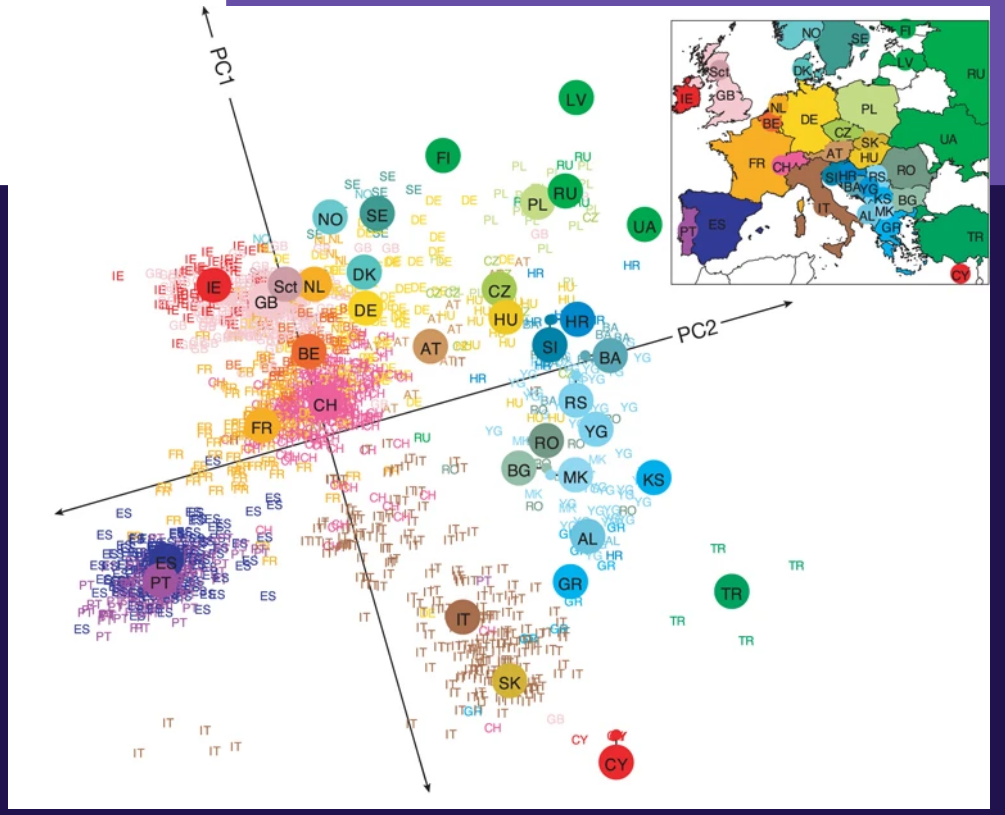
Physalia

Courses

Population structure and connectivity

27/05/2025



Physalia course

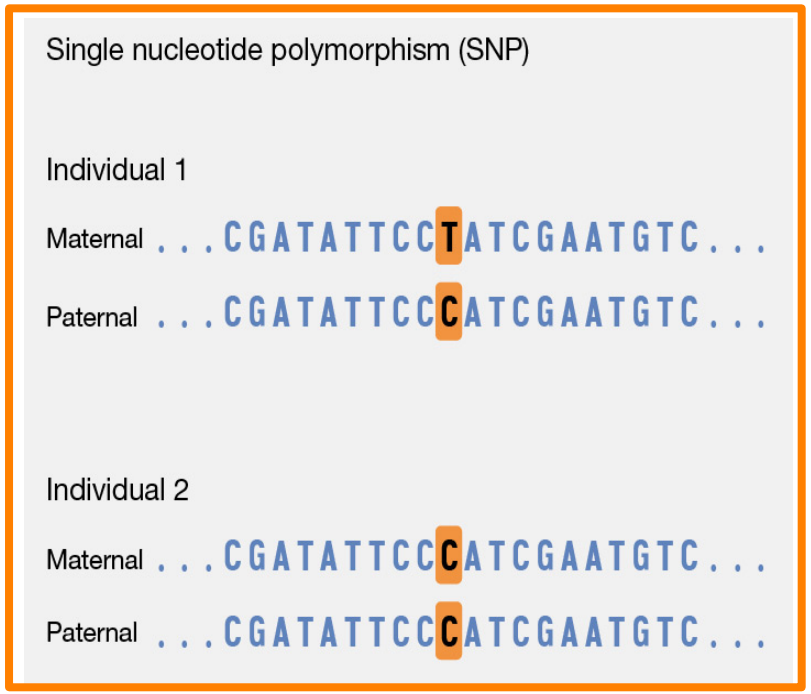
Yann Bourgeois, Thibault Leroy

Goals for today's lecture

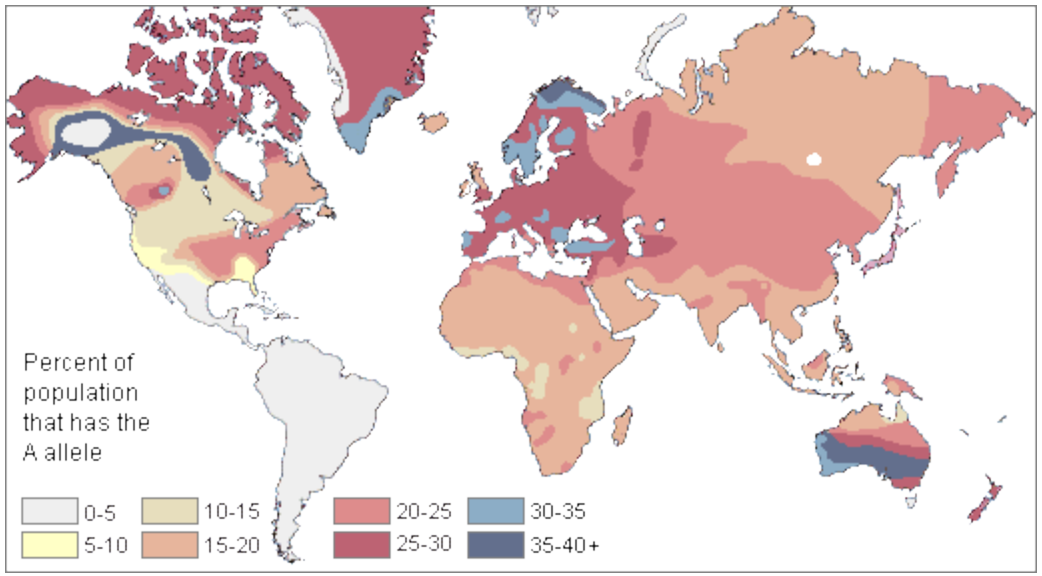
- What are some of the usual programs used to infer structure?
- On which popgen concept are they based?
- How can we detect introgression and gene flow?
- A more quantitative view of structure and gene flow: where and when?
- And throughout: a few limitations to keep in mind.

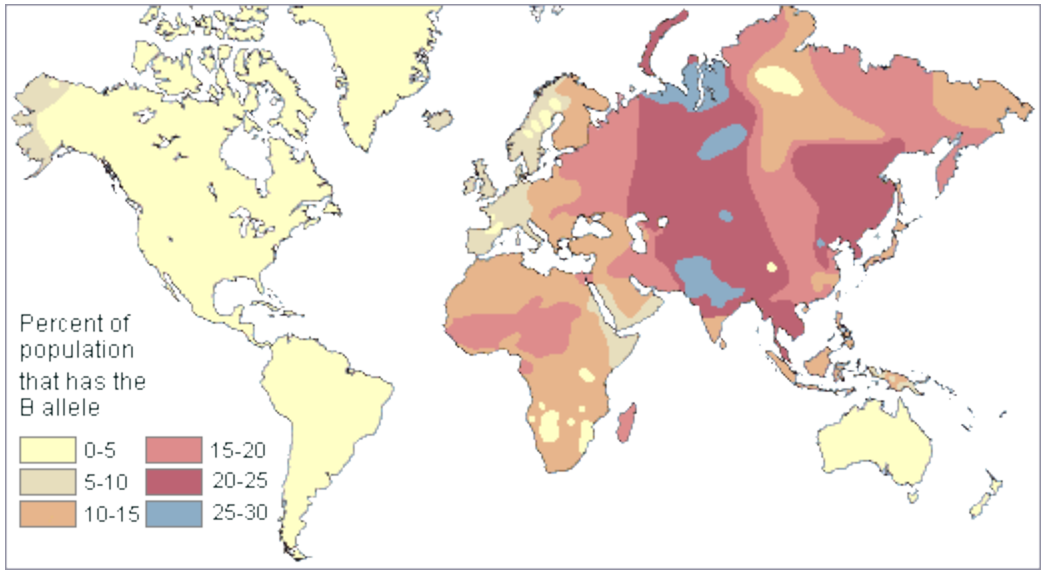
Polymorphisms: definition

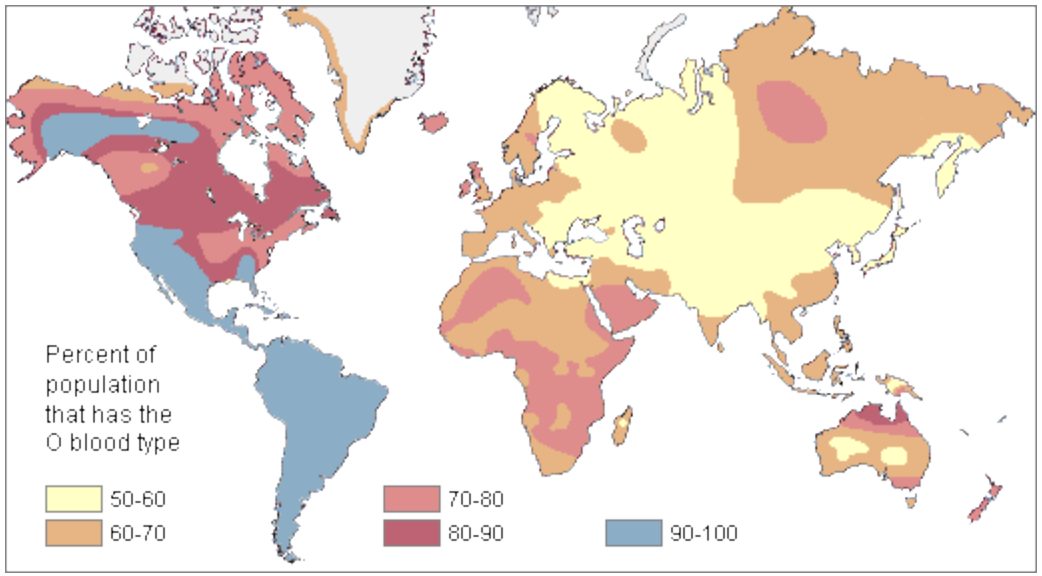
"The presence of two or more variant forms of a specific DNA sequence that can occur among different individuals or populations."



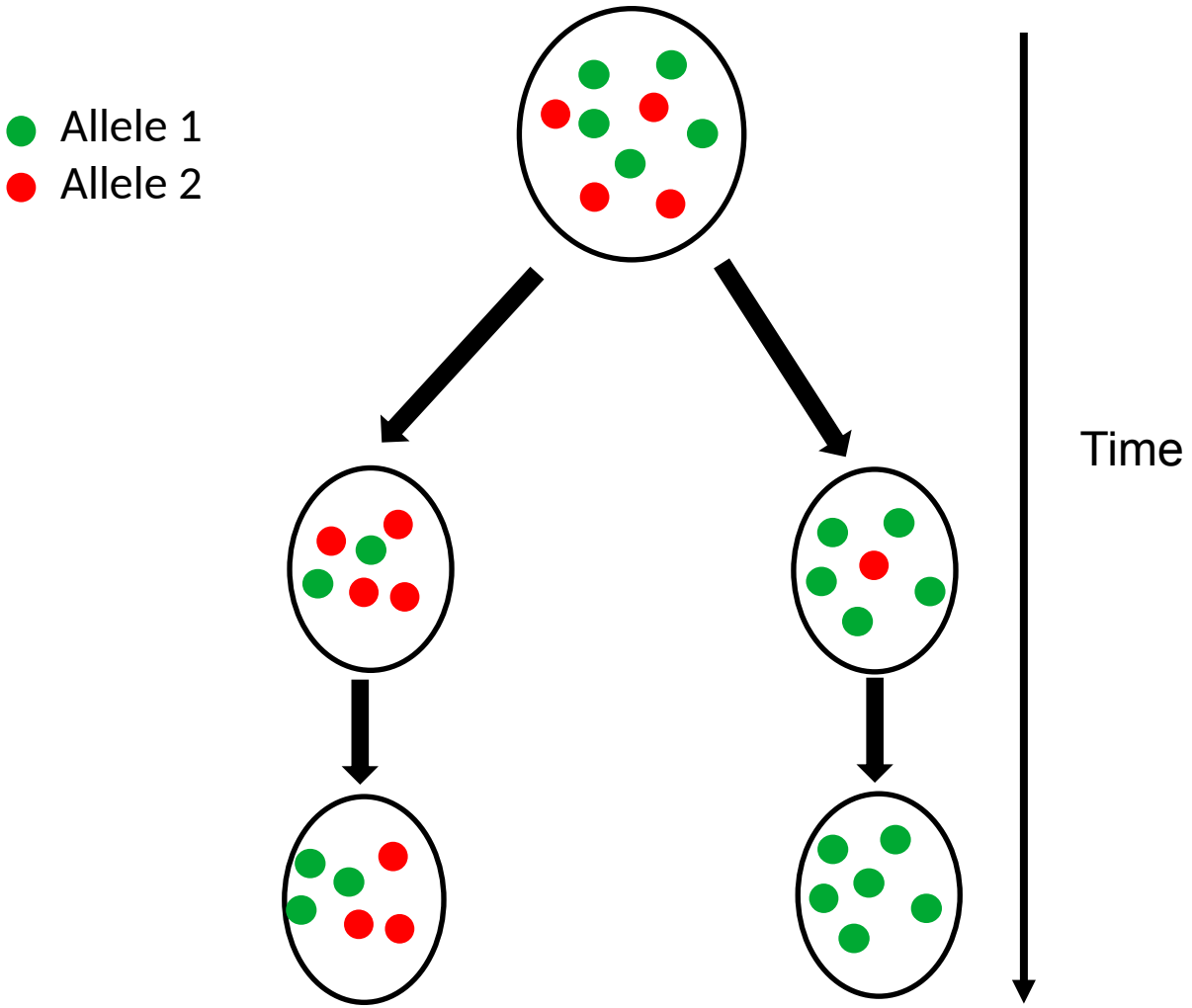
Polymorphisms segregate in populations







Why are polymorphisms segregating?



Question: What can lead to allele frequencies differences between these two populations?

Charlesworth, B., Charlesworth, D. Population genetics from 1966 to 2016. Heredity 118, 2–9 (2017). https://doi.org/10.1038/hdy.2016.55

Back to the basics: Hardy-Weinberg equilibrium

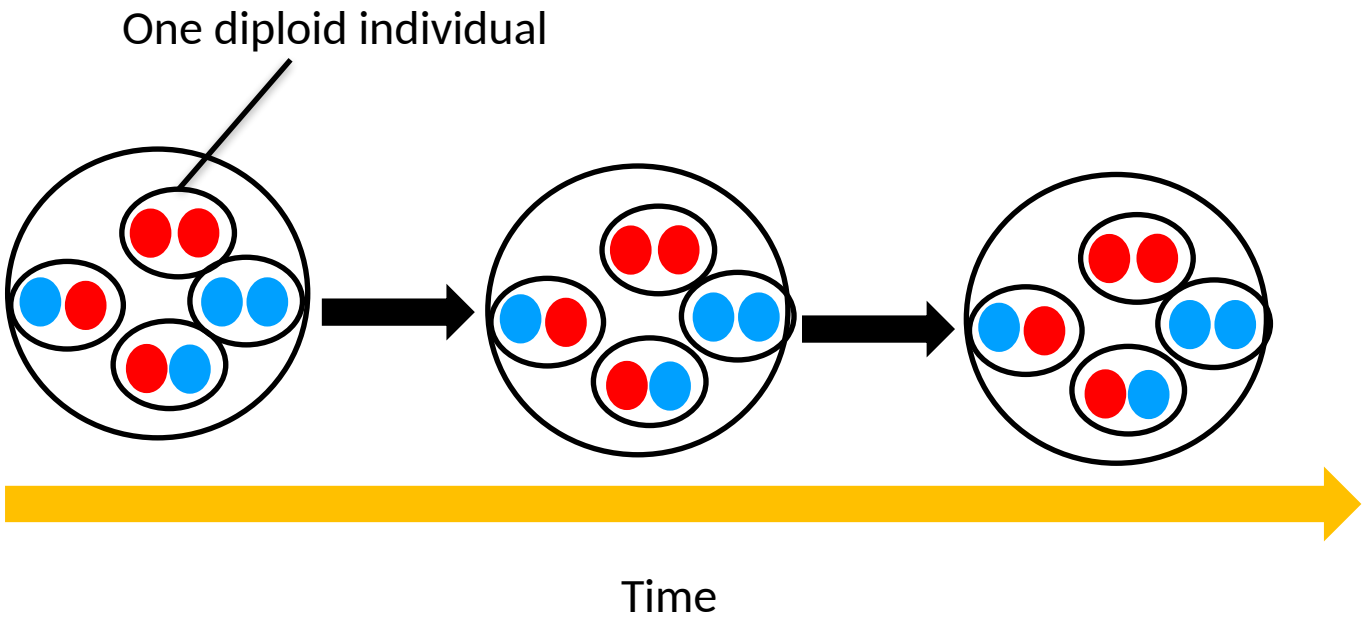
- In order to study polymorphisms and understand them we need to understand why genetic variation (in this case polymorphisms) don't get blended out of the population group.
- In other words, when will all individuals in a population carry the same phenotype/genotype? What can maintain polymorphism?

- Discovery in 1908 by Godfrey Hardy & Wilhelm Weinberg proved that individual proportions of a phenotype in a population will remain constant from generation to generation as long as the following assumptions are met:

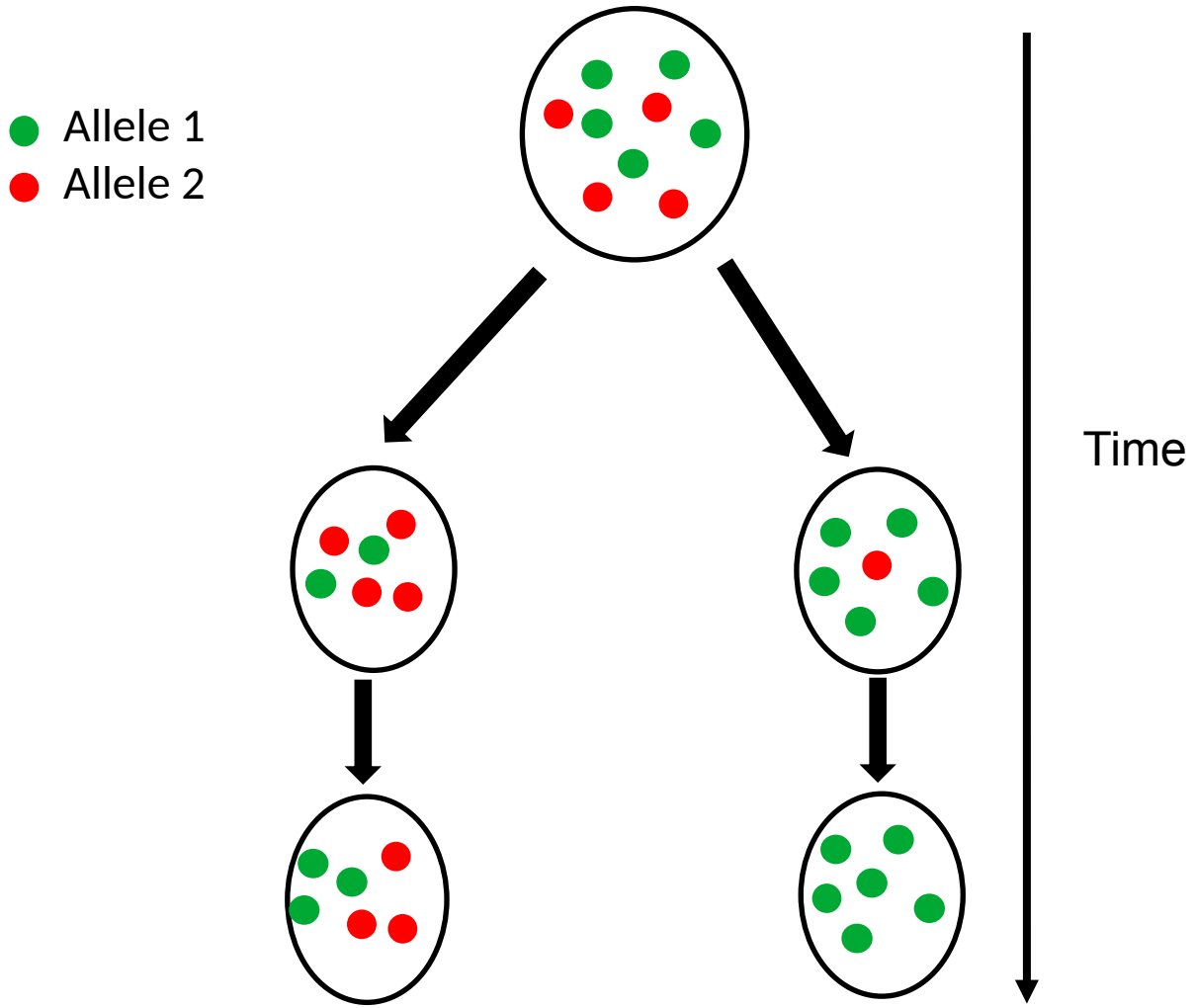
How to maintain polymorphisms forever

- No mutations take place
- No genes are transferred to or from other sources ie. no immigration or emigration
- Mating is random (individuals don't chose mates based on their phenotype or genotype
- The population size is very large (actually infinite, no genetic drift)
- No selection occurs

If those assumptions are met the genotype proportions do not change and they are said to be in Hardy-Weinberg equilibrium.



Why are polymorphisms segregating?



Deviation from H-W equilibrium:

Population structure

Assortative mating

Bottlenecks

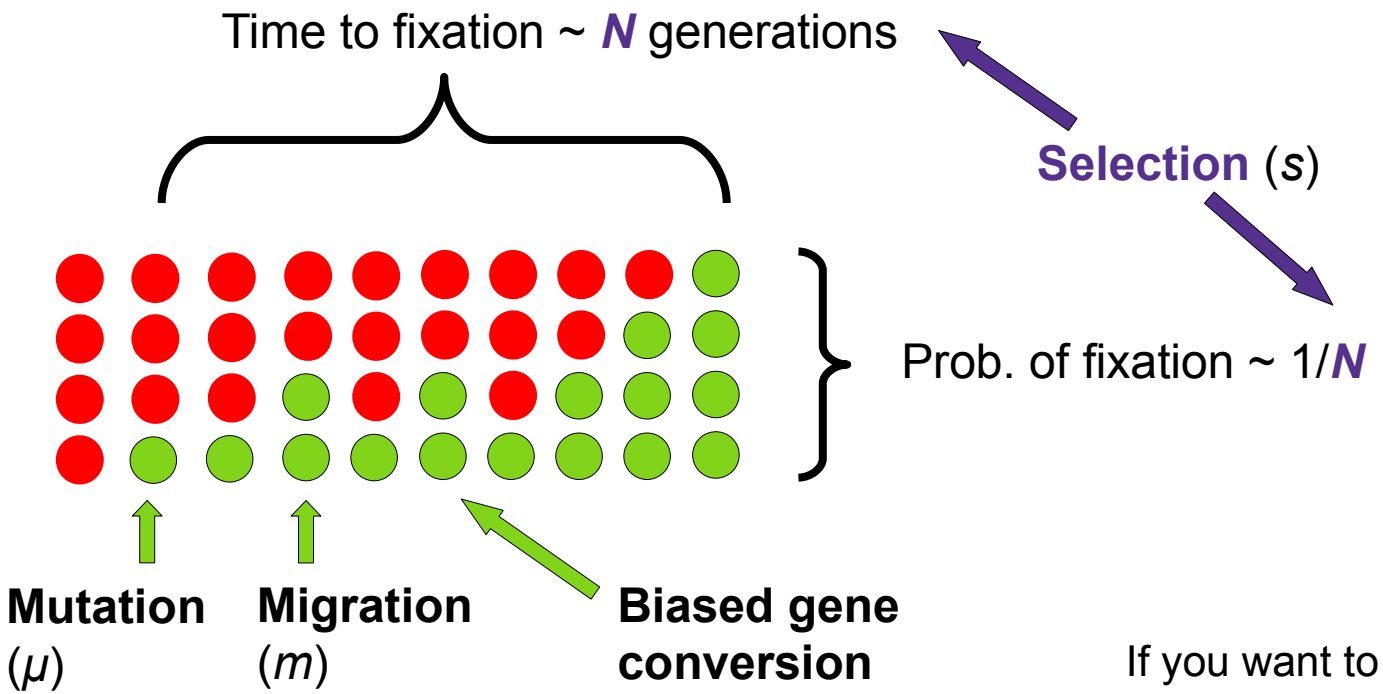
Recent selection

Overall Hardy-Weinberg equilibrium is robust to deviations as long as one does not examine long-term trends

Charlesworth, B., Charlesworth, D. Population genetics from 1966 to 2016. Heredity 118, 2–9 (2017). https://doi.org/10.1038/hdy.2016.55

Why are polymorphisms segregating?

