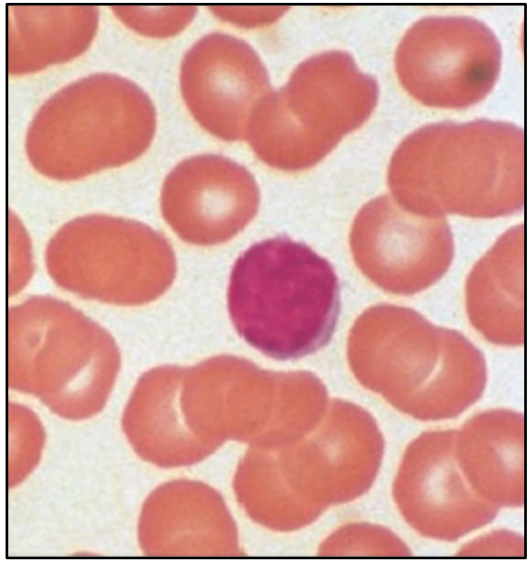
Identification of cell subsets in scRNA-Seq data

HCA Latin America

Single-Cell RNA-Seq Computational Workshop

October 24, 2023



BROAD INSTITUTE

KLARMAN CELL OBSERVATORY AT BROAD INSTITUTE

HUMAN CELL ATLAS

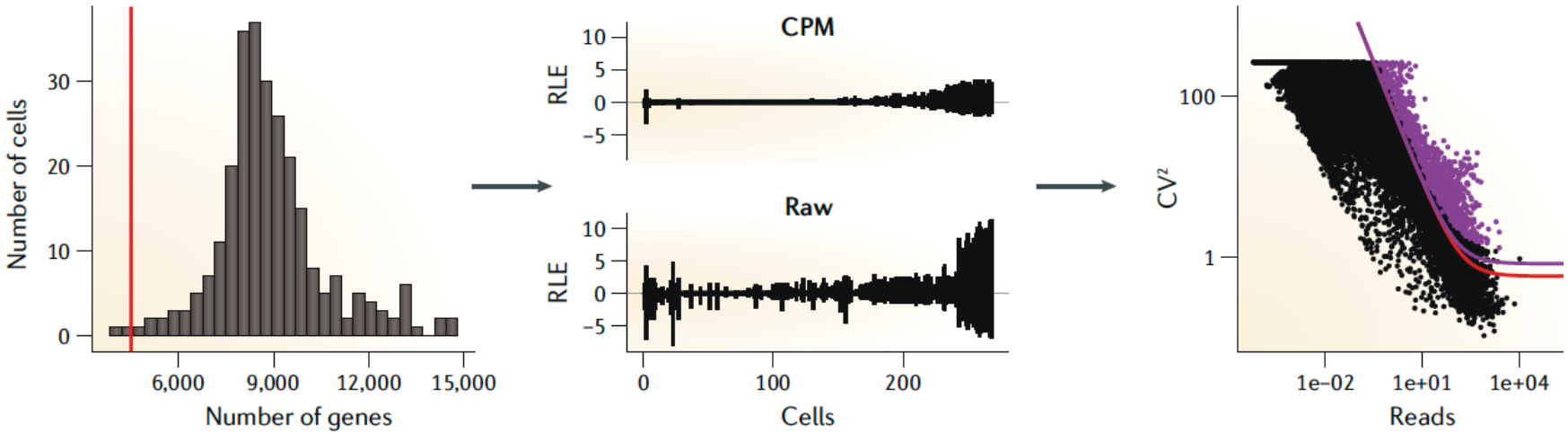
humancellatlas.org

Determining cell type, state, and function

Quality control

Normalization

Feature selection



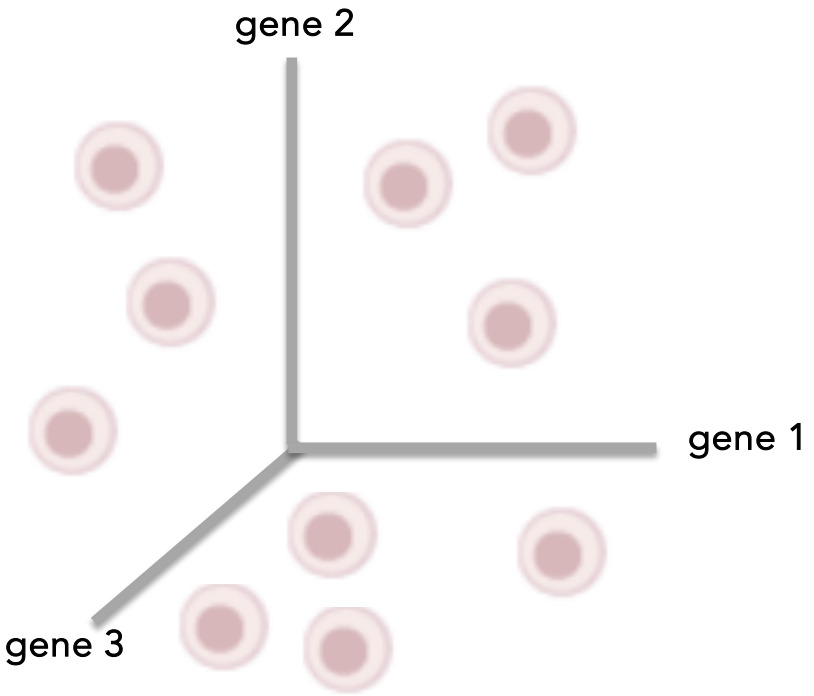
Dimensional reduction

Cell-cell distances

Unsupervised clustering

Kiselev C. at *al.* (2019) *Nature Reviews Genetics*. 20: 273-282.

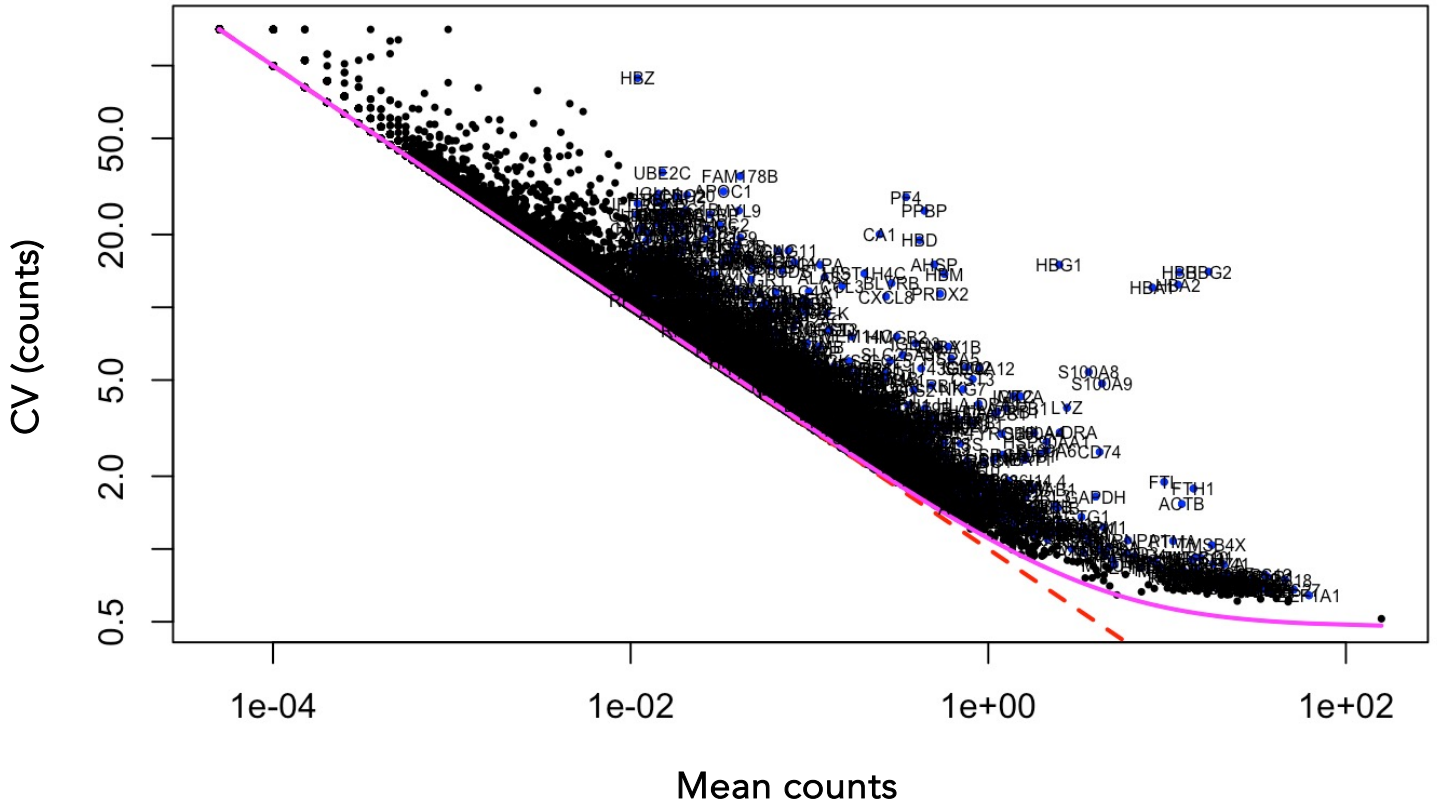
Single-cell RNA-Seq analysis: feature selection



