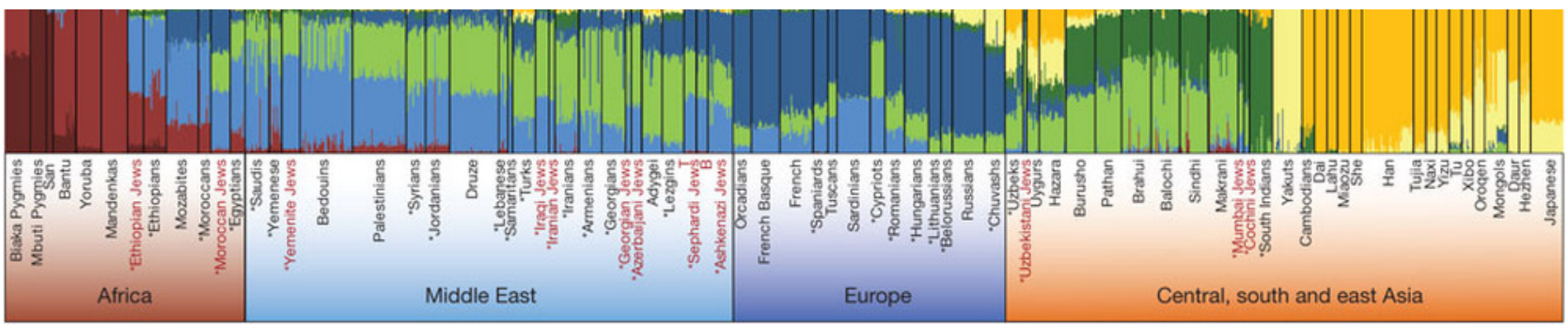
Inference of population structure and admixture

Ida Moltke, Copenhagen, Summer 2024



Outline

1. Introduction and motivation
   - What?
   - Why?
   - How?
2. ADMIXTURE and similar model-based clustering methods
   - Basic setup/problem
   - Overview of well known solutions
   - Basic idea behind inference
3. Maximum likelihood (ML) solution based on called genotypes
   - ML solution
   - Some practical problems
   - A solution to some of them - evalAdmix

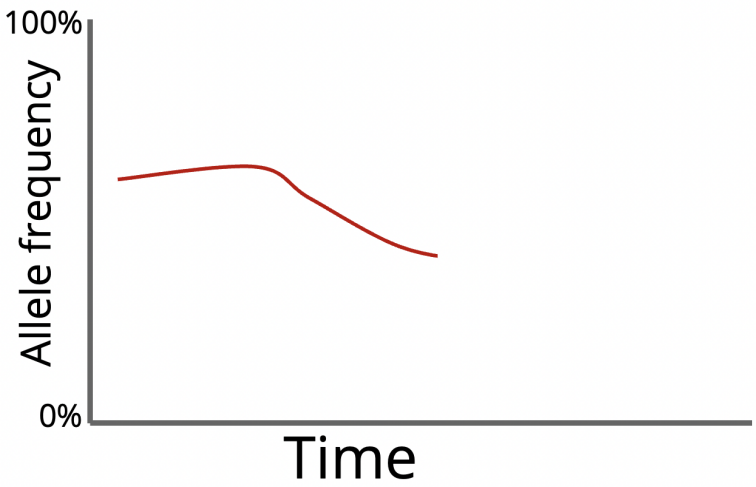
What is population structure?

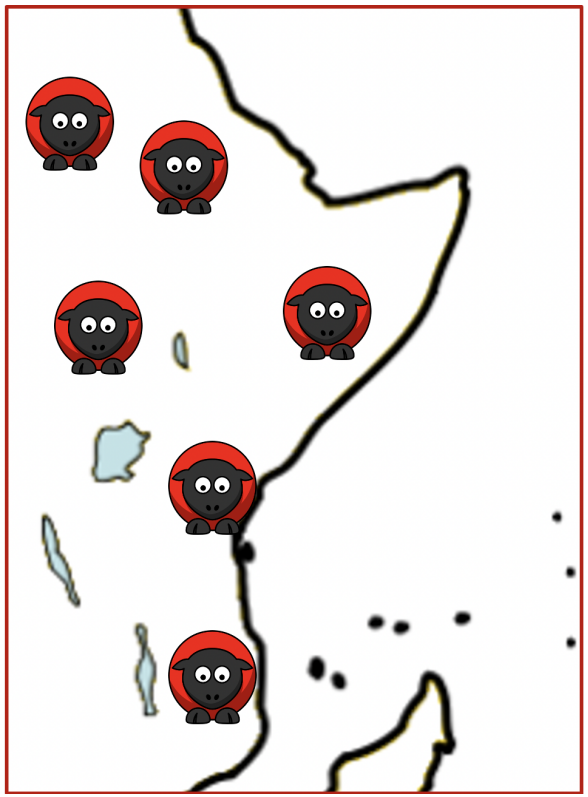
Wikipedia definition: *Population structure is the presence of a systematic difference in allele frequencies between subpopulations in a population possible due to different ancestry*

- (cur | prev) ○ 07:30, 17 Ju e 2008 Andersduck (talk | contribs . . (7,245 bytes) **(+7,245)** . . (←*Created page with 'Population stratification is the presence of difference in allele frequencies b tween subpopulations in a popu ation possible due to different ancestry...'*)

Example: A random SNP in a population w. random mating

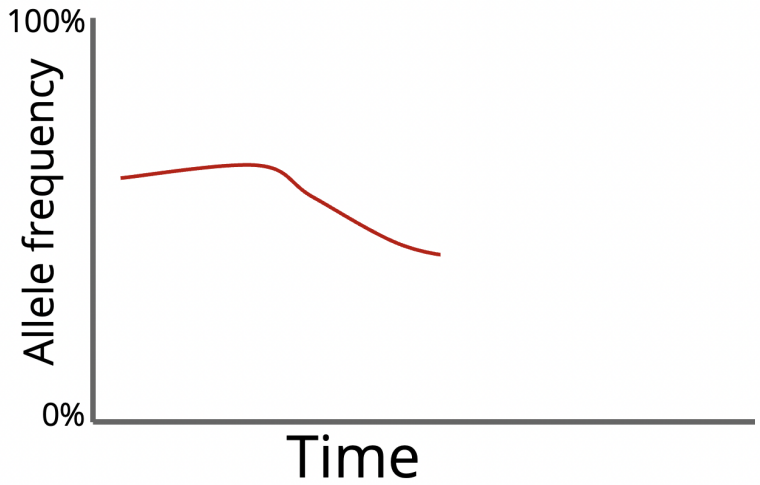
Random SNP: Frequency changes due to drift

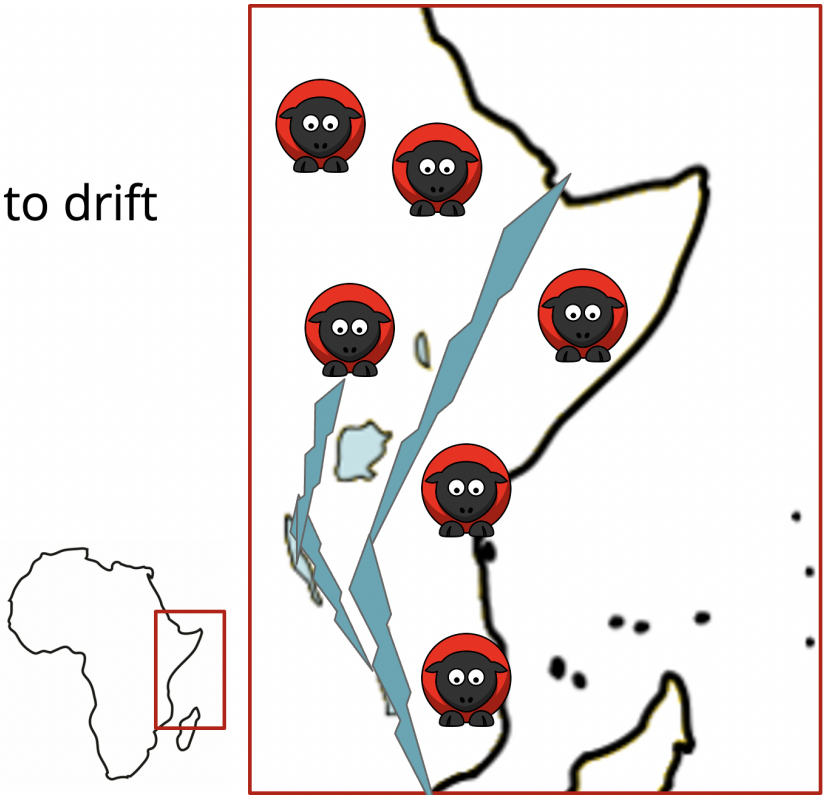




Example: A barrier to gene flow is introduced

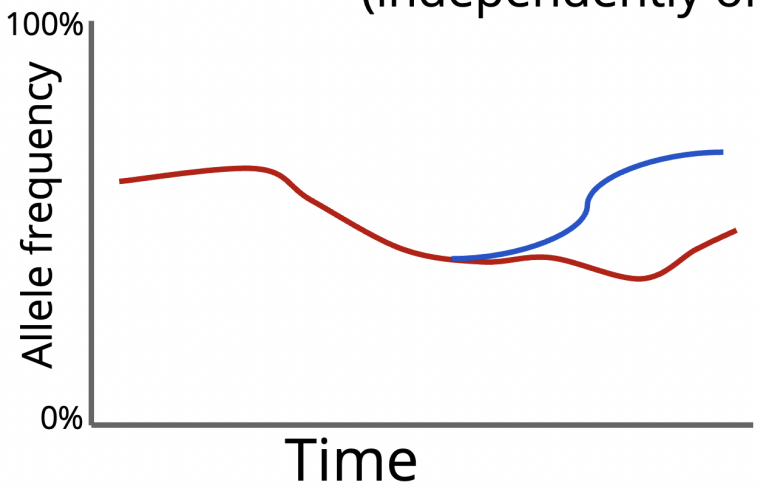
Random SNP: Frequency changes due to drift

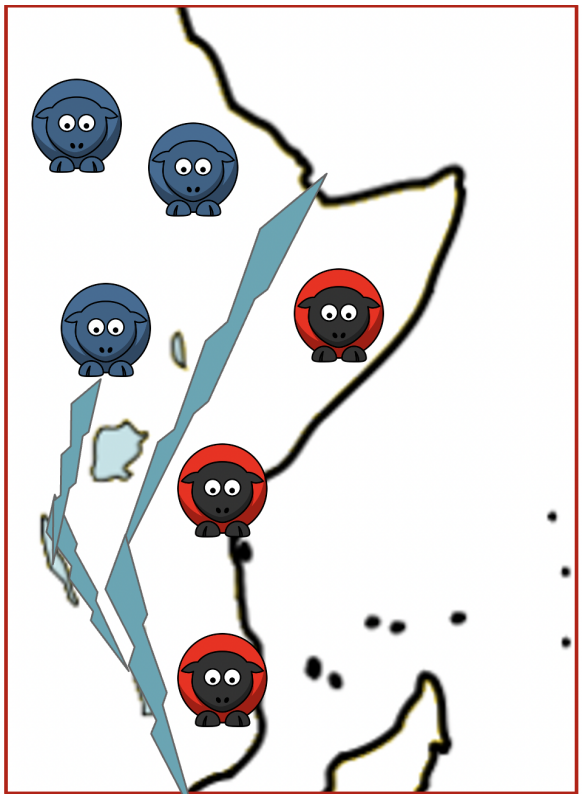




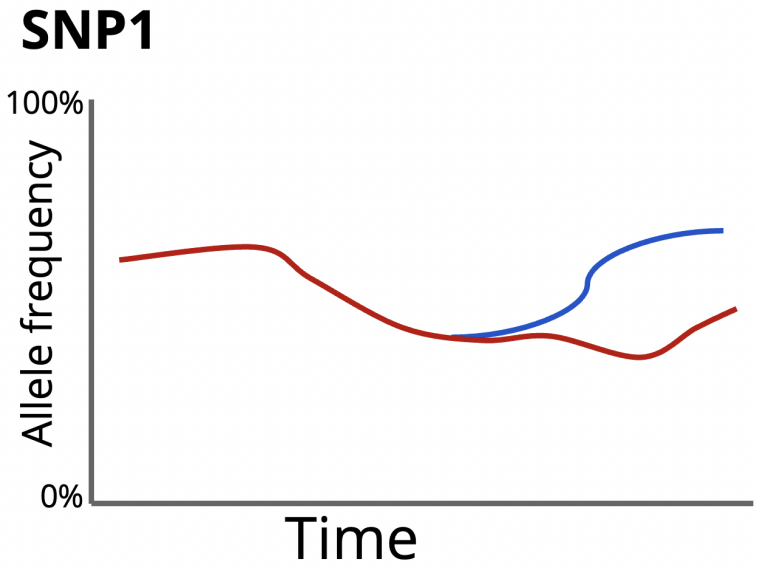
Example: The frequencies will start to differ

