

Single-cell genomics

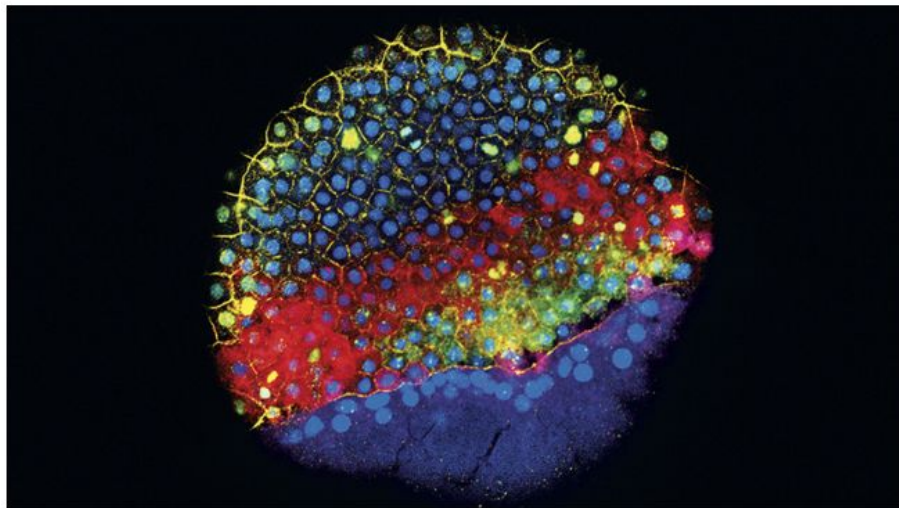
- Introduction
- Dimensionality reduction
- Supervised machine learning

Single-cell RNA-seq

BREAKTHROUGH OF THE YEAR

Development cell by cell

With a trio of techniques, scientists are tracking embryo development in stunning detail



A zebrafish embryo at an early stage of development. Fluorescent markers highlight cells expressing genes that help determine the type of cell they will become. (JEFFREY FARRELL, SCHIER LAB/HARVARD UNIVERSITY)

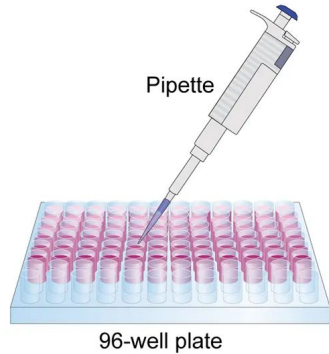
The single-cell revolution is just starting.

— Elizabeth Pennisi

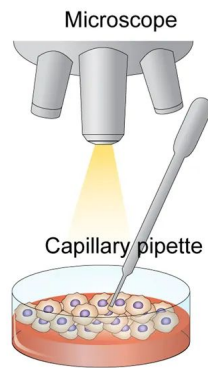
<https://vis.sciencemag.org/breakthrough2018/finalists/#cell-development>

Single-cell isolation and library preparation

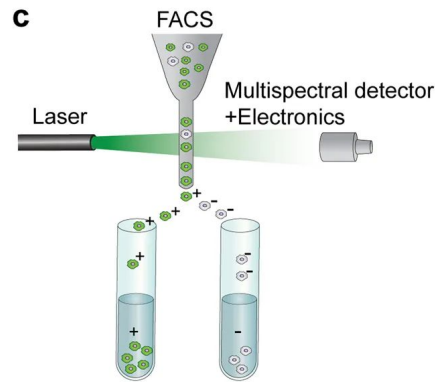
a



b



c

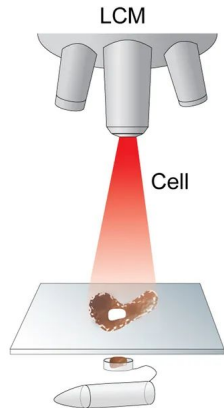


a) Limiting dilution method isolates individual cells.

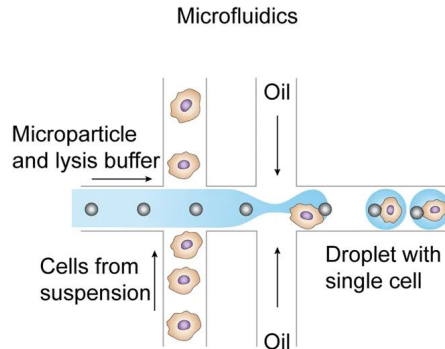
b) Collecting single cells using microscope-guided capillary pipettes.

c) FACS isolates highly purified single cells by tagging cells w/ markers.

d



e



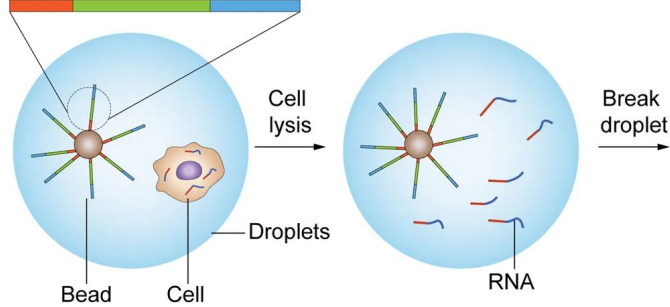
d) Laser capture microdissection (LCM) isolates cells from solid samples.

e) Microfluidics for isolating single-cells (nanoliter-sized vol; e.g. Drop-Seq).

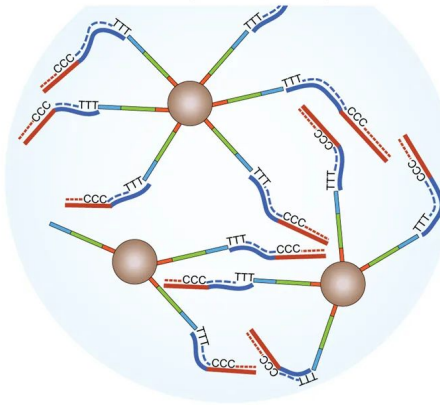
Single-cell isolation and library preparation

Structure of the barcode primer bead

PCR
handle Cell barcode UMI



Reverse transcription with template switching



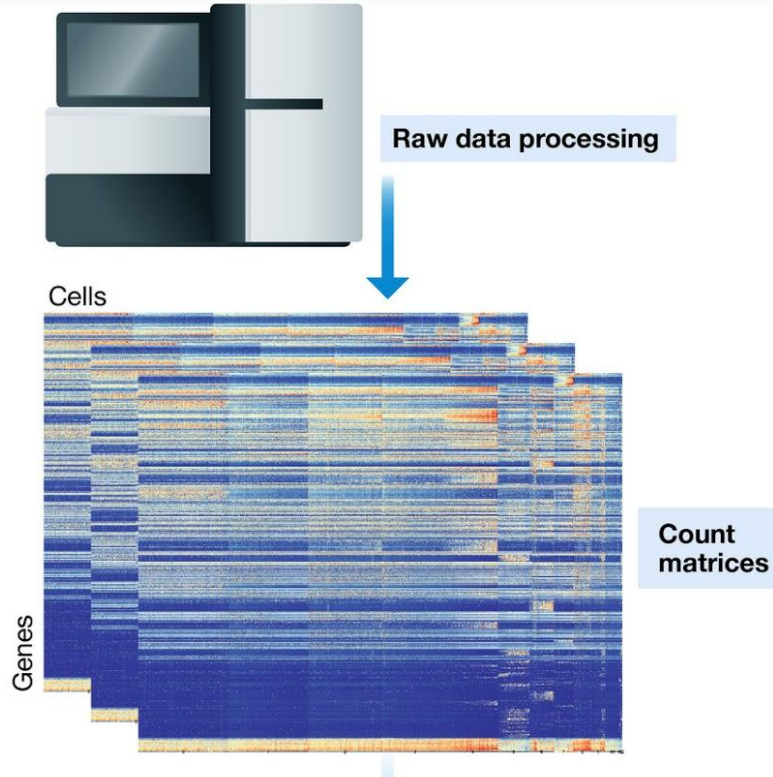
A schematic example of droplet-based library generation.

Libraries for scRNA-seq are typically generated via:

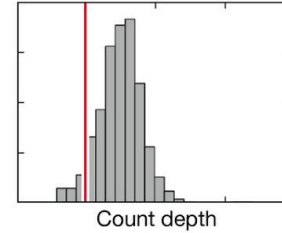
- Cell lysis
- Reverse transcription into first-strand cDNA using uniquely barcoded beads
- Second-strand synthesis, &
- cDNA amplification.

