Population (Pan)Genomics

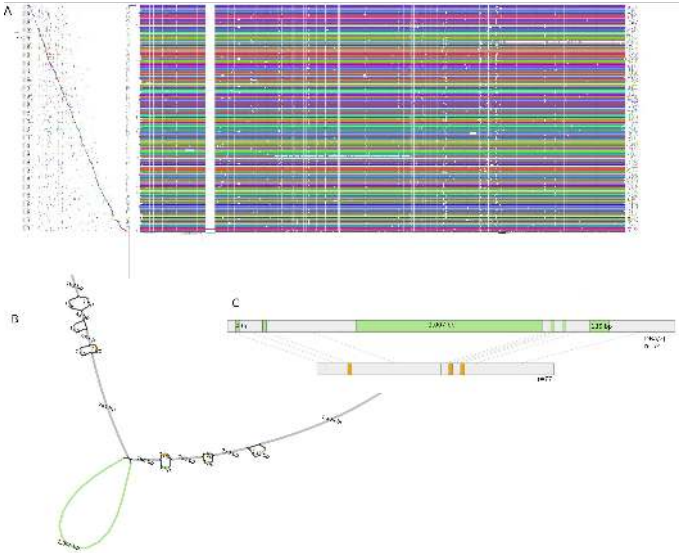
(Vinc)Enza Colonna
enza.colonna@gmail.com

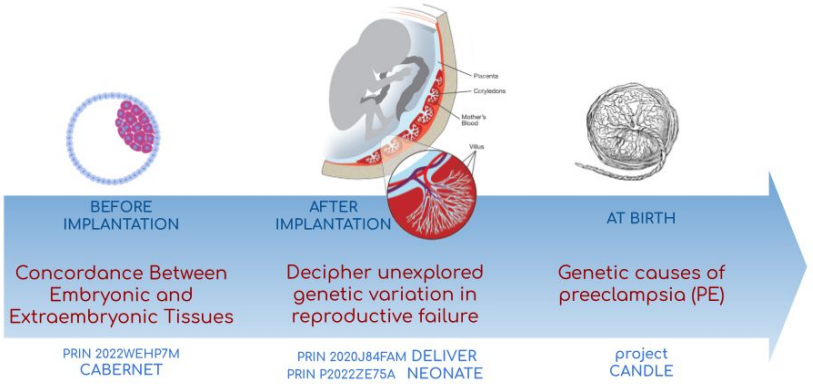
2024 WORKSHOP ON GENOMICS
Český Krumlov - 17 Jan 2024

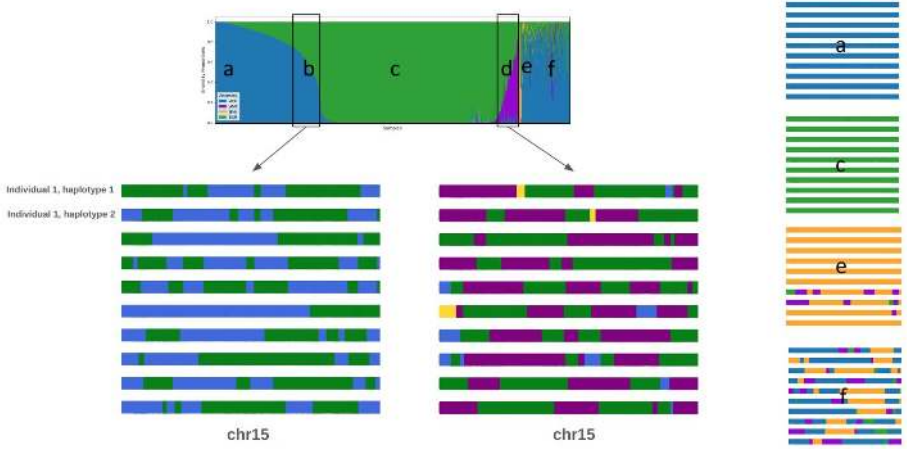
Consiglio Nazionale delle Ricerche
THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER

2024-	Associate Professor / University of Tennessee / US
2024-	Director of the biobank of the Biorepository and Integrative Genomics (BIG) Initiative Cohort / University of Tennessee / US
2012-	Researcher / Consiglio Nazionale delle Ricerche / Italy
2022-2023	Assistant Professor / University of Tennessee / US
2010-2012	Postdoc / Human Evolution Team / Wellcome Trust Sanger Institute / UK
2007-2010	Postdoc / Dep Genetics and Evolution / University Ferrara / Italy
2006	PhD / Evolutionary Biology / University of Naples Federico II / Italy

Going on in my lab







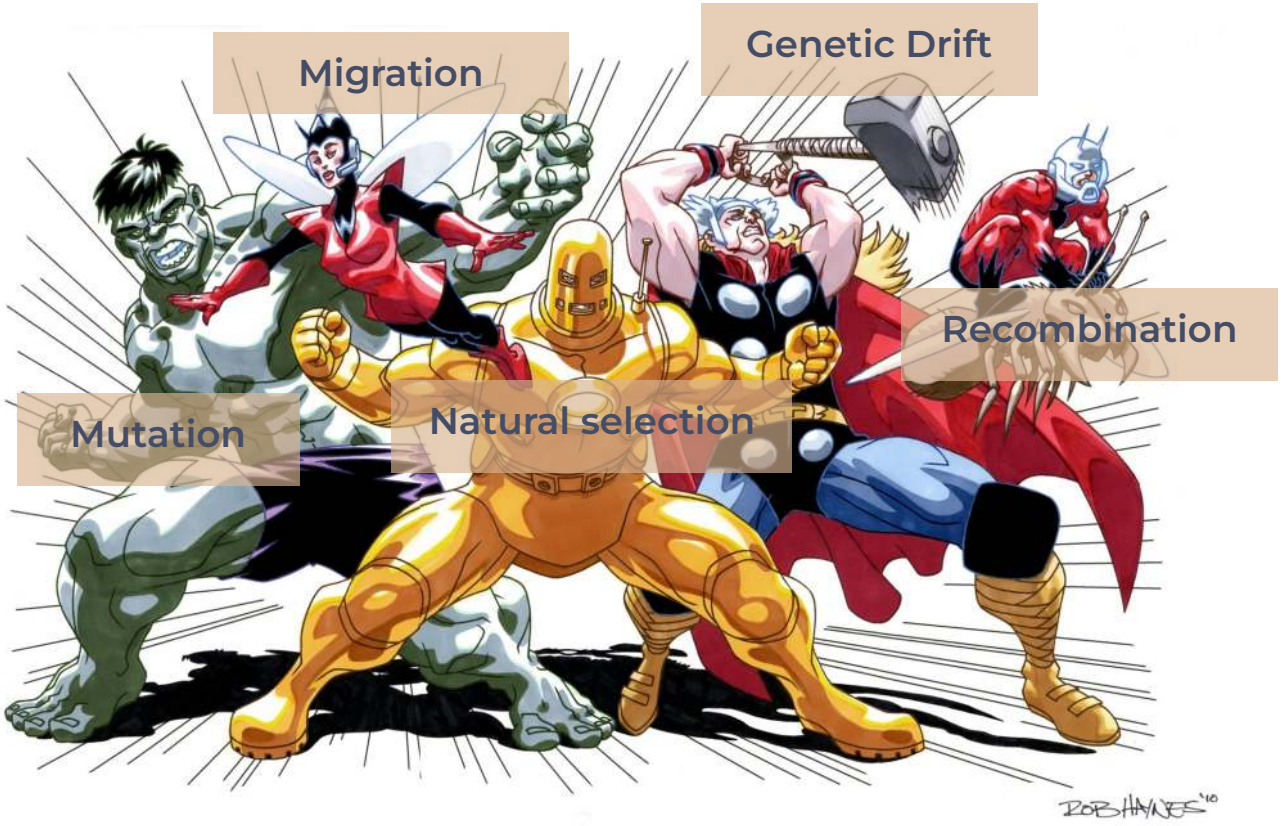
Outline

- Theory: forces shaping genetic diversity
- Population gen-etics -omics questions
- Allele Frequency Spectrum: the basic
- Fst: population differentiation
- PCA: compact representation

Theory: forces shaping genetic diversity

Forces shaping genetic diversity

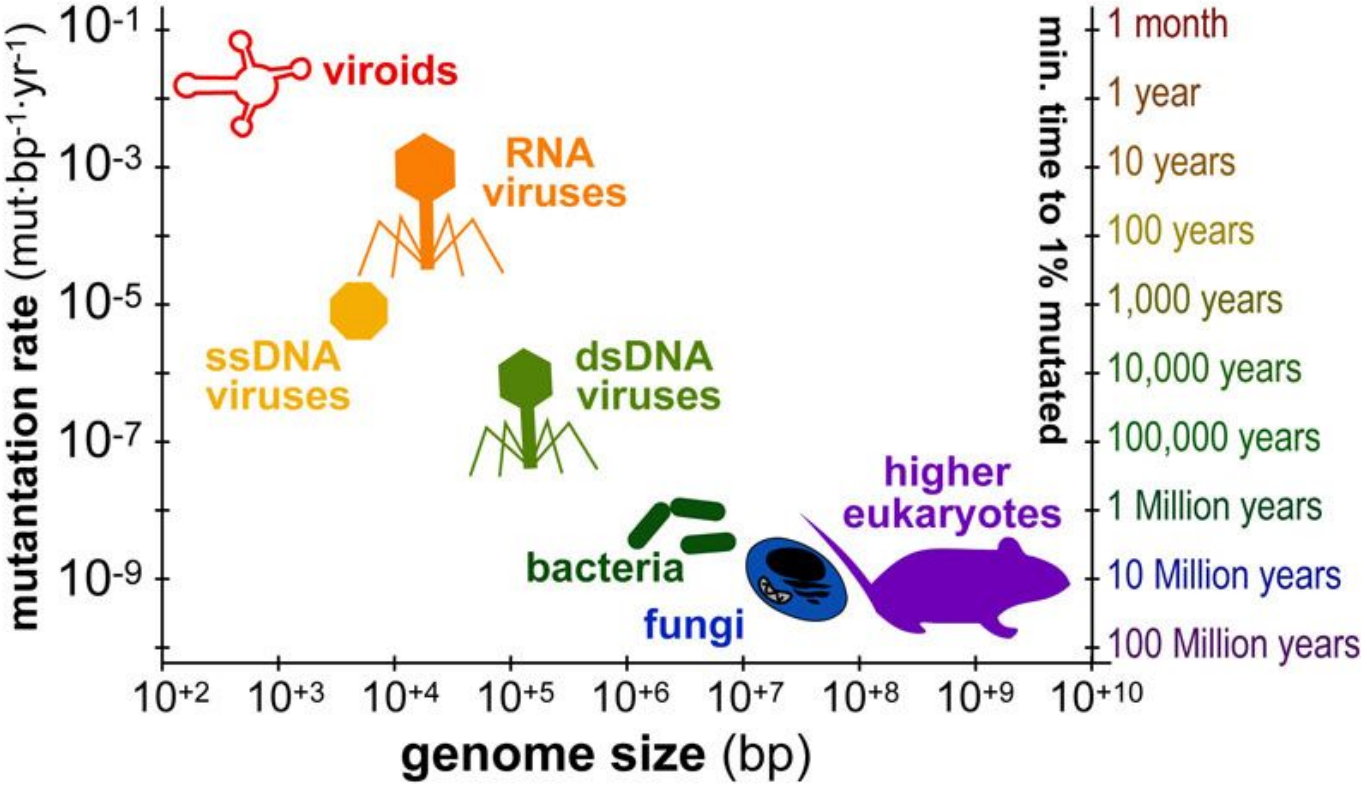
Forces shaping genetic diversity

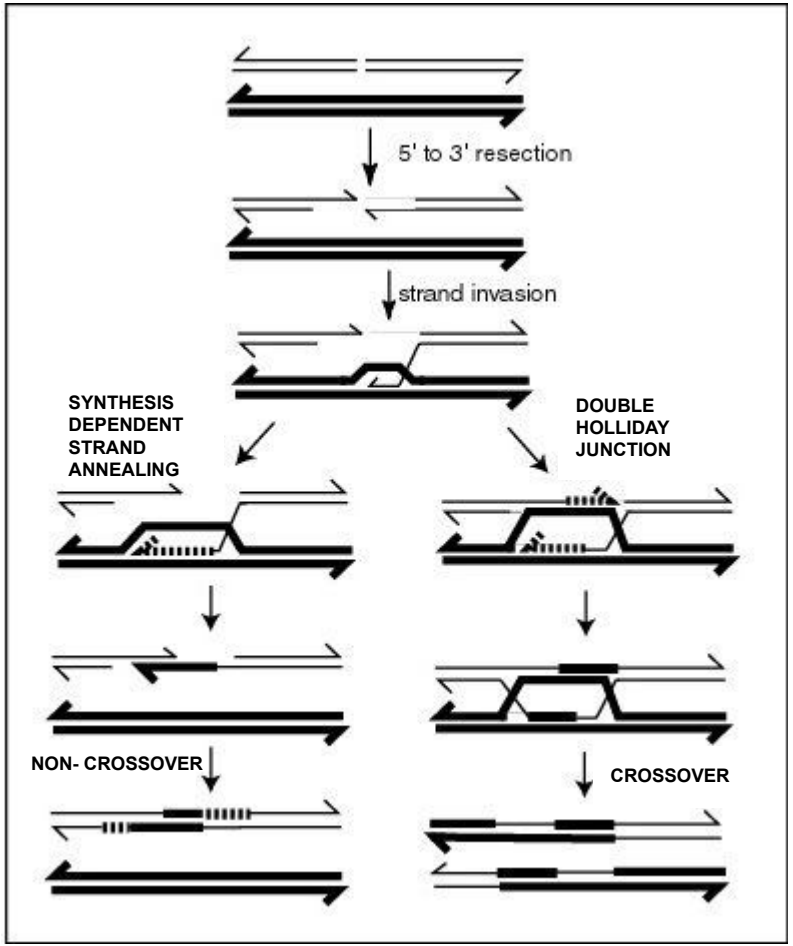


Forces shaping genetic diversity

Mutation	generation of new variants
Recombination	horizontal shuffling of alleles
Genetic Drift	stochastic fluctuations of frequencies due to finite populations
Migration	introduction of migrant alleles
Natural Selection	prevalence of alleles due to fitness

Mutation rate





Recombination

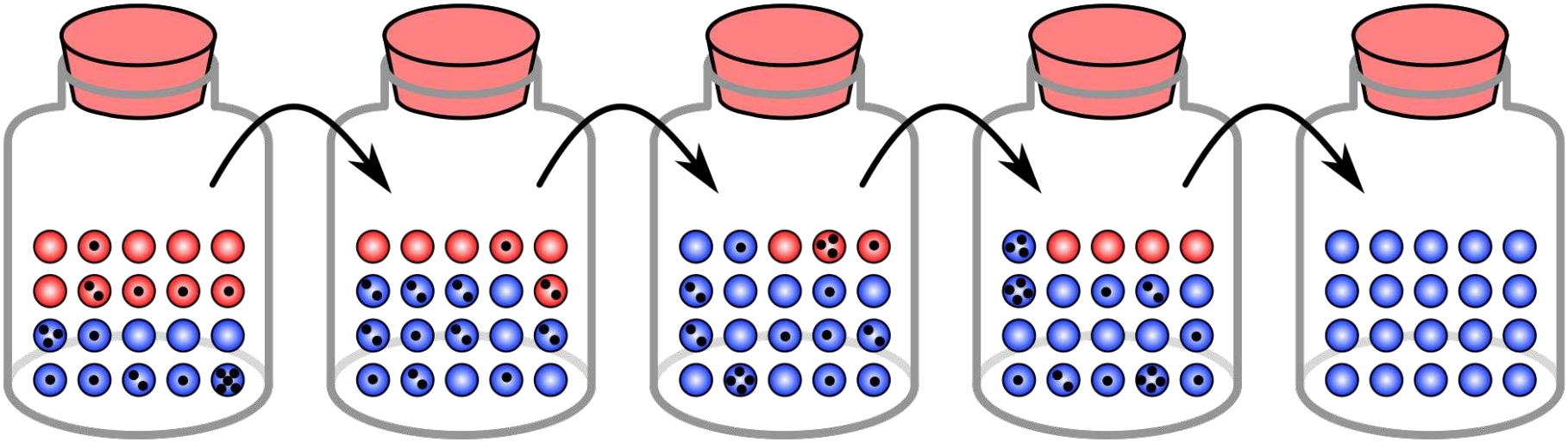
→ Creates new combinations of variation (not new variation)

→ Does not change allele frequency

→ Re-arranges mutations

https://en.wikipedia.org/wiki/Genetic_recombination

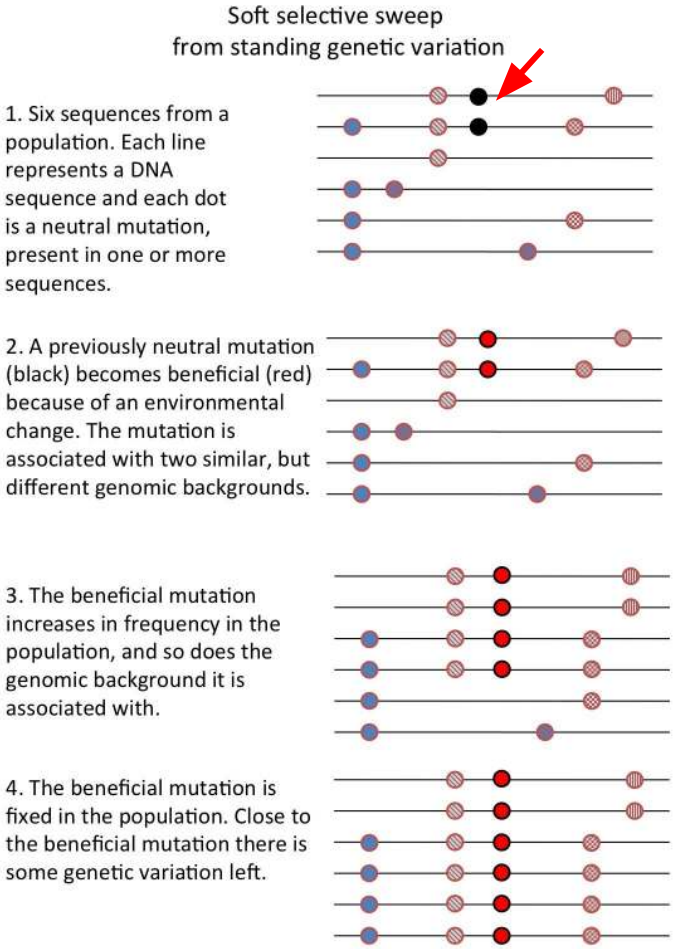
Genetic Drift



In this simulation, each black dot on a marble signifies that it has been chosen for copying (reproduction) one time.

