Population structure

Fernando Racimo

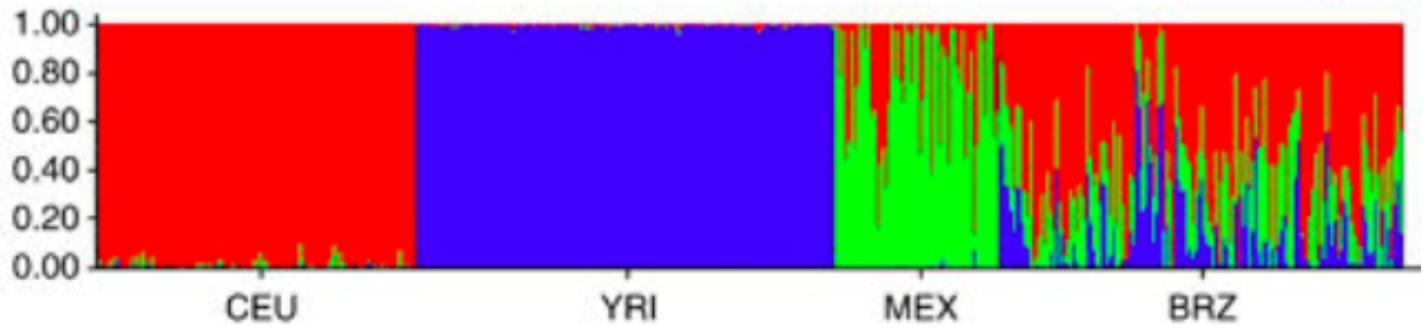
Adelaide, January 2018

Today

- F_{ST} and population structure
- The mixed-membership clustering model ("Structure")
- Leveraging haplotype information

Questions

- Is there population structure in a population?
- Can we identify subpopulation clusters of shared ancestry?
- Are individuals best modeled as mixtures of ancestral populations?
- How much admixture was passed on from each population?



Objectives

- Learn something about the pasty genetic history of a population under study
- Learn something about ourselves

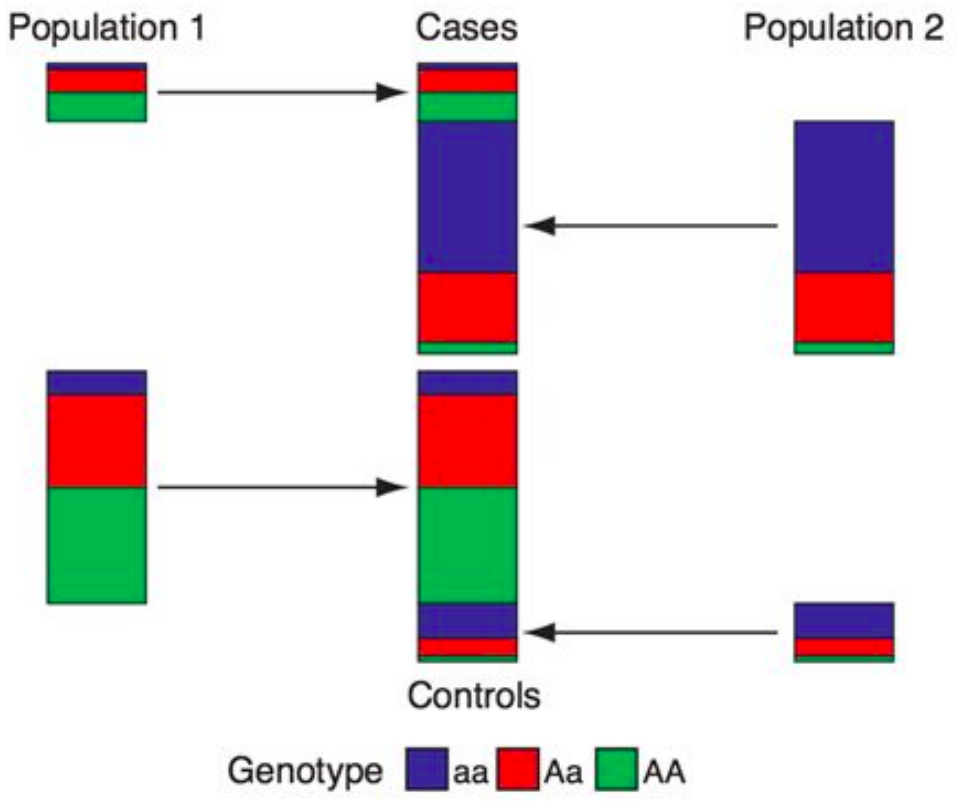
Fernando Racimo	100%
European	99.1% ›
Middle Eastern & North African	0.4%
South Asian	0.3%
Sub-Saharan African	0.1%
Unassigned	< 0.1%

See all 31 tested populations

European	99.1%
Southern European	68.4%
Iberian	38.5%
Italian	22.1%
Balkan	0.6%
Broadly Southern European	7.2%
Northwestern European	25.2%
French & German	13.4%
British & Irish	3.0%
Finnish	0.1%
Broadly Northwestern European	8.7%

Objectives

- Correct for population stratification in downstream analyses (e.g. GWAS)

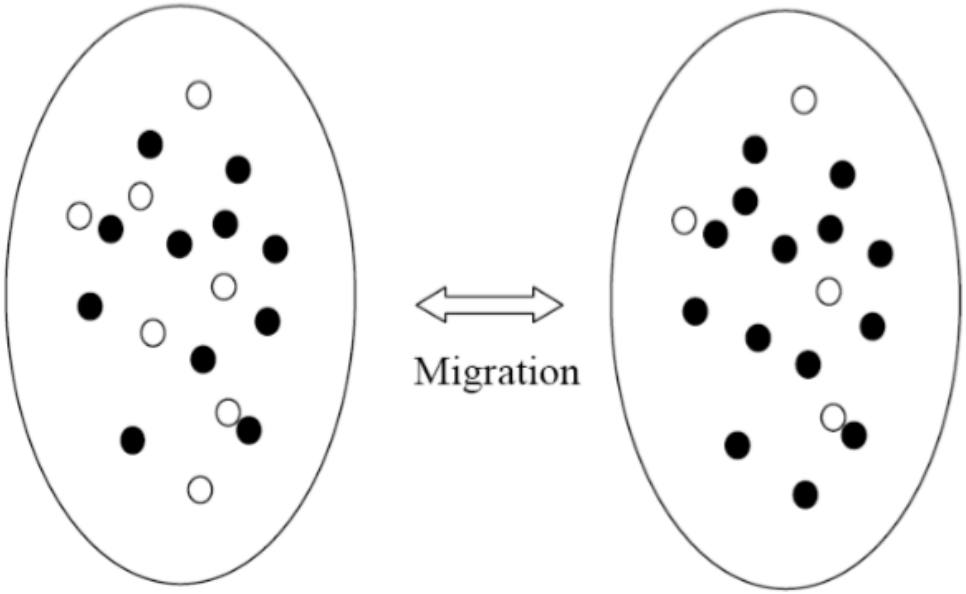


Today

- **F_{ST} and population structure**
- The mixed-membership clustering model ("Structure")
- Leveraging haplotype information

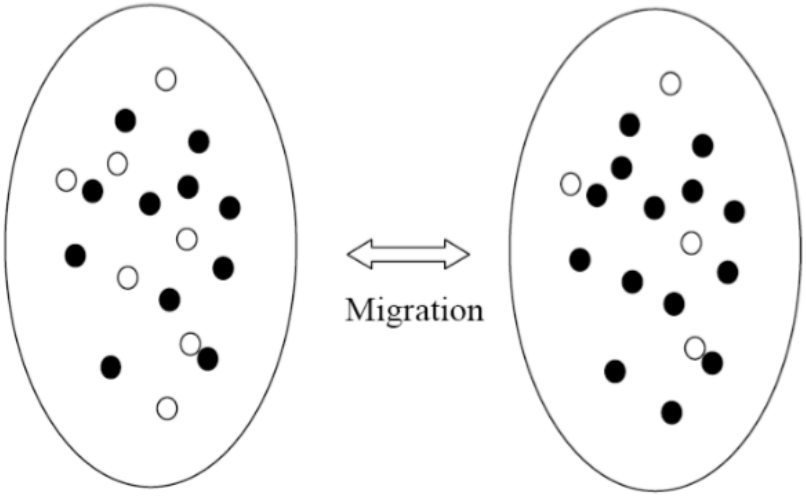
Population structure

- When a population is not randomly mating due to geographic structure, we say there is **population structure**.
- F_{st} is way to measure how much structure there is in a population.
- Before we define it, let us start by examining a simple model with two subpopulations, each of which contains the same number of individuals and is under HWE.



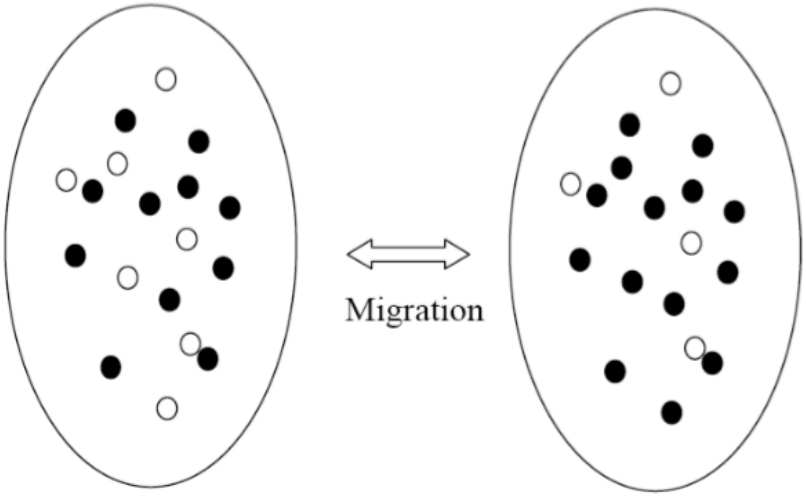
Population structure

- The frequency of an allele (A) in subpopulation 1 is f_{A1}. The frequency of the same allele in subpopulation 2 is f_{A2}.



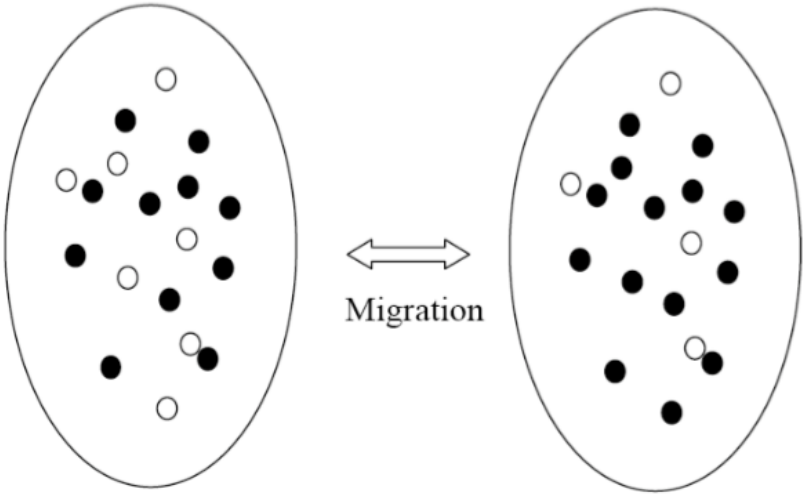
Population structure

- The frequency of an allele (A) in subpopulation 1 is f_{A1}. The frequency of the same allele in subpopulation 2 is f_{A2}.
- The frequency of the allele in the whole population is therefore $f_A = \frac{Nf_{A1}+Nf_{A2}}{2N} = \frac{f_{A1}+f_{A2}}{2}$.



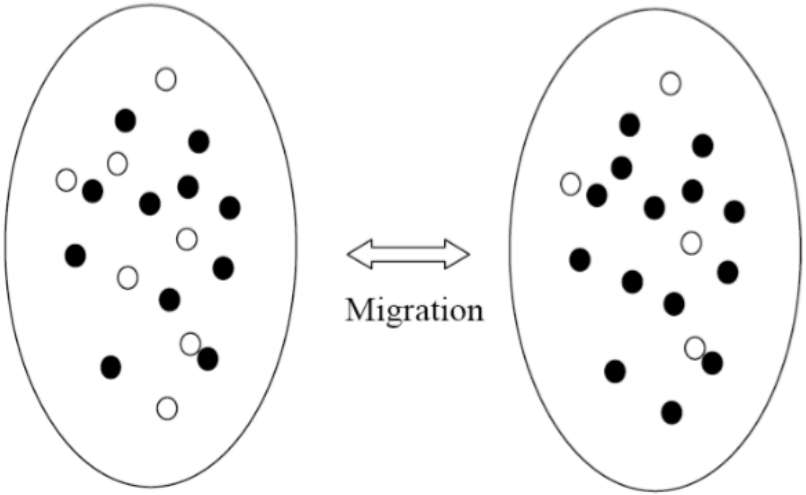
Population structure

- The frequency of an allele (A) in subpopulation 1 is f_{A1}. The frequency of the same allele in subpopulation 2 is f_{A2}.
- The frequency of the allele in the whole population is therefore $f_A = \frac{Nf_{A1}+Nf_{A2}}{2N} = \frac{f_{A1}+f_{A2}}{2}$.
- What is the expected heterozygosity under HW?