Pre-processing, QC, & normalization of scRNA-seq data

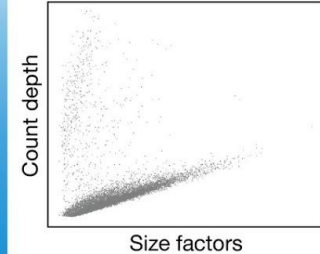
PRE-PROCESSING



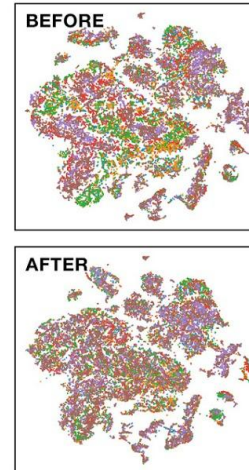
Quality control



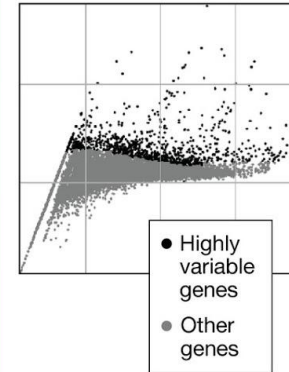
Normalization



Data correction (e.g. batch)

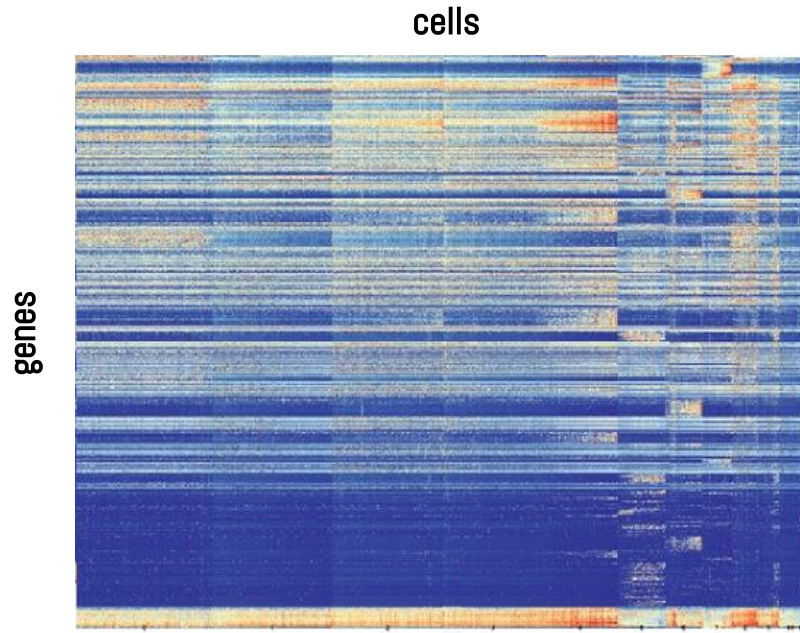


Feature selection

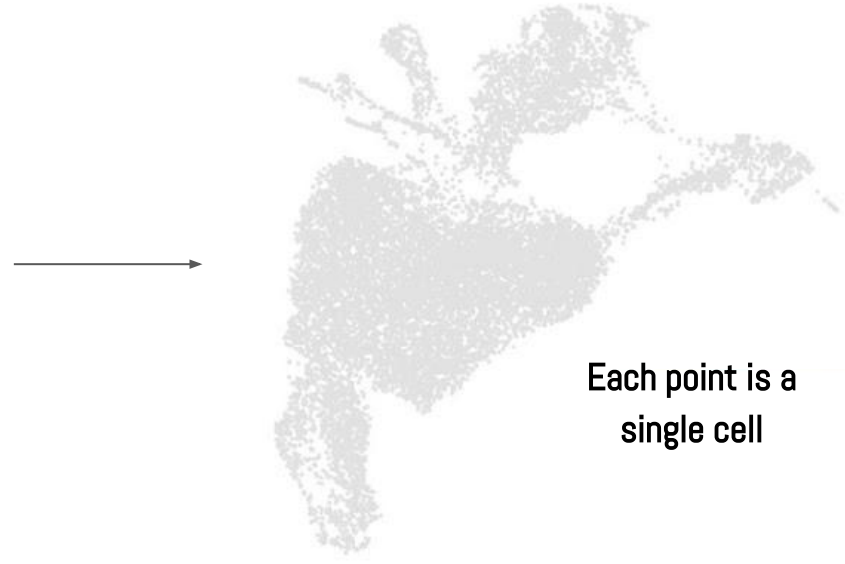


Dimensionality reduction of scRNA-seq data

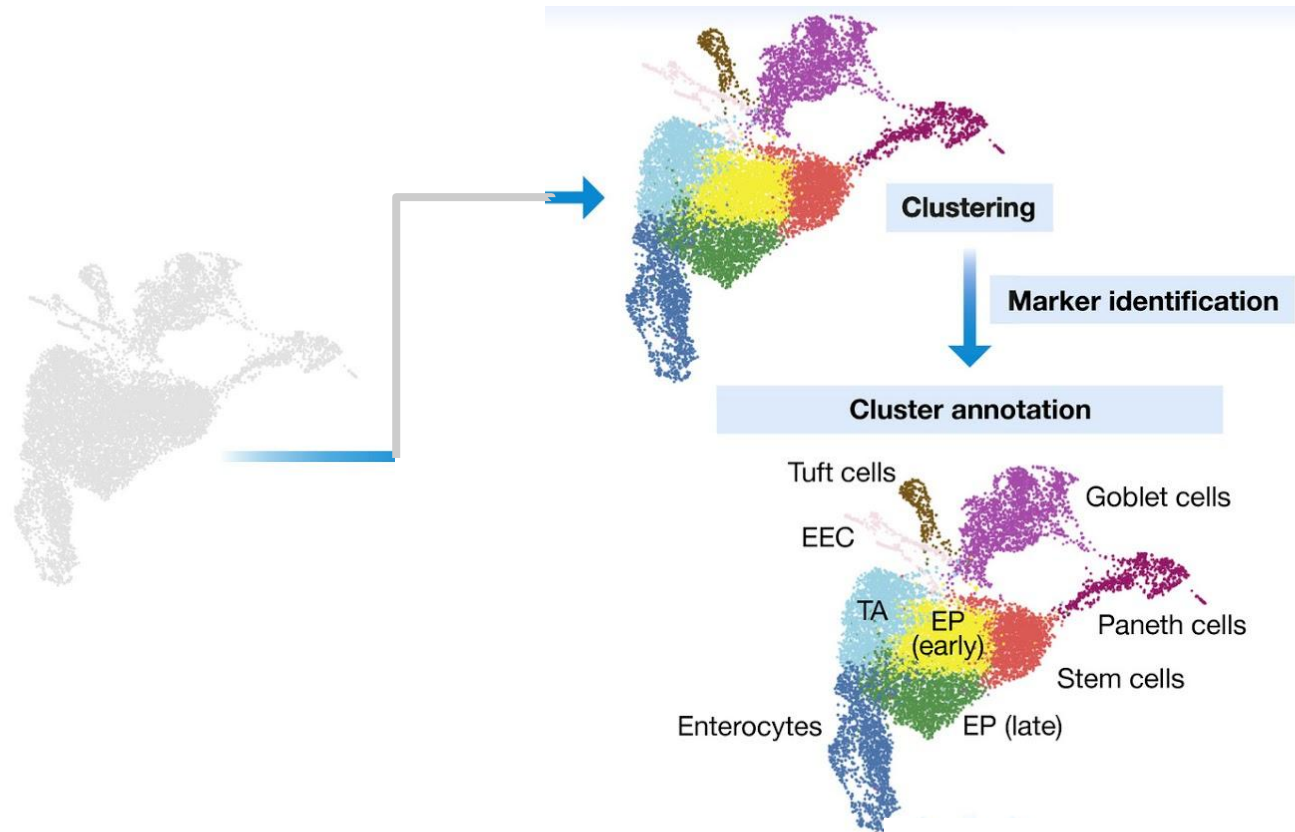
Count matrix



