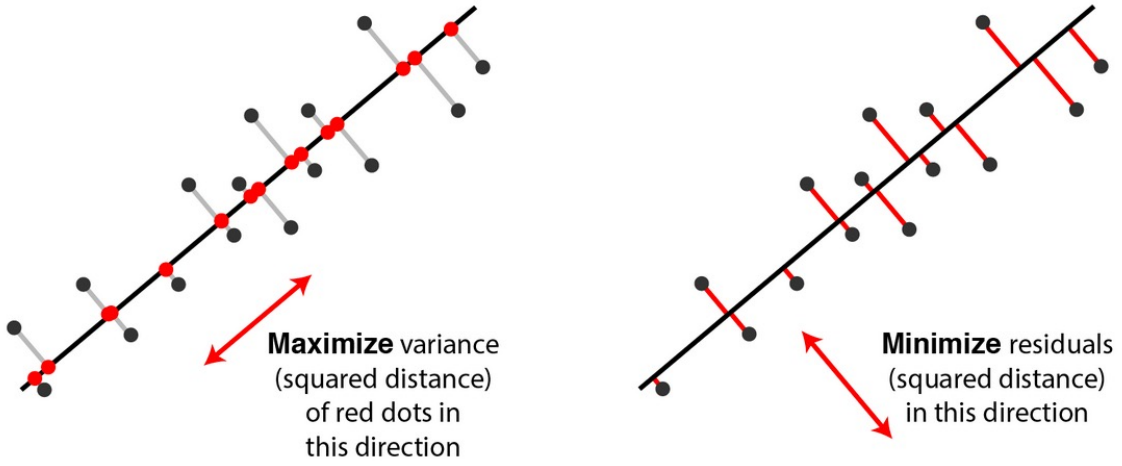
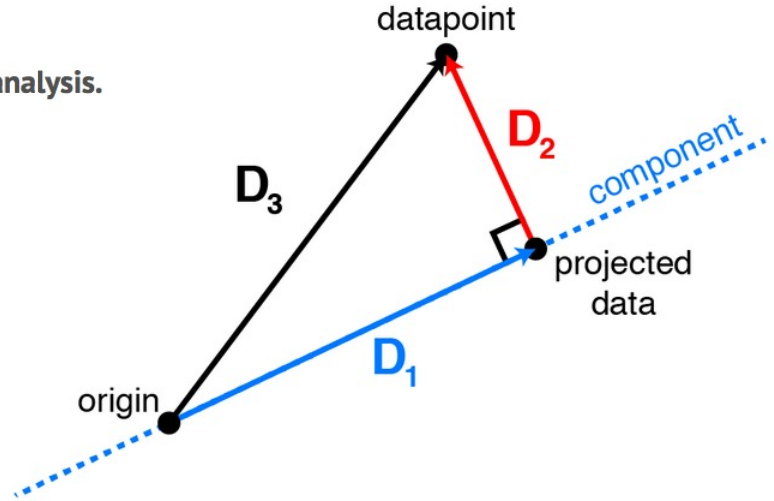
PRINCIPAL COMPONENTS ANALYSIS (PCA)

http://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/



Two equivalent views of principal component analysis.



$$D_3^2 = D_1^2 + D_2^2$$

$$\text{initial variance} = \text{remaining variance} + \text{lost variance}$$

$$\|a_i\|^2 = \|w_i c\|^2 + \|a_i - w_i c\|^2$$

this is constant | maximize this **or** minimize this

Maximizing variance in principal component space is equivalent to minimizing least-squares reconstruction error. Consider a datapoint $\mathbf{a}_i$ (row i of the data matrix $\mathbf{X}$). Assuming the data are mean-centered, the projection of $\mathbf{a}_i$ onto the principal components relates the remaining variance to the squared residual by the Pythagorean theorem. Choosing the components to maximize variance is the same as choosing them to minimize the squared residuals.

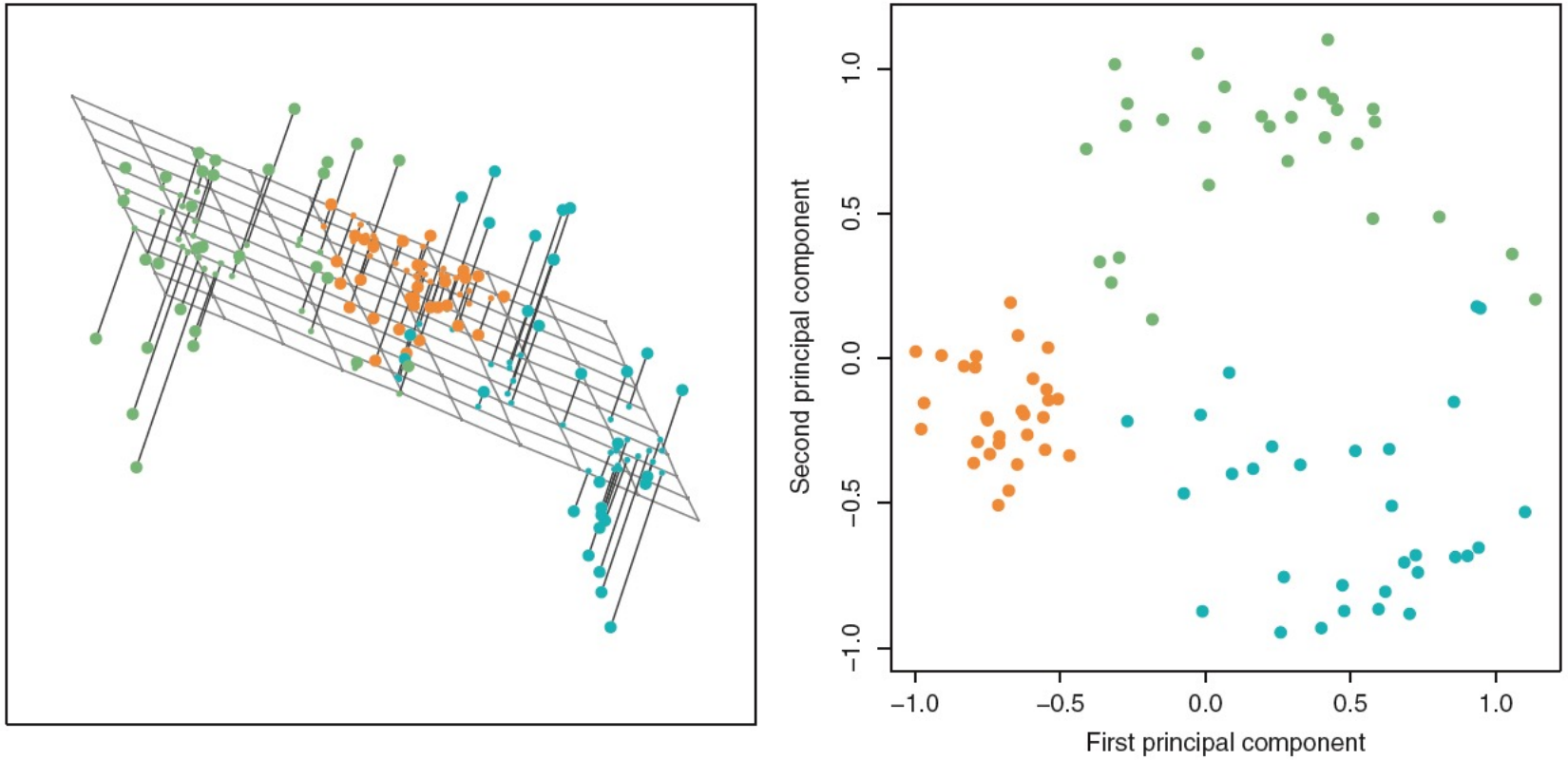


FIGURE 10.2. *Ninety observations simulated in three dimensions.* Left: *the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.*