Fixation in the blue "allele" occurs within five generations.

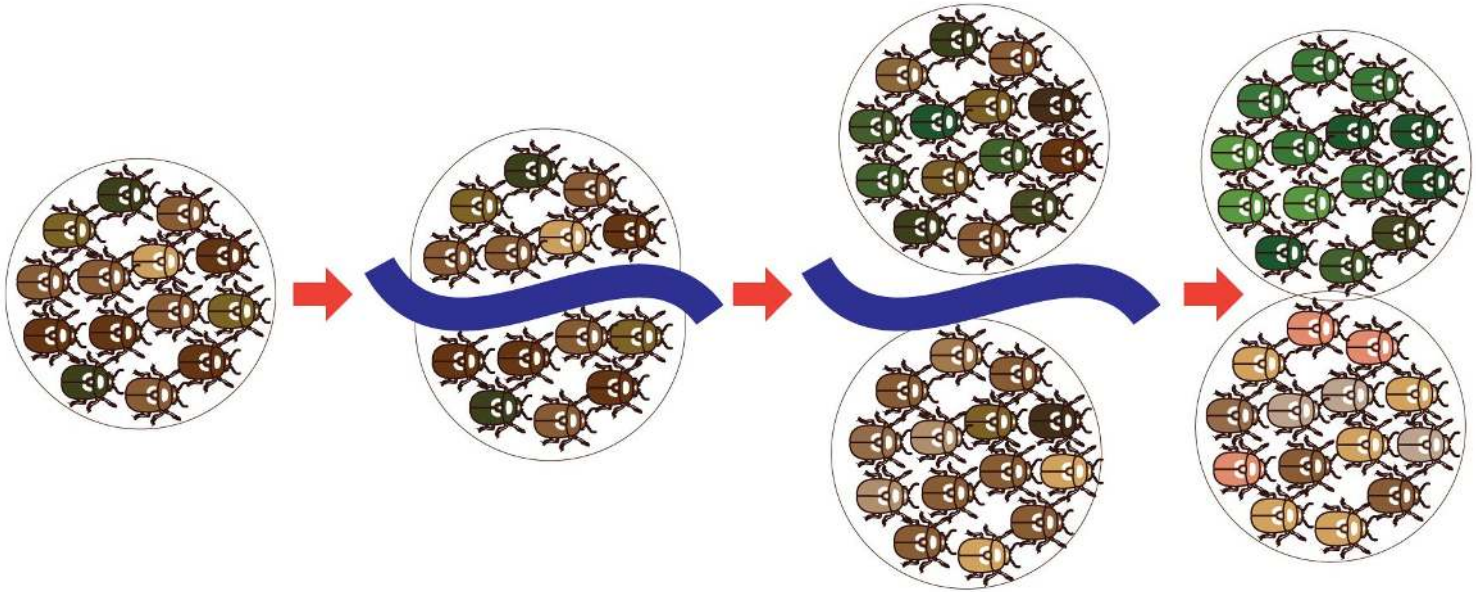
Natural Selection



By Chiswick Chap - Self-published work by Chiswick Chap, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=880130

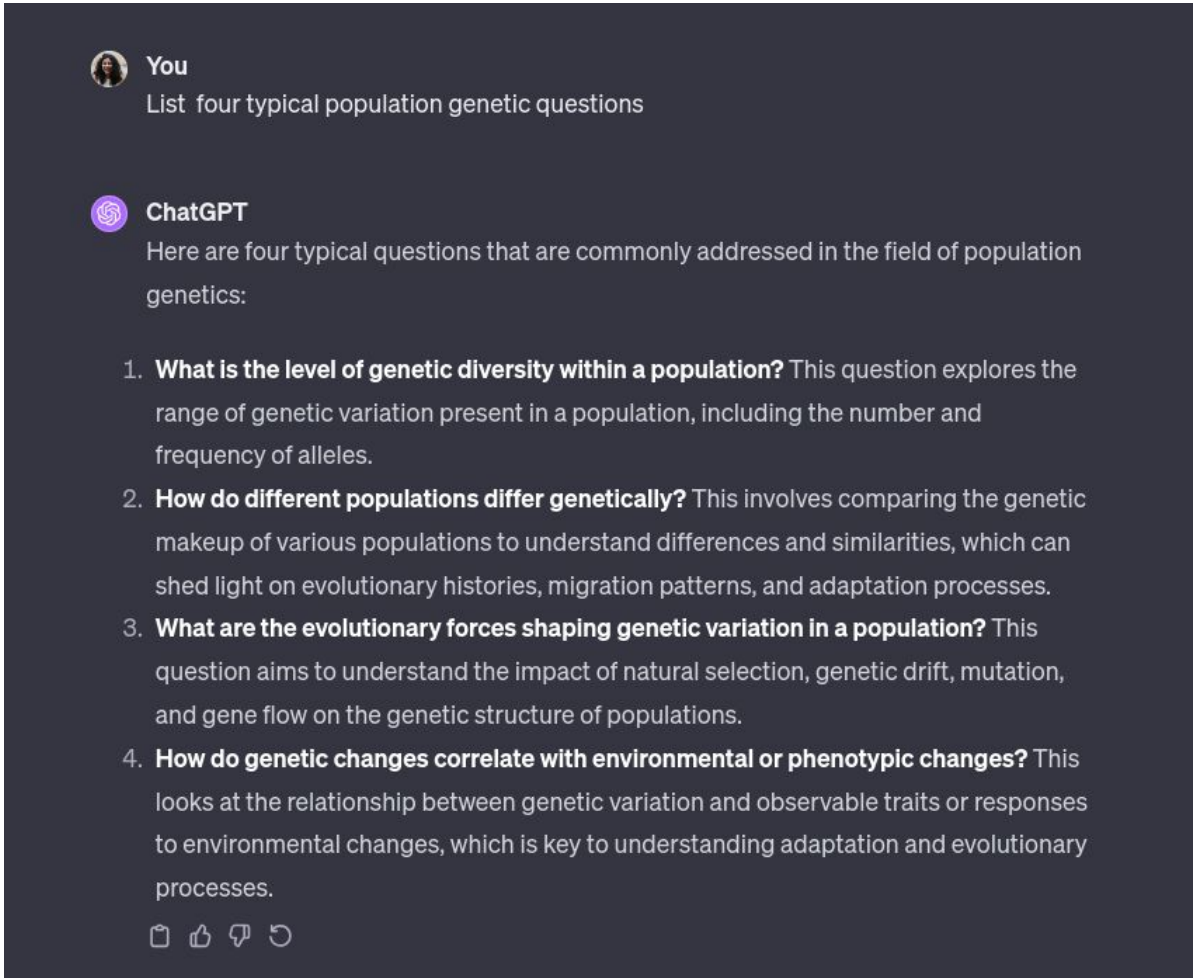
https://en.wikipedia.org/wiki/Selective_sweep

(lack of) Migration



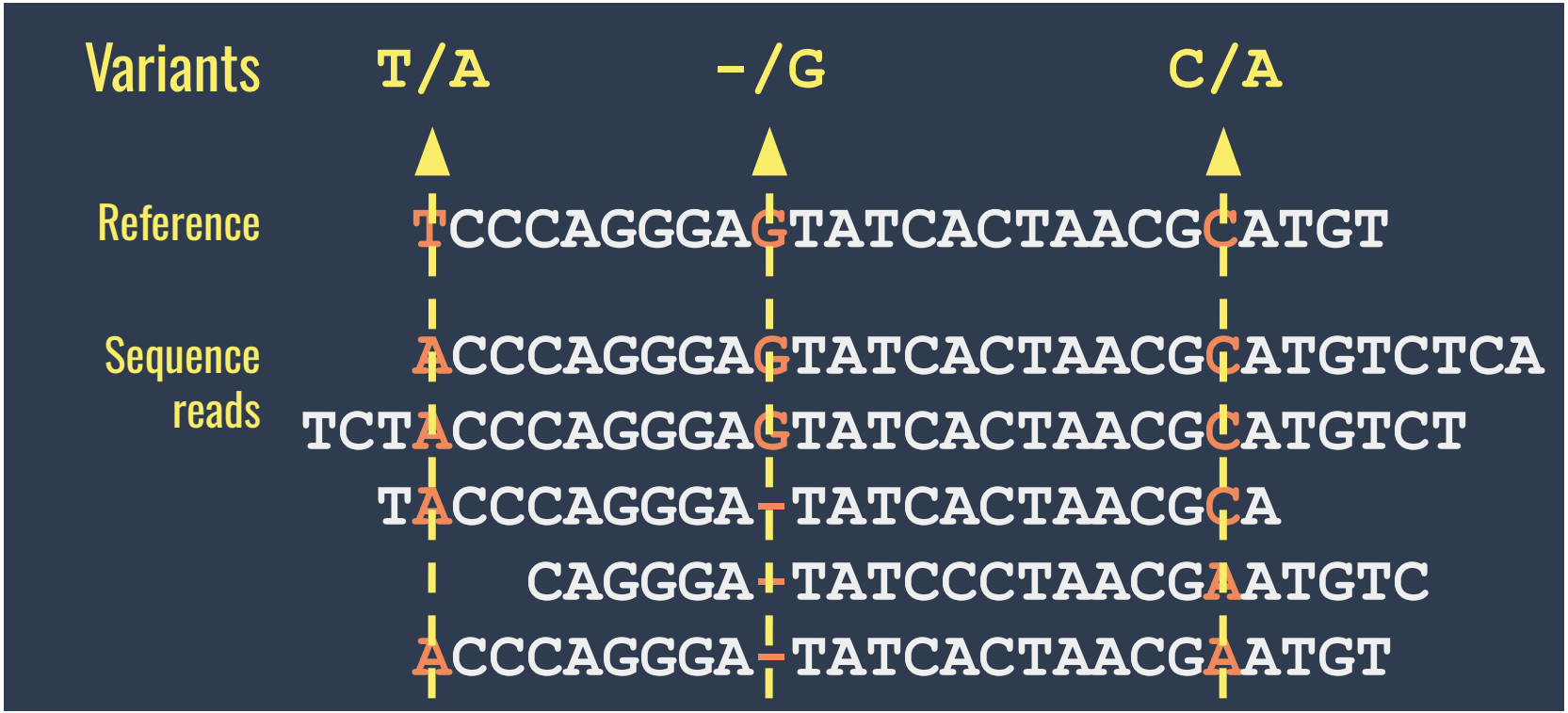
7th Edition - 10–16 June 2024 | Naples, Italy

Population gen-etics -omics questions

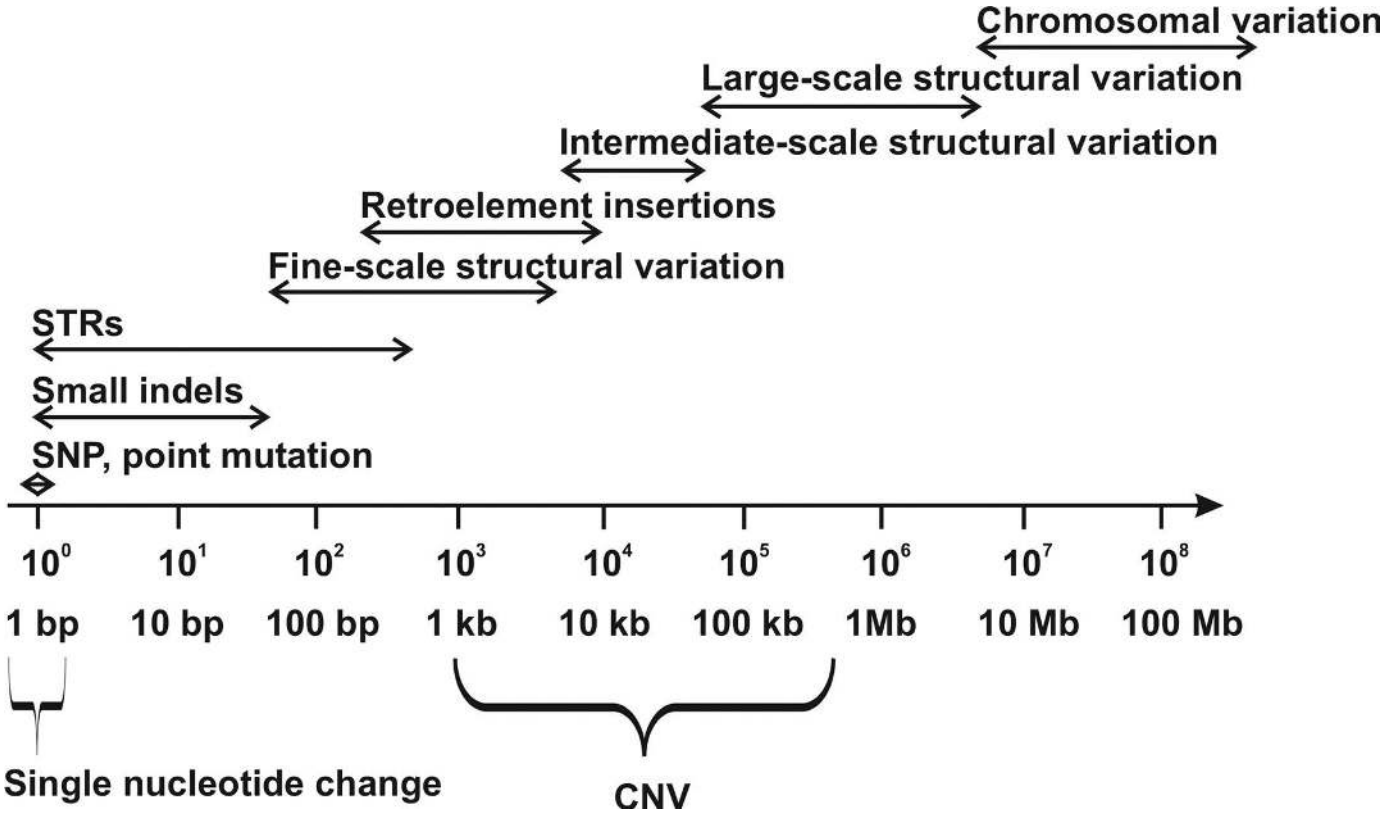


What is the focus of research in population genetics?

