CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- Piazza counts as participation (both asking AND **answering**!)
- Tomorrow (Friday): **zoom office hours** 3:30-5pm
- **Lab 7 and project proposal** (both short)
  - Due Wednesday Nov 8

Outline for November 2

- Discuss final project
- Dimensionality reduction
- PCA for data visualization

Outline for November 2

- Discuss final project
- Dimensionality reduction
- PCA for data visualization

Timeline and Logistics

- November 8: project proposal due
- November 8 - December 7: working on projects
- December 7, 12, 14: oral project presentations during class
- December 22: github repos must be finalized

Outline for a typical project:

- Find a dataset (see project writeup)
- Run an algorithm we've discussed on the dataset
- Try to do a comparison
  - run the algorithm in multiple ways
  - different data pre-processing
  - try a different algorithm
- Evaluate, interpret, and visualize the results

Project Proposal

- **Title** and **names** of both partners
  - Pair work is required!
- A **dataset** (what is n? what is p?)
- An **algorithm** or set of algorithms you will develop and/or apply to this dataset
- A **scientific question** you are trying to answer
  - "Will Naive Bayes or logistic regression perform better on my dataset?"
  - "How will pre-processing a dataset or subsampling features affect the results?"
- A way to **evaluate, interpret, and *visualize*** the results
- **References**

Project Group Options

- If you would like a random partner, please email me ASAP!

- If you *really* prefer to work individually or in a group of 3, email me ASAP!

Final Project Deliverables

- Main deliverable: presentation
  - In class Dec 7, 12, 14 (last 1.5 weeks of classes)
  - 10 min per pair
  - Peer feedback

- On git (by Dec 22)
  - Lab Notebook (in README.md)
  - Project Code
  - Slides

Project Lab Notebook

- As you accept your git repo, start creating a "lab notebook" in your **README.md**
- This should say:
  - who was working (which partner)
  - date
  - how long
  - briefly what what accomplished

Sara: 03-07-18 (2hrs)

- now averaging the Markov chain, fixed all the results
- combined ancestral 1000 genomes still running (need to start similar for SGDP)
- started new runs with filtering to only have selected alleles in the "selected pop" and only have ancestral alleles in the "reference panel"

Outline for November 2

- Discuss final project
- Dimensionality reduction
- PCA for data visualization

Principal Components Analysis (PCA)

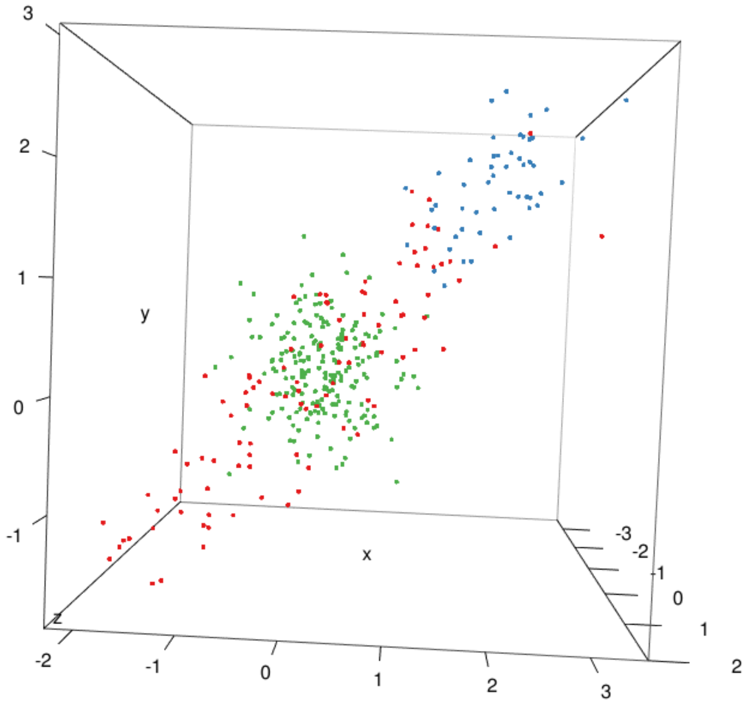
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

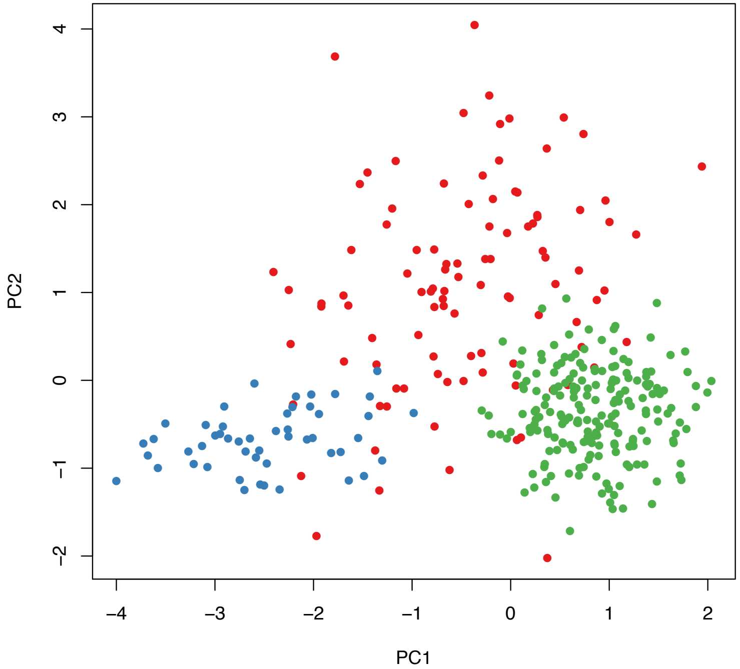
- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction and visualization

