Introduction to population genetics

leo.planche@universite-paris-saclay.fr

Today's Overview

i. What is population genomics

ii. Variation between populations

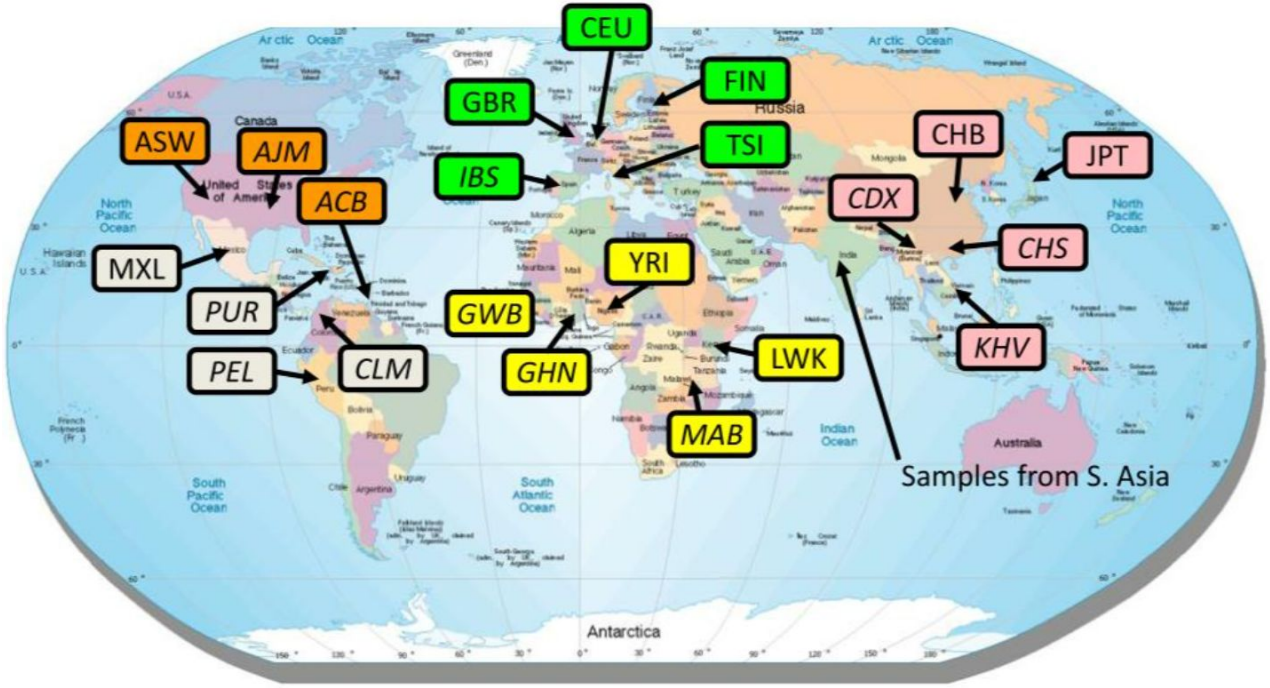
iii. Population size

iv. Sapiens meets Neandertal and Denisovan

What is population genetics?

Population genetics is a subfield of genetics that deals with genetic differences within and among populations.

What is a populations? A group of freely interbreeding individuals.



How many labeled populations are displayed on this map?

Why do population genetics?

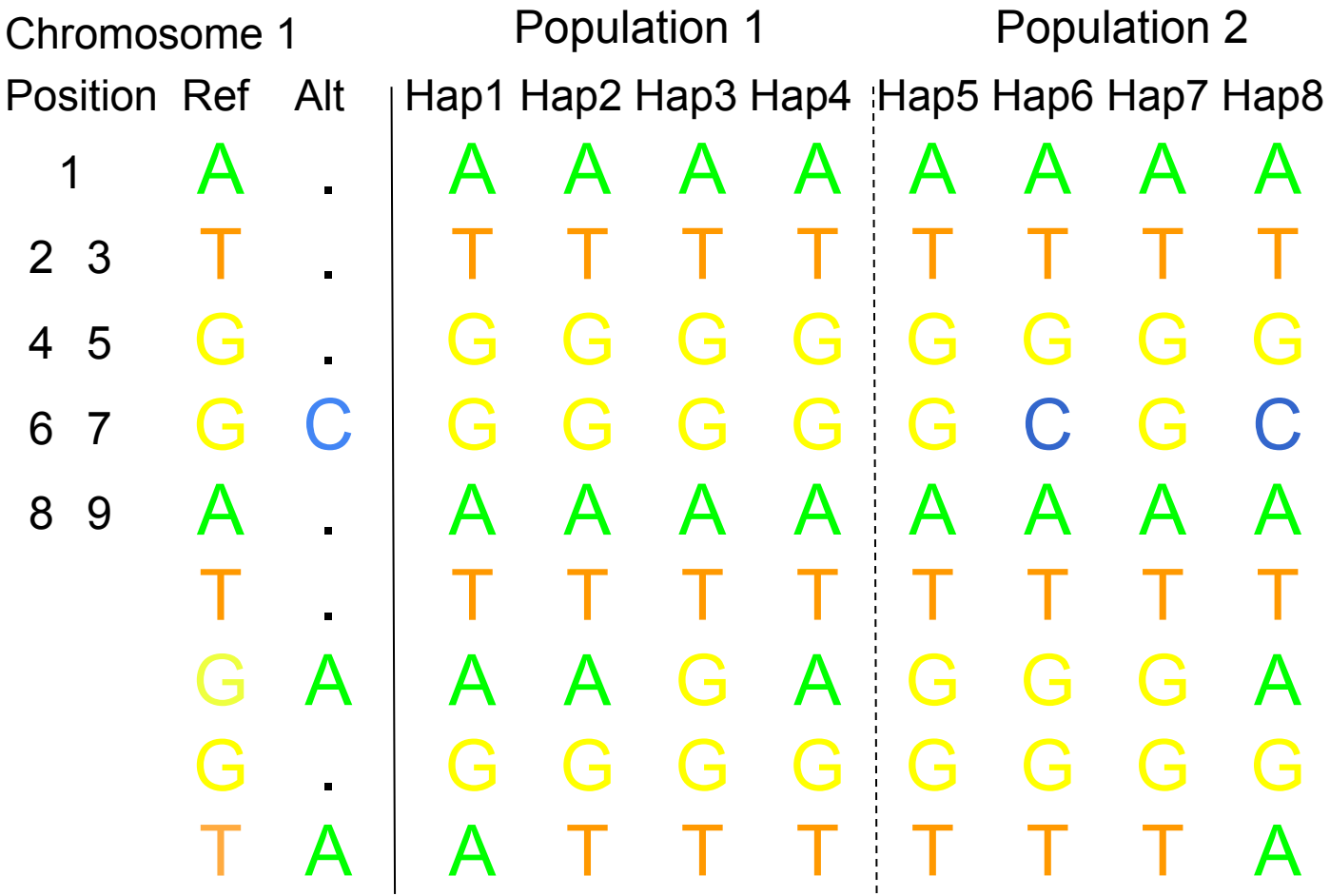
- For the love of history (human or non human)
- For medical research
- For research in ecology (conservation genetics)

Genetic variations

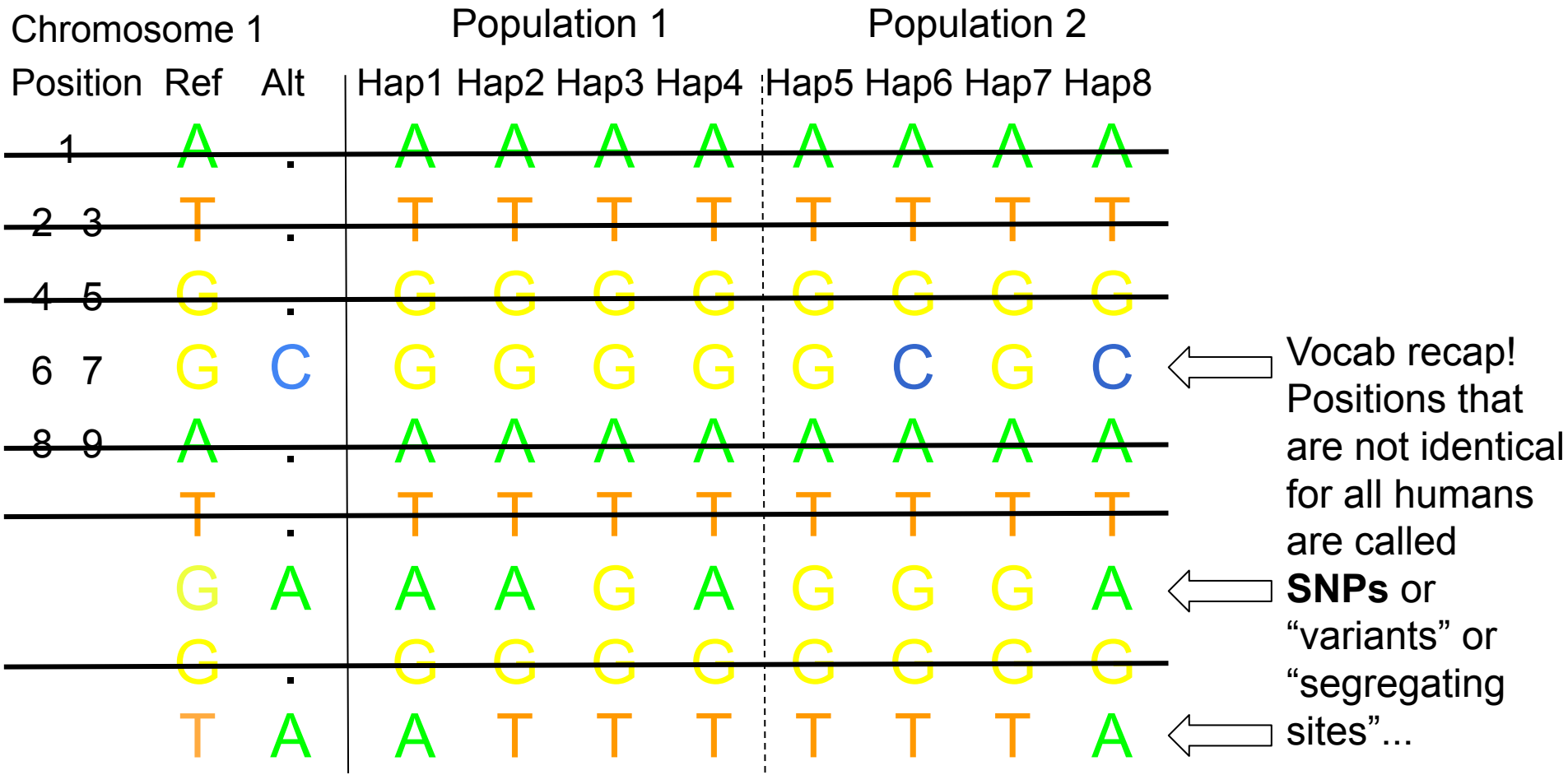
Mutations

Human mutation rate is around 1.5×10^{-8} per base pair (bp). Meaning that around 30 new variants with be introduced in each gamete.
The vast majority of mutations are **neutral**.

