Week 11: Single-cell genomics

- Introduction
- Dimensionality reduction
- Supervised machine learning

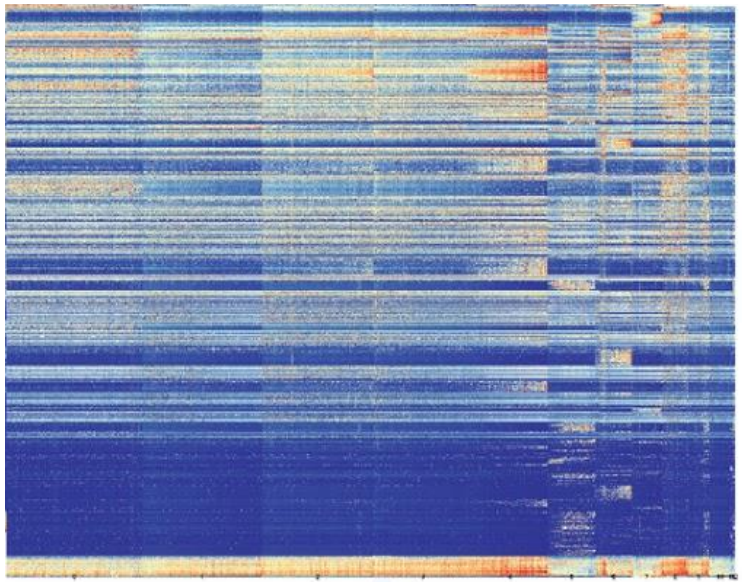
Dimensionality reduction of scRNA-seq data for visualization

Count matrix

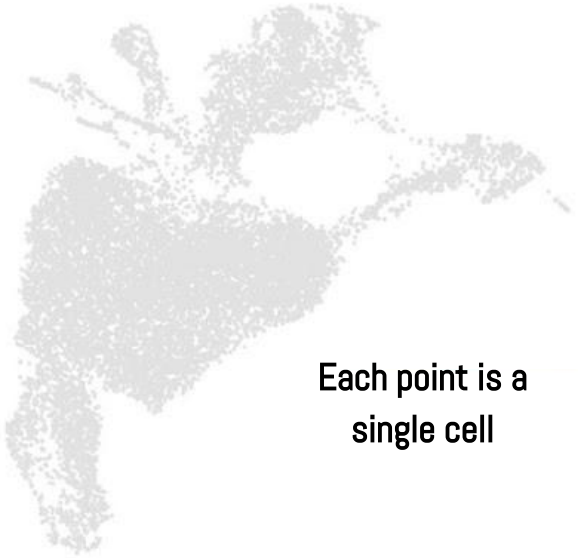
Visualization

cells

genes



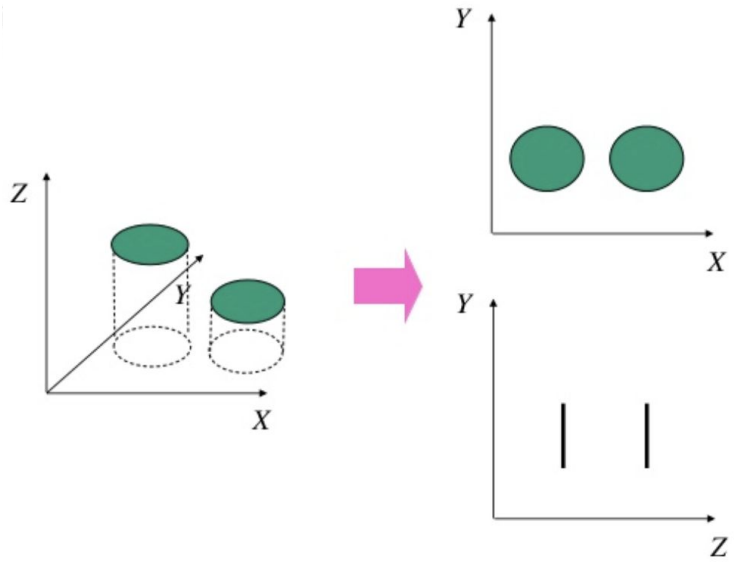
?



Each point is a single cell

Luecken & Theis (2019) Mol. Sys. Biol.

Dimensionality reduction – Projecting data into low-dim space

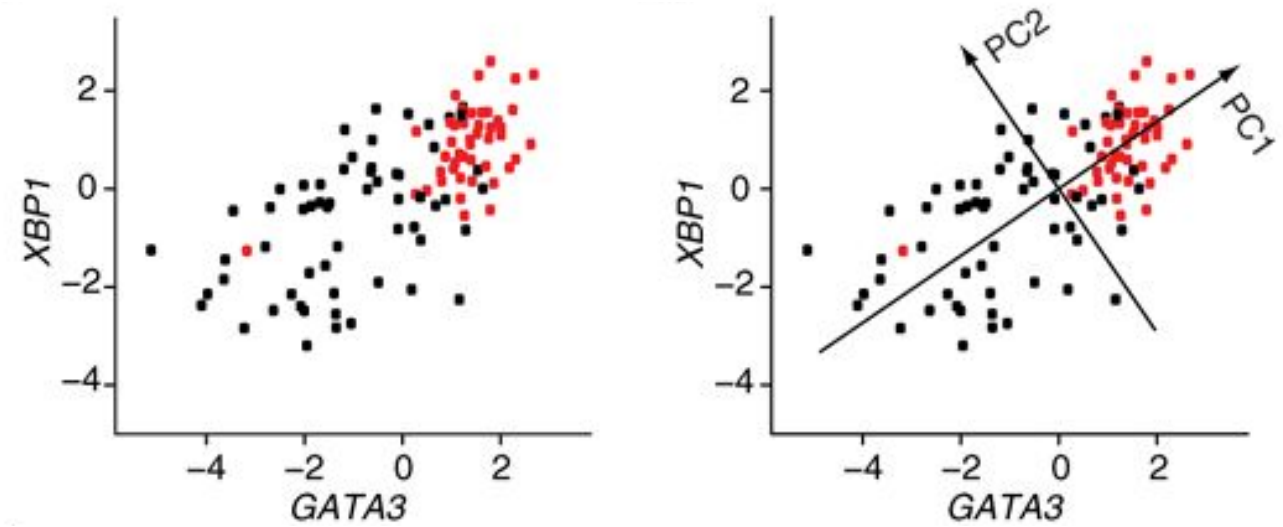


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction using Principal Components Analysis

PCA geometrically projects data onto a lower-dimensional space

- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.

