Population structure and admixture proportions

Anders Albrechtsen

Population structure I

Outline for this session

- What is population structure and admixture proportions?
  - What can we use it for
- Methods for inferring population structure
  - Admixture proportions
  - PCA (next session)
- Inferring admixture proportions from data
  - Individual allele frequences

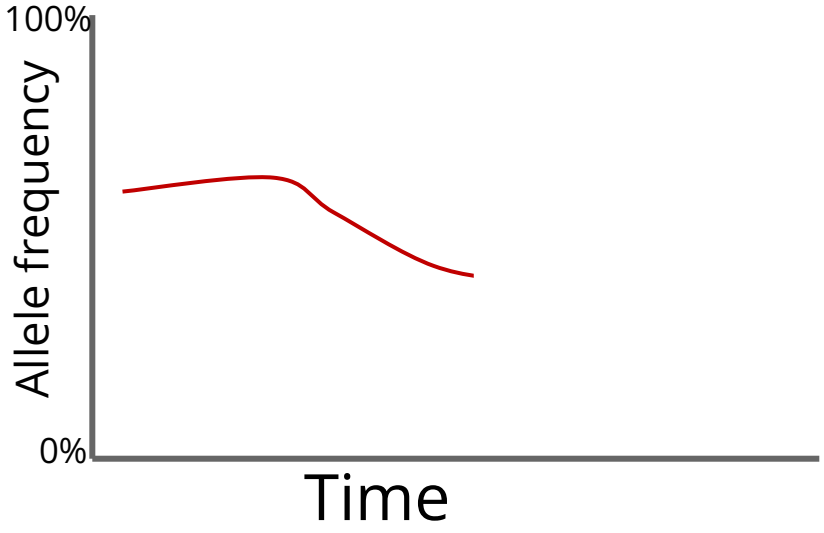
What is population structure?

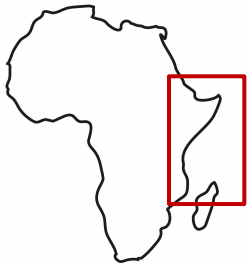
Wikipedia definition: *Population structure is the presence of a systematic difference in allele frequencies between subpopulations in a population possible due to different ancestry*

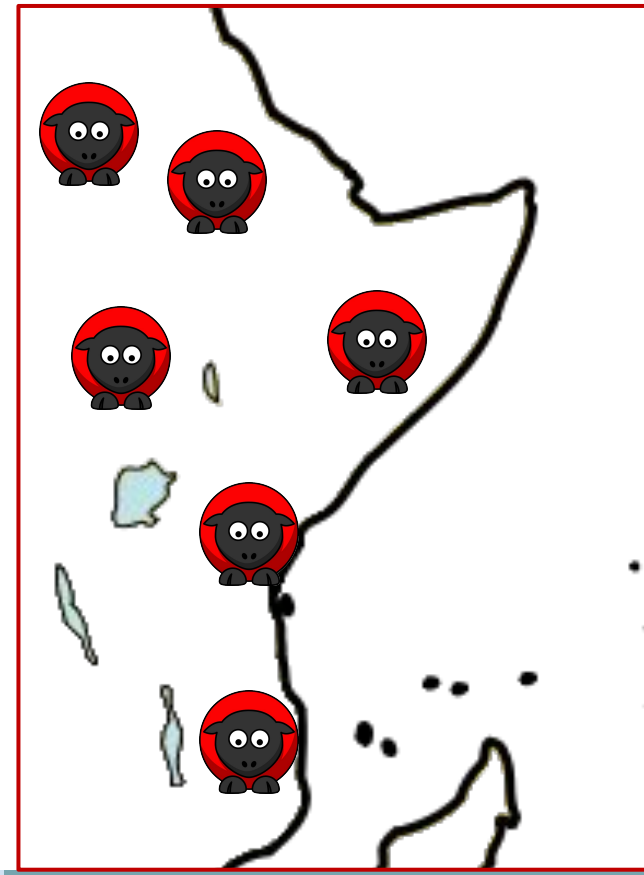
- (cur | prev) ○ 07:30, 17 June 2008 Andersduck (talk | contribs) . . (7,245 bytes) **(+7,245)** . . *(←Created page with 'Population stratification is the presence of a systematic difference in allele frequencies between subpopulations in a population possible due to different ancestry...')*