ISL

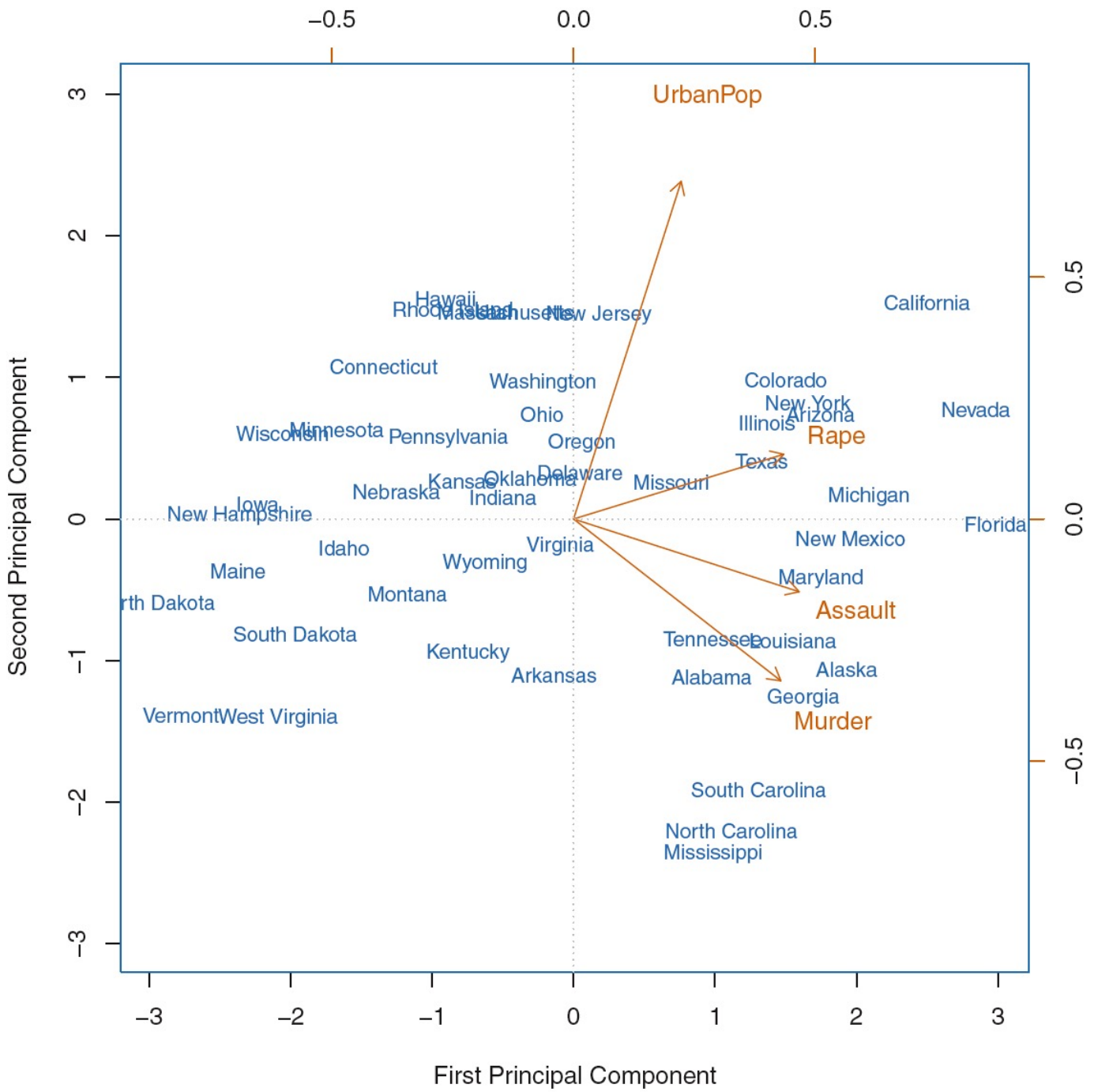


FIGURE 10.1. *The first two principal components for the* `USArrests` *data. The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for* `Rape` *on the first component is* 0.54*, and its loading on the second principal component* 0.17 *(the word* `Rape` *is centered at the point* $(0.54, 0.17)$*). This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.*

ISL

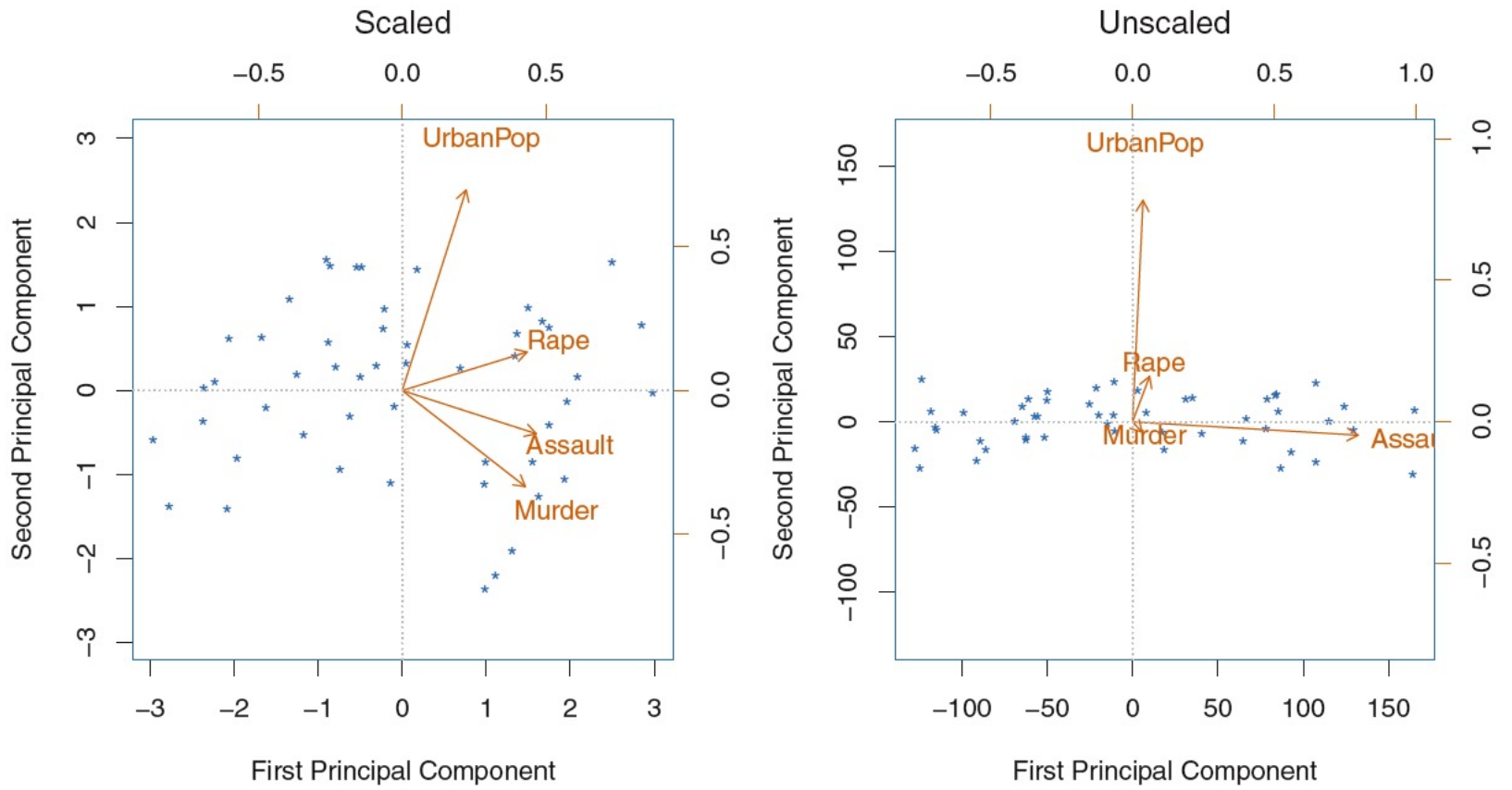


FIGURE 10.3. *Two principal component biplots for the* `USArrests` *data.* Left: *the same as Figure 10.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* `Assault` *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*

ISL

PCA EXAMPLE

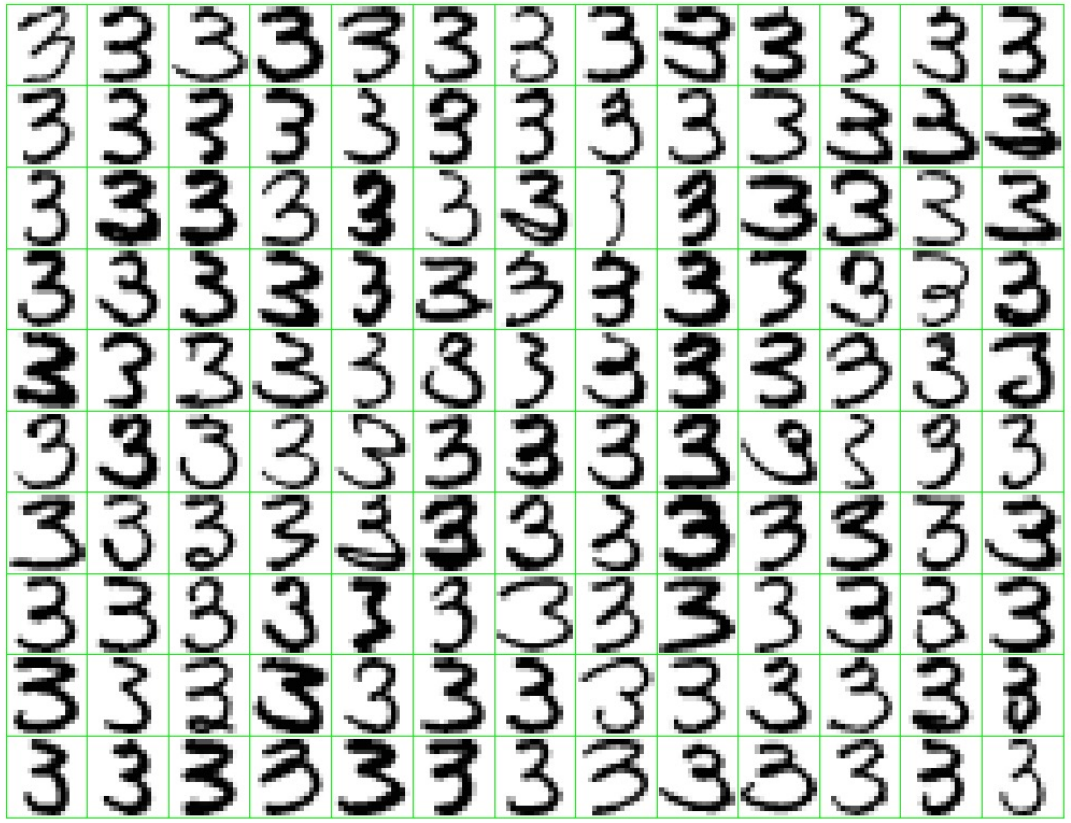


FIGURE 14.22. *A sample of* 130 *handwritten 3's shows a variety of writing styles.*