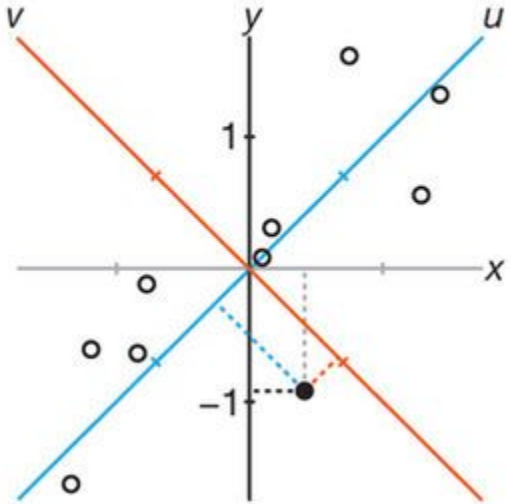


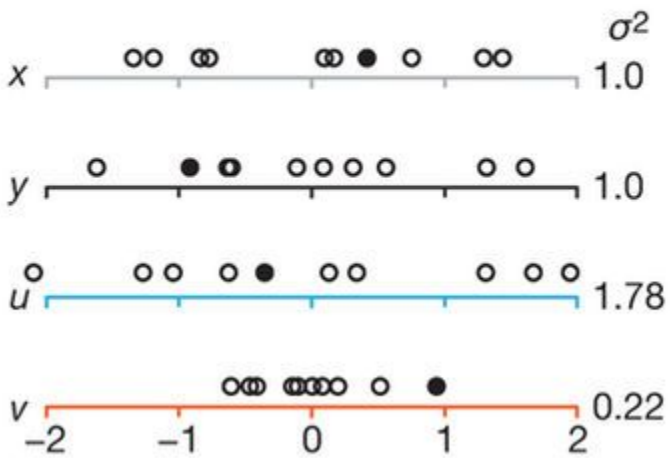
Ringnér (2008) Nat. Biotech

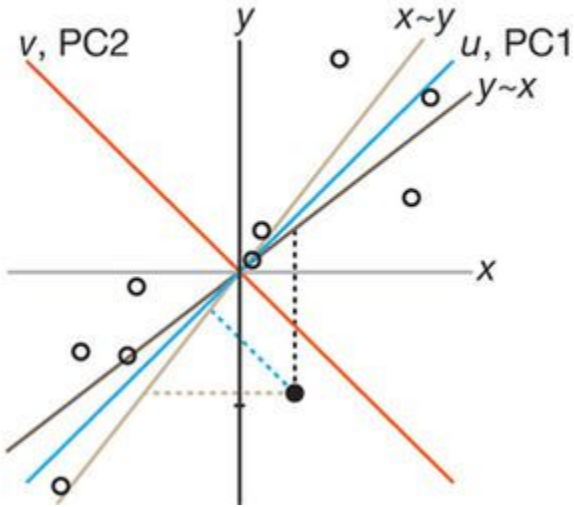
Dimensionality reduction using Principal Components Analysis

PCA geometrically projects data onto a lower-dimensional space

- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.







Lever, Krzywinski, & Altman (2017) Nat. Meth.

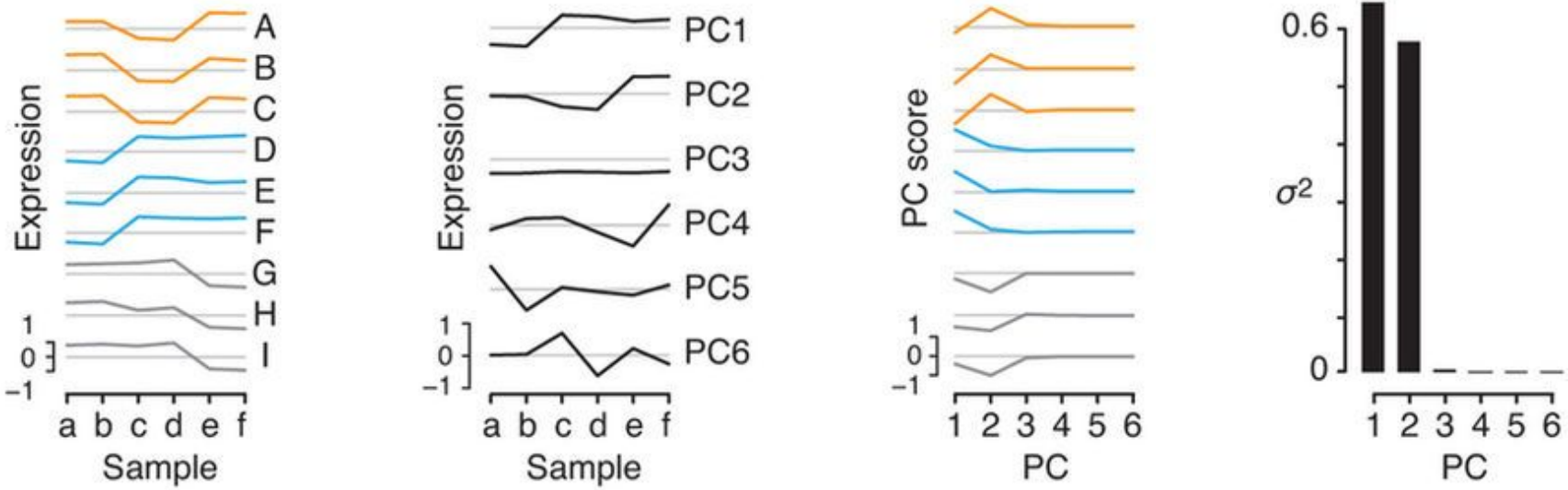
Dimensionality reduction – Principal Components Analysis

Given a dataset consisting of a set of observations representing points in a high-dimensional space, PCA finds the directions along which the observations line up best.