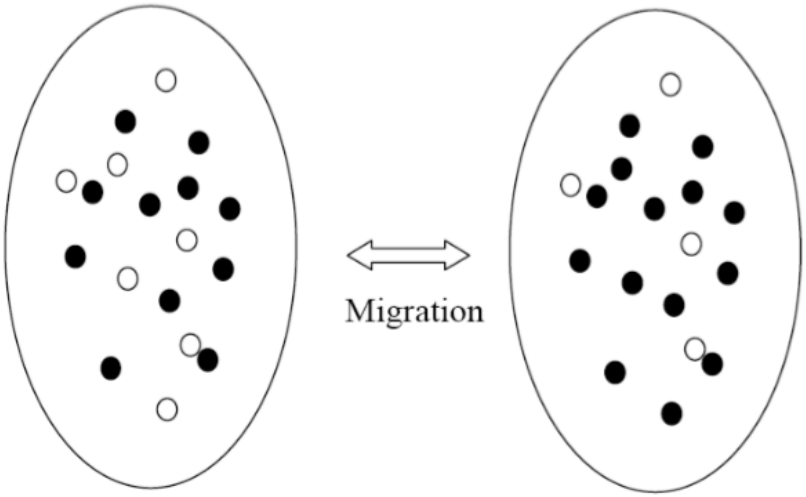
Population structure

- The frequency of an allele (A) in subpopulation 1 is f_{A1}. The frequency of the same allele in subpopulation 2 is f_{A2}.
- The frequency of the allele in the whole population is therefore $f_A = \frac{Nf_{A1}+Nf_{A2}}{2N} = \frac{f_{A1}+f_{A2}}{2}$.
- What is the expected heterozygosity under HW?
- Depends...



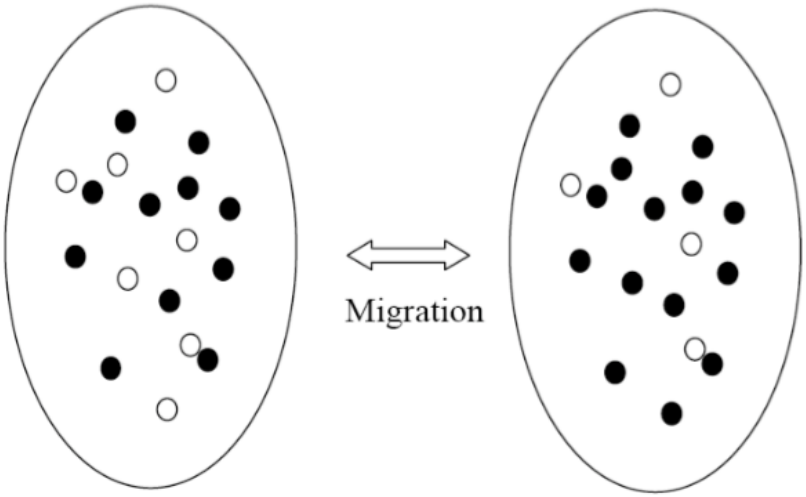
Population structure

- Recall that for a single population under HW, the expected heterozygosity is 2p(1-p).



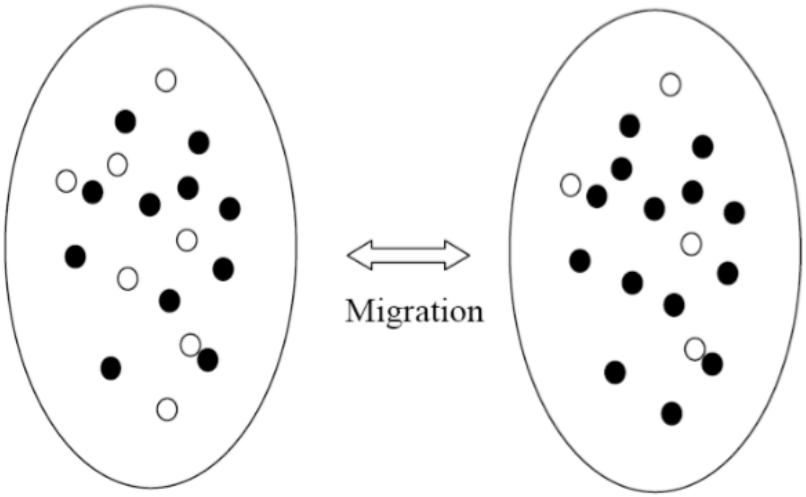
Population structure

- Recall that for a single population under HW, the expected heterozygosity is 2p(1-p).
- In this case, the expected heterozygosity is the weighted sum of each subpopulation's heterozygosity...



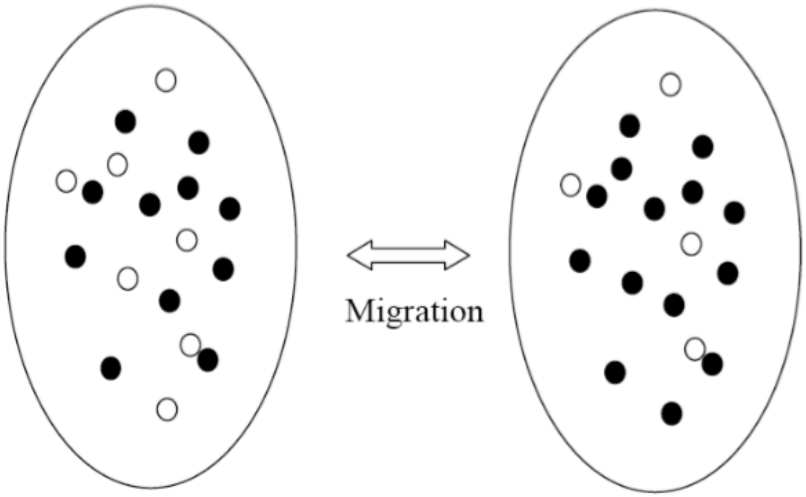
Population structure

- Recall that for a single population under HW, the expected heterozygosity is 2p(1-p).
- In this case, the expected heterozygosity is the weighted sum of each subpopulation's heterozygosity...
- $H_S = \frac{H_1 + H_2}{2}$



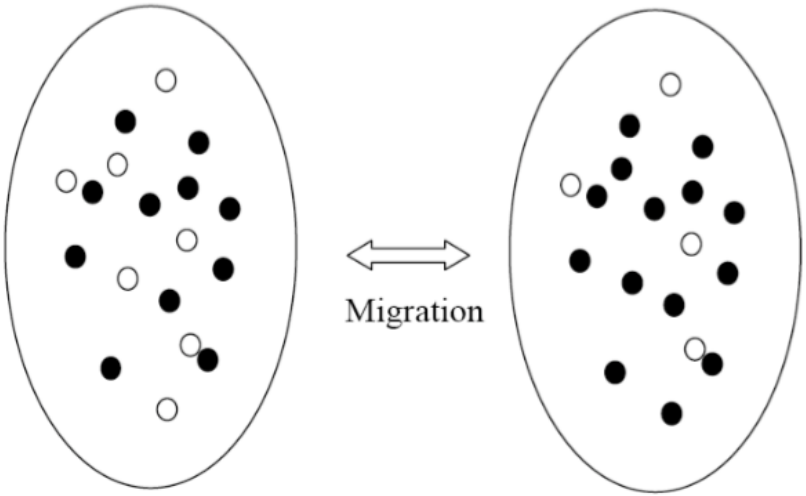
Population structure

- Recall that for a single population under HW, the expected heterozygosity is 2p(1-p).
- In this case, the expected heterozygosity is the weighted sum of each subpopulation's heterozygosity...
- $H_S = \frac{H_1 + H_2}{2}$
- $H_1 = 2f_{A1}(1 - f_{A1})$ and $H_2 = 2f_{A2}(1 - f_{A2})$



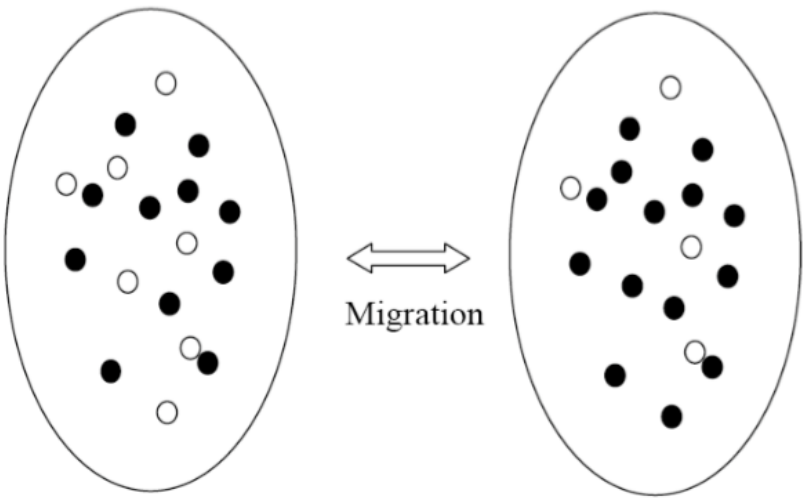
Population structure

- Recall that for a single population under HW, the expected heterozygosity is 2p(1-p).
- In this case, the expected heterozygosity is the weighted sum of each subpopulation's heterozygosity...
- $H_S = \frac{H_1 + H_2}{2}$
- $H_1 = 2f_{A1}(1 - f_{A1})$ and $H_2 = 2f_{A2}(1 - f_{A2})$
- Therefore, $H_S = (2f_{A1}(1 - f_{A1}) + 2f_{A2}(1 - f_{A2}))/2$



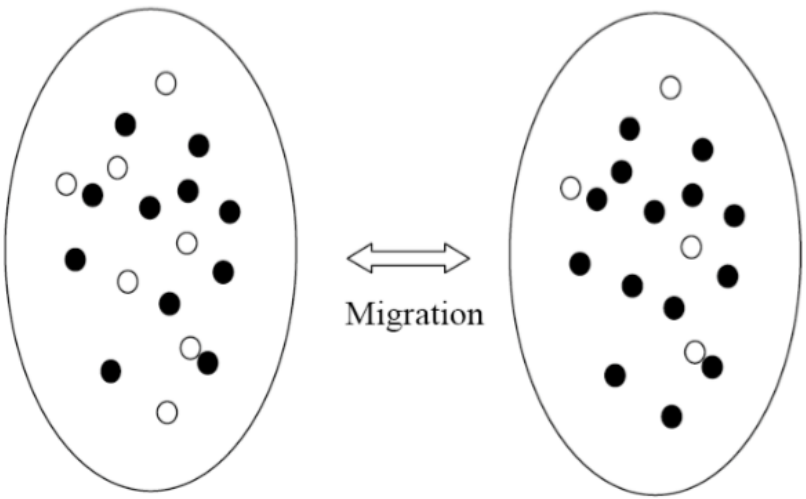
Population structure

- But if the two subpopulations are fully mixing with each other, then all we would need to calculate the expected heterozygosity is the total frequency of the A allele: f_A.