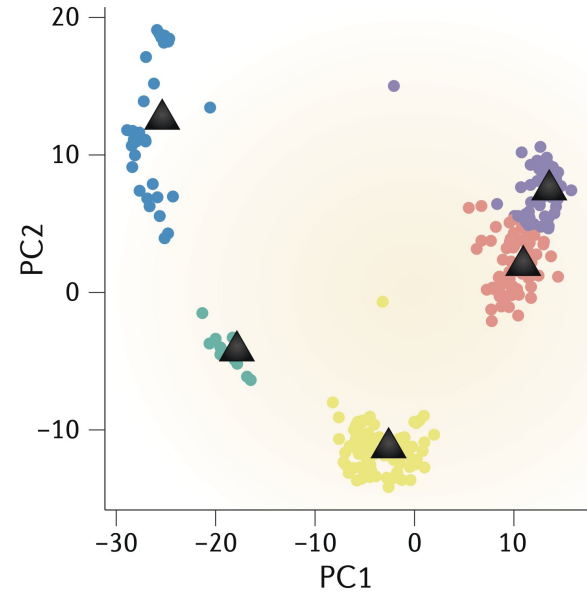
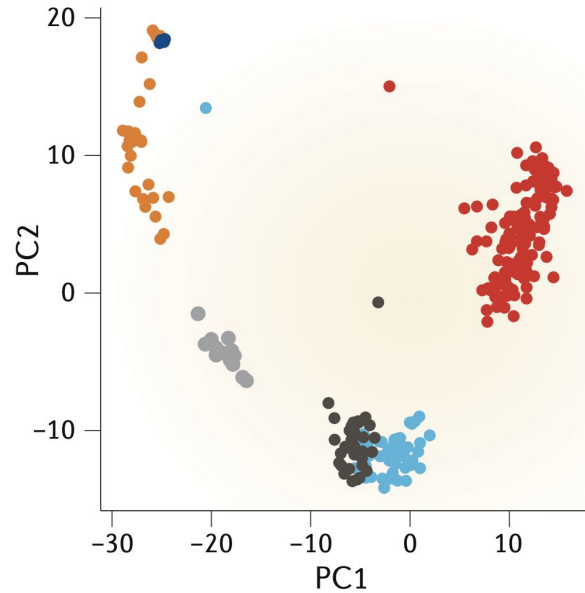
Data in reduced dimensions



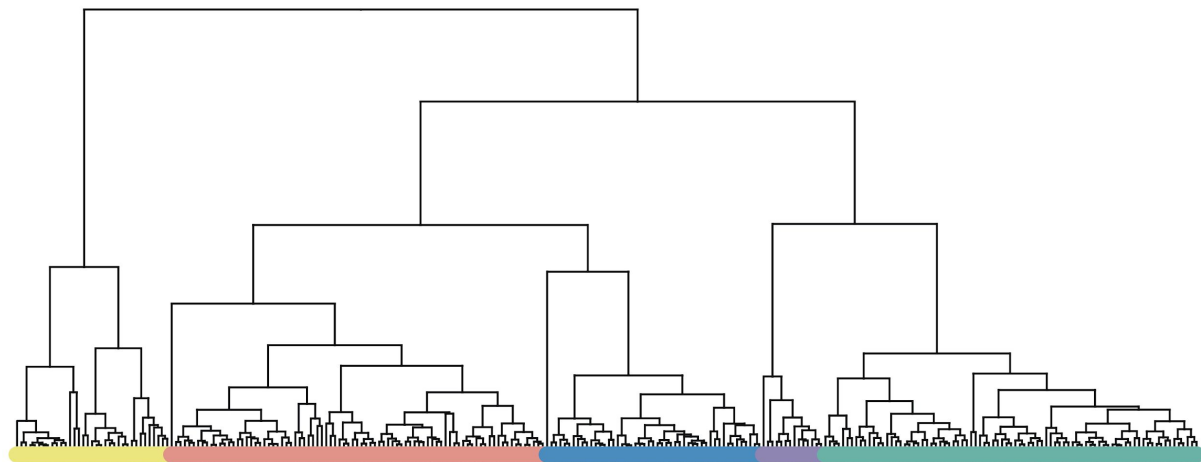
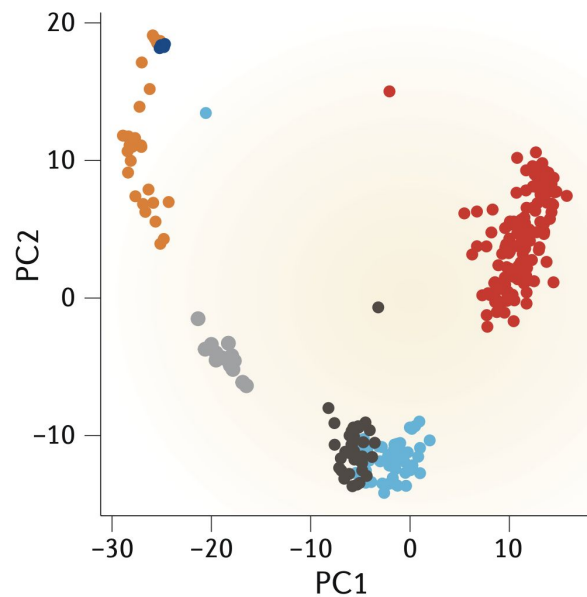
Clustering scRNA-seq data to identify cell types/states



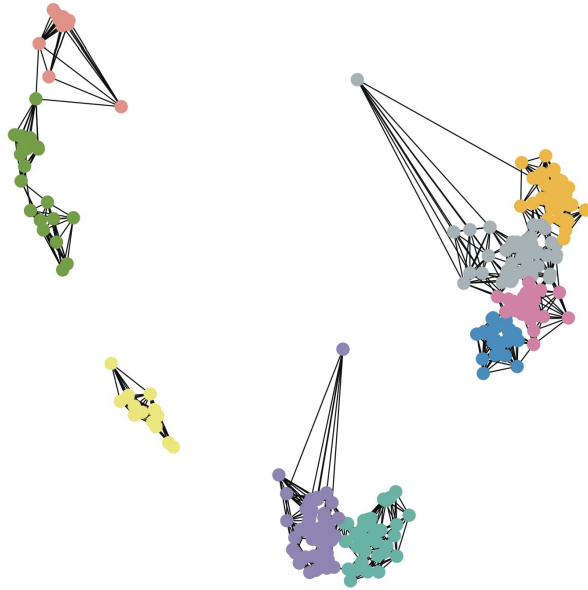
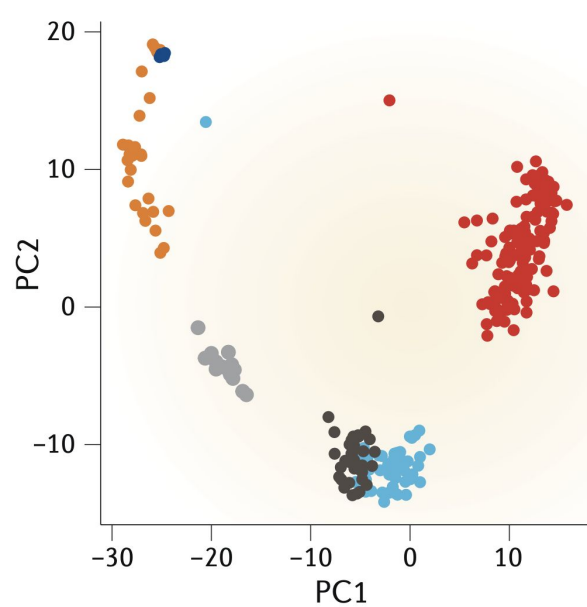
Clustering scRNA-seq data to identify cell types/states



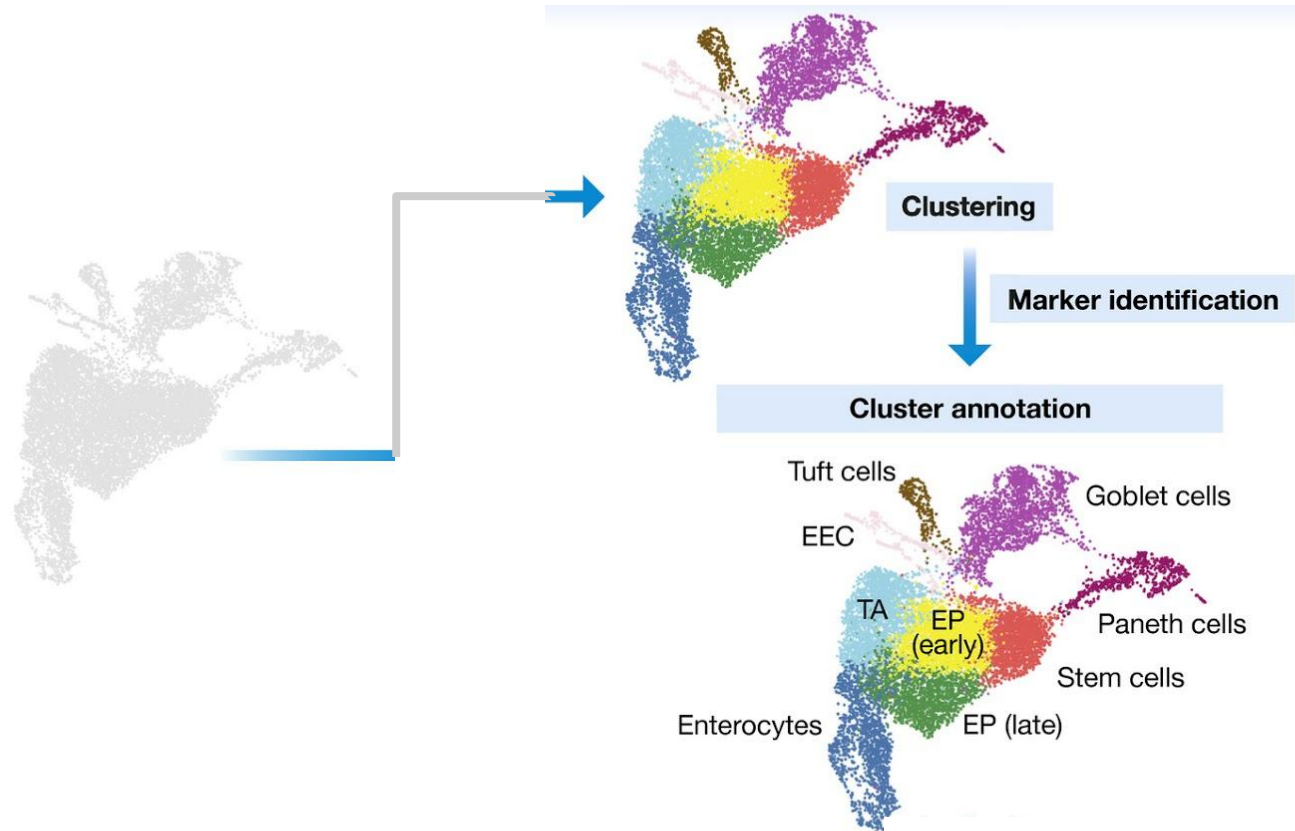
Clustering scRNA-seq data to identify cell types/states