The idea is to transform the original data matrix X by rotation and scaling into a new set of axes so that:

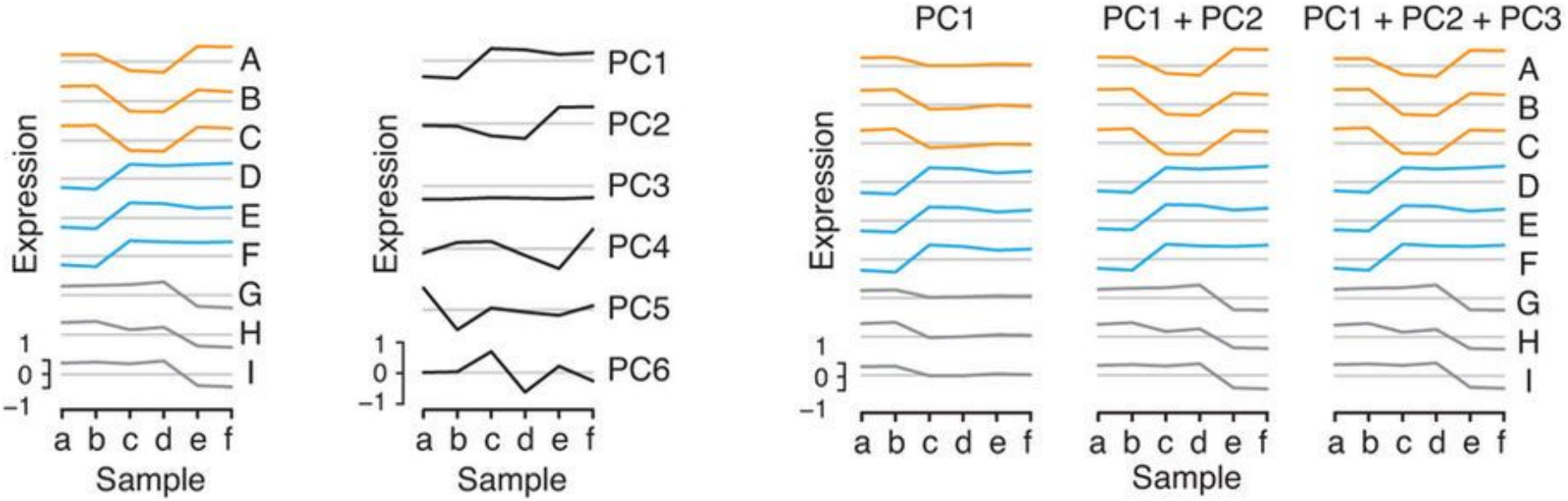
- Each new axes, termed a principal component (PC) is a linear combination of the original dimensions.
- The axis corresponding to the 1st PC satisfies the following:
  - The 1st PC is the axis along which the points are most “spread out”.
    - The axis along which the variance of the data is maximized.
  - The points can best be viewed as lying along the 1st PC, with smallest deviations from this axis.
- The axis corresponding to the 2nd PC is the axis along which the variance of distances from the first axis is greatest.
- And so on.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

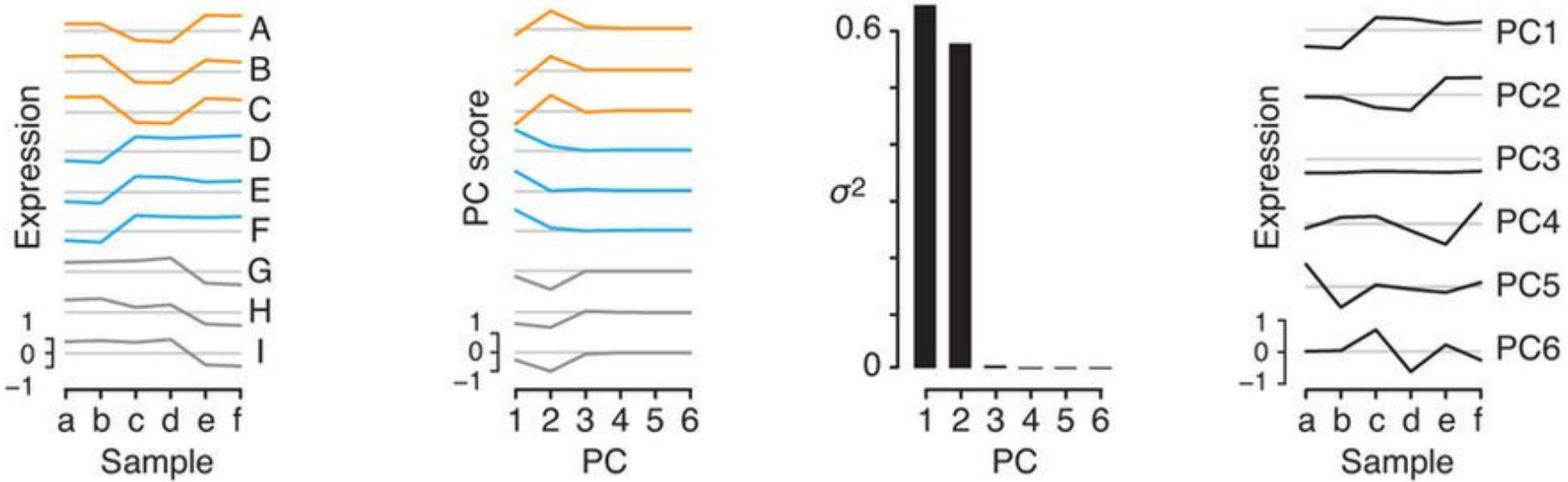
Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