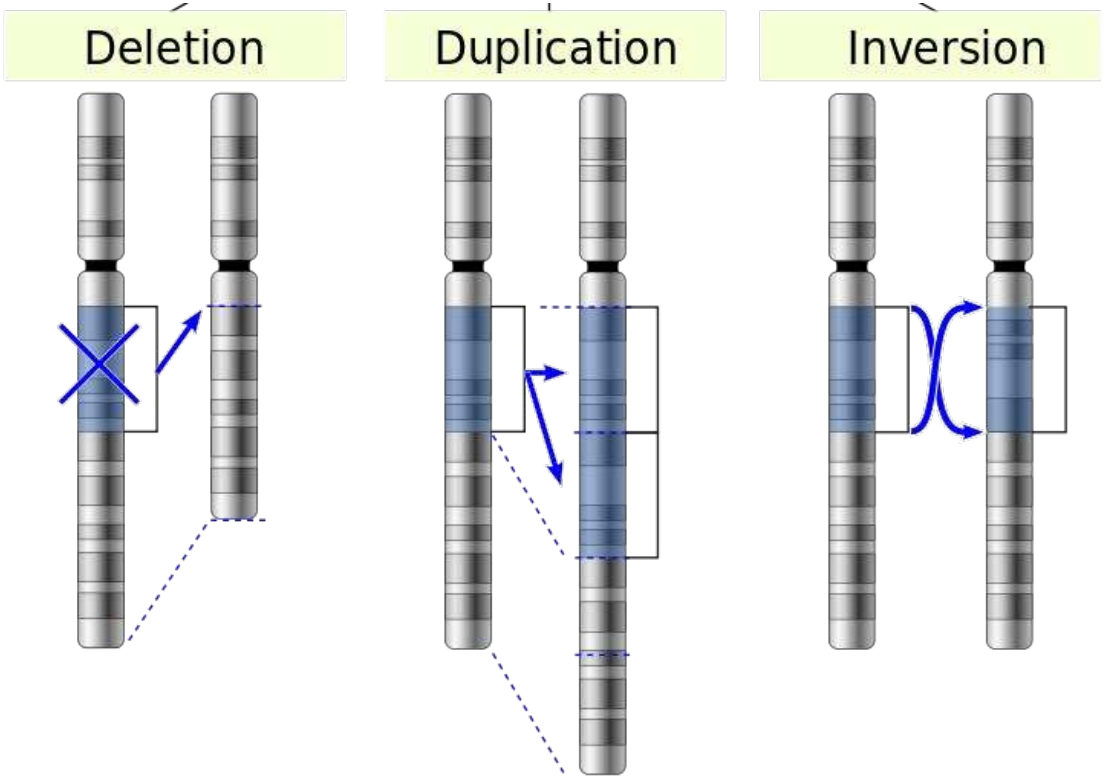
Genetic variants



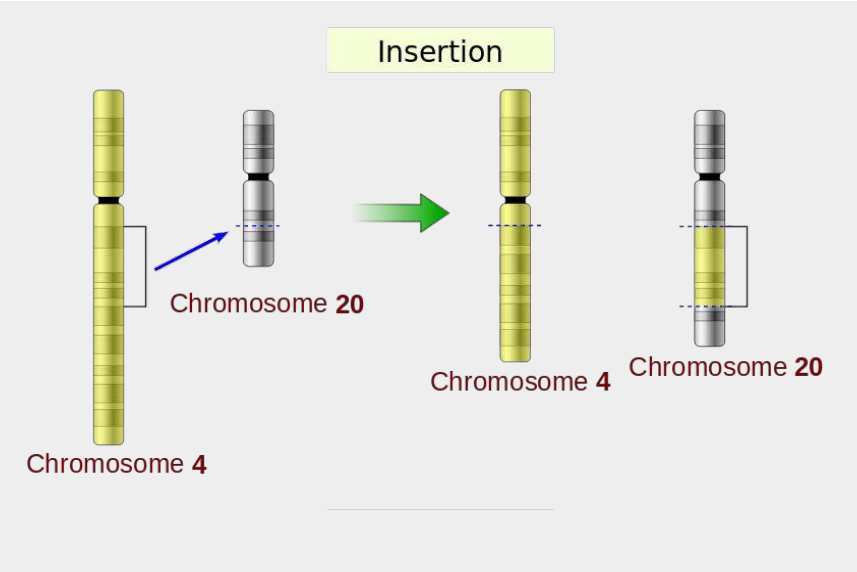
Variant types and sizes

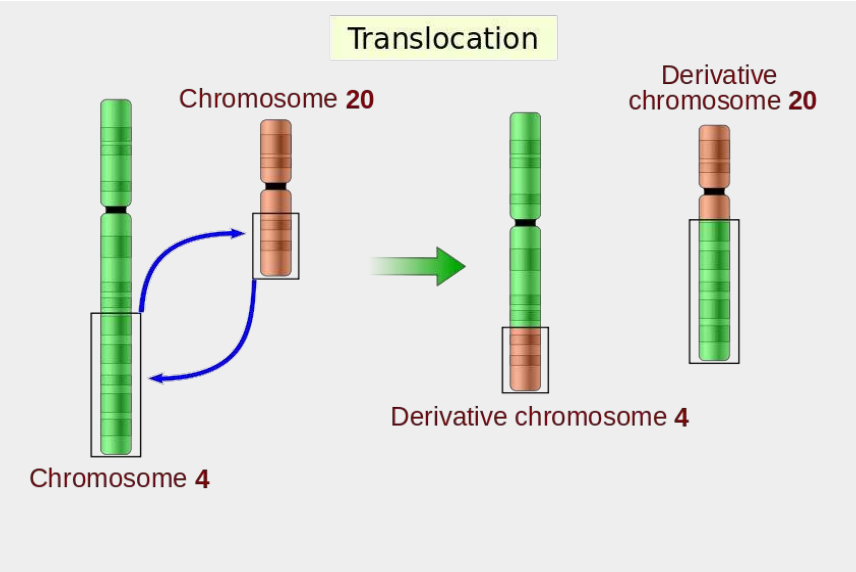


Single chromosome mutations

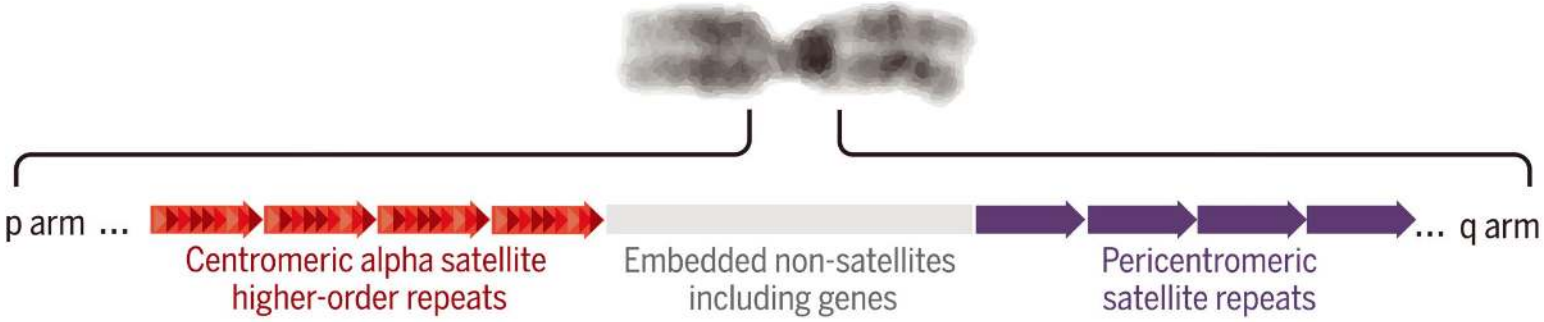


Multiple chromosomes

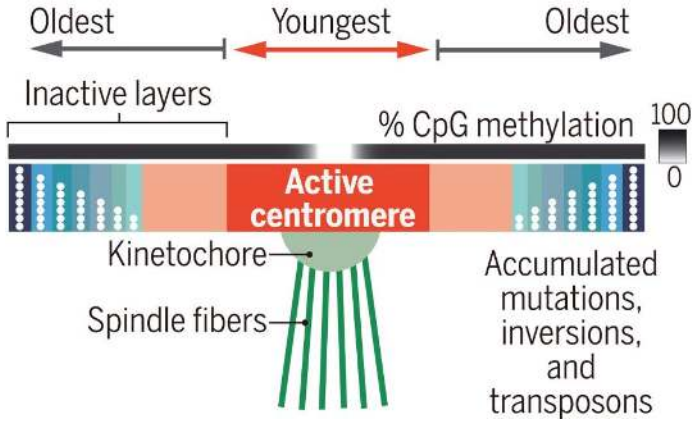




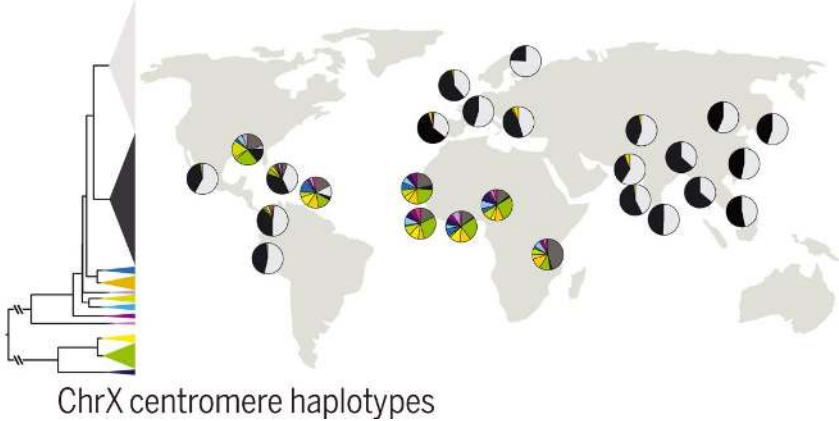
Genetic variation within centromeres



Layered expansion of kinetochore-bound centromeric sequences

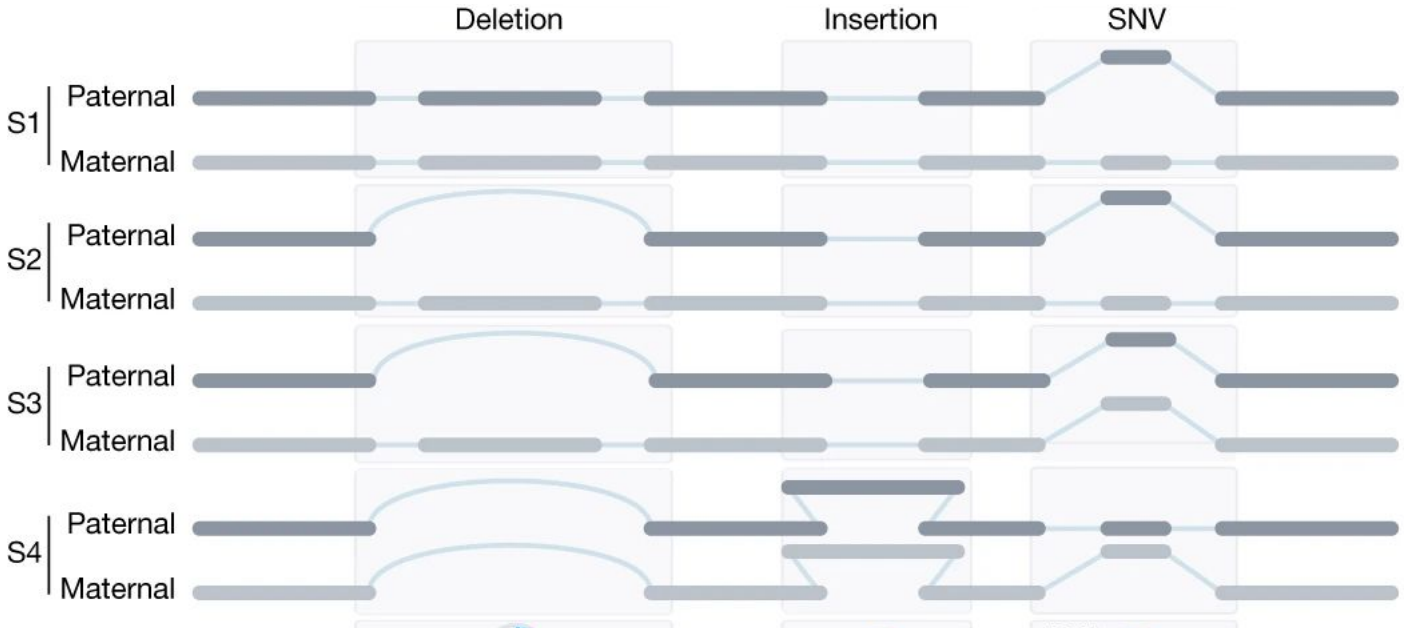


High genetic diversity in centromeres from populations with recent African ancestry



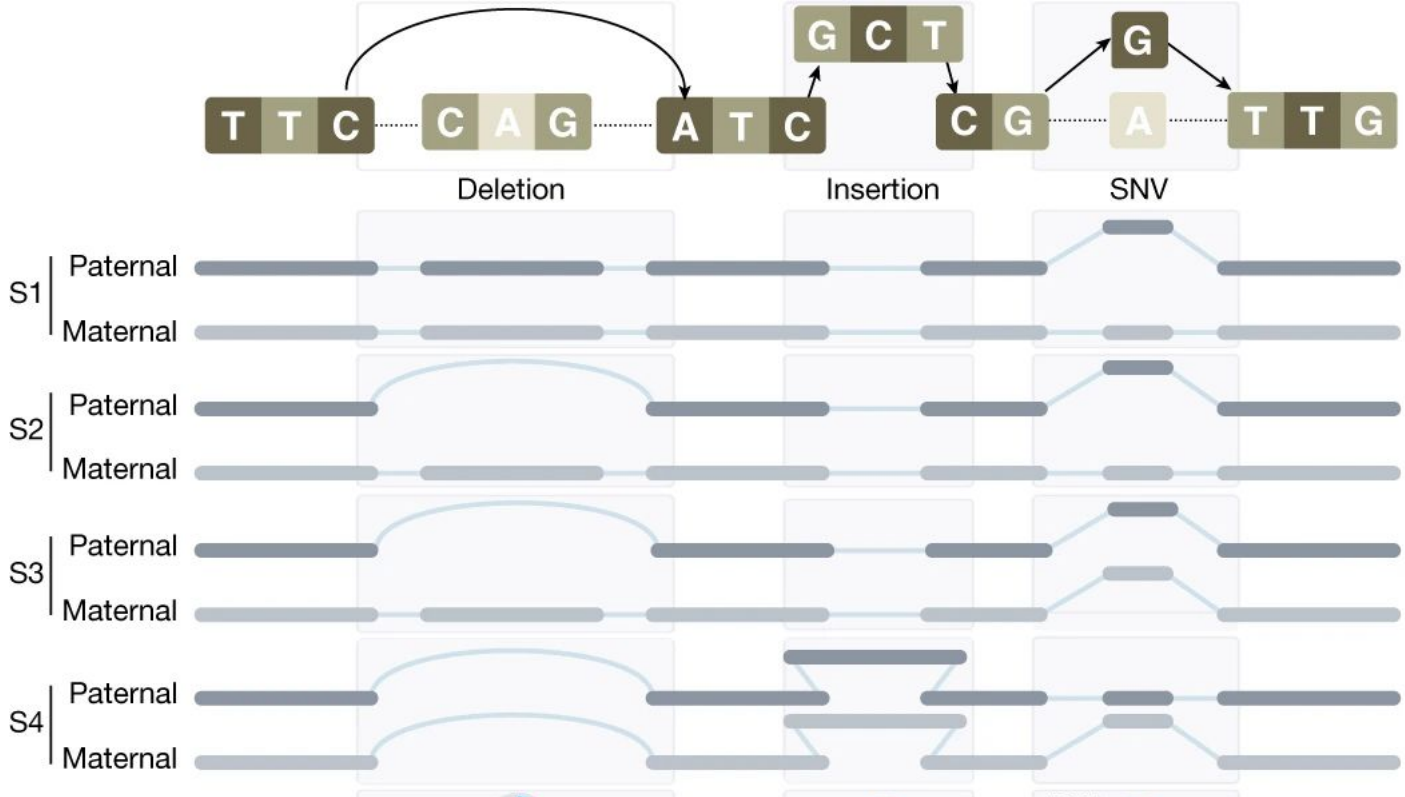
https://www.science.org/doi/10.1126/science.abl4178

Pangenomes contain genomes and variation



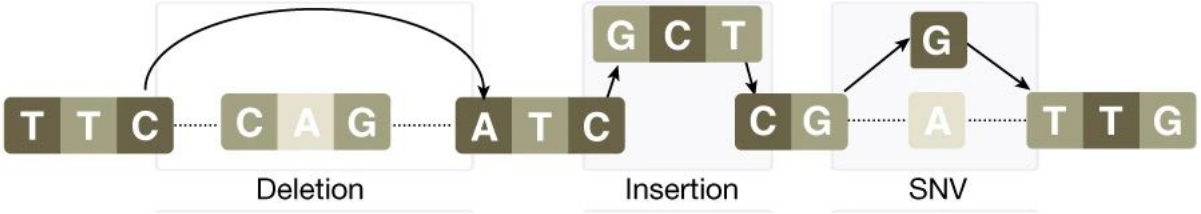
Wang et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022

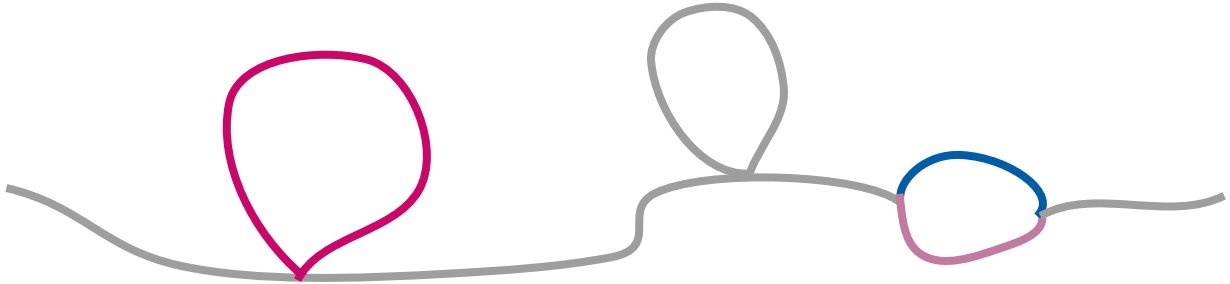
Pangenomes contain genomes and variation



Wang et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022

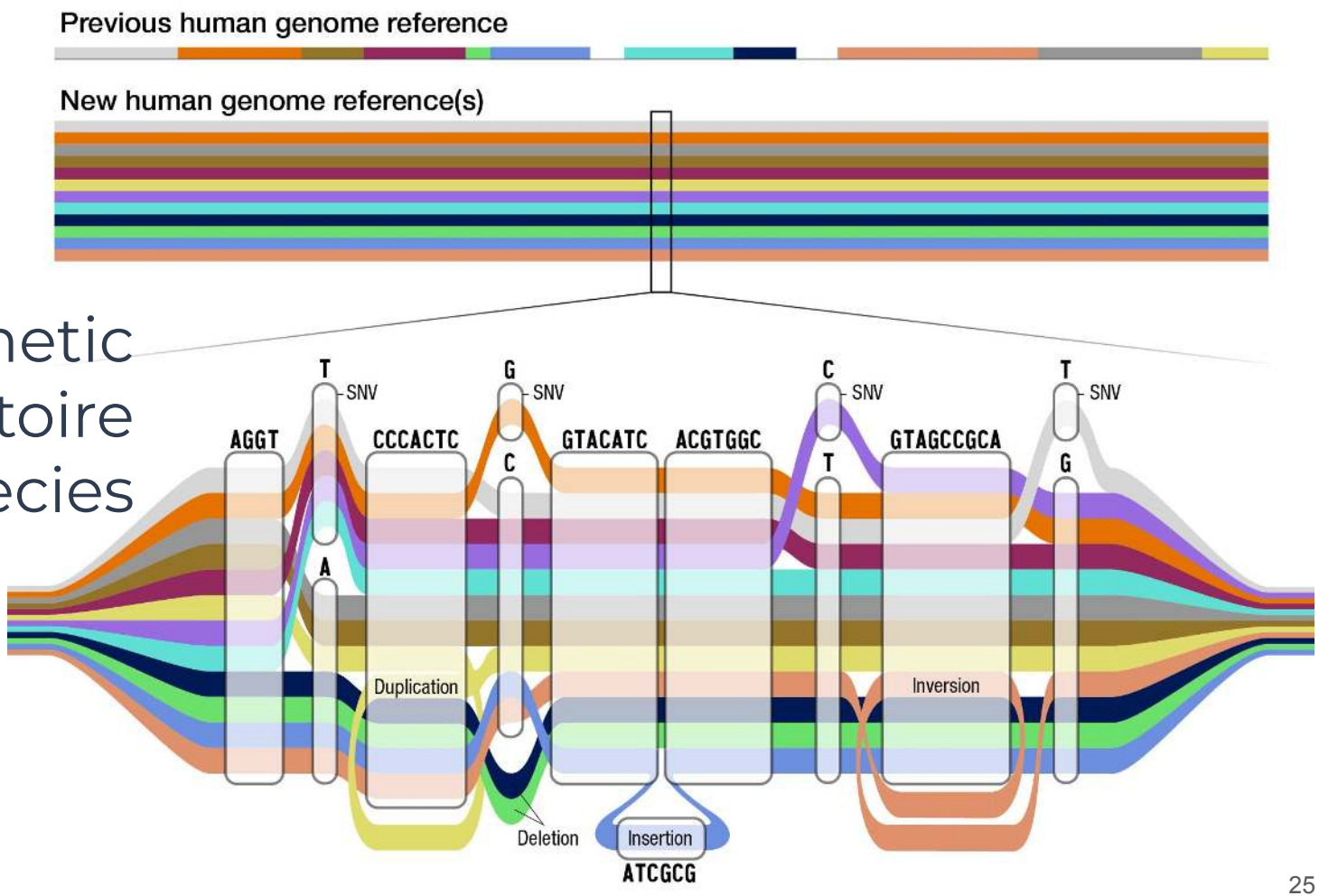
Pangenomes contain genomes and variation



