CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD
COLLEGE

Admin

- Reminder: **pair work is required** for the final project – find a partner ASAP or email me!

- **Lab 7 and project proposal** (both short)
  - Due Wednesday

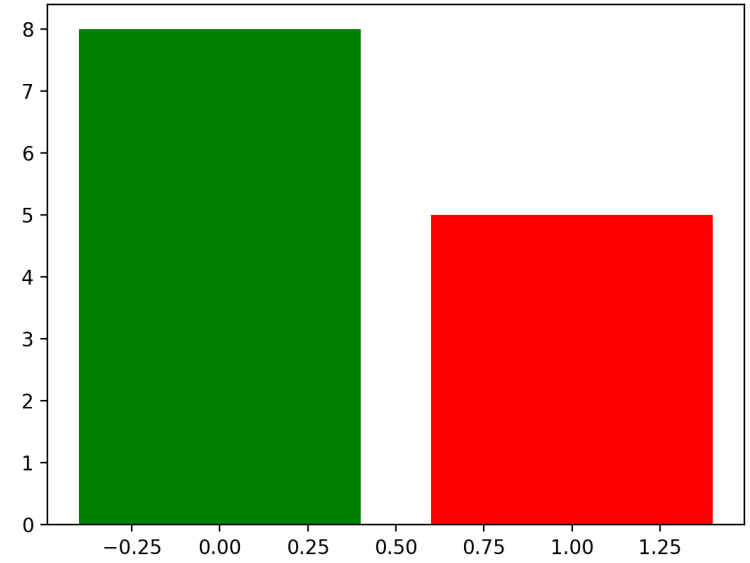
Outline

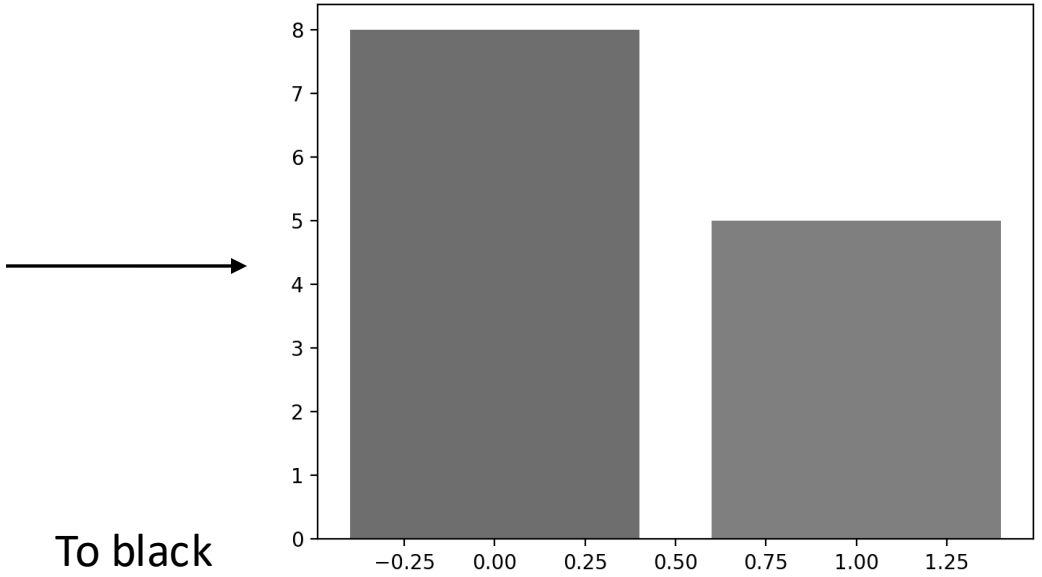
- Finish data visualization intro
- Dimensionality reduction
- PCA for data visualization

Outline

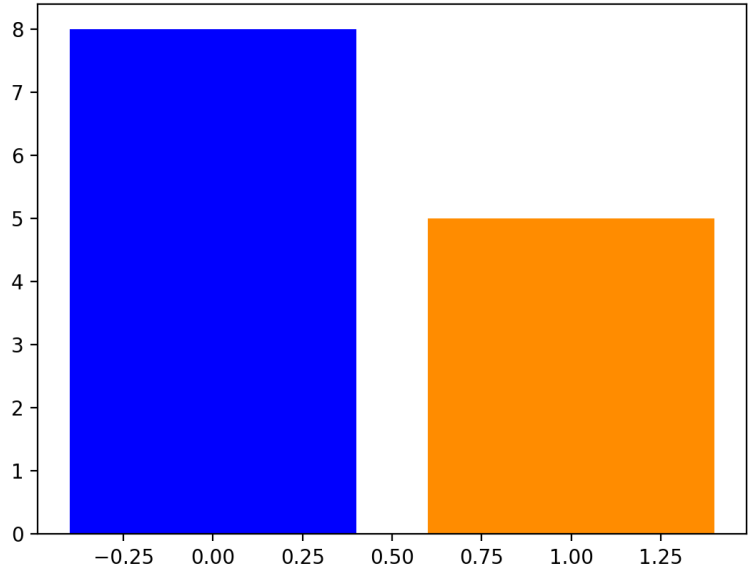
- Finish data visualization intro
- Dimensionality reduction
- PCA for data visualization