- Population genetics: studying the changes in allele frequency over time and space.
- Established in the middle 20th century by Haldane, Wright, Fisher, Moran, and other mathematicians



- The (haploid) effective population size ***N*** is related to genetic drift
- An important product: ***N*** x *s* (for later).

If you want to play a bit with the concepts: learnPopGen (check QR code)

How many polymorphisms in a genome? Why?


Tomoko Ohta

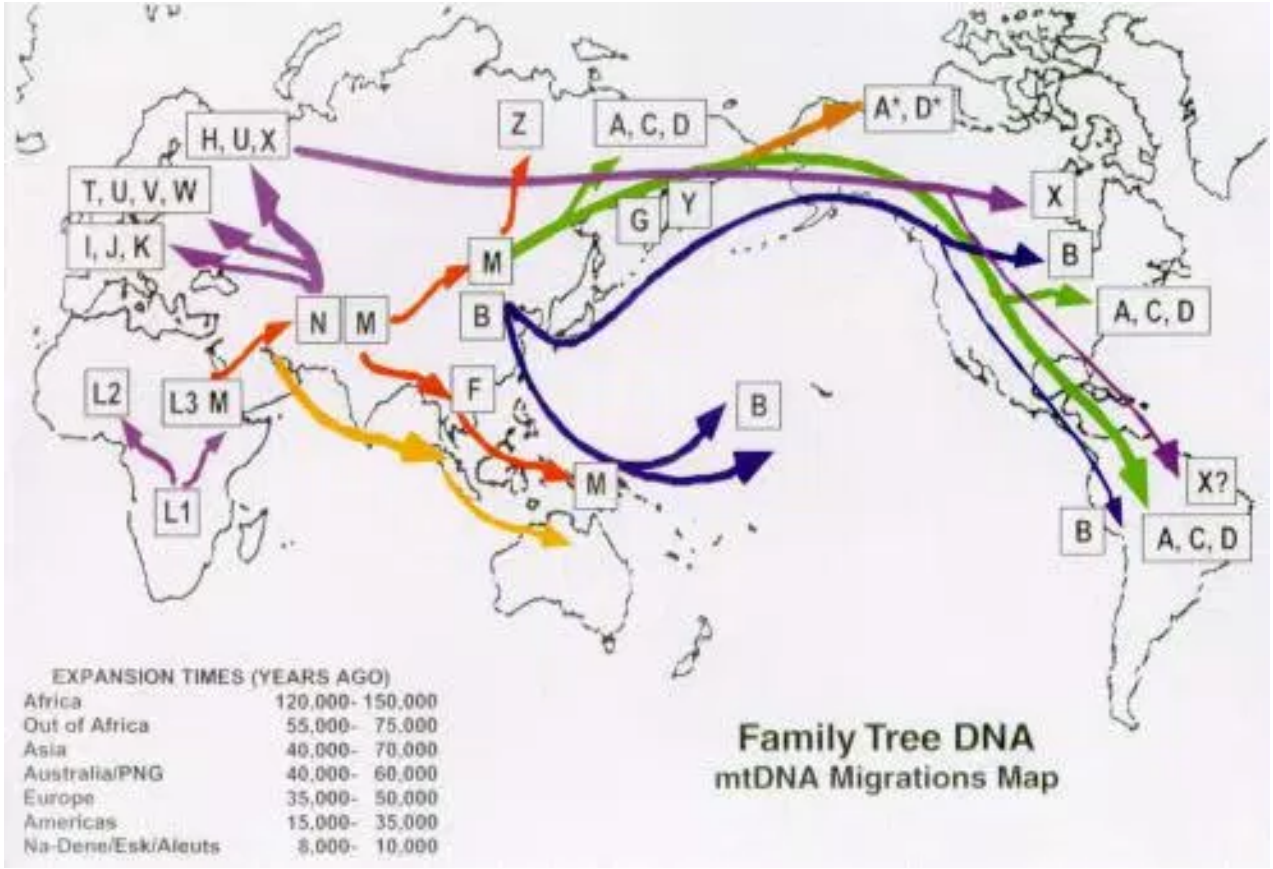
- The absolute number of polymorphic variants

- Depends on the mutation rate (e.g. for Single Nucleotide Polymorphisms, SNPs).
- Depends on the **history of a species, drift, migration.**
- Since the work from Kimura and Ohta on the (**near)-neutral theory**, the consensus is that most polymorphisms are **not** under strong selection.
- More local effects: selection (for later, read about R. Lewontin meanwhile).

Read more at:

DEALING WITH GENOME-WIDE DATA

Describing population structure: one marker



- Single markers have been very useful
- But with new technologies we can examine millions of “independent” markers along the genome
- We can also detect which polymorphisms segregate in a way that is not consistent with others (e.g. selection)

Methods to study population structure

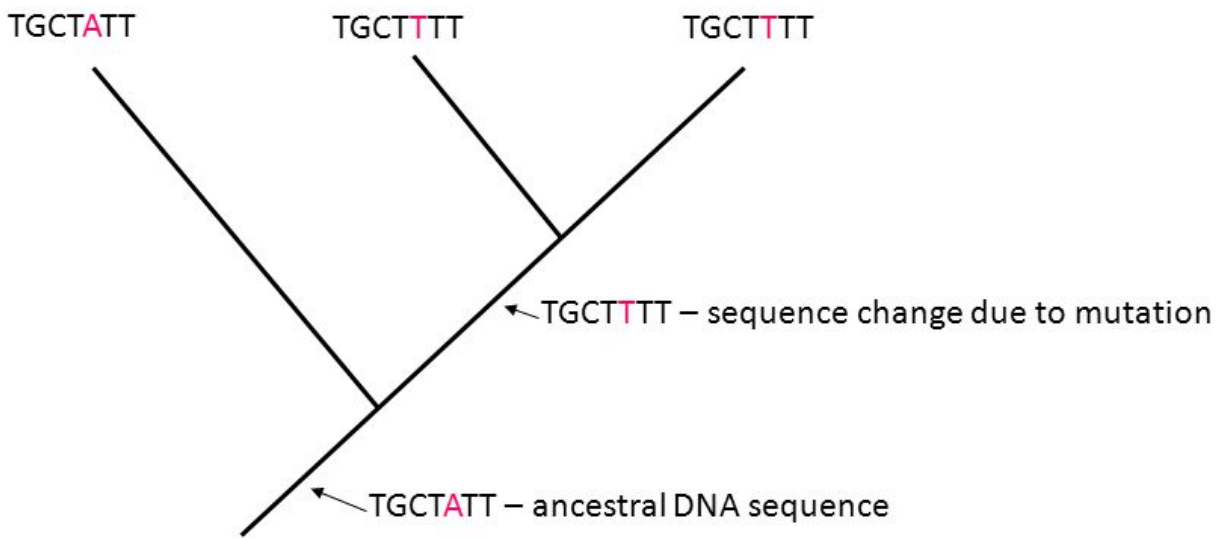
- How to deal with thousands or millions of markers?

- Measures of pairwise relatedness, pedigrees, genetic distances, summary statistics (F-statistics, ABBA-BABA…)
- Dimensionality-reduction
- Clustering of individuals (can be model-based).

WITHIN POPULATIONS: RELATEDNESS

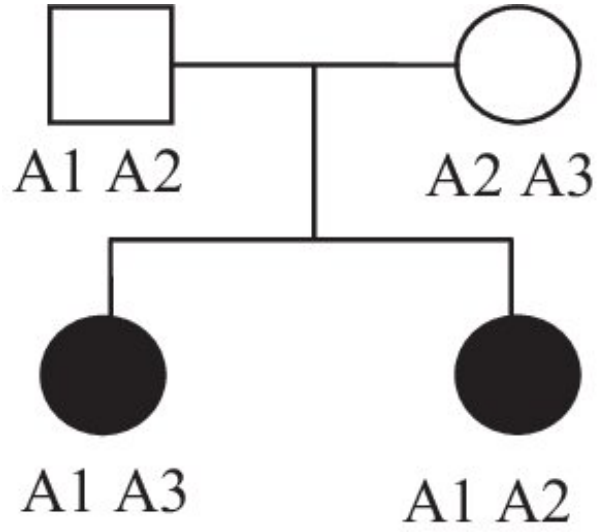
Identity by descent

- Shared polymorphisms are expected because of familial relationships
- Patterns of identity by descent (IBD) are useful for pedigree analyses
- Sometimes, only a few markers are enough
- Example: mitochondrial DNA is maternally inherited, while the Y chromosome is paternally inherited



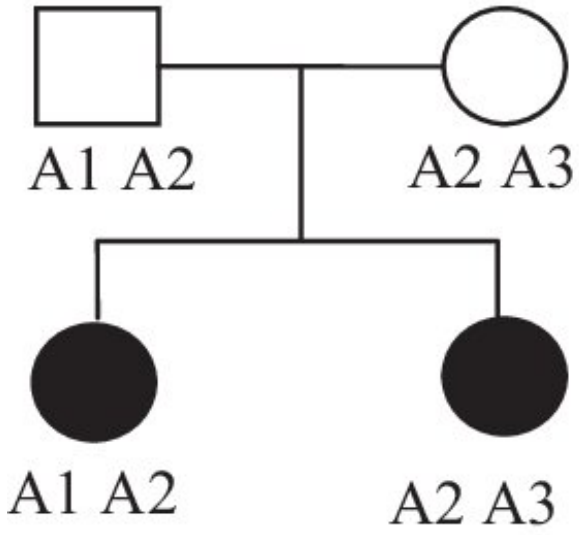
Identity by descent

- Shared polymorphisms are expected because of familial relationships
- Patterns of identity by descent (IBD) are useful for pedigree analyses
- Sometimes, only a few markers are enough
- In many analyses we identify what is identical-by-state and tend to assume a common ancestor
- More risky for alleles than for haplotypes.



An example of IBD

Allele A1 has been inherited with certainty from the father and both daughters share the same allele A1



An example of IBS

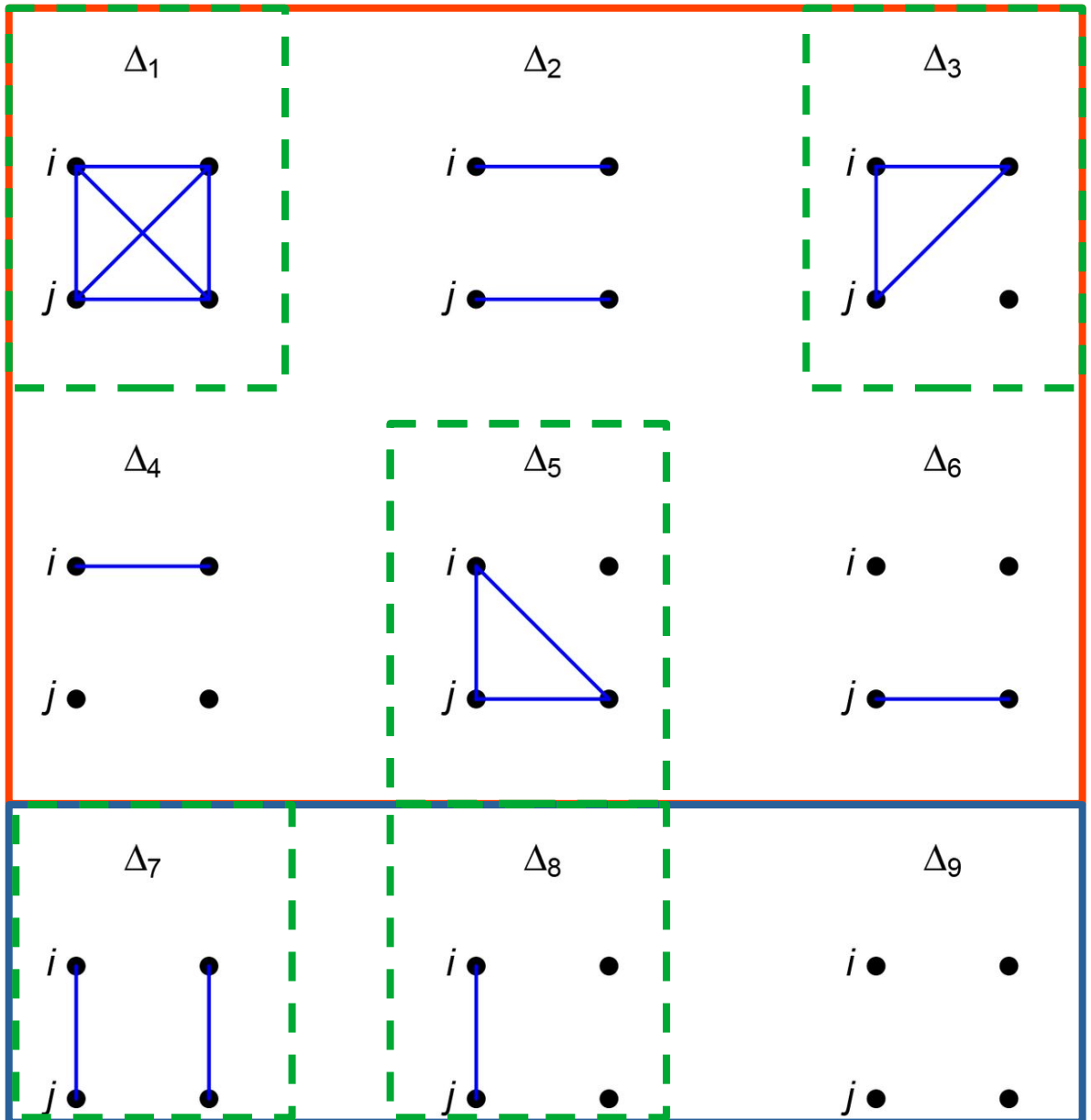
Allele A2 in the first daughter has been inherited from the mother, allele A2 in the second daughter has been inherited from the father

Jacquard's coefficients

- Two diploid individuals, i and j
- Probabilities for each configuration estimated from population allele frequencies
- Biased if population structure/admixture

$\Theta = \Delta 1 + 0.5*(\Delta 3+\Delta 5+\Delta 7) + 0.25*\Delta 8)$

$R_{AB} = \Delta 1 + \Delta 7 + 0.75*(\Delta 3+\Delta 5) + 0.5*\Delta 8)$

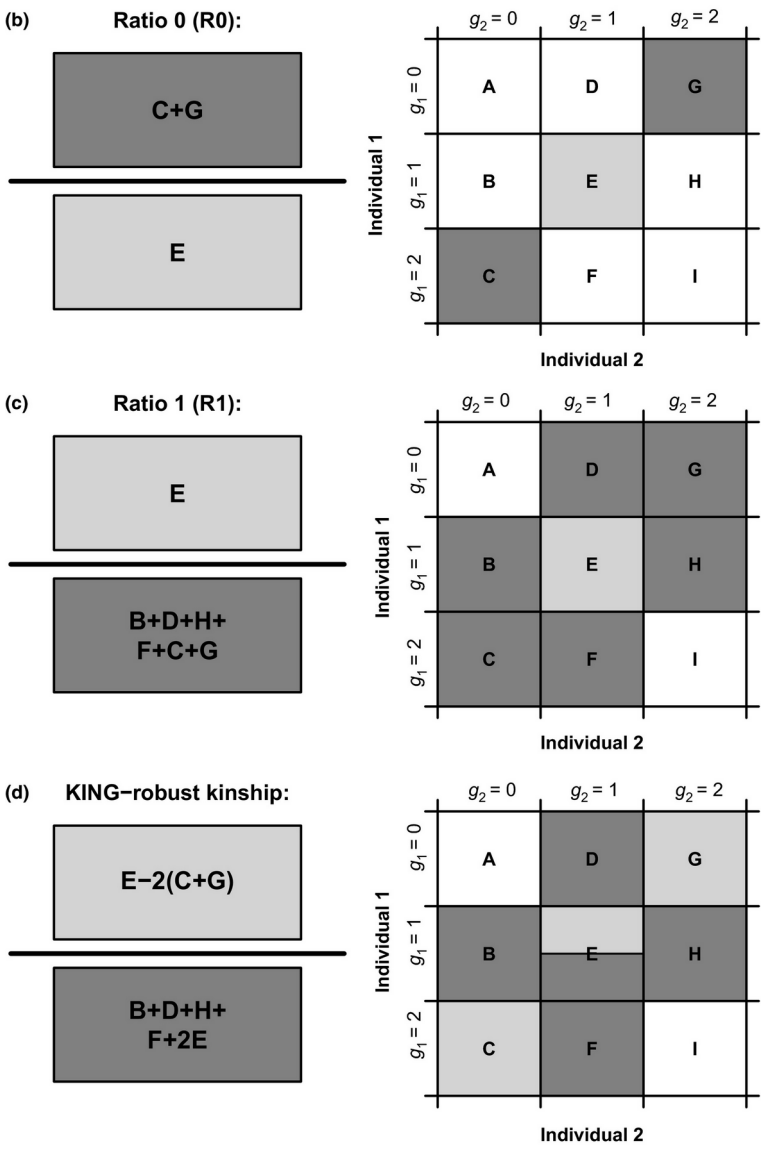


Inbreeding (identical by descent within an individual)

Configurations found in close relatives (shared identity-by-descent)

Only configurations found without inbreeding

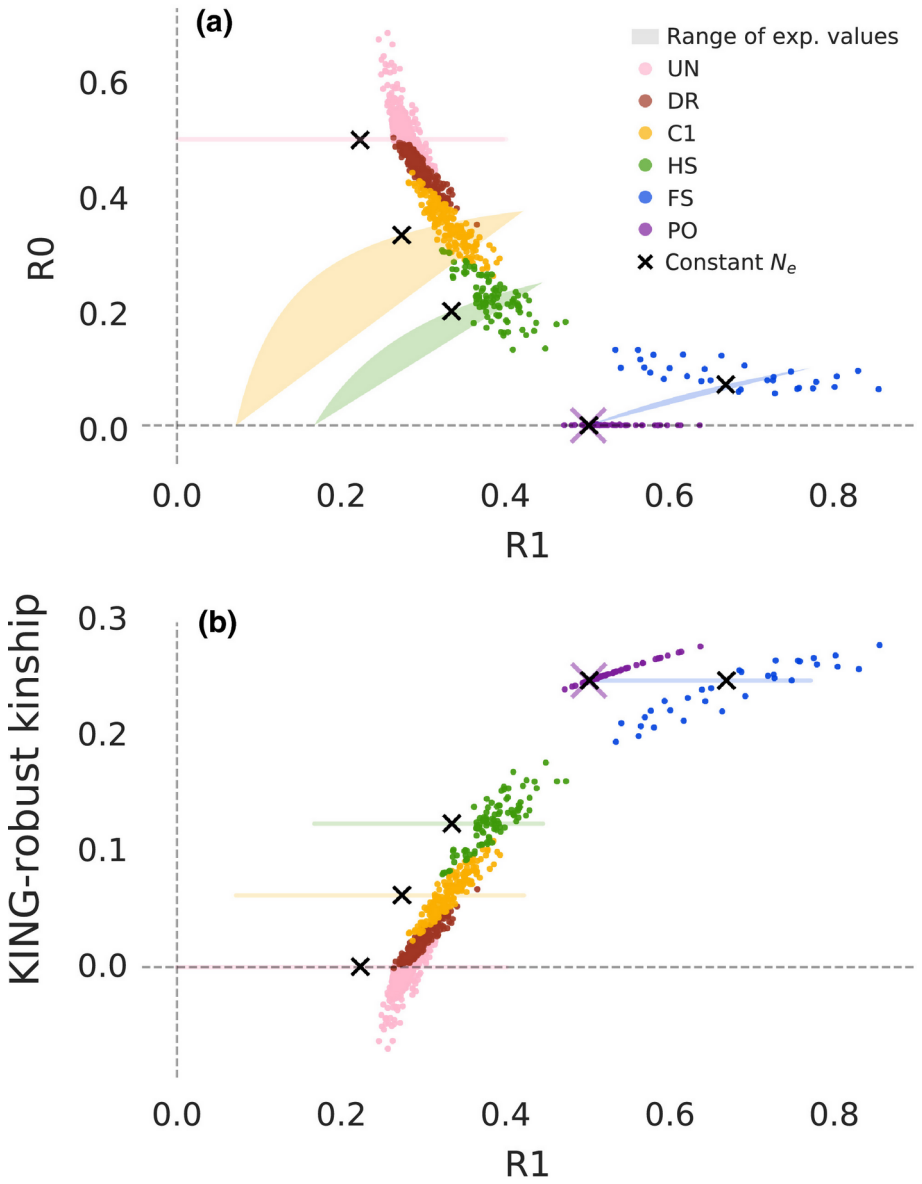
Estimate relatedness: KING-robust



- Relying on the relative proportion of shared heterozygous sites
- More robust to population structure
- Less robust to inbreeding
- Many methods addressing these issues (e.g. RelateAdmix, ngsRelate)

https://github.com/ANGSD/NgsRelate

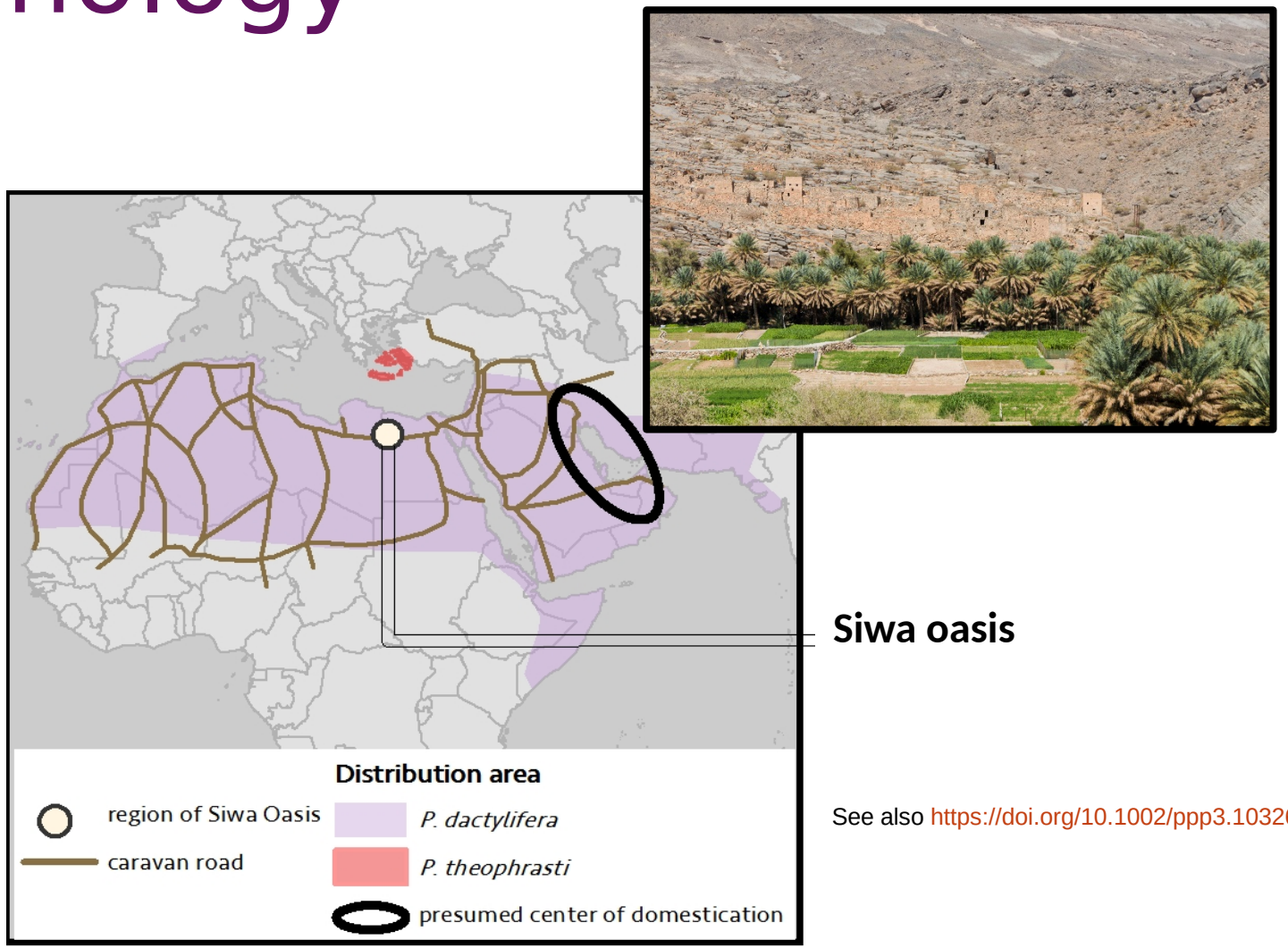
Estimate relatedness: KING-robust



- Relying on the relative proportion of shared heterozygous sites
- More robust to population structure
- Less robust to inbreeding
- Many other methods addressing these issues (e.g. RelateAdmix)
- COLONY…
- **ID related samples is very important for demographic analyses and association studies**

Using relatedness in agronomy/ethnology

- Germplasm conservation requires an accurate inventory of genetic diversity
- Siwa oasis is one of the most diverse
- Discrepancies between varieties identified by local farmers and genetic lineages
- ‘Elite’ cultivars are more genetically uniform
- Feral palms are rarely reused but constitute a source of untapped diversity

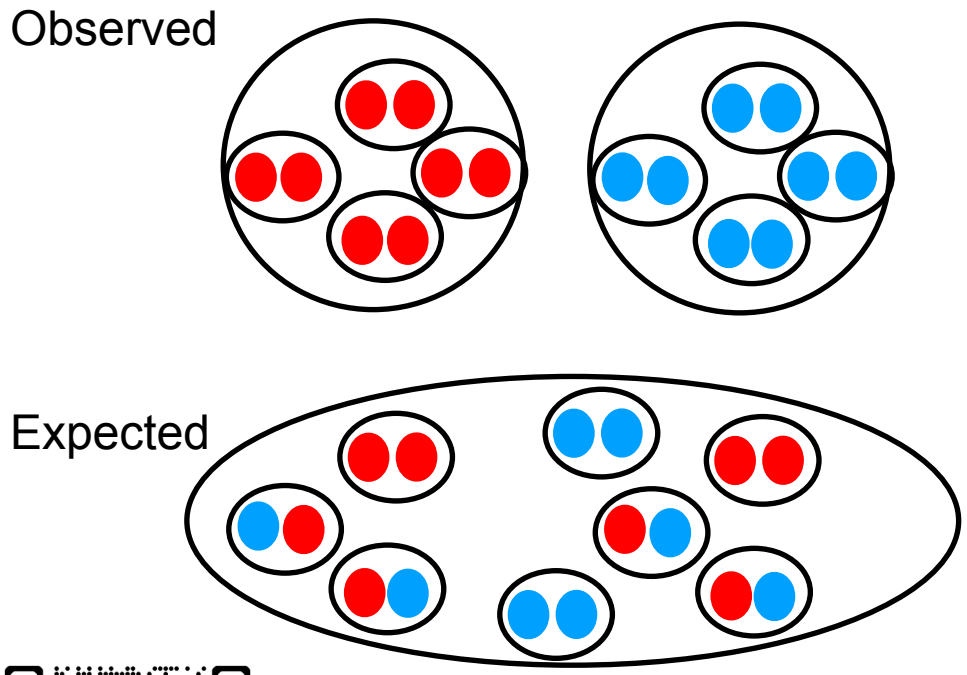


See also https://doi.org/10.1002/ppp3.10326

In prep, reproduced with the kind permission of M. Gros-Balthazard, NYUAD

BETWEEN POPULATIONS: STRUCTURE

Describing population structure using allele frequencies



$F = 1 - (H_{exp}/H_{obs})$

F_{IS} (inbreeding): looking within a population

F_{ST}: taking the entire metapopulation. **Fixation index**.