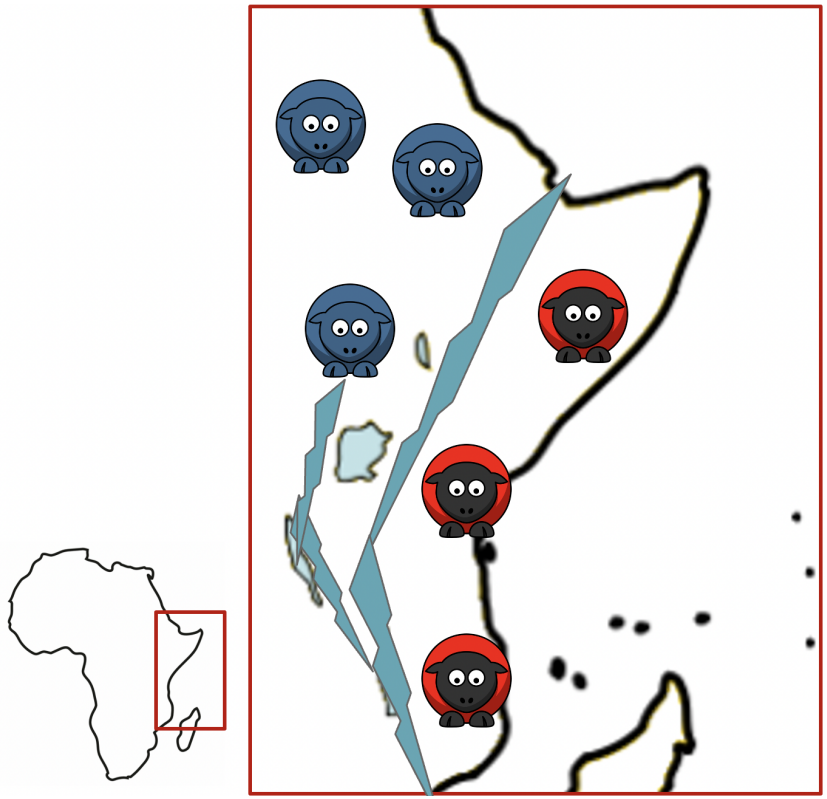
Random SNP: Frequency changes due to drift (independently on each side)



