Population structure II
PCA

Anders Albrechtsen

The idea in general

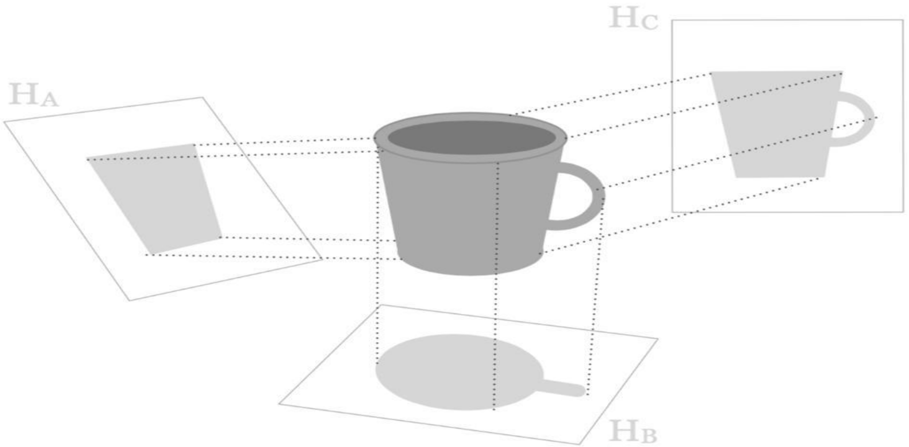
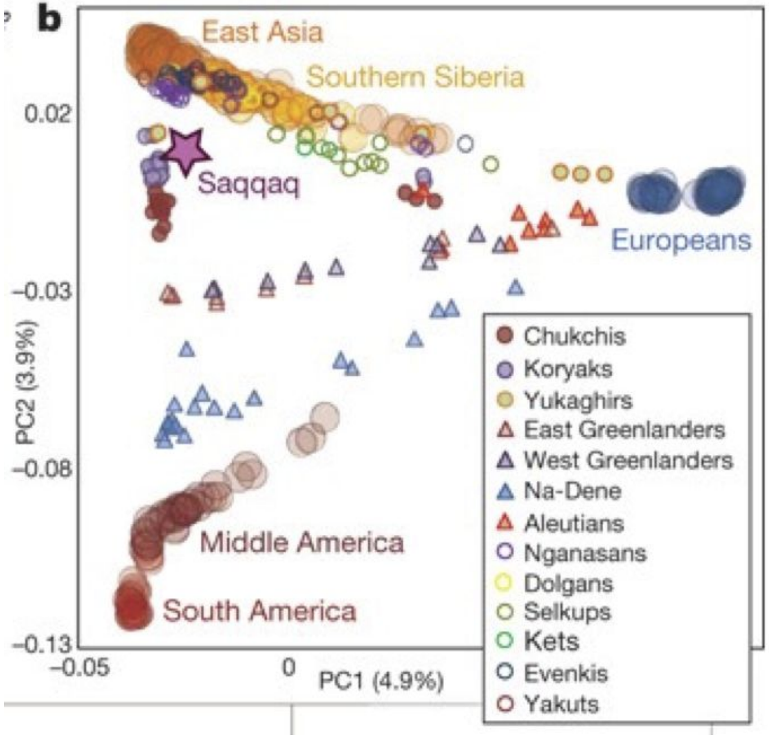