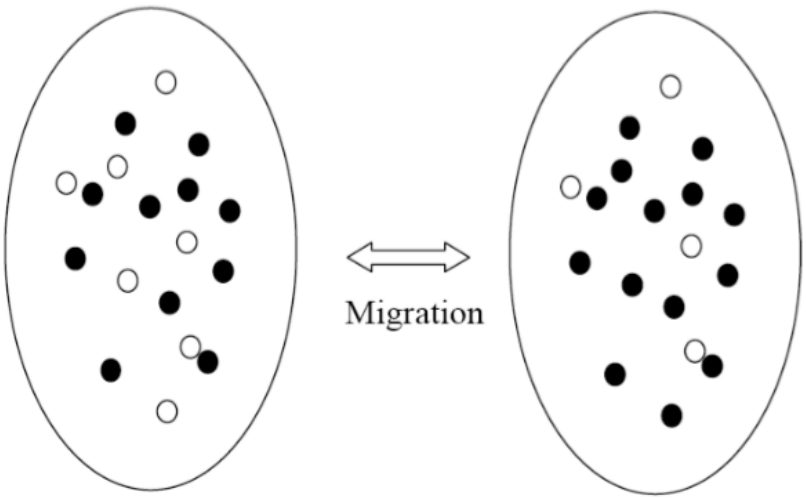
Population structure

- But if the two subpopulations are fully mixing with each other, then all we would need to calculate the expected heterozygosity is the total frequency of the A allele: f_A.
- $H_T = 2f_A(1 - f_A) = 2\frac{f_{A1}+f_{A2}}{2}(1 - \frac{f_{A1}+f_{A2}}{2})$



Wright's Fst

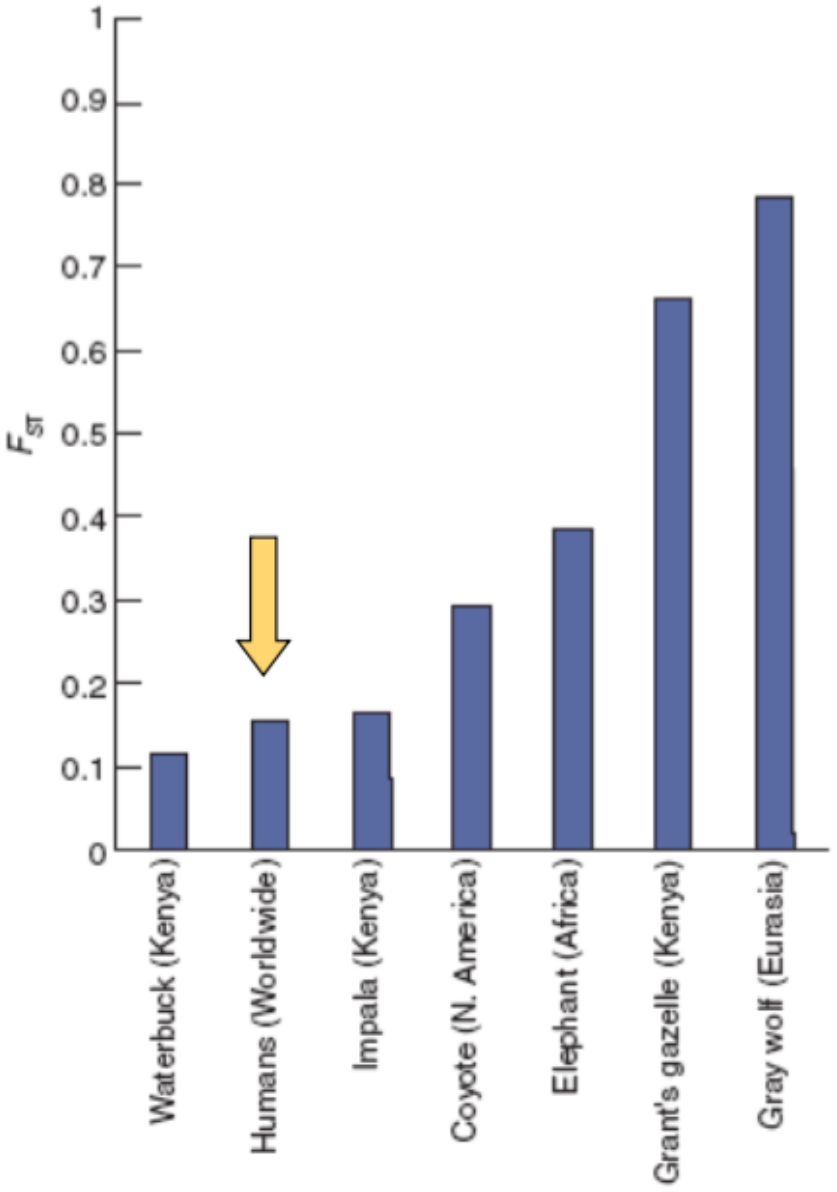
- The fixation index (F_{st}) is just the difference between H_T and H_S, weighted by H_T.
- $F_{st} = \frac{H_T - H_S}{H_T}$
- Can range from 0 (no differentiation) to 1 (fixation of different alleles in each population).



Wright's Fst

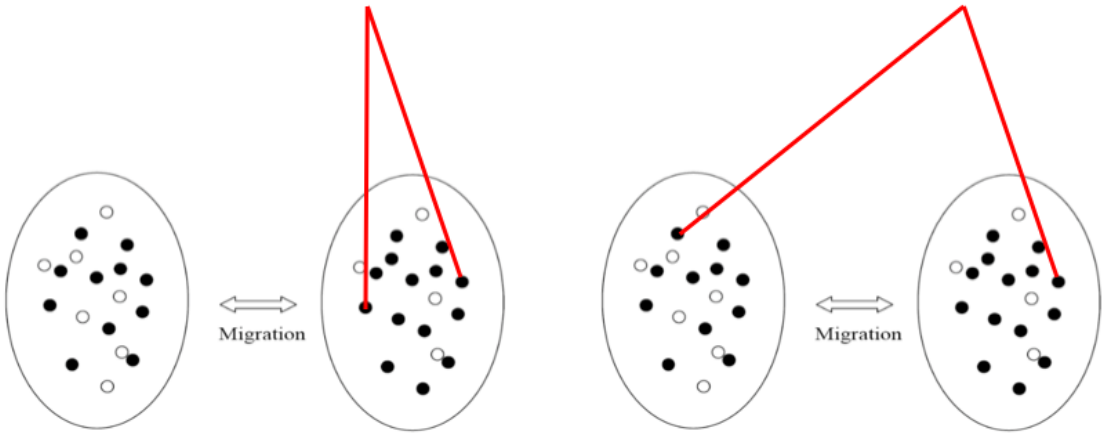
Most variation between populations

Most variation within populations



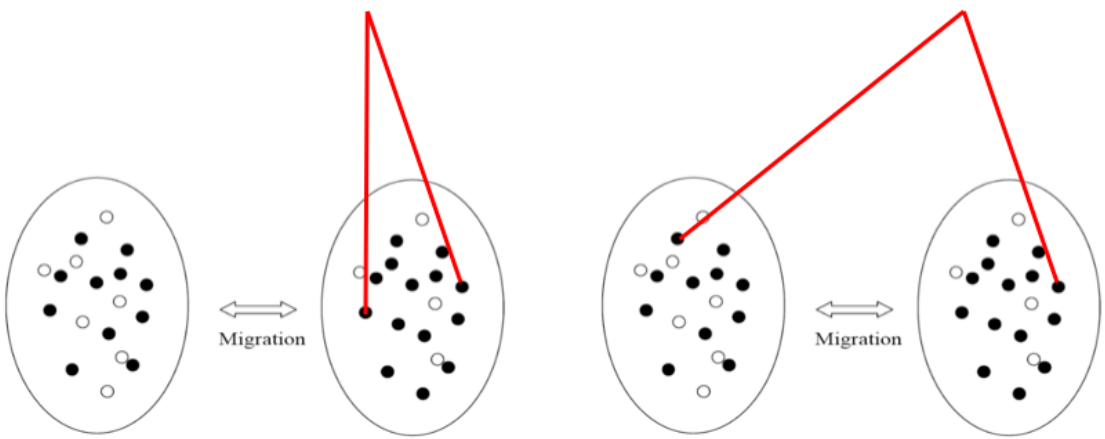
Interpreting Fst

- H_T is related to the expected time for any two samples to coalesce.
- H_S is related to the expected time for two samples FROM THE SAME SUBPOPULATION to coalesce.



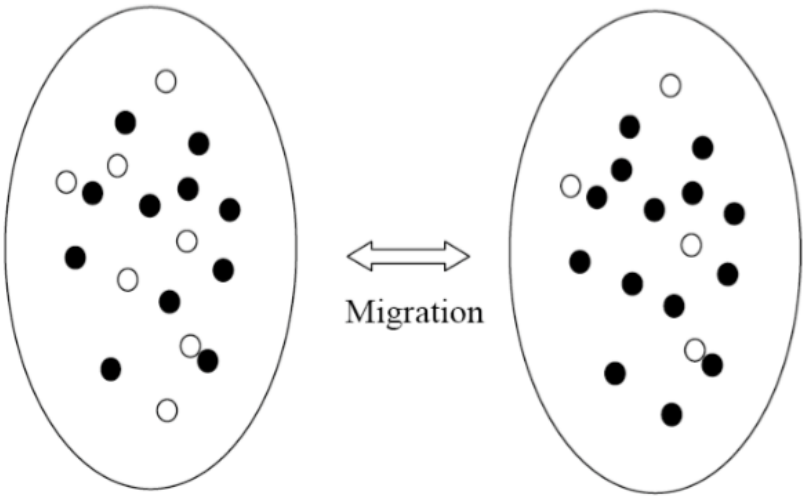
Interpreting Fst

- If $H_S < H_T$, the expected time for coalescence for two samples WITHIN THE SAME SUBPOPULATION is smaller than one would expect if the two populations were randomly mixing, so $F_{st} \approx 1$.
- If $H_S \approx H_T$, the expected time for coalescence for two samples WITHIN THE SAME SUBPOPULATION is approximately what one would expect for any two samples (regardless of which subpopulation they are sampled from), so $F_{st} \approx 0$.



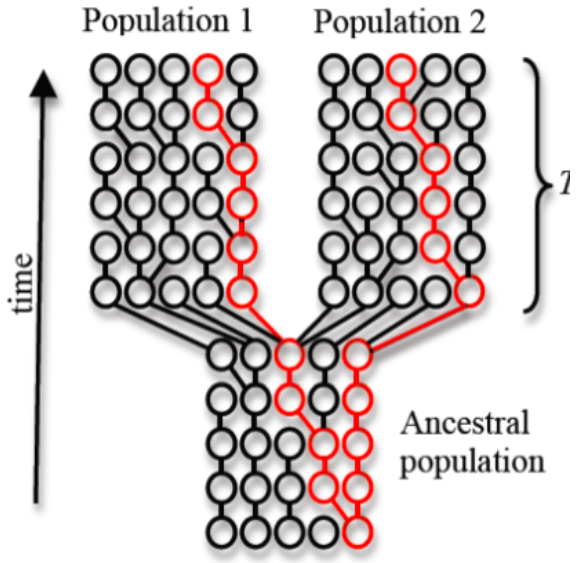
Interpreting Fst

- While F_{st} measures population structure, there are different models of structure that can generate the same F_{st} values.
- Under the two-population-with-migration model that we've been going over, one can show (Slatkin 1991) that:
- $F_{st} = \frac{1}{1+8Nm}$ where m is the migration rate between the populations.



Interpreting Fst

- Under a population-split model, F_{st} can be related to the amount of time since the split:
- $F_{st} = \frac{T}{T+2}$
- When T = 0, there is no population subdivision, and $F_{st} = 0$.
- When T is large, F_{st} approaches 1.



Wright's Fst

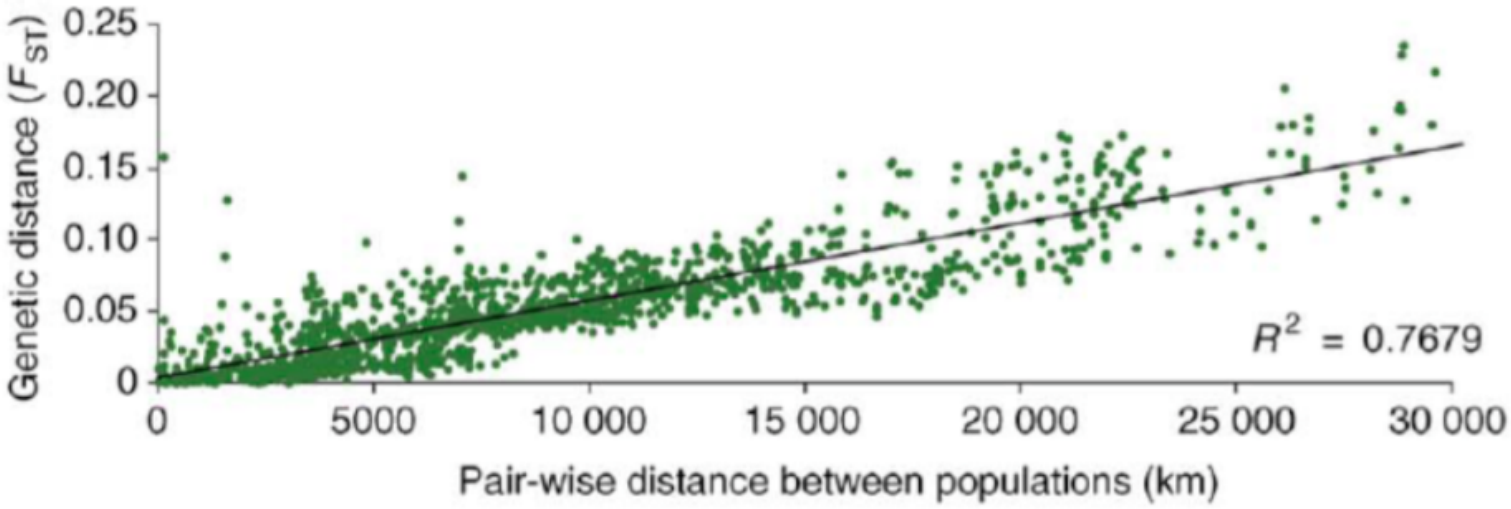


Figure 4.7. The relationship between FST and geographical distance between different pairs of human populations for more than 50 globally distributed populations. The figure is reproduced from Handley et al. 2007.