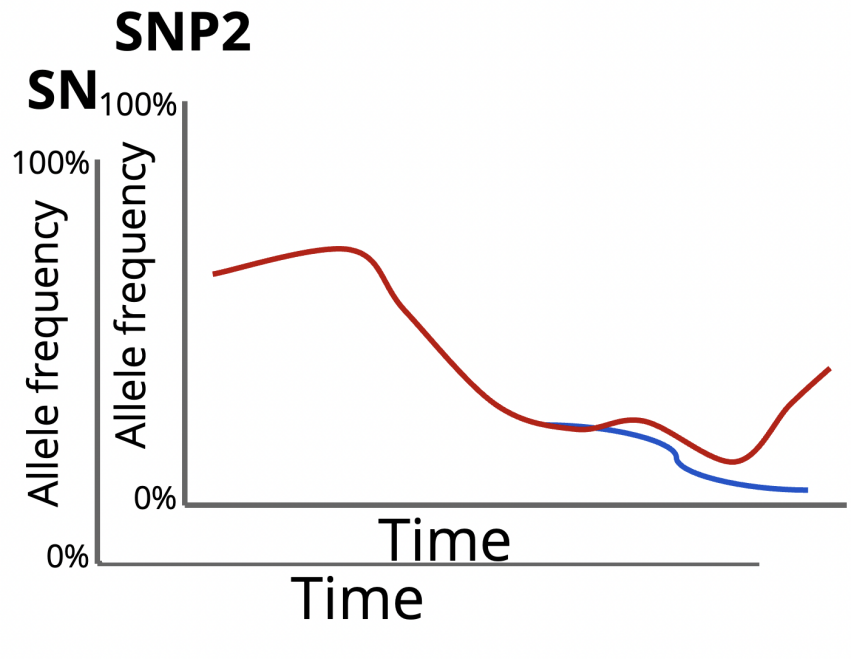


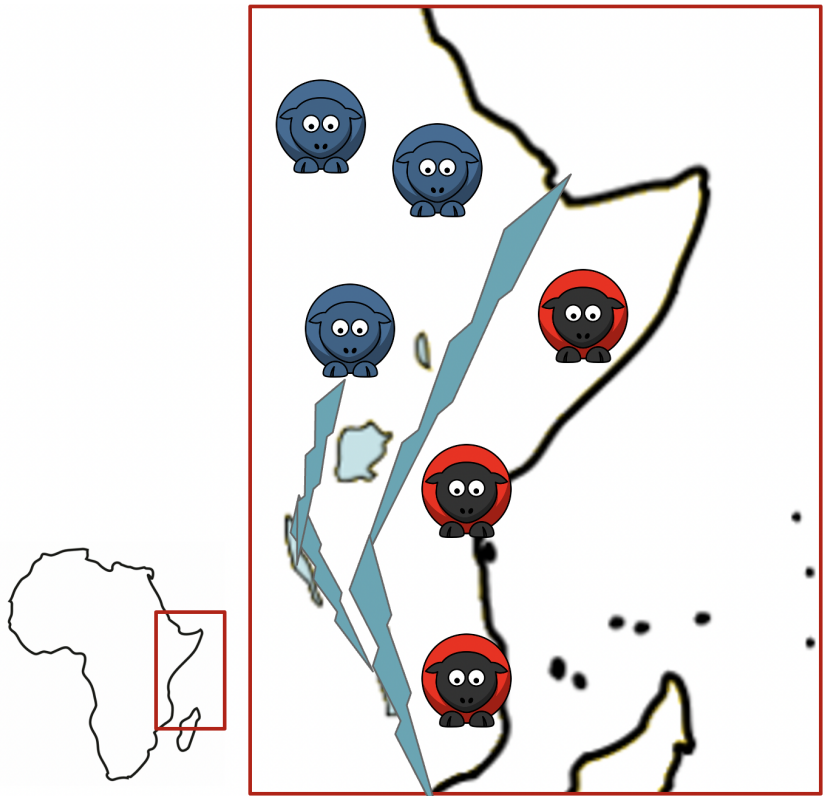
Example: this is true for all SNPs



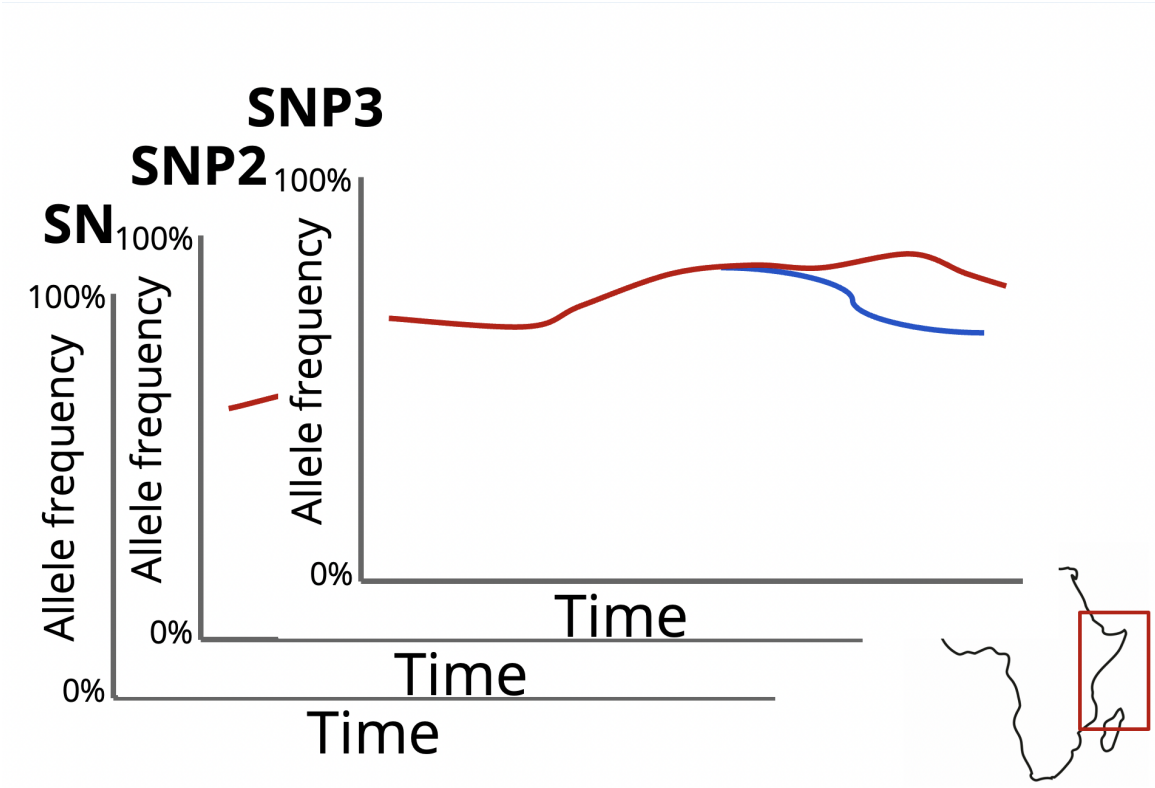


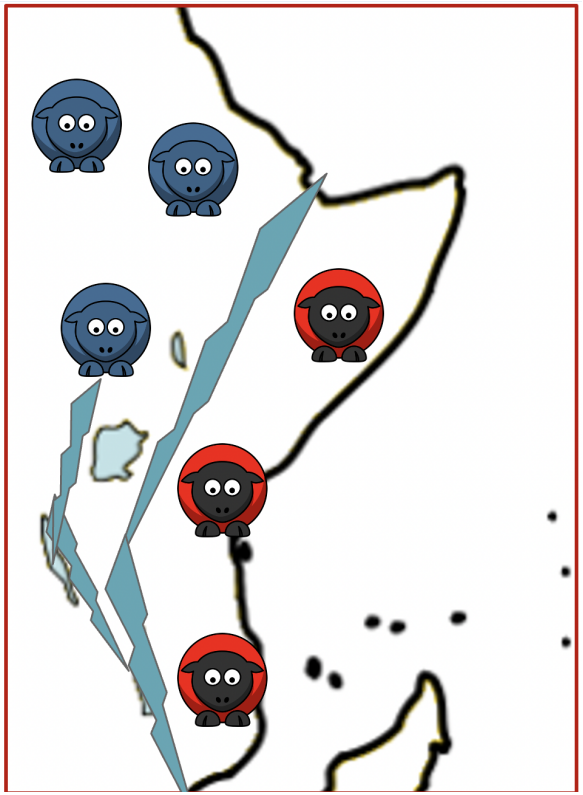
Example: this is true for all SNPs



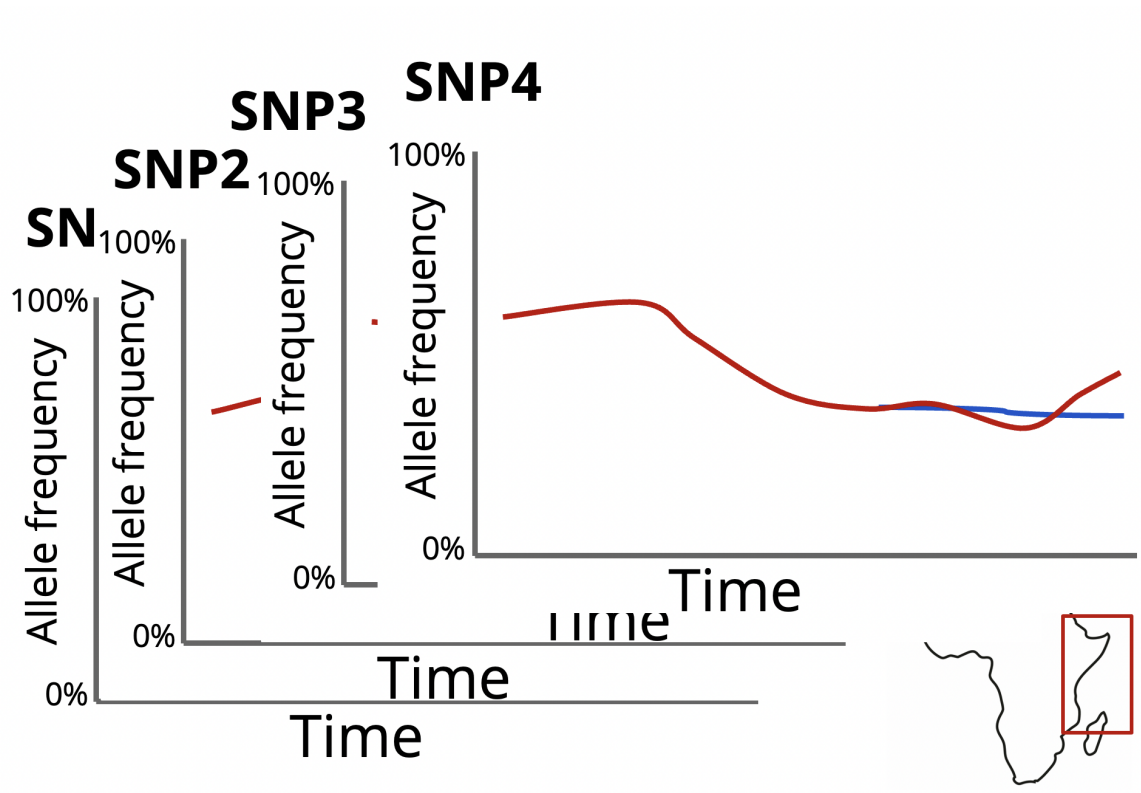


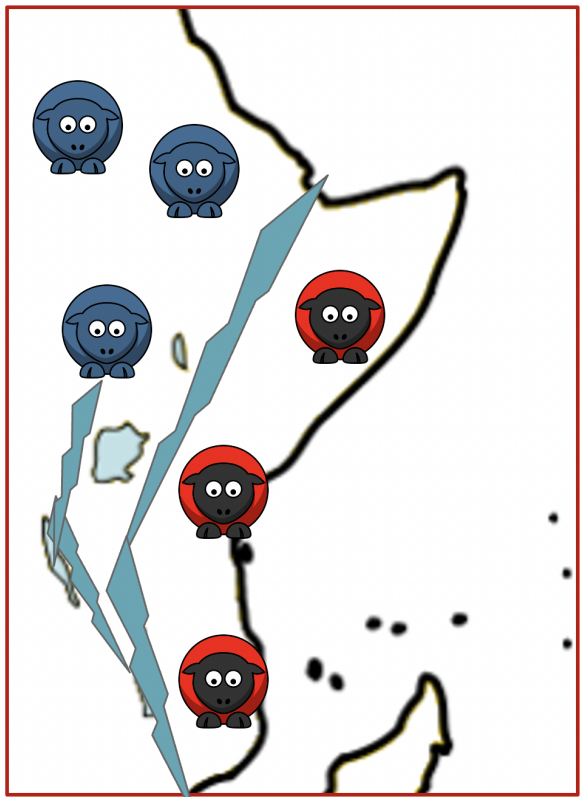
Example: this is true for all SNPs





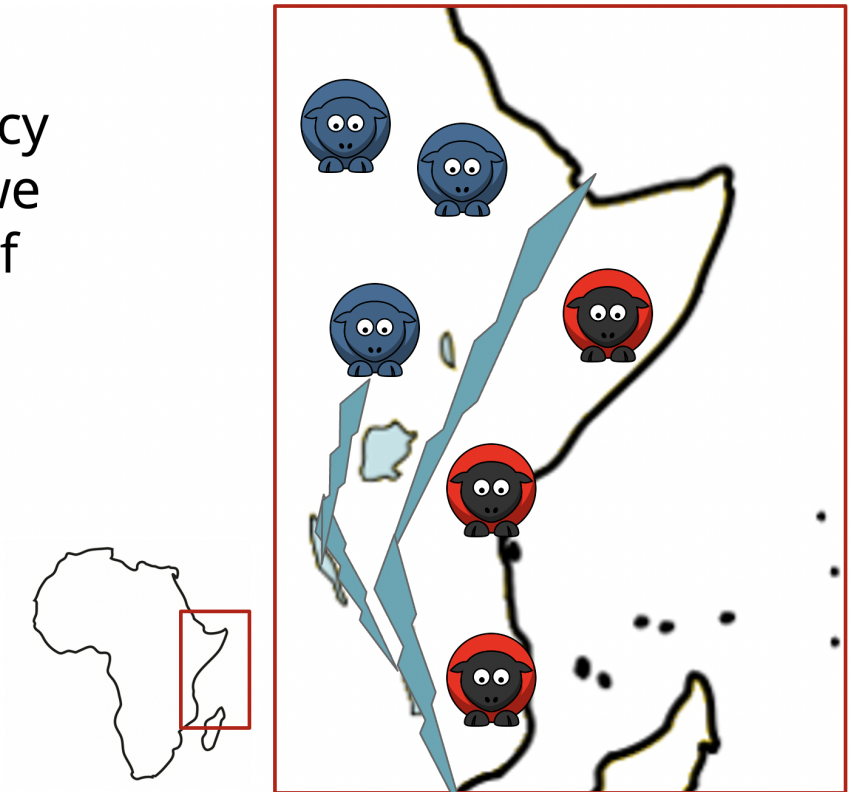
Example: this is true for all SNPs





Example: this is true for all SNPs

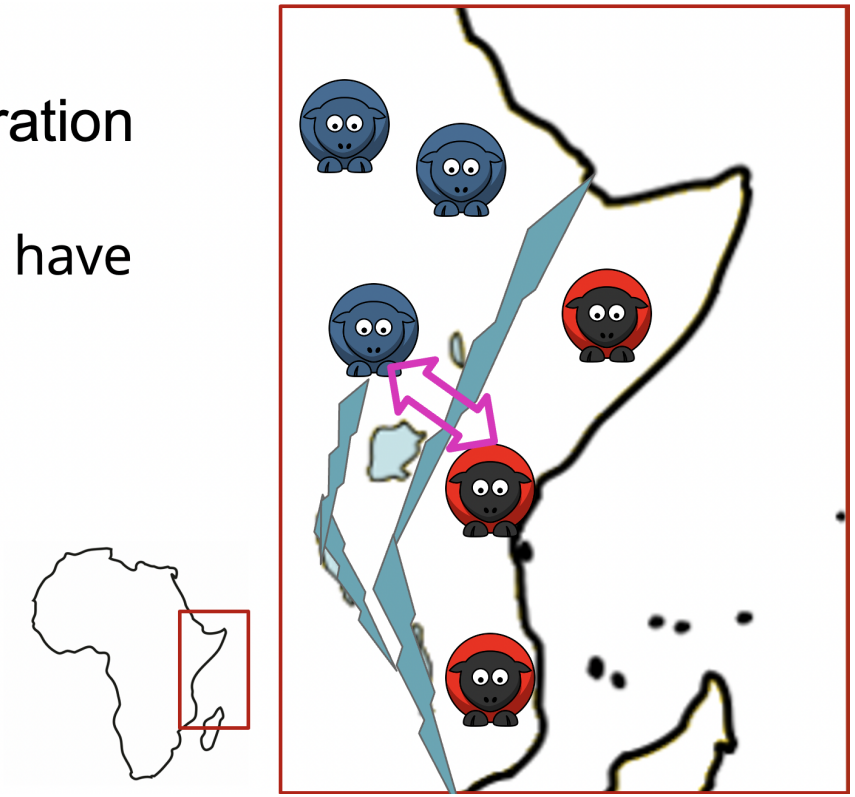
It is such systematic allele frequency differences we call structure and we typically talk about the existence of difference (sub)populations



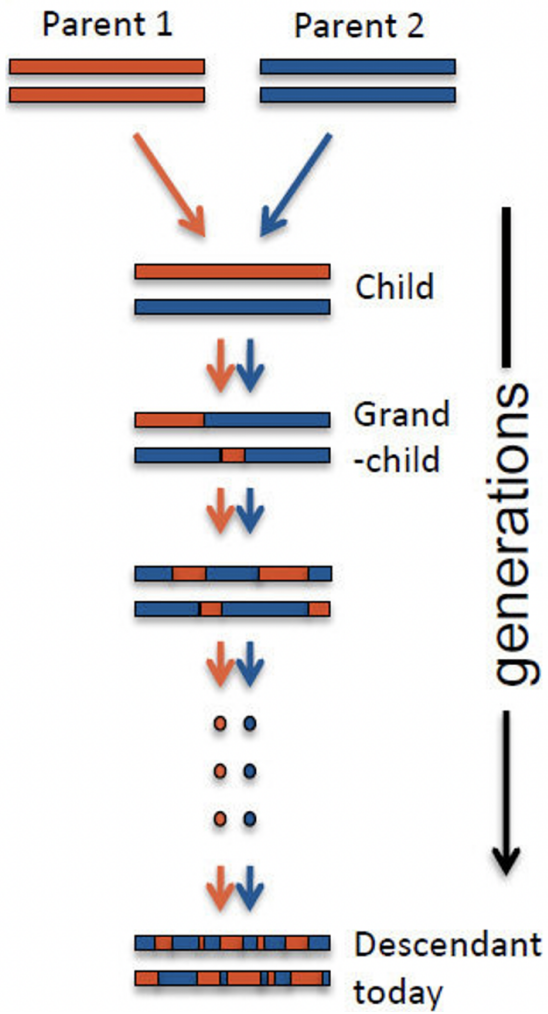
What is admixture?

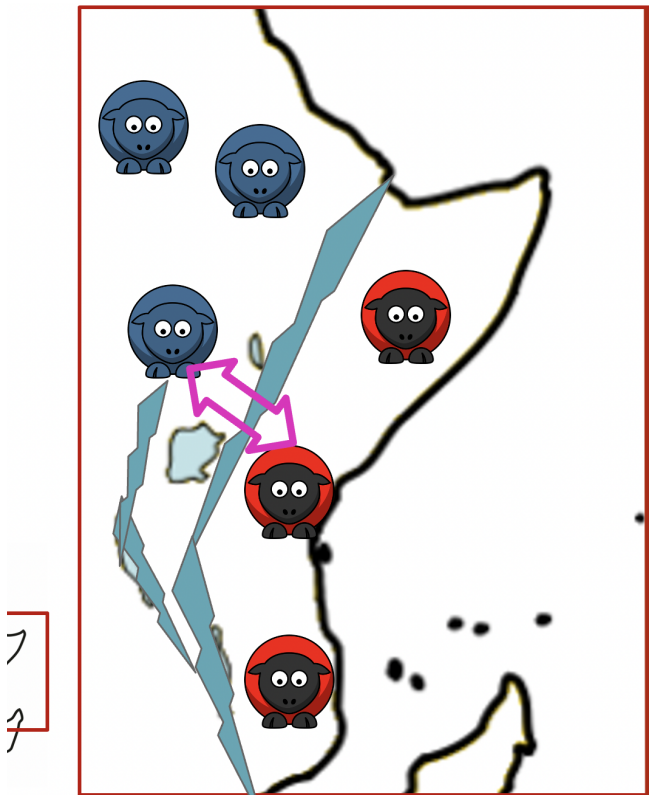
Admixture is geneflow after separation

We call individuals admixed if they have ancestry from several populations

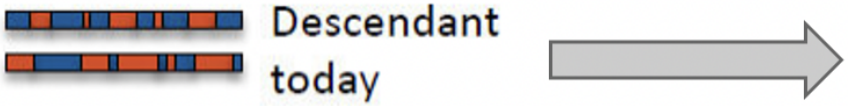


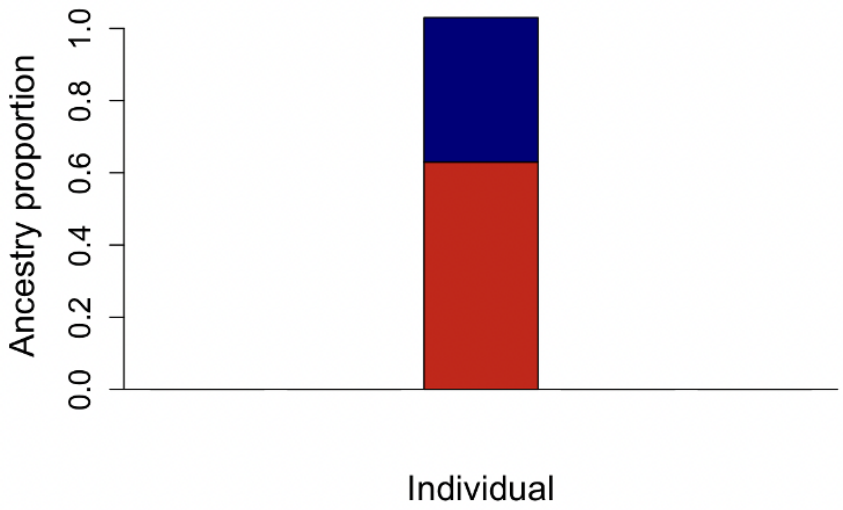
What happens in the genome?





Often used summary: ancestry/admixture proportions





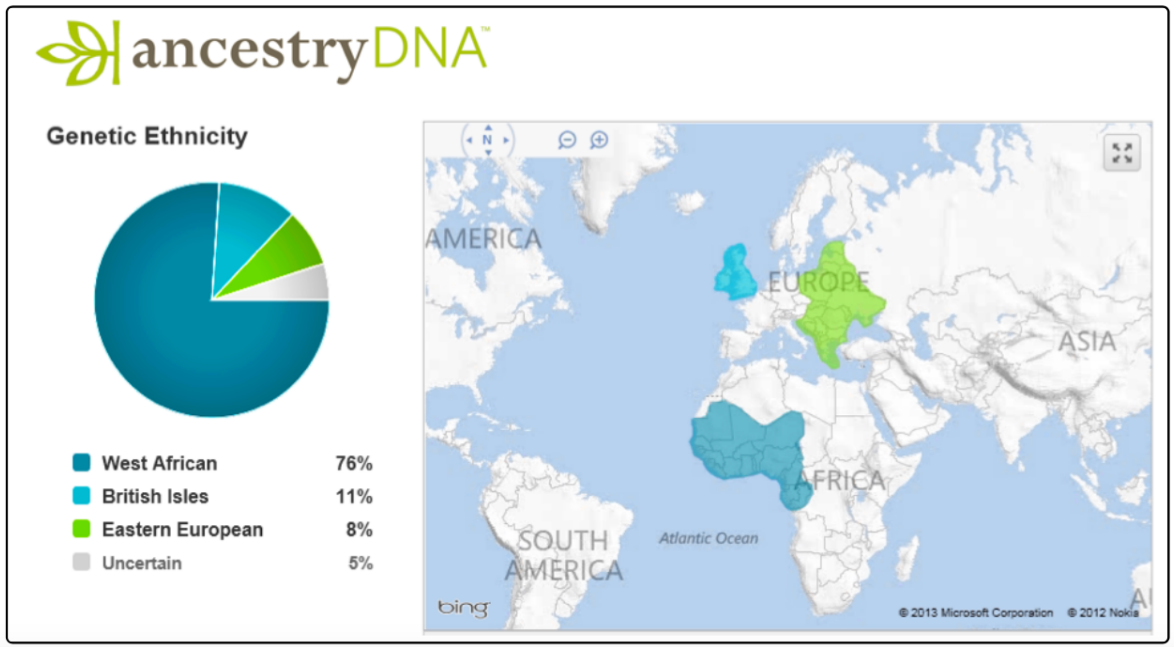
What questions we will look at

Based on DNA from a set of samples we are interested in answering:

- Is there population structure?
- Which samples are from which population?
- Are specific samples admixed?
- If so what are the ancestral populations?
- And what are the individuals' admixture proportions?

Why?

