

Population structure II

PCA

Anders Albrechtsen

The idea in general

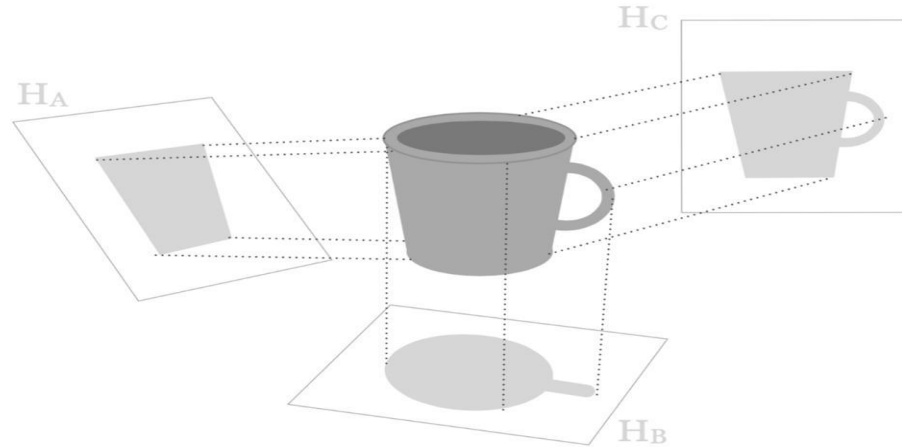
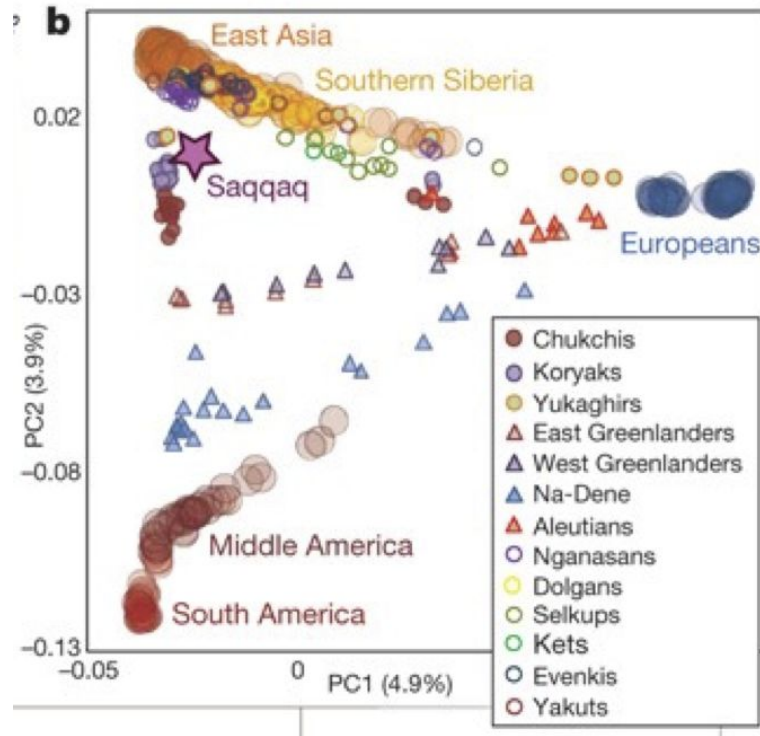


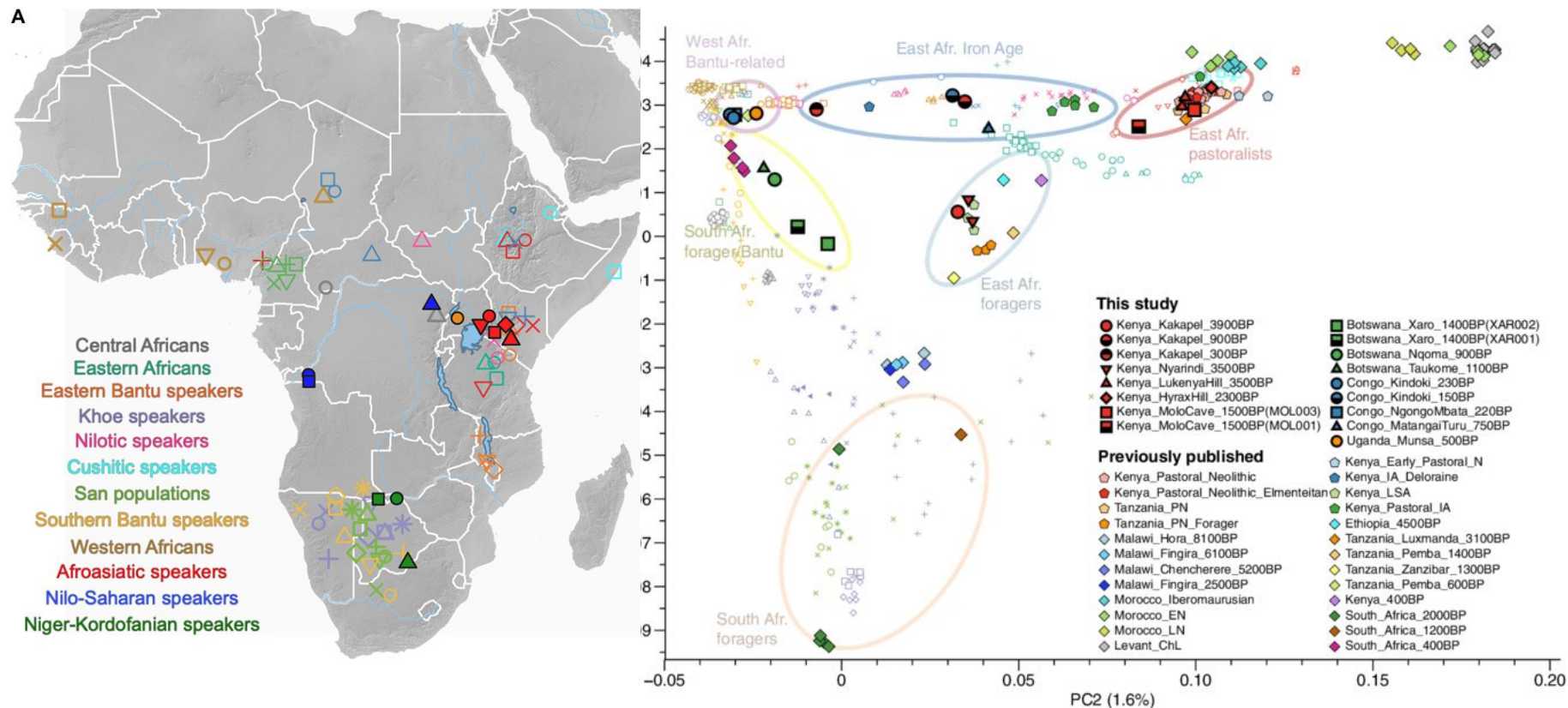
Figure 1.6: Three projections of a mug-shaped cloud points