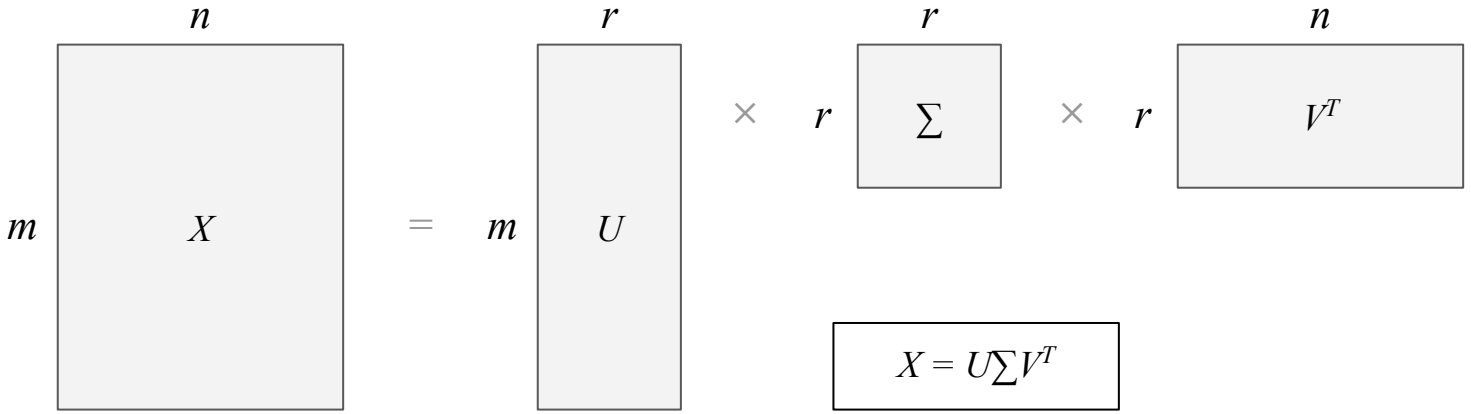
Singular Value Decomposition



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA using SVD

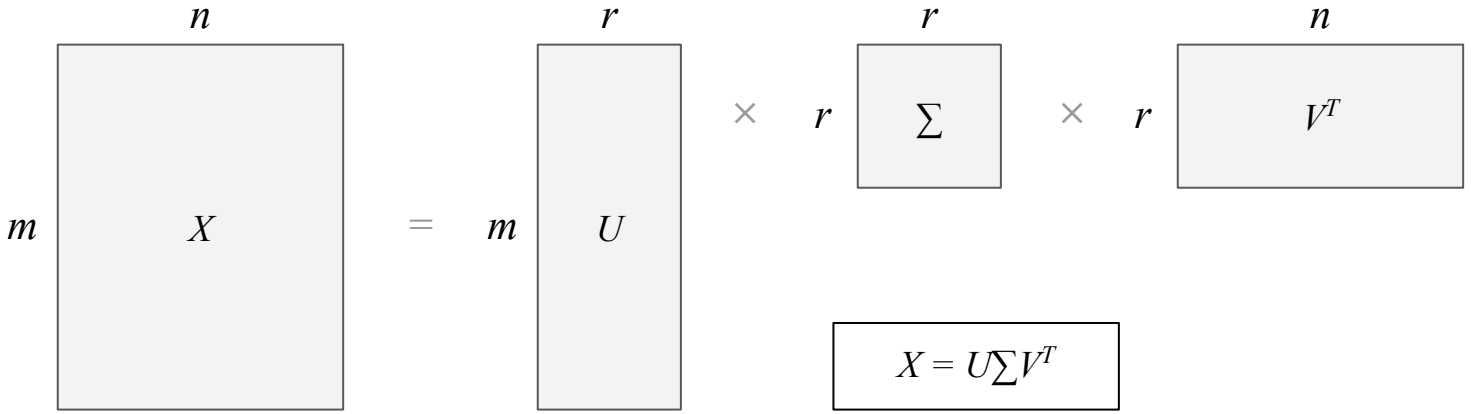
SVD theorem



- U is an $m \times r$ column-orthonormal matrix:
  - Each of its columns is a unit vector and the dot product of any two columns is 0.
- V is an $n \times r$ column-orthonormal matrix.
  - We always use V^T, so the rows of V^T are orthonormal.
- $\sum$ is a diagonal matrix with elements σ_i. (All elements not on the main diagonal are 0.)
  - The elements of $\sum$ are called the singular values of X. such that $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_i \ldots \geq \sigma_r$.
- $X = \sum_i \sigma_i u_i v_i$

Dimensionality reduction by PCA using SVD

SVD theorem



- U is an $m \times r$ column-orthonormal matrix:
  - Each of its columns is a unit vector and the dot product of any two columns is 0.
- V is an $n \times r$ column-orthonormal matrix.
  - We always use V^T, so the rows of V^T are orthonormal.
- $\sum$ is a diagonal matrix with elements σ_i. (All elements not on the main diagonal are 0.)
  - The elements of $\sum$ are called the singular values of X. such that $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_i \ldots \geq \sigma_r$.
- $X = \sum_i \sigma_i u_i v_i$

Dimensionality reduction by PCA using SVD

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1	1	1	1	0	0
Gene 2	3	3	3	0	0
Gene 3	4	4	4	0	0
Gene 4	5	5	5	0	0
Gene 5	0	2	0	4	4
Gene 6	0	0	0	5	5
Gene 7	0	1	0	2	2

	Matrix	Aliens	Star Wars	Monster Inc.	Toy Story
Customer 1	1	1	1	0	0
Customer 2	3	3	3	0	0
Customer 3	4	4	4	0	0
Customer 4	5	5	5	0	0
Customer 5	0	2	0	4	4
Customer 6	0	0	0	5	5
Customer 7	0	1	0	2	2

Dimensionality reduction by PCA using SVD

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}_{V^{\mathrm{T}}}$$

- What do the columns of U represent?
- What do the rows of V^T represent?
- What do the diagonal entries of $\sum$ represent?

Dimensionality reduction by PCA using SVD

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}_{V^{\mathrm{T}}}$$

How do we do dimensionality reduction from here?