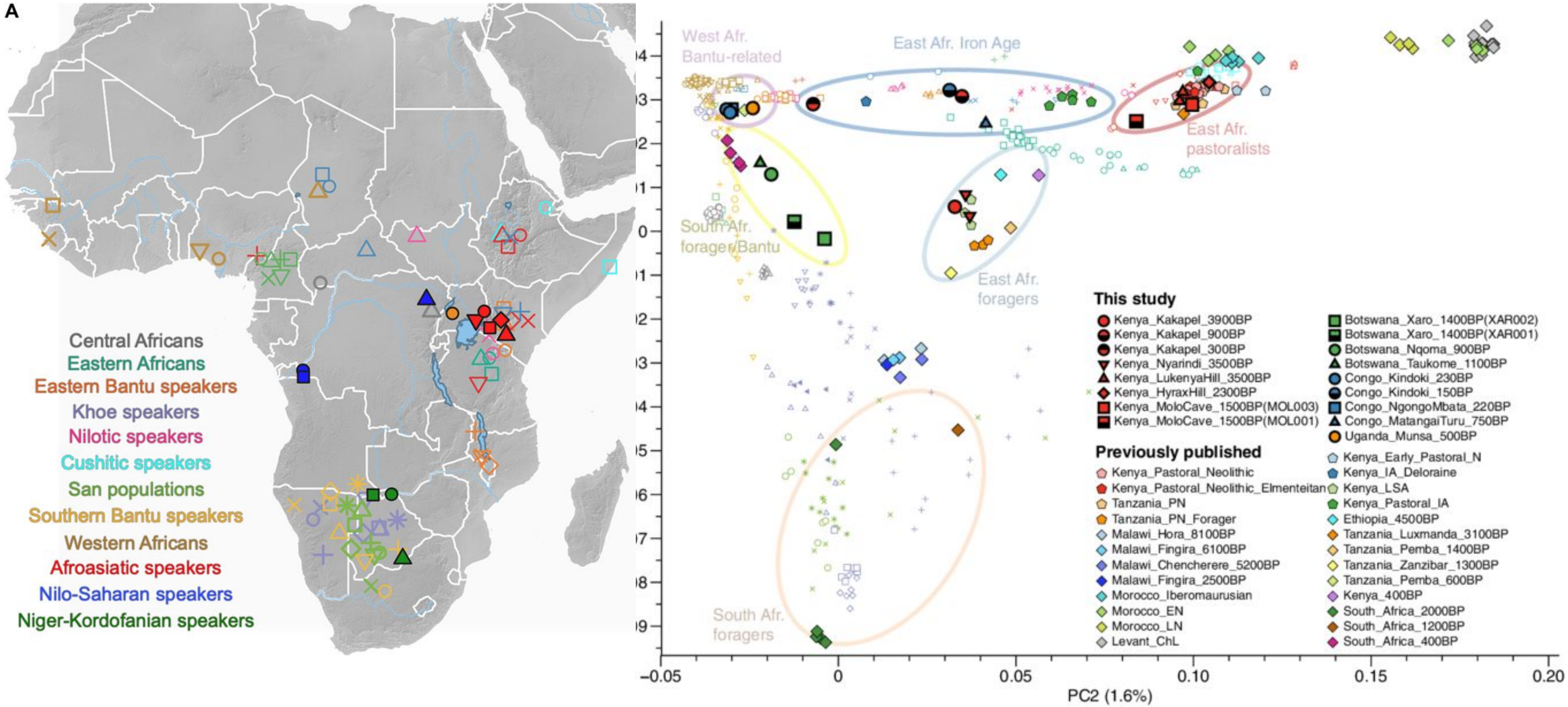
Figure 1.6: Three projections of a mug-shaped cloud points

It matters which of the three we pick, right?

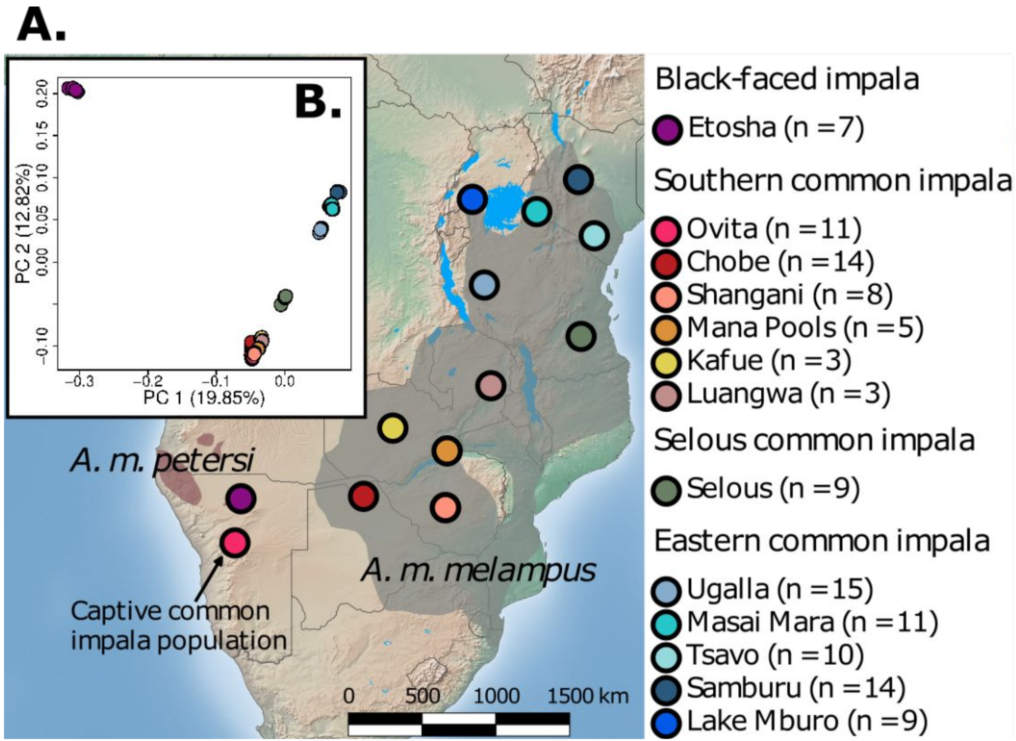
We want it to reflect the genetic relationship between all pairs

