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

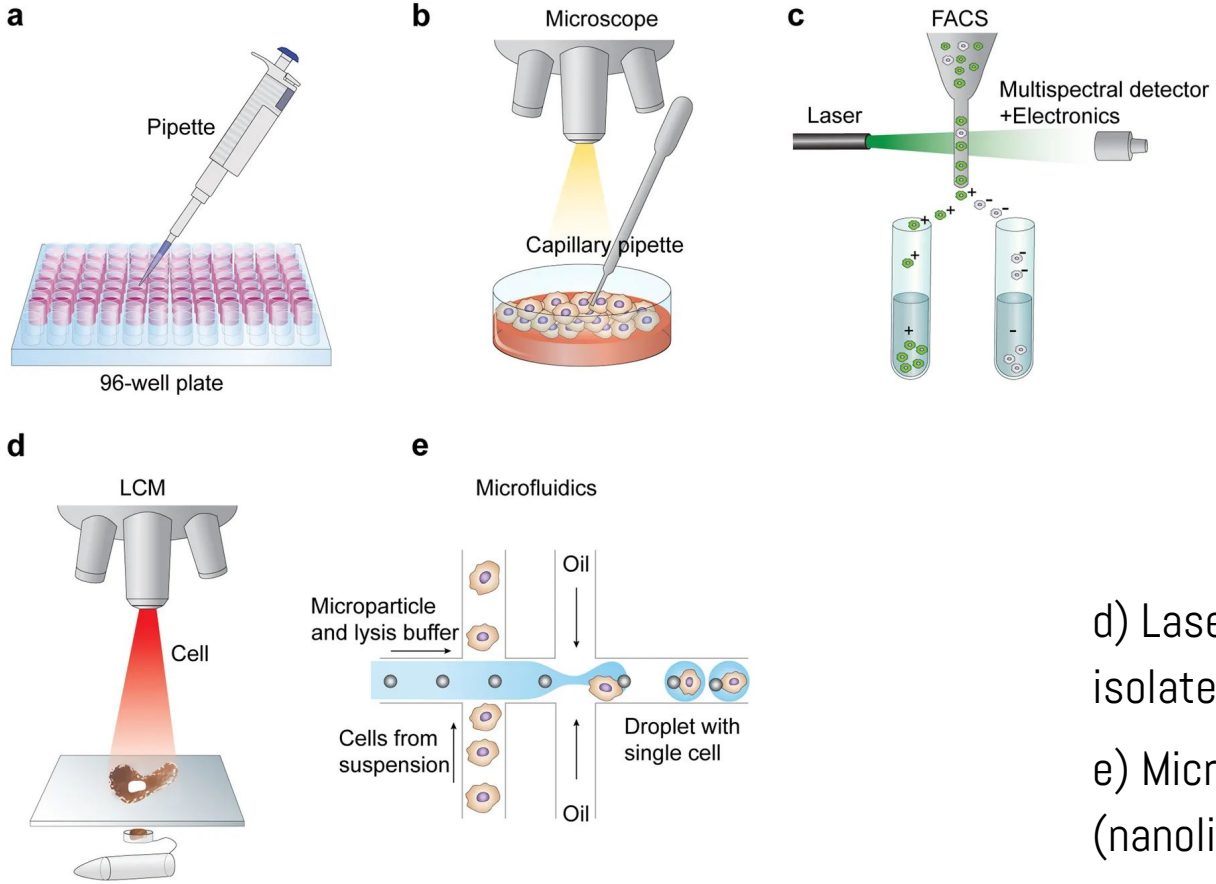
A zebrafish embryo at an early stage of development. Fluorescent markers highlight cells expressing genes that help determine the type of cell they will become. (JEFFREY FARRELL, SCHIER LAB/HARVARD UNIVERSITY)

The single-cell revolution is just starting.

— Elizabeth Pennisi

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

Single-cell isolation and library preparation



a) Limiting dilution method isolates individual cells.

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

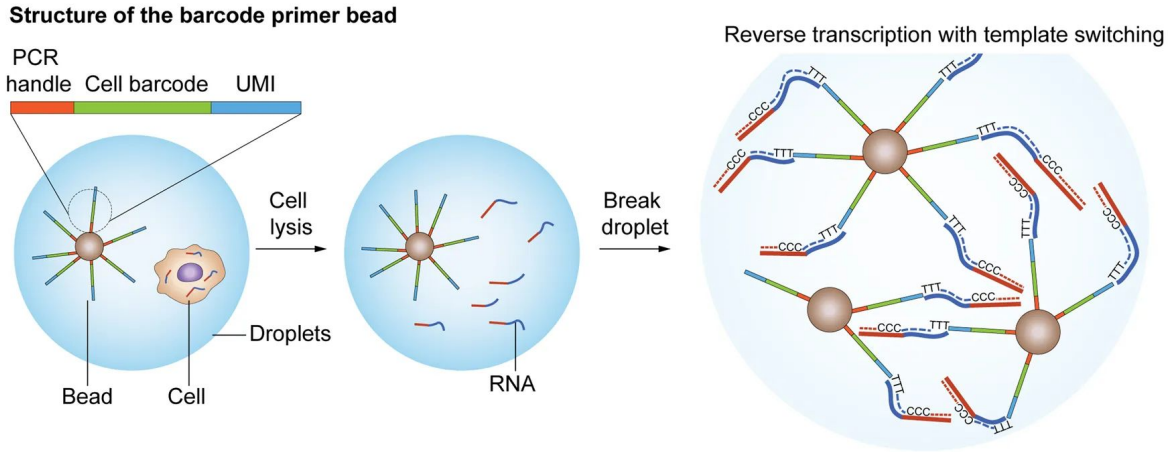
c) FACS isolates highly purified single cells by tagging cells w/ markers.

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

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

Hwang (2018) Exp. Mol. Med.

Single-cell isolation and library preparation



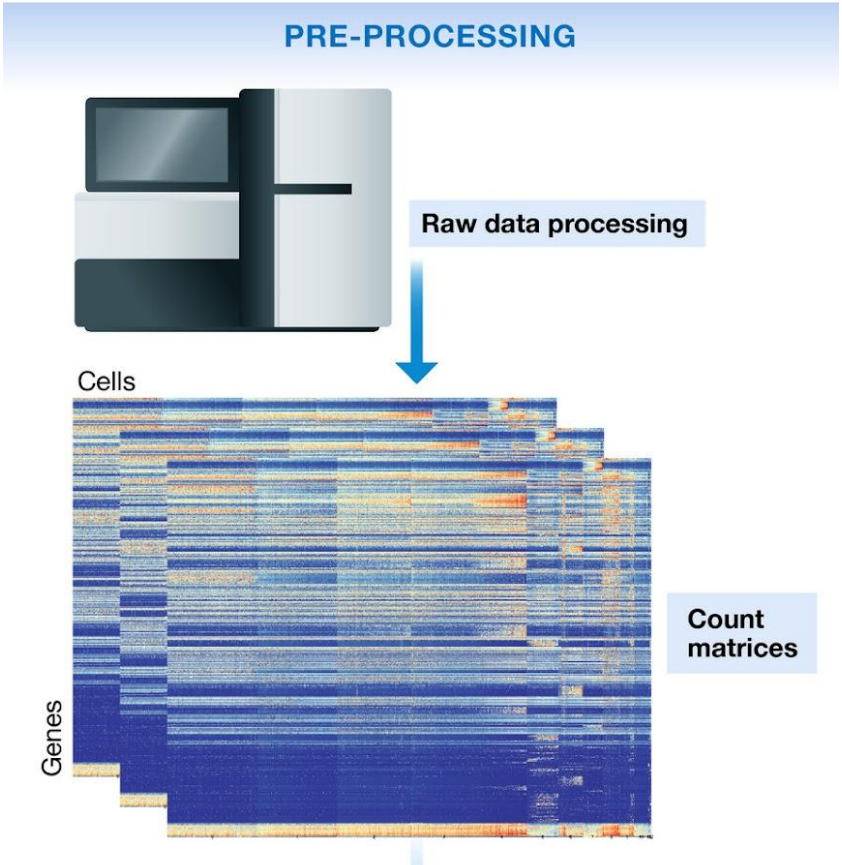
A schematic example of droplet-based library generation.

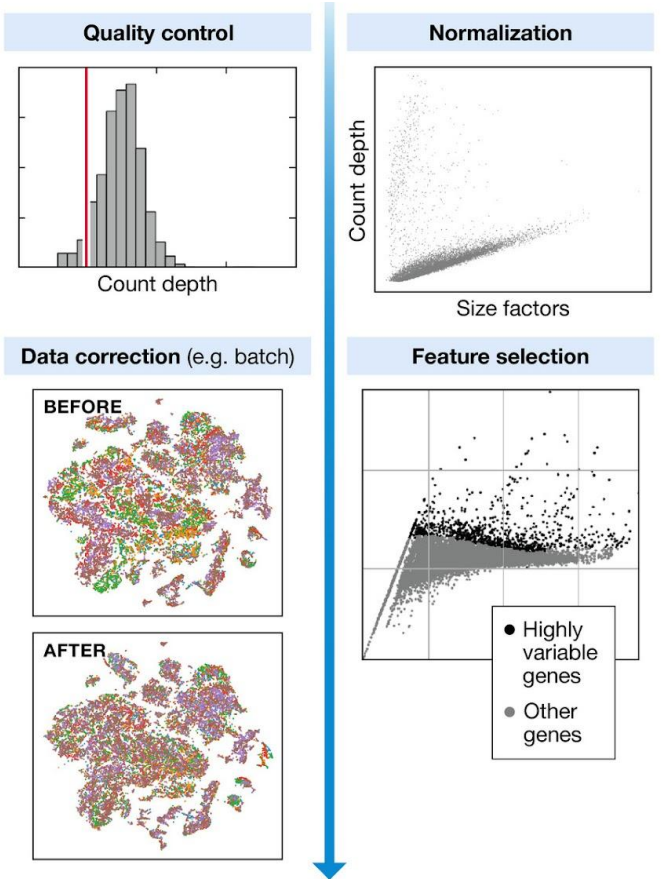
Libraries for scRNA-seq are typically generated via:

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

Hwang (2018) Exp. Mol. Med.

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





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

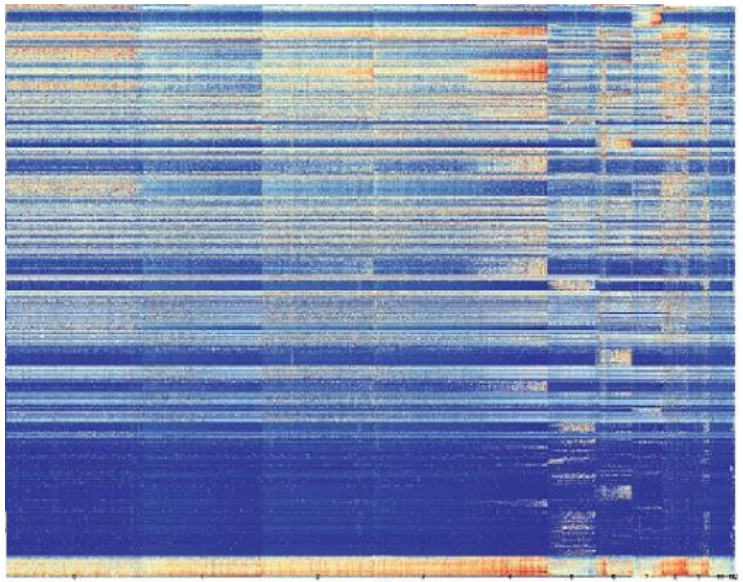
Dimensionality reduction of scRNA-seq data