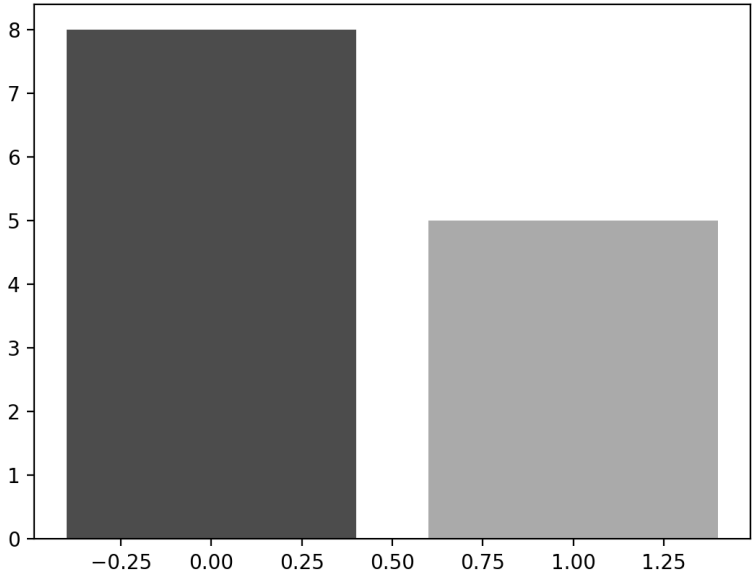
Red/green vs. blue/orange



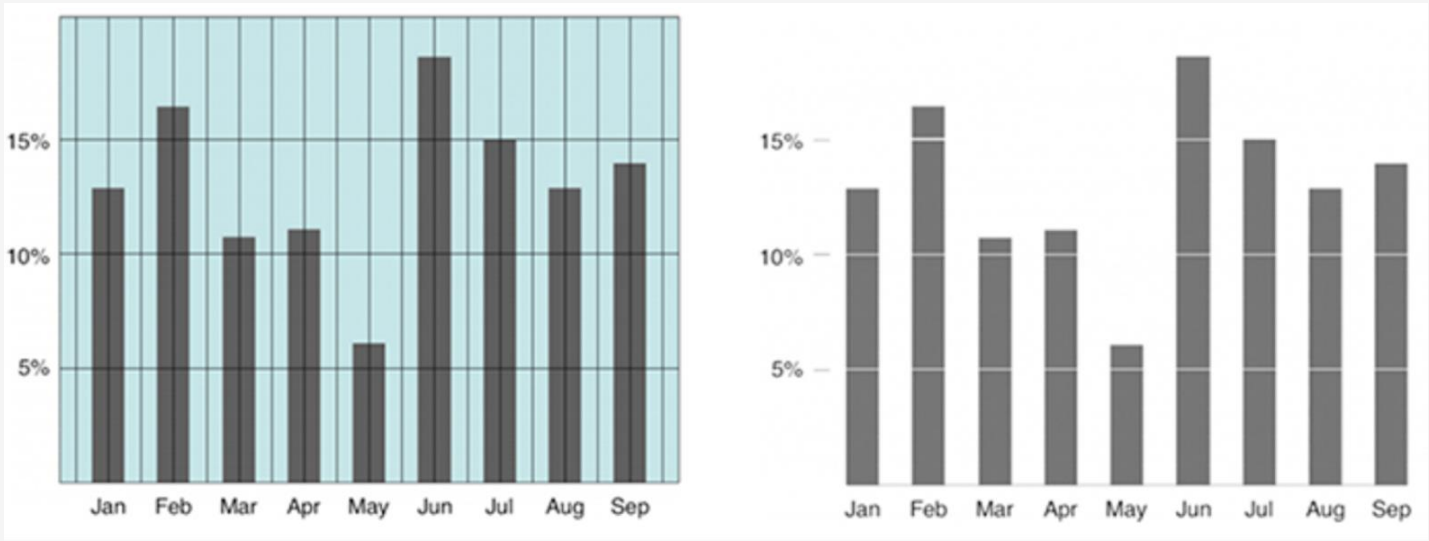


To black and white



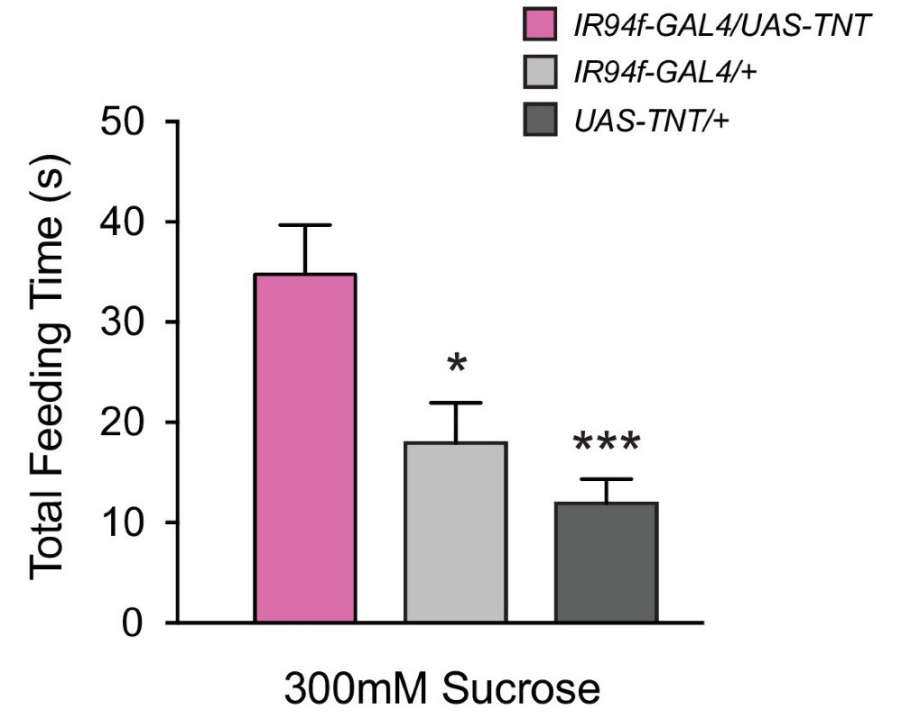


Data::Ink



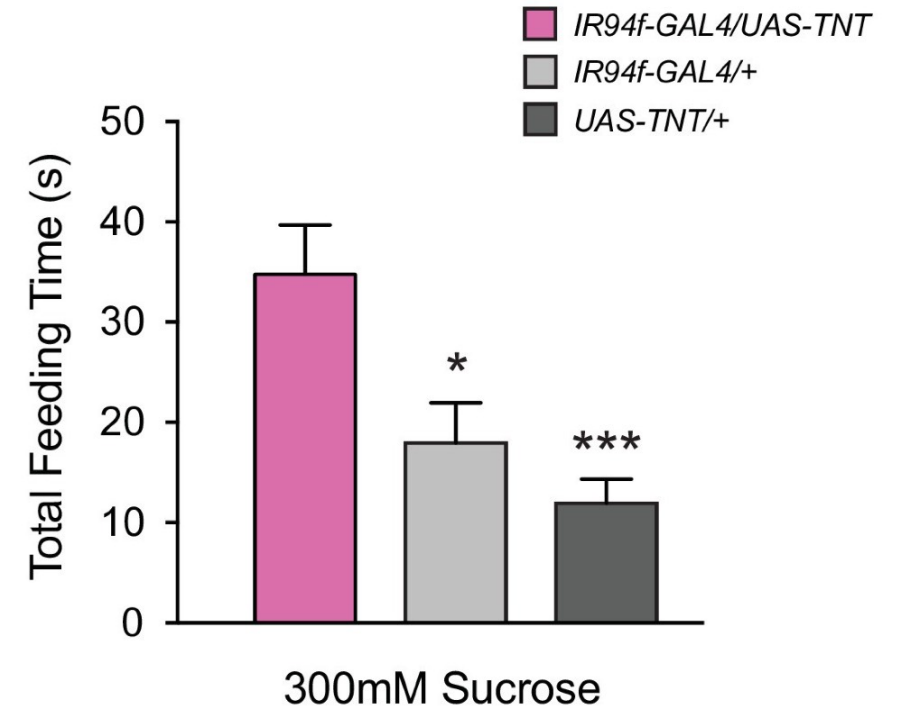
Slide from: Eric Miller

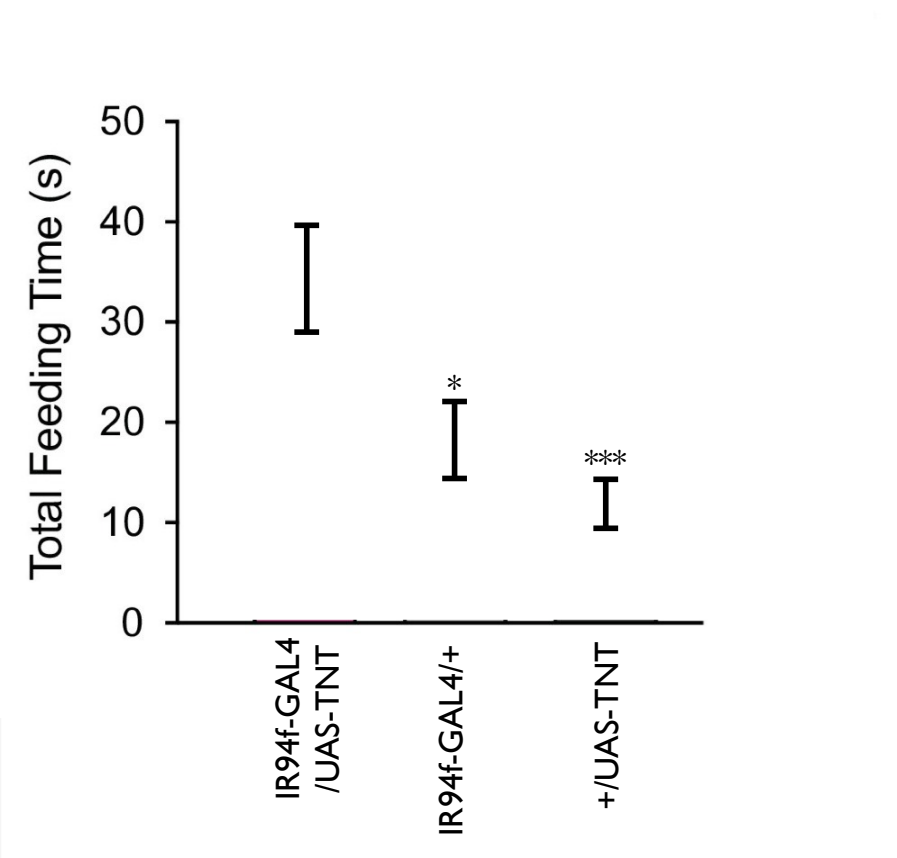
Data::Ink



Slide from: Eric Miller

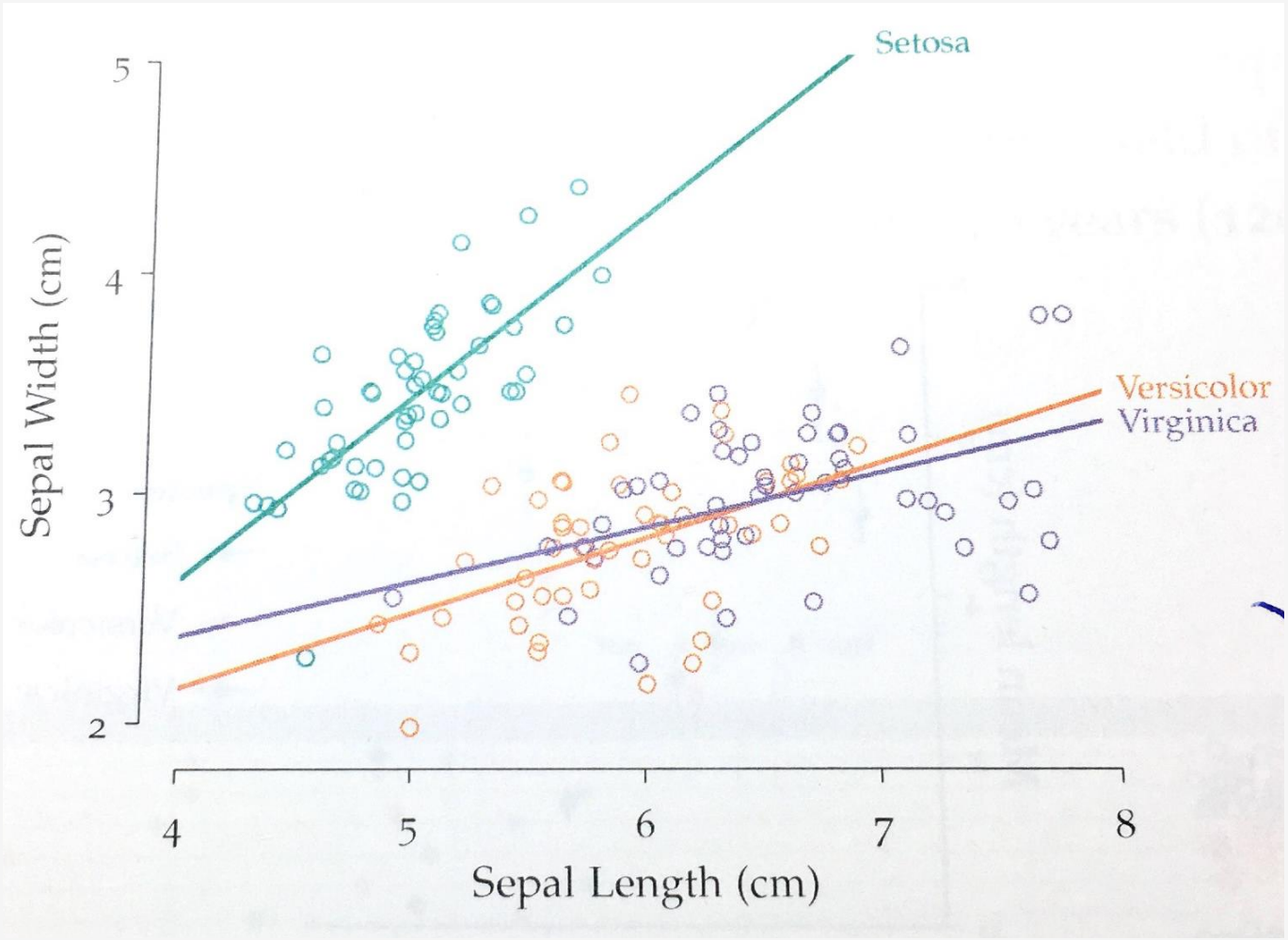
Data::Ink





Slide from: Eric Miller

Data::Ink



Where is the legend?