Cells are in 20,000 dimensional space (one dimension for each gene), but hard to visualize!

- Genes with highly variable expression contain biological signal related to cell states.
- Many genes are lowly detected or noisy measurements.
- We typically select 2,000-5,000 variable genes.

Single-cell RNA-Seq analysis: feature selection



Find genes (features) that are outliers in a plot of mean of gene expression vs variance of gene expression

Single-cell RNA-Seq analysis: Dimensionality reduction

Cells are in ~20,000 dimensional space

- many genes are lowly detected / noisy measurements
- genes are not independent of one another! rather they operate in coregulatory modules
- curse of dimensionality

Principle component analysis moves us from describing cells with 20,000 gene expression values to 10-100 principal component scores

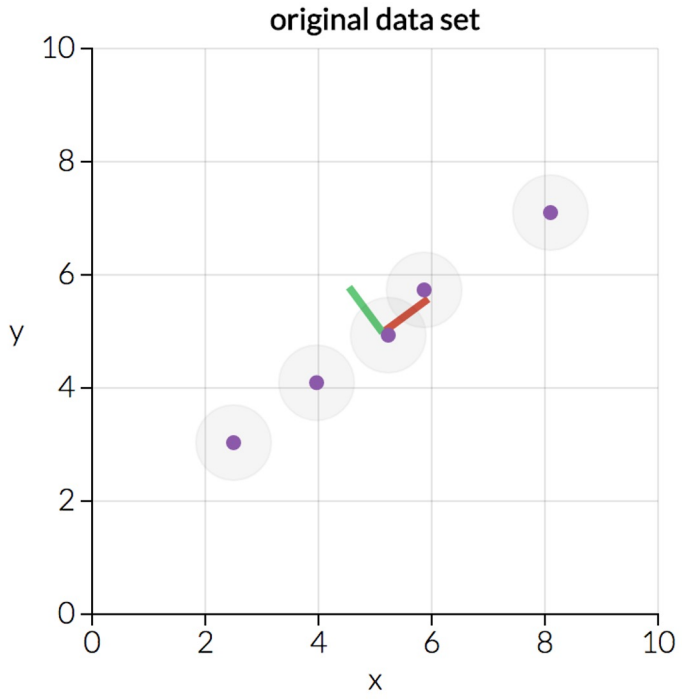
Single-cell RNA-Seq analysis: Dimensionality reduction

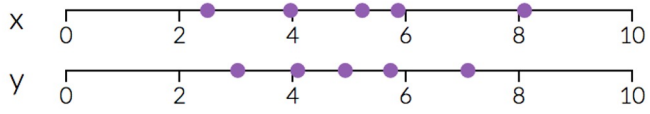
One approach to simplification is to **assume that the data of interest lies within lower-dimensional space**. If the data of interest is of low enough dimension, the data can be visualised in the low-dimensional space.

Common Techniques

- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Multidimensional Scaling (MDS)
- Non-negative Matrix Factorization (NMF)
- Probabilistic Modeling (e.g. Latent Dirichlet Allocation - LDA)