No population structure with random mating

Frequency changes due to drift

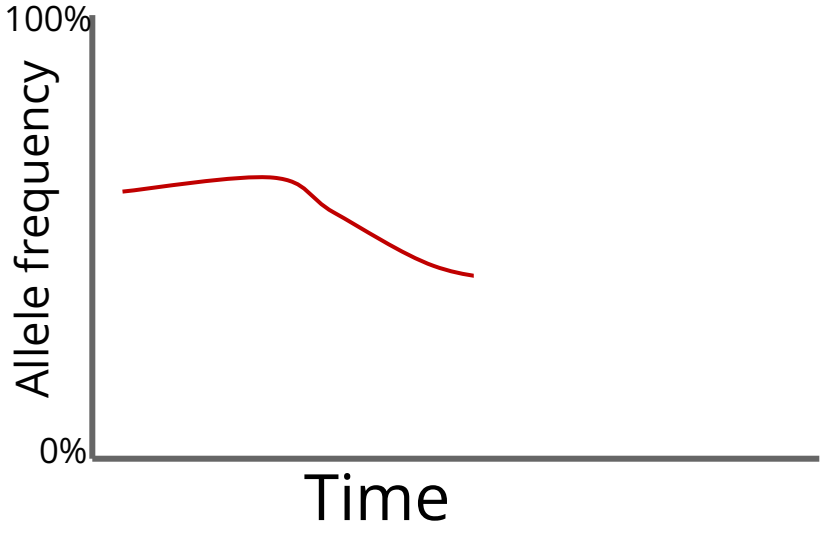


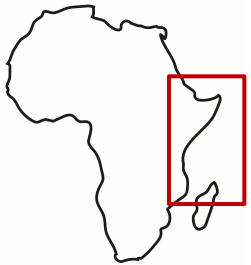


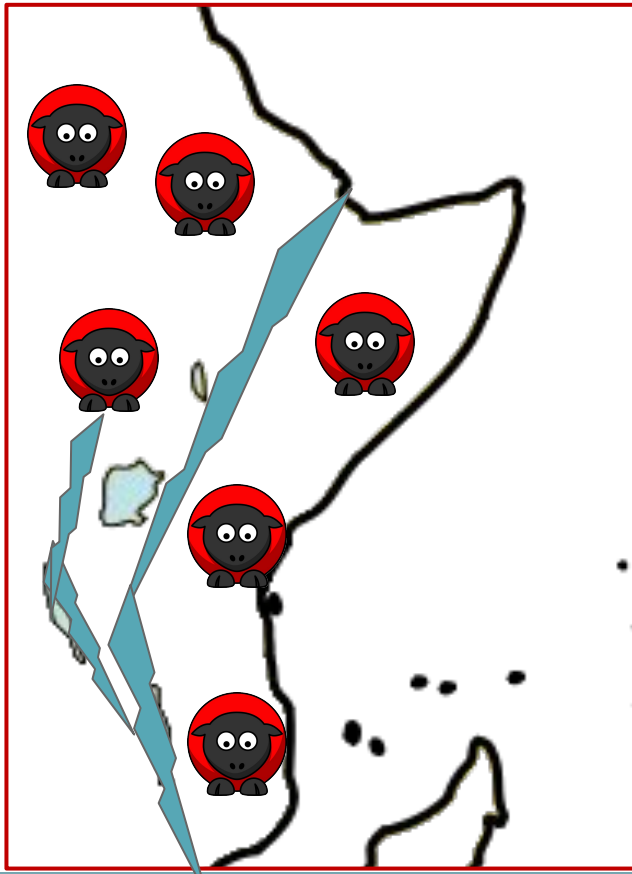


Barriers to gene flow leads to population structure

Frequency changes due to drift

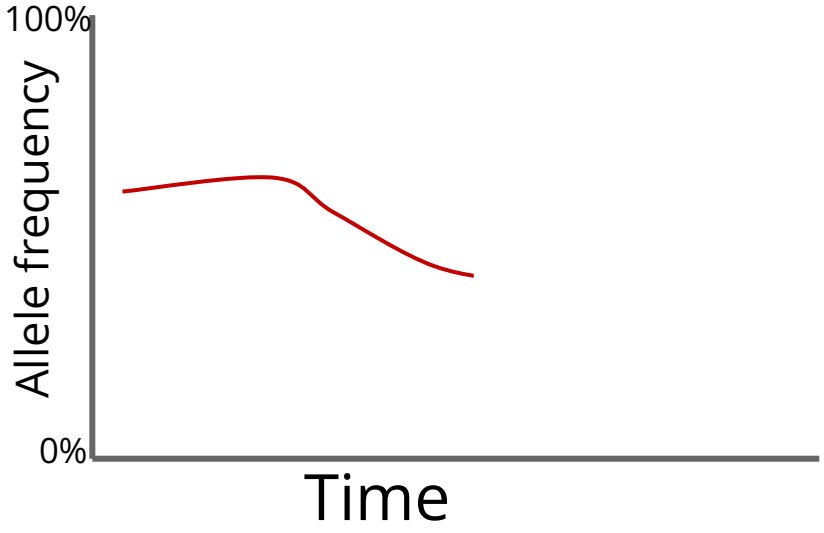


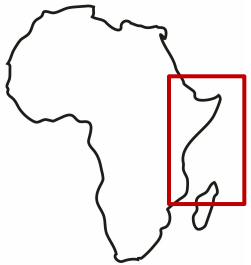


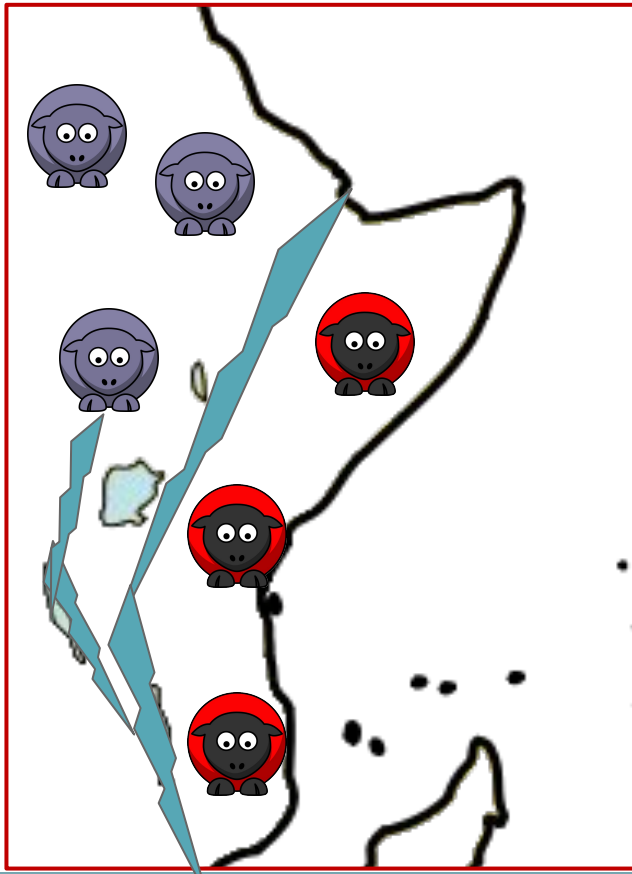


Barriers to gene flow leads to population structure

Frequency changes due to drift

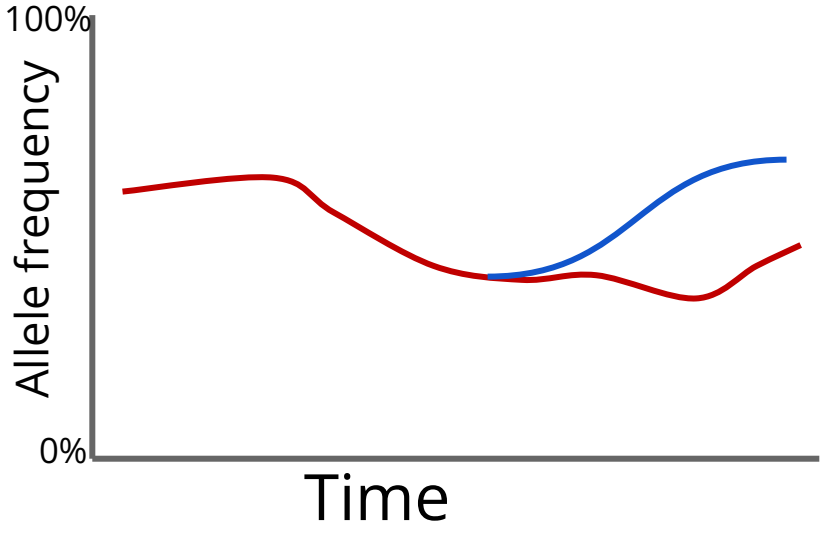


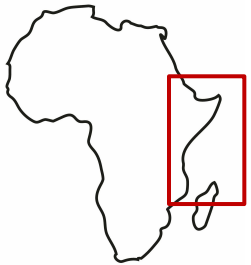


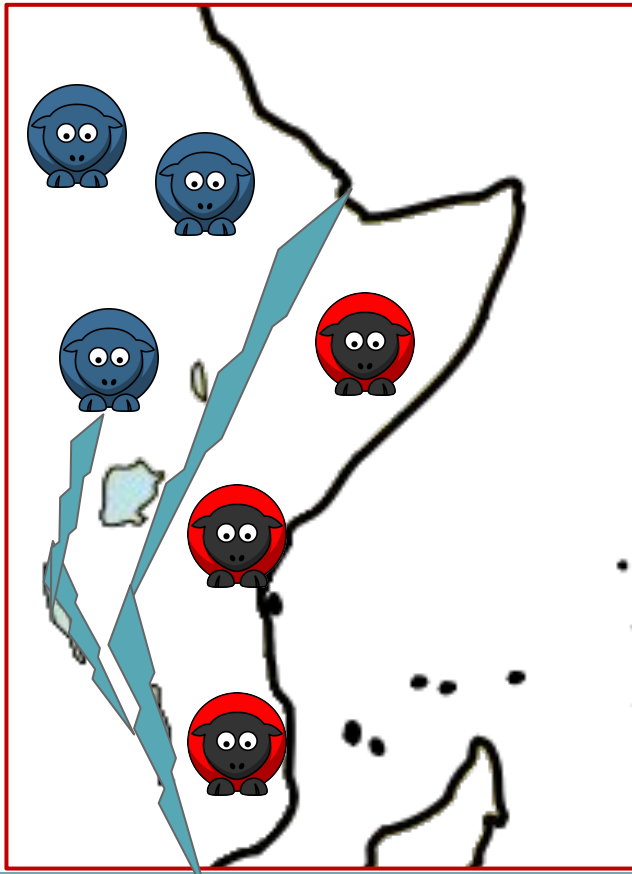


Barriers to gene flow leads to population structure

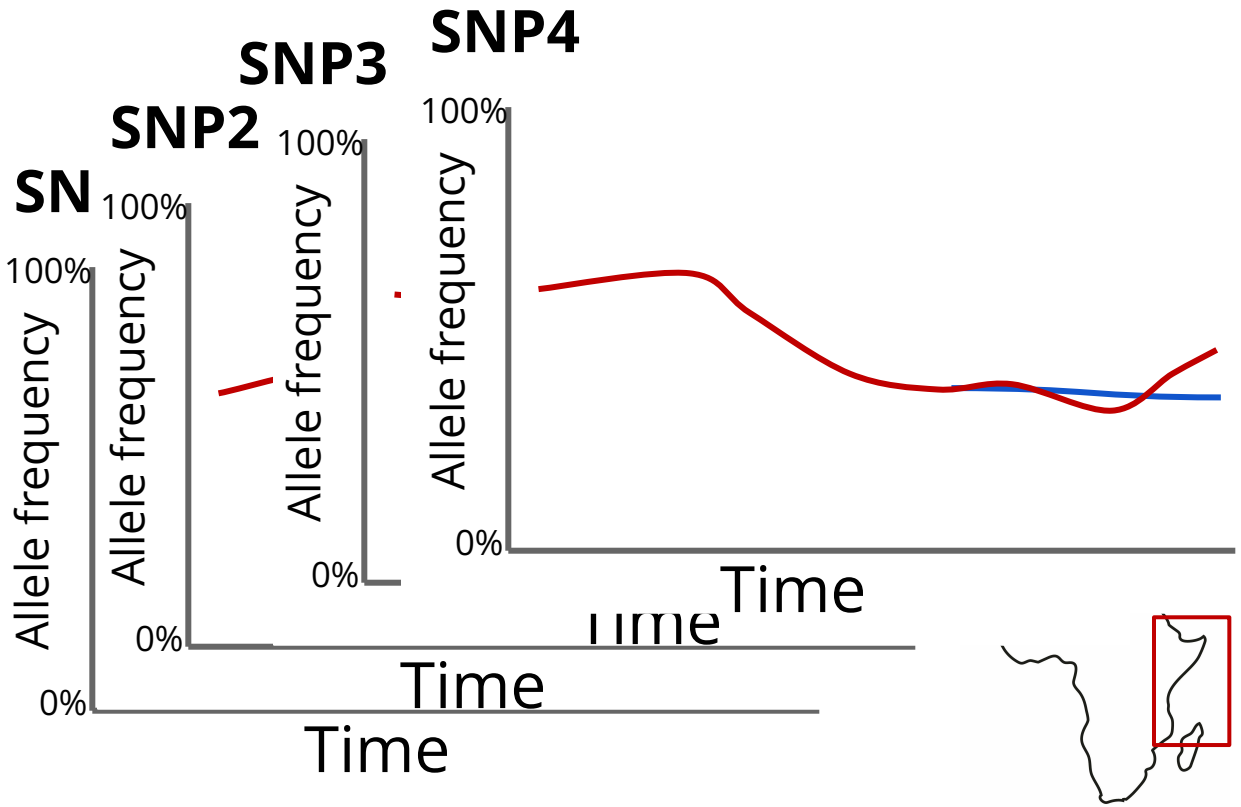
Independent drift

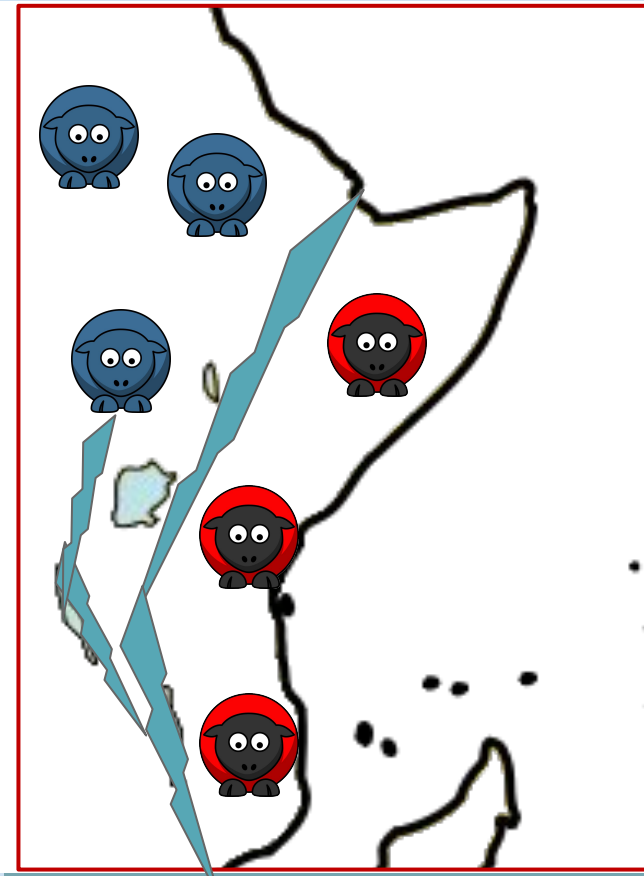






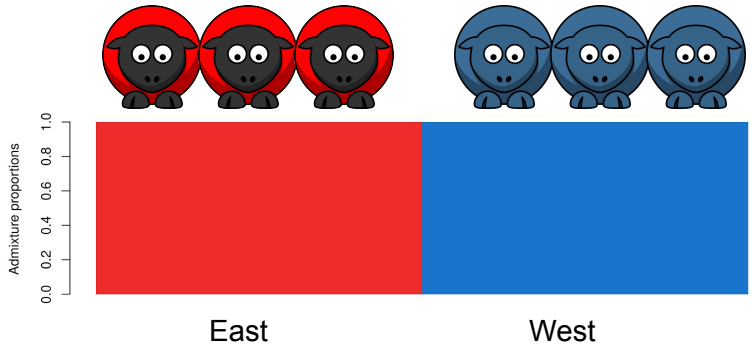
Affects all genetic variants

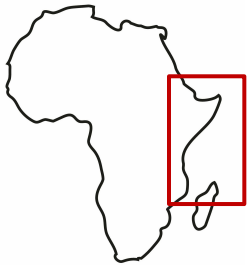


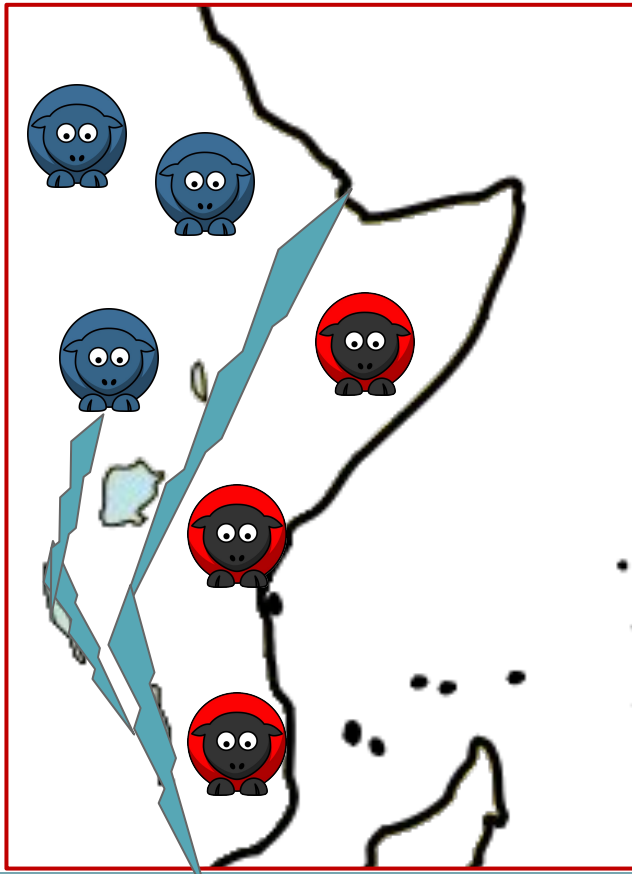


Barriers to gene flow leads to population structure

The systematic allele frequency difference allows us to separate the individuals based on their DNA

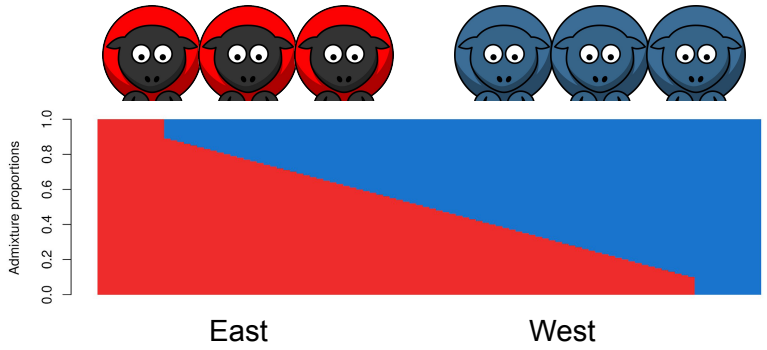


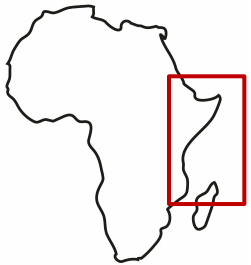


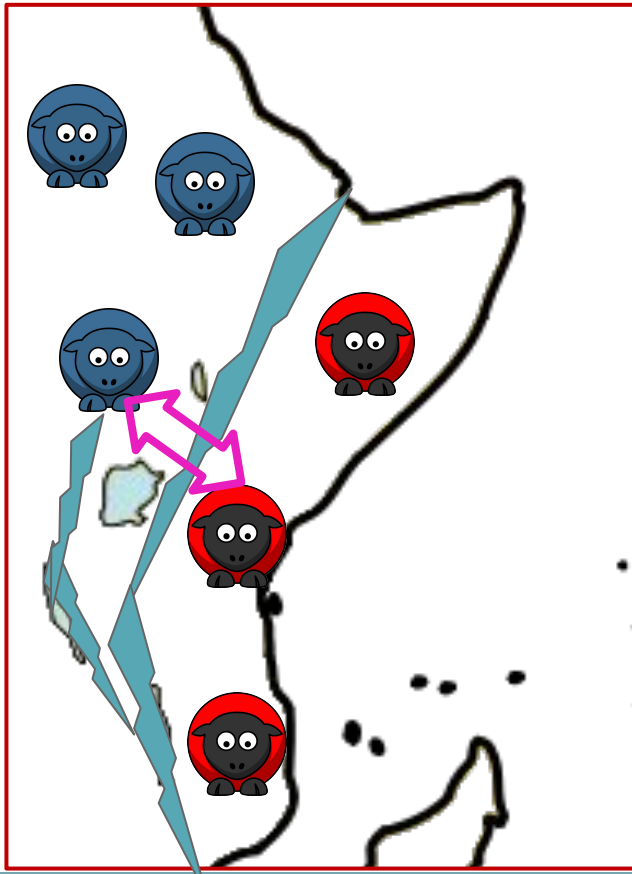


Admixture is geneflow after separation

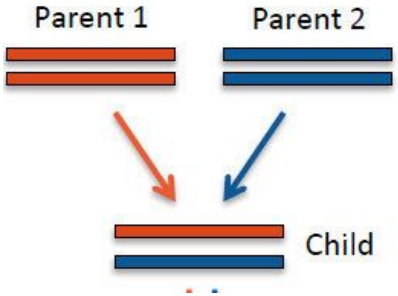
We call individuals admixed if they have ancestry from several populations