- PCA is a linear transformation

- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

Principal component analysis

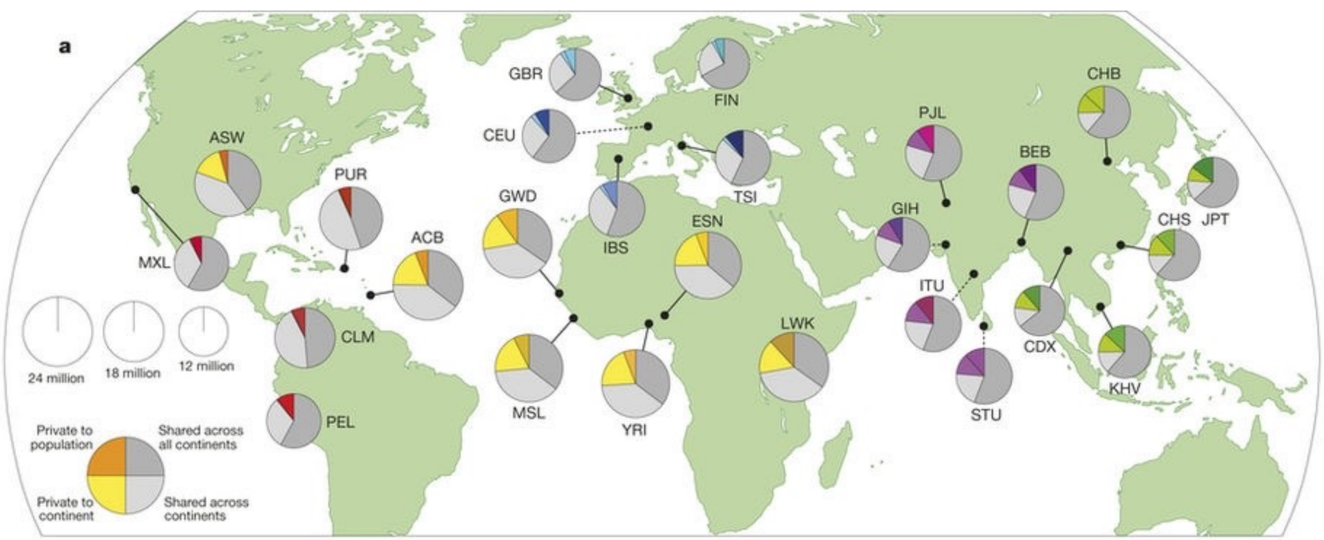




Slide: Iain Mathieson

The 1000 Genomes project

- Whole-genome sequence data from 2504 individuals from 26 populations
- A catalog of human genetic variation, useful as a reference or imputation panel
- Completely public. Download from ftp://ftp-trace.ncbi.nih.gov/1000genomes/

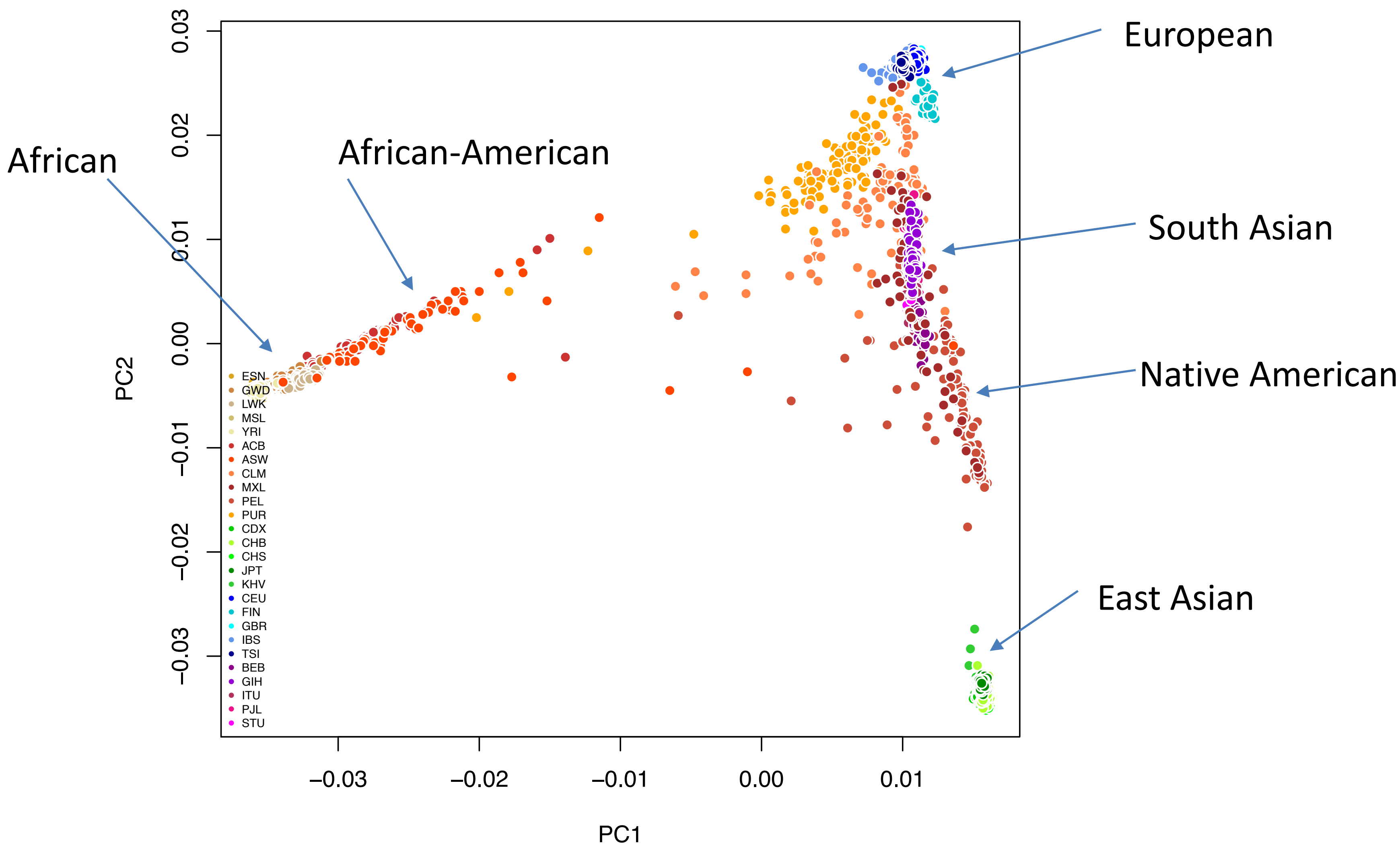


The 1000 Genomes Project Consortium, 2015; https://www.nature.com/articles/nature15393