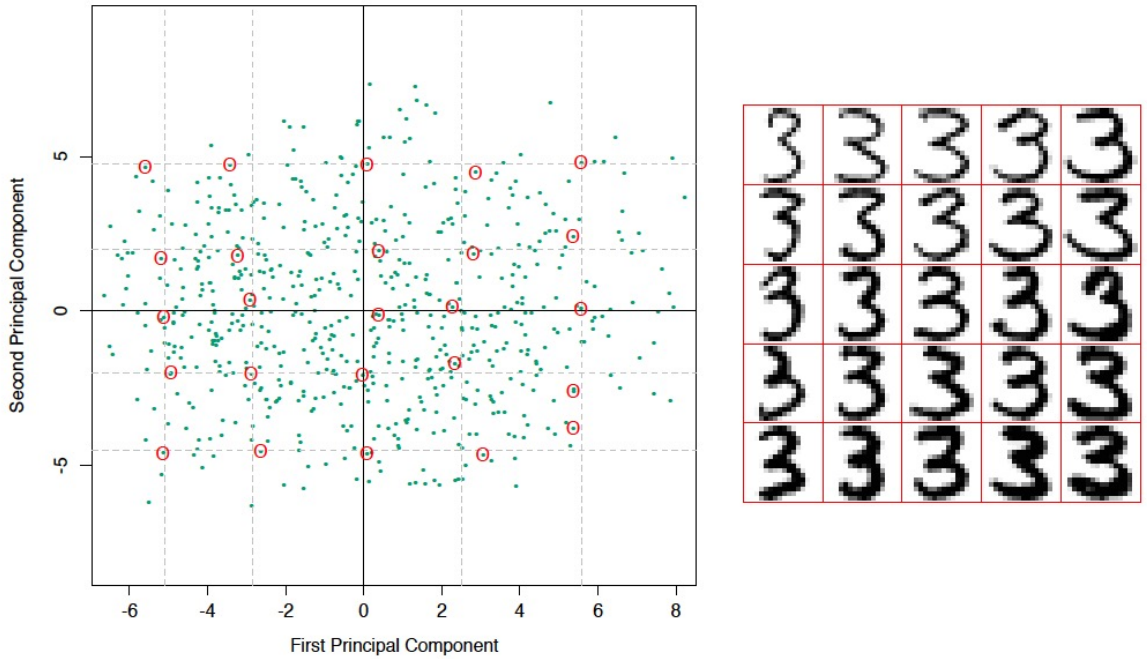
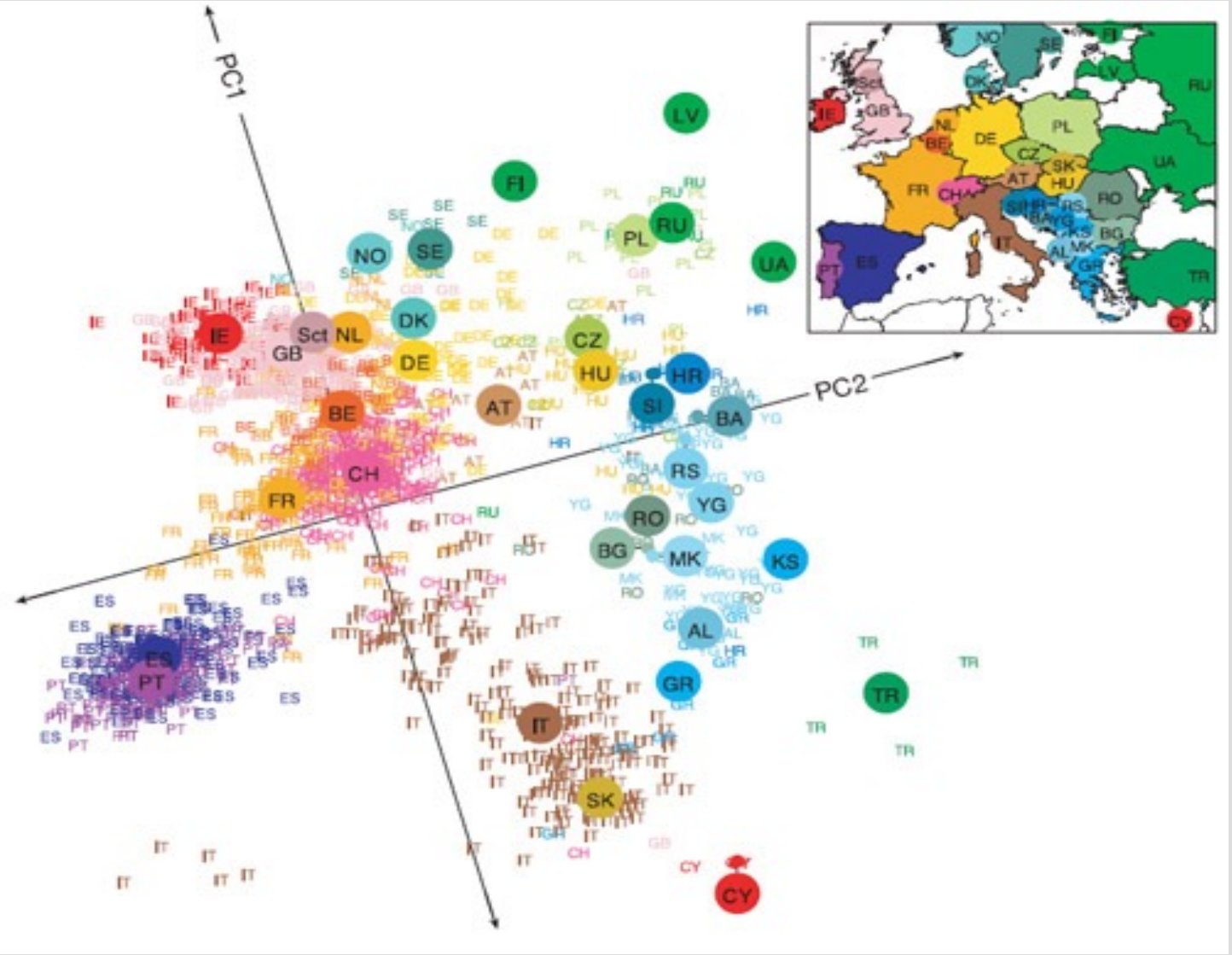


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

From ESL



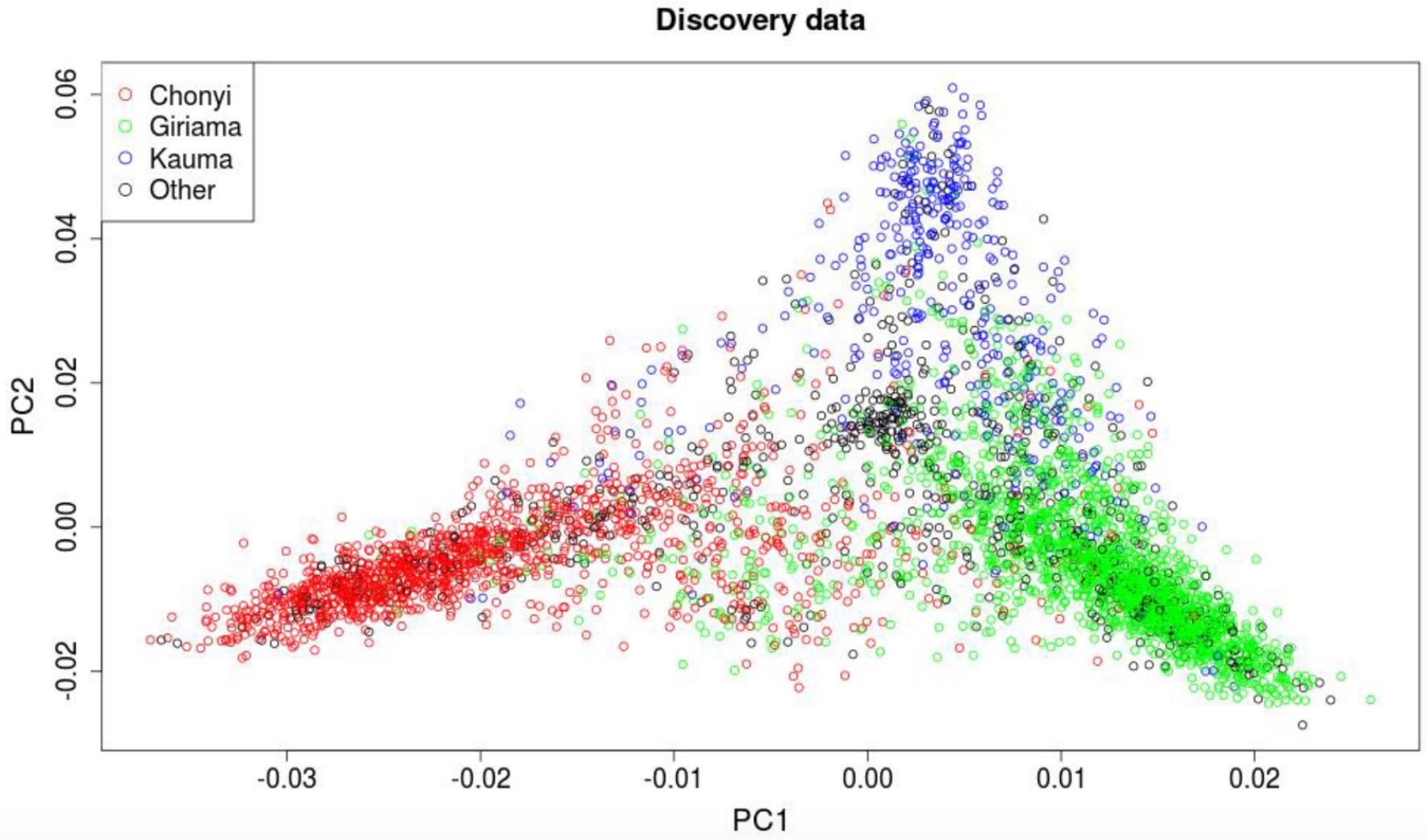
Novembre et al. 2008, Nature. "Genes mirror geography within Europe"