It matters which of the three we pick, right?

We want it to reflect the genetic relationship between all pairs

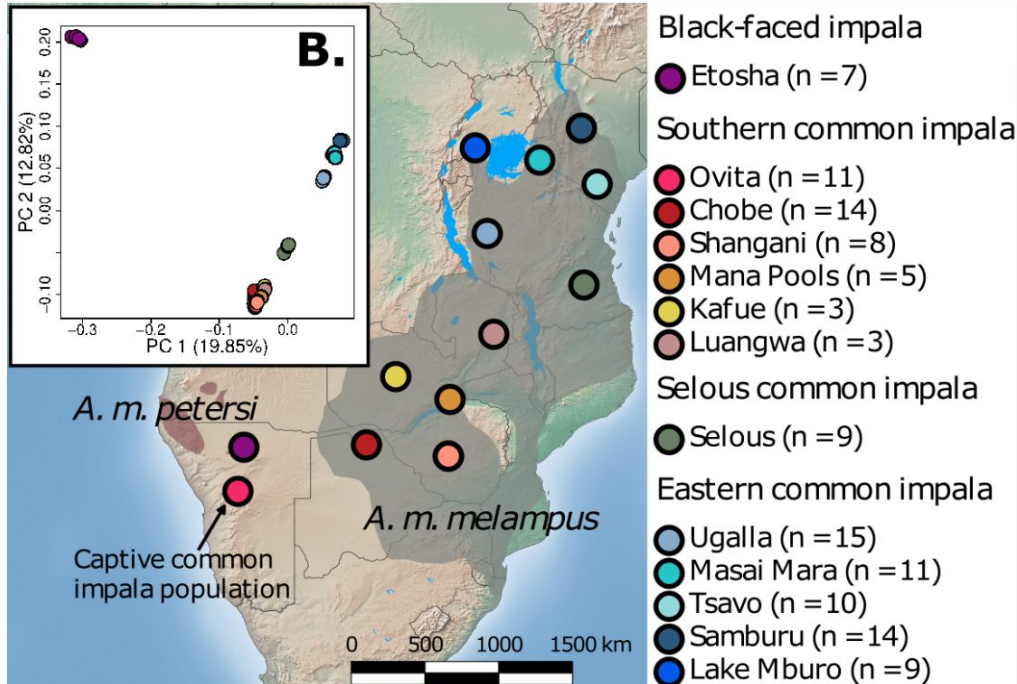
PCA for population structure





Impala

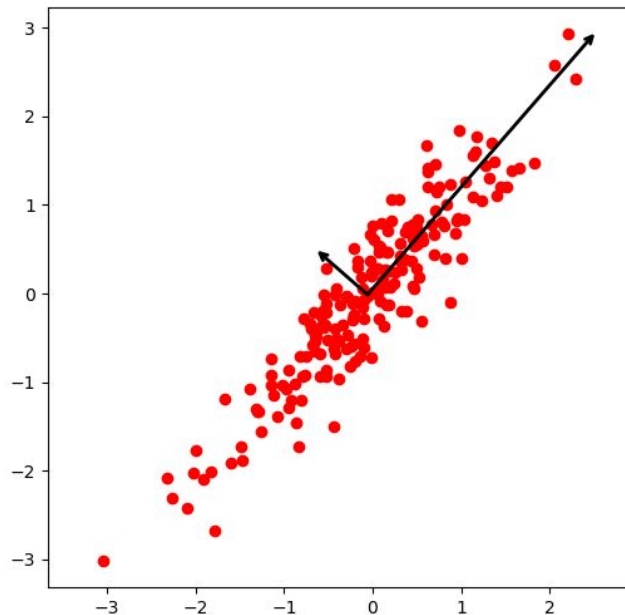
A.