Slide: Iain Mathieson

```
##ALT=<ID=CN120,Description="Copy number allele: 120 copies">
##ALT=<ID=CN121,Description="Copy number allele: 121 copies">
##ALT=<ID=CN122,Description="Copy number allele: 122 copies">
##ALT=<ID=CN123,Description="Copy number allele: 123 copies">
##ALT=<ID=CN124,Description="Copy number allele: 124 copies">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##bcftools_annotateVersion=1.6+htslib-1.6
##bcftools_annotateCommand=annotate -x INFO 20130502_phase3_final/ALL.chr20.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz; Date=Fri Jan 19 19:20:16 2018
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00097	HG00099	HG00100	HG00101	HG00102	HG00103	HG00105	HG00106	HG00107	HG00108	HG00109	HG00110	HG00111
20	60343	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60419	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60479	rs149529999	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60522	rs150241001	T	TC	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60568	.	A	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60571	rs116145529	C	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60579	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60649	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60778	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60795	rs184056664	G	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60808	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60810	.	G	GA	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60826	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60828	rs187713677	T	G	100	PASS	.	GT	0|0	0|1	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60864	.	G	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60895	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	60916	.	G	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61044	.	C	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61070	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61098	rs6078030	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|1	0|0	1|0	0|0	0|0	0|0	0|0	0|1	0|0
20	61118	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61138	rs140305189	C	CT	100	PASS	.	GT	0|0	0|1	0|0	0|0	0|1	0|0	0|0	0|1	0|0	0|0	0|0	0|0	0|0
20	61270	rs143291093	A	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61271	.	T	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61272	.	C	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61279	rs189899941	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61329	rs182162684	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61388	rs146681064	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61409	rs139103017	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61437	.	A	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61450	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61517	rs187280035	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61538	.	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61638	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61651	rs76553454	C	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61711	rs369824431	G	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61724	rs142532139	A	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61795	rs4814683	G	T	100	PASS	.	GT	1|0	0|0	0|0	0|0	0|1	0|0	1|0	0|1	1|0	0|0	0|0	0|1	0|0
20	61955	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	61972	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62100	rs6047235	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62174	.	AGATCAGTCCTTT	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62255	rs192879424	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62283	.	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62348	rs141113228	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62387	.	T	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62420	rs185326153	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62461	.	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62471	rs188652106	G	A	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62478	rs192812899	A	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62545	rs150267191	C	G	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
20	62553	rs114190700	T	C	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
```

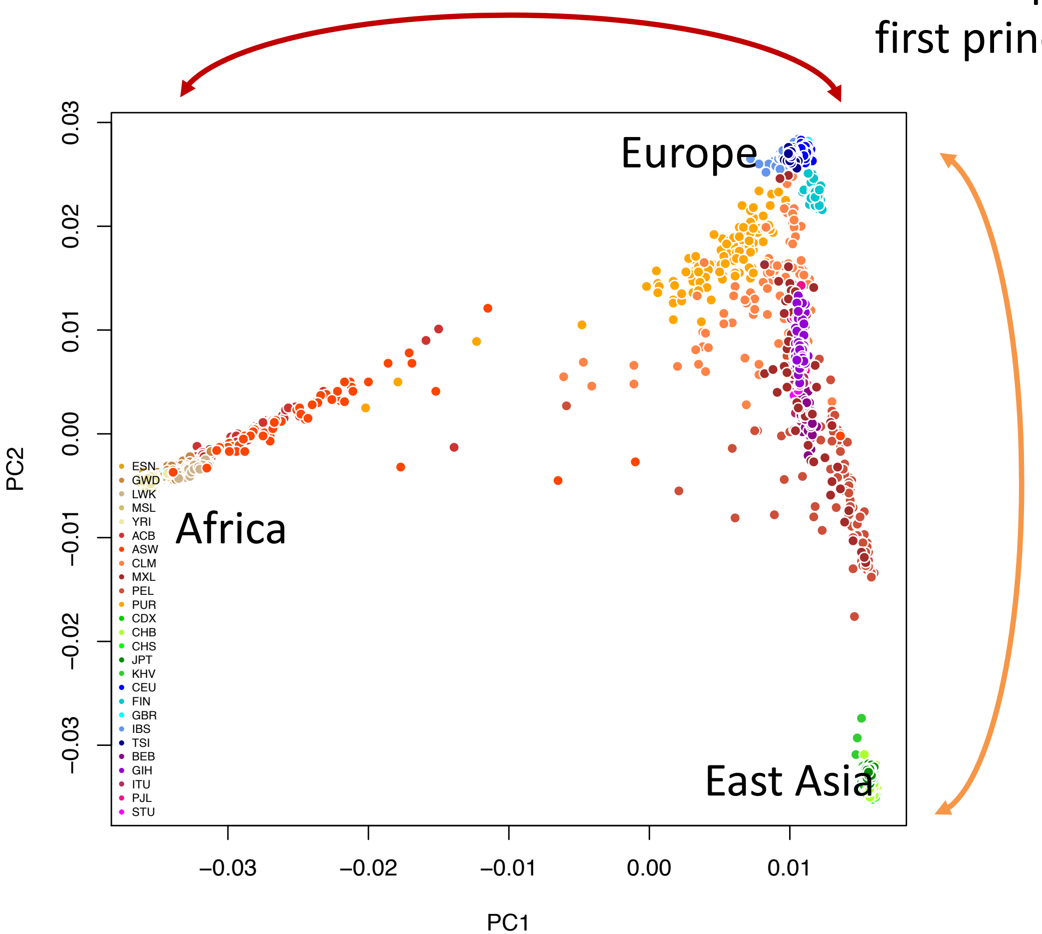
Global population structure



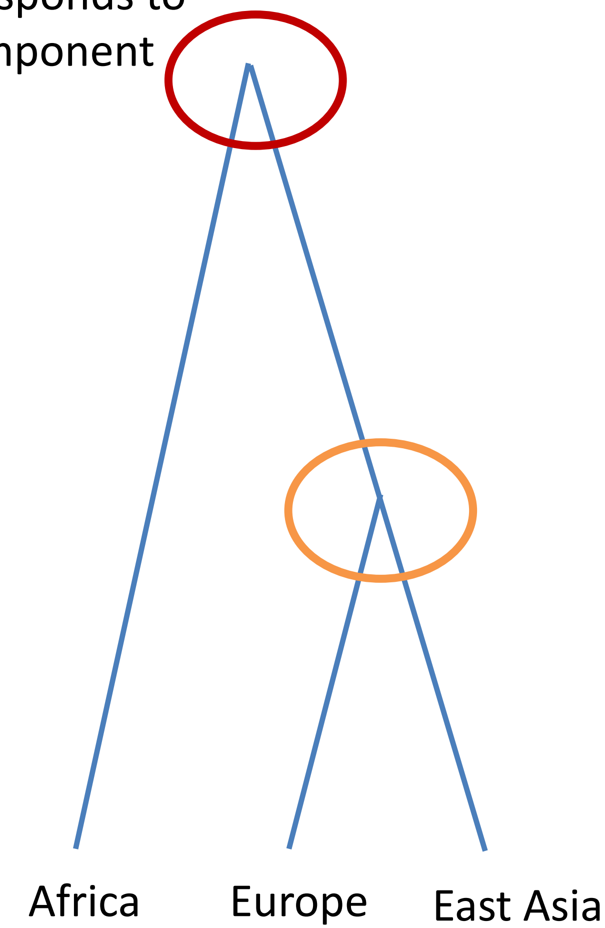
Slide: Iain Mathieson

What causes these patterns?