PCA for population structure



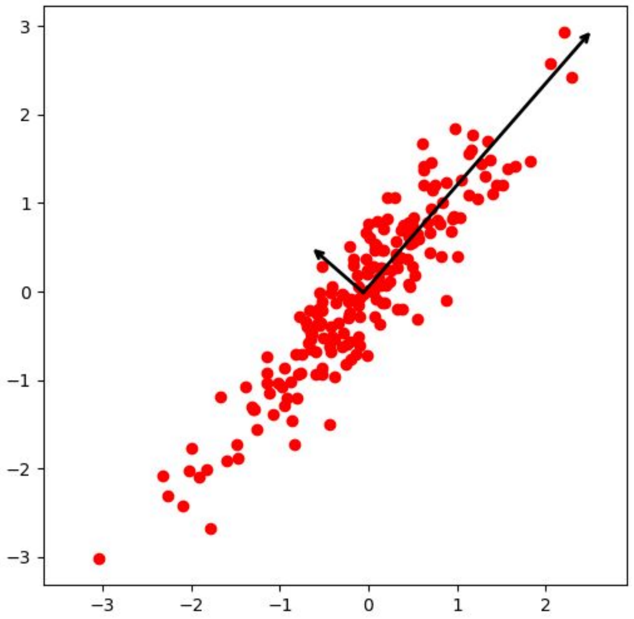

A
Central Africans
Eastern Africans
Eastern Bantu speakers
Khoe speakers
Nilotic speakers
Cushitic speakers
San populations
Southern Bantu speakers
Western Africans
Afroasiatic speakers
Nilo-Saharan speakers
Niger-Kordofanian speakers
West Afr.
Bantu-related
East Afr. Iron Age
East Afr.
pastoralists
South Afr.
forager/Bantu
East Afr.
foragers
South Afr.
foragers
This study
Kenya_Kakapel_3900BP
Kenya_Kakapel_900BP
Kenya_Kakapel_300BP
Kenya_Nyarindi_3500BP
Kenya_LukenyaHill_3500BP
Kenya_HyraxHill_2300BP
Kenya_MoloCave_1500BP(MOL003)
Kenya_MoloCave_1500BP(MOL001)
Botswana_Xaro_1400BP(XAR002)
Botswana_Xaro_1400BP(XAR001)
Botswana_Nqoma_900BP
Botswana_Taukome_1100BP
Congo_Kindoki_230BP
Congo_Kindoki_150BP
Congo_NgongoMbata_220BP
Congo_MatangaiTuru_750BP
Uganda_Munsa_500BP
Previously published
Kenya_Pastoral_Neolithic
Kenya_Pastoral_Neolithic_Elmenteitan
Tanzania_PN
Tanzania_PN_Forager
Malawi_Hora_8100BP
Malawi_Fingira_6100BP
Malawi_Chencherere_5200BP
Malawi_Fingira_2500BP
Morocco_Iberomaurusian
Morocco_EN
Morocco_LN
Levant_ChL
Kenya_Early_Pastoral_N
Kenya_IA_Deloraine
Kenya_LSA
Kenya_Pastoral_IA
Ethiopia_4500BP
Tanzania_Luxmanda_3100BP
Tanzania_Pemba_1400BP
Tanzania_Zanzibar_1300BP
Tanzania_Pemba_600BP
Kenya_400BP
South_Africa_2000BP
South_Africa_1200BP
South_Africa_400BP
–0.05
0
0.05
0.10
0.15
0.20
PC2 (1.6%)

Impala


A.
B.
PC 2 (12.82%)
0.20
0.15
0.10
0.05
0.00
−0.10
−0.3
−0.2
−0.1
0.0
PC 1 (19.85%)
A. m. petersi
A. m. melampus
Captive common impala population
0
500
1000
1500 km
Black-faced impala
Etosha (n =7)
Southern common impala
Ovita (n =11)
Chobe (n =14)
Shangani (n =8)
Mana Pools (n =5)
Kafue (n =3)
Luangwa (n =3)
Selous common impala
Selous (n =9)
Eastern common impala
Ugalla (n =15)
Masai Mara (n =11)
Tsavo (n =10)
Samburu (n =14)
Lake Mburo (n =9)

Principal component analysis (PCA)

