Population Genomics wk5

Population Structure and Admixture

Slurm Once a-More

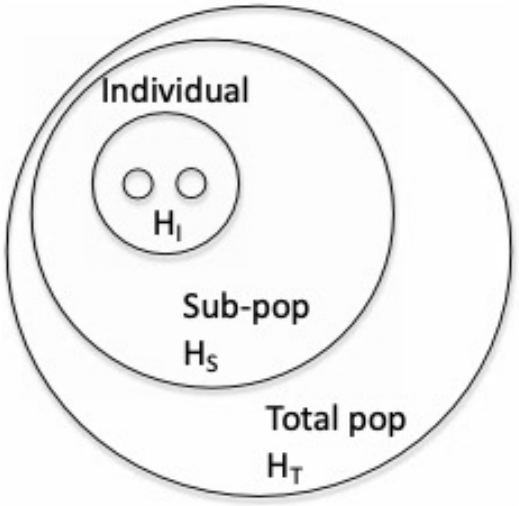
- Download an env file from cluster to your own computer
  scp username@login.genome.au.dk:populationgenomics/env/exercise_envs/jupyter_bjarke.yml <folder on your own computer>
- Install the env (conda env create –f <path to jupyter_bjarke.yml>)
- Log on the cluster and create another env using the bircproject.yml from the env/exercise_envs/ folder
- Go to you on computer and run
- `slurm-jupyter -u username -e bircproject -A populationgenomics -m 5g -t 3h`
- `Welcome to Slurm everyone`

Exercises Today

- Very breifly about population structure
- PCA and Admixture
- Exercises
- BREAK
- Exercises
- Class Wrap up

F statistics

- Individuals rarely mate completely at random
- Individuals tend to mate with individuals from the same, or closely related sets of populations.
- Populations can often differ in their allele frequencies, either due to genetic drift or selection driving differentiation among populations
- In one of the first chapters, you read about FST, FIS and the general set of F-statistics



$$F_{\mathrm{IS}} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_S q_S},$$

Fraction of individuals that are heterozygous

Expected Heterozygosity under random mating

$$F_{\mathrm{IT}} = 1 - \frac{H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T},$$

$$F_{\mathrm{ST}} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T}.$$

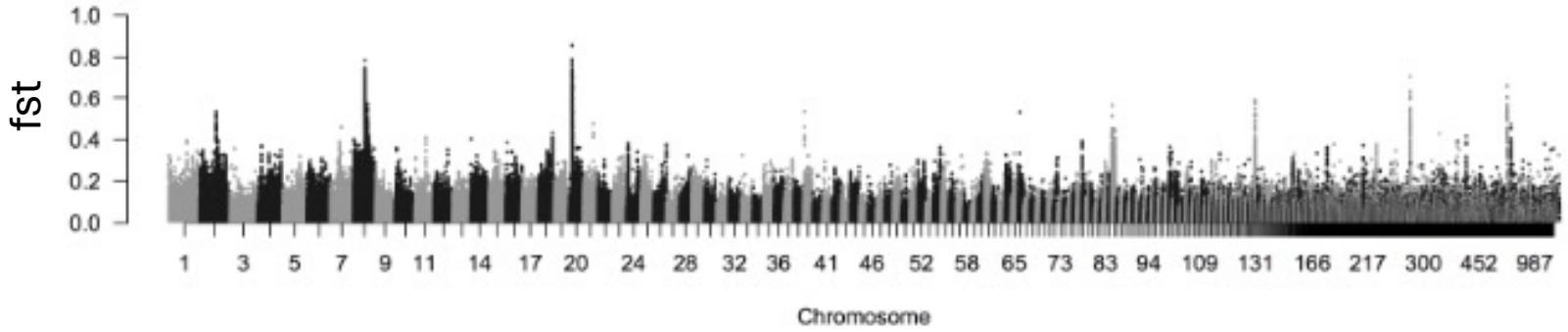
$p_s = allele\ frequency\ of\ A_1 in\ subpopulation\ s$

$q_s = allele\ frequency\ of\ A_2 in\ subpopulation\ s$

The expected in sub/ the expected in total

F Statistics

Fst example from the book



F is the covariance between pairs of alleles found in an individual, divided by the expected variance under binomial sampling.

