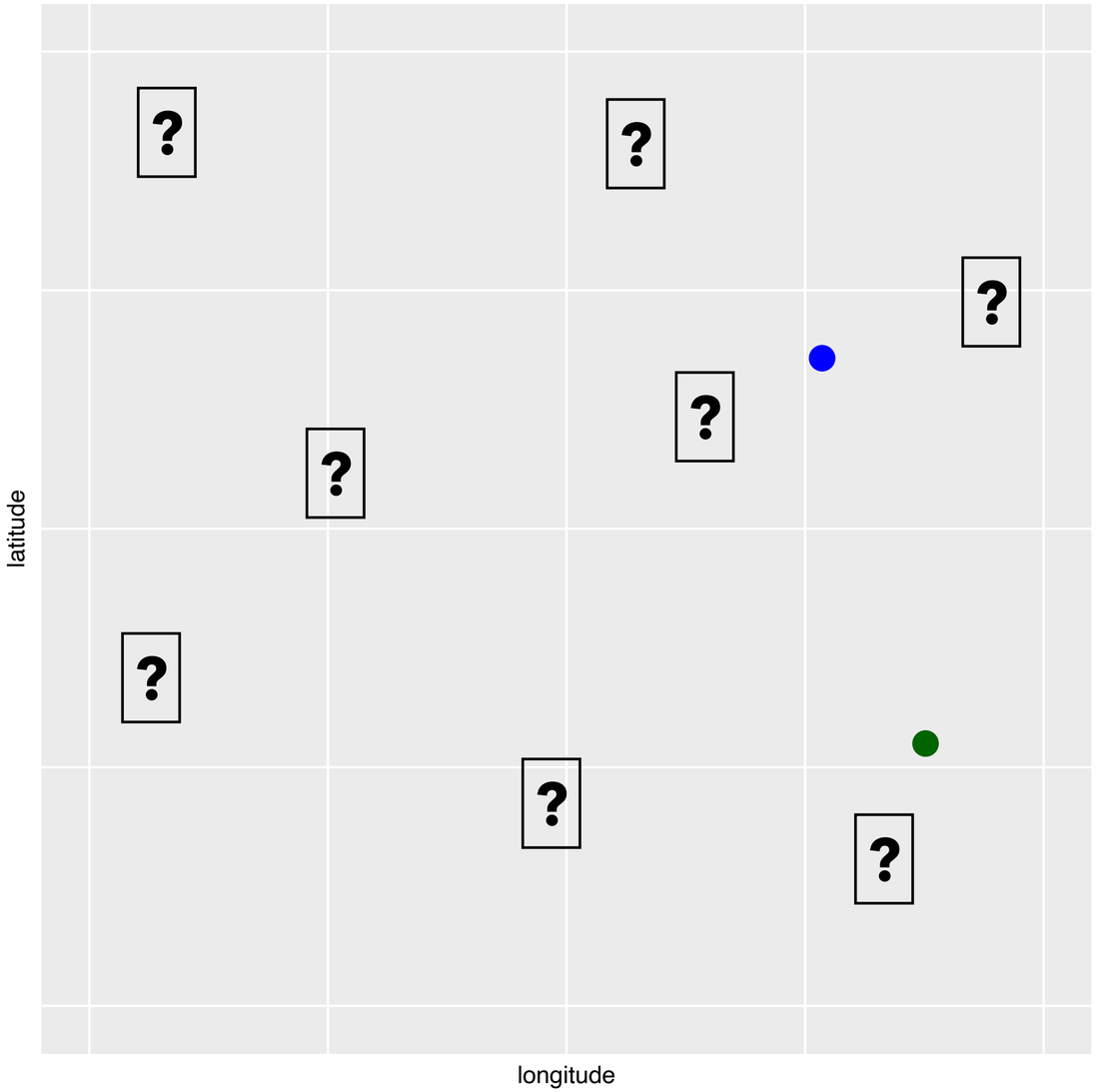


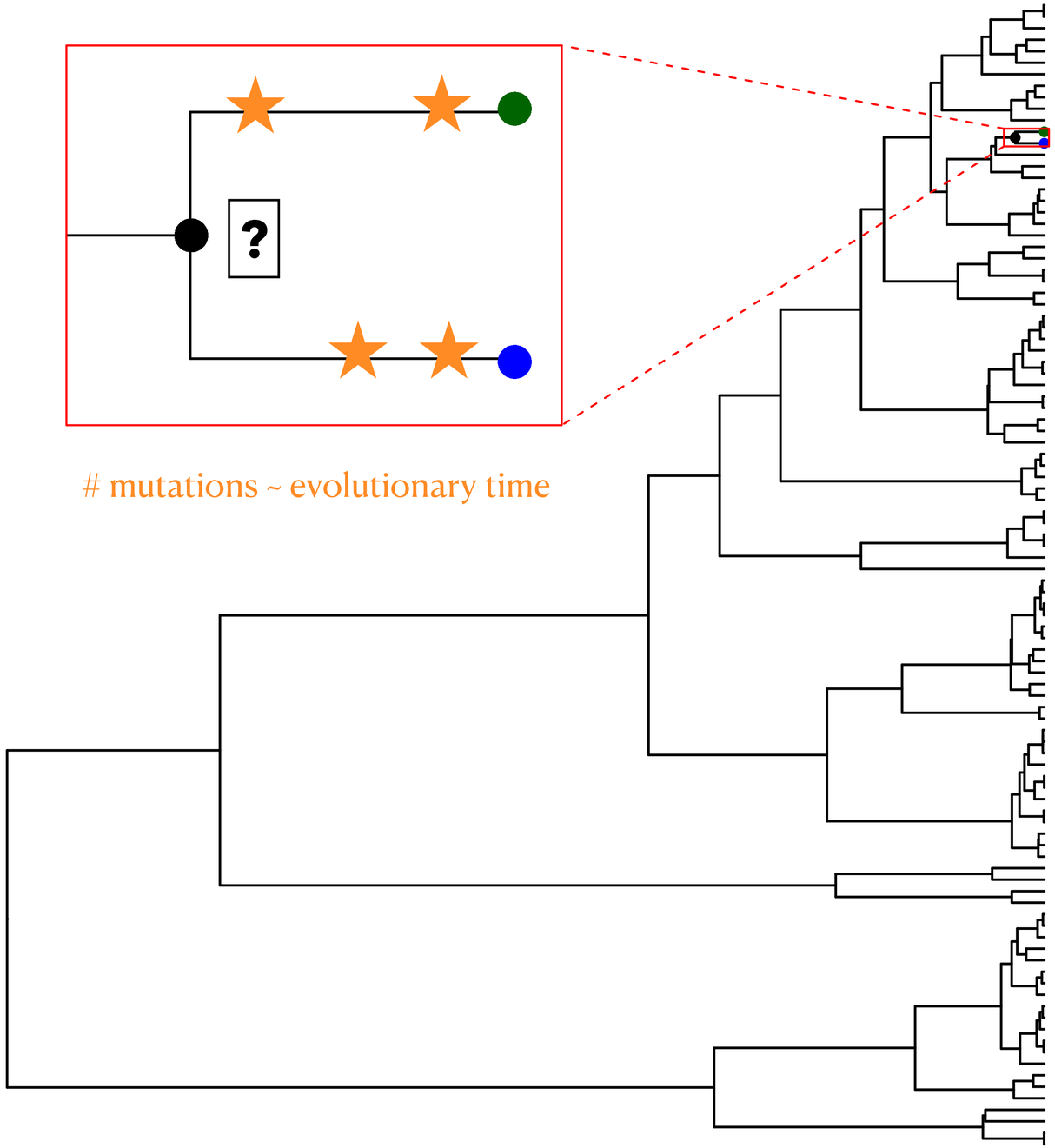

latitude
longitude

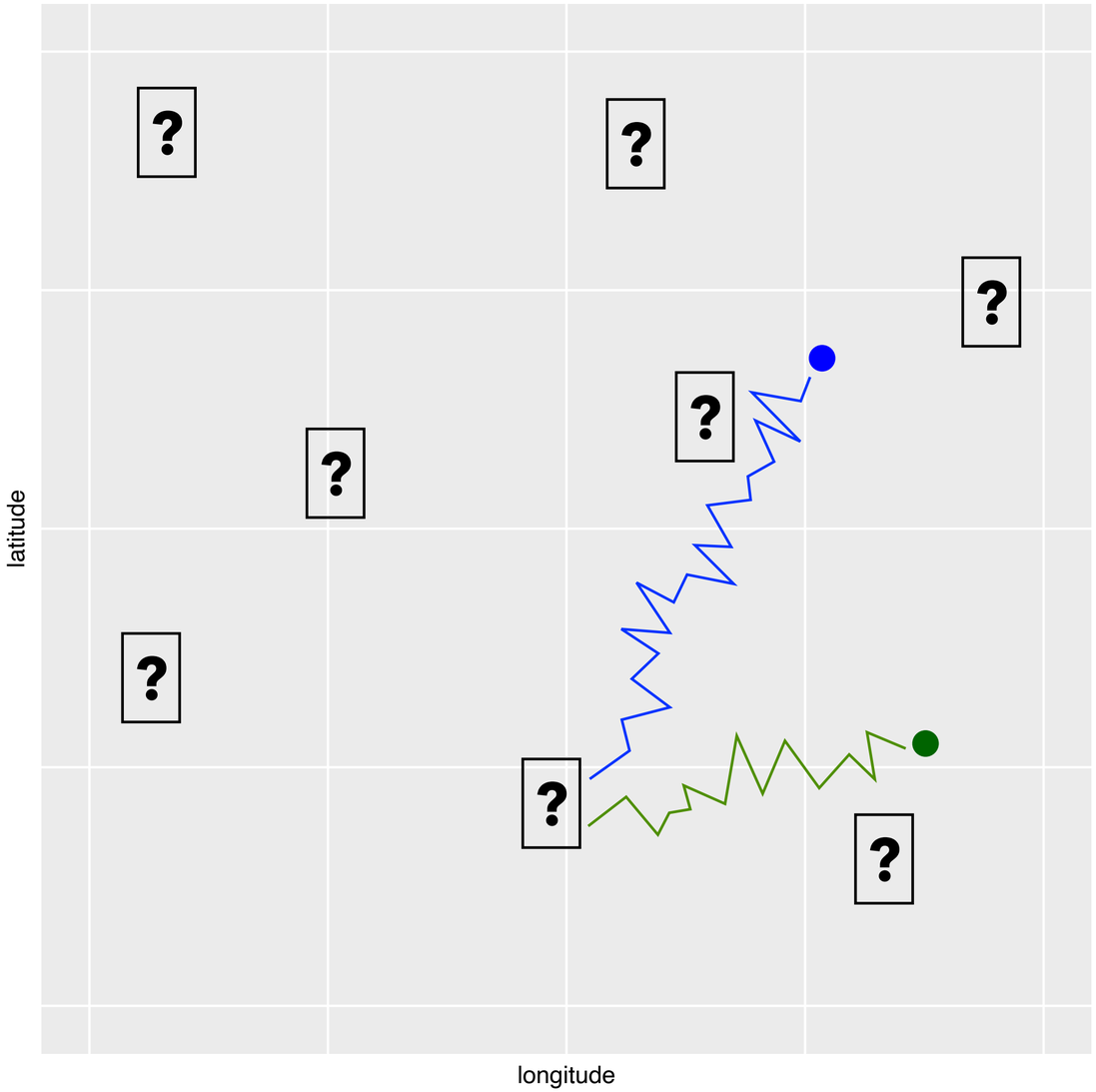

?

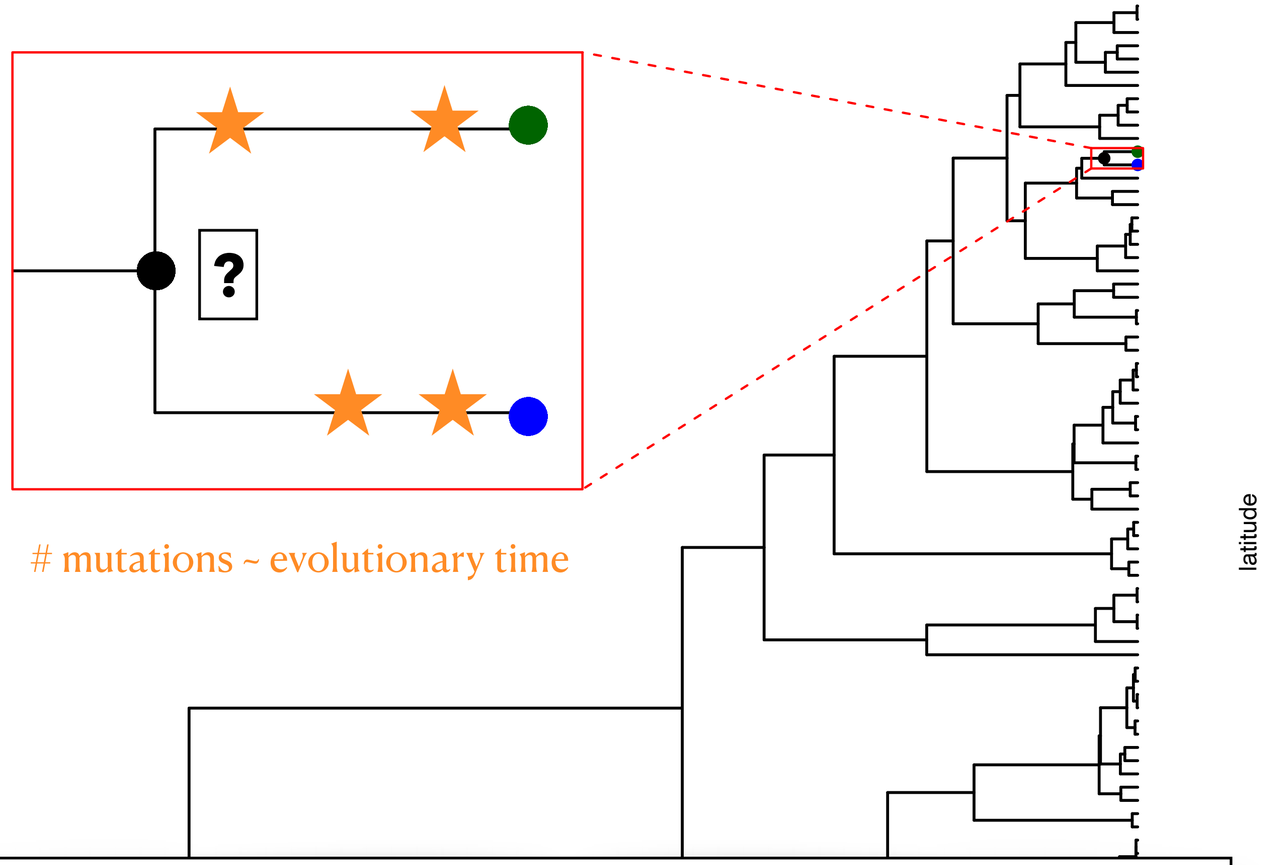

?
?
?
?
?
?
?
latitude
longitude

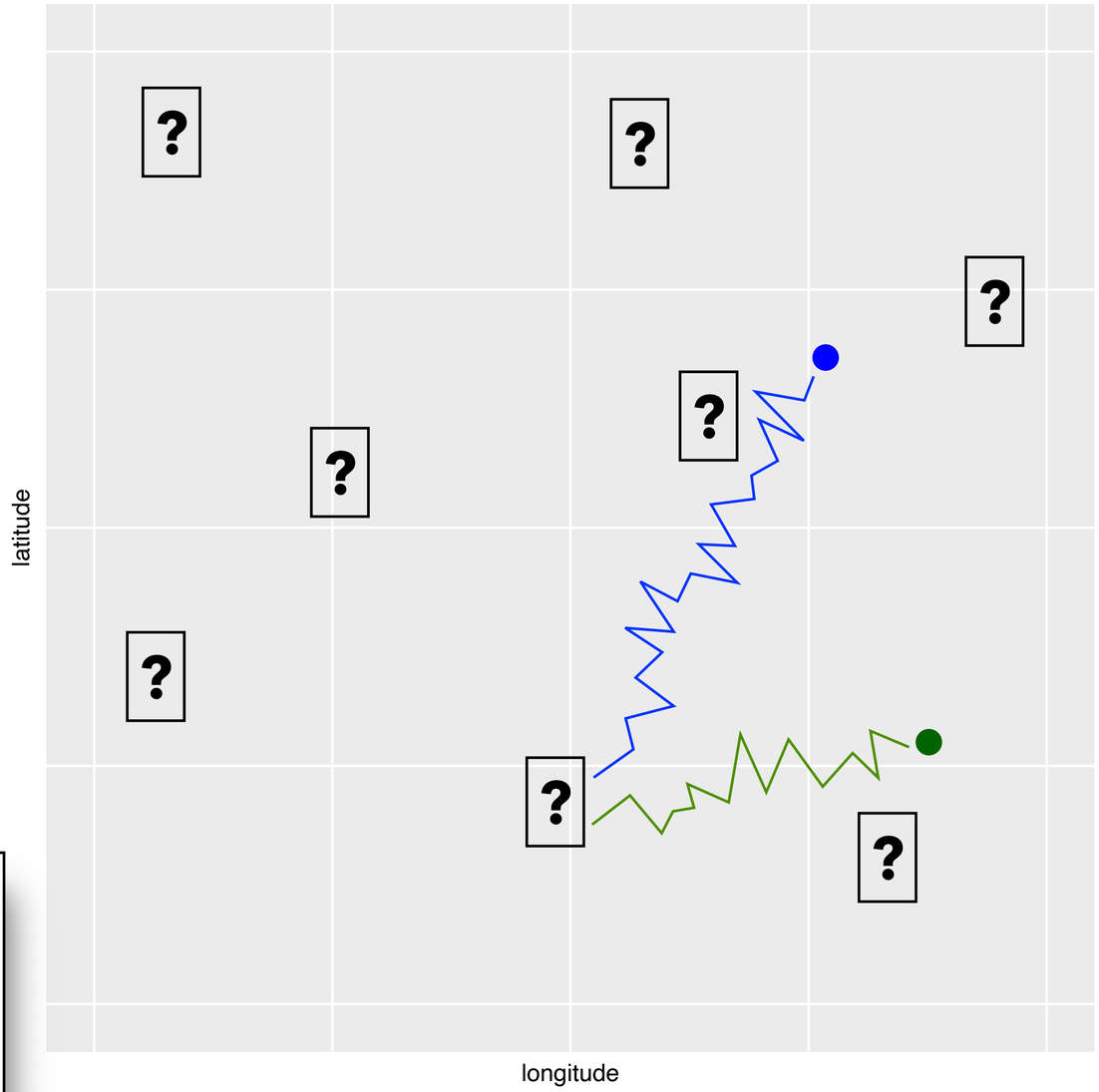

?
mutations ~ evolutionary time


?
?
?
?
?
?
?
latitude
longitude


?
mutations ~ evolutionary time


?
?
?
?
?
?
?
latitude
longitude

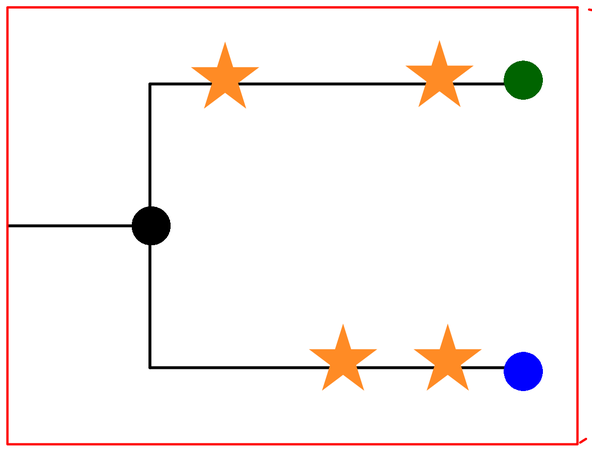




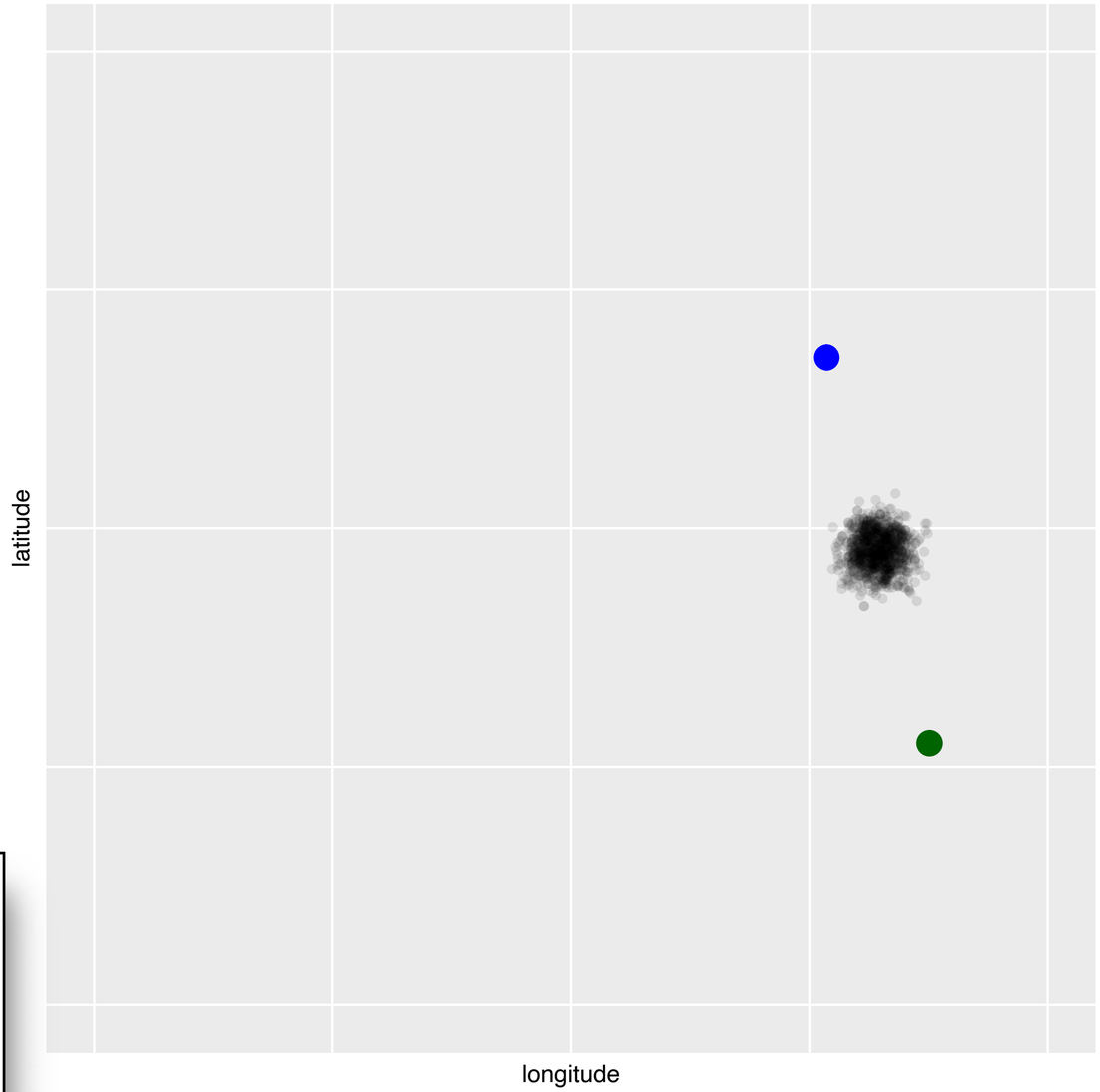
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro



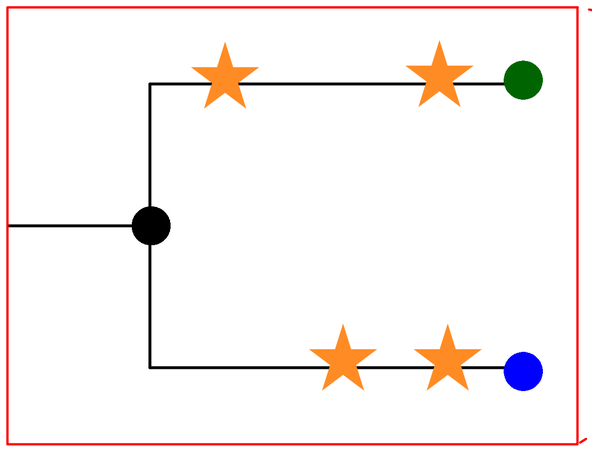
mutations ~ evolutionary time



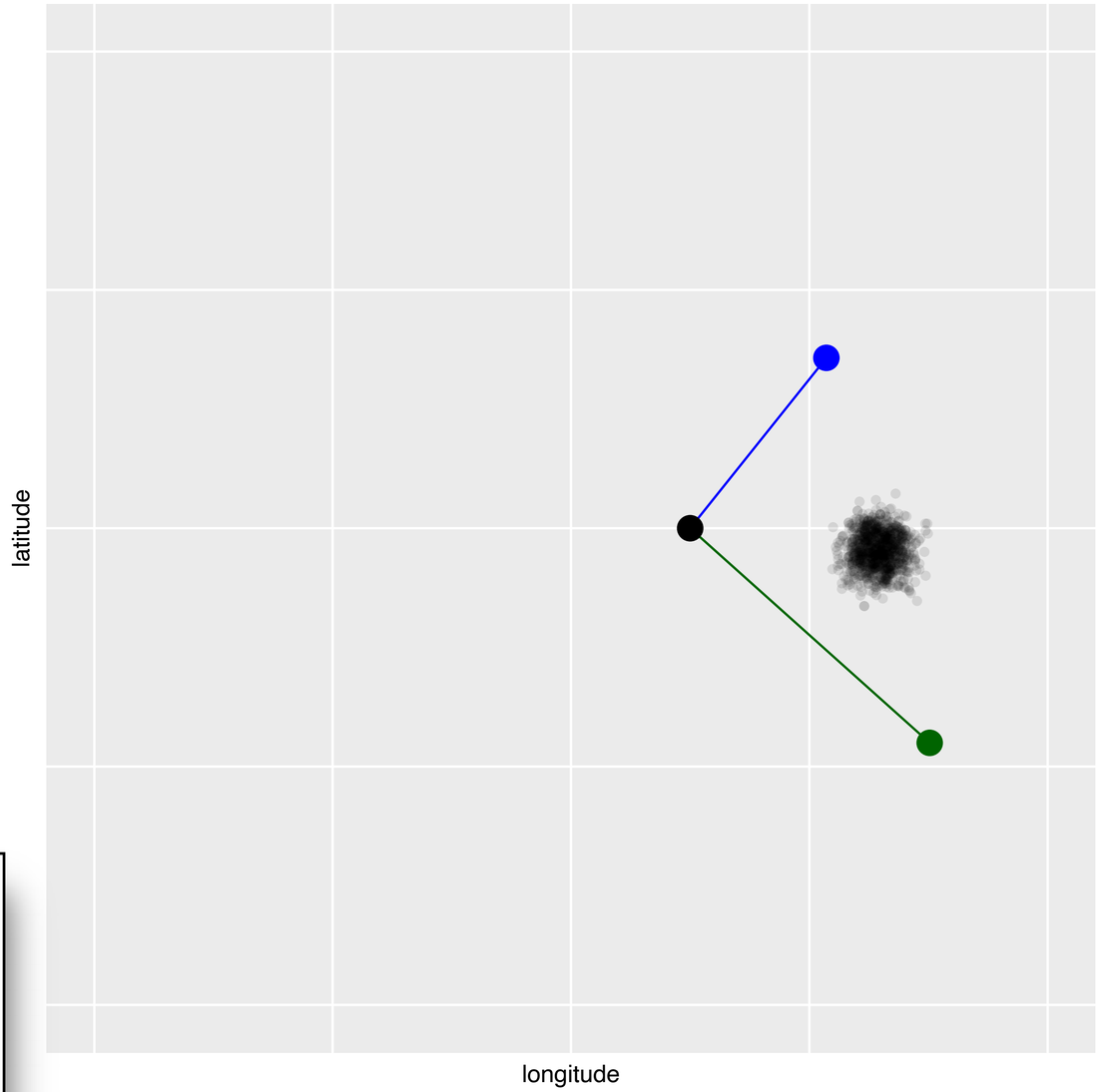
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro



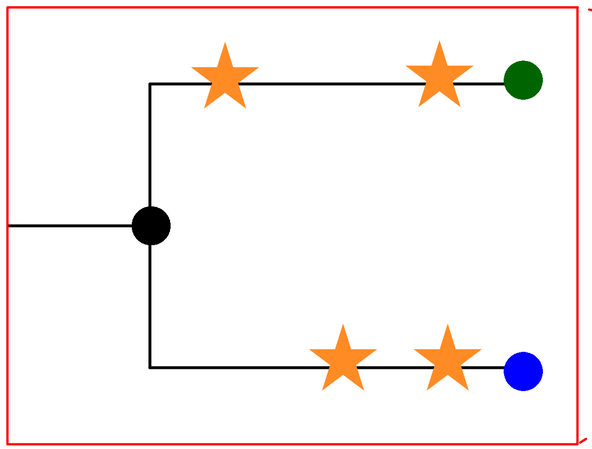
mutations ~ evolutionary time



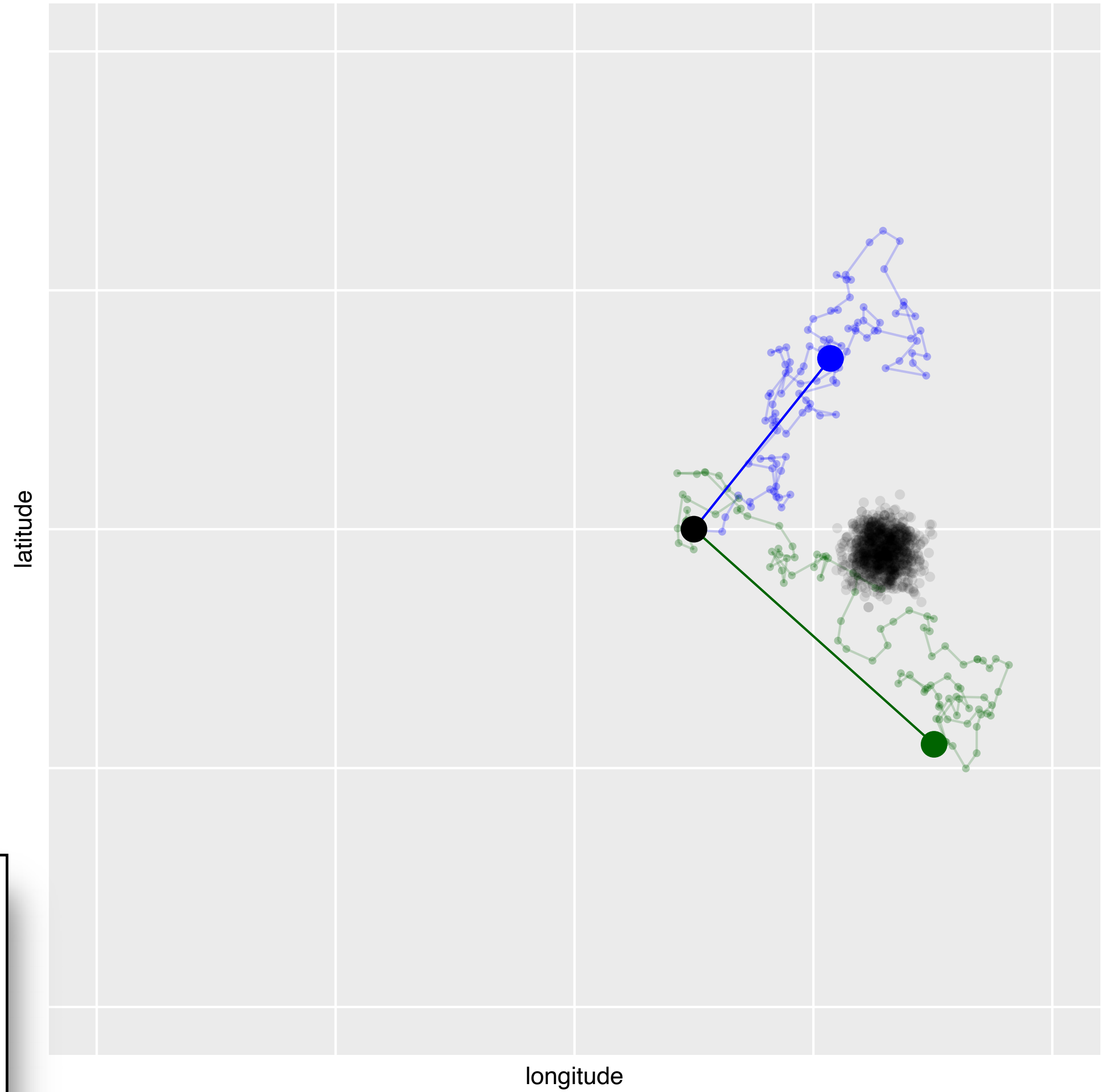
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro



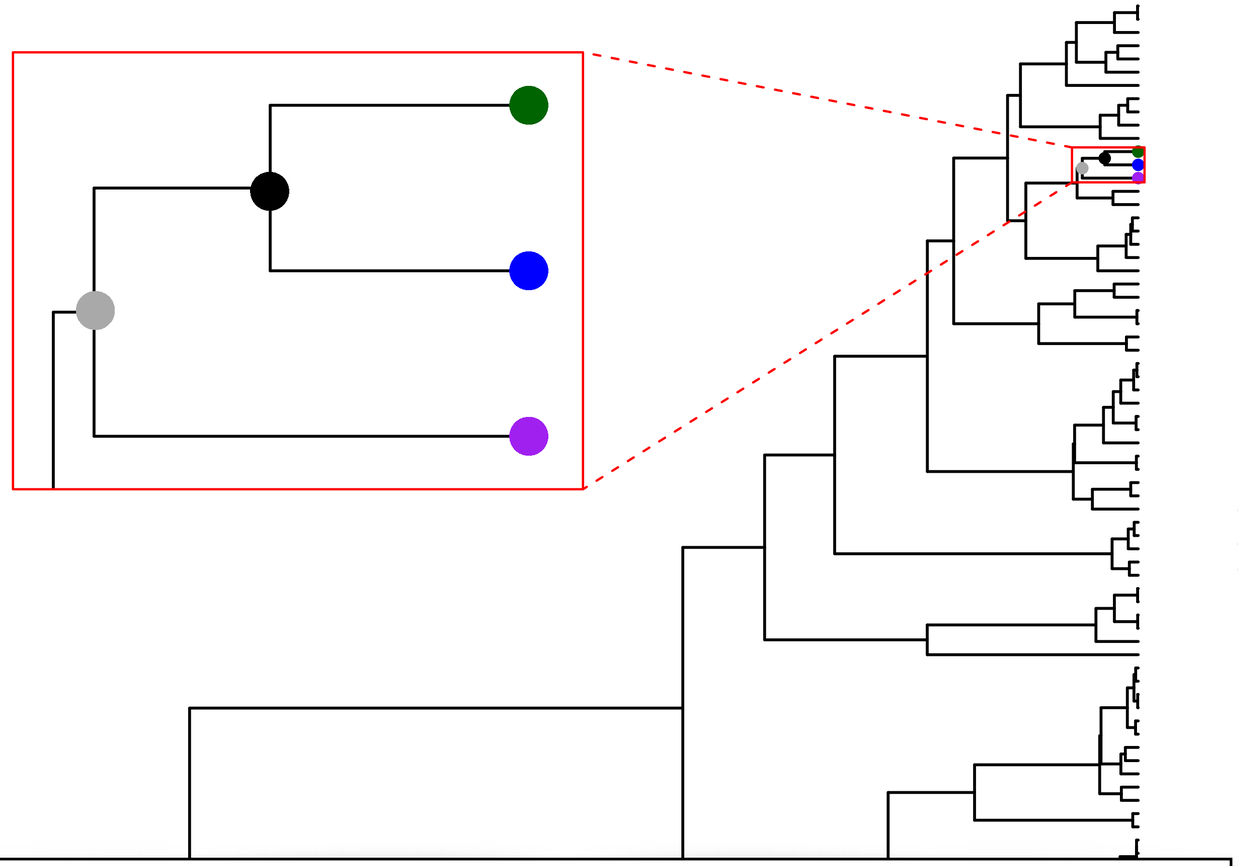
mutations ~ evolutionary time

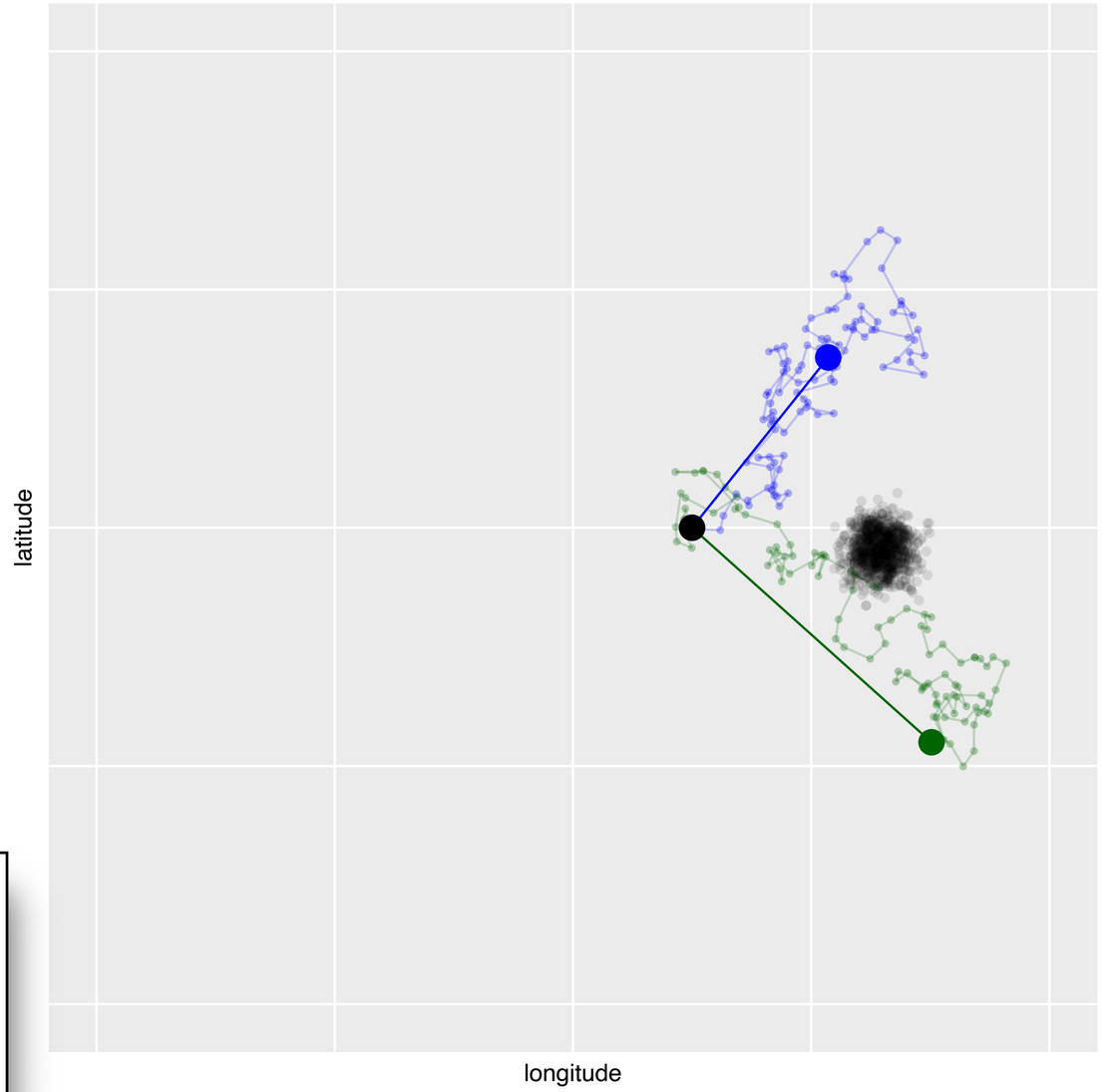


Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

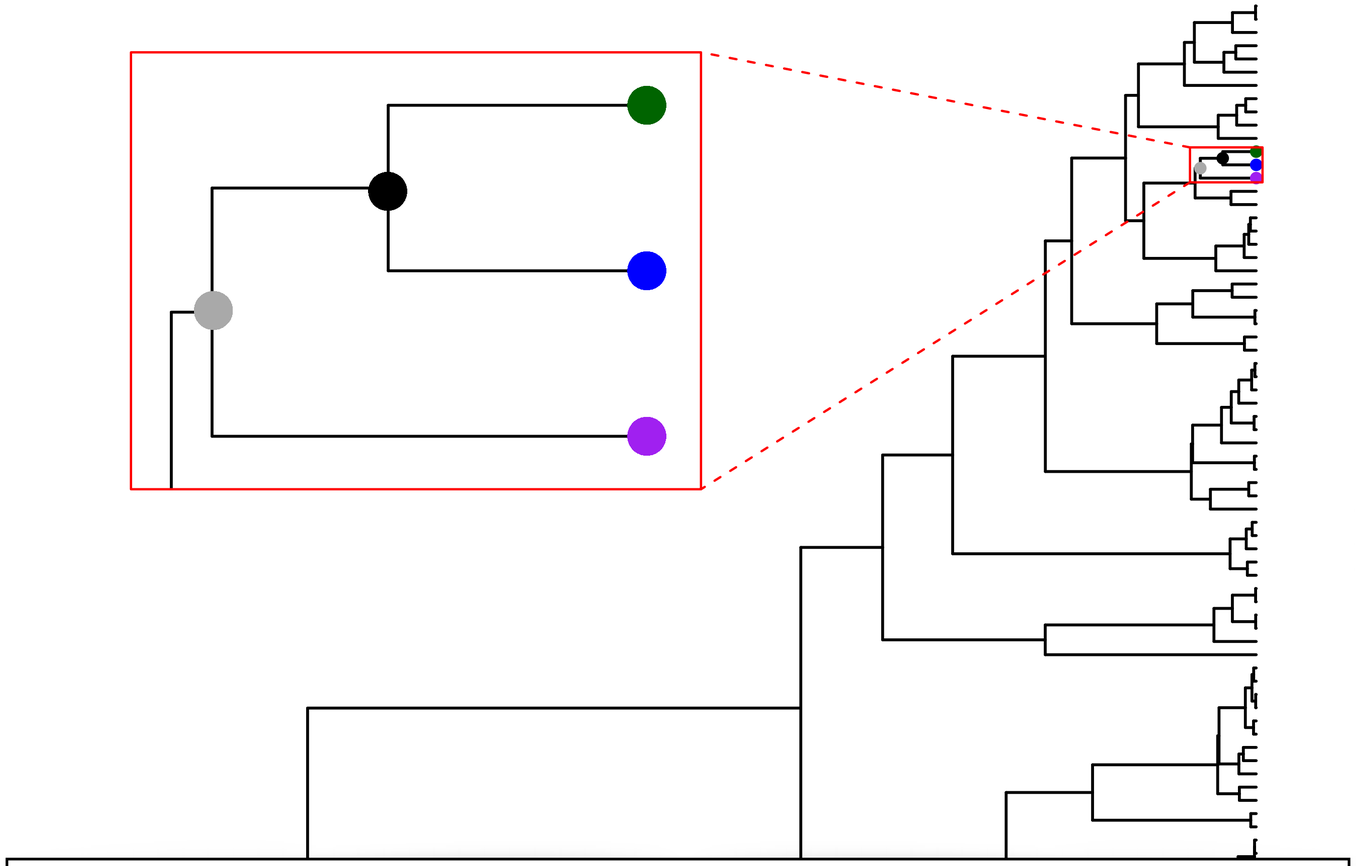


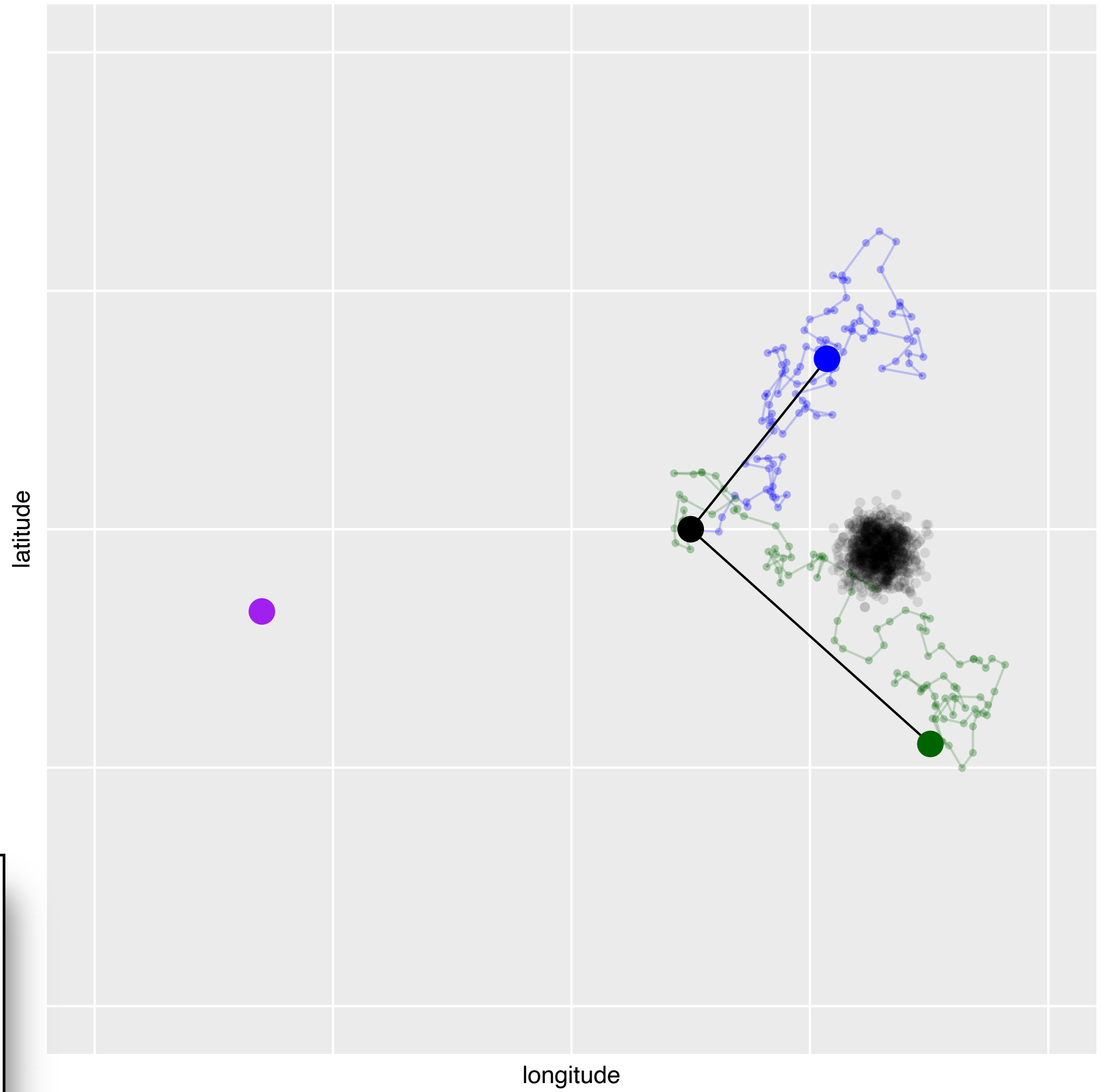


Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

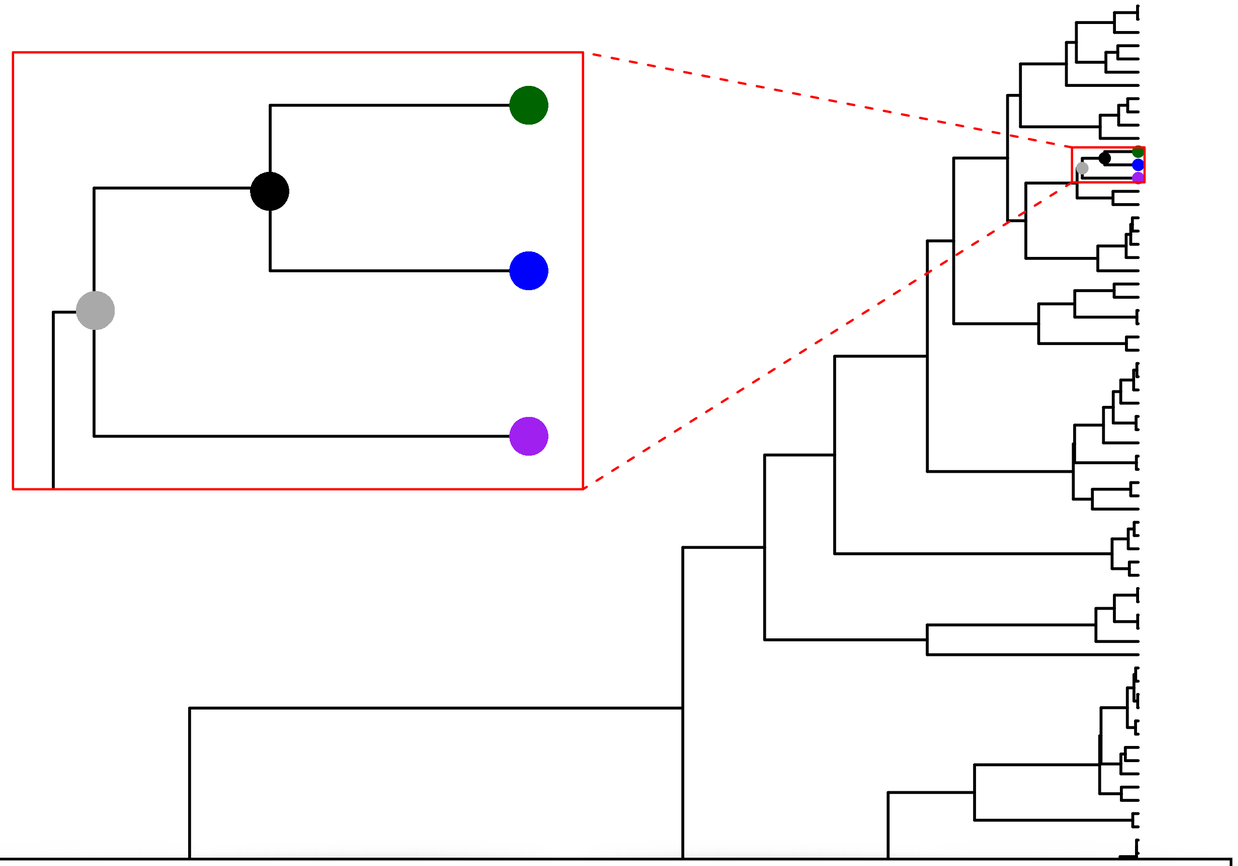


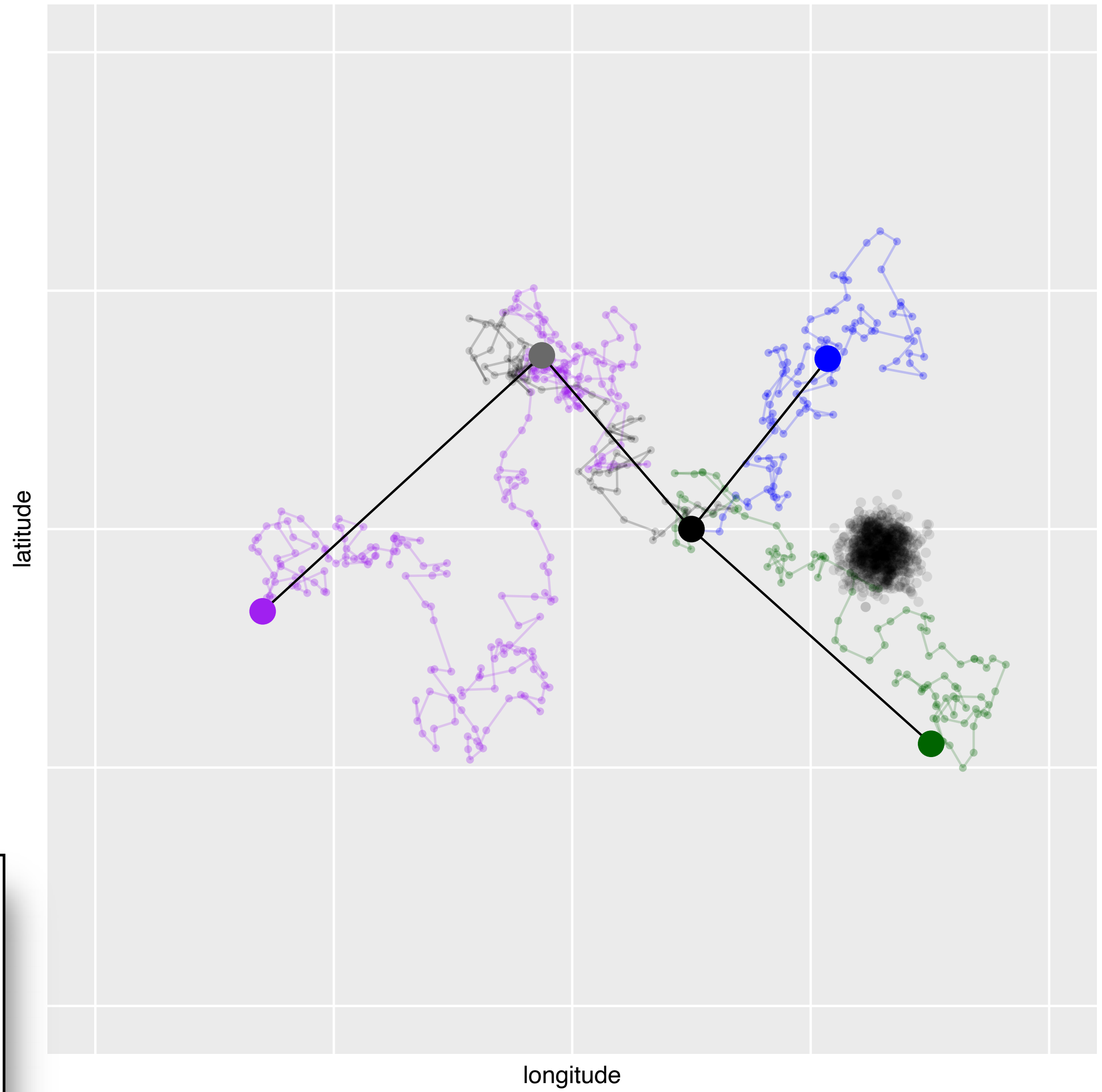


Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

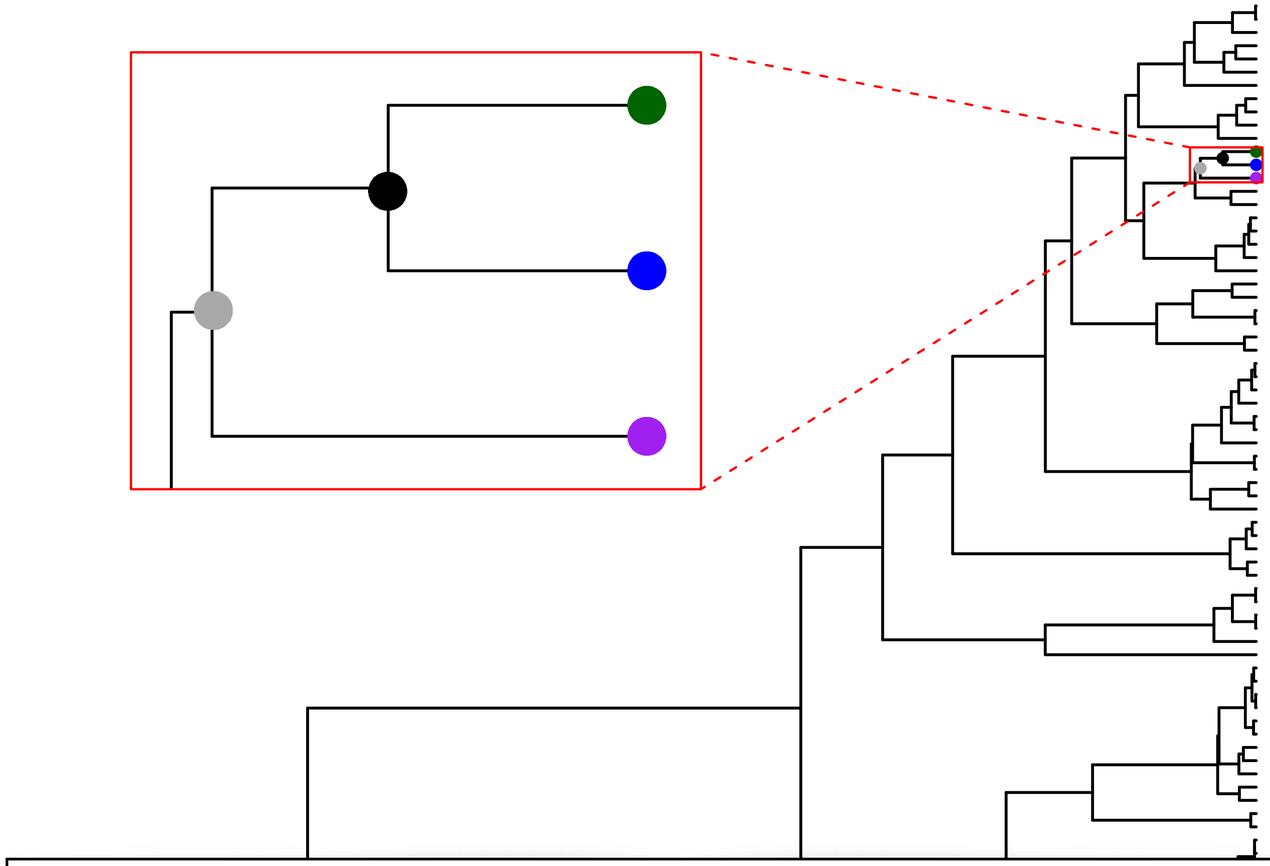


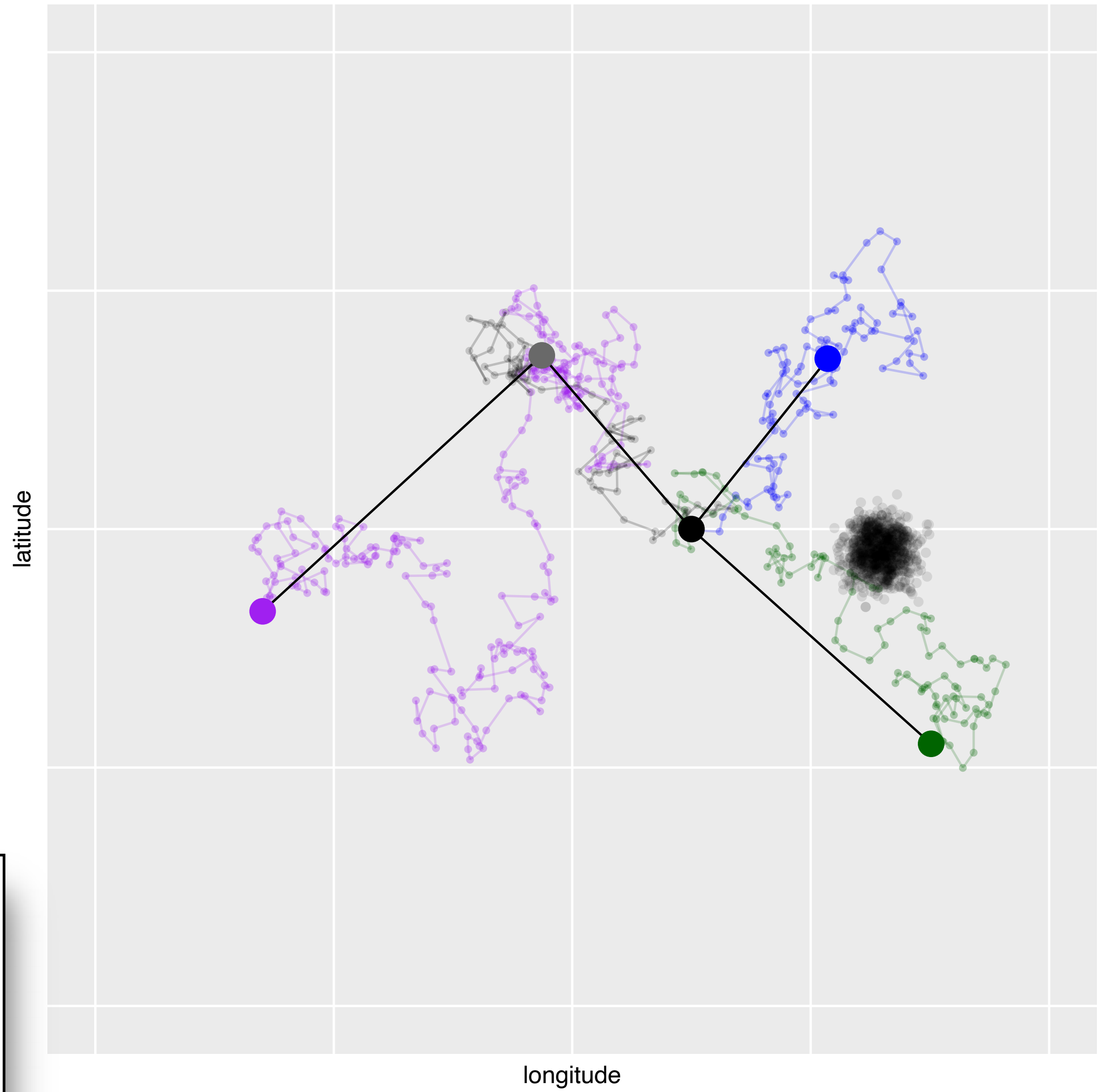


Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro



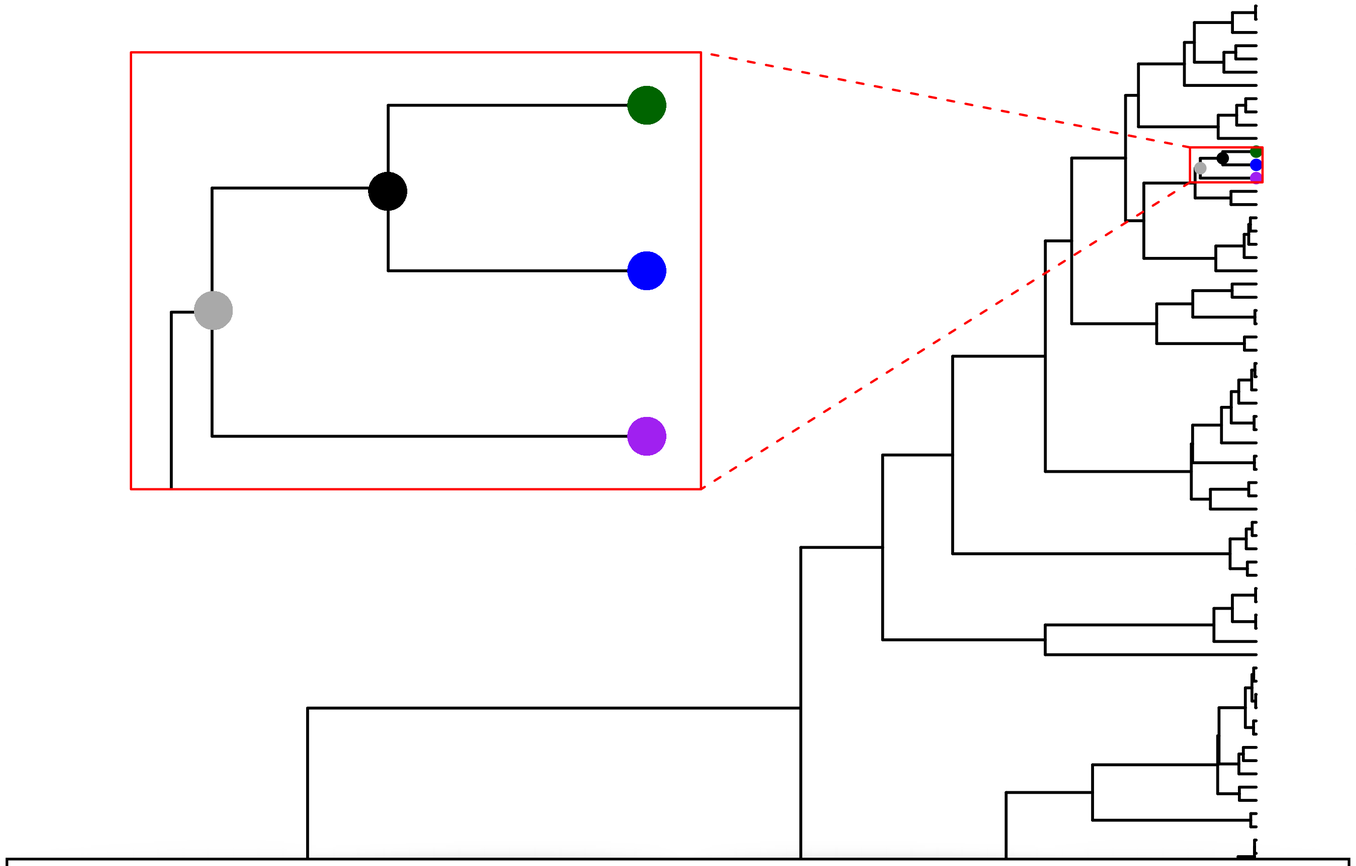


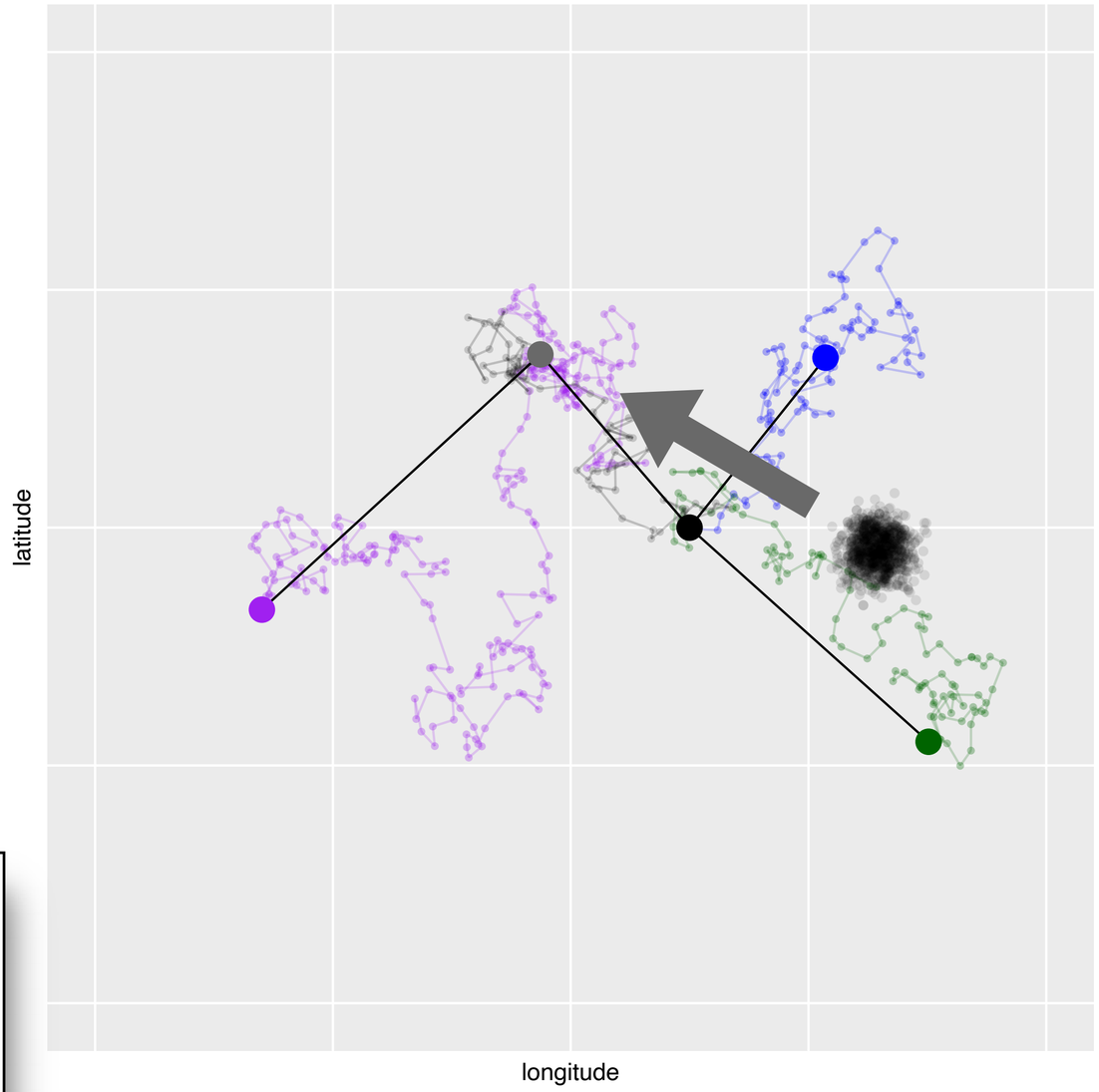
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

... all locations and times serve as constraints on one another.



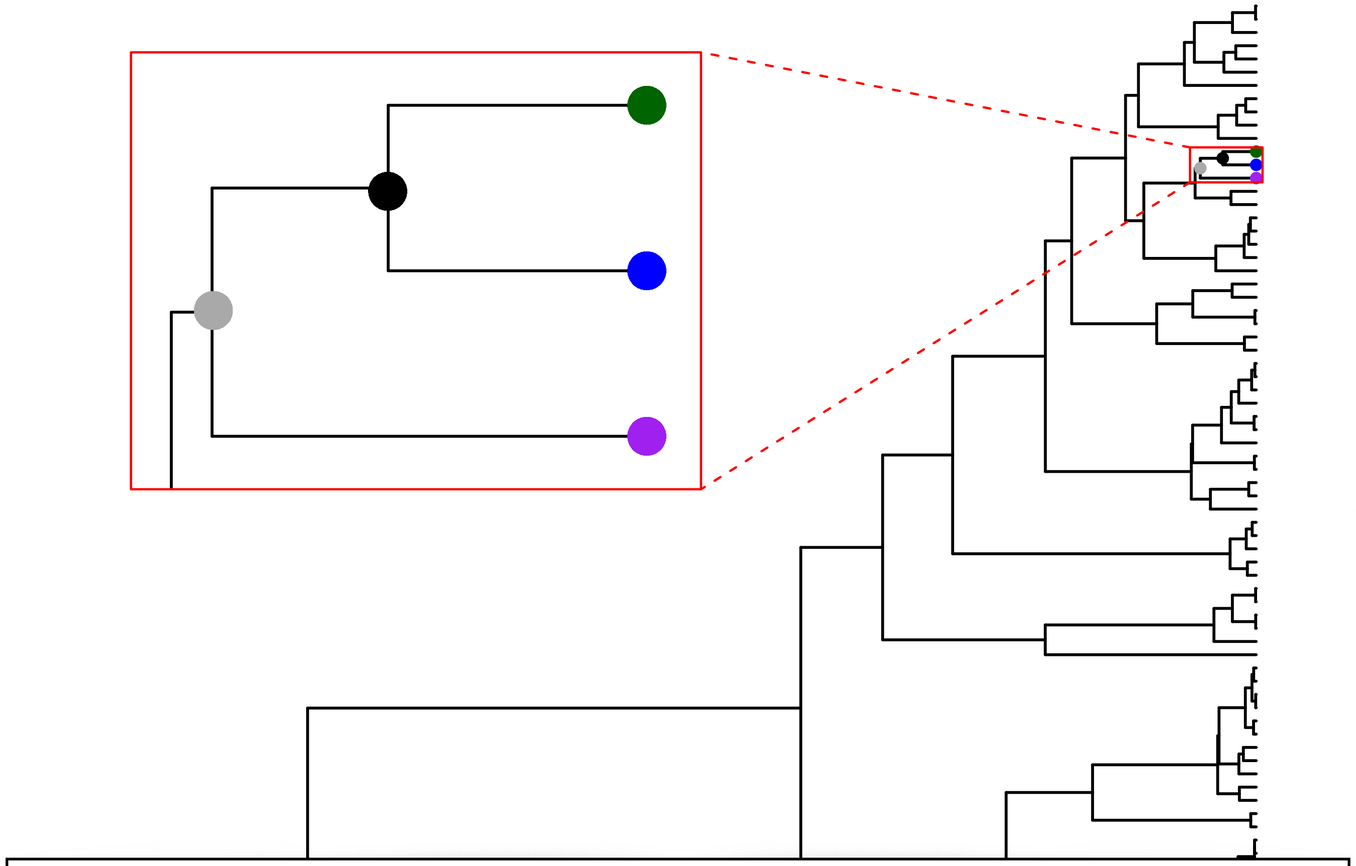


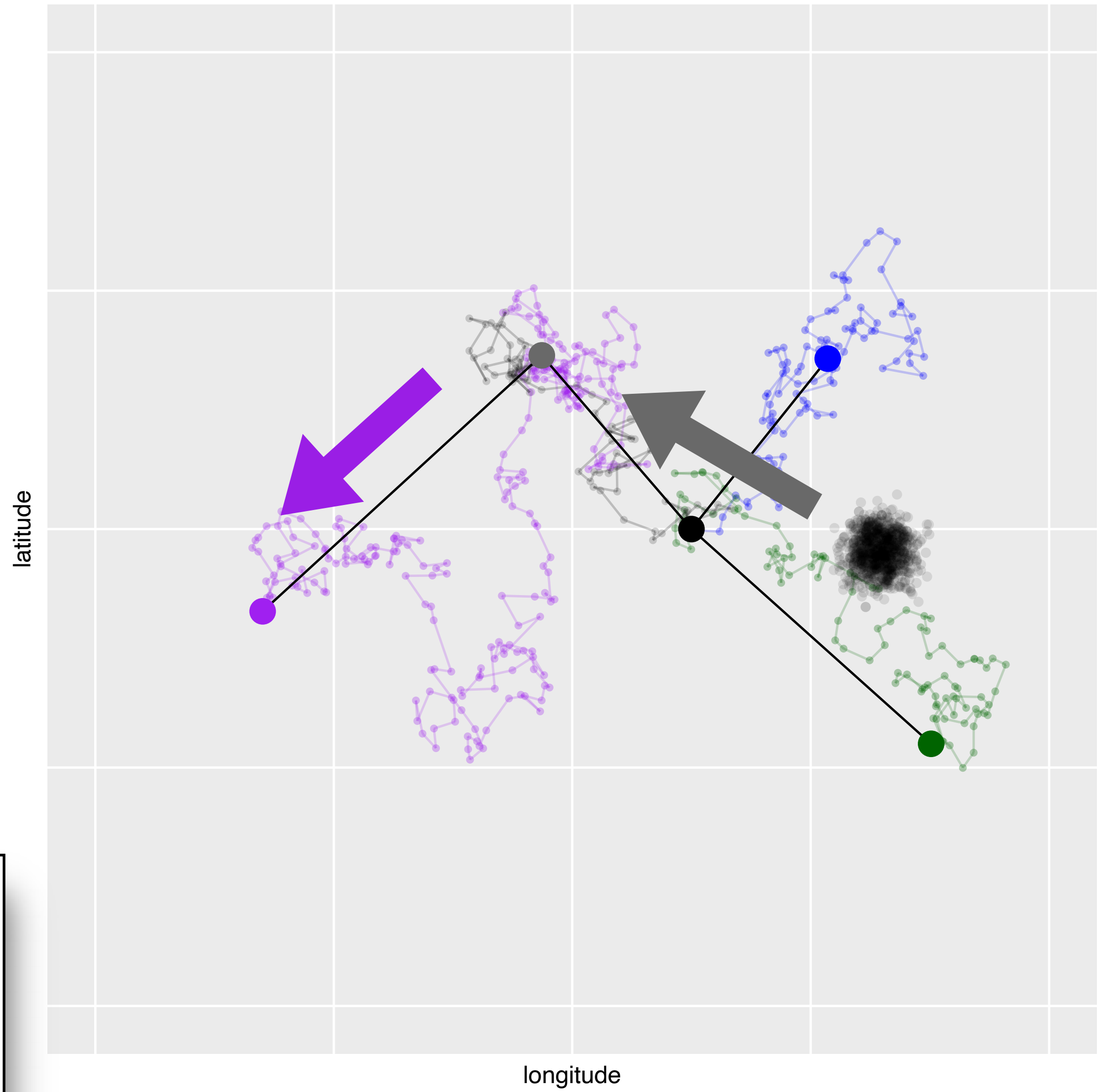
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

... all locations and times serve as constraints on one another.



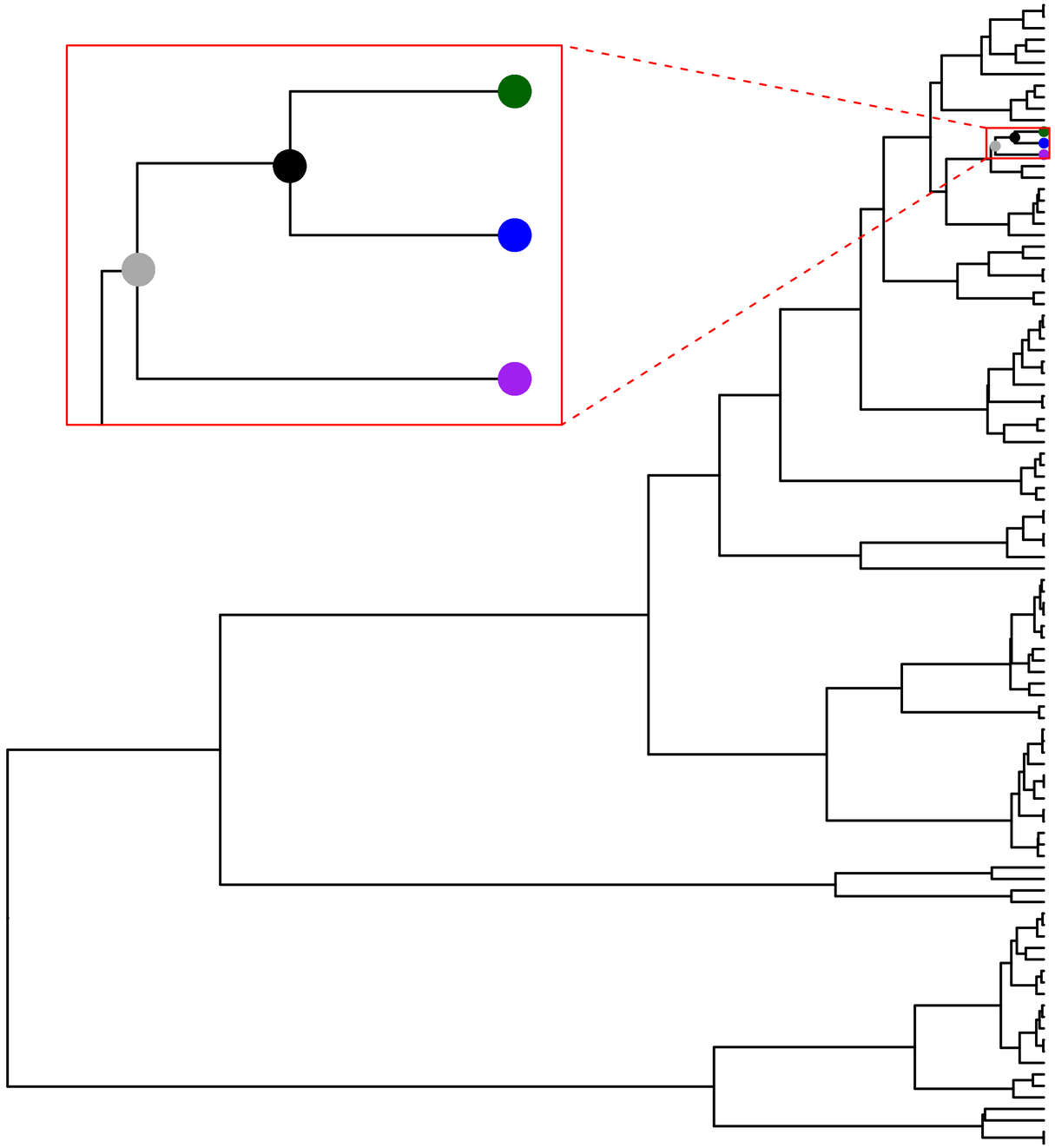


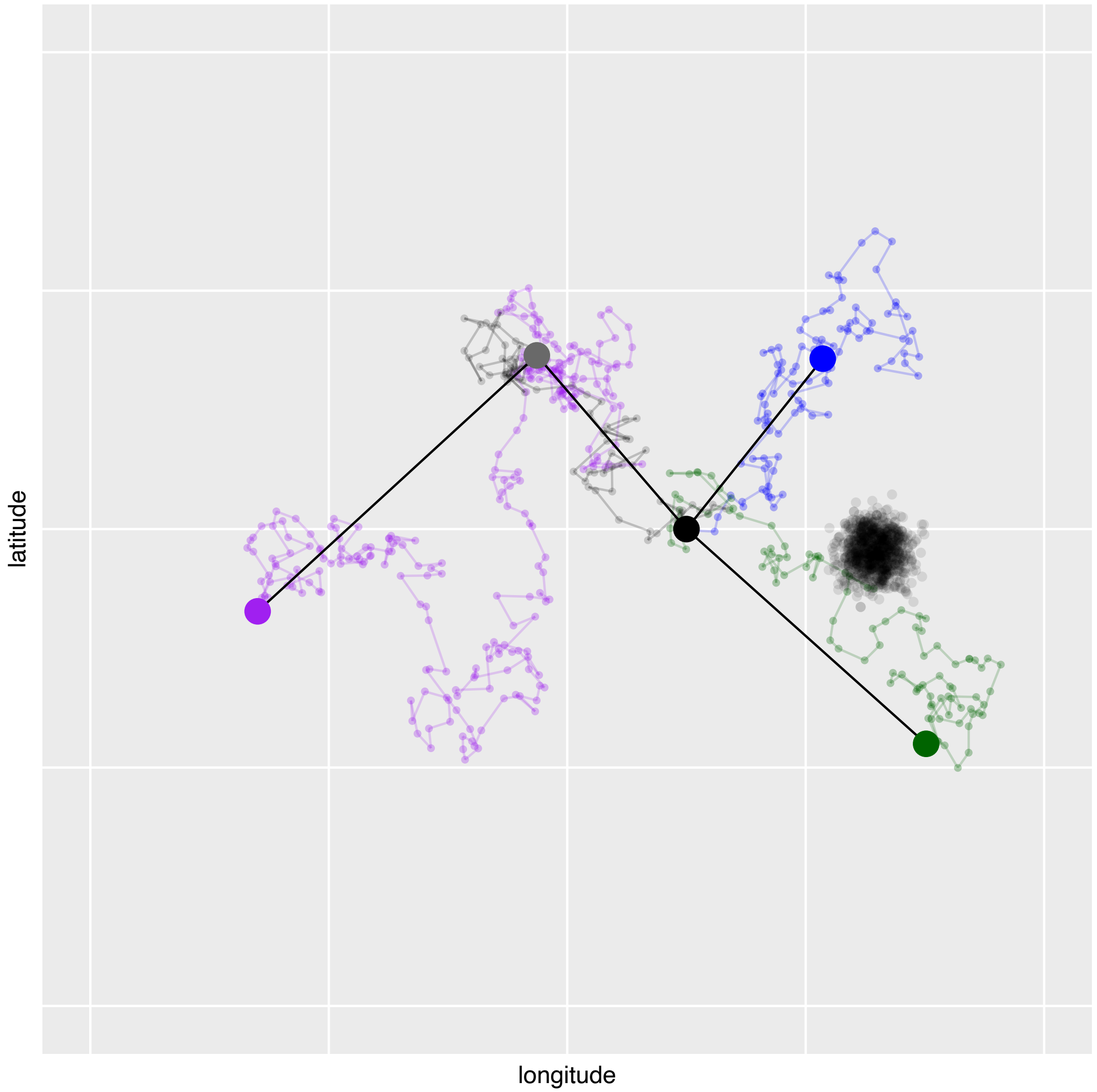
Bayesian model:

"**What is the joint posterior distribution of ancestral positions & ages** given known locations of samples and mutation counts on branches?"

MCMC using NumPyro

... all locations and times serve as constraints on one another.

