

Overview of current methods for population structure analysis

Journal club

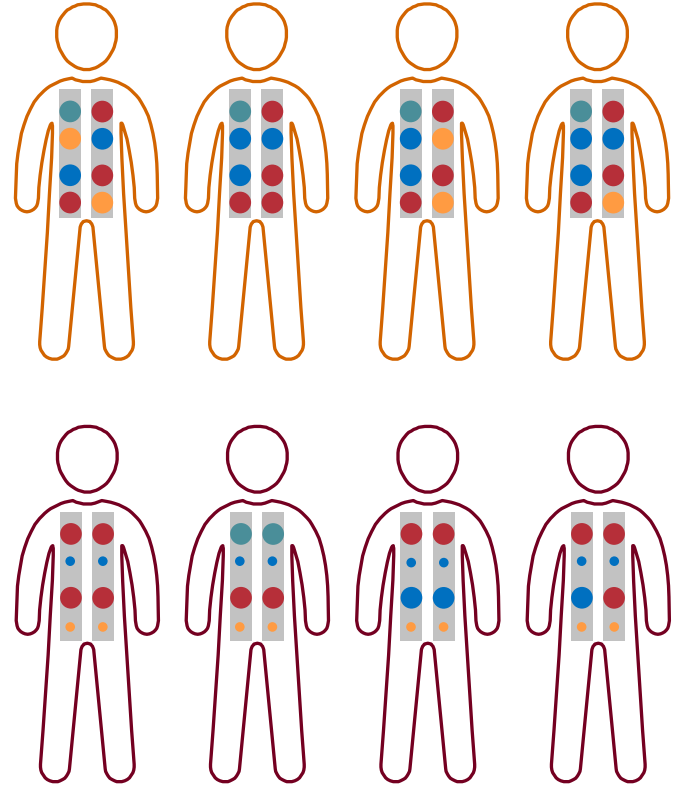
Imperial College
London



Human genetic diversity

Studying the human genetic diversity

Summarizing the genetic diversity within a population

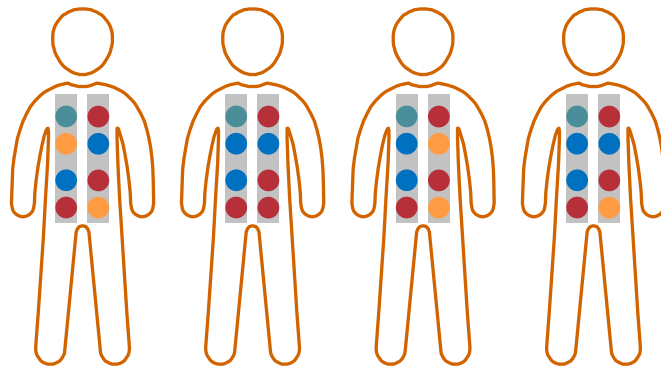


Summarizing the genetic diversity within a population

- **Watterson estimator**

Number of segregating sites

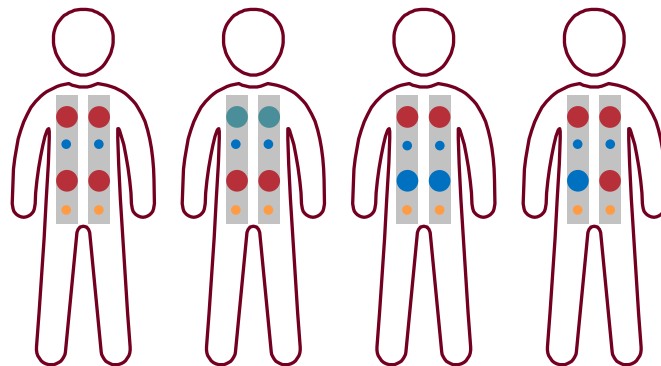
$$\hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$



- **Tajima's estimator**

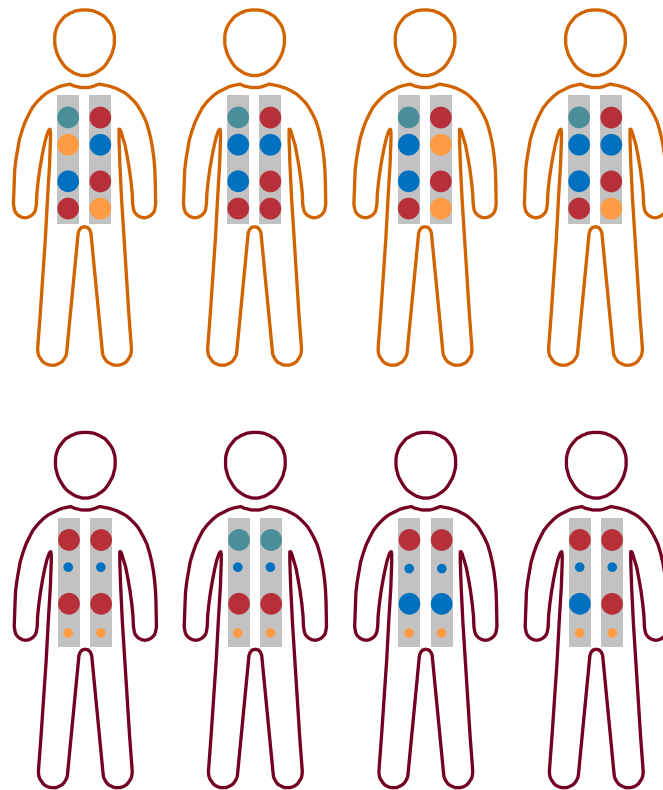
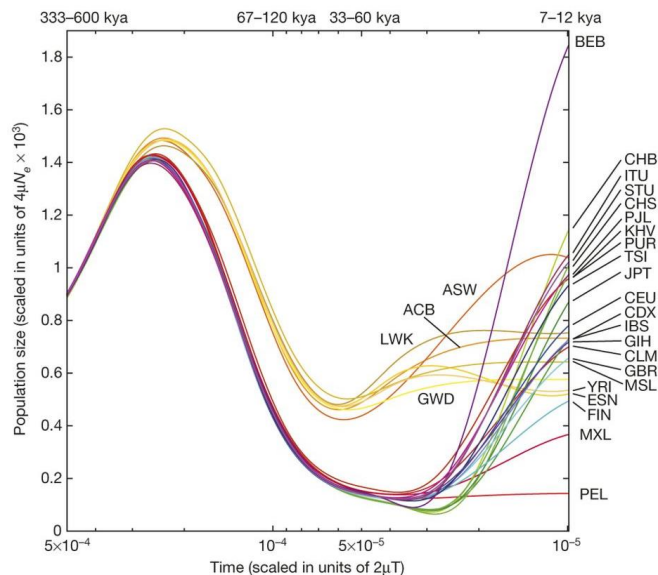
Nucleotide pairwise differences