Slide from: Eric Miller

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?

Slide from: Eric Miller

Outline

- Finish data visualization intro
- Dimensionality reduction
- PCA for data visualization

Principal Components Analysis (PCA)

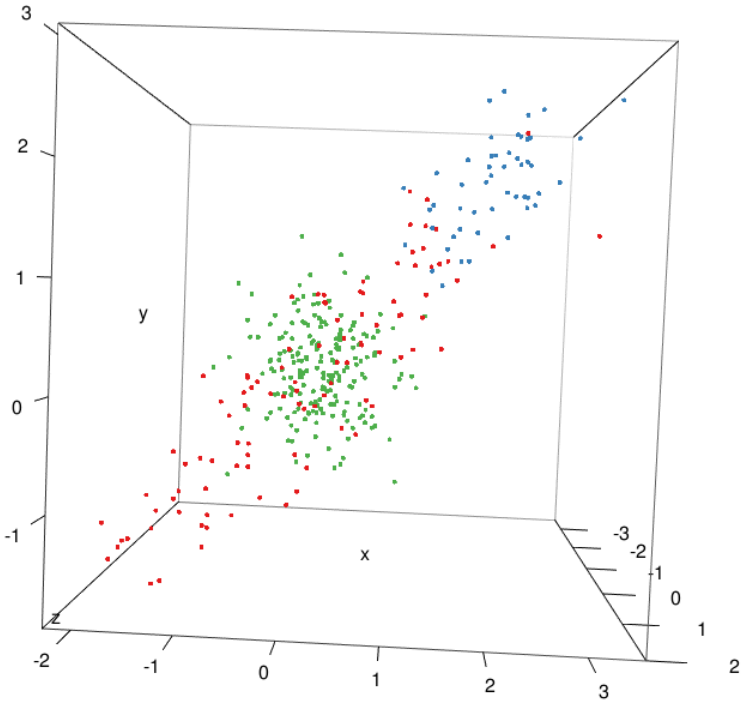
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

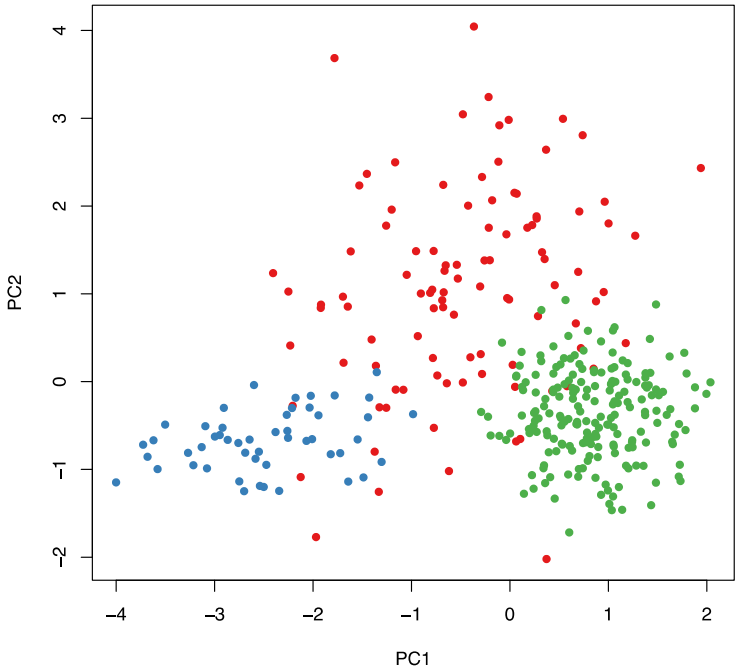
- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction and visualization