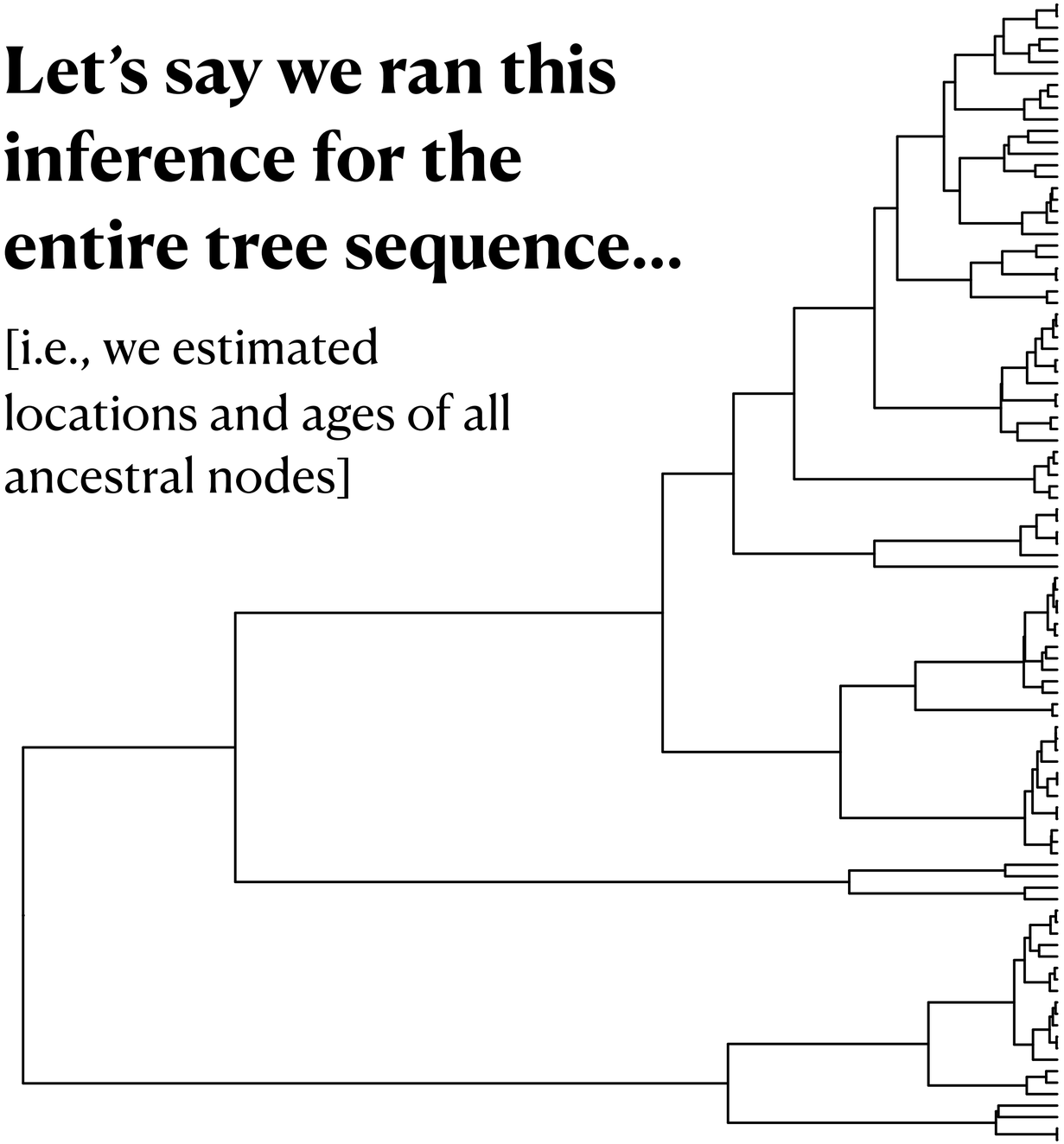




... all locations and times serve as constraints on one another.

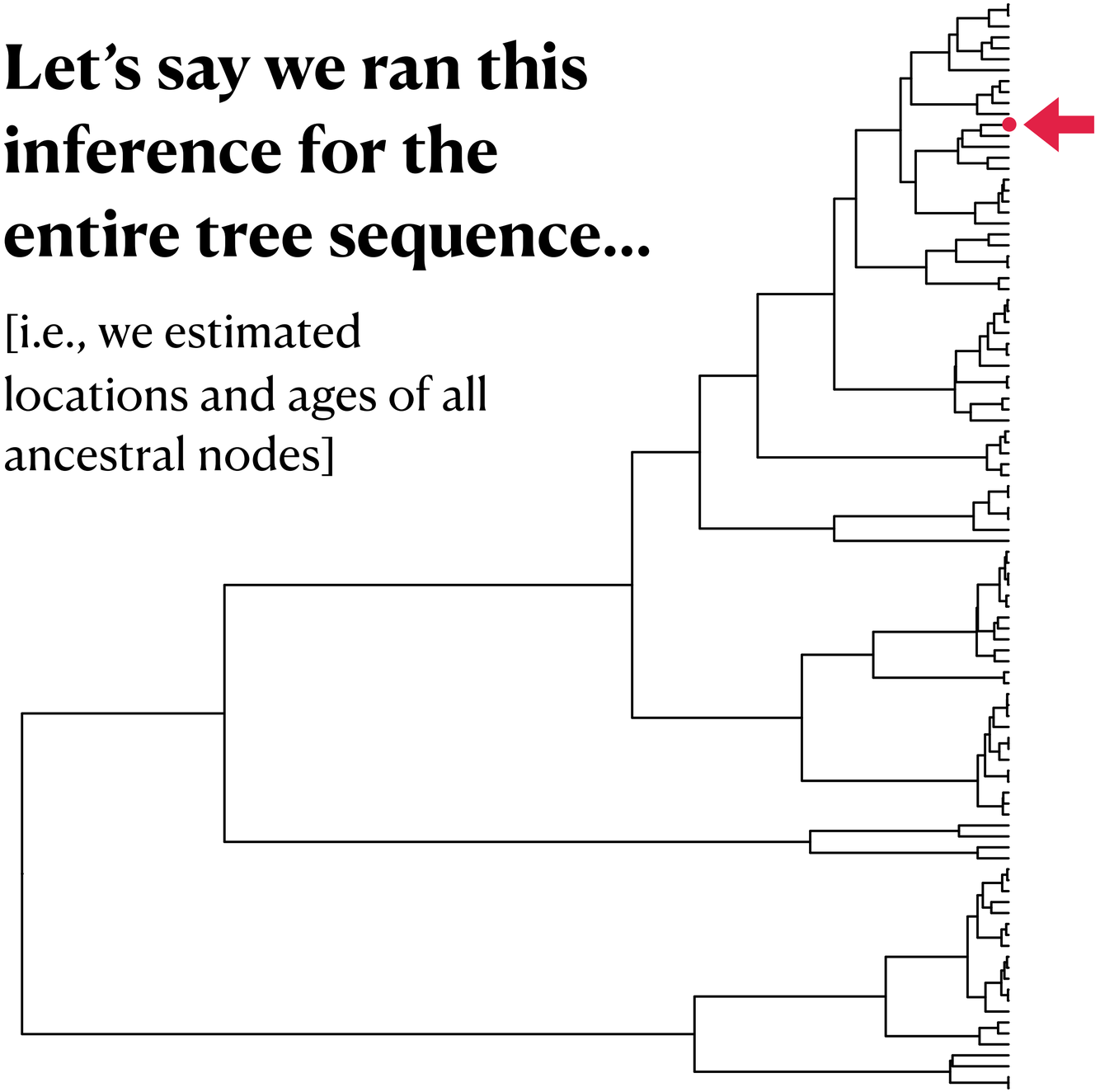
Let’s say we ran this inference for the entire tree sequence...

[i.e., we estimated locations and ages of all ancestral nodes]



Let's say we ran this inference for the entire tree sequence...

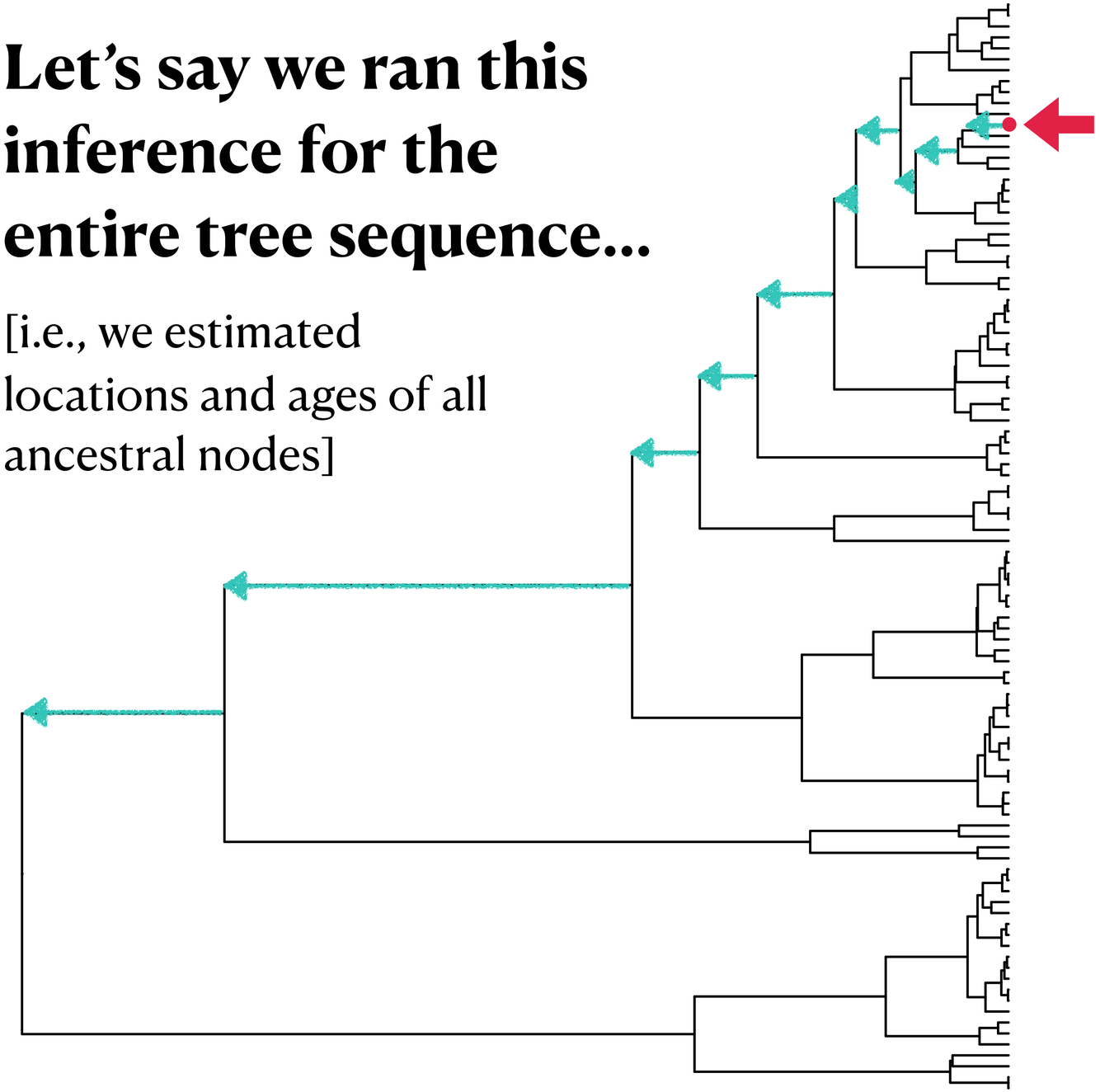
[i.e., we estimated locations and ages of all ancestral nodes]



For a **given individual**

Let's say we ran this inference for the entire tree sequence...

[i.e., we estimated locations and ages of all ancestral nodes]

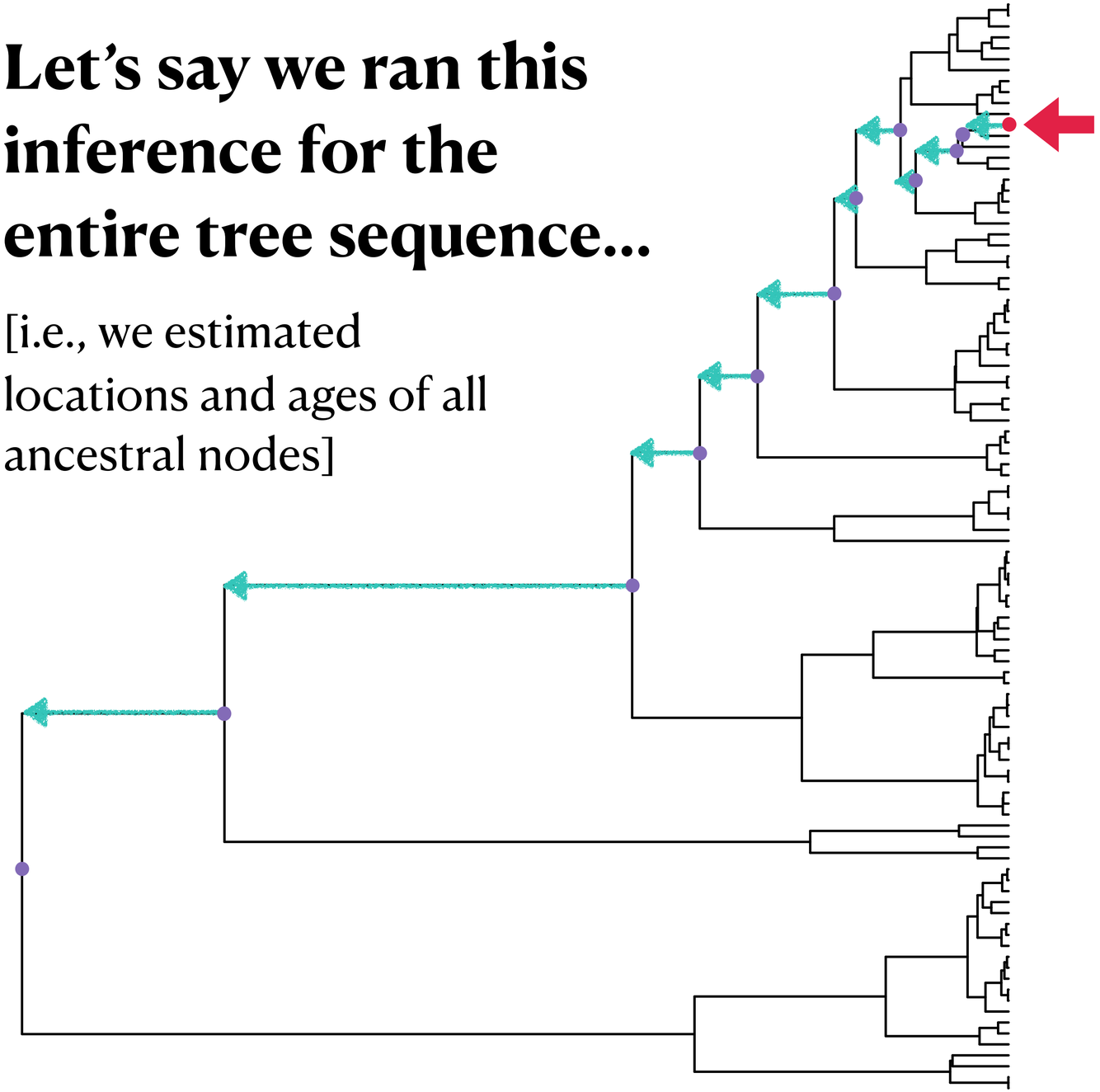


For a **given individual**

we can **"climb up"** each tree

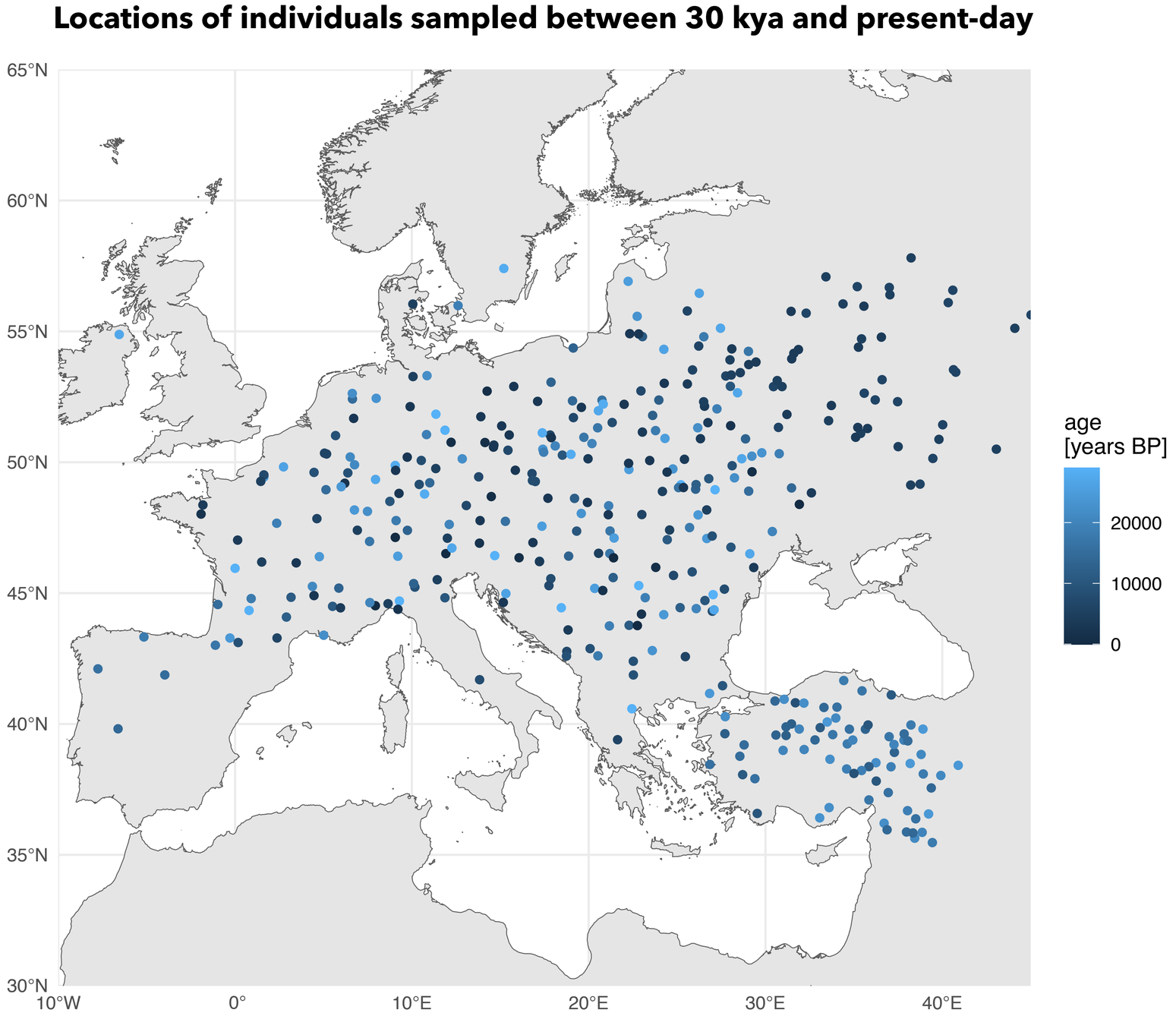
Let’s say we ran this inference for the entire tree sequence...

[i.e., we estimated locations and ages of all ancestral nodes]

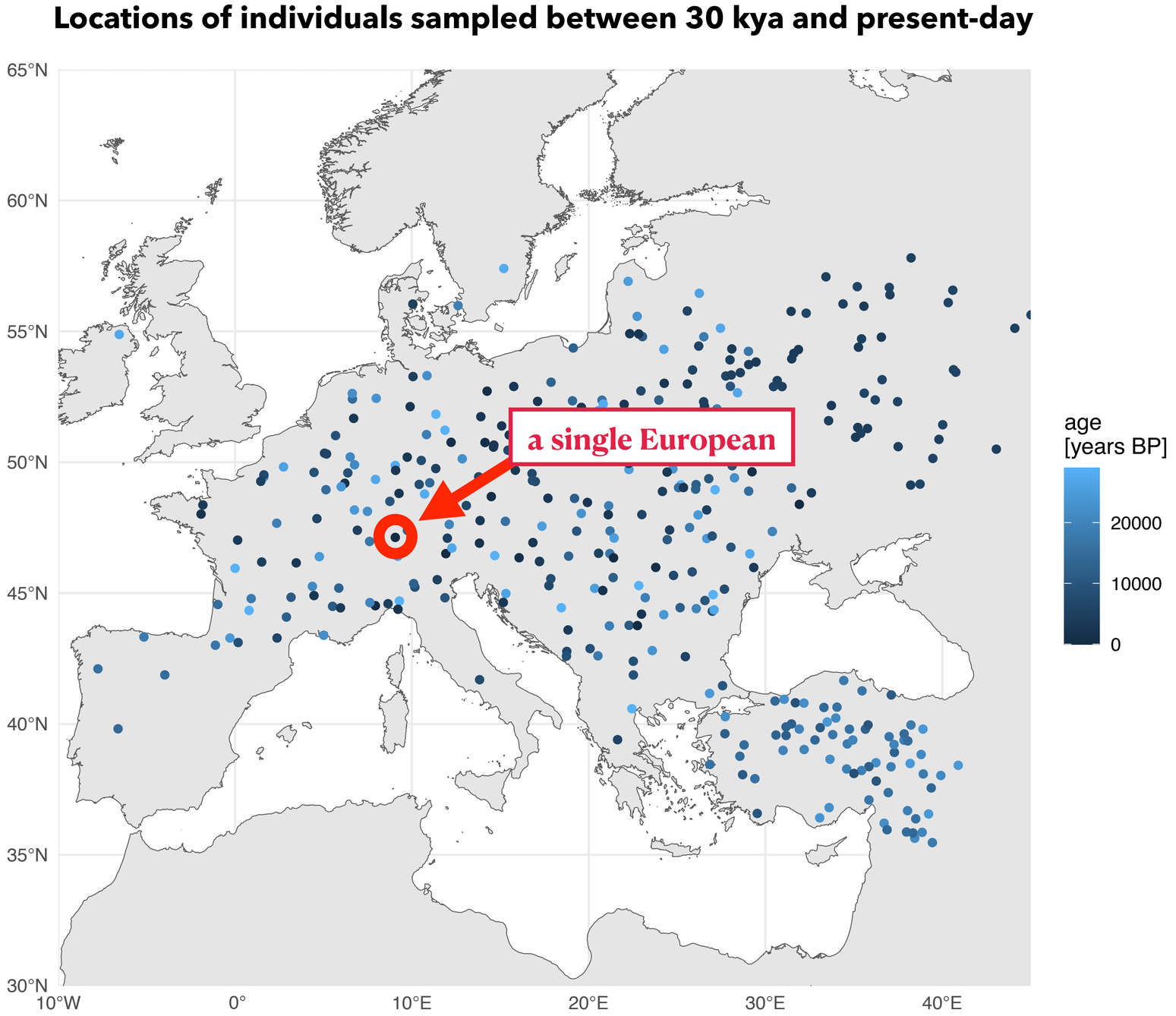


For a **given individual** we can **"climb up"** each tree and collect "paths" through inferred **ancestral locations**.

The Holy Grail?

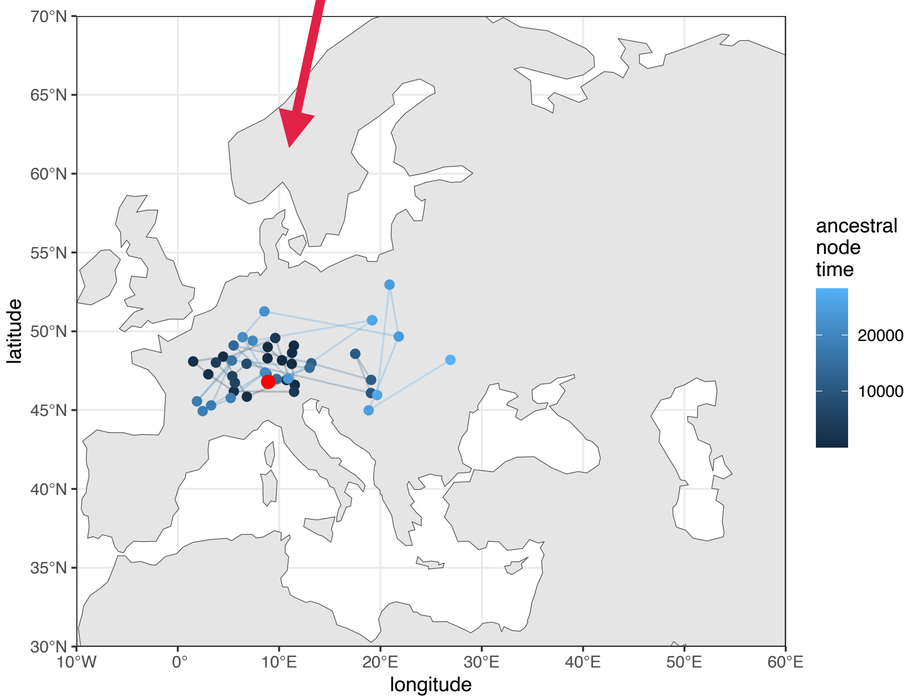


The Holy Grail?