In real life, we use different estimators (Weir and Cockerham, Bhatia, G_{ST}...)

Correct sampling biases, multiple (>2) alleles, weight of rare alleles, focus on % of variance...

A confusing list of notations.

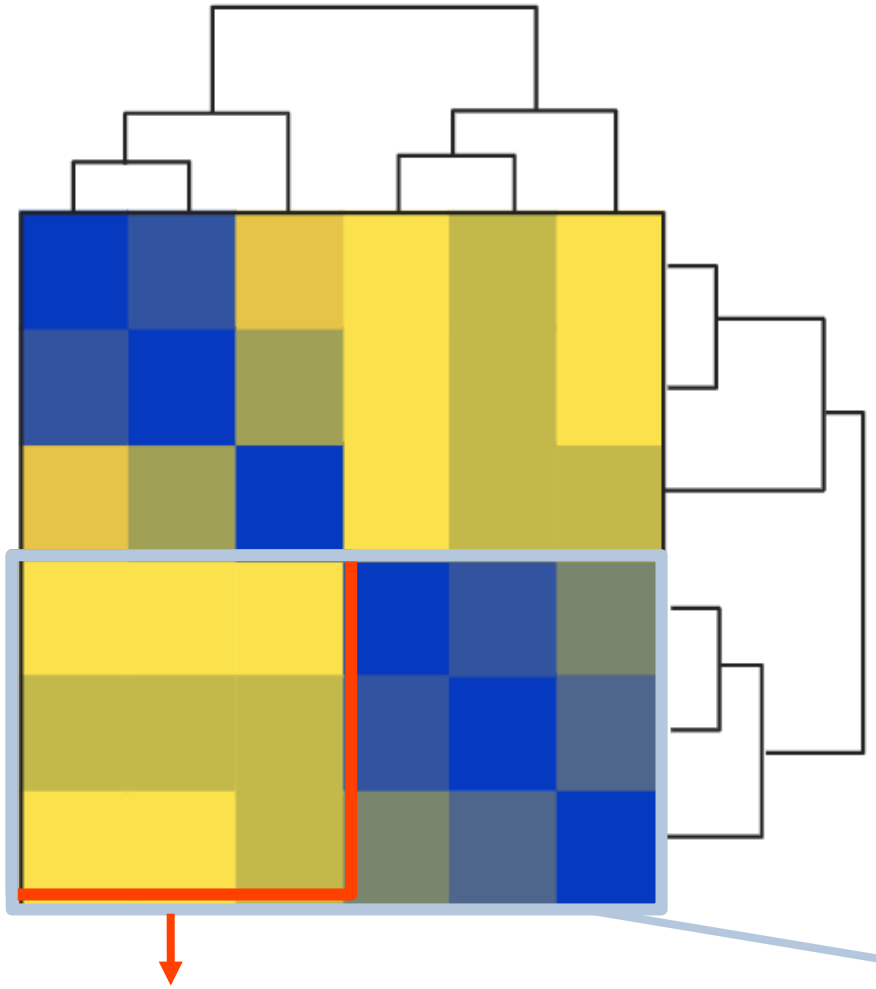
Box 1 | **Mathematical notation**

In this box, we provide definitions for the mathematical symbols used throughout the Review.

Parameter	Definition
Among-population allele frequency distribution	
π	Mean allele frequency
σ_{π}^2	Variance in allele frequency
Wright's F-statistics and Cockerham's θ-statistics	
F_{IS}	Correlation of alleles within an individual relative to the subpopulation in which it occurs; equivalently, the average departure of genotype frequencies from Hardy–Weinberg expectations within populations
F_{ST}	Correlation of randomly chosen alleles within the same subpopulation relative to the entire population; equivalently, the proportion of genetic diversity due to allele frequency differences among populations
F_{IT}	Correlation of alleles within an individual relative to the entire population; equivalently, the departure of genotype frequencies from Hardy–Weinberg expectations relative to the entire population
f	Co-ancestry for alleles within an individual relative to the subpopulation in which it occurs; equivalent to F_{IS}
θ	Co-ancestry for randomly chosen alleles within the same subpopulation relative to the entire population; equivalent to F_{ST}
F	Co-ancestry for alleles within an individual relative to the entire population; equivalent to F_{IT}
Φ-statistics and R_{ST}*	
Φ_{IS}	Excess similarity of alleles within an individual relative to the subpopulation in which it occurs; analogous to F_{IS}
Φ_{ST}	Excess similarity among randomly chosen alleles within the same subpopulation relative to the entire population; equivalently, the proportion of genetic diversity (measured as the expected squared evolutionary distance between alleles) due to differences among populations; analogous to F_{ST}
Φ_{IT}	Excess similarity of alleles within an individual relative to the entire population; analogous to F_{IT}
R_{ST}	Excess similarity among randomly chosen alleles within the same subpopulation relative to the entire population; equivalently, the proportion of genetic diversity (measured as the expected squared difference in repeat numbers between alleles) due to differences among populations; analogous to F_{ST}
Measuring genetic differentiation among populations in quantitative traits	
σ_{GI}^2	Additive genetic variance within populations
σ_{GP}^2	Additive genetic variance among populations
Q_{ST}	Proportion of additive genetic variation in the entire population due to differences among populations; analogous to F_{ST}

*Φ_{ST} from analysis of molecular variance (AMOVA) is used for haplotype data (for example, nucleotide sequence data or mapped restriction site data) and requires a measure of evolutionary distance among all pairs of haploytpes. R_{ST} is used for microsatellite data and requires that alleles are labelled according to the number of repeat units that they contain.

Describing population structure using allele frequencies



Very low variance within this group

A lot of variance between groups

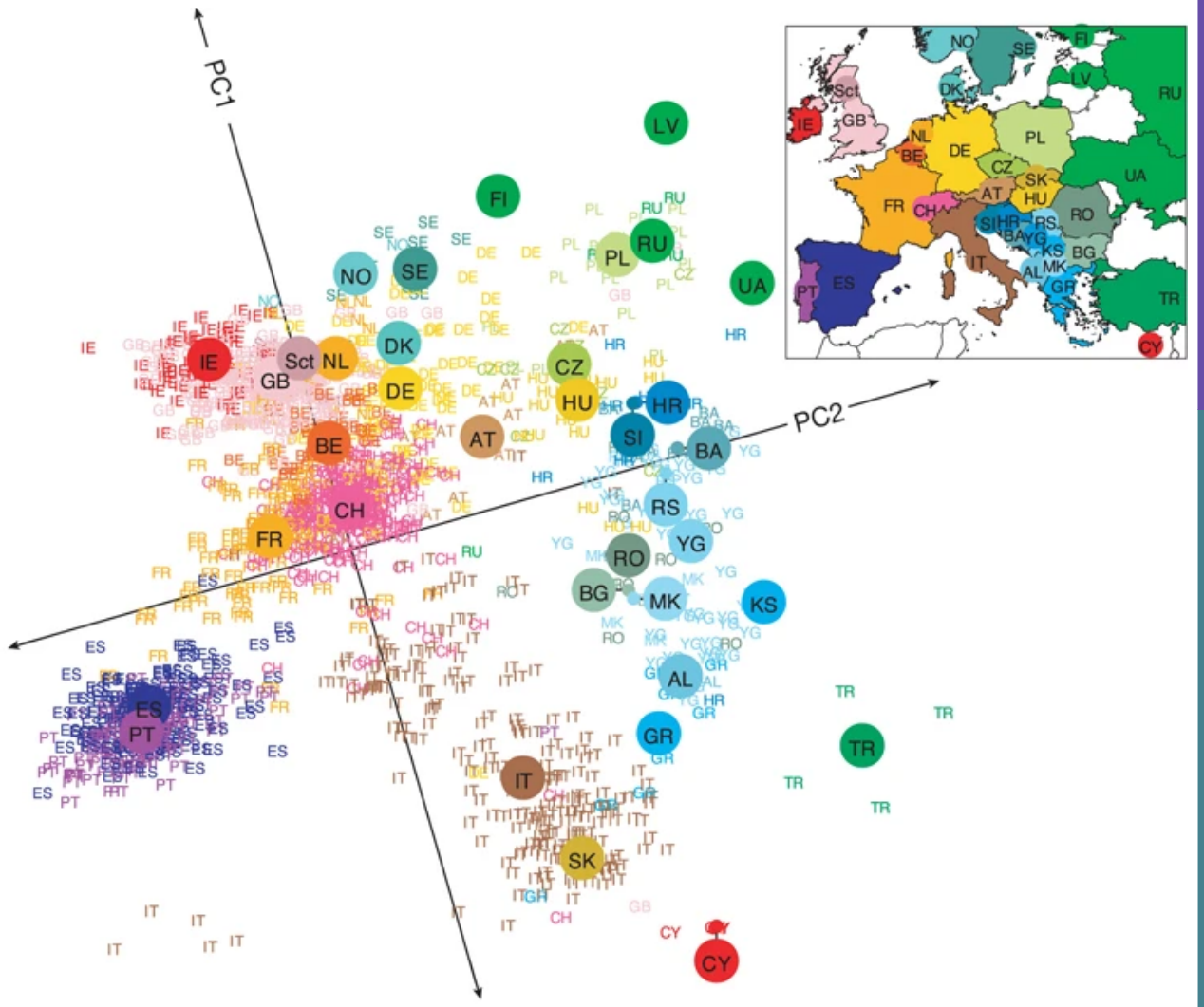
Analyses of Molecular Variance (AMOVA): how is genetic variance partitioned?

Like a (non-parametric, perm-) ANOVA, with individuals, populations, metapopulations as **pre-defined** categories.

F statistics can be related to proportions of variance.

Dimensionality reduction

- Principal Component Analyses (PCA)
- Summarizing covariance in genotypes across markers.
- Fast, convenient for exploring datasets.
- No mutation model.
- Non-supervised (“naïve”).
- Can be interpreted in terms of differentiation and gene flow.



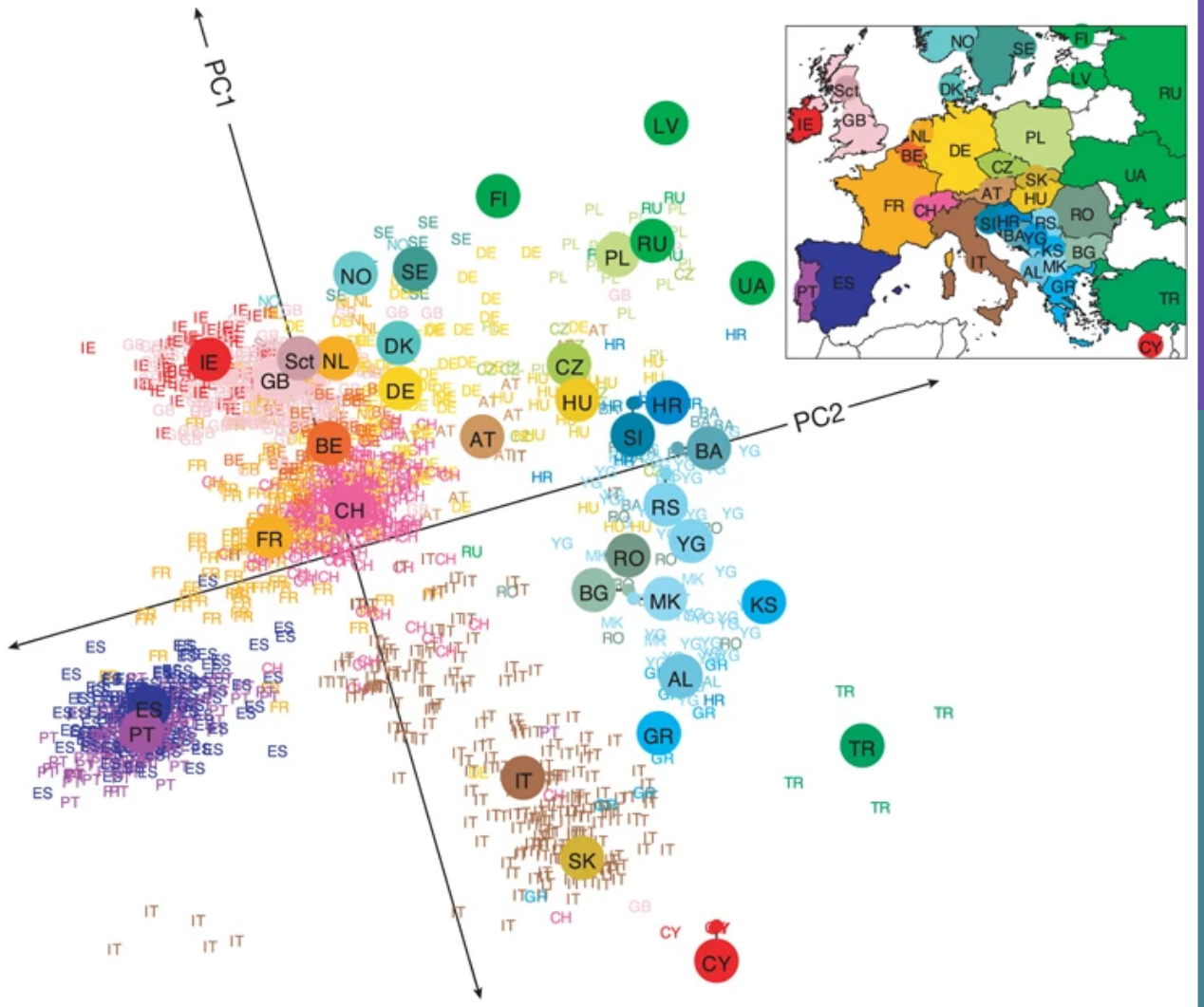
Dimensionality reduction

Principal Component Analyses (PCA)

Genetic drift leads to populations showing more differences in allele frequencies and genotypes with distance

Concept of “Isolation by distance”

Example of a Principal Component Analysis on human genotypes ->

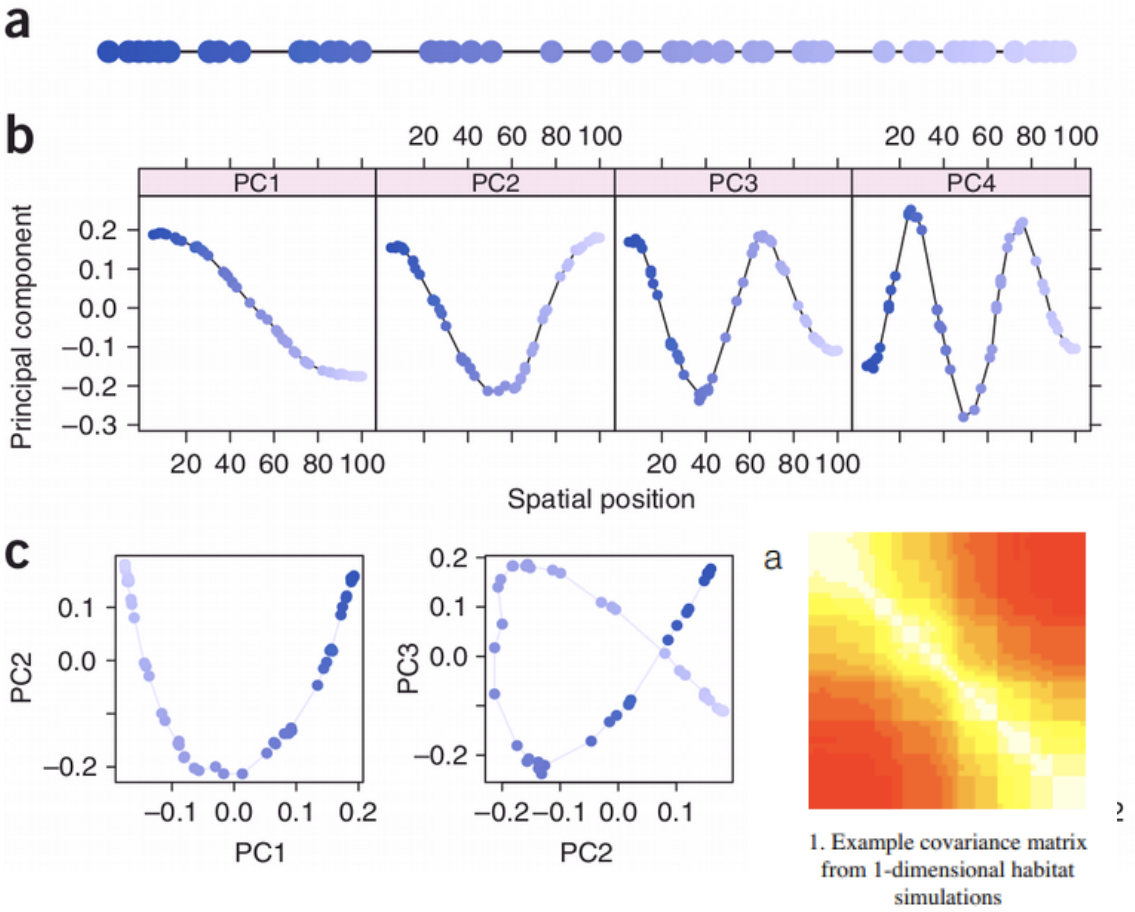


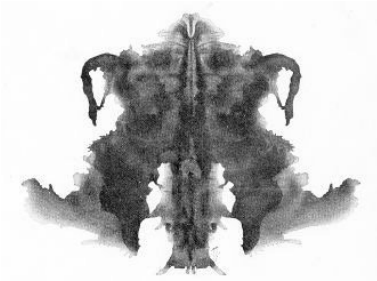
Dimensionality reduction

Caveat: Horseshoe/arch effect

A common effect when examining gradients

- Due to autocorrelation
- Impact of unbalanced sample sizes
- Impact of the choice of markers
- **Take home message**: do not overinterpret PCAs. Use other approaches.





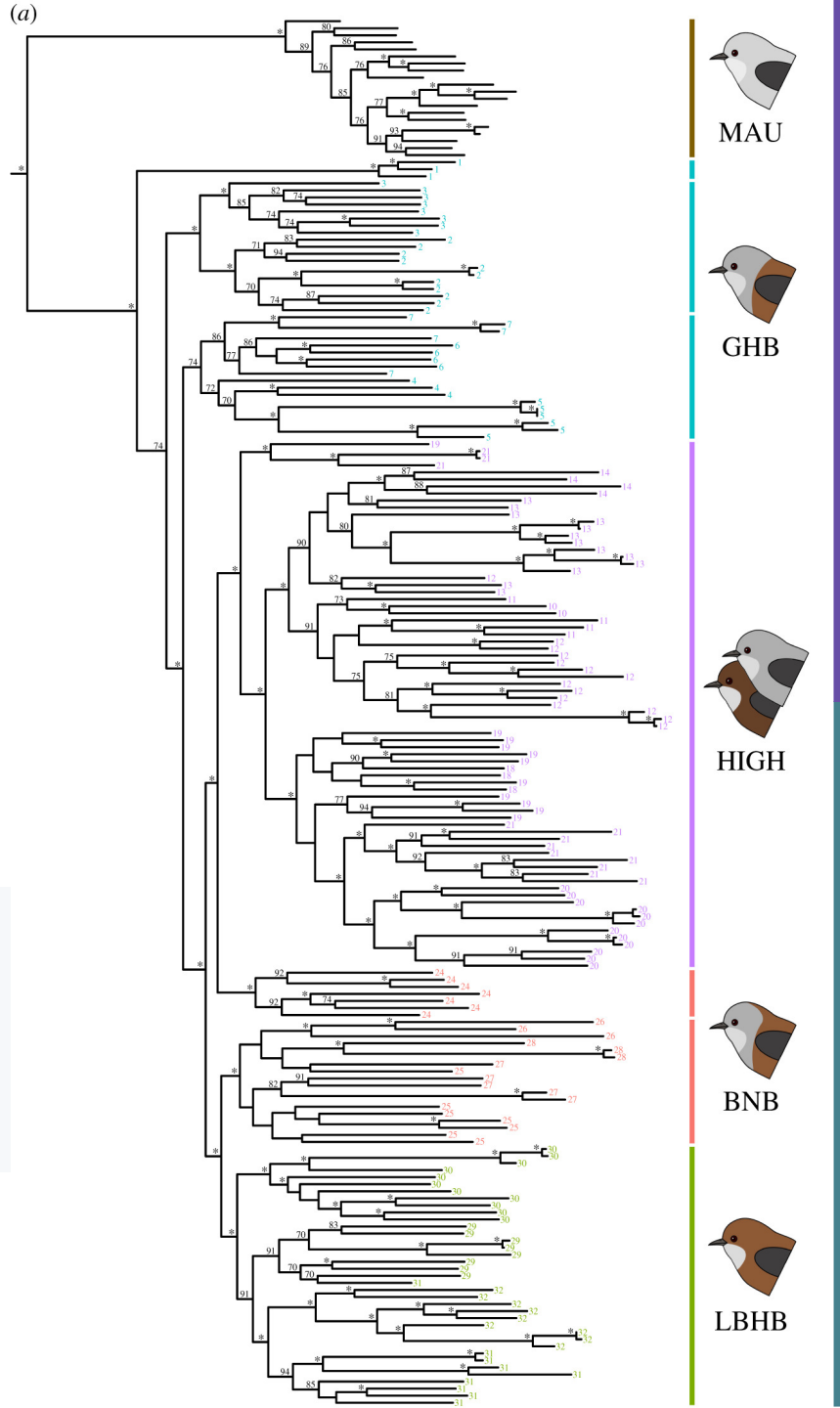
François, O., Jay, F. Factor analysis of ancient population genomic samples. Nat Commun 11, 4661 (2020). https://doi.org/10.1038/s41467-020-18335-6

Gauch HG Jr, Qian S, Piepho HP, Zhou L, Chen R (2019) Consequences of PCA graphs, SNP codings, and PCA variants for elucidating population structure. PLOS ONE 14(6): e0218306. https://doi.org/10.1371/journal.pone.0218306

Tree representation

- Phylogenies (concatenation of SNPs)
- Even simpler: Trees based on pairwise genetic distances
- Just counting the number of different alleles between individuals (p-distance).