Single-cell RNA-Seq analysis: Dimensionality reduction



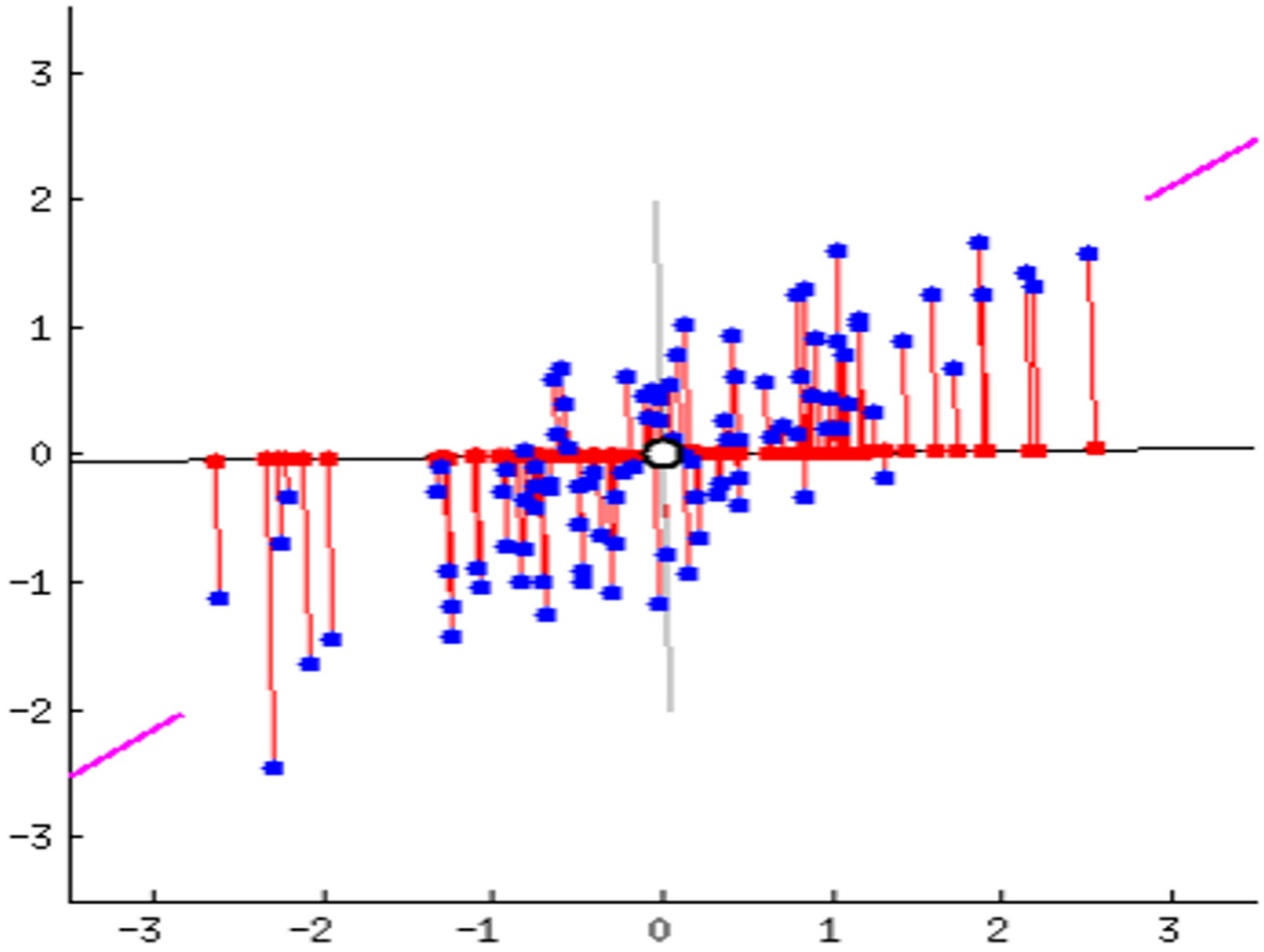


http://setosa.io/ev/principal-component-analysis/

Single-cell RNA-Seq analysis: Dimensionality reduction

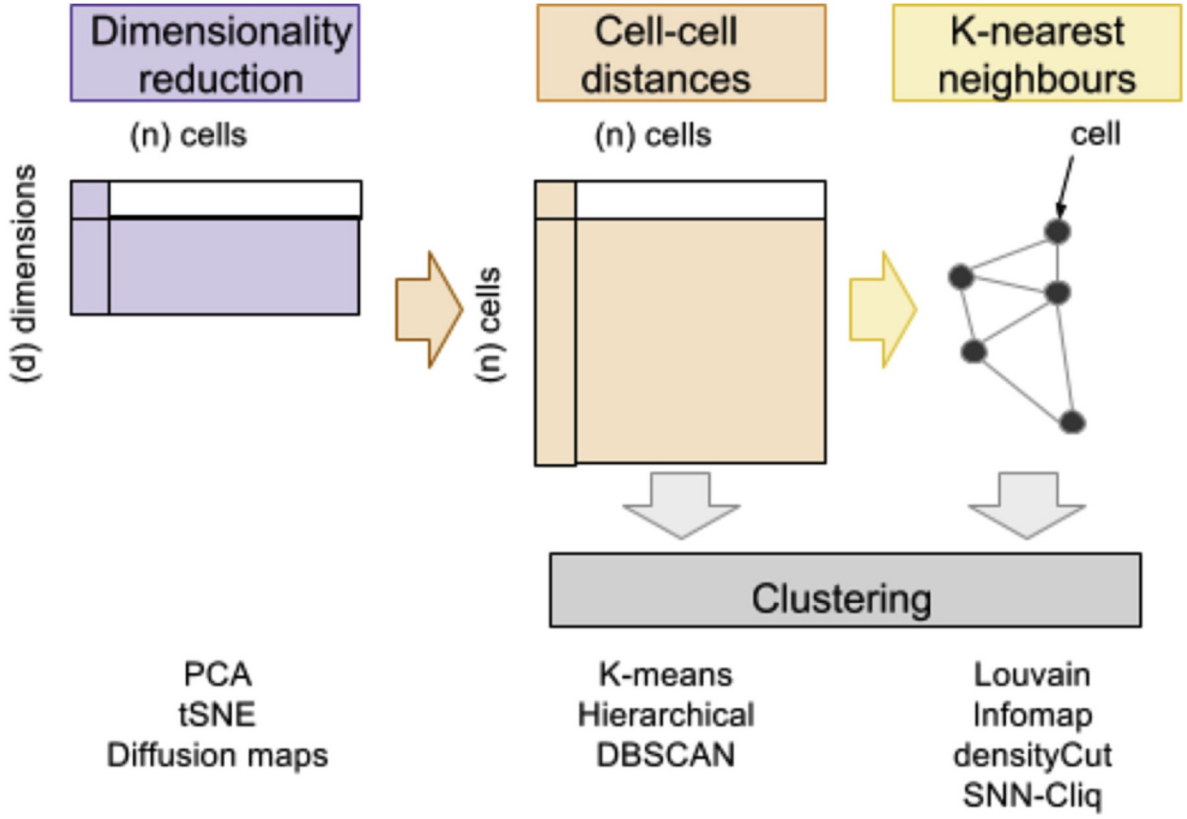
- PCA is a dimensionality reduction method that transforms a **set of features** into a set of **linearly uncorrelated variables** called principal components

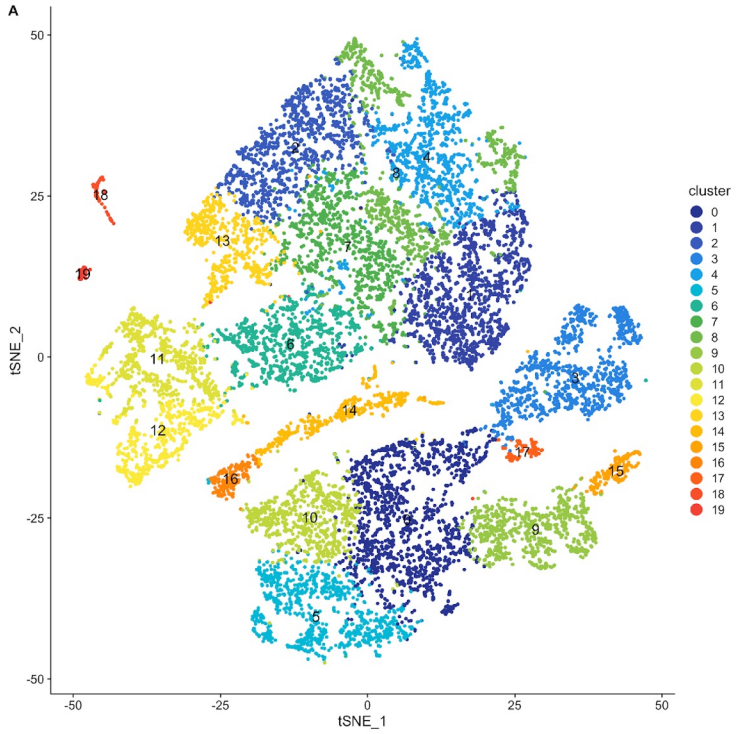
- The first principal component contains the most variance, and each component after contains as much variance while still being orthogonal to other components