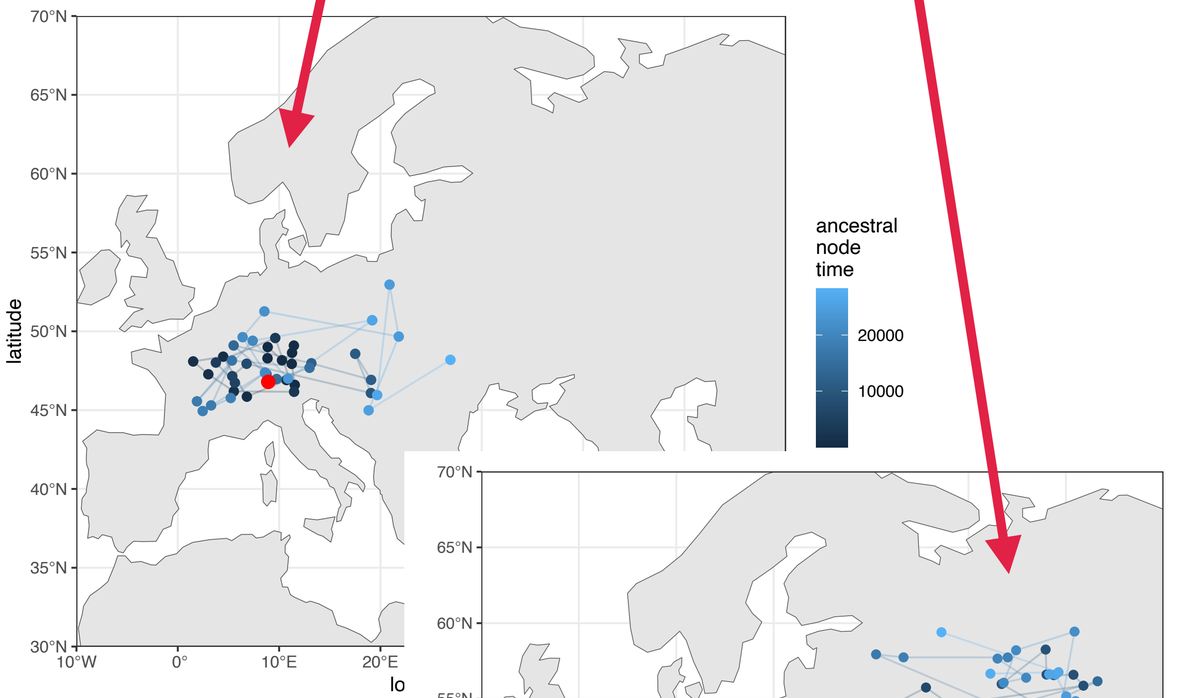


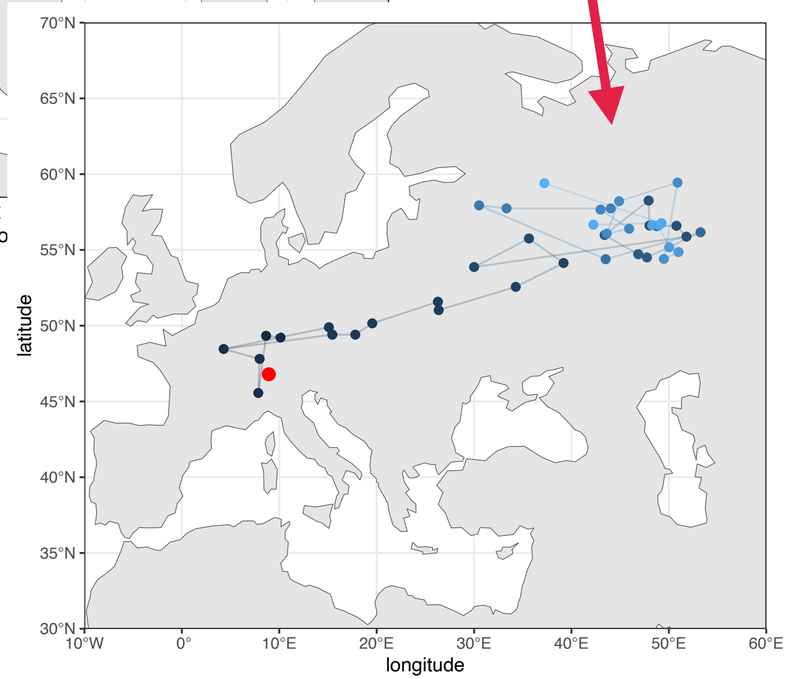
DNA sequence of that individual's genome (just 1 Mb)

DNA sequence of that individual's genome (just 1 Mb)

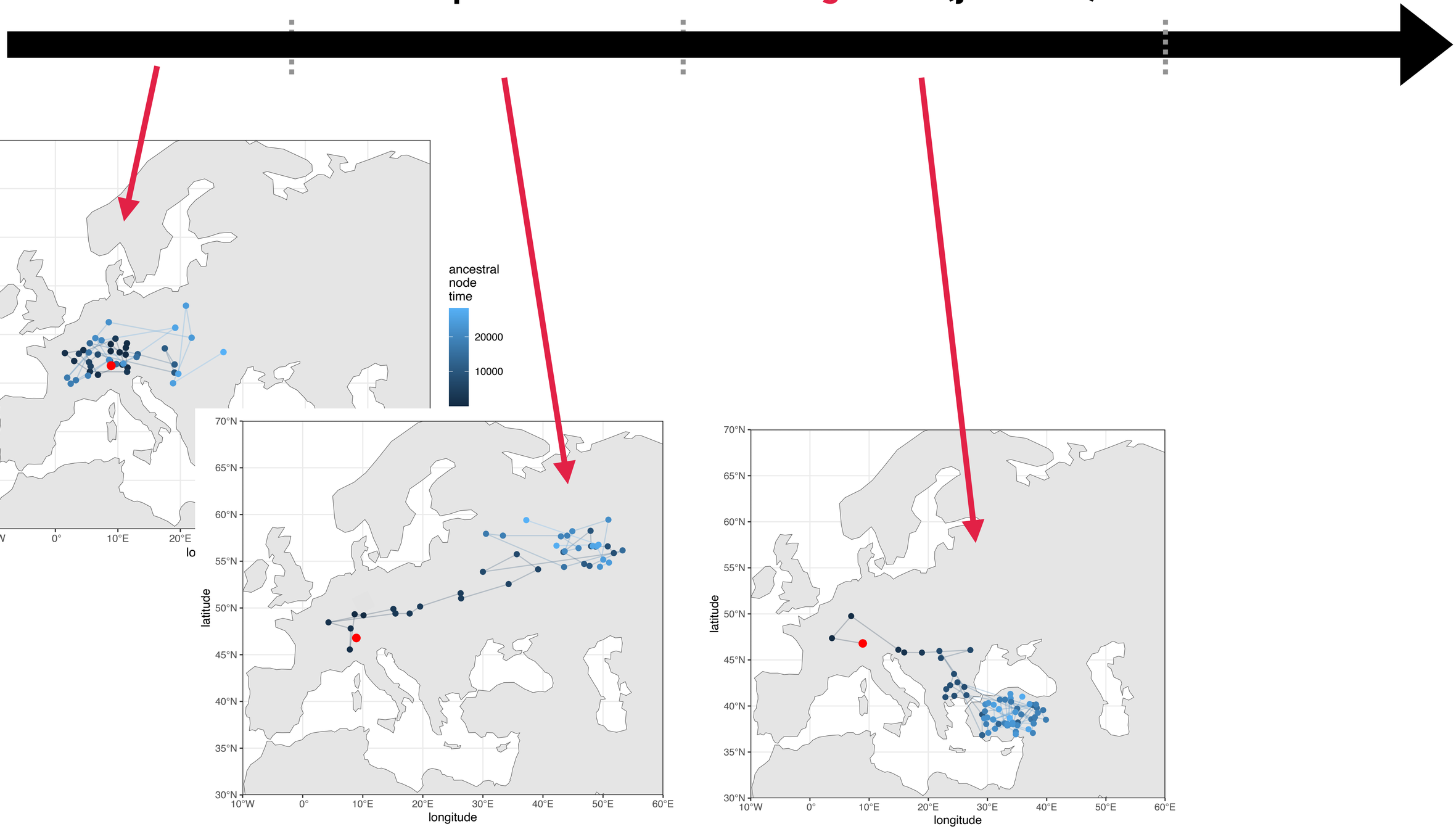


DNA sequence of that individual's genome (just 1 Mb)

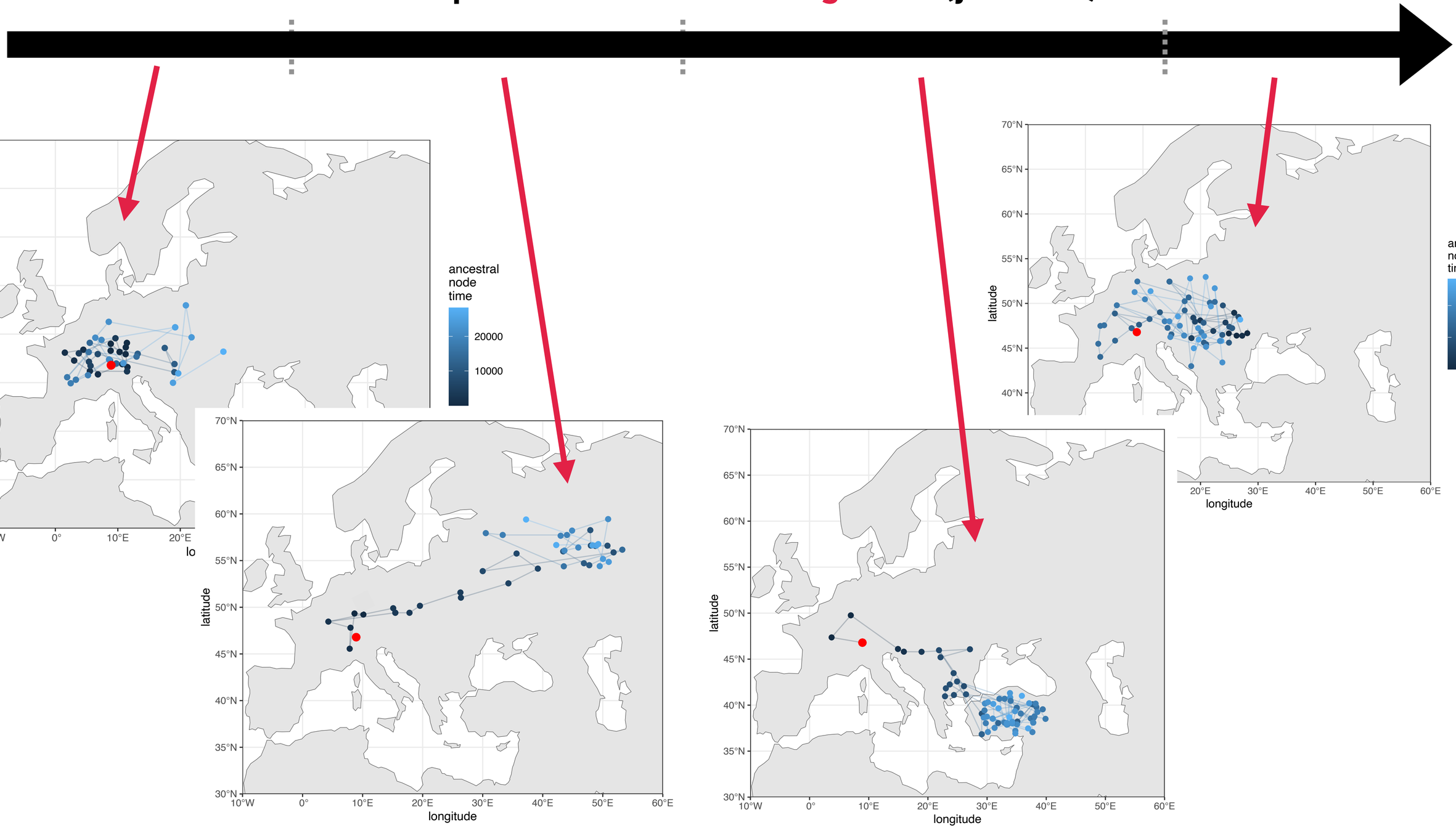




DNA sequence of that individual's genome (just 1 Mb)



DNA sequence of that individual's genome (just 1 Mb)



Summary

Summary

1. We have reached sufficient density of aDNA sampling to facilitate **spatiotemporal analyses** of our evolutionary history.

Summary

1. We have reached sufficient density of aDNA sampling to facilitate **spatiotemporal analyses** of our evolutionary history.

2. **Tree sequence** is a true game-changer in population genetic simulation and facilitates novel approaches for analysis of large-scale data sets.

Summary

1. We have reached sufficient density of aDNA sampling to facilitate **spatiotemporal analyses** of our evolutionary history.

2. **Tree sequence** is a true game-changer in population genetic simulation and facilitates novel approaches for analysis of large-scale data sets.

3. We are close to entirely **new kinds of geographically-explicit studies** of the history of genetic ancestry across space and time.

Thanks!

Fernando Racimo

& the whole Racimo group

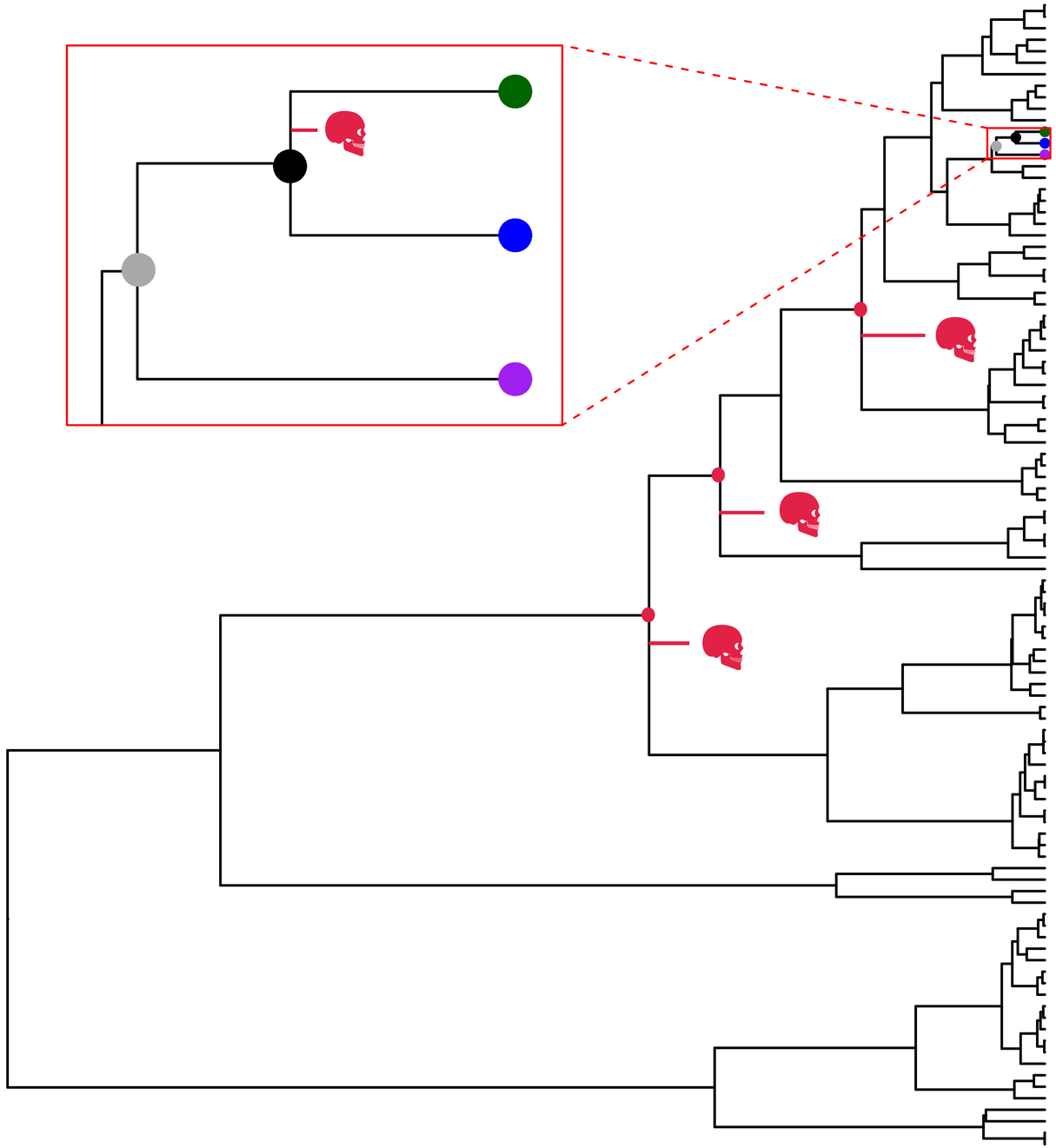
Ben Haller, Peter L. Ralph

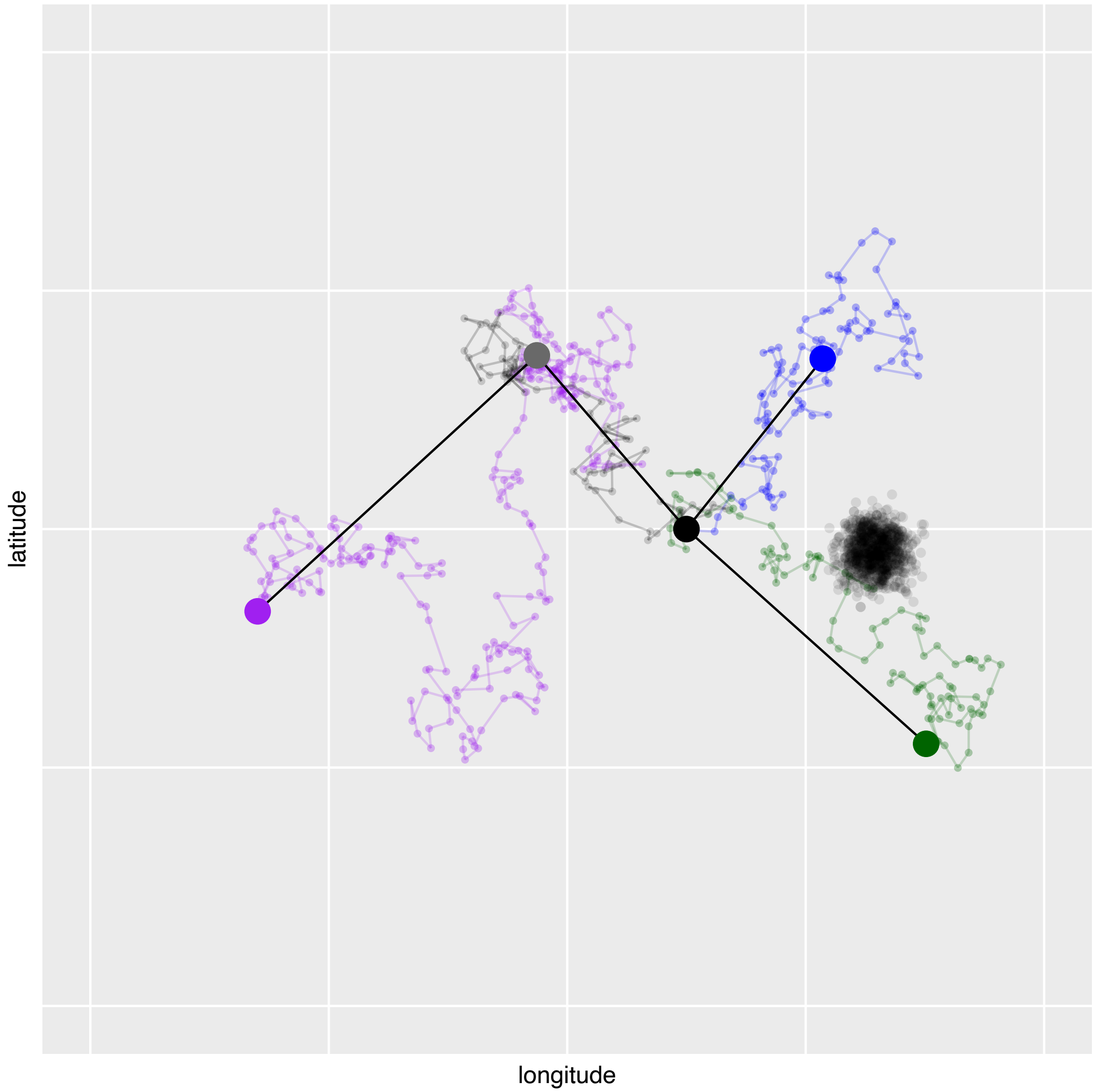
UNIVERSITY OF COPENHAGEN

LUNDBECKFONDEN

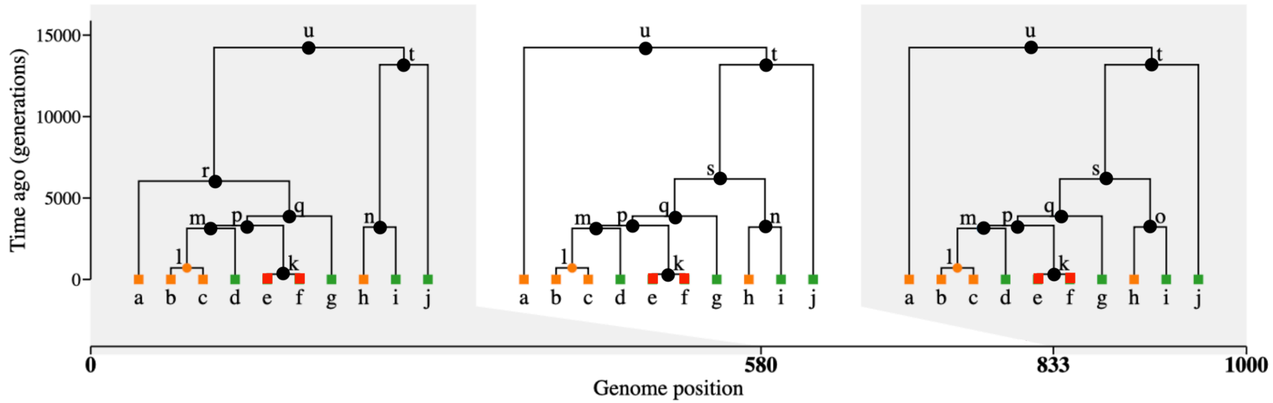
novo nordisk fonden

email: mp@bodkan.net




latitude
longitude

Genealogical nearest neighbours (GNN)



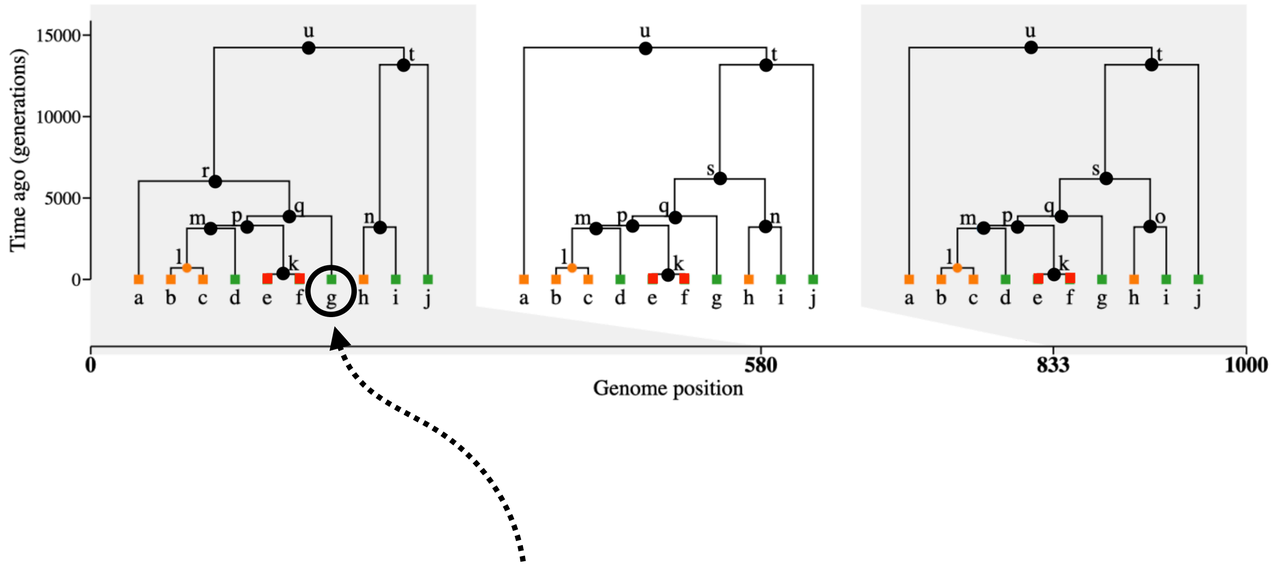
population 1

population 1

population 1

Kelleher *et al.*, Nat Genetics 2019

Genealogical nearest neighbours (GNN)

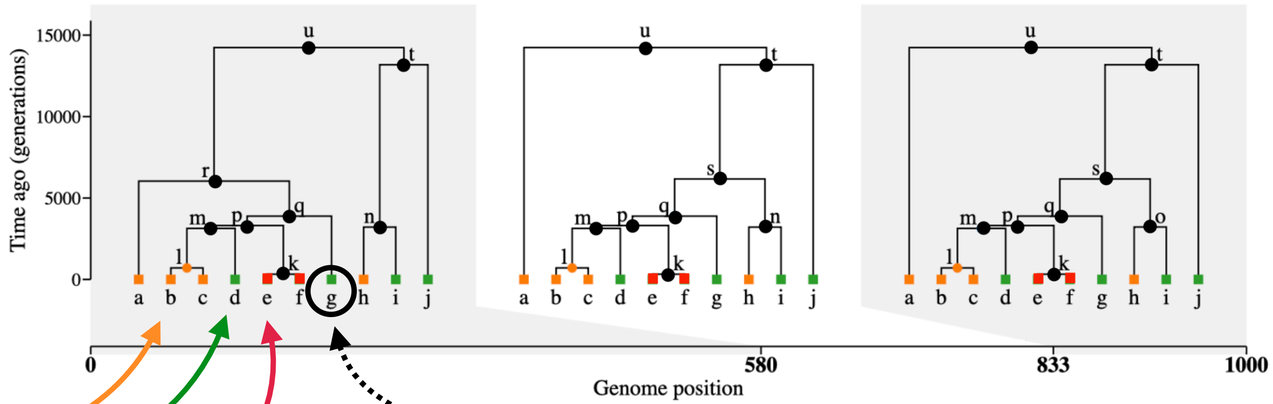


population 1
population 1
population 1

"What proportion of a given chromosome from one population has chromosomes from all sampled populations as immediate siblings?

Kelleher *et al.*, Nat Genetics 2019

Genealogical nearest neighbours (GNN)

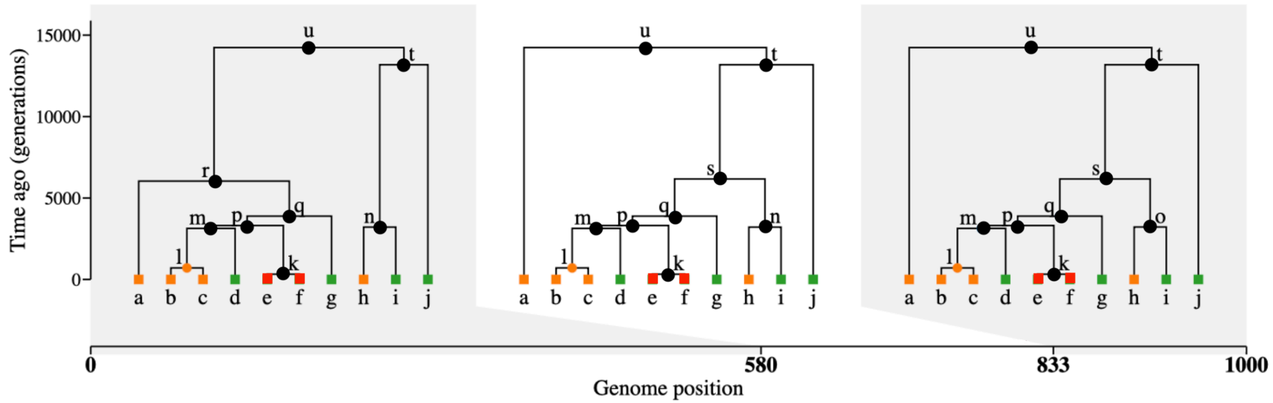


population 1
population 1
population 1

"What proportion of a given chromosome from one population has chromosomes from all sampled populations as immediate siblings?