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

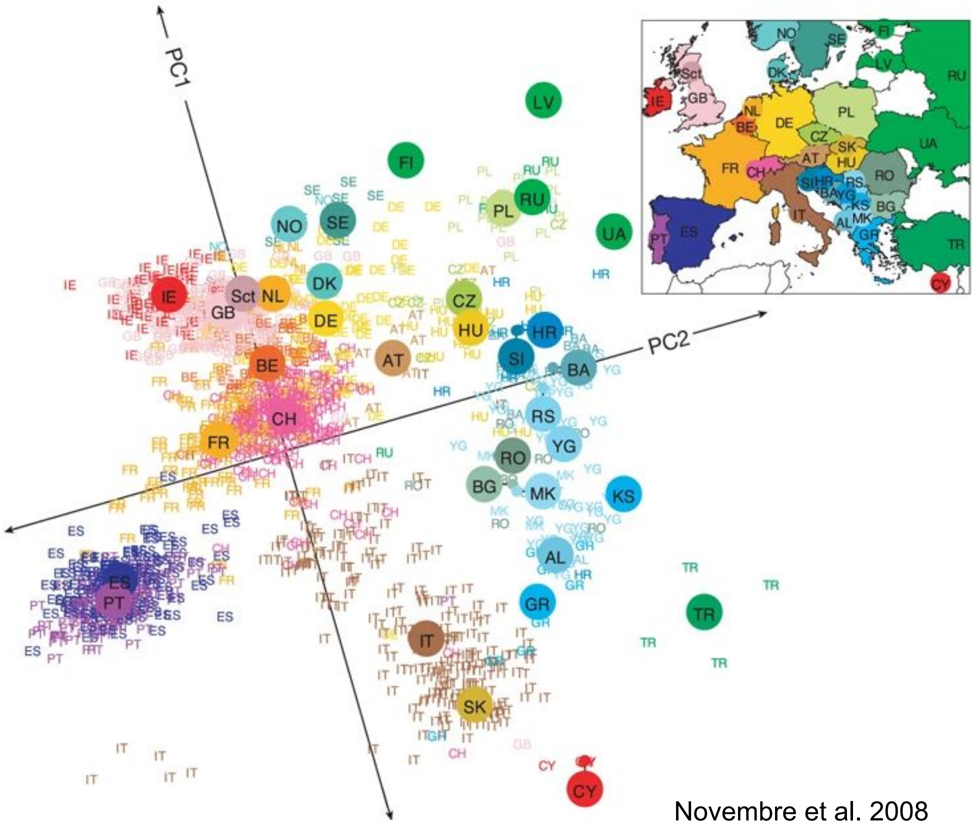


***not really true**

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

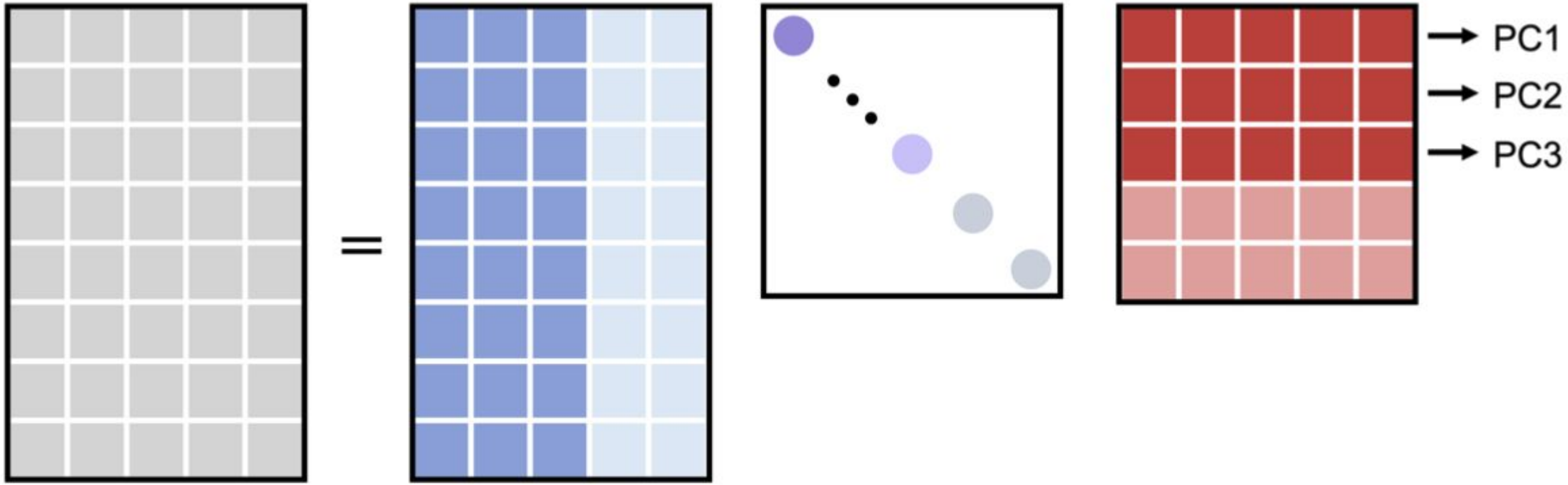
	Ind1	Ind2	Ind3	Ind4	Ind5
SNP1	AG	AG	AG	AA	AA
SNP2	TT	TA	AA	AT	AA
SNP3	AA	AC	AC	CC	AC
SNP4	GG	GG	GC	CC	CC
SNP5	TT	TC	TC	CC	CC
SNP6	AA	AA	AC	AC	AC
SNP7	TT	TT	TC	TC	CC

→

SNPs

Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	1	1	0

Genotype	left eigenvectors	singular values	right eigenvectors
G	U	Σ	V^T
$m \times n$	$m \times n$	$n \times n$	$n \times n$