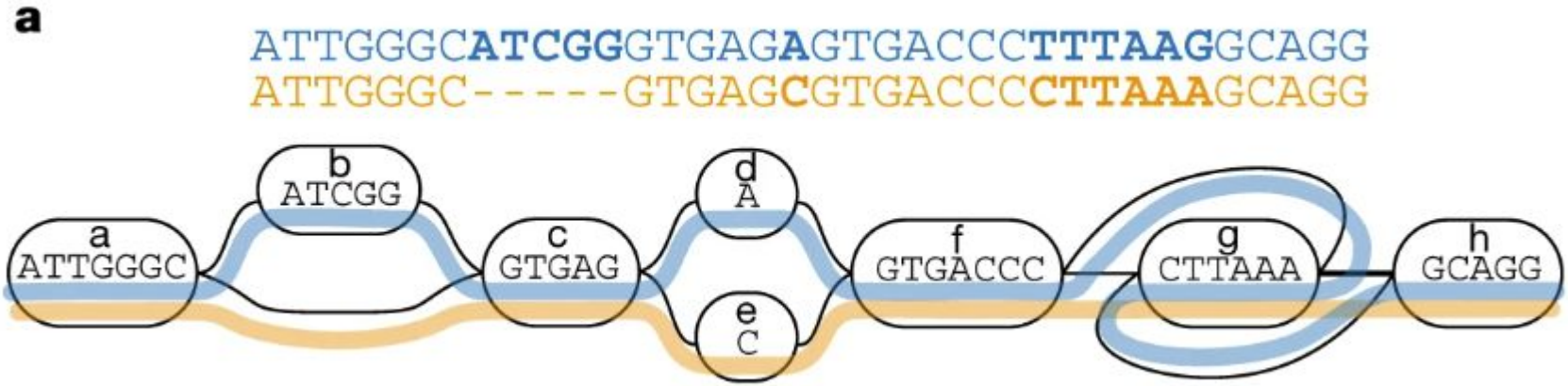


Wang et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022

The full genetic repertoire of a species



Haplotype resolved complete representation of genomes



Nested variation

Allele Frequency Spectrum: the basic

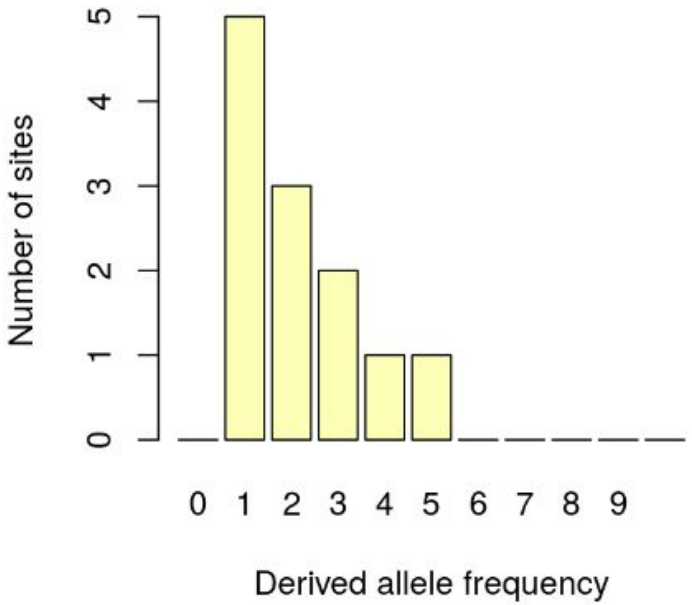
Slides borrowed from:

Victor C. Sousa	Evomics 2020	https://evomics.org/wp-content/uploads/2022/06/WoG_PopGenTutorial2_DemographicInference.pdf
Emiliano Trucchi Josephine Paris	Evomics 2022	https://evomics.org/learning/population-and-speciation-genomics/2020-population-and-speciation-genomics/inferring-demography/

	INDIVIDUALS					DERIVED ALLELE FREQUENCY
	IND1	IND2	IND3	IND4	IND5	ABSOLUTE FREQUENCY
snp1	0	1	0	1	0	**2**
snp2	0	1	0	1	0	**2**
snp3	0	1	1	1	2	**5**
snp4	0	1	1	0	0	**2**
snp5	0	1	0	0	2	**3**
snp6	1	0	0	1	1	**3**
snp7	1	0	0	0	0	**1**
snp8	0	0	1	0	0	**1**
snp9	0	0	0	0	1	**1**
snp10	0	0	0	0	1	**1**
snp11	0	0	1	0	0	**1**
snp12	0	1	2	1	0	**4**

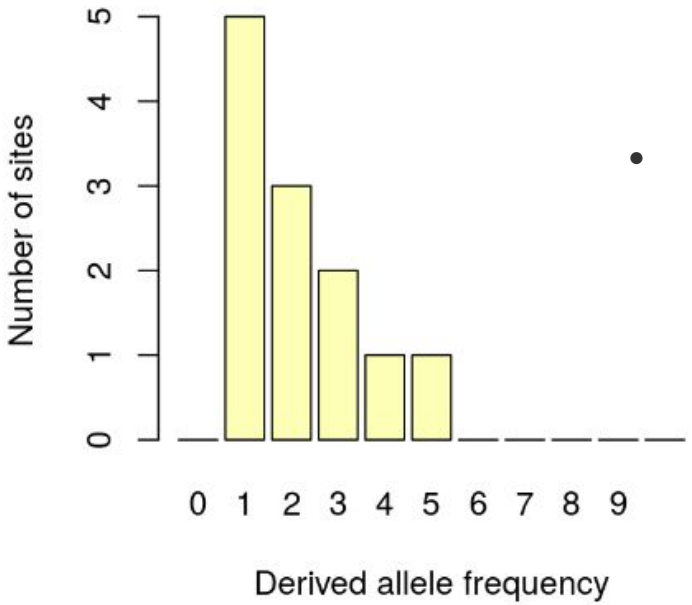
	0	1	2	3	4	5	6	7	8	9	10
Number of sites	0	5	3	2	1	1	0	0	0	0	0

Abs. Frequency of derived allele



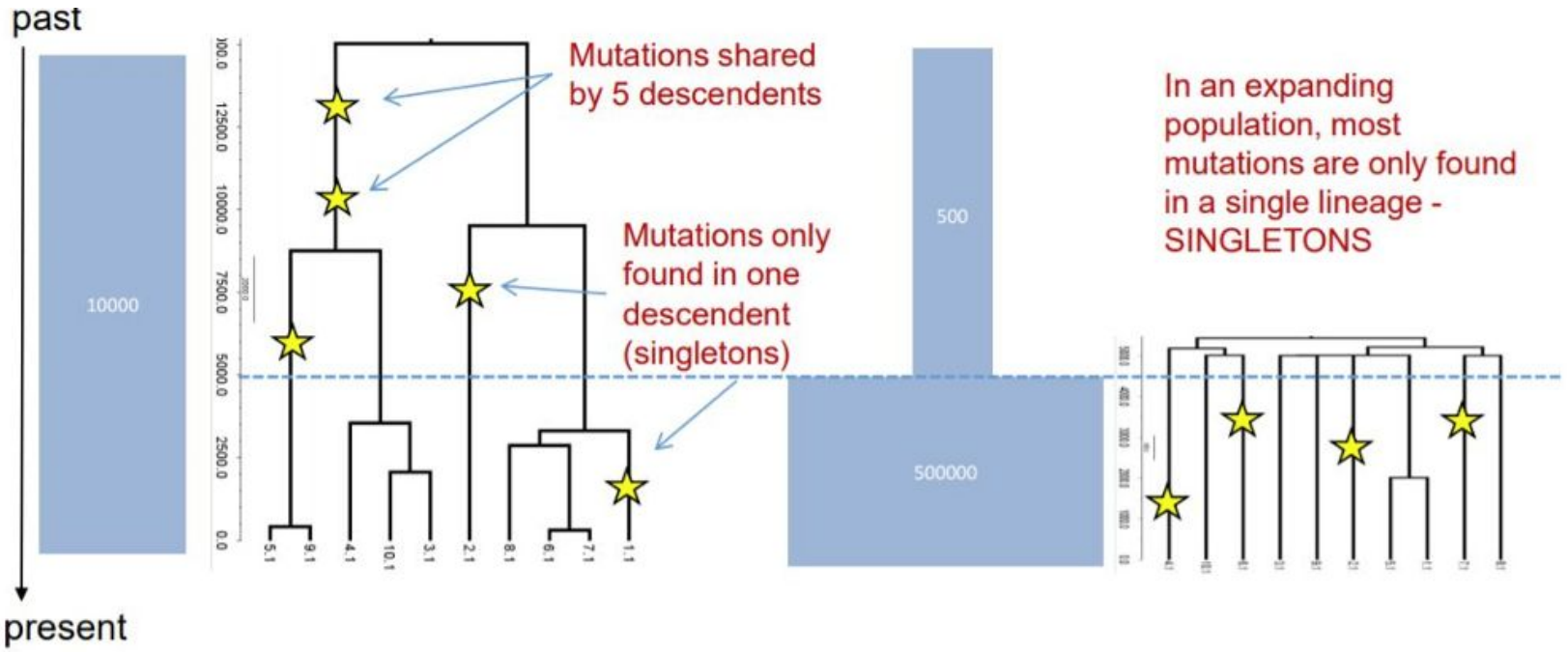
	0	1	2	3	4	5	6	7	8	9	10
Number of sites	0	5	3	2	1	1	0	0	0	0	0

Abs. Frequency of derived allele



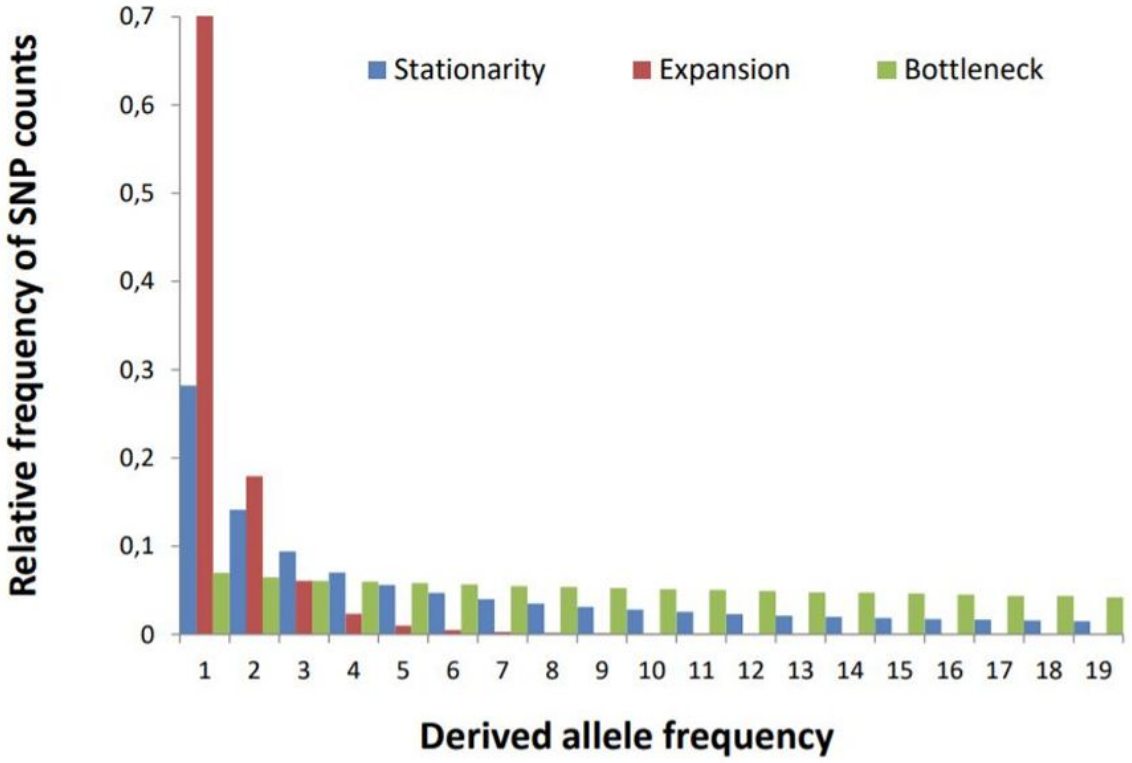
- Q1: What is the x-axis of the SFS? What is the range of possible values for the x-axis?
  - The x-axis of the SFS contains the absolute frequency of the derived allele in the sample. Thus, it can take values from zero (if at a site all individuals are homozygote for ancestral allele) to 2*n (if all the n individuals are homozygote for the derived allele). Note that we need to multiply by 2 because individuals are diploid

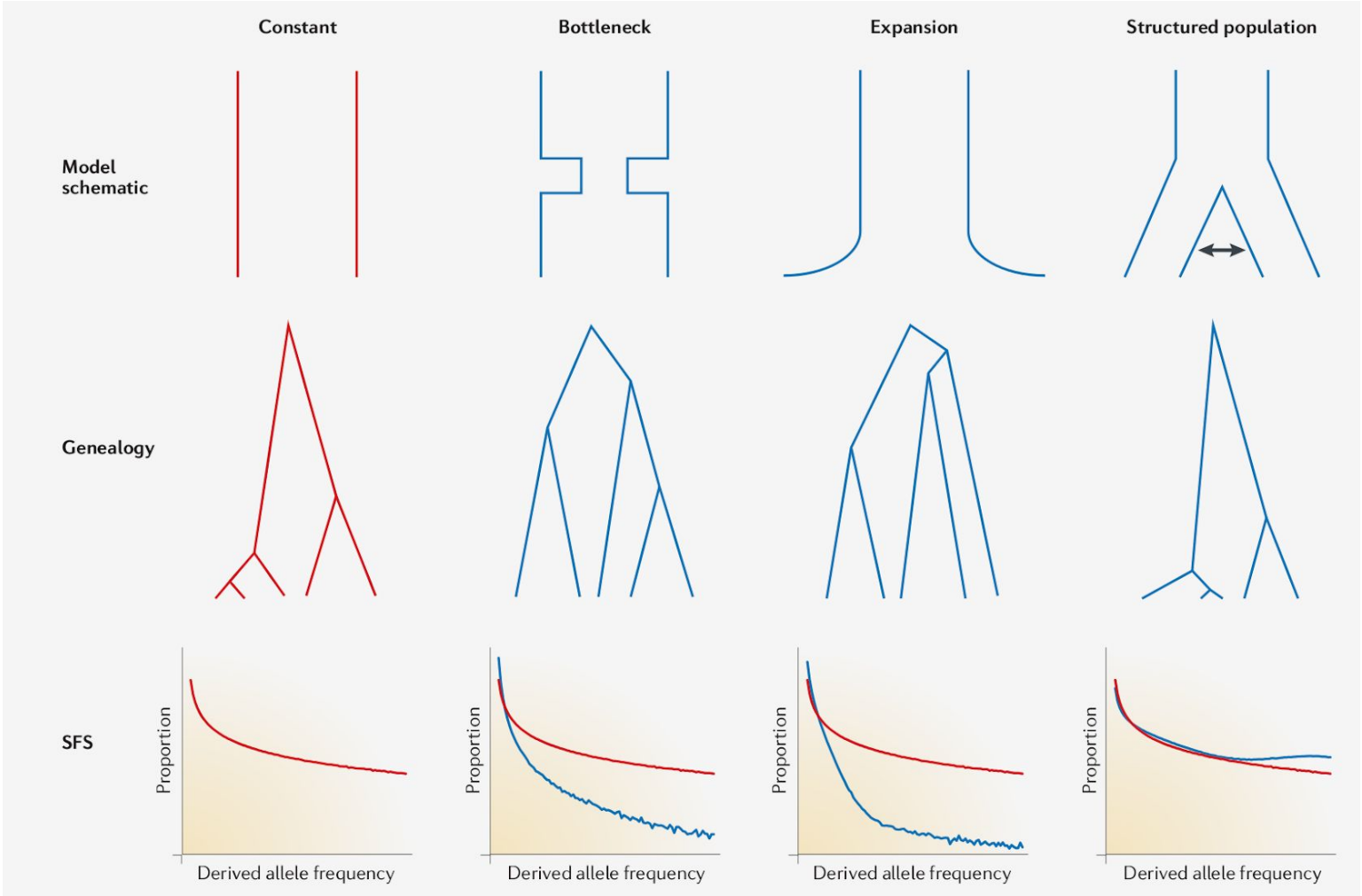
- Q2: What is the y-axis of the SFS? What is the range of possible values for the y-axis?
  - The y-axis of the SFS contains the number of sites in the genome with a given absolute frequency. Thus, it can take values between zero (if there are no sites in the genome with a given frequency) to S, where S is the number of sites.


past
10000
500
500000
present
Mutations shared by 5 descendents
Mutations only found in one descendent (singletons)
In an expanding population, most mutations are only found in a single lineage - SINGLETONS