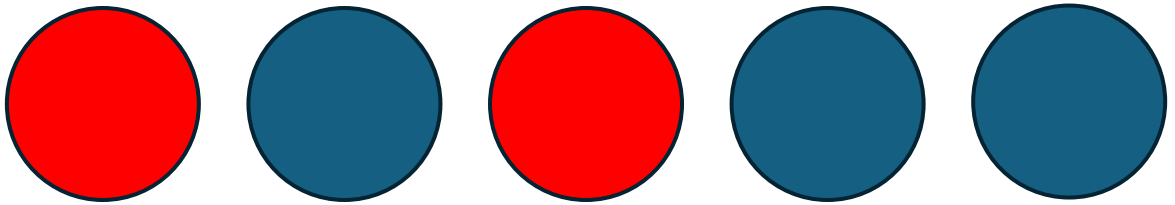
F-statistics can be understood as the correlation between alleles drawn from a population (or an individual) above that expected by chance (i.e. drawing alleles sampled at random from some broader population)

Assignment Methods

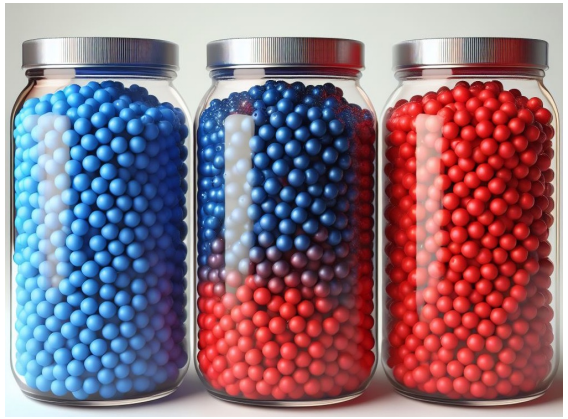
Find the probability that an individual of unknown population comes from one of K predefined populations.

Calculate the probability of an individual genotype coming from population k.

Genotype at a locus



Your Populations



Clustering

- STRUCTURE assigns ever individual to one of k populations and then calculates the allele frequency in each k population. Using these recalculated frequencies to reassign individuals

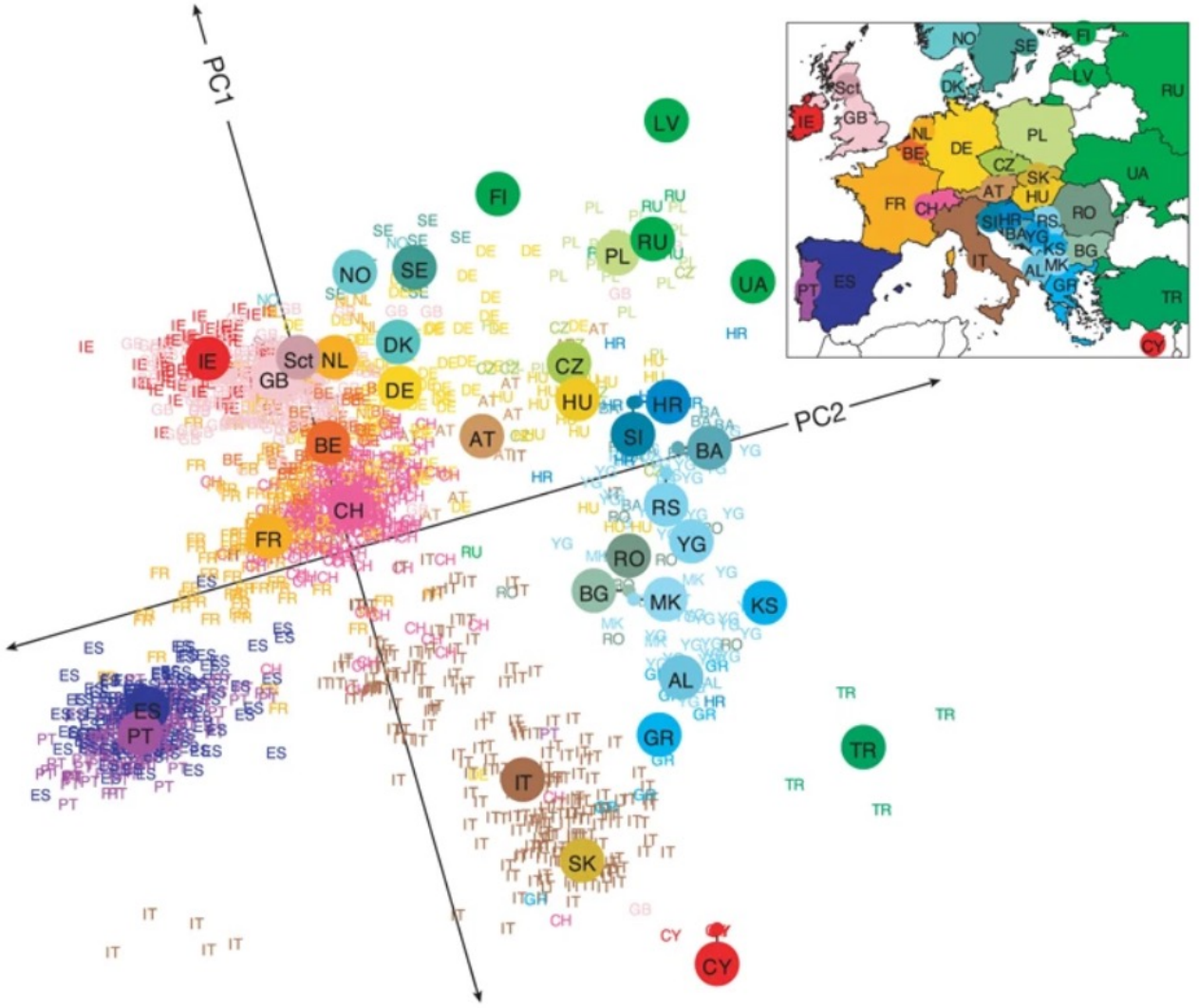
1. Given these assignments we estimate the allele frequencies at all of our loci in each population.