$$\pi = \frac{\sum_{i < j} d_{i,j}}{n(n-1)/2}$$

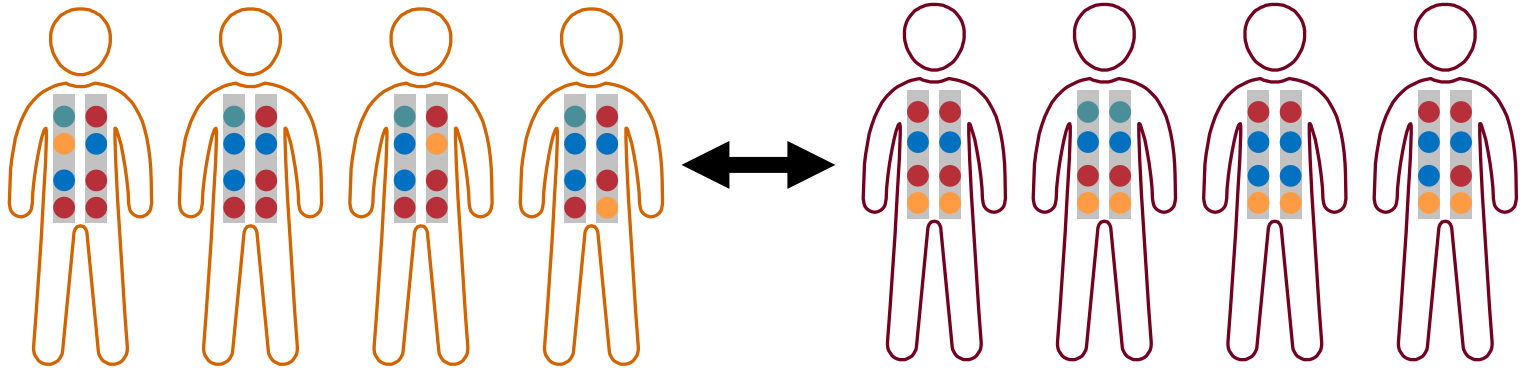


Summarizing the genetic diversity within a population

Effective population size (N_e)



Genetic distances between populations

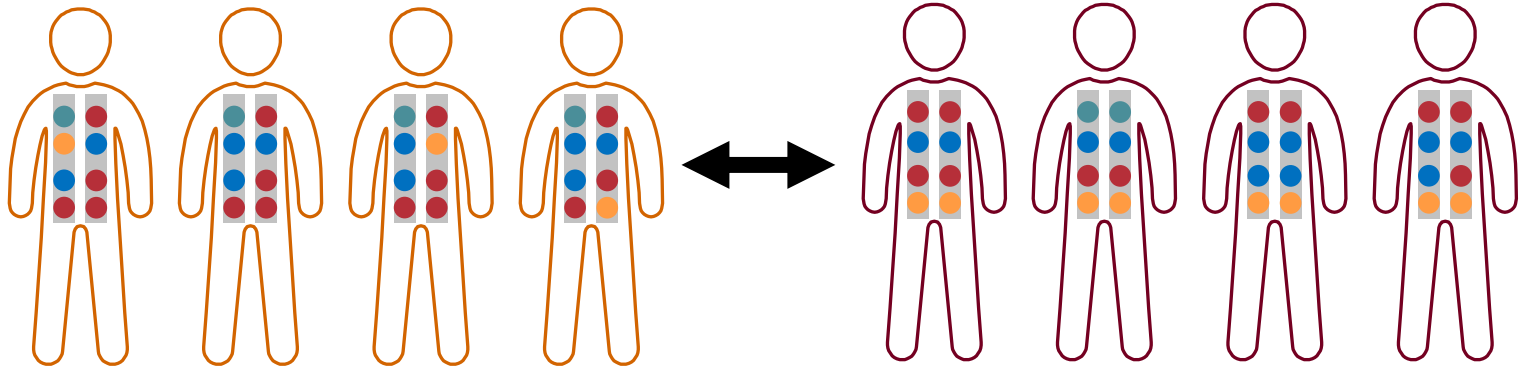


Genetic distances between populations

Allele frequency methods

- **Nei's D** $D = -\ln\left(\frac{J_{12}}{(\sqrt{J_{11}J_{22}})}\right)$

- **F_{ST}** $F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$

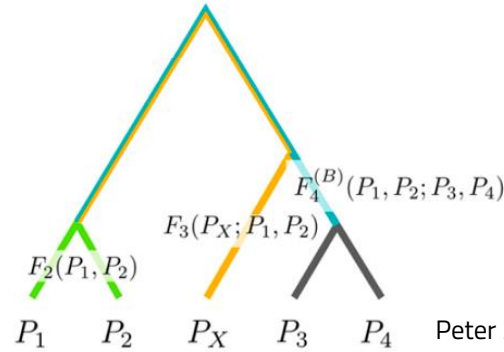


Genetic distances between populations

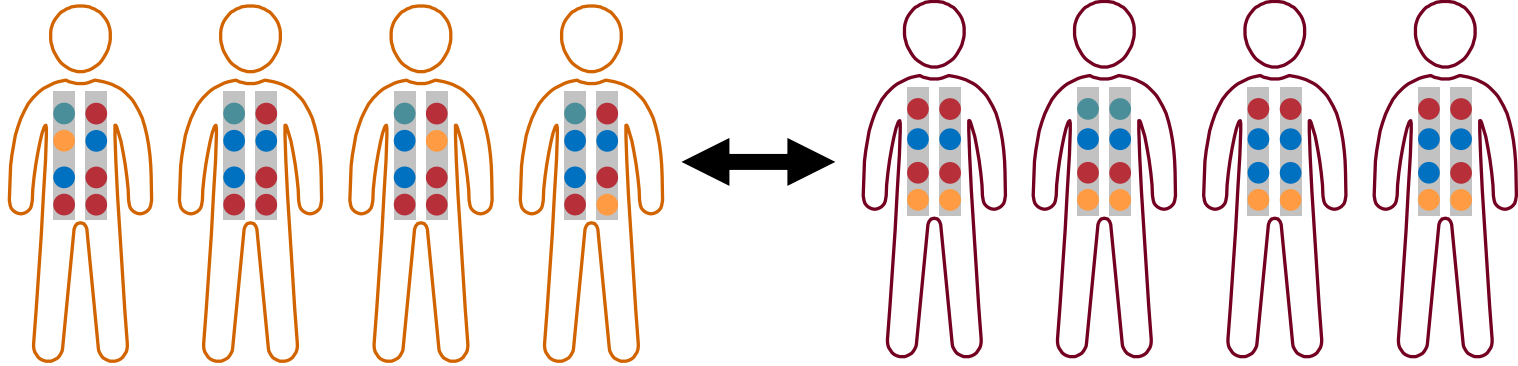
Allele frequency methods

F-statistics (Reich et al. 2009, Nature)