Principal component analysis (PCA)

- Dimensionality reduction
- Axis of variation
- Principal components
- Models more ***Continuous**** population structure than ADMIXTURE

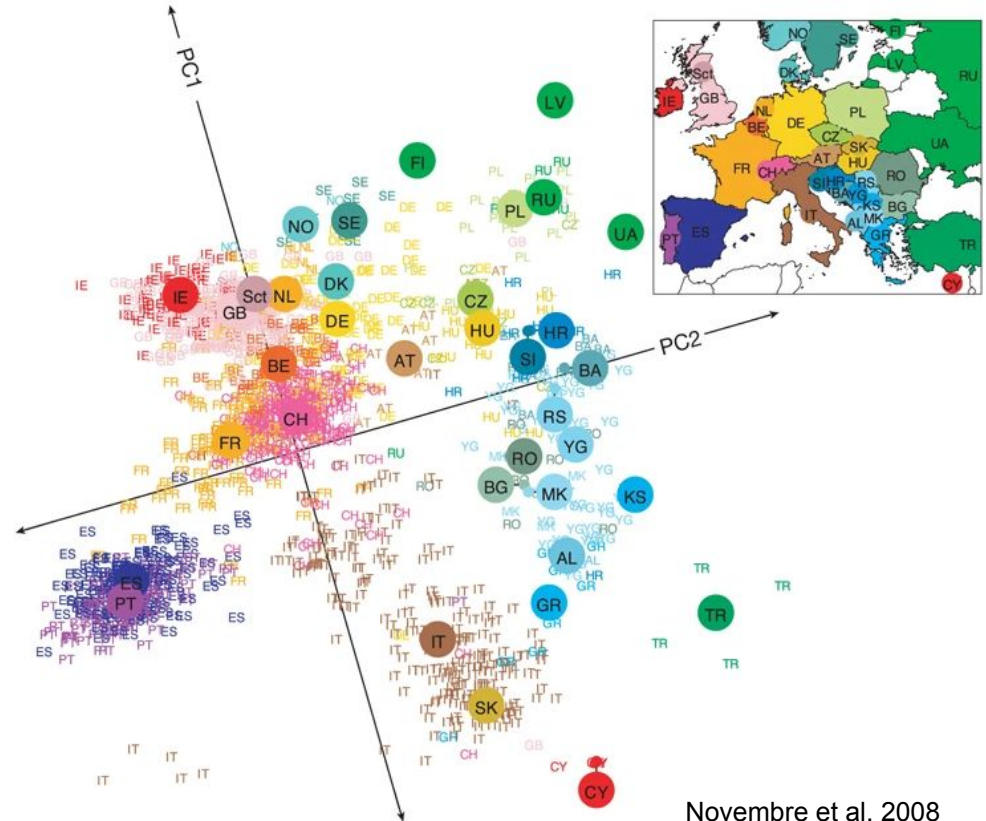
****not really true***



Principal component analysis (PCA)

Genetic data

- $m > 1$ million
- Captures genetic structure



Novembre et al. 2008

Today you will learn

- The underlying “model” of PCA and MDS
 - What these two methods are trying to achieve
- The relationship between admixture proportions and PCA
- How PCA predict genotypes
- Issues with missingness
 - For call genotypes and for low depth sequencing
- How to deal with missingness
- How PCA can be used for selection scan (teaser)