Genetic variations



Genetic variations



Genetic variations: allele frequency

Chromosome 1			Population 1				Population 2			
Position	Ref	Alt	Hap1	Hap2	Hap3	Hap4	Hap5	Hap6	Hap7	Hap8
47	G	C	G	G	G	G	G	C	G	C
9	G	A	A	A	G	A	G	G	G	A
	T	A	A	T	T	T	T	T	T	A

Genetic variations: allele frequency

Chromosome 1			Population 1				Population 2			
Position	Ref	Alt	Hap1	Hap2	Hap3	Hap4	Hap5	Hap6	Hap7	Hap8
4	G	C	0	0	0	0	0	1	0	1
7	G	A	1	1	0	1	0	0	0	1
9	T	A	1	0	0	0	0	0	0	1

The allele frequency is the frequency of an allele in a population

Frequency of the reference allele at position 4:
In population 1 : 4/4 = 1
In population 2 : 2/4 = 0.5

Frequency of the reference allele at position 7:
In population 1 : 1/4 = 0.25
In population 2 : 3/4 = 0.75

Allele frequencies:

Position	Pop1	Pop2
4	1	0.5
7	0.25	0.75
9	0.75	0.75

Genetic variations: allele frequency
Fixation index (FST)

The FST is a statistic to measure differentiation between two population.

$H_T = 2*\sum f_i.(1-f_i)$ with f_i the allele frequency at position i in both populations. This measure the allele diversity taking both populations together.

$H_S = \sum(f^1_i.(1-f^1_i) + f^2_i.(1-f^2_i))$ with f^1_i (resp. f^2_i) the allele frequency at position i in populations 1 and (resp. 2). This measure the allele diversity in both populations separately.

$$FST = \frac{H_T - H_S}{H_T}$$

Allele frequencies:

Position	Pop1	Pop2	Pop1&Pop2
4	1	0.5	0.75
7	0.25	0.75	0.5
9	0.75	0.75	0.75

Genetic variations: allele frequency
Fixation index (FST)

The FST is a statistic to measure differentiation between two population.

$H_T = 2*\sum f_i.(1-f_i)$ with f_i the frequency of allele i in both populations. This measure the allele diversity taking both populations together.

$H_S = \sum(f^1_i.(1-f^1_i) + f^2_i.(1-f^2_i))$ with f^1_i (resp. f^2_i) the frequency of allele i in populations 1 and (resp. 2). This measure the allele diversity in both populations separately.

$$FST = \frac{H_T - H_S}{H_T}$$

Allele frequencies:

Position	Pop1	Pop2	Pop1&Pop2
4	1	0.5	0.75
7	0.25	0.75	0.5
9	0.75	0.75	0.75

$H_T = 2*(0.75*0.25+0.5*0.5+0.75*0.25) = 1.25$

Genetic variations: allele frequency
Fixation index (FST)

The FST is a statistic to measure differentiation between two population.

$H_T = 2*\sum f_i.(1-f_i)$ with f_i the frequency of allele i in both populations. This measure the allele diversity taking both populations together.

$H_S = \sum(f^1_i.(1-f^1_i) + f^2_i.(1-f^2_i))$ with f^1_i (resp. f^2_i) the frequency of allele i in populations 1 and (resp. 2). This measure the allele diversity in both populations separately.

$$FST = \frac{H_T - H_S}{H_T}$$

Allele frequencies:

Position	Pop1	Pop2	Pop1&Pop2
4	1	0.5	0.75
7	0.25	0.75	0.5
9	0.75	0.75	0.75

H_T=2*(0.75*0.25+0.5*0.5+0.75*0.25)=1.25

H_S=1*0+0.25*0.75+0.75*0.25+...

Genetic variations: allele frequency
Fixation index (FST)

The FST is a statistic to measure differentiation between two population.

$H_T = 2*\sum f_i.(1-f_i)$ with f_i the frequency of allele i in both populations. This measure the allele diversity taking both populations together.

$H_S = \sum(f^1_i.(1-f^1_i) + f^2_i.(1-f^2_i))$ with f^1_i (resp. f^2_i) the frequency of allele i in populations 1 and (resp. 2). This measure the allele diversity in both populations separately.

$$FST = \frac{H_T - H_S}{H_T}$$

Allele frequencies:

Position	Pop1	Pop2	Pop1&Pop2
4	1	0.5	0.75
7	0.25	0.75	0.5
9	0.75	0.75	0.75

H_T=2*(0.75*0.25+0.5*0.5+0.75*0.25)=1.25
H_S=1*0+0.25*0.75+0.75*0.25+0.5*0.5+0.75*0.25+0.75*0.25=1

Genetic variations: allele frequency
Fixation index (FST)

The FST is a statistic to measure differentiation between two population.

$H_T = 2*\sum f_i.(1-f_i)$ with f_i the frequency of allele i in both populations. This measure the allele diversity taking both populations together.

$H_S = \sum(f^1_i.(1-f^1_i) + f^2_i.(1-f^2_i))$ with f^1_i (resp. f^2_i) the frequency of allele i in populations 1 and (resp. 2). This measure the allele diversity in both populations separately.

$$FST = \frac{H_T - H_S}{H_T}$$

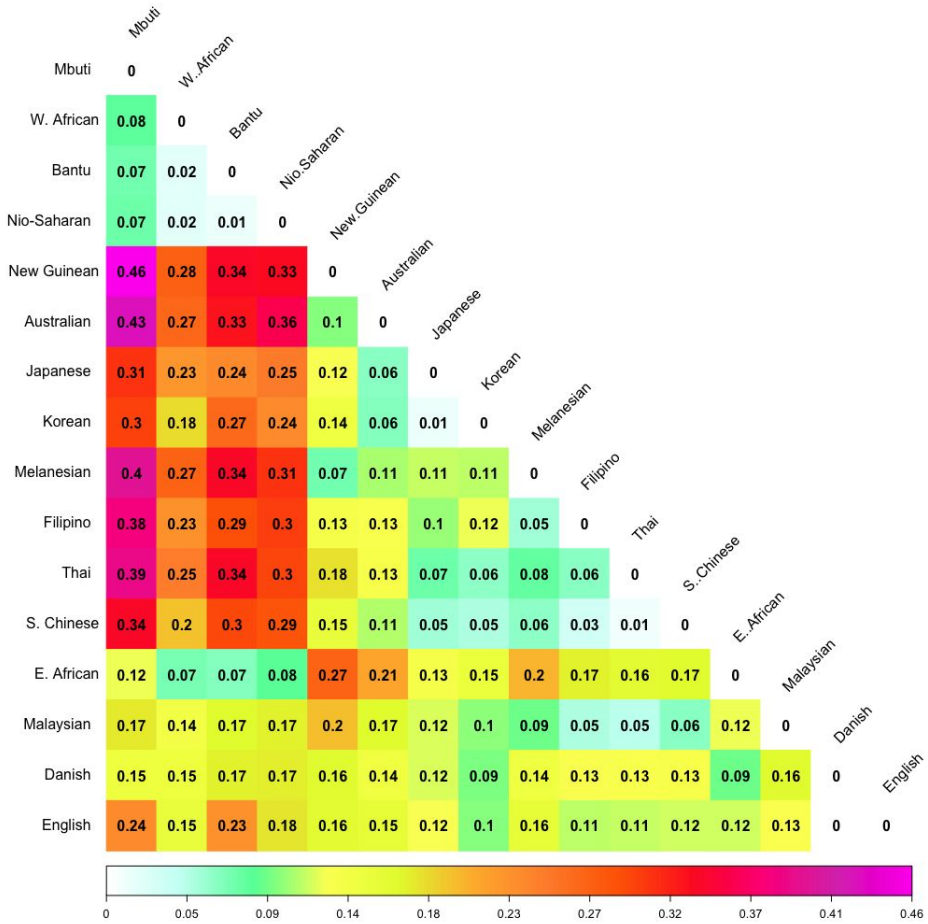
Allele frequencies:

Position	Pop1	Pop2	Pop1&Pop2
4	1	0.5	0.75
7	0.25	0.75	0.5
9	0.75	0.75	0.75

H_T=2*(0.75*0.25+0.5*0.5+0.75*0.25)=1.25
H_S=1*0+0.25*0.75+0.75*0.25+0.5*0.5+0.75*0.25+0.75*0.25=1
FST = (1.25-1)/1.25=**0.2**

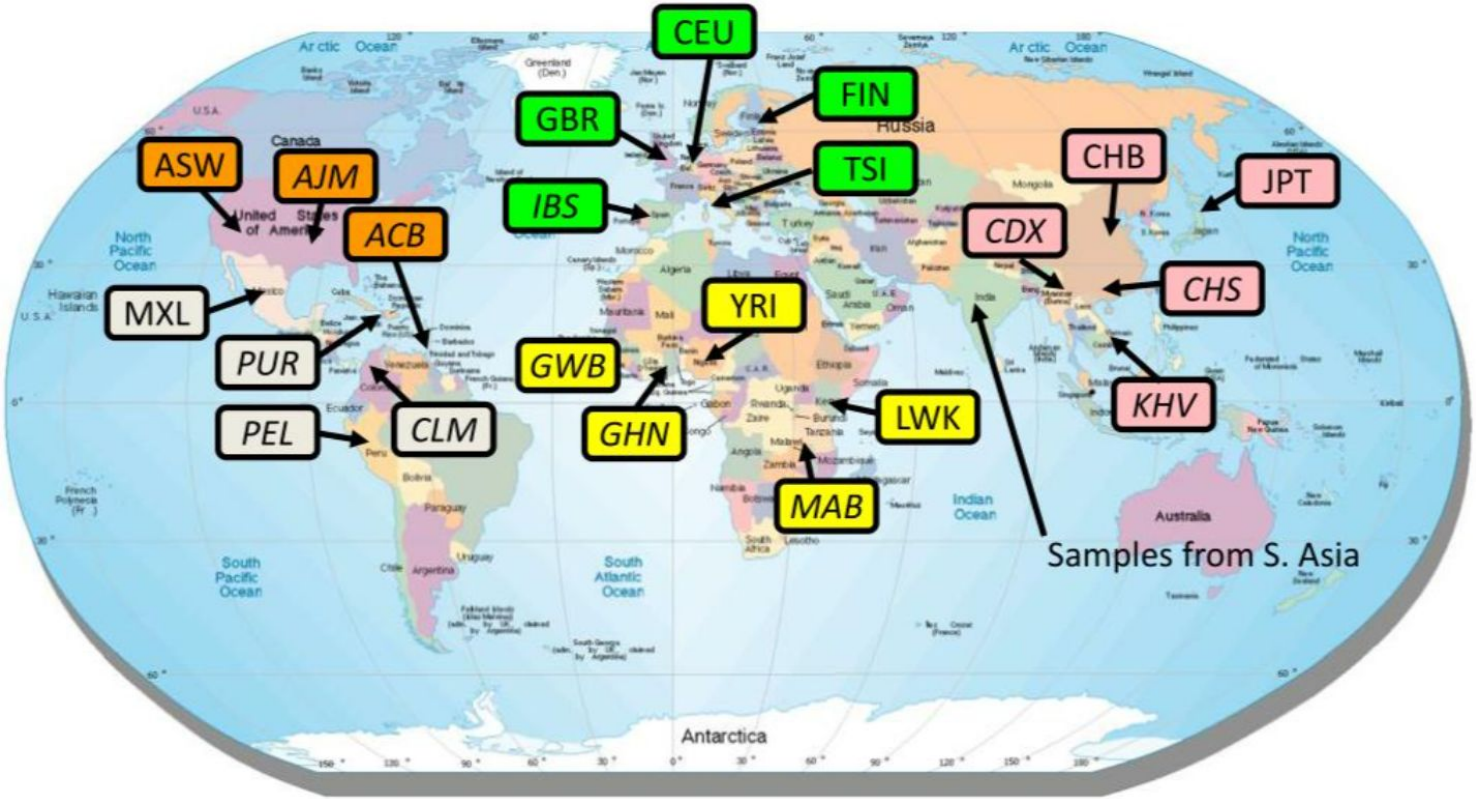
Conclusion: There is 20% more differences between the two populations than within each one.

