

# CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021



**HVERFORD**  
COLLEGE

- **Note-taker:** Pelagia
- Piazza counts as participation (both asking AND **answering!**)
- **Lab 6** due TODAY
  - Will discuss Lab 6 during lab today (or start Lab 7)
  - TA hours 4:30-6:30pm today (Yuxuan)
- **Lab 7 and project proposal** (both short)
  - Due next Thurs Nov 4

- **Plan for the rest of the semester**
  - assuming nothing unexpected happens!
- In-person for the next 3 classes
  - today and next week
- 3 classes over zoom
- Prof. Farias will take over after that
  - One week on unsupervised learning, clustering, etc
  - One week on neural networks, deep learning, applications, etc
  - Last week: final project presentations

# Outline for October 28

- Discuss final project
- Finish intro to visualization
- Dimensionality reduction
- PCA for data visualization

# Outline for October 28

- Discuss final project
- Finish intro to visualization
- Dimensionality reduction
- PCA for data visualization

# Timeline and Logistics

- November 4: project proposal due
- November 4 - December 7: working on projects
- December 7 & 9: oral project presentations during class

## 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**

# Final Project Deliverables

- Main deliverable: presentation
  - In class Dec 7 and 9 (last week of classes)
  - 10 min per pair
  - Peer feedback
- On git:
  - 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

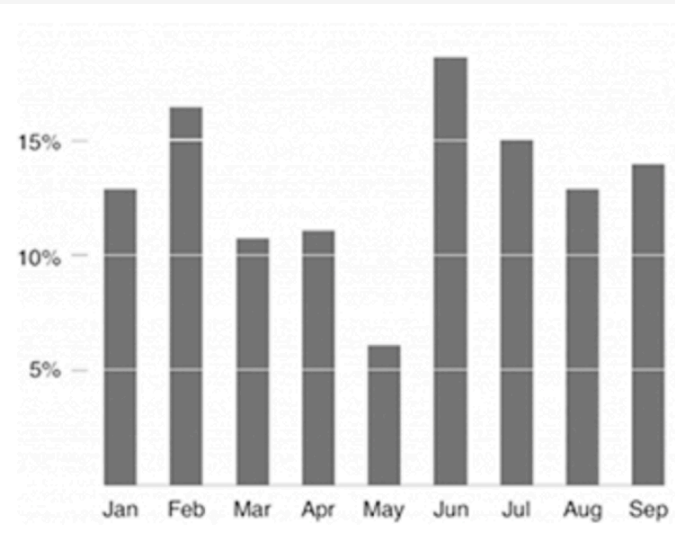
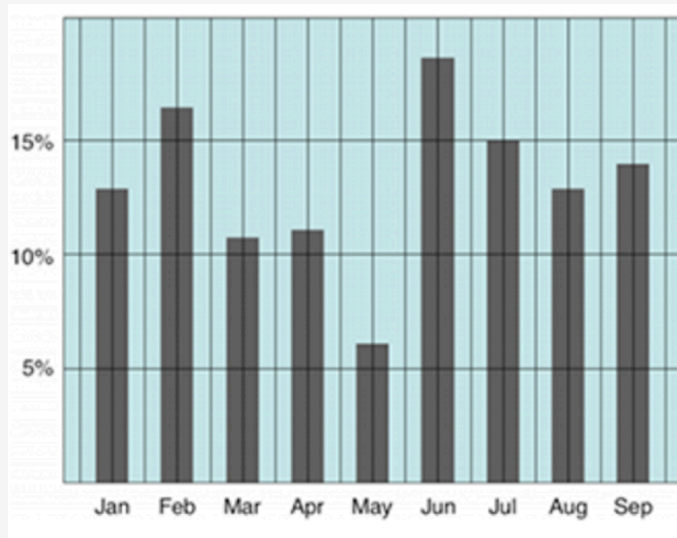
## **Sara: 03-07-18 (2hrs)**

- 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”

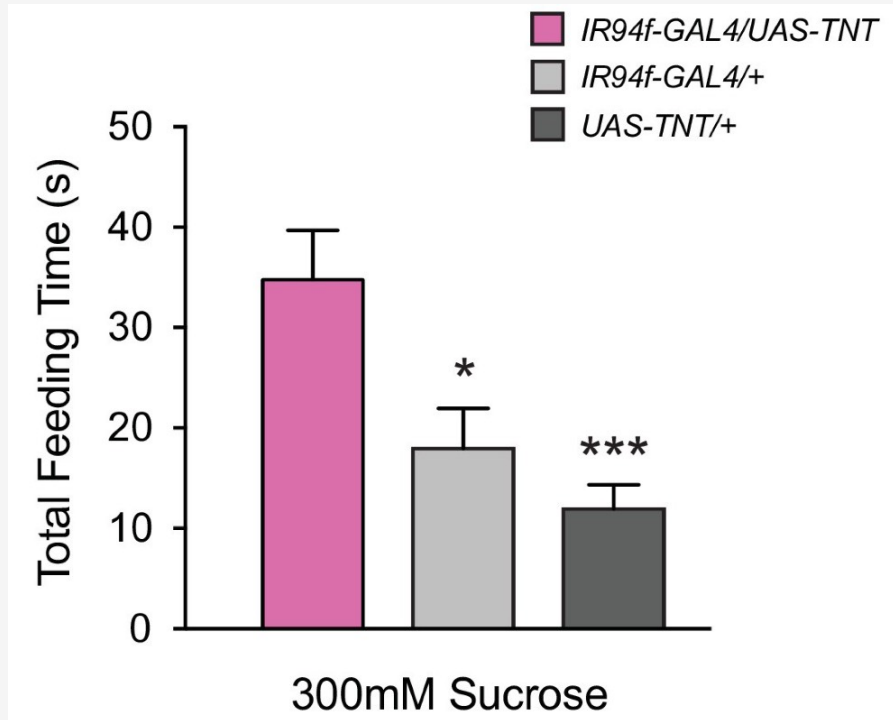
# Outline for October 28

- Discuss final project
- **Finish intro to visualization**
- Dimensionality reduction
- PCA for data visualization