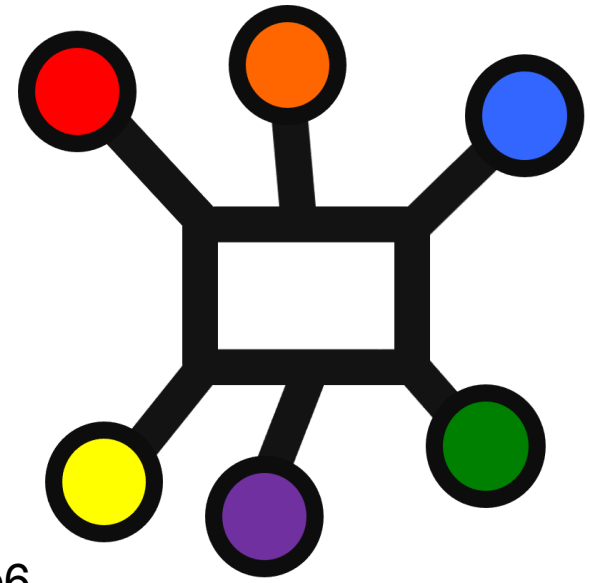
```
d(l)_ij=0.0     if the genotypes of the two individuals were AA and AA;
d(l)_ij=0.5     if the genotypes of the two individuals were AA and AC;
d(l)_ij=0.0     if the genotypes of the two individuals were AC and AC;
d(l)_ij=1.0     if the genotypes of the two individuals were AA and CC;
d(l)_ij=0.0     if the genotypes of the two individuals were CC and CC;
```

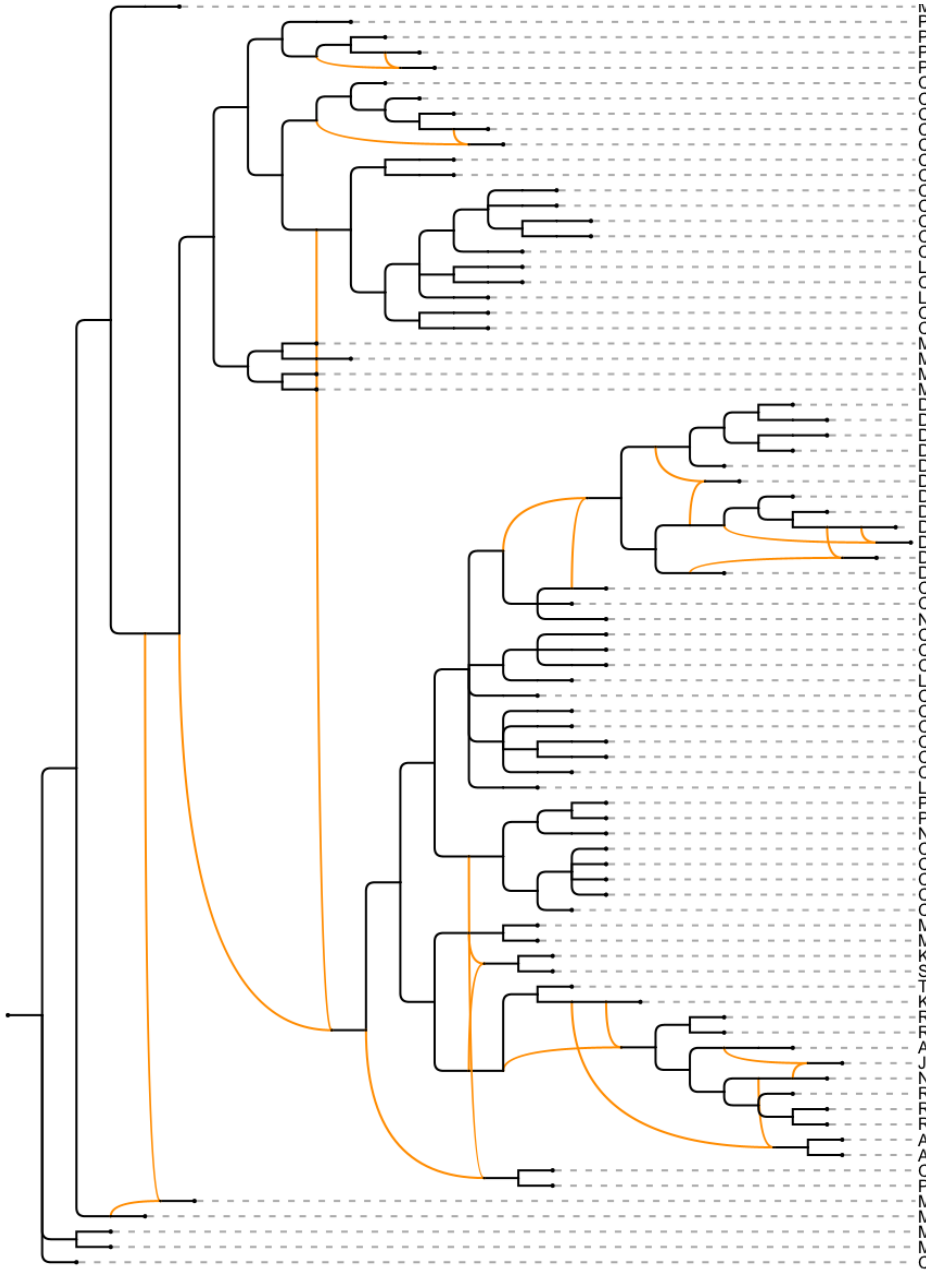


BEYOND DISCRETE CLUSTERS

Reticulation

- Networks based on pairwise genetic distances
- We can resolve conflicting signals in the data by assigning leaves or clades to two or more ancestral branches.
- Descriptive, not very much quantitative.

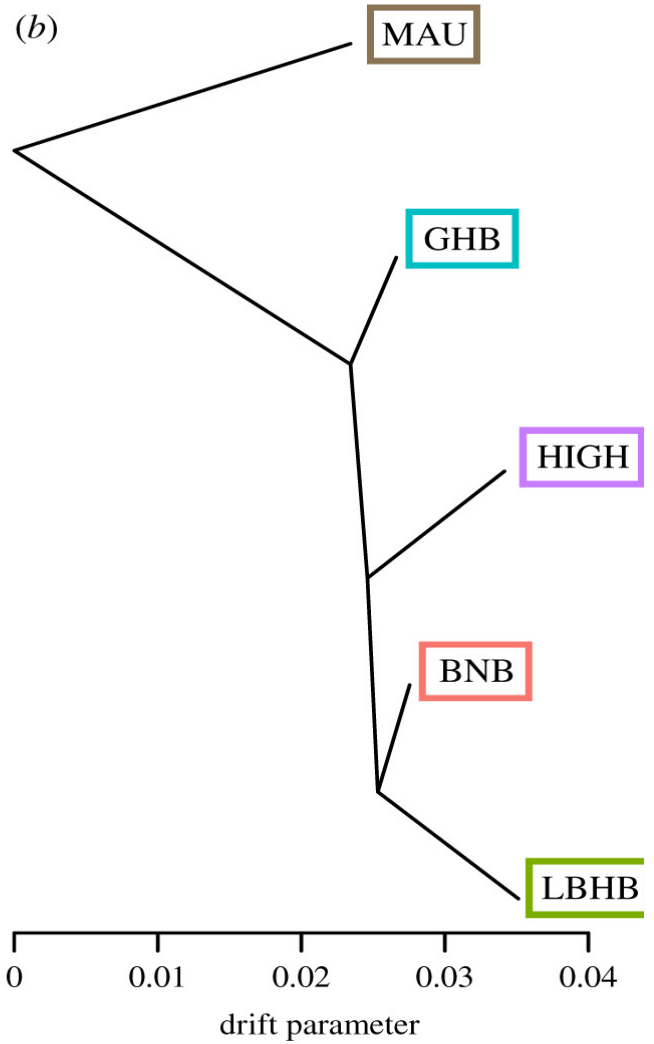




https://github.com/husonlab/splitstree6

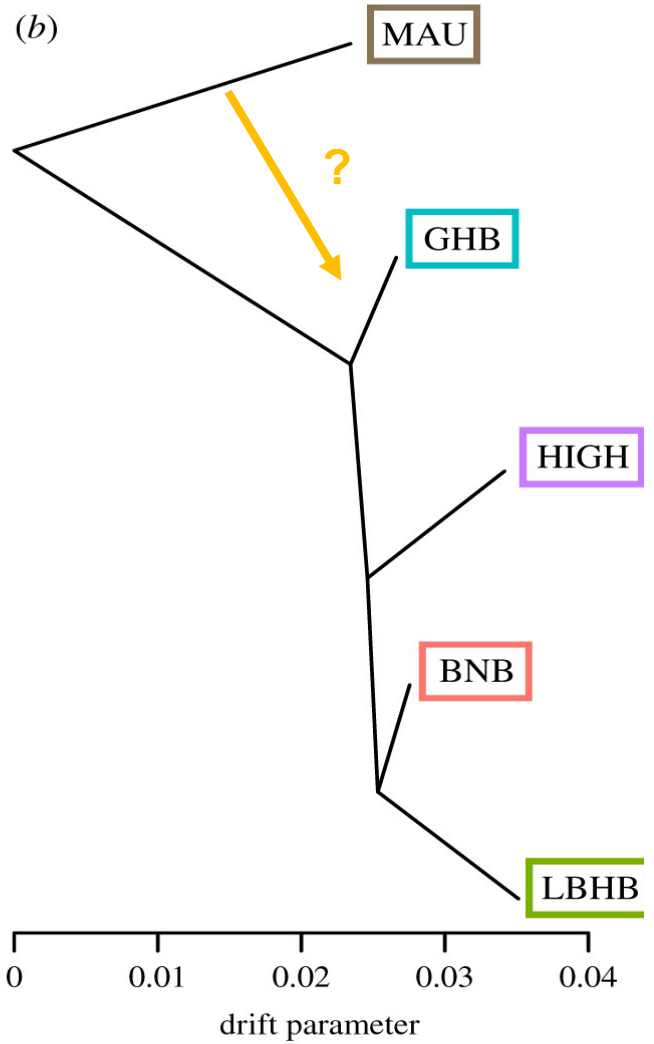
Tree and reticulation of populations

- The same general philosophy can be applied to populations.
- This time, one looks at correlations in allele frequencies
- Possible to introduce reticulation through punctual introgression events
- Comparison of pure tree v. reticulated tree
- Example: TreeMix



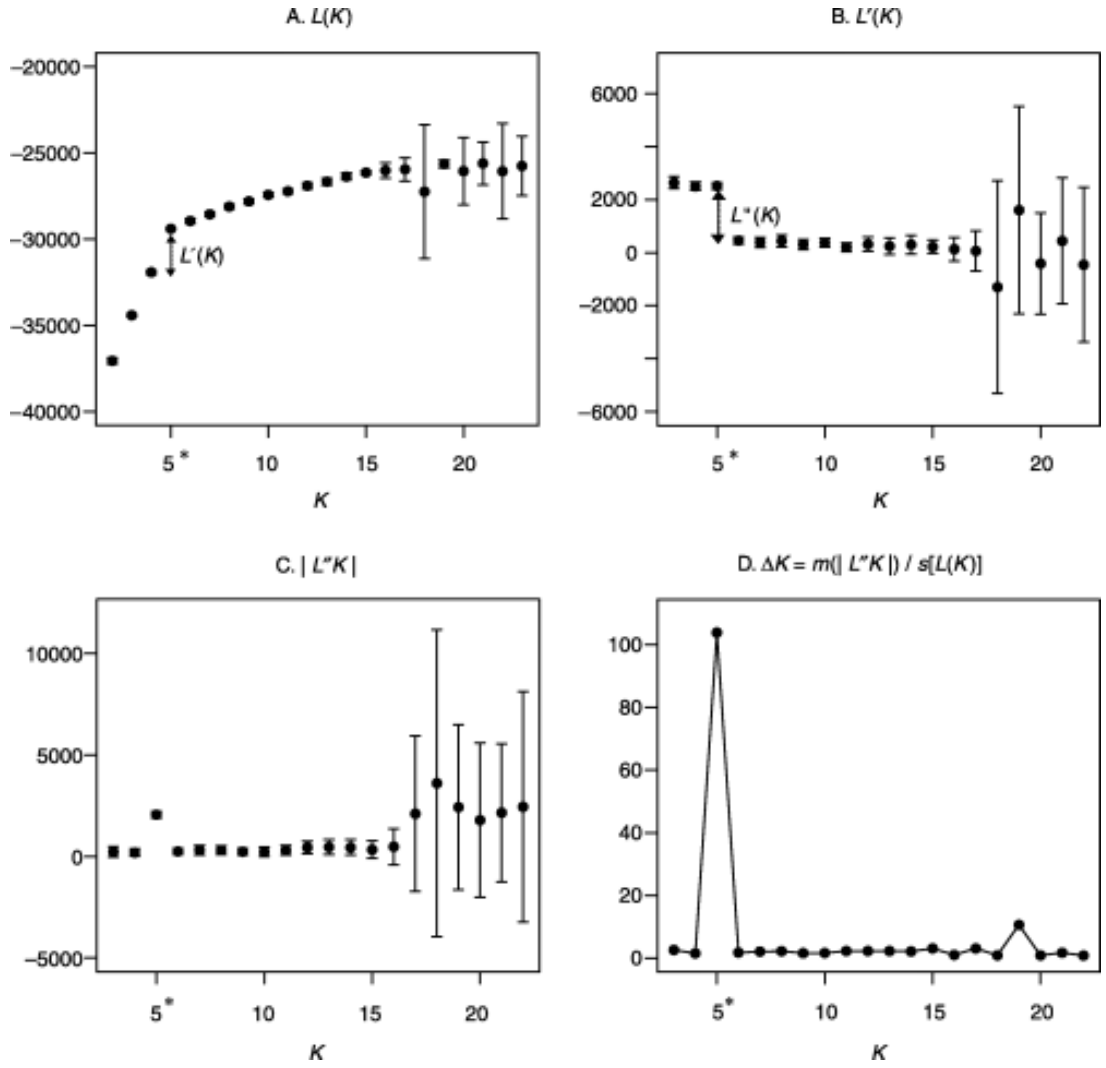
Tree and reticulation of populations

- The same general philosophy can be applied to populations.
- This time, one looks at correlations in allele frequencies
- Possible to introduce reticulation through punctual introgression events
- Comparison of pure tree v. reticulated tree
- Example: TreeMix
- A difficulty: how to identify the best model?



Tree and reticulation of populations

- Each model is associated with a likelihood
- Evanno et al. suggested to examine the second-order change in likelihoods to pick the model that captures most of the signal while avoiding overfitting.
- In today's workshop this is the method we will use with SplitsTree
- Many other methods (cross-validation, Information Criteria…). Usually the documentation provides advice.



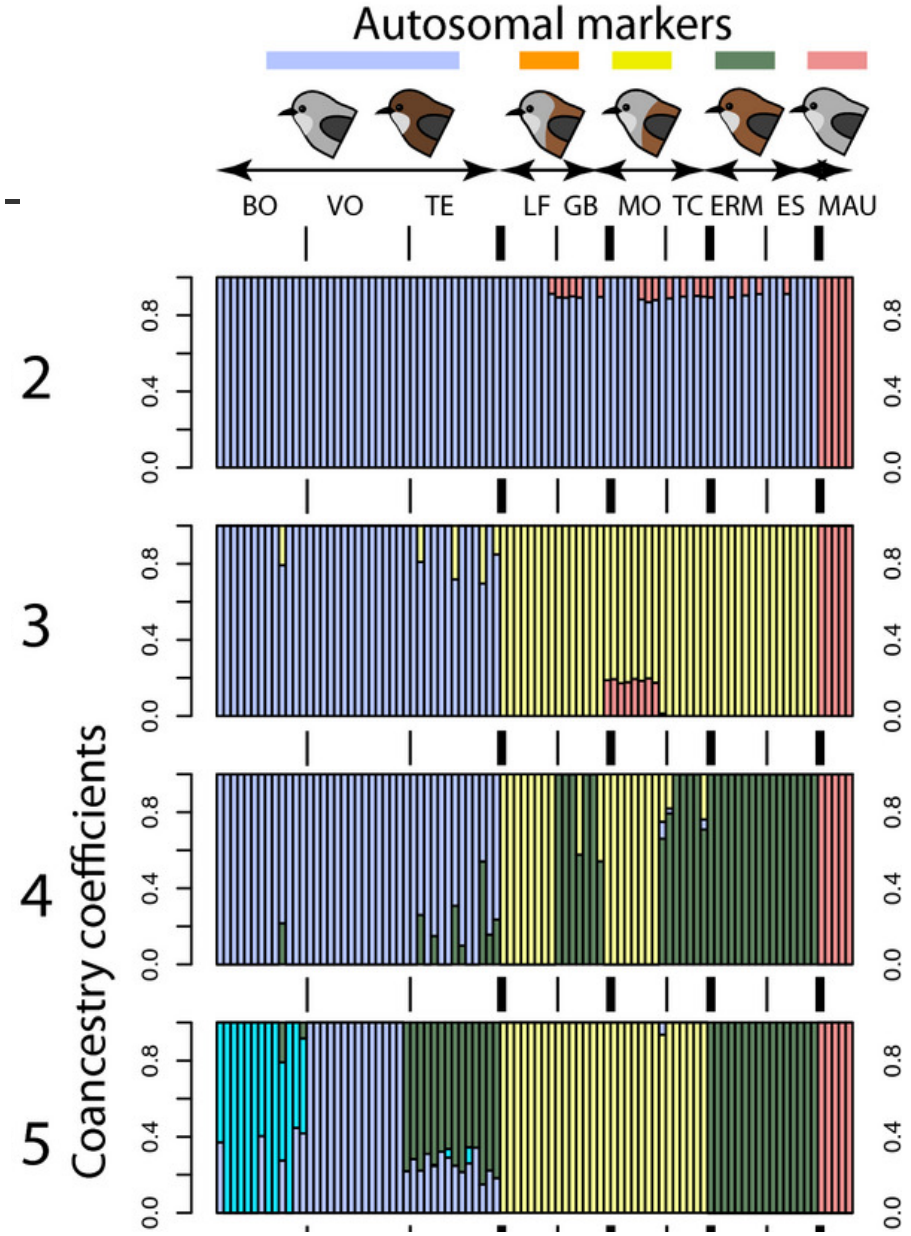
Clustering-based methods

- The model looks for the split into K clusters that fits H-W equilibrium the best
- Each individual can derive a proportion q_k of its genome from one to K clusters.
- No mutation process assumed.
- Assume Linkage Equilibrium (independent markers)
- A classical example for genomic data: ADMIXTURE

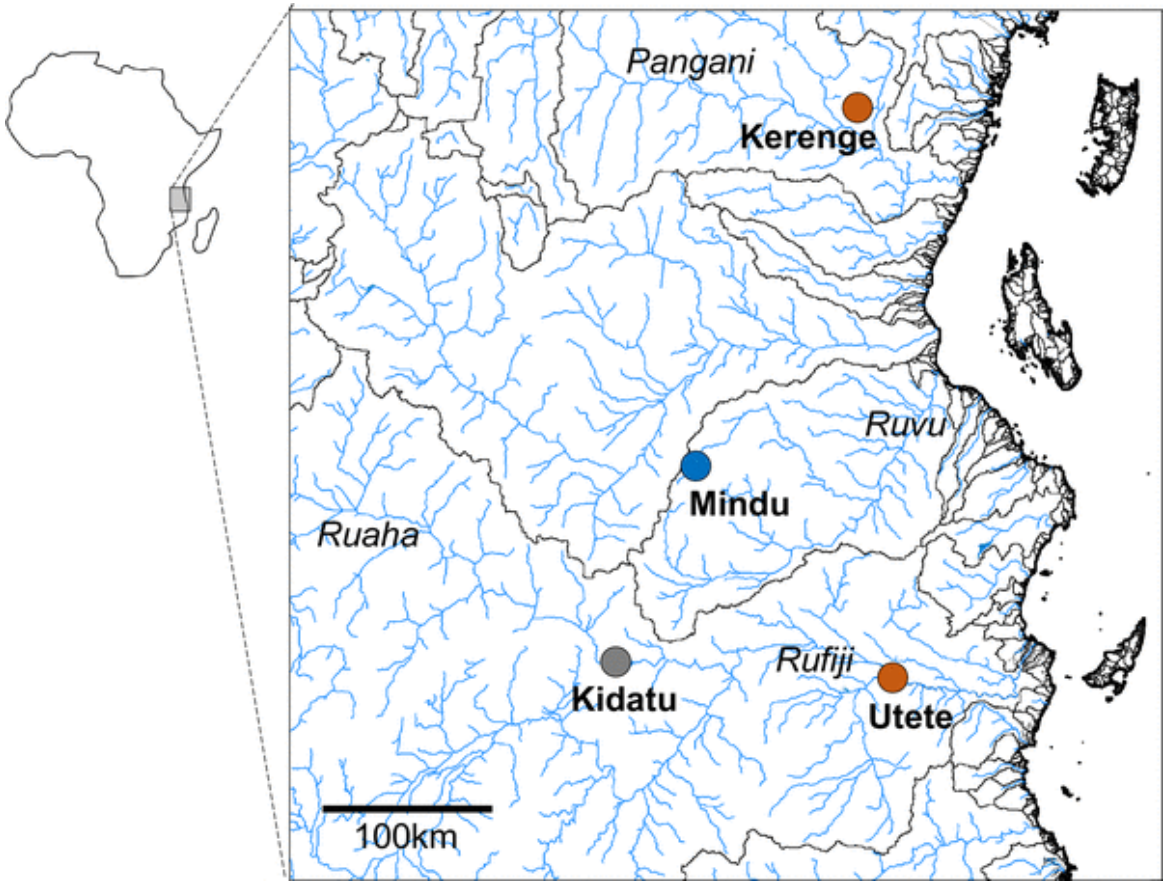
$$\Pr(1/1 \text{ for i at SNP j}) = \left[\sum_k q_{ik} f_{kj}\right]^2$$