THIS
IS
THE
WAY

AFS depends on past demography




Constant
Bottleneck
Expansion
Structured population
Model schematic
Genealogy
SFS
Proportion
Derived allele frequency
Proportion
Derived allele frequency
Proportion
Derived allele frequency
Proportion
Derived allele frequency

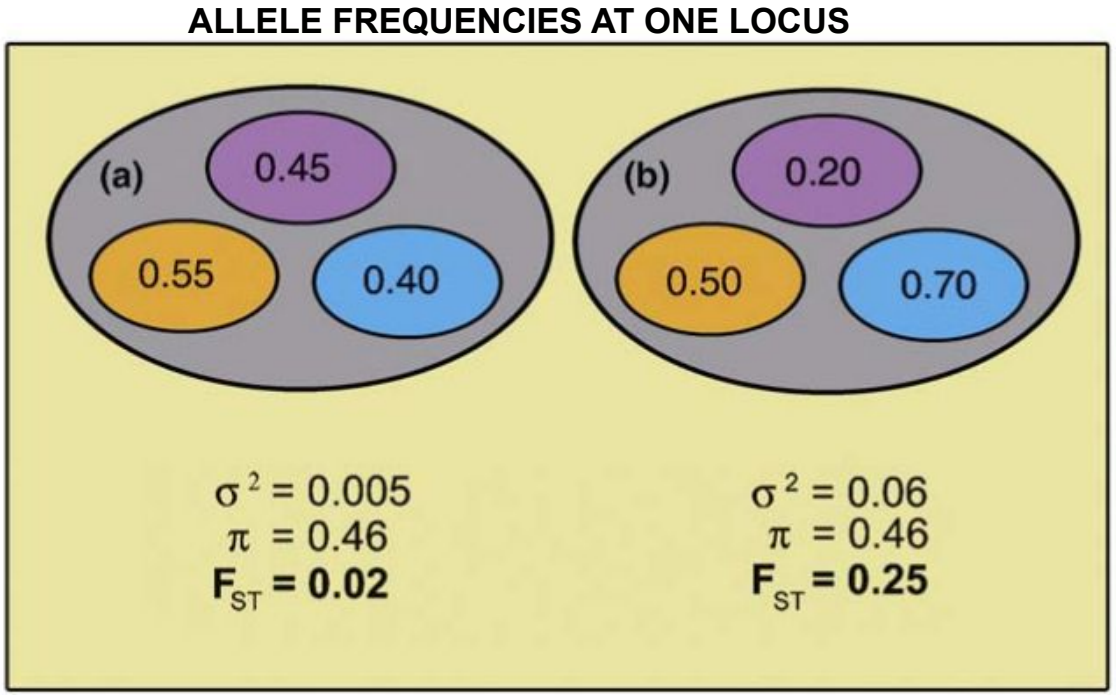
From vcf to allele frequencies

```
##fileformat=VCFv4.3
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20240116
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HG00438 HG00621 HG00673 HG00733 HG00735 HG00741 HG01071 HG01106 HG01109 HG01123 HG01175 HG01243 HG01258
HG01358
chm13#chr15 17710067 chm13#chr15:17710067 C T 60 . AC=2;AF=0.0229885;AN=87 GT 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 1|0 0|0 0|0
chm13#chr15 17710174 chm13#chr15:17710174 C G 60 . AC=35;AF=0.416667;AN=84 GT 1|1 0|0 0|1 1|0 0|0 1|0 0|0 1|0 0|0 0|0 0|1 .|. 0|0 1|1
chm13#chr15 17710214 chm13#chr15:17710214 A G 60 . AC=35;AF=0.416667;AN=84 GT 1|1 0|0 0|1 1|0 0|0 1|0 0|0 1|0 0|0 0|0 0|1 .|. 0|0 1|1
chm13#chr15 17710218 chm13#chr15:17710218 A C 60 . AC=35;AF=0.416667;AN=84 GT 1|1 0|0 0|1 1|0 0|0 1|0 0|0 1|0 0|0 0|0 0|1 .|. 0|0 1|1
chm13#chr15 17710227 chm13#chr15:17710227 G A 60 . AC=35;AF=0.416667;AN=84 GT 1|1 0|0 0|1 1|0 0|0 1|0 0|0 1|0 0|0 0|0 0|1 .|. 0|0 1|1
```

STRING OF GENOTYPES	COUNT OF ‘1’	COUNT OF ALLELES	ALLELE FREQUENCIES
0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 0|0 1|0 0|0 0|0	1	28	1/28 = 0.04
1|1 0|0 0|1 1|0 0|0 1|0 0|0 1|0 0|0 0|0 0|1 .|. 0|0 1|1	9	26	9/26 =0.35

Fst: population differentiation

One of many ways to calculate Fst



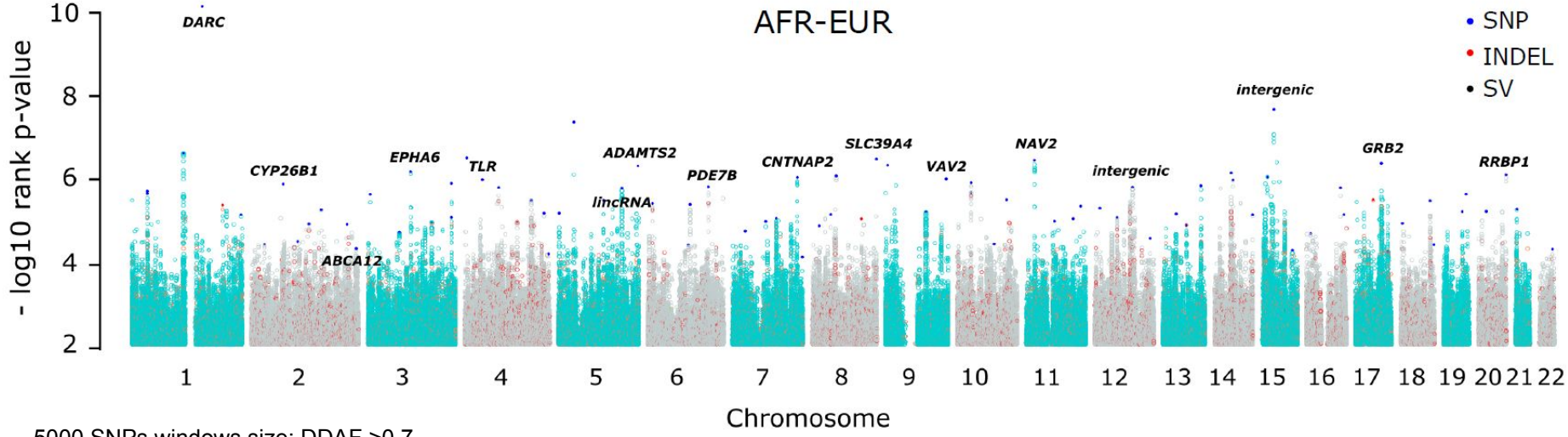
standardized variance of allele frequencies among subpopulations

FST = s^2 / p(1-p)

s^2 = variance of allele frequencies

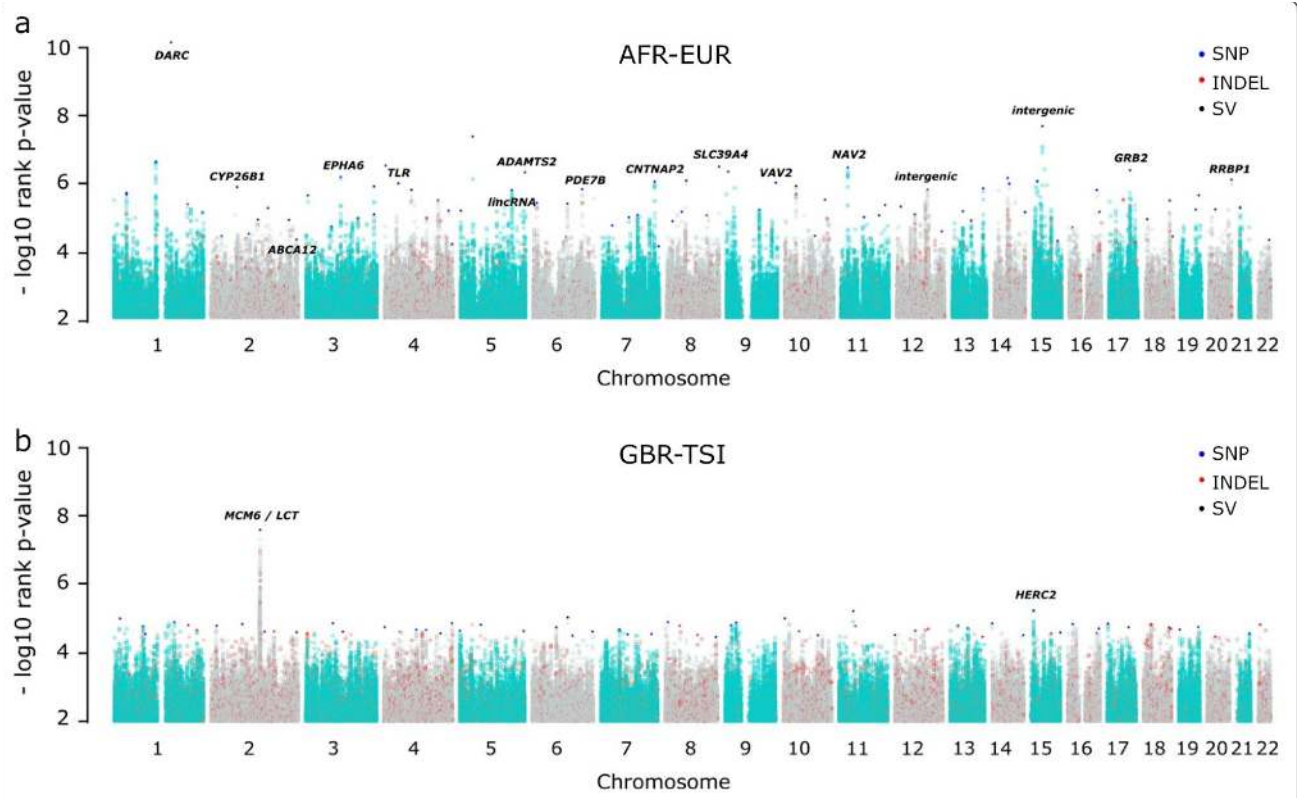
p = mean of the allele frequency

A catalog of highly differentiated sites



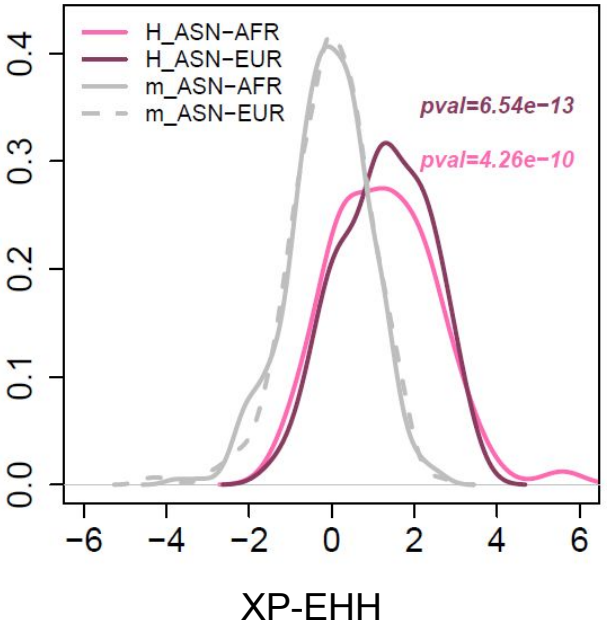
5000 SNPs windows size; DDAF >0.7

Between-continents *vs* **within**-continents



Colonna 2014 Genome Biology

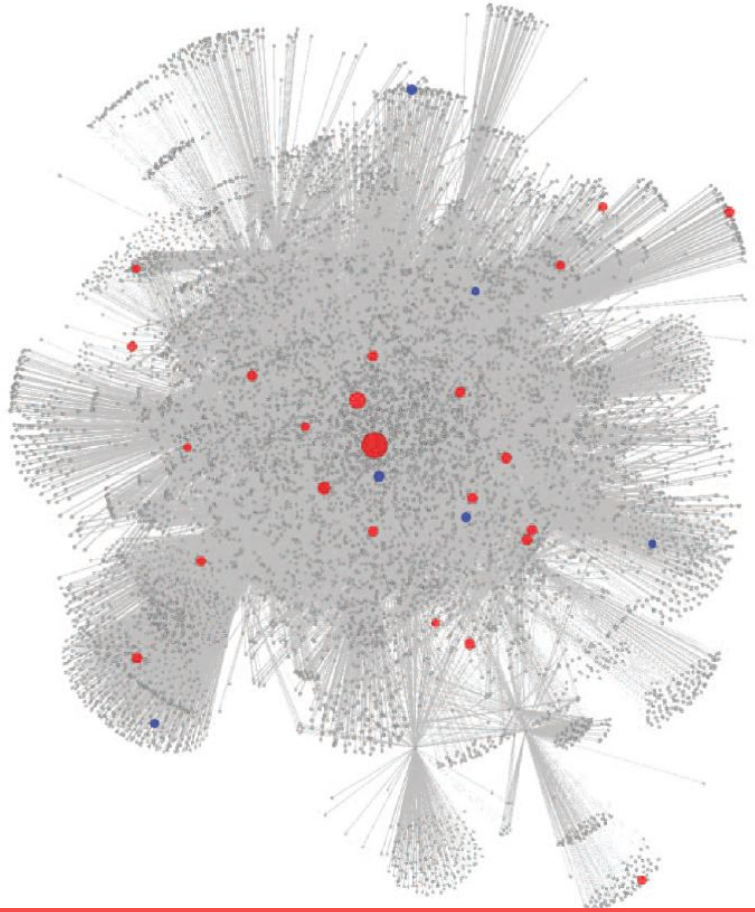
Highly differentiated sites are enriched for sites under positive selection