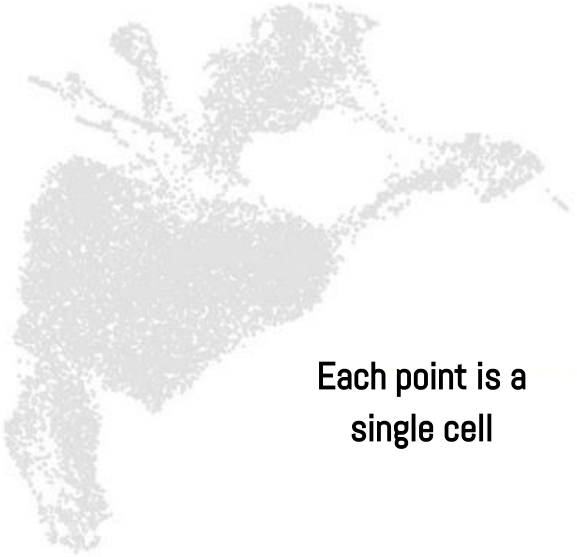
Count matrix

Data in reduced dimensions

cells

genes

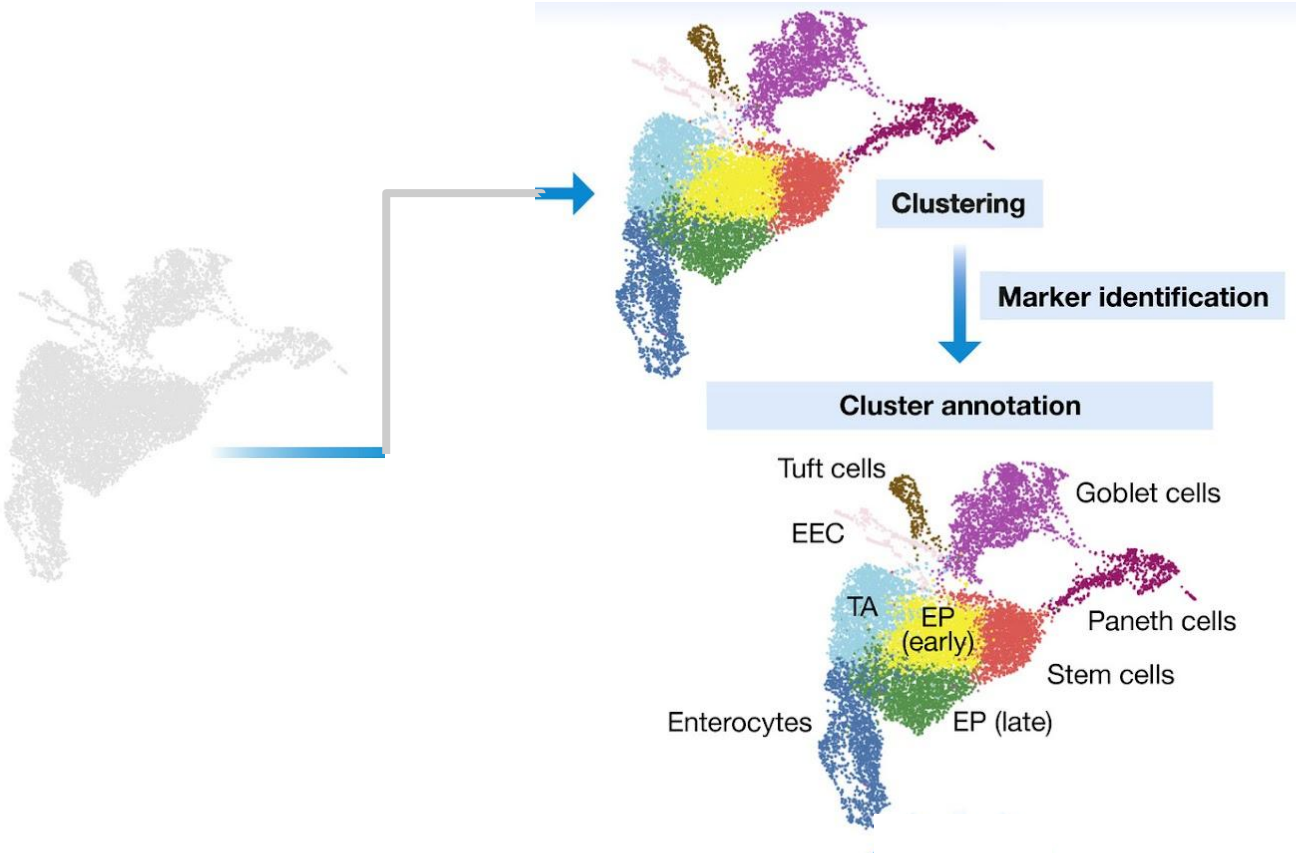




Each point is a single cell

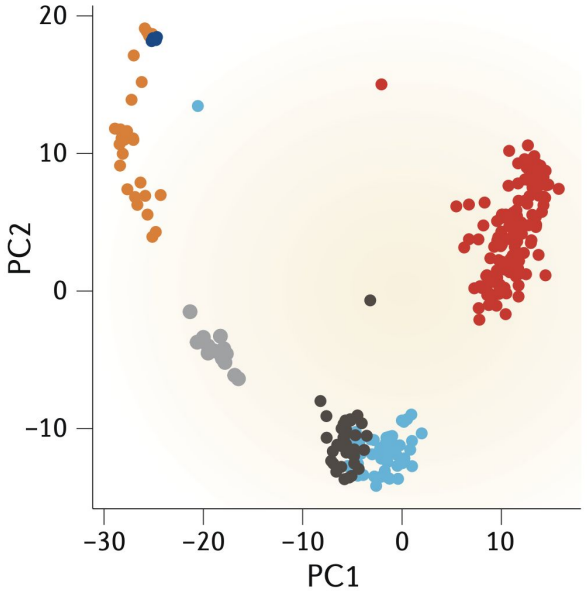
Luecken & Theis (2019) Mol. Sys. Biol.

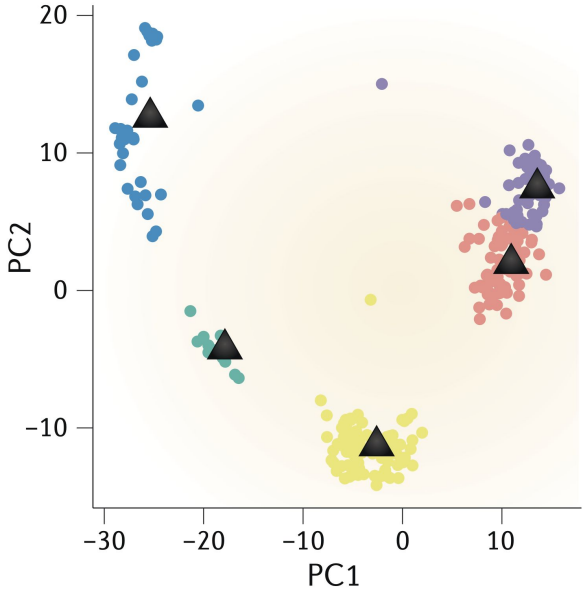
Clustering scRNA-seq data to identify cell types/states



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

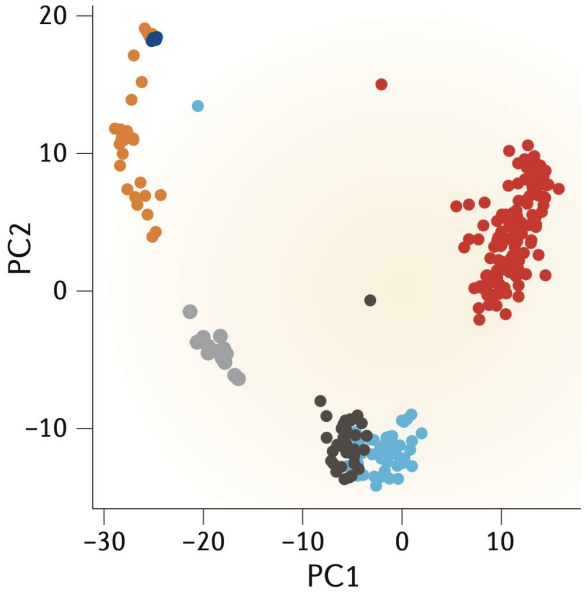
Clustering scRNA-seq data to identify cell types/states

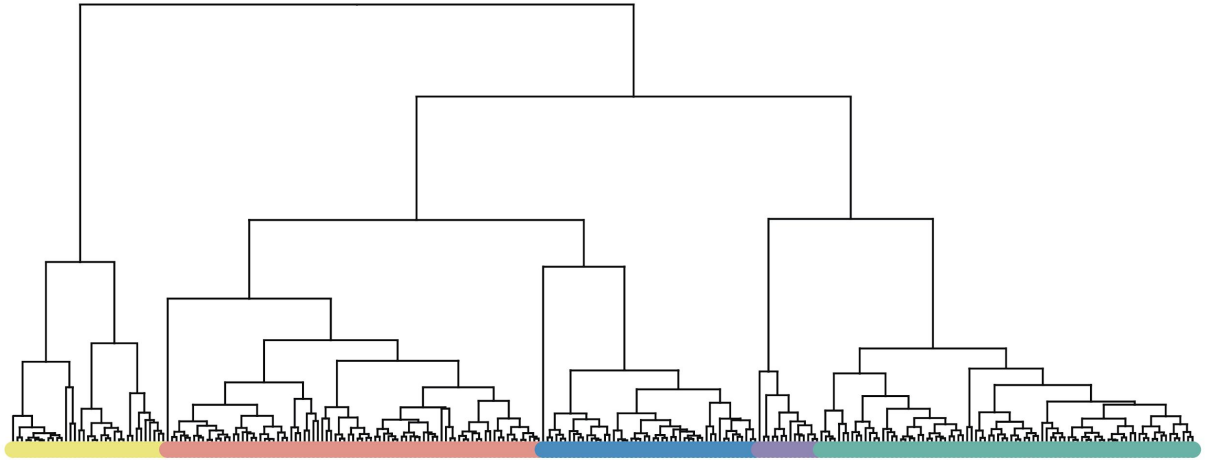




Kiselev (2019) Nat. Rev. Genet..

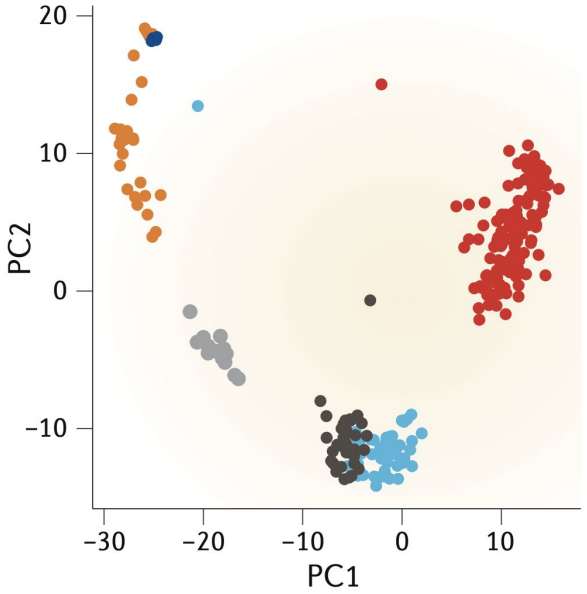
Clustering scRNA-seq data to identify cell types/states

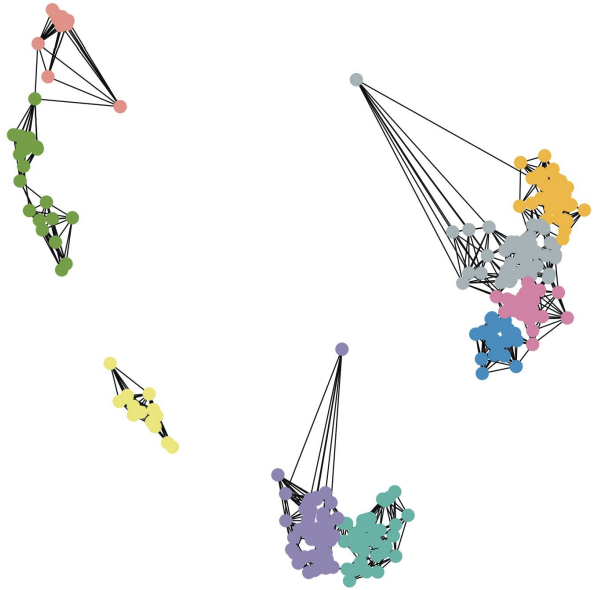




Kiselev (2019) Nat. Rev. Genet..