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

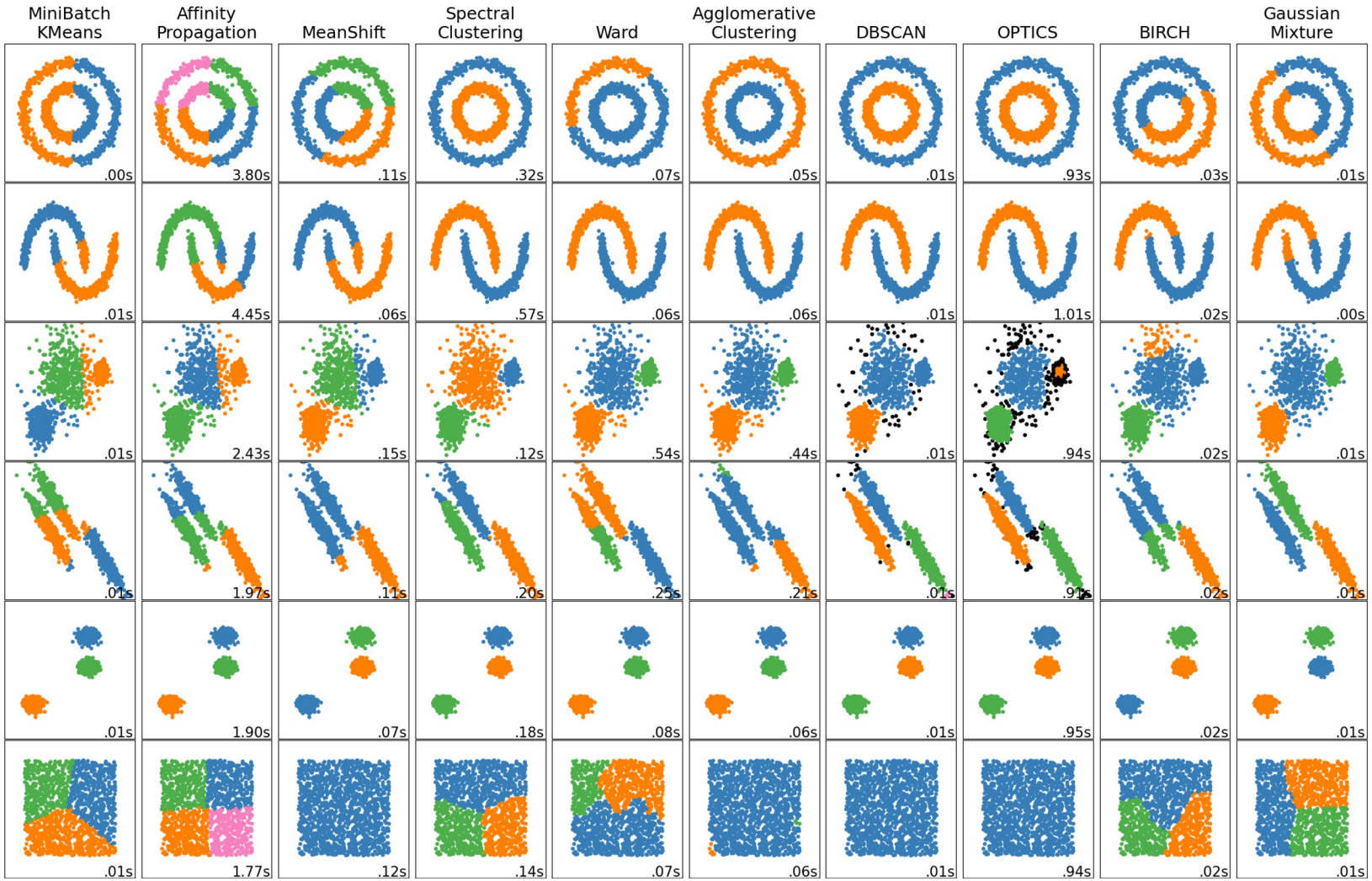
Single-cell RNA-Seq analysis: Clustering principal components





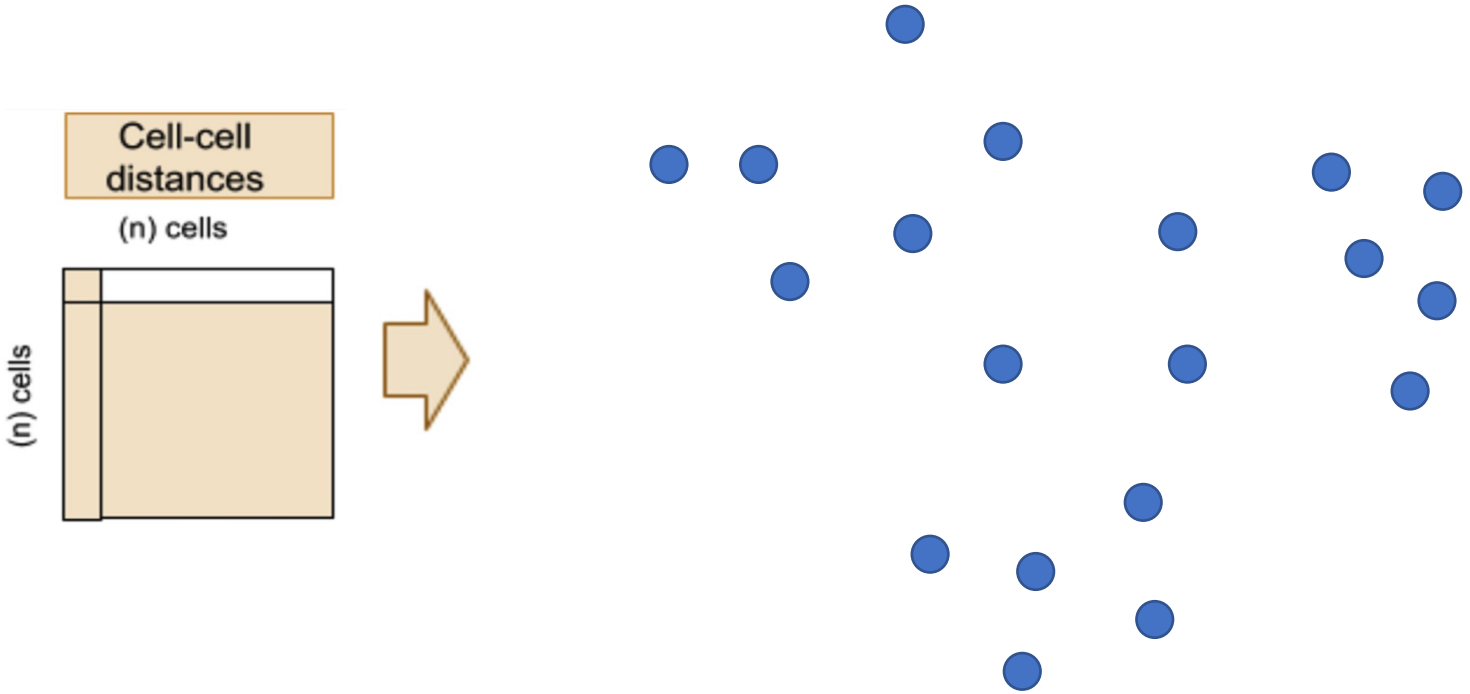
Andrews, Tallulah S. at *al.* (2017) *Molecular aspects of medicine.*