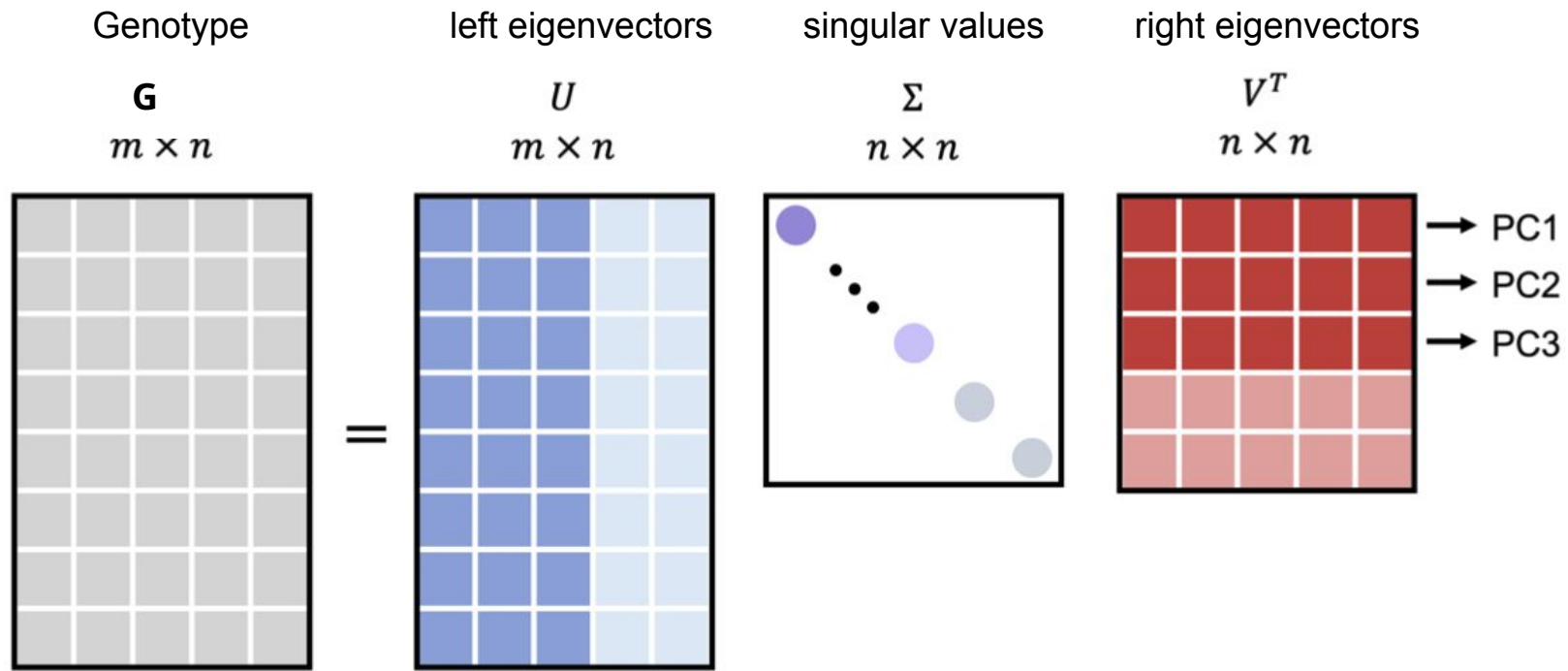
Genotype data

	Ind1	Ind2	Ind3	Ind4	Ind5
SNP1	AG	AG	AG	AA	AA
SNP2	TT	TA	AA	AT	AA
SNP3	AA	AC	AC	CC	AC
SNP4	GG	GG	GC	CC	CC
SNP5	TT	TC	TC	CC	CC
SNP6	AA	AA	AC	AC	AC
SNP7	TT	TT	TC	TC	CC



SNPs

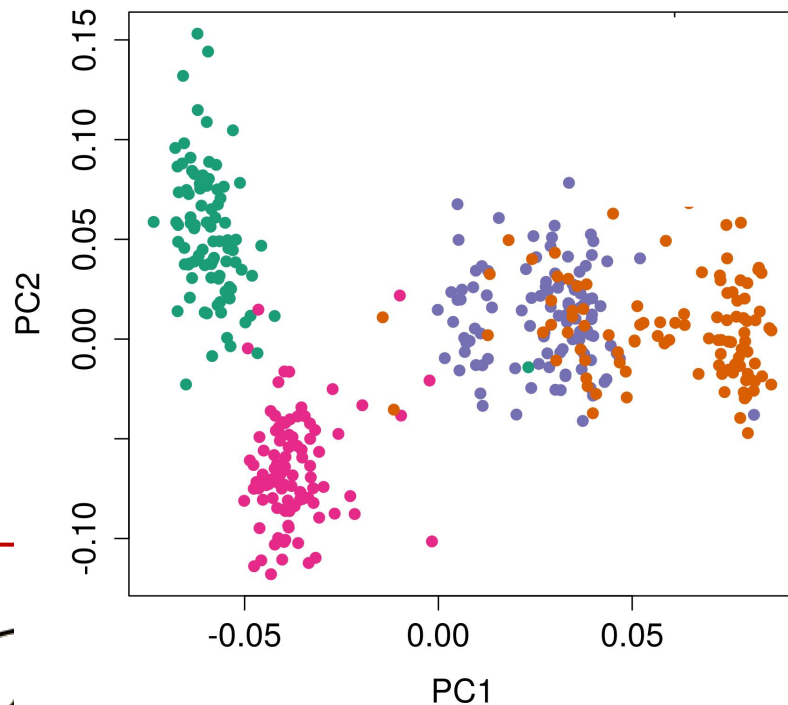
	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	1	0	0
0	1	2	1	1	2
2	1	1	0	0	1
0	0	1	2	2	2
2	1	1	0	0	0
0	0	1	1	1	1
2	2	1	1	1	0



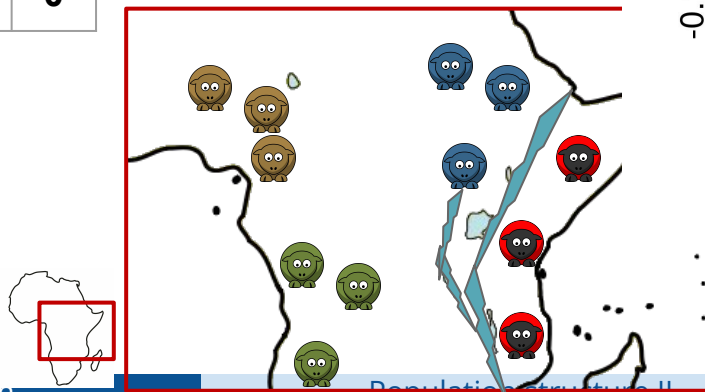
\mathbf{G} is a genotype matrix, n is the number of samples, m is the number of SNPs

SNPs

	I ₁ d	I ₂ d	I ₃ d	I ₄ d	I ₅ d
1	1	1	1	0	0
0	1	2	1	1	2
2	1	1	1	0	1
0	0	1	2	2	2
2	1	1	0	0	0
0	0	1	1	1	1
2	2	1	1	1	0