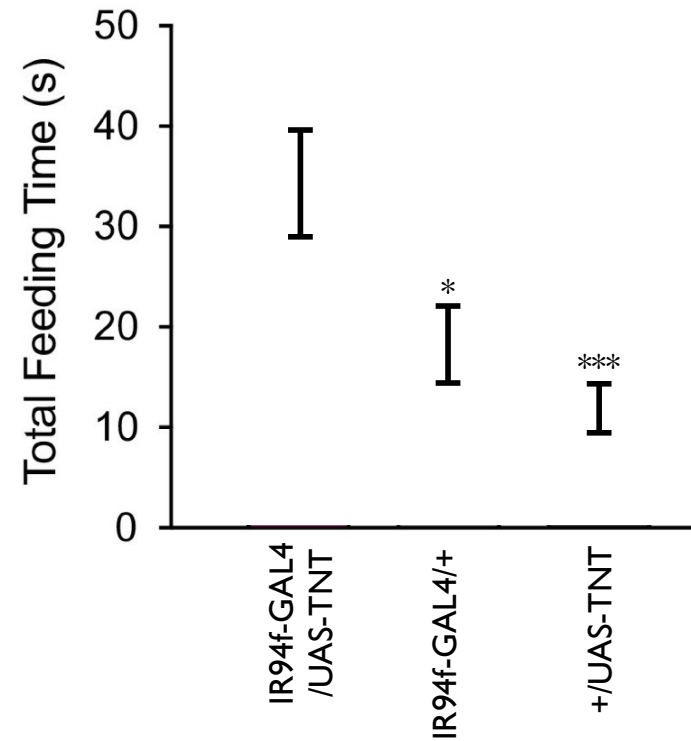
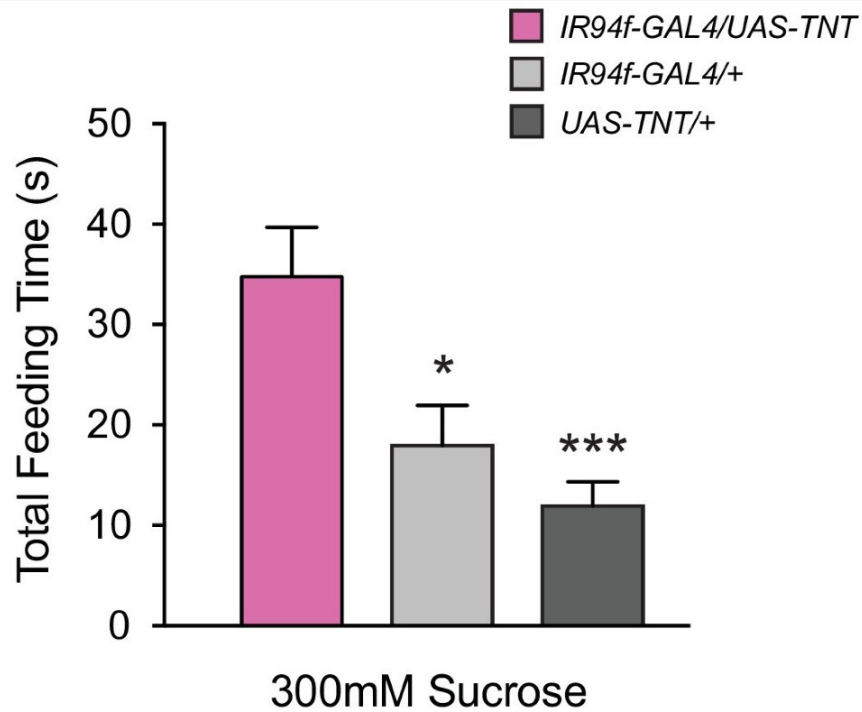
# Data::Ink



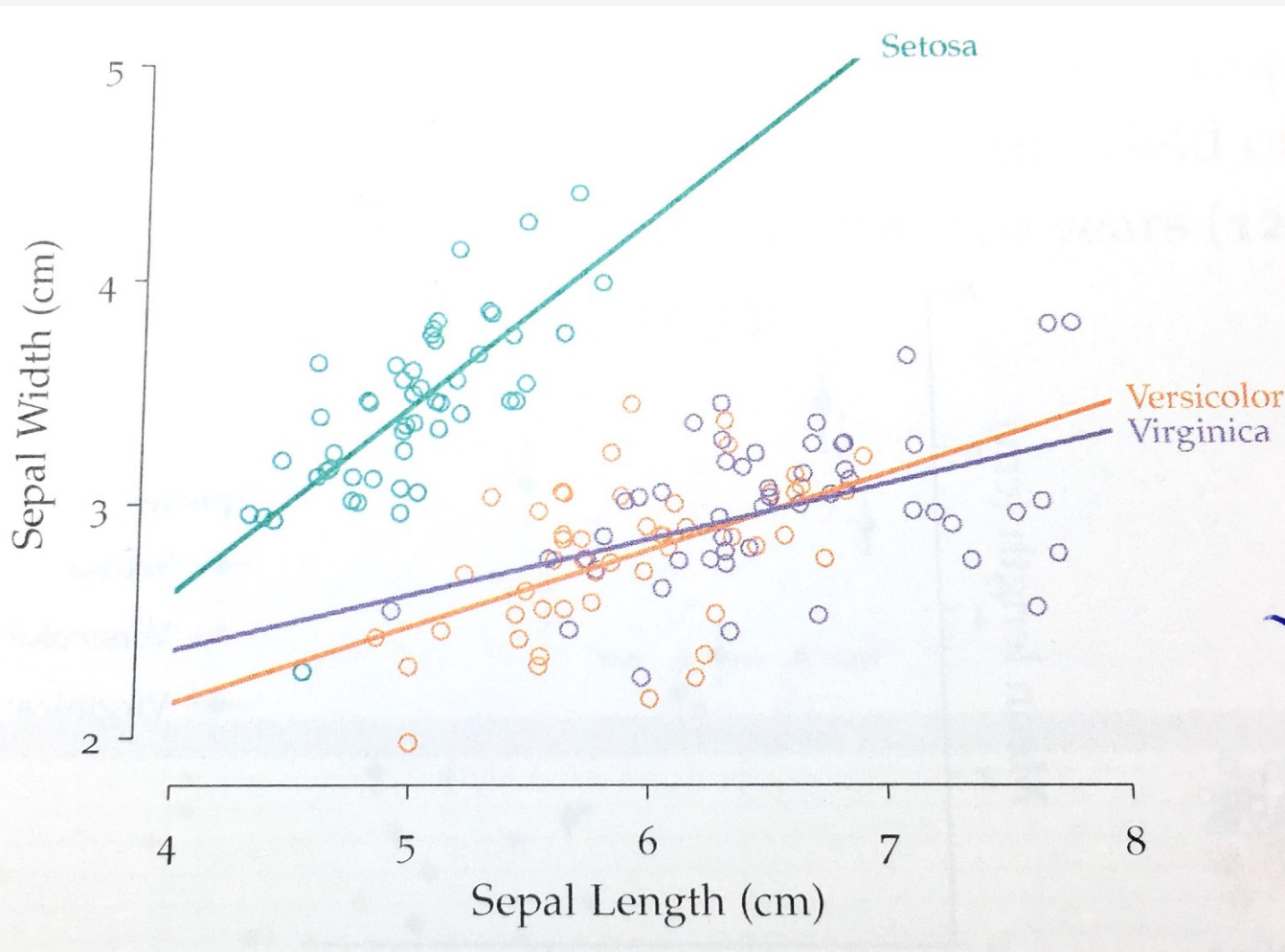
# Data::Ink



# Data::Ink



# Data::Ink



Where is the legend?

# Data::Ink

---

- Remove excess ink
- Show distributions, instead of bars
- Can you remove the legend?
- Remove double encodings
- Is a log scale appropriate?
- What do the 'error bars' represent?