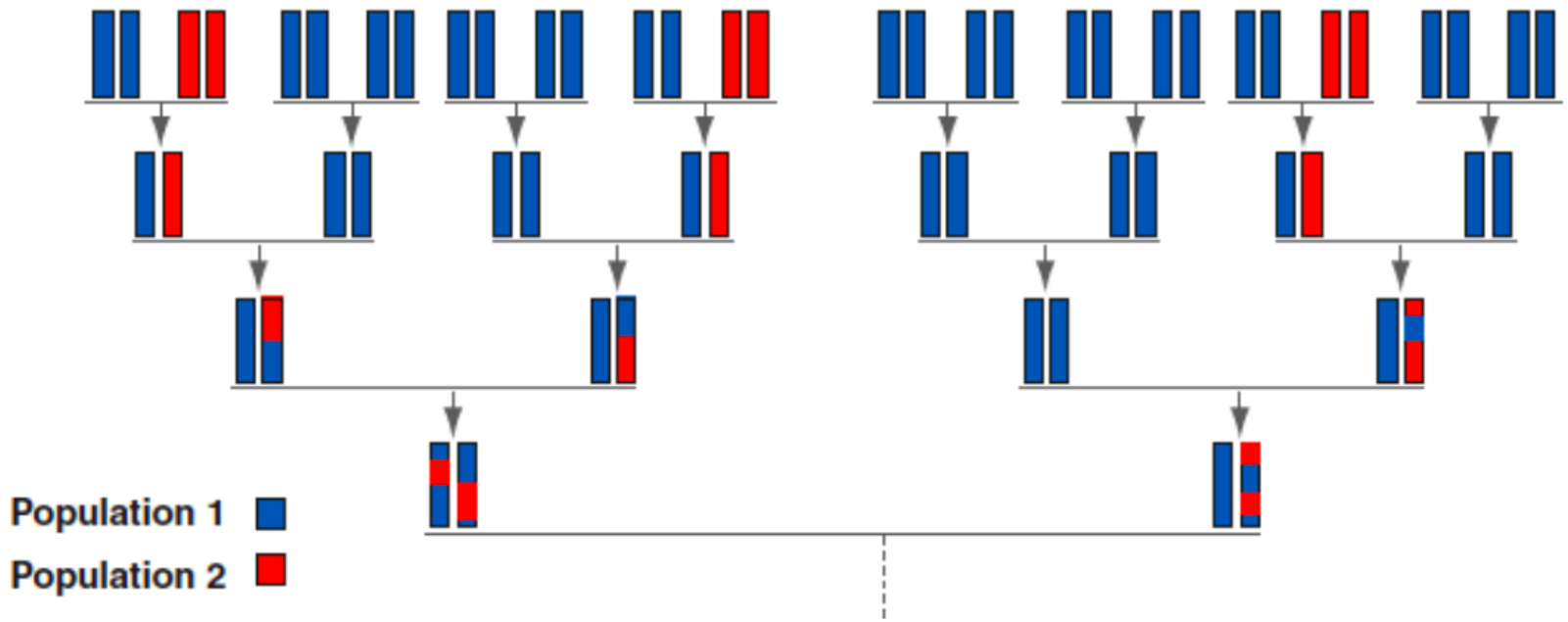
Today

- F_{ST} and population structure
- **The mixed-membership clustering model ("Structure")**
- Leveraging haplotype information

The "Structure" model

- The original model was first proposed by Pritchard et al. (2000)
- **Assumption 1**: each individual can be modeled as a mixture of one or more ancestral **"source populations"**
- **Assumption 2**: each locus is independent
- The proportion of genetic matrerial from each source in each individual is called the **"admixture proportion"**
- **Problem 1**: we don't know the identity and number of these source populations
- **Problem 2**: we don't know the admixture proportions
- **Objective**: find best-fitting sources and their proportions

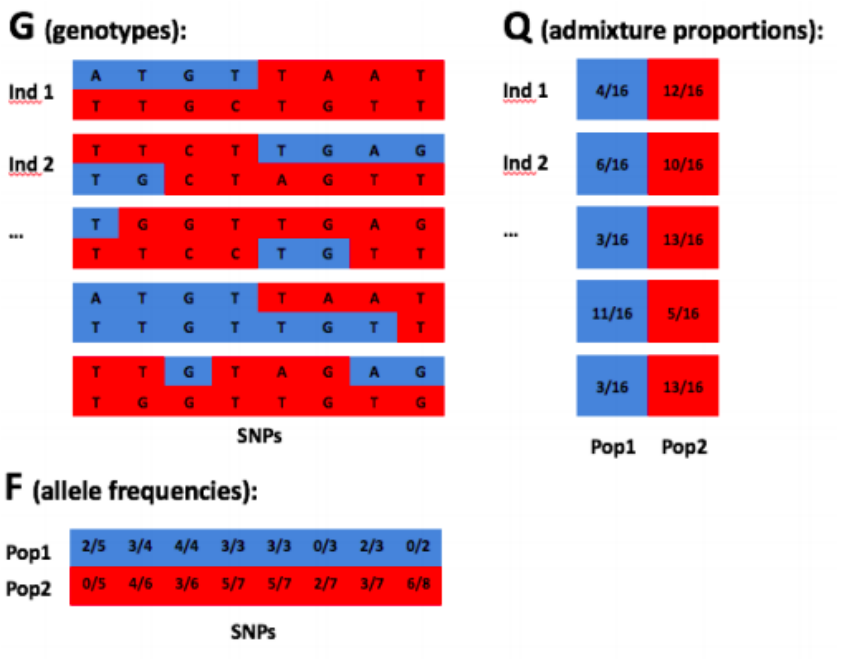
The "Structure" model



[0]Darvasi and Shifman 2005

The "Structure" model

- Known: genotypes (G)
- Unknown:
  - admixture proportions (Q)
  - allele frequencies in source populations (F)
- Need to estimate Q and F, given that we know G.
- Objective: Maximize likelihood function: $P[G|Q,F]$



[0] Ida Moltke pers. comm.

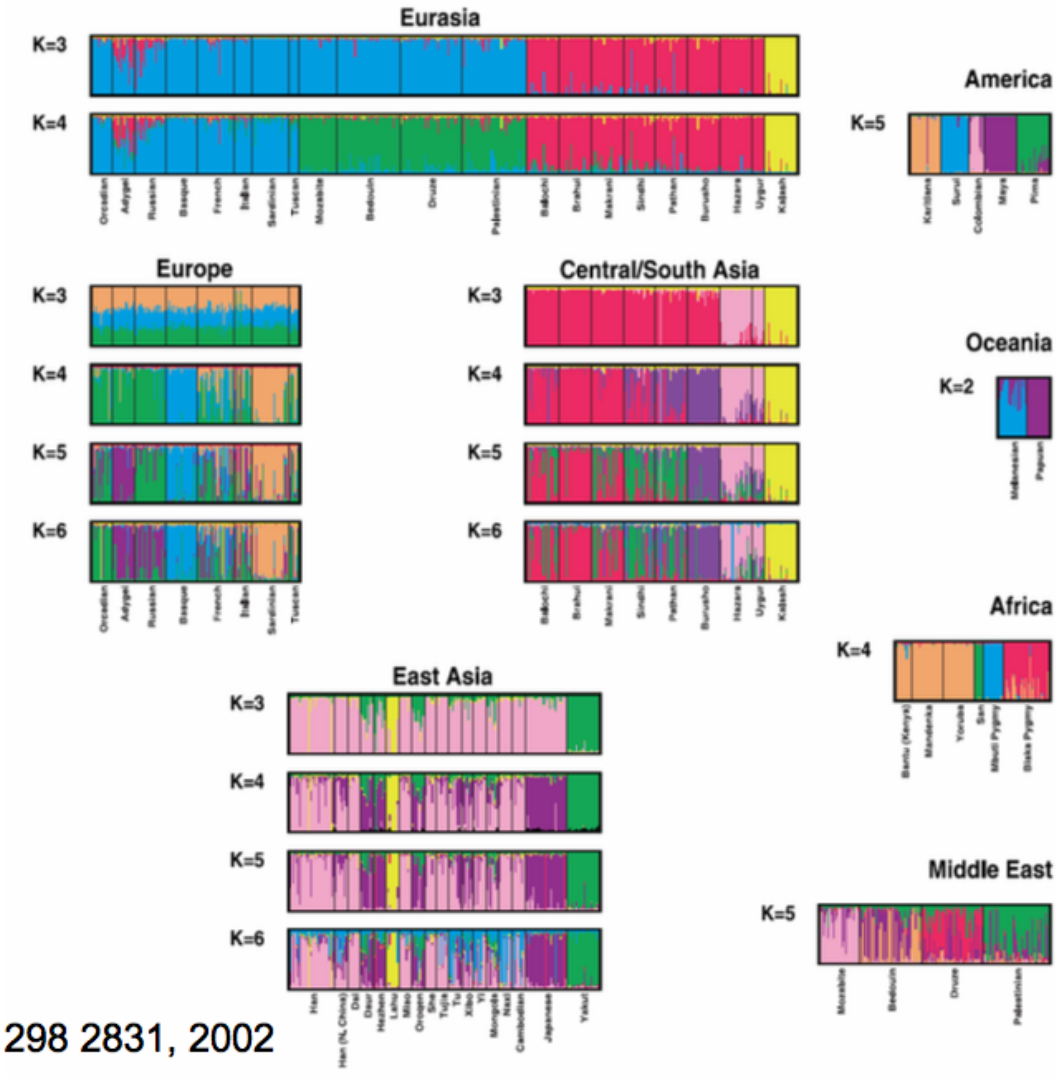
Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.
- Let $F^j = (f^{j,1}, f^{j,2}, ..., f^{j,K})$ be the allele frequencies of allele a in the K "source populations"
- Let $Q^i = (q^{i,1}, q^{i,2}, ..., q^{i,K})$ be the K admixture proportions of individual i.
- Then, for one of the allele copies in locus j in idividual i:
  $P[b|Q, F] = q^{i,1} f^{j,1} + q^{i,2} f^{j,2} + ... + q^{i,K} f^{j,K} = h^{i,j}$
- Assuming Hardy-Weinberg equilibrium:
  $P[G_{ij} = 2|Q^i, F^j] = (h^{ij})^2$
  $P[G_{ij} = 1|Q^i, F^j] = 2(h^{ij})(1 - h^{ij})$
  $P[G_{ij} = 0|Q^i, F^j] = (1 - h^{ij})^2$

Likelihood function: N individuals, M loci

- Assuming loci are independent and individuals are unrelated...
- $P[G|Q,F] = \prod_i^N \prod_j^M P[G_{ij}|Q^i, Fj]$
- Q is a matrix of admixture proportions for each of the N individuals: $Q = (Q^1, Q^2, ..., Q^N)$
- F is a matrix of ancestral allele frequencies for each of the M loci: $F = (F^1, F^2, ..., F^M)$
- Structure-lke methods try to find the **parameters Q and F that maximize** $P[G|Q,F]$

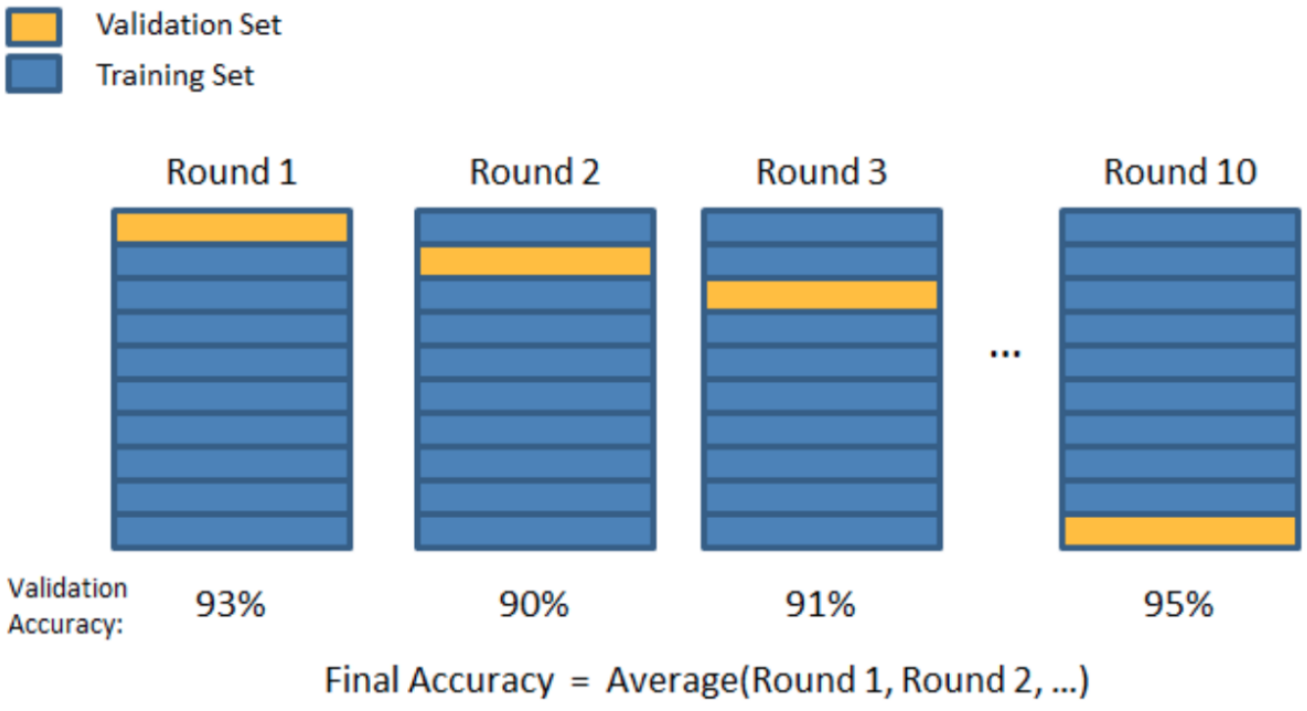
Structure model applied to human populations