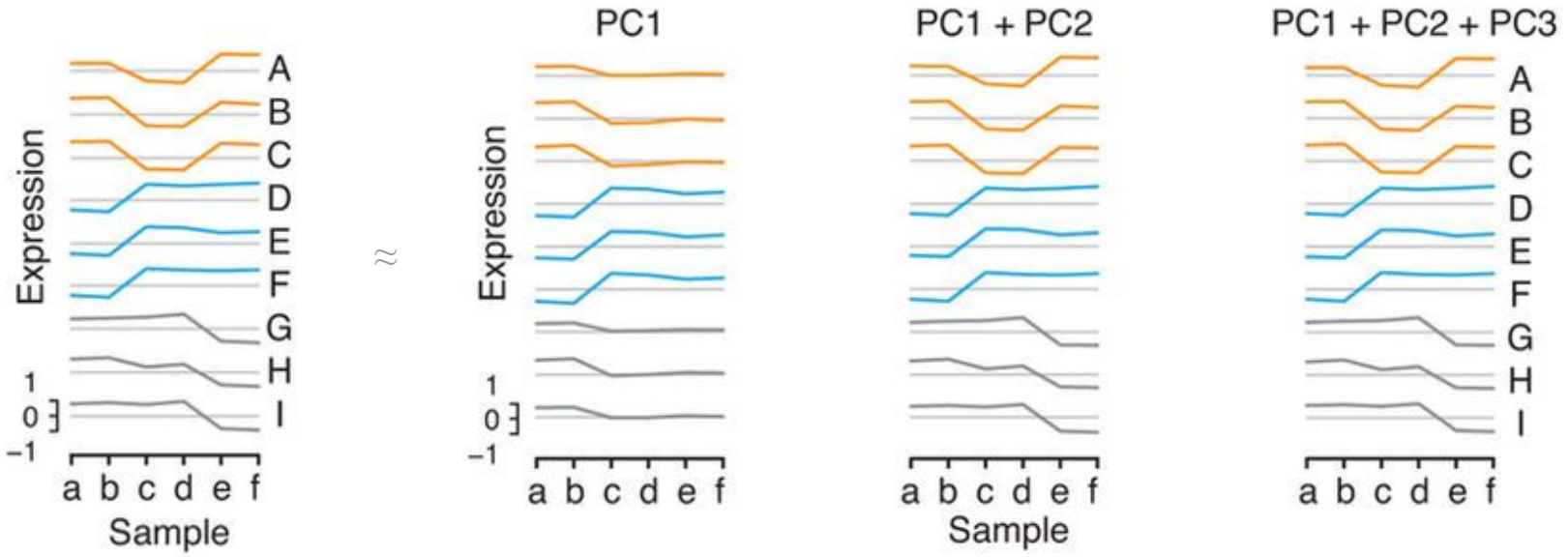
Dimensionality reduction by PCA using SVD

$$\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .07 & -.73 \\ .07 & -.29 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \end{bmatrix} = \begin{bmatrix} 0.93 & 0.95 & 0.93 & .014 & .014 \\ 2.93 & 2.99 & 2.93 & .000 & .000 \\ 3.92 & 4.01 & 3.92 & .026 & .026 \\ 4.84 & 4.96 & 4.84 & .040 & .040 \\ 0.37 & 1.21 & 0.37 & 4.04 & 4.04 \\ 0.35 & 0.65 & 0.35 & 4.87 & 4.87 \\ 0.16 & 0.57 & 0.16 & 1.98 & 1.98 \end{bmatrix}$$