# Outline for October 28

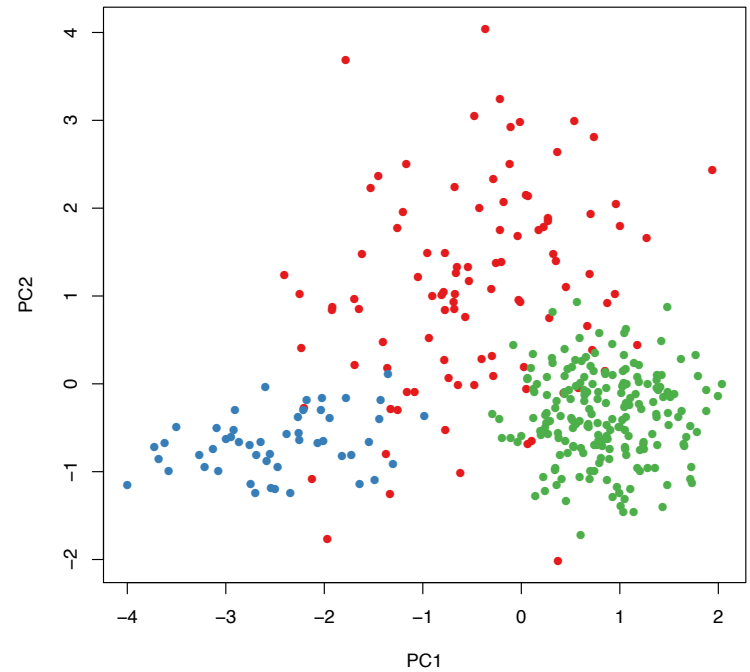
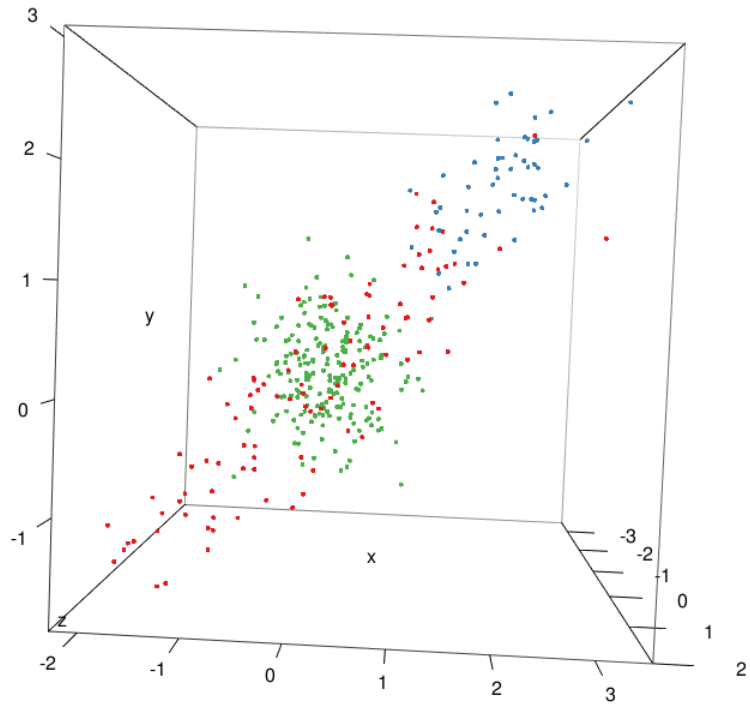
- Discuss final project
- Finish intro to visualization
- **Dimensionality reduction**
- PCA for data visualization



# Principal Components Analysis (PCA)

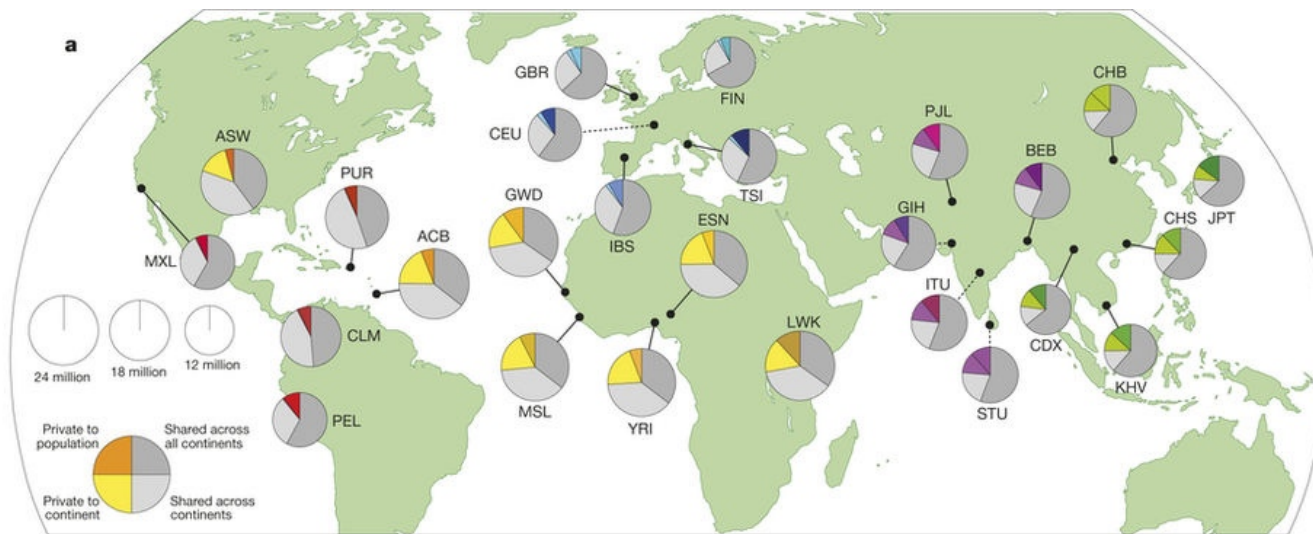
- Transforms  $p$ -dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on
- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction and visualization
- PCA is a linear transformation
- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