Each individuals is a dot in PCA plot



Multi-dimensional scaling (MDS)

Goal: Project the data into a low dimensional space that preserves distances

- Choose a distance
- Choose a dimension (K)

Multi-dimensional scaling

SNPs

	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	1	0	0
2	0	1	2	1	2
3	2	1	1	0	1
4	0	0	1	2	2
5	2	1	1	0	0
6	0	0	1	1	1
7	2	2	1	1	0

Pairwise distance



Manhattan distance

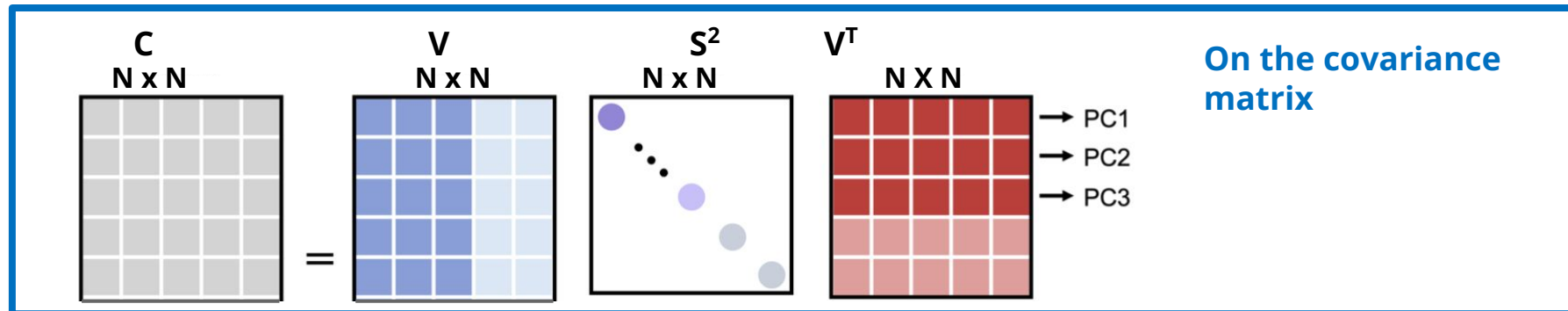
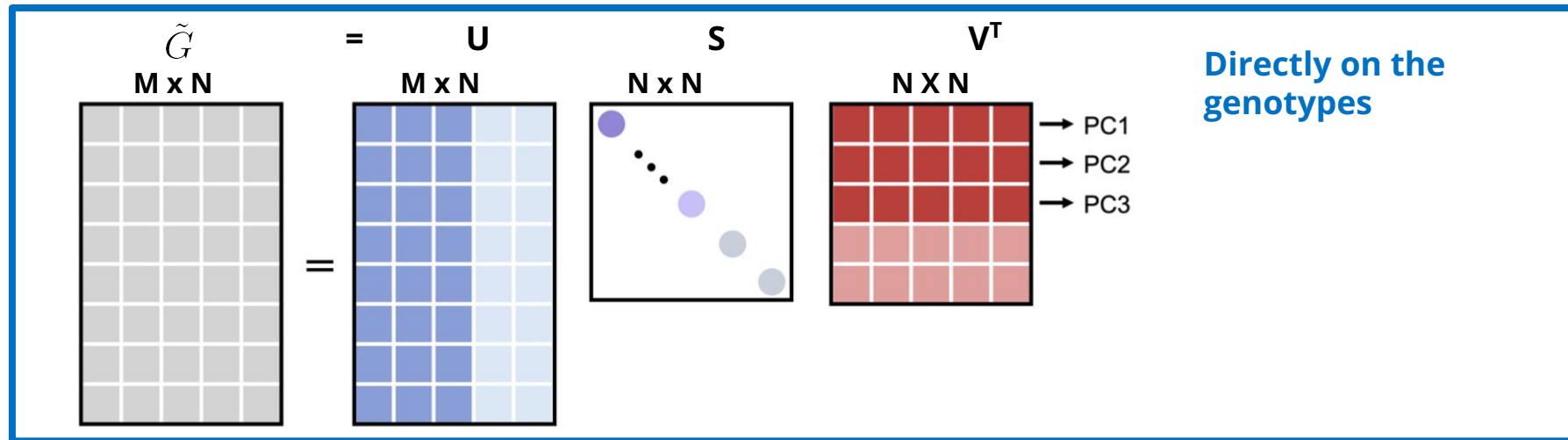
	Ind1	Ind2	Ind3	Ind4	Ind5
Ind1	0	3	7	10	11
Ind2	3	0	4	7	8
Ind3	7	4	0	5	4
Ind4	10	7	5	0	6
Ind5	11	8	4	3	0

Project into 1 dimension



	Ind1	Ind2	Ind3	Ind4	Ind5
Dim 1	6.1	3.08	-0.62	-3.7	-4.85

Two ways to do PCA



Principal component analysis

Goal: Project the data into a low dimensional space that explains the largest amount of variance

- Choose a dimension (K)

Principal component analysis

SNPs

	Ind1	Ind2	Ind3	Ind4	Ind5
1	1	1	1	0	0
0	1	2	1	1	2
2	1	1	0	0	1
0	0	1	2	2	2
2	1	1	0	0	0
0	0	1	1	1	1
2	2	1	1	1	0