$$X = \sum_i \sigma_i u_i v_i$$

$$X \approx \sum_{i \text{ in } 1:r} \sigma_i u_i v_i$$

Reconstructing the original data matrix from PCs



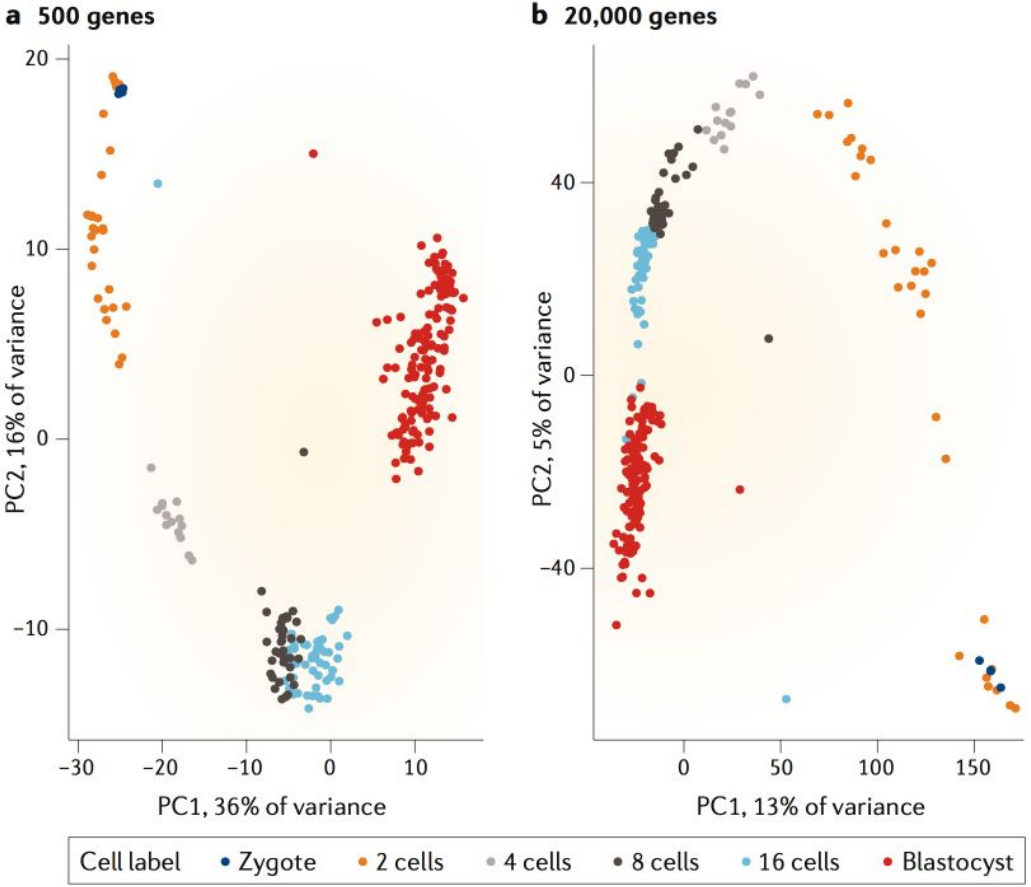
$X = \sum_i \sigma_i u_i v_i$

$X \approx \sigma_1 u_1 v_1$

$X \approx \sigma_1 u_1 v_1 + \sigma_2 u_2 v_2$

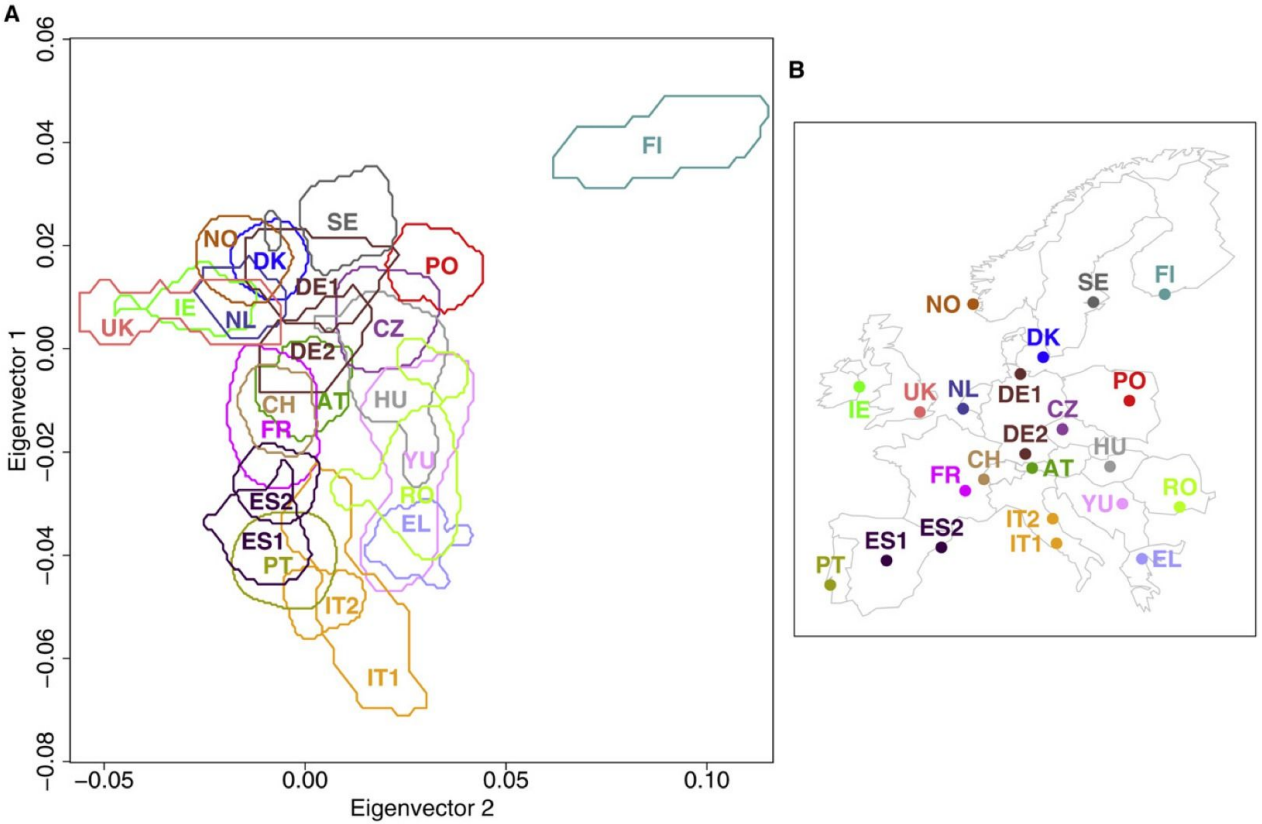
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



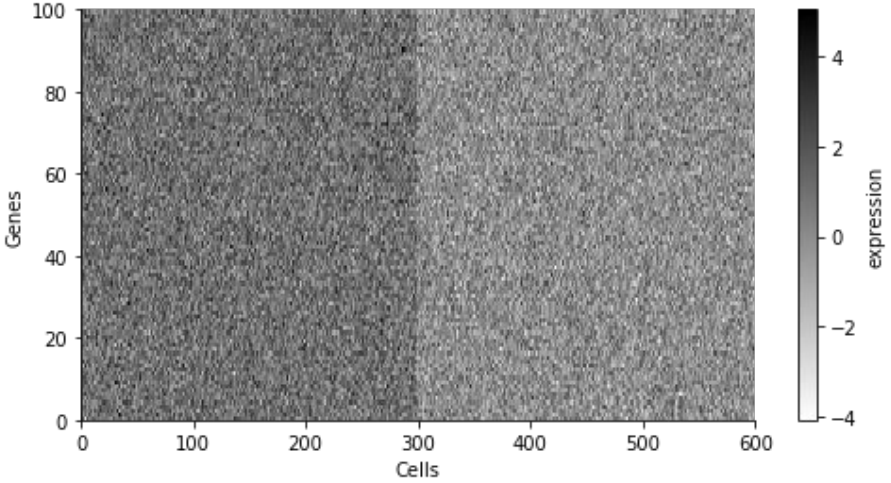
Lever, Krzywinski, & Altman (2017) Nat. Meth.