- PCA is a linear transformation

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

Principal component analysis

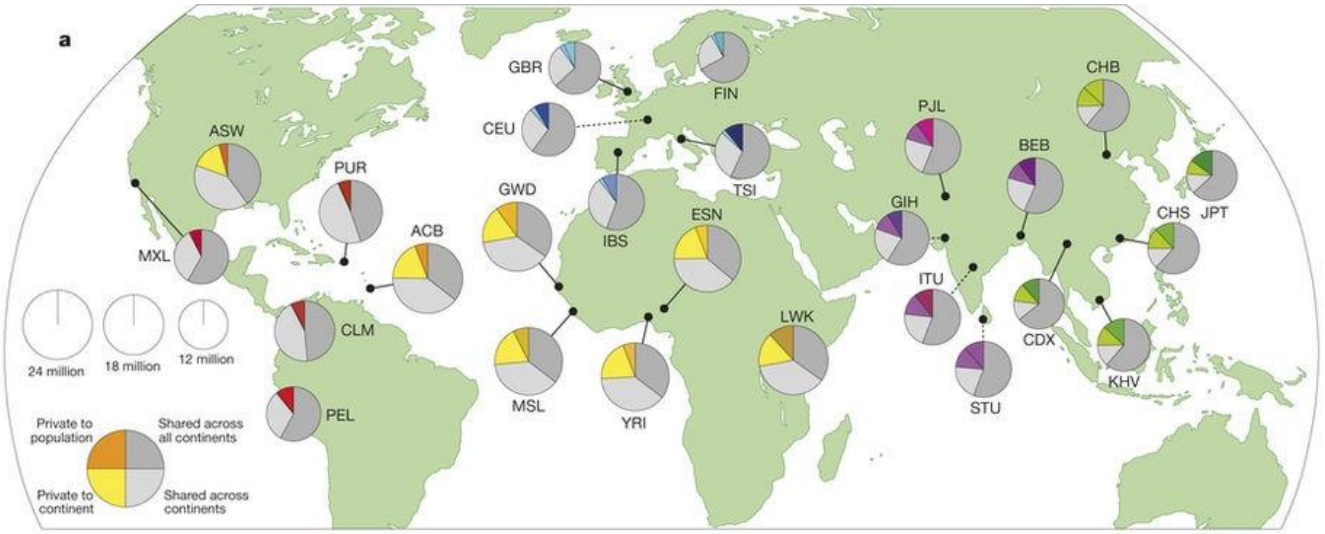




Slide: Iain Mathieson

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/

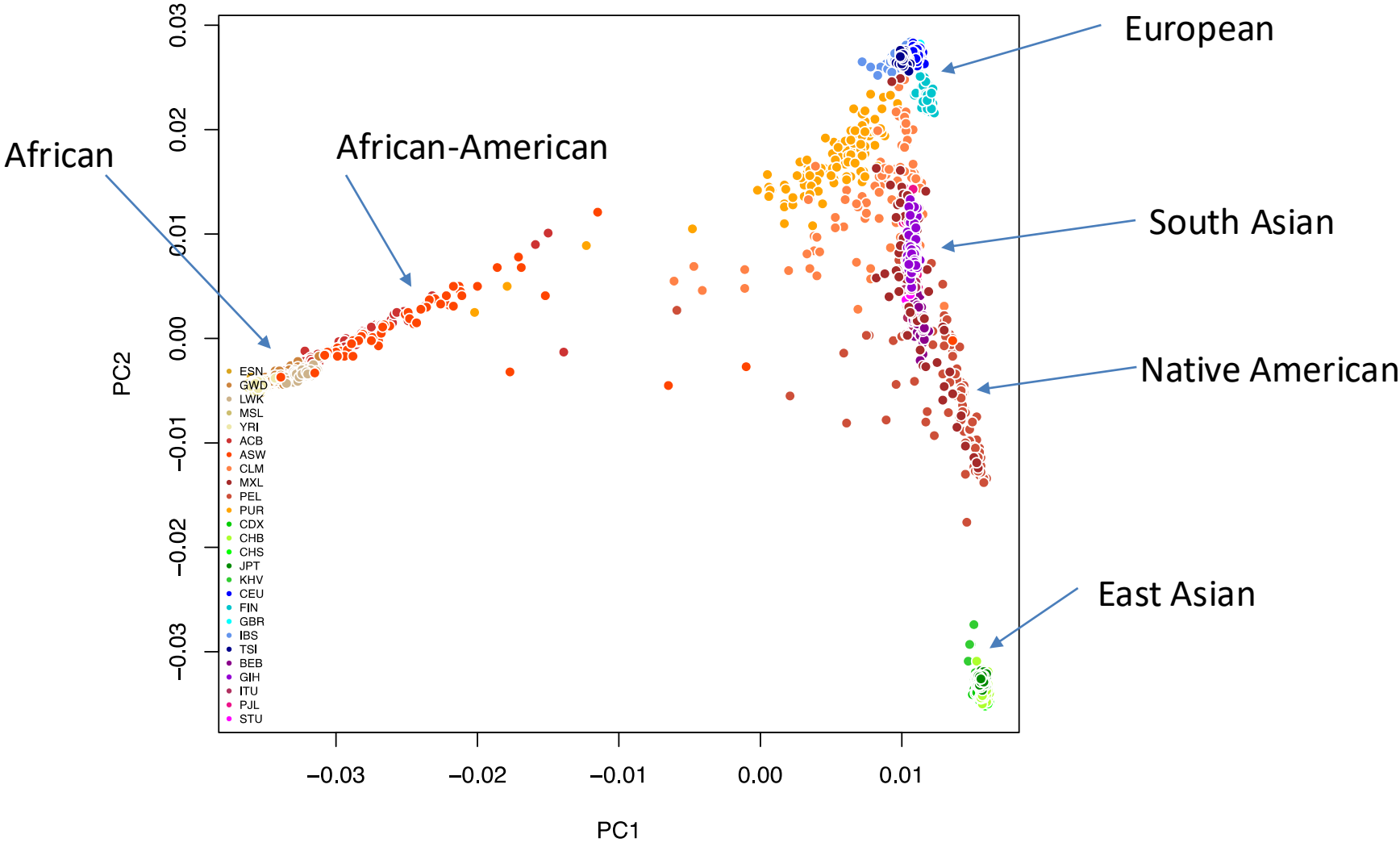


The 1000 Genomes Project Consortium, 2015; https://www.nature.com/articles/nature15393