Science 298 2831, 2002

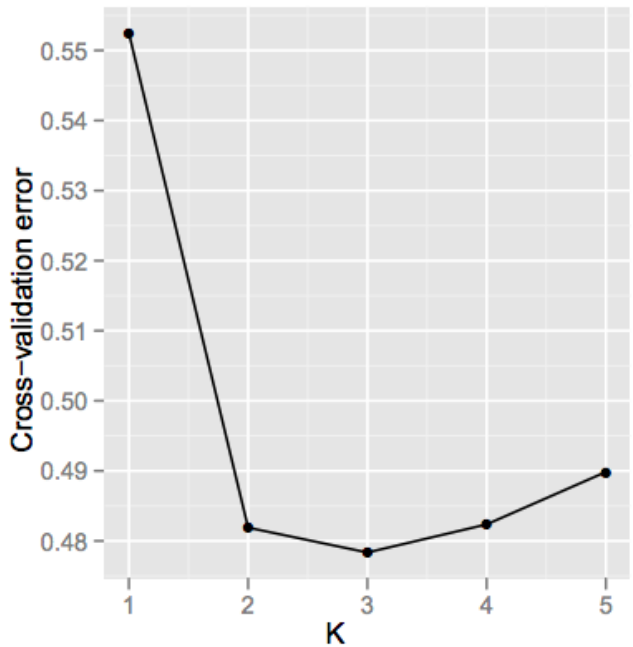
Choosing K

- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



Choosing K

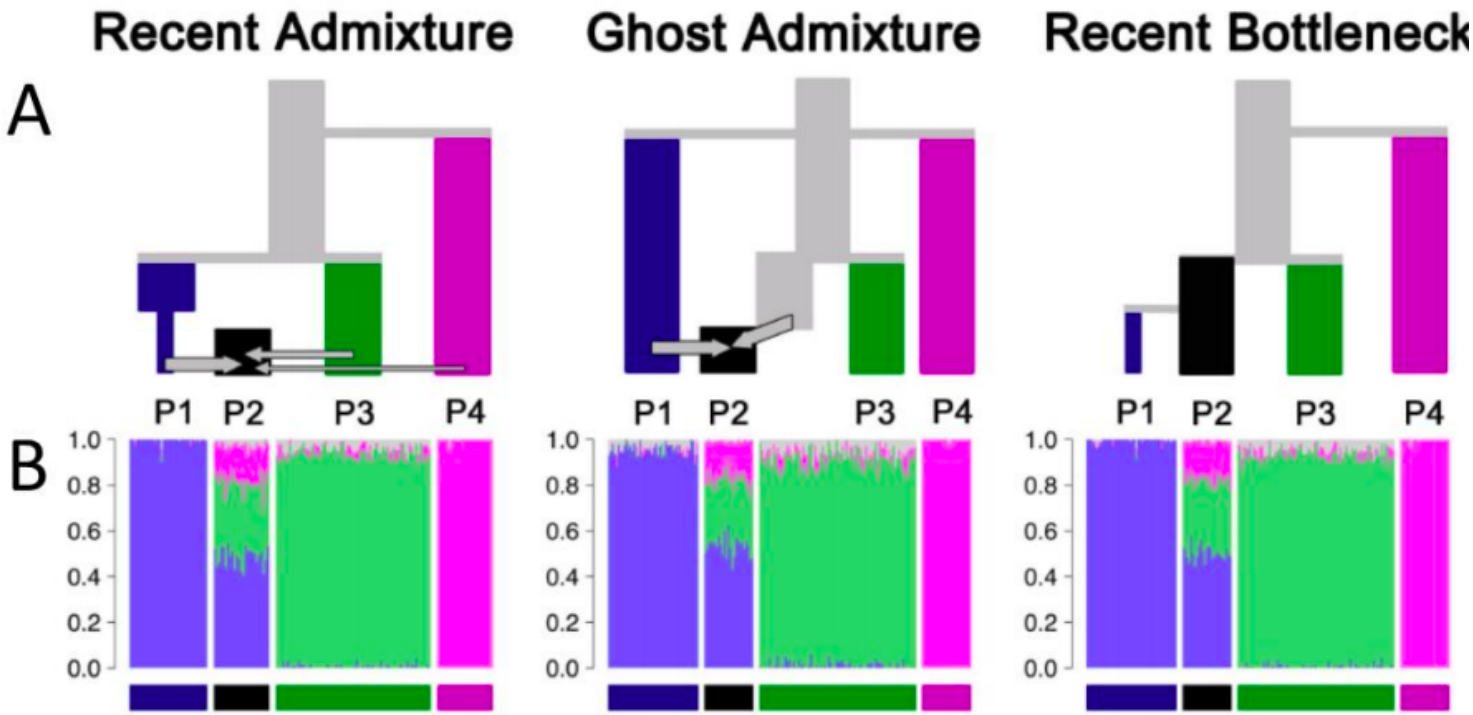
- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



[0] Alexander et al. 2011

Over-interpreting Structure results

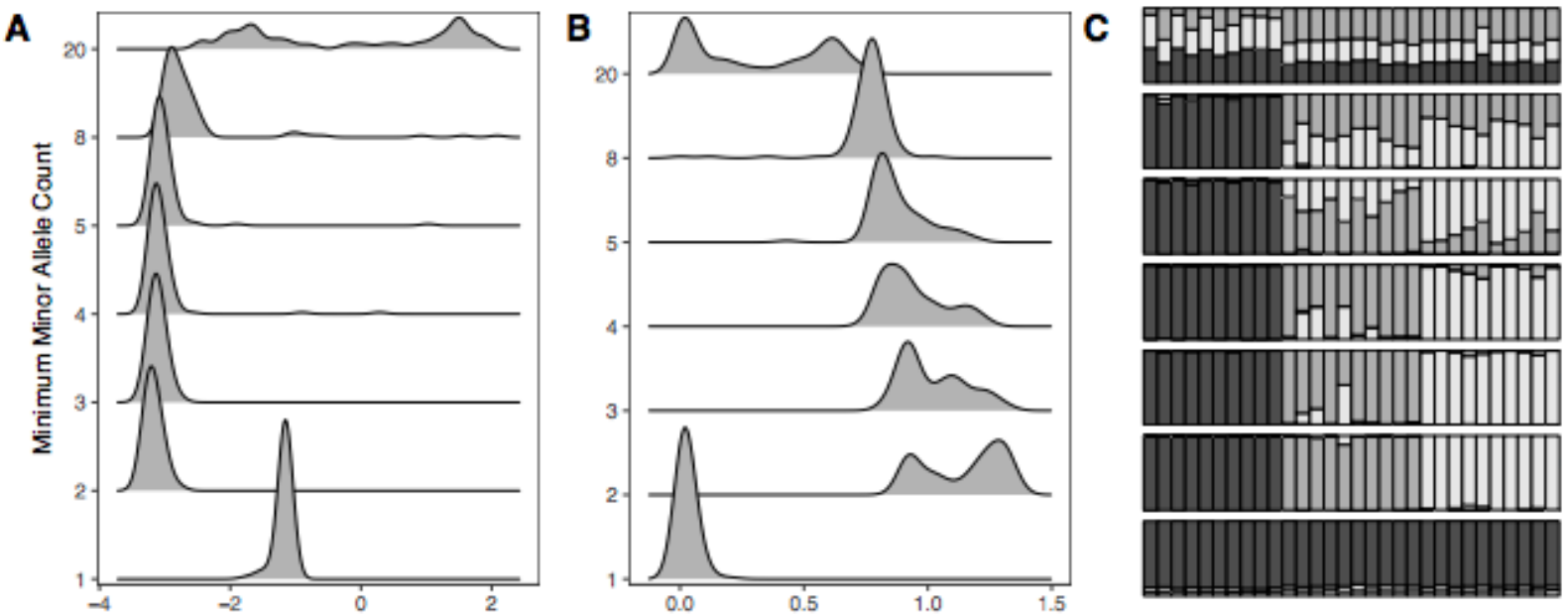
- Structure does not necessarily pick up admixture events![1]
- "Source populations" need not be real populations that ever existed!
- A population that is highly drifted will be assigned its own cluster at high enough K



[1] Falush et al. 2016

Minor allele frequency cutoff

- MAF cutoffs can strongly inference under the Structure model[2]