Calculate
Covariance



Covariance matrix

	Ind1	Ind2	Ind3	Ind4	Ind5
Ind1	12.6	5.6	-2.0	-7.0	-9.1
Ind2	5.6	4.7	-0.8	-3.7	-5.8
Ind3	-2.0	-0.8	2.3	-0.8	1.3
Ind4	-7.0	-3.7	-0.8	6.7	4.7
Ind5	-9.1	-5.8	1.3	4.7	8.9

Project into 1 dimension

	Ind1	Ind2	Ind3	Ind4	Ind5
Dim 1	0.65	0.36	-0.08	-0.4	-0.53



Genotype covariance matrix

M number of sites

G genotype

G_i genotype for individual i

G_{ij} genotype for individual i site j

f_j frequency for site j

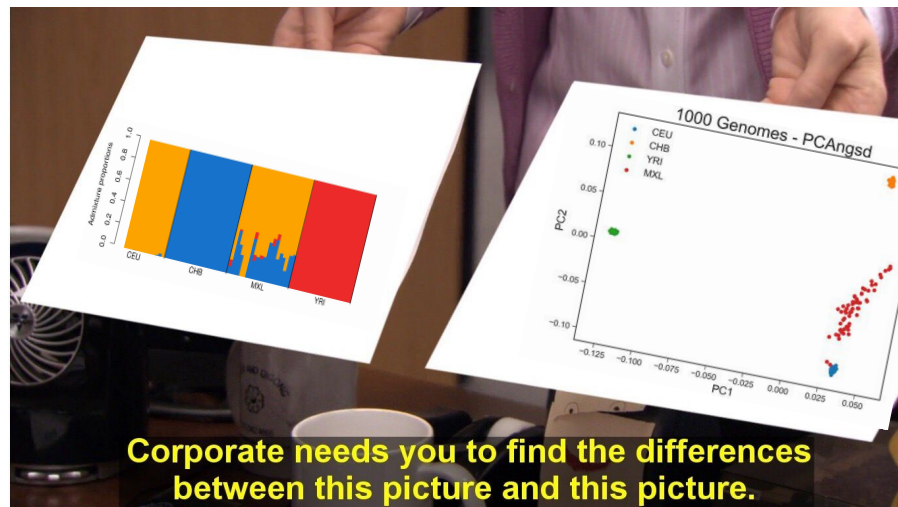
$$\tilde{G}_{ij} = \frac{G_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}$$