2. Given these allele frequencies we chose to reassign each individual to a population k with a probability given by eqn(3.9).

It is tempting to think of these clusters as representing ancestral populations, which themselves are not the result of admixture. How- ever, that is not the case, for example, running STRUCTURE on world-wide human data identifies a cluster that contains many Euro- pean individuals, however, on the basis of ancient DNA we know that modern Europeans are a mixture of distinct ancestral groups.

Principal Components Analysis (PCA)

- Find SNPS for a lot of people
- Cliffhanger: PCAs are not that great, if not careful



EXERCISES

```
%%R
info <- read.csv("sample_infos_accessionnb.csv", header = T, sep = ';')
```

```
%%R
info
```

	ID	ENA.RUN	population	region	country	latitude	longitude	Sex	Illumina.ID
1	ERS1042176	ERR1019075	Ju_hoan_North	Africa	Namibia	-18.90000	21.5000	male	LP6005441-DNA_B11
2	ERS1042177	ERR1019076	Ju_hoan_North	Africa	Namibia	-18.90000	21.5000	male	LP6005441-DNA_A11
3	ERS1042248	ERR1025622	Esan	Africa	Nigeria	6.50000	6.0000	female	LP6005442-DNA_B10
4	ERS1042265	ERR1025639	Luhya	Africa	Kenya	1.30000	36.8000	male	LP6005442-DNA_E11
5	ERS1042266	ERR1025640	Mandenka	Africa	Senegal	12.00000	-12.0000	male	LP6005441-DNA_E07
6	ERS1042267	ERR1025641	Mandenka	Africa	Senegal	12.00000	-12.0000	female	LP6005441-DNA_F07
7	ERS1042283	ERR1025657	Yoruba	Africa	Nigeria	7.40000	3.9000	female	LP6005442-DNA_B02
8	ERS1042284	ERR1025658	Yoruba	Africa	Nigeria	7.40000	3.9000	male	LP6005442-DNA_A02
9	ERS1042265	ERR1347655	Luhya	Africa	Kenya	1.30000	36.8000	male	LP6005442-DNA_E11
10	ERS1255044	ERR1347660	Luo	Africa	Kenya	-0.10000	34.3000	female	LP6005442-DNA_F09
11	ERS1042124	ERR1019060	Miao	EastAsia	China	28.00000	109.0000	male	LP6005441-DNA_C08
12	ERS1042175	ERR1019074	Japanese	EastAsia	Japan	38.00000	138.0000	female	LP6005441-DNA_D06
13	ERS1042141	ERR1019039	Naxi	EastAsia	China	26.00000	100.0000	male	LP6005441-DNA_A09
14	ERS1042157	ERR1019055	Han	EastAsia	China	32.30000	114.0000	female	LP6005441-DNA_D05
15	ERS1042224	ERR1025598	Atayal	EastAsia	Taiwan	24.61171	121.2964	male	LP6005442-DNA_E07
16	ERS1042236	ERR1025610	Ami	EastAsia	Taiwan	22.84314	121.1854	male	LP6005442-DNA_C07
17	ERS1042243	ERR1025617	Cambodian	EastAsia	Cambodia	12.00000	105.0000	male	LP6005441-DNA_G03
18	ERS1042244	ERR1025618	Cambodian	EastAsia	Cambodia	12.00000	105.0000	female	LP6005441-DNA_H03
19	ERS1255084	ERR1347700	Korean	EastAsia	Korea	37.60000	127.0000	female	LP6005443-DNA_C06
20	ERS1042264	ERR1025638	Kinh	EastAsia	Vietnam	21.00000	105.9000	male	LP6005442-DNA_C11
21	ERS1042240	ERR1025614	Bulgarian	WestEurasia	Bulgaria	42.20000	24.7000	male	LP6005442-DNA_A03
22	ERS1042241	ERR1025615	Bulgarian	WestEurasia	Bulgaria	42.20000	24.7000	male	LP6005442-DNA_B03
23	ERS1042245	ERR1025619	Druze	WestEurasia	Israel(Carmel)	32.00000	35.0000	male	LP6005441-DNA_G04
24	ERS1042246	ERR1025620	English	WestEurasia	England	51.20000	0.7000	male	LP6005442-DNA_E10
25	ERS1042249	ERR1025623	Georgian	WestEurasia	Georgia	42.50000	41.9000	male	LP6005442-DNA_B04
26	ERS1042250	ERR1025624	Georgian	WestEurasia	Georgia	42.50000	41.9000	male	LP6005442-DNA_A04
27	ERS1042255	ERR1025629	Hungarian	WestEurasia	Hungary	47.50000	19.1000	female	LP6005442-DNA_B08
28	ERS1042256	ERR1025630	Hungarian	WestEurasia	Hungary	47.50000	19.1000	male	LP6005442-DNA_A08
29	ERS1042257	ERR1025631	Icelandic	WestEurasia	Iceland	64.10000	-21.9000	female	LP6005442-DNA_D08
30	ERS1042258	ERR1025632	Iranian	WestEurasia	Iran	35.60000	51.5000	male	LP6005442-DNA_C04