1. Populations splits separate populations



Oldest split corresponds to first principal component

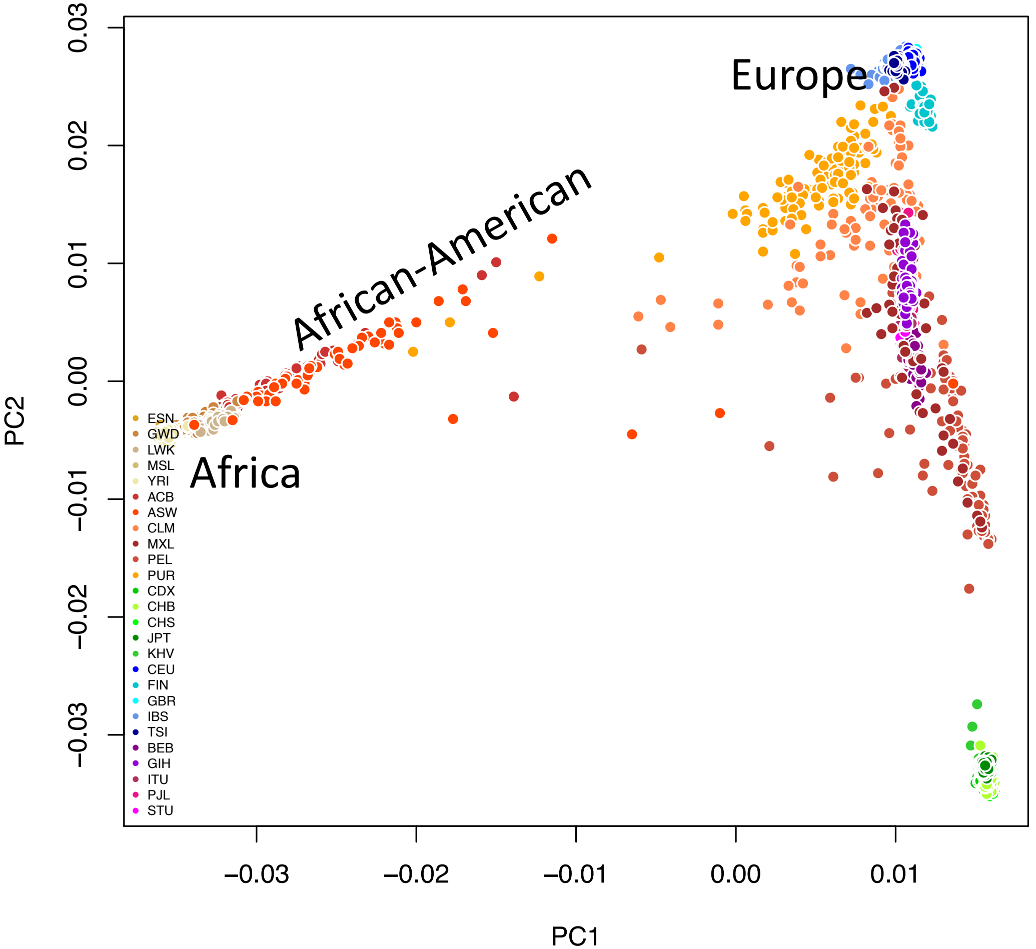


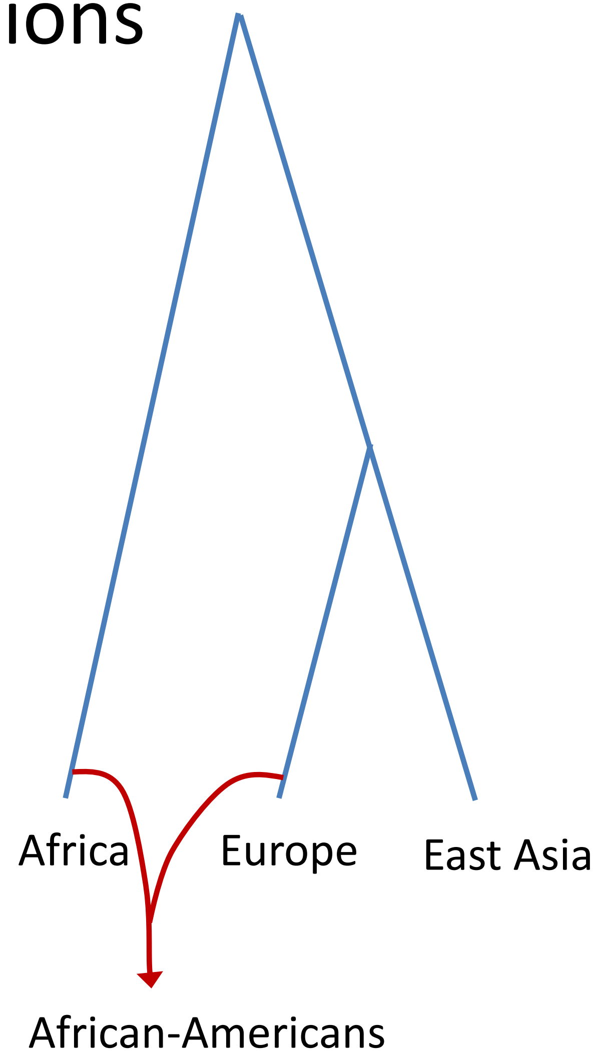
Slide: Iain Mathieson

McVean 2009: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686

What causes these patterns?