Single-cell RNA-Seq analysis: Clustering principal components

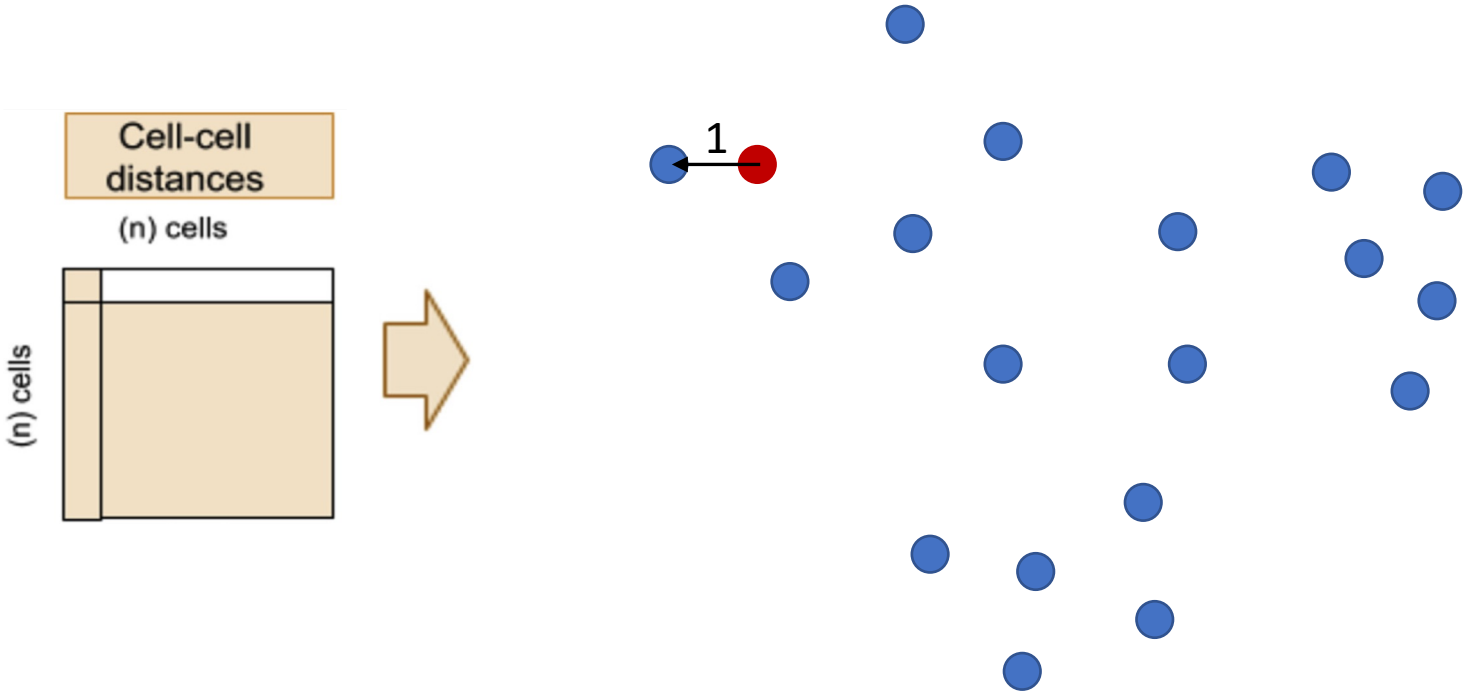


https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

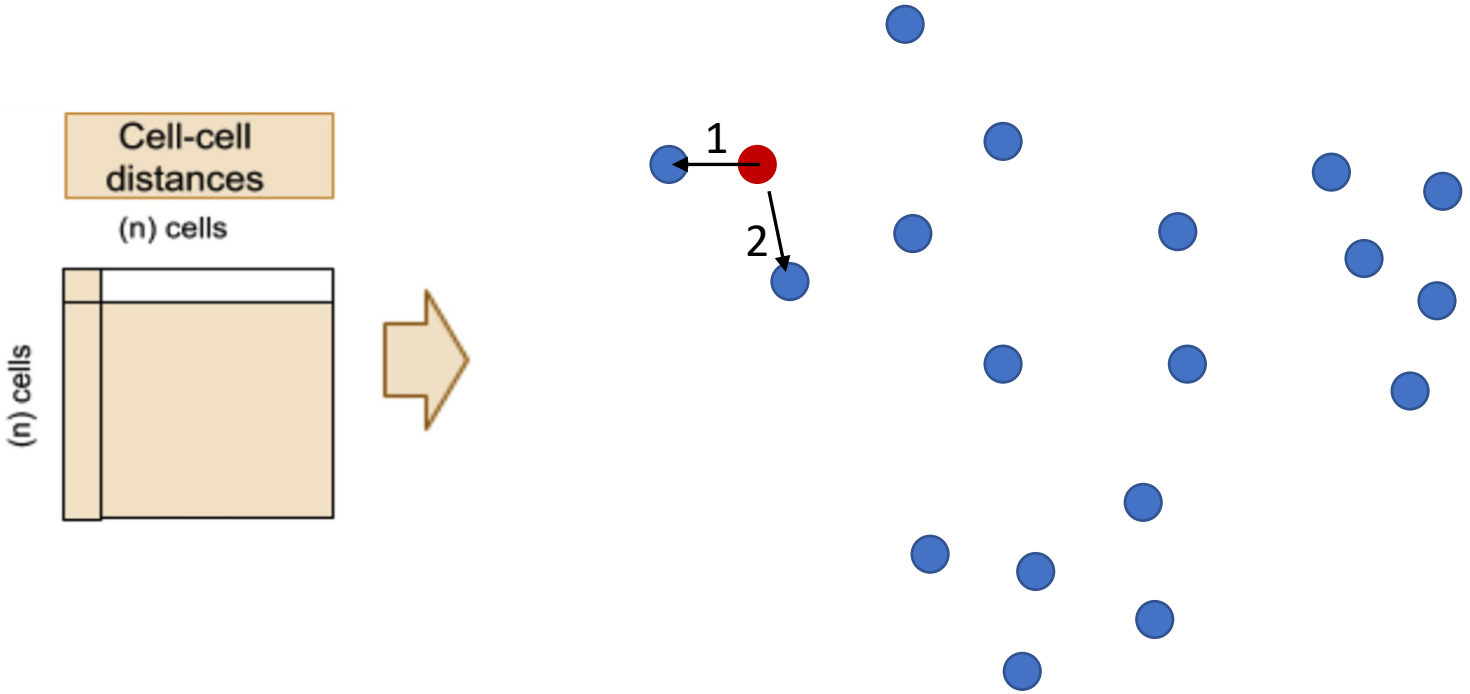
Single-cell RNA-Seq analysis: Clustering principal components



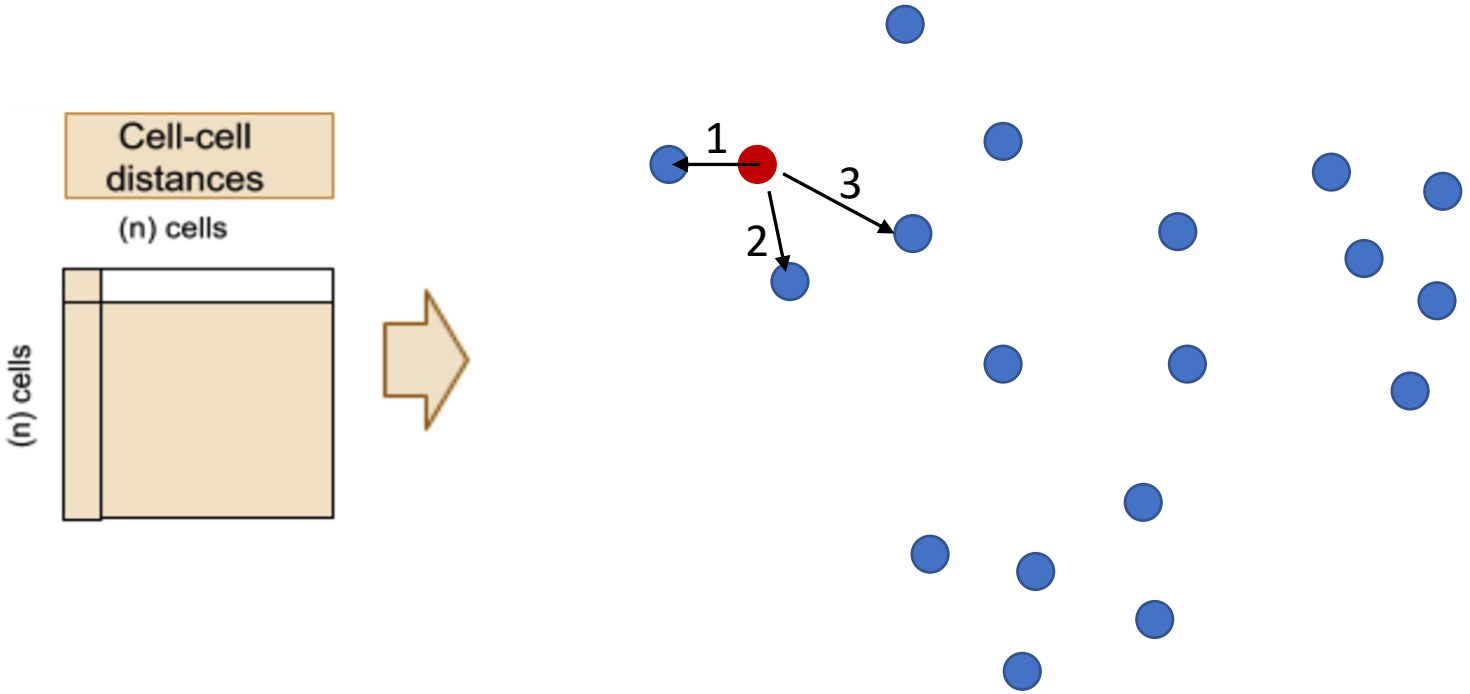
Single-cell RNA-Seq analysis: Clustering principal components