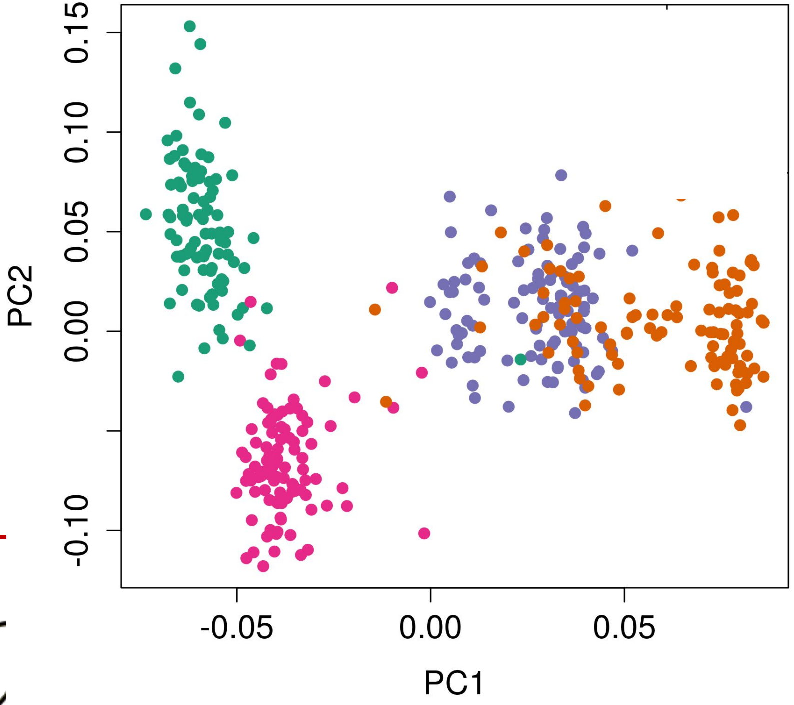


G is a genotype matrix, n is the number of samples, m is the number of SNPs

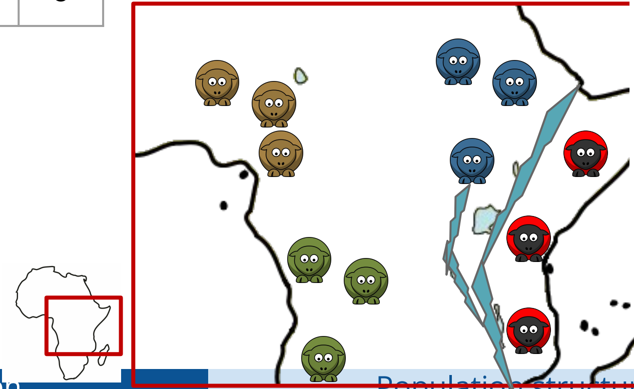
SNPs

Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	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

Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	1	1	0

Pairwise distance

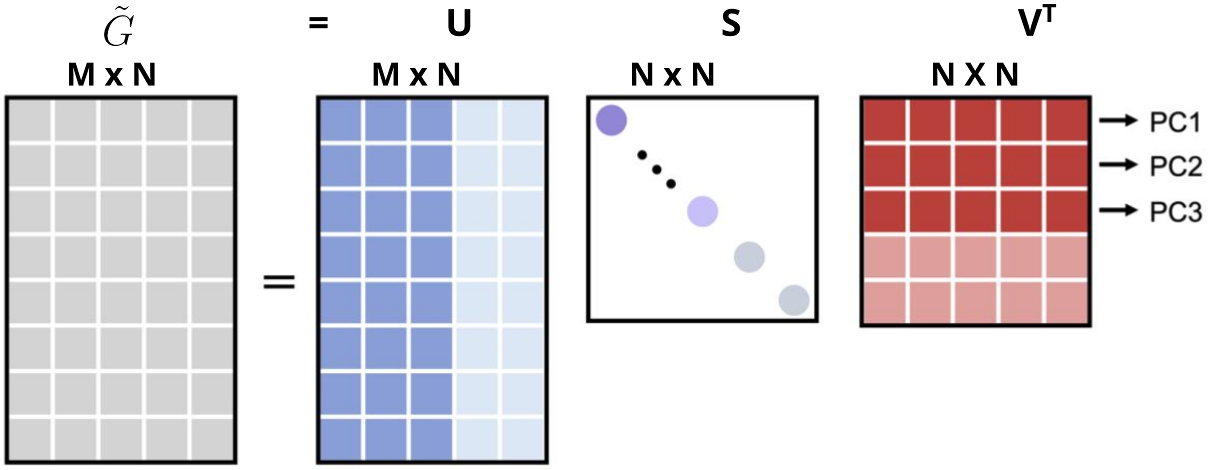
Manhattan distance

	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
Ind 1	0	3	7	10	11
Ind 2	3	0	4	7	8
Ind 3	7	4	0	5	4
Ind 4	10	7	5	0	6
Ind 5	11	8	4	3	0

