



CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018



Outline: Apr 13

- Finish PCA (see Handout 25)
- PCA on 1000 genomes data (human)
- Genealogical interpretation of PCA

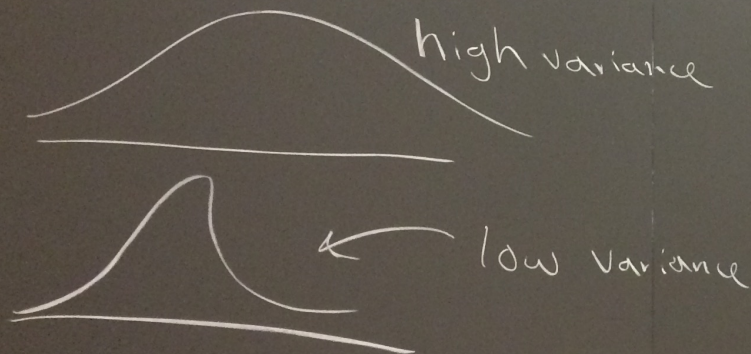
PCA

step 2: subtract off column-wise mean

step 3: compute covariance matrix A

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

↑
mean



Covariance

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

2 columns/
features

covariance
matrix A

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

← (f_1, \square)
← (f_2, \square)
← (f_3, \square)

$p \times p$

Symmetric

($p=2$
here)

step 4
 f_i

[cov

Step 4 find eigenvectors ^{PCs} & eigenvalues of A

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

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

(f_1, \square)

(f_2, \square)

(f_3, \square)

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

$$\begin{bmatrix} \text{cov}(f_1, f_1) & \text{cov}(f_1, f_2) & \text{cov}(f_1, f_3) \\ \vdots & \vdots & \vdots \end{bmatrix}_{3 \times 3}$$

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

$f \quad g$

$$\text{var}(f) = \frac{1}{5} \left(\frac{1}{2} \right)^2 \cdot 6$$

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

$$\text{var}(g) = \frac{3}{10}$$

$$\text{cov}(f, g) = \frac{1}{5} \left(-\frac{1}{2} \right) \left(\frac{1}{2} \right) 6$$

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

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

eigenvalues

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

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

$$\boxed{\lambda_2 = 0}$$

$$-\frac{3}{10} + \lambda = \frac{3}{10} \rightarrow \boxed{\lambda_1 = \frac{3}{5}}$$

Sort eigenvalues high \rightarrow low

eigenvectors

$$\begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix} \vec{v}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow$$

$$\boxed{\vec{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}}$$

$$\begin{bmatrix} -3/10 & -3/10 \\ -3/10 & -3/10 \end{bmatrix} \vec{v}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow$$

$$\boxed{\vec{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}}$$

principal
component

1
PC1

2 genetic data $p = \text{millions}$
Lab 9: $p \approx 600,000$

PCs = \boxed{r} $r \approx 5-6$

dimensionality reduction: $\boxed{p \rightarrow r}$

start

$X: n \times p \xrightarrow{\text{end}} X_{\text{transform}} = T_r: n \times r$

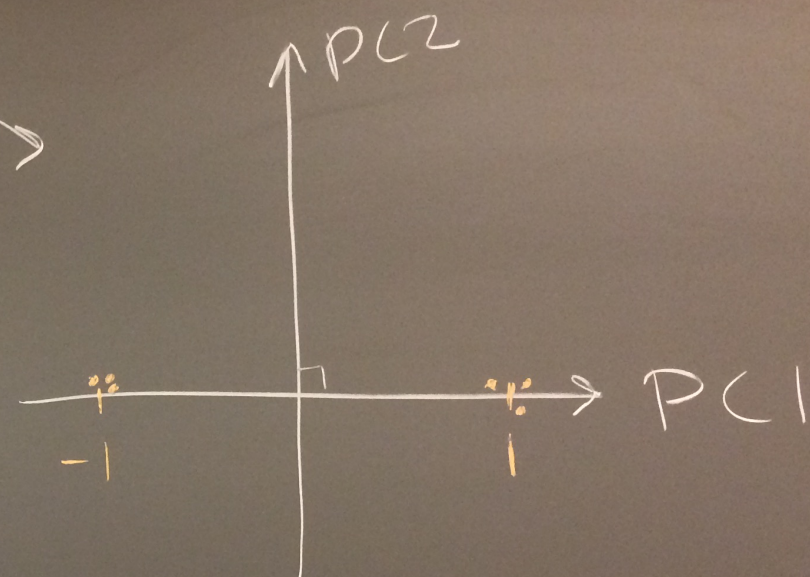
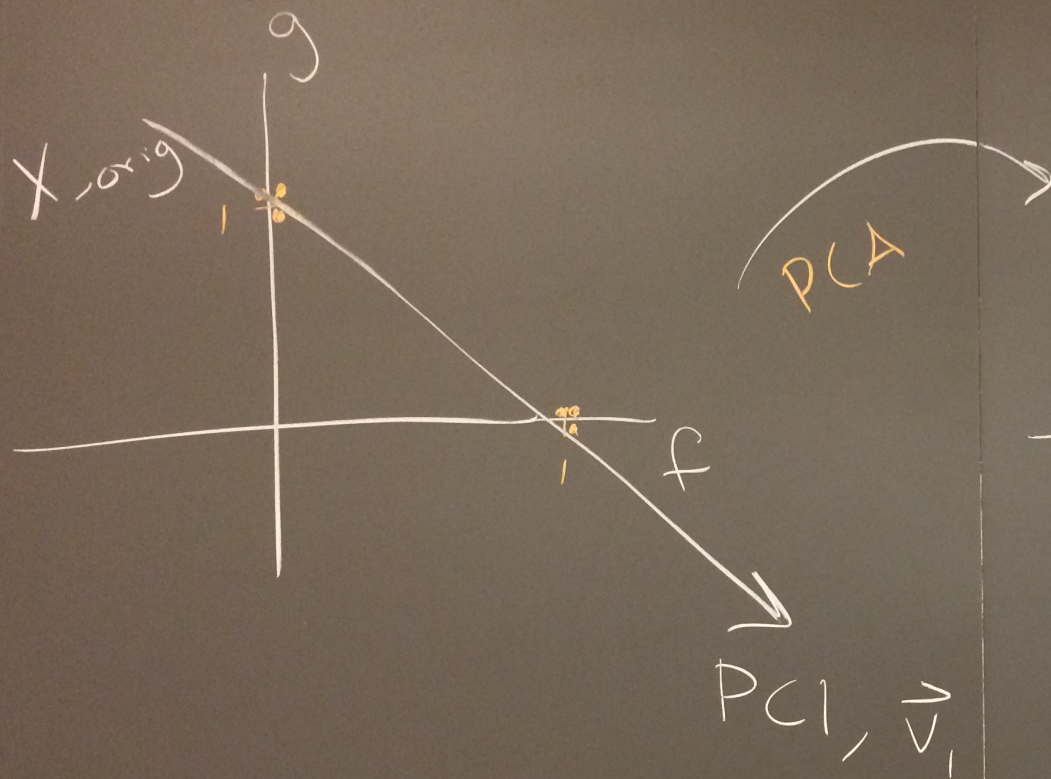
$$X \begin{matrix} \text{eigenvectors} \\ W_r \end{matrix} = T_r$$