Clustering scRNA-seq data to identify cell types/states

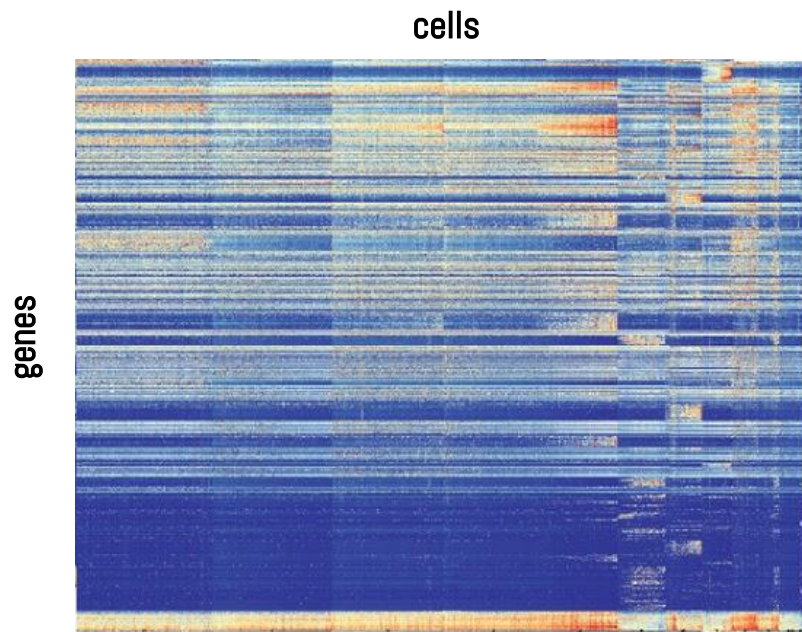


Clustering scRNA-seq data to identify cell types/states

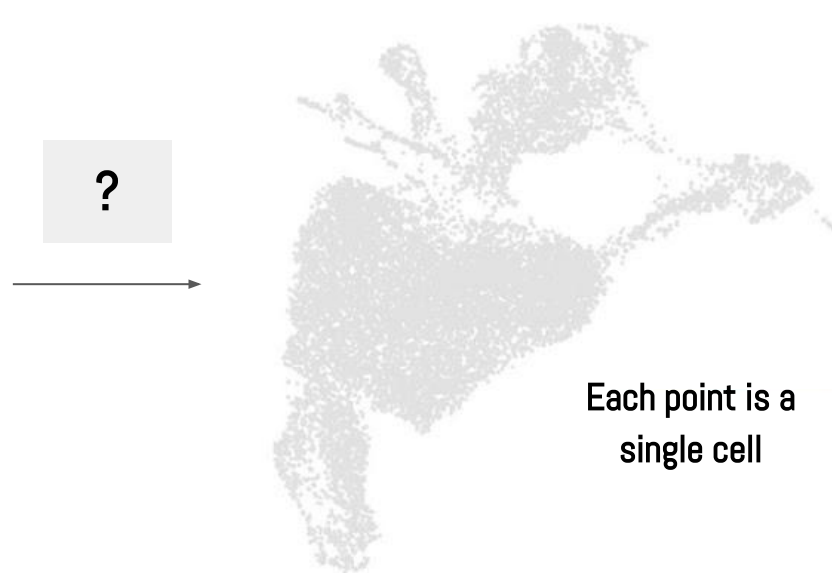


Dimensionality reduction of scRNA-seq data

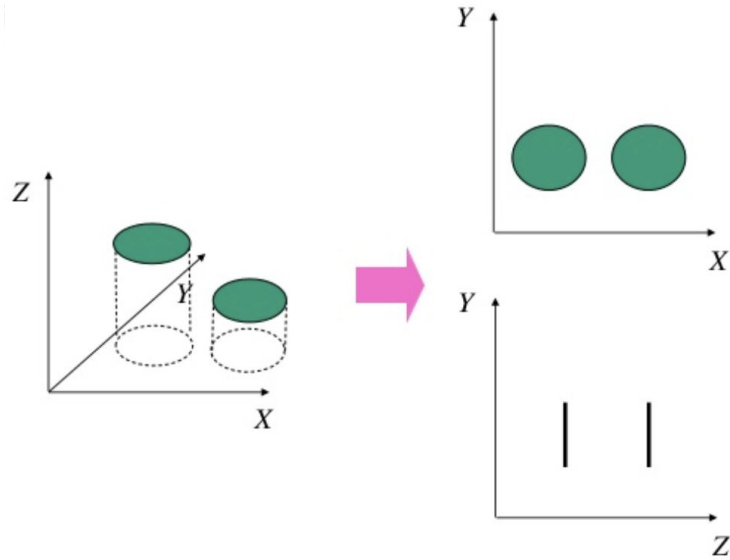
Count matrix



Data in reduced dimensions



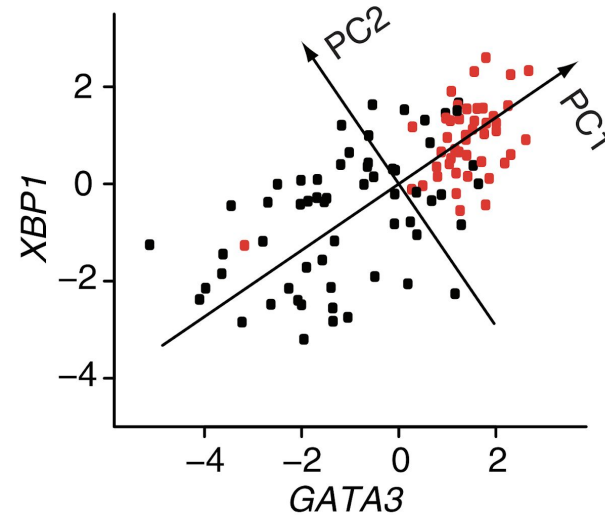
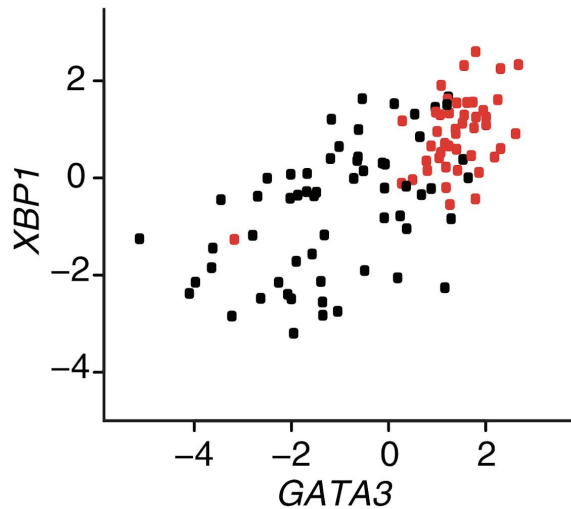
Dimensionality reduction – Projecting data into low-dim space