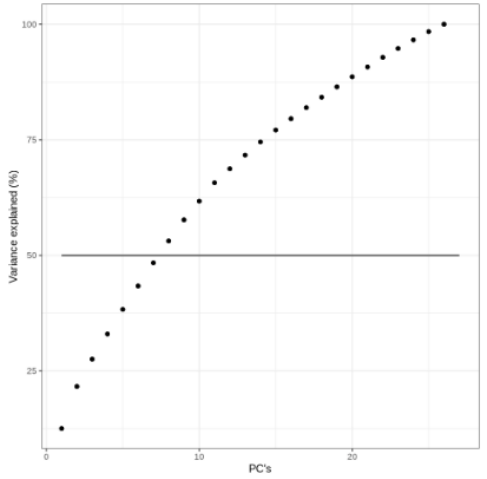
Q.1 How many individuals and snps does this dataset have? What is an eigenvector and an eigenvalue?

```
%%R
# Setting the directory of the VCF file
vcf.fn <- "chr2_135_145_flt.vcf.gz"

# Transforming the vcf file to gds format
snpgdsVCF2GDS(vcf.fn, "chr2_135_145_flt.gds", method="biallelic.only")
```

```
Start file conversion from VCF to SNP GDS ...
Method: extracting biallelic SNPs
Number of samples: 27
Parsing "chr2_135_145_flt.vcf.gz" ...
        import 49868 variants.
+ genotype   { Bit2 27x49868, 328.7K } *
Optimize the access efficiency ...
Clean up the fragments of GDS file:
    open the file 'chr2_135_145_flt.gds' (519.6K)
    # of fragments: 50
    save to 'chr2_135_145_flt.gds.tmp'
    rename 'chr2_135_145_flt.gds.tmp' (519.2K, reduced: 360B)
    # of fragments: 20
```

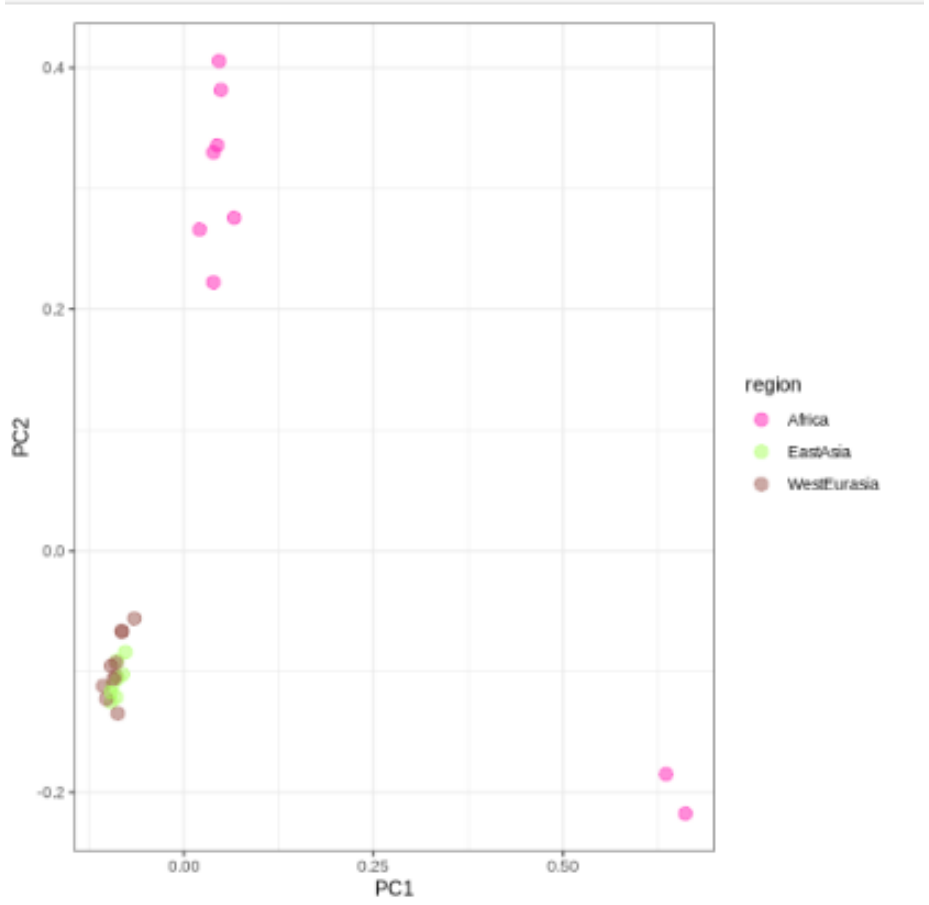
Q.2 How many PC's do we need in order to explain 50% of the variance of the data? Can you make a cumulative plot of the variance explained PC?