Genetic variations: allele frequency - (FST)



- The average FST between continent is at around 10%.
- If were were to do an FST-like statistic on skin color, that number would be around 90%.
- The reason is strong selection on genes responsible for the skin color.

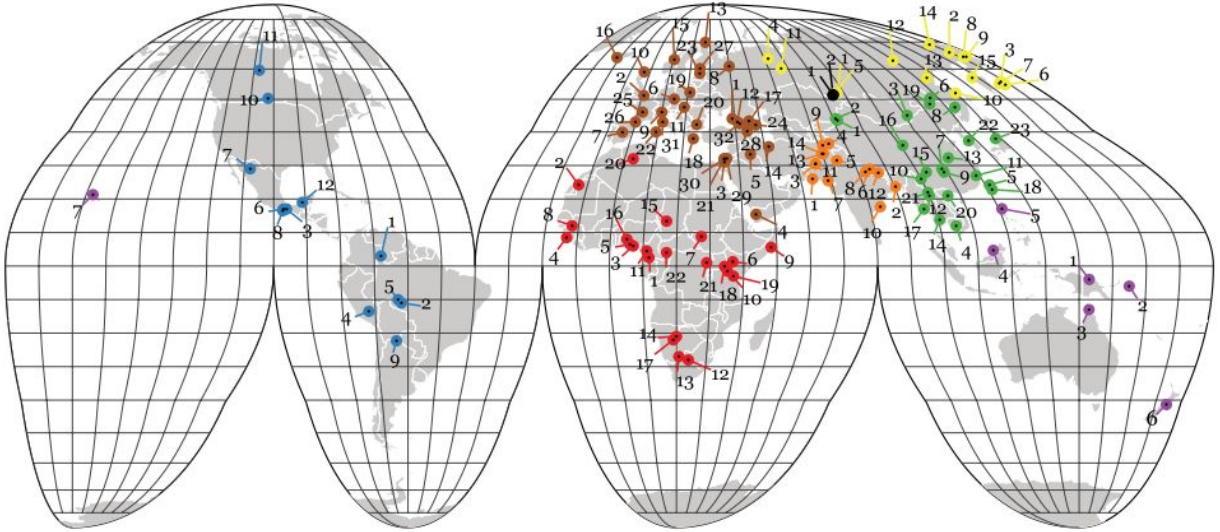
Data availability - 1000 Genome project



Sequencing of around 100 genomes per 12 different populations around the globe.

Data is publicly available.

Data availability - SGDP



236 individual genomes (total) from 125 distinct human populations

1 - Neanderthal
2 - Denisova
1 - Piapoco
2 - Surui
3 - Mixe
4 - Quechua
5 - Karitiana
6 - Mixtec
7 - Pima
8 - Zapotec
9 - Chane
10 - Cree
11 - Chipewyan
12 - Maya
1 - Xibo
2 - Uygur
3 - Mongola
4 - Cambodian
5 - Atayal
6 - Oroqen
7 - Tujia
8 - Hezhen
9 - Miao
10 - Naxi
11 - She
12 - Dai
13 - Han
14 - Thai
15 - Yi
16 - Tu
17 - Burmese
18 - Ami
19 - Daur
20 - KinhKVH
21 - Lahu
22 - Korean
23 - Japanese
1 - Papuan
2 - Bougainville
3 - AustraliaECCAC
4 - Dusun
5 - Igorot
6 - Maori
7 - Hawaiian
1 - Adygei
2 - EnglandGBR
3 - Bedouin
4 - YemeniteJew
5 - IraqiJew
6 - Czechoslovakian
7 - SpainIBS
8 - Russian
9 - Bergamo
10 - Orcadian
11 - Hungarian
12 - Abkhasian
13 - Saami
14 - Iranian
15 - Norwegian
16 - Icelander
17 - Chechen
18 - Greek
19 - Polish
20 - Bulgarian
21 - Palestinian
22 - Sardinian
23 - Estonian
24 - Lezgin
25 - French
26 - Basque
27 - FinlandFIN
28 - Armenian
29 - Jordanian
30 - Druze
31 - Tuscan
32 - Georgian
1 - Kongo
2 - Saharawi
3 - Igbo
4 - MendeMSL
5 - EsanESN
6 - LuhyaLWK
7 - Dinka
8 - Mandenka
9 - Somali
10 - MasaiMKK
11 - Lemande
12 - BantuSETswana
13 - KhomaniSan
14 - San
15 - GambiaGWD
16 - Yoruba
17 - BantuSEHerero
18 - Luo
19 - BantuKenya
20 - Mozabite
21 - Mbuti
22 - Biaka
1 - Makrani
2 - BengaliBEB
3 - Balochi
4 - Burusho
5 - PunjabiPJL
6 - Tibetan
7 - Sindhi
8 - Kusunda
9 - Kalash
10 - Madiga
11 - Brahui
12 - Sherpa
13 - Hazara
14 - Pathan
1 - Tubalar
2 - EskimoNaukan
3 - AleutBeringIsland
4 - MansiSosvaRiver
5 - Altaian
6 - TlingitPreobrazhenskoye
7 - TlingitNikolskoye
8 - EskimoChaplino
9 - EskimoSireniki
10 - Ulchi
11 - MansiKondaRiver
12 - Yakut
13 - Even
14 - Chukchi
15 - Itelman

Genetic distances

Allele frequencies:

Position	Pop1	Pop2	Pop3
4	1	0.5	0.85
7	0.25	0.75	0.35
9	0.75	0.75	0.8

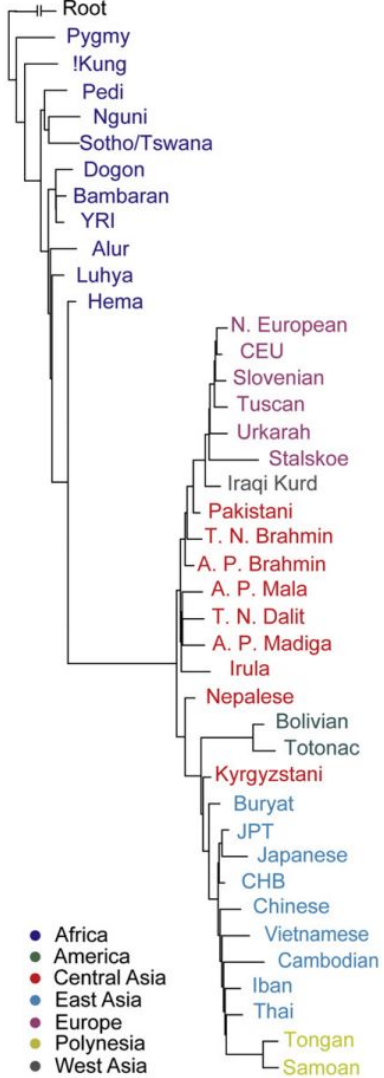
How to compute the distance between population i and j at some position:

$d_{ij} = |p_i - p_j|$

With p_i and p_j allele frequencies.

We then compute mean by taking the average value for all positions.

Genetic distances - Population trees



Allele frequencies:

Position	Pop1	Pop2	Pop3
4	1	0.5	0.85
7	0.25	0.75	0.35
9	0.75	0.75	0.8

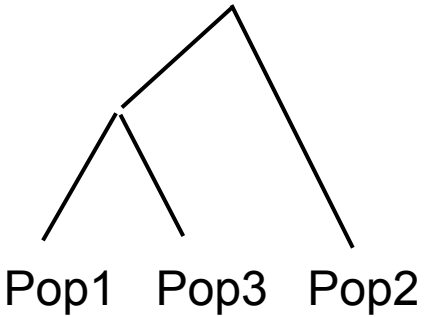
d(pop1,pop2) = (|1-0.5|+|0.25-0.75|+|0.75-0.75|)/3 = 0.33

d(pop1,pop3)=0.1
d(pop2,pop3)=0.26

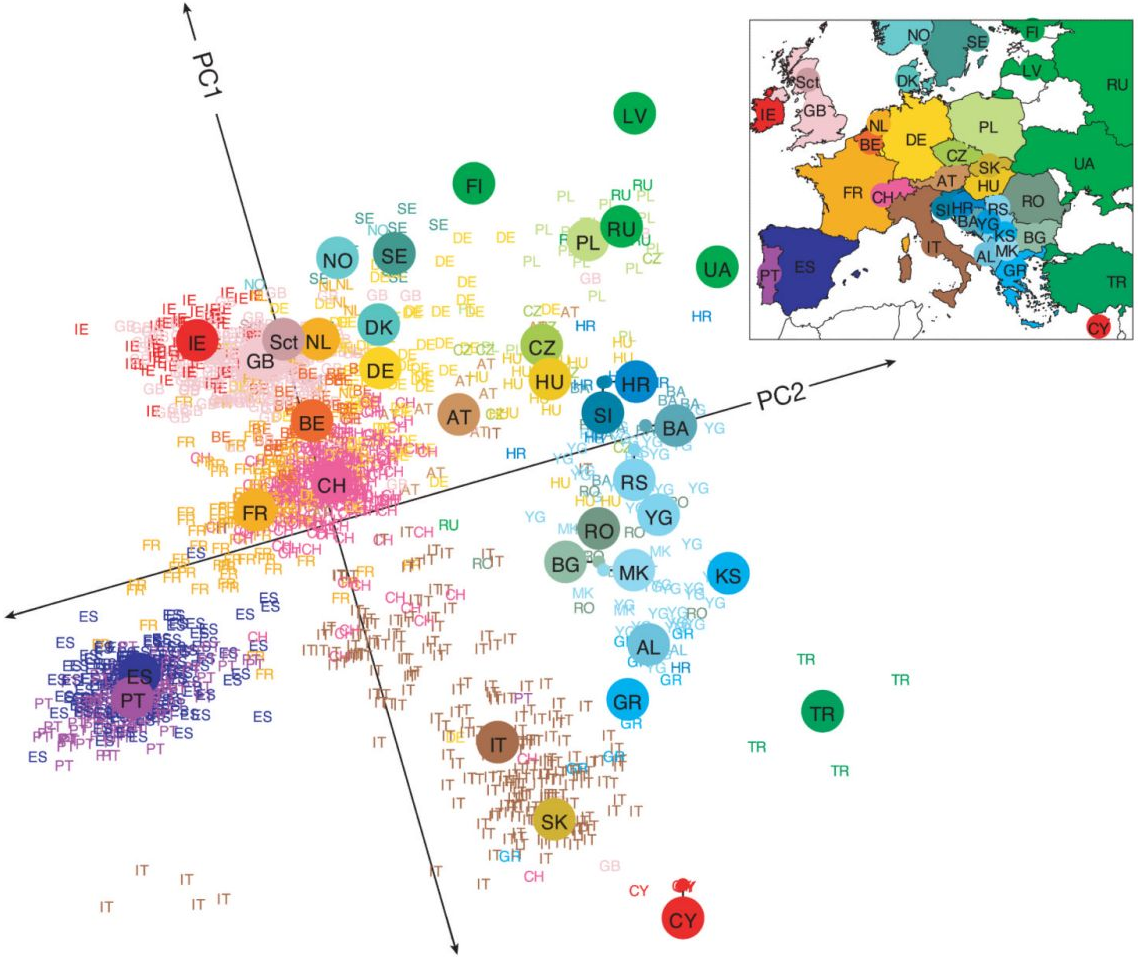
How to compute the distance between population i and j at some position:

$d_{ij} = |p_i - p_j|$

With p_i and p_j allele frequencies.
We then compute mean by taking the average value for all positions.