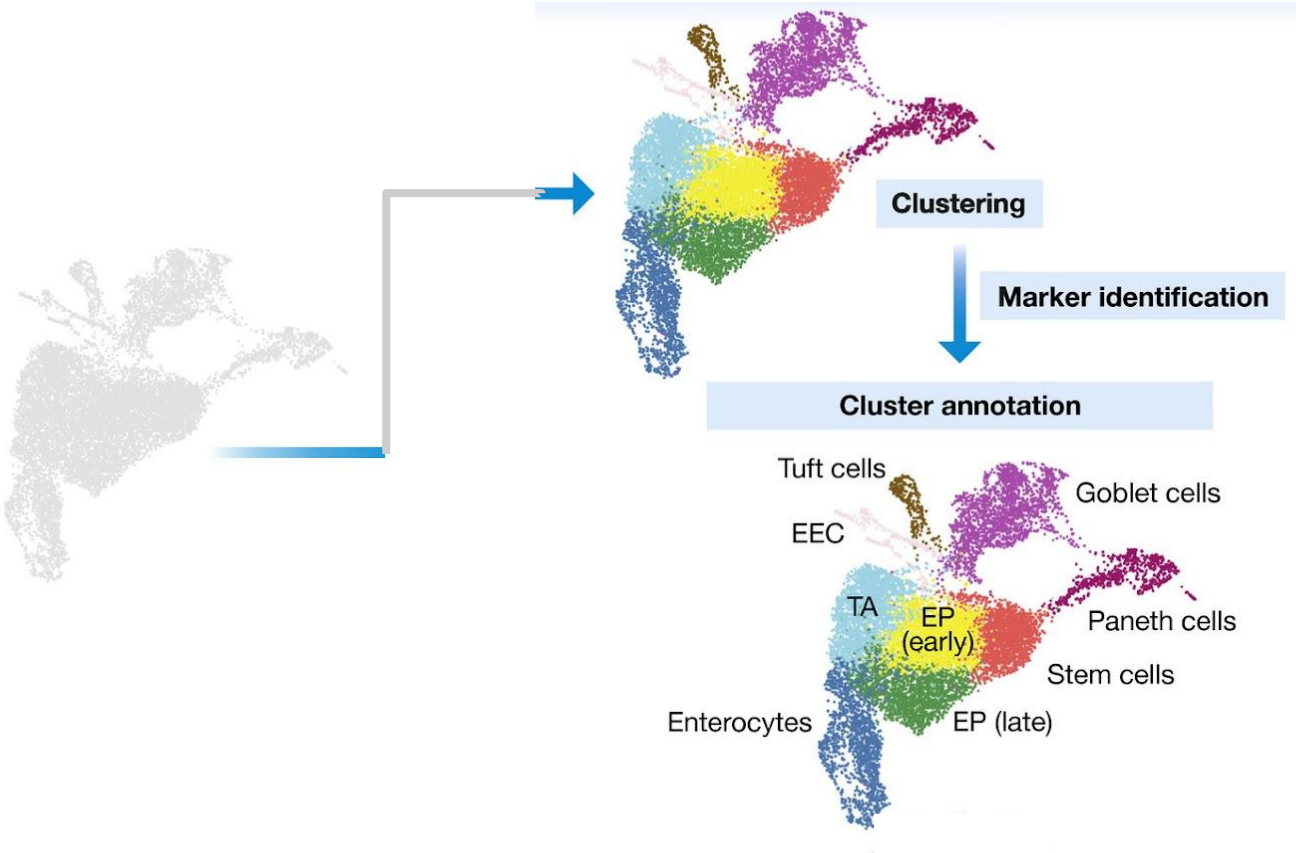
Clustering scRNA-seq data to identify cell types/states





Kiselev (2019) Nat. Rev. Genet..

Clustering scRNA-seq data to identify cell types/states



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

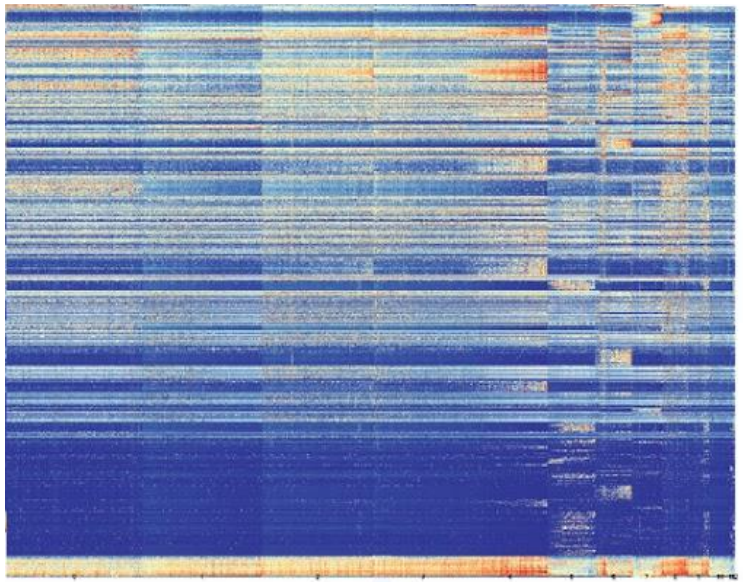
Dimensionality reduction of scRNA-seq data

Count matrix

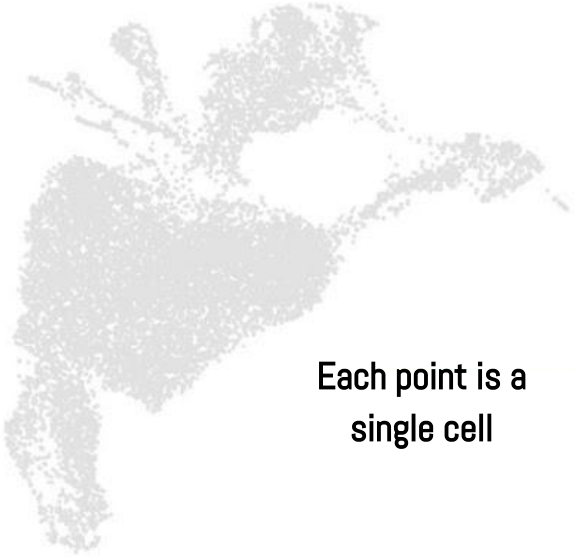
Data in reduced dimensions

cells

genes



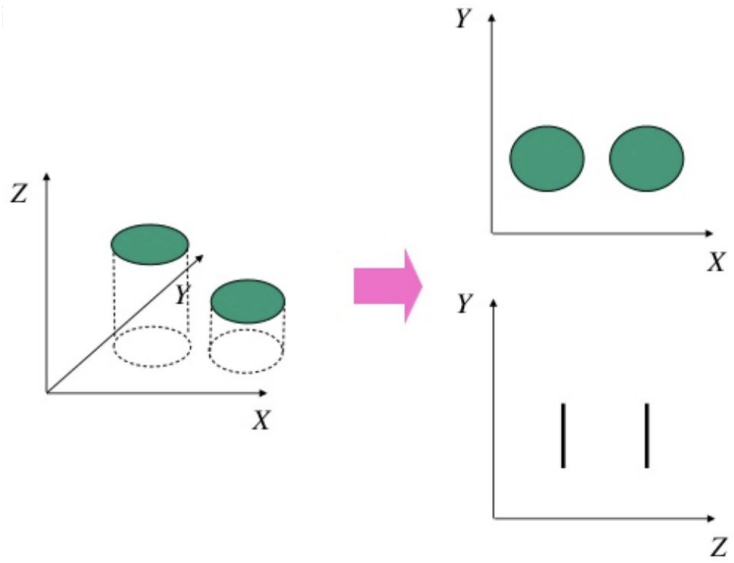
?



Each point is a single cell

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

Dimensionality reduction – Projecting data into low-dim space

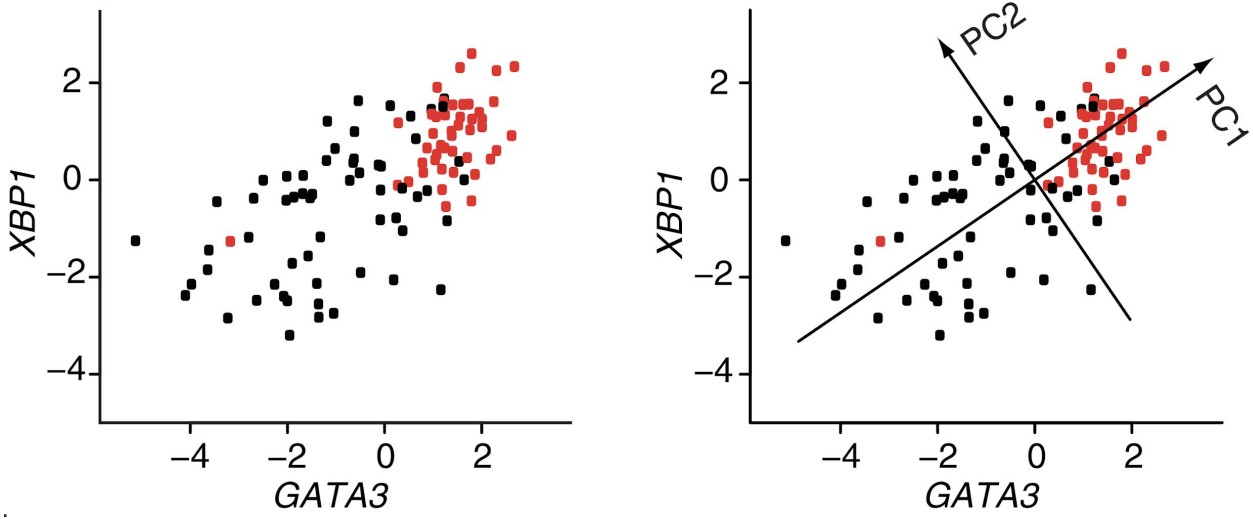


Lever, Krzywinski, & Altman (2017) Nat. Meth.

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.

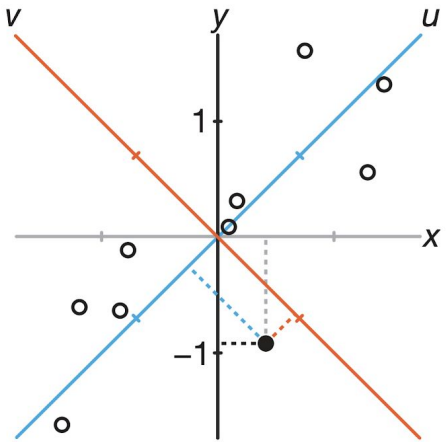


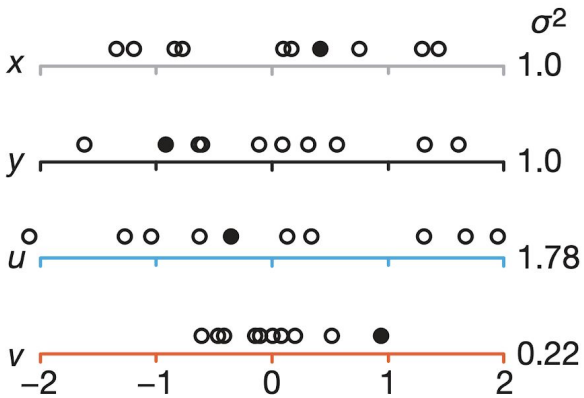
Ringnér (2008) Nat. Biotech

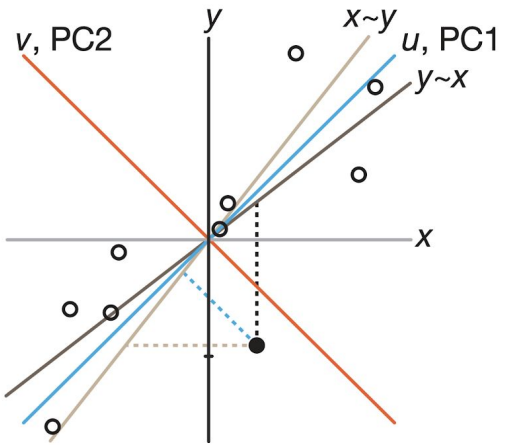
Dimensionality reduction using Principal Components Analysis

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.