- Collected n=3000 Europeans and genotyped at p=500,000 variants
- "Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans."
- This "2D summary" = PCs 1 & 2
- "The results emphasize that when mapping the genetic basis of a disease phenotype, **spurious associations** can arise if genetic structure is not properly accounted for."
- "In addition, the results are relevant to the prospects of genetic ancestry testing; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres."

Figure S5. The first and second axes of the principal components analysis in the discovery dataset using 168,217 SNPs. Color coding represents A) reported ethnicity and B) case control status.

A)

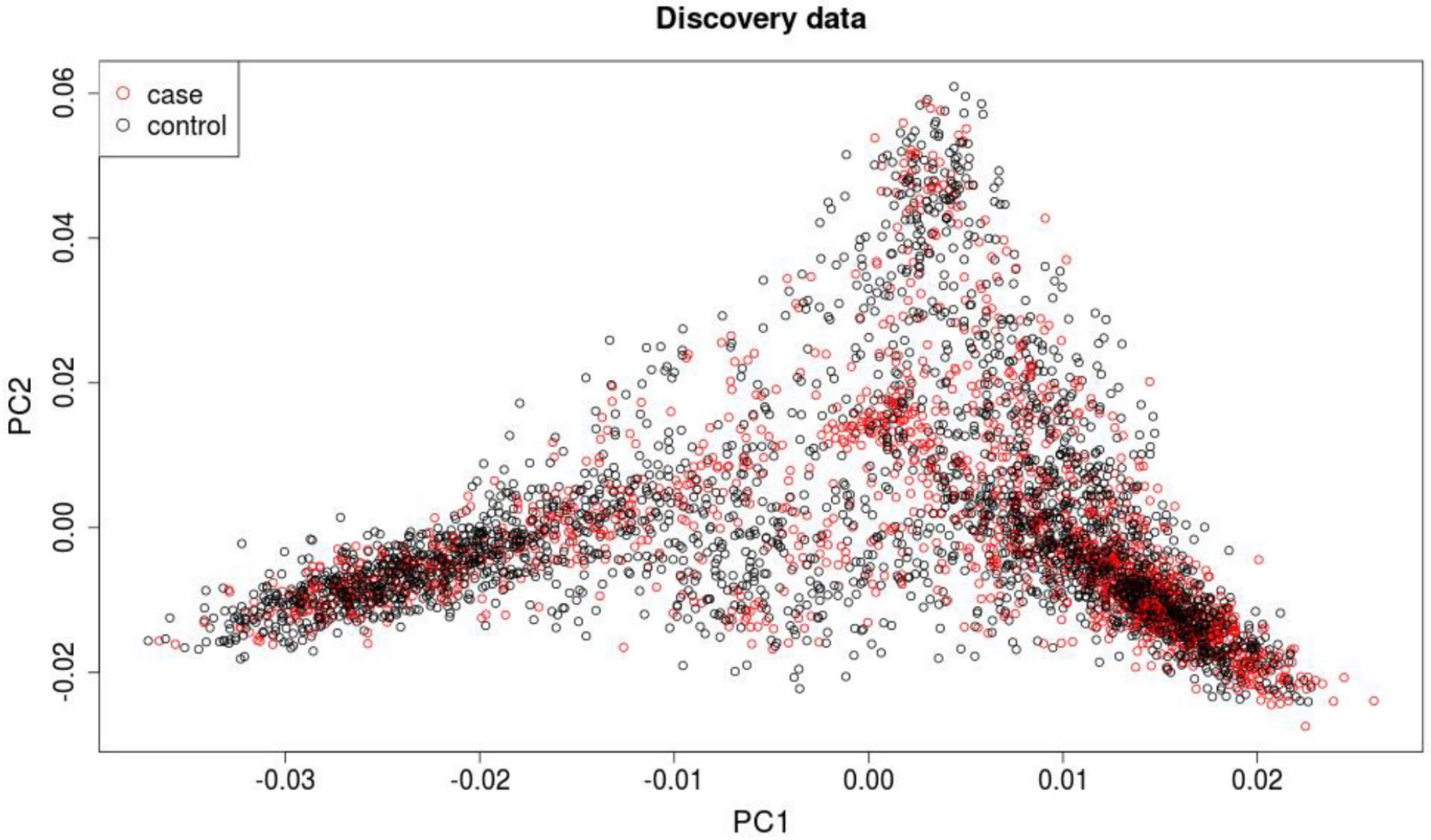
B on next slide



From a study of genetics of Bacteraemia in Kenyan Children

(Rautanen et al. 2016 AJHG)

B)



Very important that cases and controls are matched w.r.t their genome-wide ancestry!

Otherwise spurious genetic associations would arise.

PCs can help us here.

From a study of genetics of Bacteraemia in Kenyan Children

(Rautanen et al. 2016 AJHG)

CONFOUNDING BY ANCESTRY

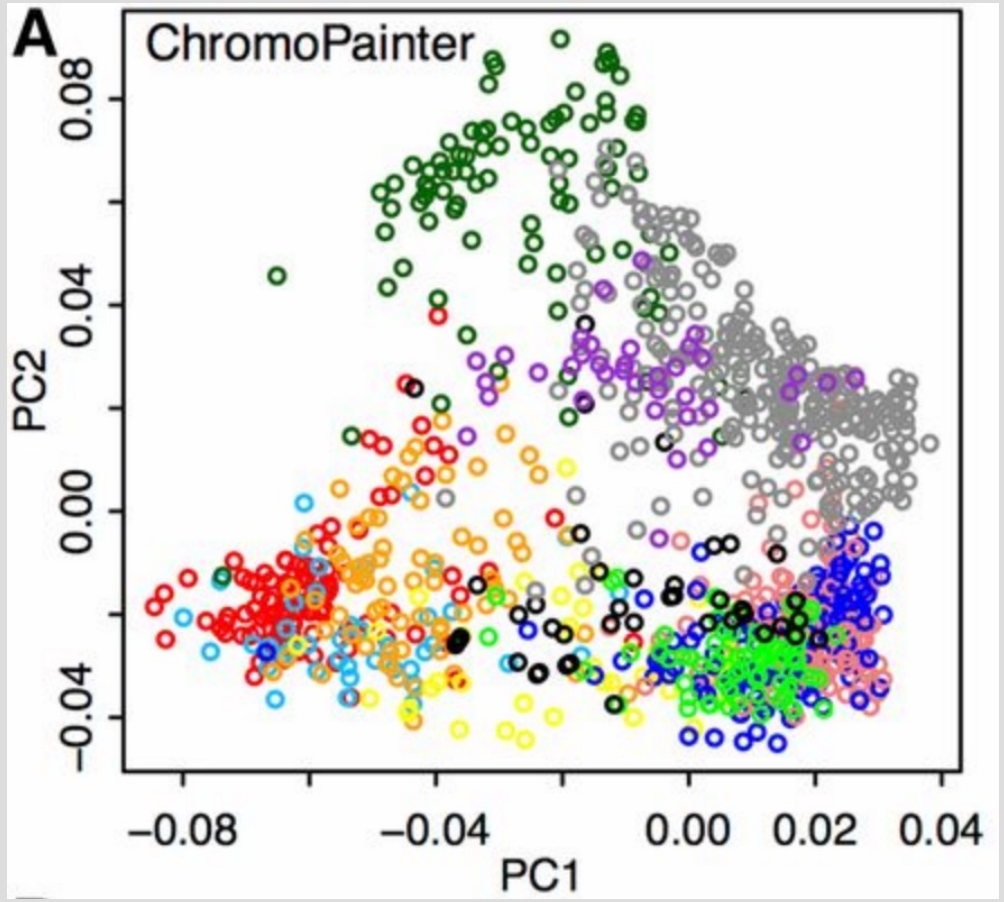
- Consider a genetic variant that has no effect on heart disease but has different regional frequencies
  - Variant ”A” frequency 0.23 in Helsinki region
  - Variant ”A” frequency 0.35 in Oulu region
- Does not show association with disease in Helsinki or in Oulu (because there is none)
- What happens if we do not match well regions of case and control ?