Kelleher *et al.*, Nat Genetics 2019

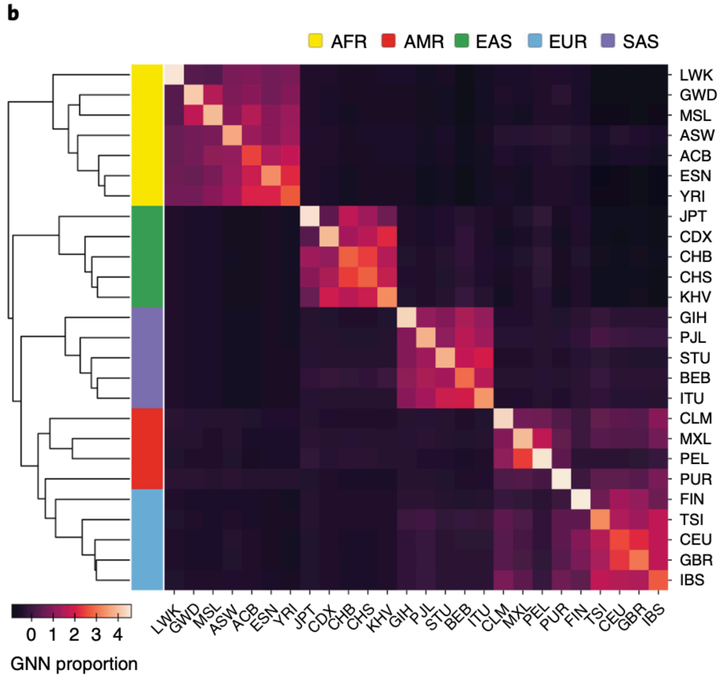
Genealogical nearest neighbours (GNN)



population 1

population 1

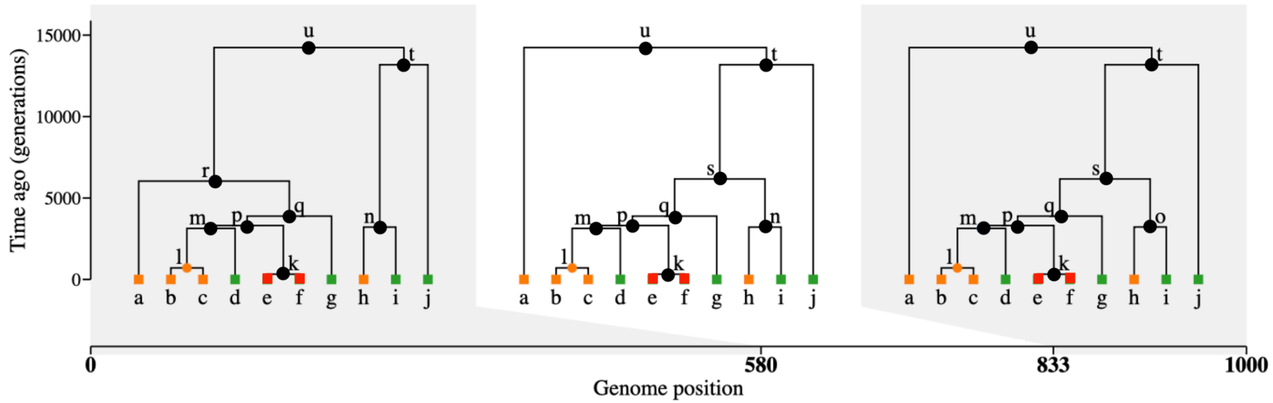
population 1



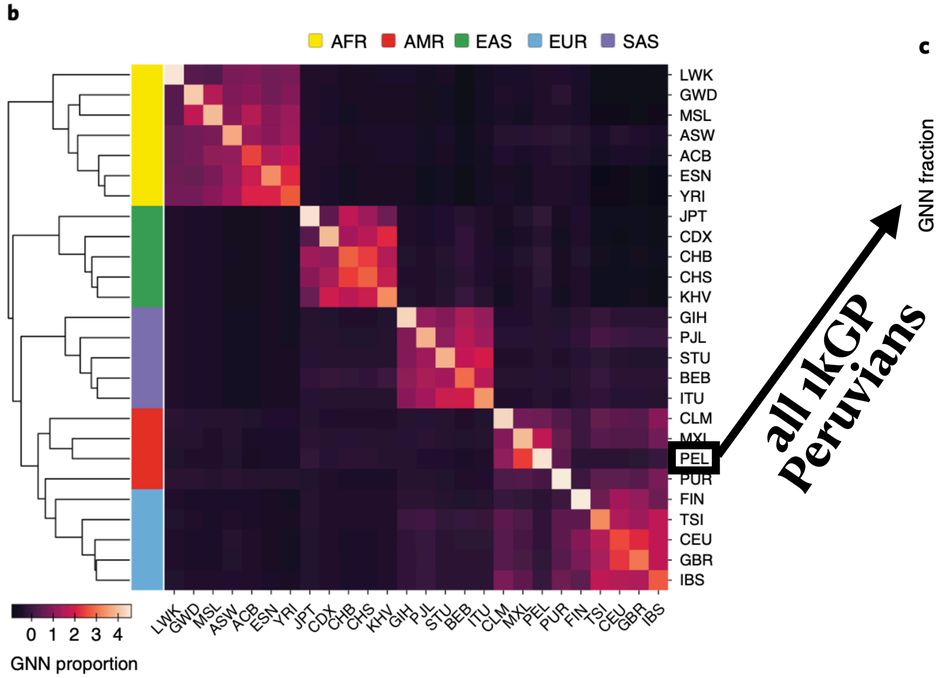
GNN clustering for all 2504 1000 Genomes Project individuals

Kelleher *et al.*, Nat Genetics 2019

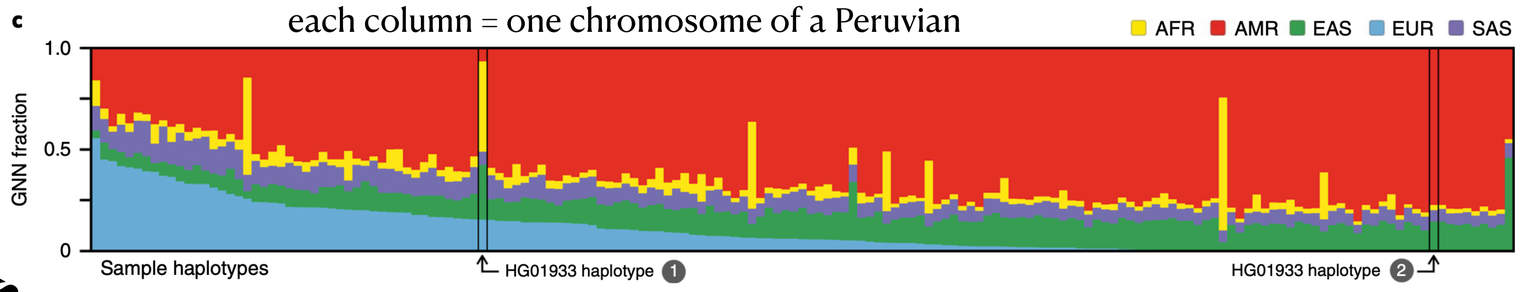
Genealogical nearest neighbours (GNN)



population 1
population 1
population 1



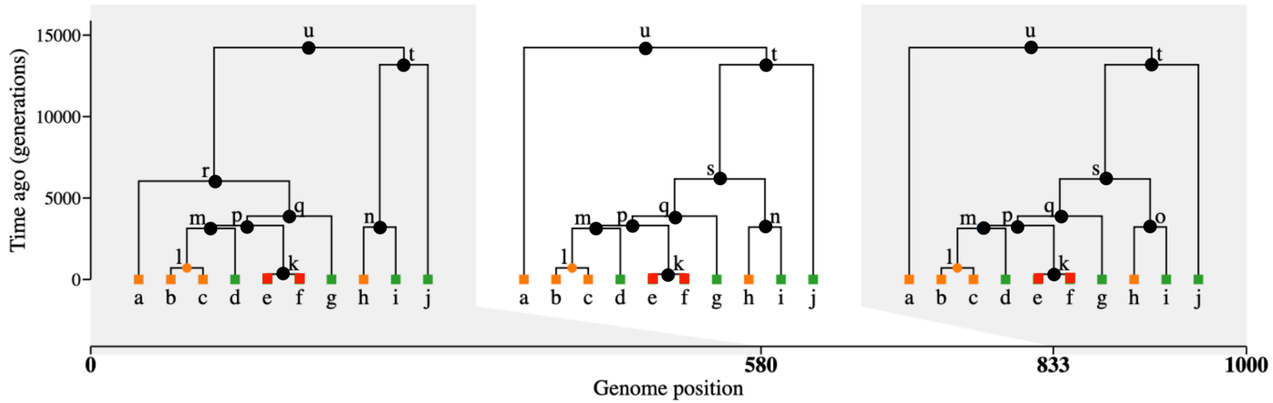
all 1kGP Peruvians



GNN clustering for all 2504 1000 Genomes Project individuals

Kelleher *et al.*, Nat Genetics 2019

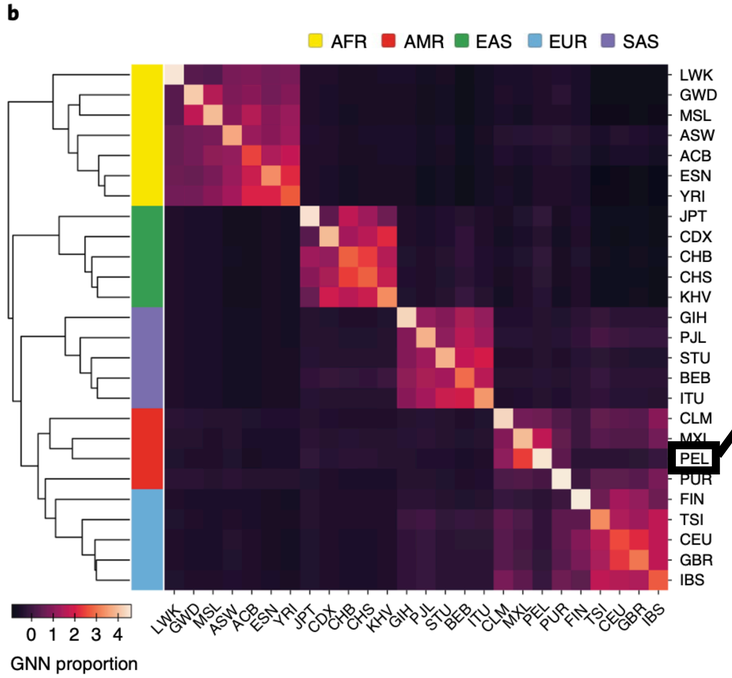
Genealogical nearest neighbours (GNN)



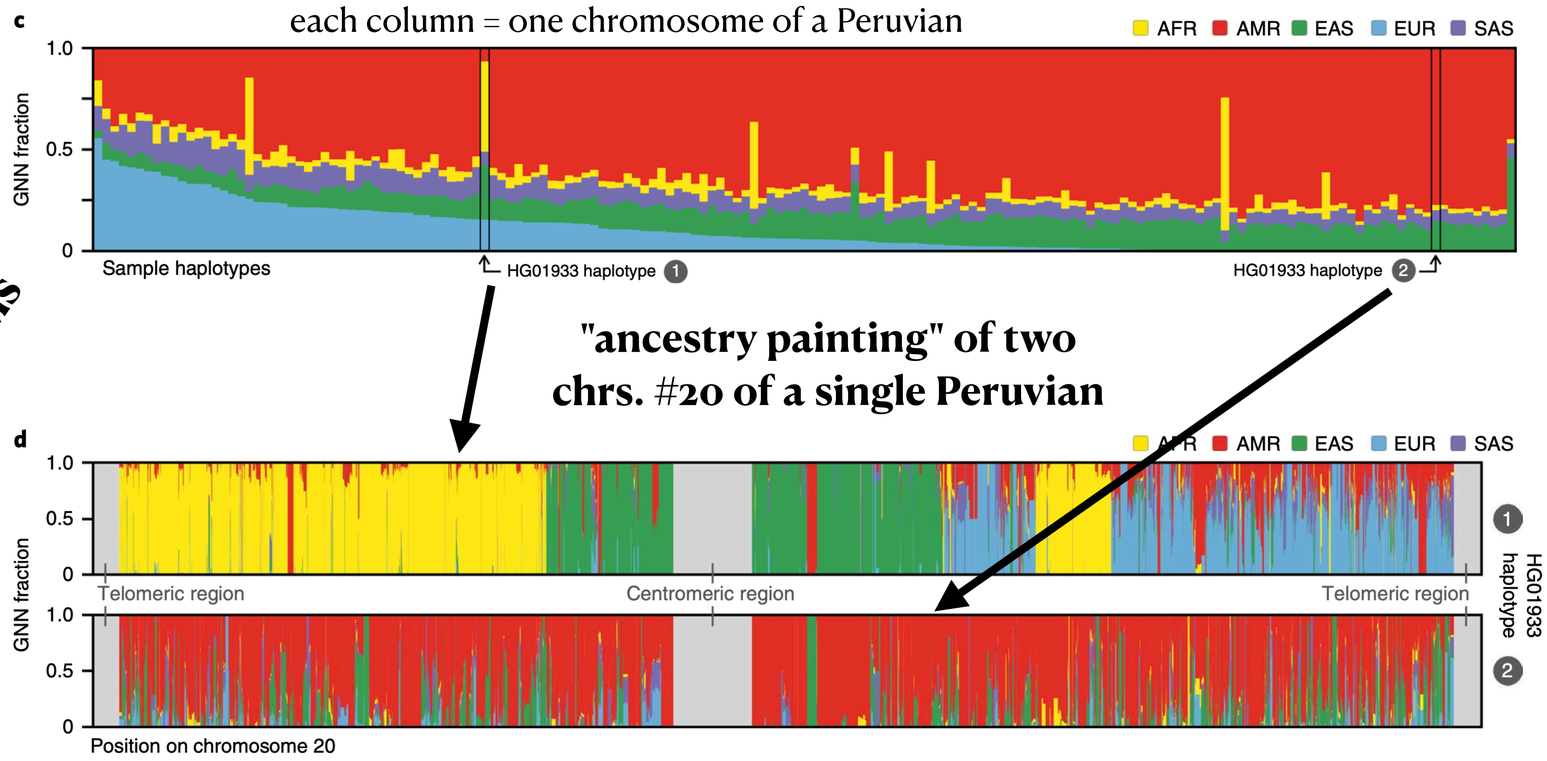
population 1

population 1

population 1



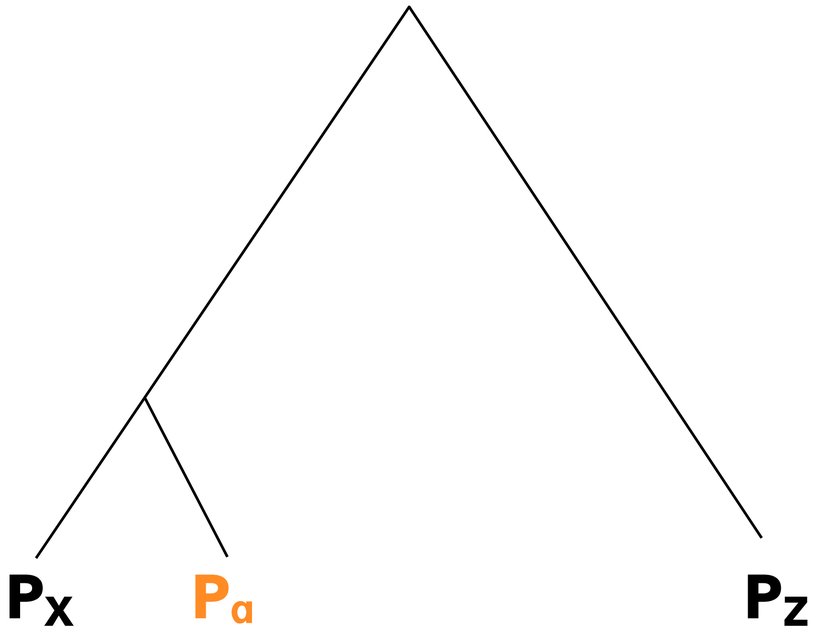
all 1kGP Peruvians



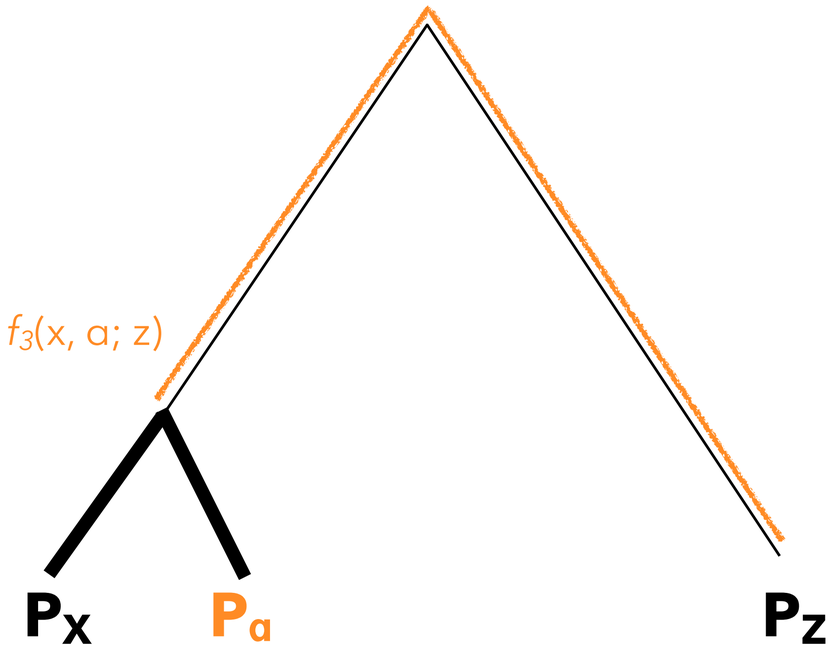
GNN clustering for all 2504 1000 Genomes Project individuals

Kelleher *et al.*, Nat Genetics 2019

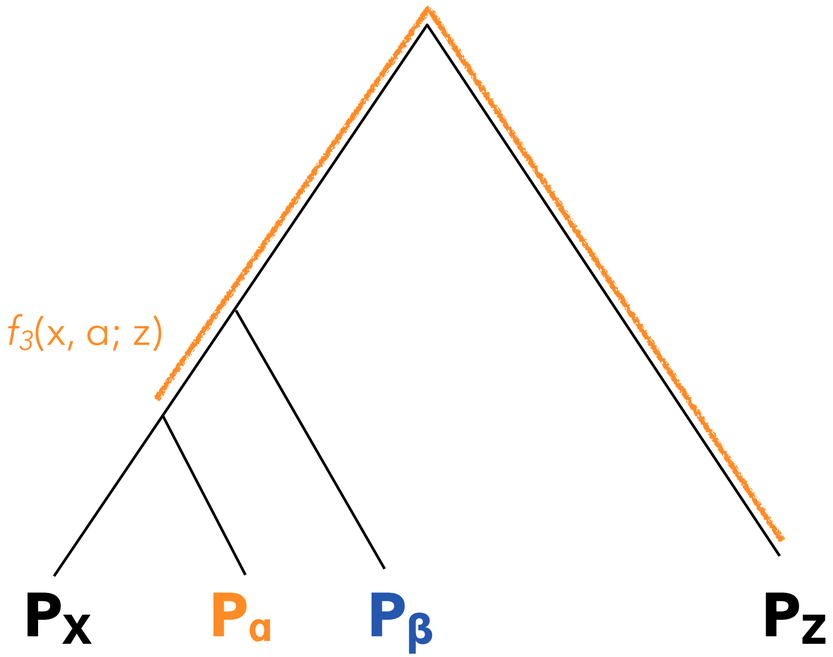
How do traditional popgen statistics work? — *f*-statistics



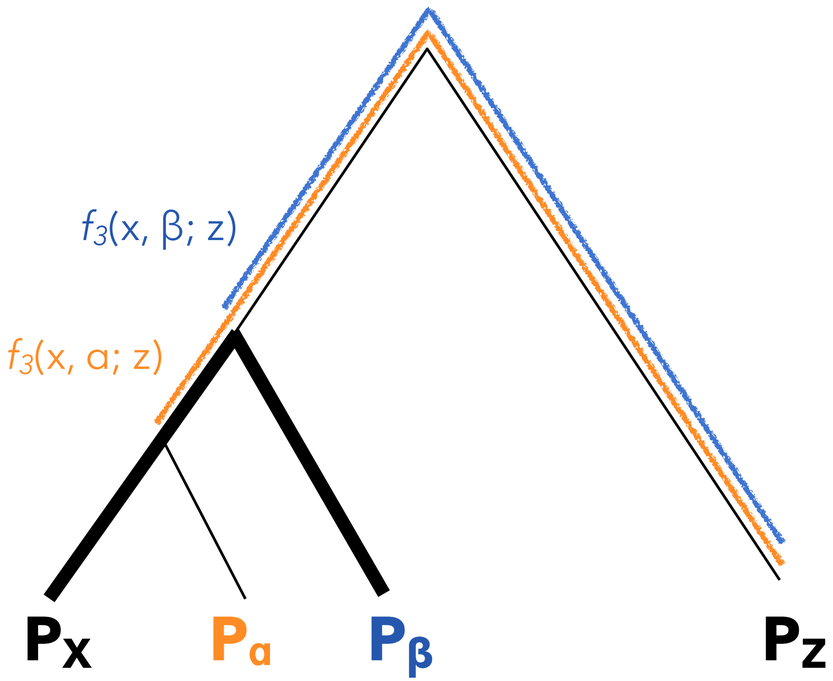
How do traditional popgen statistics work? — *f*-statistics



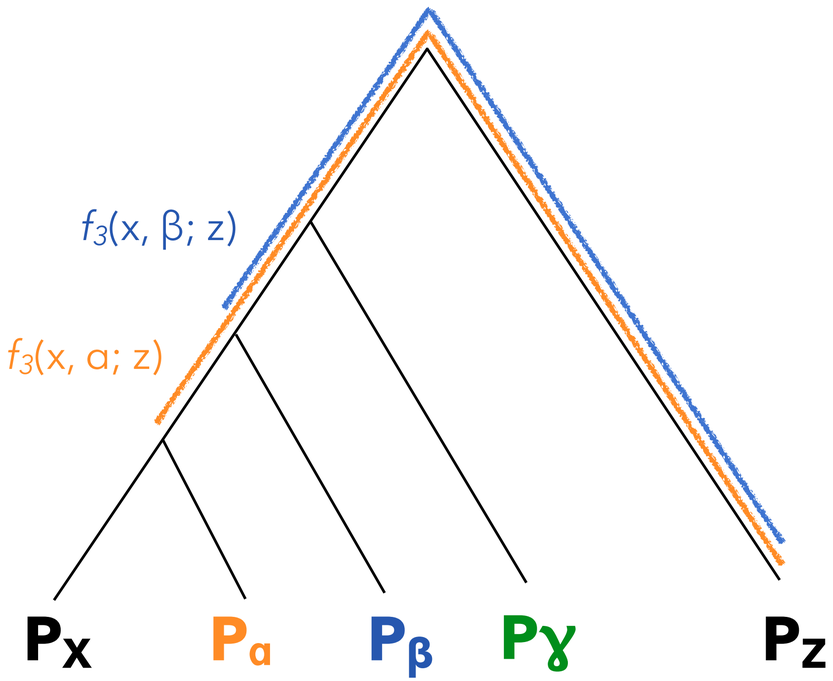
How do traditional popgen statistics work? — *f*-statistics



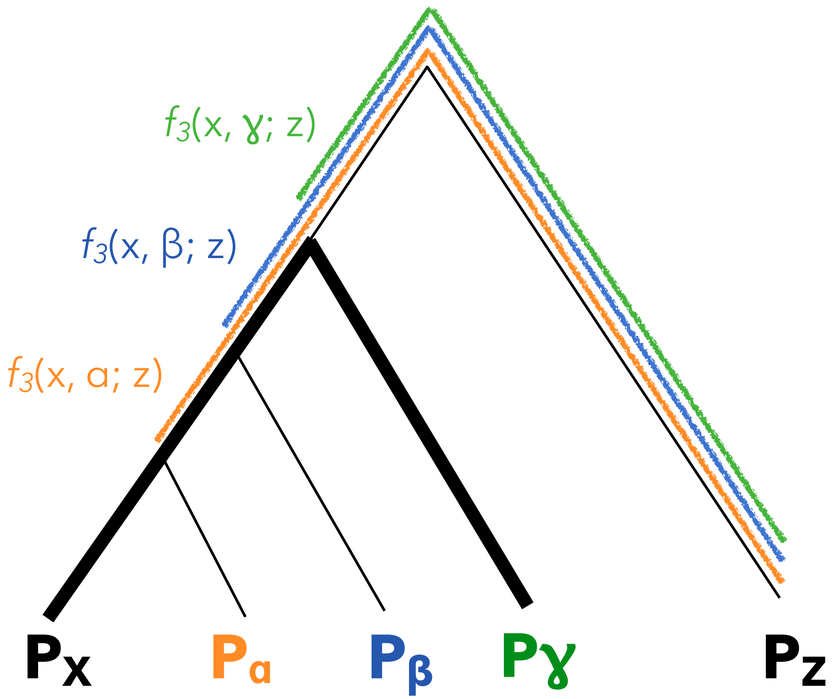
How do traditional popgen statistics work? — *f*-statistics



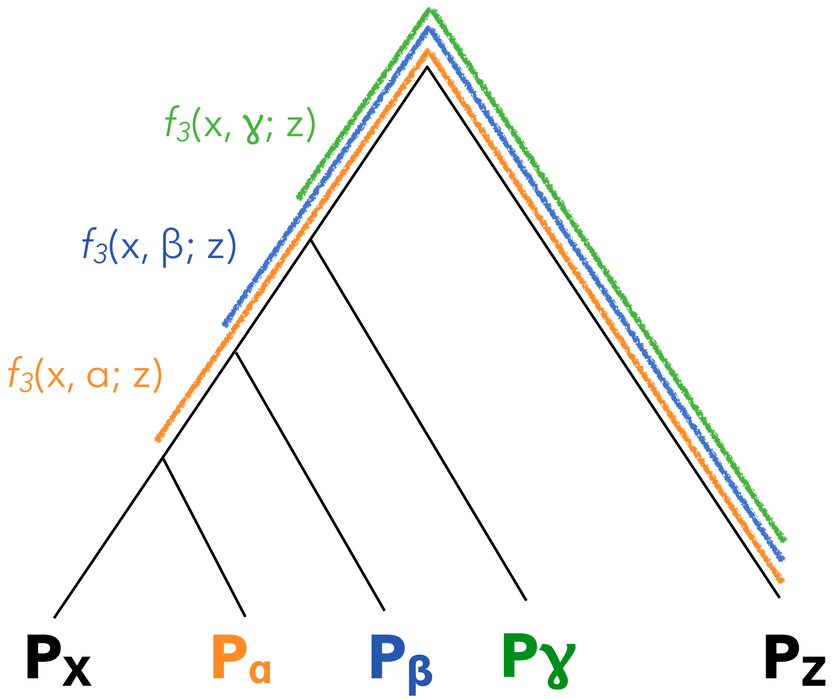
How do traditional popgen statistics work? — *f*-statistics



How do traditional popgen statistics work? — *f*-statistics

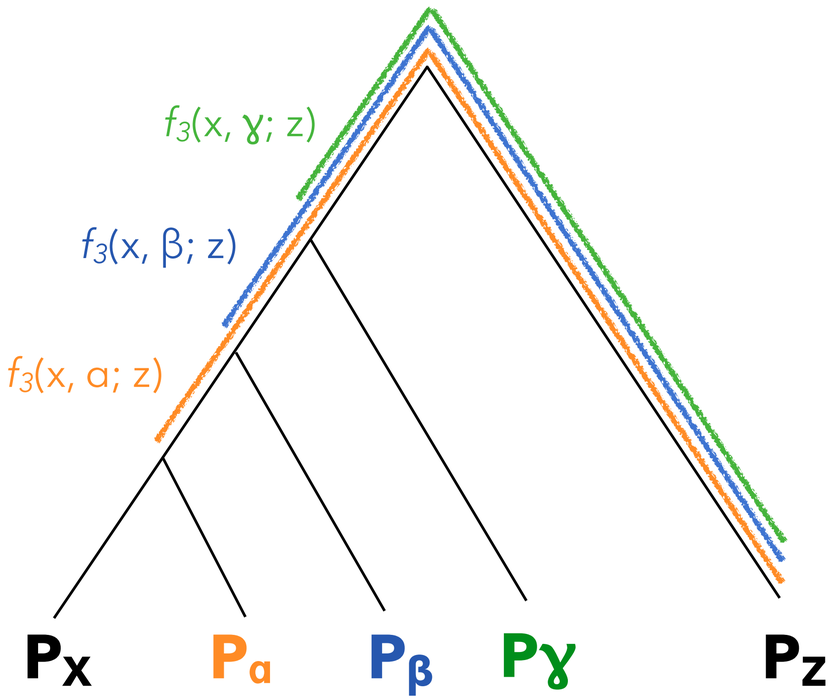


How do traditional popgen statistics work? — *f*-statistics



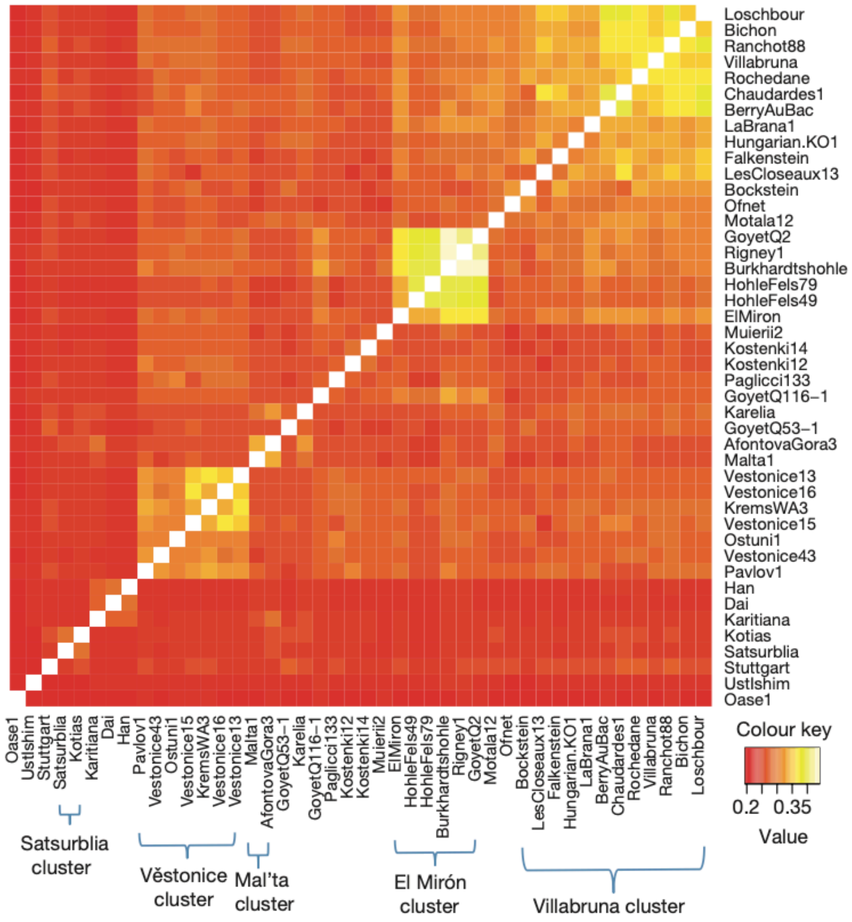
f_3(x, α; z) > f_3(x, β; z) > f_3(x, ɣ; z)