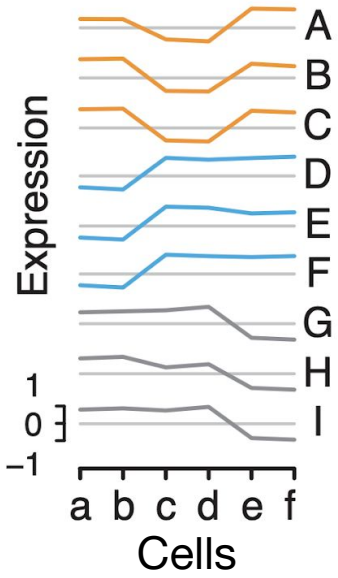
Dimensionality reduction – Principal Components Analysis

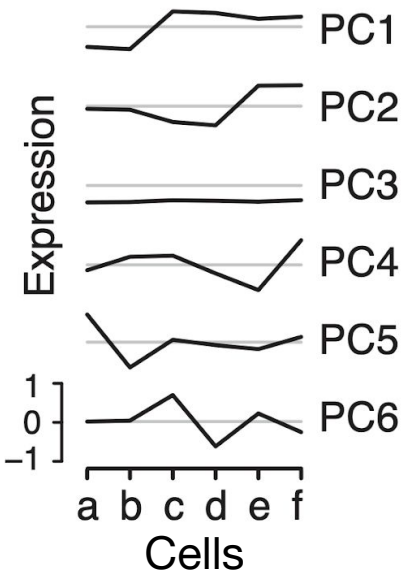
Given a dataset consisting of a set of observations representing points in a high-dimensional space, PCA finds the directions along which the observations line up best.

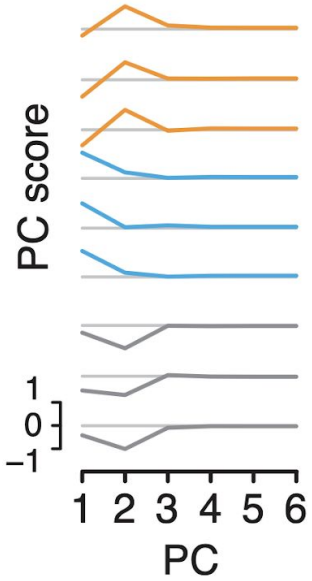
The idea is to transform the original data matrix X by rotation and scaling into a new set of axes so that:

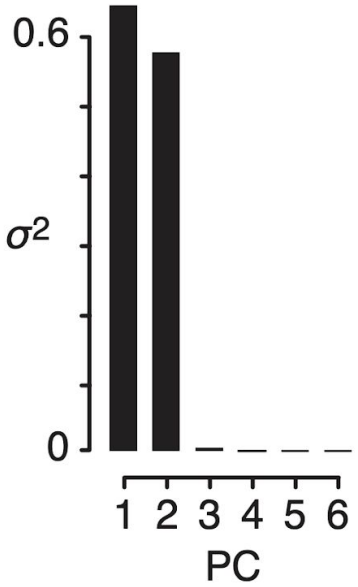
- Each new axes, termed a principal component (PC) is a linear combination of the original dimensions.
- The axis corresponding to the 1st PC satisfies the following:
  - The 1st PC is the axis along which the points are most “spread out”.
    - The axis along which the variance of the data is maximized.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

Dimensionality reduction by PCA



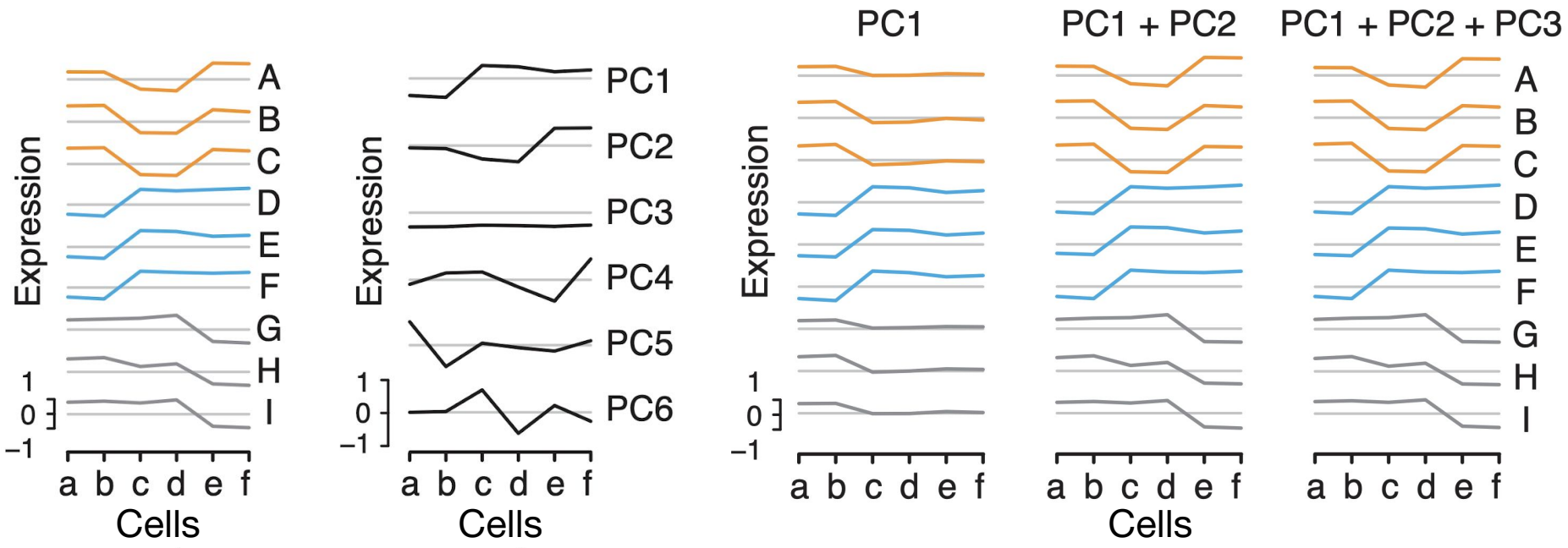






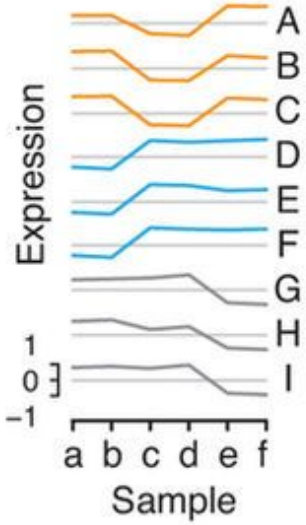
Lever, Krzywinski, & Altman (2017) Nat. Meth.