- F_2
- F_3
- F_4

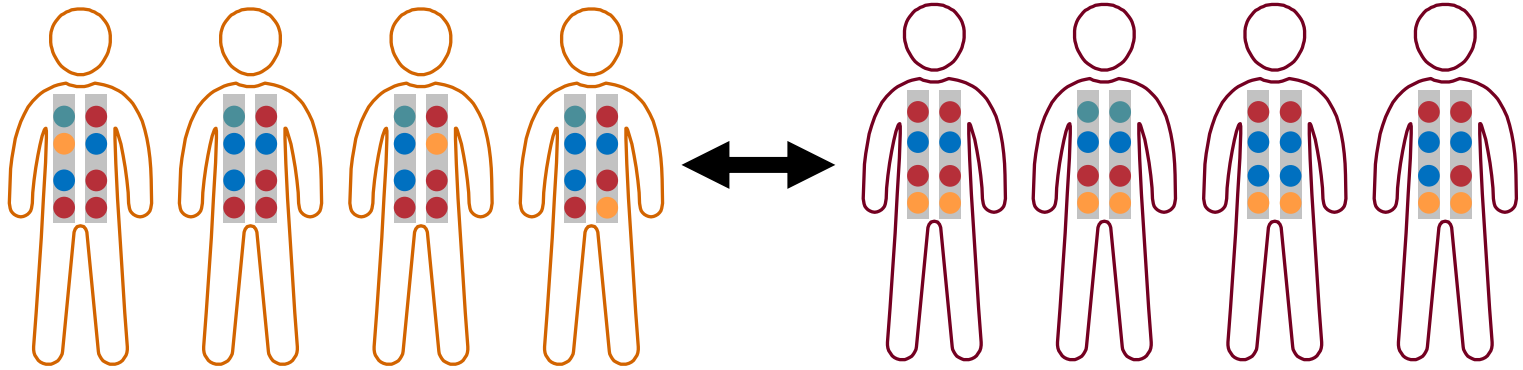


Peter 2016, Genetics



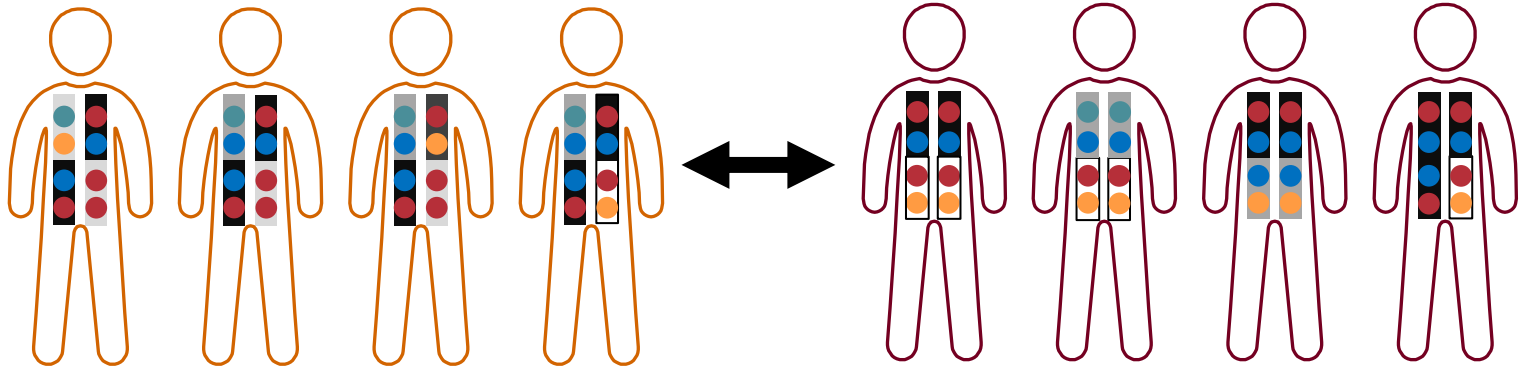
Genetic distances between populations

Haplotype-based methods



Genetic distances between populations

Haplotype-based methods

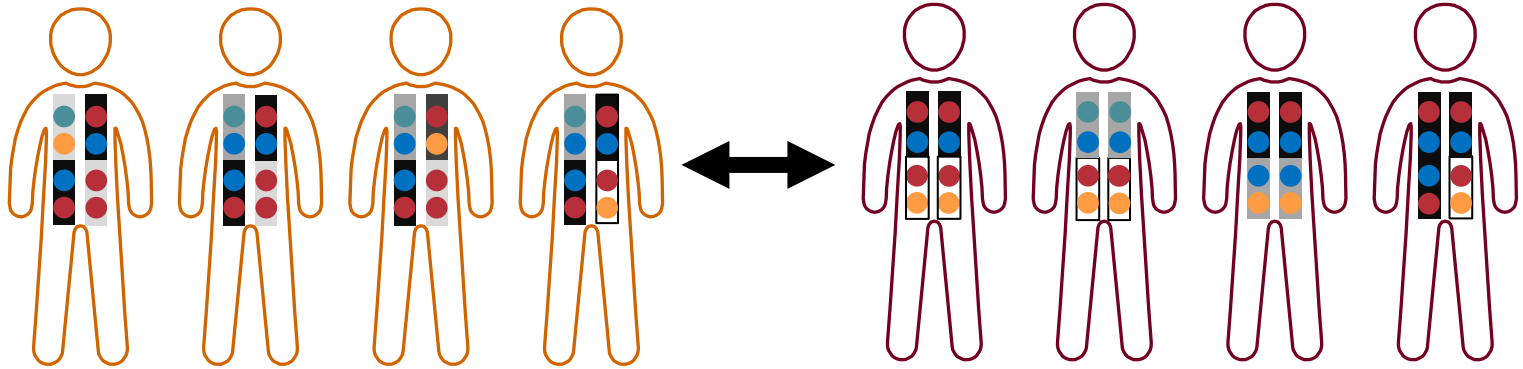


Genetic distances between populations

Haplotype-based methods

- ChromoPainter

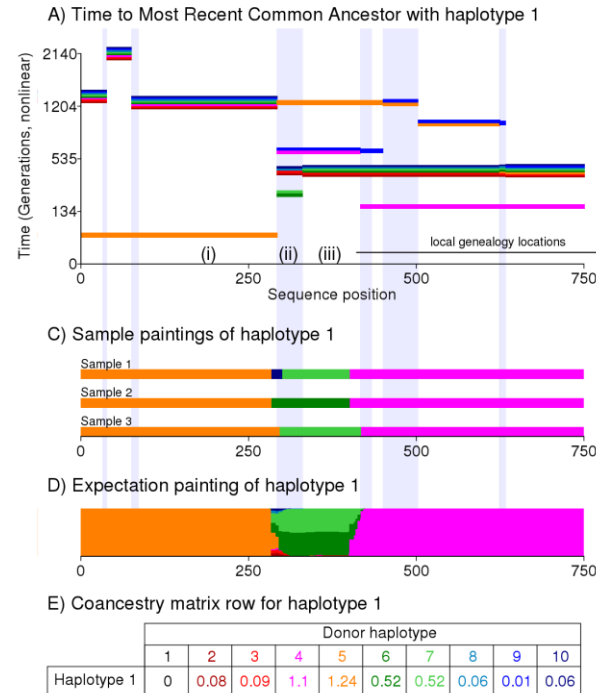
Lawson et al. 2012, PLoS Genet



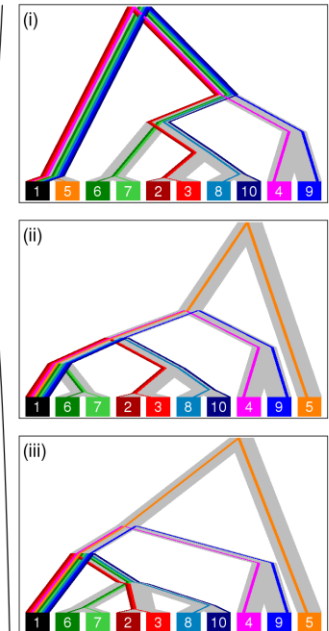
Genetic distances between populations

Haplotype-based methods

- ChromoPainter
Lawson et al. 2012, PLoS Genet
- Search for the most common recent ancestor of each individual haplotype within the haplotypes of the other individuals of the dataset