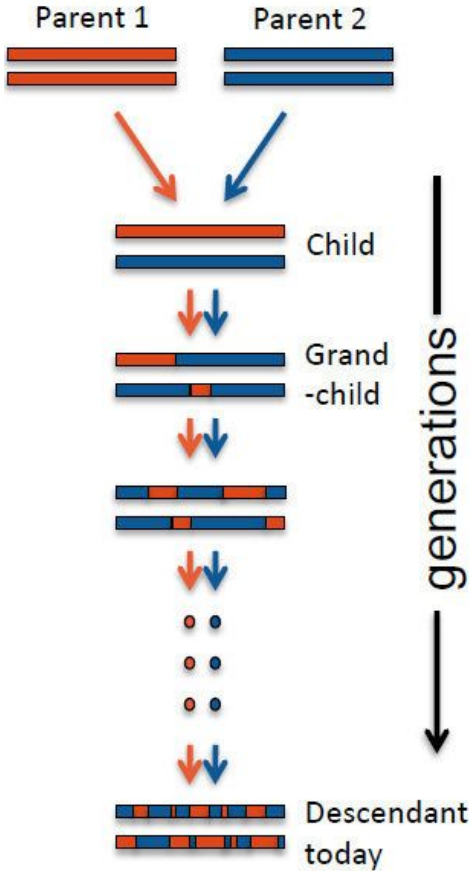




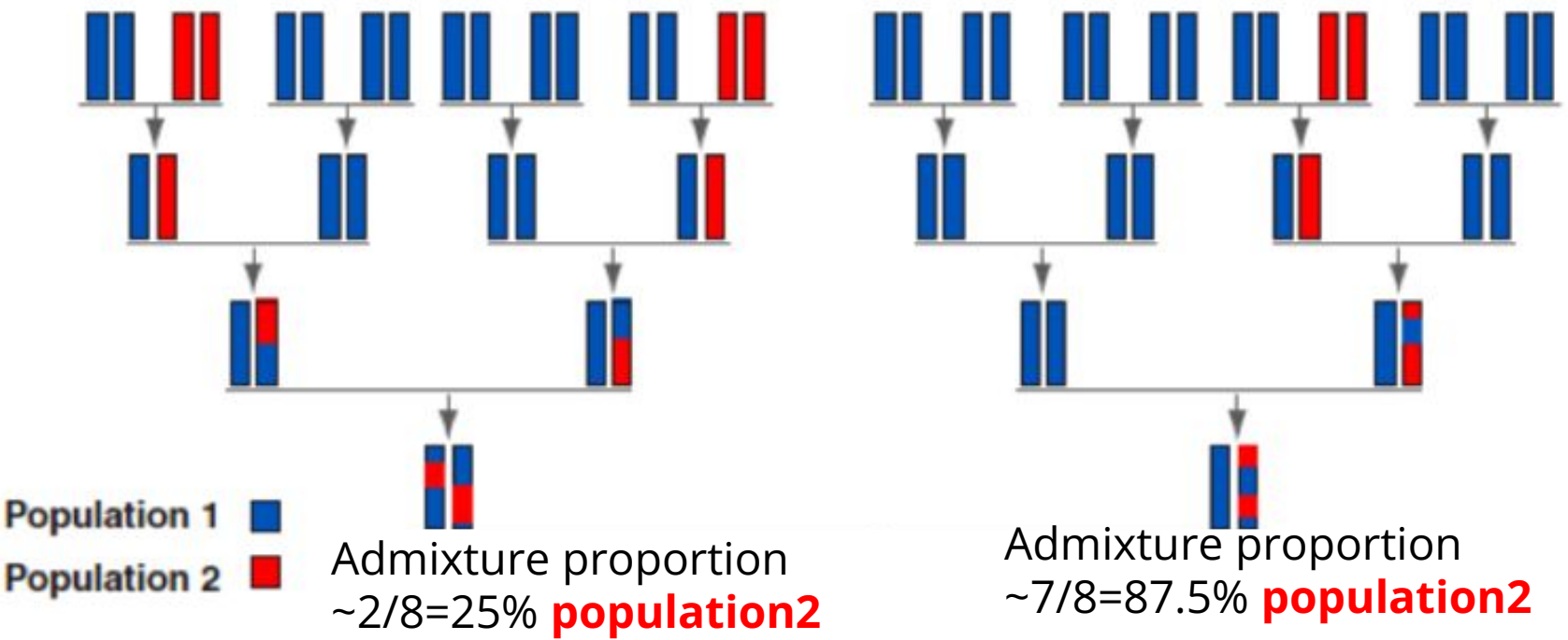
Admixture - chromosome level



Admixture - chromosome level



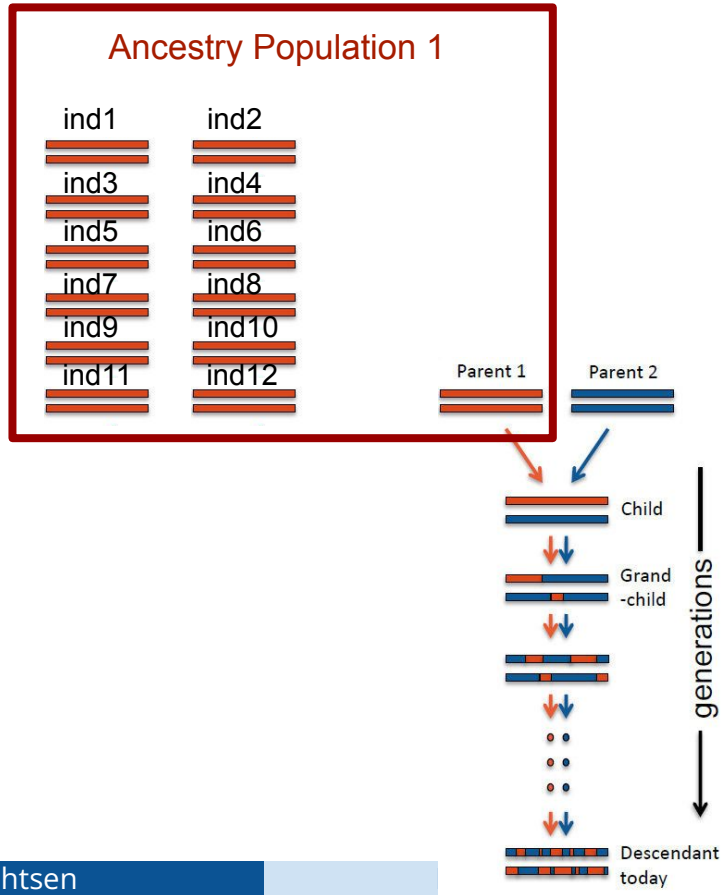
Proportions and ancestors



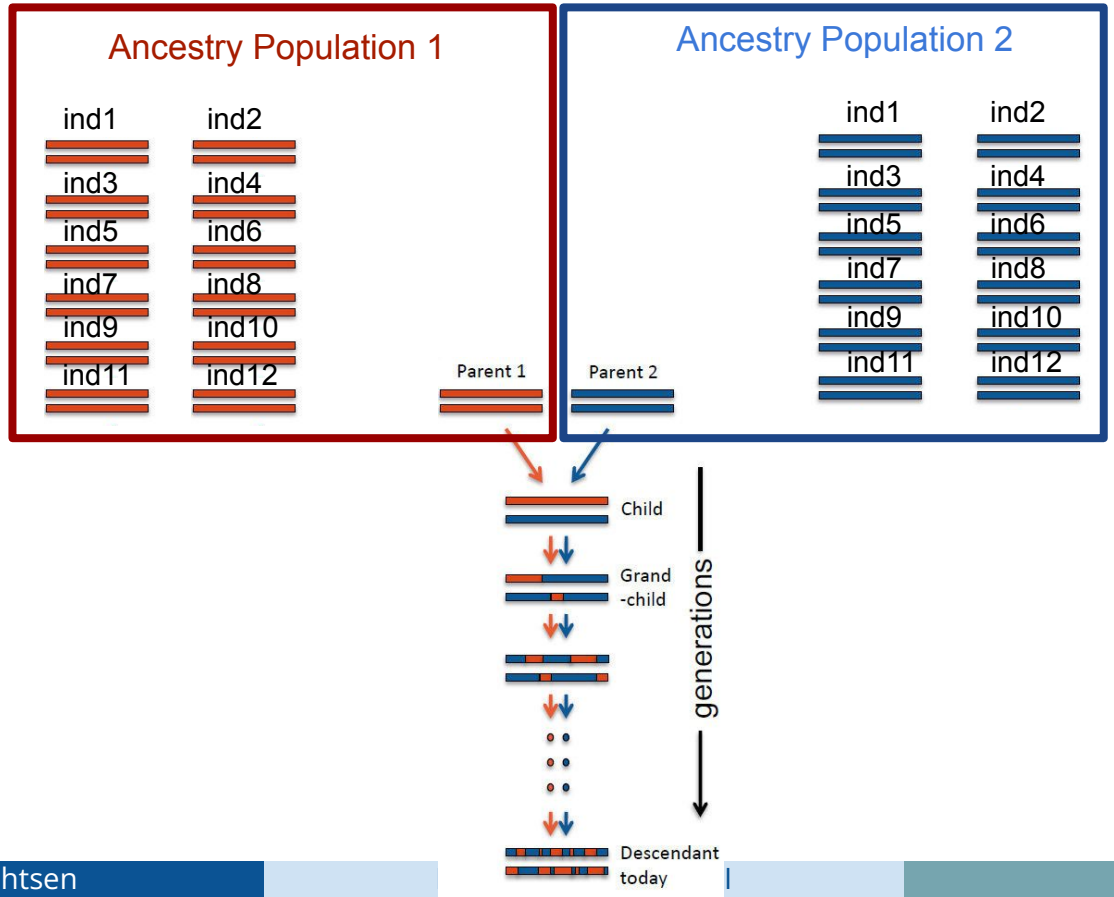
Admixture proportion
~2/8=25% **population2**