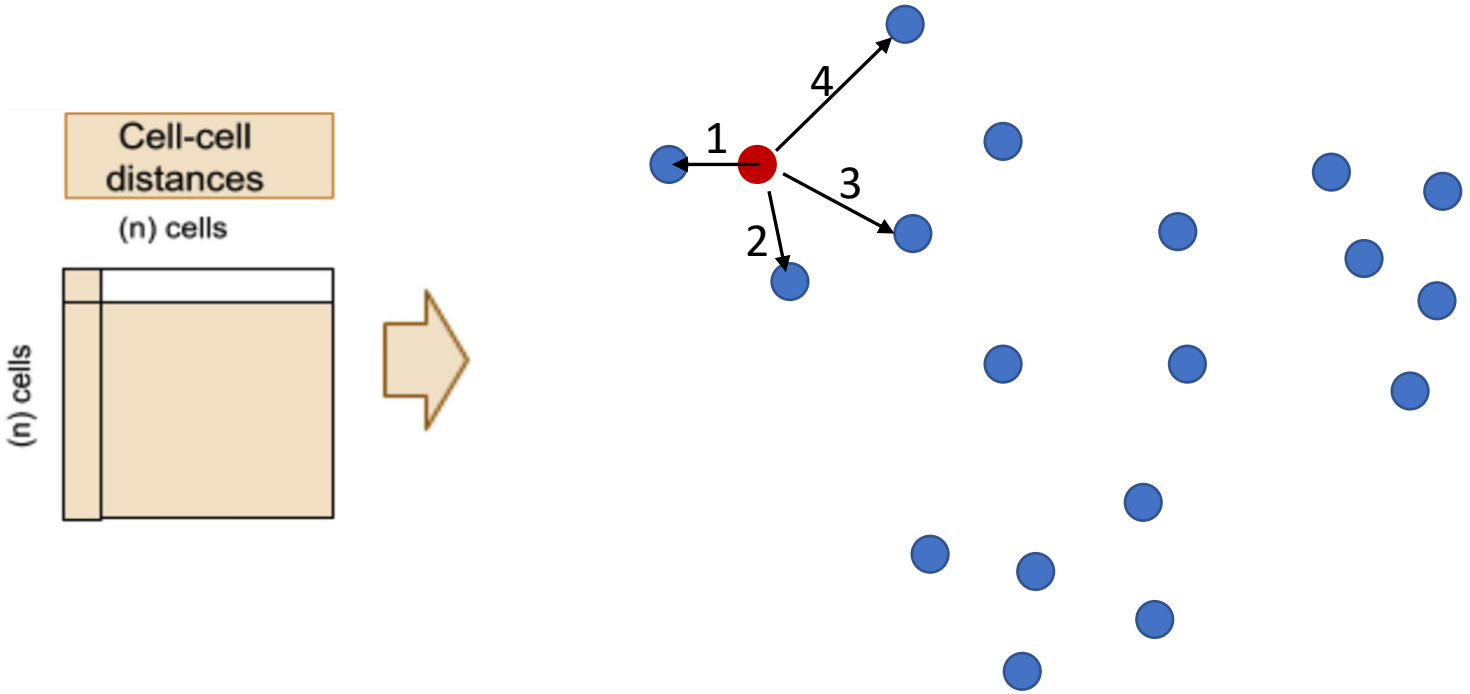
Single-cell RNA-Seq analysis: Clustering principal components



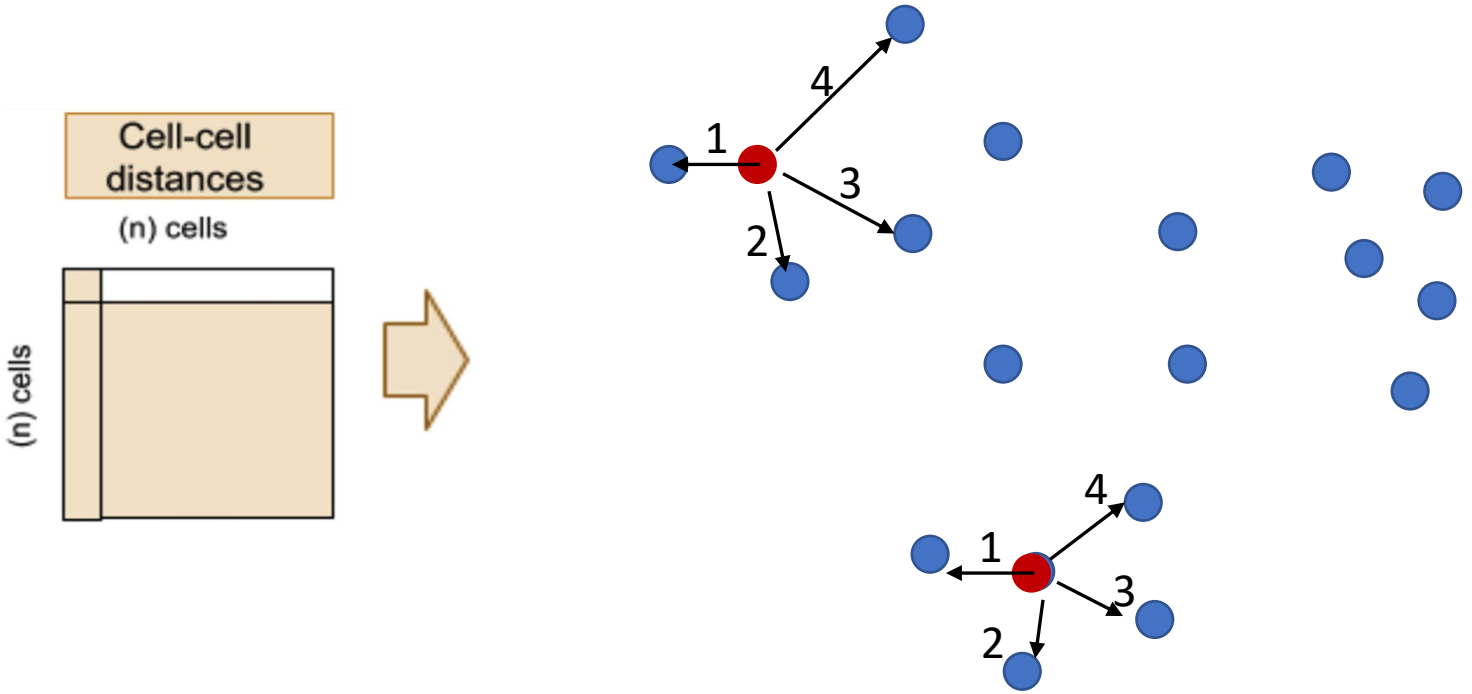
Single-cell RNA-Seq analysis: Clustering principal components