High Derived Allele Frequency in Asian

HighD in promoters

HighD is missense

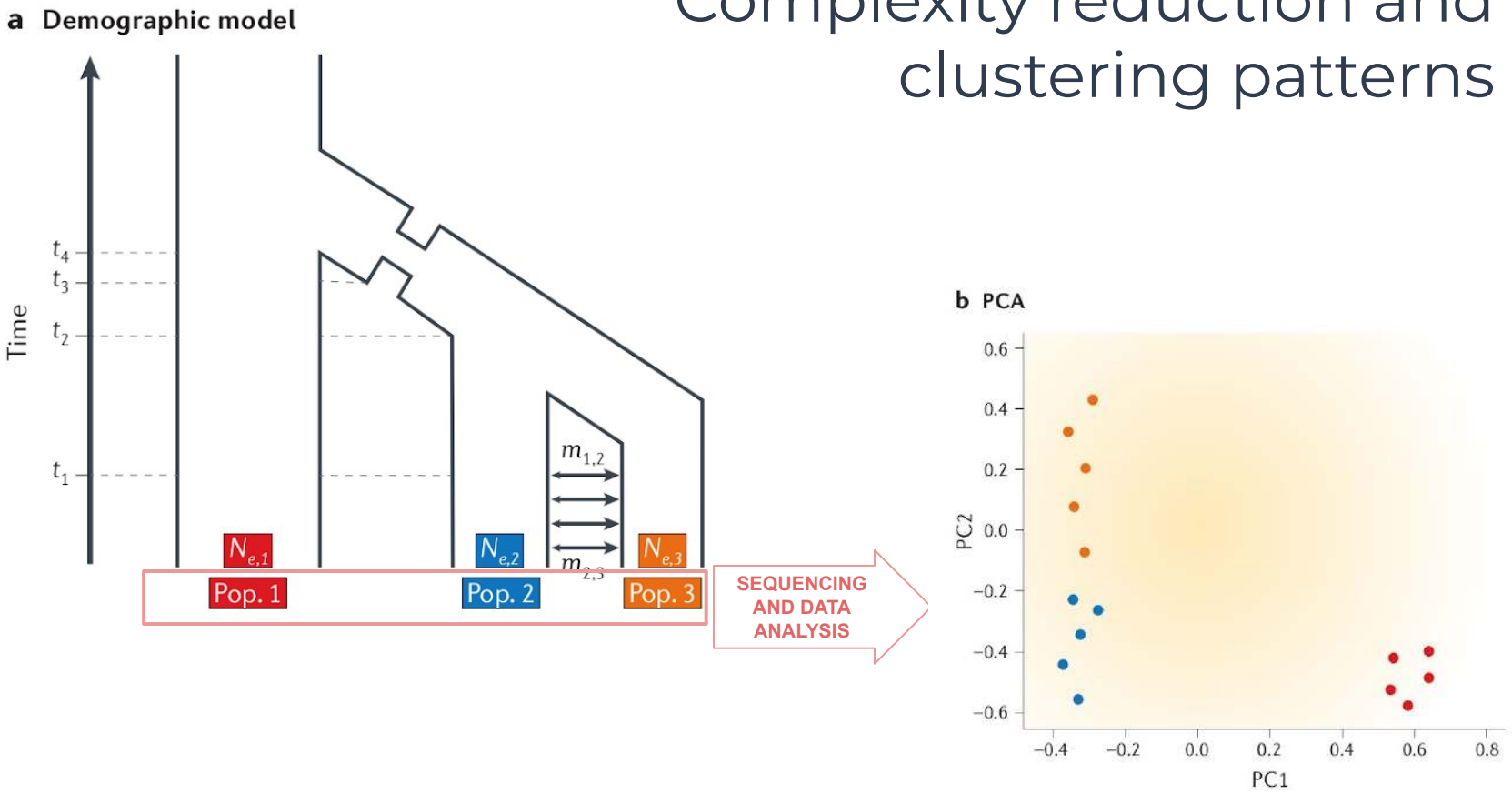


Highly differentiated sites are found in promoters of gene in central nodes

Kurana*, Fu*, Colonna* , 2012 Science

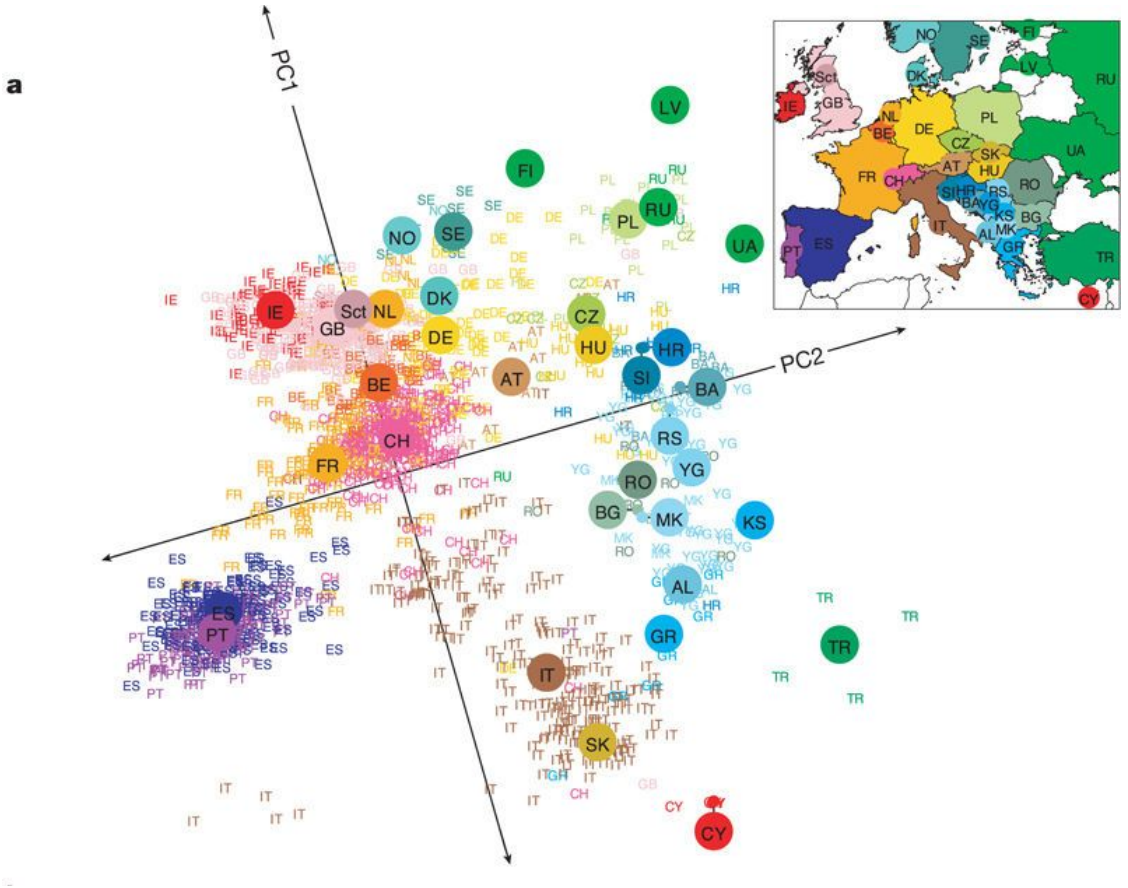
PCA: compact representation

Complexity reduction and clustering patterns



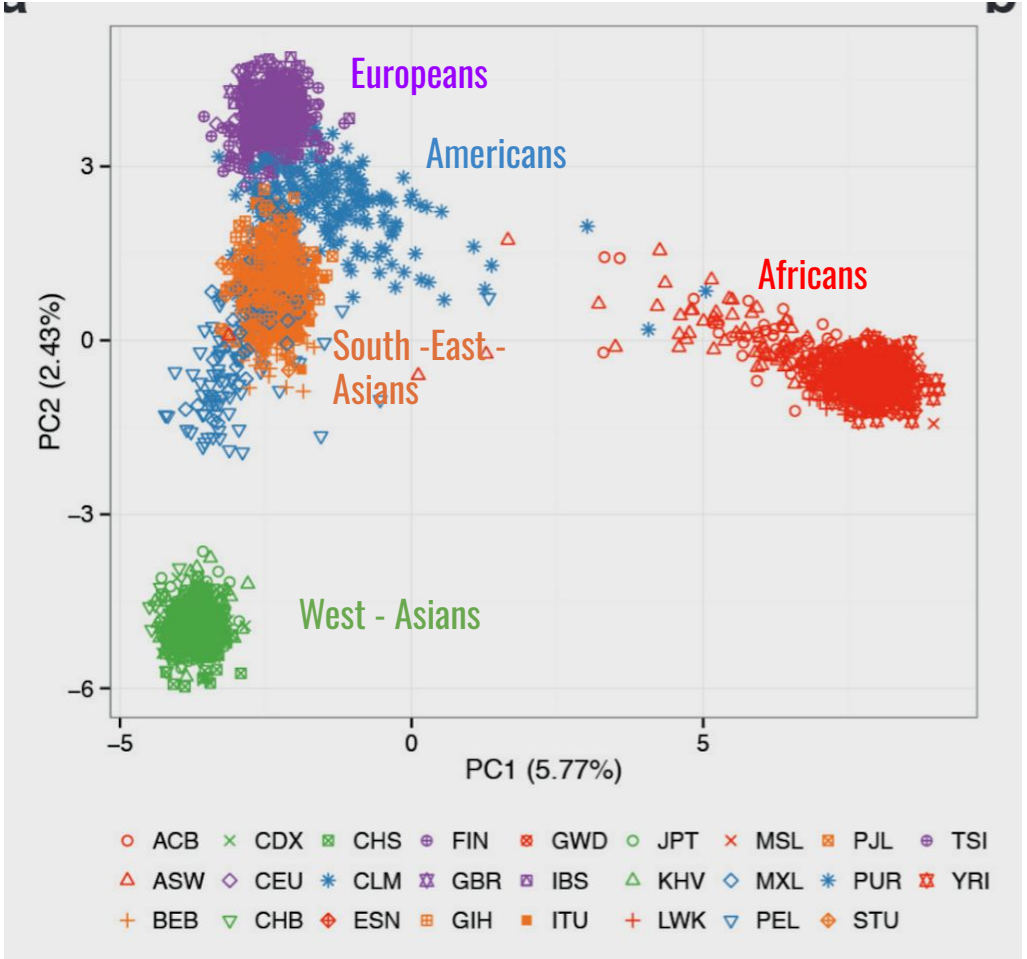
Schraiber JG, Akey JM. Nat Rev Genet. 2015 Dec;16(12):727-40. doi: 10.1038/nrg4005

Genes mirror geography within Europe



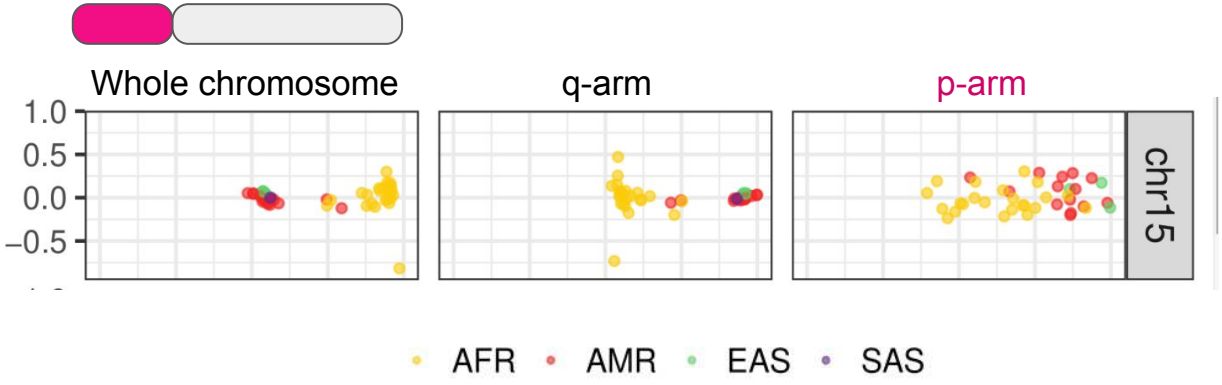
Novembre J, Johnson T, Bryc K, et al. 2008;456(7218):98-101. doi:10.1038/nature07331.

Genes mirror geography everywhere!

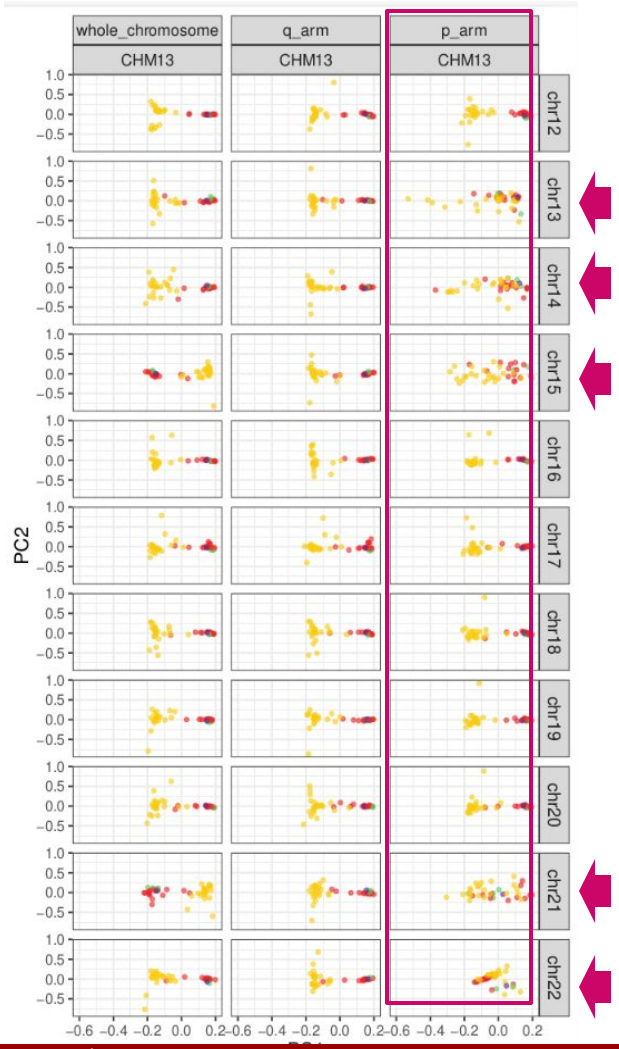


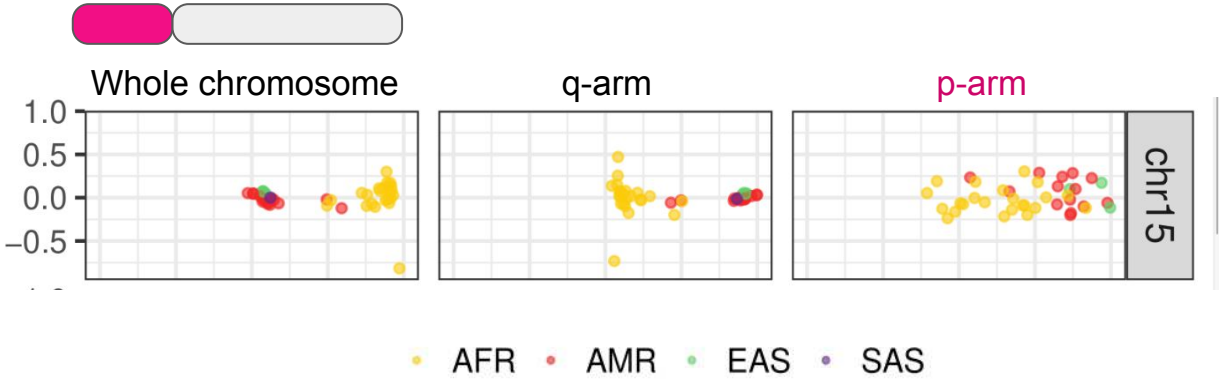
A Auton et al. Nature 526, 68-74 (2015) doi:10.1038/nature15393

Population stratification breaks down in acrocentric p-arms

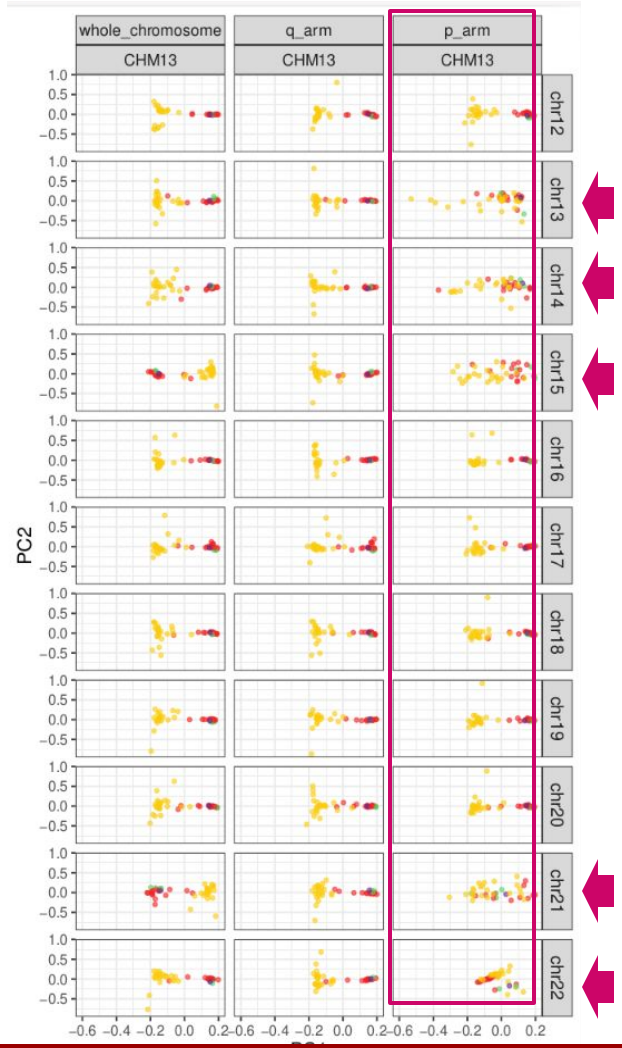


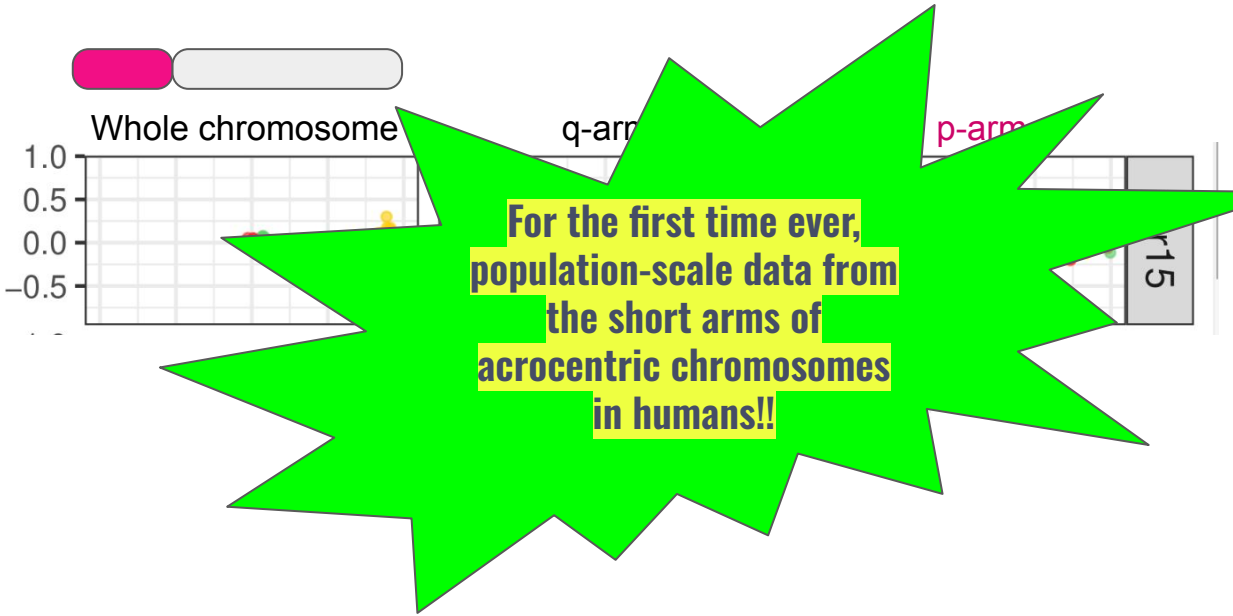
Population stratification breaks down in acrocentric p-arms