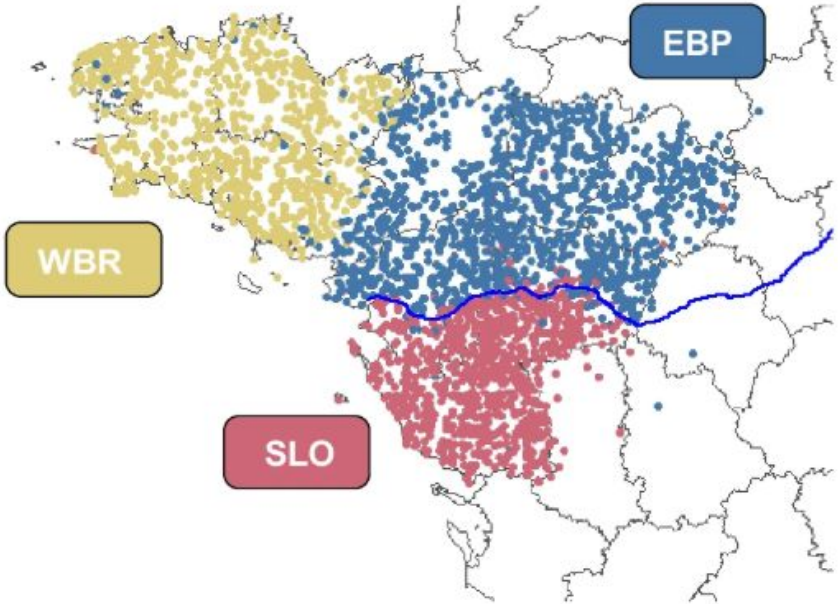


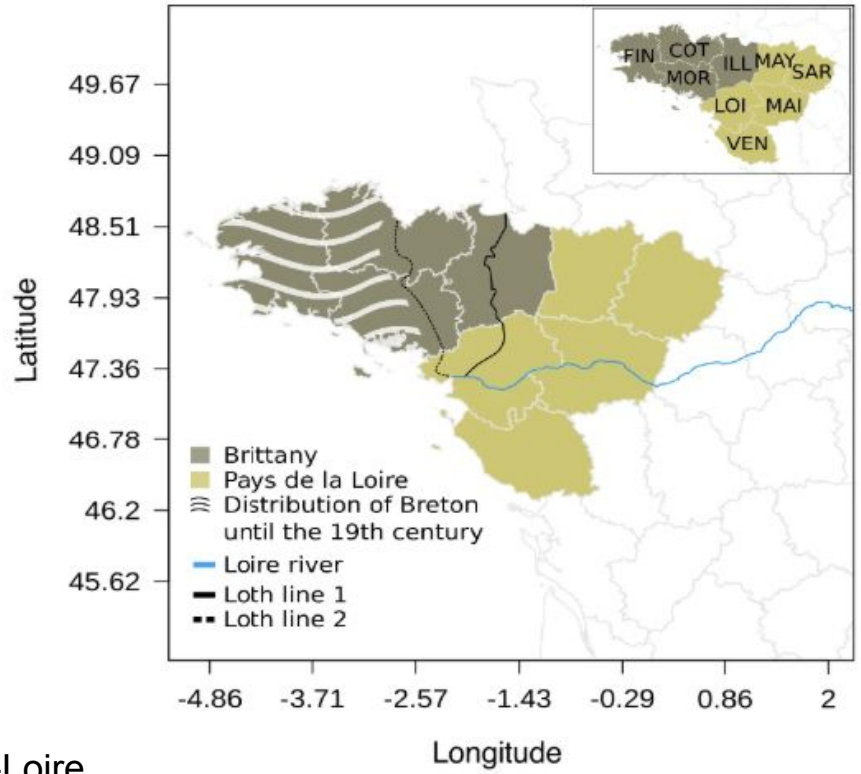
Principal Component Analysis (PCA)



Dataset of 197,146 SNPs per individual projected on 2-d space.

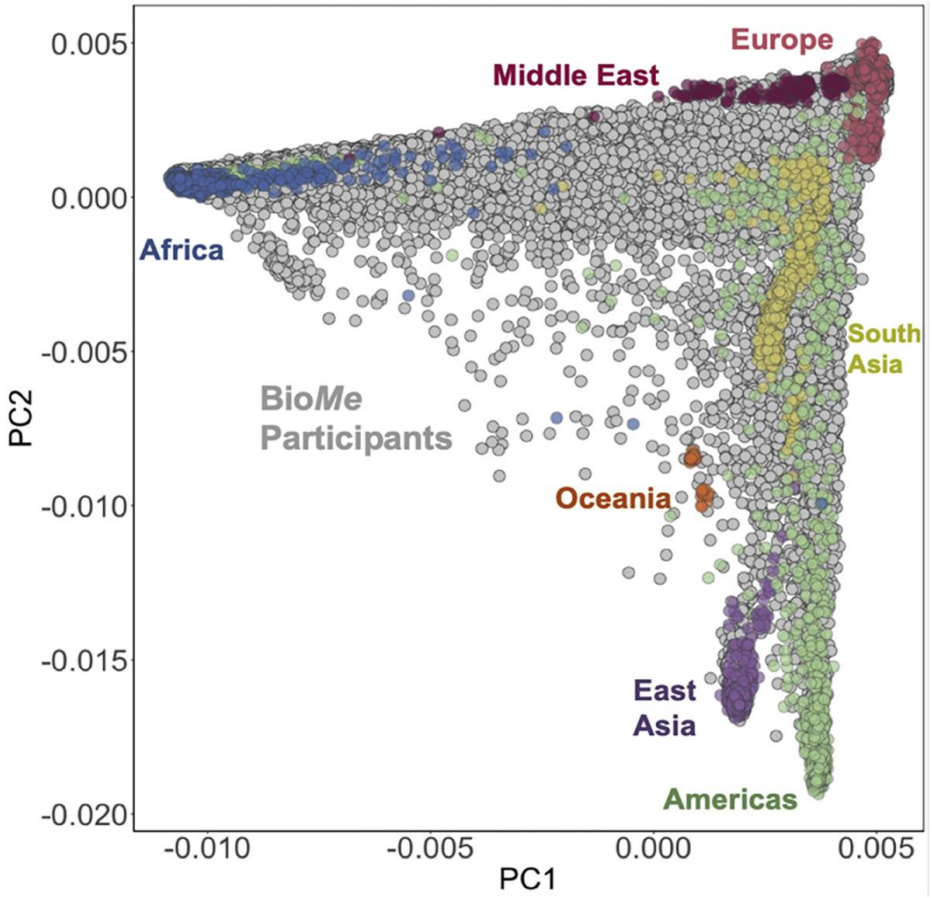
k-clustering





WBR: Western Brittany, EBP: Eastern Brittany and Pays-de-la-Loire, SLO:South Loire. 3234 present day individuals.
Alves et al. (2024)

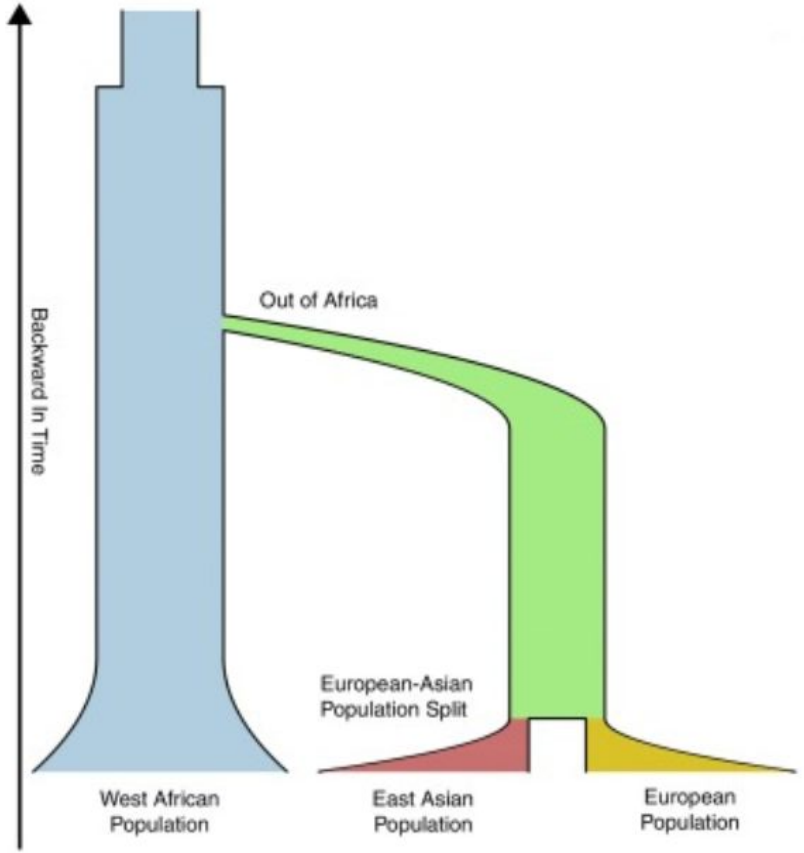
Population as strict categories?



Lewis et al. (2022)

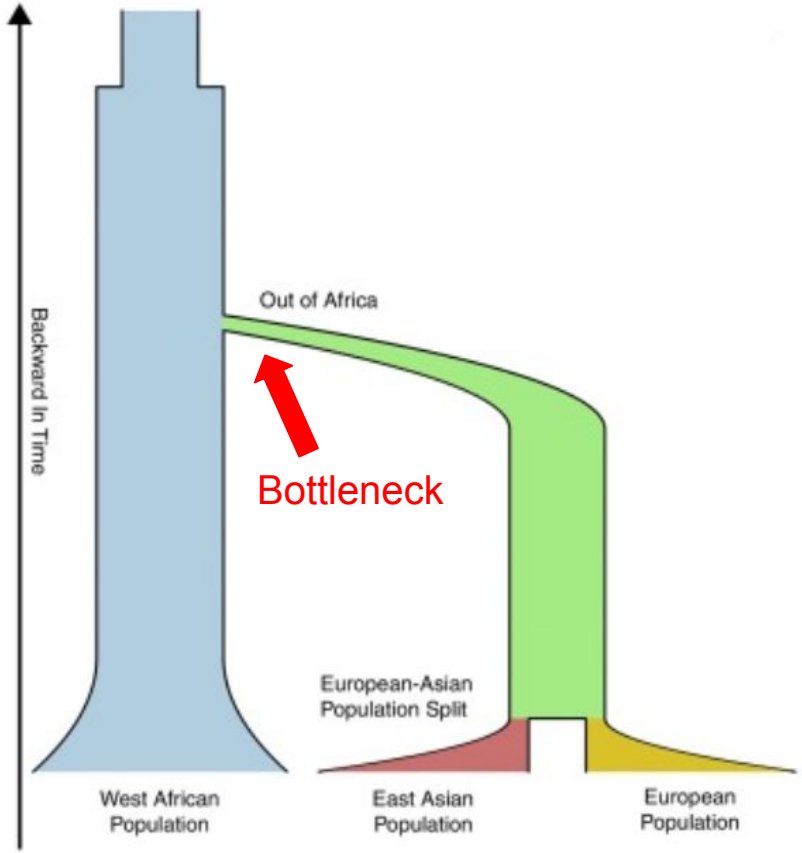
Population size

Population size



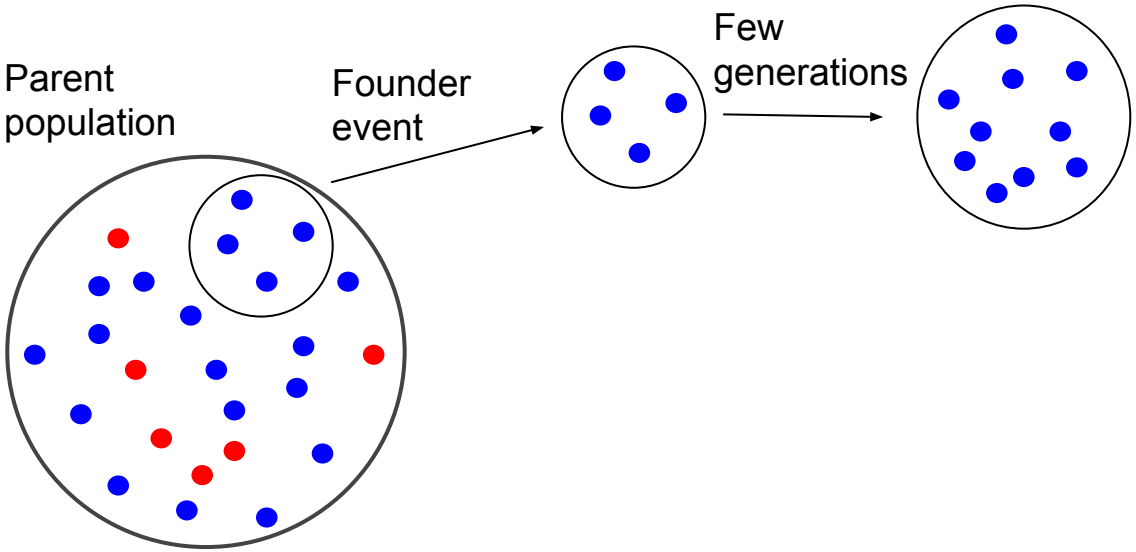
- Population sizes vary through time
- This can be inferred from present day genomes
- First idea:
  low genetic diversity ⇔ low population size