Admixture proportion
~7/8=87.5% **population2**

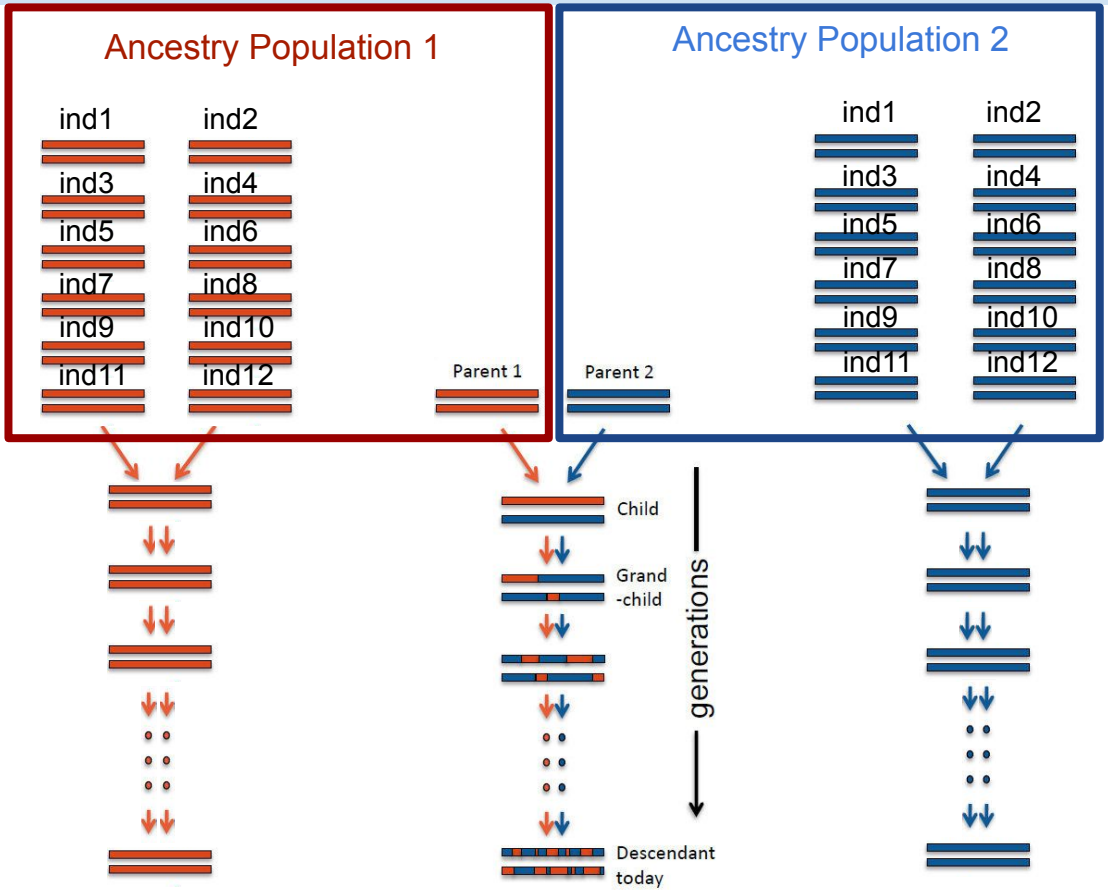
Admixture - chromosome level



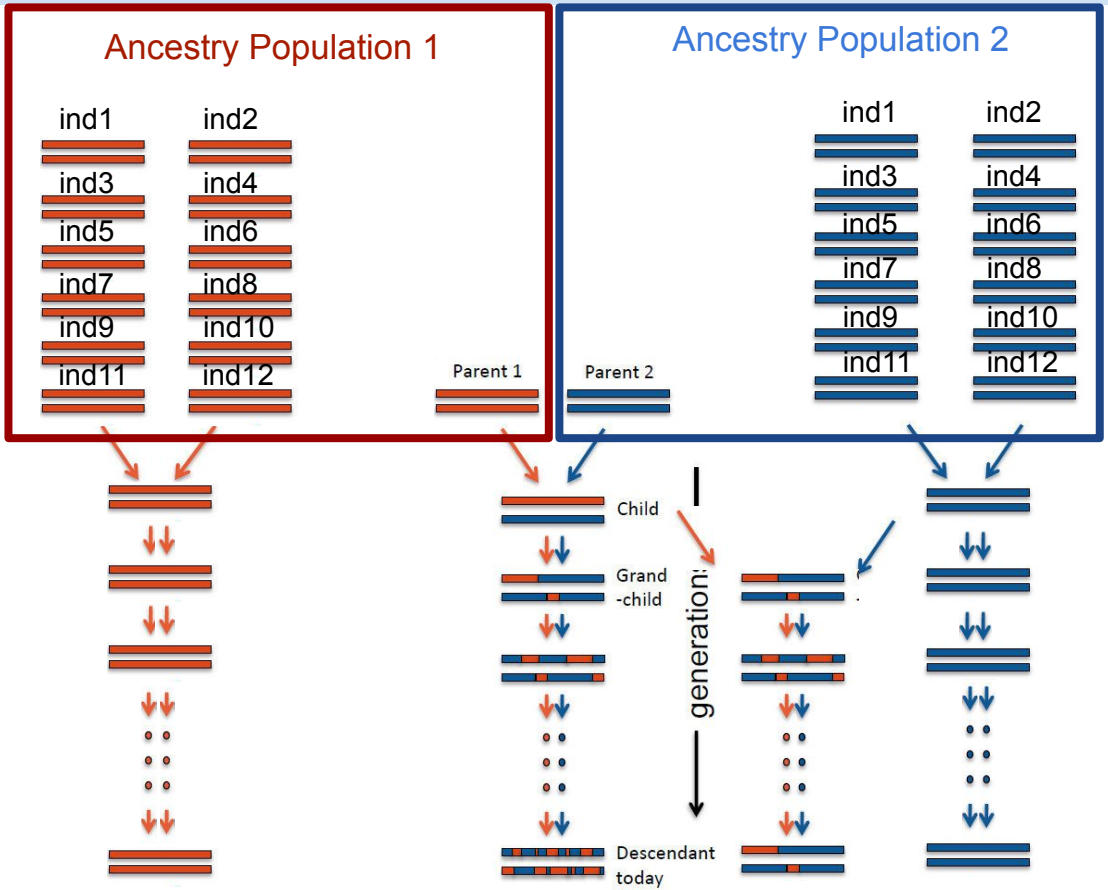
Admixture - chromosome level



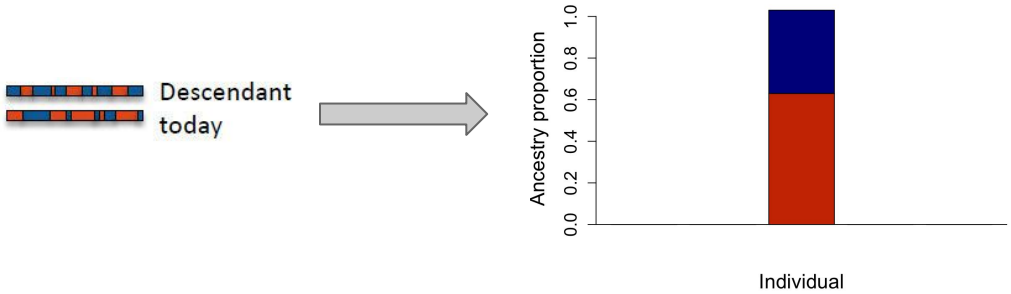
Admixture - chromosome level



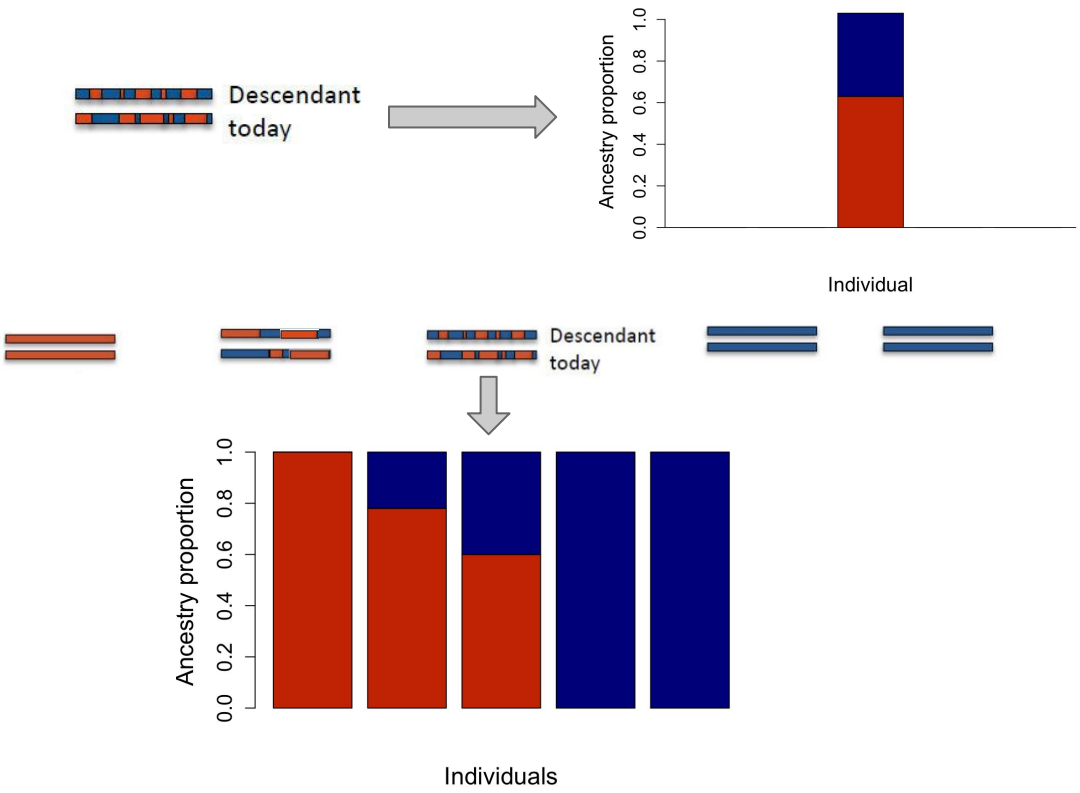
Admixture - chromosome level



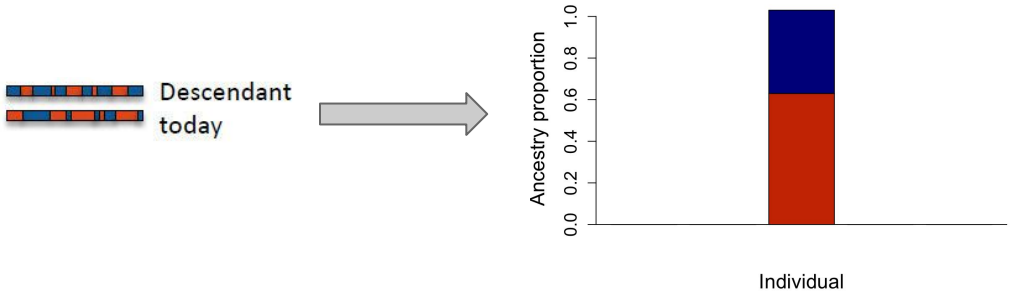
Chromosome level to Global level

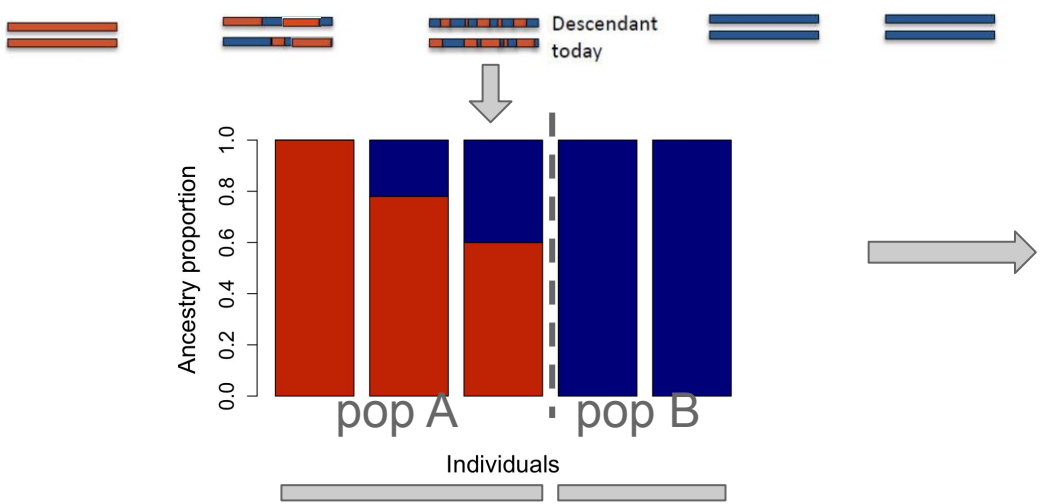


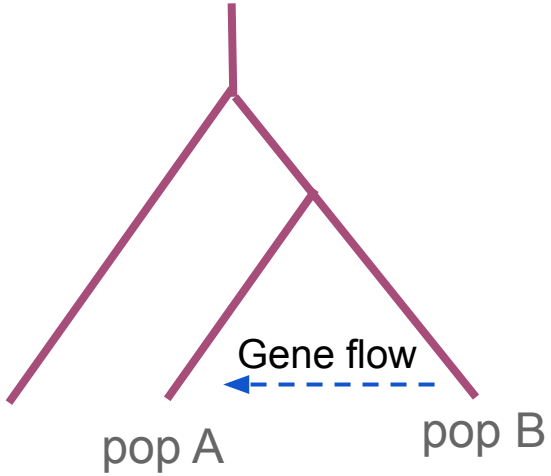
Chromosome level to Global level



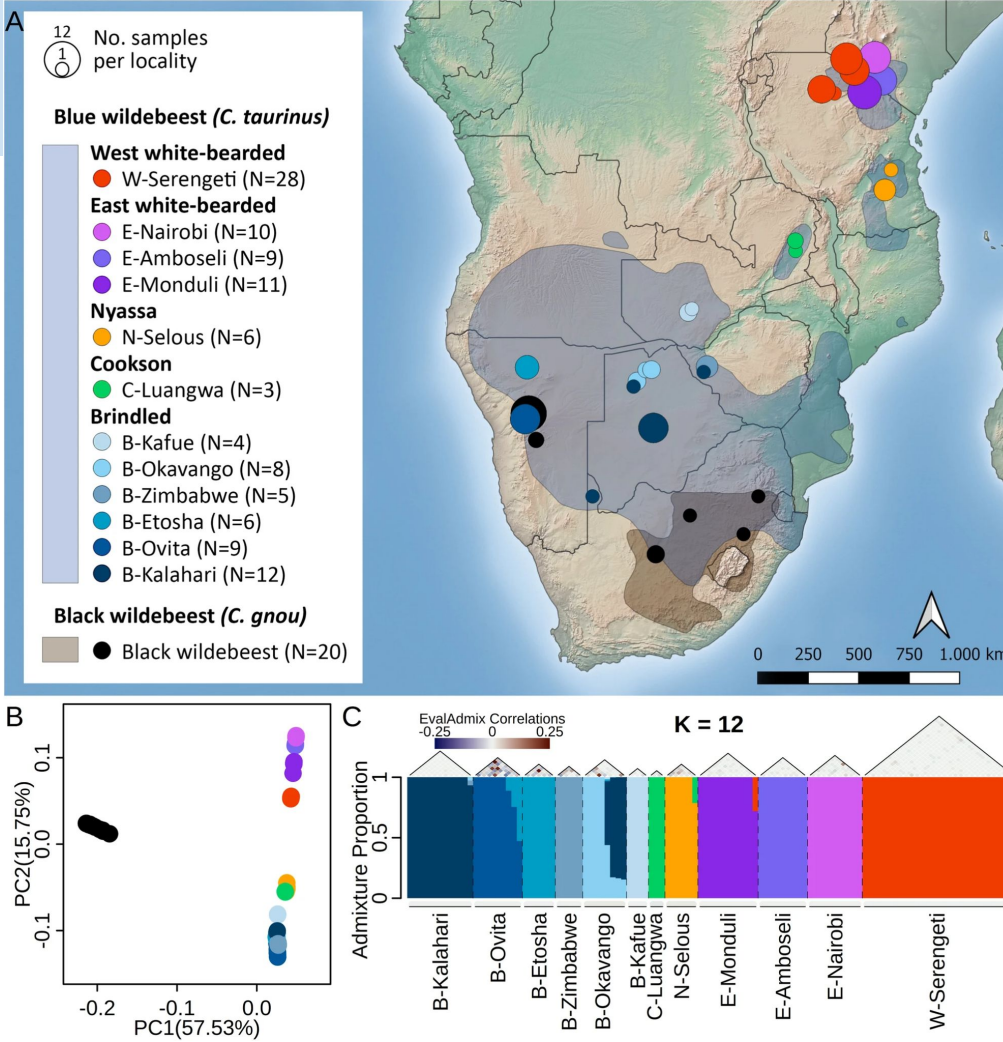
Inference from admixture proportion





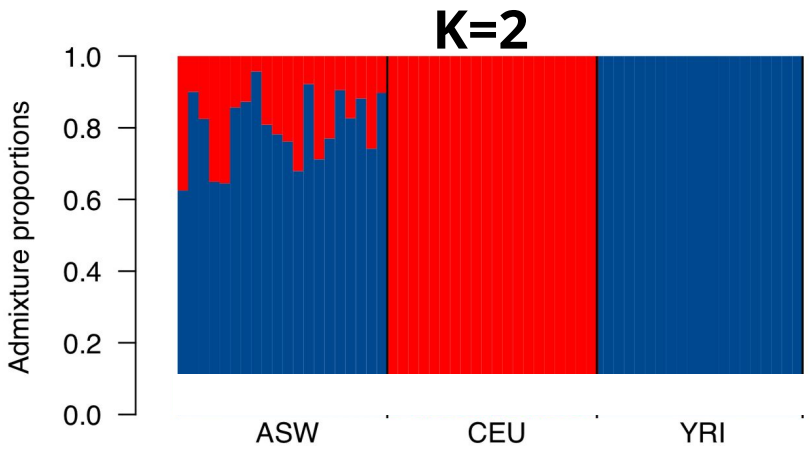


Real example



How to estimate admixture proportions

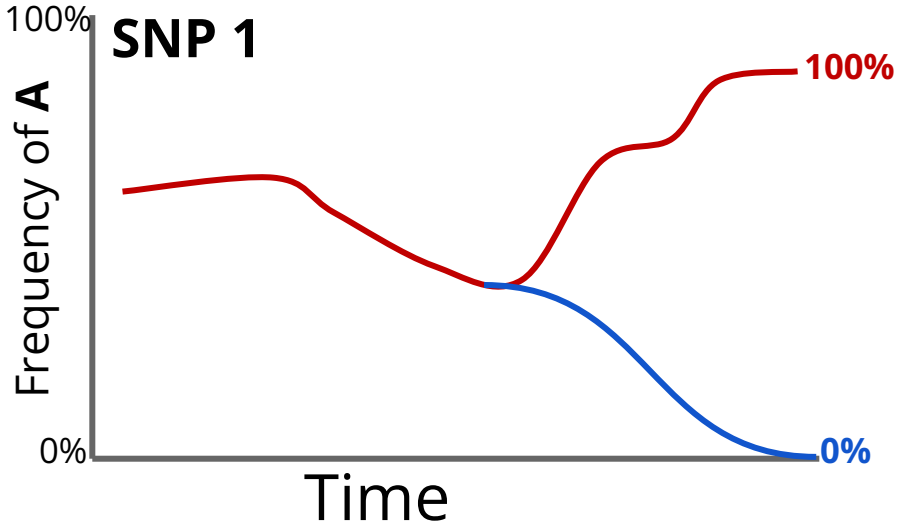
The admixture proportions, Q (n×K matrix):



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

Solution 2: ADMIXTURE

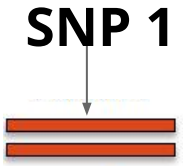
We exploit that allele frequencies differs



We genotype an individual

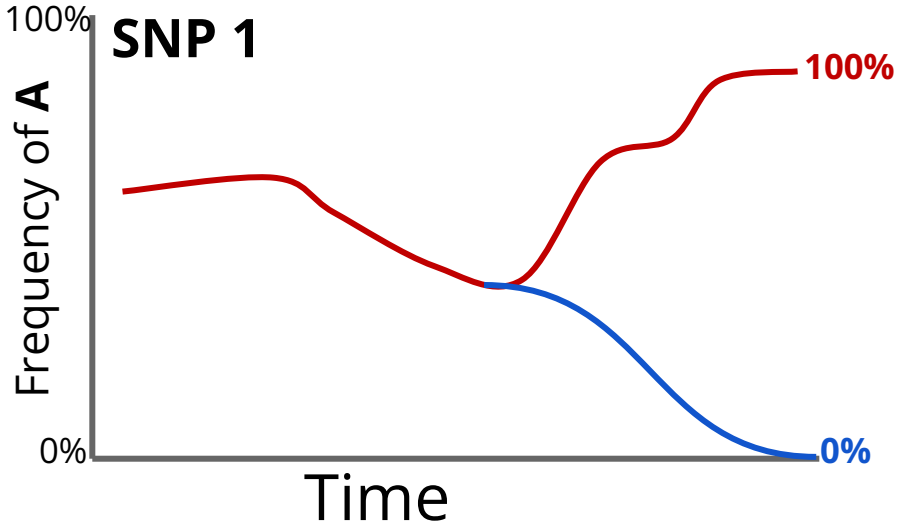
SNP 1 Genotype =AA

What can we say?



Solution 2: ADMIXTURE

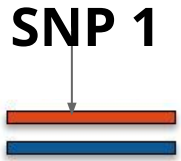
We exploit that allele frequencies differs



We genotype an individual

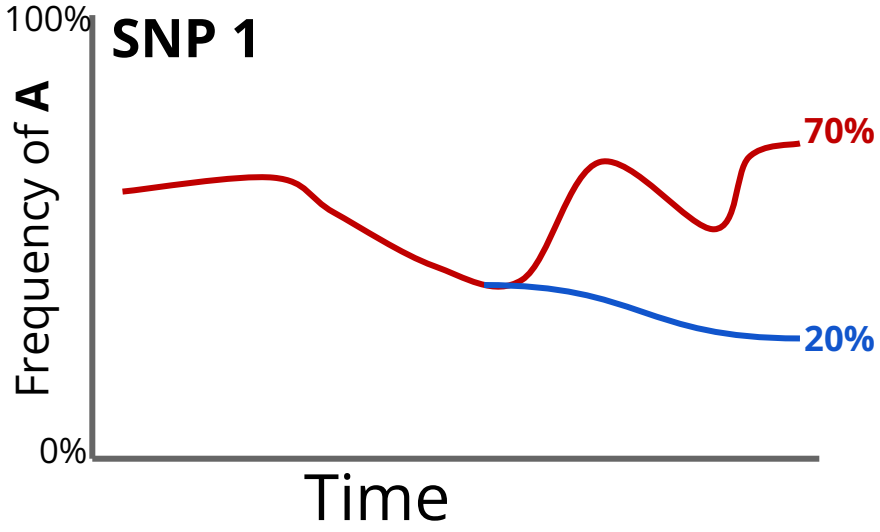
SNP 1 Genotype =AG

What can we say?



Solution 2: ADMIXTURE

We exploit that allele frequencies differs

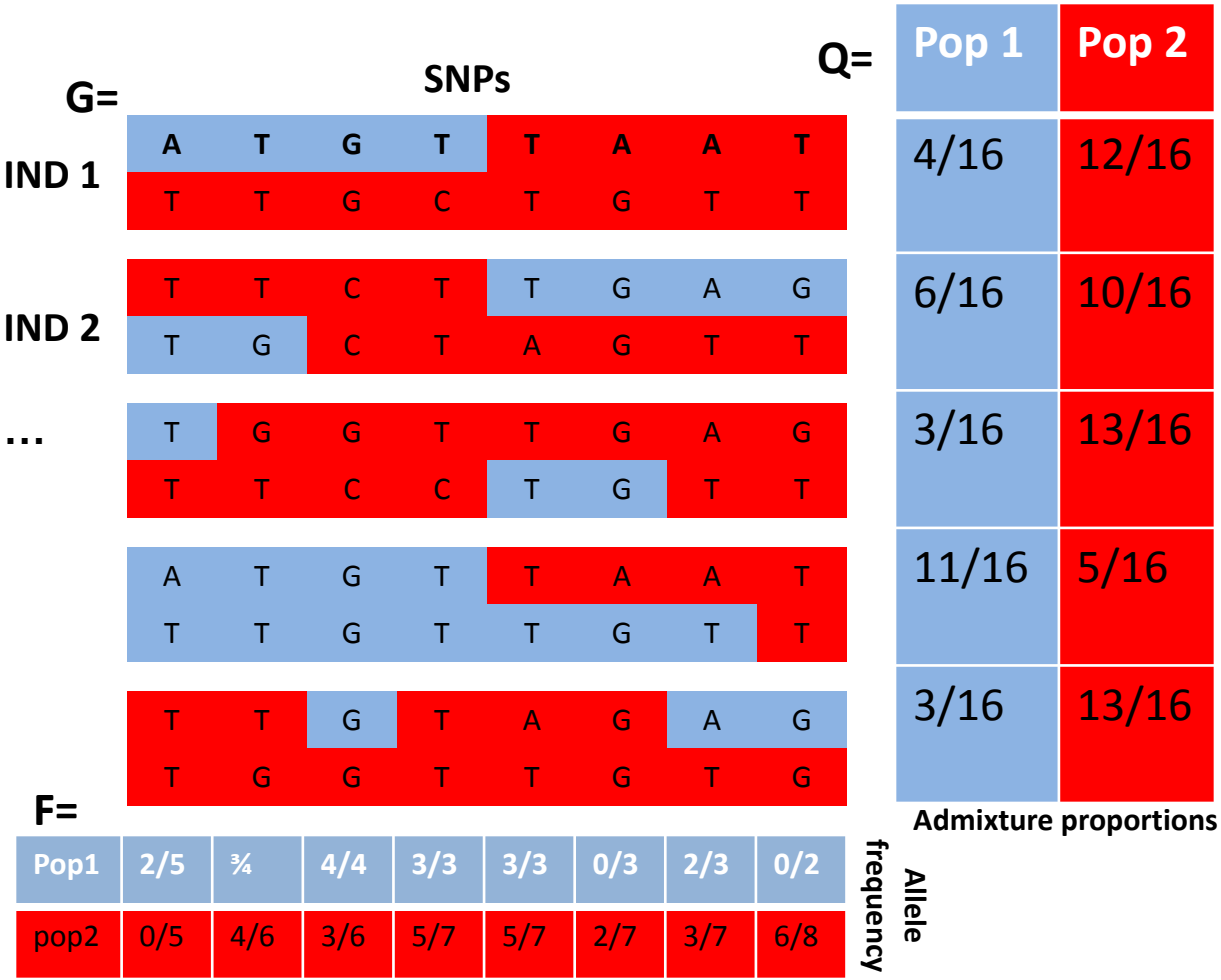


We genotype an individual

SNP 1 Genotype =AA

What can we say?

??


G=
SNPs
Q=
Pop 1
Pop 2
IND 1
A T G T T A A T
T T G C T G T T
4/16
12/16
IND 2
T T C T T G A G
T G C T A G T T
6/16
10/16
...
T G G T T G A G
T T C C T G T T
3/16
13/16
A T G T T A A T
T T G T T G T T
11/16
5/16
T T G T A G A G
T G G T T G T G
3/16
13/16
Admixture proportions
F=
Pop1 2/5 ¾ 4/4 3/3 3/3 0/3 2/3 0/2
pop2 0/5 4/6 3/6 5/7 5/7 2/7 3/7 6/8
Allele frequency

For one of i's alleles:

$$p(allele|Q^i, F^j) = q_{i1} f_{j1} + q_{i2} f_{j2} + \ldots q_{ik} f_{jk} = h_{ij}$$

$$p(G_{ij}|Q_i, F_j) = \begin{cases} (h_{ij})^2 & \text{if } G_{ij} = 2, \\ 2h_{ij}(1 - h_{ij}) & \text{if } G_{ij} = 1, \\ (1 - h_{ij})^2 & \text{if } G_{ij} = 0. \end{cases}$$

$$p(G|Q, F) = \prod_i^N \prod_j^M p(G_{ij}|Q_i, F_j)$$

Q=

Pop 1	Pop 2
4/16	12/16
6/16	10/16
3/16	13/16
11/16	5/16
3/16	13/16

Admixture proportions

F=

Pop1	2/5	¾	4/4	3/3	3/3	0/3	2/3	0/2
pop2	0/5	4/6	3/6	5/7	5/7	2/7	3/7	6/8

Allele frequency

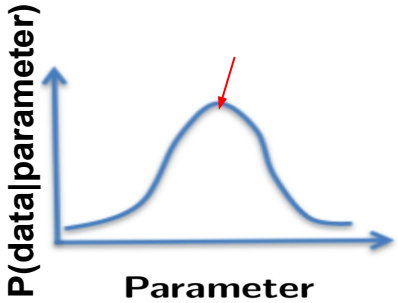
Solution 2: ADMIXTURE

A model with frequencies (F) and admixture proportions (Q) as parameters and Genotype (G) as the data

It is a maximum likelihood method

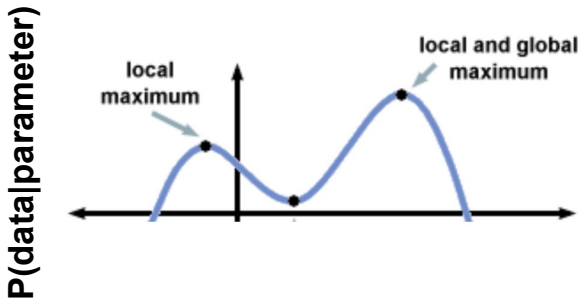
- This means it in principle provides the parameter values that make the observed data the most likely to observe

$$\hat{F}, \hat{Q} = argmax_{F,Q} P(G|F,Q)$$



Things to be aware of with ADMIXTURE

Uses numerical optimization because there is not an analytical solution!