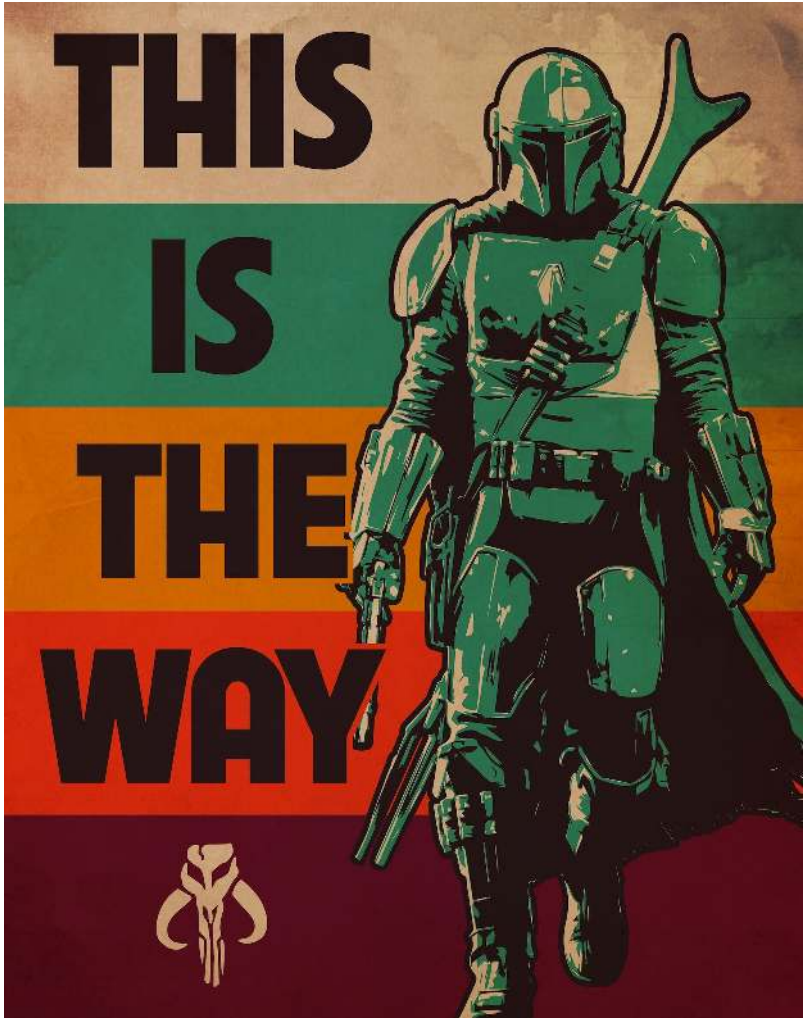


Population stratification breaks down in acrocentric p-arms





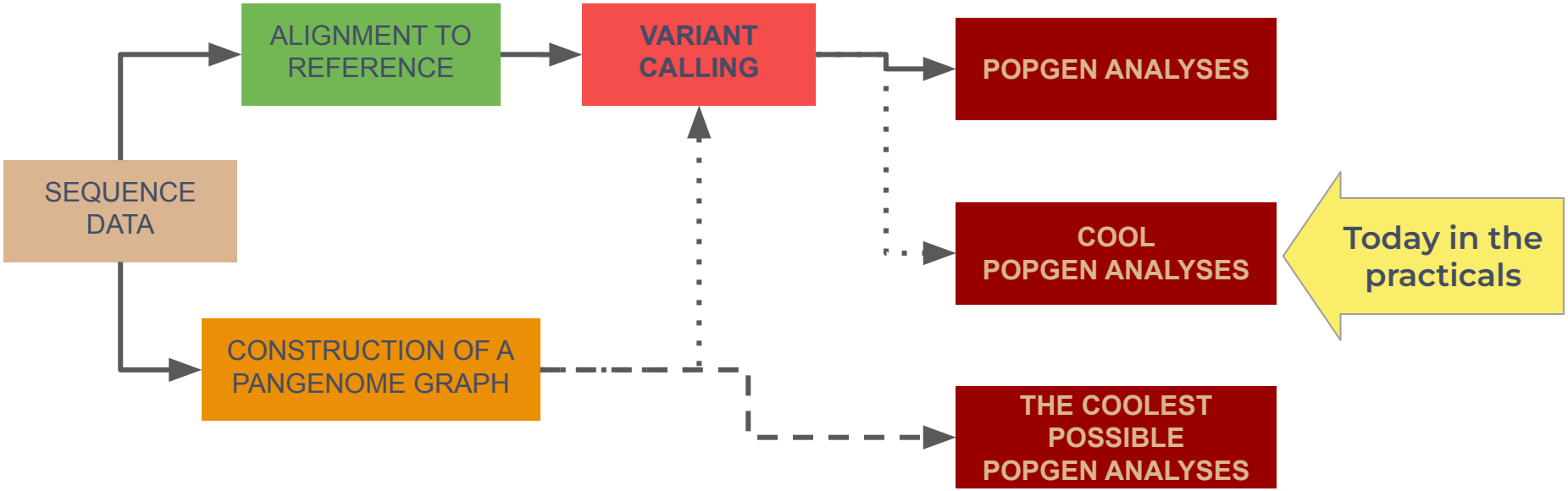

THE UNIVERSITY OF TENNESSEE
HEALTH SCIENCE CENTER
THE UNIVERSITY OF
TENNESSEE
HEALTH SCIENCE CENTER



YOUR TURN!

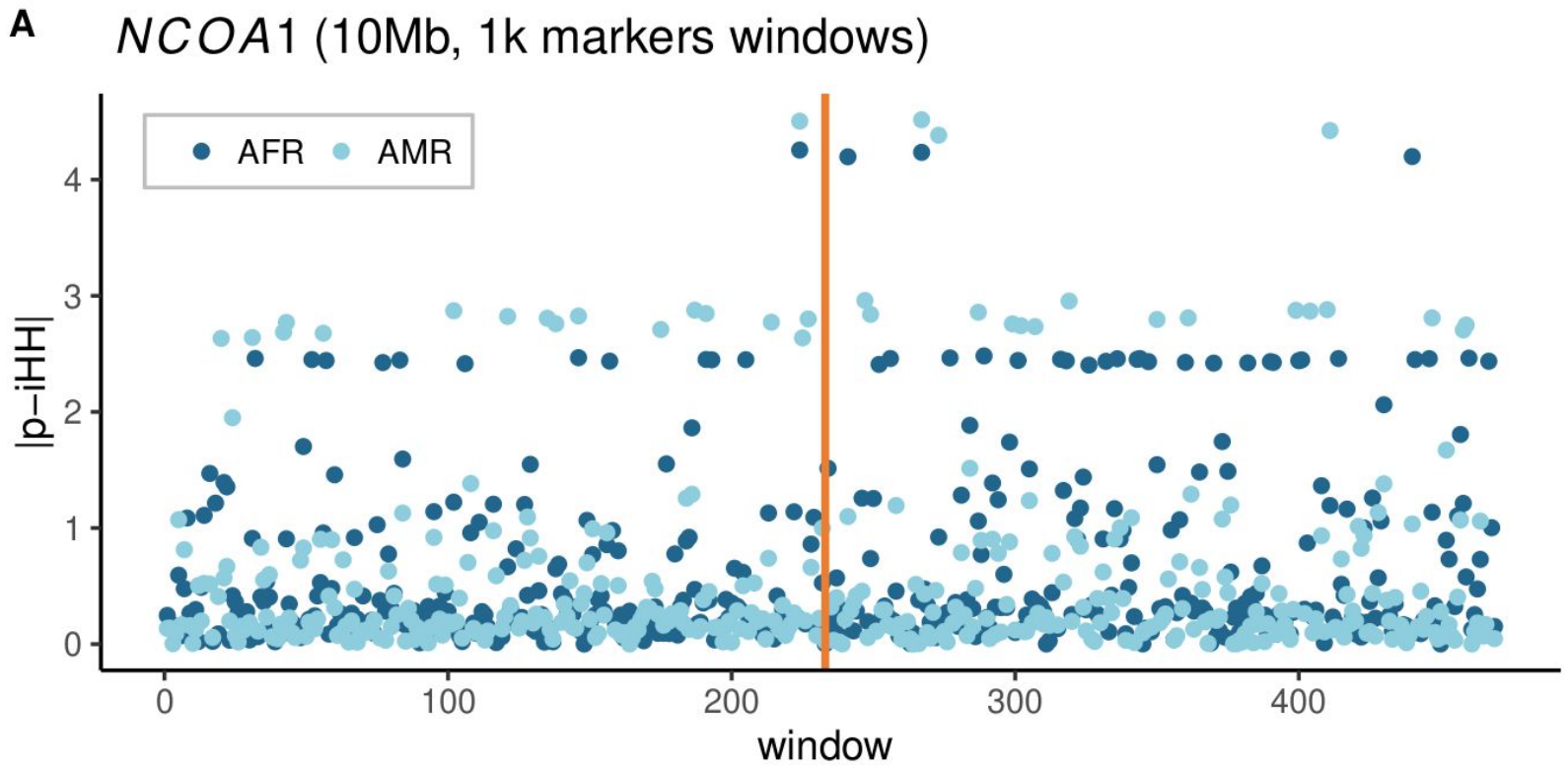
Move to the tutorial

(Traditional) genomics vs pangenomics

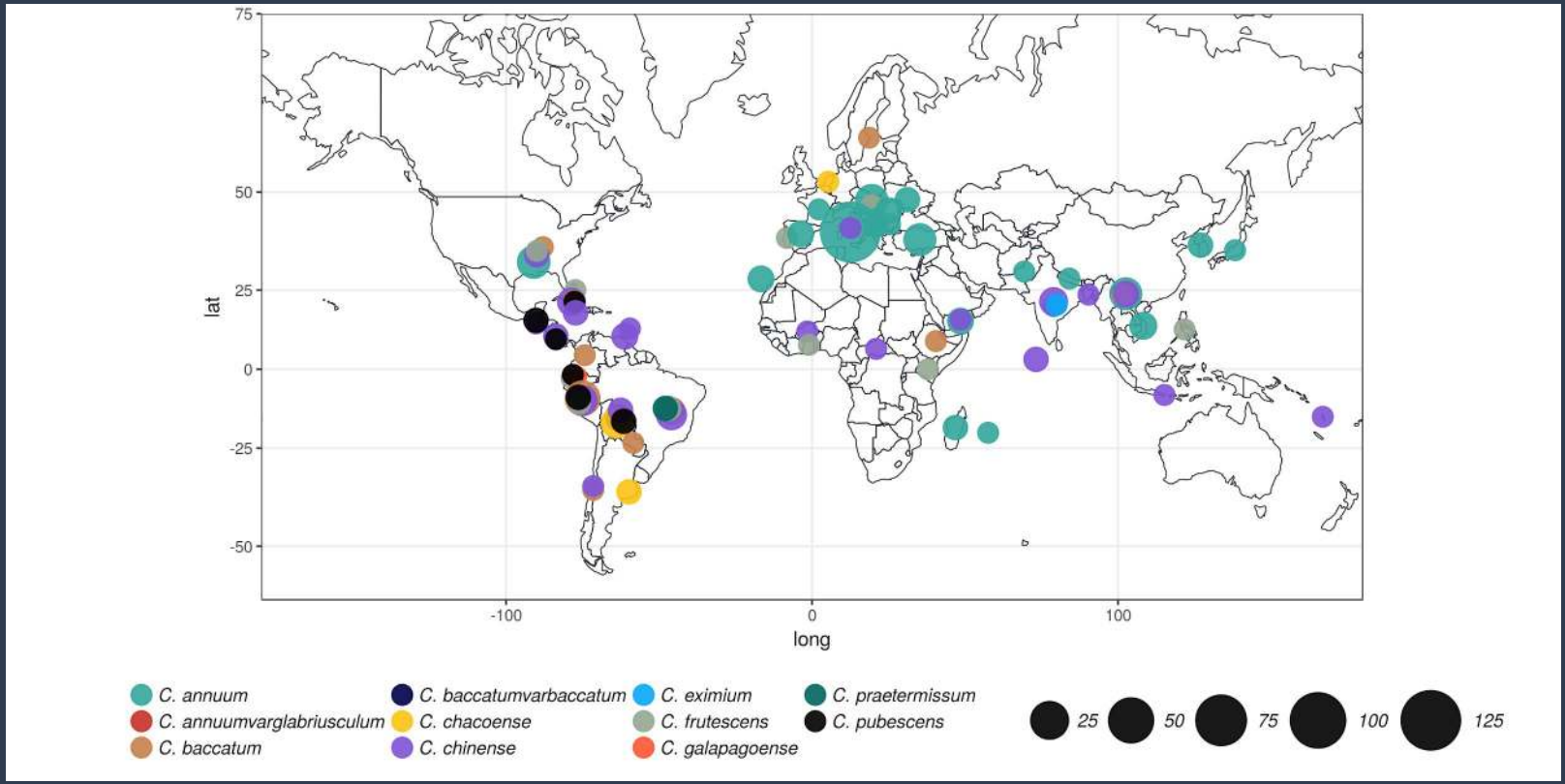


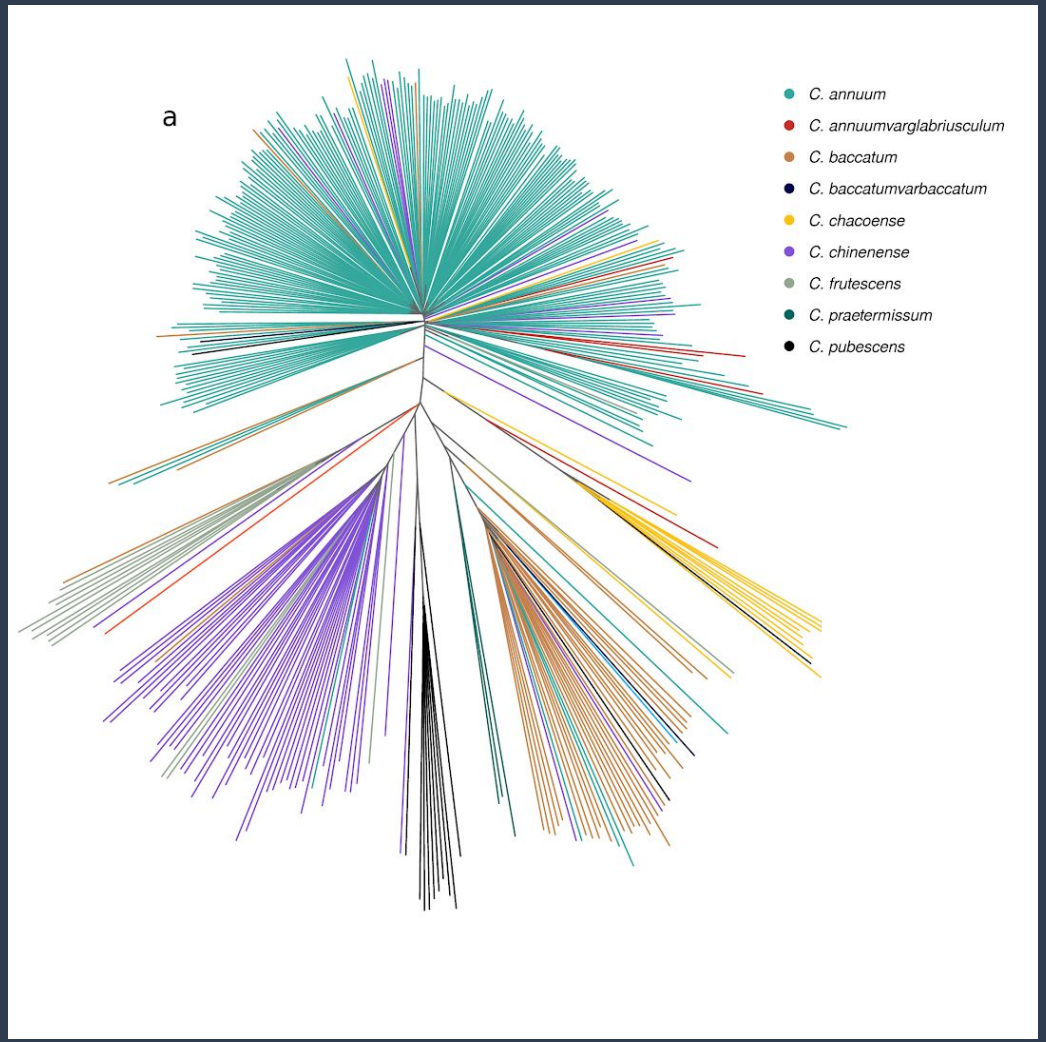
PRACTICALS: https://github.com/ColonnaLab/popgen-evomics2024