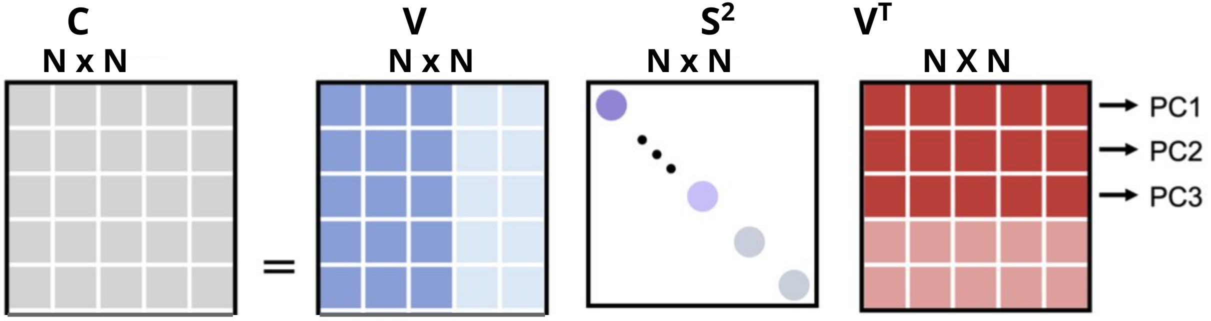
Project into 1 dimension

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

Two ways to do PCA



Directly on the genotypes



On the covariance matrix

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

Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	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

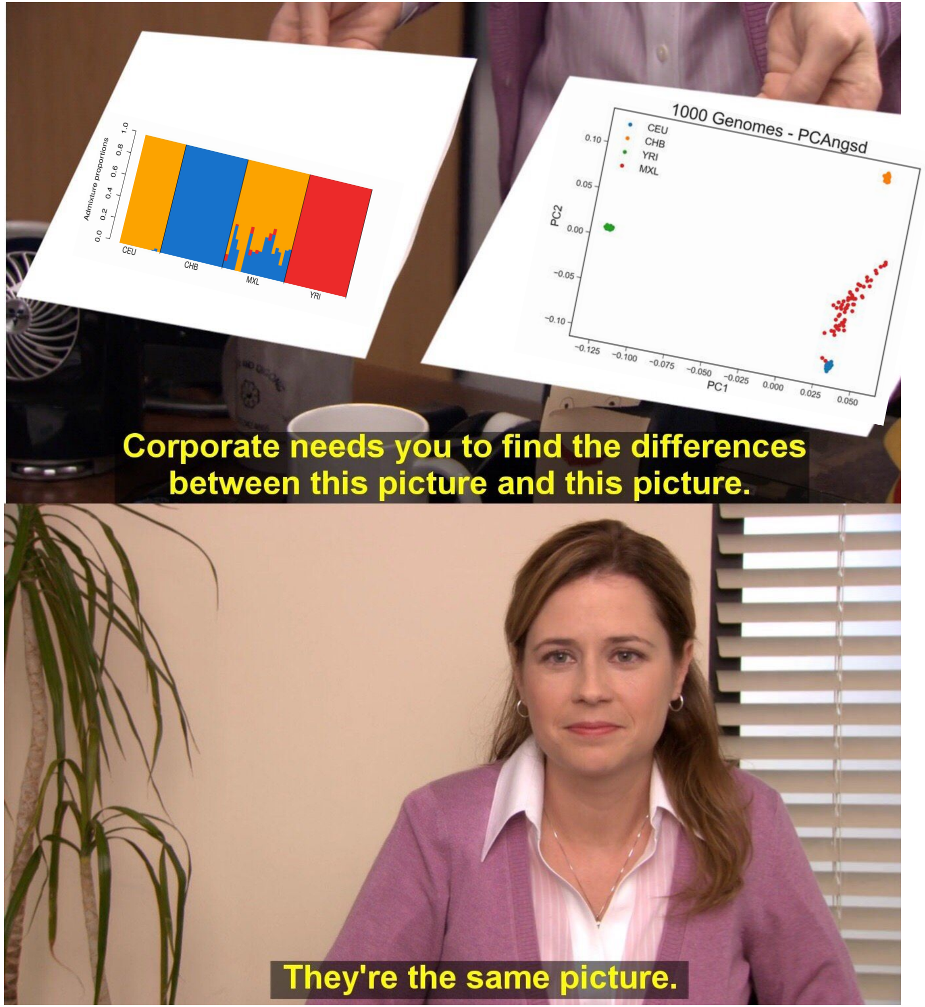
$$\tilde{G}_{ij} = \frac{G_{ij} - 2f_j}{\sqrt{2f_j(1-f_j)}}$$

