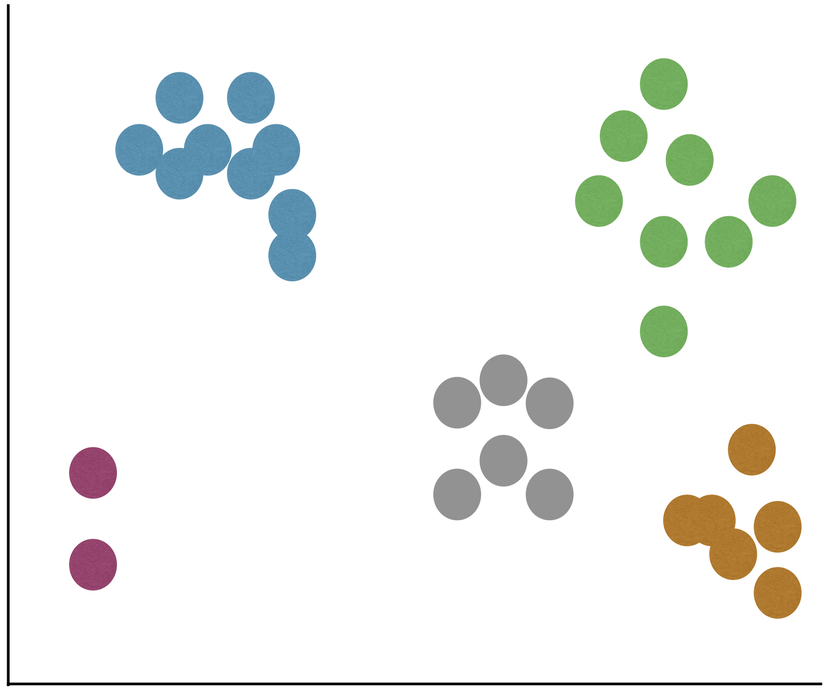
Week-4

Discovery Analyses - Principal Component Analysis (PCA)



Aim: *Understanding the PCA, learning eigenvectors and eigenvalues, running a simple PCA*

Hands-on: *Running smartpca tool of AdmixTools for PCA and plotting the results in R.*

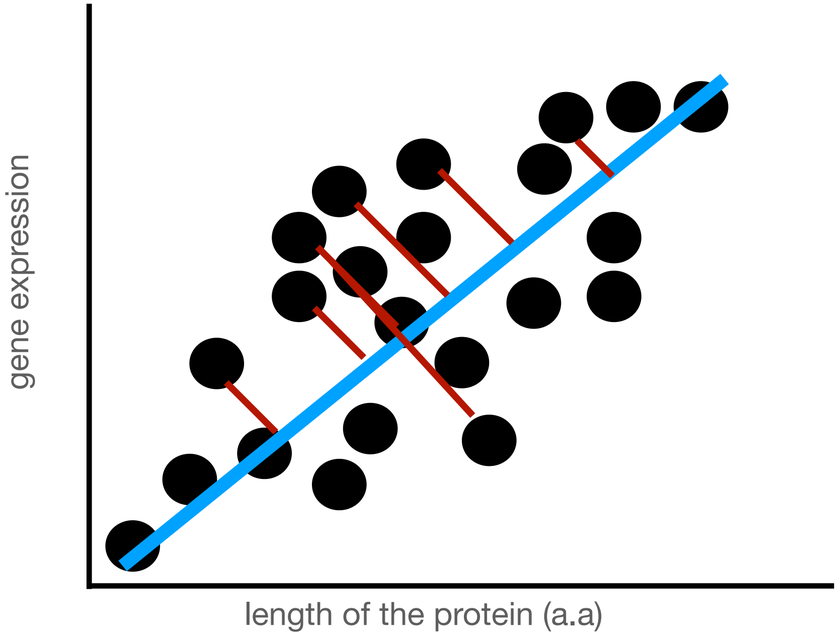
Reading suggestions: *Patterson N, Price A, Reich D (2006) Population Structure and Eigen Analysis doi: https://doi.org/10.1371/journal.pgen.0020190*

Menozzi, Piazza and Cavalli-Sforza (1978) Synthetic Maps of Human Gene Frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic, doi: https://doi.org/10.1126/science.356262

https://github.com/chrchang/eigensoft/blob/master/POPGEN/README

Principal component analysis (PCA) - *basically*:

- Invented by Karl Pearson in 1901
- A method not just specific to genetic data but to summarize any kind of data - dimension reduction
- A discovery method - exploring the patterns of data
- Eigenvectors of the covariance matrix
- Find axis of the largest variation (first PC)
- Each point has two values -> converted to only one value (how far the point is to the blue line
- Second PC is orthogonal to first PC

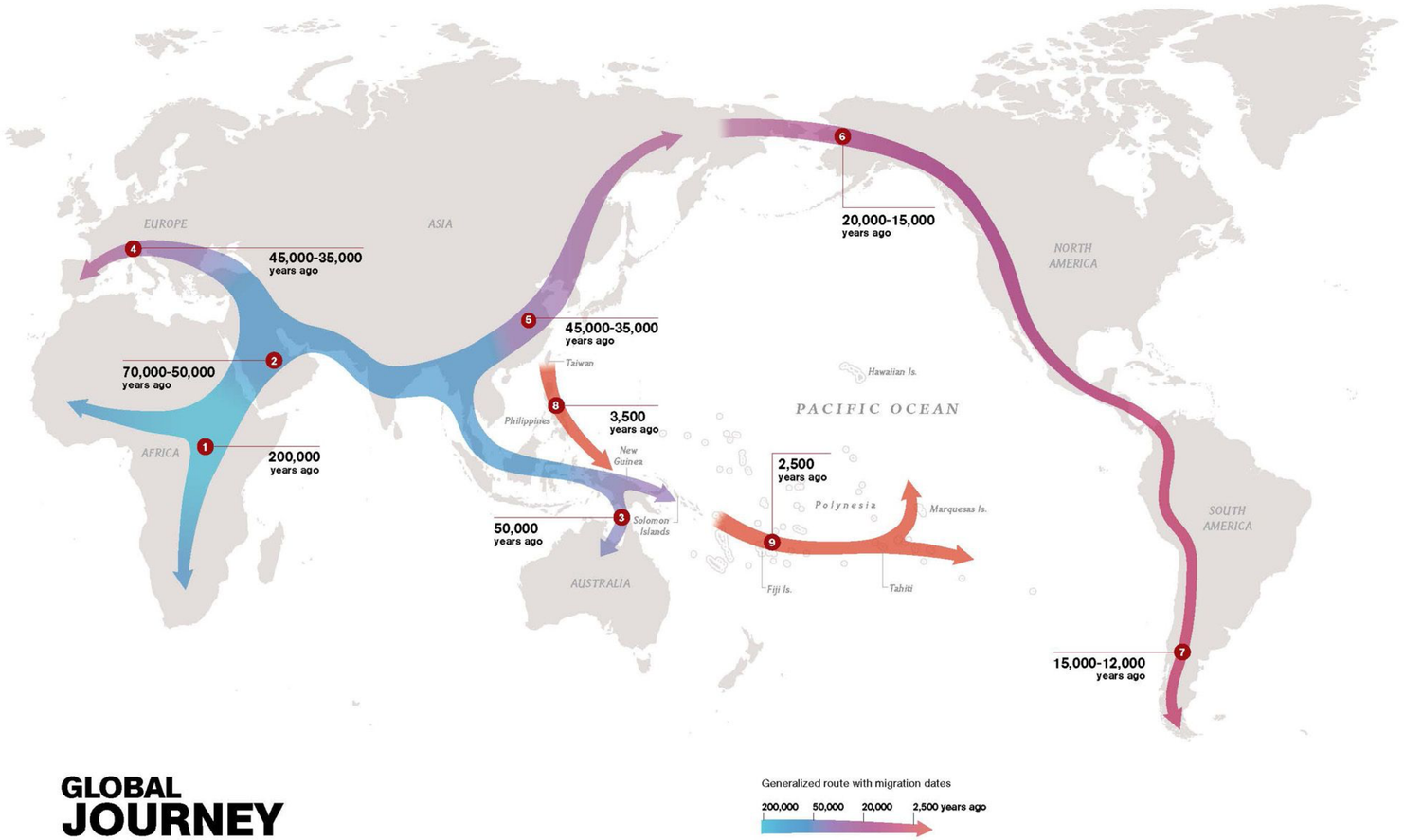


$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

How do we use it in population genetics? -> Genetic similarity matrices

Understanding population structure: Principal Component Analysis

Cavalli-Sforza - studying genetic data using Principal Component Analysis (*1978*)



https://www.nationalgeographic.org/media/global-human-journey/

Genetic diversity:

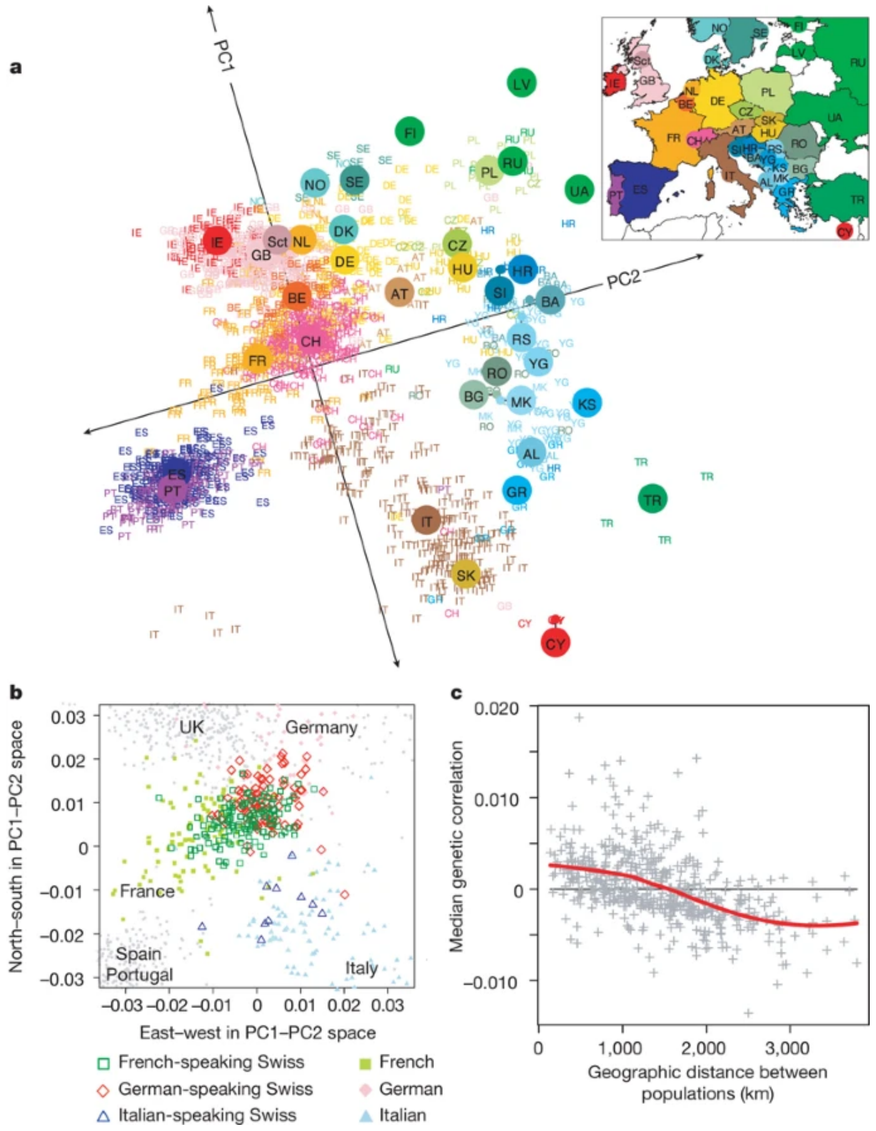
- Mutation & Selection
- Migration
- Genetic drift

↓

Population structure

↓

An Example PCA based on European human populations (Novembre 2008)



Novembre 2008, Nature

In population genetics:
Genotype data for a large number of SNPs
Thousands of individuals

Reduce the dimension into 2 (First 2 PCs)
Infer about:

Population structure
Genetic simmilarities between populations
If populations are homogeneous or heterogeneous
Discover the outliers in your dataset

PCA on human population genetic data

Tool: smartpca

Files: .geno, .snp and .ind

500,000 SNPs
122 individuals