Dimensionality reduction using Principal Components Analysis

PCA geometrically projects data onto a lower-dimensional space

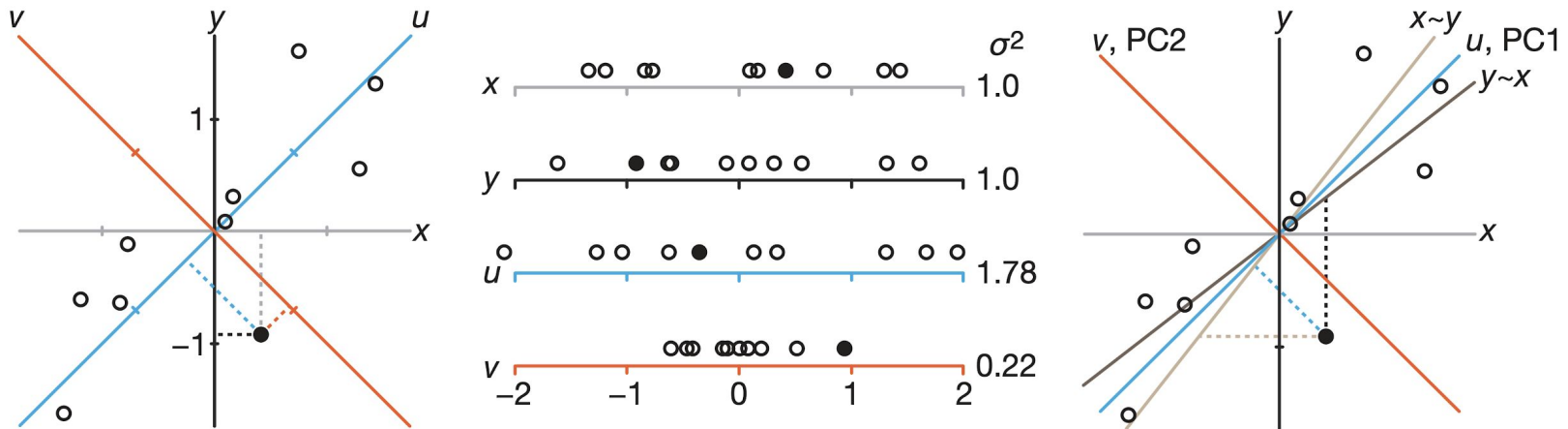
- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.



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.



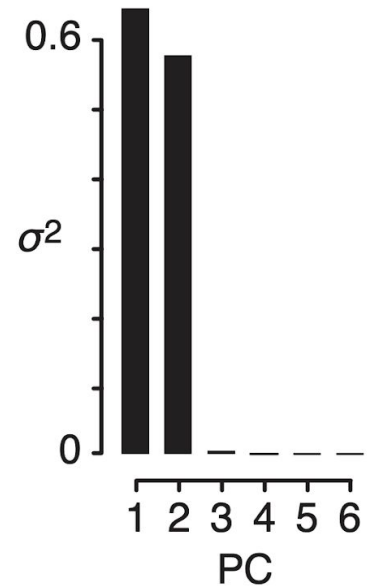
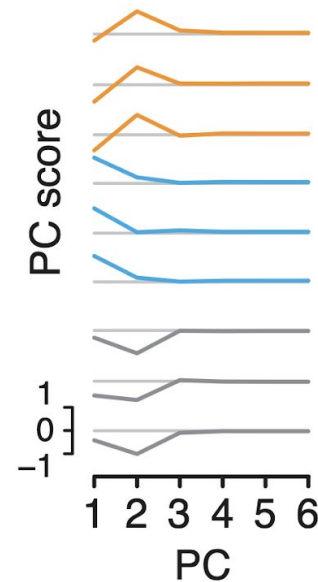
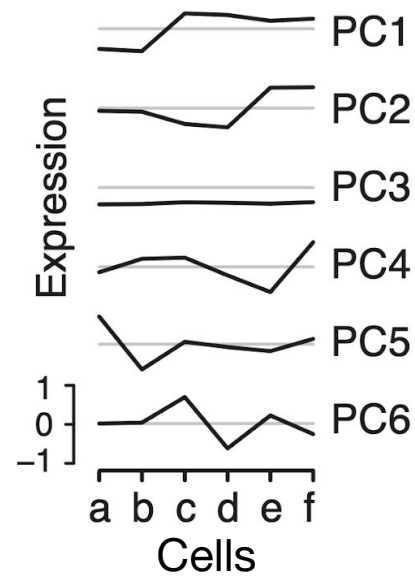
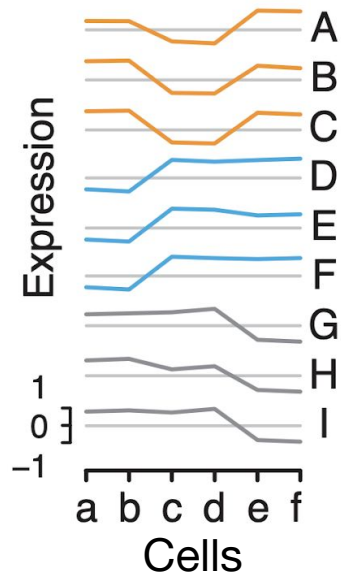
Dimensionality reduction – Principal Components Analysis

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.

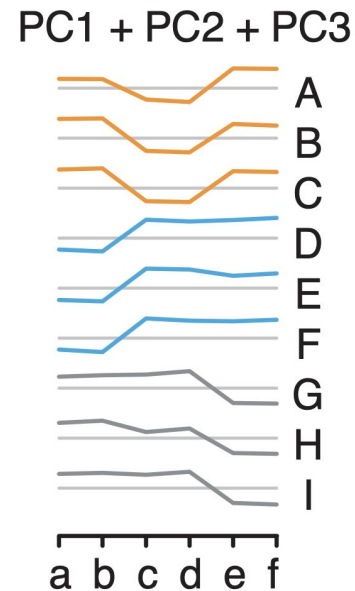
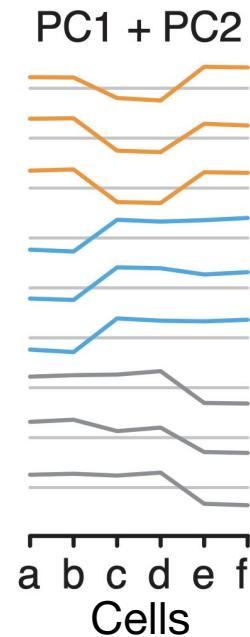
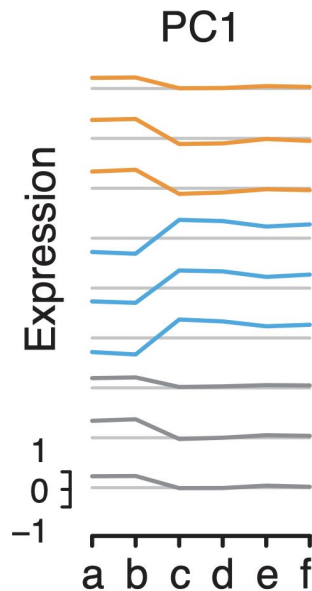
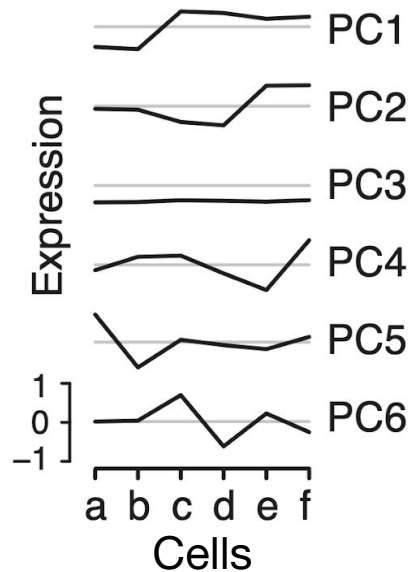
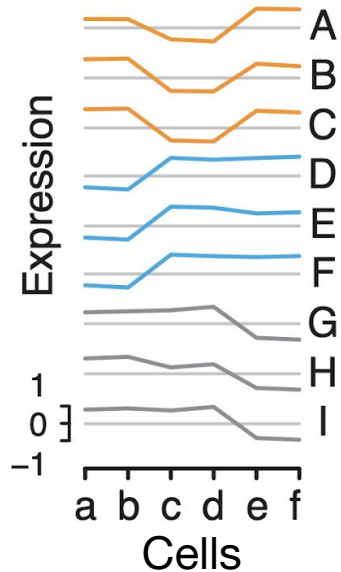
The idea is to transform the original data matrix X by rotation and scaling into a new set of axes so that:

- 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

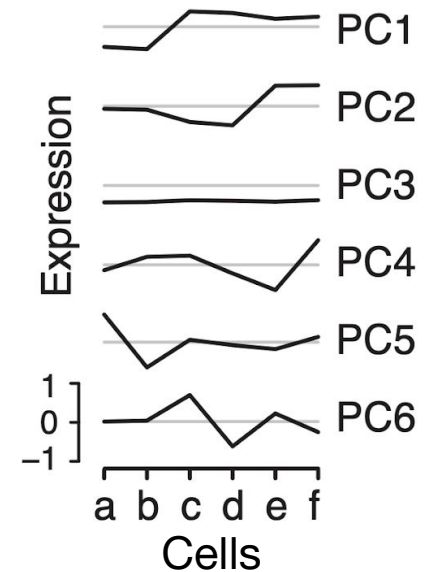
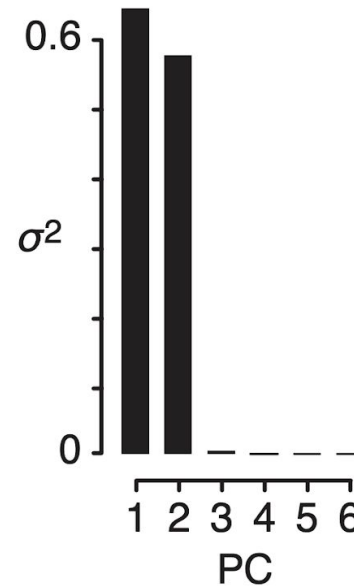
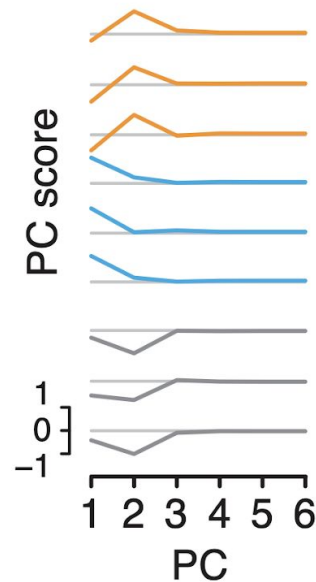
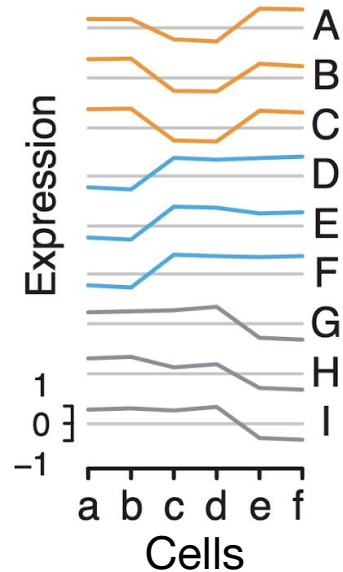


Dimensionality reduction by PCA



Dimensionality reduction by PCA

Singular Value Decomposition



Dimensionality reduction by PCA using SVD

SVD theorem

$$\begin{matrix} & n \\ & \boxed{X} \\ m & \end{matrix} = \begin{matrix} & r \\ & \boxed{U} \\ m & \end{matrix} \times \begin{matrix} & r \\ & \boxed{\Sigma} \\ r & \end{matrix} \times \begin{matrix} & n \\ & \boxed{V^T} \\ r & \end{matrix}$$

$$X = U \Sigma V^T$$

- 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.
- Σ is a diagonal matrix with elements σ_i . (All elements not on the main diagonal are 0.)
 - The elements of Σ are called the singular values of X . such that $\sigma_1 \geq \sigma_2 \geq \dots \sigma_i \dots \geq \sigma_r$.
- $X = \sum_i \sigma_i u_i v_i$

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

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

X
 U
 Σ
 V^T

- What do the columns of U represent?
- What do the rows of V^T represent?
- What do the diagonal entries of Σ represent?

	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

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

X
 U
 Σ
 V^T

How do we do dimensionality reduction from here?

	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

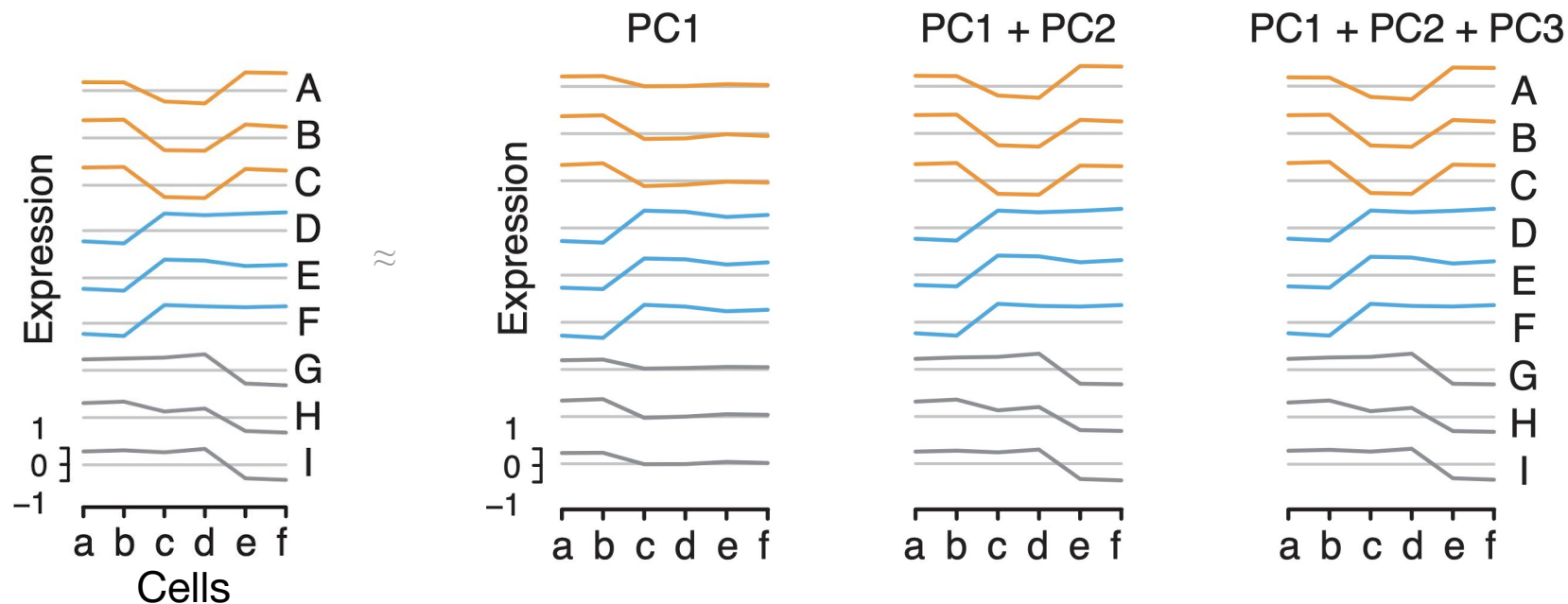
Dimensionality reduction by PCA using SVD