$\begin{matrix} (n \times p) & (p \times r) & (n \times r) \end{matrix}$

$$\begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} \vec{v}_1 & \vec{v}_2 \end{bmatrix} = \begin{bmatrix} \boxed{-1} & 0 \\ -1 & 6 \\ -1 & 6 \\ -1 & 6 \\ -1 & 6 \\ -1 & 6 \end{bmatrix}$$

$X \qquad W_r \qquad T_r$

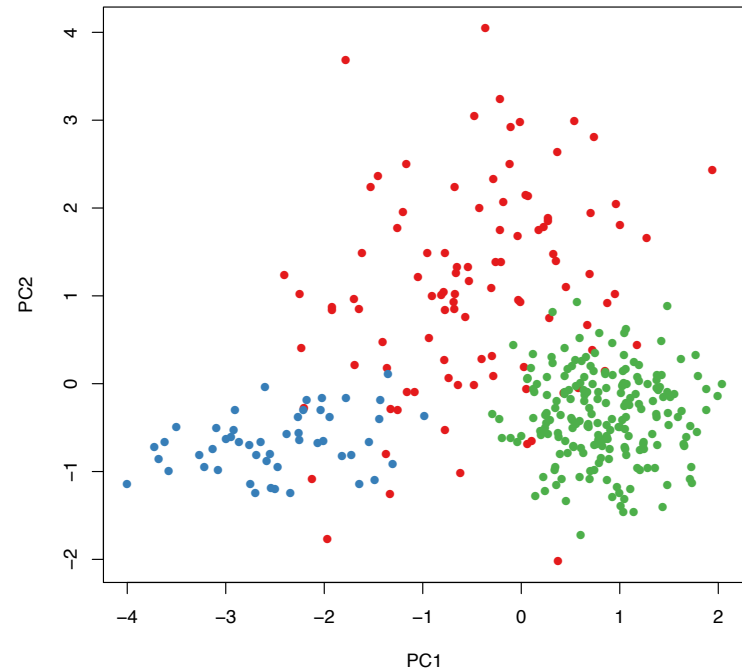
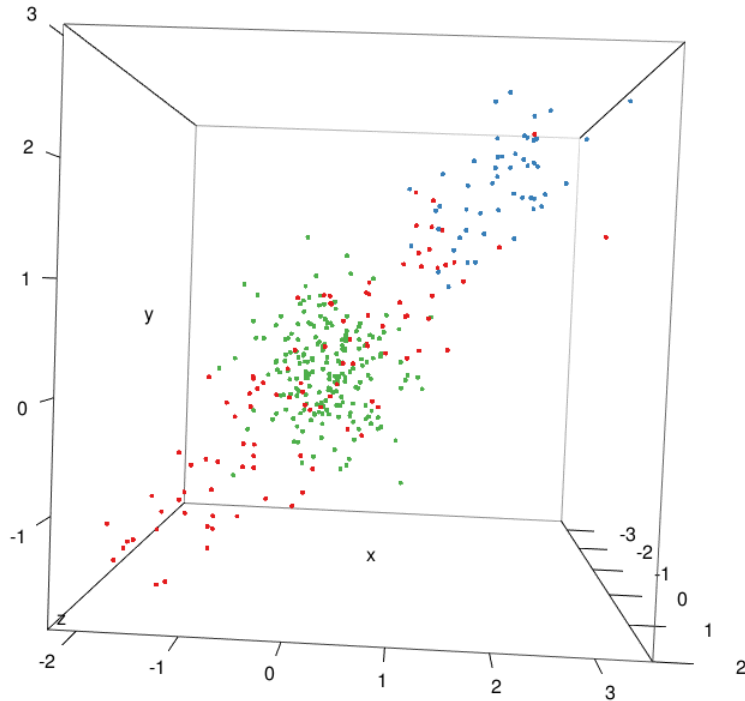
goal
linear comb.
of feature