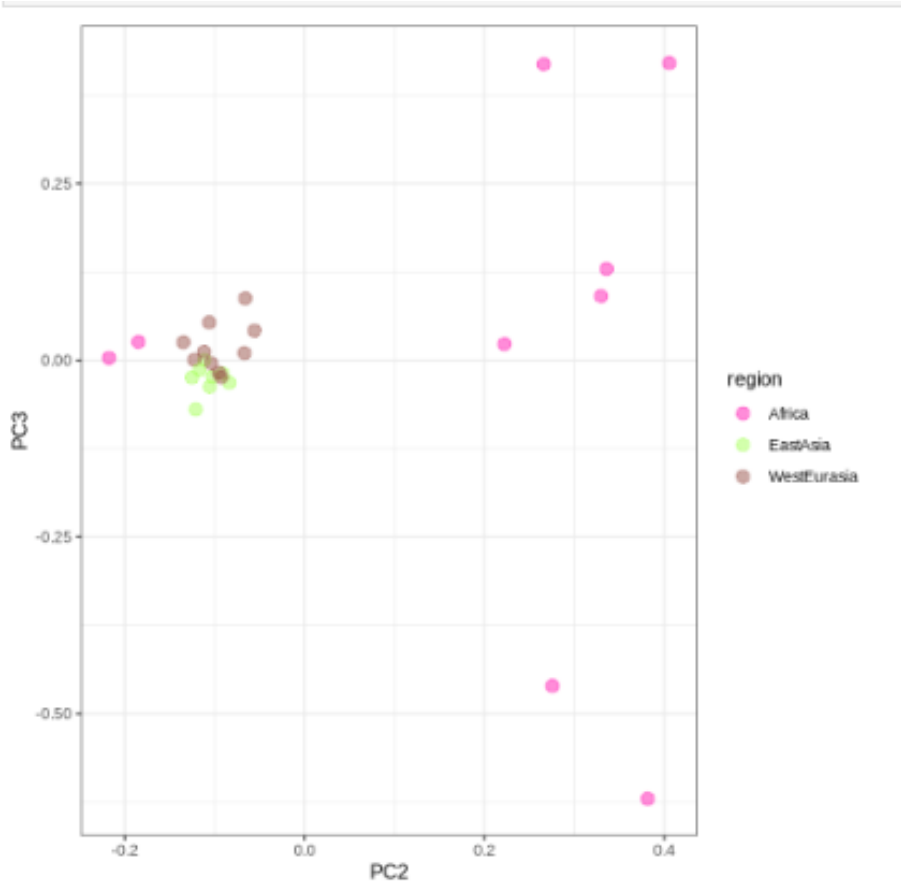


EXERCISES

Q.3 Try to plot PC2 and PC3. Do you see the same patterns? What is the correlation between PC2 and PC3 (hint use the function cor())?

Q.4 Try also to color the graph based on population. What do you observe?