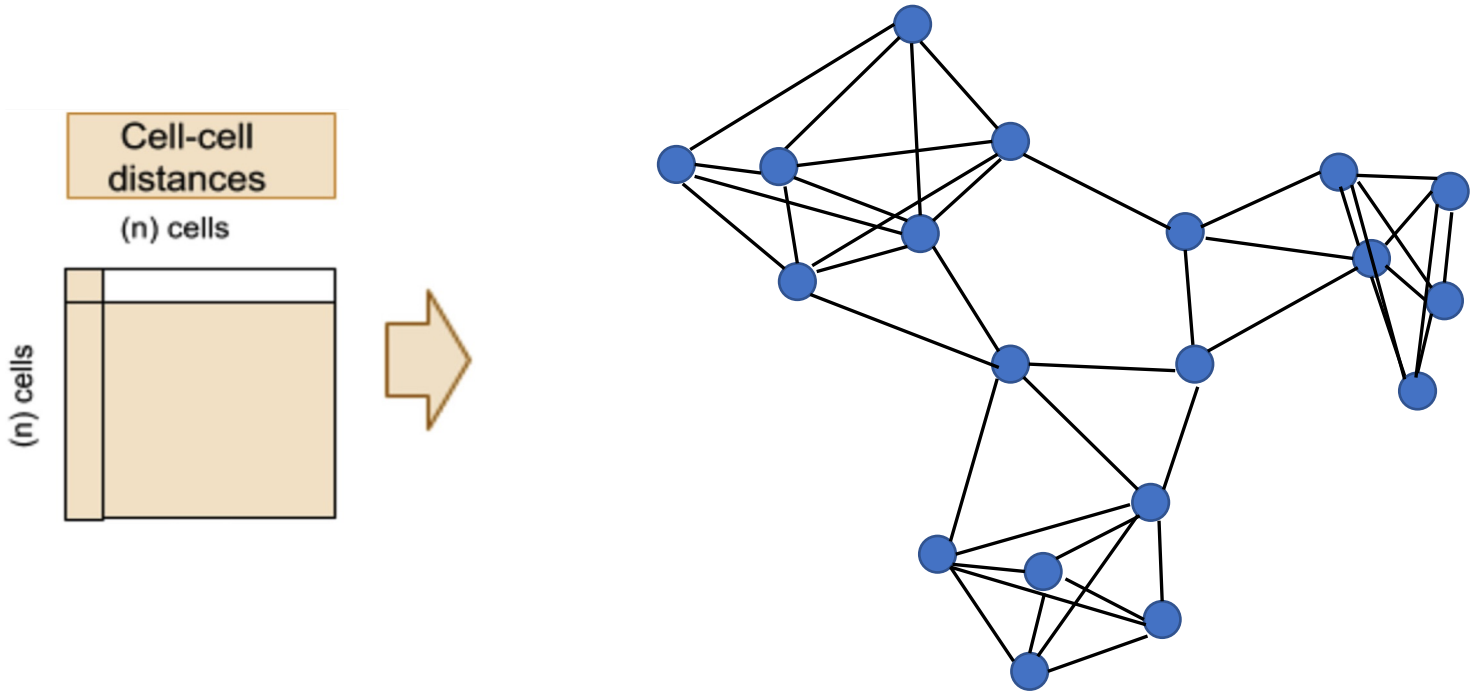
Single-cell RNA-Seq analysis: Clustering principal components



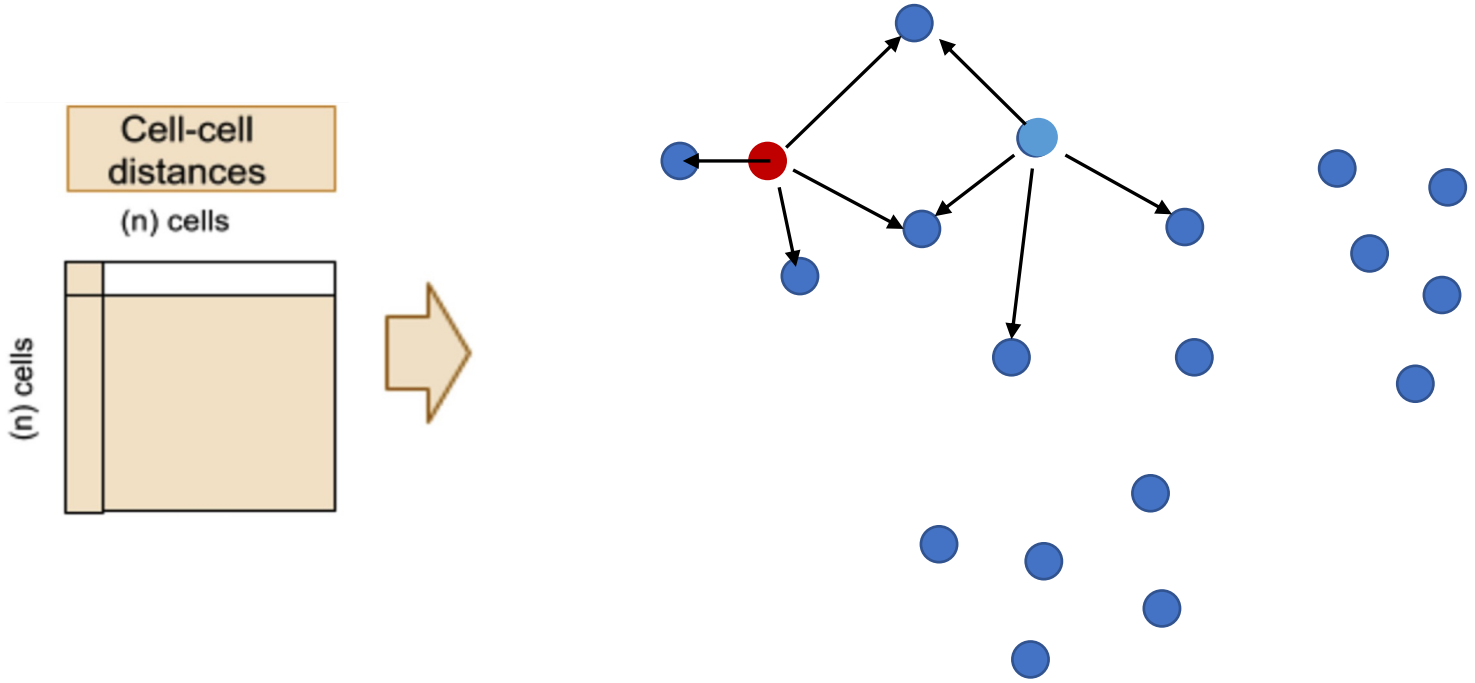
Single-cell RNA-Seq analysis: Clustering principal components



Single-cell RNA-Seq analysis: Clustering principal components

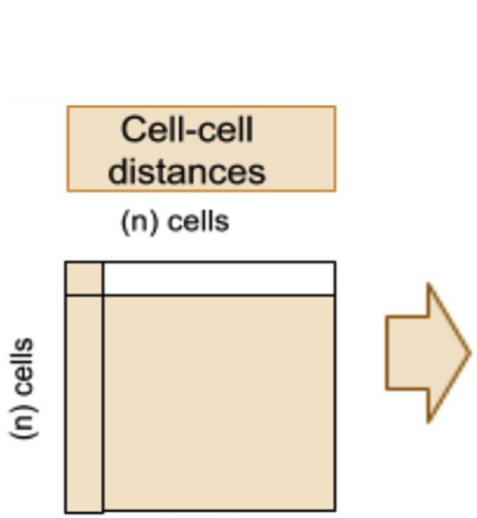


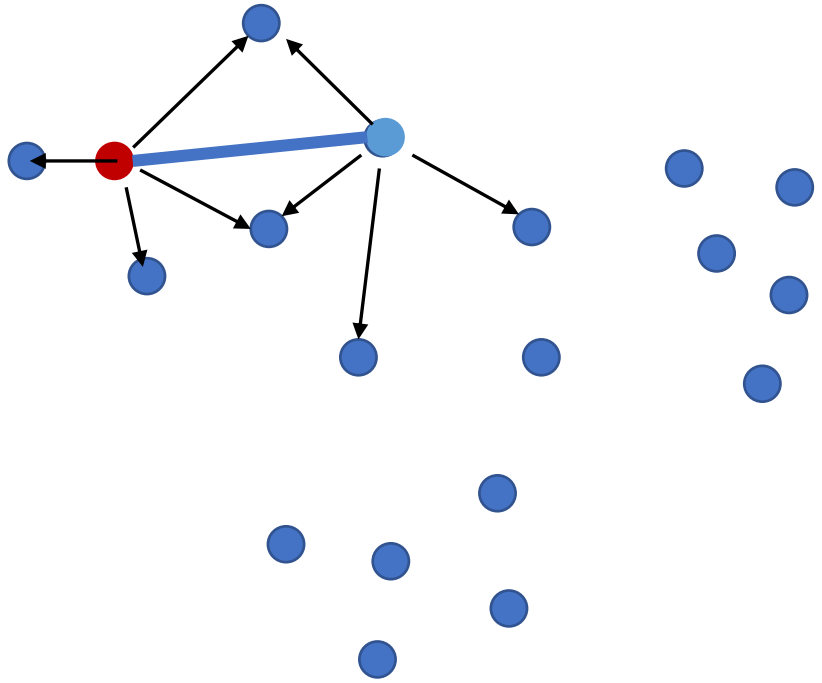
Single-cell RNA-Seq analysis: Clustering principal components