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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     0|0
```

Global population structure

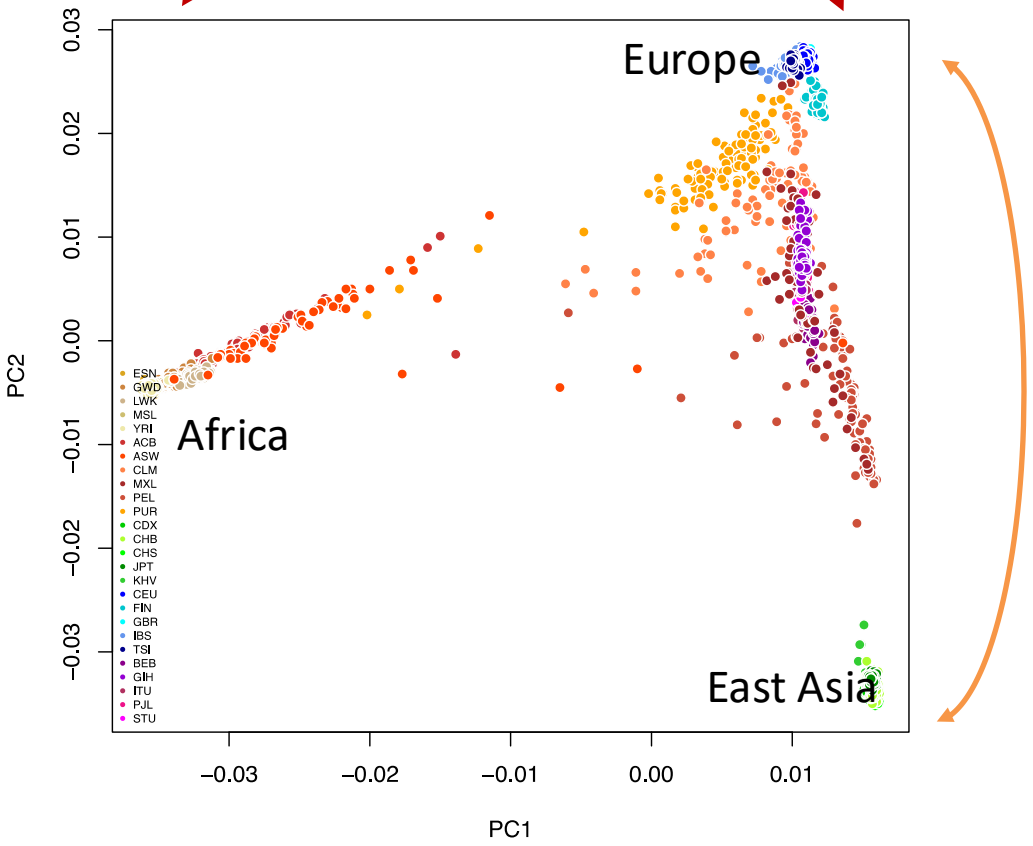


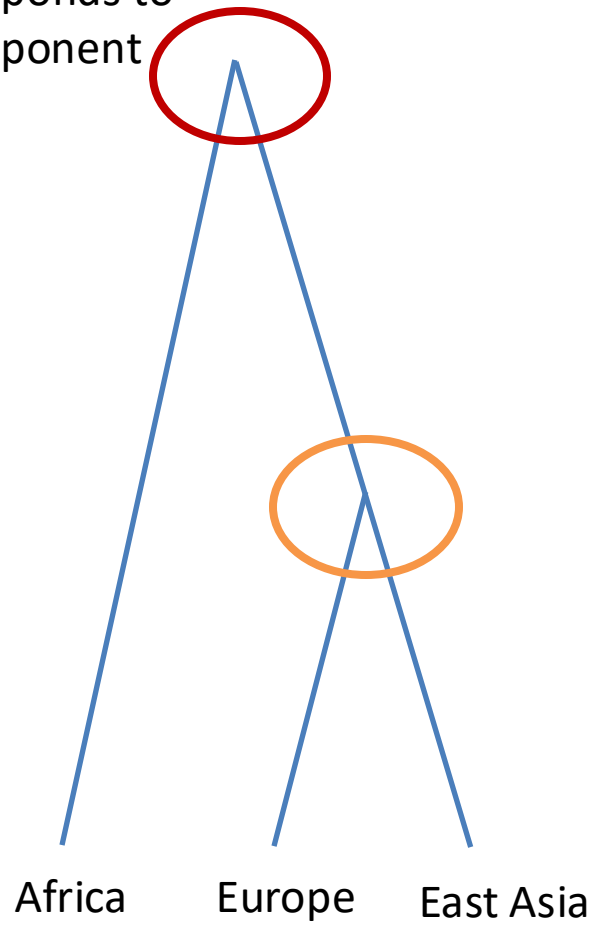
Slide: Iain Mathieson

What causes these patterns?

1. Populations splits separate populations

Oldest split corresponds to first principal component



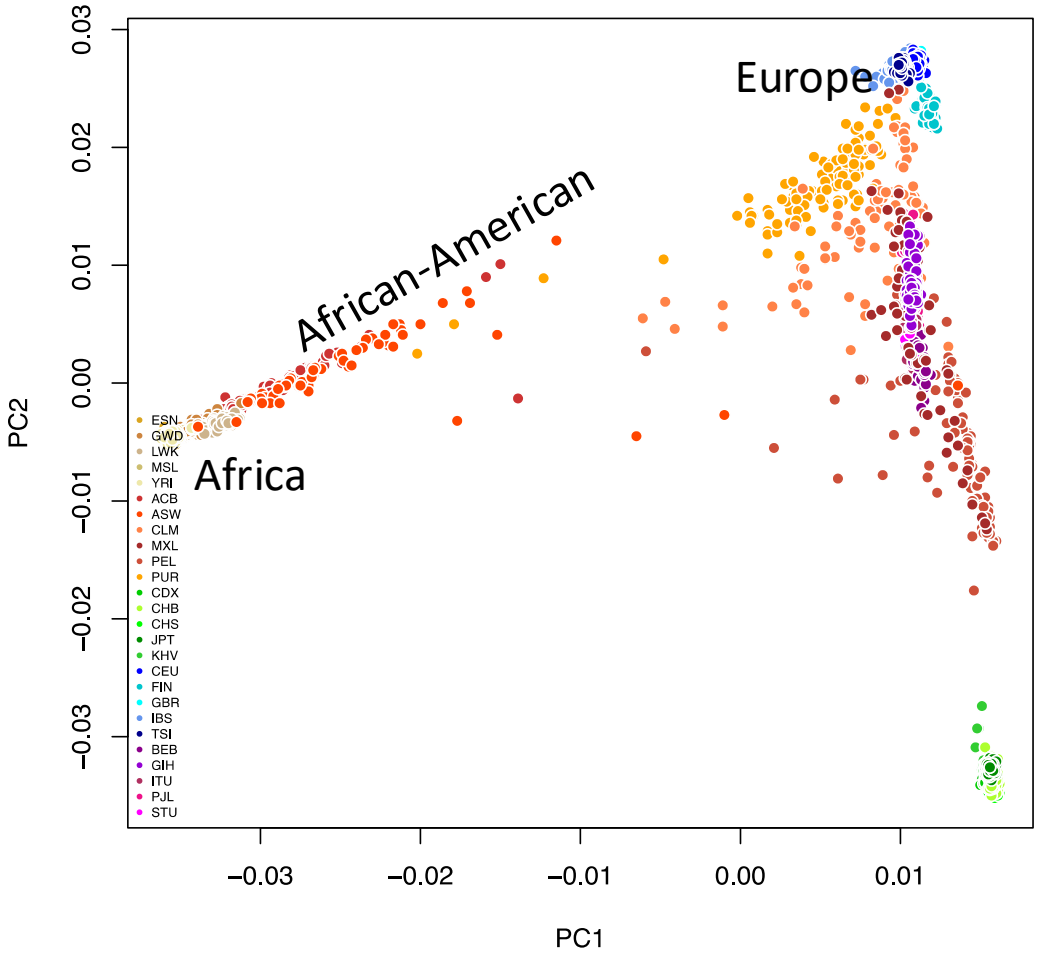


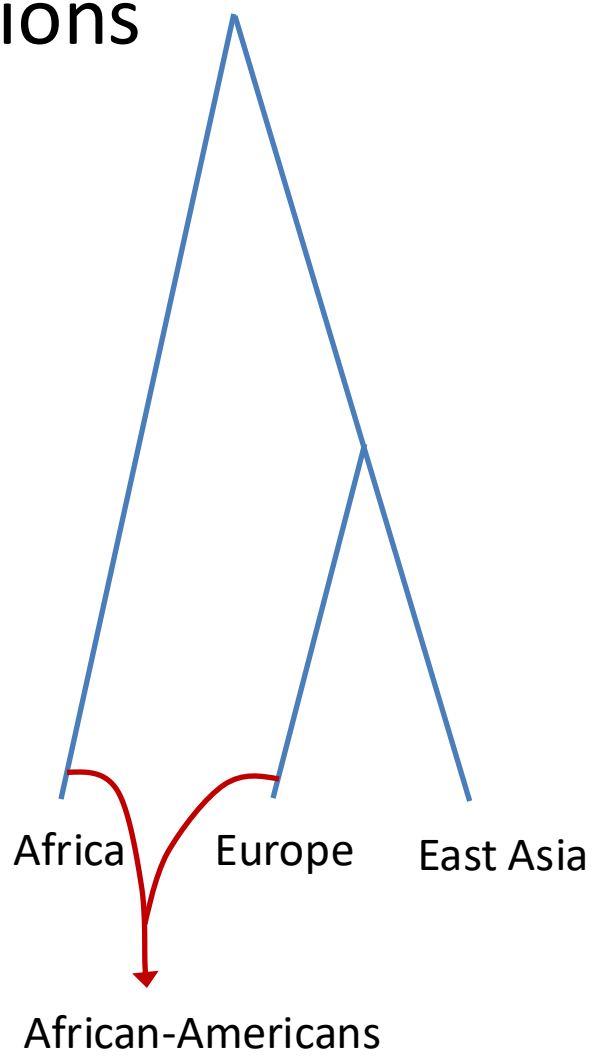
Slide: Iain Mathieson

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

What causes these patterns?

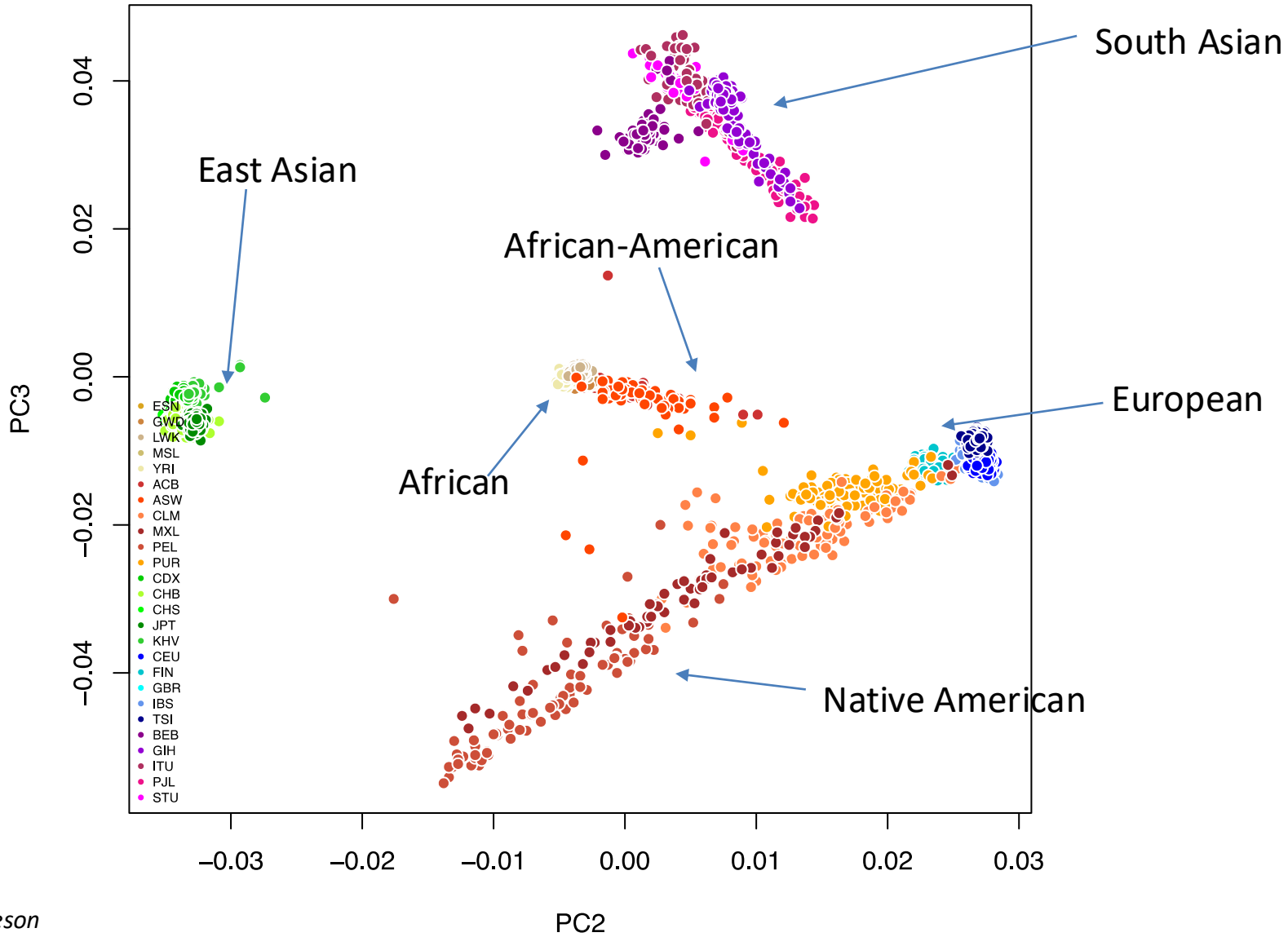
2. Admixture merges populations





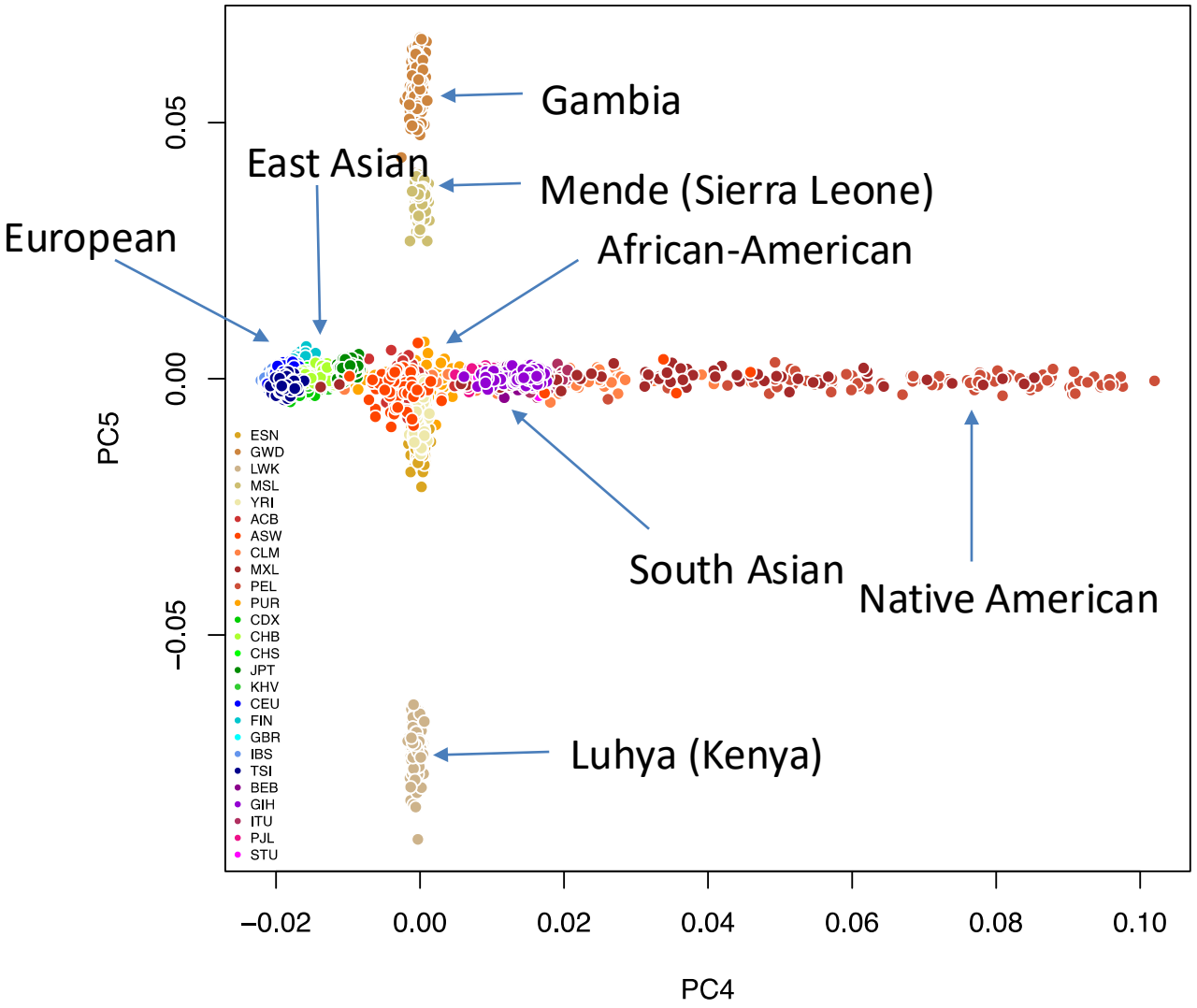
Slide: Iain Mathieson

Global population structure



Slide: Iain Mathieson

Global population structure

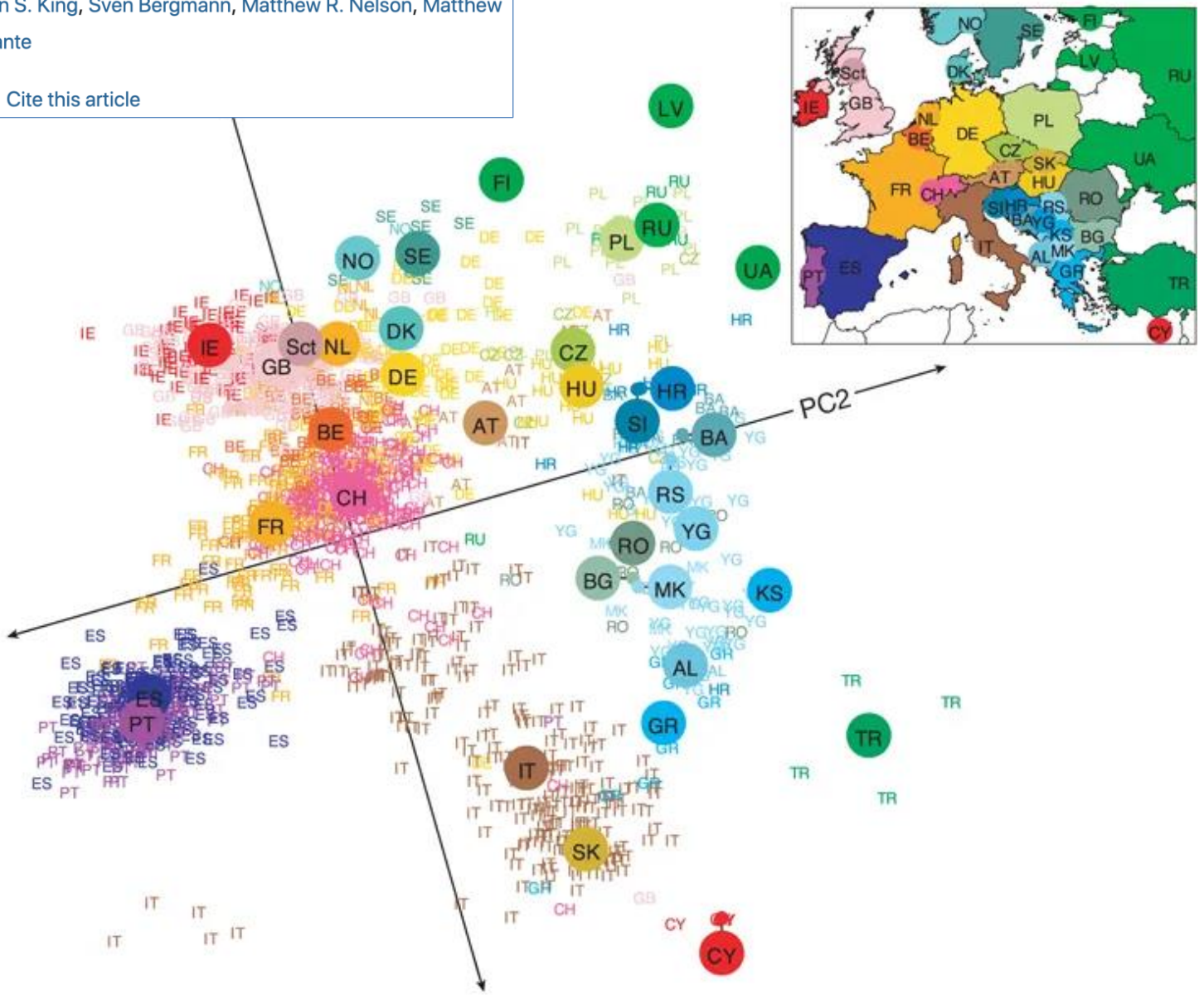


Slide: Iain Mathieson

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

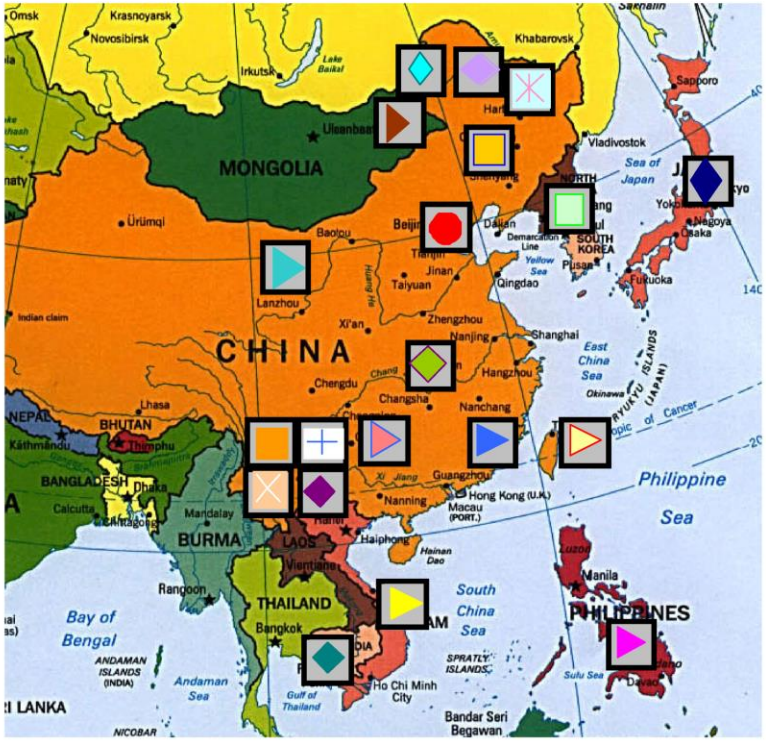


Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays

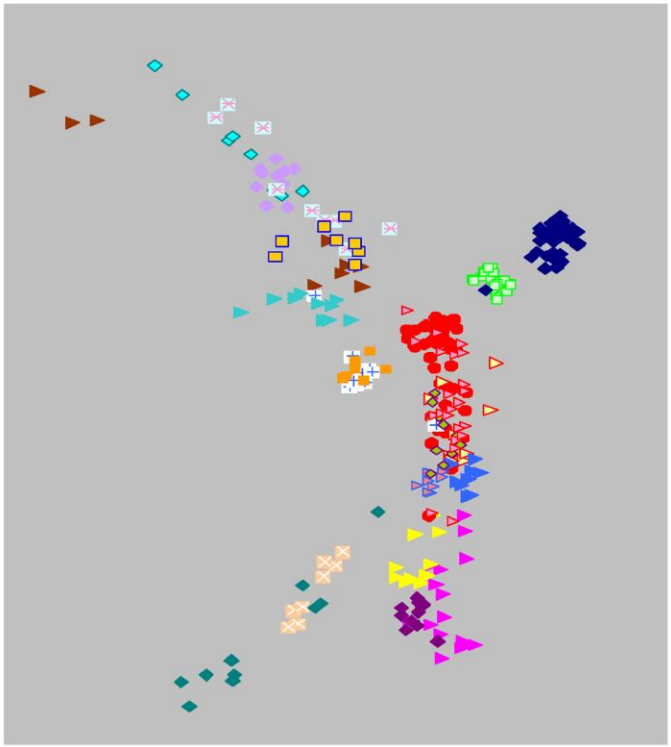
Chao Tian, Roman Kosoy, Annette Lee, Michael Ransom, John W. Belmont, Peter K. Gregersen, Michael F. Seldin