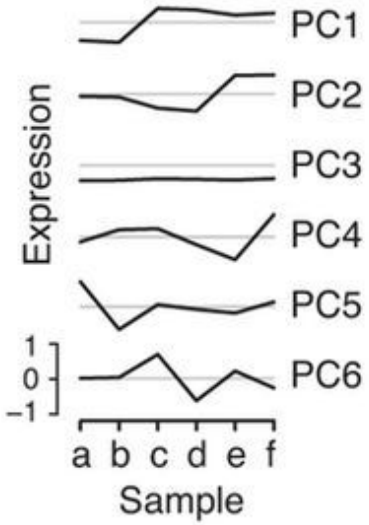
Dimensionality reduction by PCA

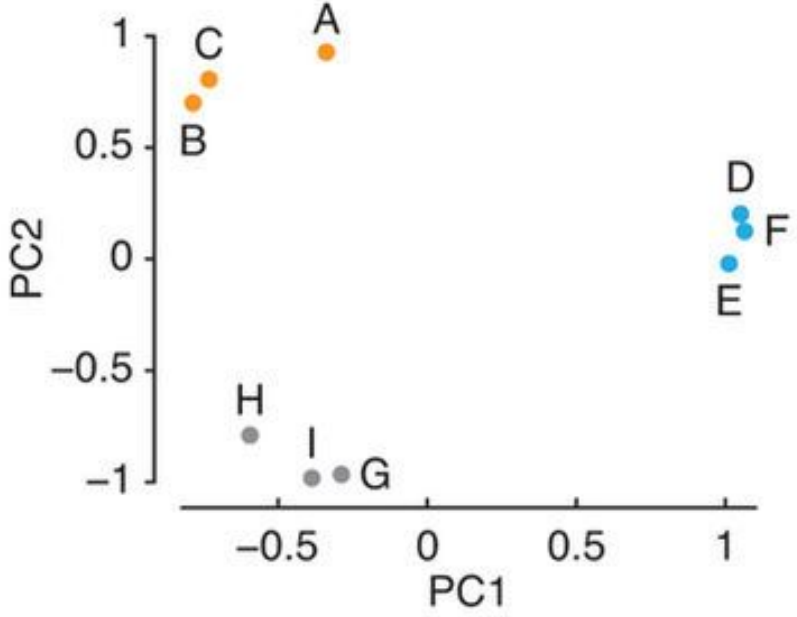


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



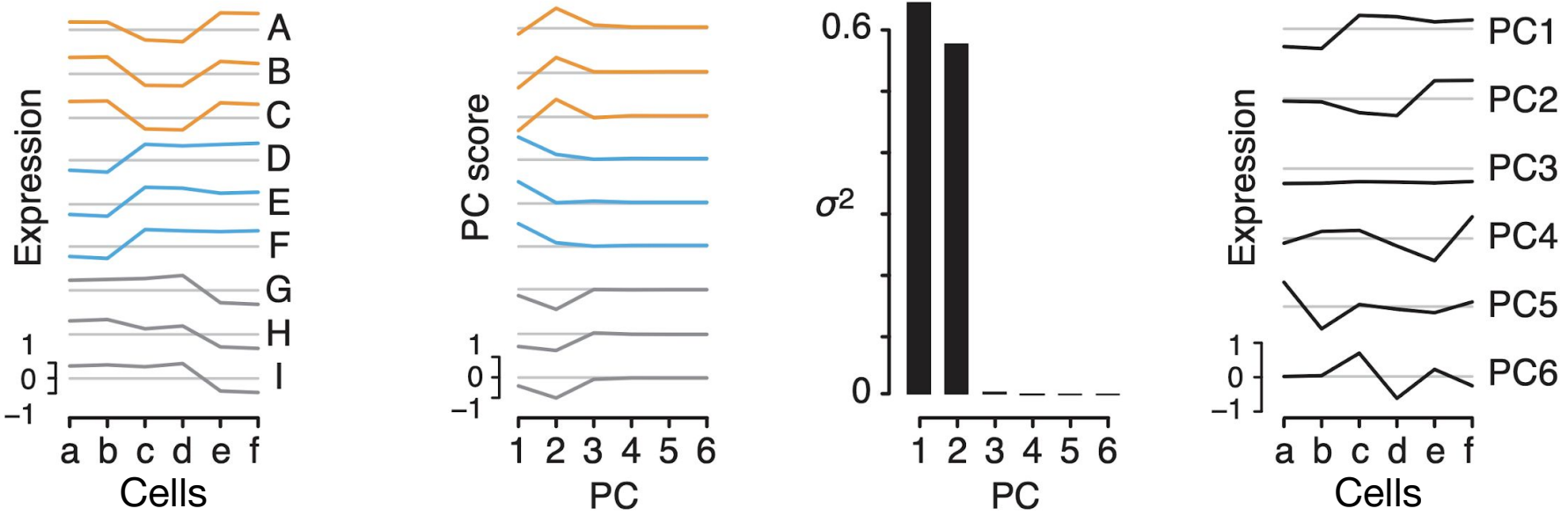




Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

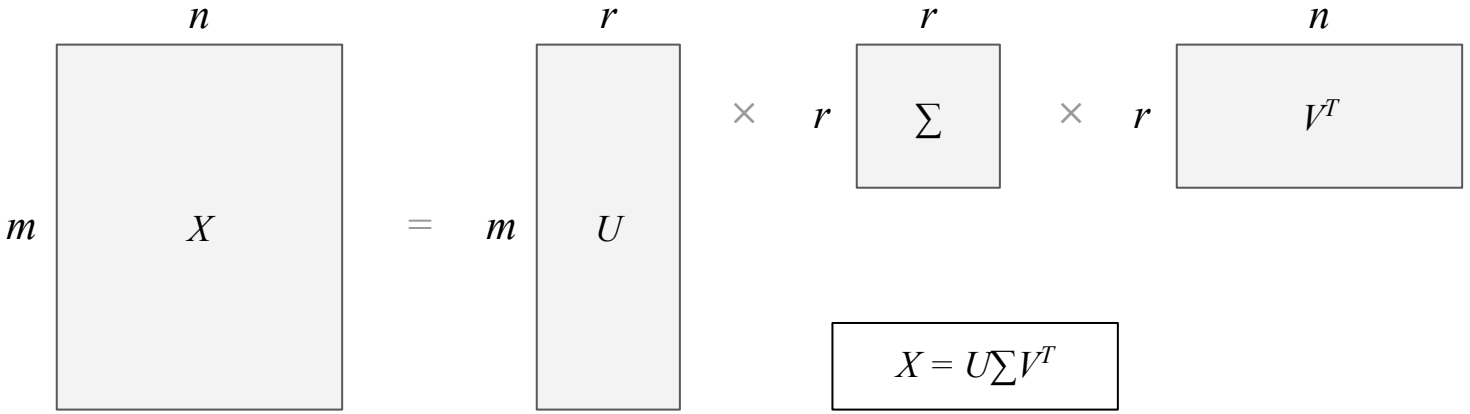
Singular Value Decomposition



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

	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

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

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

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

	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

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

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

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

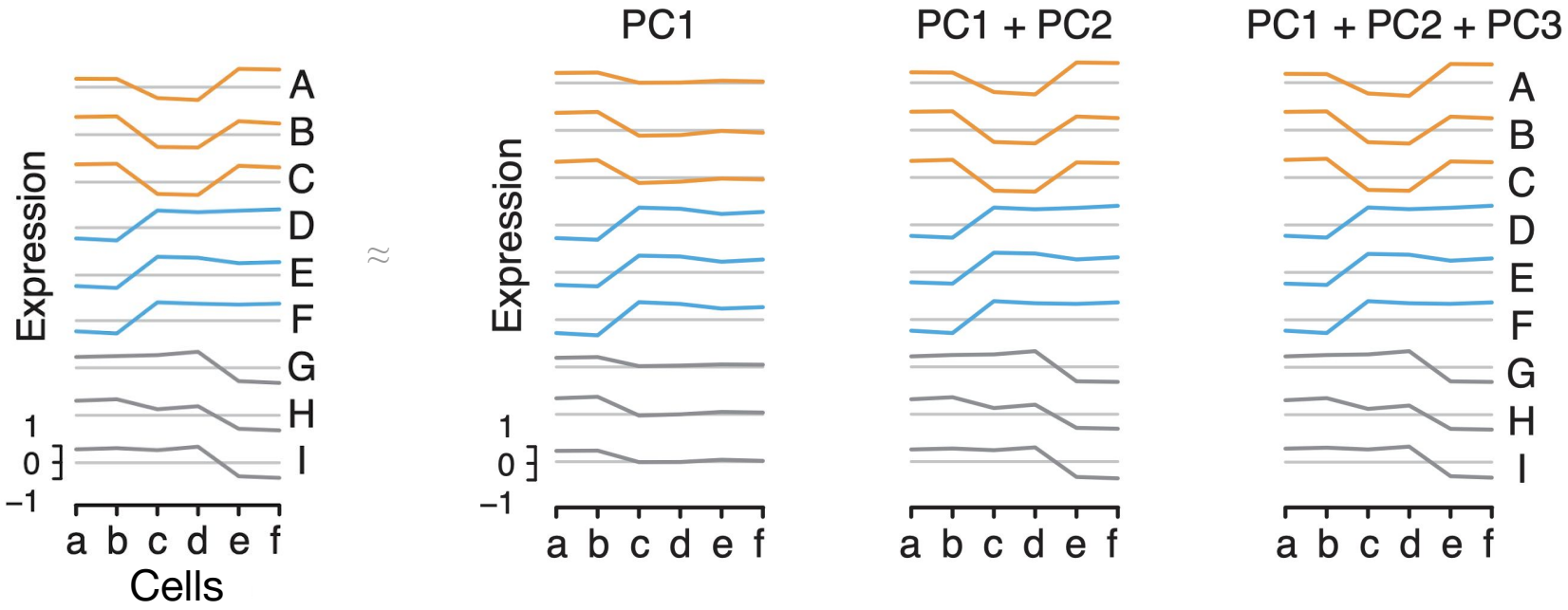
$$X = U \sum V^T$$

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

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

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

Reconstructing the original data matrix from PCs



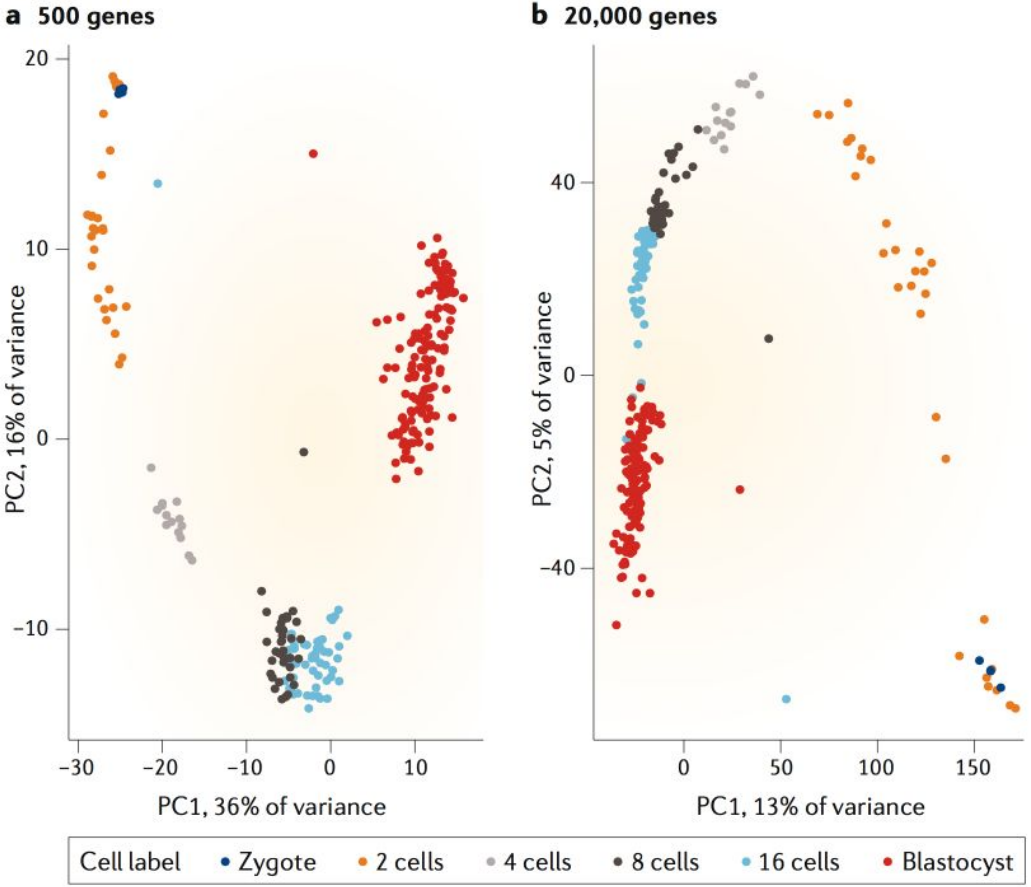
$X = \sum_i \sigma_i u_i v_i$

$X \approx \sigma_1 u_1 v_1$

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

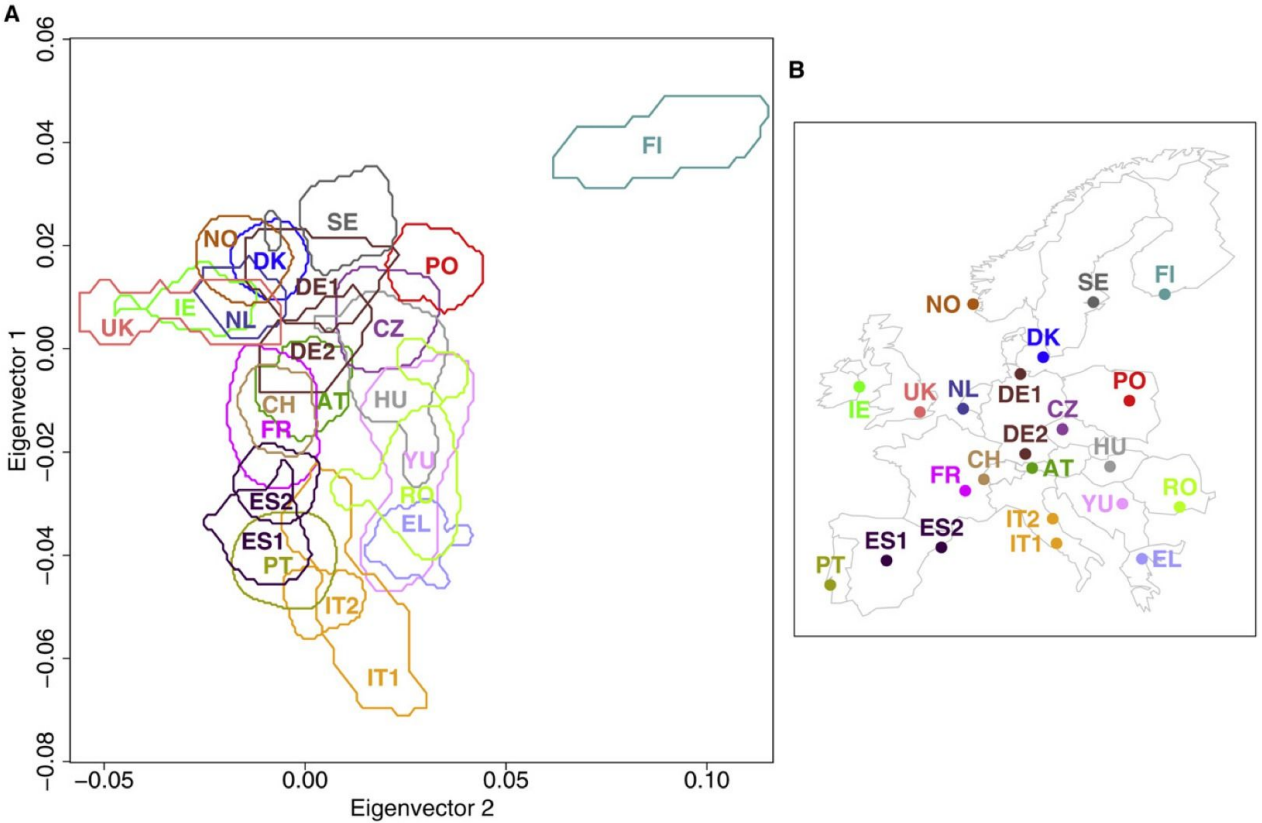
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



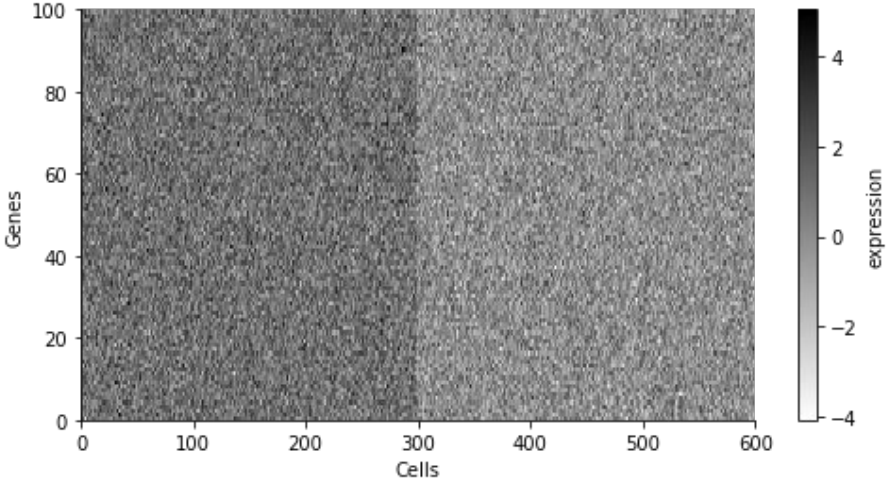
Kiselev (2019) Nat. Rev. Genet..