# Principal component analysis



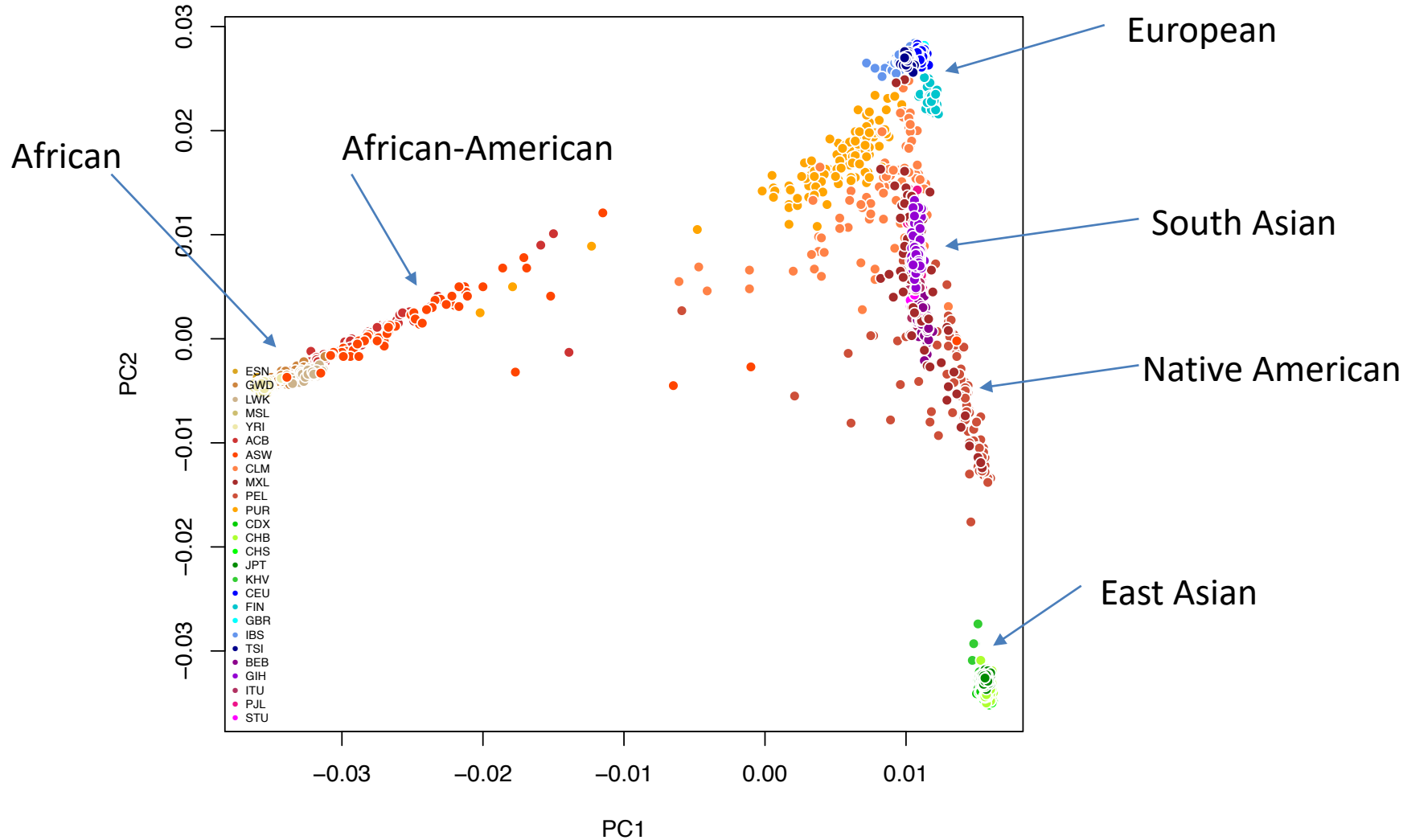
# The 1000 Genomes project

- 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/>