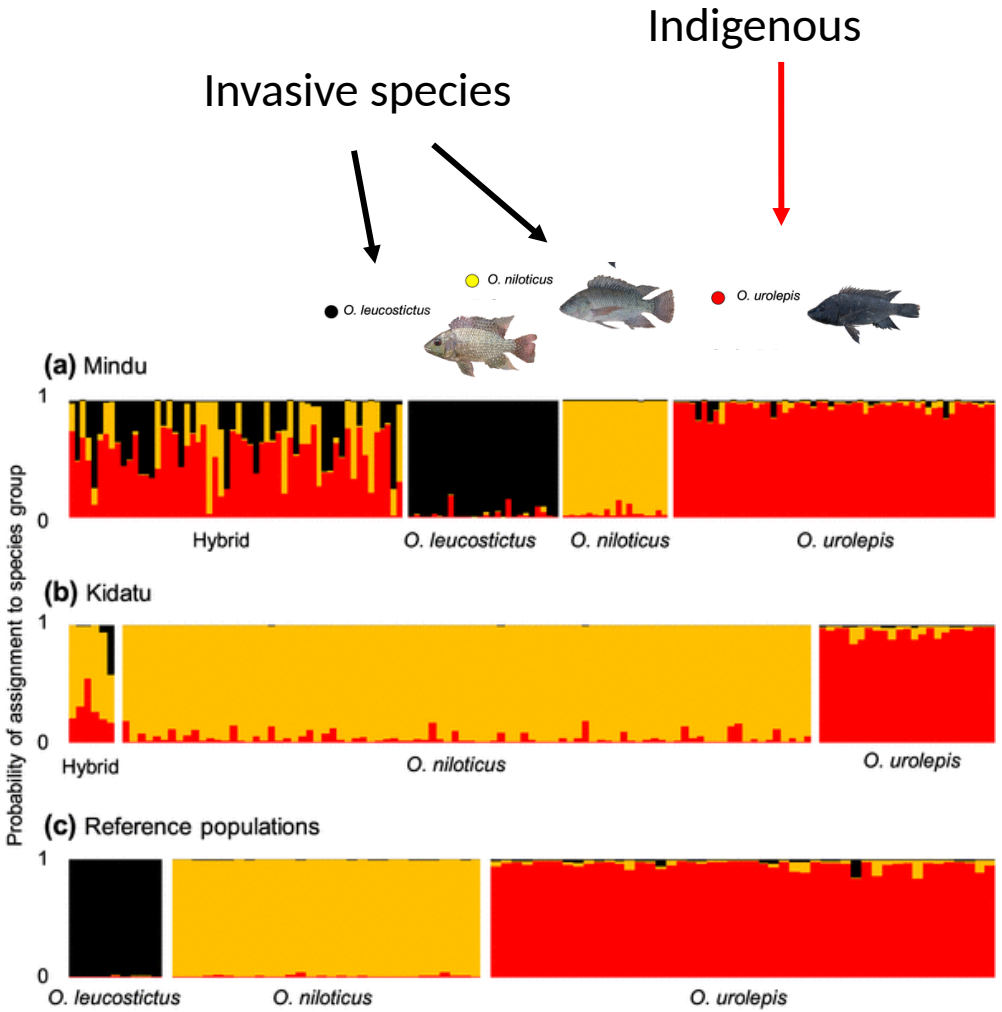
$$\Pr(1/2 \text{ for i at SNP j}) = 2\left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1-f_{kj})\right]$$

$$\Pr(2/2 \text{ for i at SNP j}) = \left[\sum_k q_{ik}(1-f_{kj})\right]^2.$$



Natural populations hybridize and exchange alleles

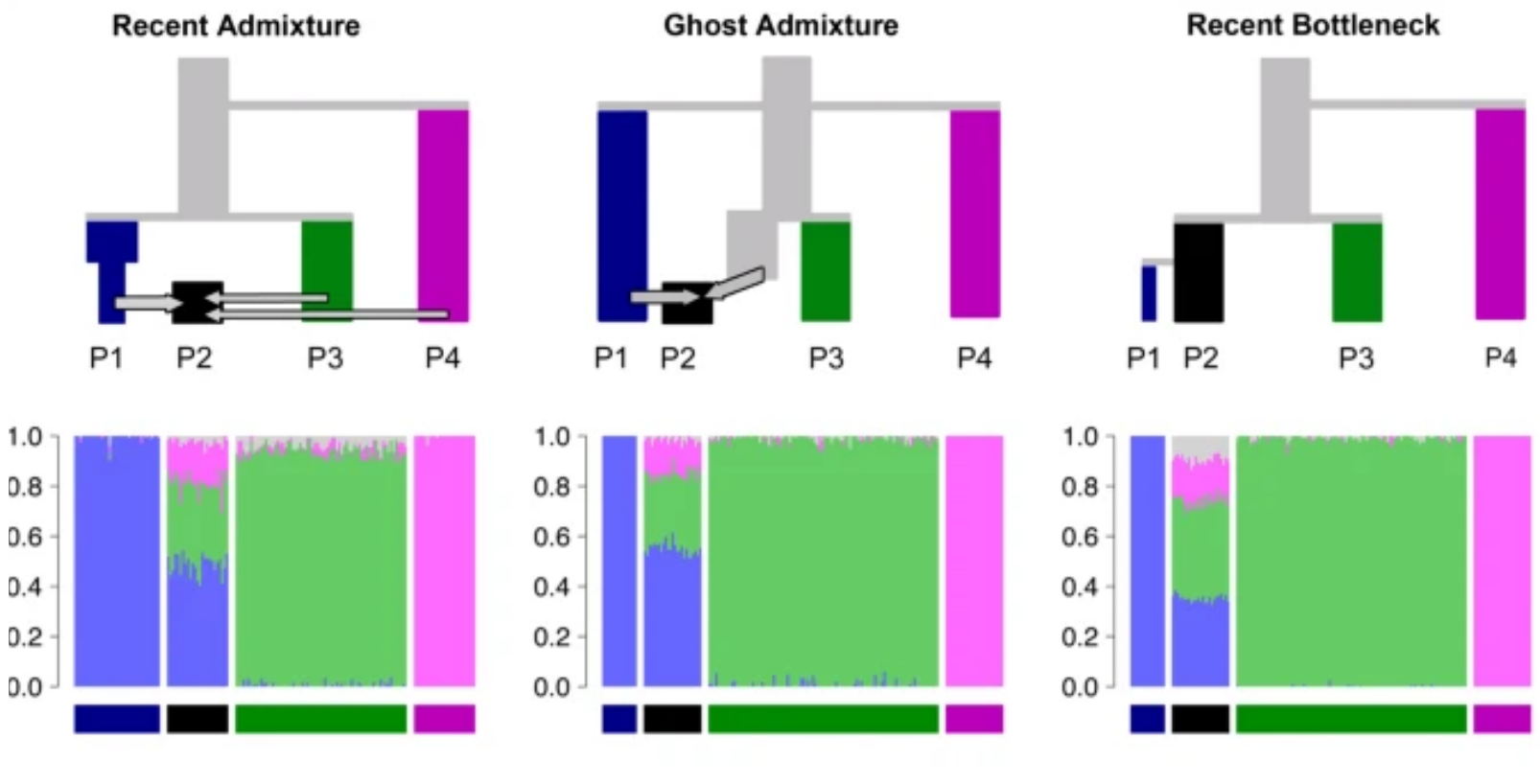




Shechonge et al. 2018, Cons. Genet.

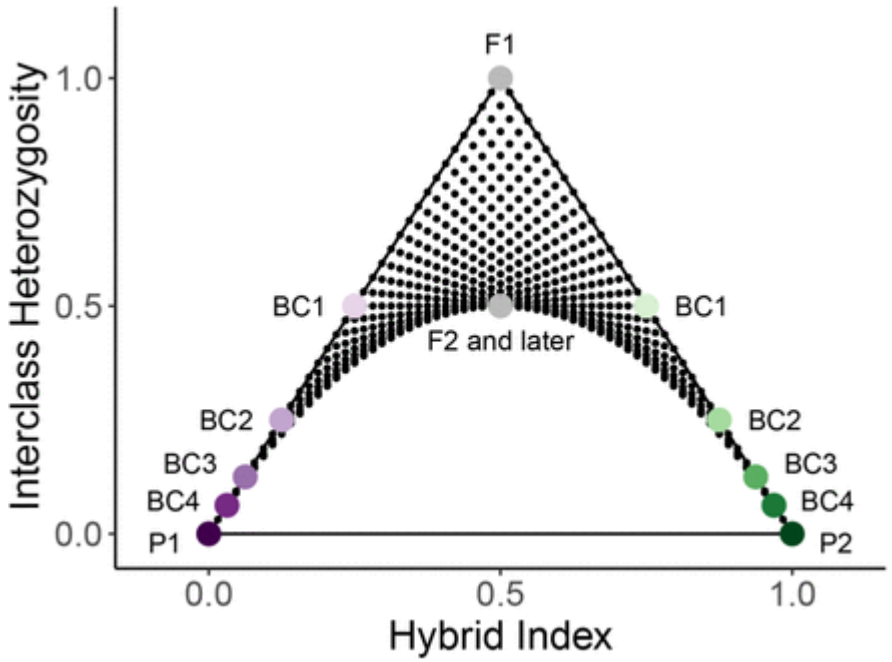
Clustering-based methods: pitfalls

- Avoid LD (independent markers) and close relatives (H-W assumption)
- A few extensions of the method can handle deviations, such as inbreeding or sex chromosomes.
- TEST FOR ADMIXTURE WITH MODELS!



Triangle plots

- Identify parental populations
- Estimate the relative contribution of each genetic background to the focal individual
- Estimate heterozygosity.
- Fast packages in R (triangulaR)
- Same pitfalls as clustering-based methods (see below)



A FEW QUESTIONS FOR YOU...

Which factors can cause deviations from Hardy-Weinberg equilibrium in natural populations?

A - Assortative mating
B - Very large population size
C - Recent selection
D - Population bottlenecks
E - High mutation rate

Which factors can cause deviations from Hardy-Weinberg equilibrium in natural populations?

A - Assortative mating
B - Very large population size
C - Recent selection
D - Population bottlenecks
E - High mutation rate

Which statements about Principal Component Analysis (PCA) in population genomics are true?

A - PCA assumes a model of mutation and recombination
B - PCA can detect isolation by distance in genetic data
C - The horseshoe effect can distort PCA interpretation
D - PCA is a supervised method that uses prior groupings

Which statements about Principal Component Analysis (PCA) in population genomics are true?

A - PCA assumes a model of mutation and recombination
B - PCA can detect isolation by distance in genetic data
C - The horseshoe effect can distort PCA interpretation
D - PCA is a supervised method that uses prior groupings

What are the assumptions or limitations of clustering-based methods like ADMIXTURE?

A - Assume linkage equilibrium across loci

B - Optimizes Hardy-Weinberg equilibrium within clusters

C - Require a predefined number of populations (K)

D - Can accurately infer direction of gene flow

E – Is not biased by related individuals or LD

What are the assumptions or limitations of clustering-based methods like ADMIXTURE?

A - Assume linkage equilibrium across loci

B - Optimizes Hardy-Weinberg equilibrium within clusters

C - Require a predefined number of populations (K)

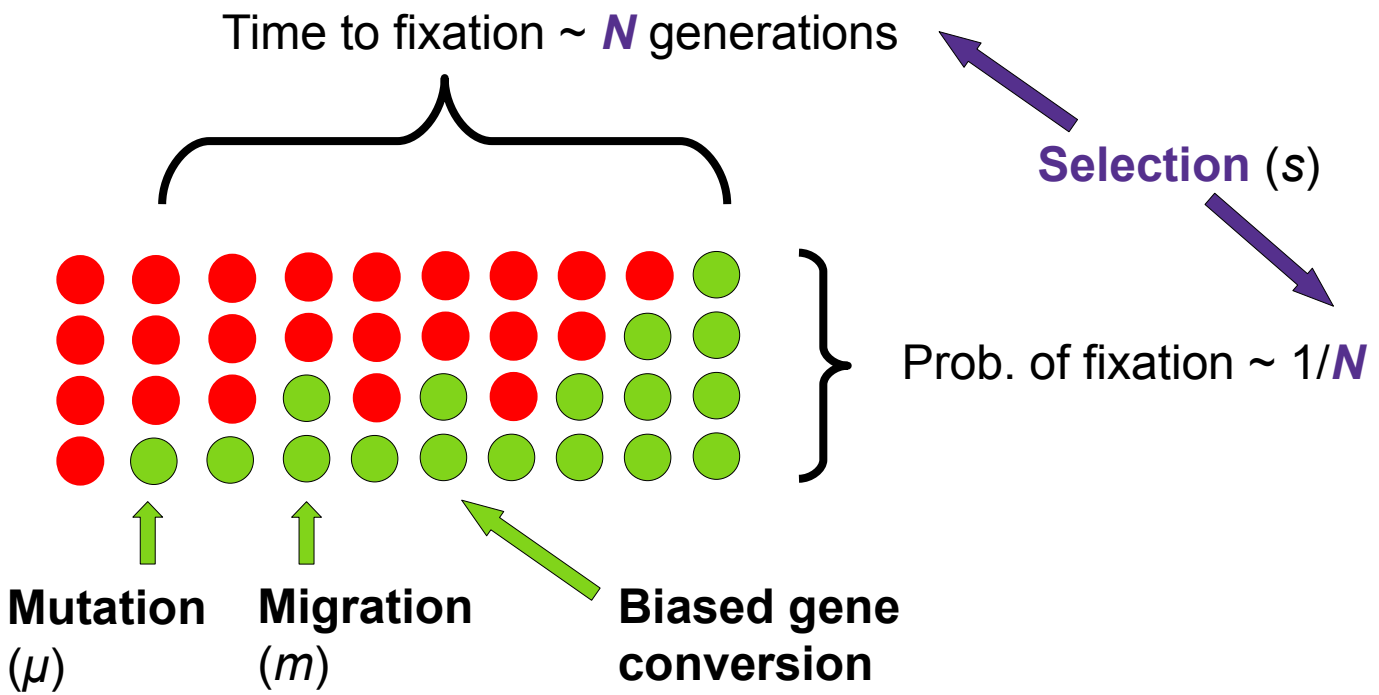
D - Can accurately infer direction of gene flow

E – Is not biased by related individuals or LD

EXCHANGE OF ALLELES BETWEEN POPULATIONS

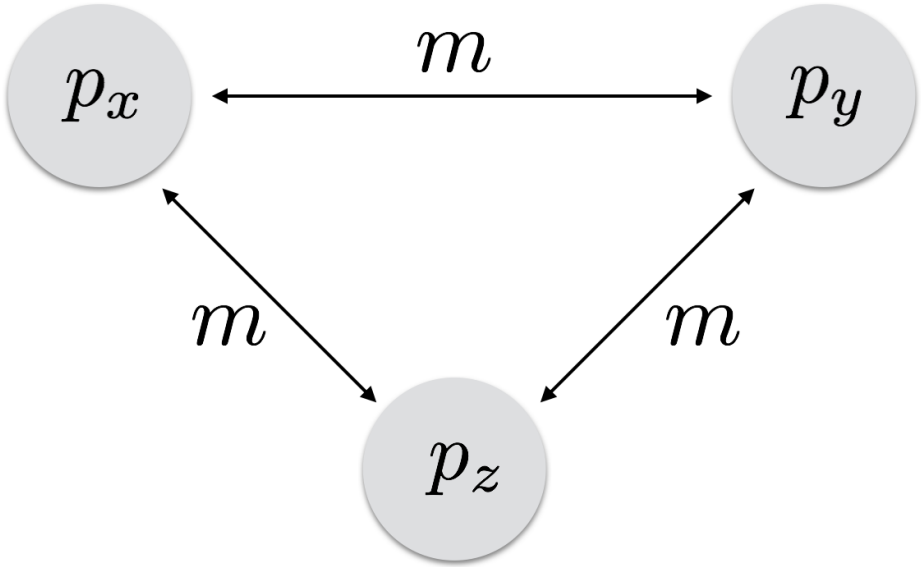
Methods to study admixture/gene flow.

- Populations are often connected
- How do we quantify connectivity?



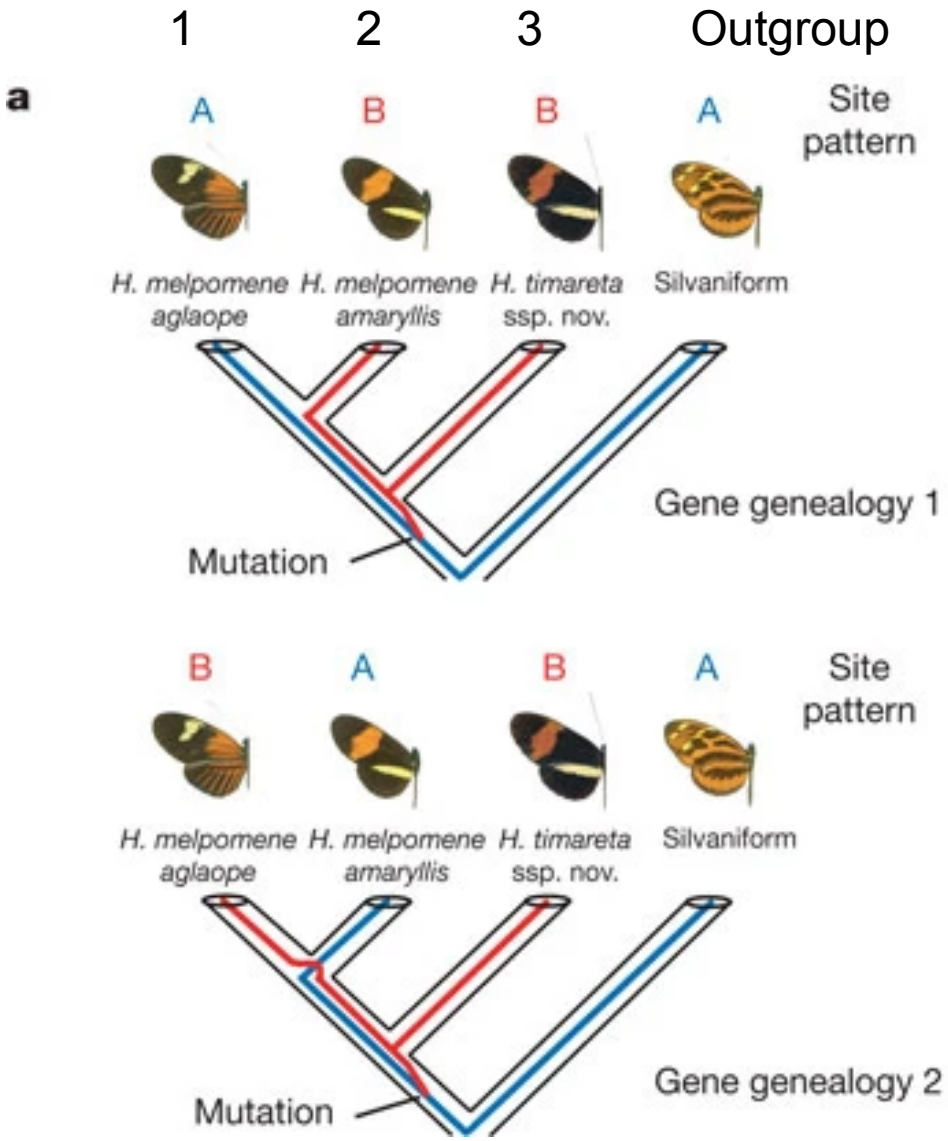
Summary statistics: Wright island model

- The "fantasy" island model



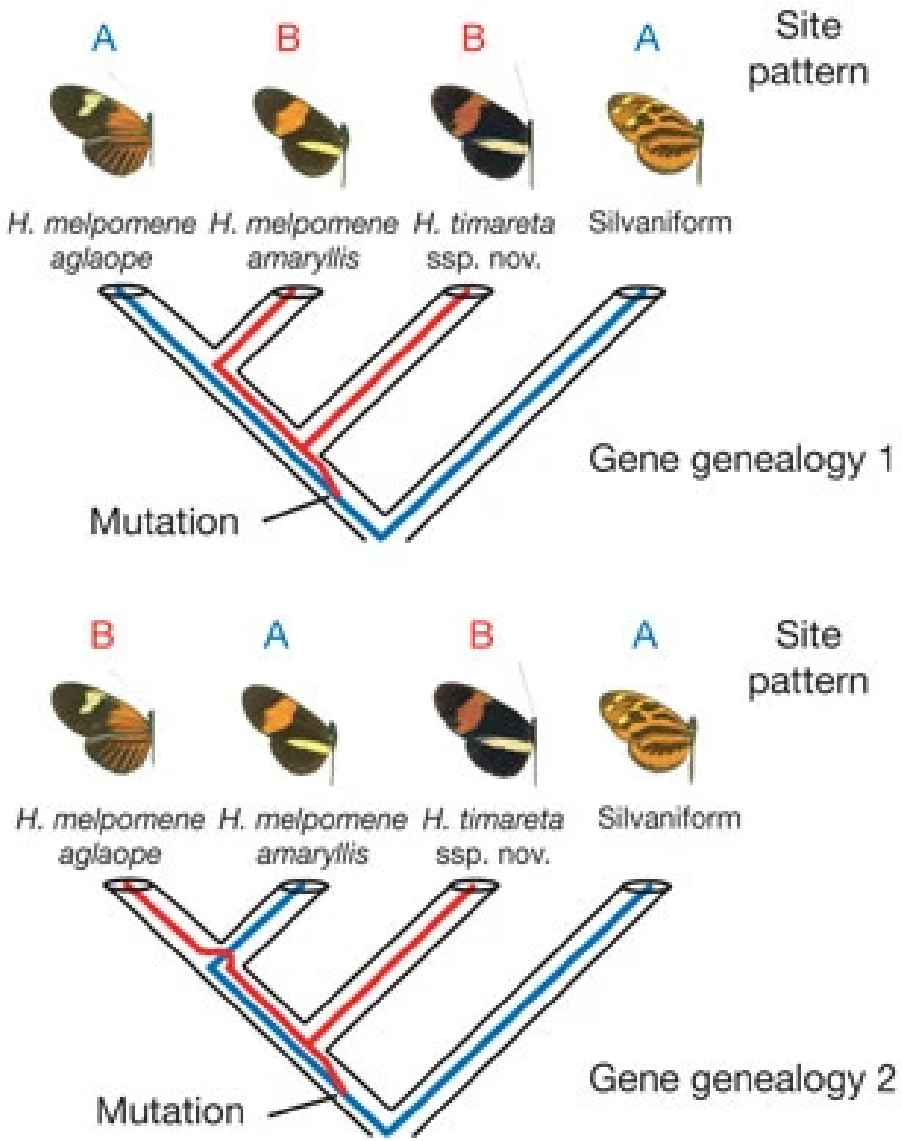
- $F_{ST} \approx 1/(1 + 2Nm)$ with N in # of haploids
- Simple link between the fixation index and migration rate.
- Assumes equilibrium between migration and drift.
- Rarely that simple, but useful rule of thumb.

Summary statistics: ABBA-BABA tests



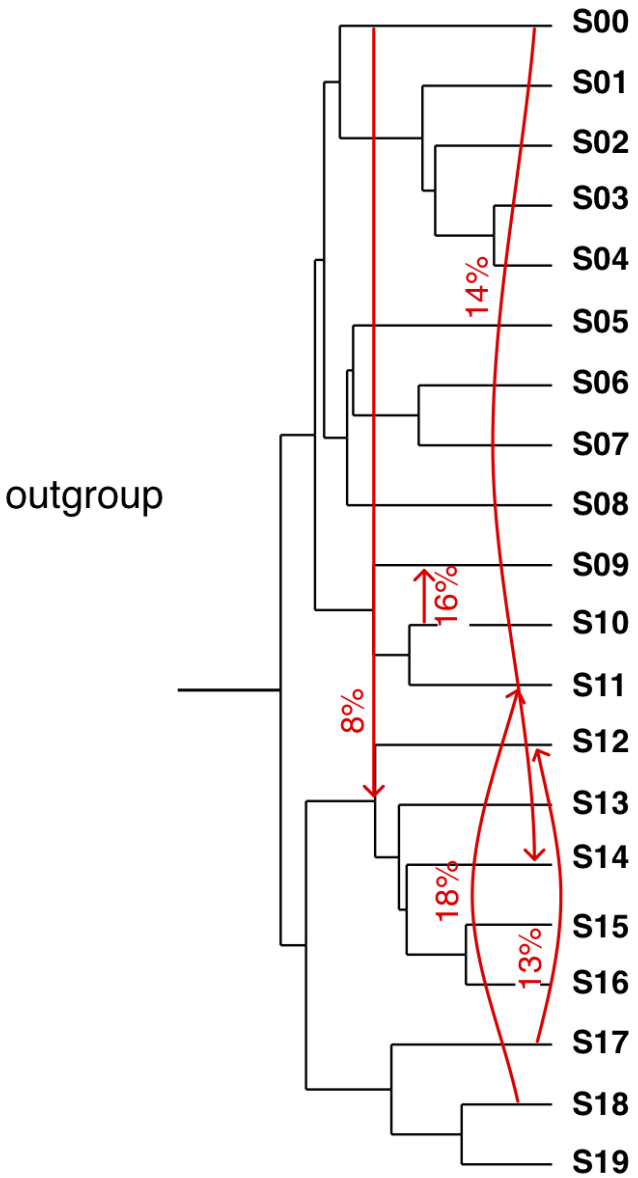
- If no gene flow between 2 and 3, BABA=BABA
- If gene flow between 2 and 3, ABBA > BABA
- If gene flow between 1 and 3, ABBA < BABA
- Summarized as Patterson's D: (ABBA-BABA)/(ABBA+BABA)

Summary statistics: ABBA-BABA tests



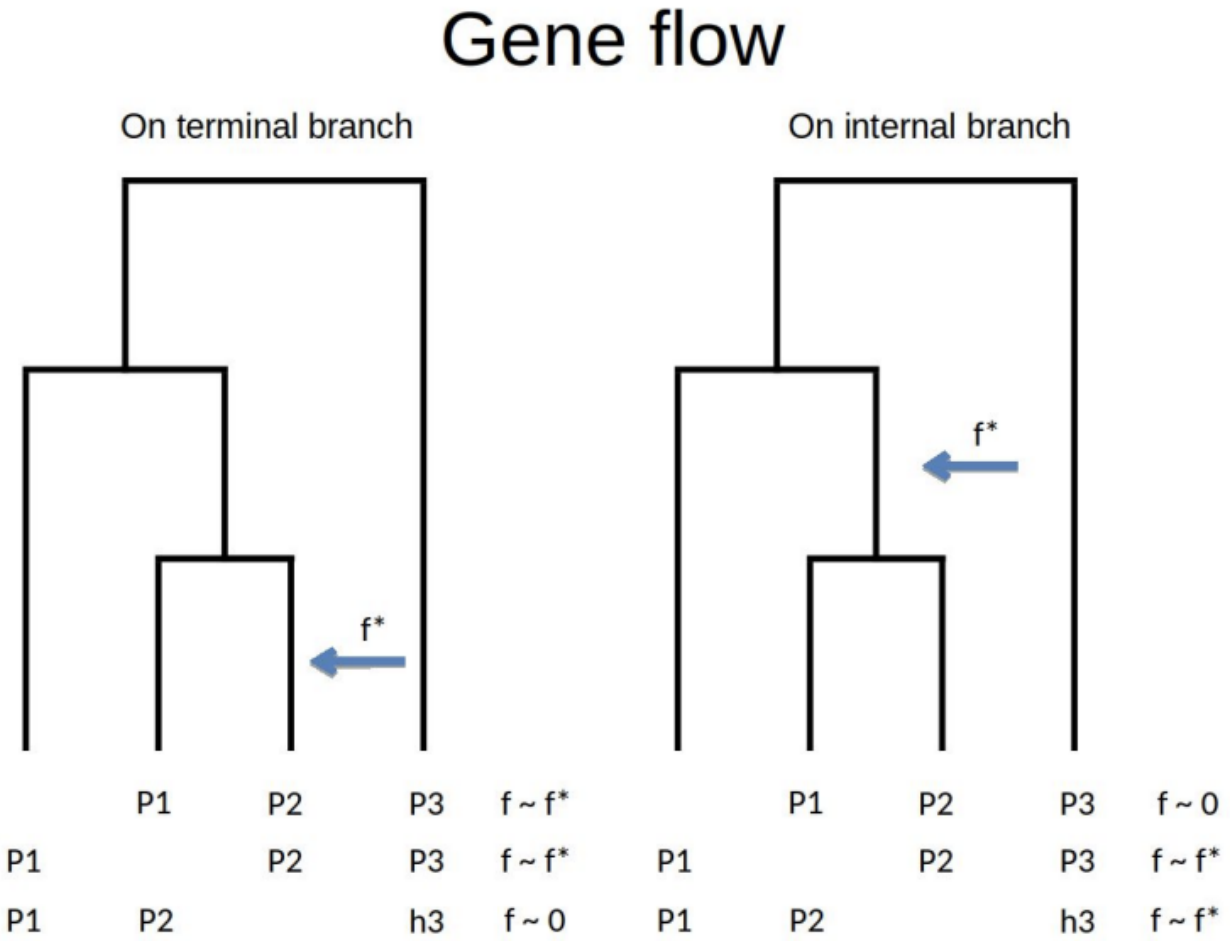
- Fast.
- Can be adapted to populations/polymorphisms.
- Requires the right set of species/populations.
- Does not give information on the direction of introgression

More complicated...