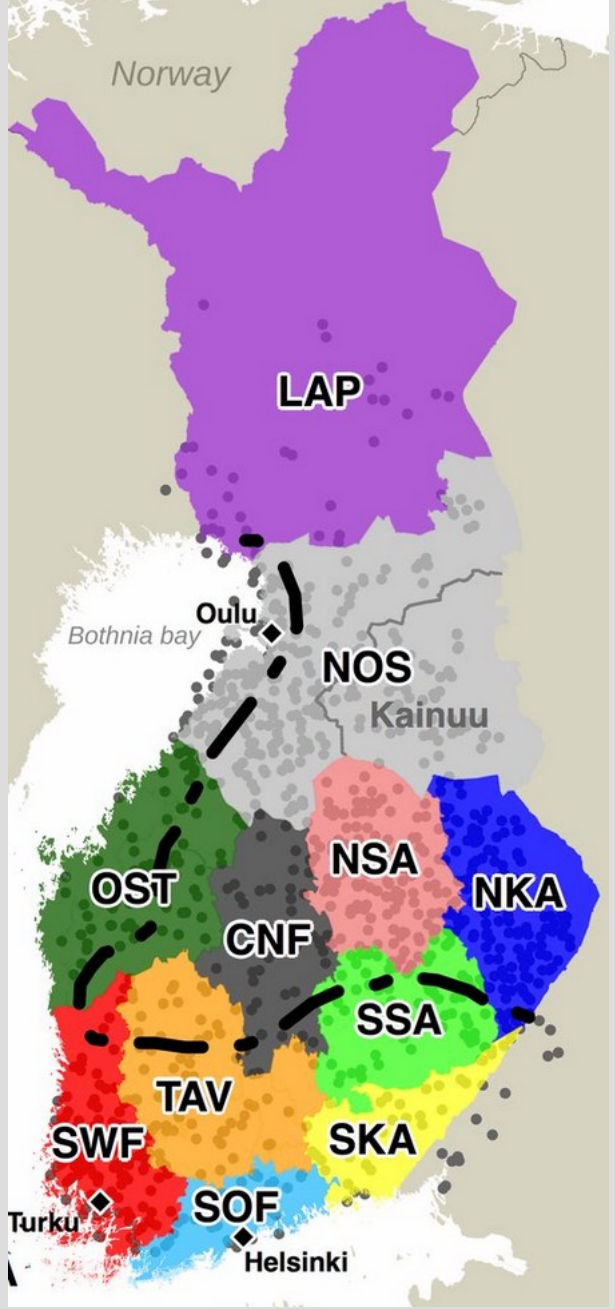
CONFOUNDING BY ANCESTRY

- SNP that has no effect on heart disease but has different regional frequencies
  - Variant ”A” frequency 0.23 in Helsinki region
  - Variant ”A” frequency 0.35 in Oulu region
- Consider sampling
  - 2000 cases (500 from H and 1500 from O).
    - ”A” frequency in cases is 0.32
  - 2000 controls (1500 from H and 500 from O).
    - ”A” frequency in cases is 0.26
- False association that variant ”A” increases risk for heart disease !
- Cases and controls must be matched !

USING LEADING PCS TO MATCH CASES & CONTROLS

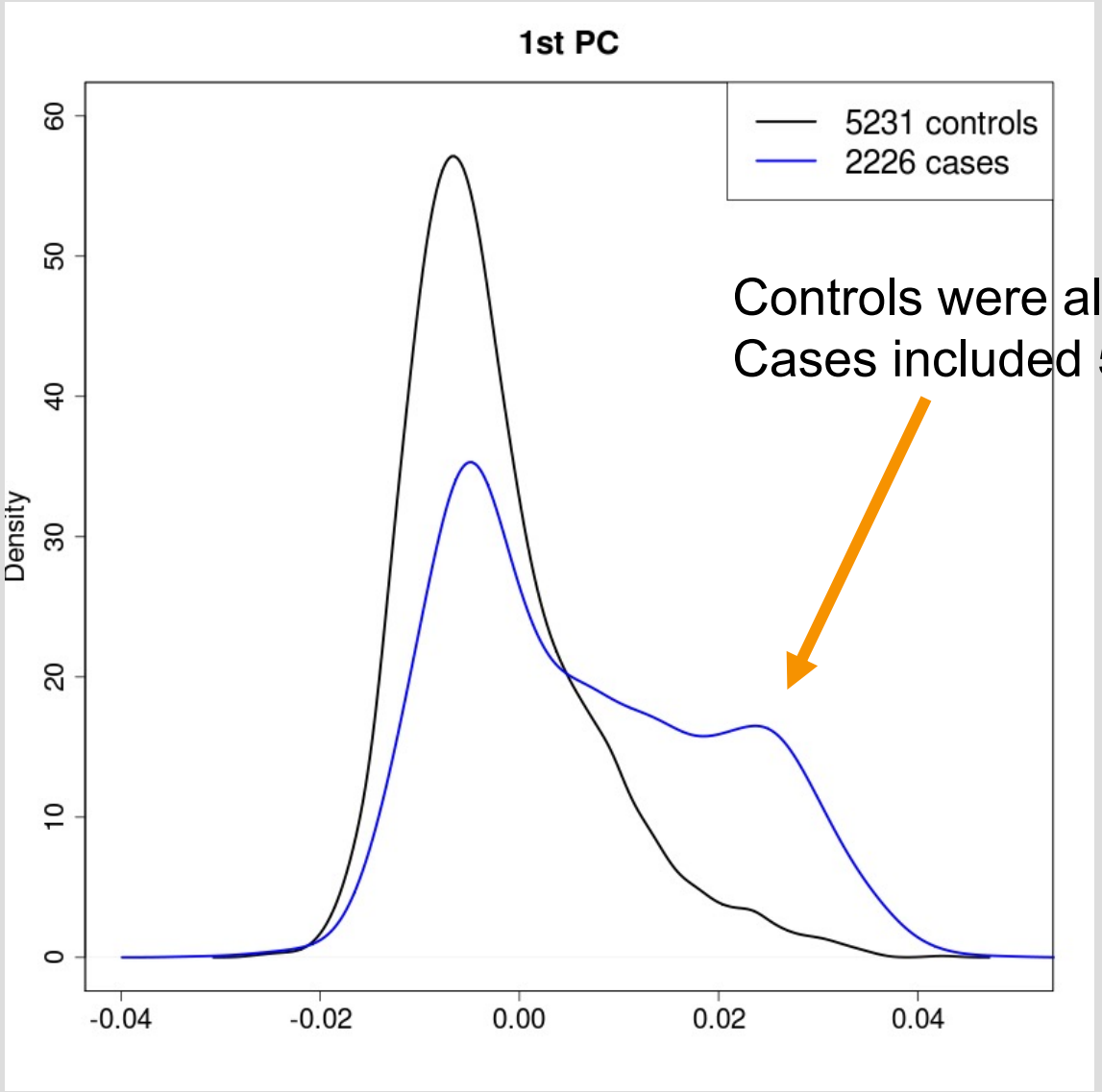
- Often we do not know regional origins of samples or they may not be informative of genetic background
- But we can infer genetic similarity and adjust the analyses for that by taking leading PCs of the genetic correlation matrix and use them as covariates (= additional predictors) in the regression model to remove confounding





Kerminen et al. 2017 G3: http://www.g3journal.org/content/7/10/3459

EXAMPLE FROM A PSORIASIS STUDY IN UK

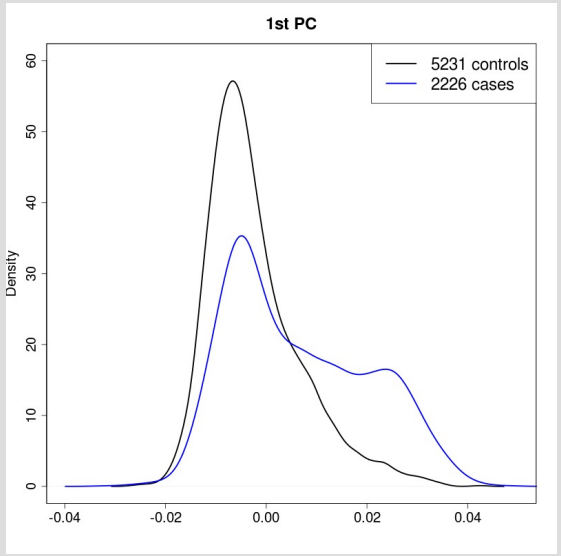


Controls were all from the UK.
Cases included 500 Irish samples.