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

C



D

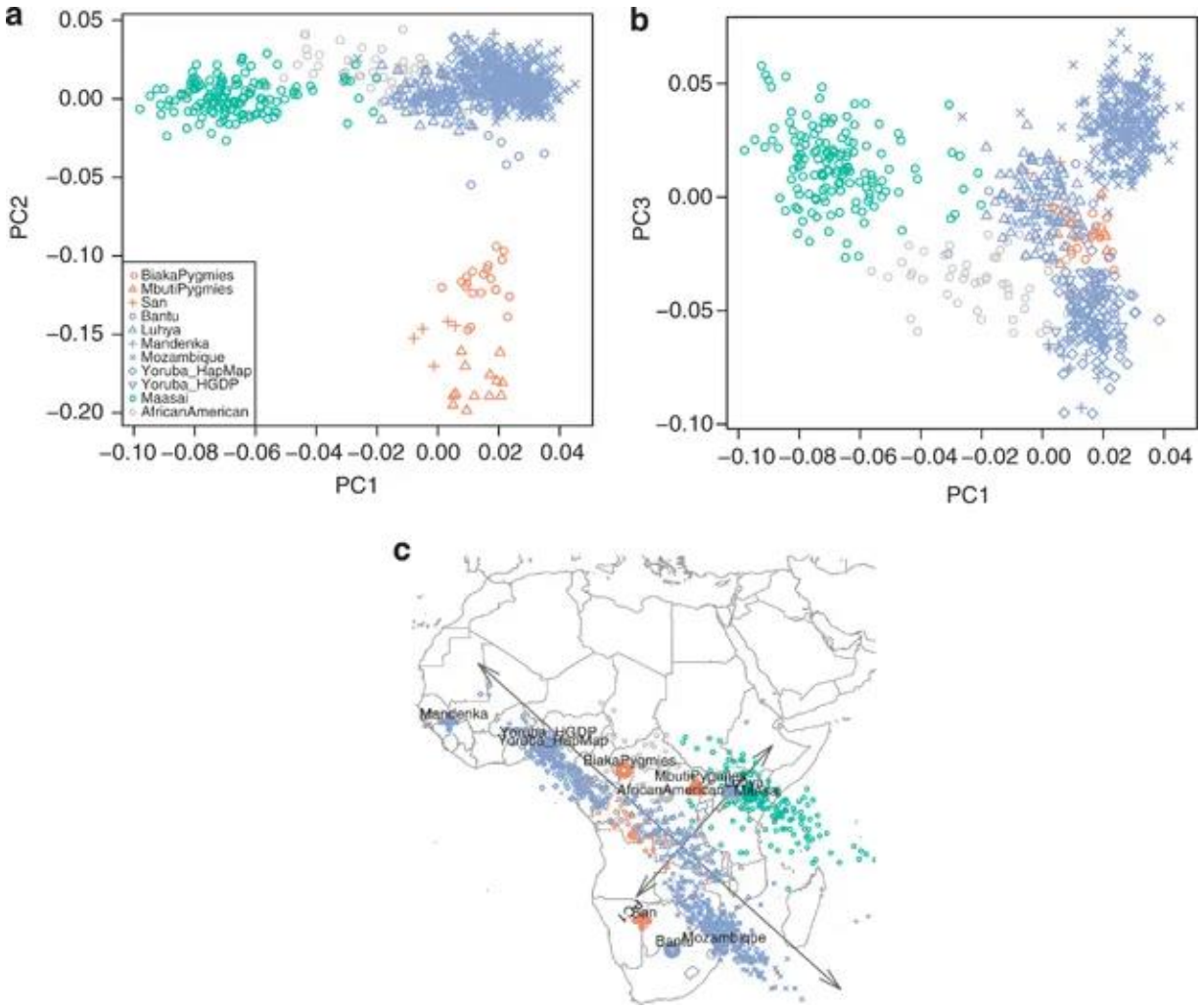


FIL
VIET
LAHU
DAI
CAMB
CHB
MGL
ORQ
DAUR
KOR
TWN
YI
HEZ
MIAO
NAXI
SHE
TU
TUJ
XIBO
CHA
JPT
YAK

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



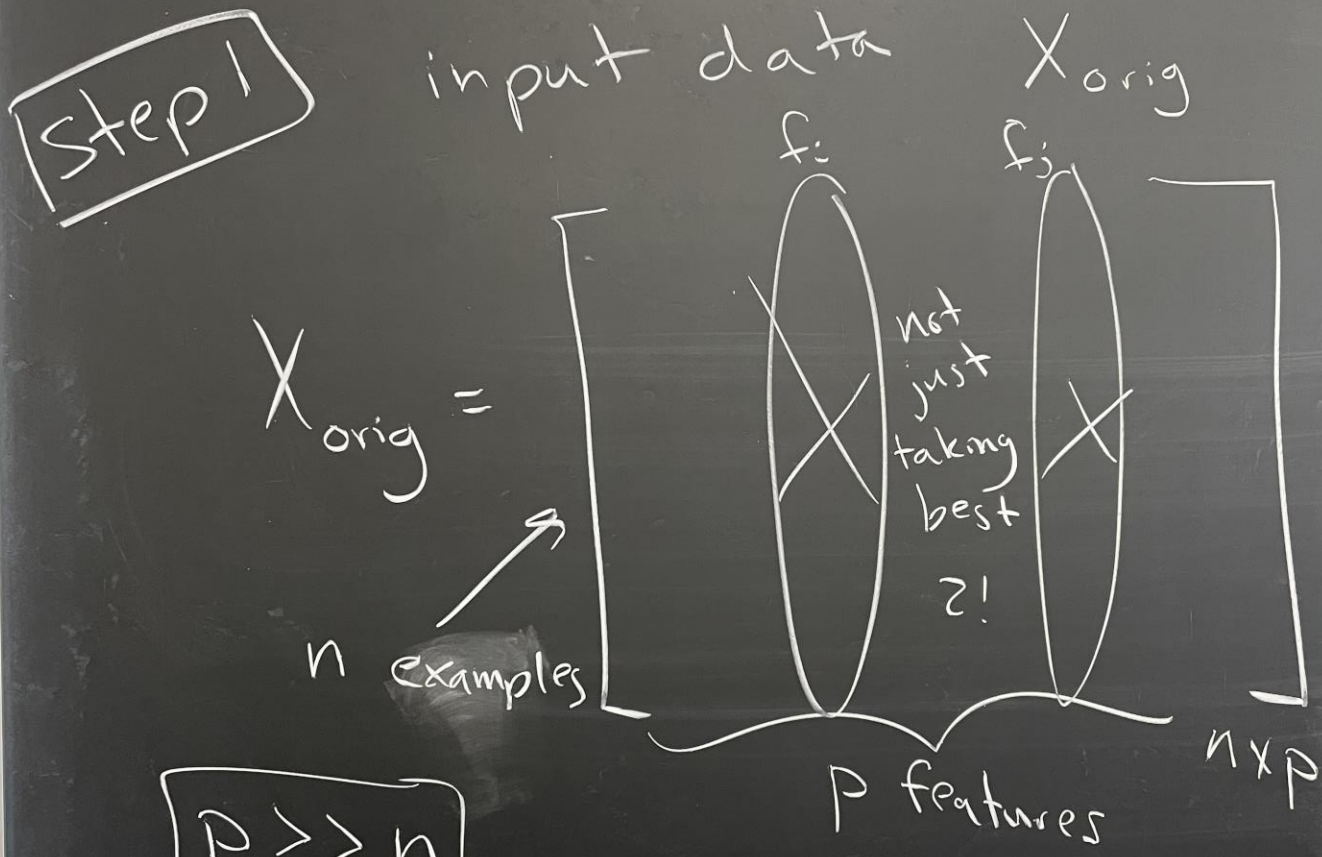
Outline

- Finish data visualization intro
- Dimensionality reduction
- PCA for data visualization

Principal Component Analysis

Step 1

input data X_{orig}



$p >> n$

sometimes!

(PCA)

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

$r = 2$ (often)

no label! (i.e. $\vec{y}$)

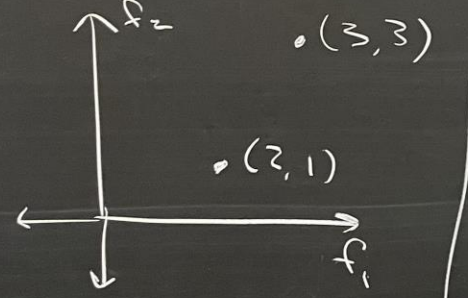
unsupervised

Step 2 Subtract of column-wise mean

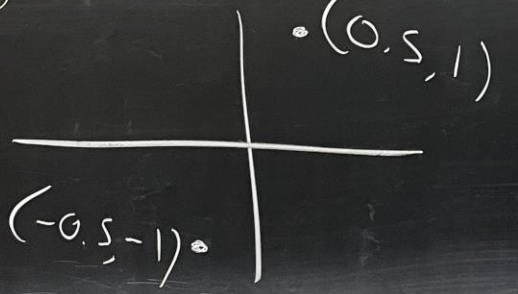
$$X_{orig} = \begin{matrix} f_1 & f_2 \\ \end{matrix} \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix}$$

$\bar{f}_1 = 2.5$

$\bar{f}_2 = 2$



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



Step 3 compute covariance matrix A

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

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

$(n-1$ to make unbiased$)$

for 2 features

$A^T = A$ Symmetric

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