- You could compute all possible D values.
- Clearly not easy.
- One issue: introgression into the ancestor of a clade will be reflected in all the descending tips.

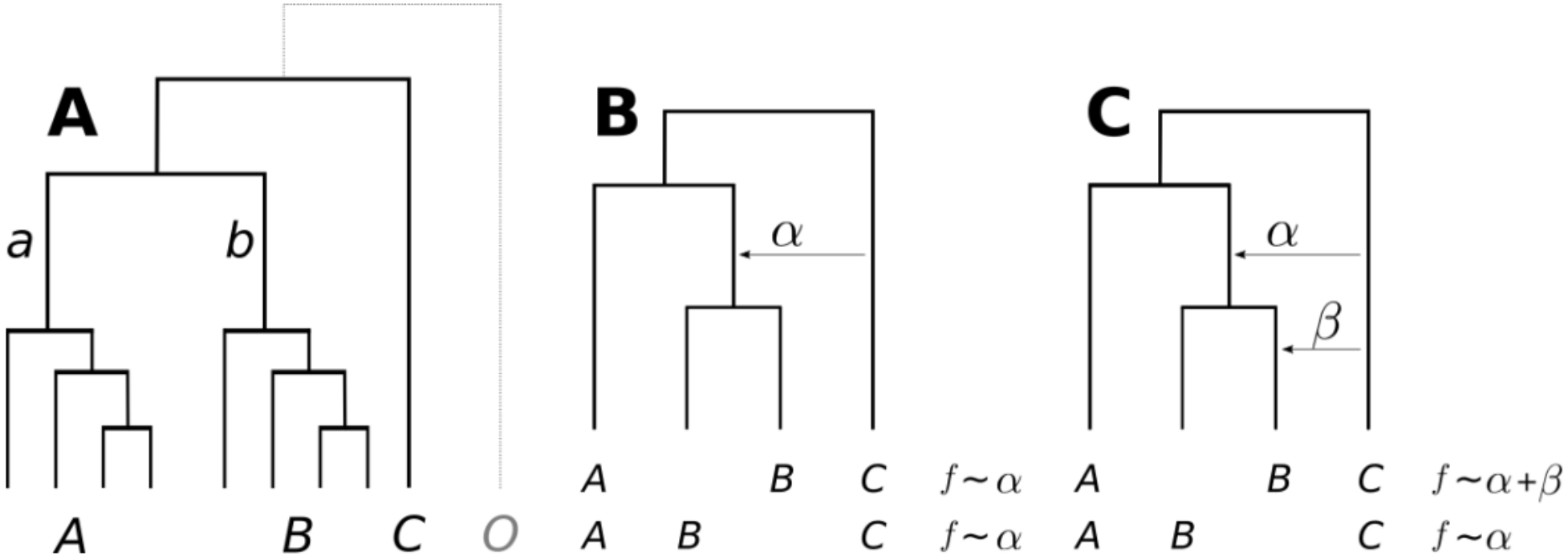
More complicated...



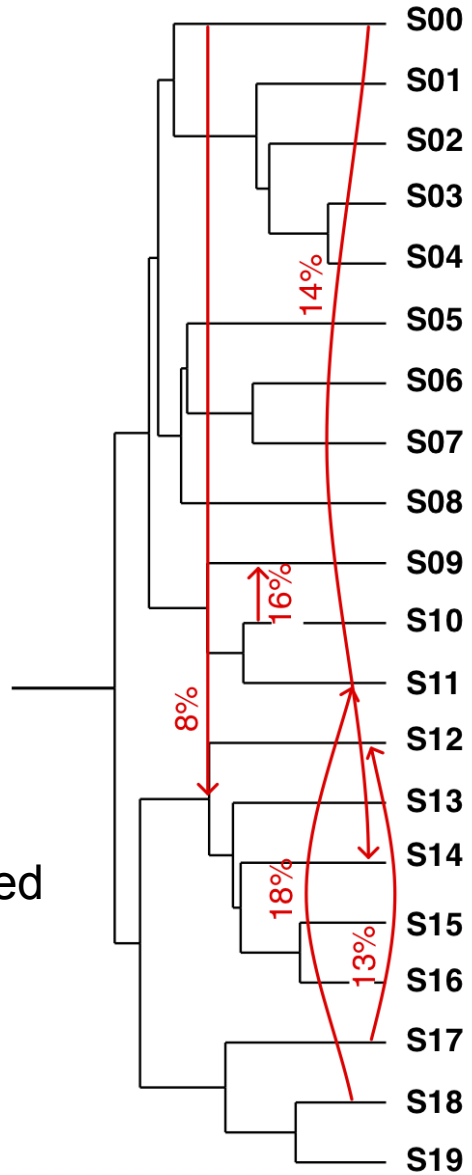
- You could compute all possible D values.
- Clearly not easy.
- One issue: introgression into the ancestor of a clade will be reflected in all the descending tips.

More complicated...

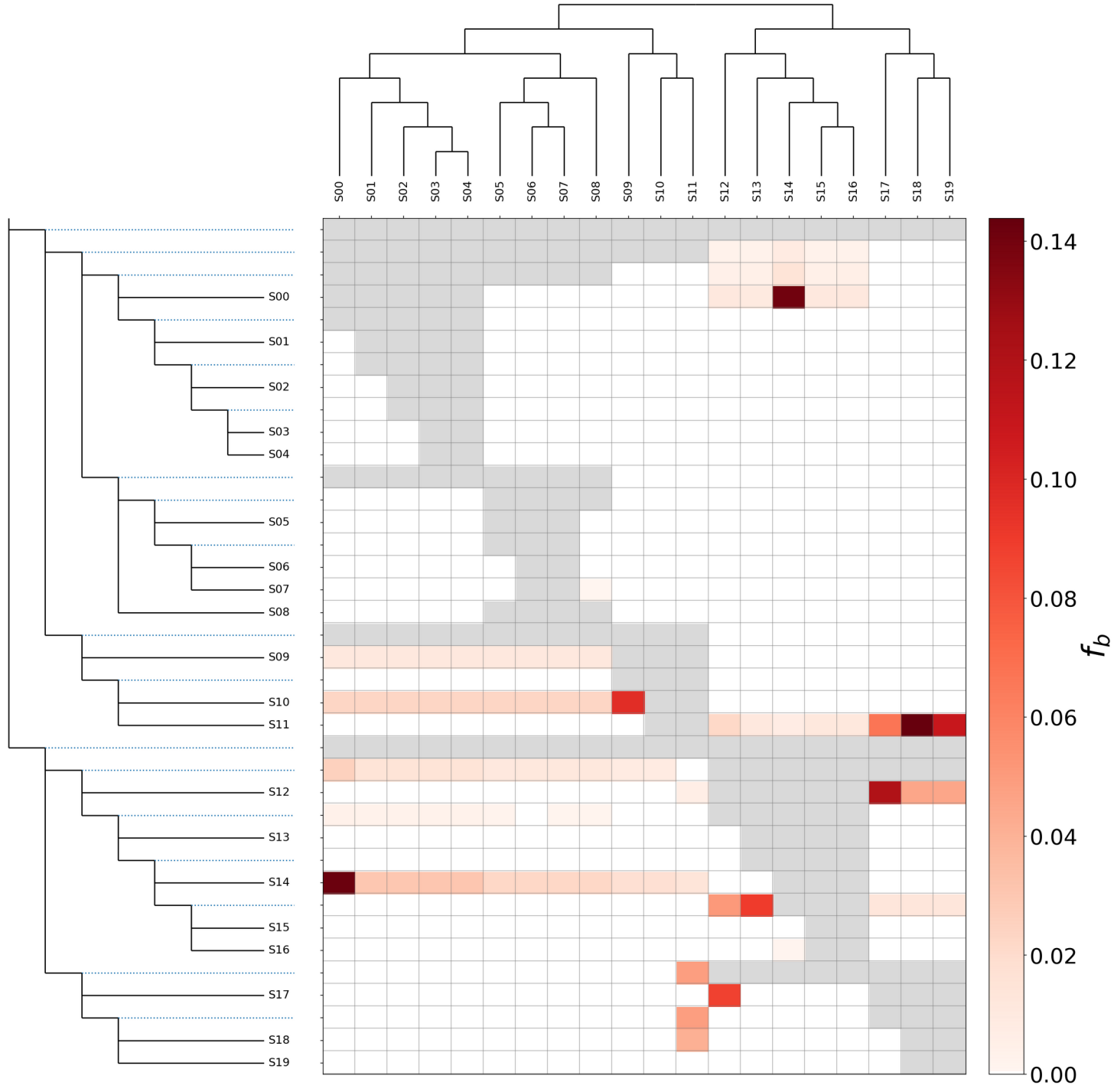
- F_{branch} statistics can tackle this issue
- Helps identifying the origin of gene flow (if not its direction).



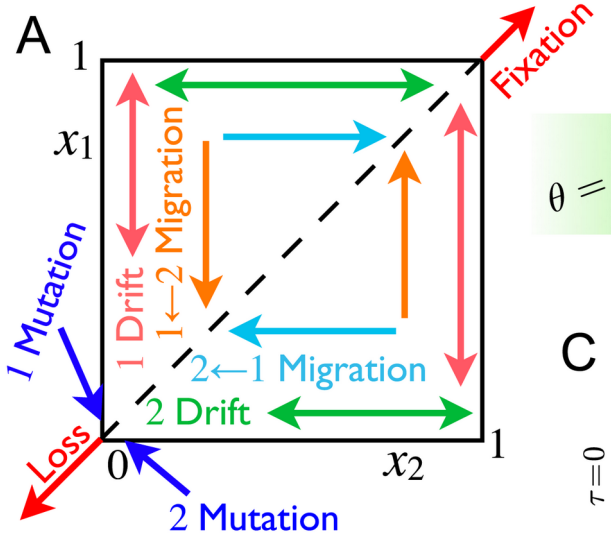
outgroup

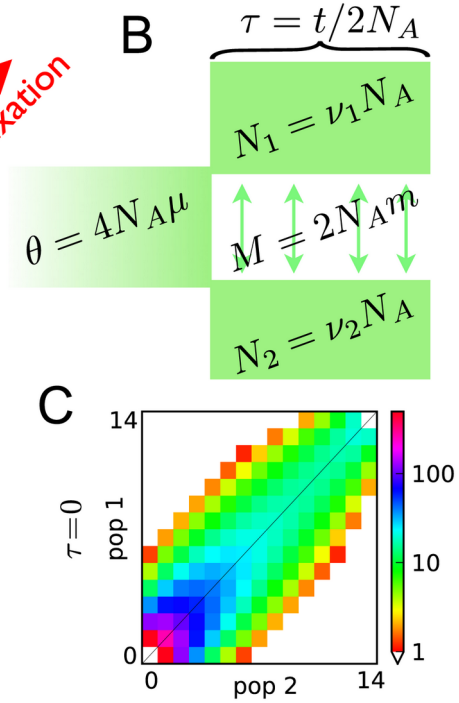


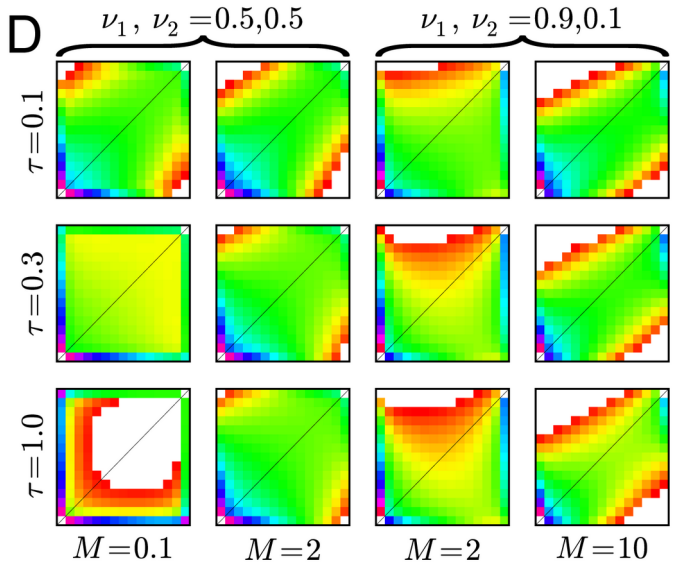
Method implemented in Dsuite

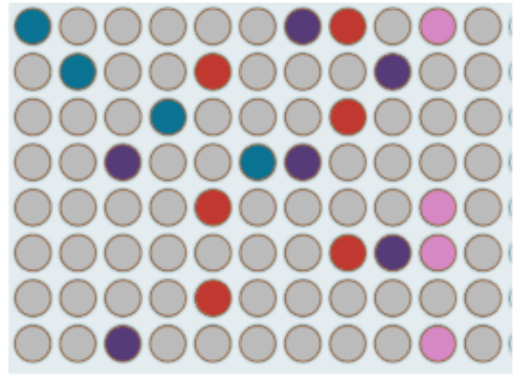


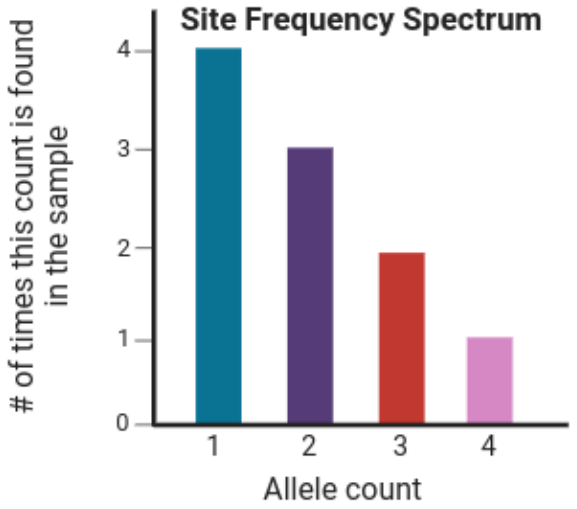
All these statistics describe the frequency spectrum (more tomorrow)









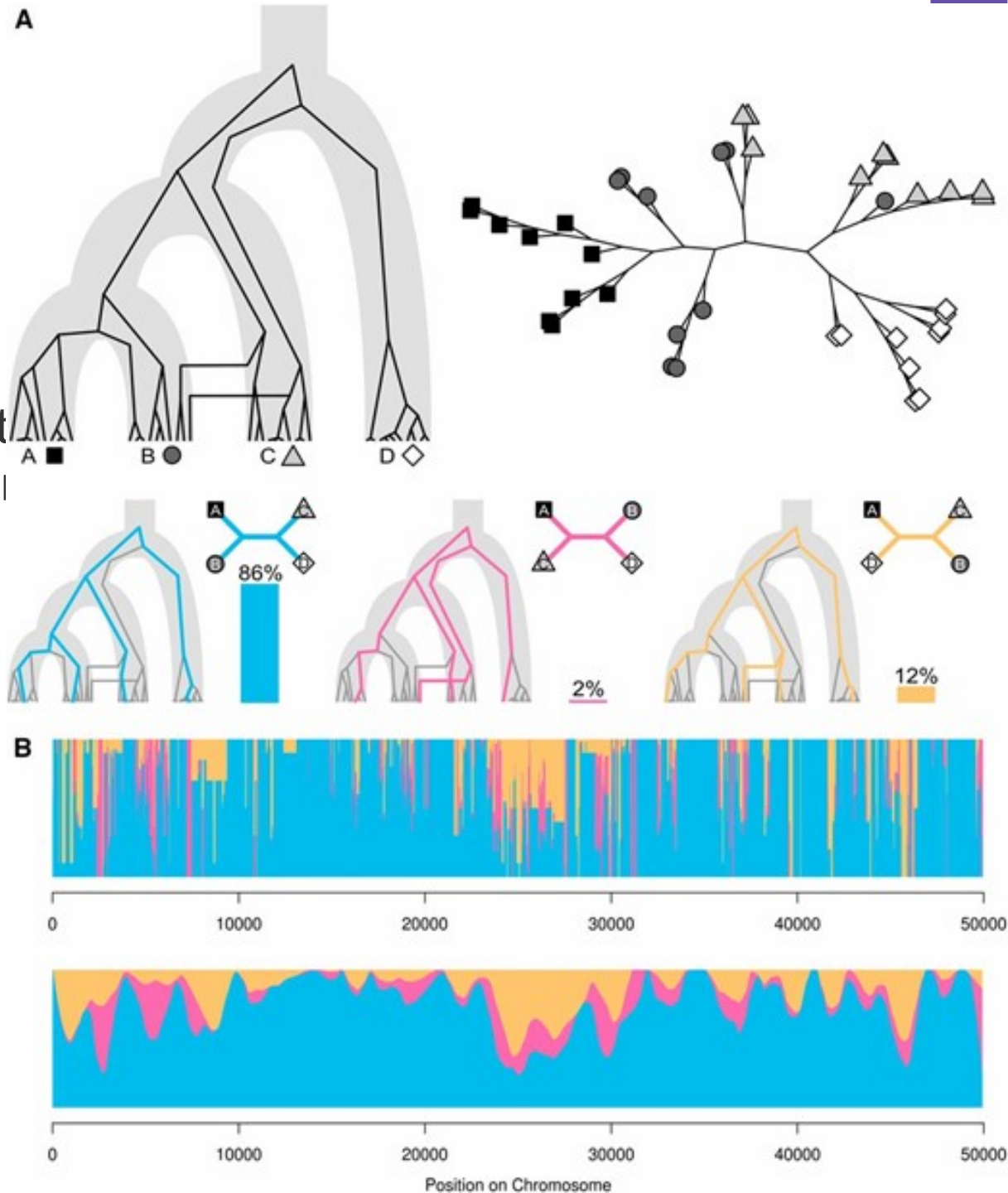


SCANNING STRUCTURE ALONG GENOMES

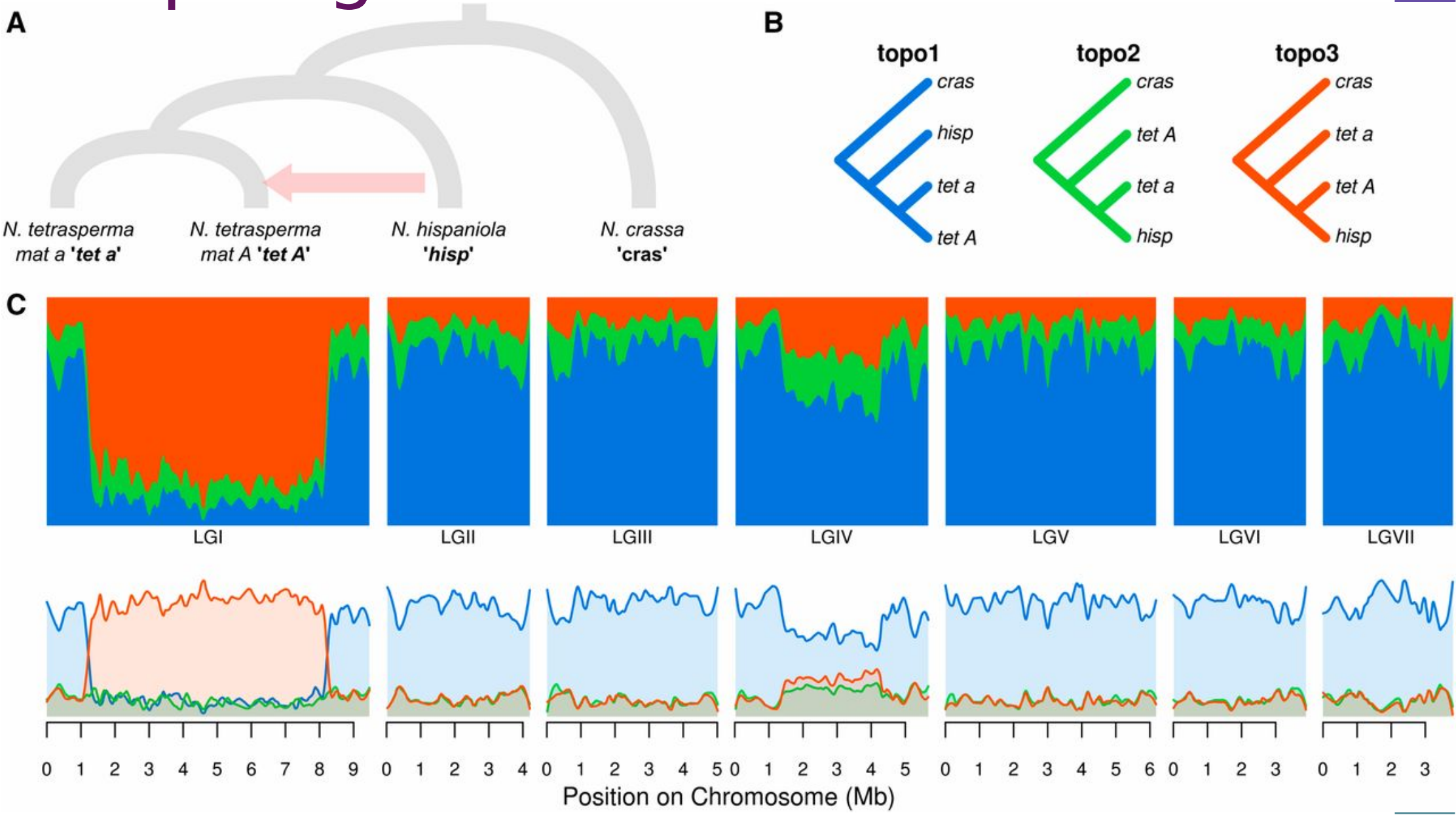
Local topologies

TWISST (https://github.com/simonhmartin/twisst)

- Topology weighting by iterative sampling of subtrees
- Estimates the relative contributions of predefined topologies to the observed local tree
- Useful to detect (adaptive?) introgression