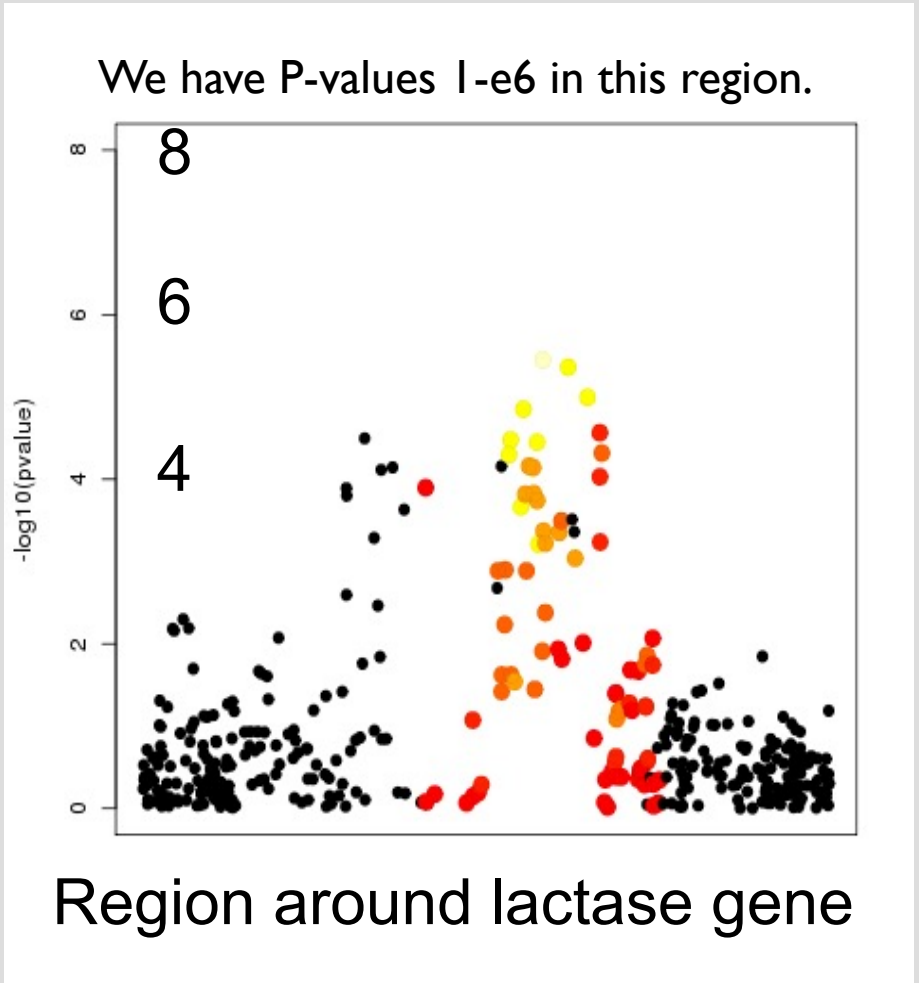
- Clear mismatch in ancestry profiles btw cases / controls!
- If we just analyze these data for association between genetic variants and psoriasis what comes up?

Strange et al. 2011 Nature Genetics

EXAMPLE FROM A PSORIASIS STUDY IN UK



Controls were all from the UK.
Cases included 500 Irish samples.



Region around lactase gene

Does lactase persistance variant really affect psoriasis susceptibility ?

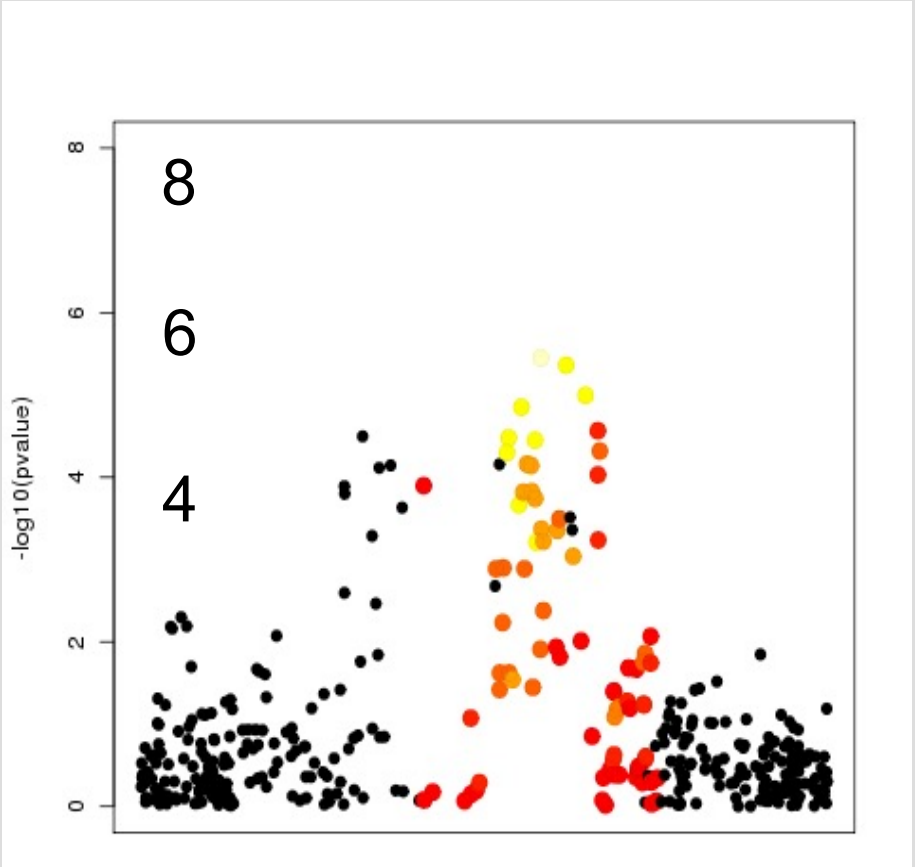
(Or is it just in different frequencies in the UK and Ireland, and we are seeing a spurious association with psoriasis in this unmatched sample?)