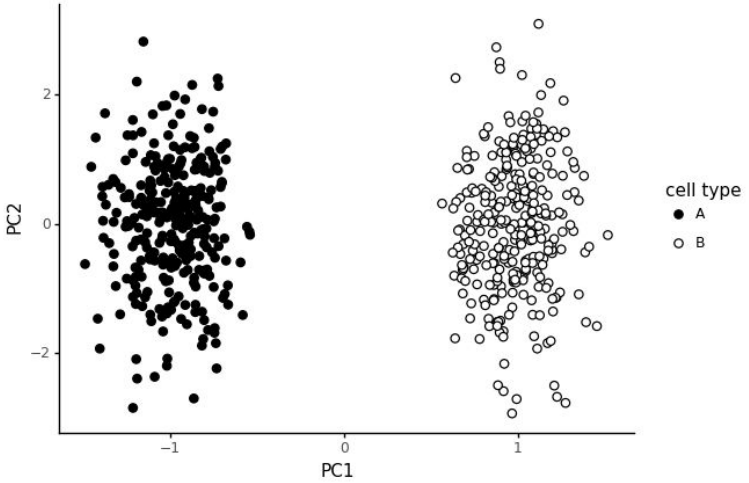
Dimensionality reduction



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

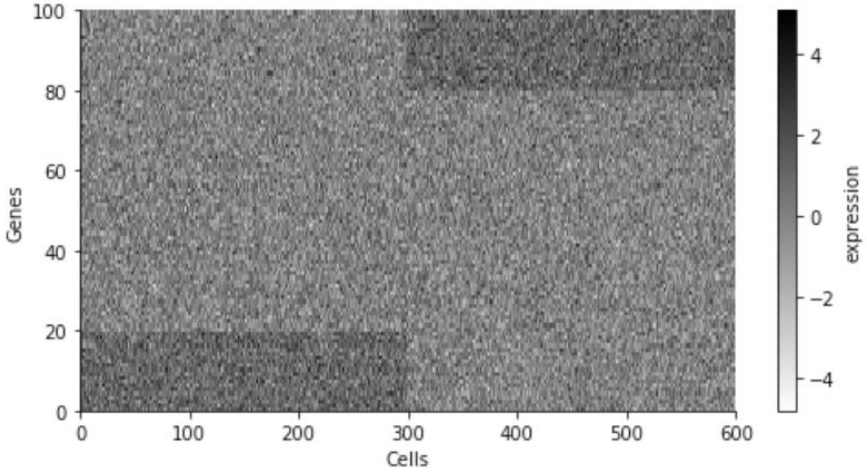
Dimensionality reduction by PCA

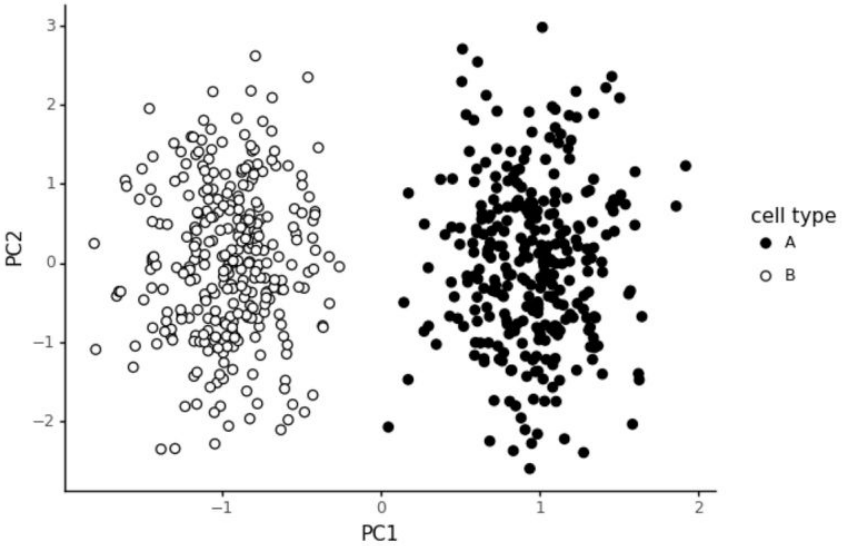




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

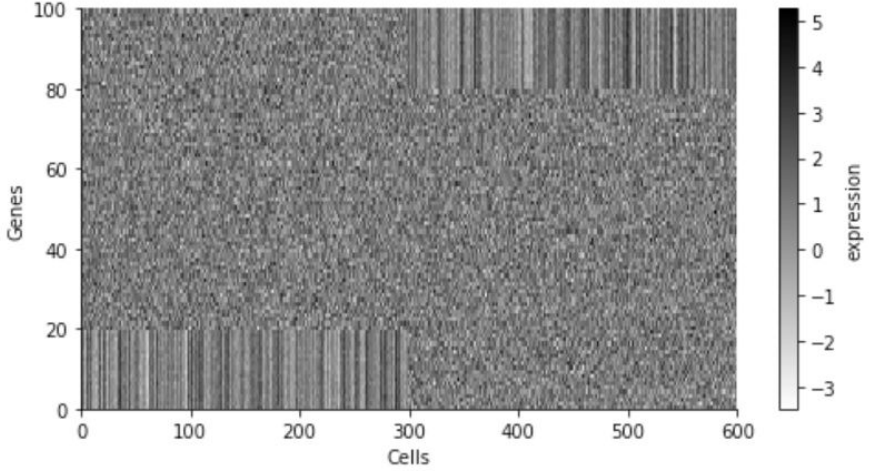
Dimensionality reduction by PCA

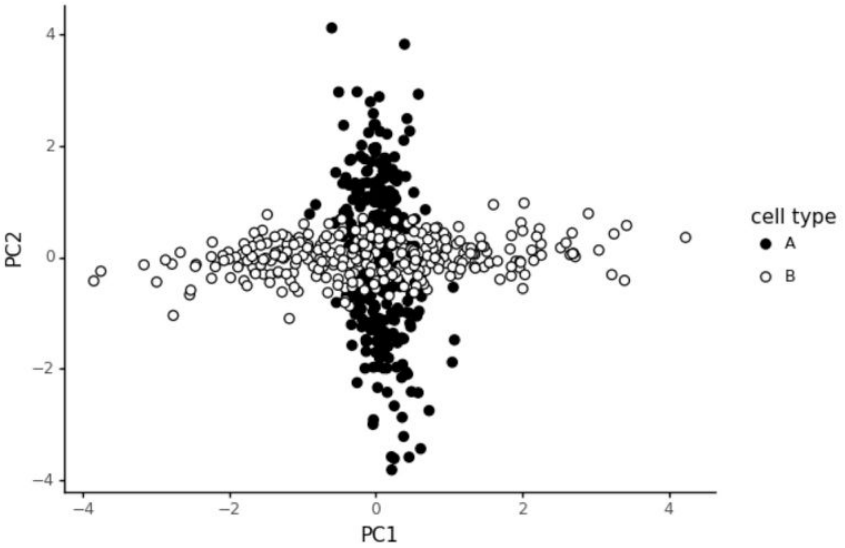




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

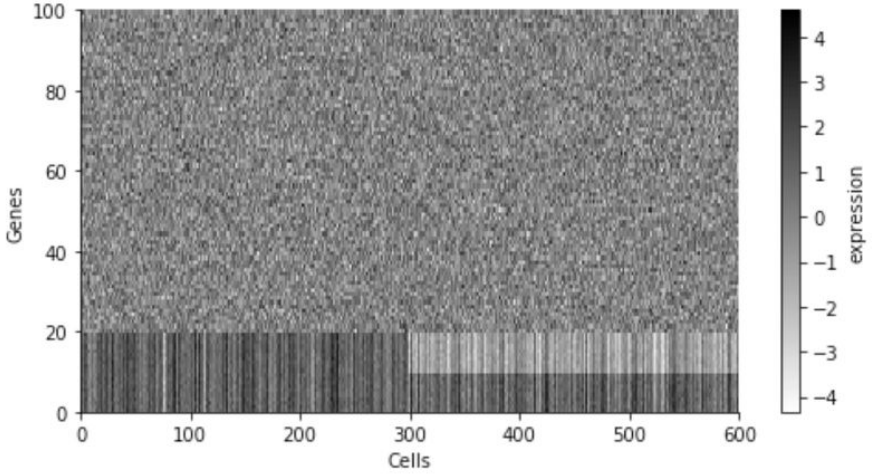
Dimensionality reduction by PCA

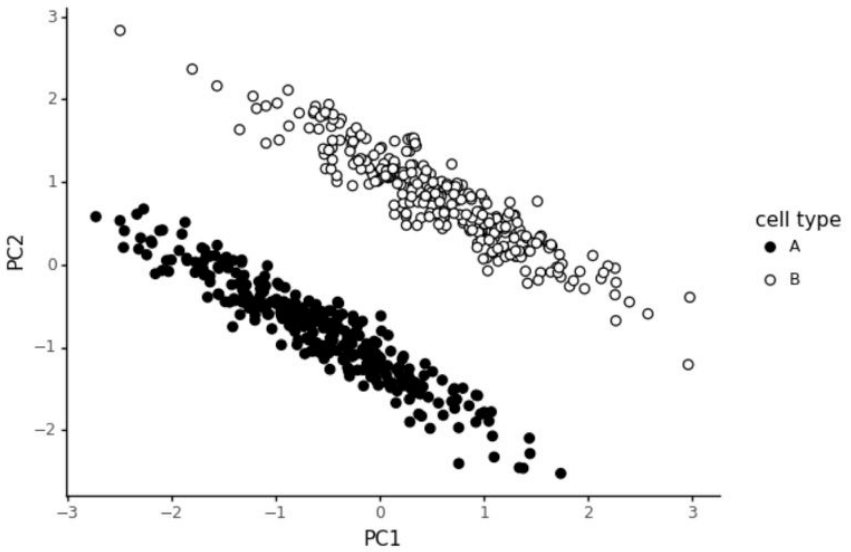




http://www.nxn.se/valent/2017/6/12/how-to-read-pca-plots

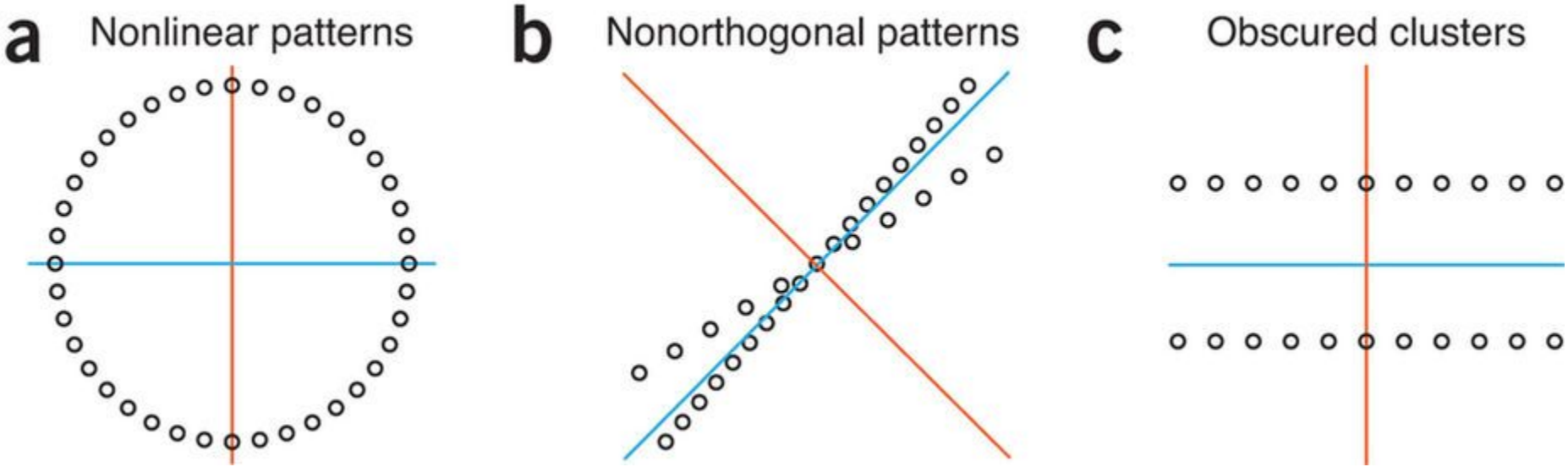
Dimensionality reduction by PCA





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

Limitations of PCA, alternatives



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

Imputing missing values

scRNA-seq data has:

- **a high frequency of zero values**, often referred to as dropouts, and
- **high levels of noise** due to the low amounts of input RNA obtained from individual cells.