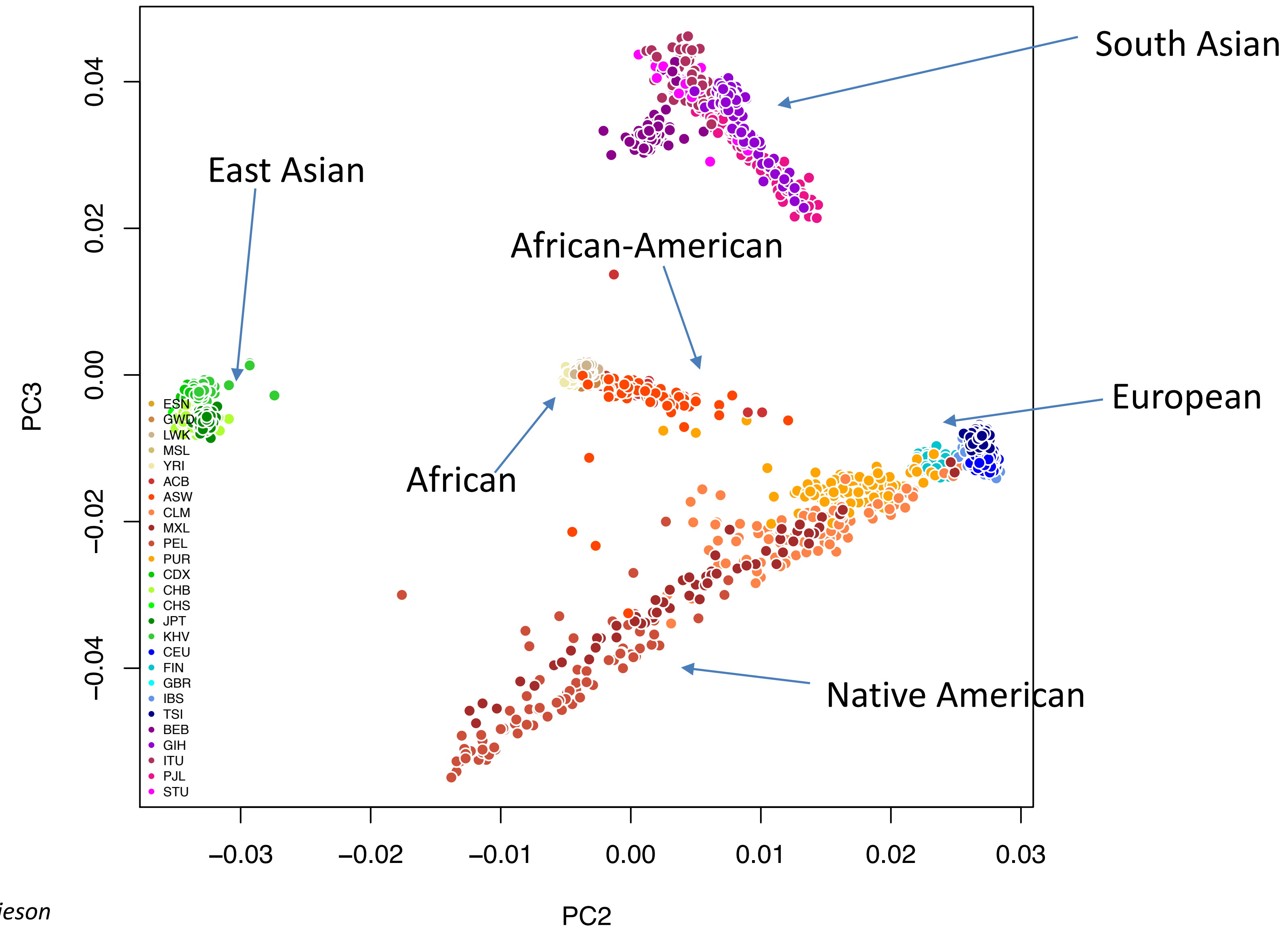
2. Admixture merges populations





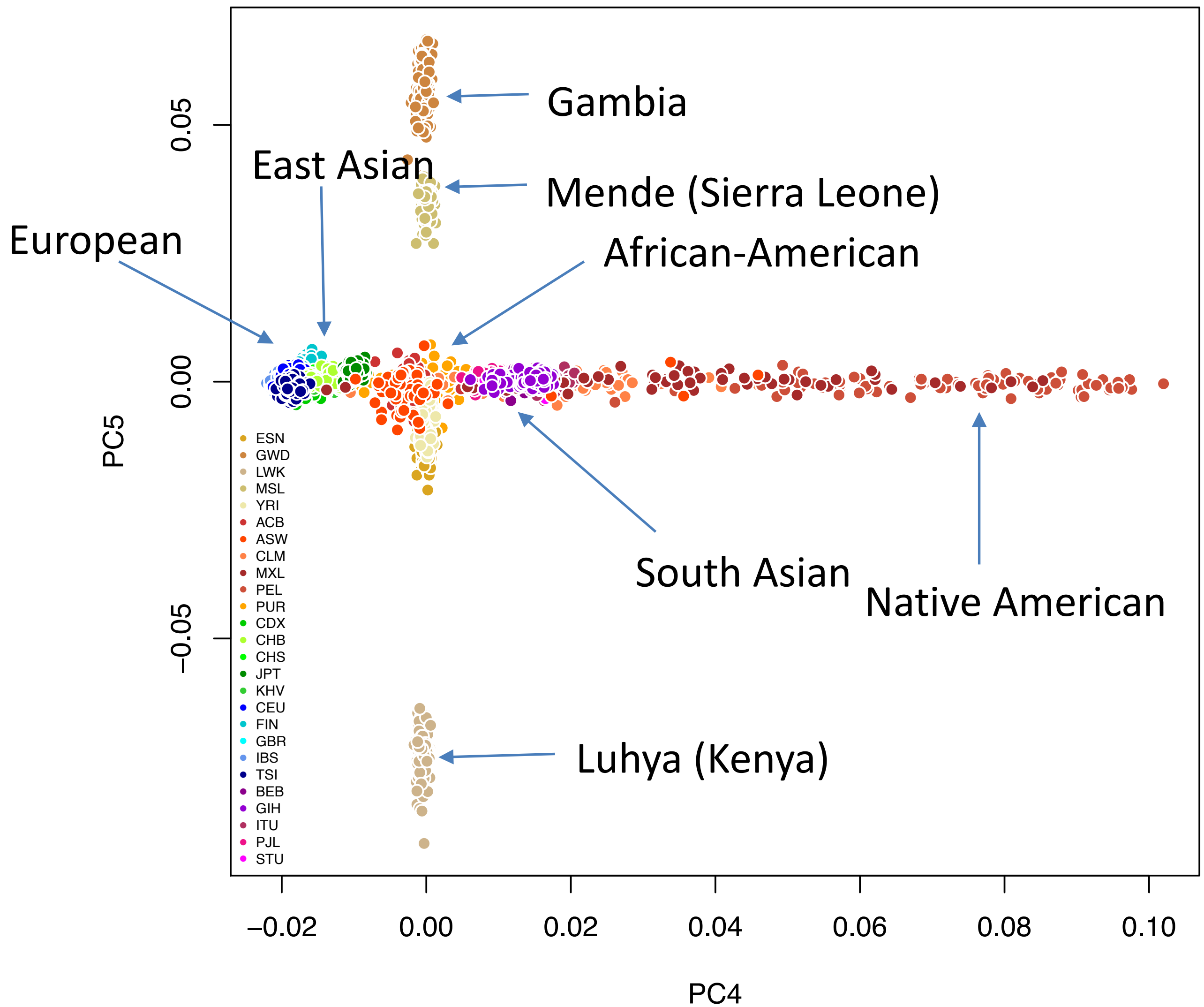
Slide: Iain Mathieson

Global population structure



Slide: Iain Mathieson

Global population structure

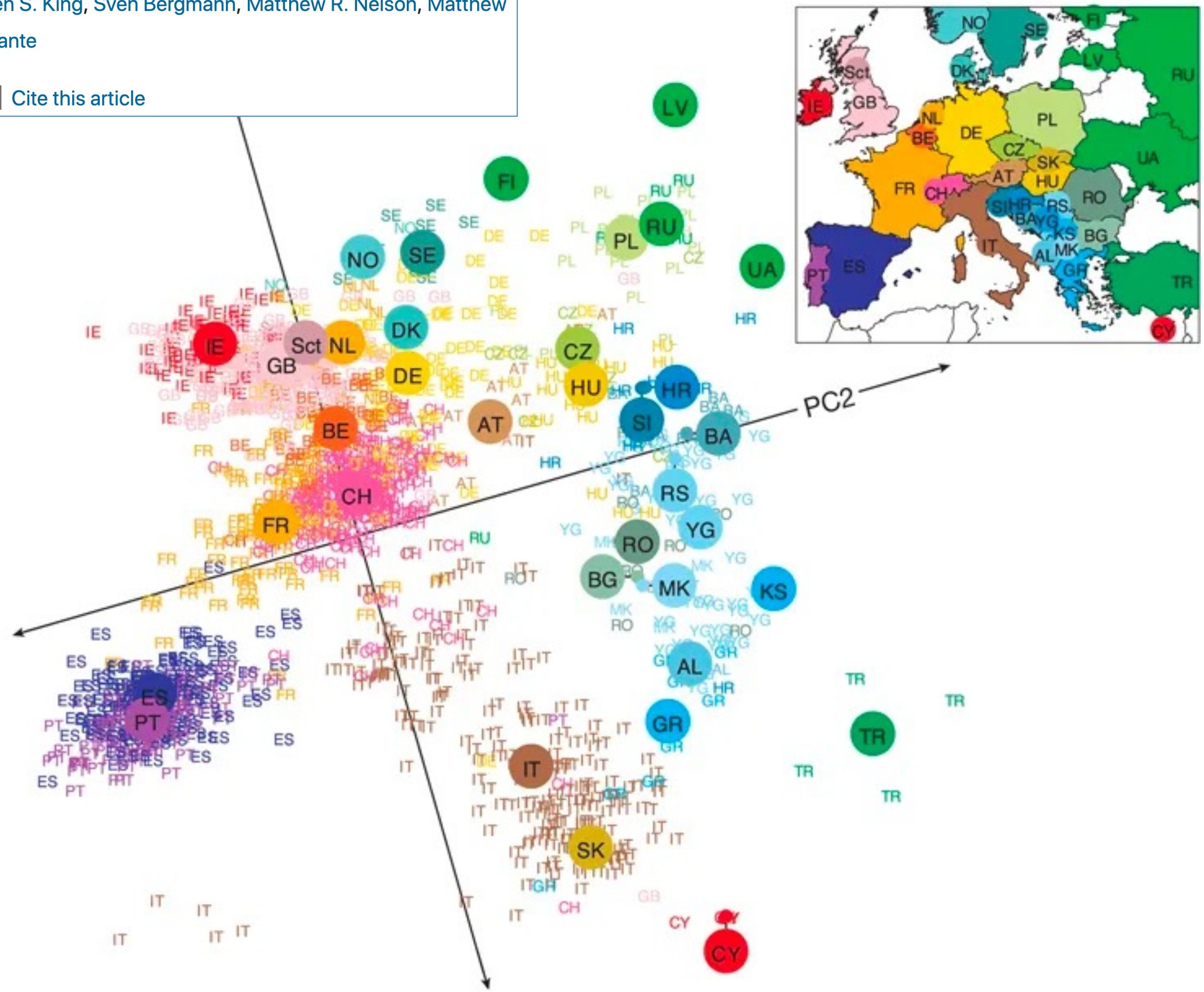


Slide: Iain Mathieson

Genes mirror geography within Europe