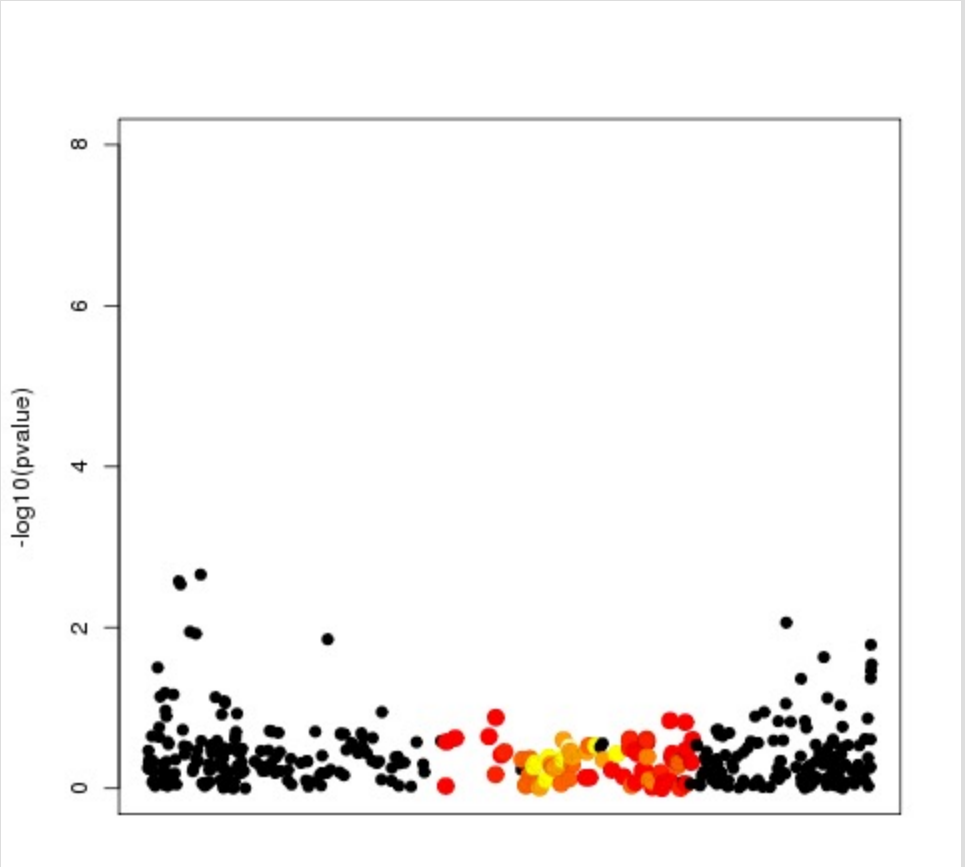
Strange et al. 2011 Nature Genetics

EXAMPLE FROM A PSORIASIS STUDY IN UK

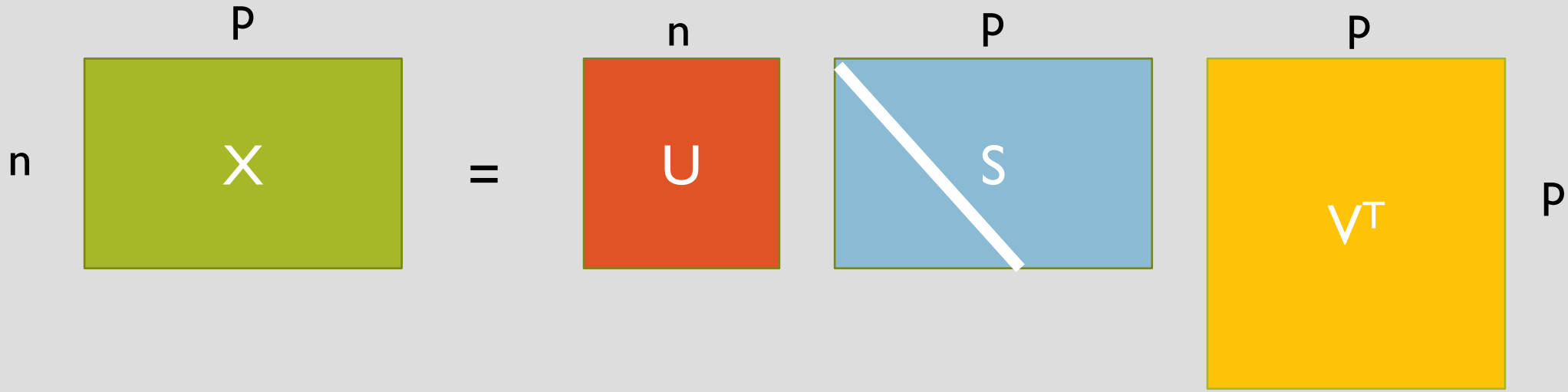
Does lactase gene really affect psoriasis susceptibility?

Probably not, since the signal can be completely explained by ancestry (1st PC) and goes away when PC1 is included in the logistic regression model



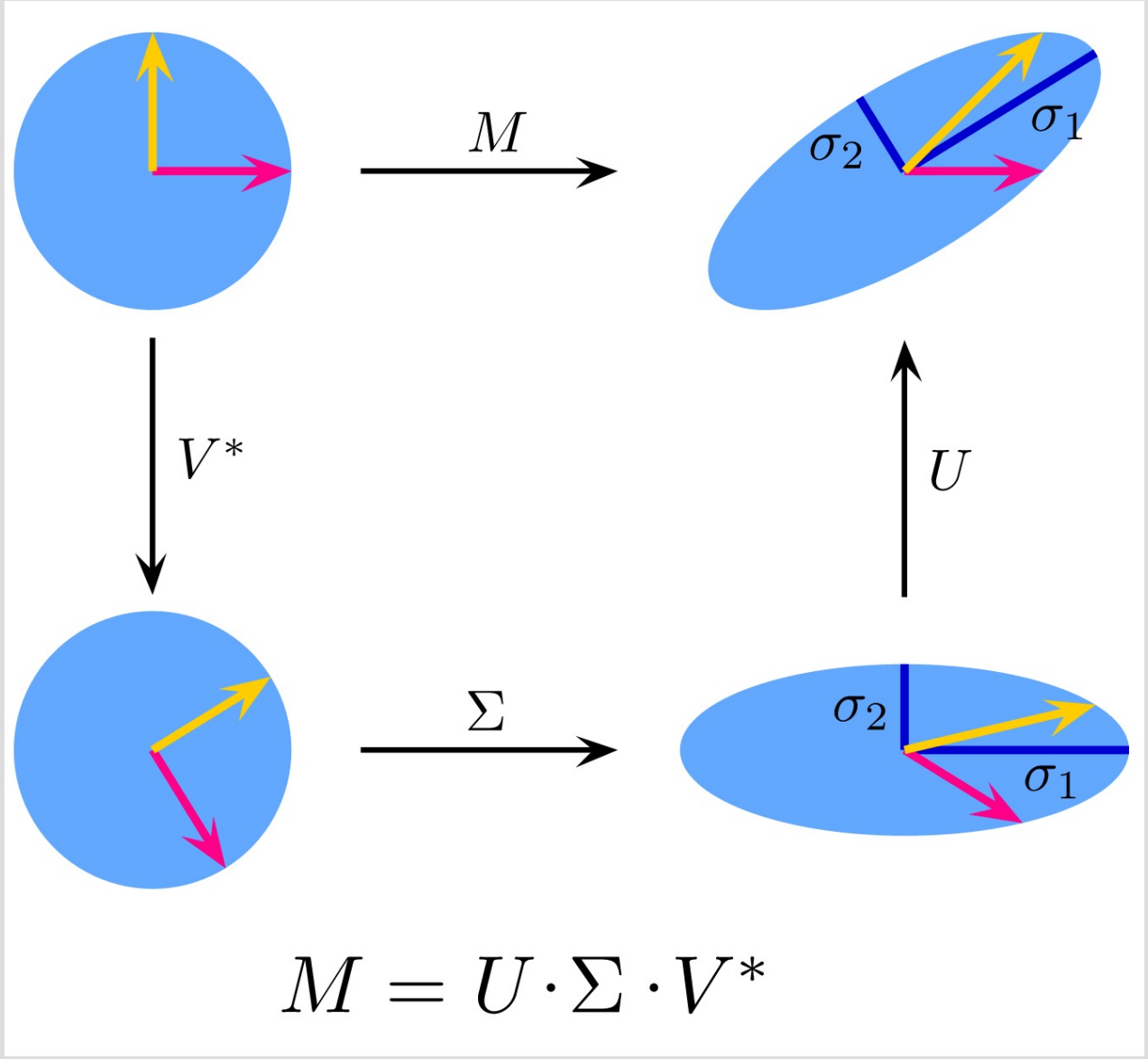


SINGULAR VALUE DECOMPOSITION (SVD)



- U has orthonormal basis of R^n
- V has orthonormal basis of R^p
- S ("sigma") is rectangular diagonal matrix with non-negative elements on the diagonal

Drawn here with n < p
(Could also be n >= p)



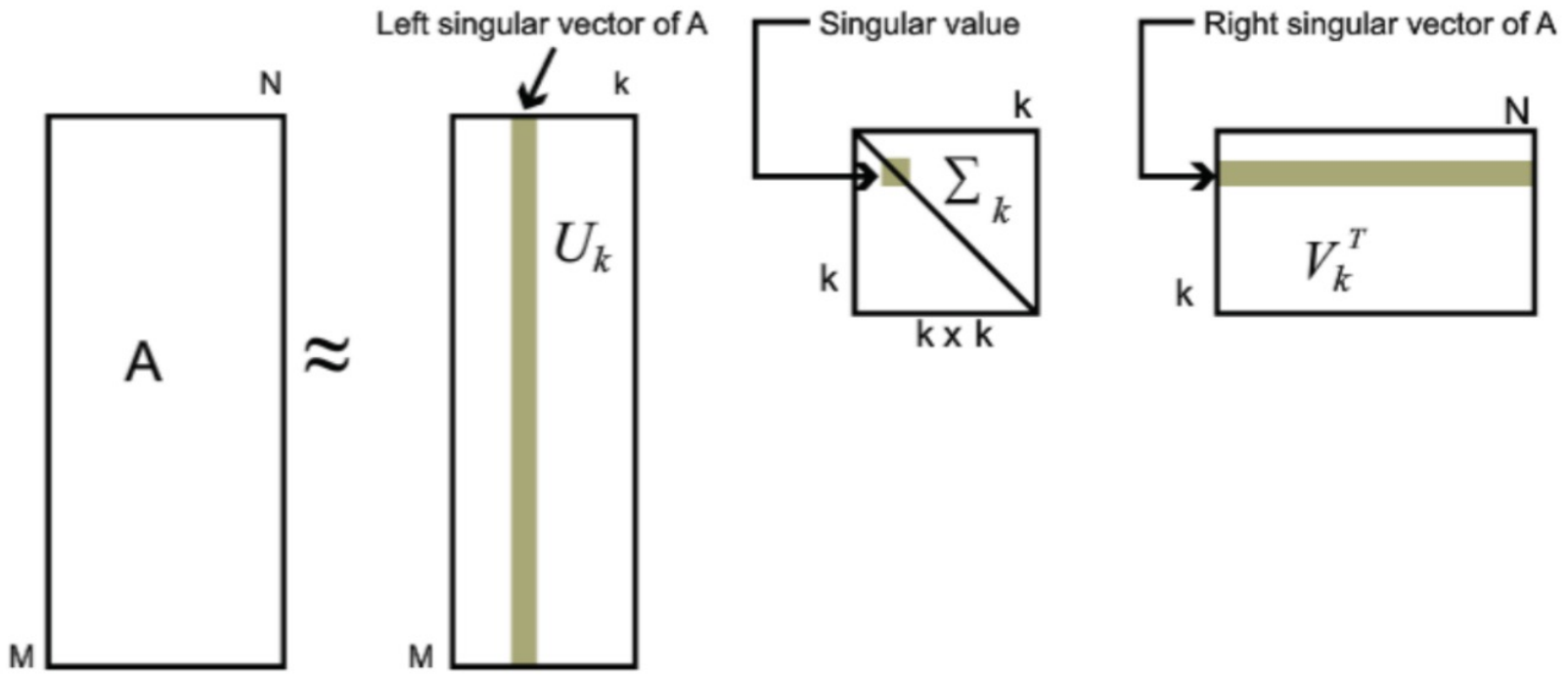
MATRIX AS A LINEAR MAPPING

- The image of a matrix is a composition of
  - Rotation in input space (V*)
  - Scaling along the new coordinates in input space and projection to output space (Sigma)
  - Rotation in output space (U)
- SVD is applicable to **any** matrix