Zero values in scRNA-seq may arise due to:

- low experimental sensitivity, e.g. sequencing sampling noise, technical dropouts during library preparation, or
- biologically the gene is not expressed in the particular cell.

Imputing missing values

Zero values in scRNA-seq may arise due to:

- low experimental sensitivity, e.g. sequencing sampling noise, technical dropouts during library preparation, or
- because biologically the gene is not expressed in the particular cell.

Imputation is a common approach when dealing with sparse genomics data: predict missing values from the rest of the measured values.

One challenge when imputing expression values is to **distinguish true zeros from missing values**.

scRNA-seq data imputation methods use information internal to the dataset to be imputed.

- Some degree of circularity → false positive results when identifying marker genes, gene-gene correlations, or testing differential expression.

Imputing missing values

Many imputation methods:

- **SAVER, DrImpute & scImpute**: use models of the expected gene expression distribution to distinguish true biological zeros from zeros originating from technical noise.
    - Assume homogenous cell populations → identify clusters of similar cells to which an appropriate mixture model is fitted.
    - Values falling above a given probability threshold to originate from technical effects are subsequently imputed.
- **MAGIC & knn-smooth**: perform data smoothing.
    - Infer values of missing data + reduces noise present in observed values (using information from neighbouring data points).
    - Use each cell's k nearest neighbours either through the application of diffusion models or weighted sums respectively.

Imputing missing values

Many imputation methods:

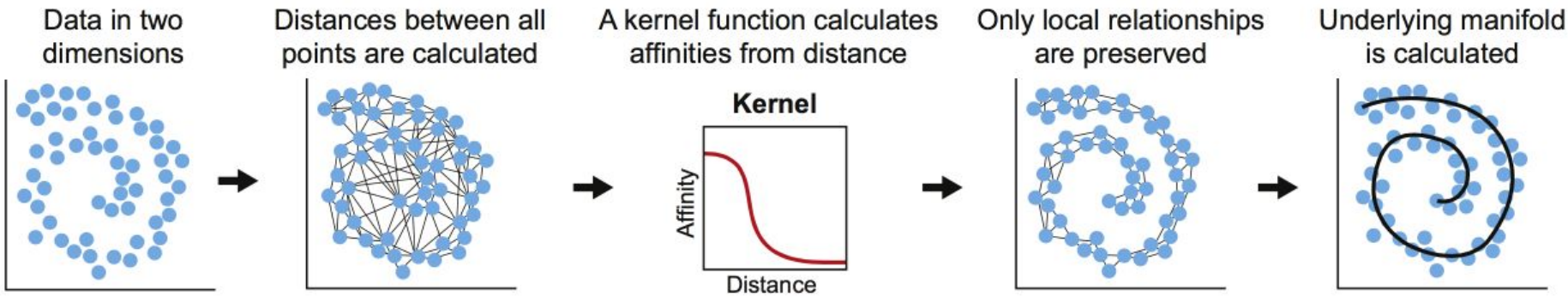
	Designed for single cell	Local or global	Beyesian method	Need other information	Imputation strategy
LLSimpute	N	local	N	No. of nearest genes	1
Low-rank	N	global	N	error tolerance δ	2
BISCUIT	Y	global	Y	dispersion parameter	1 and 2
scUnif	Y	global	Y	cell labels	2
MAGIC	Y	global	N	diffusion time	2
scImpute	Y	local	N	dropout rate cutoff	2
DrImpute	Y	local	N	cluster numbers	2
SAVER	Y	global	Y	size factor	1

Strategy 1 represents imputing dropout based on co-expressed or similar genes, while strategy 2 denotes imputing dropout by borrowing information from similar cells.

Zhang & Zhang (2017) bioRxiv

Imputing missing values

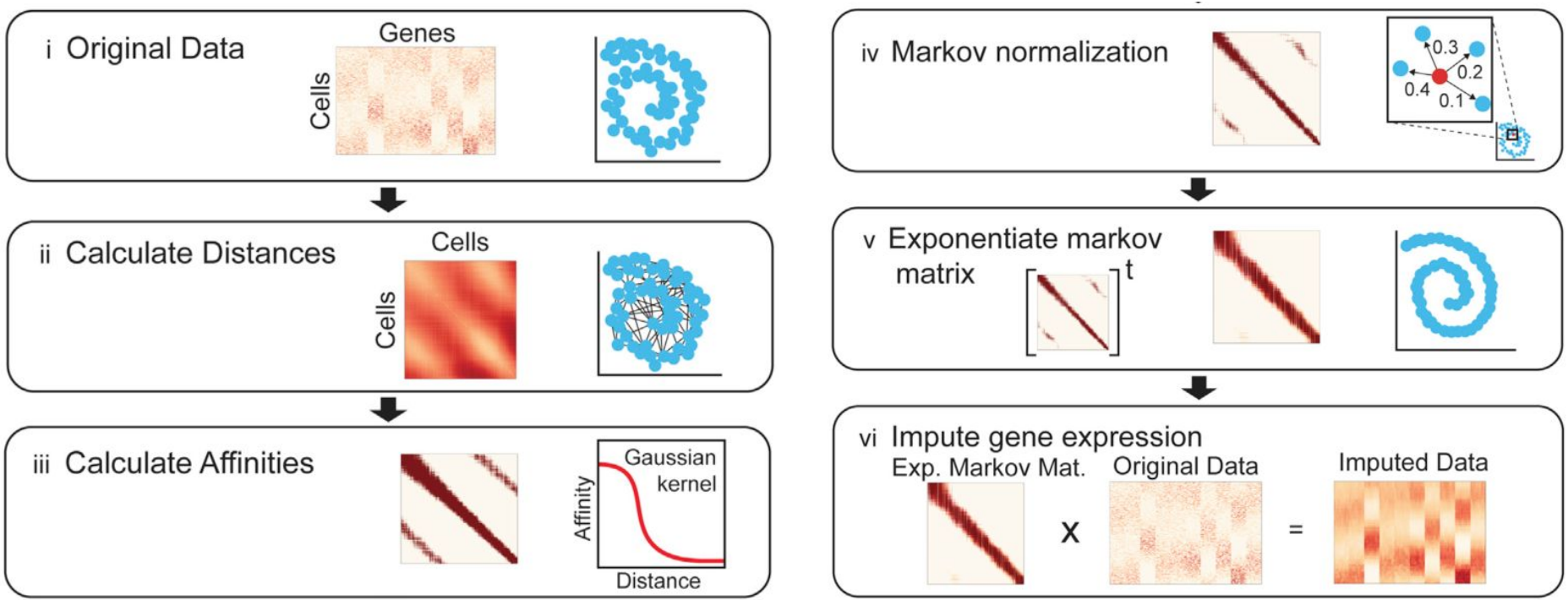
Manifold learning



Moon (2018) Curr. Opin. Sys. Biol.