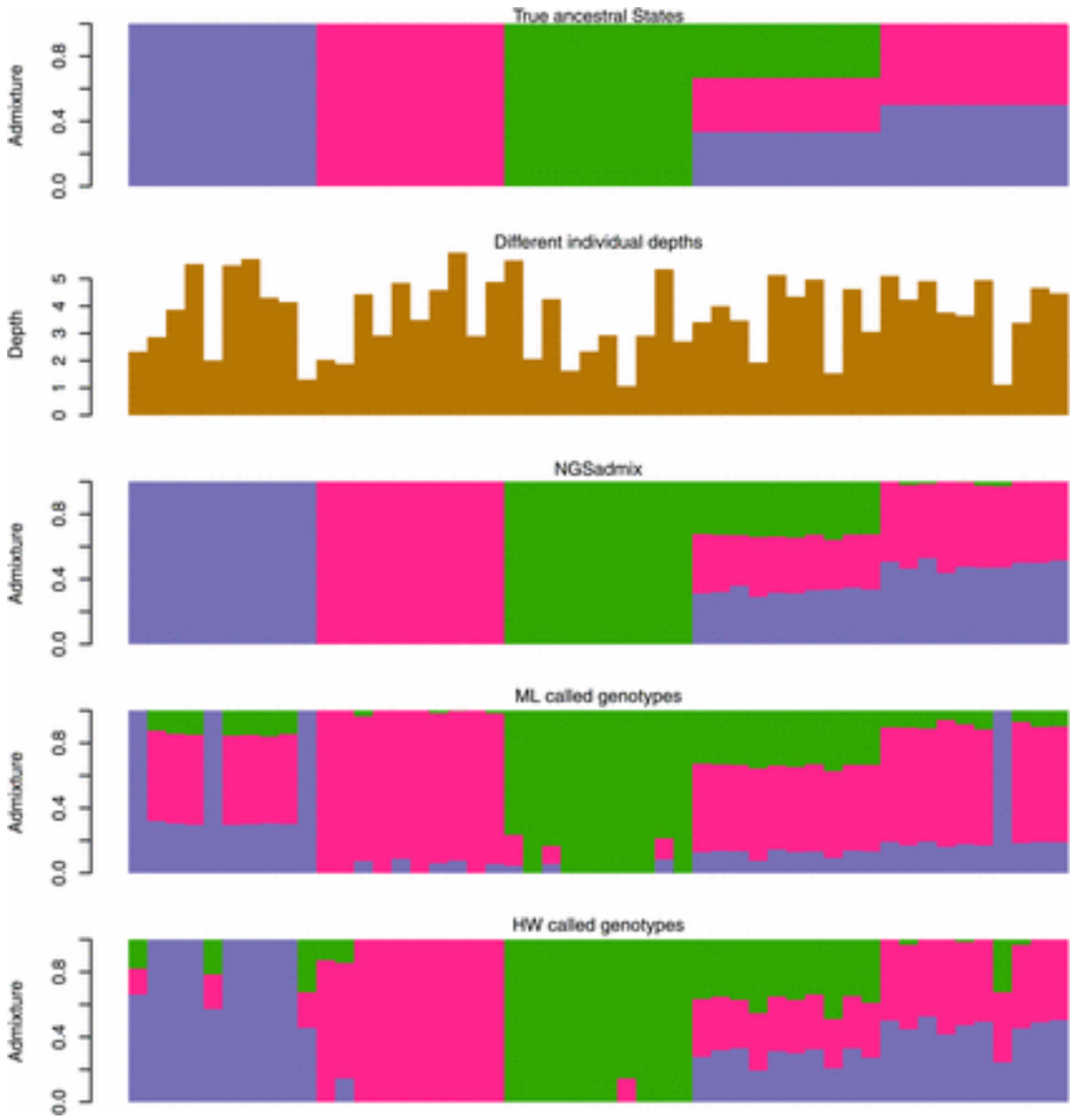


[2]Linck and Battey 2017

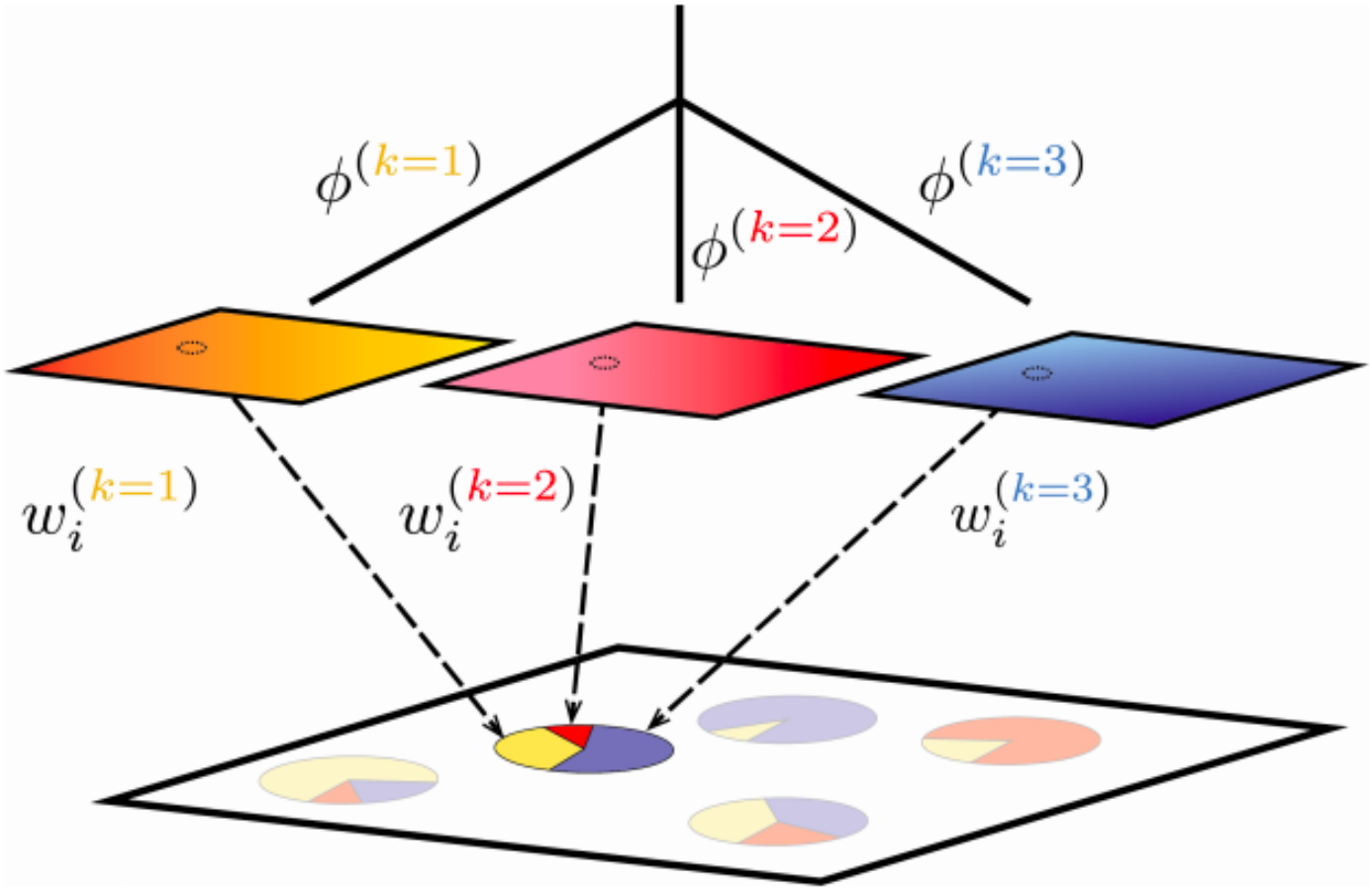
Variations on a theme...

- Structure (Pritchard et al. 2000): original model; uses Bayesian priors to obtain posterior estimates of Q and F
- Admixture (Alexander et al. 2011): faster than Structure; uses a maximum likelihood model rather than a Bayesian model; uses cross-validation to choose K
- fastStructure (Raj et al. 2014): faster than Structure; uses variational inference to choose K; can detect weak structure
- ngsAdmix (Skotte et al. 2013): can work with genotype likelihoods; better for low coverage data
- Ohana (Cheng et al. 2016): uses Gaussian approximation to model drift in each ancestry component; can detect selection by testing for local deviations from genome-wide model

ngsAdmix (Skotte et al. 2013)



consStruct: structure + isolation-by-distance (Bradburd et al. 2017)

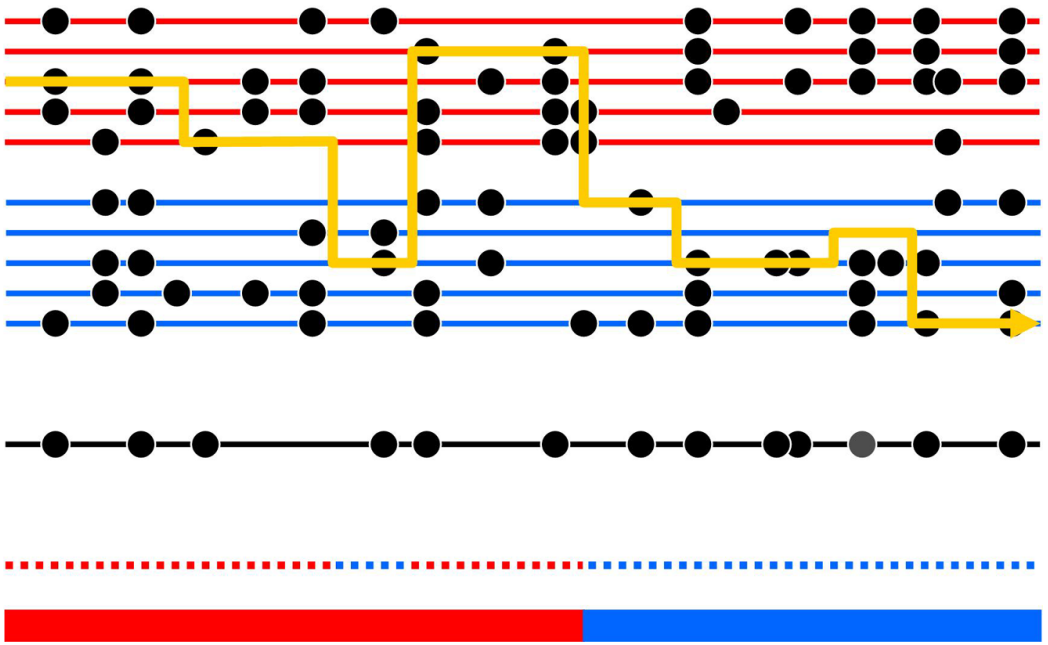


SNP vs. Haplotype information

- All these methods ingnore the spatial distribution of SNPs along the genome
- They require LD pruning: a lot of SNPs sit in the same haplotype and have redundant information
- Advantage: can model each SNP independently (simple model)
- Disadvantage: we ignore haplotype information

Chromosome painting

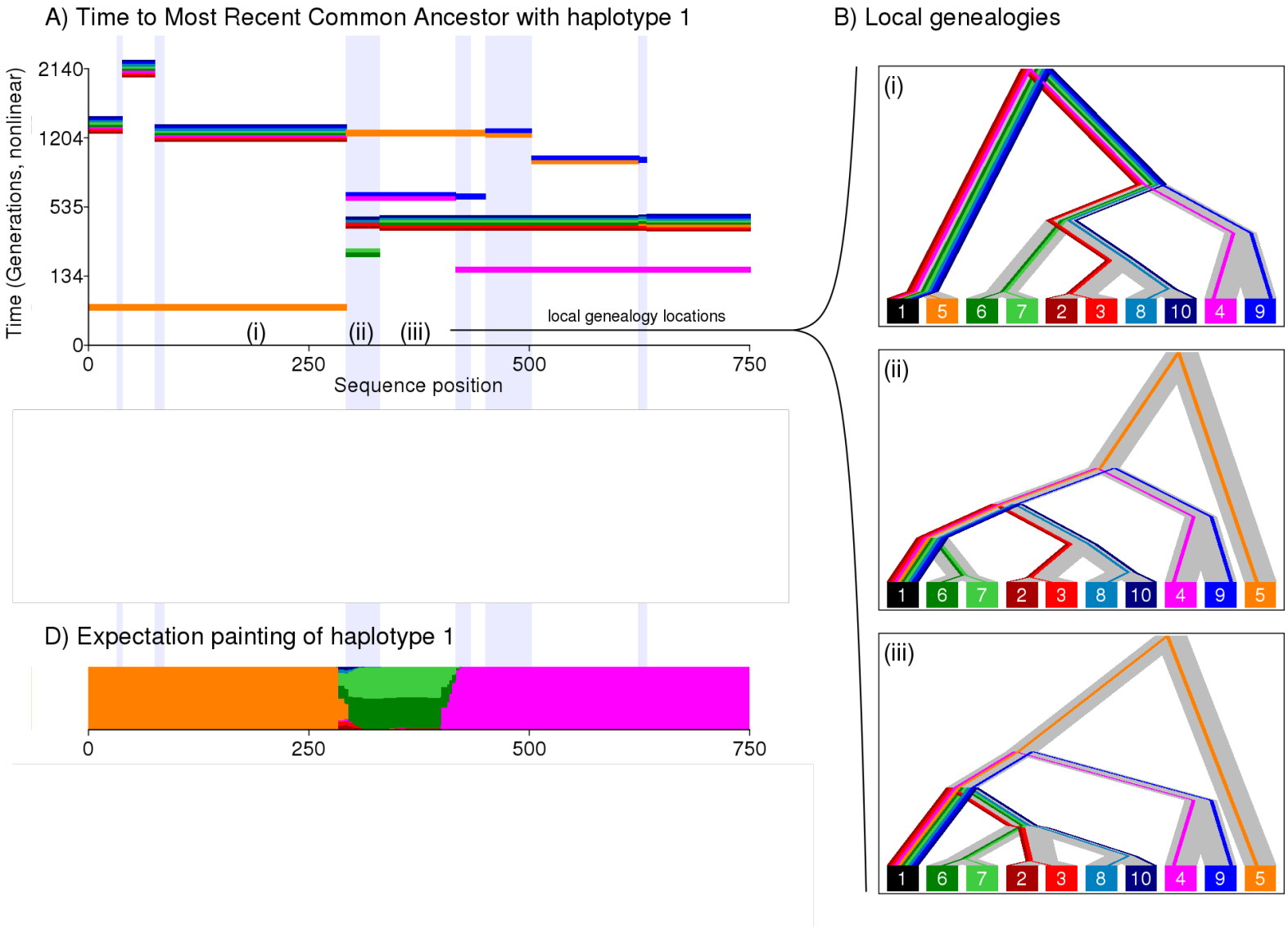
- Painting a chromosome by "copying" segments from other individuals[3]
- Can be done very fast, thanks to Li and Stephens algorithm [4]