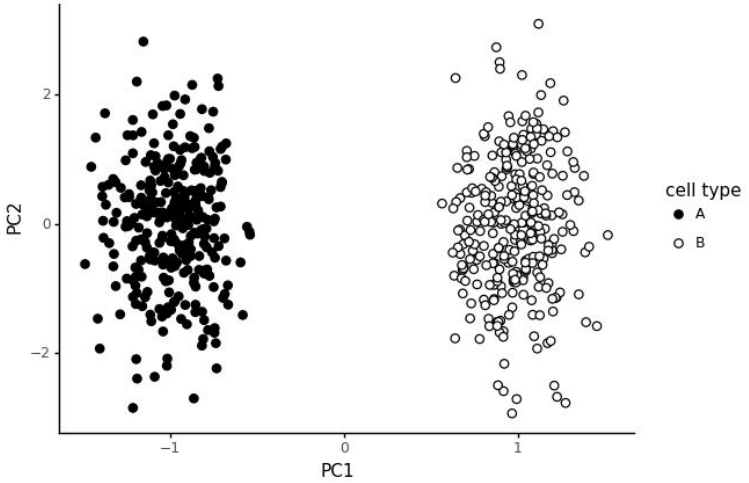
Dimensionality reduction



"Genes mirror geography within Europe." Novembre (2008) Nat.