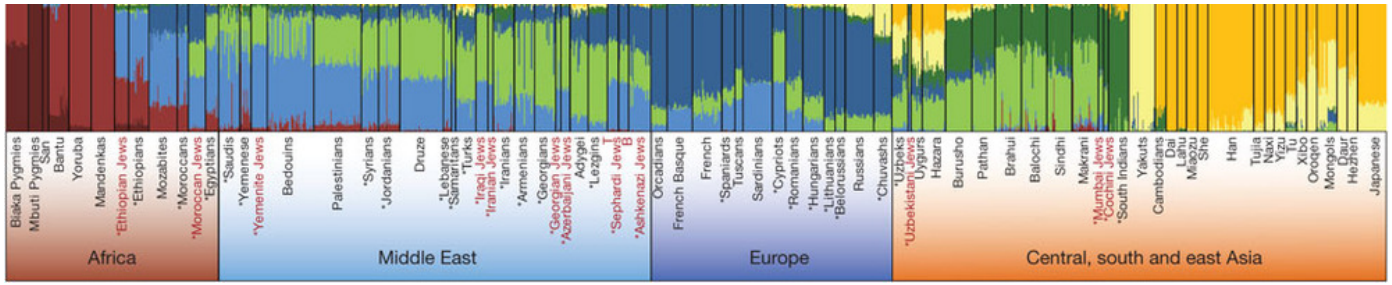
- Can be of interest in itself, e.g.
  - to learn about the history of a population
  - to learn about the history of specific individuals



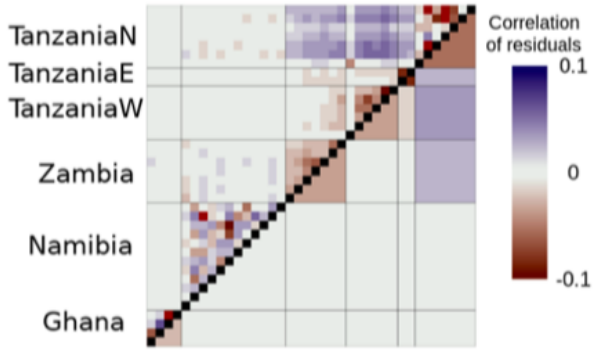
- But also used for QC and to ensure analyses are performed properly, e.g.
  - if not properly dealt with can lead to a lot of false positives in GWAS

How will we try to answer the questions?

- Many difference methods!
- This afternoon:
  - ADMIXTURE and similar model-based clustering methods



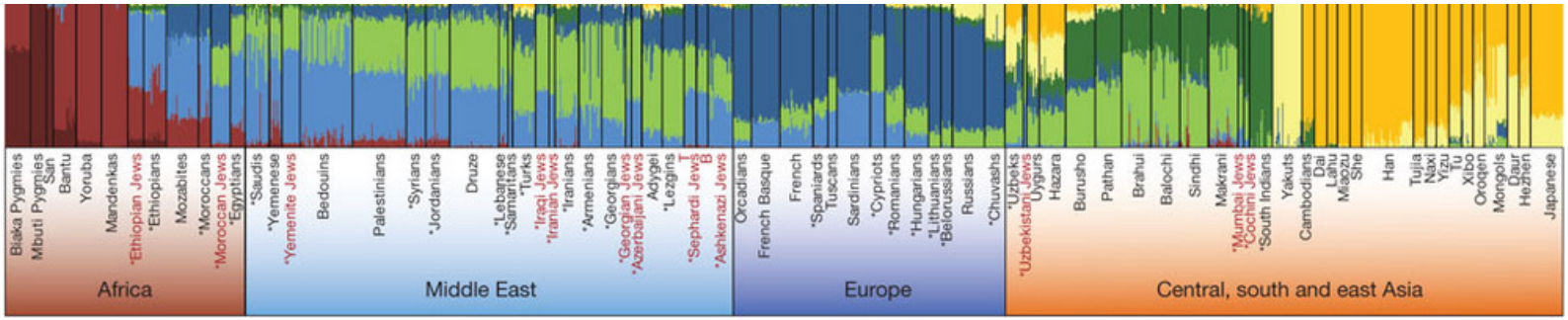
  - Problems and a few solutions including EvalAdmix



- Later this week: PCA, D/F statistics, finestructure

Outline

ADMIXTURE and similar model-based clustering methods

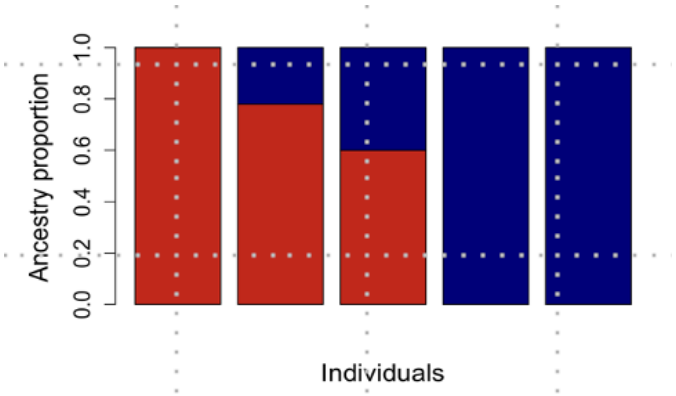


Can help answer the questions if

- admixture is recent or not present at all
- the (sub)populations differ sufficiently in terms of allele frequencies

Basic setup/problem

- **Underlying assumption**: All analyzed individuals have DNA that originates from one or more of $K \geq 1$ ancestral source populations.
- **Consequence**: they will have some proportion of DNA from each of the K populations (admixture proportions)
- **Note:** can be 0 for all but 1 population (individuals can be unadmixed):

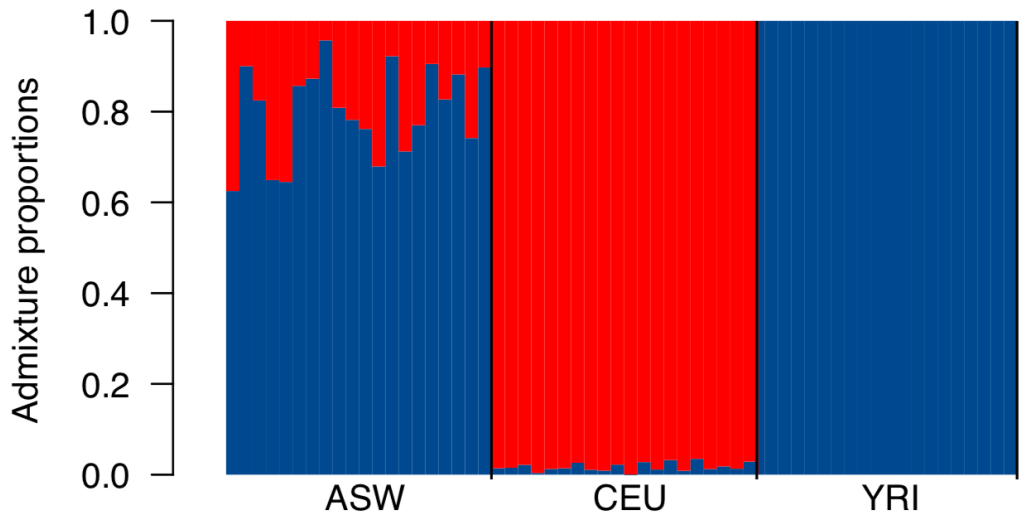


- **Goal:** to infer admixture proportions

Overview of well known solutions

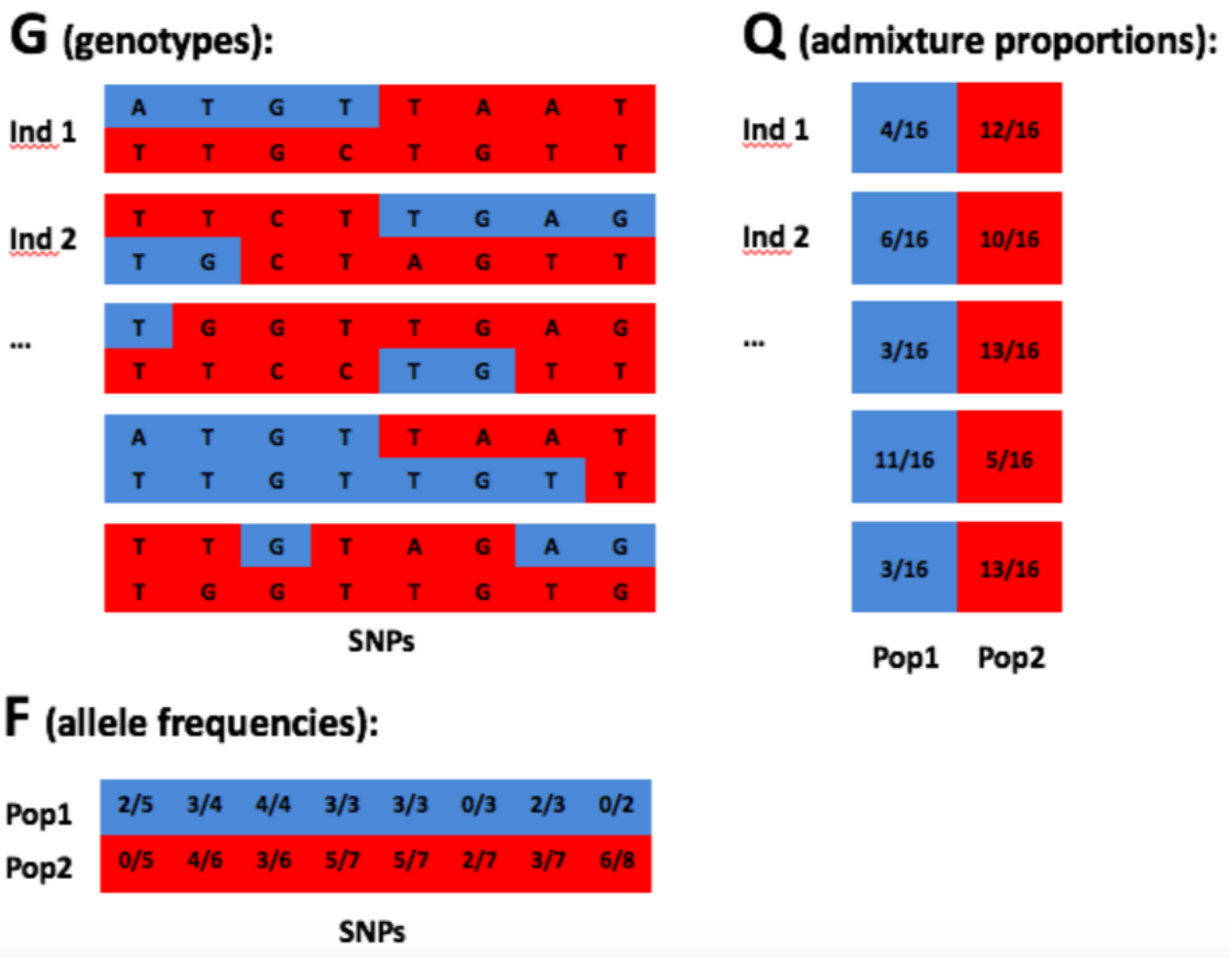
- Several methods:
  - Maximum Likelihood (ML): e.g ADMIXTURE (Alexander et al 2009)
  - Bayesian: e.g. STRUCTURE (Pritchard et al. 2000)
- They all base their inference on called genotypes G and infer
  1. Admixture proportions for all samples, Q, e.g.:

$$Q = \begin{pmatrix} 0.62 & 0.38 \\ 0.90 & 0.10 \\ 0.86 & 0.14 \\ \dots & \\ 0.00 & 1.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \\ \dots & \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{pmatrix}$$



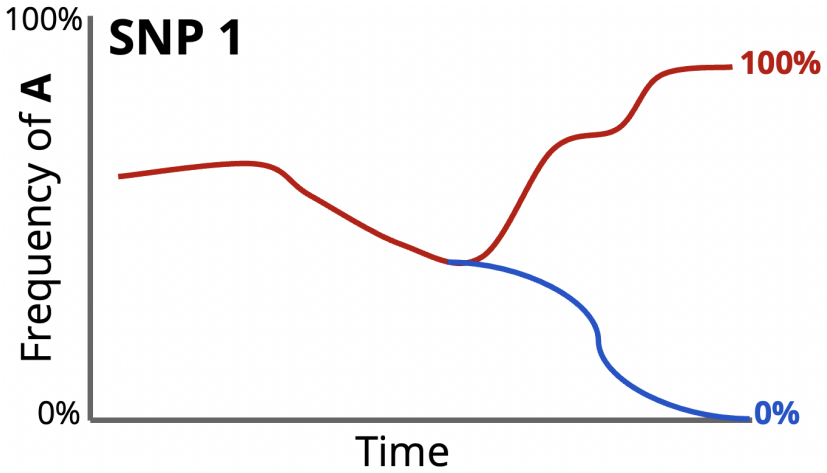
  2. Allele frequencies for all loci for all K populations, F

Easy if we knew the underlying ancestry...



Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- Let's think about a simple, clearcut example:

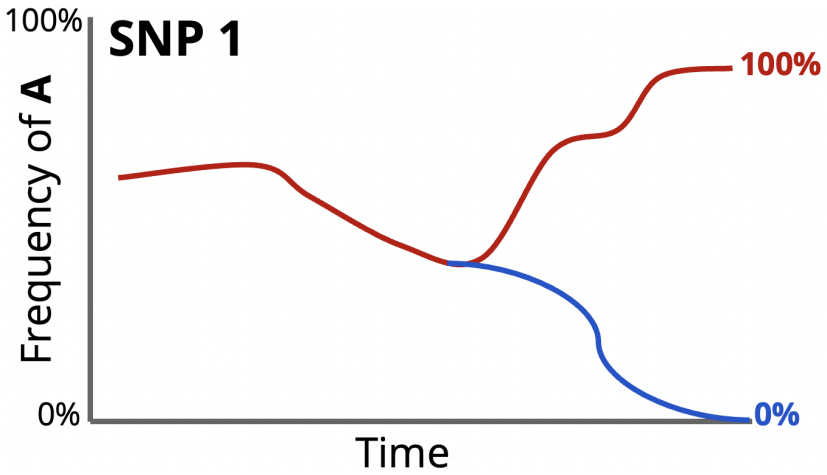


In SNP 1 individual has genotype AA

What does that tell us about the individuals ancestry in that SNP?