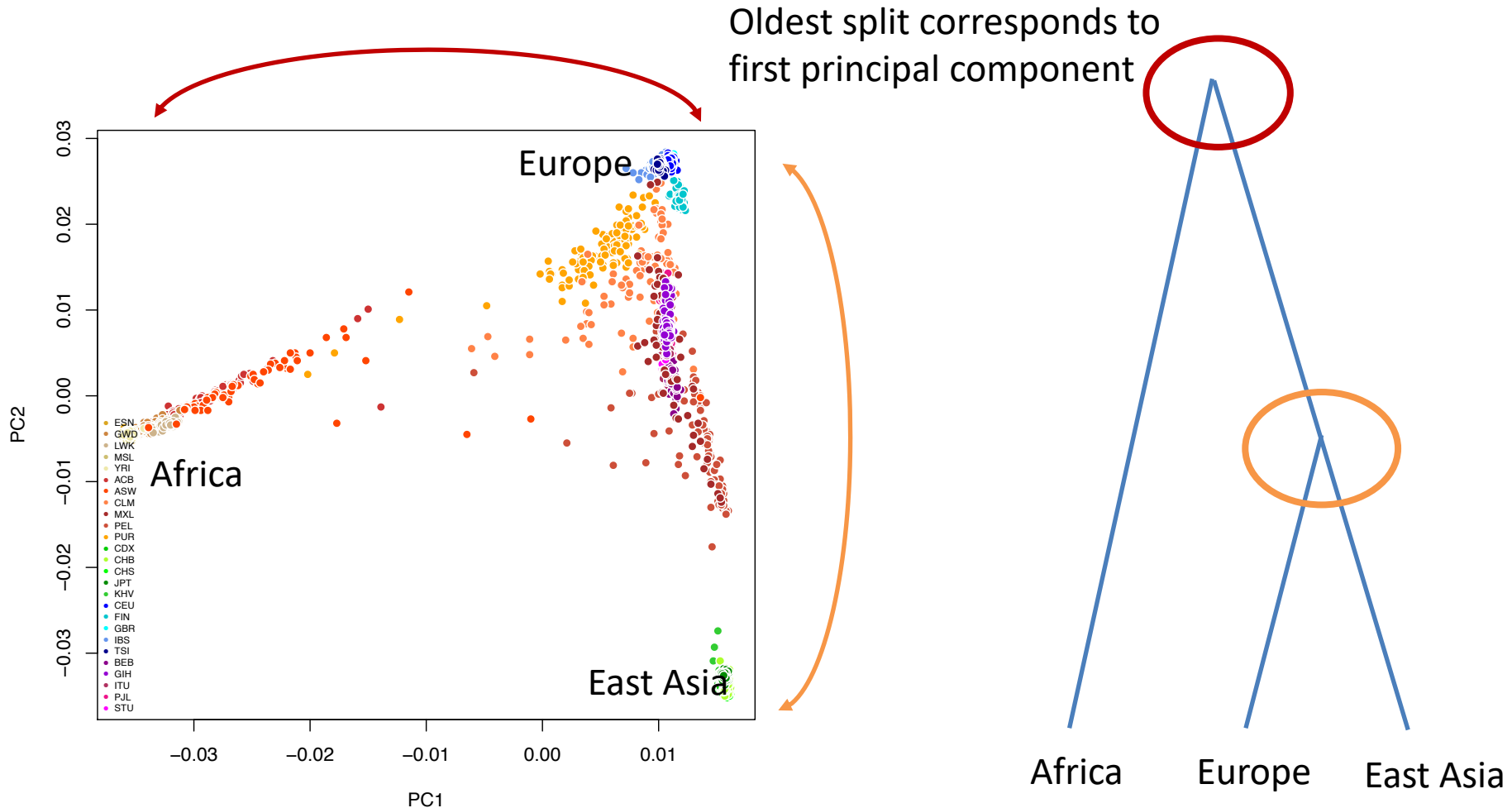


# Global population structure



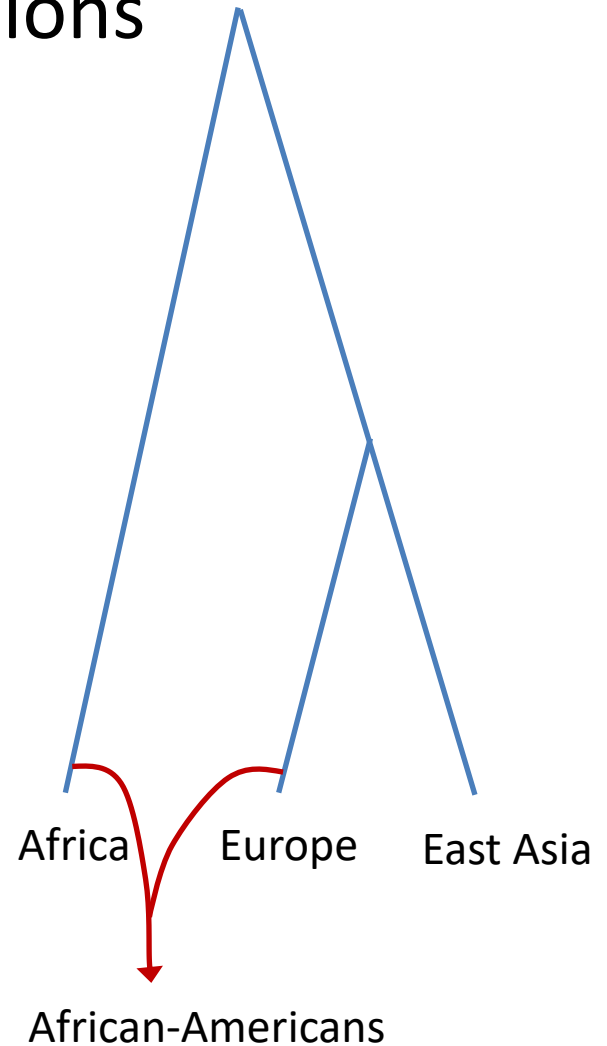
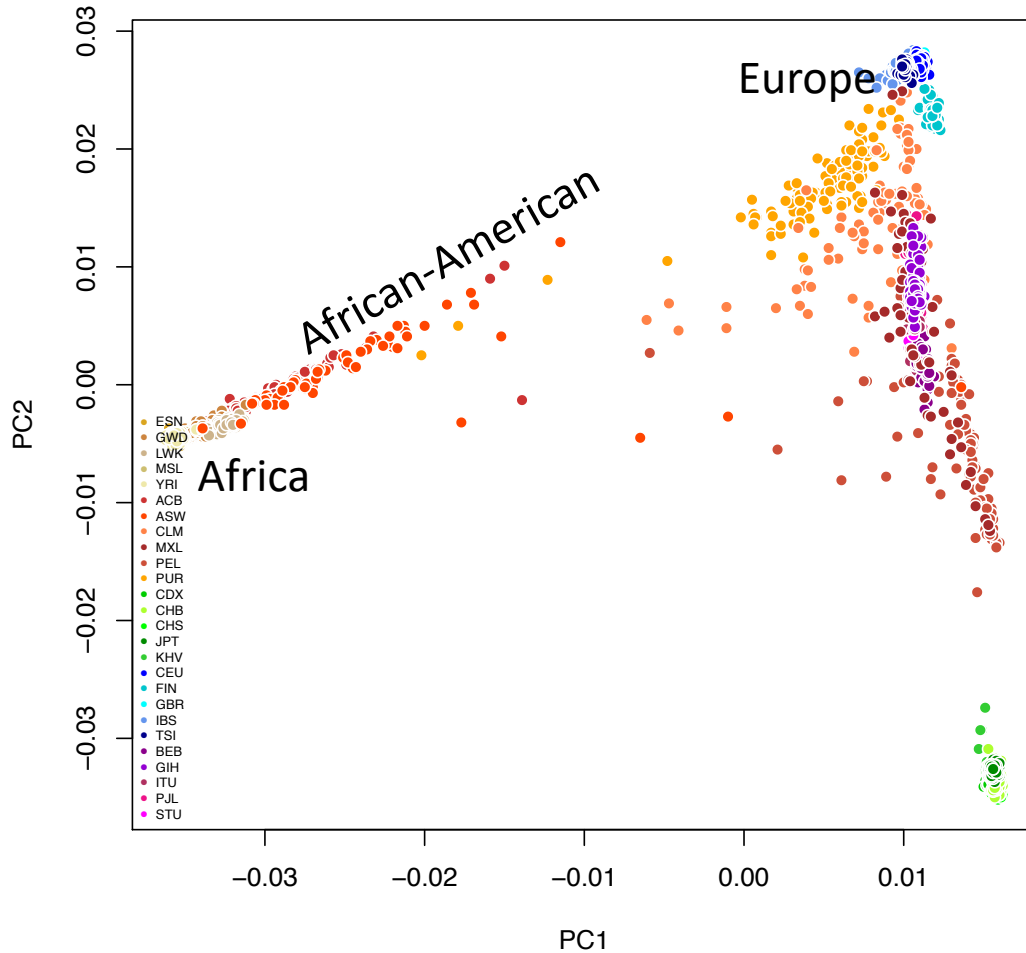
# What causes these patterns?