[4] Price et al. 2009

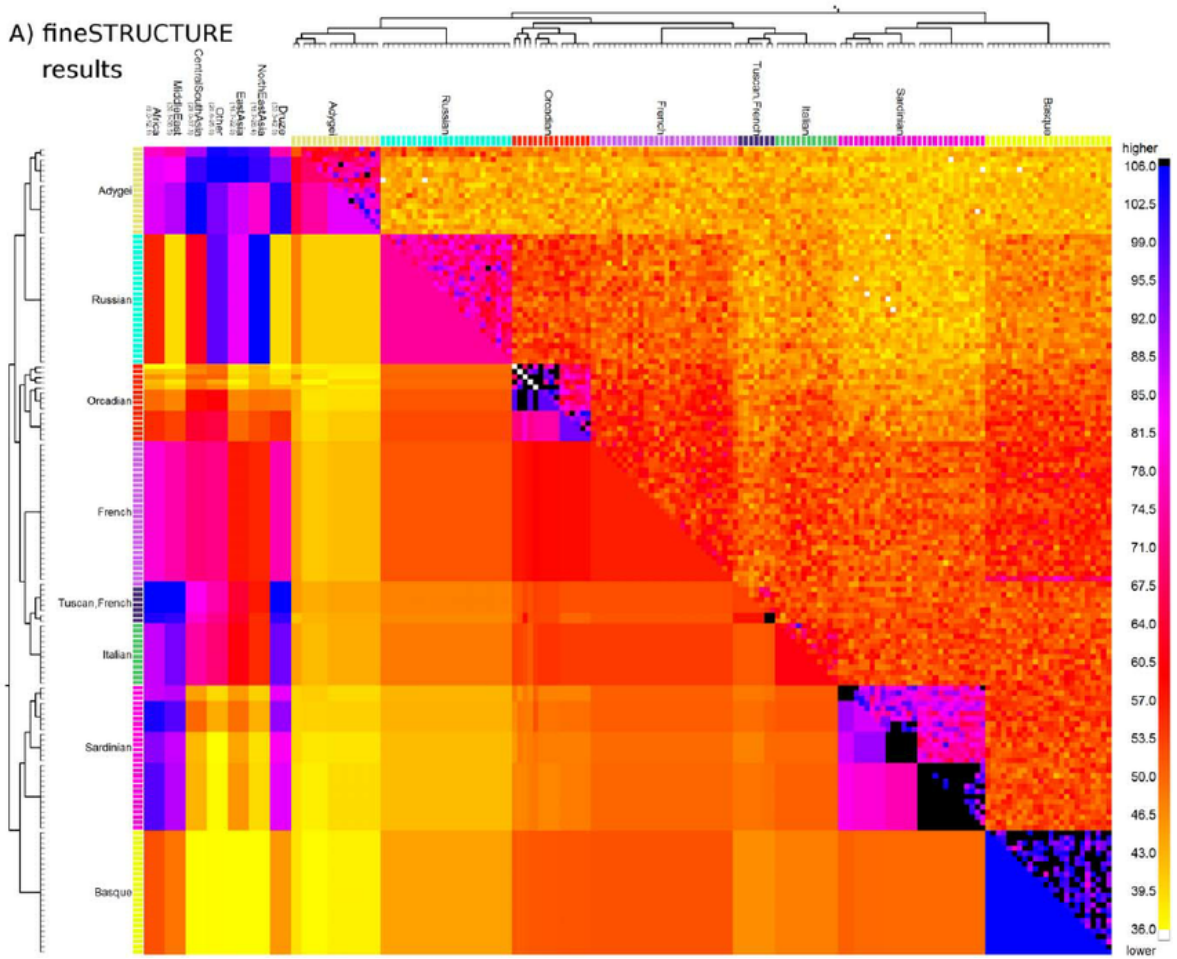
[4] Li and Stephens 2003

Chromosome painting



[4] Hellenthal et al. 2014

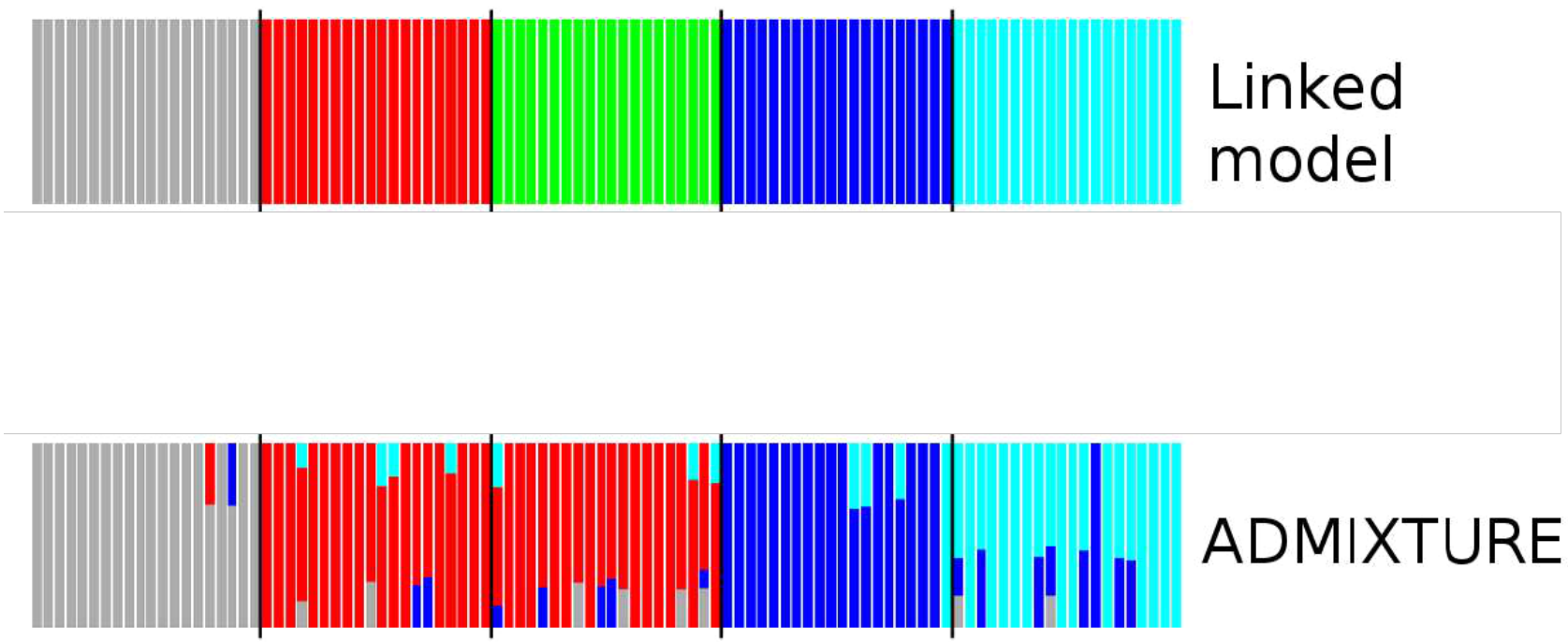
Coancestry matrix based on painting patterns



[4] Hellenthal et al. 2014

Chromosome "palettes"

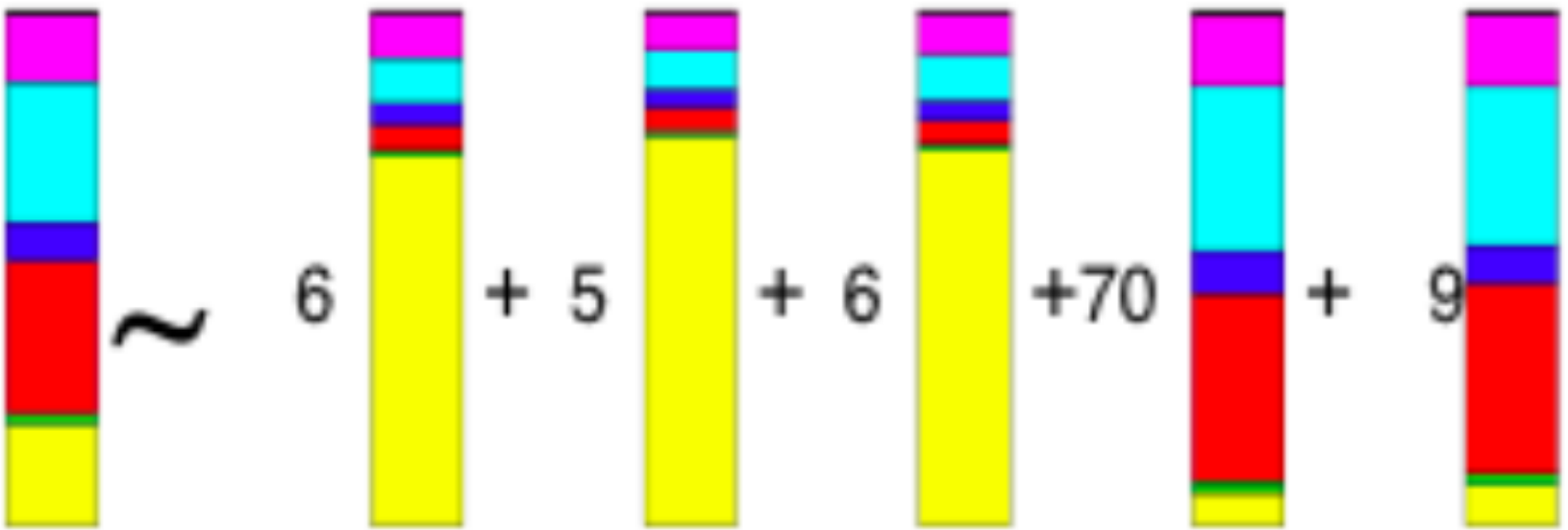
- We can create a chromosome palette by aggregating the ancestry from each of the chromosome segments
- This is comparable (but not the same!) as a Structure / Admixture barplot



[4] Hellenthal et al. 2014

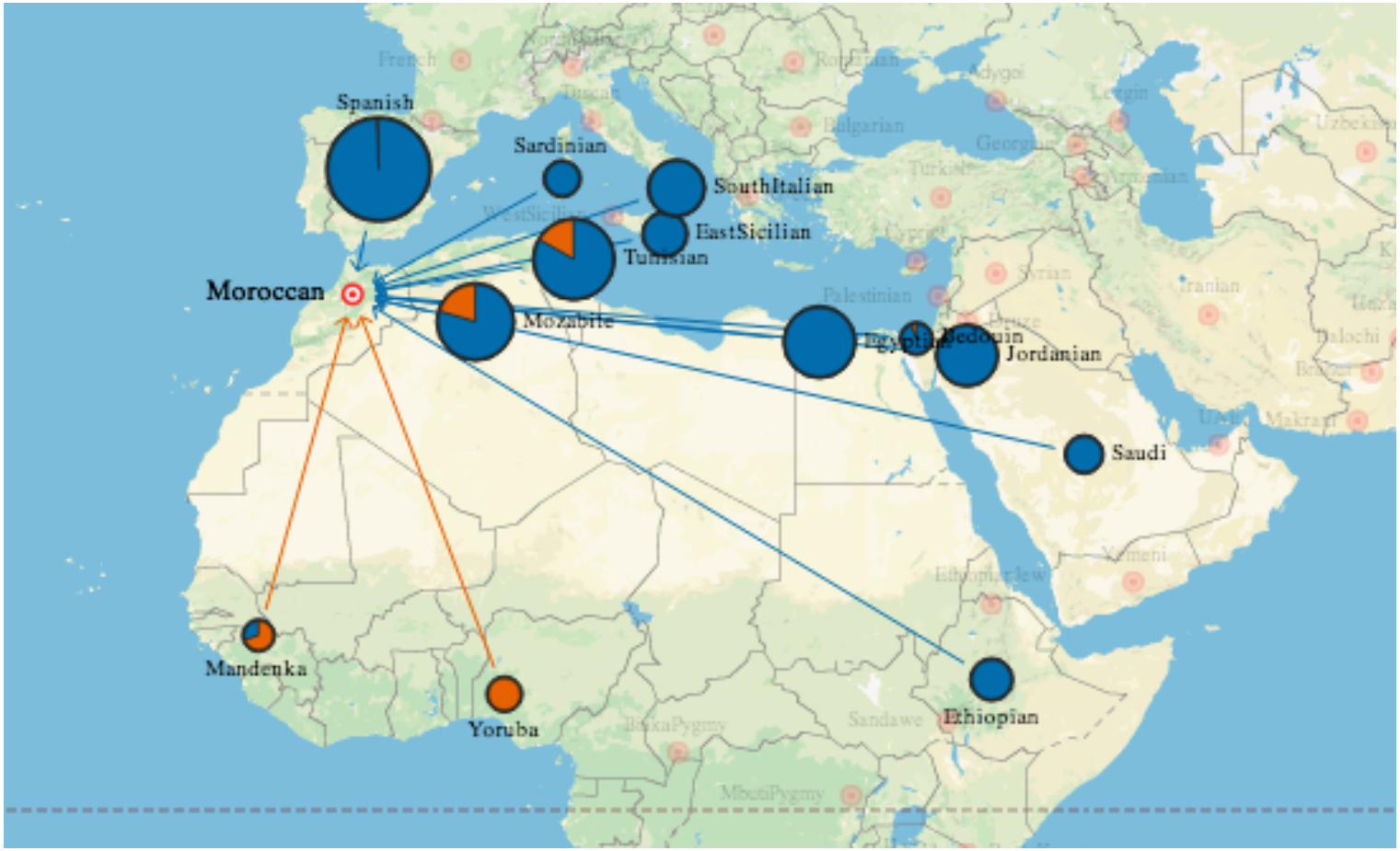
Chromosome “palettes”

- We can model an individual as a linear mixture of the palettes from other individuals