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101(2008) | Cite this article

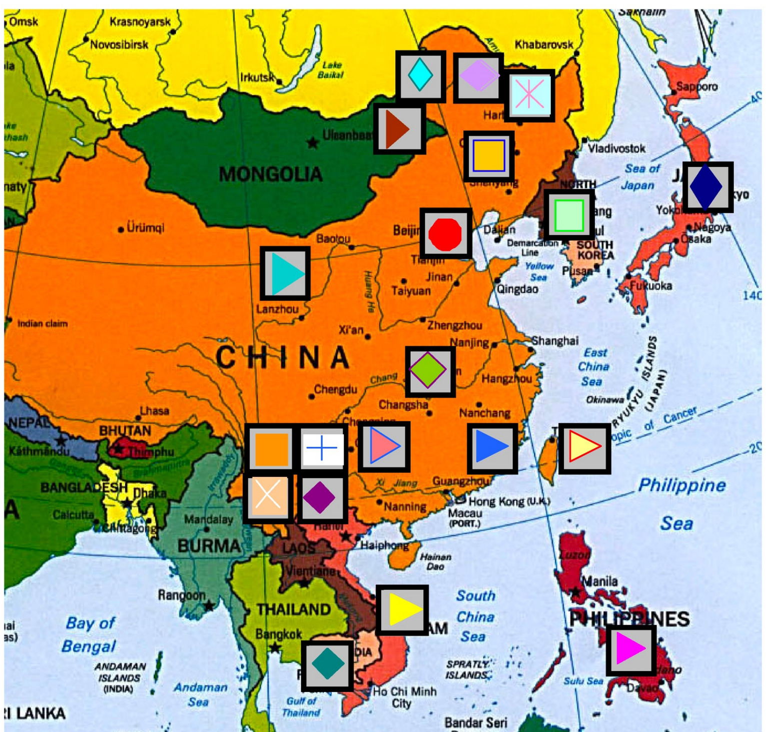


Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays

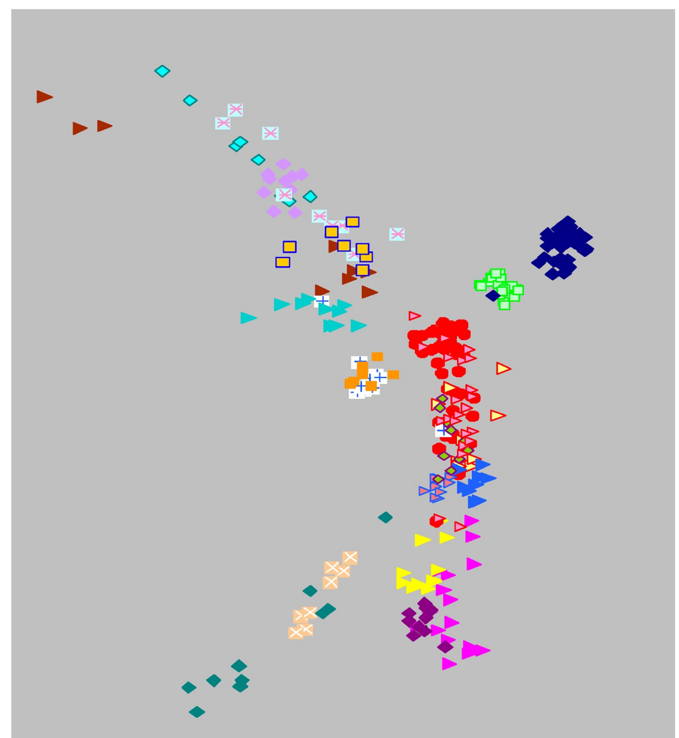
Chao Tian, Roman Kosoy, Annette Lee, Michael Ransom, John W. Belmont, Peter K. Gregersen, Michael F. Seldin