- So you need to run it several times to be able to assess whether it has found the maximum likelihood estimate (“reached convergence”)!

Solution: Assessing convergence

1. Run it ~10 times and sort them according to their likelihood:

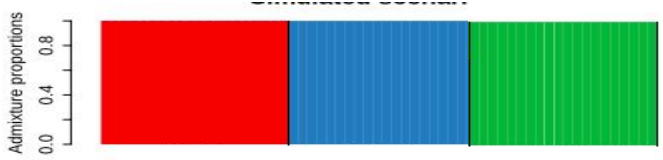
```
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles15K3.log -3865964.394787
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles18K3.log -3865964.414354
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles12K3.log -3865964.414777
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles5K3.log -3865964.415840
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles3K3.log -3865964.415979
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles7K3.log -3865964.419814
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles6K3.log -3865964.422308
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles13K3.log -3865964.424335
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles10K3.log -3865964.427415
/home/ida/teaching/popgen19/admixexercise/NGSadmixoutput/myoutfiles9K3.log -3865964.428677
```

2. If the top 5 have very similar likelihood pick the top one
3. If not, then run it 10 more times and go to step 2 again

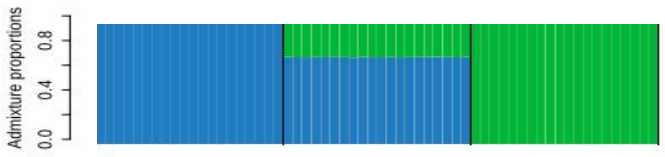
Things to be aware of with ADMIXTURE

Choose a K. If K is too little you risk getting wrong results e.g.

Truth is K=3

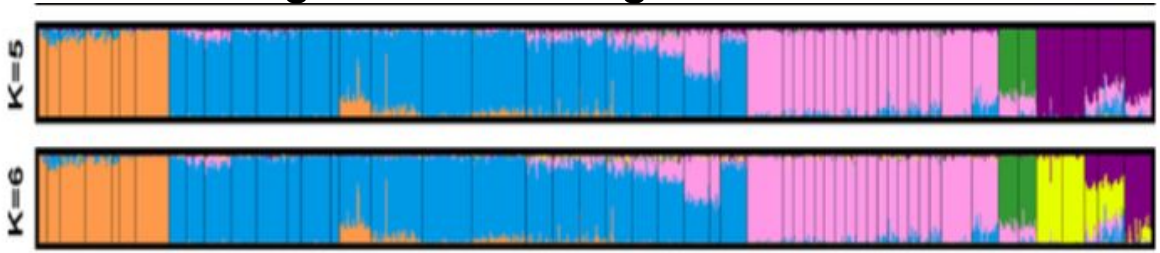


You choose K=2:



ADMIXTURE manual suggests to use cross-validation to find the “correct K”, but this approach has several issues, so I would never use that

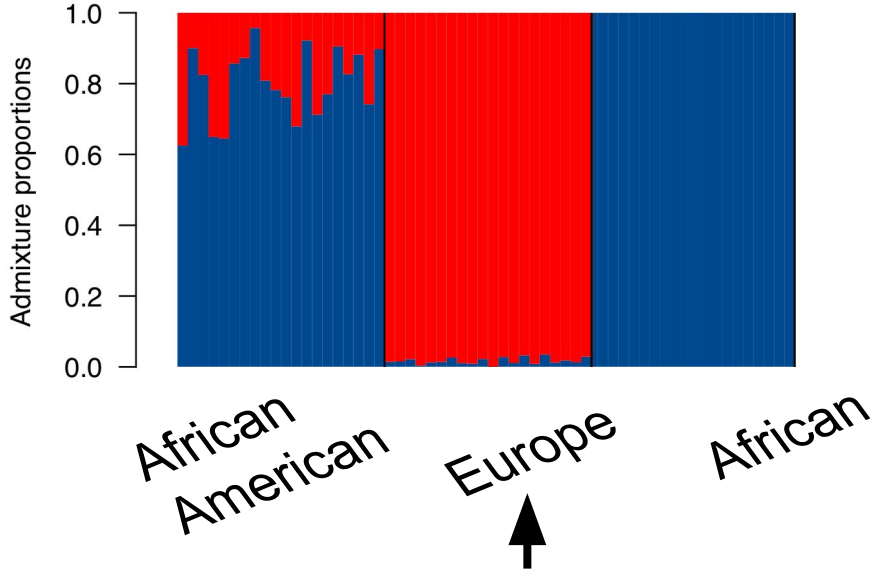
- Several Ks might be meaningful…



Things to be aware of with ADMIXTURE

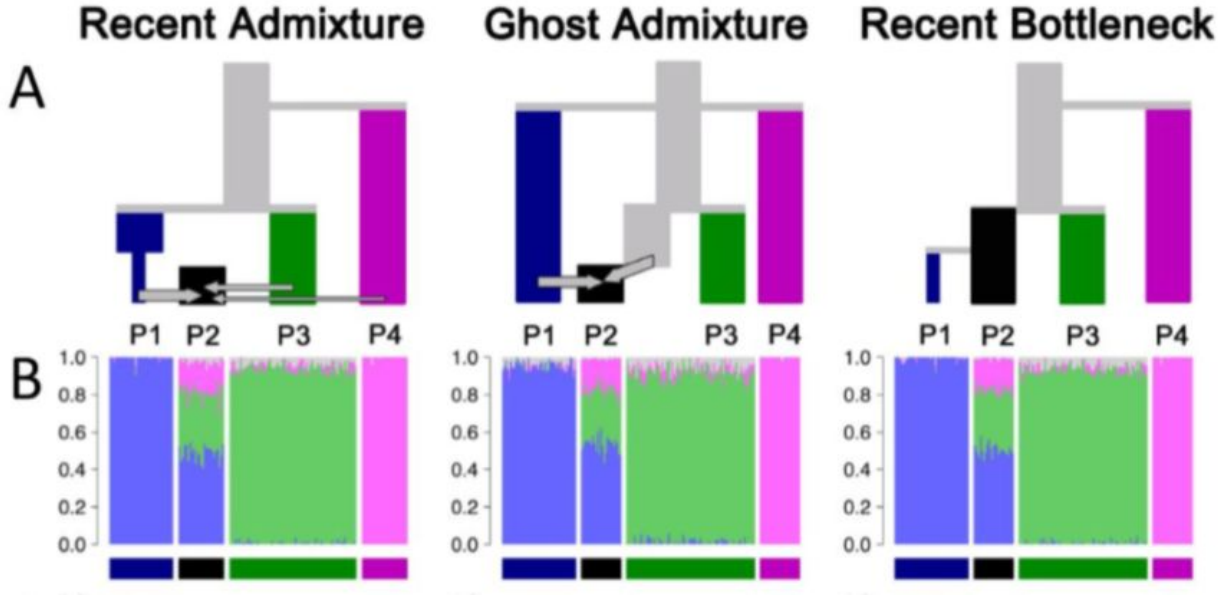
You may not have the actual source populations and that can affect both the estimates and the interpretation

E.g. what if we had samples from China and not Europe?



Things to be aware of with ADMIXTURE

4. Several demographic scenarios can lead to similar results!



Potential solution: evalAdmix

- There is a tool, **evalAdmix**, that can help assess the fit and suggest a lower bound for K
- For each individual i and locus j it looks at the difference between the observed genotype and the ADMIXTURE based estimate
- $$\hat{g}_{ij} = 2\sum_{k=1}^{K} q_{ik} f_{jk} \qquad r_{ij} = g_{ij} - \hat{g}_{ij}$$
- More specifically it calculates the mean correlation between these differences for each pair of individual a and b

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

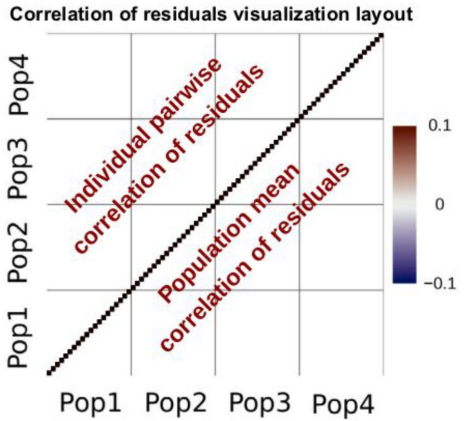
Potential solution issues: evalAdmix

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:

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



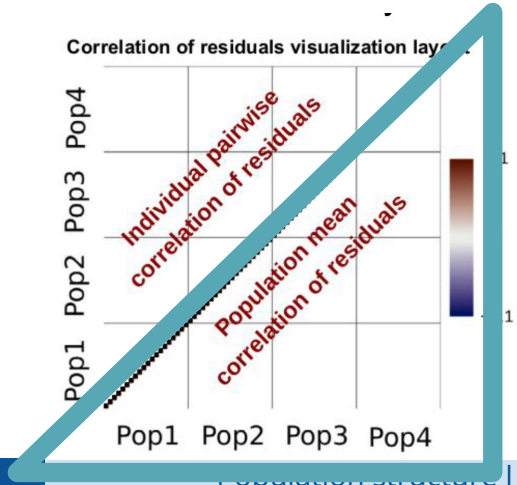
Potential solution issues: evalAdmix

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:

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

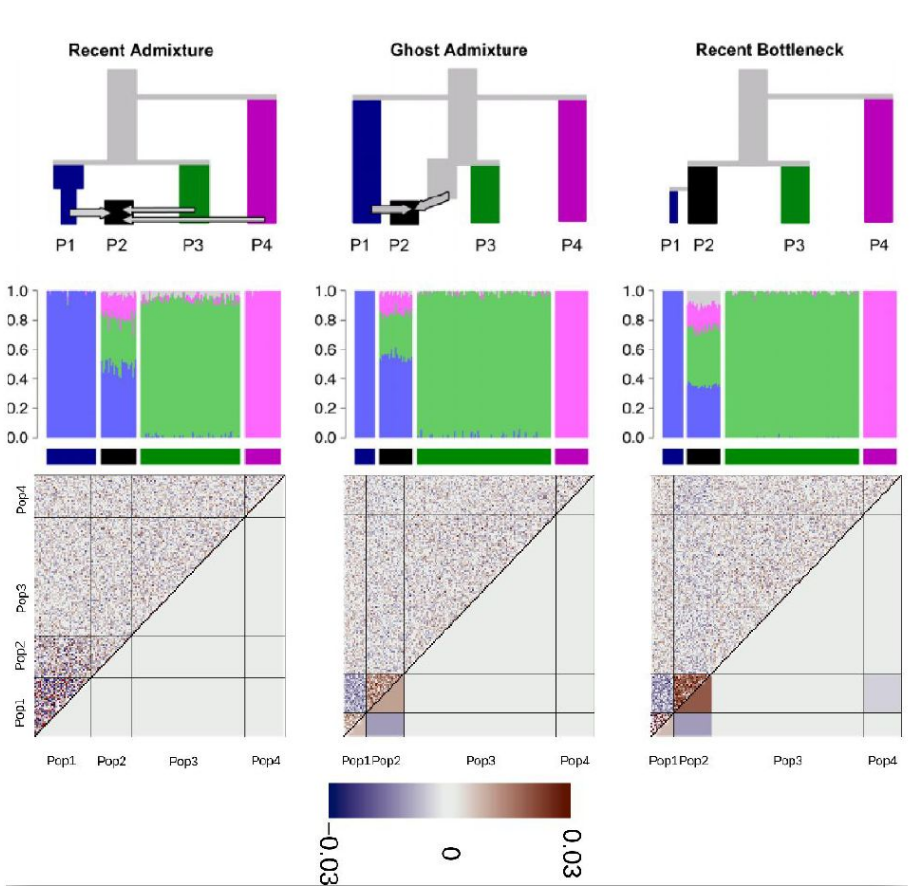
If individual *a* and *b* are sampled from the same population AND they have a bad model fit, they will share a systematic error and their residuals will be positively correlated:

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

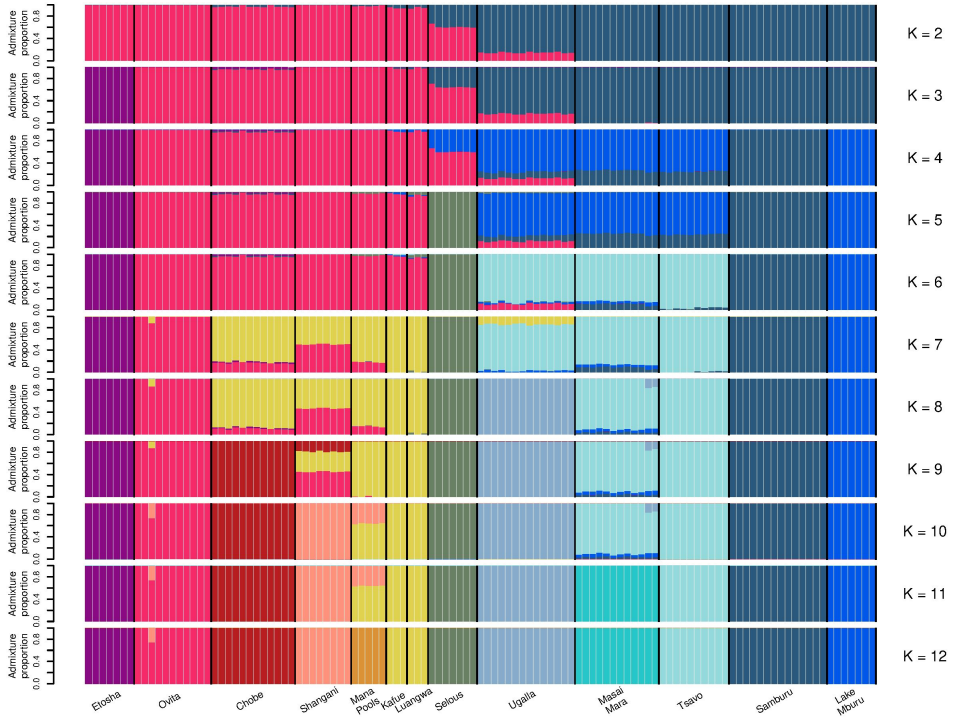


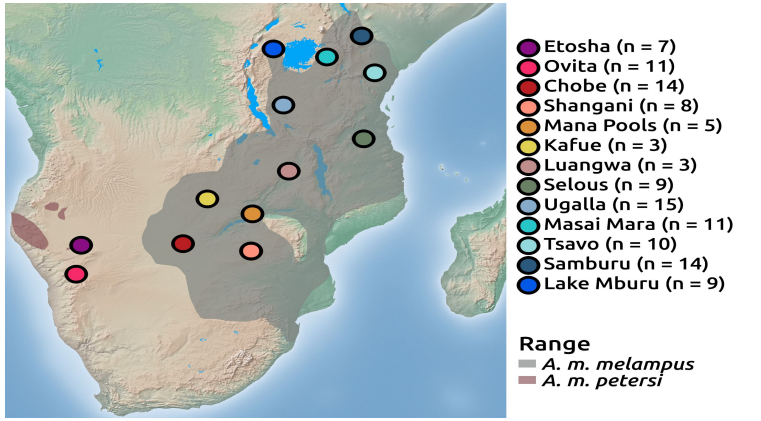
← Main focus

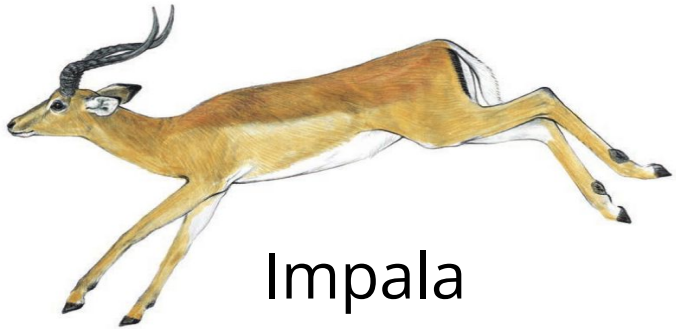
Potential solution issues: evalAdmix



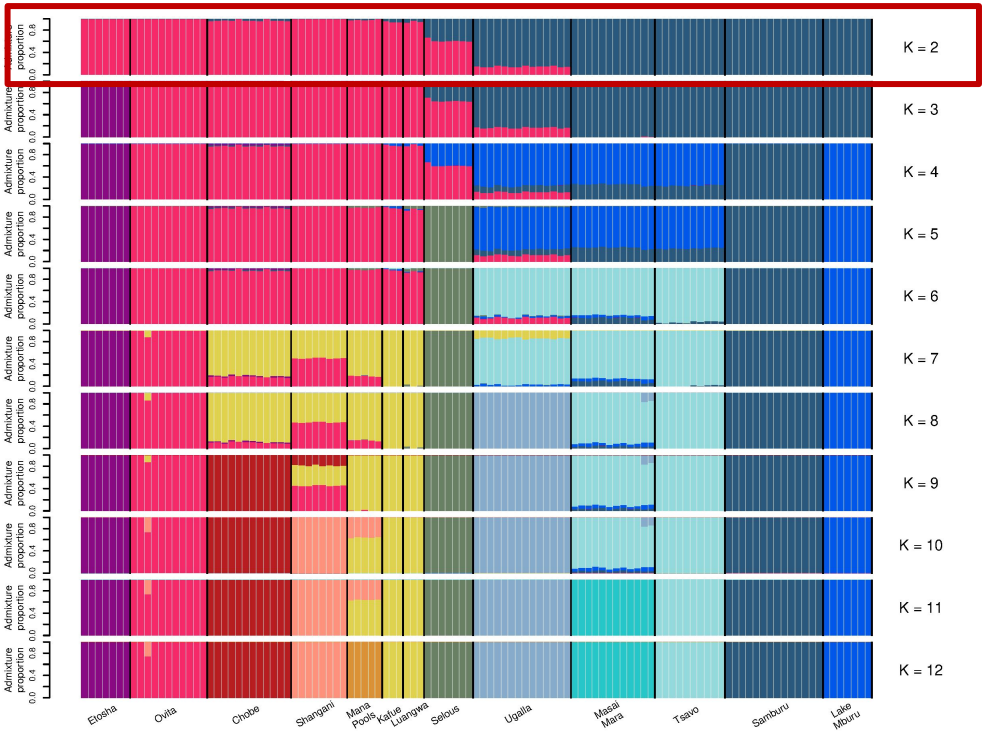
Evaluating the fit of admixture analyses: value of K