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

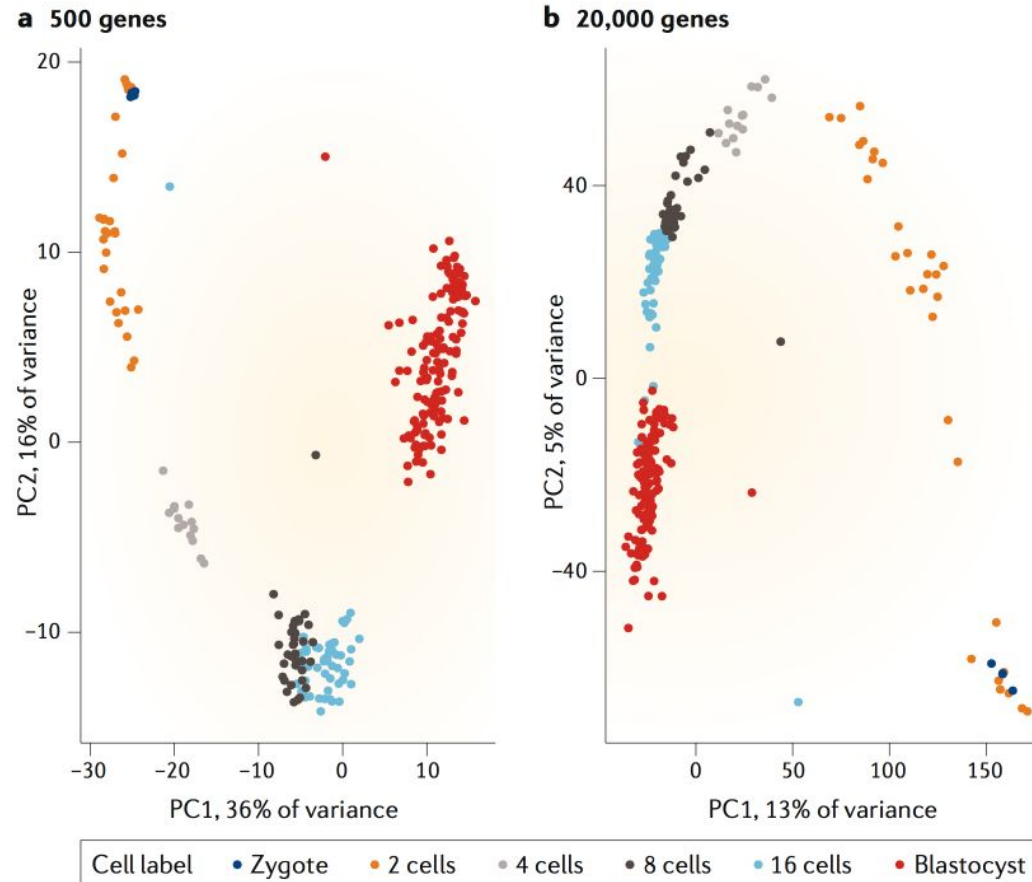
$$X = \sum_i \sigma_i u_i v_i$$

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

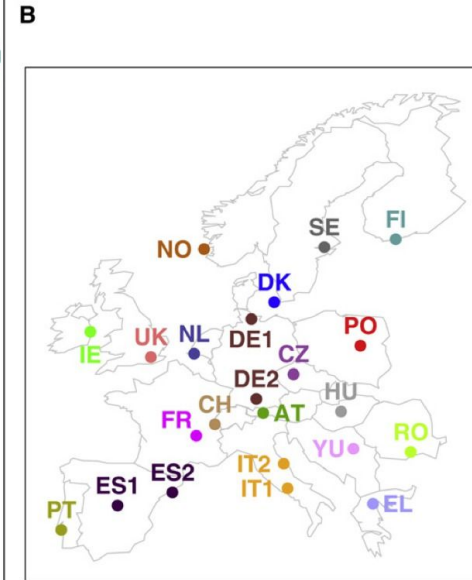
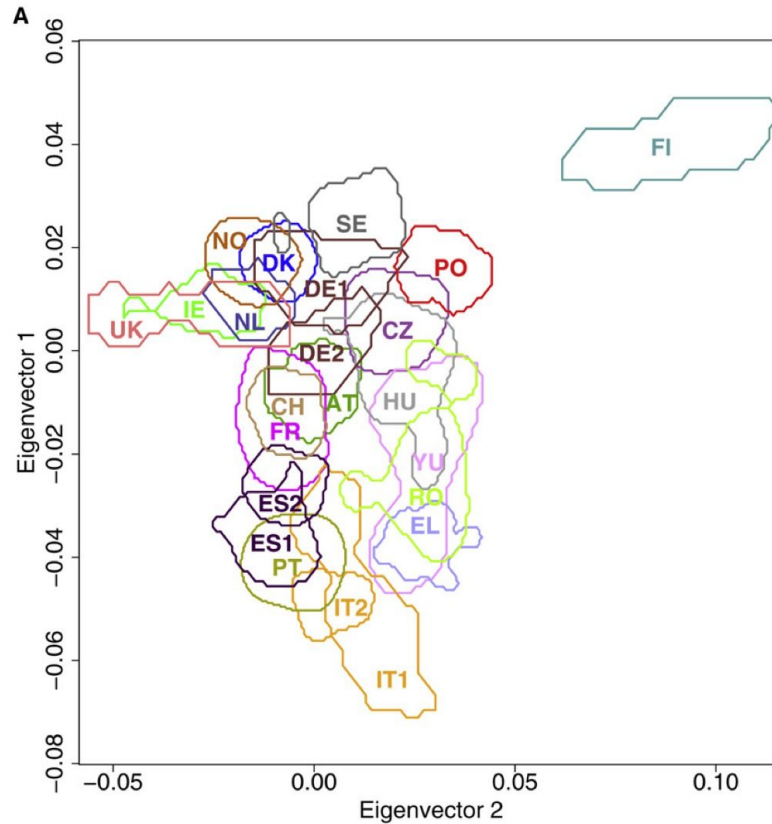
Reconstructing the original data matrix from PCs



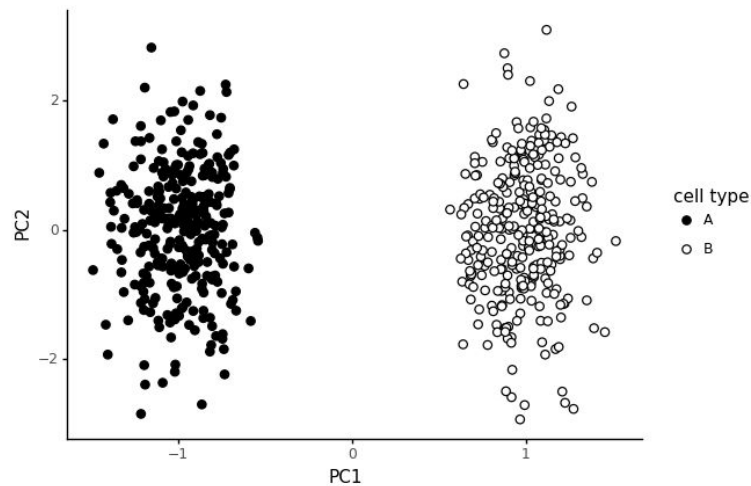
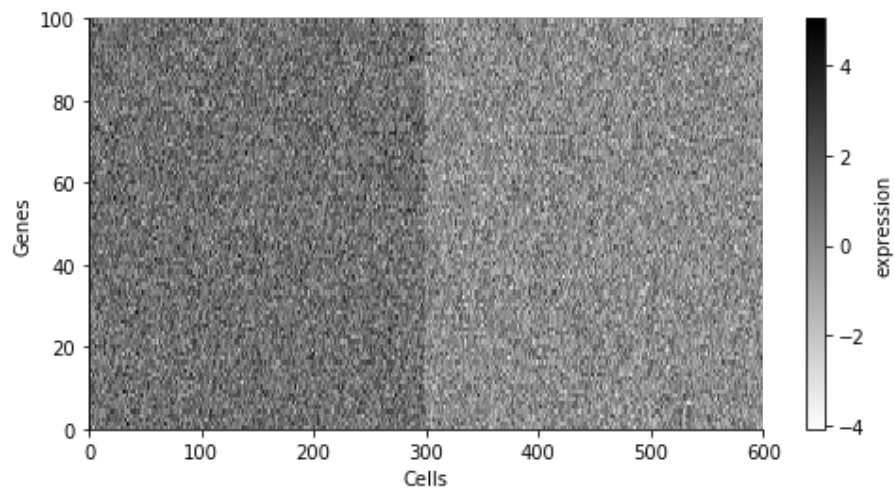
Dimensionality reduction by PCA



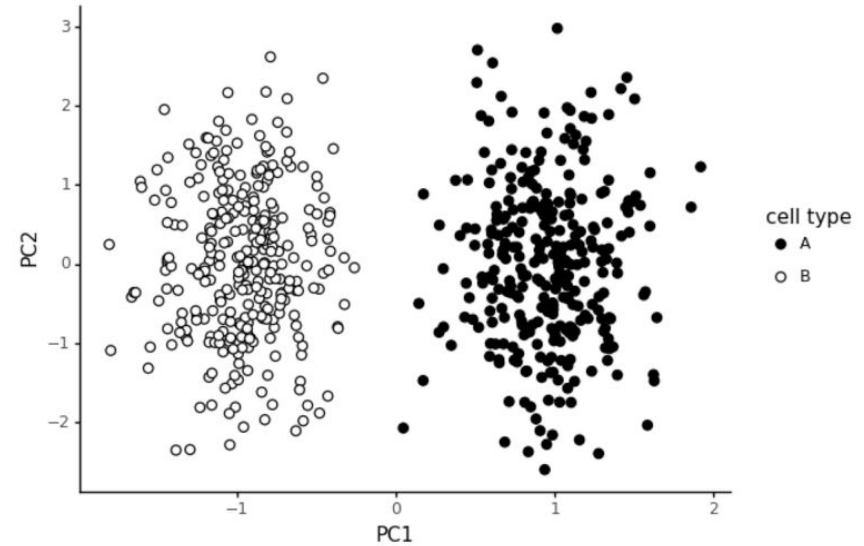
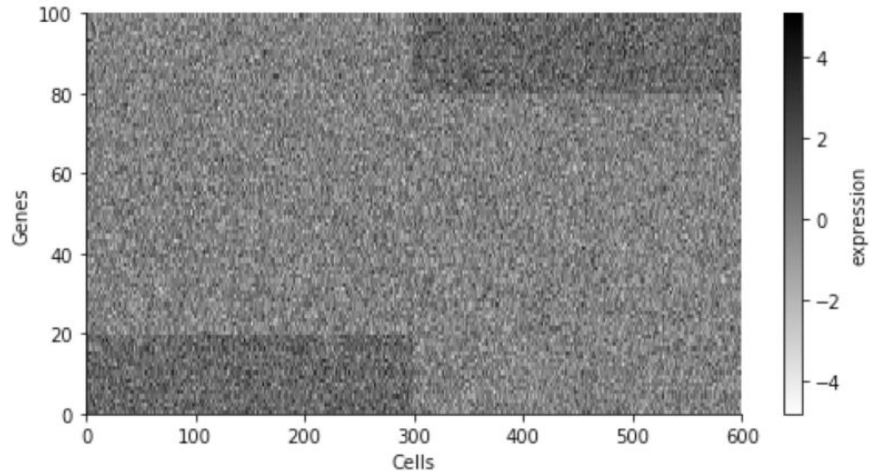
Dimensionality reduction