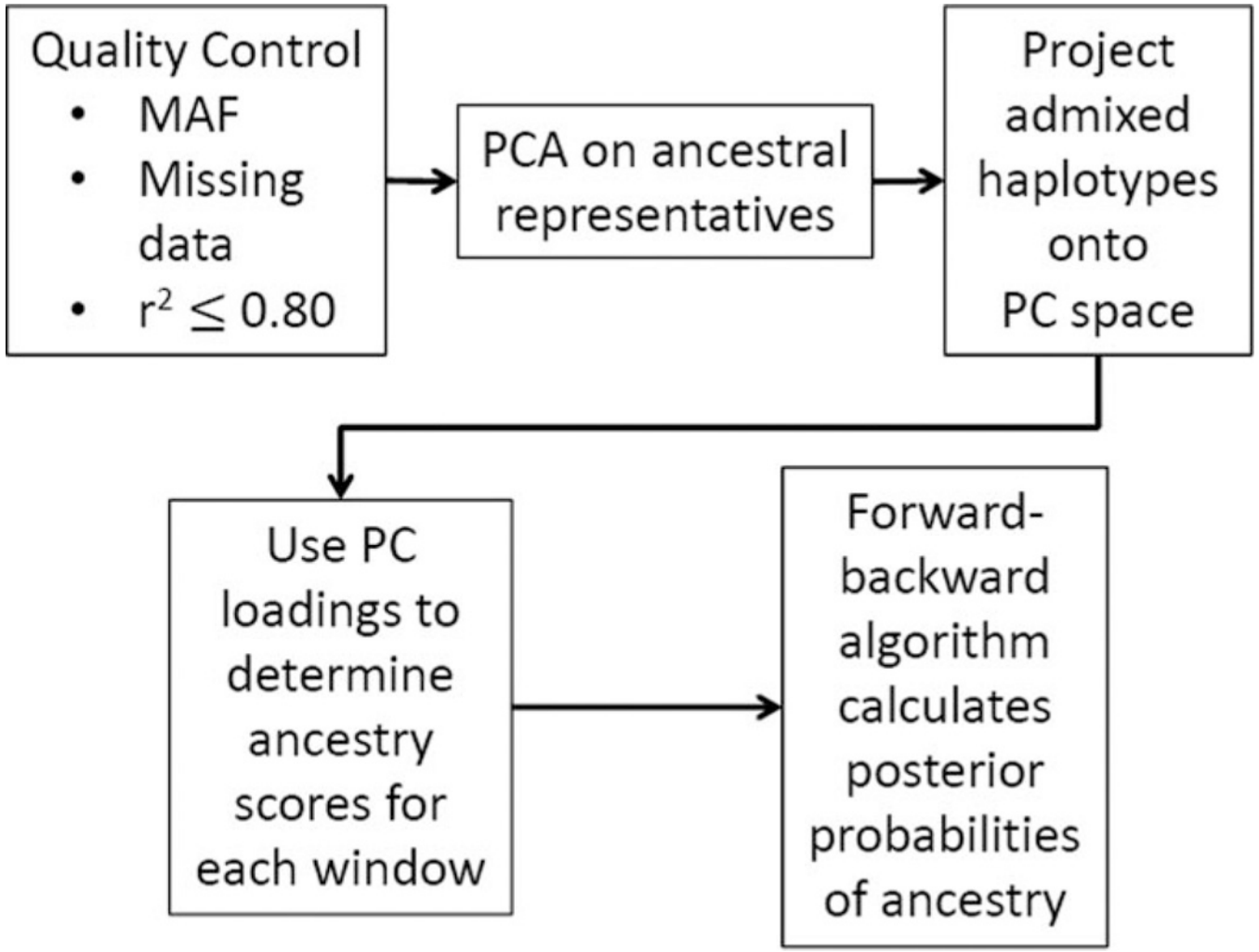


Local topologies

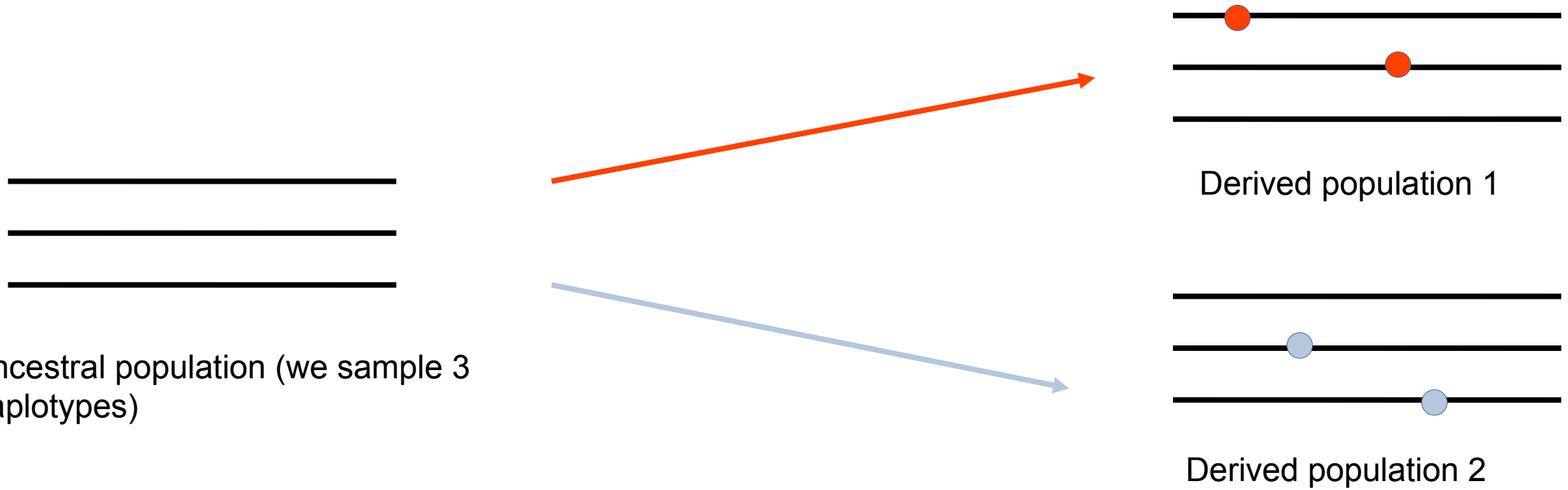


Local ancestry

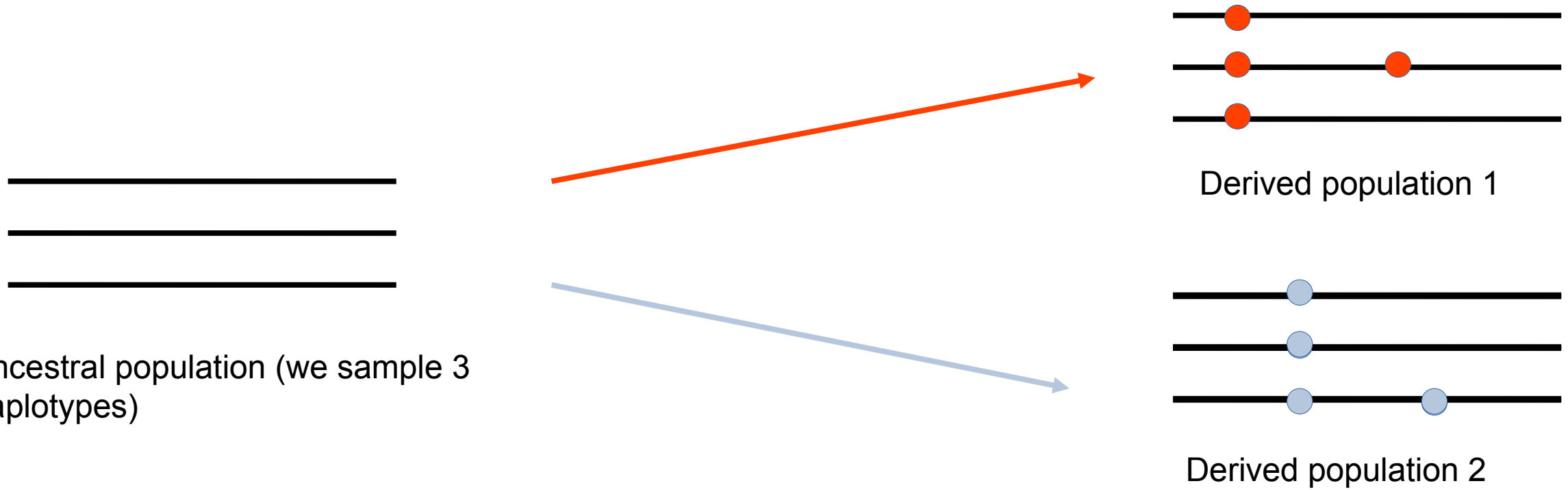
- PCAdmix
- Using local PCAs to contrast haplotype proximity between samples
- Assigns individuals to predefined populations (but are those representative?)
- Same principle behind some tests for selection (e.g. PCAdapt).



Measures of ‘absolute’ and ‘relative’ divergence

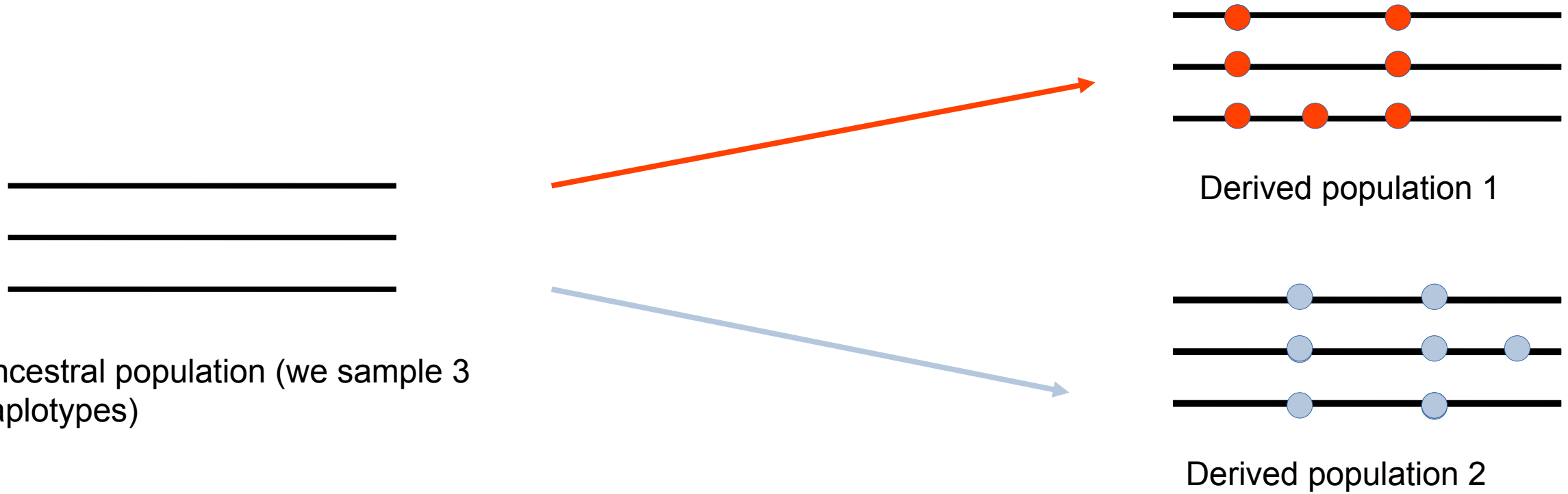


Measures of ‘absolute’ and ‘relative’ divergence



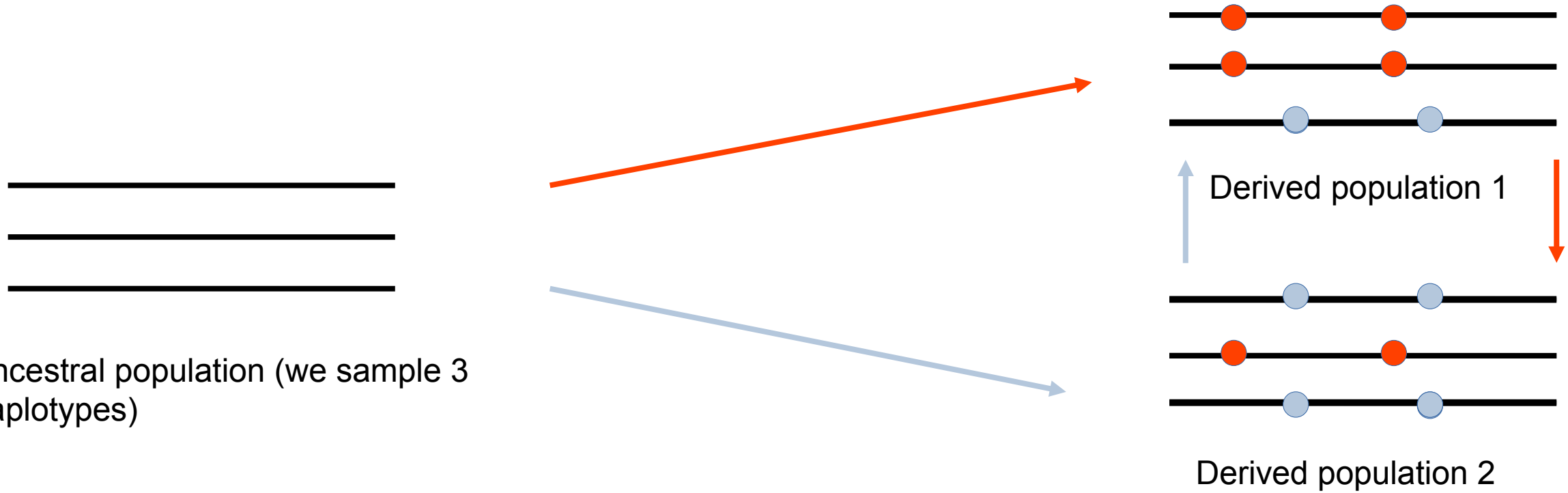
Ancestral population (we sample 3 haplotypes)

Measures of ‘absolute’ and ‘relative’ divergence



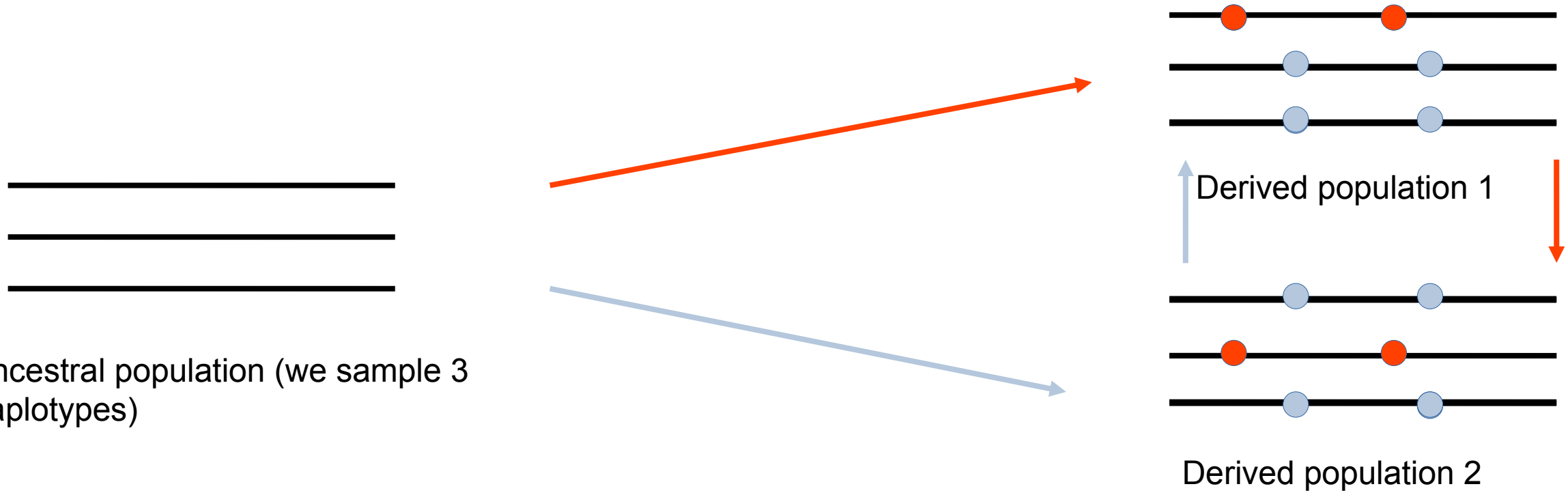
Ancestral population (we sample 3 haplotypes)

Measures of ‘absolute’ and ‘relative’ divergence



Ancestral population (we sample 3 haplotypes)

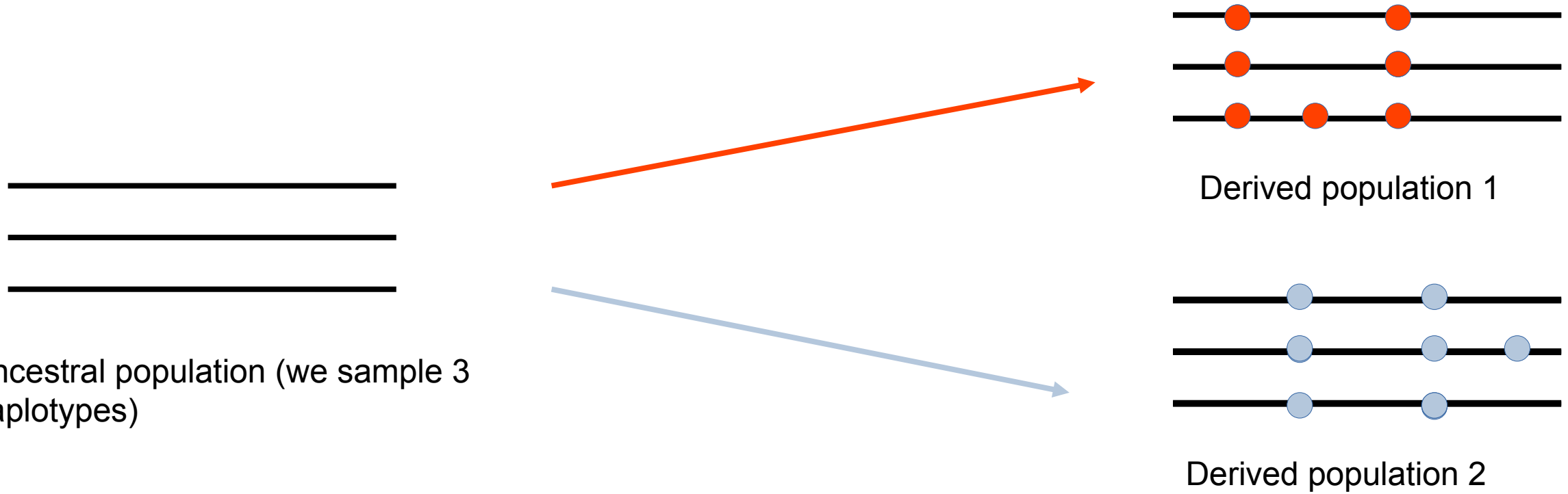
Measures of ‘absolute’ and ‘relative’ divergence



Ancestral population (we sample 3 haplotypes)

We can see that alleles are not fixed anymore : F_{ST} drops
There are also fewer differences between pairs of haplotypes taken from the two populations.
This statistics, d_{XY}, drops too

Islands resisting gene flow



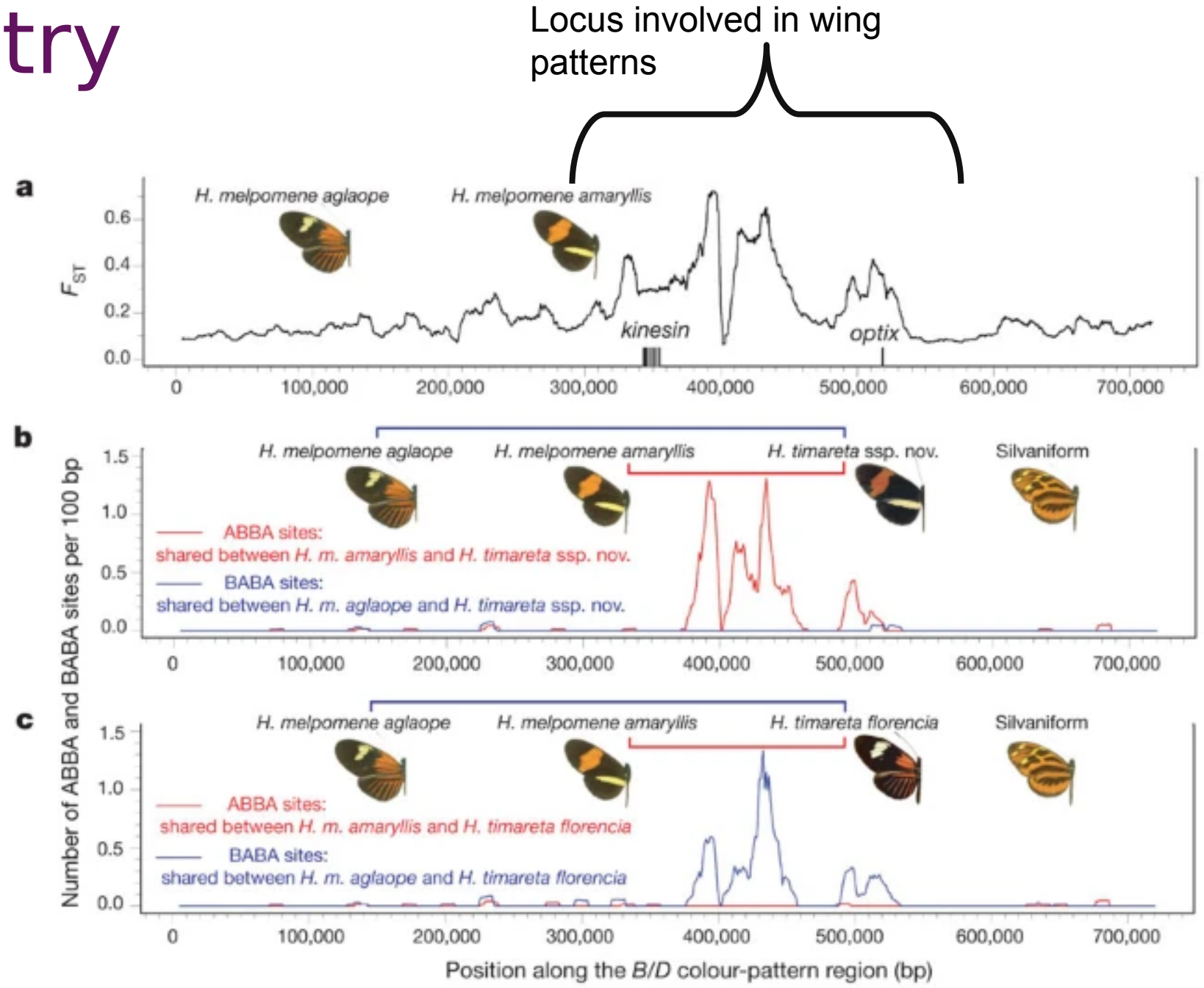
Ancestral population (we sample 3 haplotypes)

We can see that alleles are fixed : F_{ST} is high
There are also more differences between pairs of haplotypes taken from the two populations.
This statistics, d_{XY}, should be high too

More about these measures here

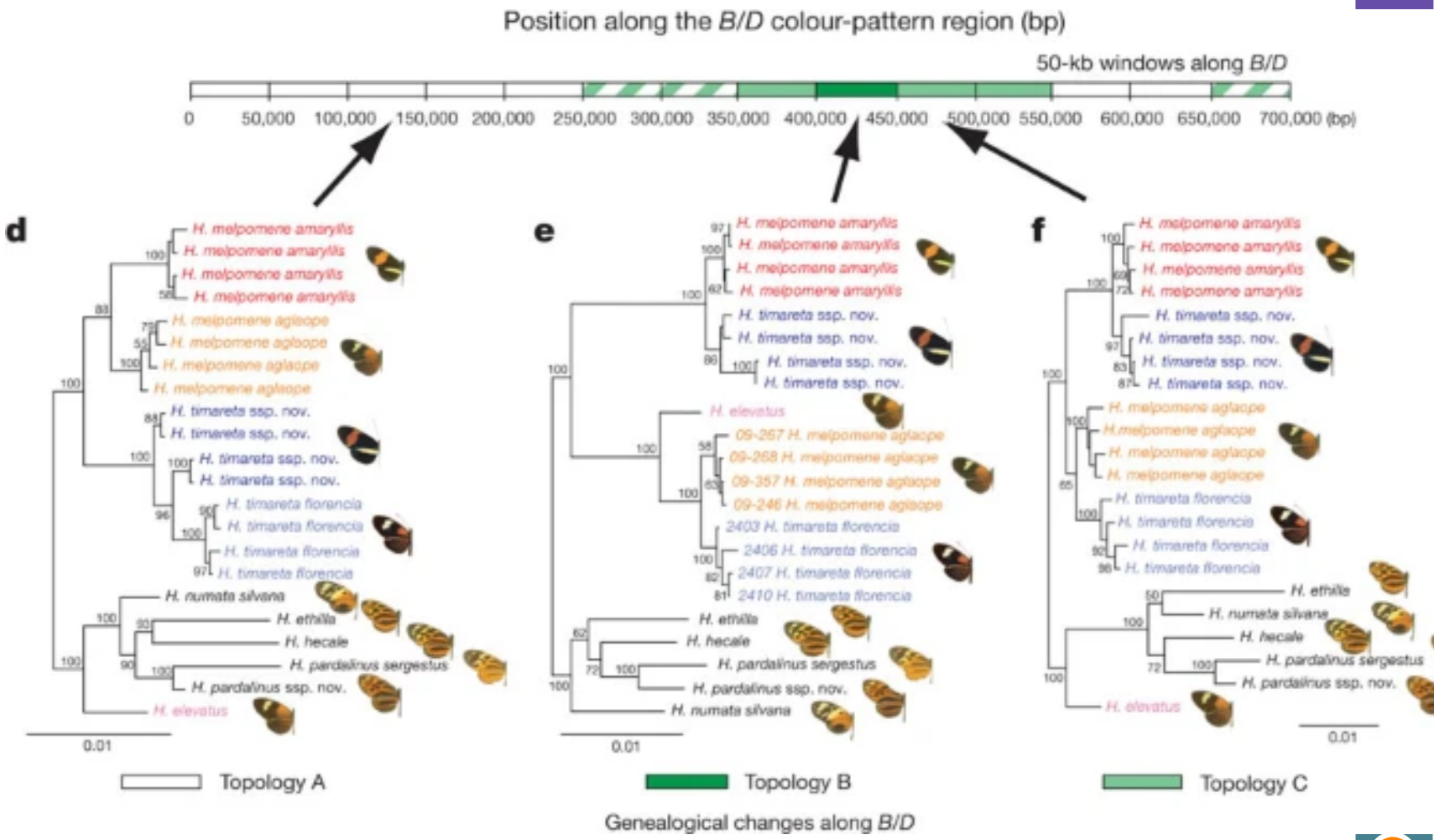
Local ancestry

- Scan for local ABBA-BABA statistics.
- Scan for F_{ST}
- Examine topologies
- In the next lectures, we will see how we can refine even more these analyses.