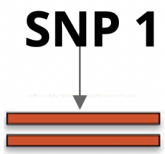
Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- Let's think about a simple, clearcut example:



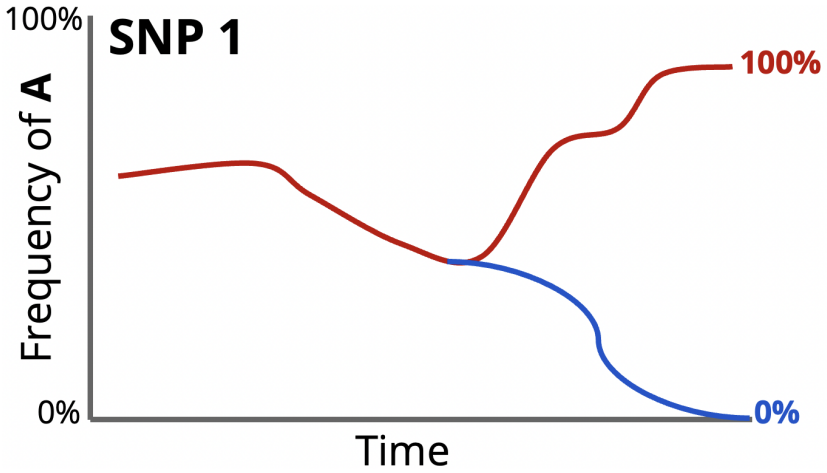
In SNP 1 individual has genotype AA

What does that tell us about the individuals ancestry in that SNP?



Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- Let's think about a simple, clearcut example:

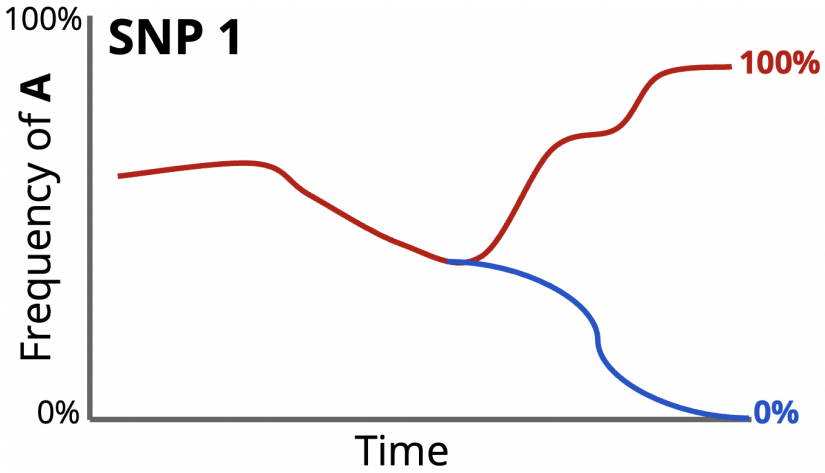


In SNP 1 individual has genotype AG

What does that tell us about the individuals ancestry in that SNP?

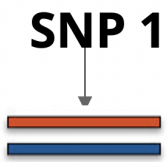
Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- Let's think about a simple, clearcut example:



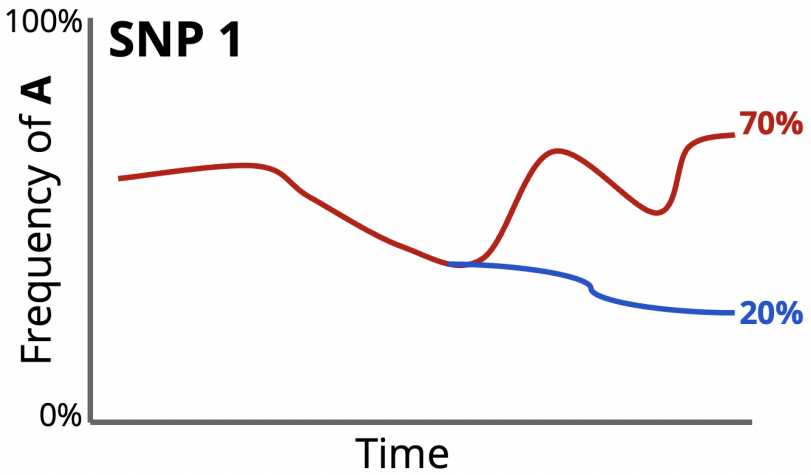
In SNP 1 individual has genotype AG

What does that tell us about the individuals ancestry in that SNP?



Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- What about a more realistic example:

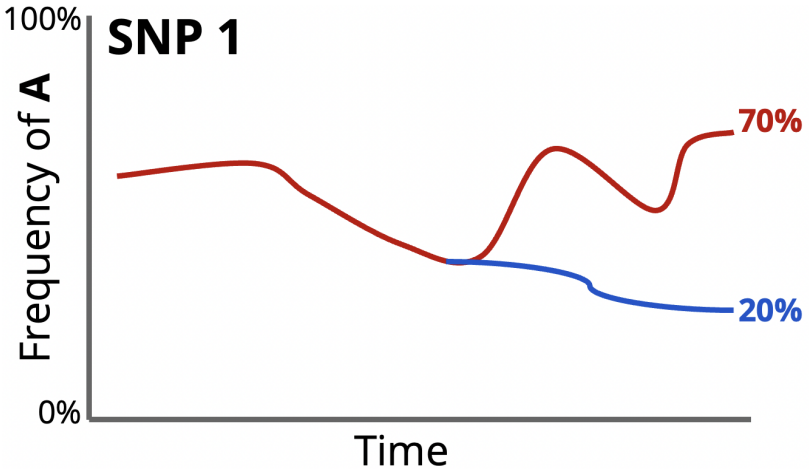


In SNP 1 individual has genotype AA

What does that tell us about the individuals ancestry in that SNP?

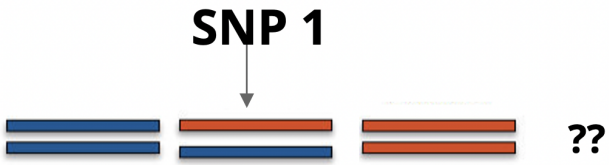
Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- What about a more realistic example:



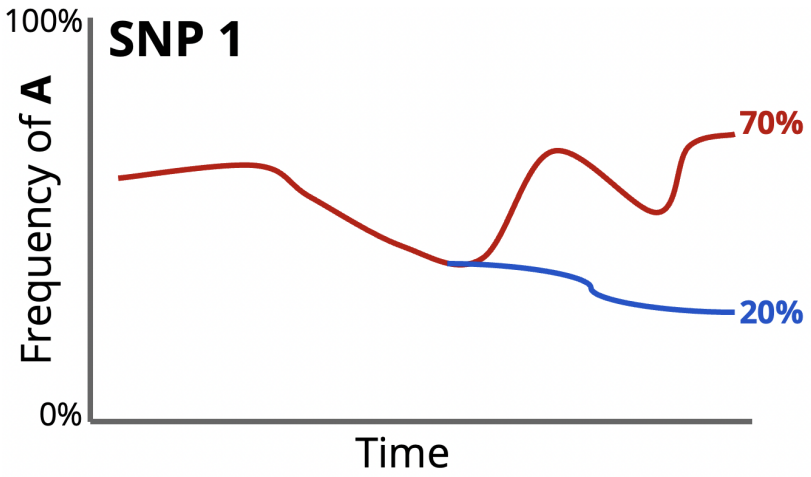
In SNP 1 individual has genotype AA

What does that tell us about the individuals ancestry in that SNP?



Basic idea/intuition behind if we do not

- To exploit that the allele frequencies differ in different (sub)populations
- What about a more realistic example:



In SNP 1 individual has genotype AA

What does that tell us about the individuals ancestry in that SNP?

SNP 1

↓

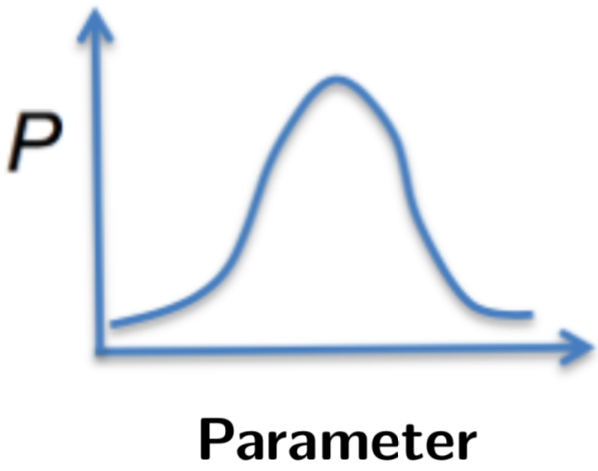
Probability of observing AA is higher in red population

- This is still informative - and this is the information we will use
- Most often enough when many SNPs are combined

Outline

ML solution overview

- We will for now assume we have called genotypes for several samples, G
- To find an ML solution, i.e. MLE for Q and F, we have to
  - Define a model/likelihood function $P(data|Q,F)$
  - Find (an efficient way to find) $\underset{(Q,F)}{\text{argmax}}\ P(data|Q,F)$



- The latter is usually solved using EM which I will not focus on
- Instead I will spend time describing the model/likelihood function

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- A and B denote the possible alleles at locus j
- G_{ij} be the genotype of i at locus j (measured in counts of allele A)
- $F^j = (f^{j1}, f^{j2}, ..., f^{jK})$ denote the allele frequencies of allele A
- $Q^i = (q^{i1}, q^{i2}, ..., q^{iK})$ denote i's genome-wide admixture proportions

Then the likelihood function is $P(G_{ij}|Q^i, F^j)$.

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- A and B denote the possible alleles at locus j
- G_{ij} be the genotype of i at locus j (measured in counts of allele A)
- $F^j = (f^{j1}, f^{j2}, ..., f^{jK})$ denote the allele frequencies of allele A
- $Q^i = (q^{i1}, q^{i2}, ..., q^{iK})$ denote i's genome-wide admixture proportions

Then the likelihood function is $P(G_{ij}|Q^i, F^j)$. And to get that we notice

- for one of i's allele copies: $P(A|Q^i, F^j) = q^{i1}f^{j1} + q^{i2}f^{j2}... + q^{iK}f^{jK} = h^{ij}$

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- A and B denote the possible alleles at locus j
- G_{ij} be the genotype of i at locus j (measured in counts of allele A)
- $F^j = (f^{j1}, f^{j2}, ..., f^{jK})$ denote the allele frequencies of allele A
- $Q^i = (q^{i1}, q^{i2}, ..., q^{iK})$ denote i's genome-wide admixture proportions

Then the likelihood function is $P(G_{ij}|Q^i, F^j)$. And to get that we notice

- for one of i's allele copies: $P(A|Q^i, F^j) = q^{i1}f^{j1} + q^{i2}f^{j2}... + q^{iK}f^{jK} = h^{ij}$
- and therefore $P(B|Q^i, F^j) = 1 - h^{ij}$

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- A and B denote the possible alleles at locus j
- G_{ij} be the genotype of i at locus j (measured in counts of allele A)
- $F^j = (f^{j1}, f^{j2}, ..., f^{jK})$ denote the allele frequencies of allele A
- $Q^i = (q^{i1}, q^{i2}, ..., q^{iK})$ denote i's genome-wide admixture proportions

Then the likelihood function is $P(G_{ij}|Q^i, F^j)$. And to get that we notice

- for one of i's allele copies: $P(A|Q^i, F^j) = q^{i1}f^{j1} + q^{i2}f^{j2}... + q^{iK}f^{jK} = h^{ij}$
- and therefore $P(B|Q^i, F^j) = 1 - h^{ij}$
- so assuming HWE we get the likelihood function:

$$P(G_{ij}|Q^i, F^j) = \begin{cases} (h^{ij})^2 & \text{if } G_{ij} = 2 \ (AA), \\ 2h^{ij}(1-h^{ij}) & \text{if } G_{ij} = 1 \ (AB), \\ (1-h^{ij})^2 & \text{if } G_{ij} = 0 \ (BB). \end{cases}$$

Likelihood function (1 individual i, 1 diallelic locus j)

Assume K source populations and let

- A and B denote the possible alleles at locus j
- G_{ij} be the genotype of i at locus j (measured in counts of allele A)
- $F^j = (f^{j1}, f^{j2}, ..., f^{jK})$ denote the allele frequencies of allele A
- $Q^i = (q^{i1}, q^{i2}, ..., q^{iK})$ denote i's genome-wide admixture proportions

Then the likelihood function is $P(G_{ij}|Q^i, F^j)$. And to get that we notice

- for one of i's allele copies: $P(A|Q^i, F^j) = q^{i1}f^{j1} + q^{i2}f^{j2}... + q^{iK}f^{jK} = h^{ij}$
- and therefore $P(B|Q^i, F^j) = 1 - h^{ij}$
- so assuming HWE we get the likelihood function:

$$P(G_{ij}|Q^i, F^j) = \begin{cases} (h^{ij})^2 & \text{if } G_{ij} = 2 \ (AA), \\ 2h^{ij}(1-h^{ij}) & \text{if } G_{ij} = 1 \ (AB), \\ (1-h^{ij})^2 & \text{if } G_{ij} = 0 \ (BB). \end{cases}$$

- notice e.g. for an individual with G_{ij}=AA:
  if frequency of A is higher in say pop1, a higher admixture proportion for pop1 will lead to a higher likelihood - so we are capturing what we wanted

Likelihood function (N individuals, M diallelic loci)

- If we assume:
  - the individuals are unrelated and thus independent
  - loci are independent

  we can write the likelihood as

$$P(G|Q,F) = \prod_{i}^{N}\prod_{j}^{M} P(G_{ij}|Q^i, F^j)$$

with $G = (G_{ij})$, $Q = (Q^1, Q^2, ...Q^N)$ and $F = (F^1, F^2, ...F^M)$.

Likelihood function (N individuals, M diallelic loci)

- If we assume:
  - the individuals are unrelated and thus independent
  - loci are independent

  we can write the likelihood as

$$P(G|Q,F) = \prod_{i}^{N}\prod_{j}^{M} P(G_{ij}|Q^i, F^j)$$

  with $G = (G_{ij})$, $Q = (Q^1, Q^2, ...Q^N)$ and $F = (F^1, F^2, ...F^M)$.