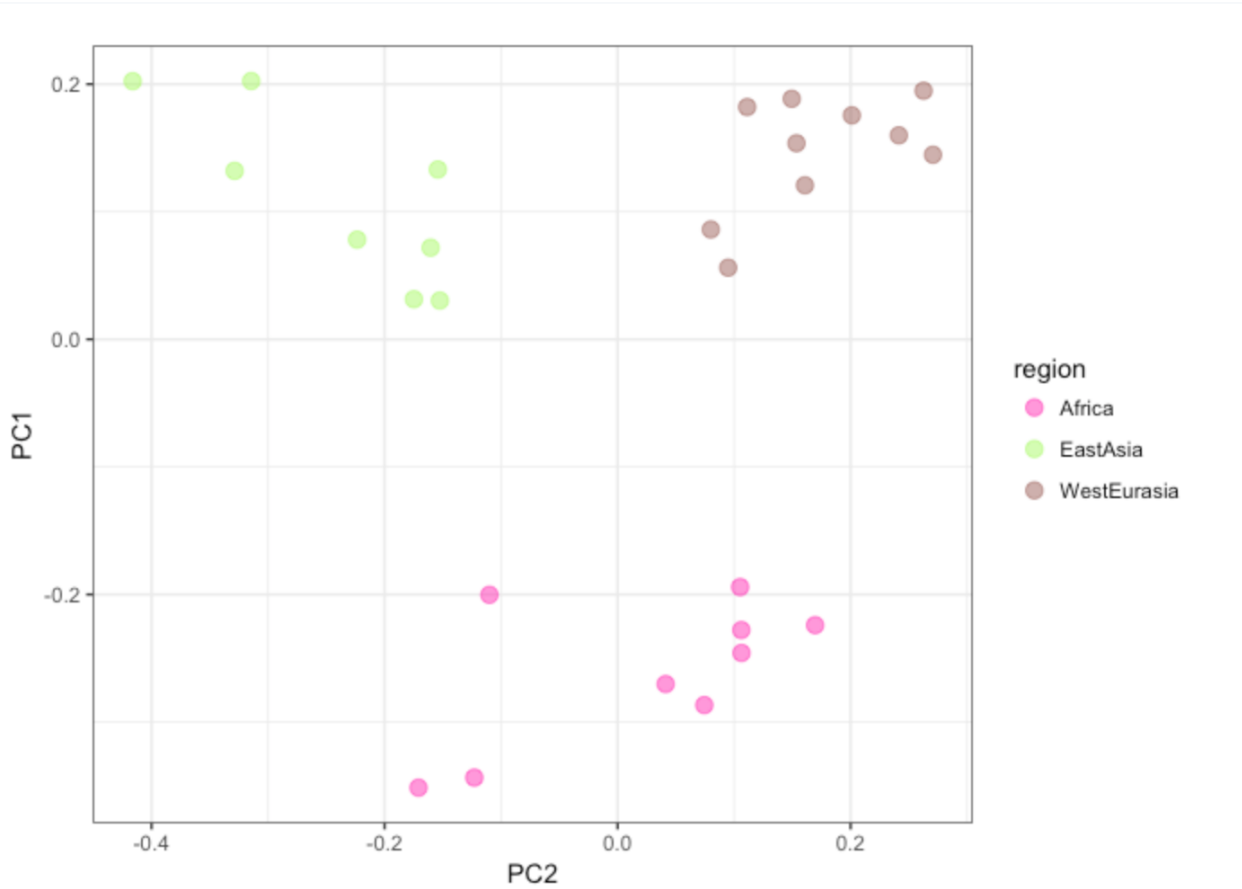




EXERCISES

```
%%R
# Get all selected snp's ids
snpset.id <- unlist(snpset)
pca_pruned <- snpgdsPCA(genofile, snp.id=snpset.id, num.thread=2, ,eigen.cnt=n_pcs)
```

```
Principal Component Analysis (PCA) on genotypes:
Excluding 49,261 SNPs (non-autosomes or non-selection)
Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
    # of samples: 27
    # of SNPs: 607
    using 2 threads
    # of principal components: 26
PCA:    the sum of all selected genotypes (0,1,2) = 29826
CPU capabilities: Double-Precision SSE2
Thu Feb 29 11:11:39 2024    (internal increment: 75092)
[==================================================] 100%, completed, 0s
Thu Feb 29 11:11:39 2024    Begin (eigenvalues and eigenvectors)
Thu Feb 29 11:11:39 2024    Done.
```