My first popgen analysis made directly from a variation graph …



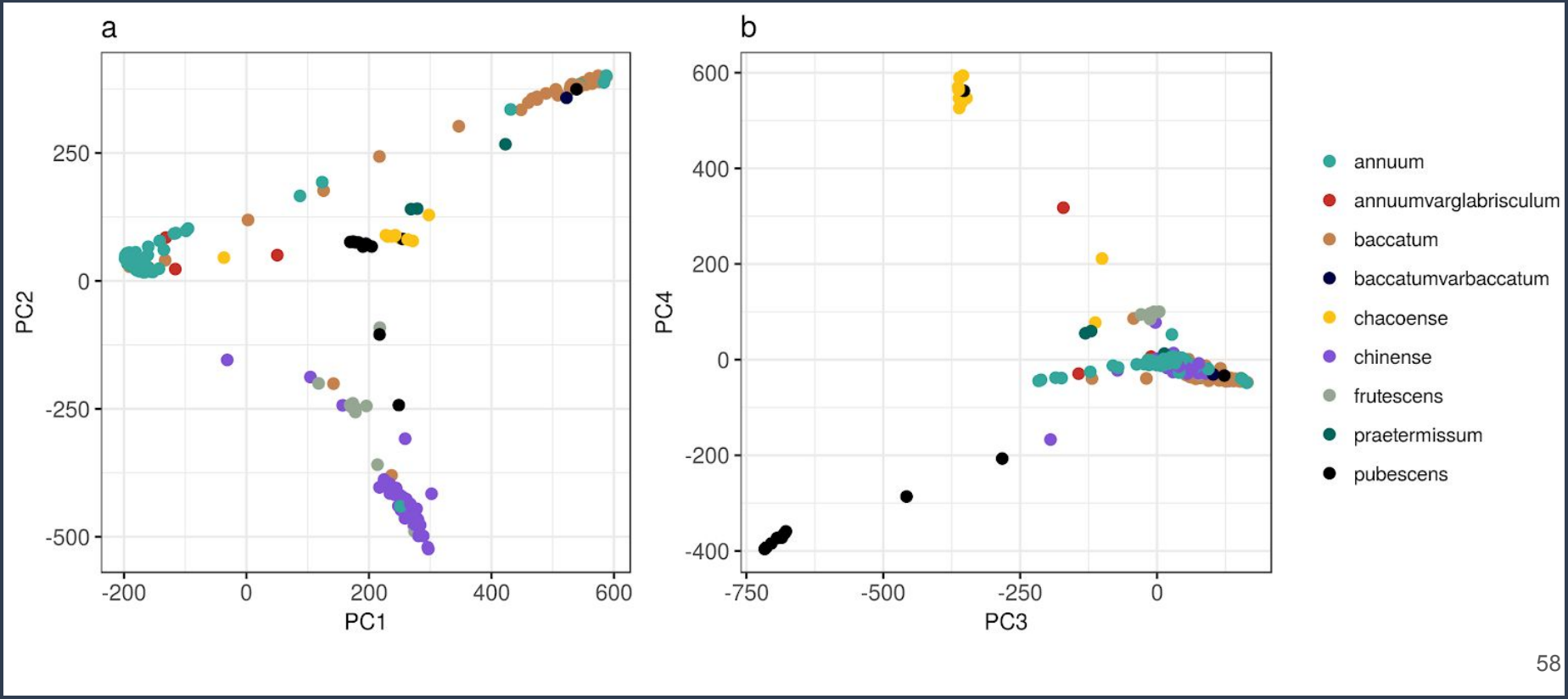
373 peppers, 11 species, 51 countries



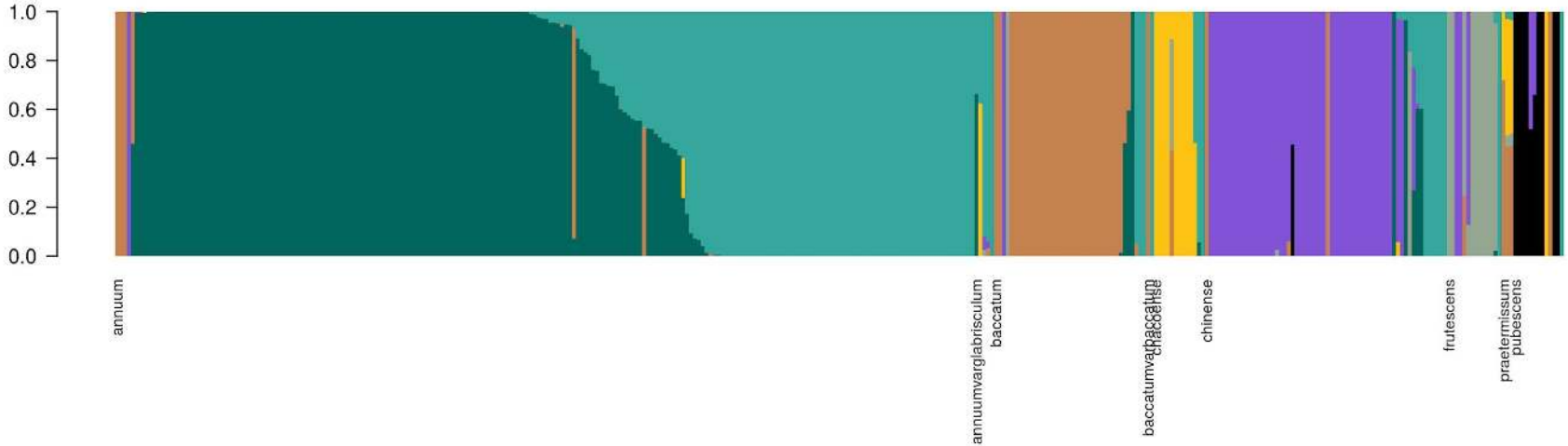


Clustering by species

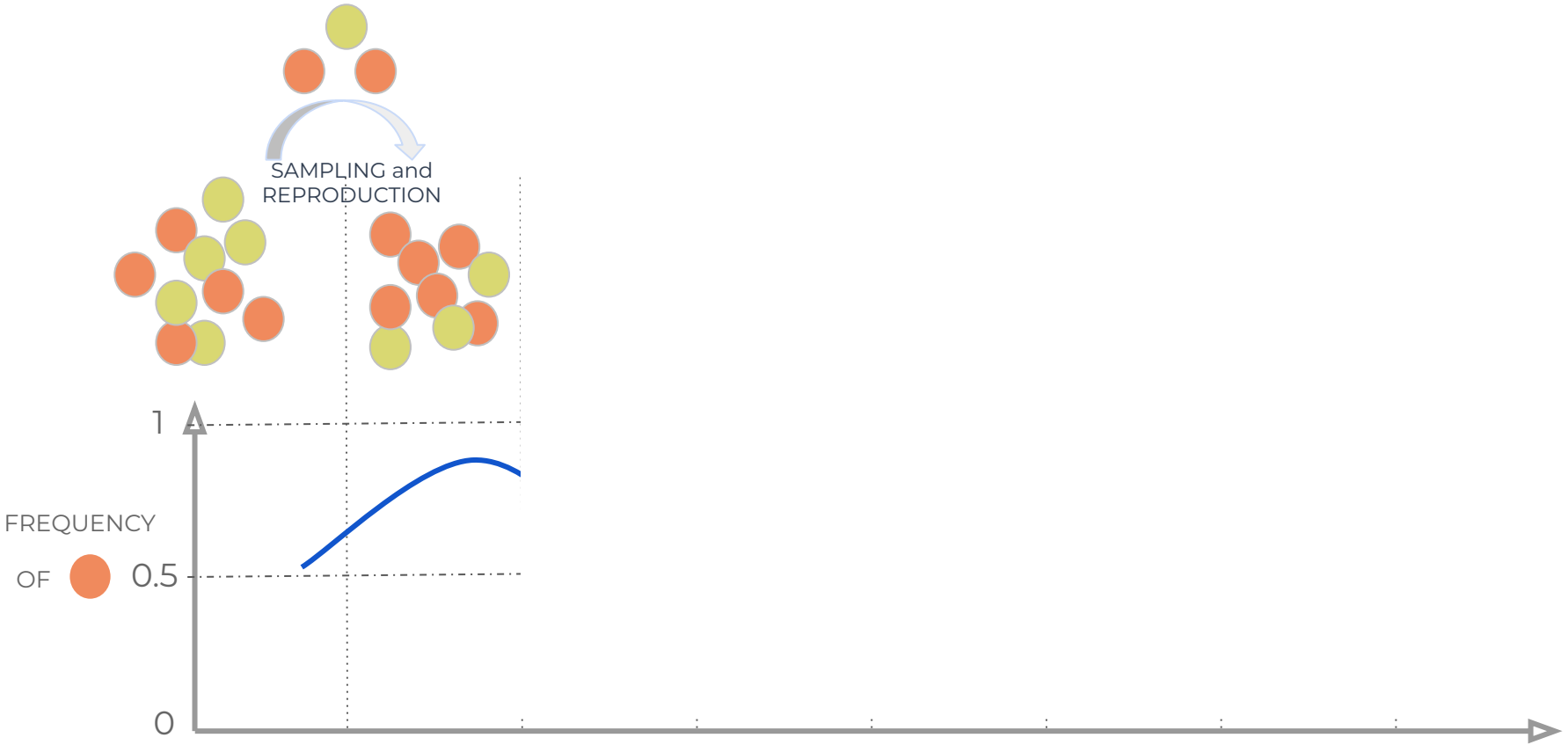
Clustering by species with some admixture

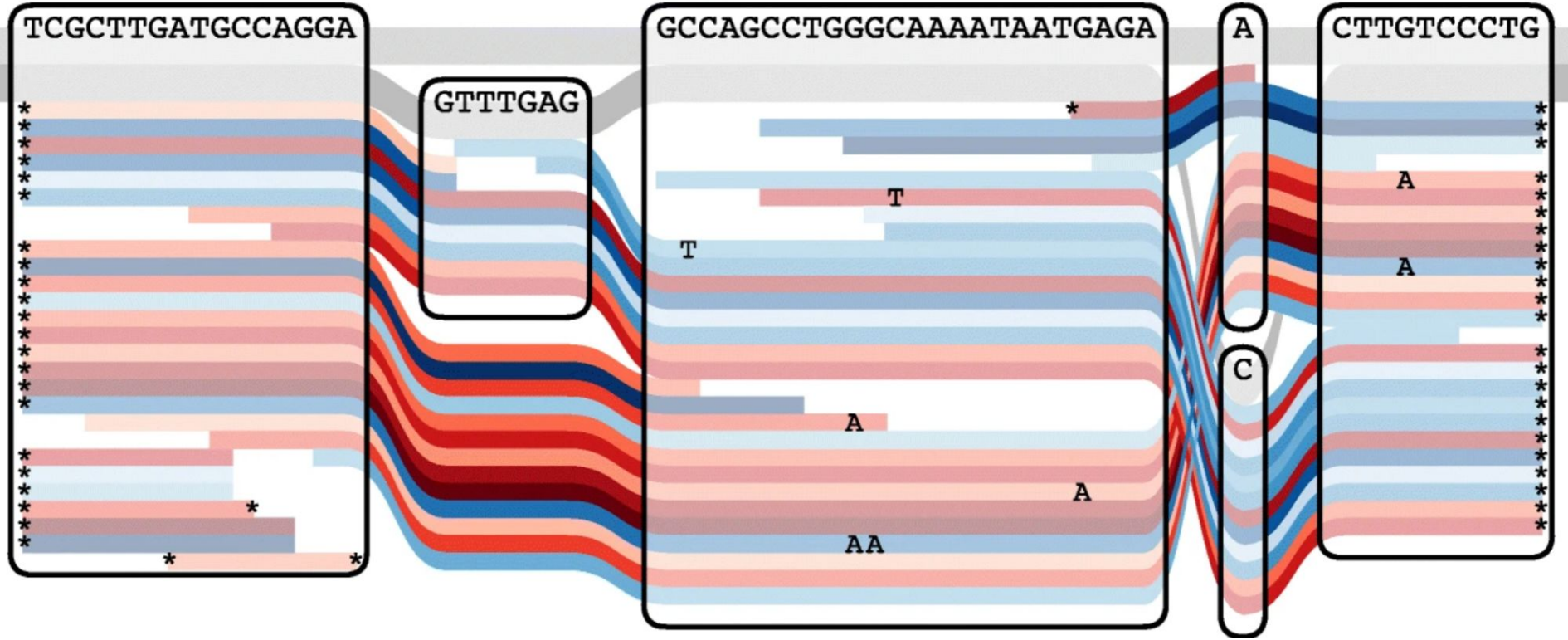


Admixture patterns



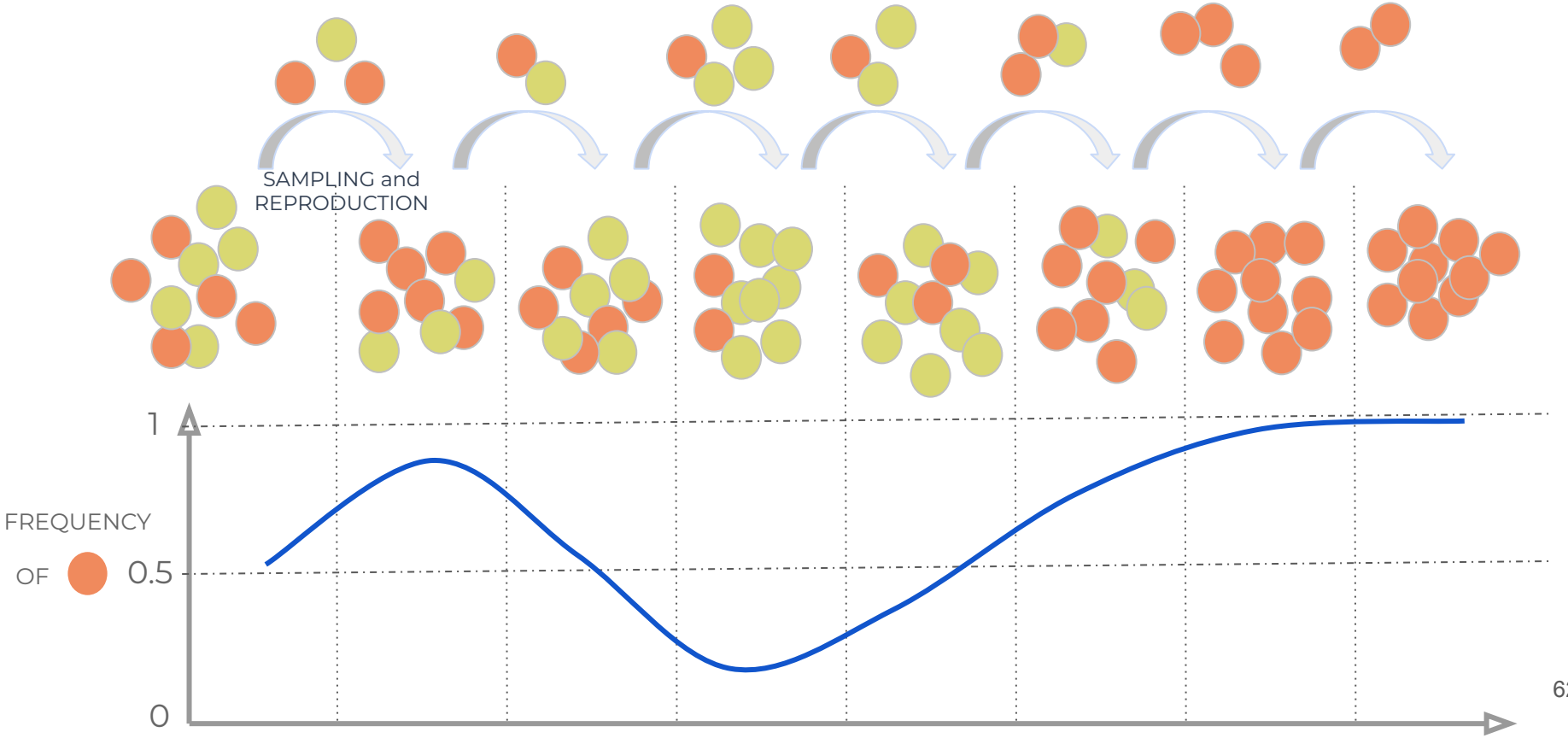
Genetic Drift



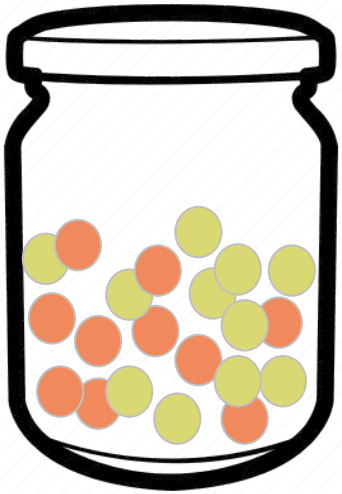


reads aligned to a variation graph

Genetic Drift

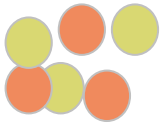


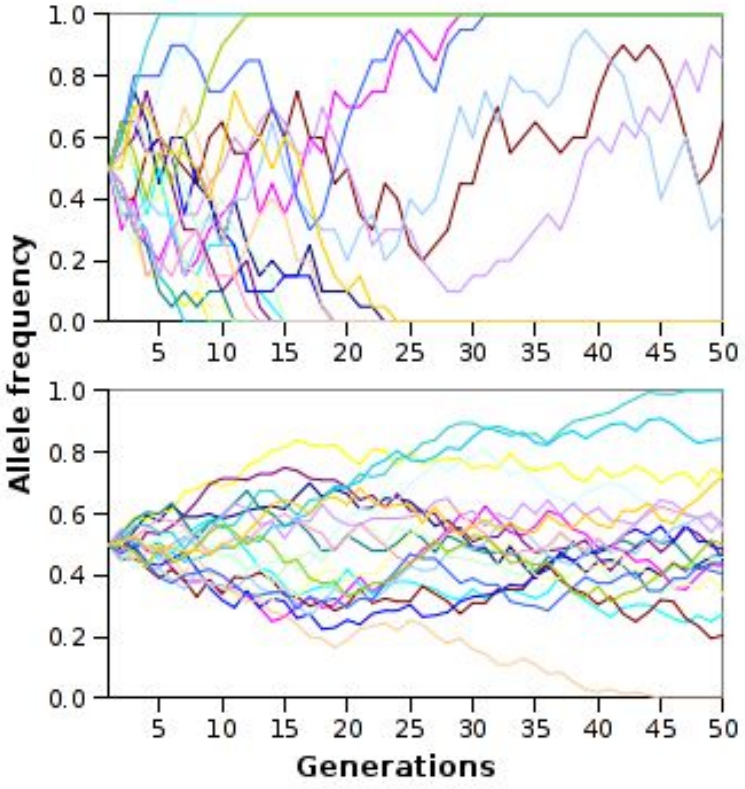
Genetic Drift and population size



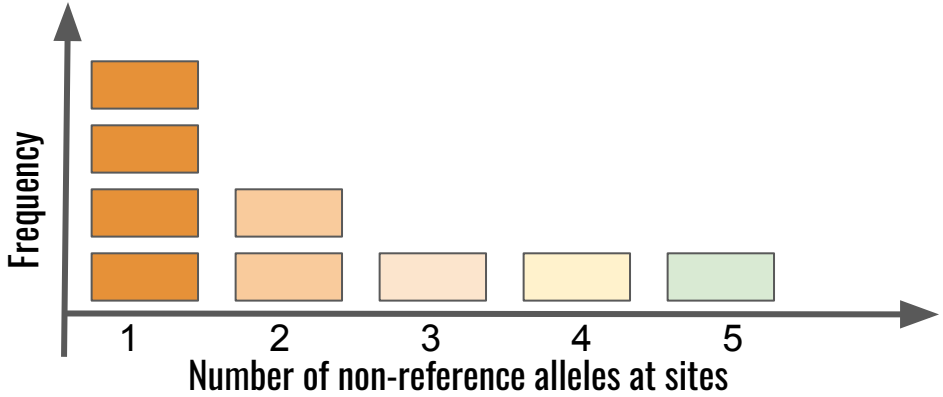
Small population

Large population





Site Frequency Spectrum



	locus1	locus2	locus3	locus4	locus5	locus6	locus7	locus8	locus9
Total	1	5	4	2	1	2	1	3	1
ID1	0|0	1|0	0|0	1|0	0|0	1|0	0|0	1|0	0|0
ID2	0|0	1|0	1|1	1|0	0|0	0|0	0|0	0|0	1|0
ID3	0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0
ID4	1|0	1|1	1|0	0|0	0|0	1|0	0|0	1|1	0|0
ID5	0|0	0|0	1|0	0|0	0|0	0|0	1|0	0|0	0|0