How do traditional popgen statistics work? — *f*-statistics



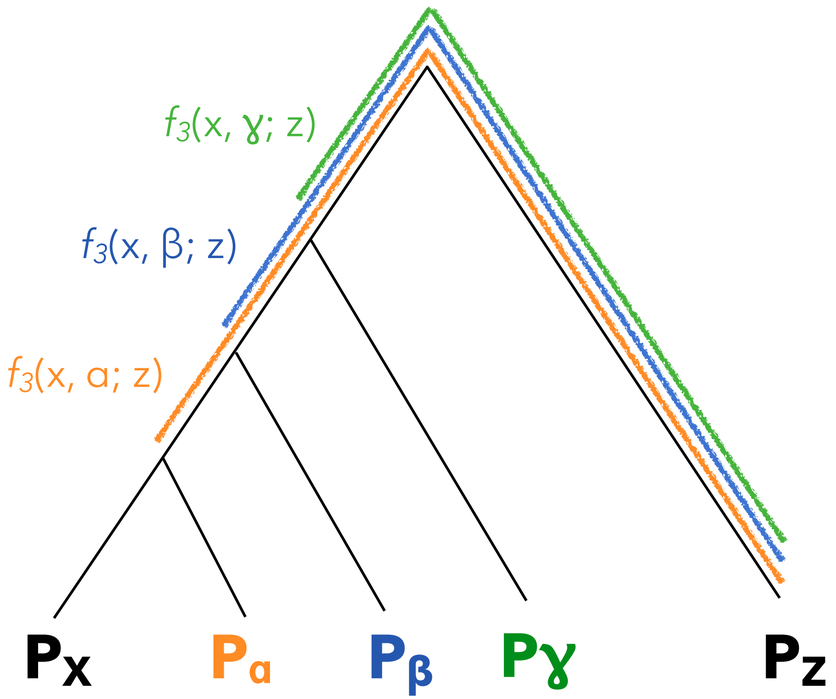
f_3(x, α; z) > f_3(x, β; z) > f_3(x, ɣ; z)

pairwise f_3 heatmap matrix



Fu *et al.*, Nature 2016

How do traditional popgen statistics work? — *f*-statistics

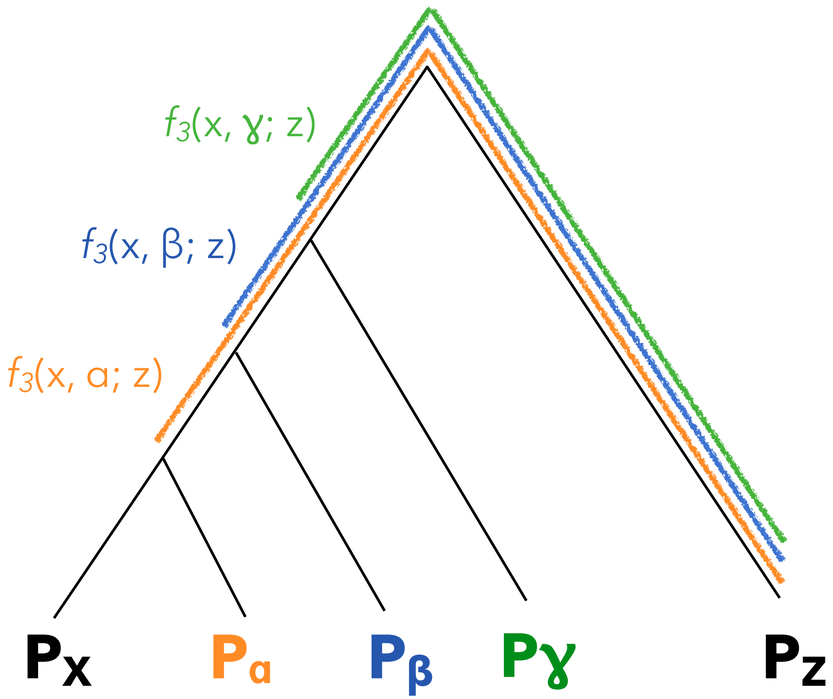


$f_3(x, \alpha; z) > f_3(x, \beta; z) > f_3(x, \gamma; z)$

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00099	HG00101
12	60076	.	A	C	100	PASS	.	GT	1|0	0|0	0|0
12	60252	.	A	G	100	PASS	.	GT	0|0	0|0	0|1
12	60317	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	60344	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60383	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	60405	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60474	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	60614	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	60628	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	60654	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61021	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61107	.	G	T	100	PASS	.	GT	0|0	0|0	0|0
12	61172	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61220	.	G	A	100	PASS	.	GT	0|0	0|0	0|1
12	61258	.	C	T	100	PASS	.	GT	0|0	0|0	0|1
12	61272	.	T	C	100	PASS	.	GT	0|0	0|0	0|0
12	61329	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61341	.	G	A	100	PASS	.	GT	0|0	0|1	0|1
12	61368	.	C	T	100	PASS	.	GT	0|0	0|1	0|1
12	61392	.	T	A	100	PASS	.	GT	0|0	0|0	0|0
12	61405	.	G	C	100	PASS	.	GT	0|0	0|0	0|0
12	61411	.	C	A	100	PASS	.	GT	0|0	0|0	0|0
12	61416	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61422	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61476	.	C	G	100	PASS	.	GT	0|0	0|0	0|0
12	61510	.	G	A	100	PASS	.	GT	0|0	0|0	0|0
12	61516	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61552	.	C	T	100	PASS	.	GT	0|0	0|0	0|0
12	61604	.	T	G	100	PASS	.	GT	0|0	0|0	0|0
12	61687	.	G	A	100	PASS	.	GT	1|0	0|1	0|1
12	61700	.	C	T	100	PASS	.	GT	0|0	0|0	0|0

This evaluation of "tree-ness" requires lots of time, disk space, and memory.

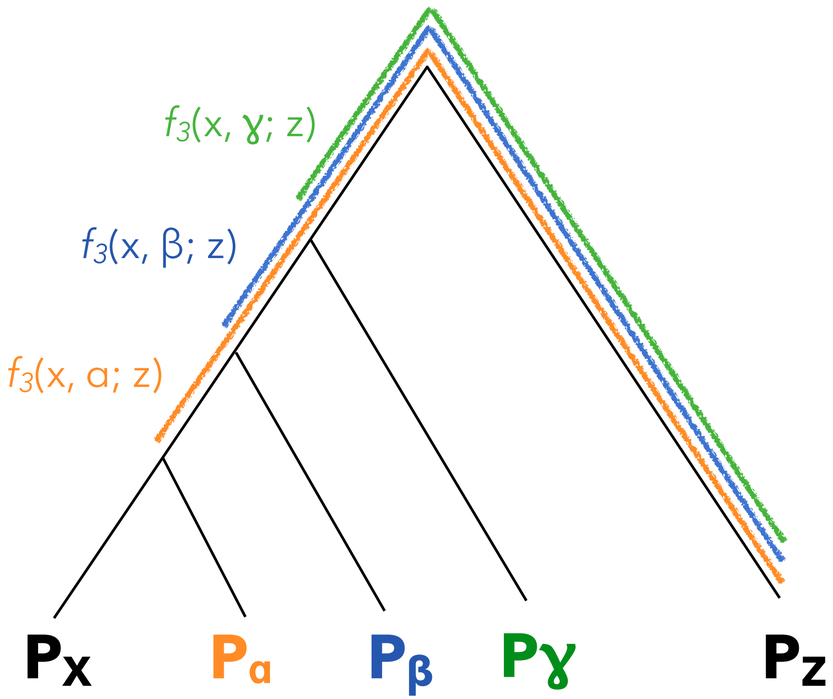
With a tree sequence, computing statistics is trivial...

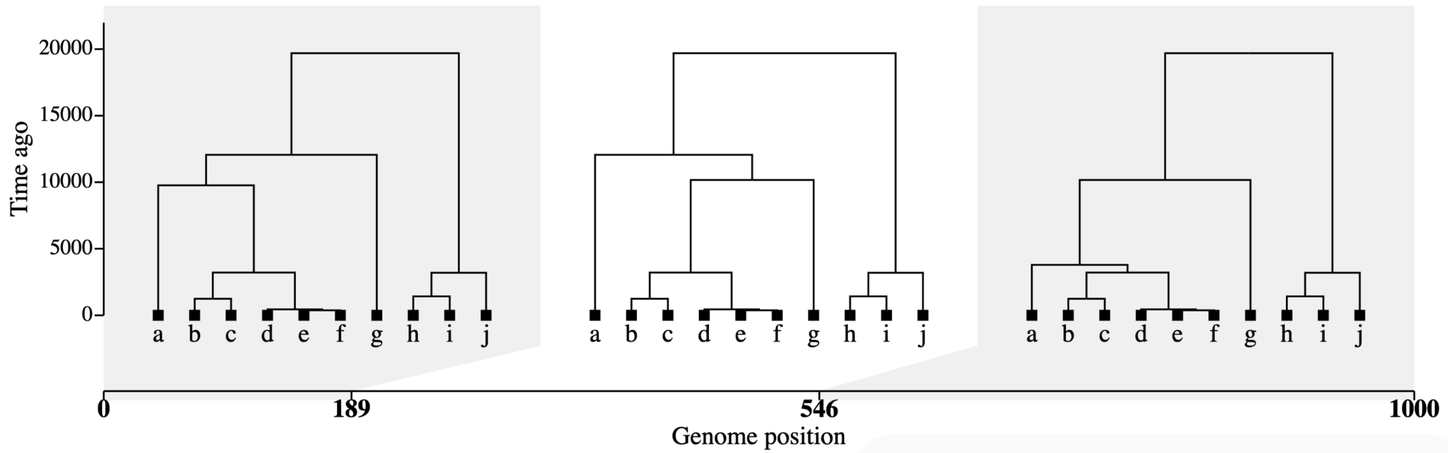


f_3(x, α; z) > f_3(x, β; z) > f_3(x, ɣ; z)

With a tree sequence, computing statistics is trivial...

... <u>because we have all trees to begin with!</u>

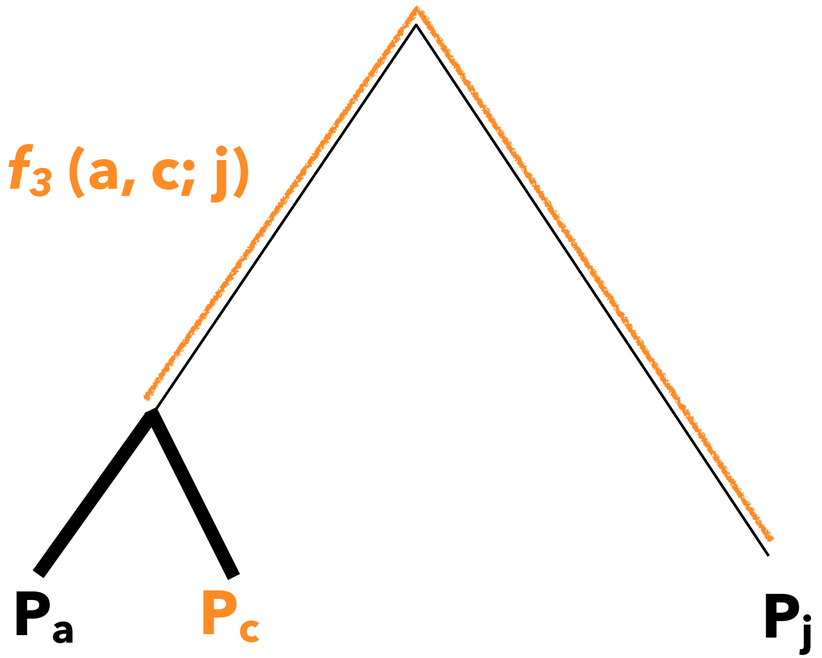


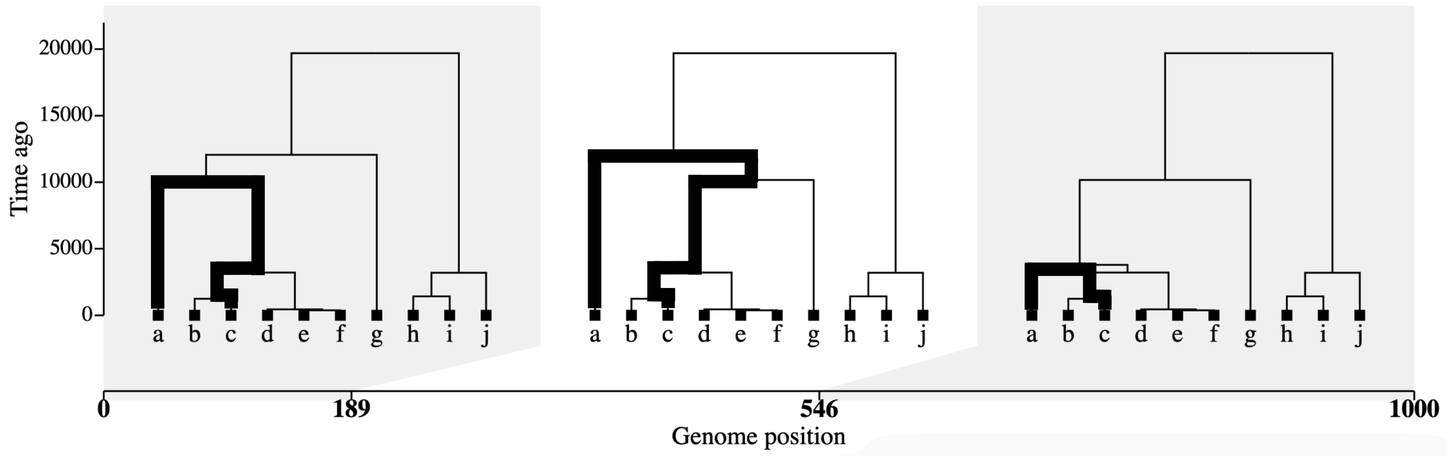


f_3(x, α; z) > f_3(x, β; z) > f_3(x, γ; z)

With a tree sequence, computing statistics is trivial...

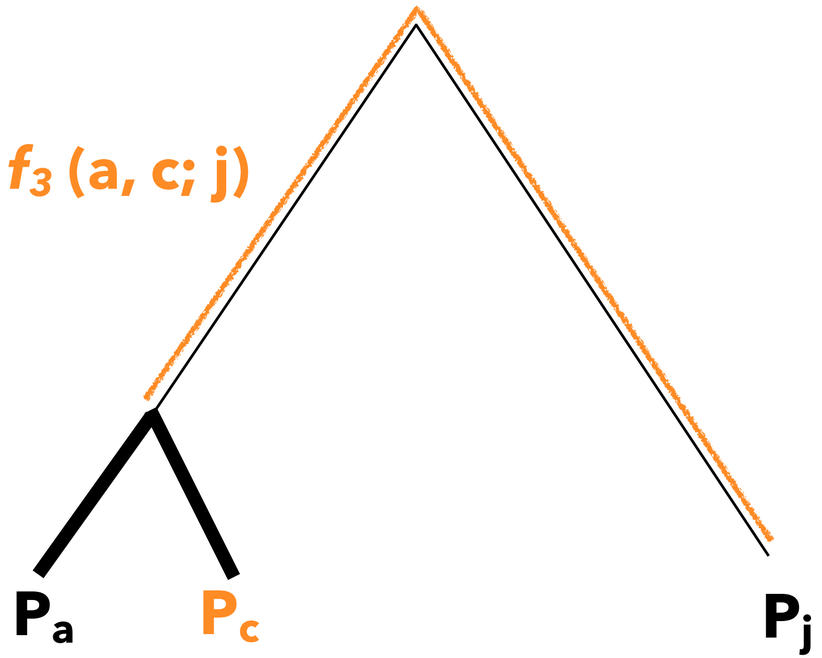
... <u>because we have all trees to begin with!</u>

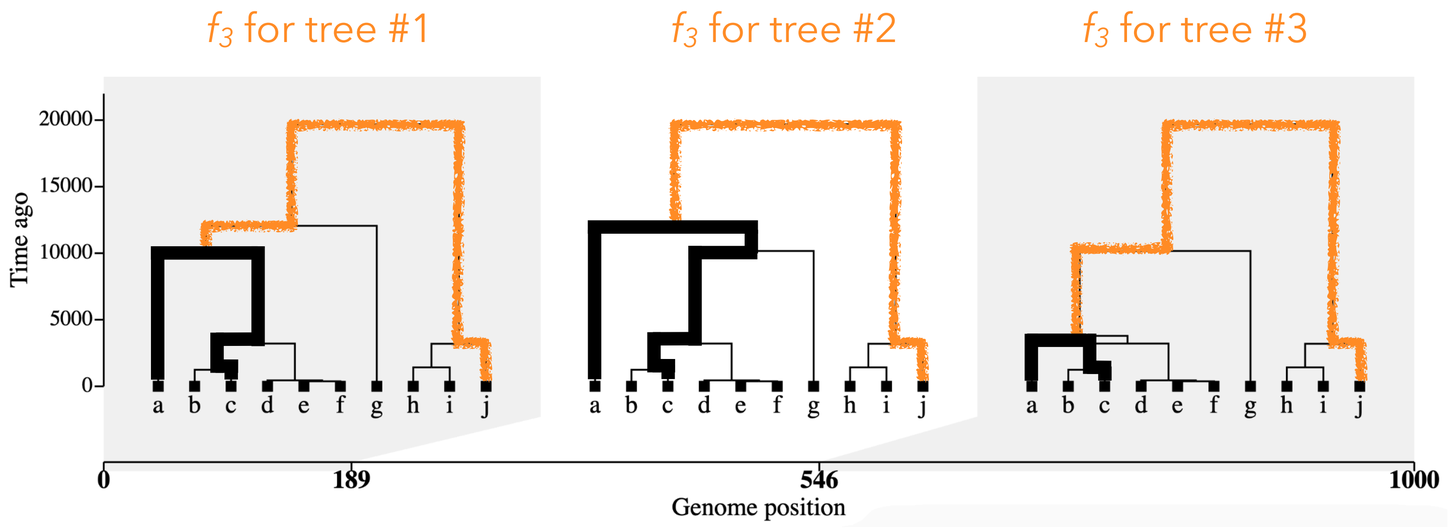




With a tree sequence, computing statistics is trivial...

... <u>because we have all trees to begin with!</u>

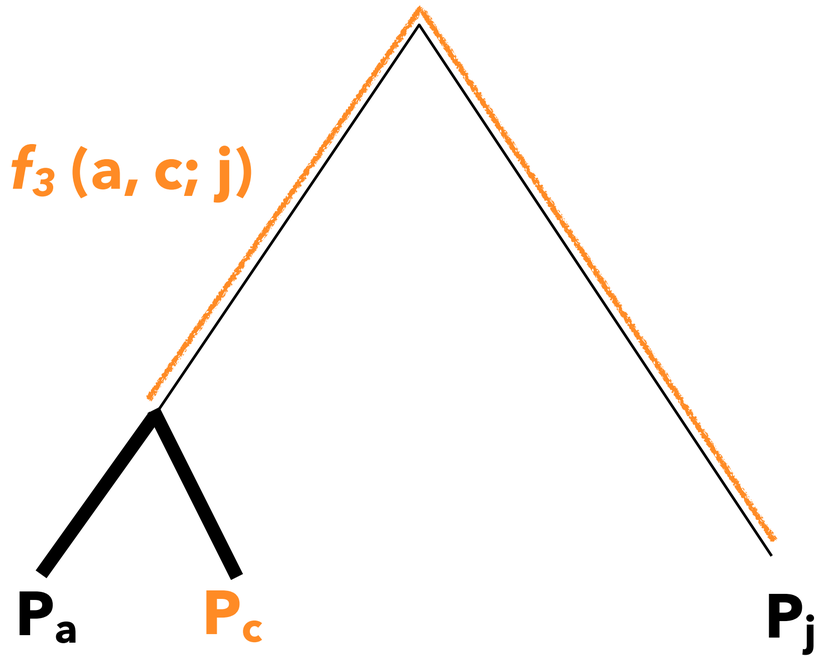


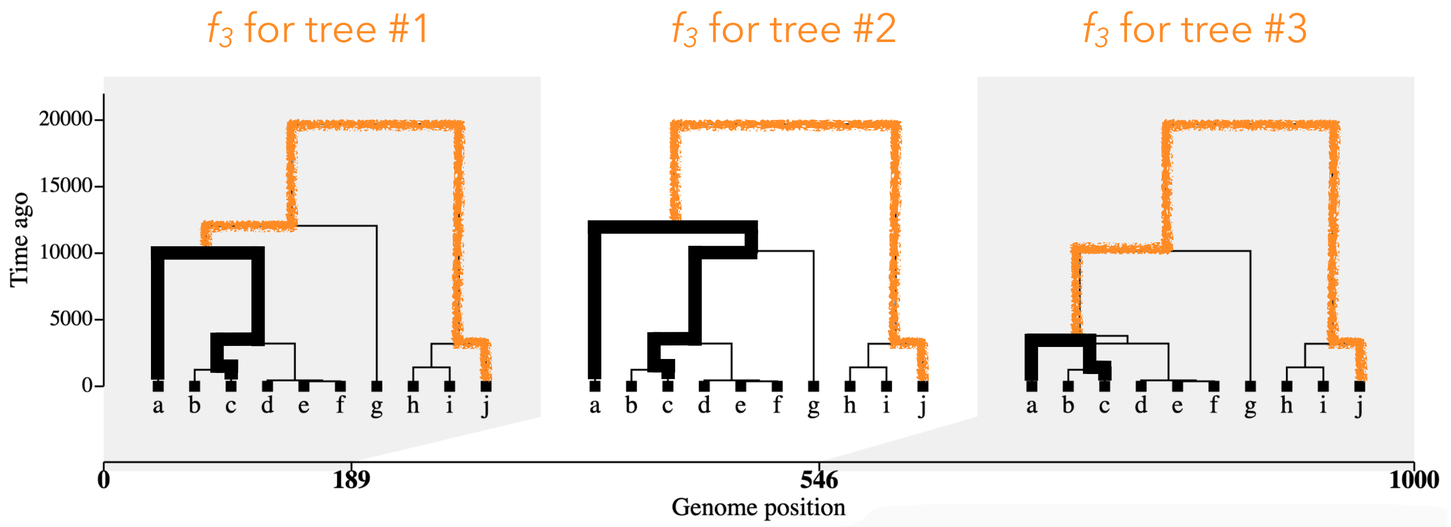


Then aggregate f_3 values across all trees.

With a tree sequence, computing statistics is trivial...

... <u>because we have all trees to begin with!</u>





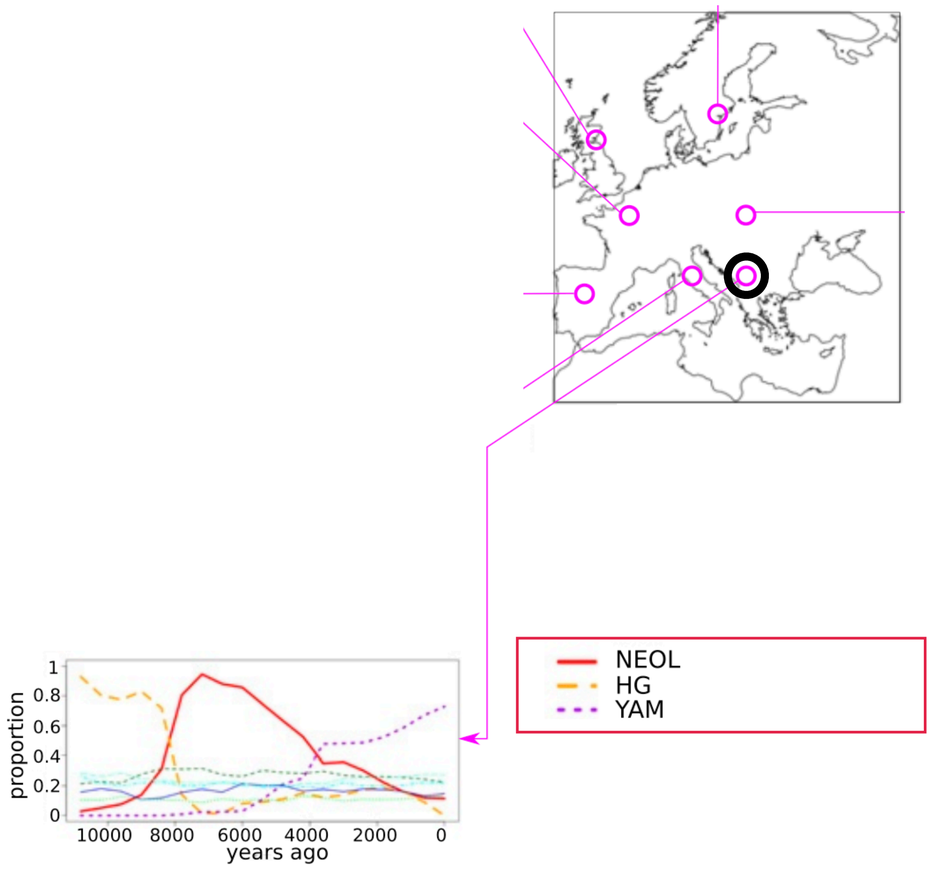
Then aggregate f_3 values across all trees.

This is implemented in *tskit* (tskit.dev) & the R package *slendr* (slendr.net).

Were human migrations associated with changes in landscape?

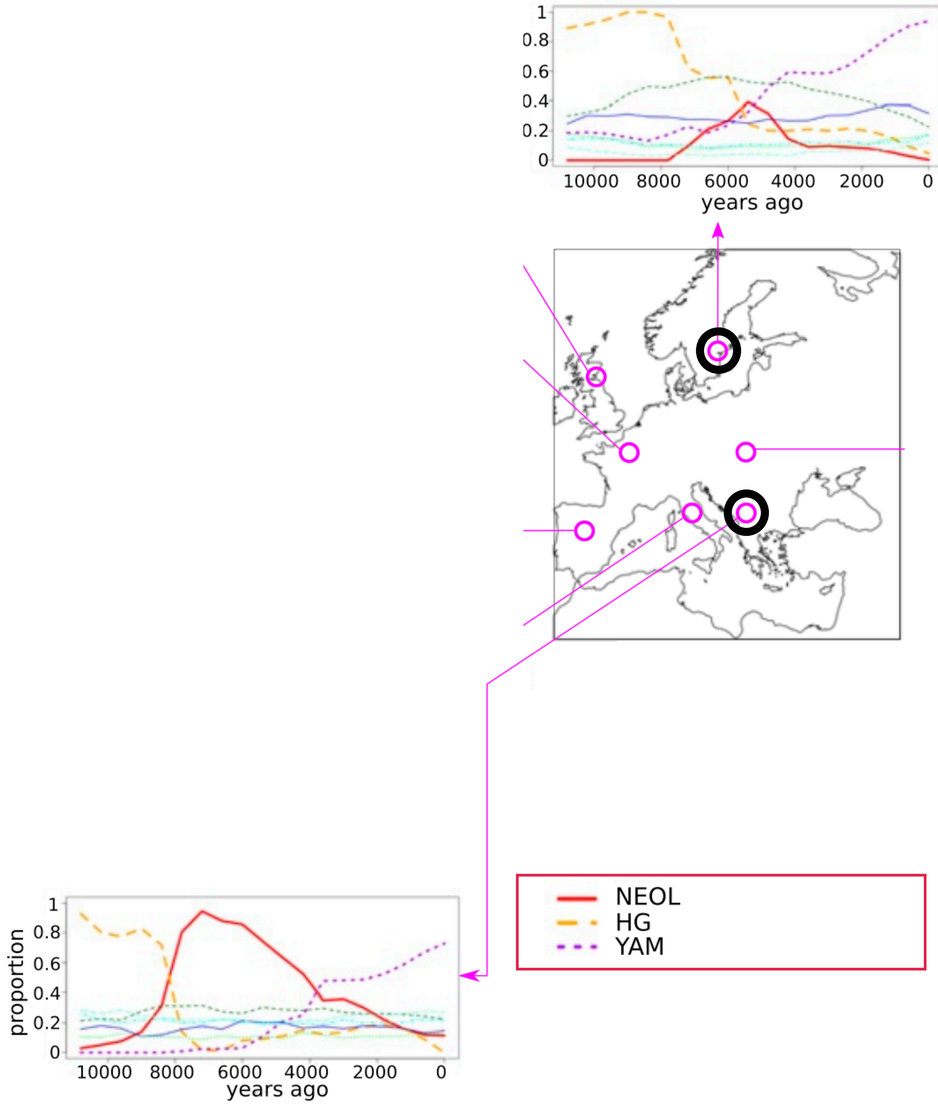
Racimo *et al.*, PNAS 2020

Were human migrations associated with changes in landscape?