Population size: bottlenecks

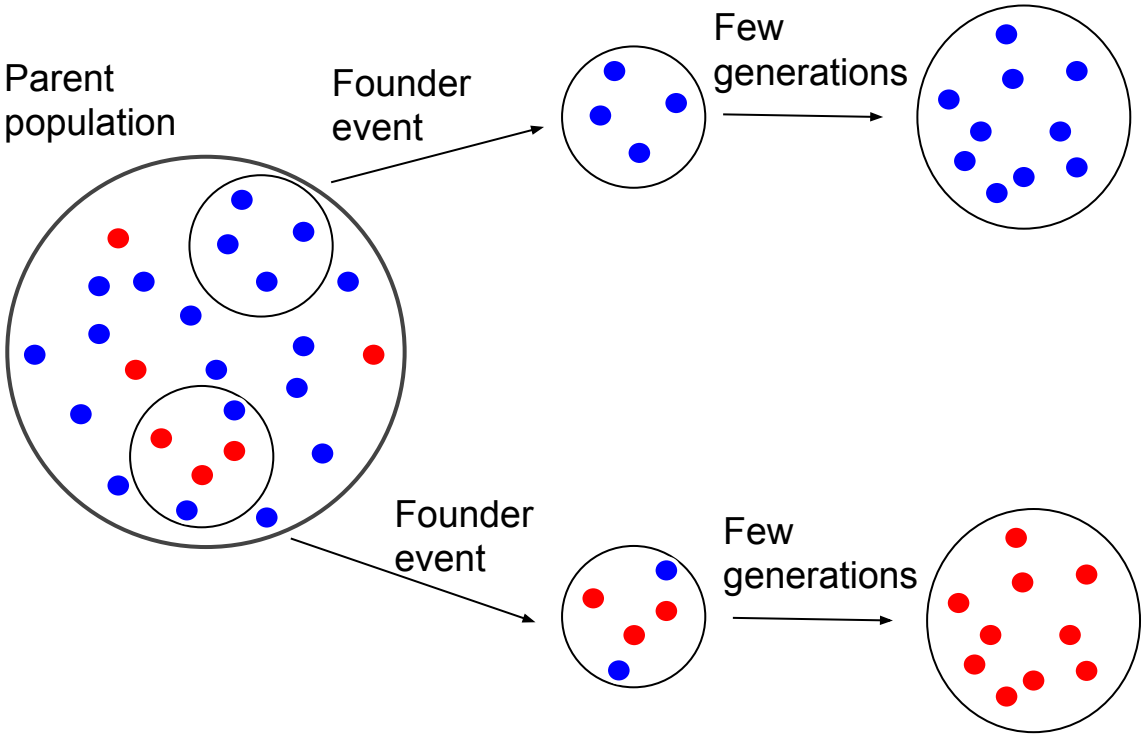


- A bottleneck is a large reduction in population size
- During such events, lots of variants are lost and genetic diversity suddenly decreases

Population size - bottlenecks and founder effect

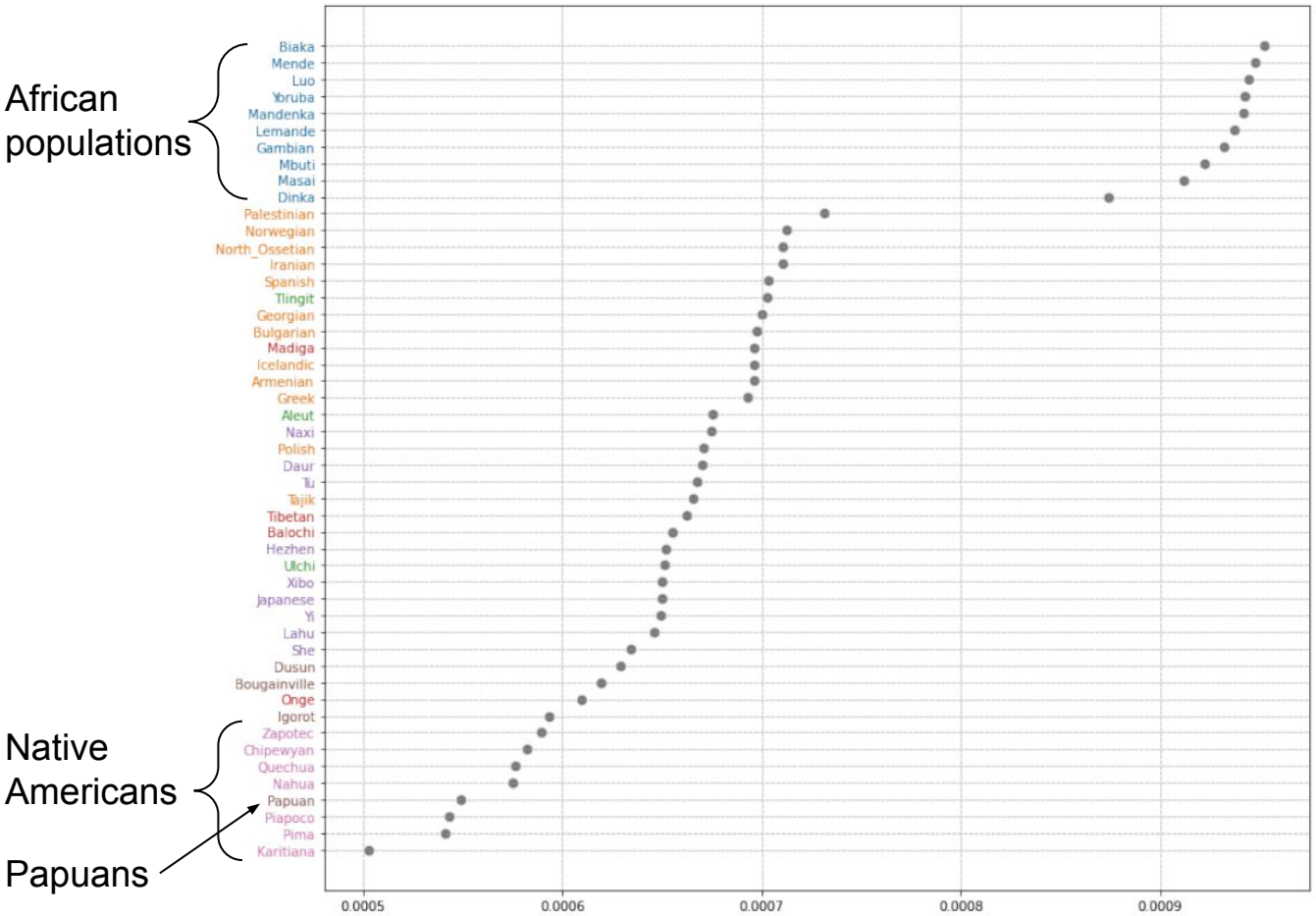


Population size - bottlenecks and founder effect



- A bottleneck can be caused by migrations (e.g. the Out of Africa event or the peopling of the Americas) or by environmental events.
- Due to the **founder effect**, a bottleneck will cause a loss of genetic diversity that cannot be recovered even if the population grows afterward.

Population size - bottlenecks consequences

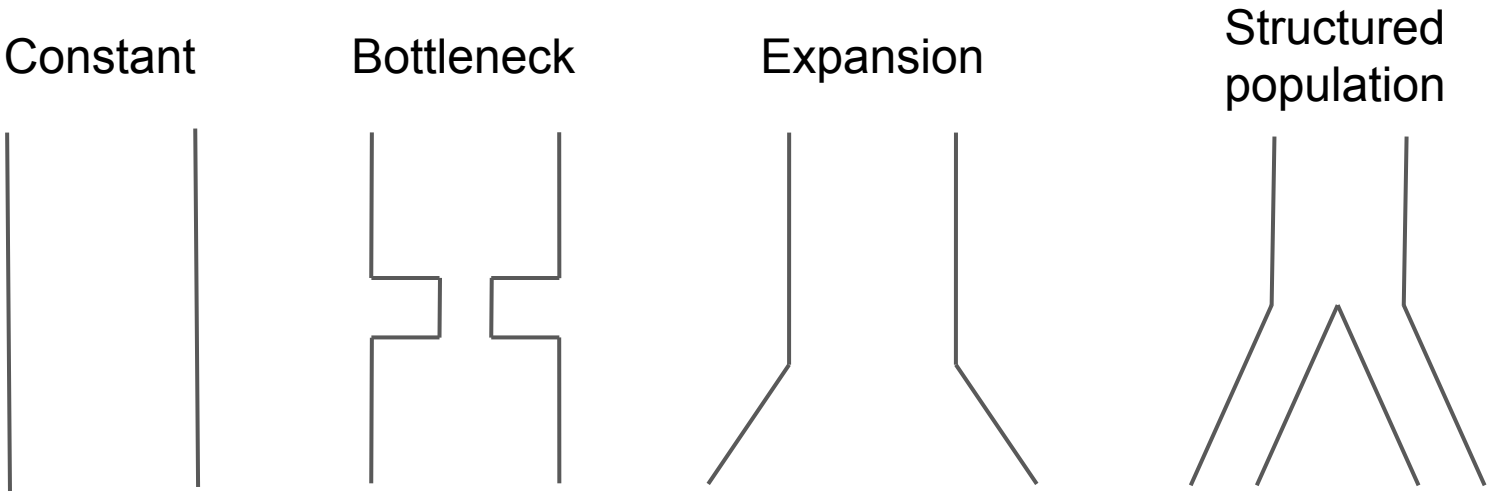


- African populations have the highest amount of genetic diversity.
- Native Americans have the least genetic diversity due to multiple consecutive bottlenecks.