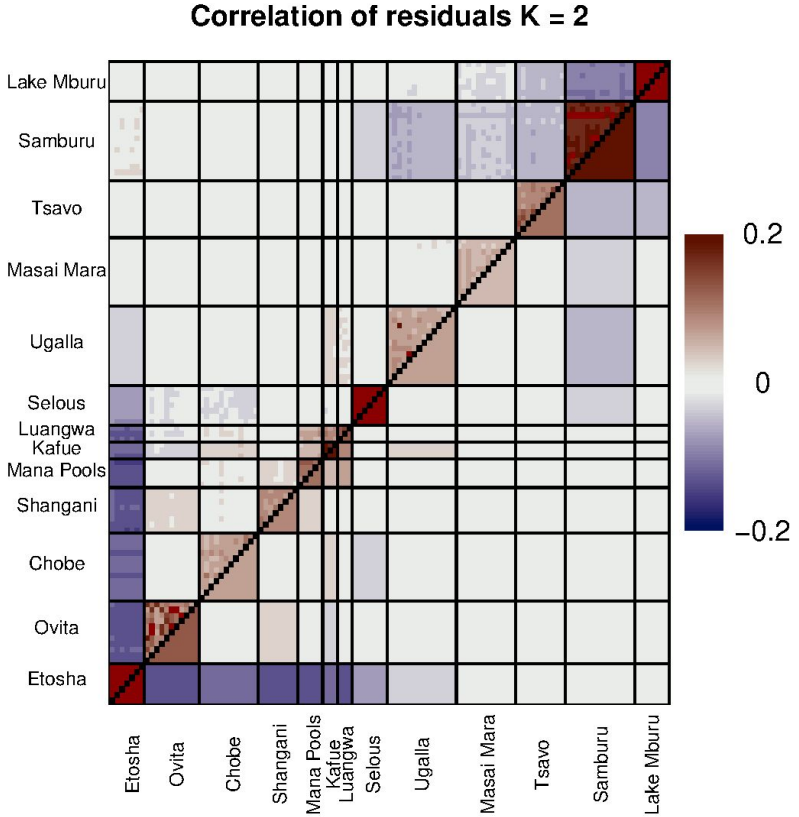




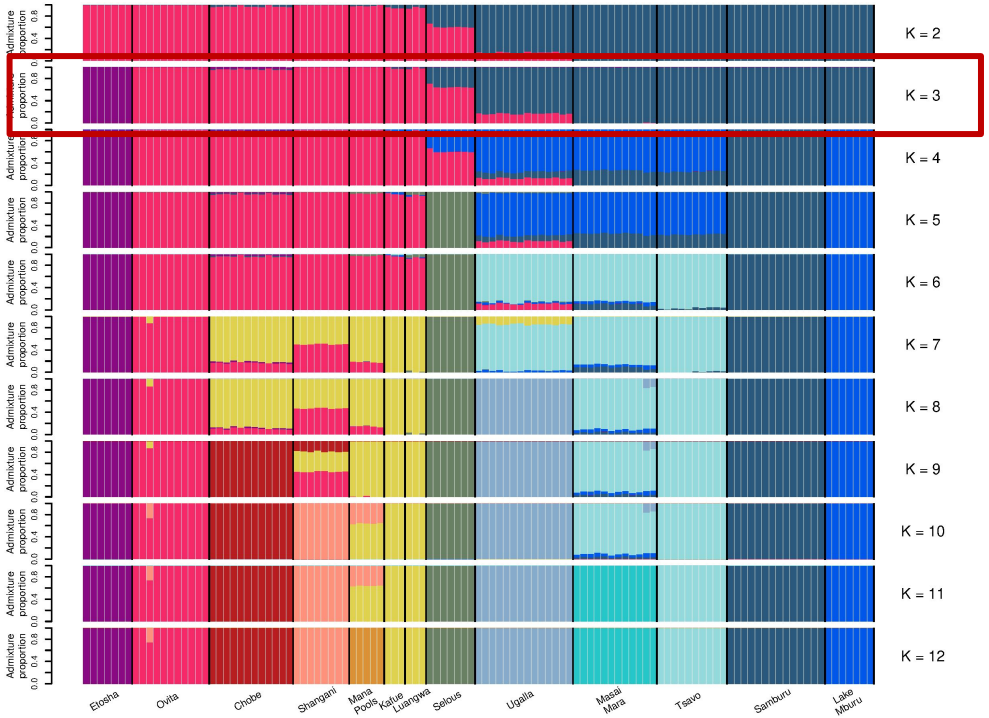


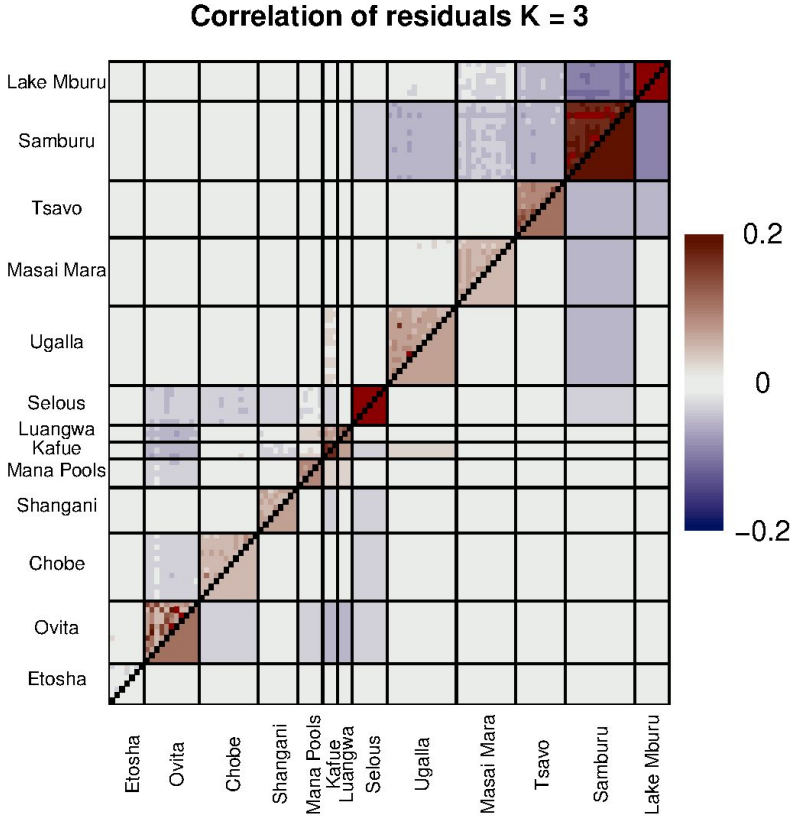
Impala


K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Admixture proportion
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

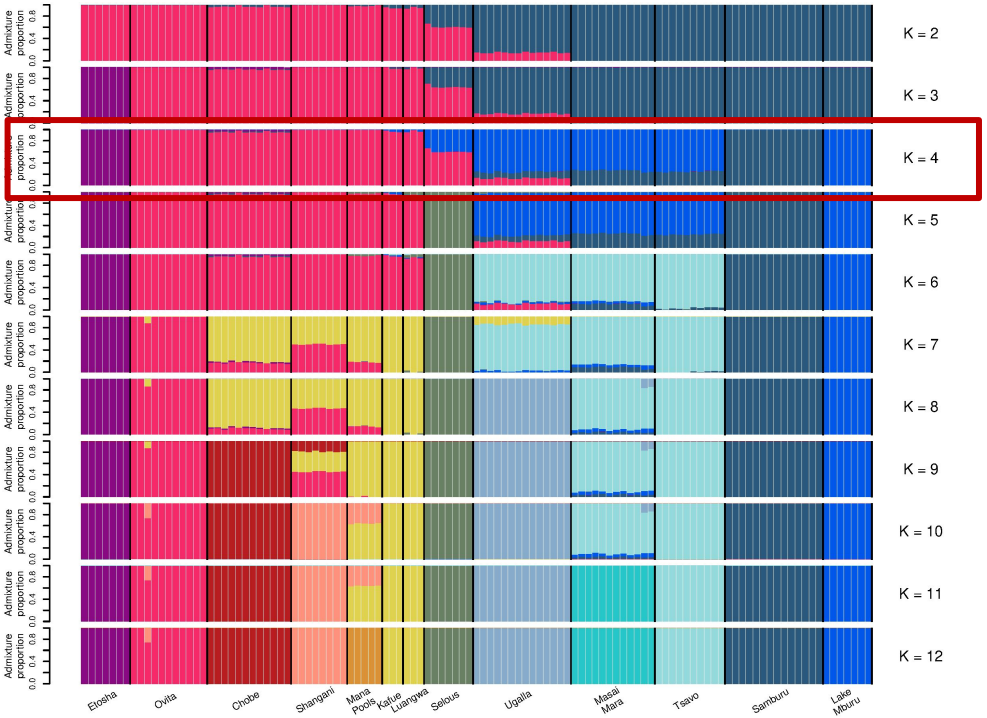

Correlation of residuals K = 2
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

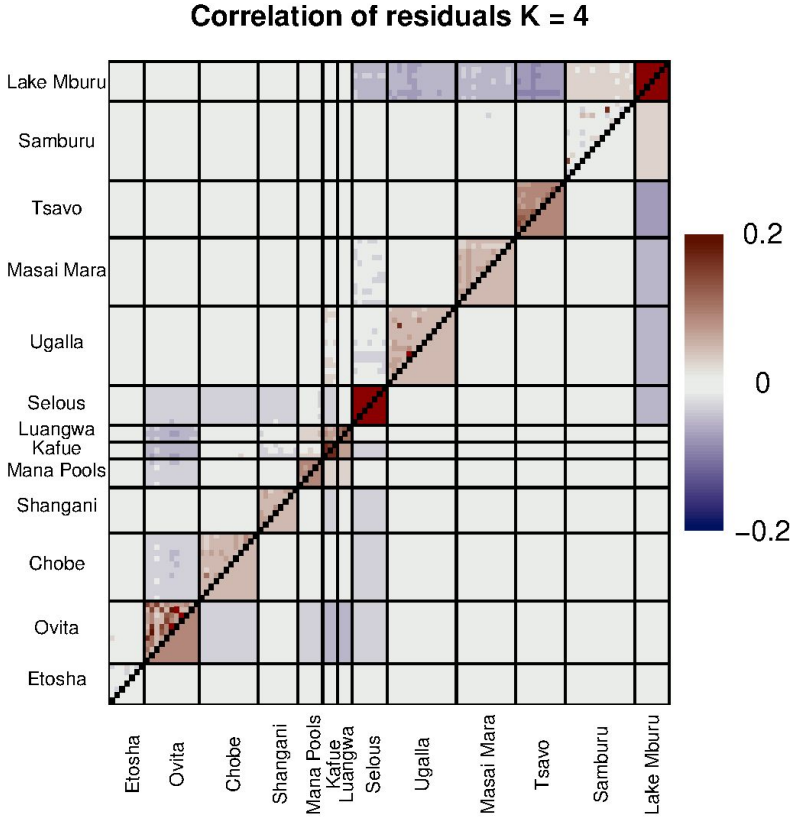
K=2


K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Admixture proportion
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

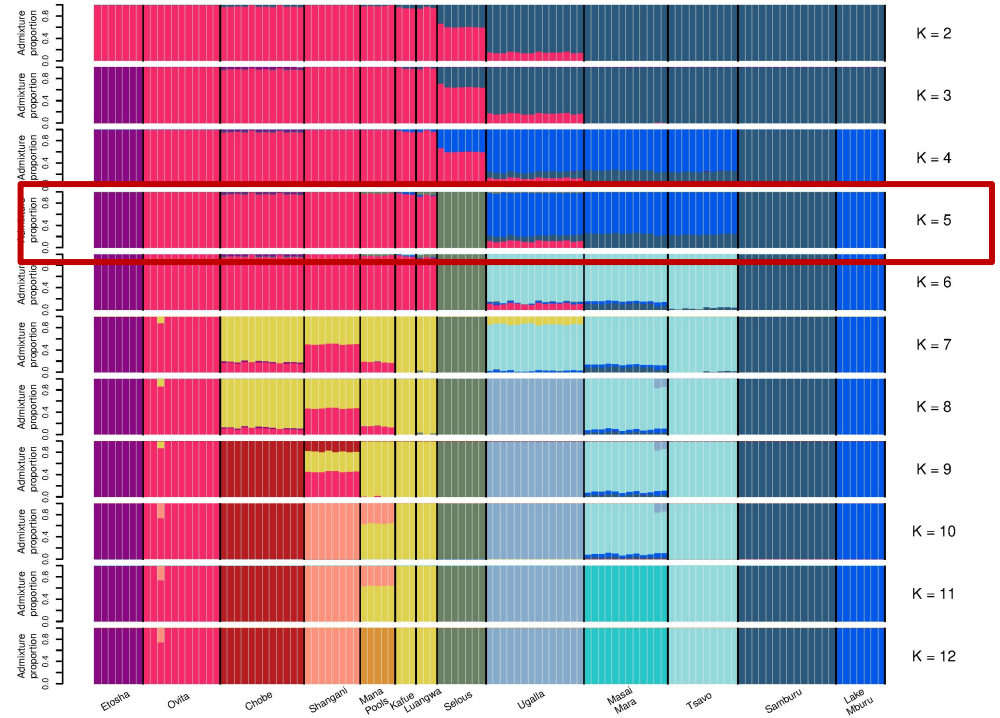

Correlation of residuals K = 3
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