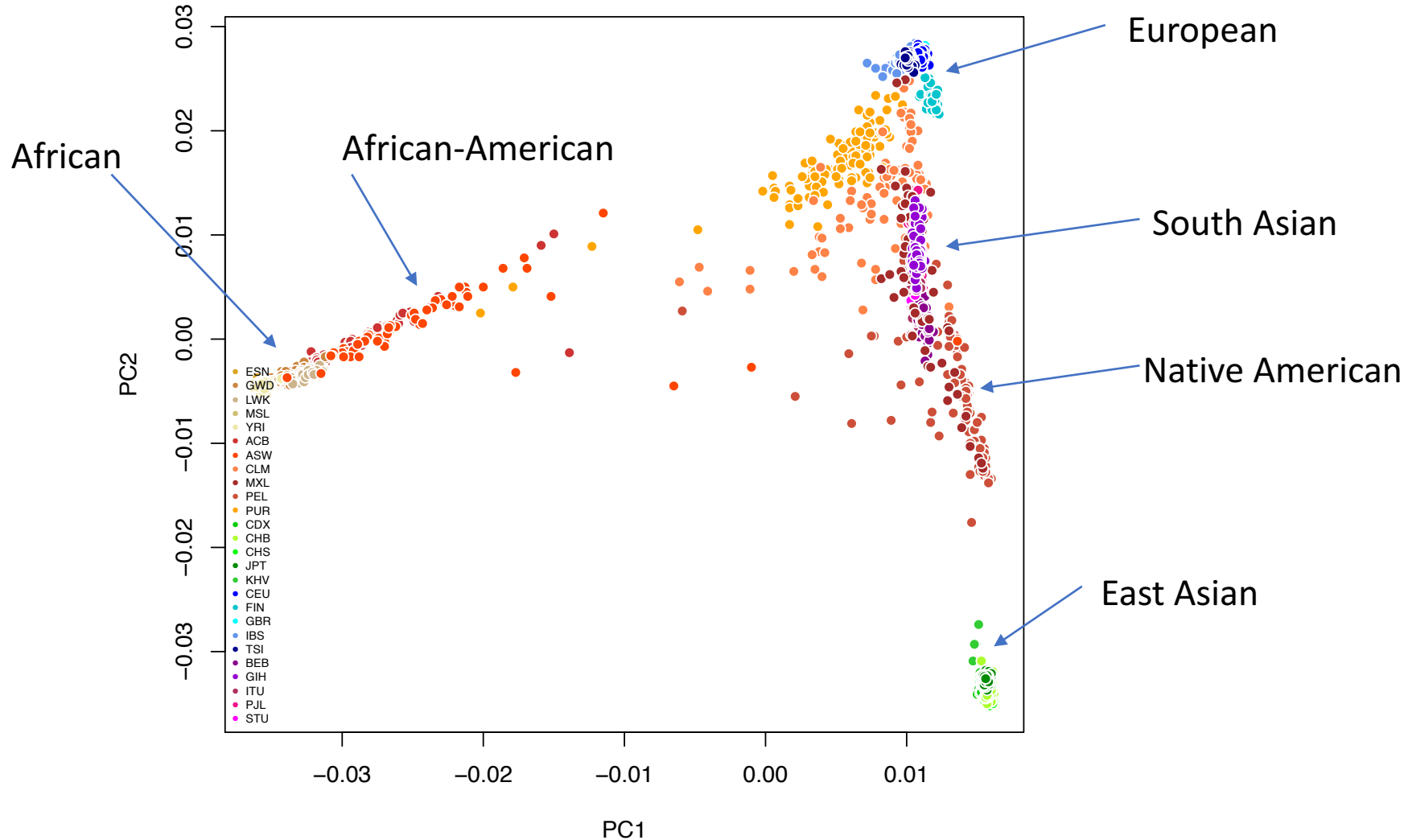
Principal component analysis

- Transforms **n-dimensional** data so that the the new first dimension explains as much of the variation as possible, the 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**.
- PCA is a **linear** transformation.