Step 4 Compute eigenvalues & eigenvectors of A

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

Solve for λ

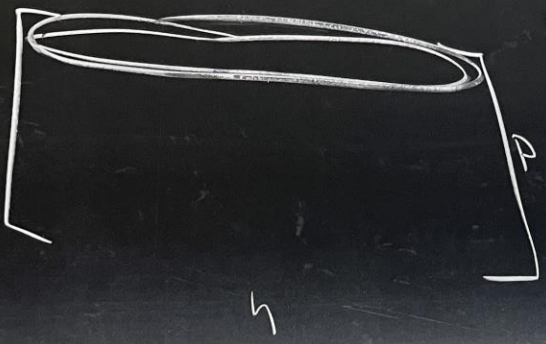
plug back in

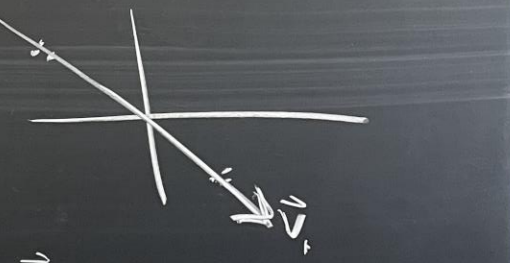
identity

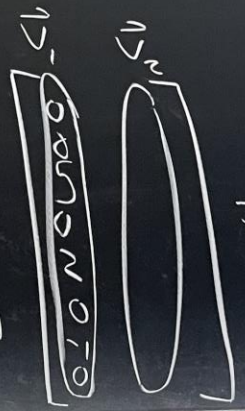
$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

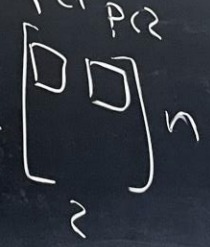
$p = 2$

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









Step 5

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

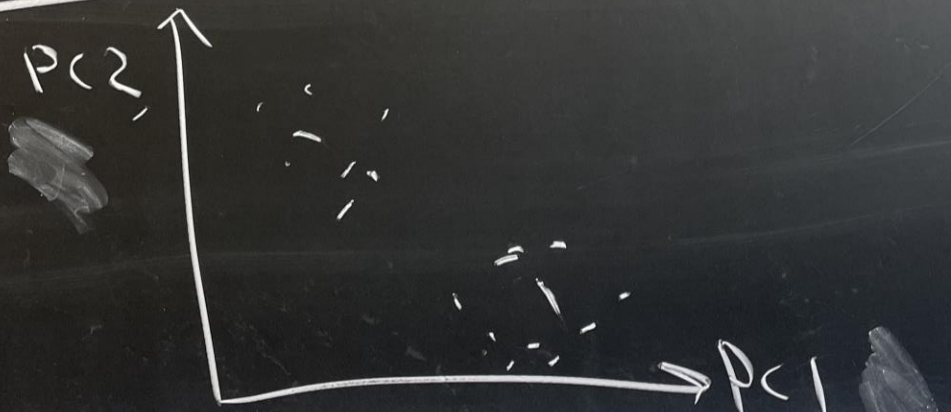
matrix of eigenvectors

$\lambda_1 \quad \lambda_2 \cdots \lambda_r$

$$T_{nxr} = X_{nxp} W_{pxr}$$

transformed data

Step 6 plot to visualize!



Handout 16

Step 1

$$X_{orig} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ \\ \\ \\ (f_1) \end{matrix} \quad \begin{matrix} \\ \\ \\ \\ \\ (f_2) \end{matrix}$$

$\bar{f}_{1,orig} = \frac{1}{2}$, $\bar{f}_{2,orig} = \frac{1}{2}$

Step 2

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

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

Step 3

$$\text{var}(f_1) = \frac{1}{6-1}\left(\left(-\tfrac{1}{2}\right)\left(-\tfrac{1}{2}\right)\cdot 6\right) = \frac{1}{5}\cdot\frac{1}{4}\cdot 6$$