B) Local genealogies

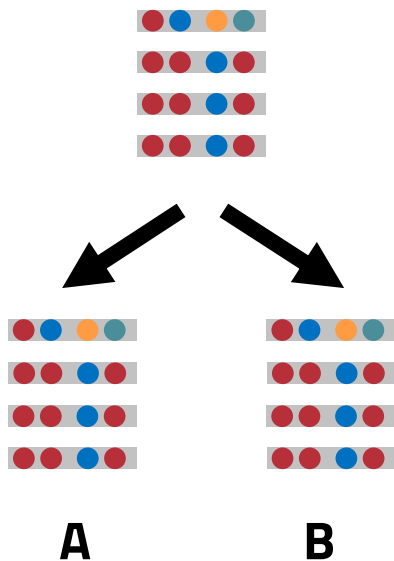


Genetic distances between populations

Haplotype-based methods

- ChromoPainter

Lawson et al. 2012, PLoS Genet

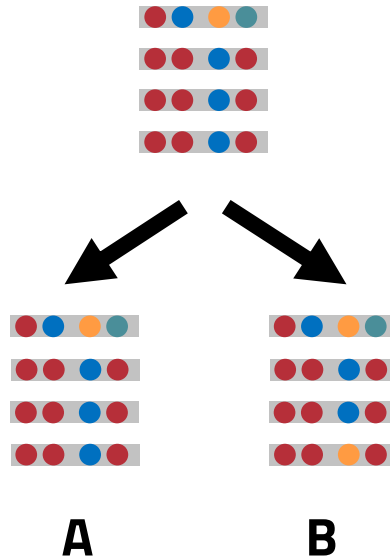


Genetic distances between populations

Haplotype-based methods

- ChromoPainter

Lawson et al. 2012, PLoS Genet

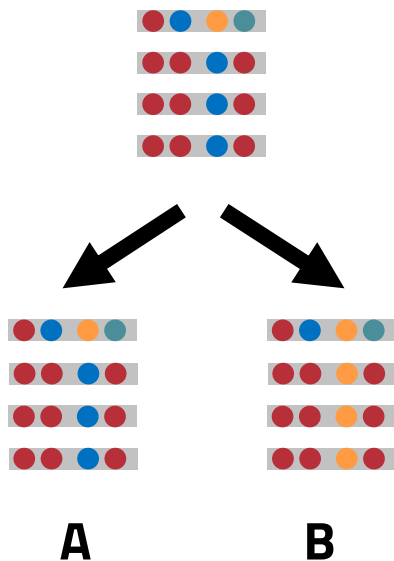


Genetic distances between populations

Haplotype-based methods

- ChromoPainter

Lawson et al. 2012, PLoS Genet

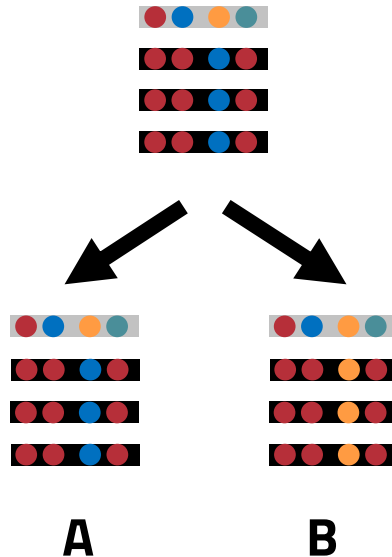


Genetic distances between populations

Haplotype-based methods

- ChromoPainter

Lawson et al. 2012, PLoS Genet



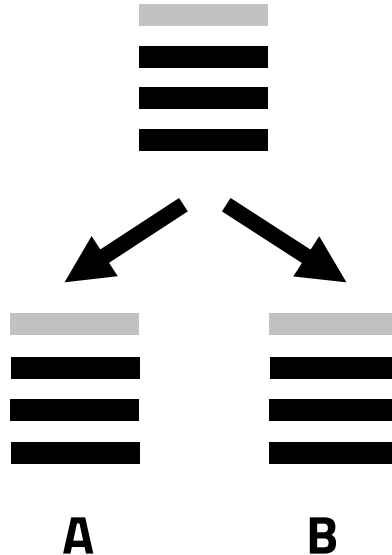
Analyzing the context of the snps, **the haplotypes**, we can smooth the weight of increased allele frequencies after genetic drift and **relate the populations through shared haplotypes**

Genetic distances between populations

Haplotype-based methods

- ChromoPainter

Lawson et al. 2012, PLoS Genet



Analyzing the context of the snps, **the haplotypes**, we can smooth the weight of increased allele frequencies after genetic drift and **relate the populations through shared haplotypes**



Genetic distances between populations

Population structure through clustering methods



Genetic distances between populations

Population structure through clustering methods

From allele frequencies

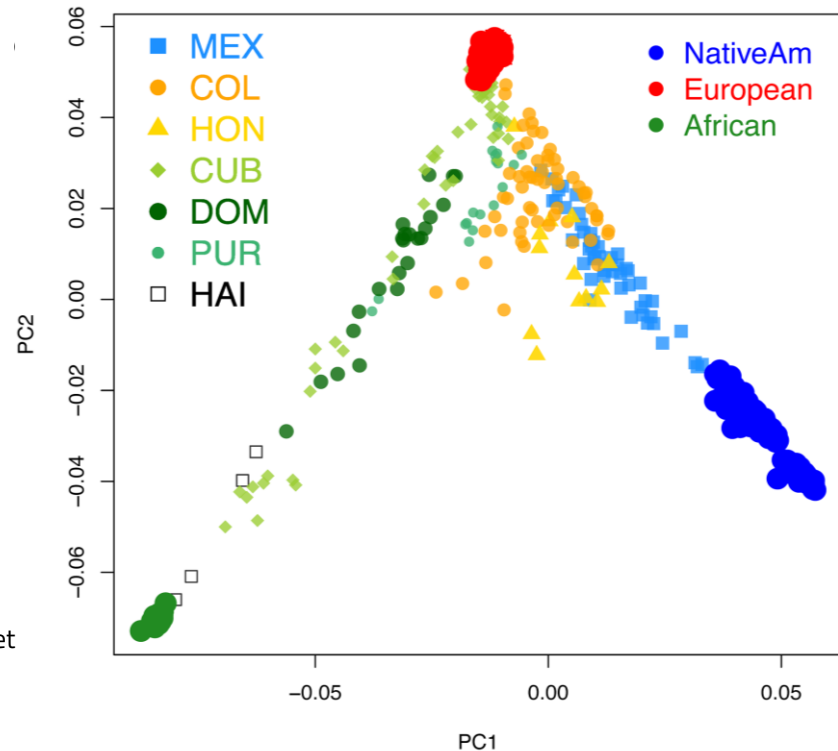
Genetic distances between populations

Population structure through clustering methods

From allele frequencies

- **Principal Component Analysis**
 - Transforms allele frequencies to a set of linearly uncorrelated variables called principal components.
 - The visualization of the individuals as points of two principal component coordinates clusters them based on their genetic distances.