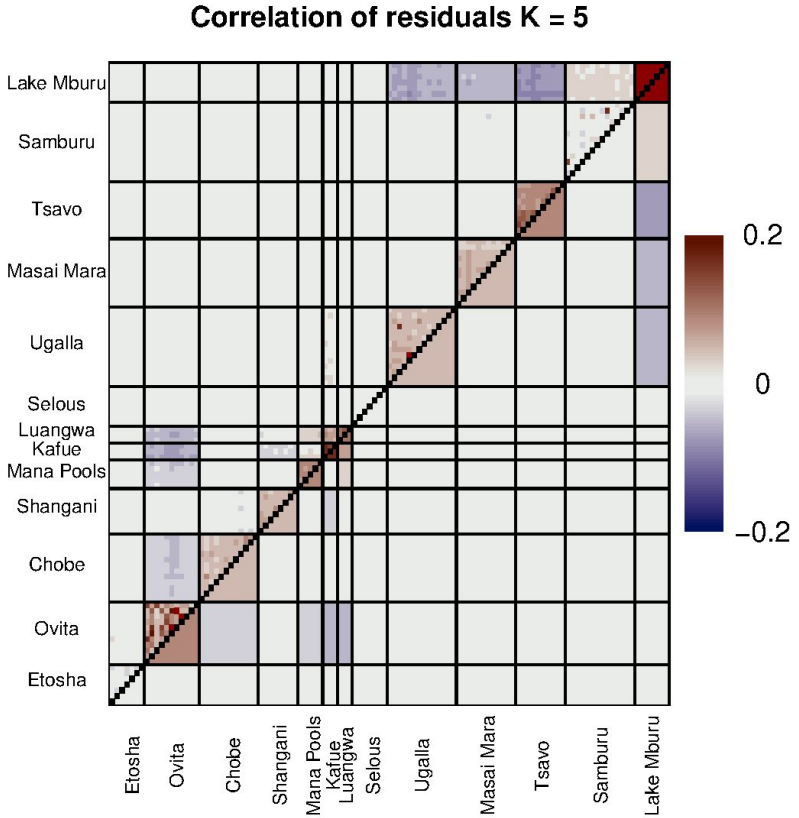
K=3


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

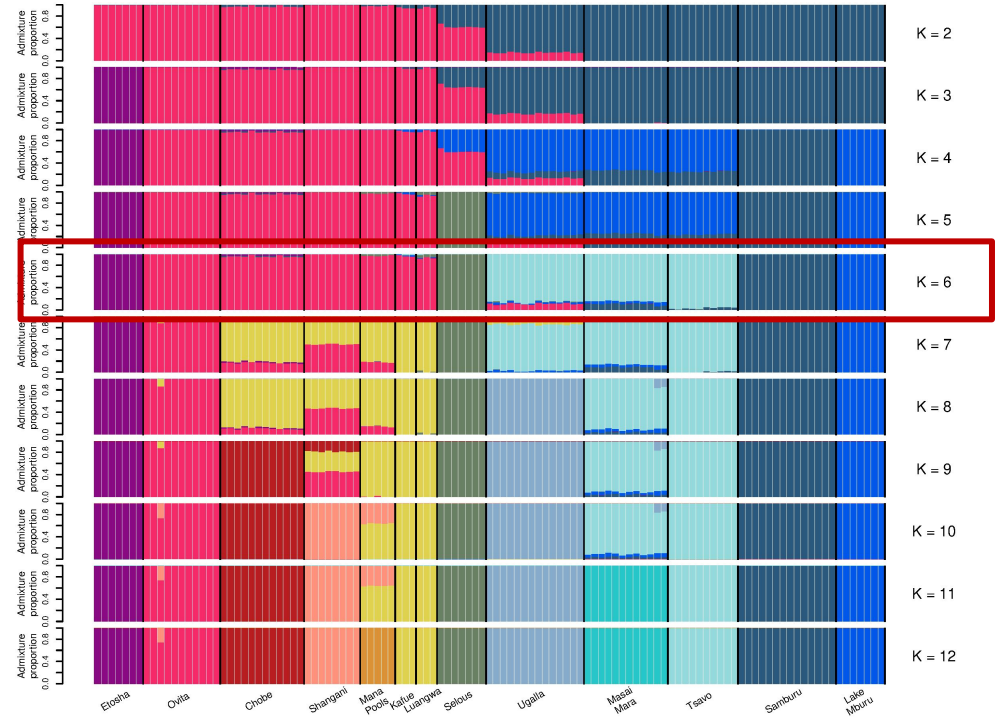

Correlation of residuals K = 4
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

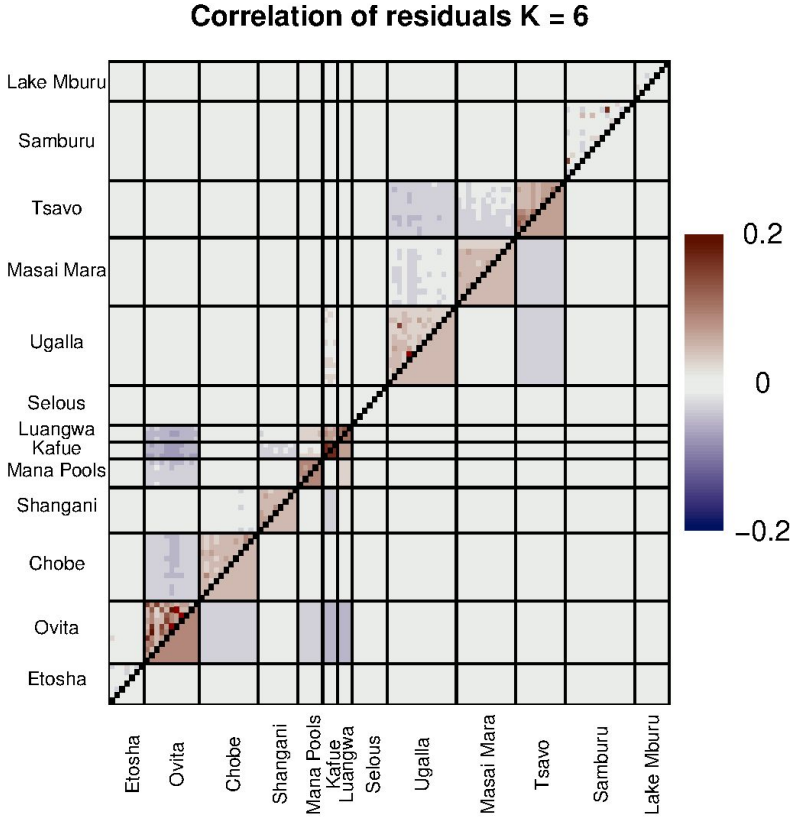
K=4


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

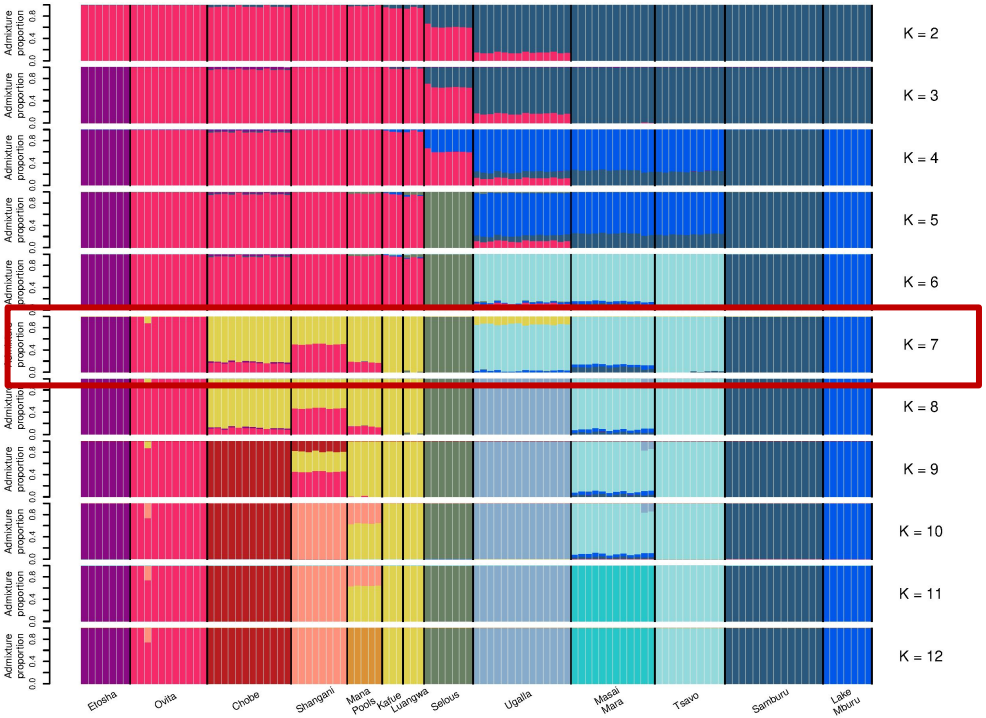

Correlation of residuals K = 5
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

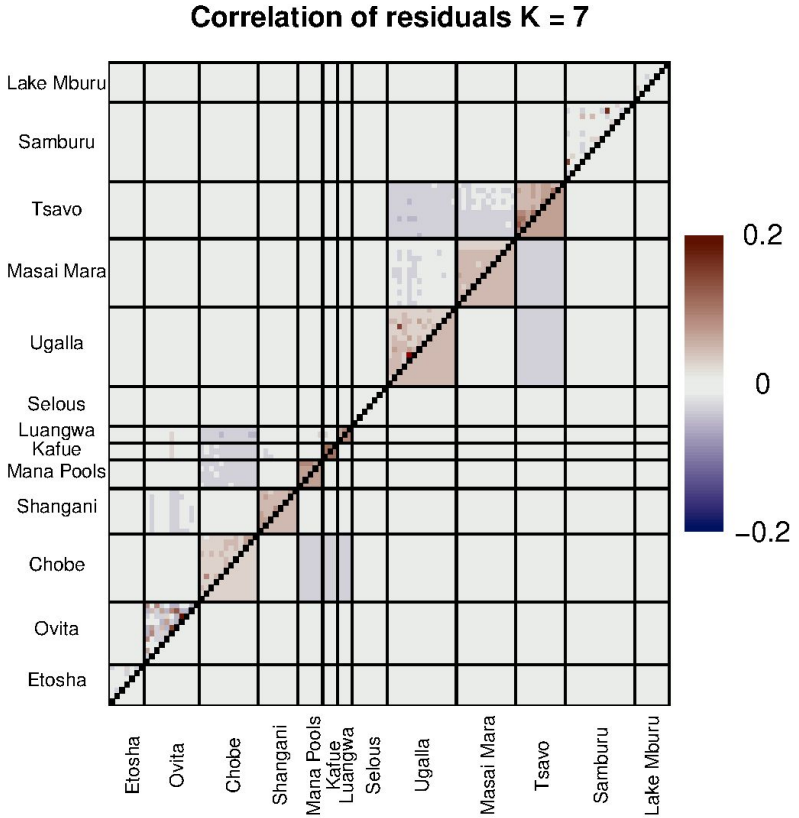
K=5


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

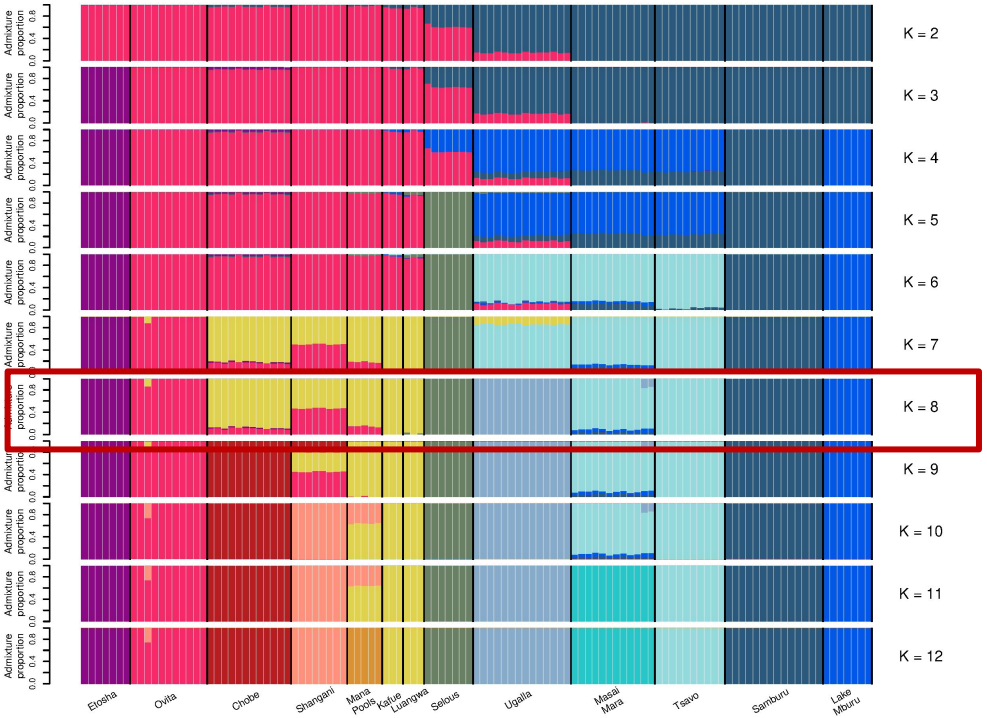

Correlation of residuals K = 6
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

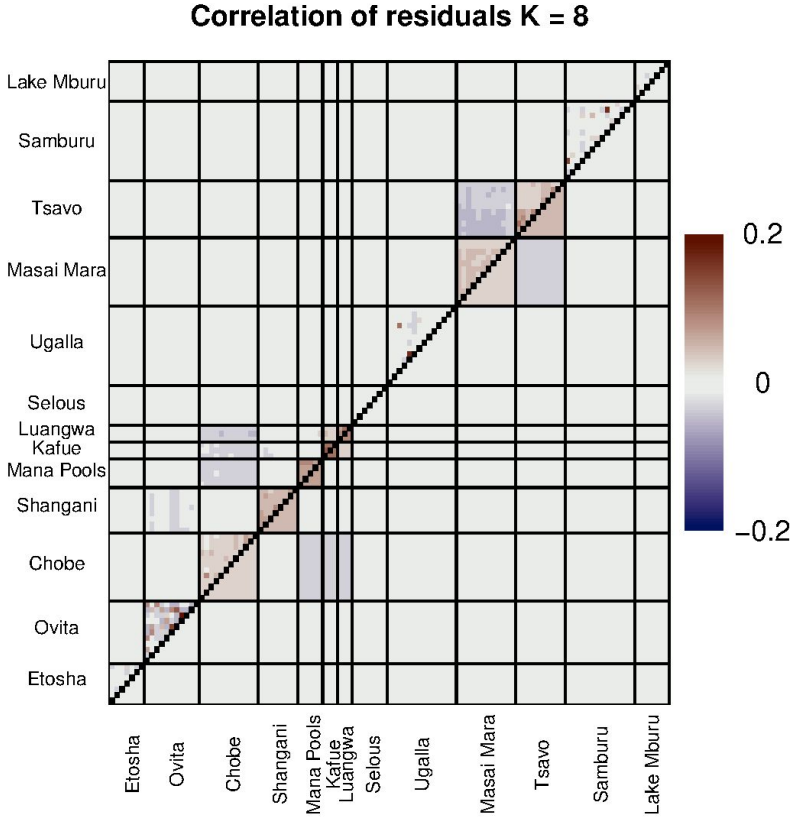
K=6


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

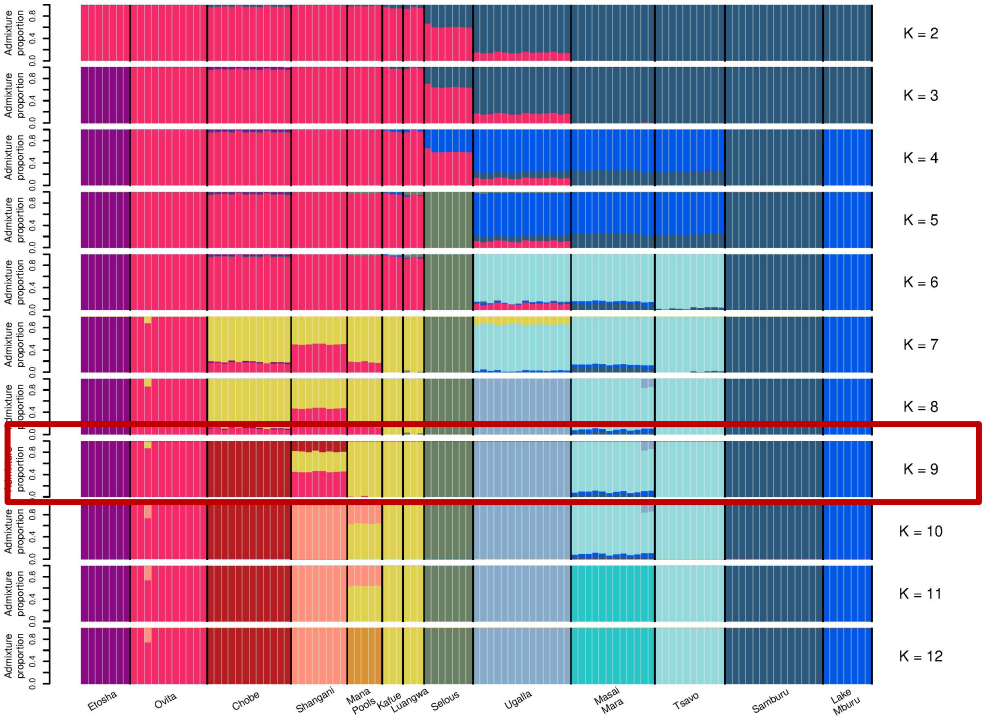

Correlation of residuals K = 7
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