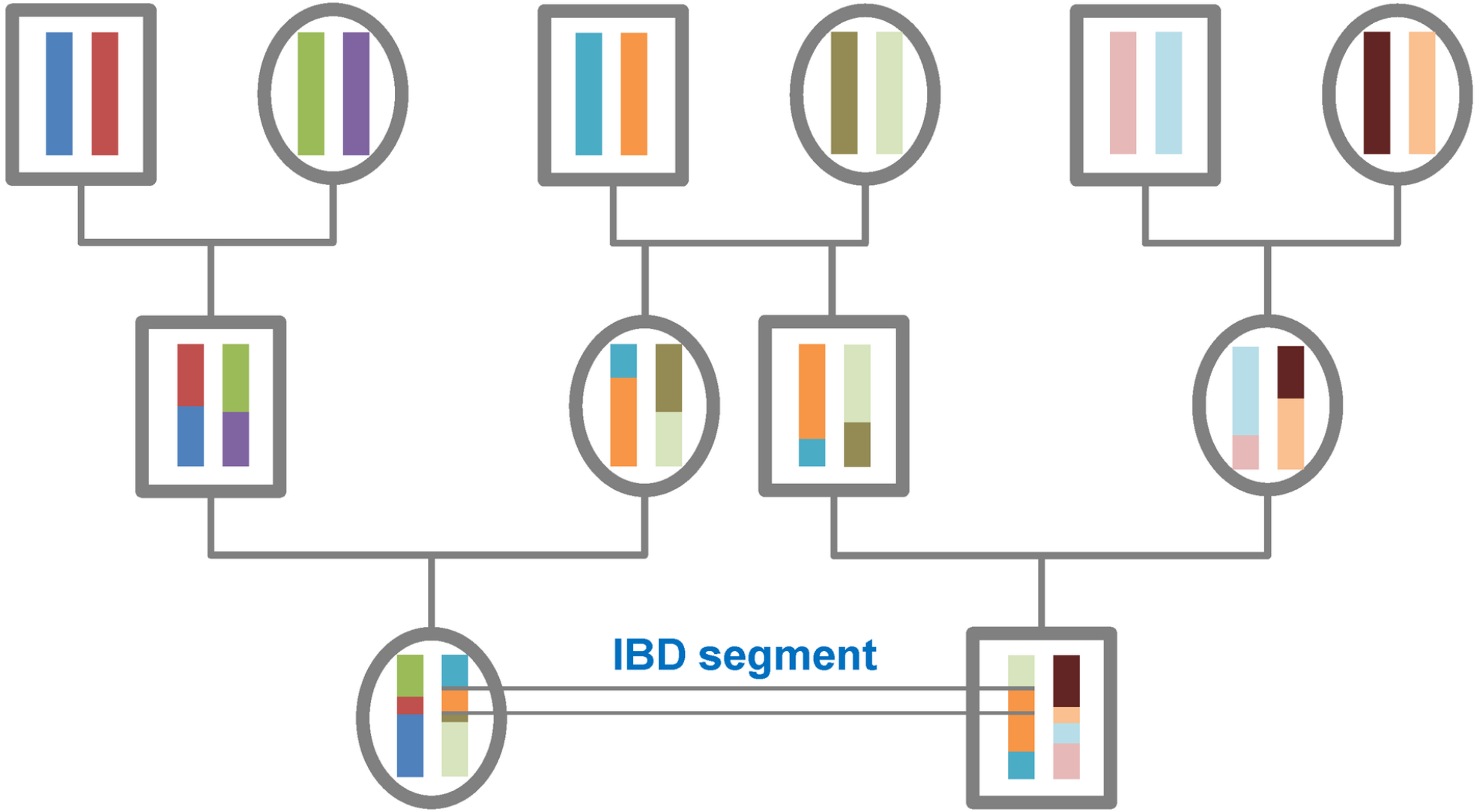


Local ancestry

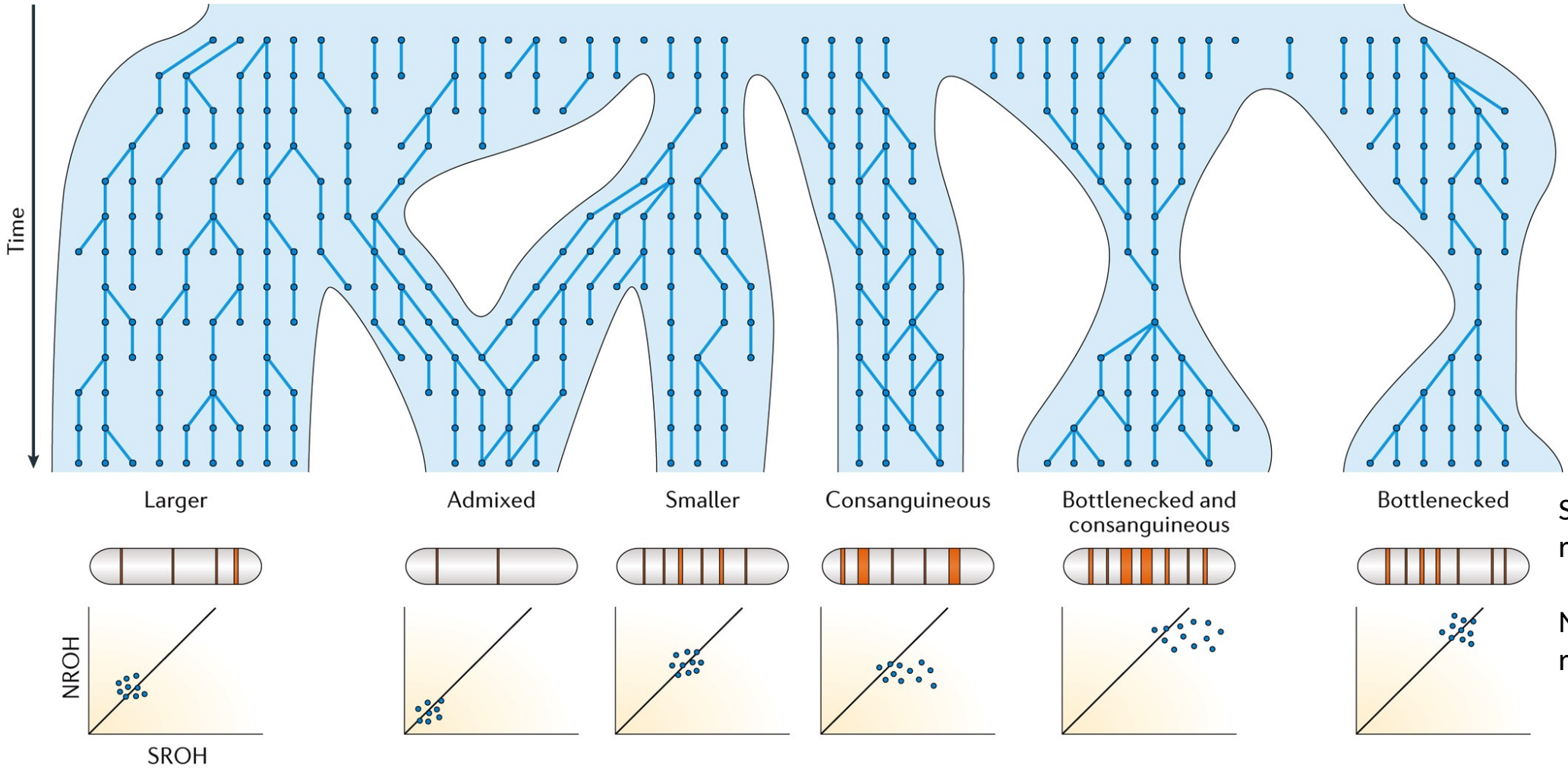
- Scan for local ABBA-BABA statistics.
- Scan for F_{ST}
- Examine topologies
- In the next lectures, we will see how we can refine even more these analyses.



Back to Identity by descent



Back to Identity by descent



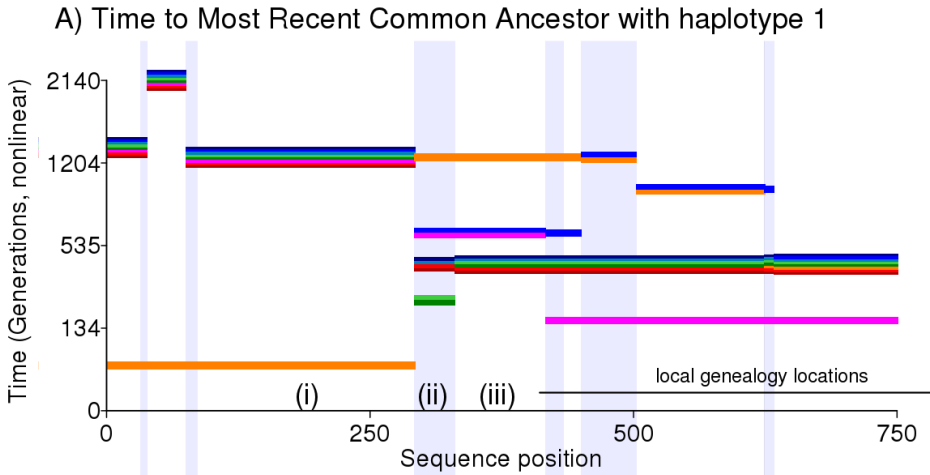
SROH: total length of runs of homozygosity

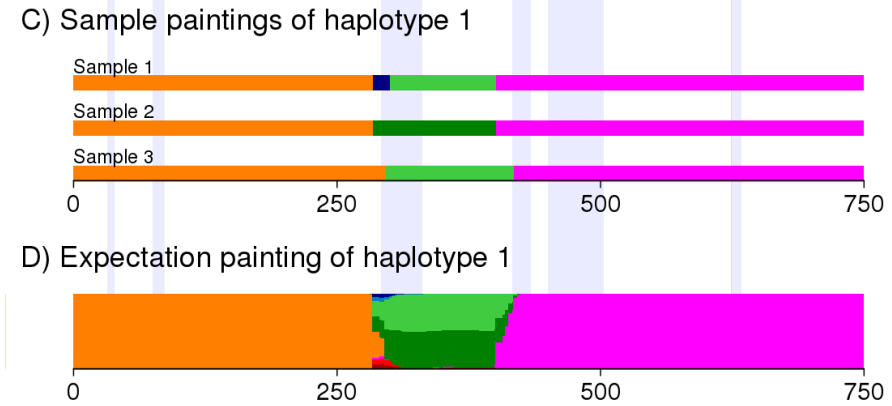
NROH: total number of runs of homozygosity

Ceballos 2018 Nat Rev Gen

Chromosome painting helps interpreting clustering.

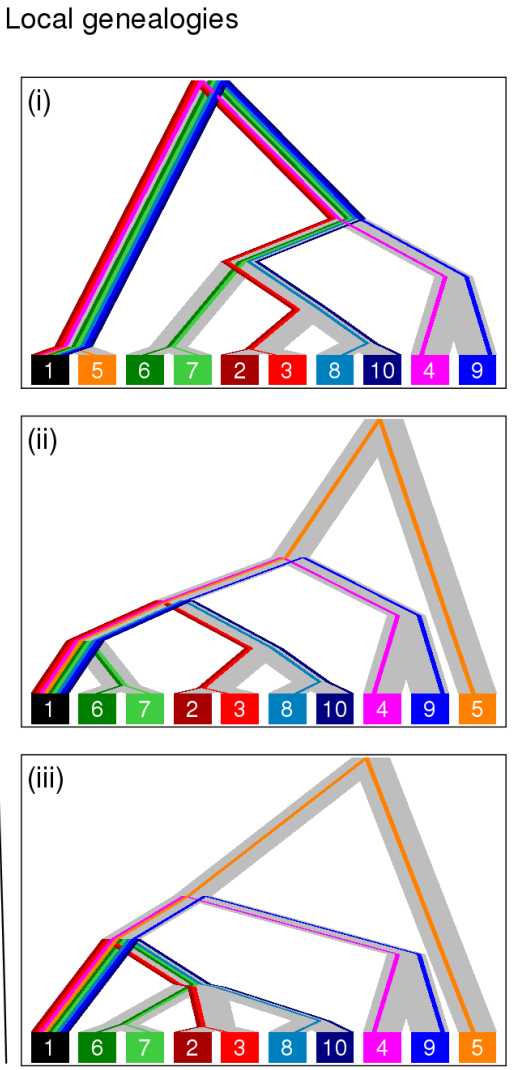
- Chromopainter
- BadMixture, Globetrotter, FineStructure.
- Uses patterns of genome-wide similarities
- Can quantify and provide time estimates of admixture (Globetrotter)





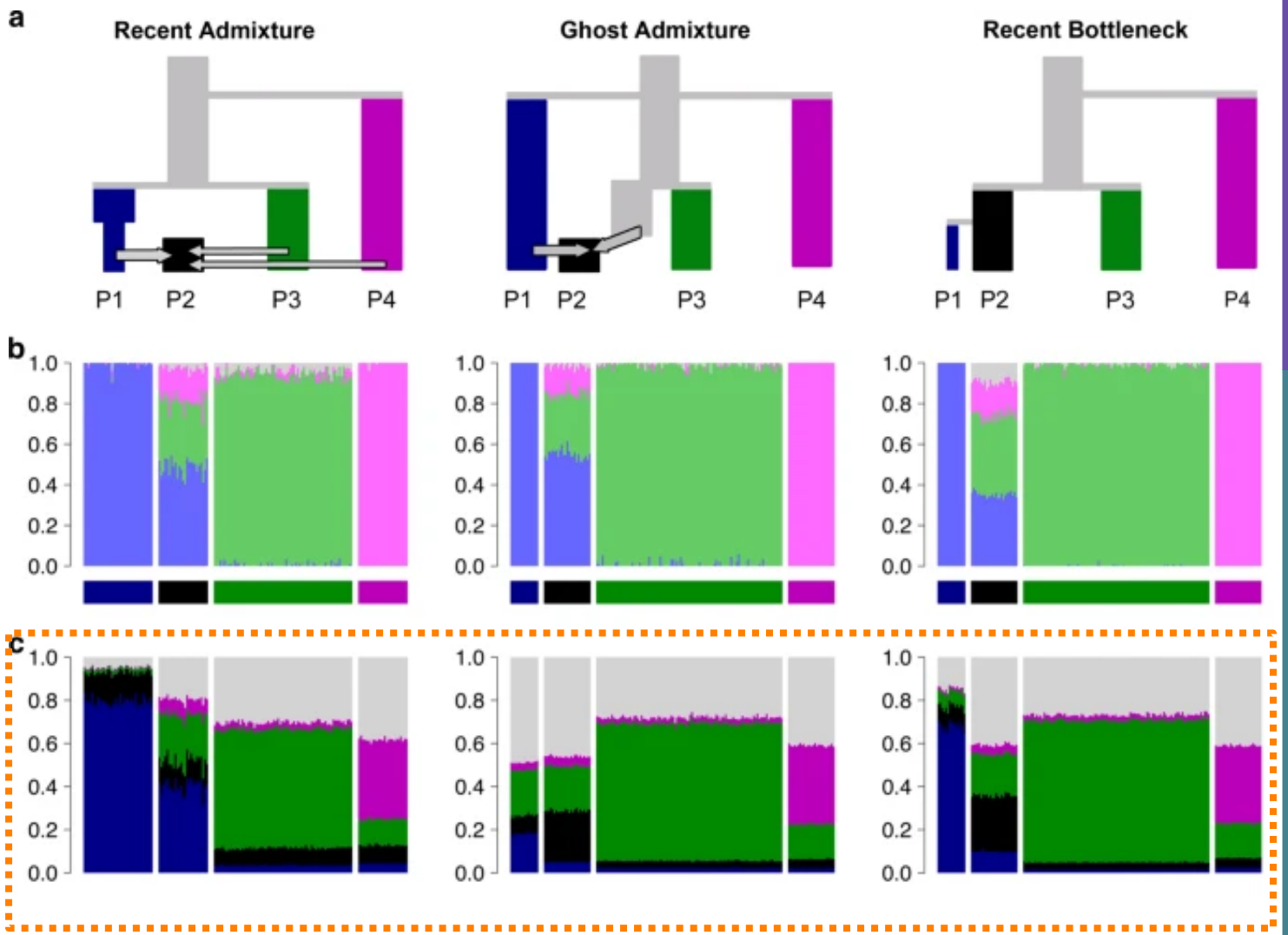
E) Coancestry matrix row for haplotype 1

	Donor haplotype									
	1	2	3	4	5	6	7	8	9	10
Haplotype 1	0	0.08	0.09	1.1	1.24	0.52	0.52	0.06	0.01	0.06



Making clustering a bit better

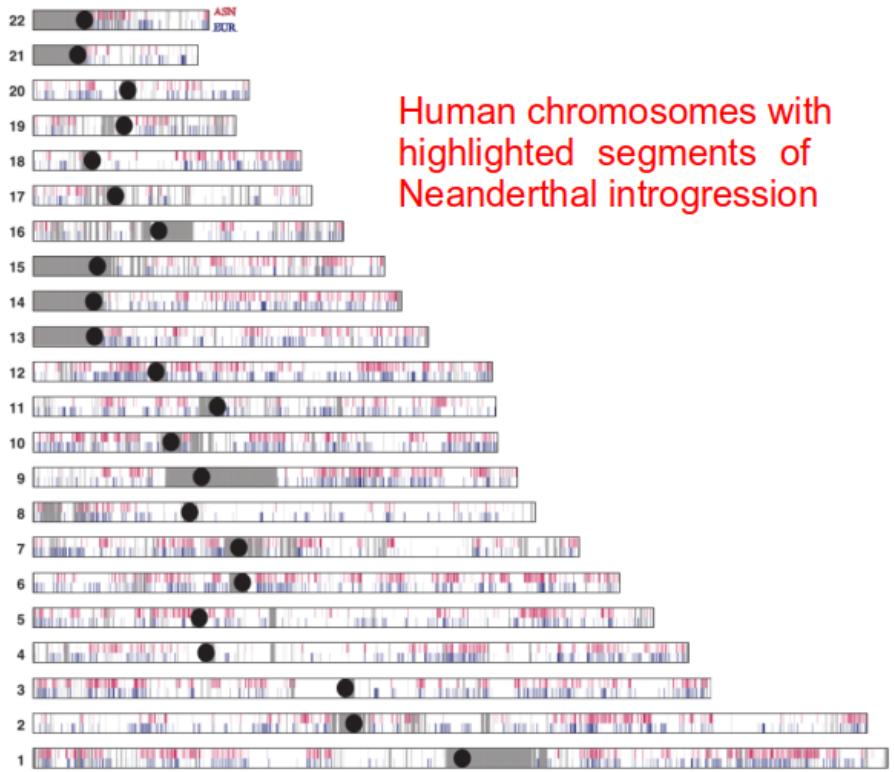
- An example: BadMixture
- Possible to incorporate information about genome-wide allele sharing.
- An unlinked version can also be used for reduced representation data (allele-sharing instead of haplotype-sharing).



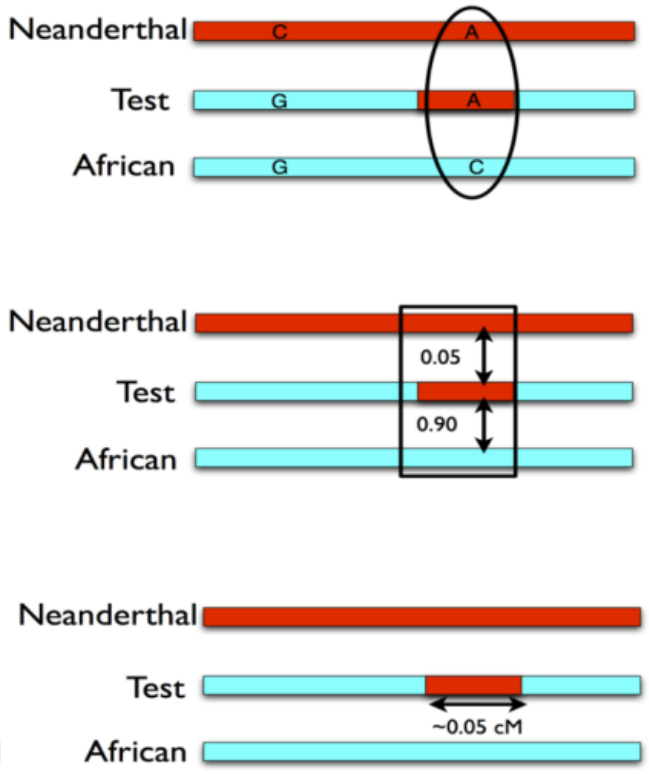
‘Colour palette’: Proportions of closest haplotypes found along the genome

Application of chromosome-painting

- Possible to identify genes under recent or ancient selection, having allele frequencies inconsistent with the majority of genotypes along the genome
- The length of IBD fragments depends on their age
- Introgression from other species or populations

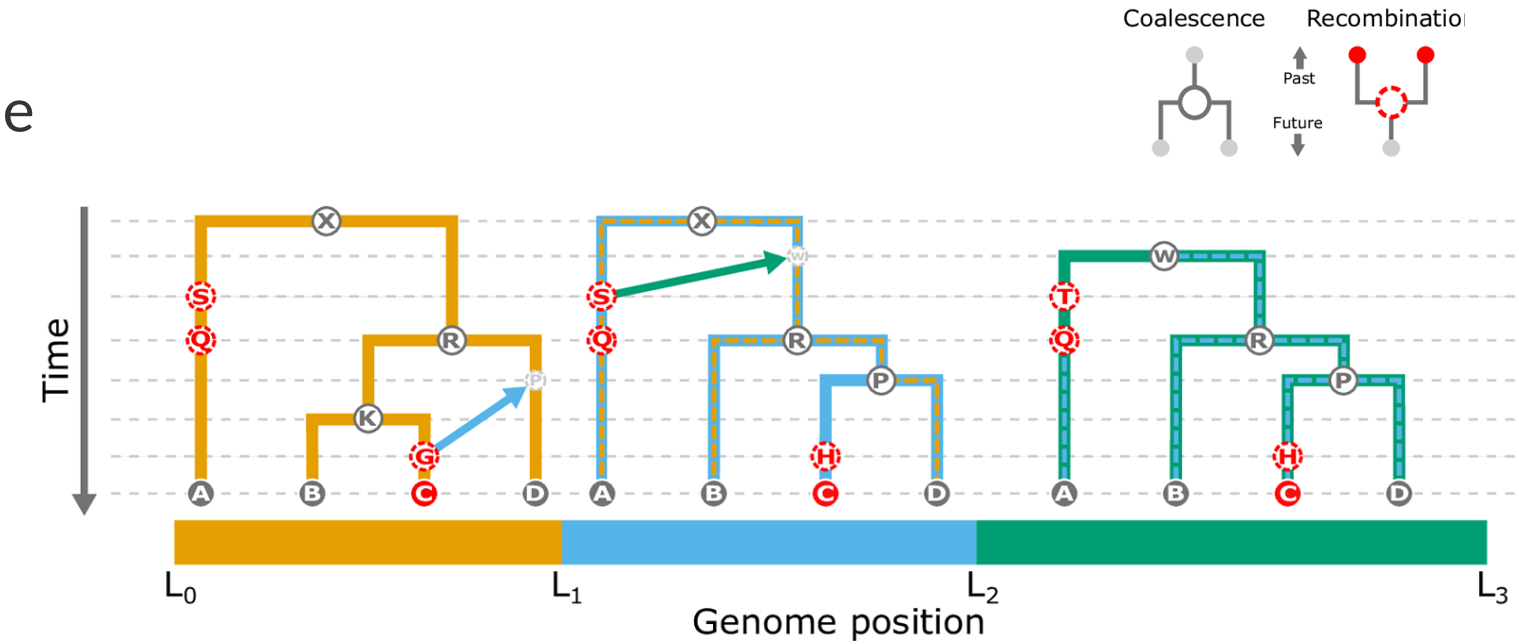


Human chromosomes with highlighted segments of Neanderthal introgression



Ancestral recombination graphs

- The complete summary of the recombination and coalescence process along a sequence
- ARGWeaver (50-100 individuals at most, more precise).
- Relate (1000s of individuals).



Which of the following are valid approaches to study admixture and gene flow?

A - ABBA-BABA (D-statistics)
B - Chromosome painting (e.g., Globetrotter, FineStructure)
C - Estimating F_{ST} from local genomic windows
D - Maximum-likelihood tree assuming no introgression
E - PCAdmix for haplotype ancestry along the genome

Which of the following are valid approaches to study admixture and gene flow?

A - ABBA-BABA (D-statistics)
B - Chromosome painting (e.g., Globetrotter, FineStructure)
C - Estimating F_{ST} from local genomic windows
D - Maximum-likelihood tree assuming no introgression
E - PCAdmix for haplotype ancestry along the genome

So far...

- All of this is quite descriptive
- No clear idea of the underlying process
- What about time and space?
- A way around : explicitly test evolutionary models. This is what we will do tomorrow (Thibault Leroy).
- Another way around: introduce a spatial context (Friday's lecture)