Moreno-Estrada et al. 2013, PLoS Genet



Genetic distances between populations

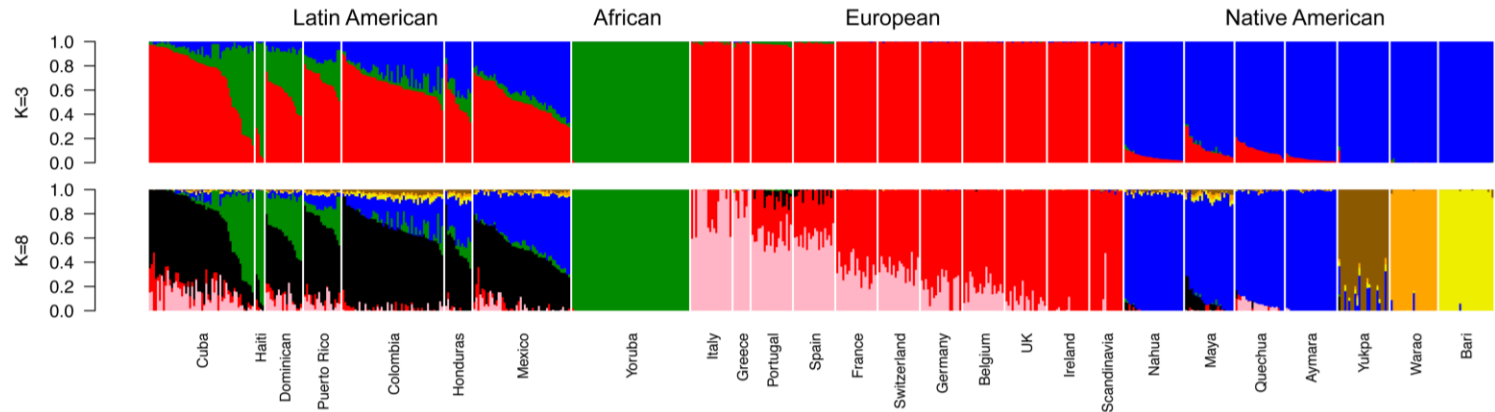
Population structure through clustering methods

From allele frequencies

■ Admixture

Alexander et al. 2019, Genome Res

Moreno et al. 2013, PLoS Genet



- analyses differences in the distribution of allele frequencies amongst individuals with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation



Genetic distances between populations

Population structure through clustering methods

From haplotype-based methods

Genetic distances between populations

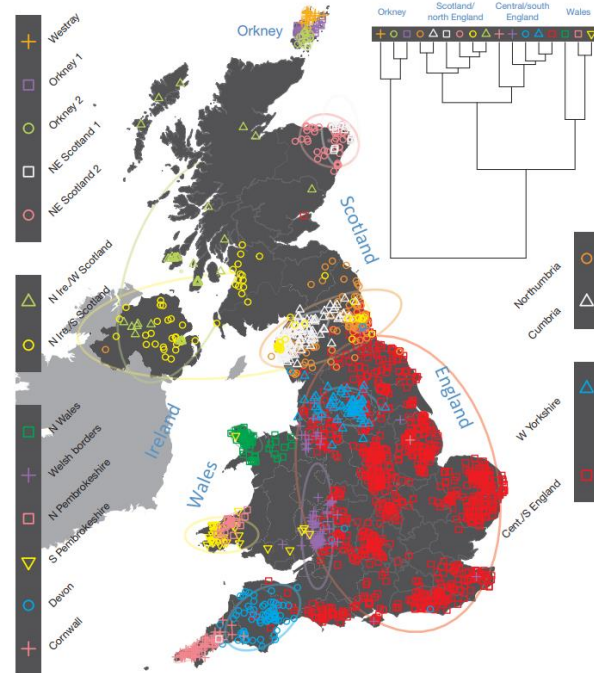
Population structure through clustering methods

From haplotype-based methods

- **FineStructure**

Lawson et al. 2012, PLoS Genet

- ▣ Clusters the individuals based on their **haplotypic similarities** computed in the ChromoPainter coancestry matrix

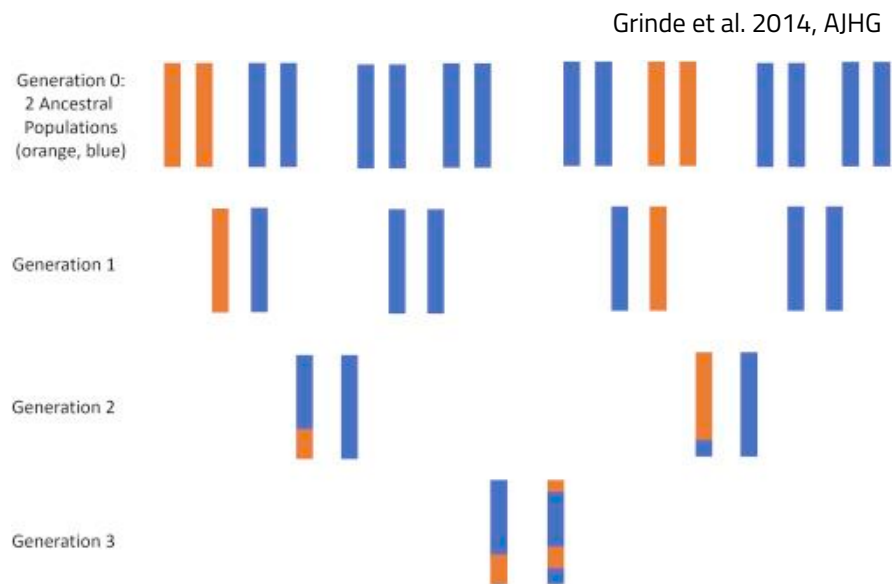
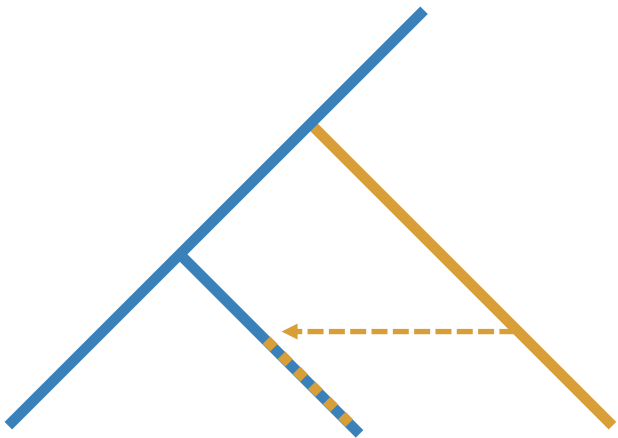