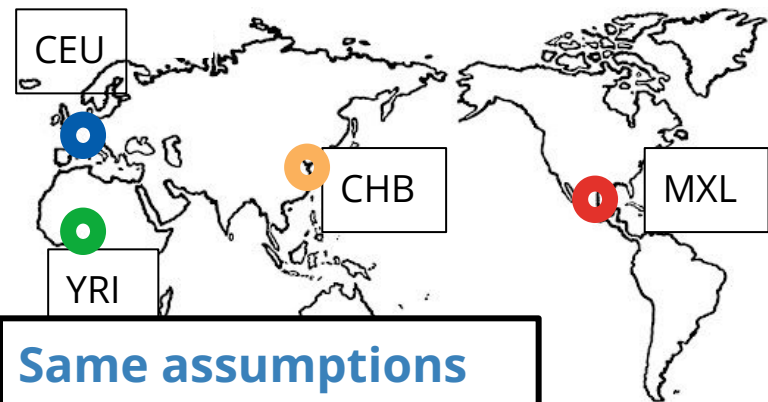
$$\text{var}(G_{ij}) = 2f_j(1 - f_j)$$

After normalization
all SNPs have the
same mean and
variance

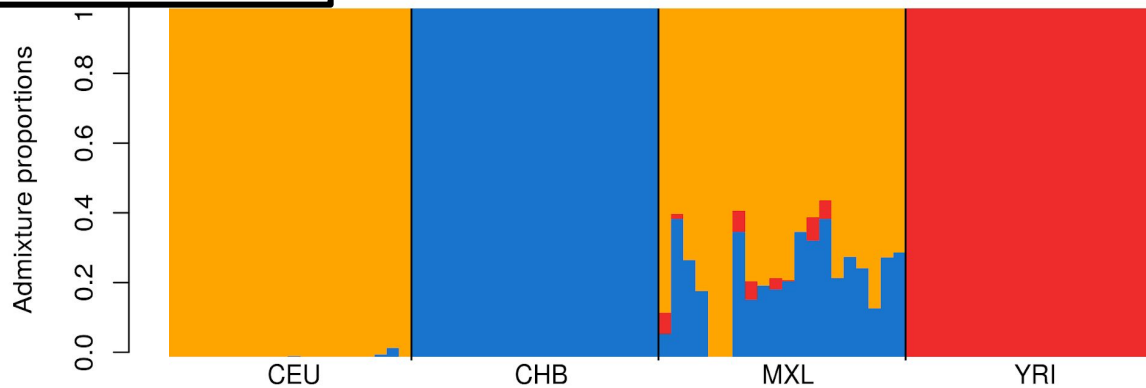
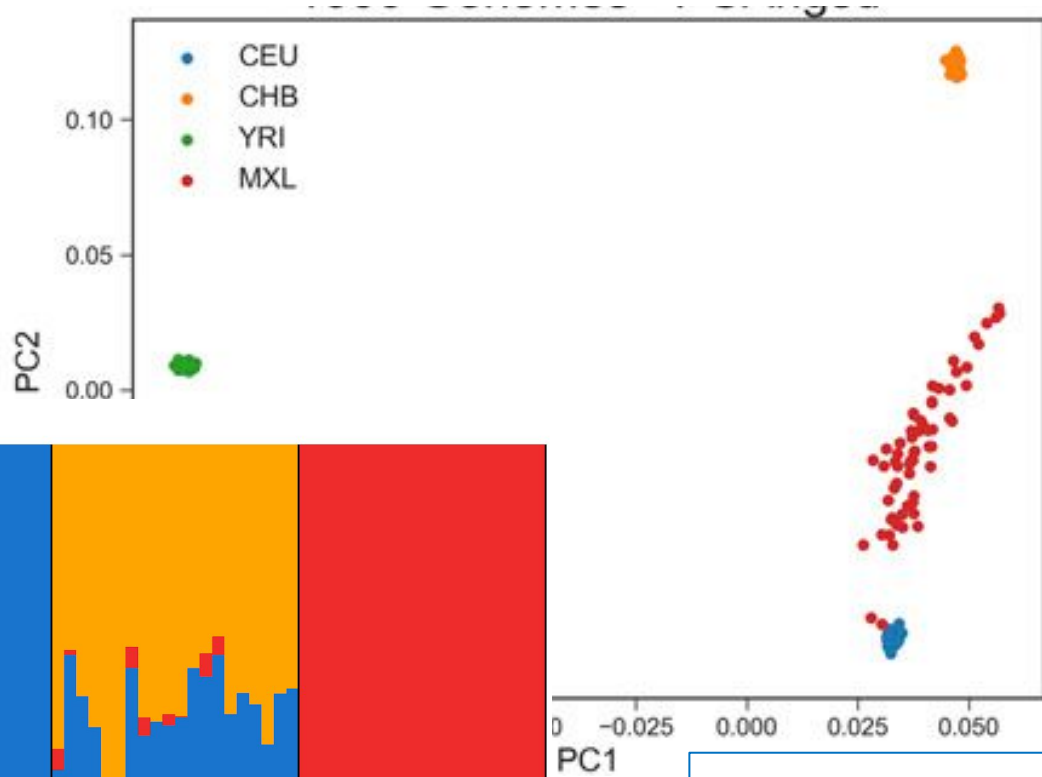
$$\text{cov}(\tilde{G}_i, \tilde{G}_l) = \frac{1}{M} \sum_{j=1}^M \frac{(G_{ij} - 2f_j)(G_{lj} - 2f_j)}{2f_j(1-f_j)} = \frac{1}{M} \tilde{G} \tilde{G}^T$$

Connection between Admixture analysis and PCA





Same assumptions
=
same issues

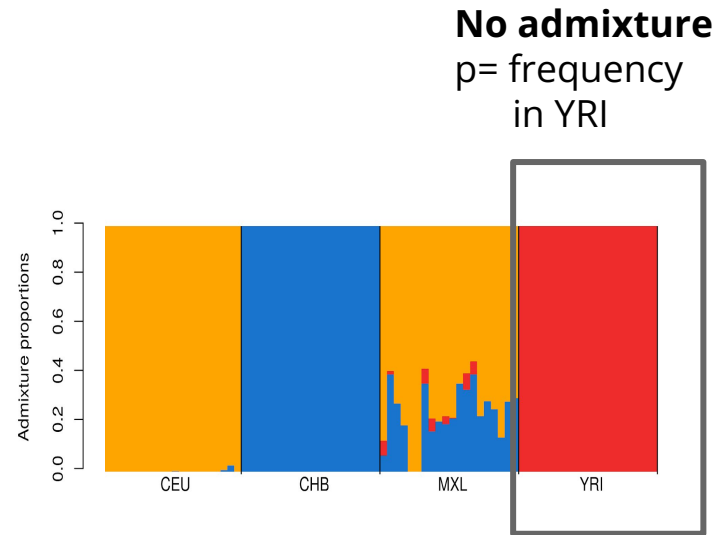


EvalAdmix

Individual allele frequencies

- Population allele frequency: $\mathbb{E}[g] = 2p$
- Individual allele frequency: $\mathbb{E}[g_i] = 2\pi_i = 2p$

$$p(g) = \begin{cases} p^2 & g = 0 \\ 2p(1 - p) & g = 1 \\ (1 - p)^2 & g = 2 \end{cases}$$