Principal component analysis

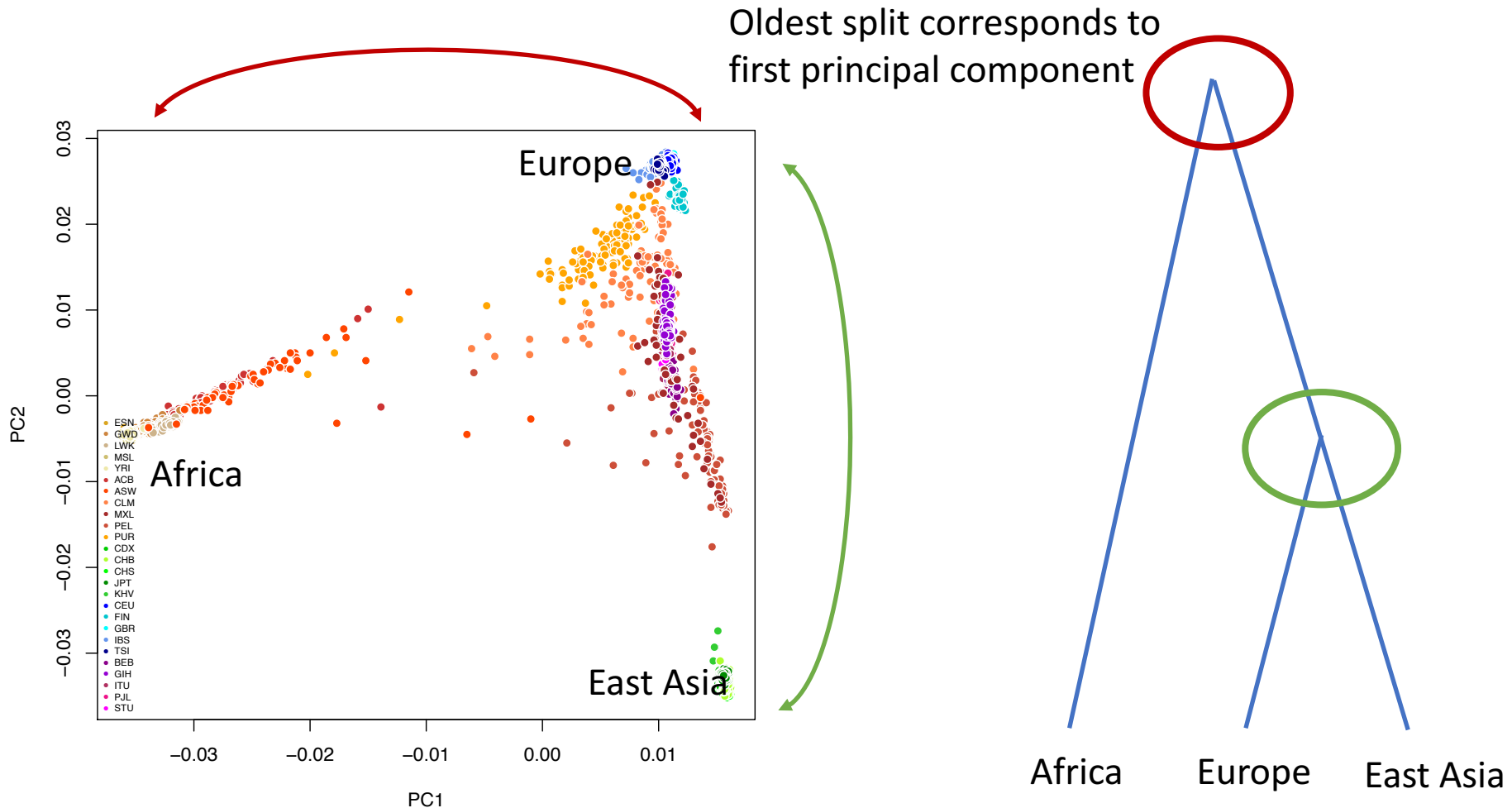


Global population structure



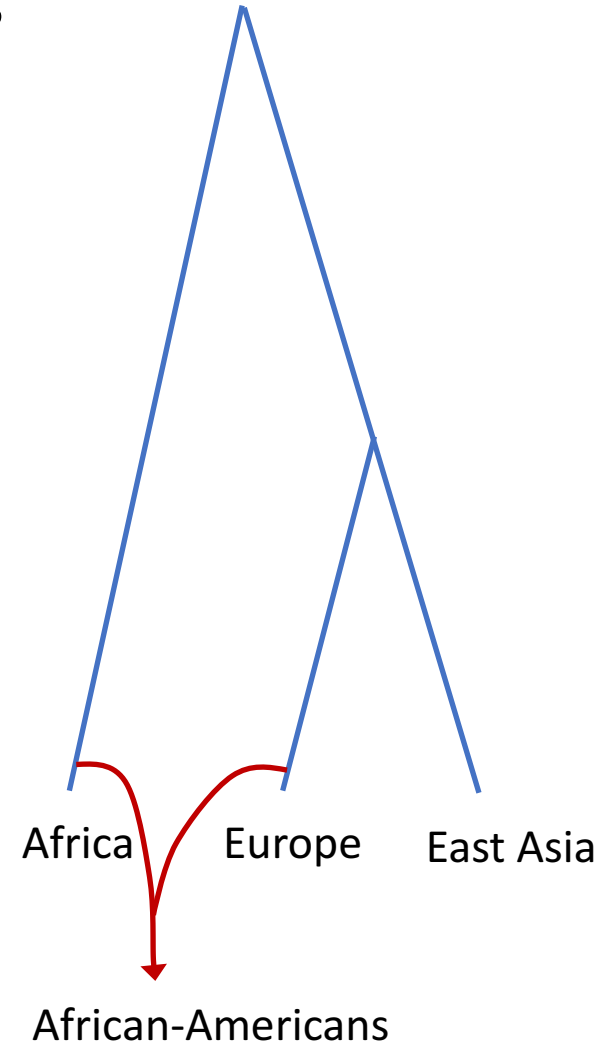
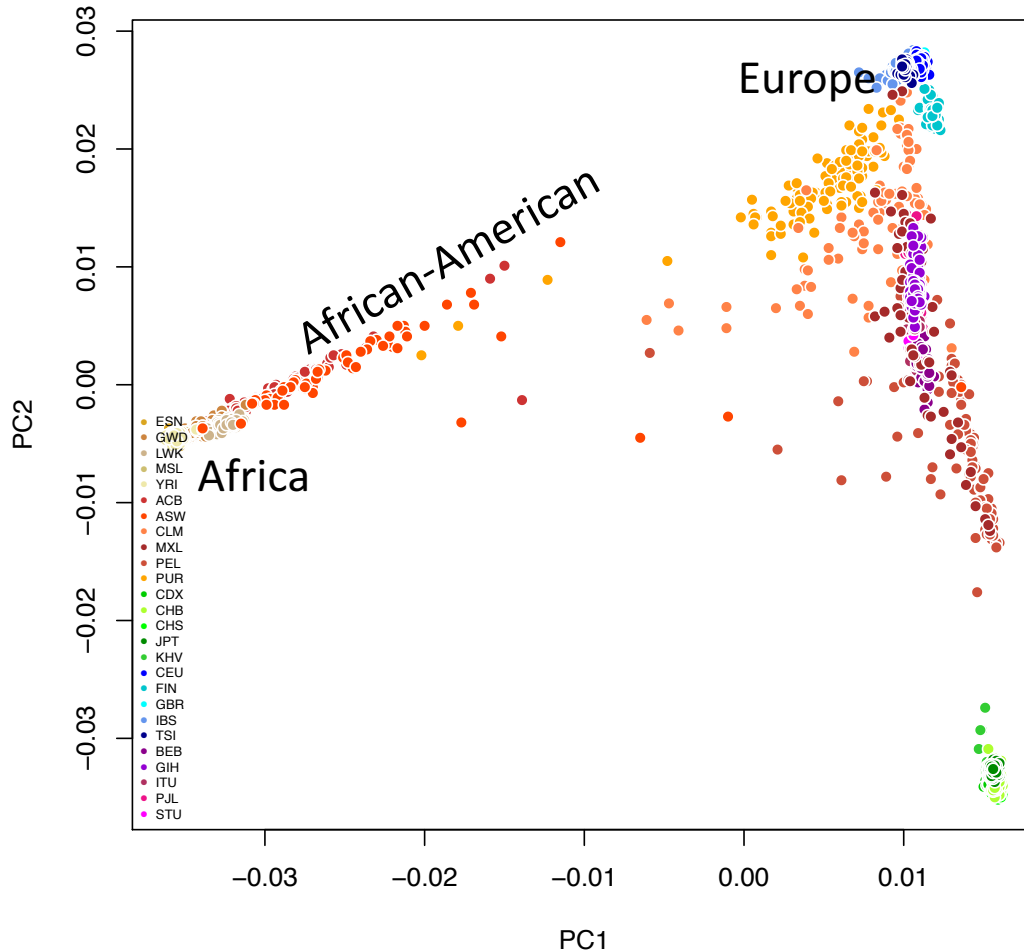
What causes these patterns?