Leslie et al. 2015, Nature



Inference of admixture between populations

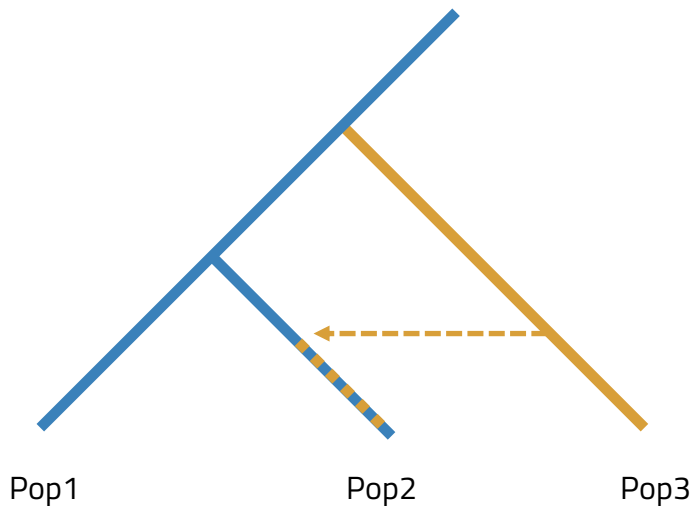
Inference of admixture between populations



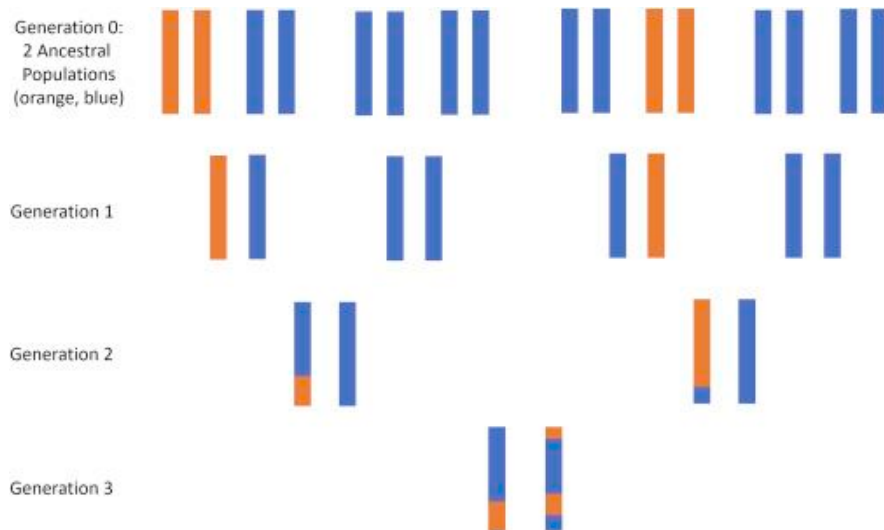
Inference of admixture between populations

From allele frequencies

■ F_3



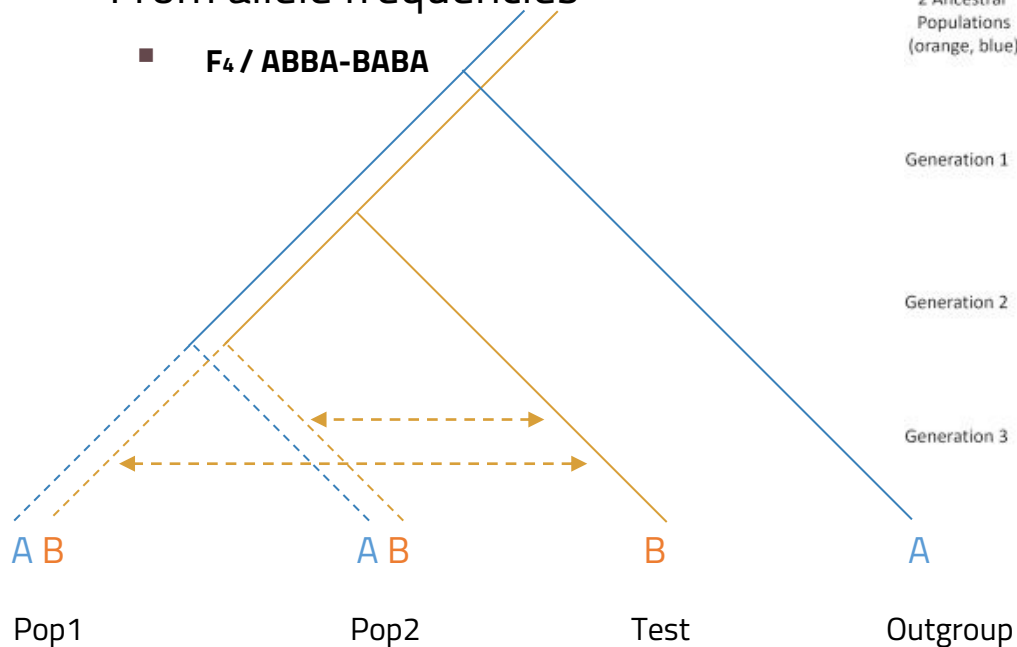
$$F_3(\text{Pop2}; \text{Pop1}, \text{Pop3}) < 0$$