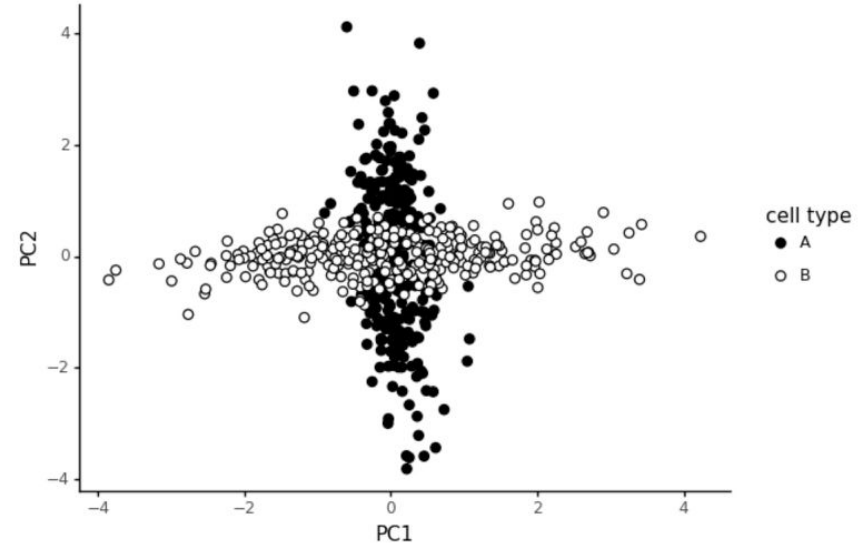
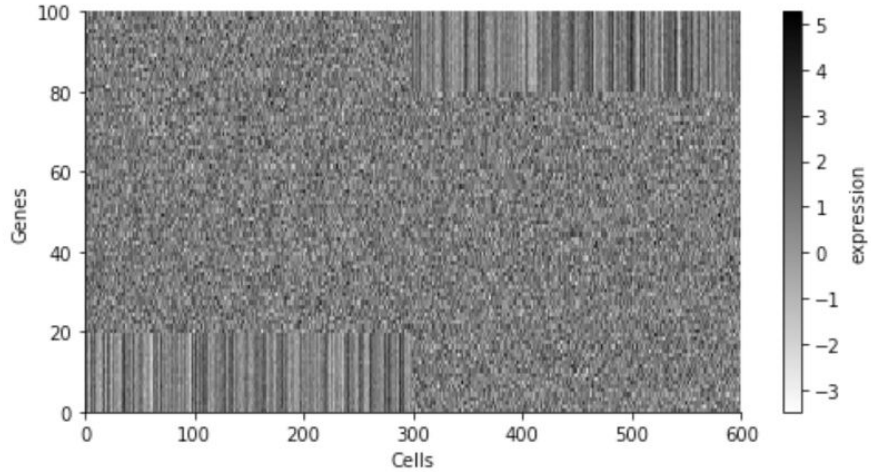
Dimensionality reduction by PCA



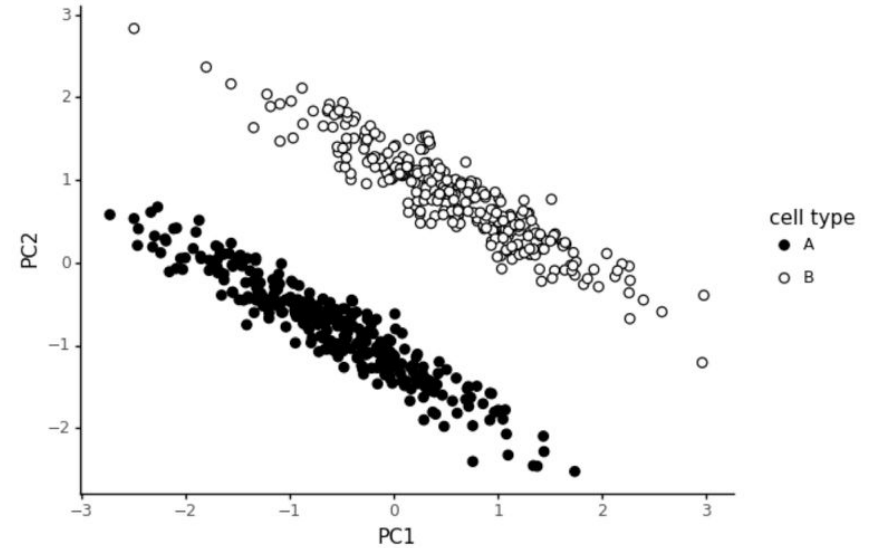
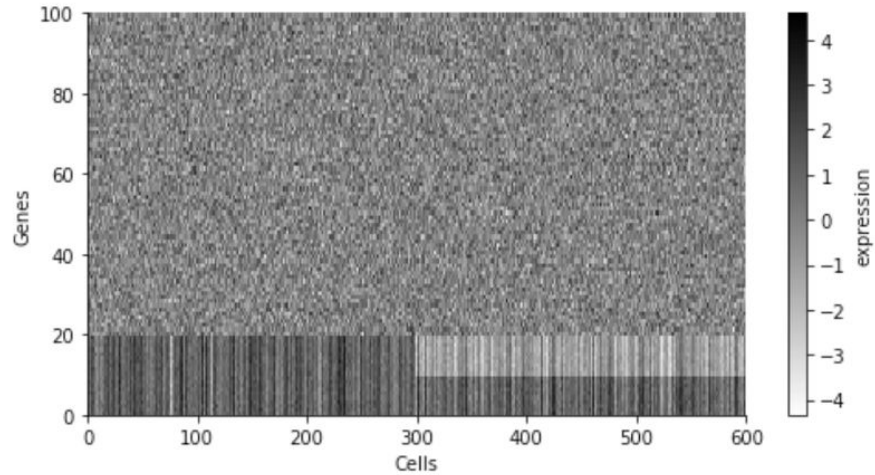
Dimensionality reduction by PCA



Dimensionality reduction by PCA

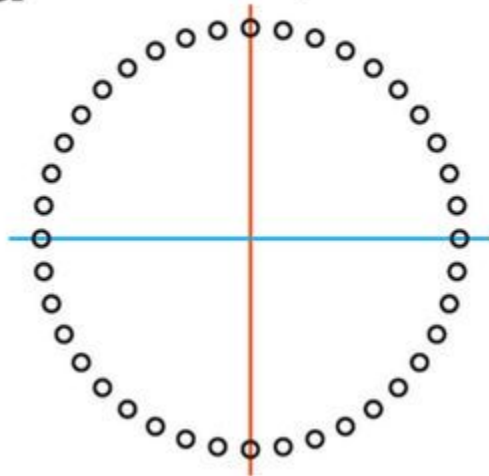


Dimensionality reduction by PCA

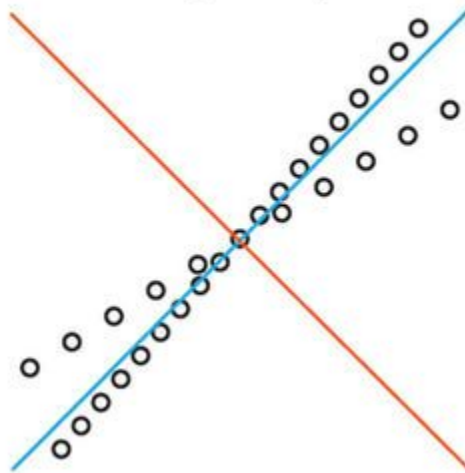


Limitations of PCA, alternatives

a Nonlinear patterns



b Nonorthogonal patterns



c Obscured clusters