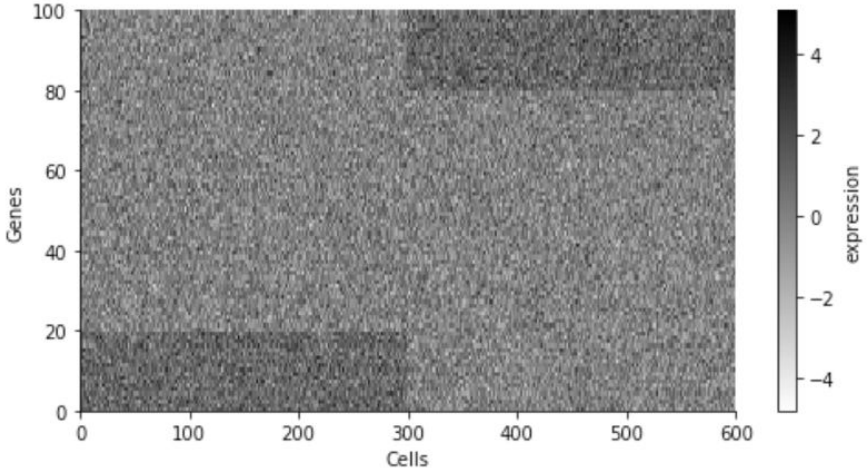
Dimensionality reduction by PCA

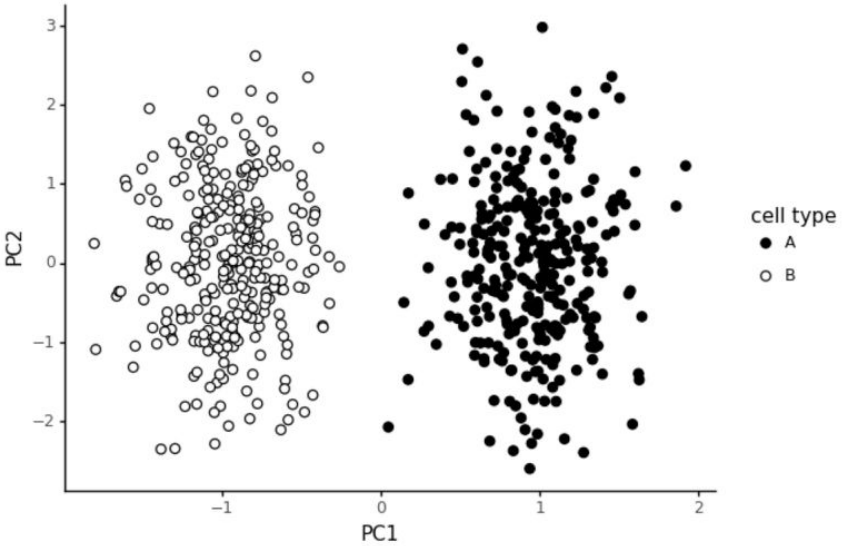




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

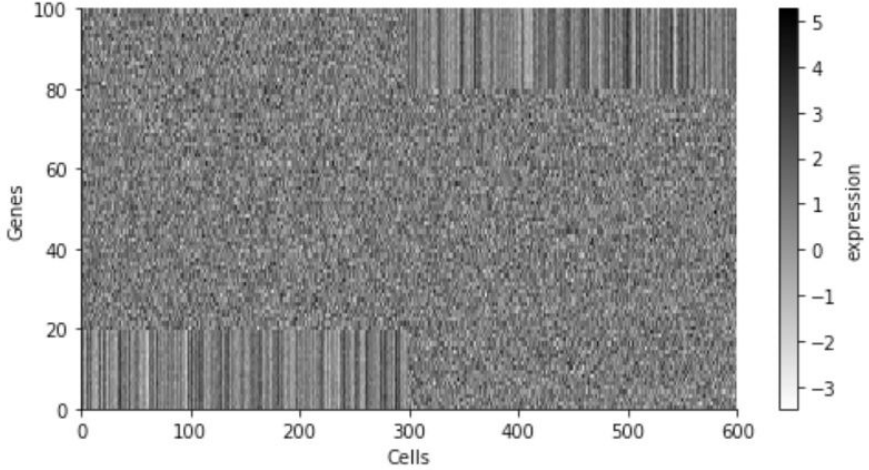
Dimensionality reduction by PCA

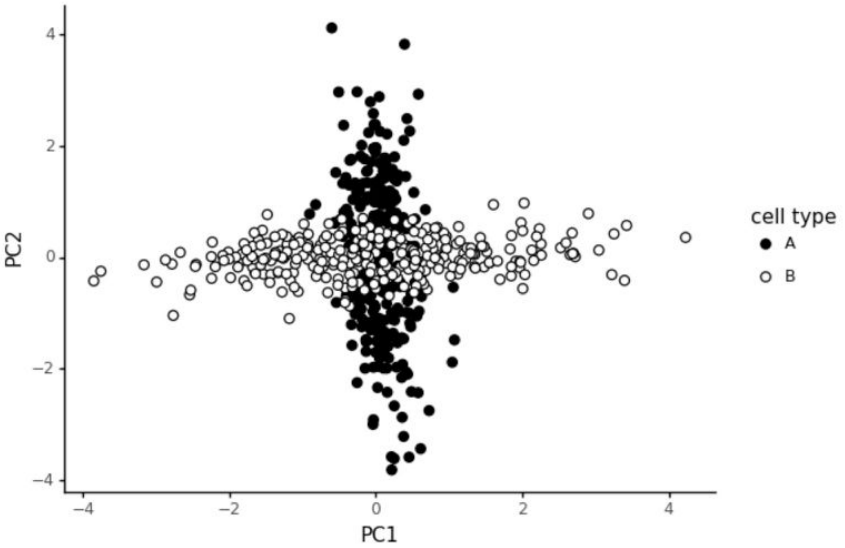




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

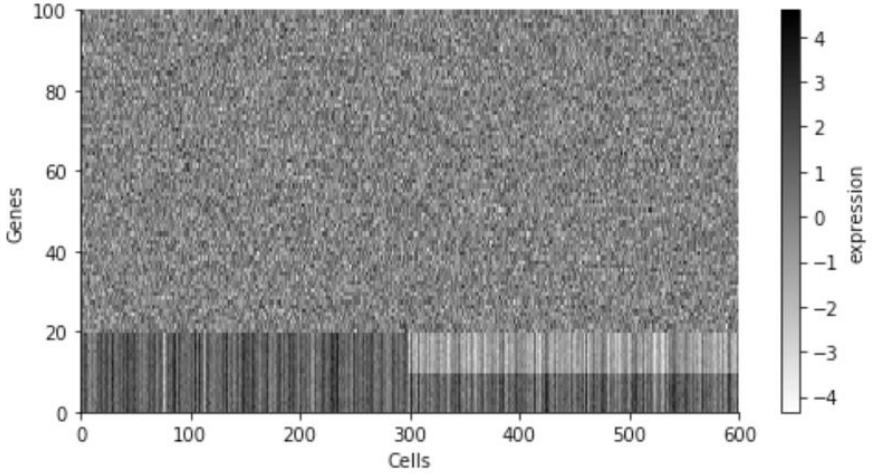
Dimensionality reduction by PCA

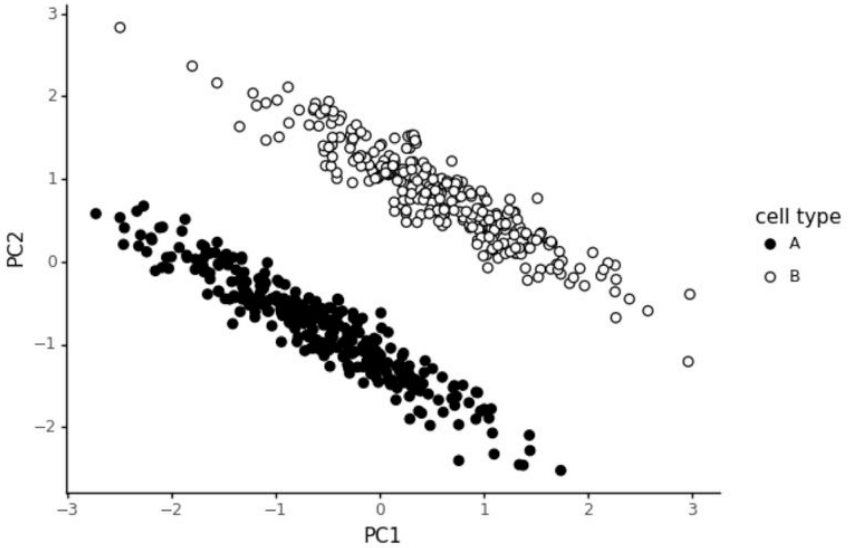




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

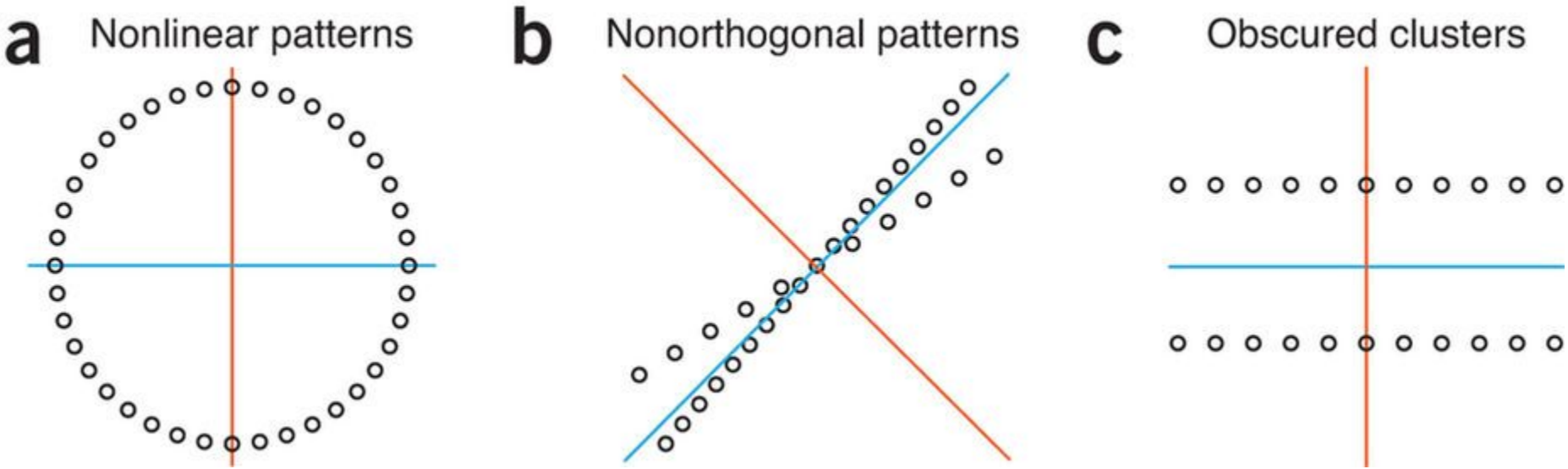
Dimensionality reduction by PCA





http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

Limitations of PCA, alternatives



Lever, Krzywinski, & Altman (2017) Nat. Meth.