Inference of admixture between populations

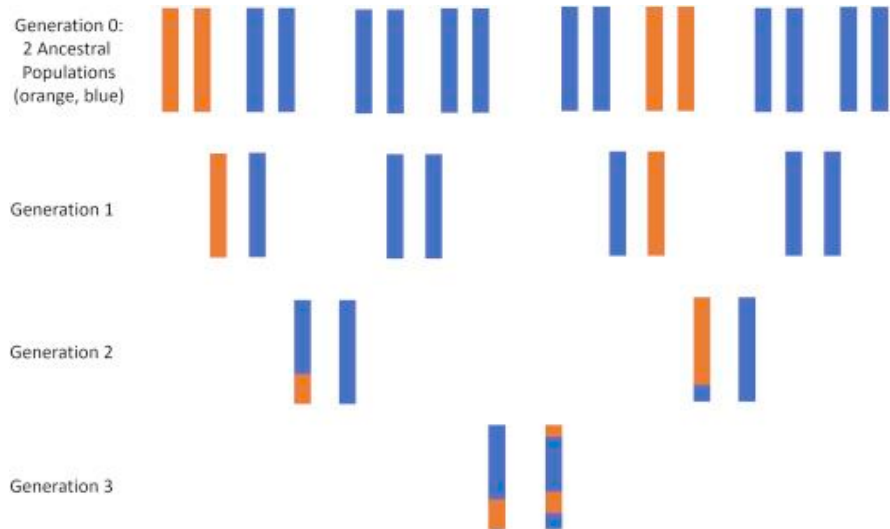
From allele frequencies

■ F_4 / ABBA-BABA



$F_4 > 0$: Admixture in Pop2

$F_4 < 0$: Admixture in Pop1



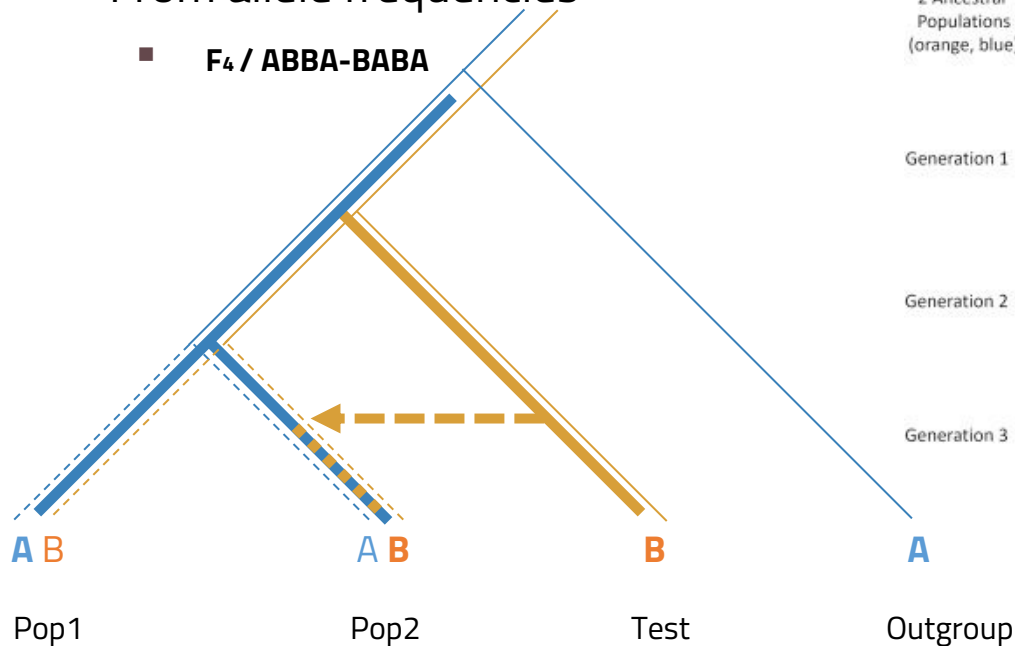
Alleles for a given SNP

Population

Inference of admixture between populations

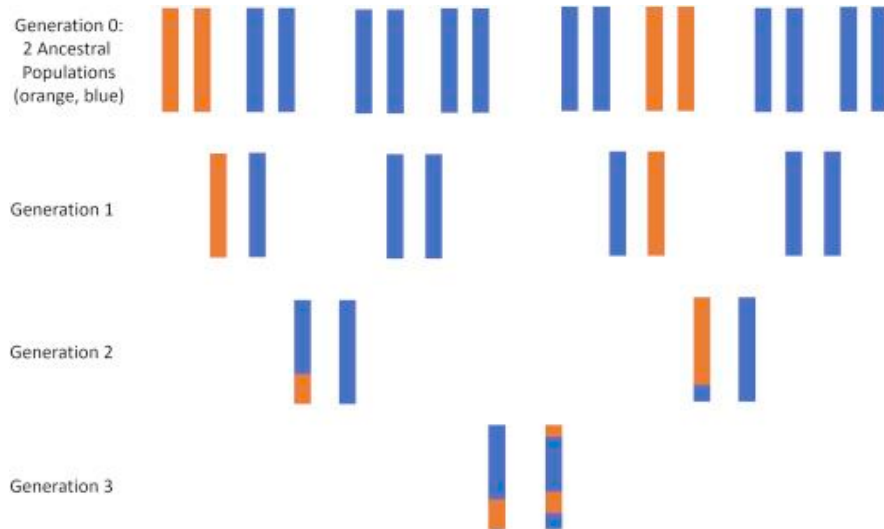
From allele frequencies

■ F_4 / ABBA-BABA



$F_4 > 0$: Admixture in Pop2

$F_4 < 0$: Admixture in Pop1



Alleles for a given SNP

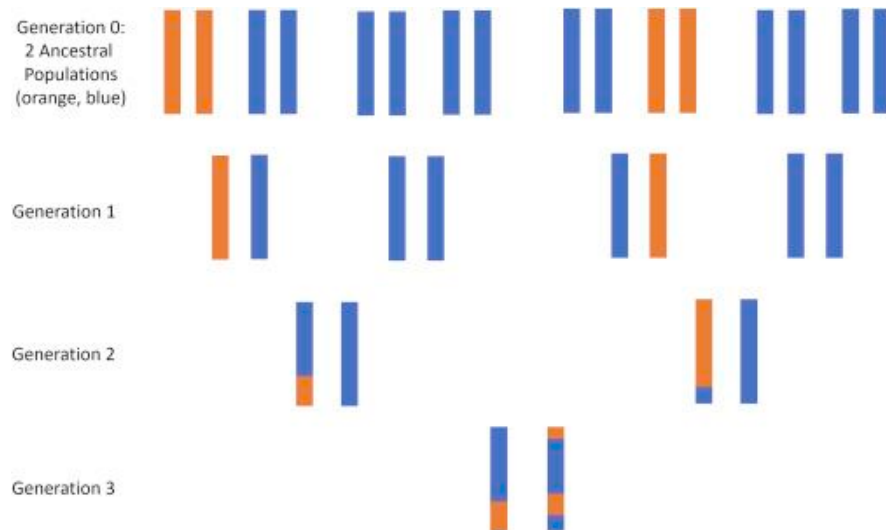
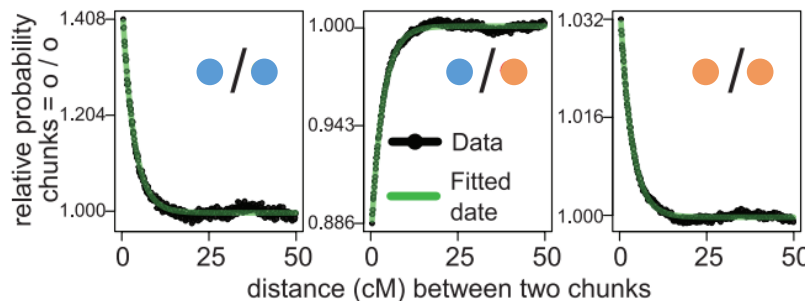
Population

Inference of admixture between populations

From haplotype-based methods

■ Globetrotter

Hellenthal et al. 2014, Science



Generation 4 Longer fragments, longer distance between fragments



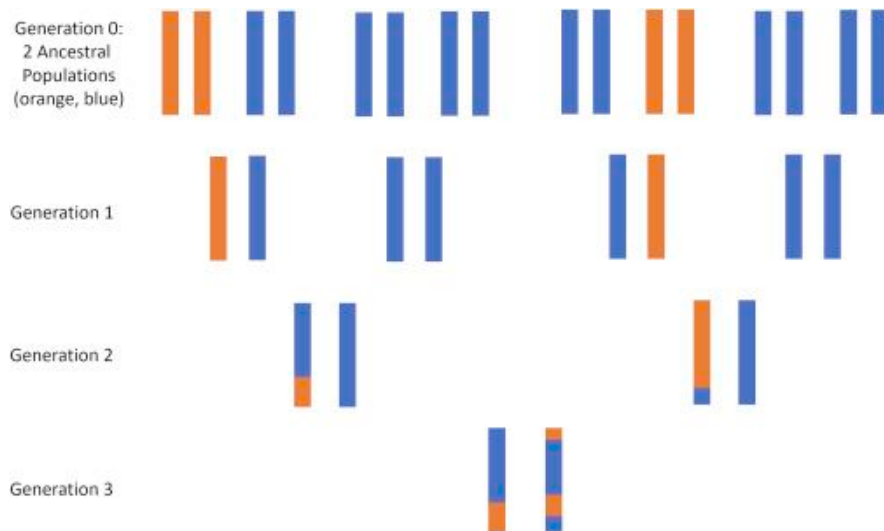
Generation 8 Shorter fragments, shorter distance between fragments