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

$$\text{cov}(f_1, f_2) = \frac{1}{5}\left(-\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)\cdot 6 = \frac{-3}{10}$$

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

Step 4

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

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

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

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

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

$$\lambda\left(\lambda - \frac{3}{5}\right) = 0$$

Sort by high → low magnitude

$$\lambda_1 = \frac{3}{5},\ \lambda_2 = 0$$

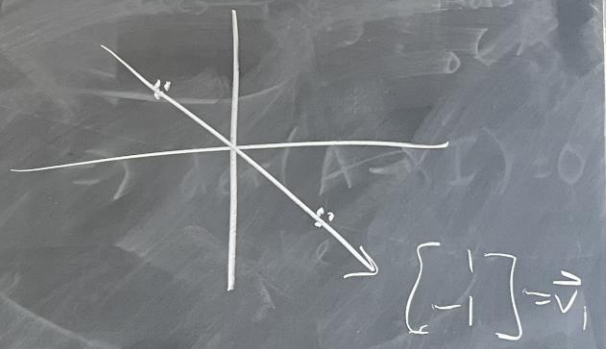
$$A\vec{v} = \lambda\vec{v}$$

$W = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 \\ 1 & 1 \\ -1 & 1 \end{bmatrix}$ matrix

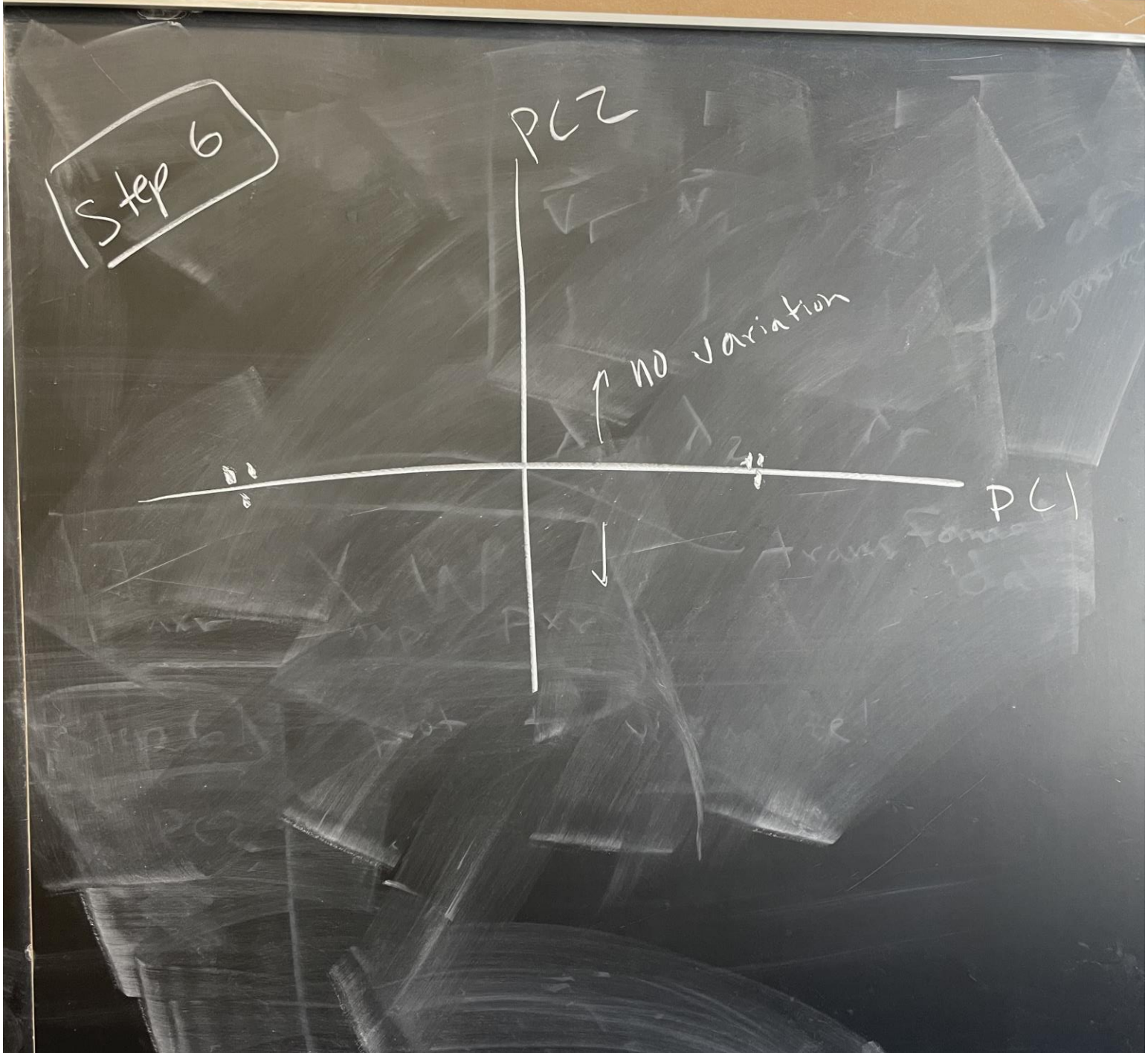
$\lambda_1 = \frac{3}{5}$ $\lambda_2 = 0$

$T_2 = X W_2 = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ \vdots & \vdots \\ \frac{1}{2} & -\frac{1}{2} \\ \vdots & \vdots \end{bmatrix}$ dot product $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$

$P(1)$ $P(2)$



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


Step 6
PC2
↑ no variation
PC1
↓