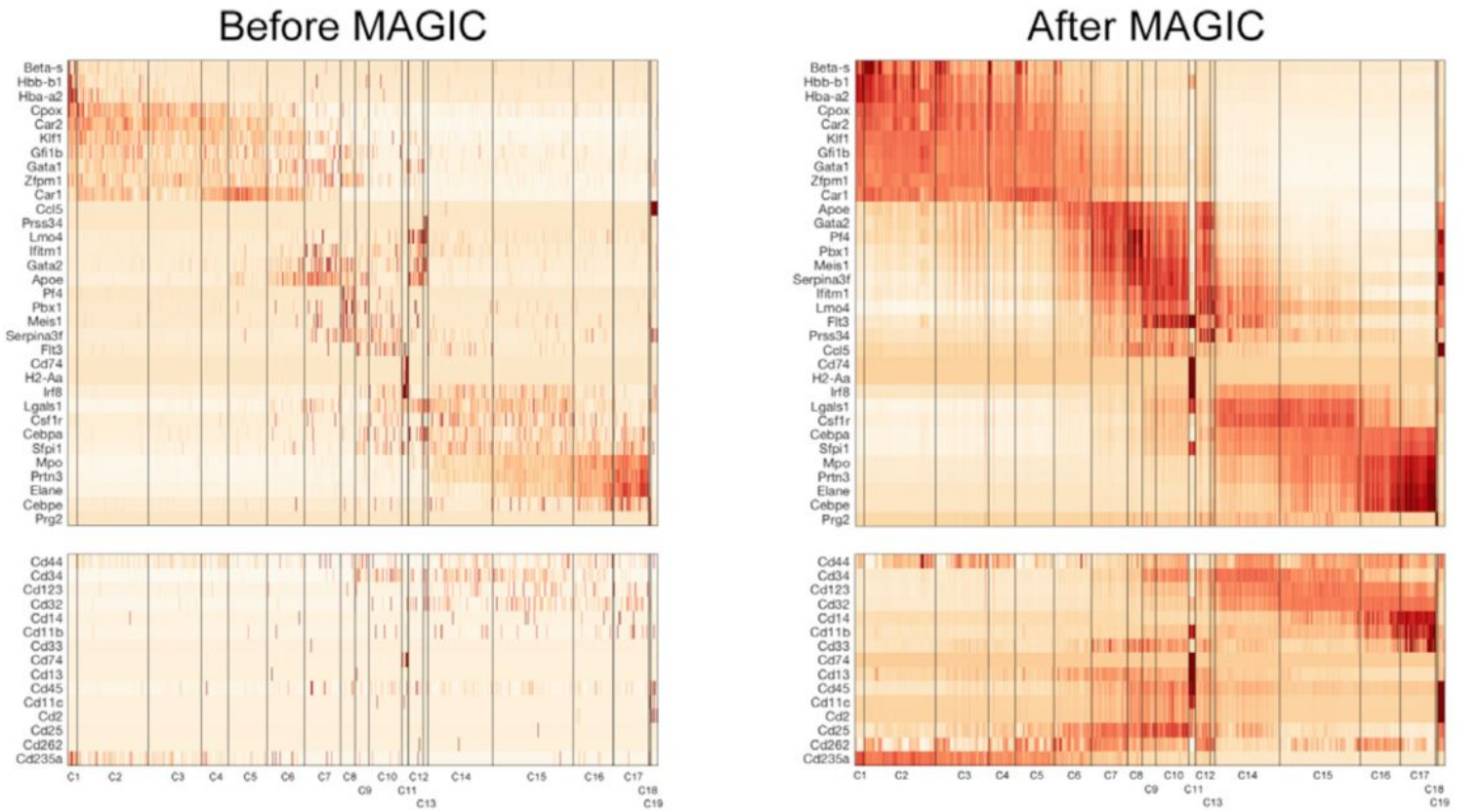
Imputing missing values

MAGIC



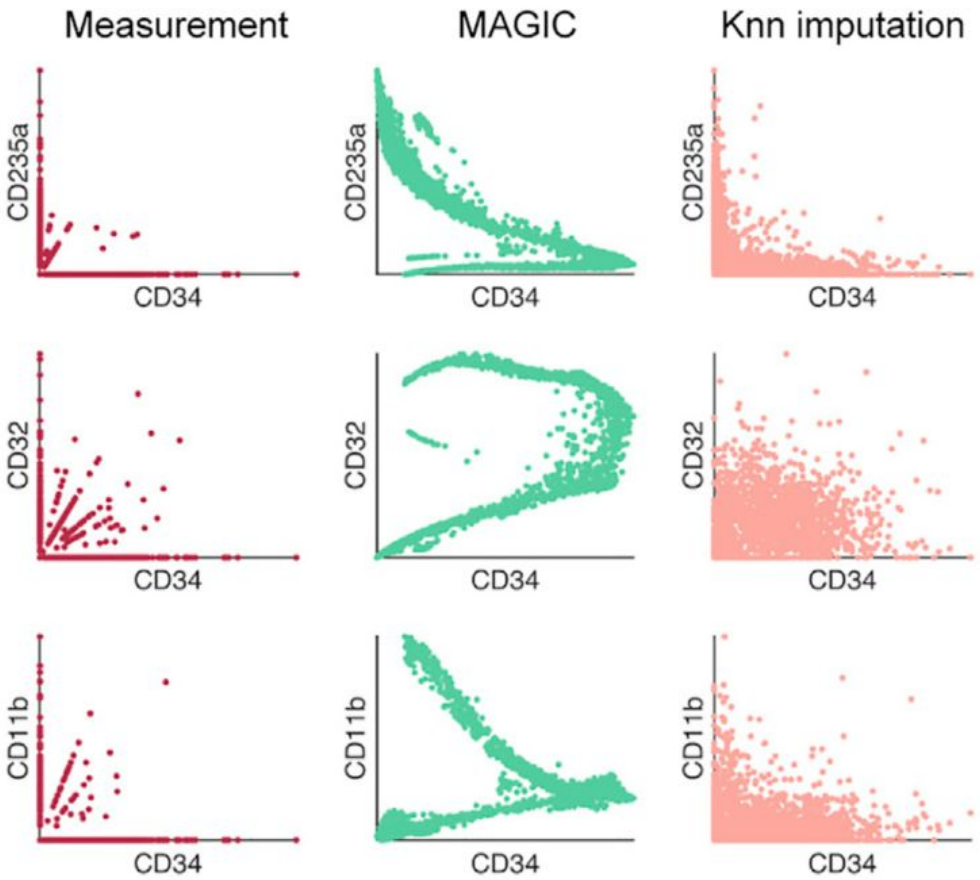
Vandijk (2018) Cell

Imputing missing values



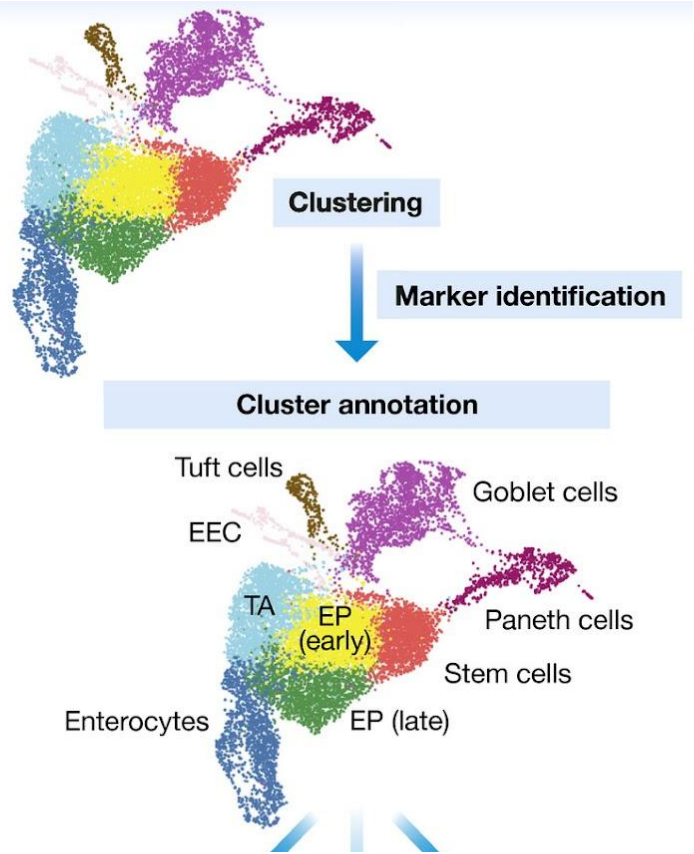
Vandijk (2018) Cell

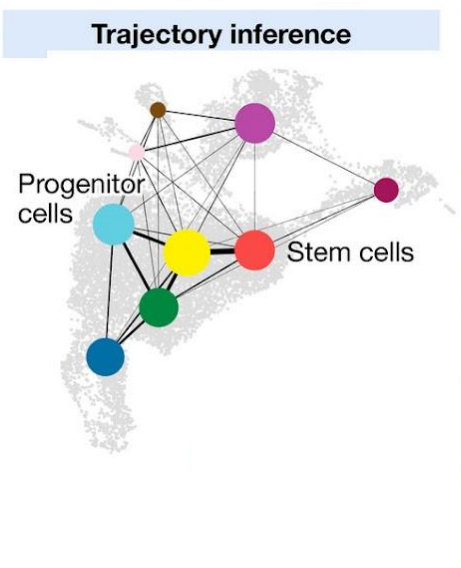
Imputing missing values

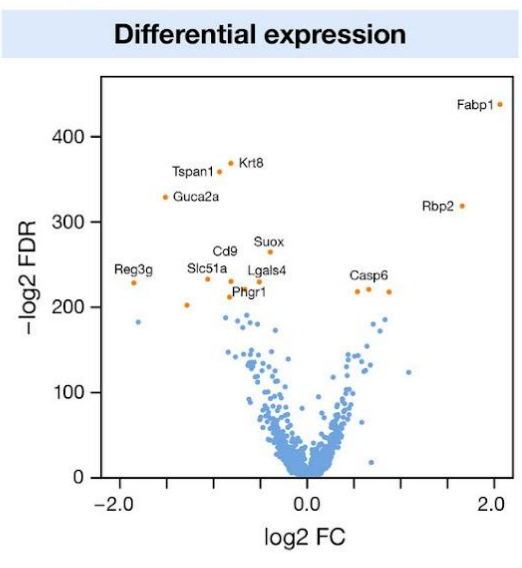


Vandijk (2018) Cell

Clustering scRNA-seq data to identify cell types/states

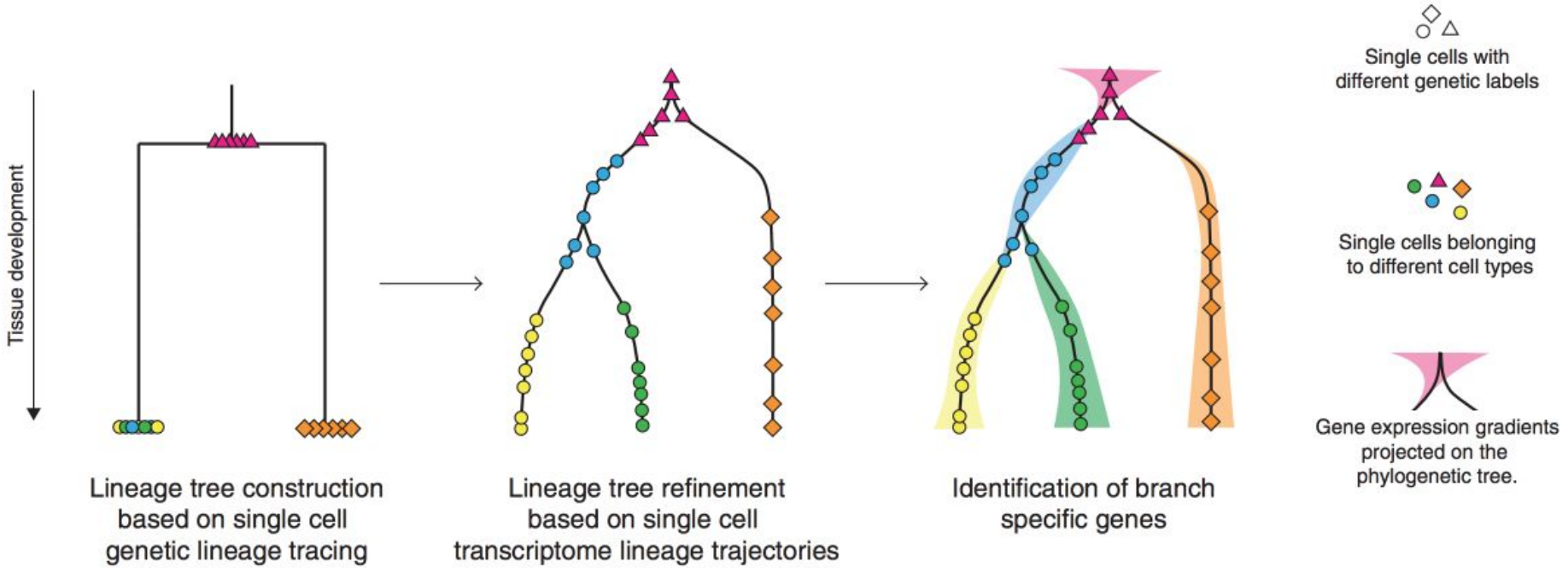






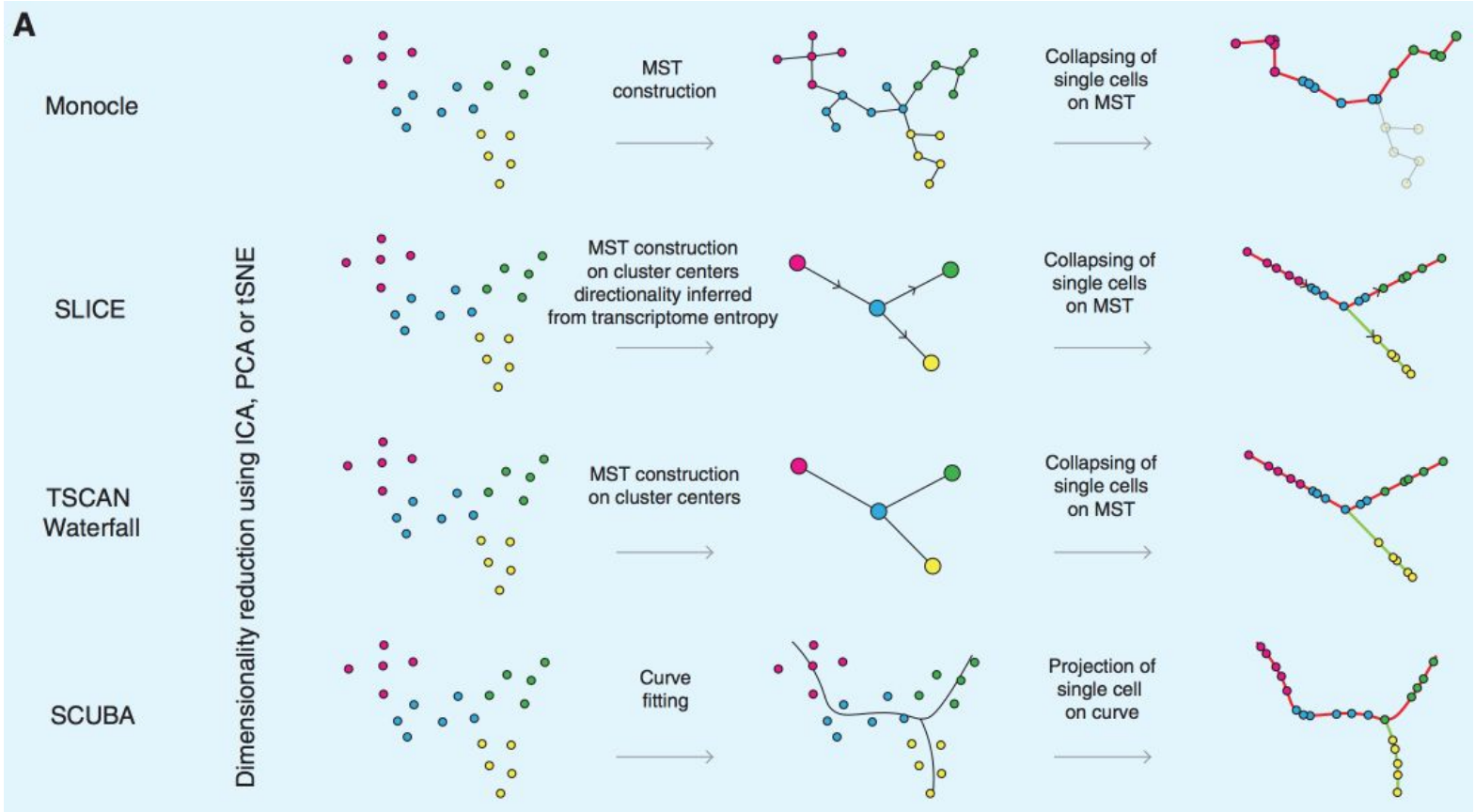
Luecken & Theis (2019) Mol. Sys. Biol.

Lineage tracing



Vandijk (2018) Cell

Lineage tracing



Kester (2018) Cell Stem Cell