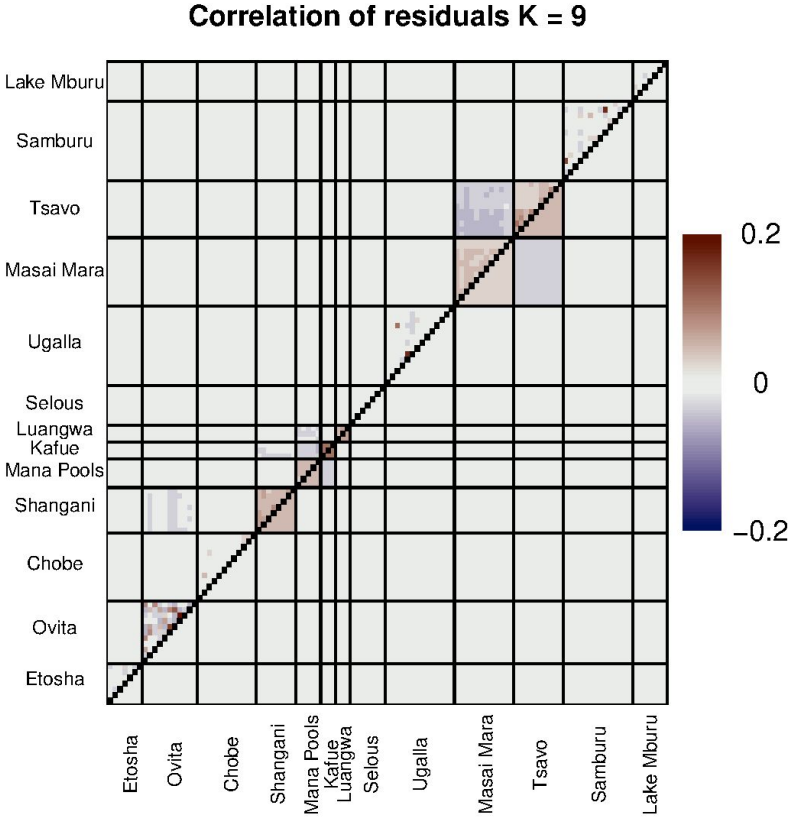
K=7


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

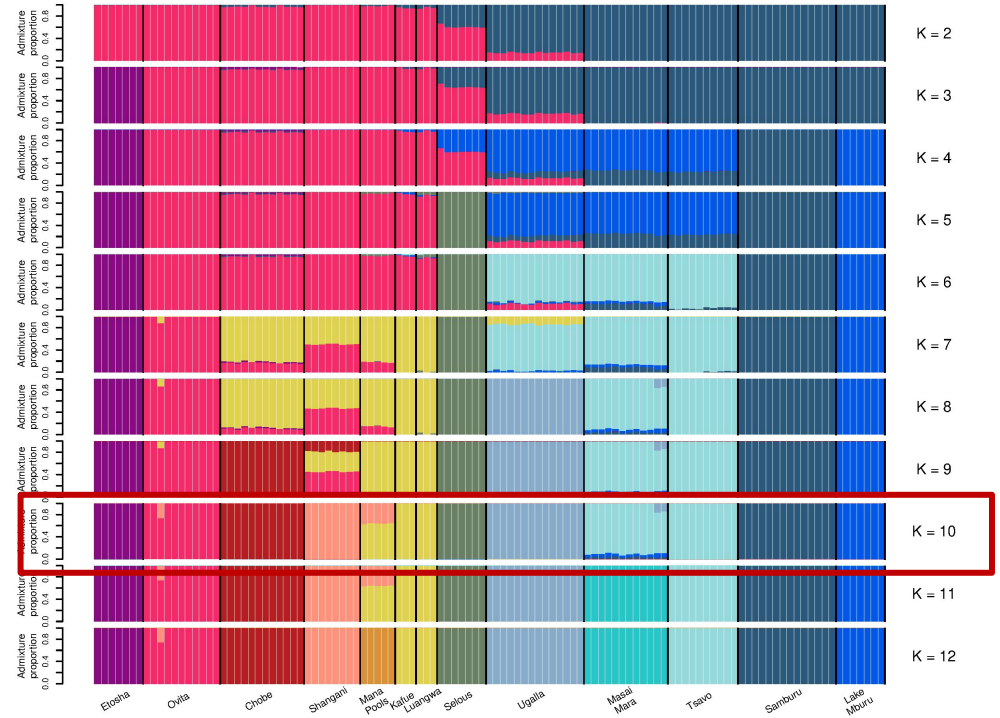

Correlation of residuals K = 8
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

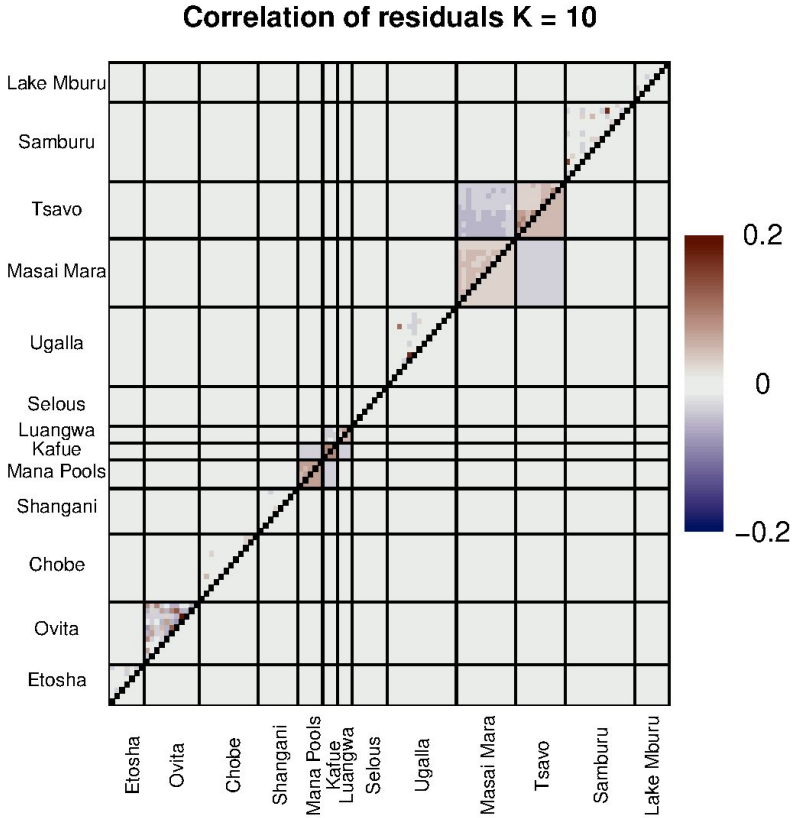
K=8


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

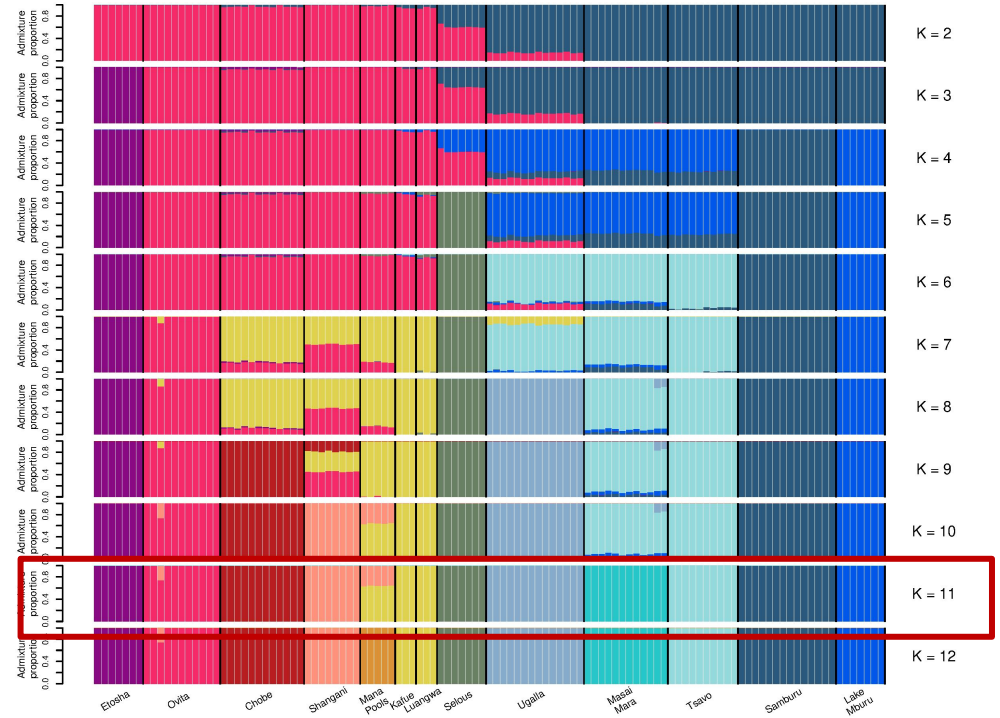

Correlation of residuals K = 9
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

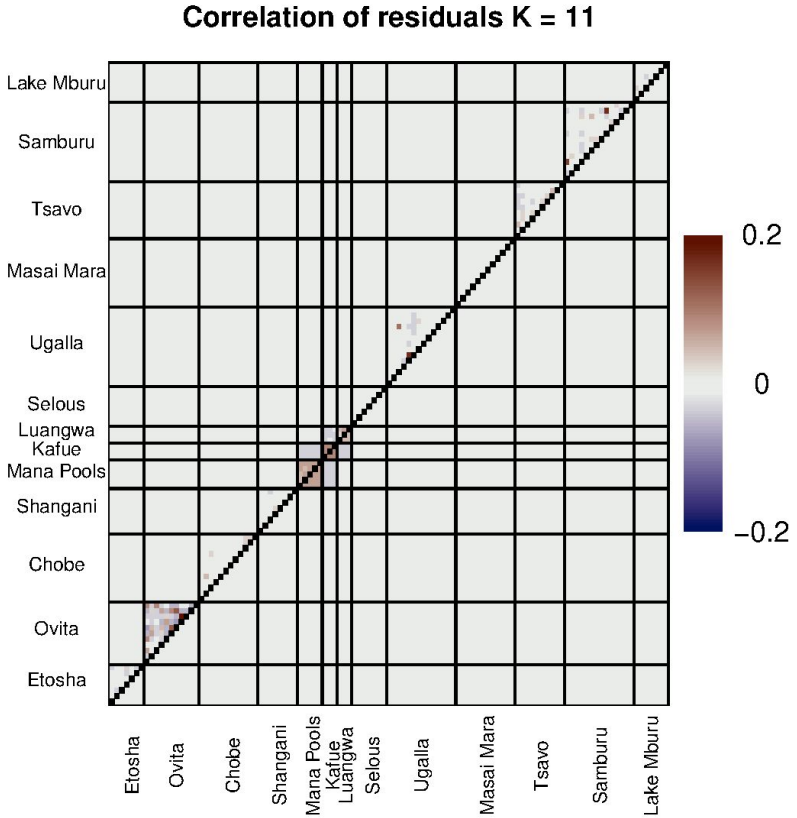
K=9


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu

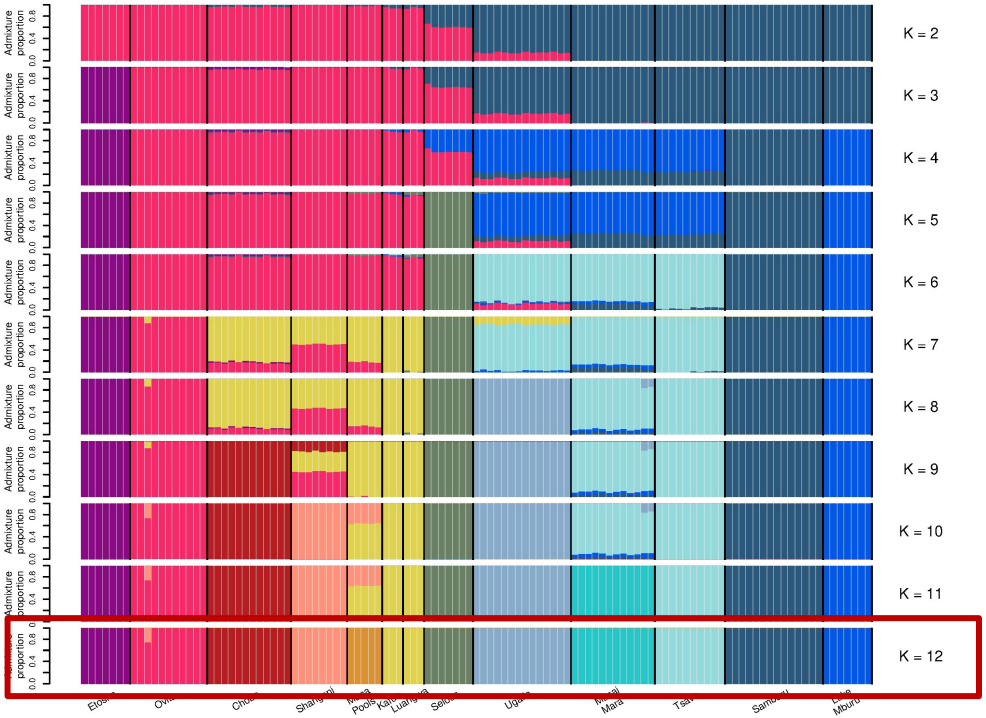

Correlation of residuals K = 10
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

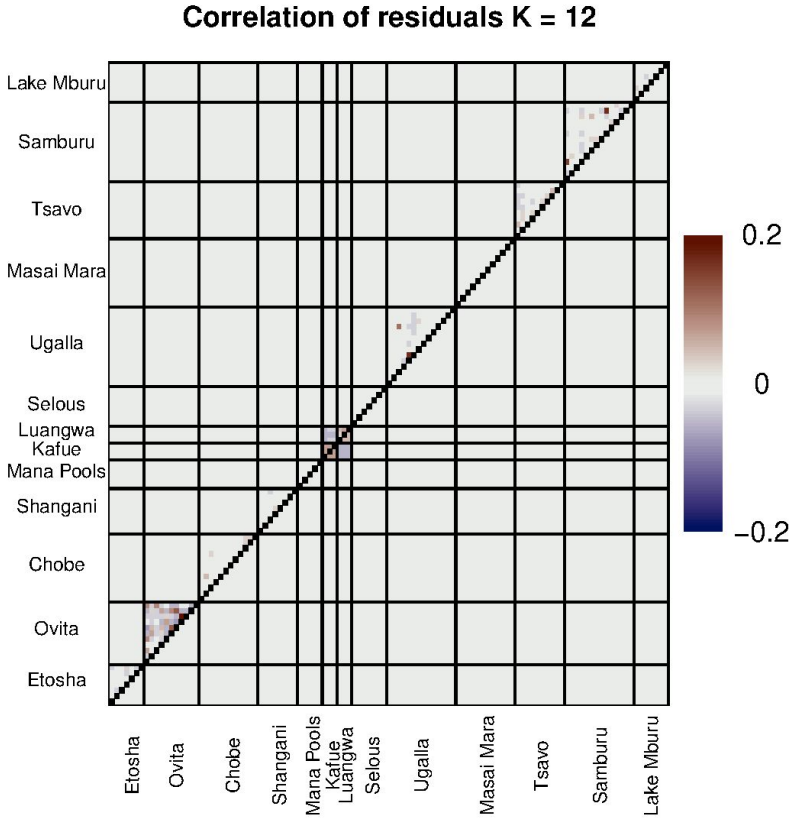
K=10


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
Etosha
Ovita
Chobe
Shangani
Mana Pools
Kafue
Luangwa
Selous
Ugalla
Masai Mara
Tsavo
Samburu
Lake Mburu


Correlation of residuals K = 11
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

K=11


Admixture proportion
K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12


Correlation of residuals K = 12
Lake Mburu
Samburu
Tsavo
Masai Mara
Ugalla
Selous
Luangwa
Kafue
Mana Pools
Shangani
Chobe
Ovita
Etosha
0.2
0
−0.2

K=12

Time for exercises

Run the admixture notebook