## 1. Populations **splits** separate populations

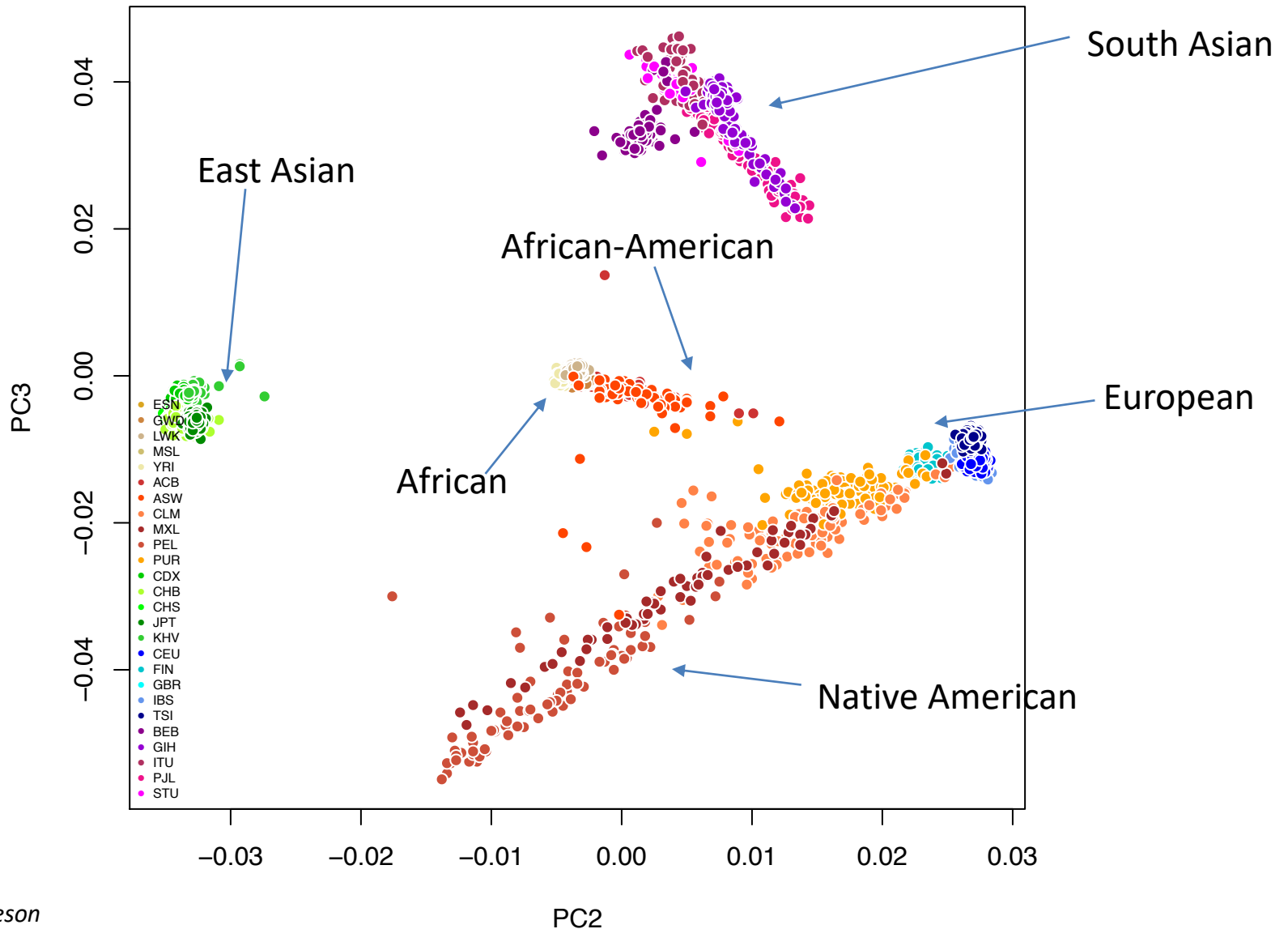


# What causes these patterns?

## 2. Admixture merges populations

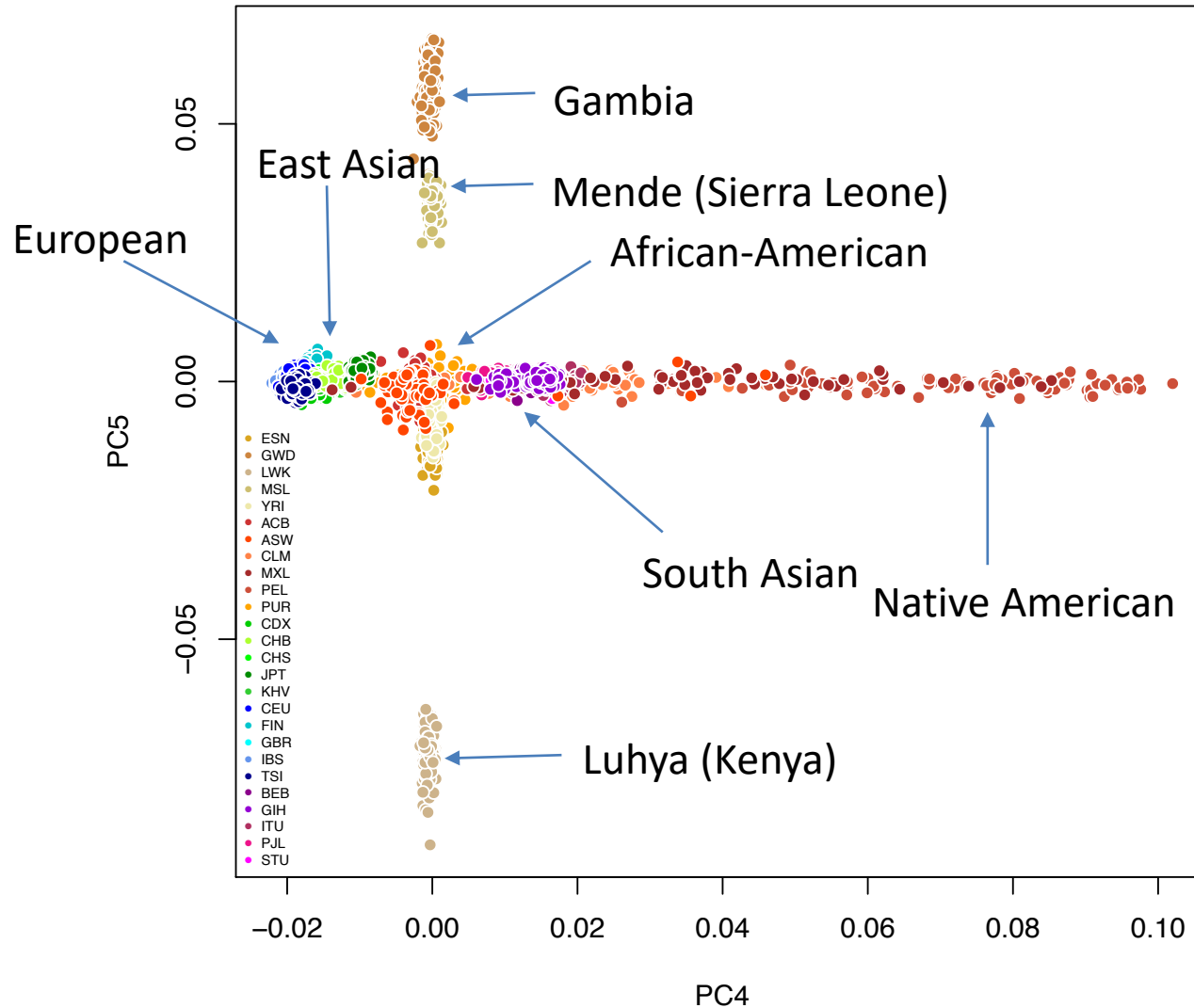