- Based on this we can find ML estimates: $(\hat{Q}, \hat{F}) = \underset{(Q,F)}{\text{argmax}}\ p(G|Q,F)$.
- Often done using some kind of EM-algorithm

Some problems (overview)

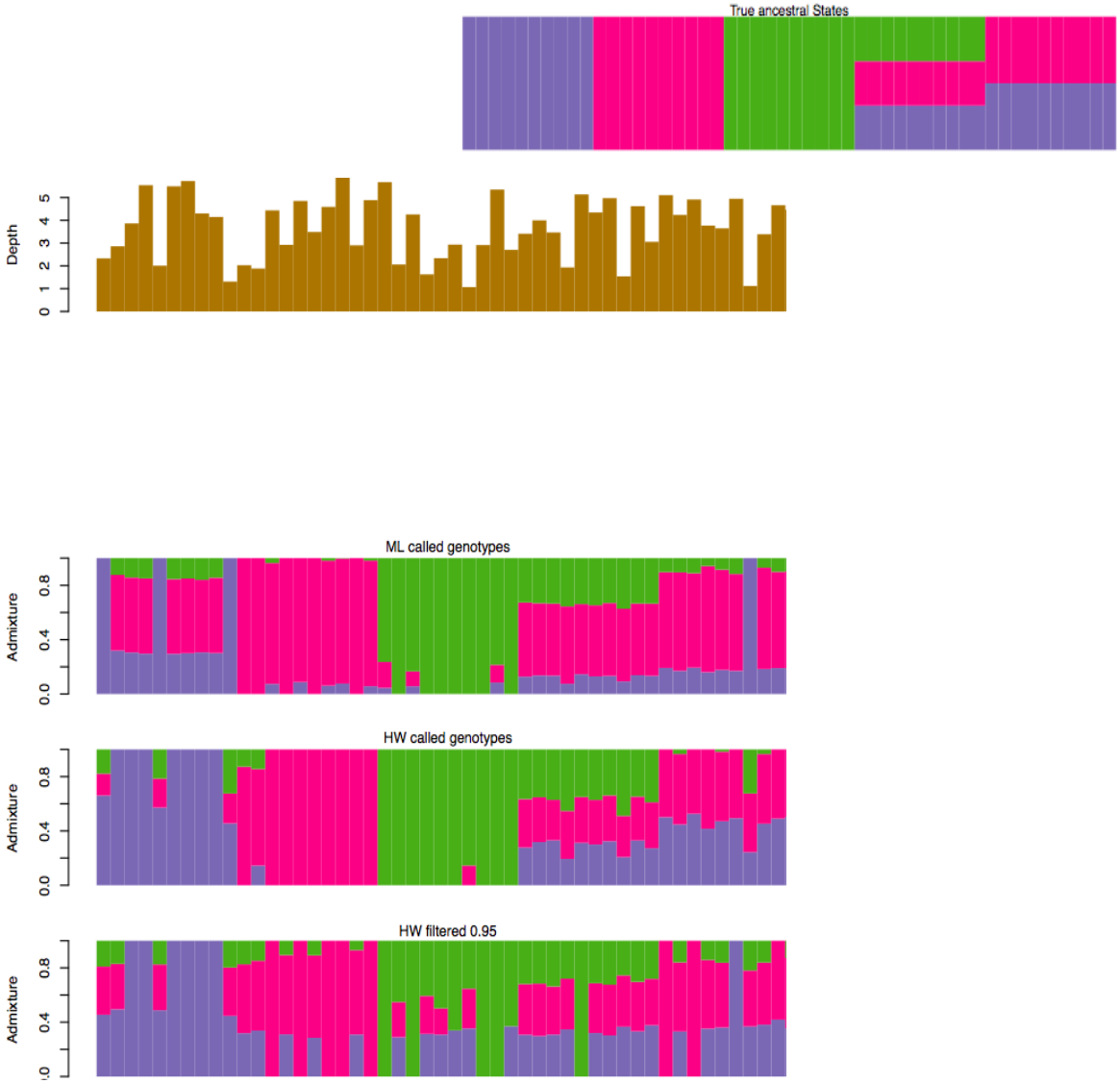
There are several potential problems in practice, among these:

- Does not always work well for NGS data
- Convergence
- What is the right value for K?
- Interpreting the admixture proportions can be tricky!

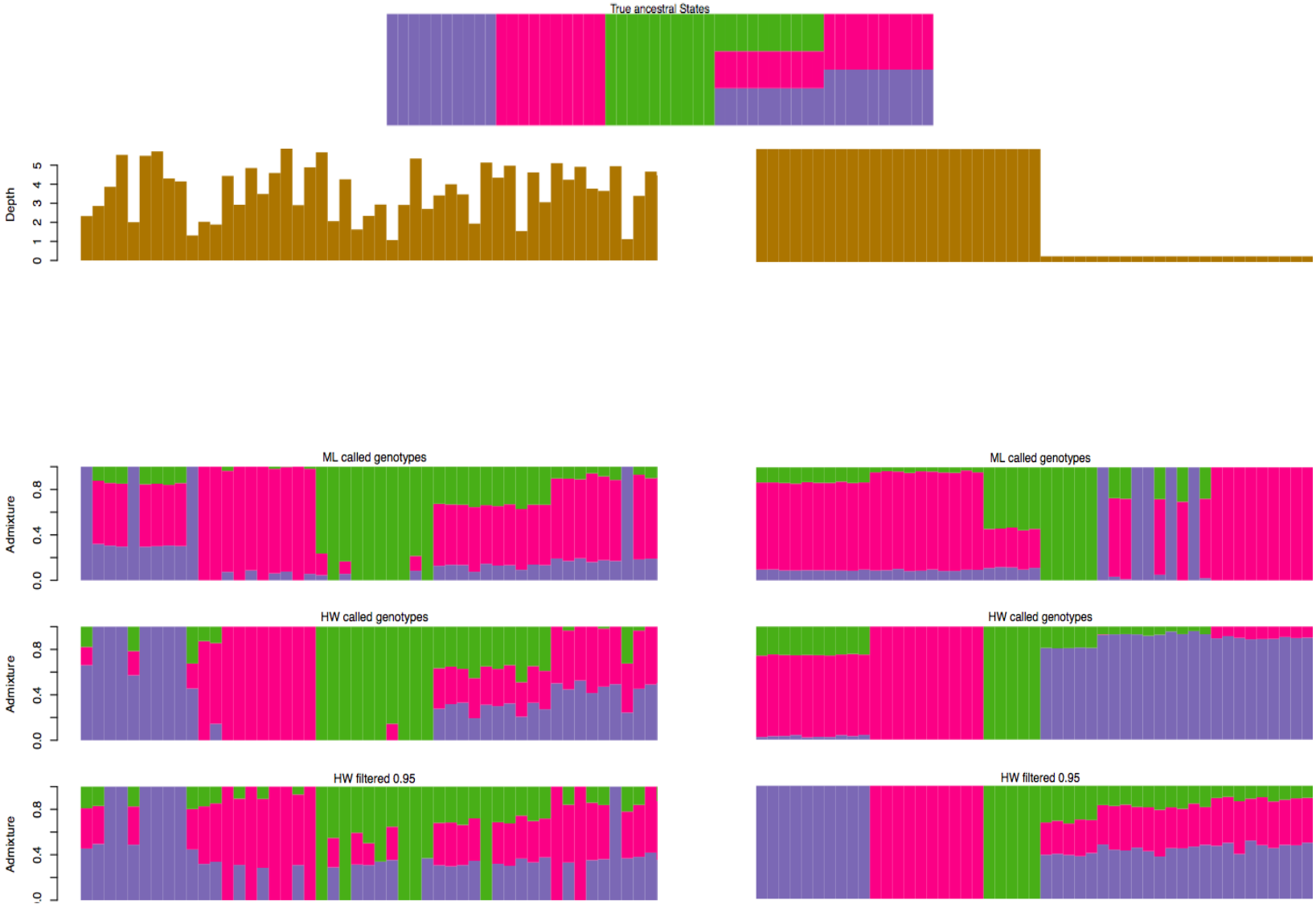
Some problems (NGS data)

- (Almost) all programs available work on called genotypes
- This can be very problematic in some cases if you have NGS data.
- Well described in this paper: Skotte et al. 2013, Genetics
- Just a few examples here

Some problems (NGS data, variable depth)



Some problems (NGS data, variable depth)



Solution: NGSadmix

- Works on genotype likelihoods instead of called genotypes
- I.e. input is $P(X_{ij}|G_{ij})$ for all 3 possible values of G_{ij}, where X_{ij} is NGS data for individual i at locus j

Solution: NGSadmix

- Works on genotype likelihoods instead of called genotypes
- I.e. input is $P(X_{ij}|G_{ij})$ for all 3 possible values of G_{ij}, where X_{ij} is NGS data for individual i at locus j
- The previous likelihood is extended from

$$P(G|Q,F) = \prod_i^N \prod_j^M P(G_{ij}|Q^i, F^j)$$

to

$$P(X|Q,F) = \prod_i^N \prod_j^M P(X_{ij}|Q^i, F^j) = \prod_i^N \prod_j^M \sum_{G_{ij} \in \{0,1,2\}} P(X_{ij}|G_{ij}) P(G_{ij}|Q^i, F^j)$$

Solution: NGSadmix

- Works on genotype likelihoods instead of called genotypes
- I.e. input is $P(X_{ij}|G_{ij})$ for all 3 possible values of G_{ij}, where X_{ij} is NGS data for individual i at locus j
- The previous likelihood is extended from

$$P(G|Q,F) = \prod_i^N \prod_j^M P(G_{ij}|Q^i, F^j)$$

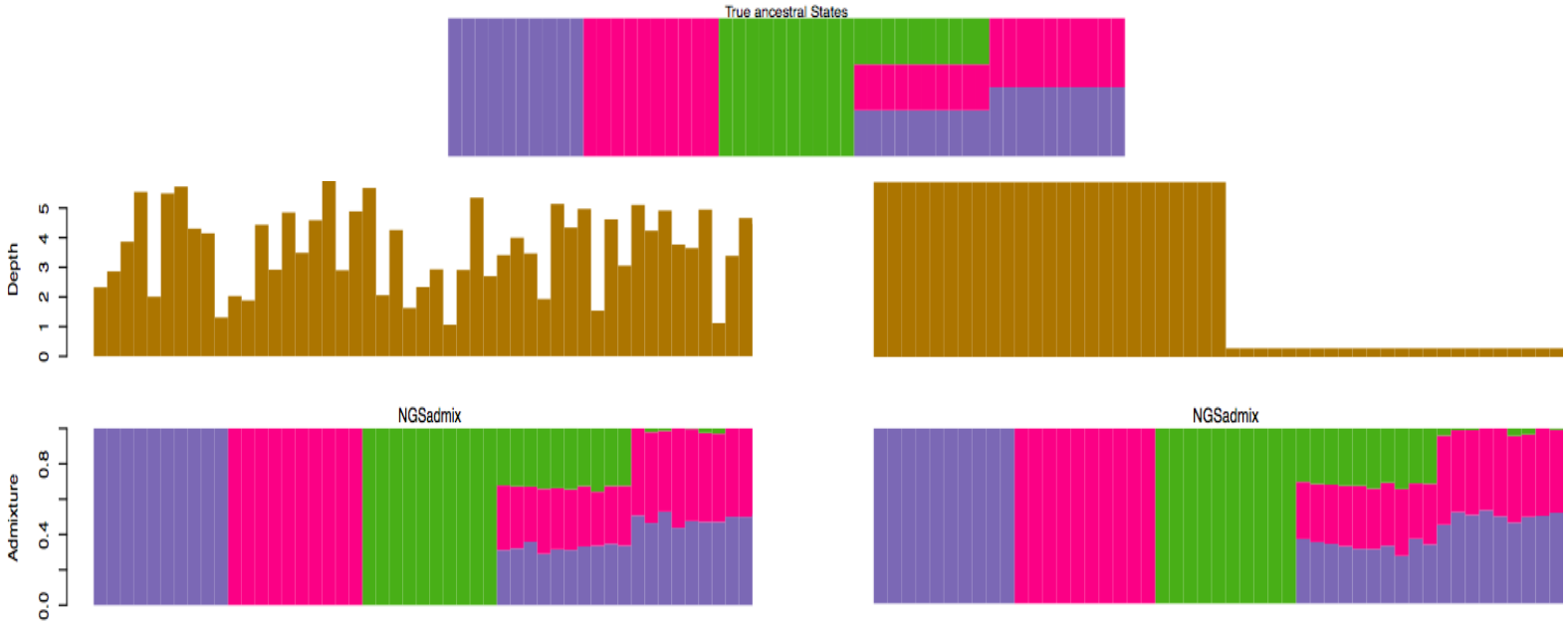
to

$$P(X|Q,F) = \prod_i^N \prod_j^M P(X_{ij}|Q^i, F^j) = \prod_i^N \prod_j^M \sum_{G_{ij} \in \{0,1,2\}} P(X_{ij}|G_{ij}) P(G_{ij}|Q^i, F^j)$$

- Note that for known genotypes the two are equivalent
- A solution is found using an EM-algorithm

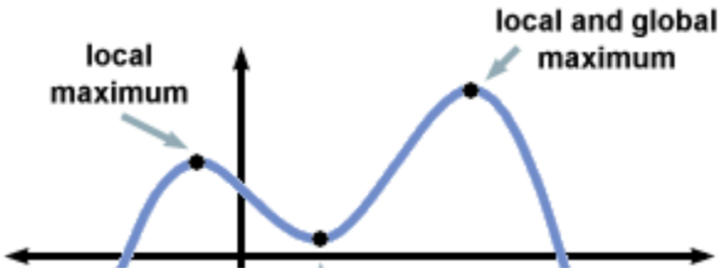
Solution: NGSadmix

- Does well even for low depth and variable depth data:



Some problems (convergence)

- Sometimes EM stops before it reaches the **global maximum** likelihood
- Either because it reaches a preset maximum number of steps to run
- Or because it instead stops in a **local maximum**



- When this happens we say that the analysis did not converge
- In this case the estimates provided are NOT ML estimates

Some problems (convergence)

- This can happen even with 1 parameter
- But tends to happen more often with many parameters
- And we are trying to optimize a lot of parameters!
- E.g. if you analyse 100 individuals with K=2 and have data from 1M loci then you are estimating 2,000,100 parameters! (why?)