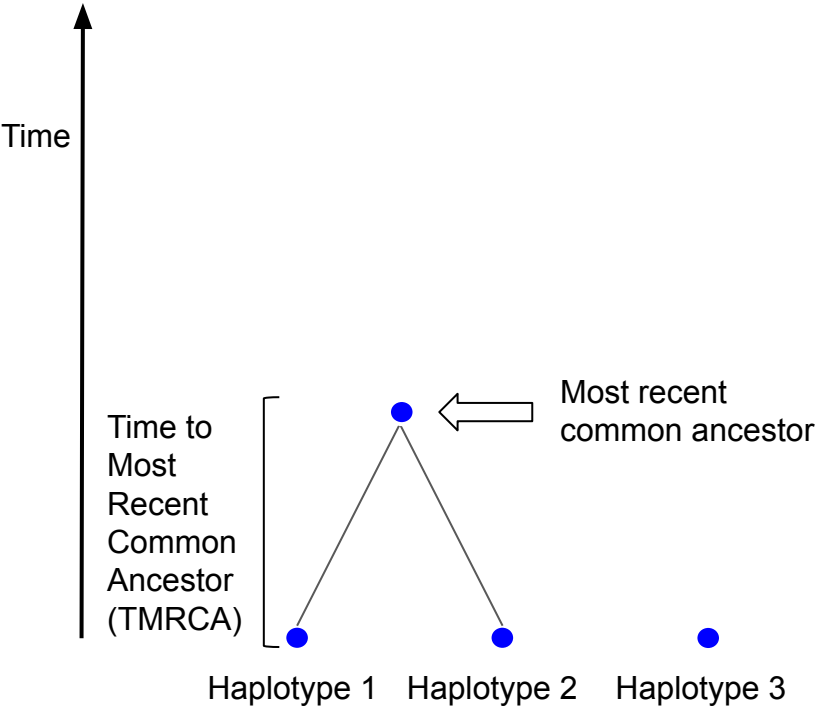
x-axis: heterozygosity proportion (higher -> more genetic diversity)

Inferring population size history

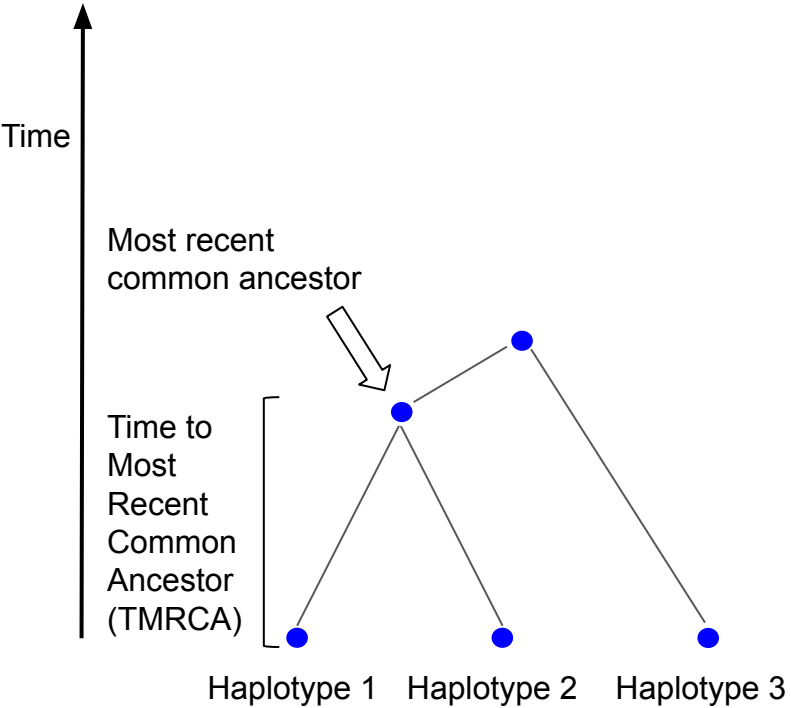
Models



Inferring population size - Coalescence trees



Inferring population size - Coalescence trees

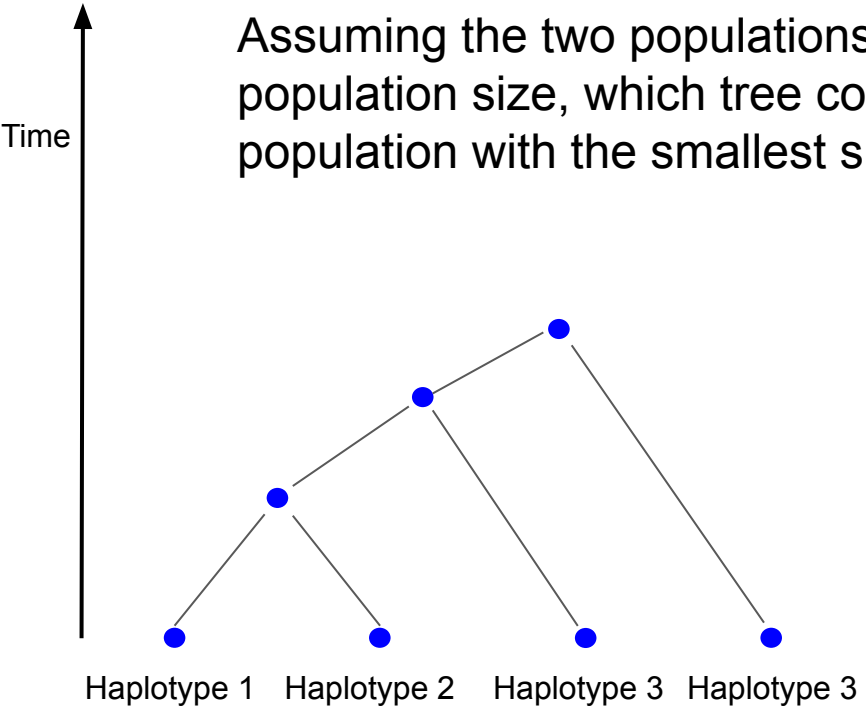


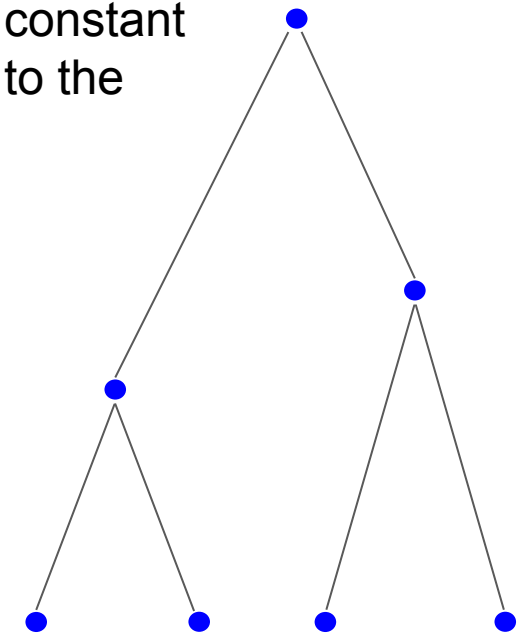
Some vocabulary:

- Any two haplotypes must come from the same ancestral haplotype at some point in the past.
- We say that the haplotypes **coalesce** (merge) together at time **TMRCA**.
- If we take multiple haplotypes, we can build a tree from this, it is called a **coalescence tree**.

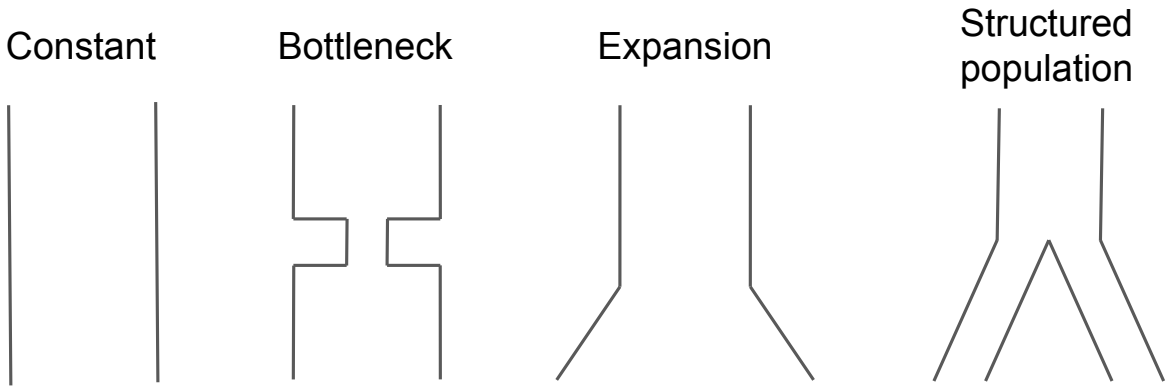
Inferring population size - Coalescence trees

Assuming the two populations are with constant population size, which tree correspond to the population with the smallest size ?