# Global population structure






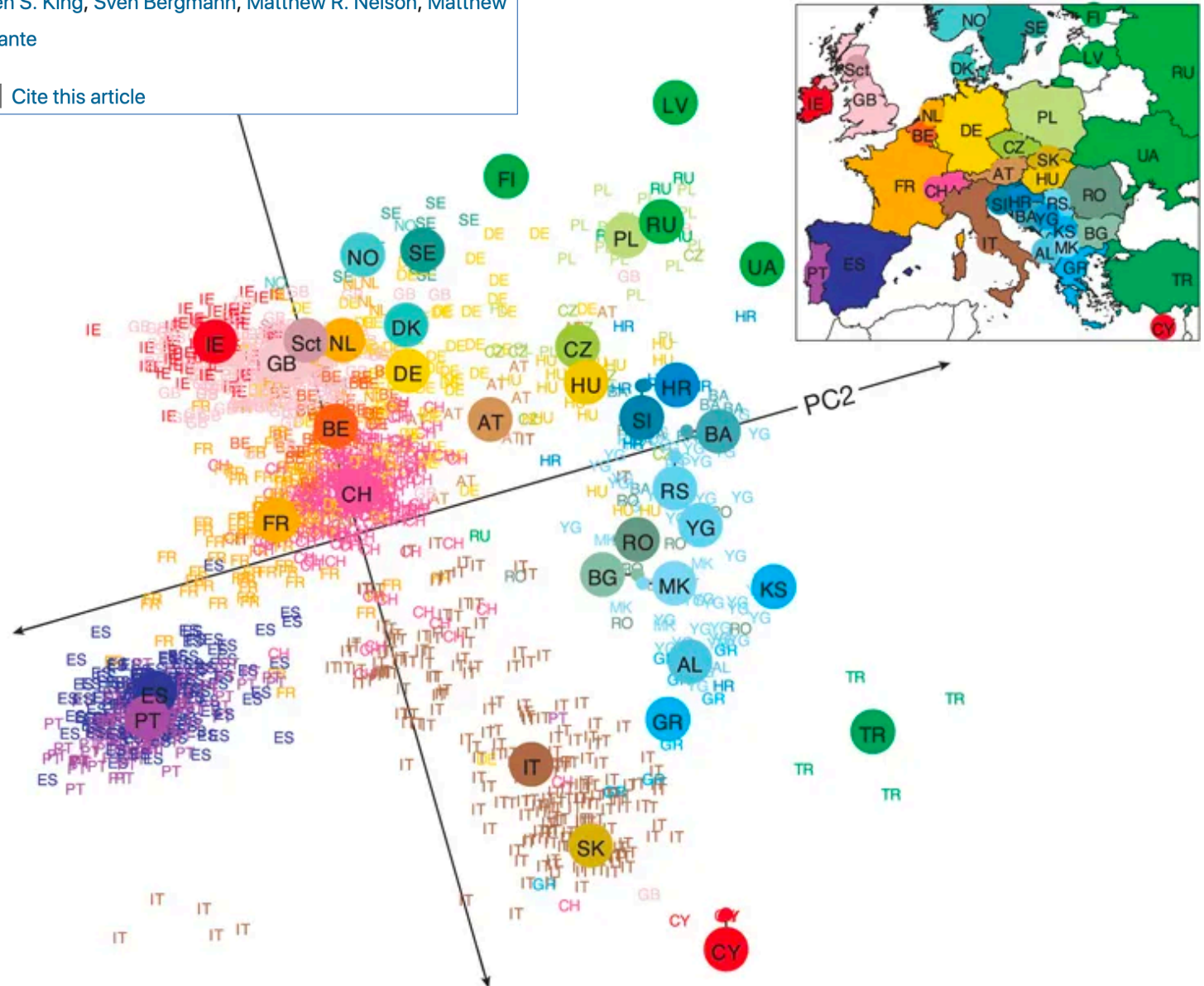
# Global population structure



# 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](#)

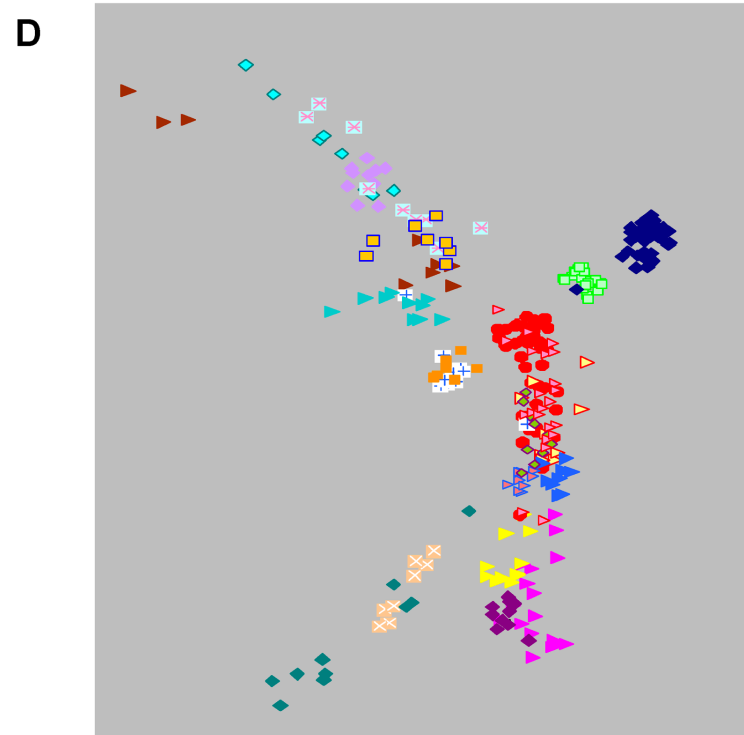
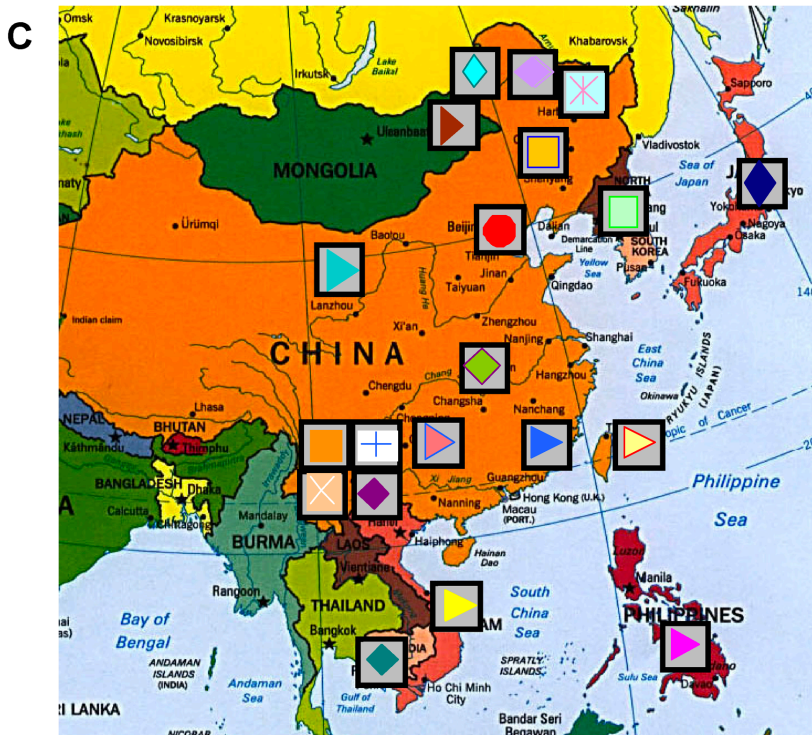


# Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays

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

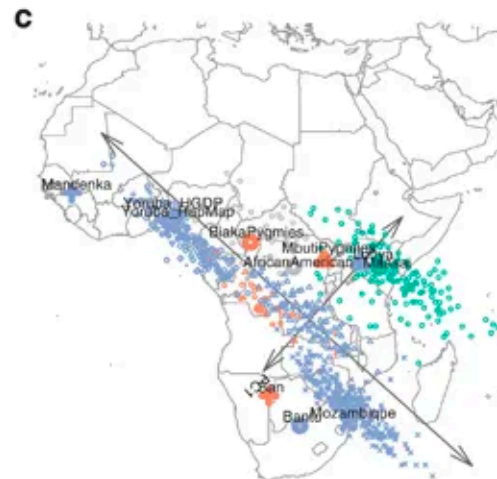
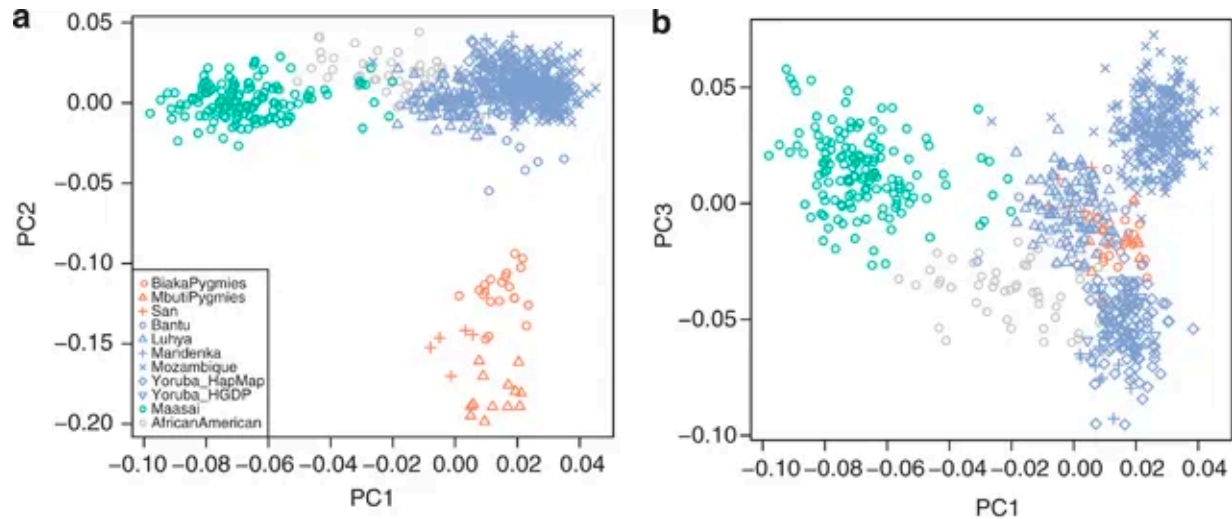
Published: December 5, 2008 • <https://doi.org/10.1371/journal.pone.0003862>

-  FIL
-  VIET
-  LAHU
-  DAI
-  CAMB
-  CHB
-  MGL
-  ORQ
-  DAUR
-  KOR
-  TWN
-  YI
-  HEZ
-  MIAO
-  NAXI
-  SHE
-  TU
-  TUJ
-  XIBO
-  CHA
-  JPT
-  YAK



# 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 October 28

- Discuss final project
- Finish intro to visualization
- Dimensionality reduction
- **PCA for data visualization**

# Principal Component Analysis (PCA)

**Step 1** input data feature  $j$

$$X_{\text{orig}} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & x_{ij} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad n \times p$$

$p \gg n$   
could happen

no label "y"!

goal: create  $n \times 2$  matrix (2D)

**Step 2** Subtract off the mean of each feature.

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

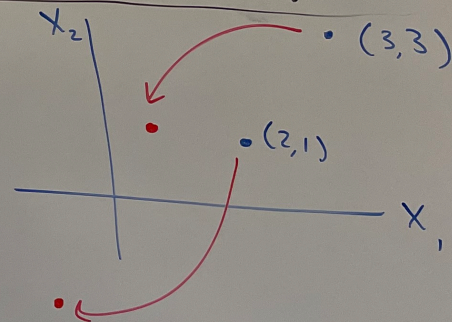
every entry:

$$x_{ij} - \bar{x}_j$$

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

$$\bar{x}_1 = \frac{2+3}{2} = 2.5$$

$$\bar{x}_2 = \frac{1+3}{2} = 2$$



Step 3 Compute covariance matrix A

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

2 features

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

$$A = \begin{bmatrix} \text{cov}(f, f) & \text{cov}(f, g) \\ \text{cov}(g, f) & \text{cov}(g, g) \end{bmatrix} \left. \begin{array}{l} \text{Symmetric} \\ A = A^T \\ p \times p \end{array} \right\}$$

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

↑                    ↑  
f                    g

Step 4

compute eigenvalues + eigenvectors of  $A$

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

$\uparrow$  eigenvector       $\leftarrow$  eigenvalue

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

solve for  $\lambda$ , plug back in  $r$   $\times$  5

Step 5

transform the data

$$W_r = \begin{bmatrix} | & | & \dots & | \\ \vec{v}_1 & \vec{v}_2 & & \vec{v}_r \\ | & | & & | \end{bmatrix}$$

high  $\lambda$   $\rightarrow$  low  $\lambda$

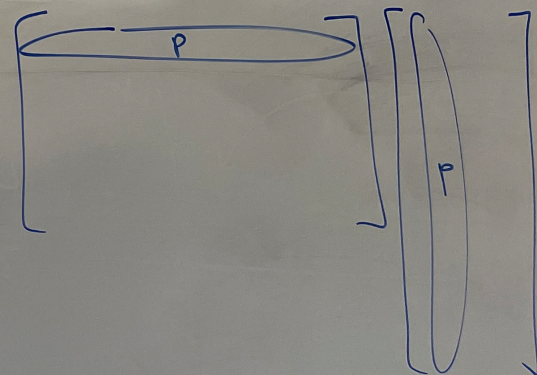
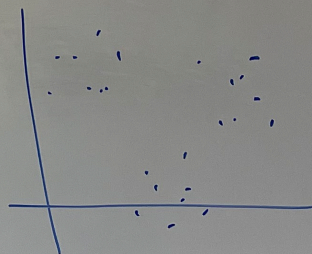
$p \times r$

$r=2$

$$\underbrace{T}_{n \times r} = X_{n \times p} W_{p \times r} =$$

Step 6

plot!





# Handout 16

Step 2

$$\bar{f}_1 = \frac{1}{2}, \bar{f}_2 = \frac{1}{2}$$

orig                  orig

Step 3

$$\bar{f}_1 = 0, \bar{f}_2 = 0$$

$$X = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix}$$

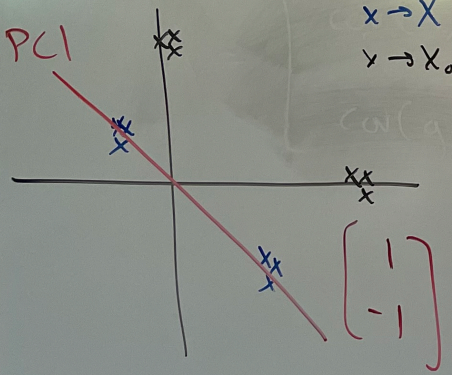
$$\left(\frac{1}{2} - 0\right)\left(\frac{1}{2} - 0\right)$$

$$\text{cov}(f_1, f_2) = \frac{1}{5} \left(-\frac{1}{4}\right) \cdot 6 = -\frac{3}{10}$$

$$\text{cov}(f_1, f_1) = \frac{1}{5} \left(\frac{1}{4}\right) \cdot 6 = \frac{3}{10}$$

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

PCI



$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$T_2 = XW_2 = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

$\lambda_1 = \frac{3}{5}$      $\lambda_2 = 0$   
 high to low