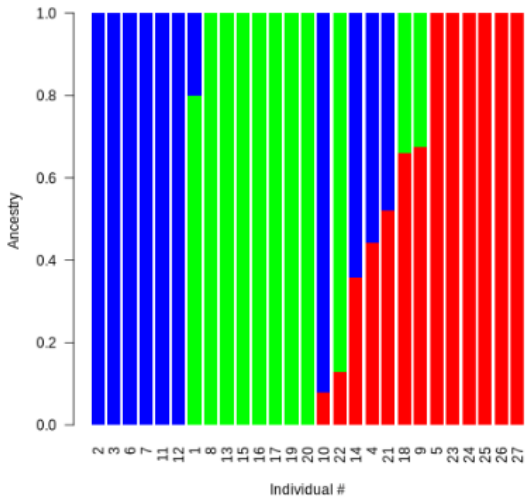


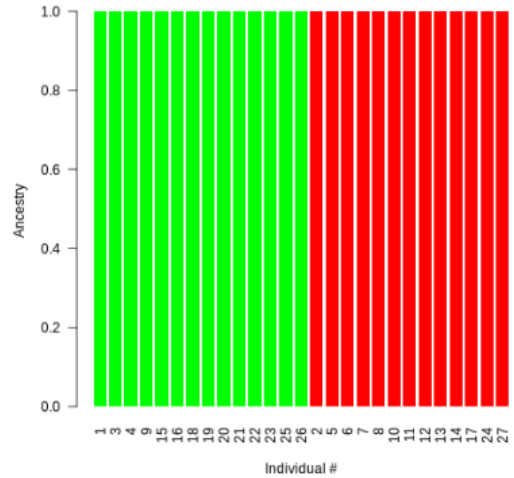
Exercises

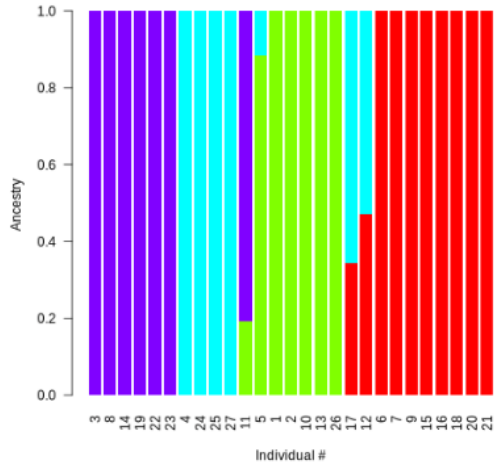
Q.6 Have a look at the Fst across populations, that is printed in the terminal. Would you guess which populations are Pop0, Pop1 and Pop2 referring to?

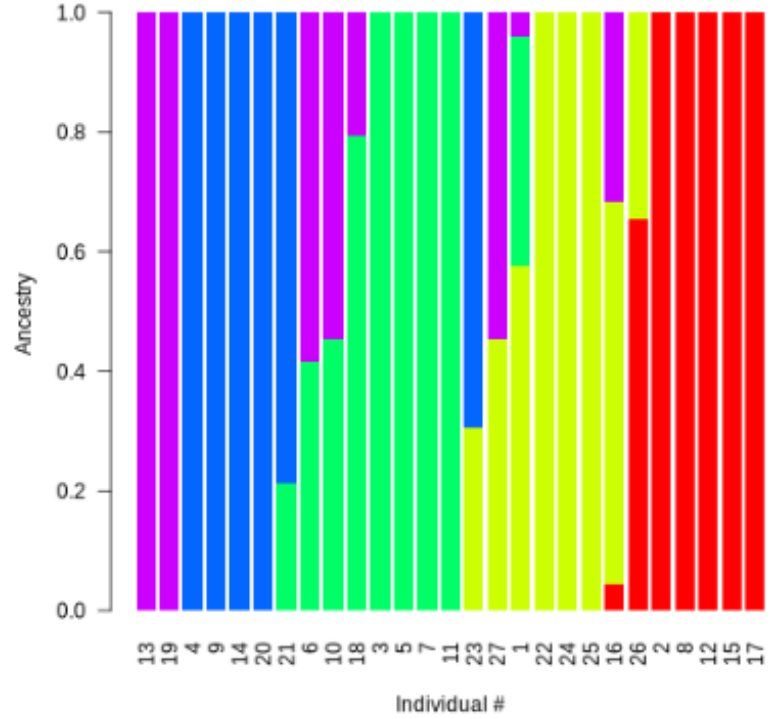
```
****                  ADMIXTURE Version 1.3.0                 ****
****                  Copyright 2008-2015                     ****
****           David Alexander, Suyash Shringarpure,          ****
****                John  Novembre, Ken Lange                 ****
****                                                          ****
****                 Please cite our paper!                   ****
****   Information at www.genetics.ucla.edu/software/admixture  ****

Random seed: 43
Point estimation method: Block relaxation algorithm
Convergence acceleration algorithm: QuasiNewton, 3 secant conditions
Point estimation will terminate when objective function delta < 0.0001
Estimation of standard errors disabled; will compute point estimates only.
Size of G: 27x607
Performing five EM steps to prime main algorithm
1 (EM)  Elapsed: 0      Loglikelihood: -6472.91 (delta): 19992
2 (EM)  Elapsed: 0      Loglikelihood: -6018.01 (delta): 454.903
3 (EM)  Elapsed: 0      Loglikelihood: -5911.84 (delta): 106.168
4 (EM)  Elapsed: 0      Loglikelihood: -5852.32 (delta): 59.5197
5 (EM)  Elapsed: 0      Loglikelihood: -5812.44 (delta): 39.8764
Initial loglikelihood: -5812.44
Starting main algorithm
1 (QN/Block)    Elapsed: 0.007  Loglikelihood: -5540.34 (delta): 272.101
2 (QN/Block)    Elapsed: 0.003  Loglikelihood: -5473.39 (delta): 66.9472
3 (QN/Block)    Elapsed: 0.007  Loglikelihood: -5452.41 (delta): 20.981
4 (QN/Block)    Elapsed: 0.009  Loglikelihood: -5415.6  (delta): 36.8176
5 (QN/Block)    Elapsed: 0.009  Loglikelihood: -5411.06 (delta): 4.53634
6 (QN/Block)    Elapsed: 0.01   Loglikelihood: -5410.85 (delta): 0.210446
7 (QN/Block)    Elapsed: 0.008  Loglikelihood: -5410.83 (delta): 0.0145321
8 (QN/Block)    Elapsed: 0.003  Loglikelihood: -5410.82 (delta): 0.0110643
9 (QN/Block)    Elapsed: 0.003  Loglikelihood: -5410.82 (delta): 7.50908e-06
Summary:
Converged in 9 iterations (0.081 sec)
Loglikelihood: -5410.823587
Fst divergences between estimated populations:
        Pop0    Pop1
Pop0
Pop1    0.106
Pop2    0.111   0.076
Writing output files.
```