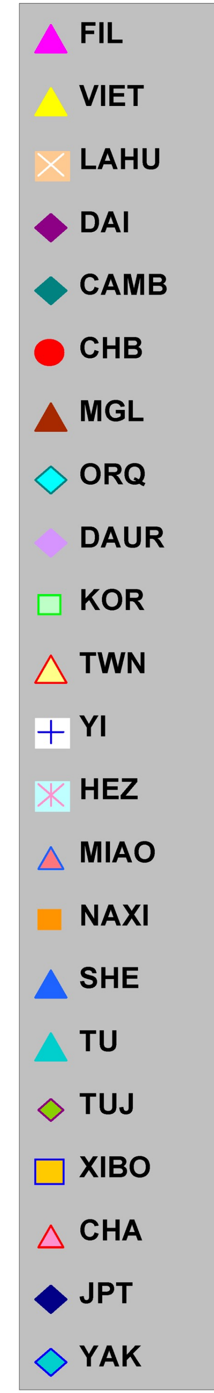
Published: December 5, 2008 • https://doi.org/10.1371/journal.pone.0003862

C



D

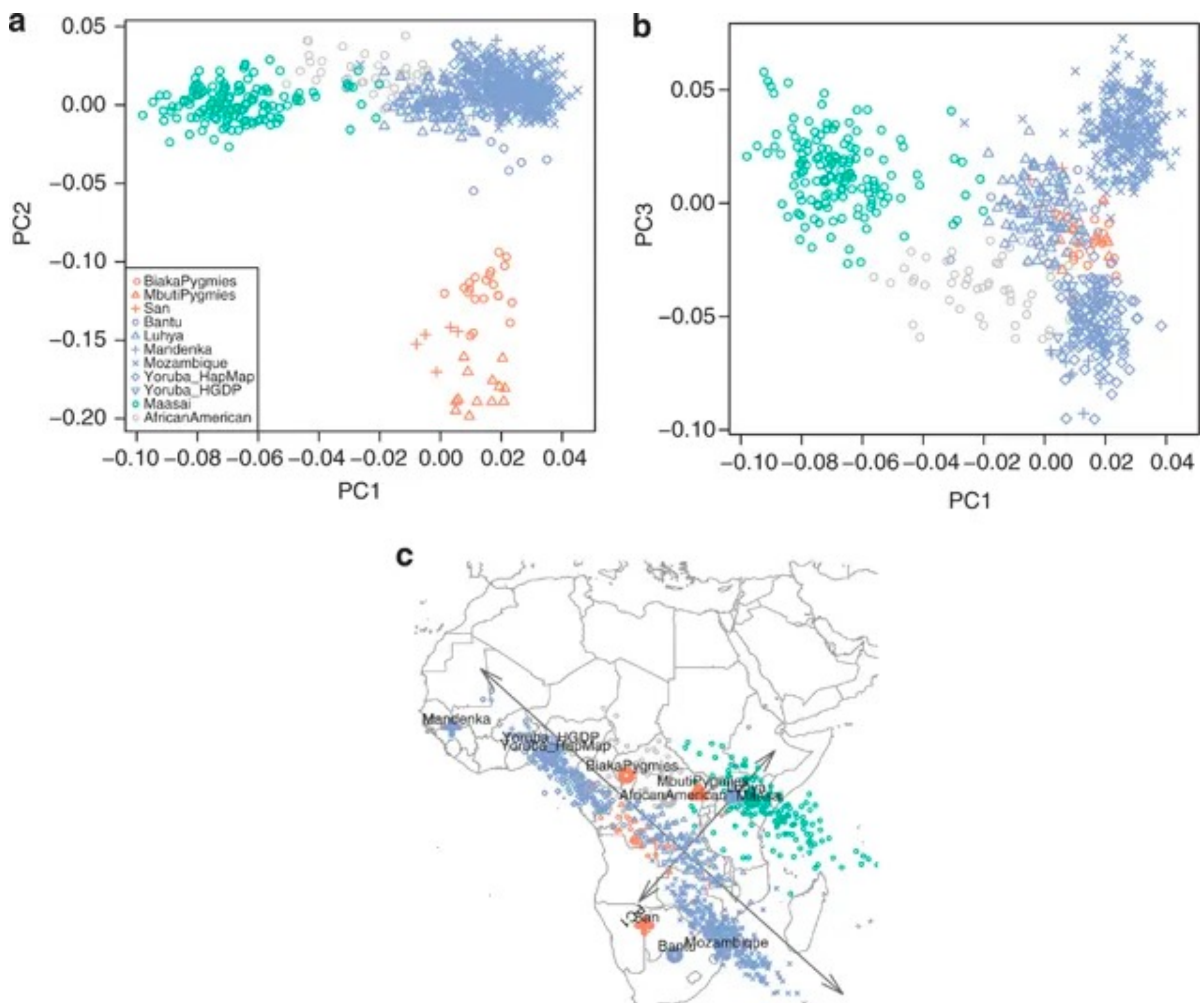




Article | Published: 25 August 2010

A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations

Martin Sikora, Hafid Laayouni, Francesc Calafell, David Comas & Jaume Bertranpetit