Individual allele frequencies

- Individual

$$\mathbb{E}[g_i] = 2$$

$$p(g_i) =$$

Individual allele frequencies

$$\pi_{ij} = \sum_{k=1}^K f_{jk} q_{ik}$$

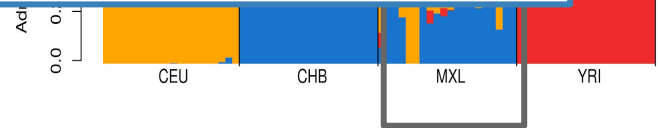
K = #populations

i = individual i

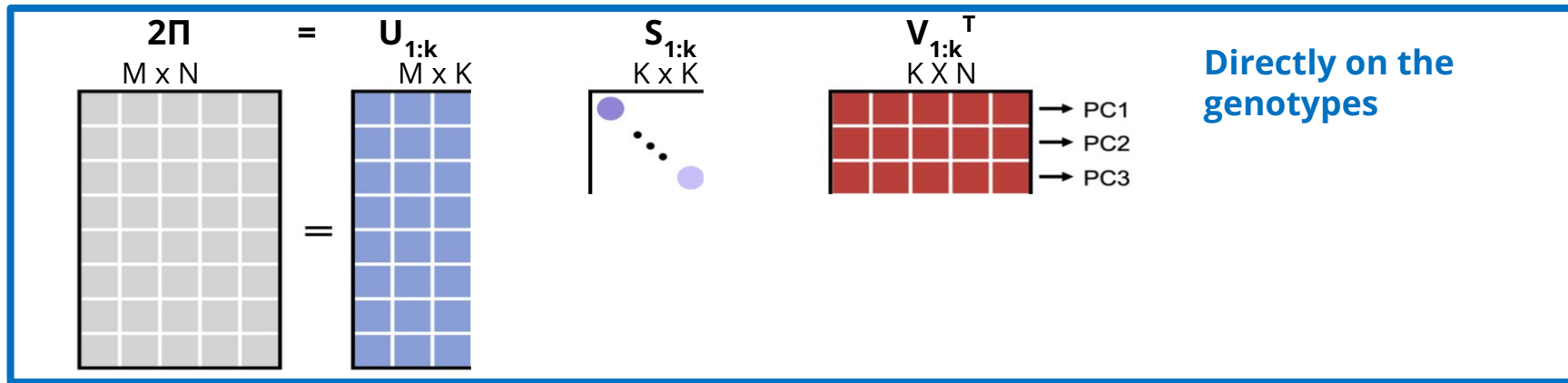
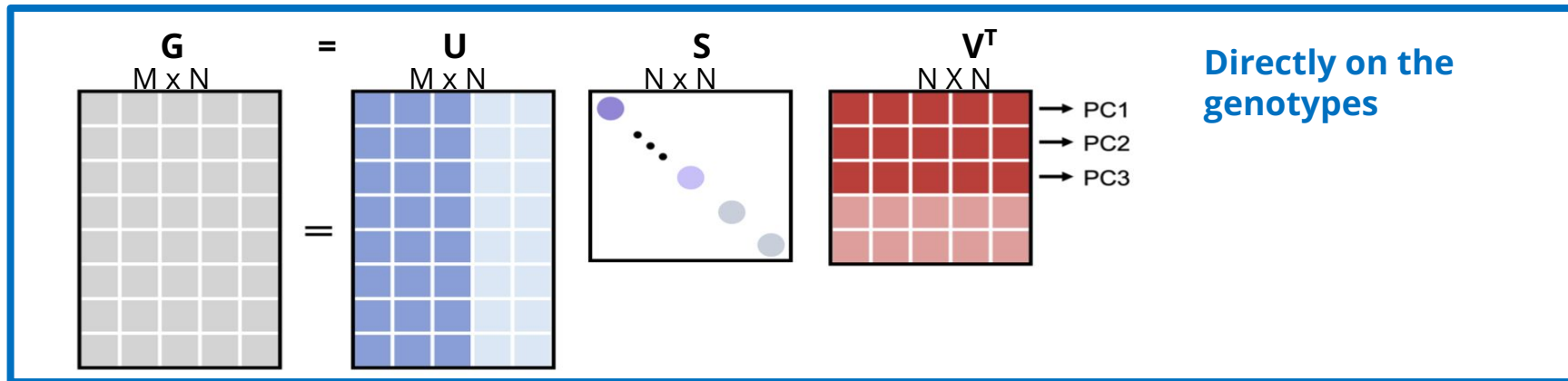
j = site j

$$\Pi = QF$$

$IB)$



Individual allele frequencies from PCA



Individual allele frequencies

- Allele frequency:

$$\mathbb{E}[g] = 2p$$

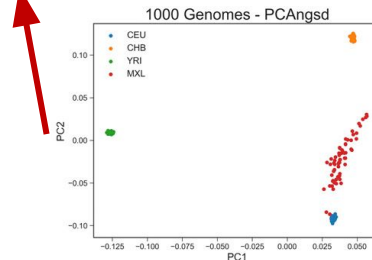
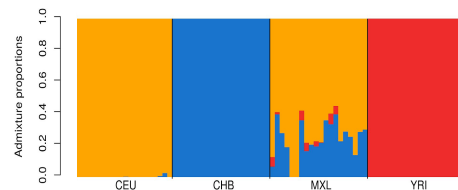
- Low-rank approximation

- Admixture

$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{Q}\mathbf{F}$$

- PCA
Truncated
SVD

$$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{U}_{[1:k]} \mathbf{S}_{[1:k]} \mathbf{V}_{[1:k]}^T$$



Admixture and PCA from Π

Π is the matrix of individual frequencies

ADMIXTURE \rightarrow PCA

$$\begin{aligned} & cov(\tilde{G}_i, \tilde{G}_l) \\ & \approx \frac{1}{M} \sum_{j=1}^M \frac{(\Pi_{ij} - f_j)(\Pi_{lj} - f_j)}{f_j(1 - f_j)} \\ & \approx \frac{1}{M} \tilde{G} \tilde{G}^T \end{aligned}$$

PCA \rightarrow ADMIXTURE

$$\operatorname{argmin}_{Q, F} \|\Pi - QF\|$$

Solved with NMF

Time for exercises

Run the admixture notebook