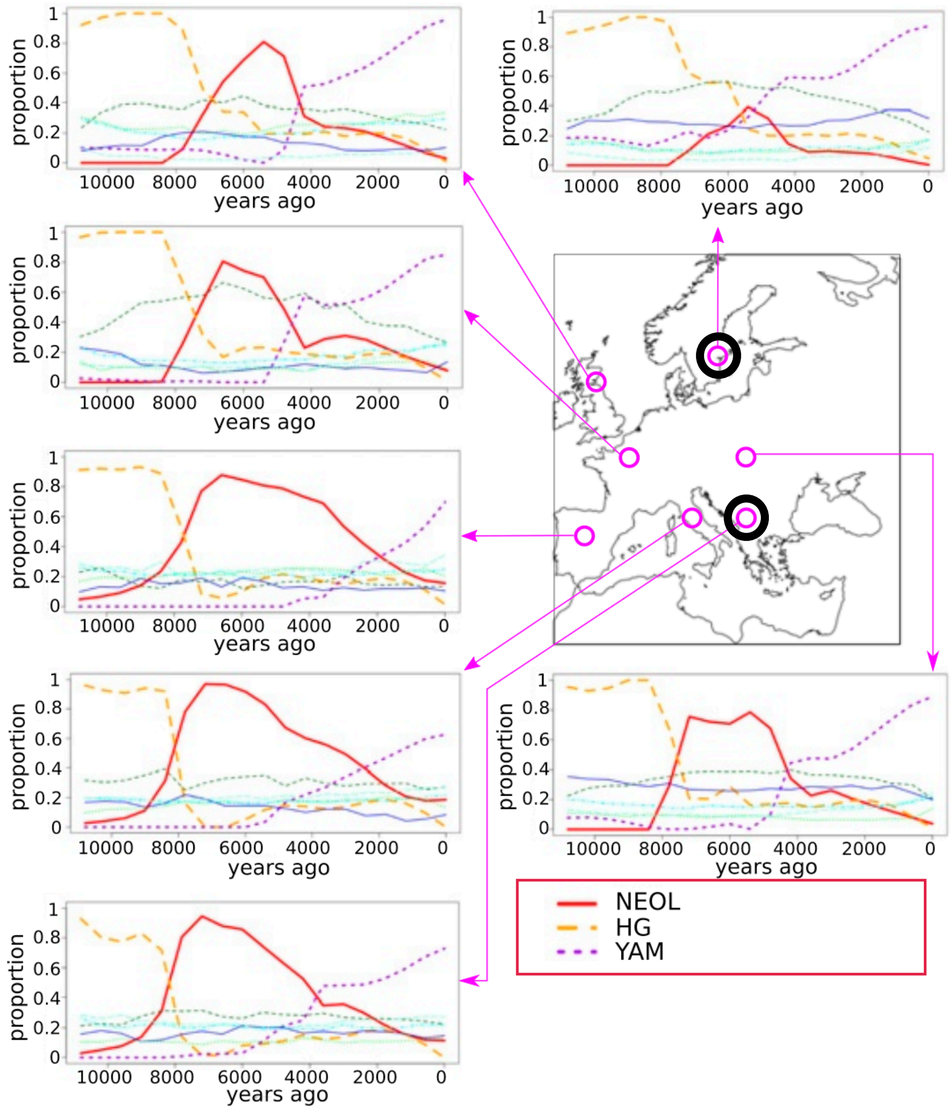
Racimo *et al.*, PNAS 2020

Were human migrations associated with changes in landscape?



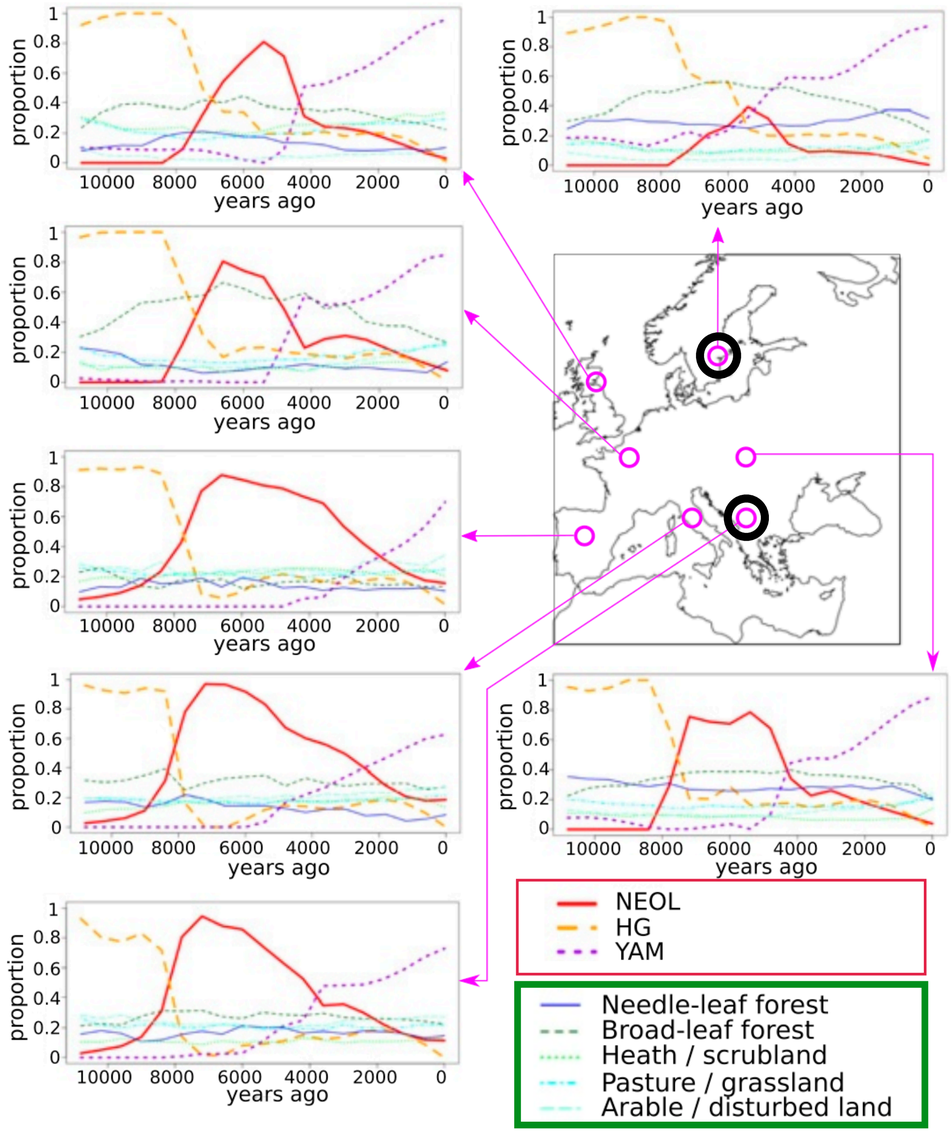
Racimo *et al.*, PNAS 2020

Were human migrations associated with changes in landscape?



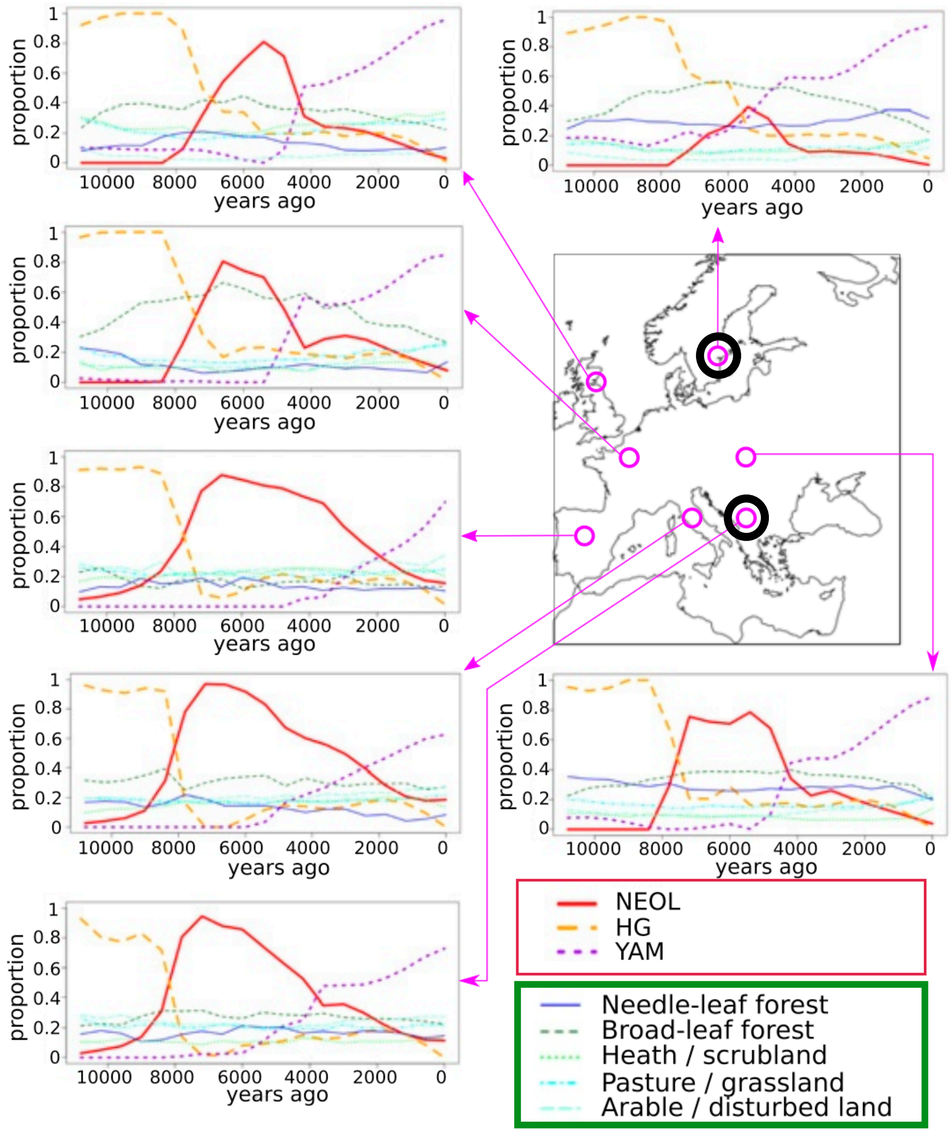
Racimo *et al.*, PNAS 2020

Were human migrations associated with changes in landscape?



Racimo *et al.*, PNAS 2020

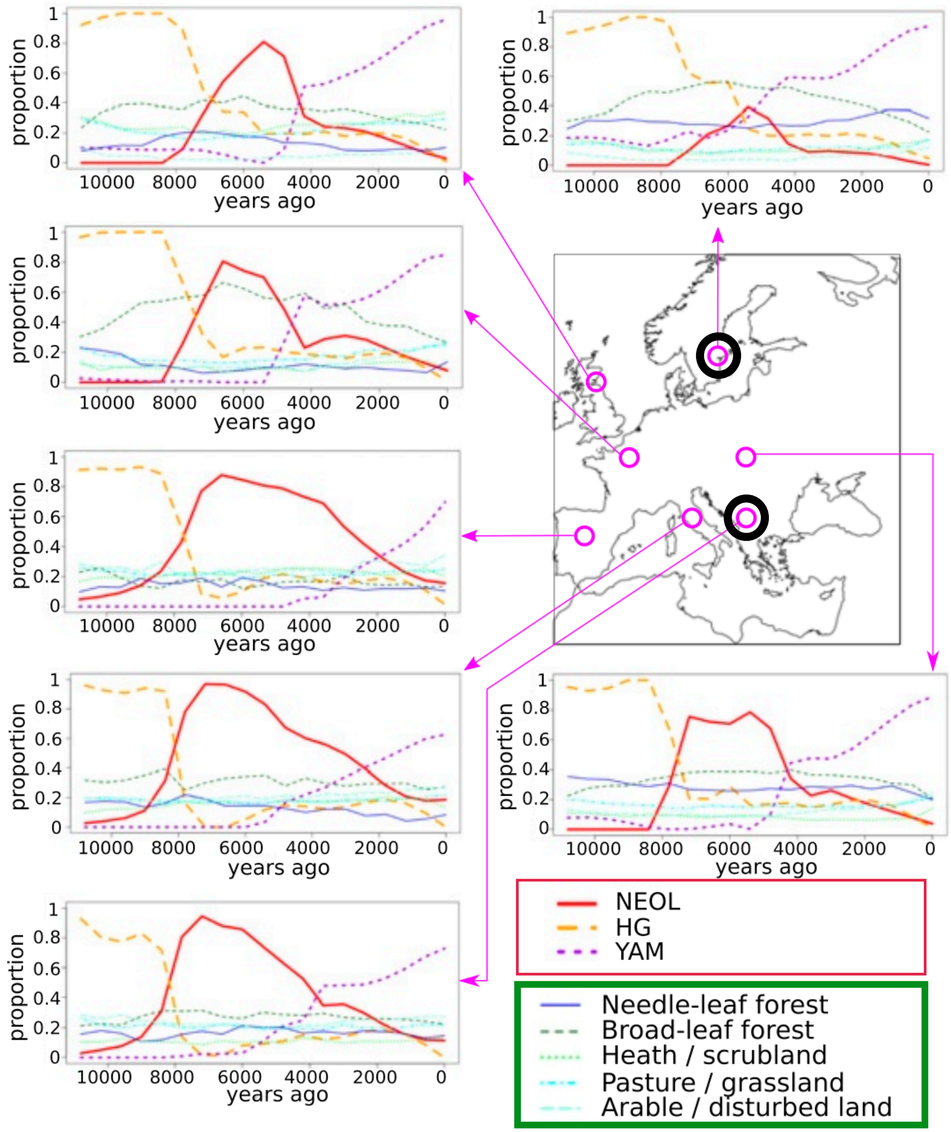
Were human migrations associated with changes in landscape?



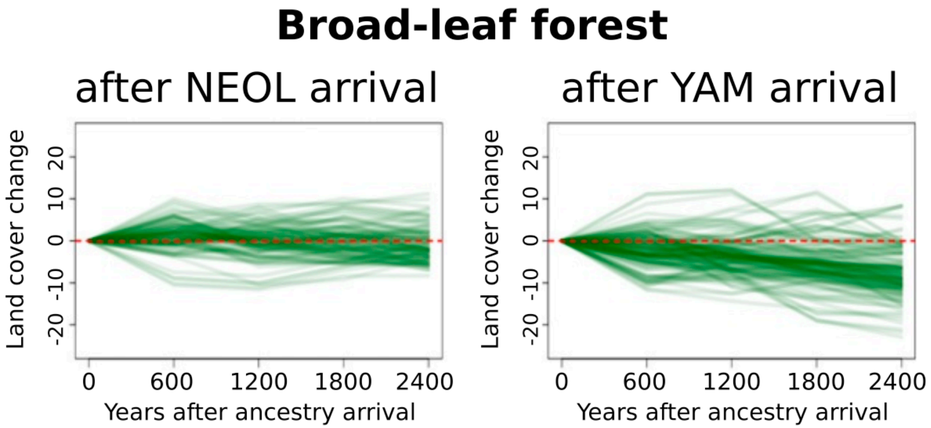
genetic and environmental metrics aggregated over each location of the entire map

Racimo *et al.*, PNAS 2020

Yamnaya coincided with strong vegetation changes

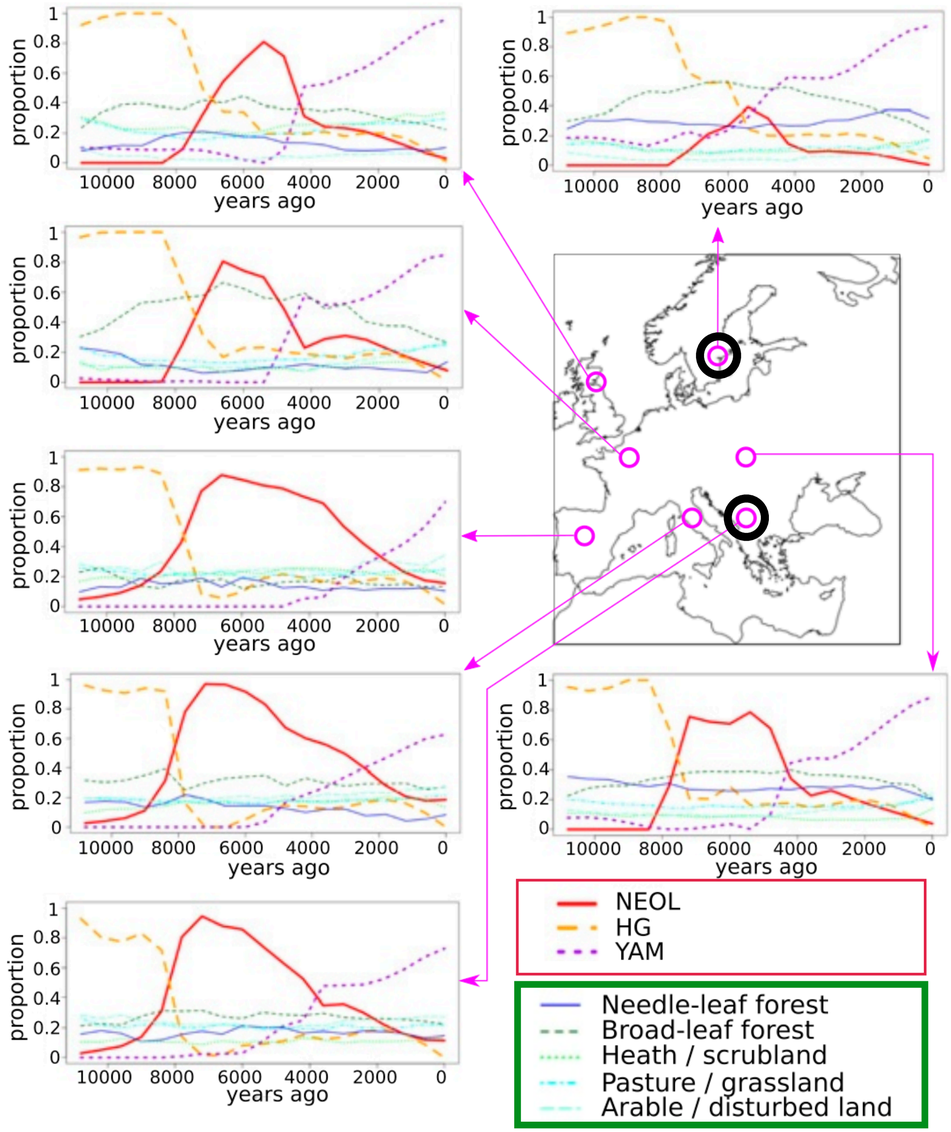


genetic and environmental metrics aggregated over each location of the entire map

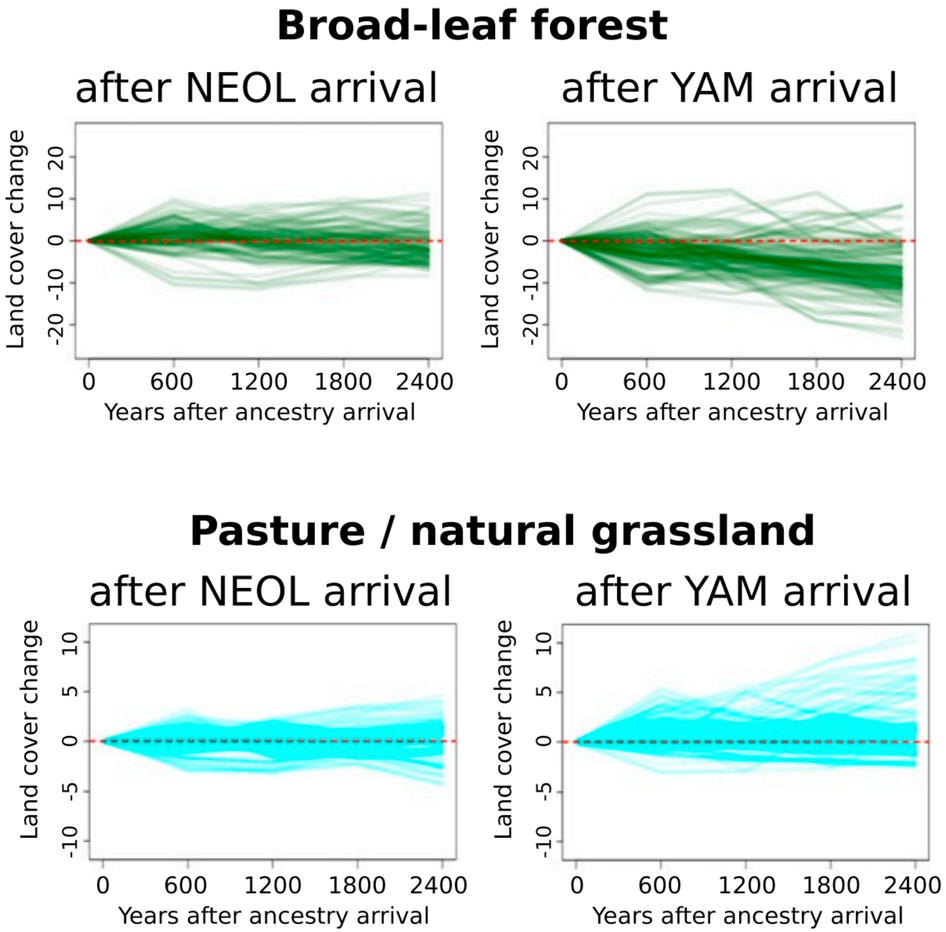


Racimo *et al.*, PNAS 2020

Yamnaya coincided with strong vegetation changes



genetic and environmental metrics aggregated over each location of the entire map



Racimo *et al.*, PNAS 2020

Every other statistic

— nucleotide diversity, IBD, split times, N_e, D statistic, ...—

works in the same way.

Every other statistic

— nucleotide diversity, IBD, split times, N_e, D statistic, ...—

works in the same way.

Inferring whole-genome histories in large population datasets

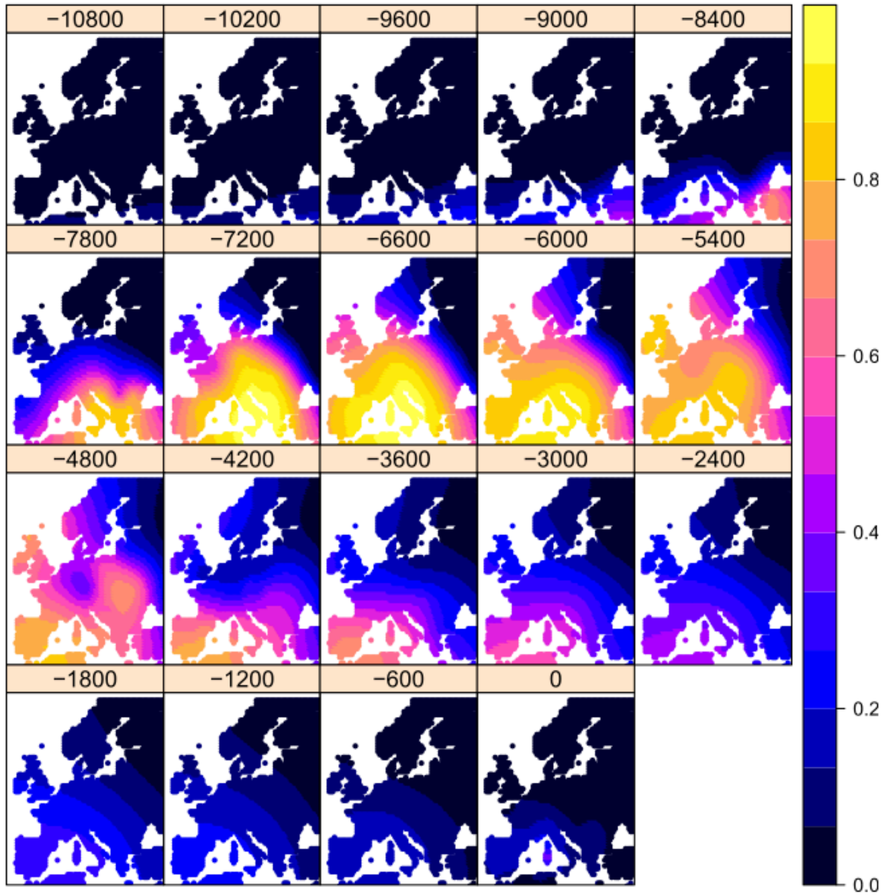
Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers and Gil McVean

A method for genome-wide genealogy estimation for thousands of samples

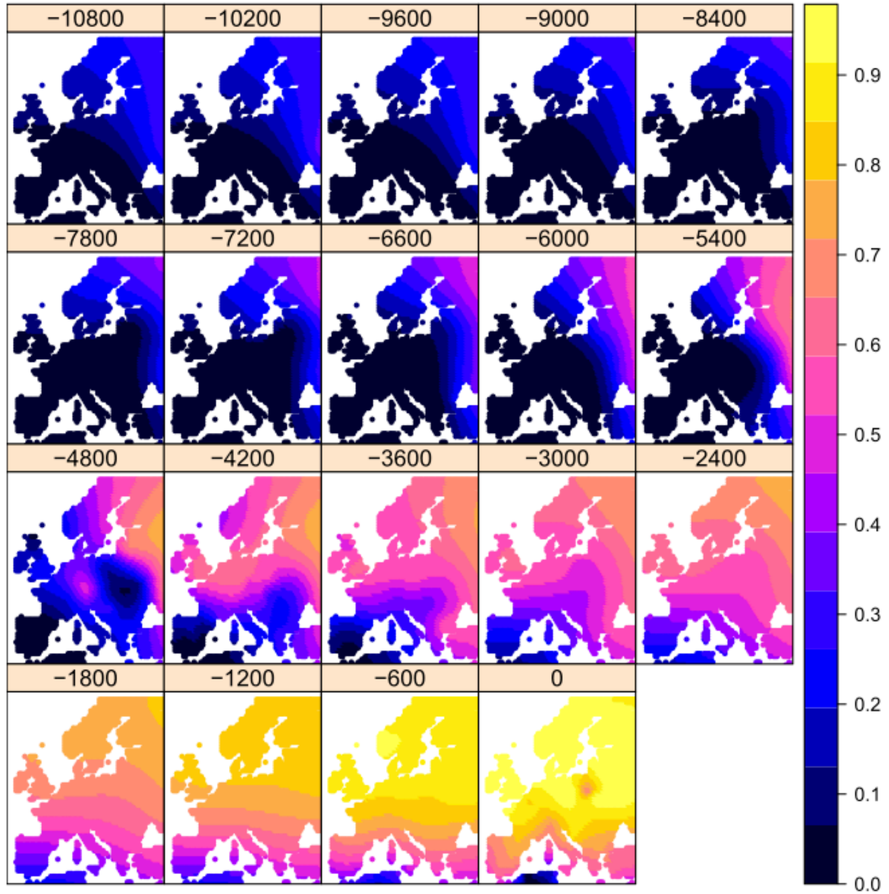
Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myers[1,3]*

Interpolated ancestry maps

"NEOL" ancestry



"YAM" ancestry



Racimo *et al.*, PNAS 2020