Inference of admixture between populations

- Other methods using similar approaches

- Tracts
 - Gravel et al. 2012, Genetics
- ALDER
 - Loh et al. 2013, Genetics



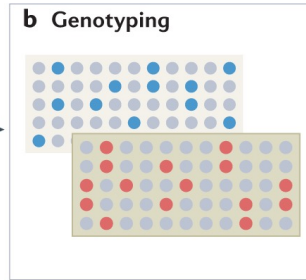
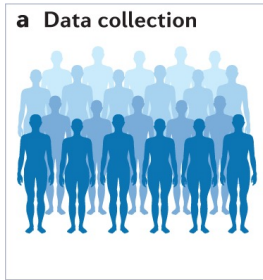
Generation 4 Longer fragments, longer distance between fragments



Generation 8 Shorter fragments, shorter distance between fragments



Genome wide association studies (GWAS)



Phenotype of
interest (Y)
Genotype (X_s)

Linear Regression

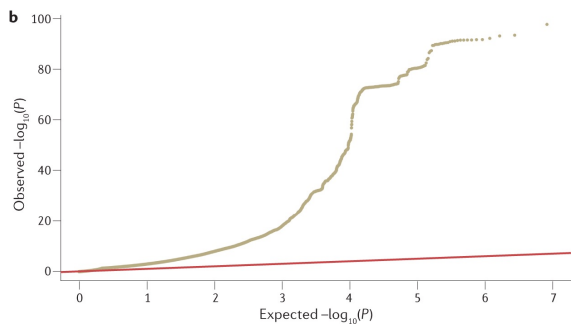
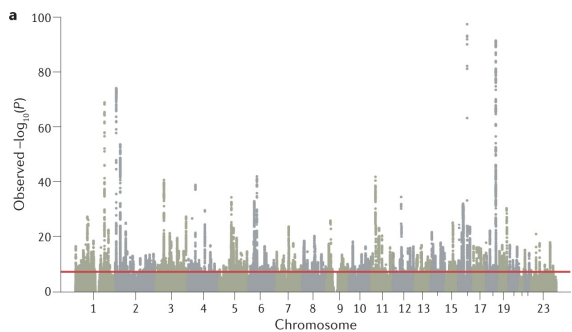
$$Y \sim X_s \beta + W\alpha + e$$

Results

- 1) β SNP effect size
- 2) P-value

Genome wide association studies (GWAS)

1) P-value → identification of associated loci

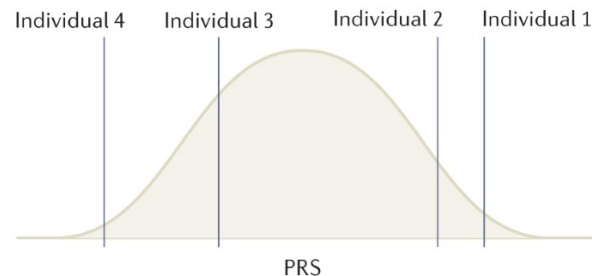


2) β SNP effect size → Polygenic Risk Score

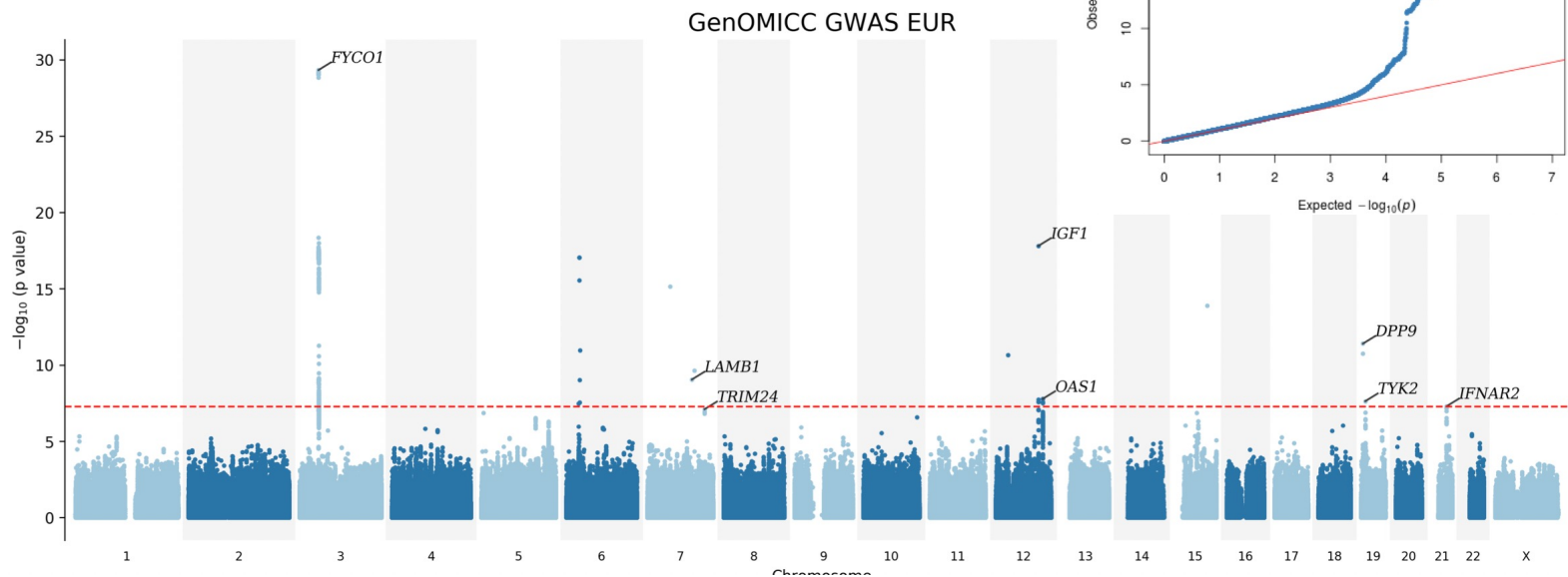
$$\text{PRS} = \sum_{i=1}^l \beta_i \text{ along the genome, per individual}$$

Express how likely an individual will present a phenotype

④ PRS distribution



Genome wide association studies (GWAS)



Païro-Castineira, E., Clohisey, S., Klaric, L. *et al.* **Genetic mechanisms of critical illness in COVID-19.** *Nature* **591**, 92–98 (2021). <https://doi.org/10.1038/s41586-020-03065-y>

Imperial College
London