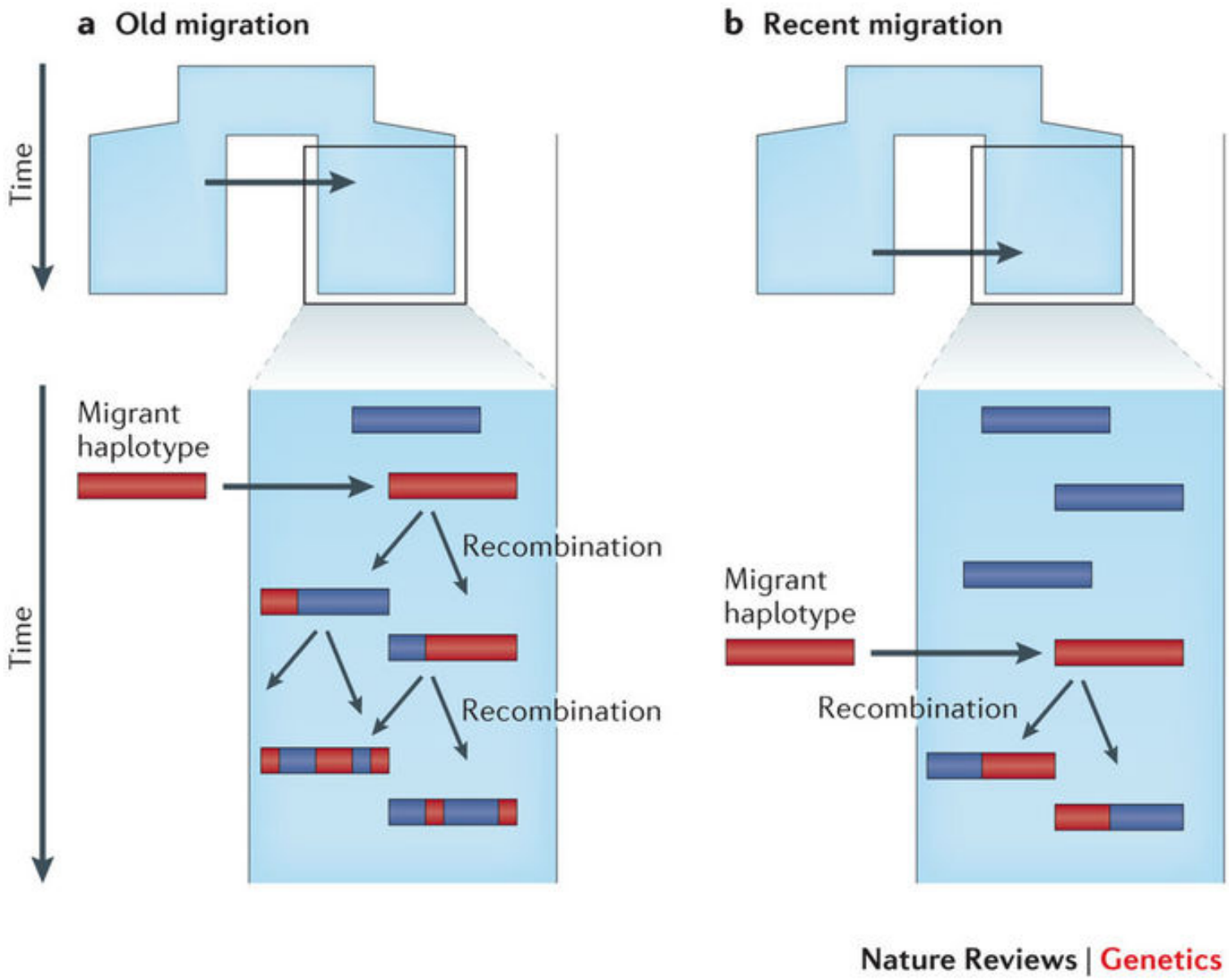
[4] Hellenthal et al. 2014

GlobeTrotter



[4] Hellenthal et al. 2014

Admixture date inference



[4] Souza and Hey 2013

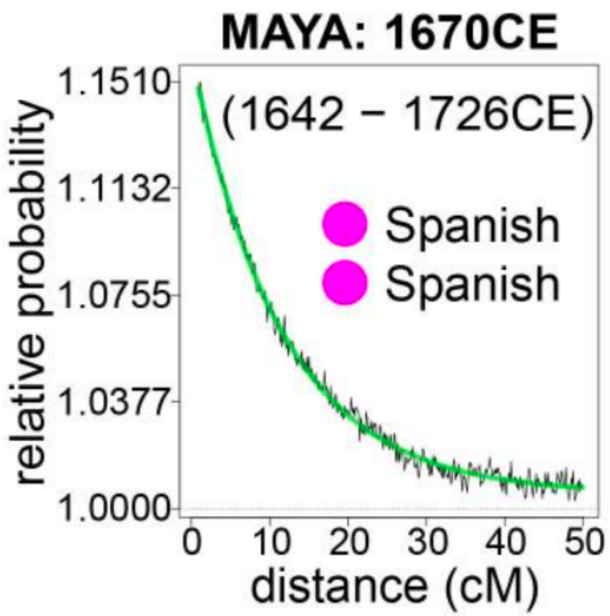
Admixture date inference

- Assume an admixture event happened at time λ between two populations A and B
- We're interested in modeling the length of tracts from A in an admixed genome
- Assume that the probability of no recombination between two points a distance g apart since admixture is $e^{-g\lambda}$ (where λ is the time since admixture scaled by the recombination rate)
- If the fraction of total ancestry from population A in the genome is α, then the probability that we'll find two loci with ancestry from A a distance g apart is:
- $p_{AA}(g) = \alpha(e^{-g\lambda} + (1 - e^{-g\lambda})\alpha) = \alpha^2 + \alpha(1 - \alpha)e^{-g\lambda}$
- This is just an exponential function of the admixture time λ! [5]

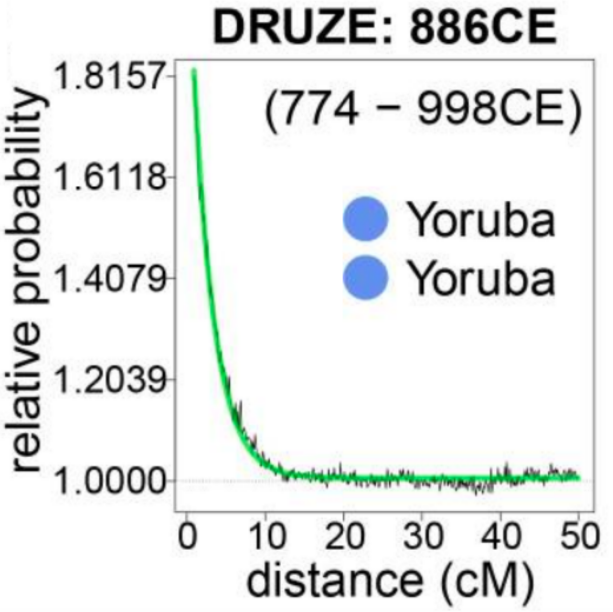
[5] Hellenthal et al. 2014

Admixture date inference

p_Spanish,Spanish as a function of distance in a Mayan genome

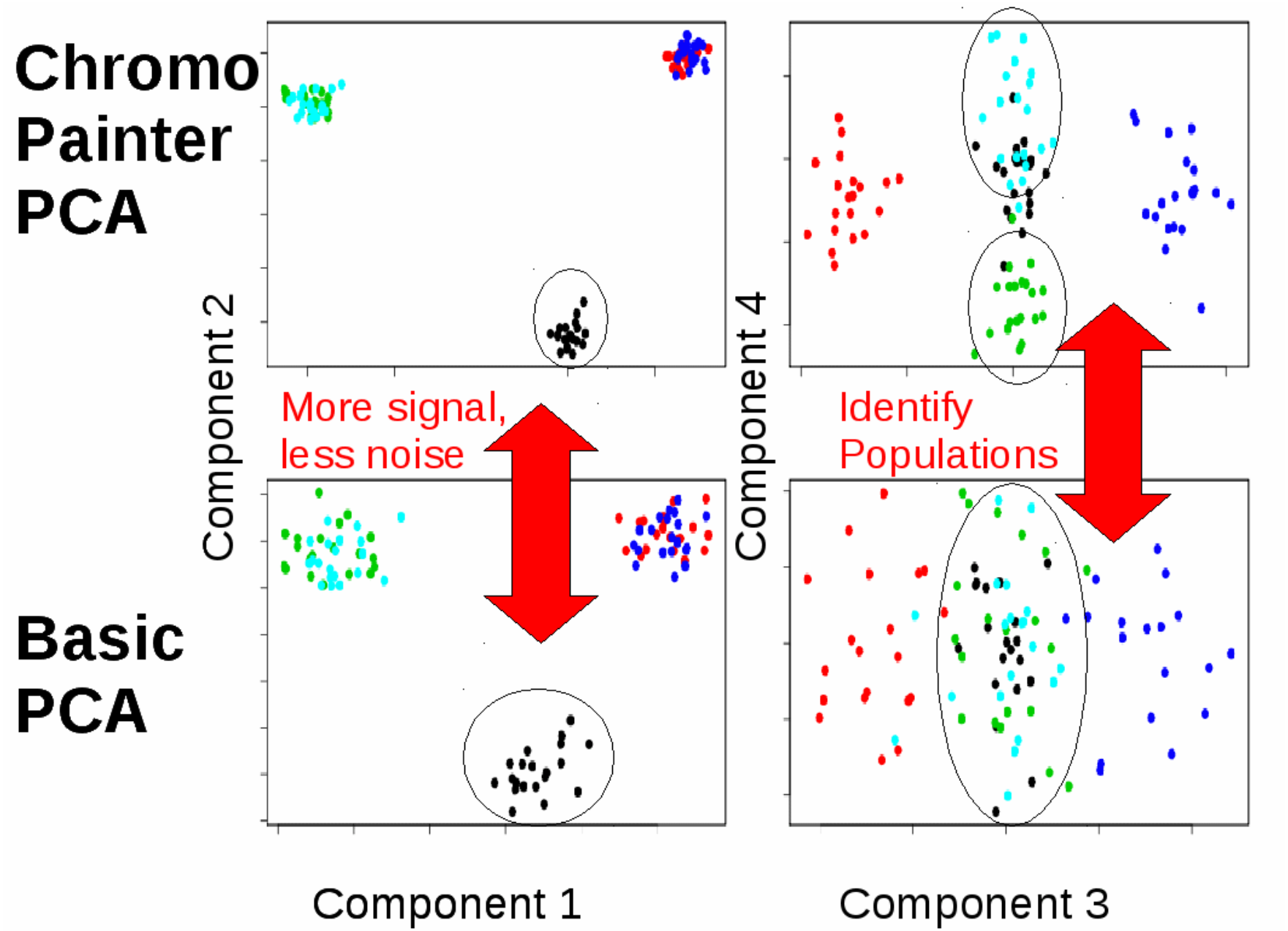


p_Yoruba,Yoruba as a function of distance in a Druze genome

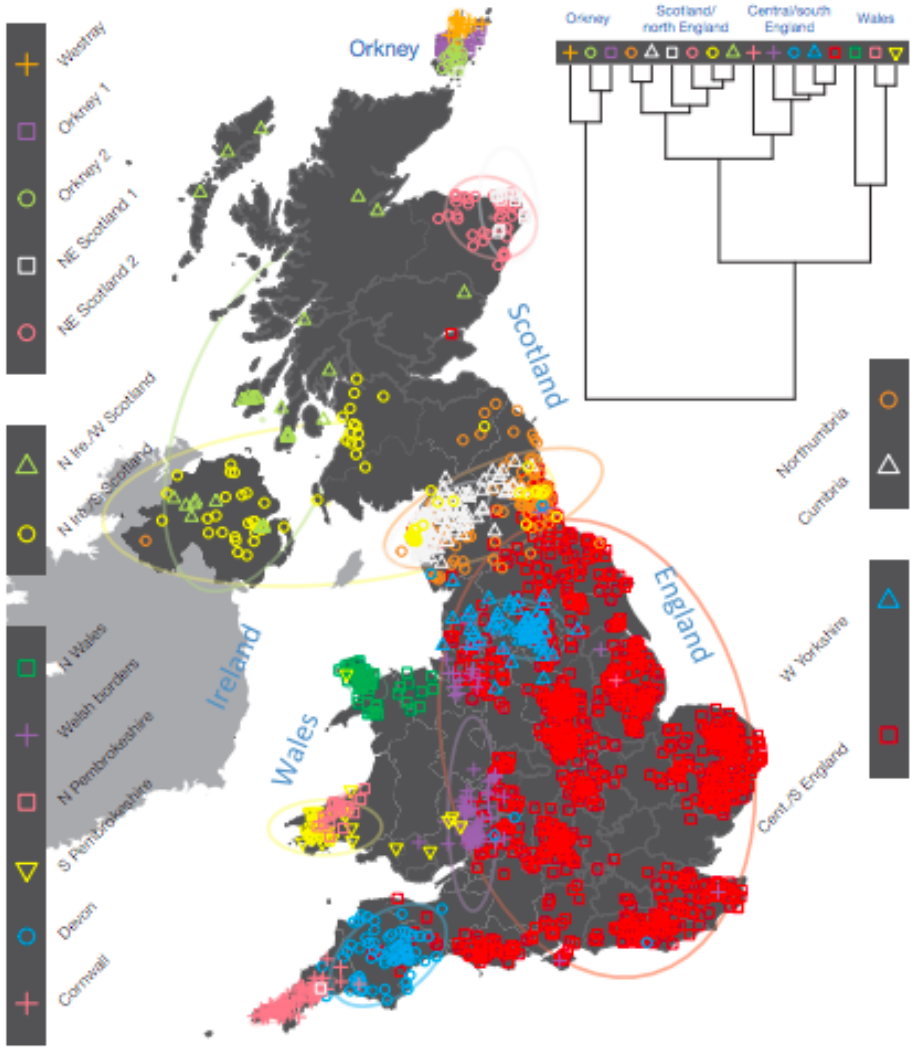


[5] Hellenthal et al. 2014

Haplotype data has more information than SNP data



Haplotype models can capture very recent history



[5] Leslie et al. 2015