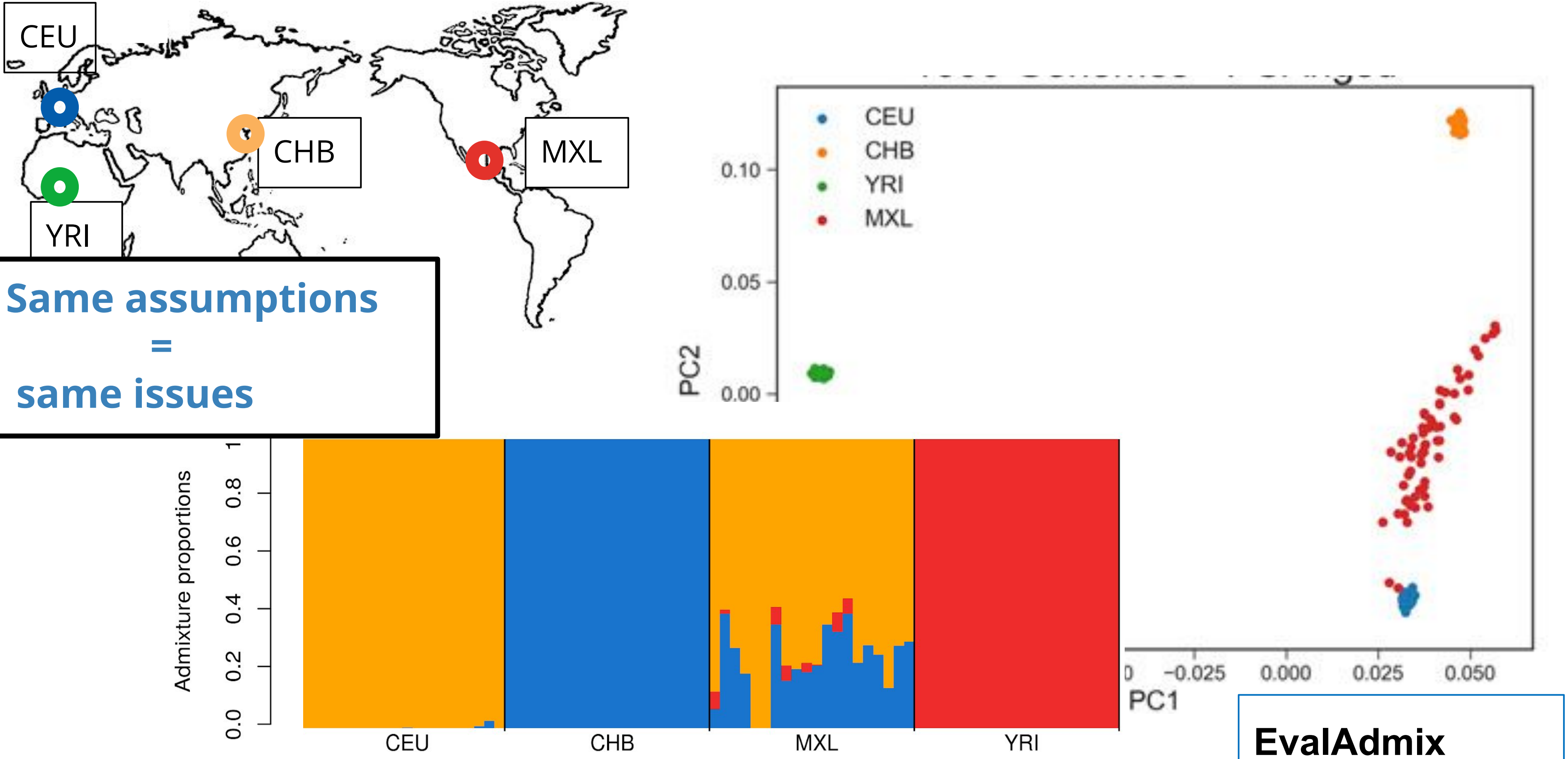
$$var(G_{ij}) = 2f_j(1-f_j)$$

After normalization all SNPs have the same mean and variance

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

Connection between Admixture analysis and PCA





Same assumptions

=

same issues

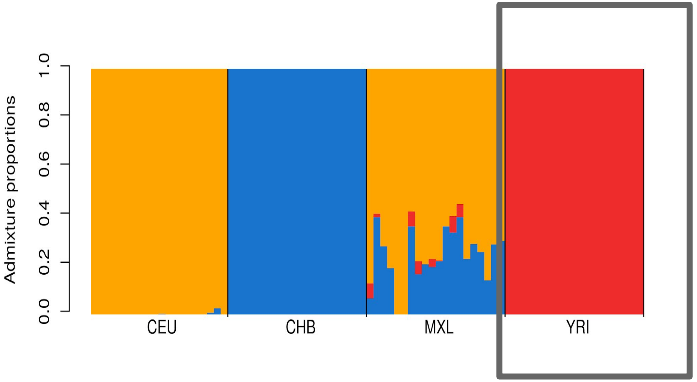
EvalAdmix

Individual allele frequencies

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

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

No admixture
p= frequency
in YRI



Individual allele frequencies

- Individu

$\mathbb{E}[g_i] = 2$ … $HB)$

$p(g_i) =$

Individual allele frequencies

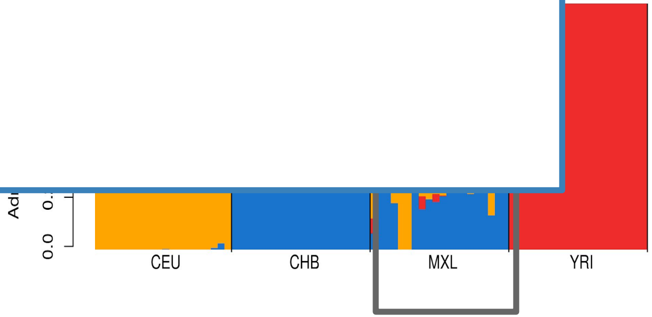
$$\pi_{ij} = \sum_{k=1}^{K} f_{jk} q_{ik}$$