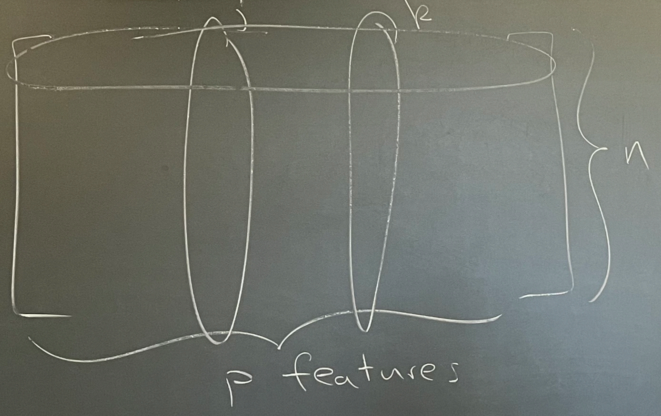


Outline for November 2

- Discuss final project
- Dimensionality reduction
- PCA for data visualization

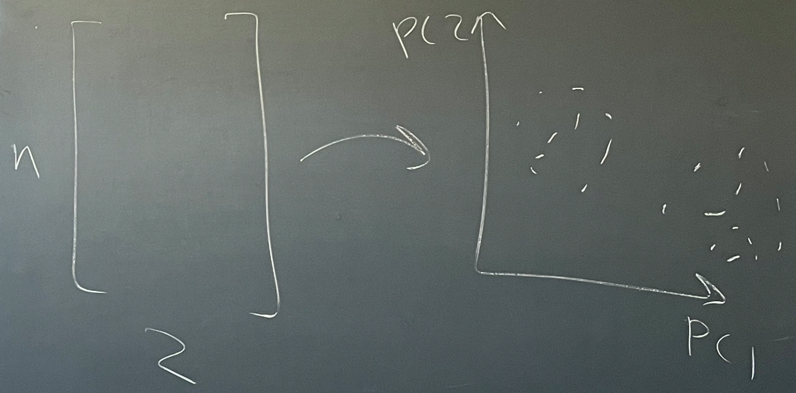
Step 1 Principal Component Analysis

$X_{orig} =$



$p >> n$

goal: create $n \times 2$ matrix for visualization

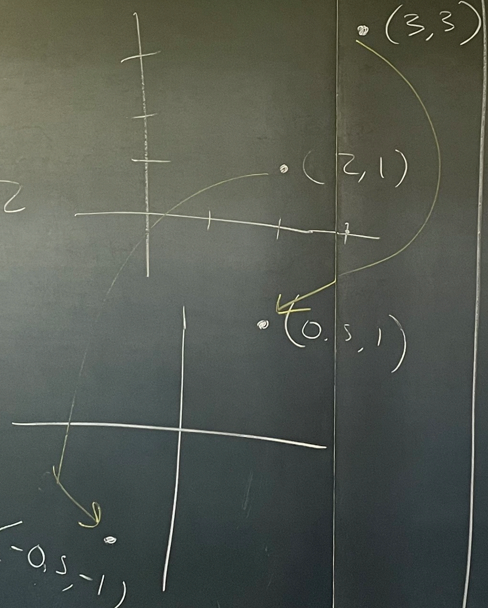


Step 2 subtract off column-wise mean

$$X_{orig} = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \end{array}$$

$$\bar{x}_1 = 2.5 \qquad \bar{x}_2 = 2$$

$$X = \begin{bmatrix} -0.5 & -1 \\ 0.5 & 1 \end{bmatrix}$$



Step 3

Compute covariance matrix A

$$\text{cov}(f,g) = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})(g_i - \bar{g})$$

$$\text{cov}(f,f) = \text{var}(f) = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2$$

2 features

$$\boxed{\text{runtime} \Rightarrow O(np^2)}$$

$$A = \begin{bmatrix} \text{cov}(f,f) & \text{cov}(f,g) \\ \text{cov}(g,f) & \text{cov}(g,g) \end{bmatrix}_{p \times p}$$

Square!

Symmetric!

Step 4

compute
eigenvalues & eigenvectors
of A

$\Rightarrow$ sort by eigenvalue
high $\rightarrow$ low

$$A\vec{v} = \lambda \vec{v}$$

eigenvalue

eigenvector

$$\det(A - \lambda I) = \vec{0}$$

Solve for λ

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$

plug back in!

Step 5

$$W = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_r \\ | & | & & | \end{bmatrix}_{p \times r}$$

λ_1, λ_2

first eigenvector

usually $r=2$

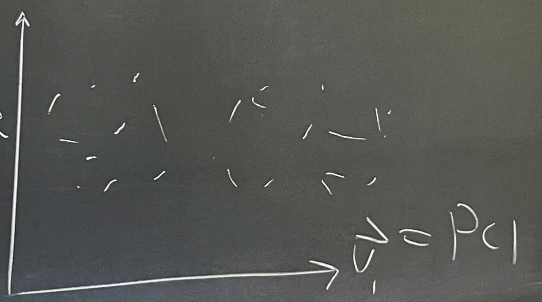
☆

$$\Rightarrow T_{n\times r} = X_{n\times p} W_{p\times r}$$

transformed data!

Step 6

$PC2 = \vec{v}_2$



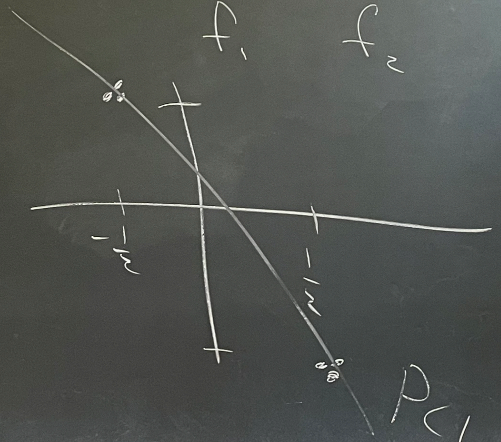
Handout 16

Handout 16

$\boxed{\text{Step 1 \& 2}}$

$$X = \begin{array}{cc} \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \\ \quad f_1 \qquad f_2 \end{array}$$

$$\text{orig}\left\{ \begin{array}{l} \bar{f}_1 = \frac{1}{2} \\ \bar{f}_2 = \frac{1}{2} \end{array} \right.$$



PC1

$\boxed{\text{Step 3}}$

$$A = \begin{bmatrix} var(f_1) & cov(f_1, f_2) \\ cov(f_2, f_1) & var(f_2) \end{bmatrix}$$

$\bar{f}_1 = 0$

$\bar{f}_2 = 0$

$$cov(f_1, f_2) = \frac{1}{6-1}\left(-\frac{1}{2} \cdot \frac{1}{2}\right) \cdot 6$$

$$= -\frac{3}{10}$$

$$\Rightarrow A = \begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix}$$

$\boxed{\text{Step 4}}$

$\det(A - \lambda I) = 0$

$\left(\begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$

$\det\left(\begin{bmatrix} 3/10 - \lambda & -3/10 \\ -3/10 & 3/10 - \lambda \end{bmatrix}\right) = 0$

$\left(\frac{3}{10} - \lambda\right)^2 - (3/10)^2 = 0$

$\cancel{\left(\frac{3}{10}\right)^2} - 2 \cdot \frac{3}{10}\lambda + \lambda^2 - \cancel{(3/10)^2} = 0$

$\frac{1}{2}) \cdot 6$

$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$

$\Rightarrow \boxed{\lambda^2 - \frac{3}{5}\lambda = 0}$

$\lambda(\lambda - \frac{3}{5}) = 0 \Rightarrow \boxed{\begin{aligned} \lambda_1 &= \frac{3}{5} \\ \lambda_2 &= 0 \end{aligned}}$

$A\vec{v} = \lambda\vec{v}$

$$T_2 = XW_2 = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \overset{\lambda_1 = 3/5 \quad \lambda_2 = 0}{\underset{\vec{v}_1 \qquad \vec{v}_2}{\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$