Fig. Georg-Johann wikipedia

APPROXIMATION FROM SVD

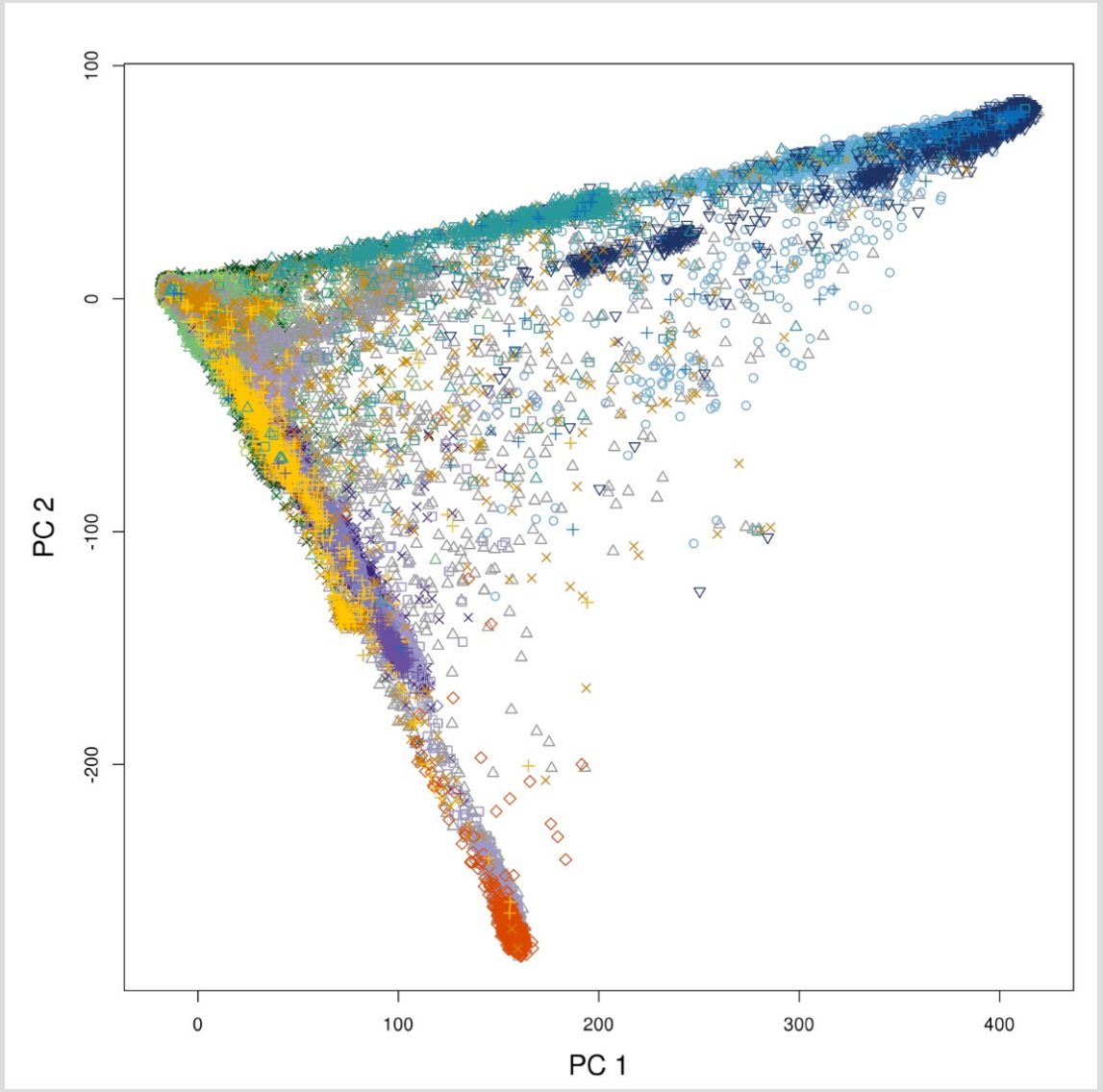
$$A \approx U_k \, \Sigma_k \, V_k^T$$

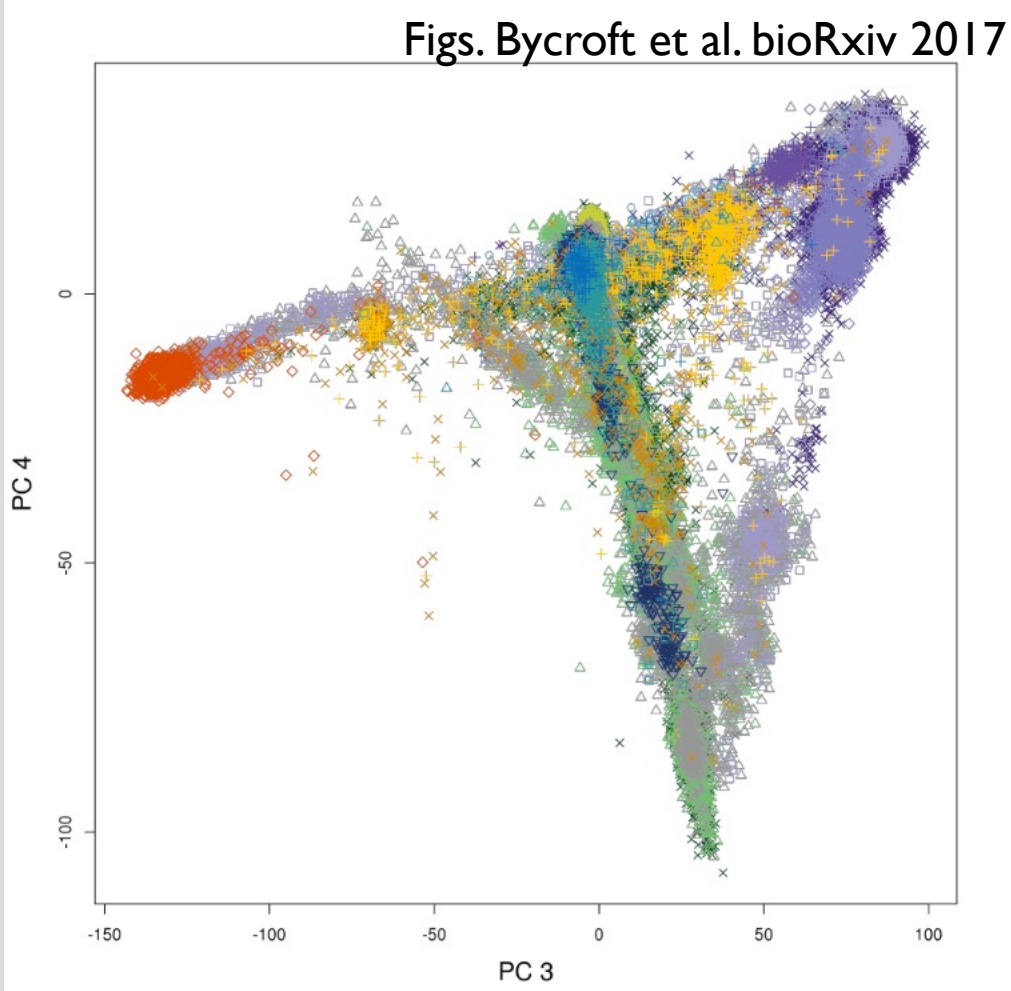


- A is put tohether as a sum of rank 1 matrices of the type $s_k \, \mathbf{u}_k \, \mathbf{v}_k^T$
- By truncating the sum after K steps we get the best (in L^2 sense) rank K approximation to A

Fig. https://rpubs.com/dsmilo/DATA643-Project3

Figs. Bycroft et al. bioRxiv 2017





Self-reported ethnic background

- British
- Irish
- Any other white background
- Indian
- Pakistani
- Bangladeshi
- Any other Asian background
- Chinese
- African
- Caribbean
- Any other Black background
- White and Asian
- White and Black African
- White and Black Caribbean
- Any other mixed background
- Other/Unknown

- UK Biobank genetic data with n=407,000 and p=150,000,; K=40 PCs
- Done with FastPCA (Galinsky et al 2015) that implements *blanczos* algorithm, a stable version of power iteration