1. Populations **splits** separate populations

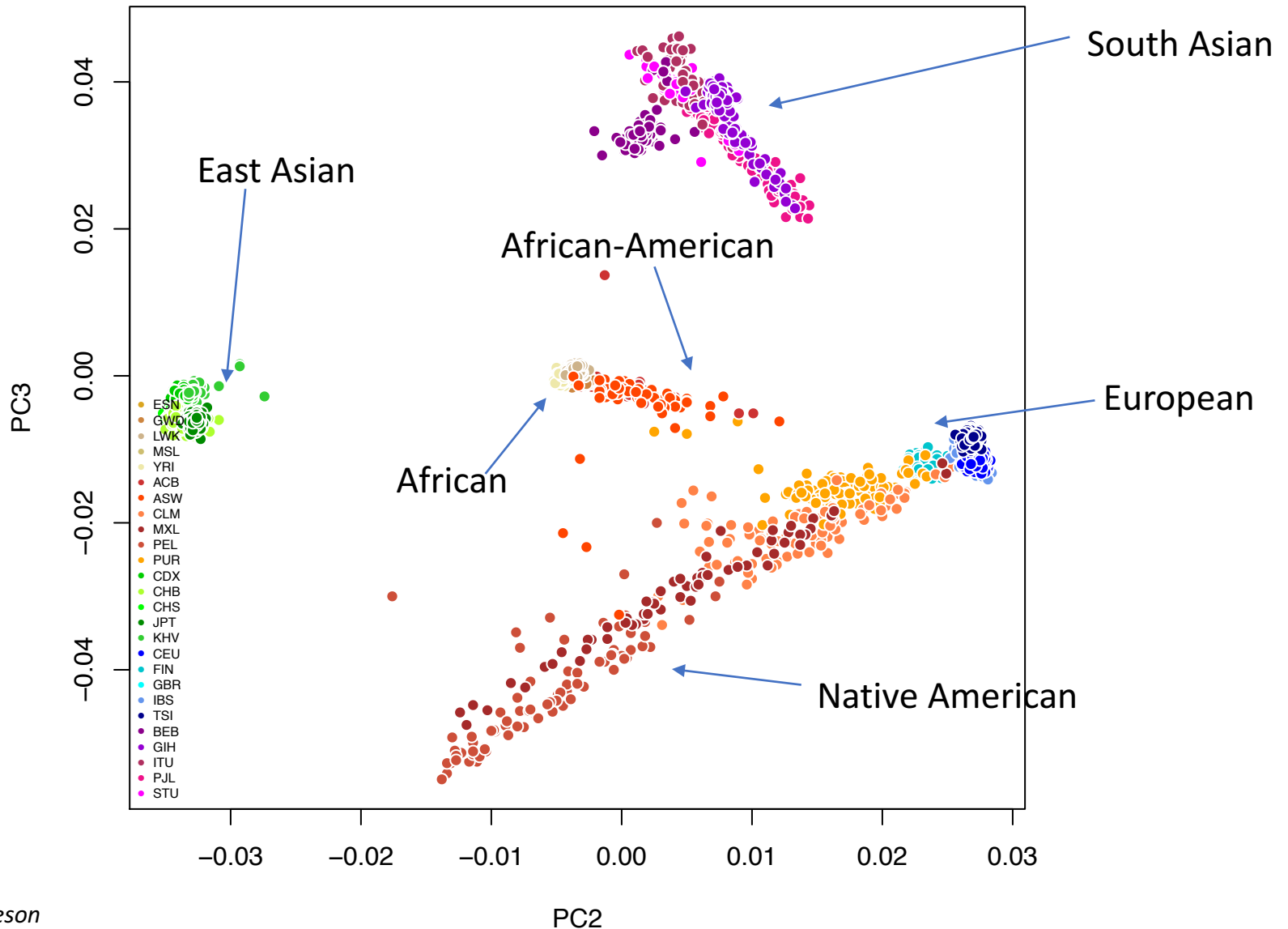


What causes these patterns?