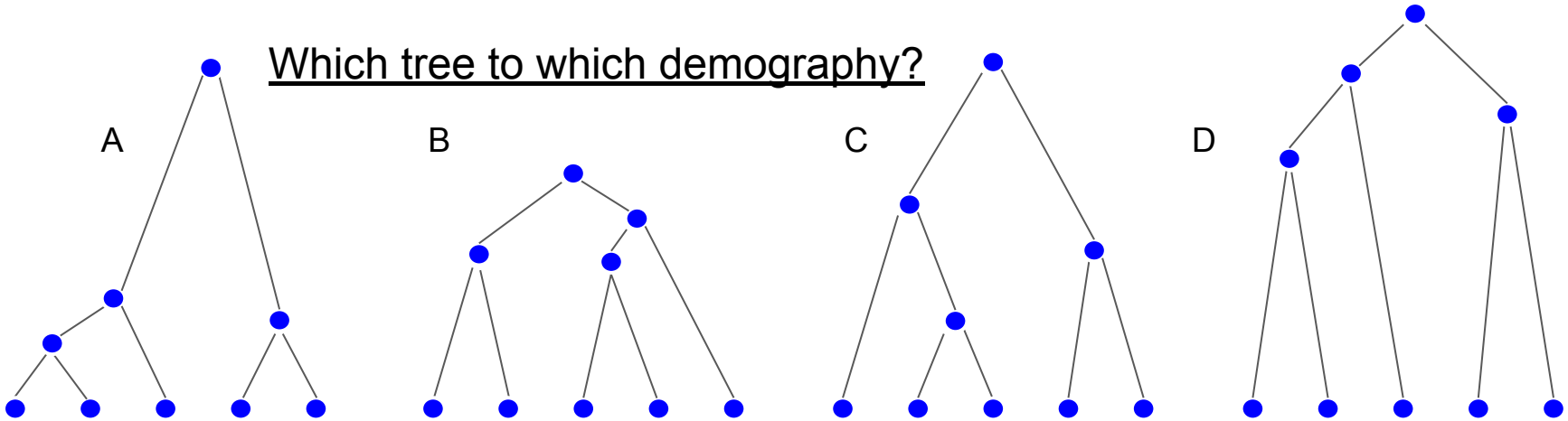




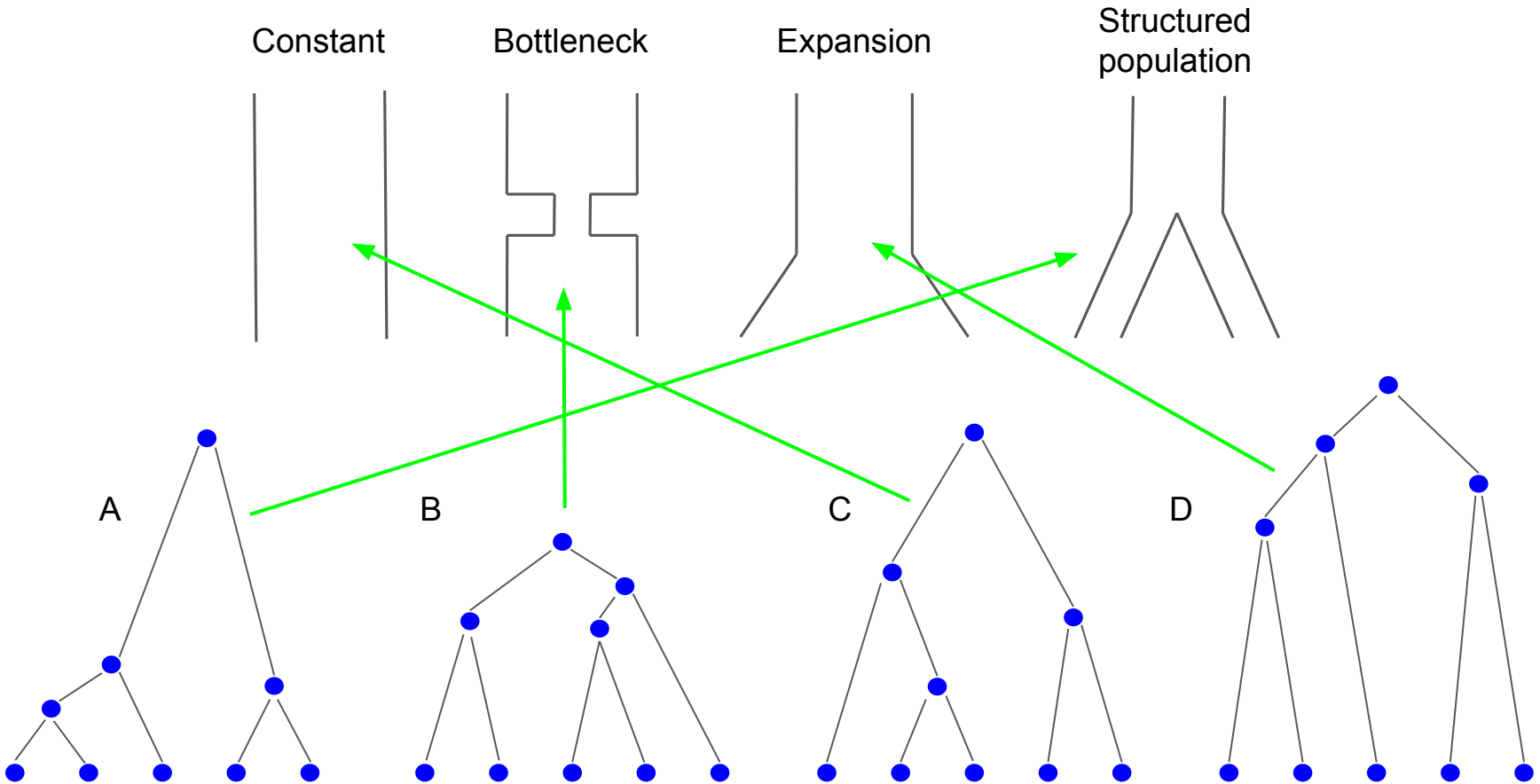
Inferring population size history



Which tree to which demography?



Inferring population size history



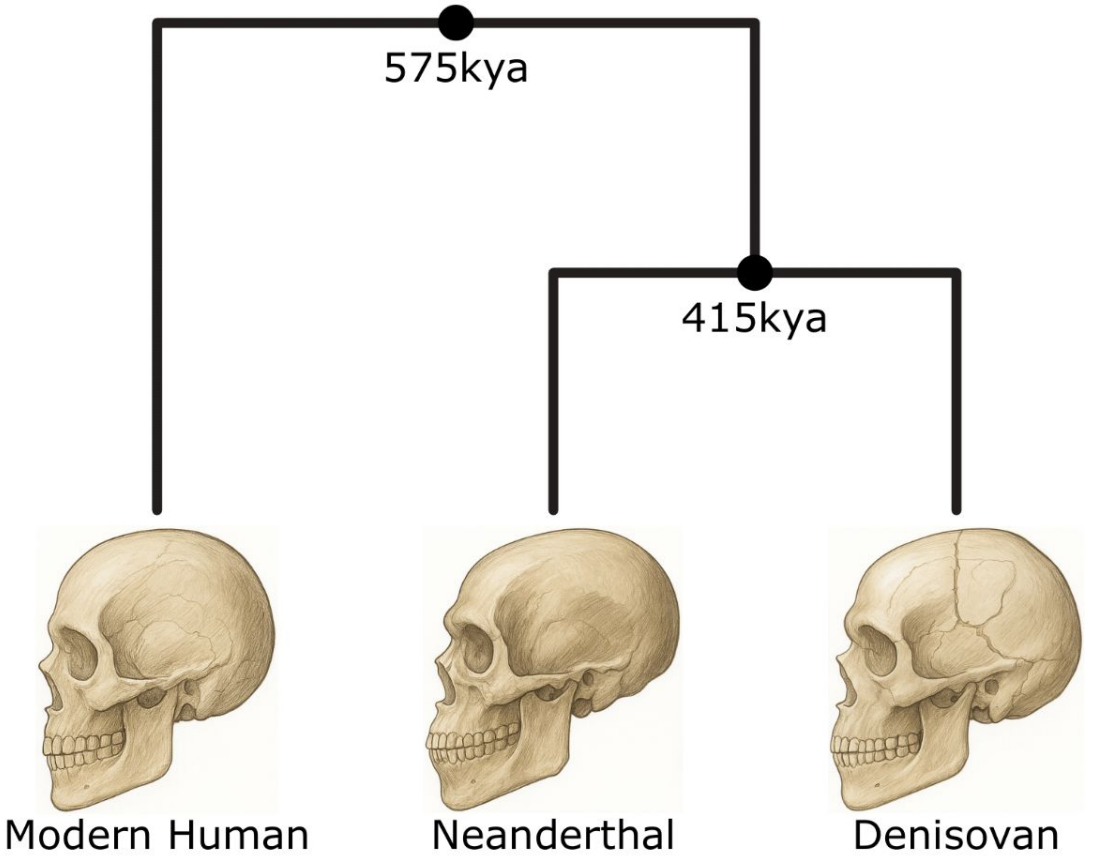
Inferring population size history

Conclusion:

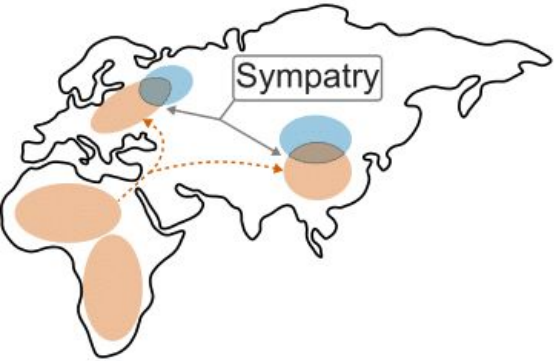
- If we are able to construct the coalescence tree from a population, we can know its whole population size history.
- Population size history can notably inform us about past migrations or ecological events.
- Still, how do we construct this coalescence tree from genomic data?
- The full answer is not for today! Key idea: If two individuals are closely genetically related, then they have a recent common ancestor and conversely.

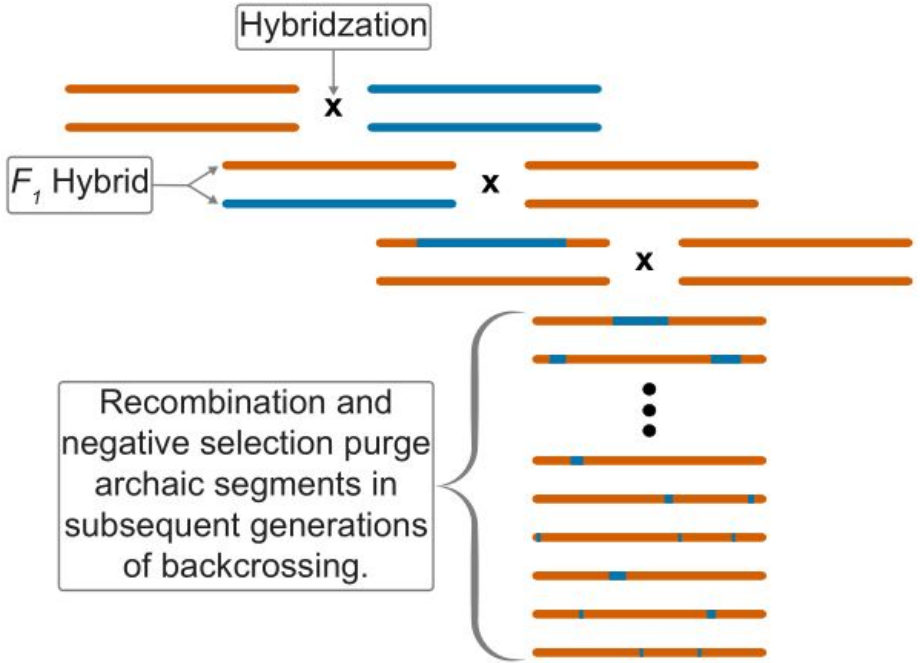
Archaic ancestry

Archaic Humans



Archaic Introgression



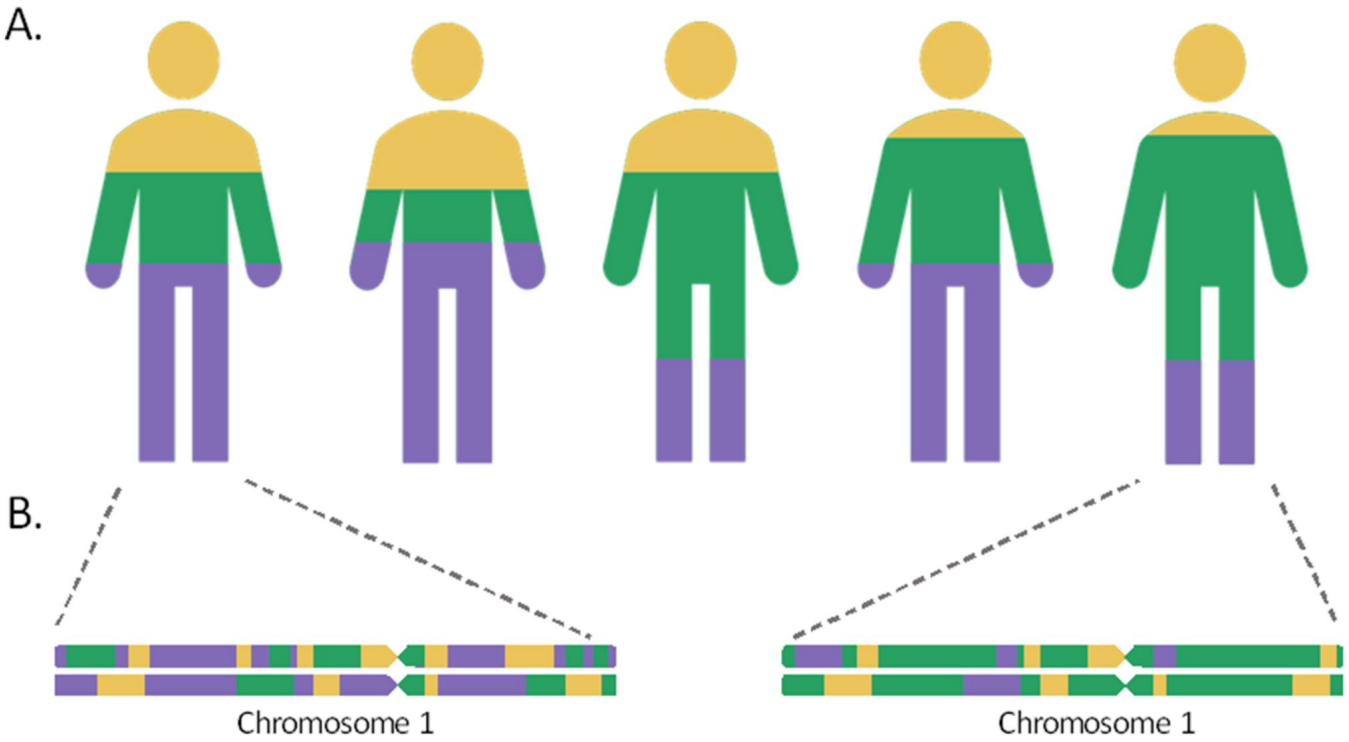


Archaic Introgression

Possible questions:

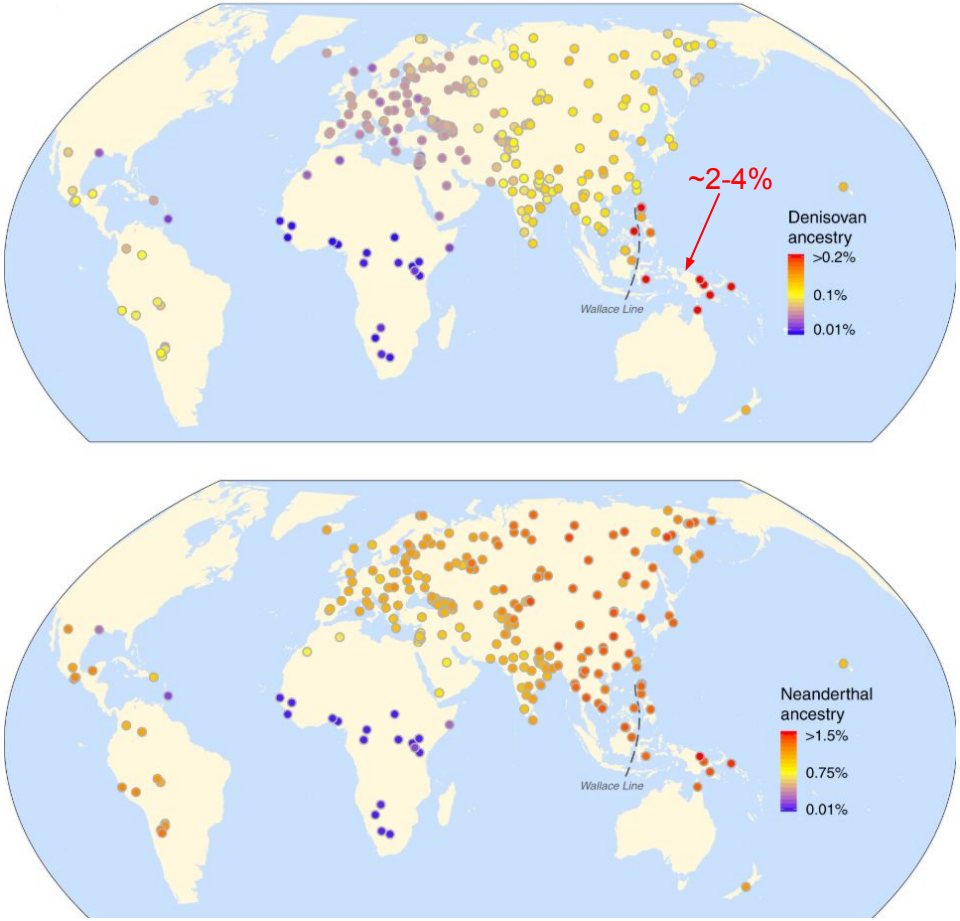
- Do we have archaic DNA, if so, how much and in which populations?
- Can we locate where archaic DNA is present in our genomes?
- Was there positive or negative selection for archaic DNA in Sapiens?
- If Archaic DNA is found in multiple contemporary populations (i.e Europe and Asia), does it have the same source?
- Can we date the introgression(s) event(s)?
- How did this archaic DNA spread trough time?

Global and Local ancestry inference



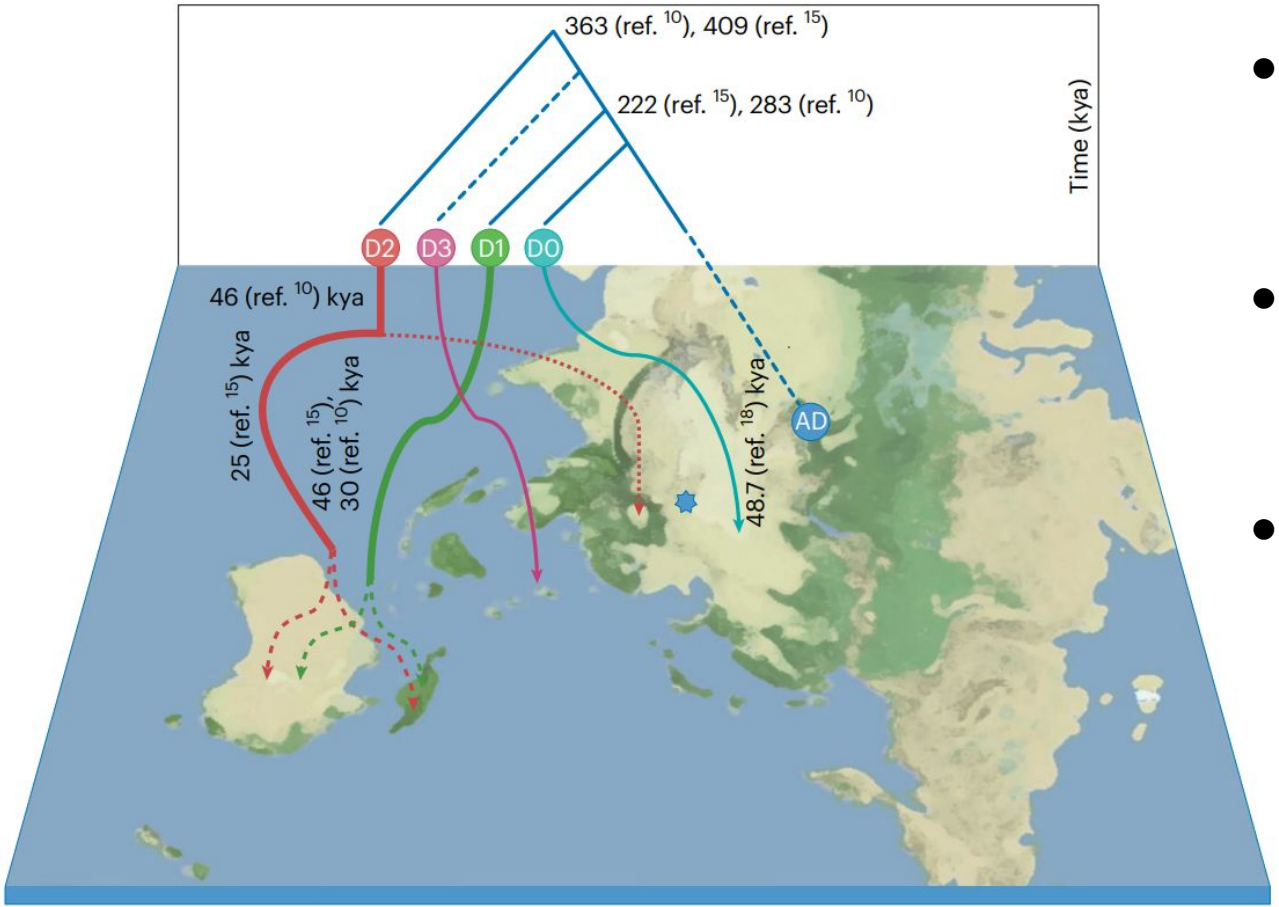
- A. Global ancestry inference.
- B. Local ancestry inference (LAI)

Archaic Introgression - Global ancestry



Yuan et al. (2021)

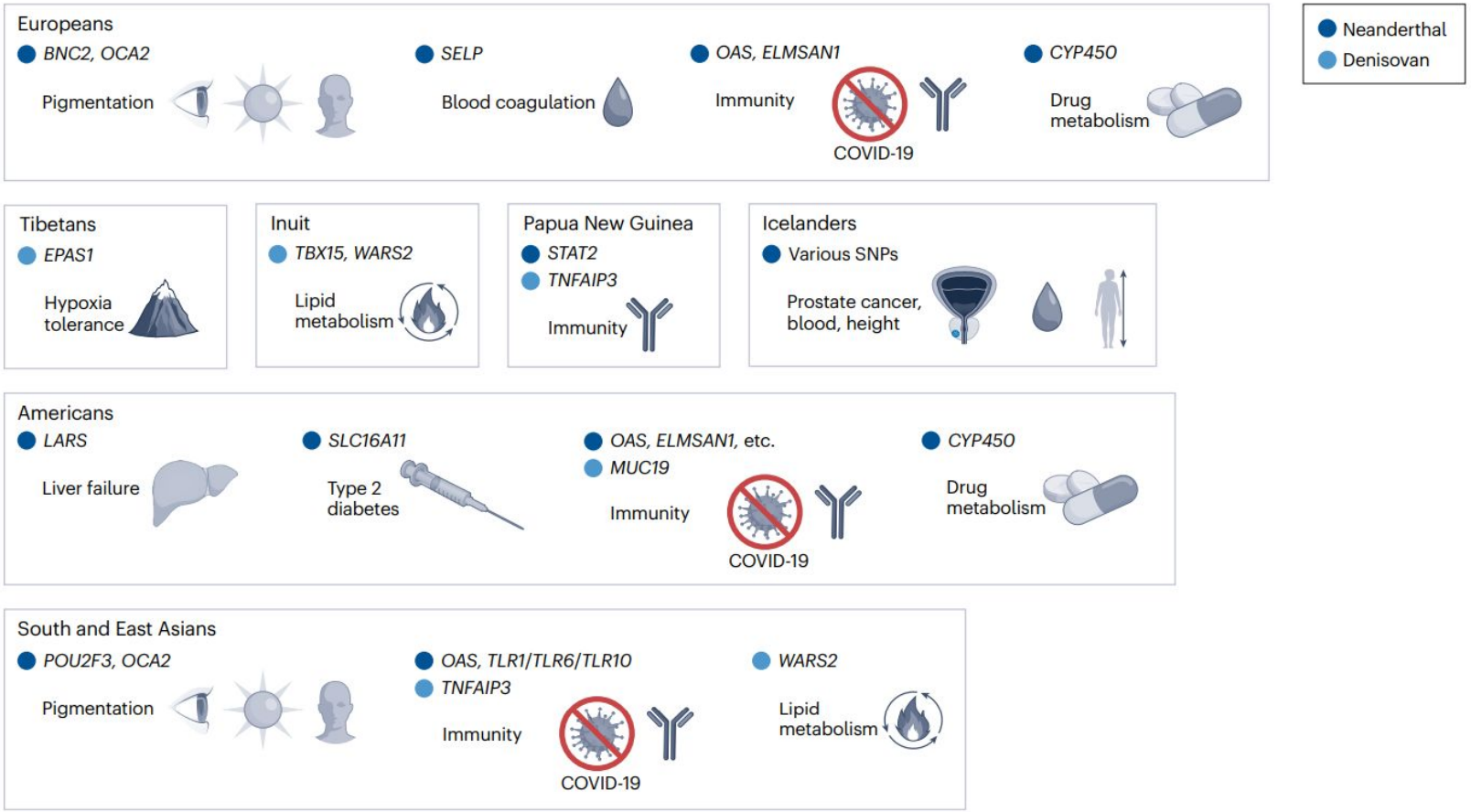
Archaic Introgression - Multiple sources



Ongaro et al. (2023)

- Most likely a single Neanderthal introgression event
- Multiple Denisovan introgression events, the number is still debated
- This can be deduced by comparing Denisovan segments in modern humans to the Denisovan reference genome

Archaic Introgression - Adaptively introgressed genes



Ongaro et al. (2023)