Some problems (convergence)

- This can happen even with 1 parameter
- But tends to happen more often with many parameters
- And we are trying to optimize a lot of parameters!
- E.g. if you analyse 100 individuals with K=2 and have data from 1M loci then you are estimating 2,000,100 parameters! (why?)
- **So the programs do not always converge to the ML estimate!**
- **The resulting solutions can differ a lot from ML estimate**
- Especially tends to be a problem when K is high (why?)

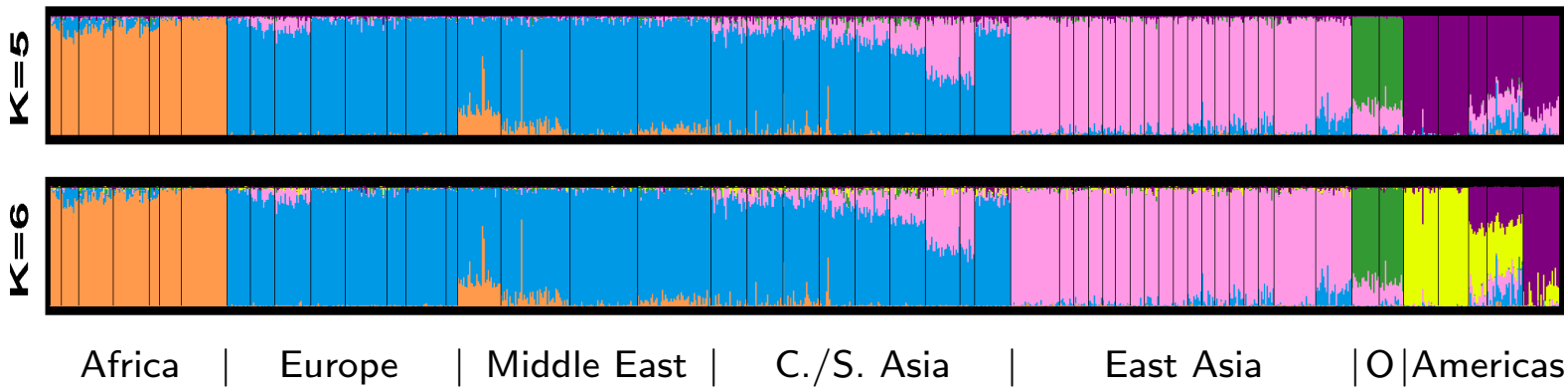
Some problems (convergence)

- This can happen even with 1 parameter
- But tends to happen more often with many parameters
- And we are trying to optimize a lot of parameters!
- E.g. if you analyse 100 individuals with K=2 and have data from 1M loci then you are estimating 2,000,100 parameters! (why?)
- **So the programs do not always converge to the ML estimate!**
- **The resulting solutions can differ a lot from ML estimate**
- Especially tends to be a problem when K is high (why?)
- One has to **run analyses multiple times** (with different starting points!) and compare solutions
- One way to assess convergence is to look at the likelihood of all your runs and see if the top 5 are very similar (no guarantee)
- If not, run the analysis more times

Some problems (choice of K)

The choice of K is still not a solved problem

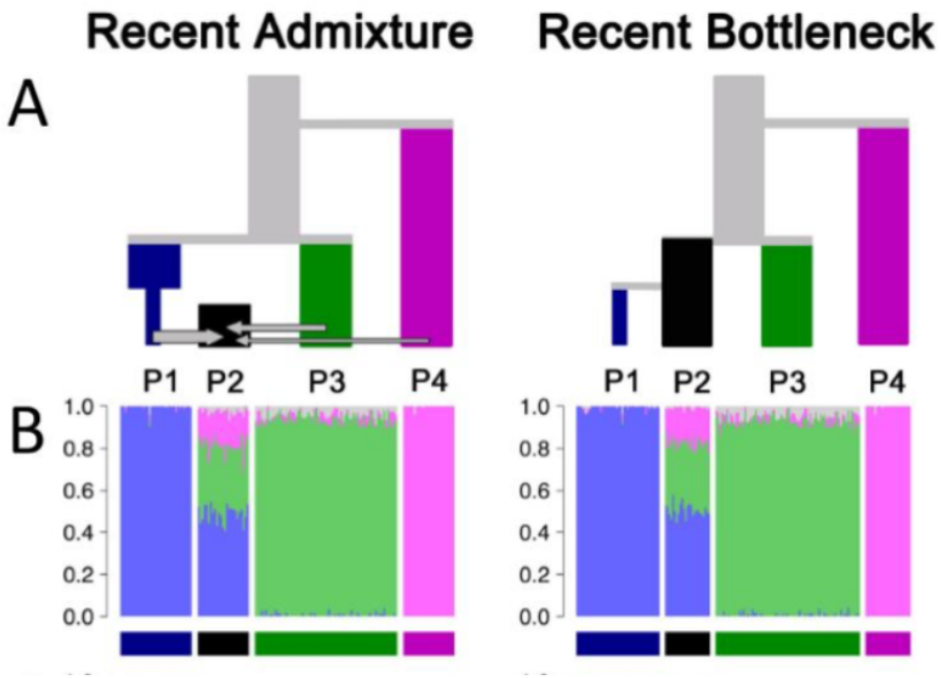
- Is there a right value? E.g. Rosenberg et al., 2002:



- One solution is to use several
- If you want to choose STRUCTURE/ADMIXTURE papers have solutions
- Also, see Garcia-Erill & Albrechtsen 2020, Molecular Ecology

Some problems (multiple scenarios can lead to same solutions)

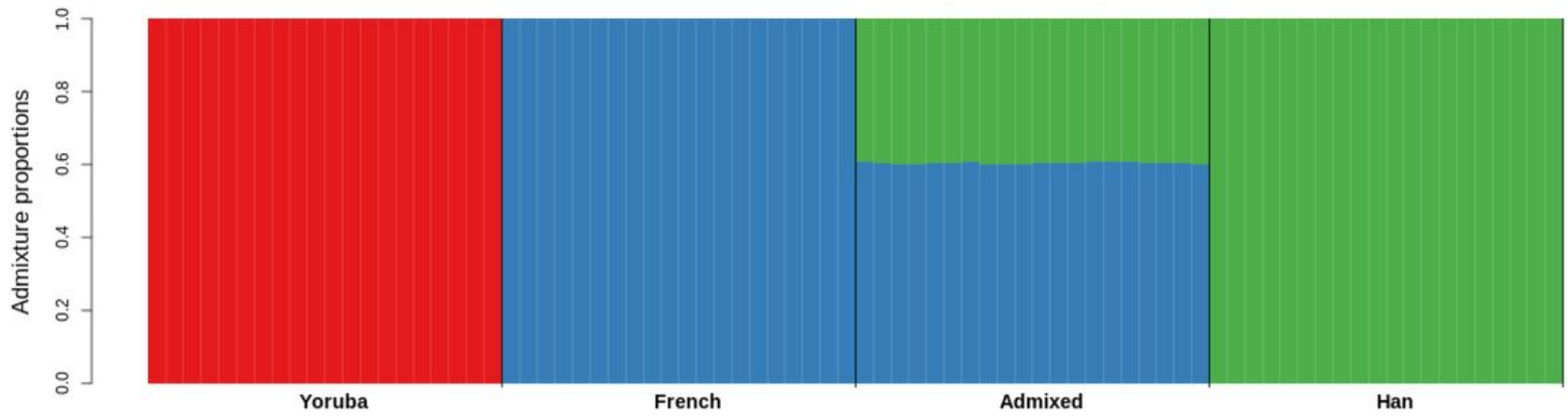
- Unfortunately, several scenarios can lead to (basically) same solutions, e.g.



- For details see Falush et al. called "A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots."
- So one has to be careful when interpreting the results!
- For a helpful tool, again see Garcia-Erill & Albrechtsen 2020

Some problems (missing source population)

- You may not have the correct source populations and that can affect both the estimates and the interpretation.
- E.g. imagine you have samples from a population where all individuals are 30% Native American and 70% European but you only have unadmixed samples from YRI, Han Chinese and French (so no Native Americans):



- So again one has to be careful when interpreting the results!
- For a helpful tool, again see Garcia-Erill & Albrechtsen 2020

EvalAdmix (Garcia-Erill & Albrechtsen 2020)

- A tool that allows you to assess the model fit
- The basic idea is to look at the difference between the expected genotype given the model, $\hat{G}_{ij}$ and the true genotype G_{ij}, also called the residuals:

$$r_{ij} = G_{ij} - \hat{G}_{ij} = G_{ij} - 2\sum_{k=1}^{K} q_{ik} f_{jk}$$

- And then the fit is measured by the mean correlation of these residuals for each pair of individuals, a and b ($E(\hat{\rho}_{ab})$)

EvalAdmix (Garcia-Erill & Albrechtsen 2020)

- A tool that allows you to assess the model fit
- The basic idea is to look at the difference between the expected genotype given the model, $\hat{G}_{ij}$ and the true genotype G_{ij}, also called the residuals:

$$r_{ij} = G_{ij} - \hat{G}_{ij} = G_{ij} - 2\sum_{k=1}^{K} q_{ik} f_{jk}$$

- And then the fit is measured by the mean correlation of these residuals for each pair of individuals, a and b ($E(\hat{\rho}_{ab})$)
- Because:

If individual a and b are sampled from the same population, and they have a good model fit, all error will be random and residuals will be uncorrelated:

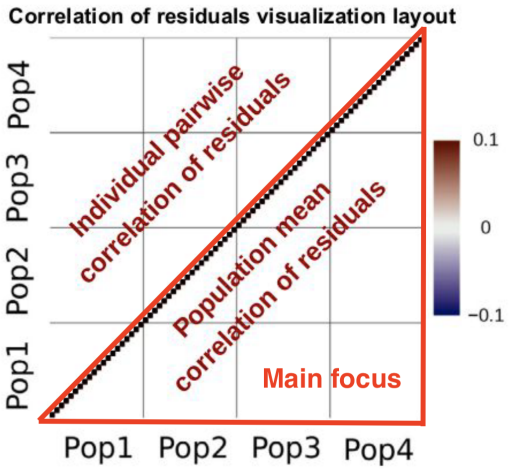
$$E[\hat{\rho}_{ab}] = 0$$

If individual a and b are sampled from the same population AND they have a bad model fit, they will share a systematic error and their residuals will be positively correlated:

$$E[\hat{\rho}_{ab}] > 0$$

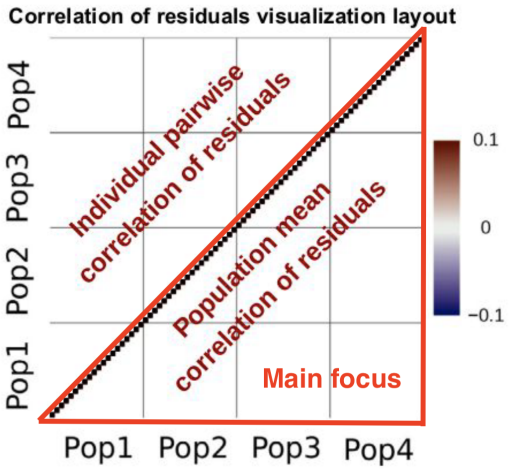
EvalAdmix (Garcia-Erill & Albrechtsen 2020)

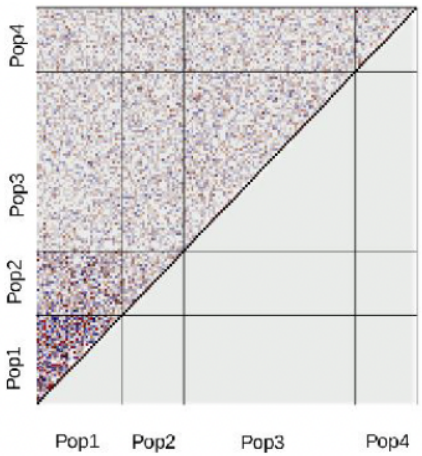
- In practice this means we can assess the fit with a plot like this:



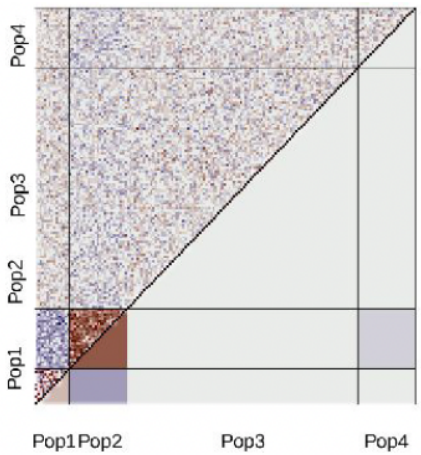
EvalAdmix (Garcia-Erill & Albrechtsen 2020)

- In practice this means we can assess the fit with a plot like this:





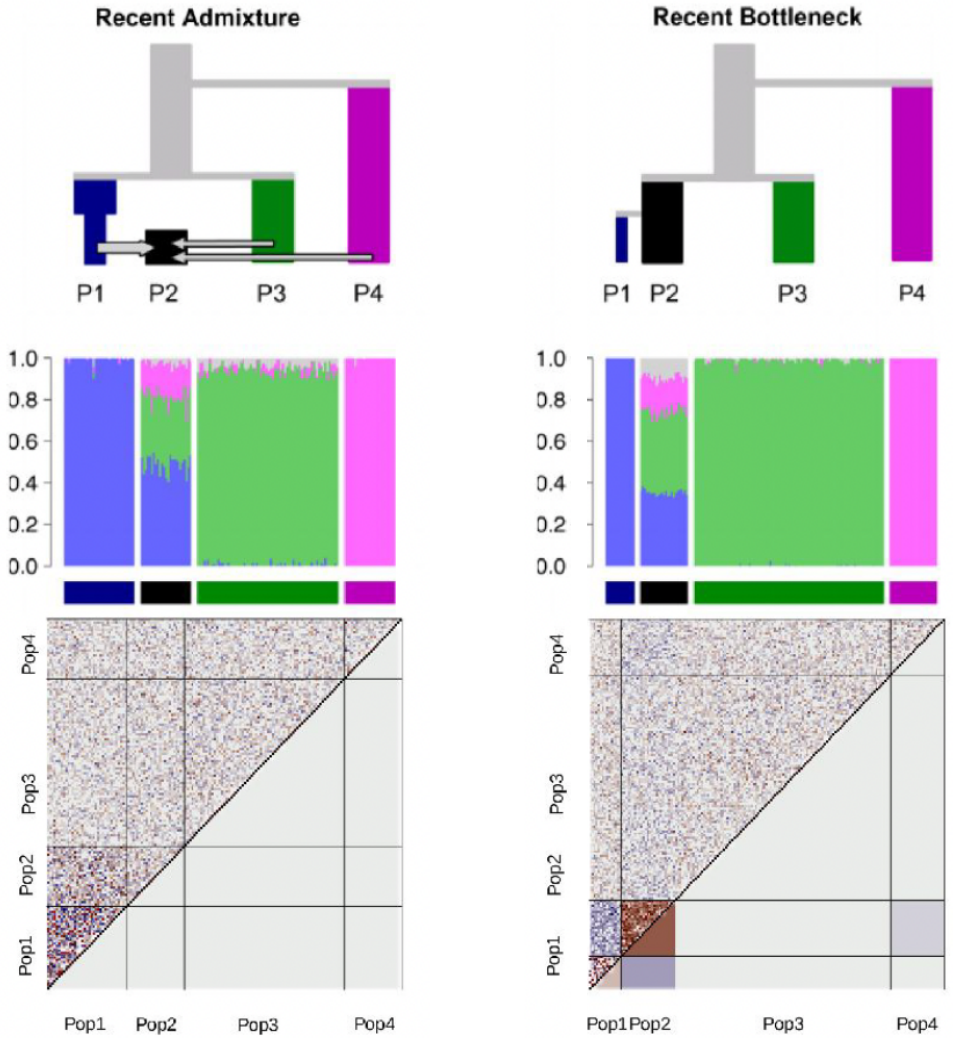
Good fit



Bad fit

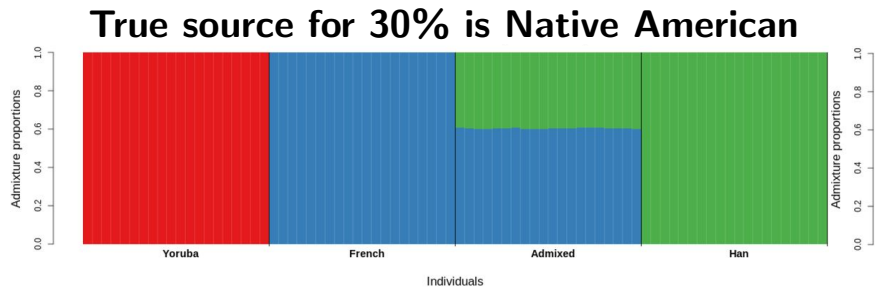
EvalAdmix (Garcia-Erill & Albrechtsen 2020)

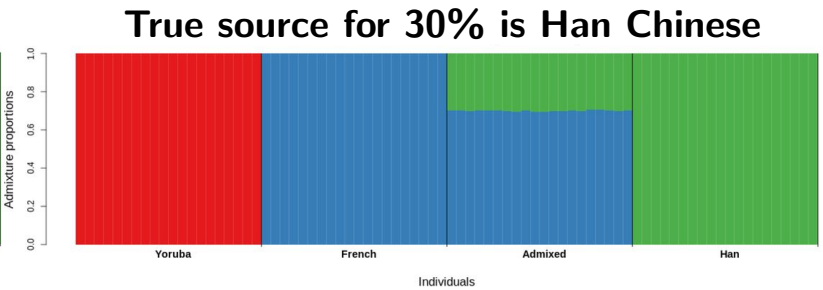
- E.g. let's look at evalAdmix plots for the three scenarios from before:

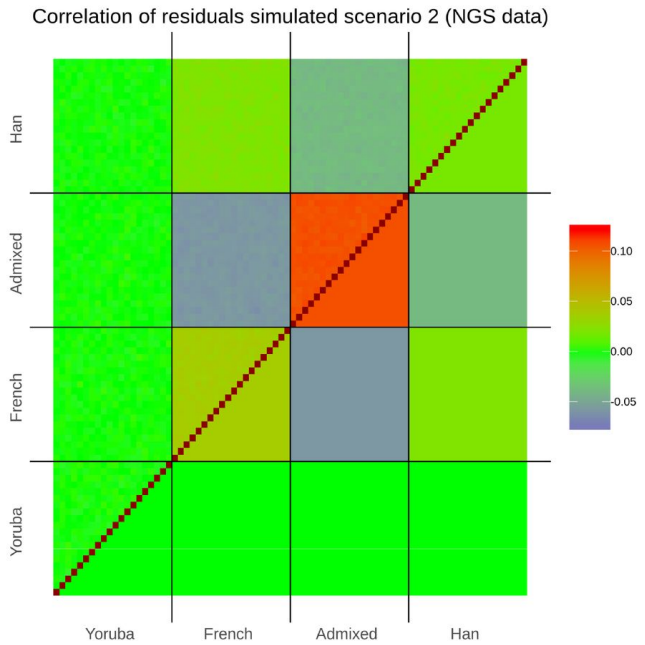


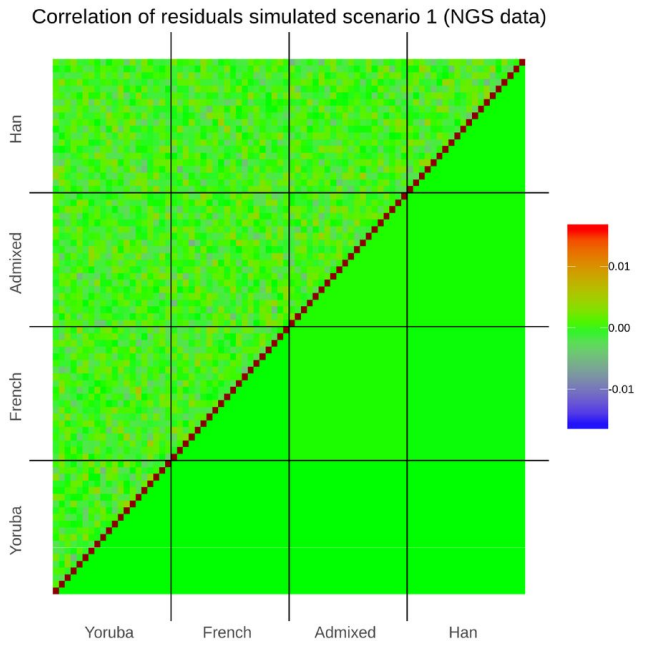
EvalAdmix (Garcia-Erill & Albrechtsen 2020)

- You can also use it to detect if a source population is not correct:



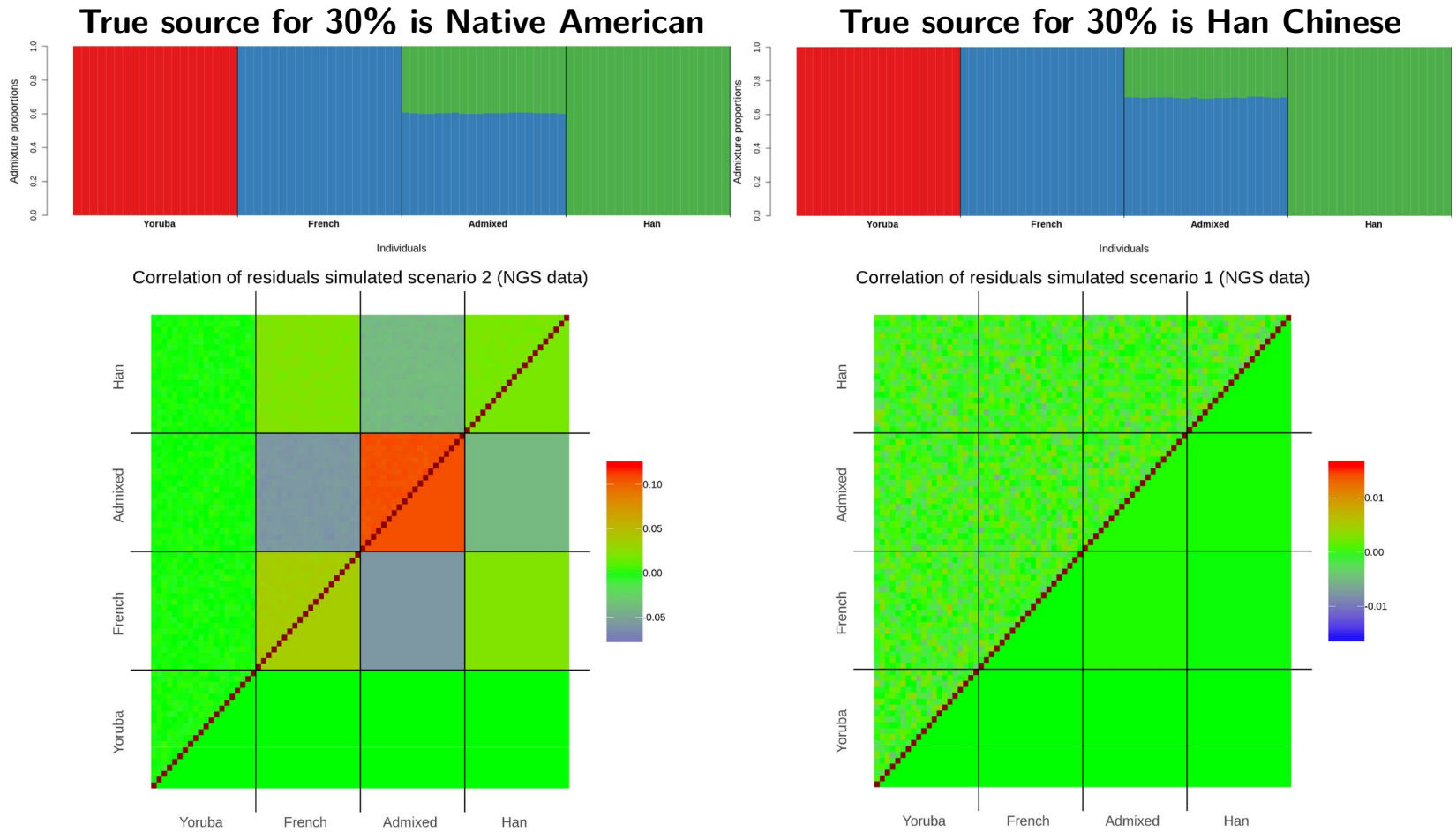






EvalAdmix (Garcia-Erill & Albrechtsen 2020)

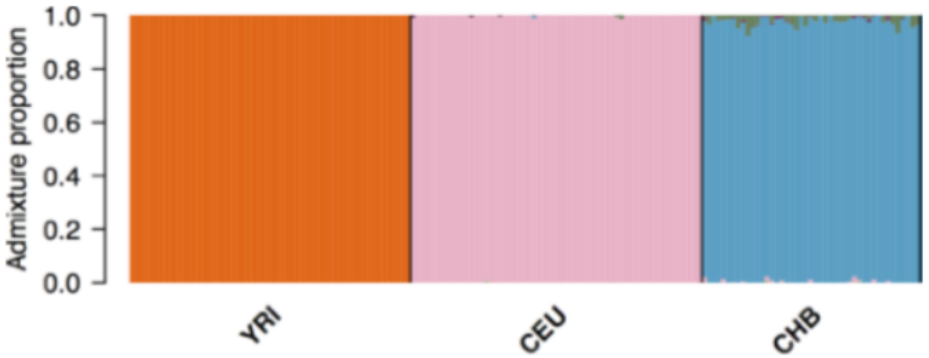
- You can also use it to detect if a source population is not correct:



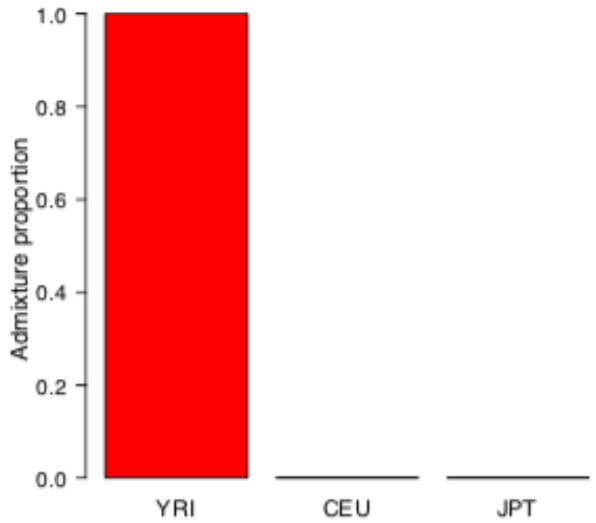
- You can find a lower bound for K by increasing K until the fit is not bad

Another related method: fastNGSadmix (Jørsboe et al. 2017)

- Similar in idea to NGSadmix, but works for single samples
- Uses allele frequencies from an appropriate reference panel, e.g.



- Gives estimates for an additional sample, e.g. for a YRI sample (if K=3):



or equivalently

Exercises

Go to the github, download the jupyter notebook and upload it on emily