$$K = \text{\#populations}$$

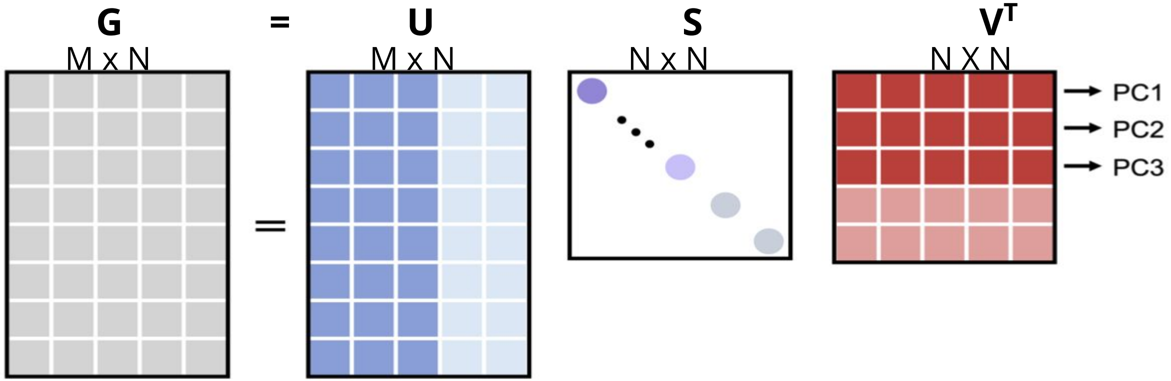
$$i = \text{individual i}$$

$$j = \text{site j}$$

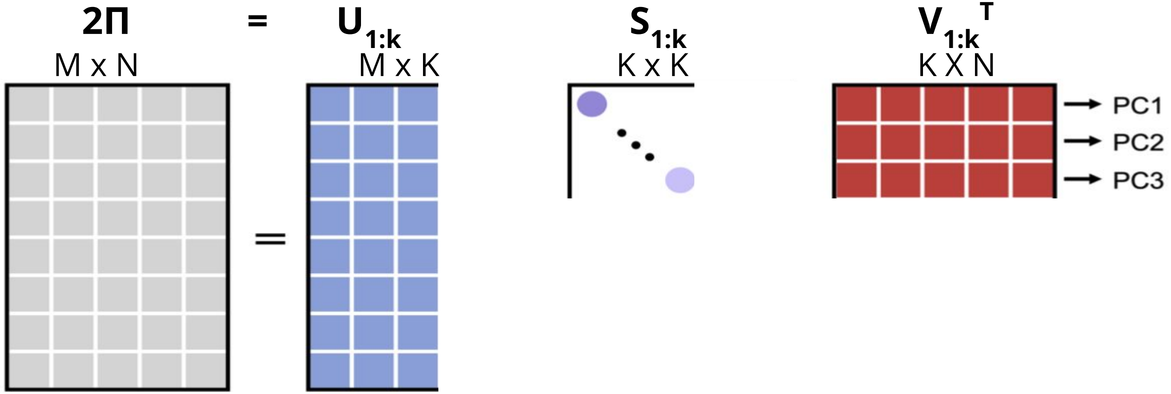
$$\Pi = QF$$



Individual allele frequencies from PCA



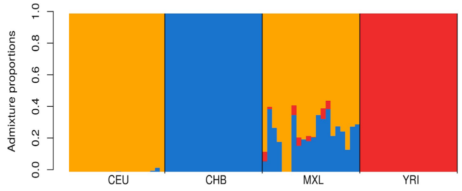
Directly on the genotypes



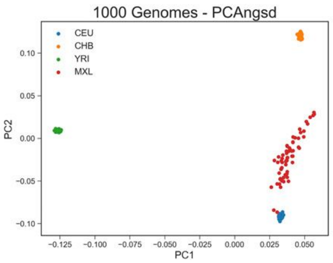
Directly on the genotypes

Individual allele frequencies

- Allele frequency: $$\mathbb{E}[g] = 2p$$



- Low-rank approximation
  - Admixture $$\frac{1}{2}\mathbb{E}[\mathbf{G}] \approx \mathbf{\Pi} = \mathbf{QF}$$
  - 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 → PCA

$$cov(\tilde{G}_i, \tilde{G}_l)$$

$$\approx \frac{1}{M} \sum_{j=1}^{M} \frac{(\Pi_{ij} - f_j)(\Pi_{lj} - f_j)}{f_j(1 - f_j)}$$

$$\approx \frac{1}{M} \tilde{G}\tilde{G}^T$$

PCA → ADMIXTURE

$$argmin_{Q,F} ||\Pi - QF||$$

Solved with NMF

Time for exercises

Run the admixture notebook