Ancestry
1.0
0.8
0.6
0.4
0.2
0.0
2 3 6 7 11 12 1 8 13 15 16 17 19 20 10 22 14 4 21 18 9 5 23 24 25 26 27
Individual #


Ancestry
1.0
0.8
0.6
0.4
0.2
0.0
1 3 4 9 15 16 18 19 20 21 22 23 25 26 2 5 6 7 8 10 11 12 13 14 17 24 27
Individual #


Ancestry
1.0
0.8
0.6
0.4
0.2
0.0
3 8 14 19 22 23 4 24 25 27 11 5 1 2 10 13 26 17 12 6 7 9 15 16 18 20 21
Individual #


Ancestry
1.0
0.8
0.6
0.4
0.2
0.0
13 19 4 9 14 20 21 6 10 18 3 5 7 11 23 27 1 22 24 25 16 26 2 8 12 15 17
Individual #

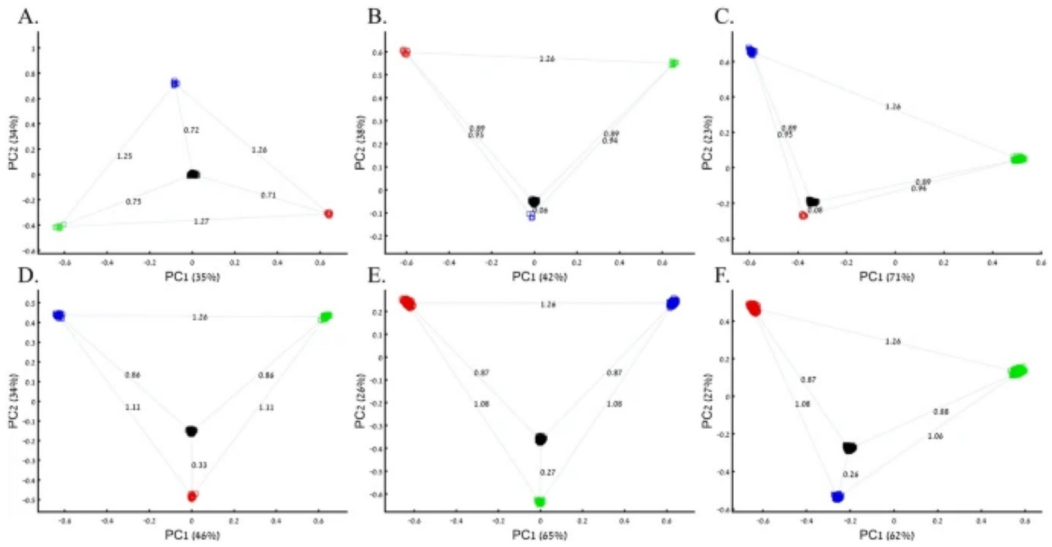
Caution is always needed when doing ADMIXTURE or PCA

- PCAs are sensitive to a lot of thing that can make interpretaions hard or even directly wrong
  - **Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated (https://www.nature.com/articles/s41598-022-14395-4)**

- ADMIXTURE/structure, have a problem of choosing K
  - K could be many different values but for some reason it always tend to be the one value that fits the authors conclusions

Studying the origin of Black using the primary colors

Figure 4



PCA of uneven-sized samples of four color populations. (**A**) $n_{Red} = n_{Green} = n_{Blue} = 10$; $n_{Black} = 200$, (**B**) $n_{Red} = n_{Green} = 10$; $n_{Blue} = 5$; $n_{Black} = 200$, (**C**) $n_{Red} = 10$; $n_{Green} = 200$; $n_{Blue} = 50$; $n_{Black} = 200$ (**D**) $n_{Red} = 25$; $n_{Green} = n_{Blue} = 50$; $n_{Black} = 200$, (**E**) $n_{Red} = 300$; $n_{Green} = 200$; $n_{Blue} = n_{Black} = 300$, and (**F**) $n_{Red} = 1000$; $n_{Green} = 2000$; $n_{Blue} = 300$; $n_{Black} = 2000$. Scatter plots show the top two PCs. The numbers on the grey bars reflect the Euclidean distances between the color populations over all PCs. Colors include Red [1,0,0], Green [0,1,0], Blue [0,0,1], and Black [0,0,0].