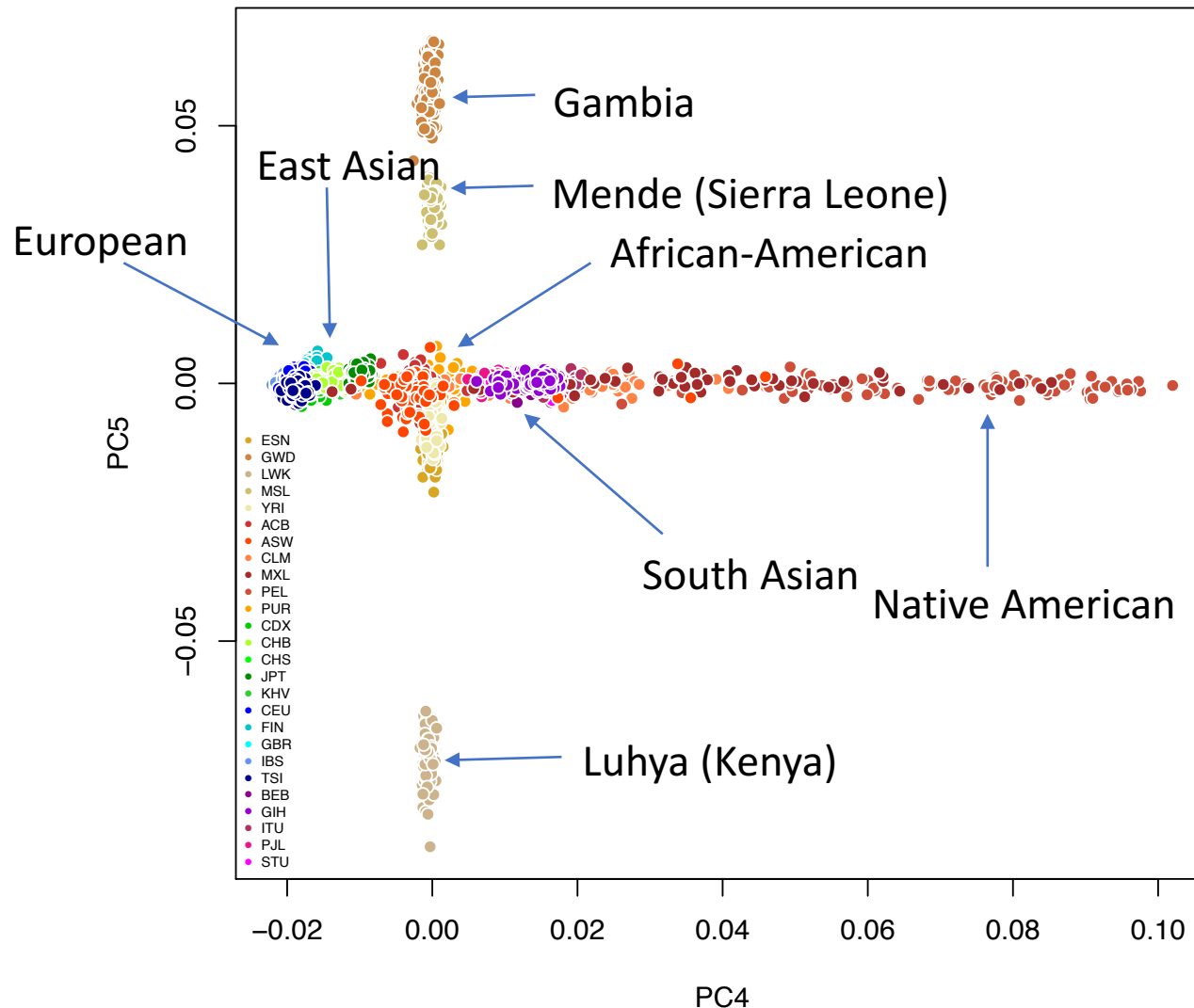
2. **Admixture** merges populations



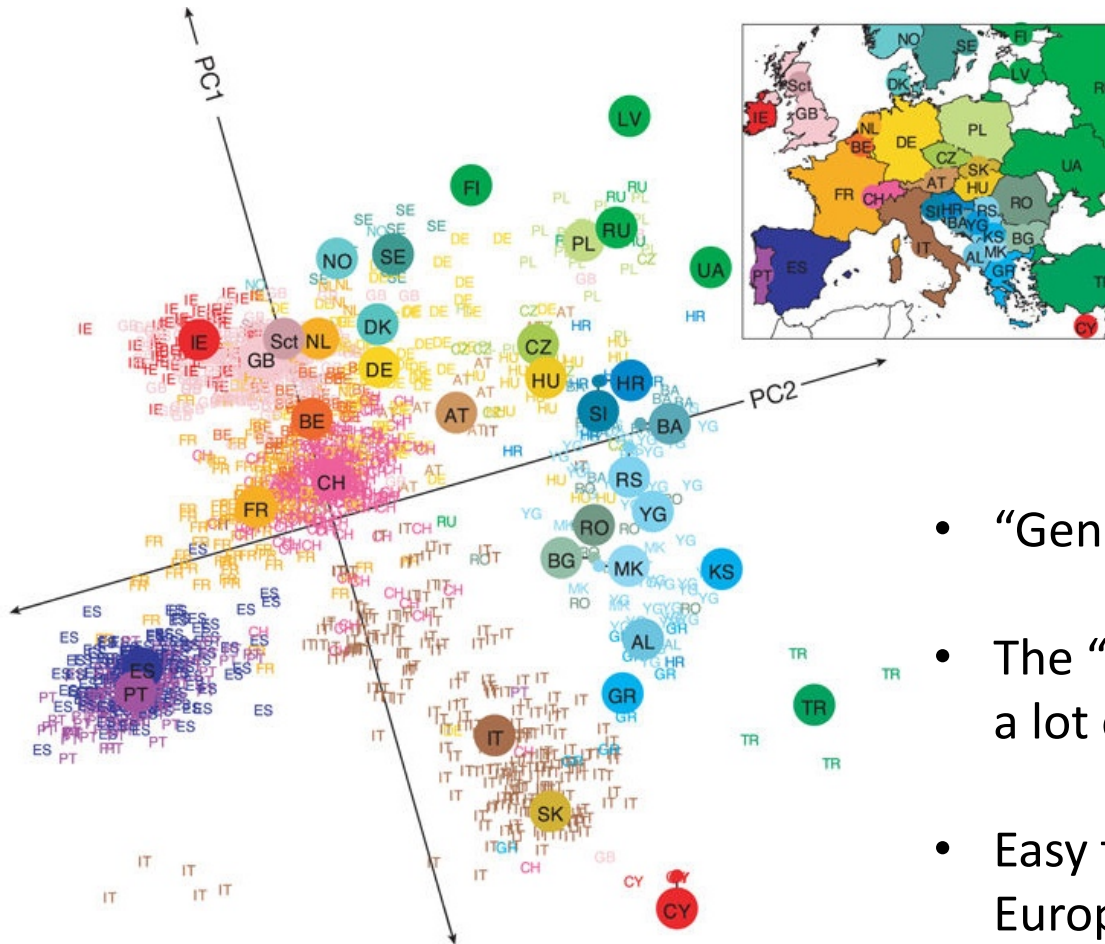
Global population structure



Global population structure

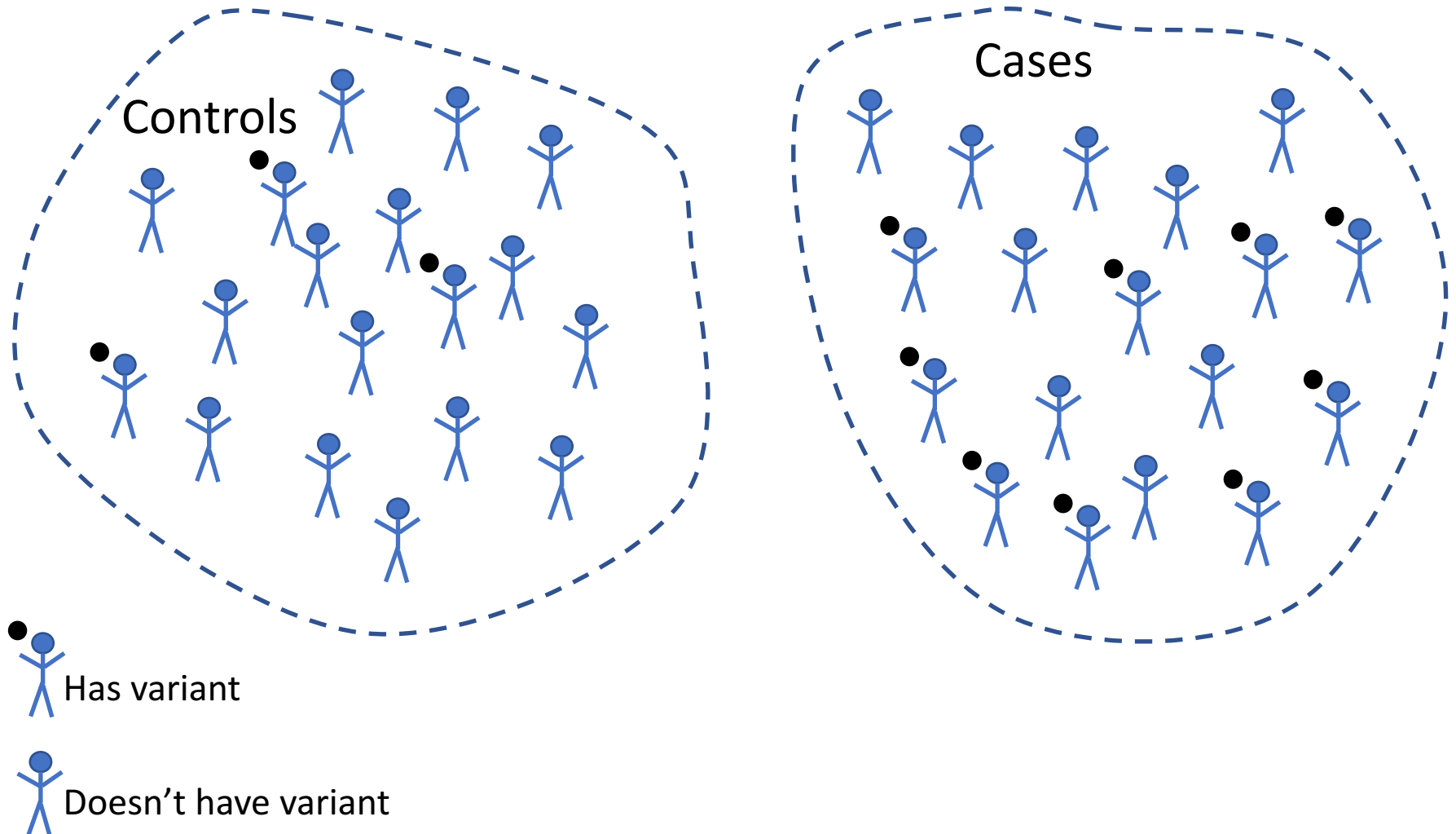


Fine-scale population structure

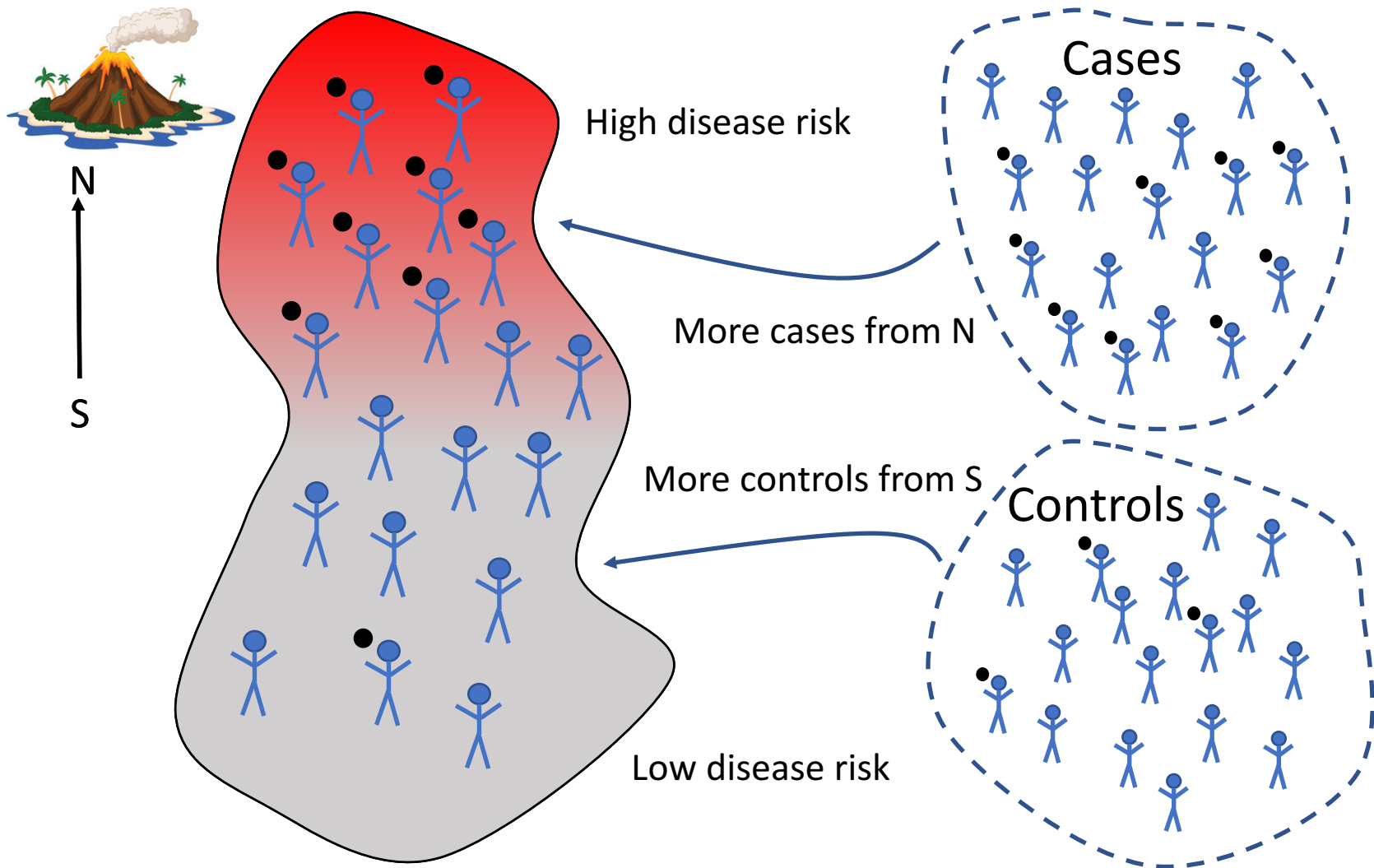


- “Genes mirror geography in Europe”
- The “European” population contains a lot of population structure.
- Easy to distinguish origin of most European-origin individuals from genetic data.

Why does population structure confound disease gene discovery?



Why does population structure confound disease gene discovery?



Population structure and association studies.

- Population structure leads to **false positives** in association studies.
- No real way to correct for population structure in **candidate gene studies**.
- This is one the advantages of **genome-wide association studies** (GWAS).
- With genome-wide data, we can estimate genome-wide population structure and use it to correct statistical tests.

Further reading on human data

Papers:

- The 1000 Genomes Project Consortium **A global reference for human genetic variation** *Nature* 2015
- Mallick et al **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations** *Nature* 2017