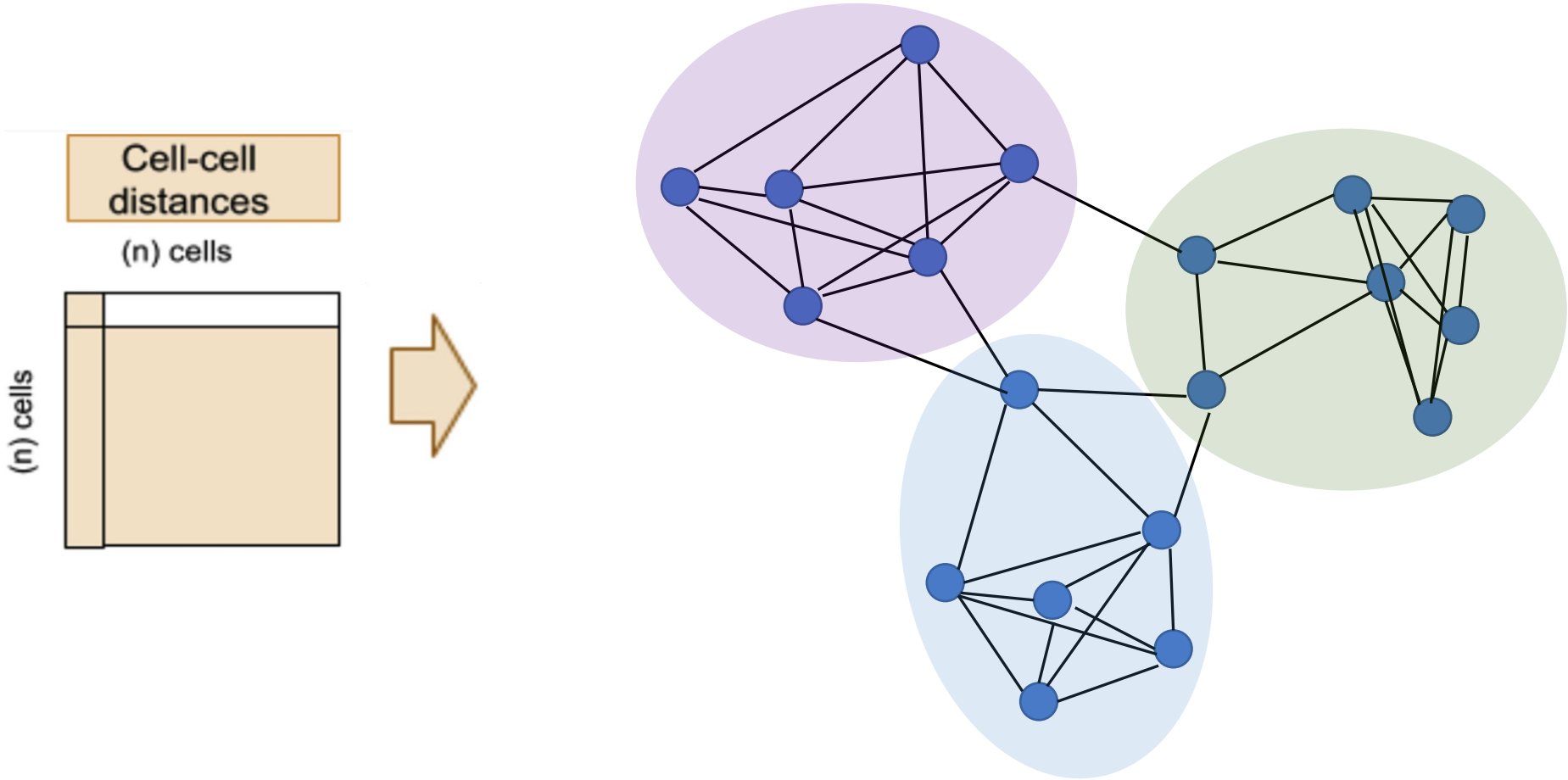
Single-cell RNA-Seq analysis: Clustering principal components

Two cells are connected by an edge if any of their nearest neighbors are shared.



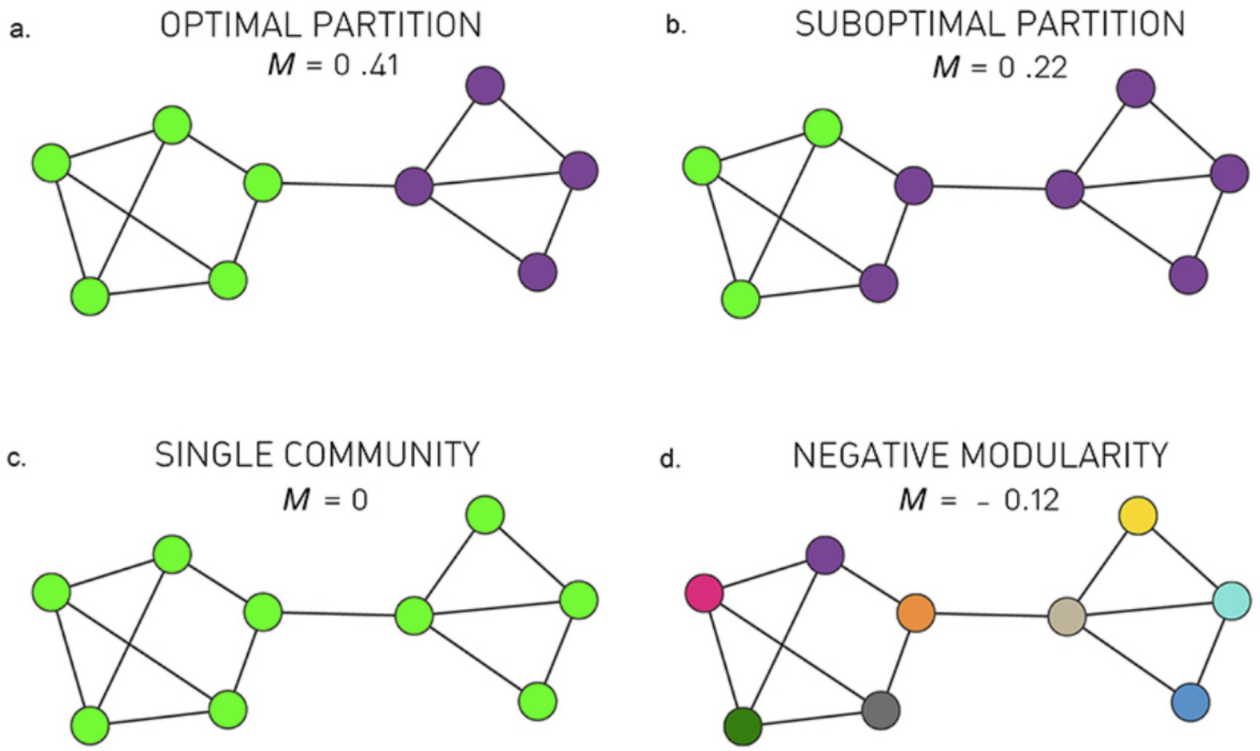


Single-cell RNA-Seq analysis: Clustering principal components

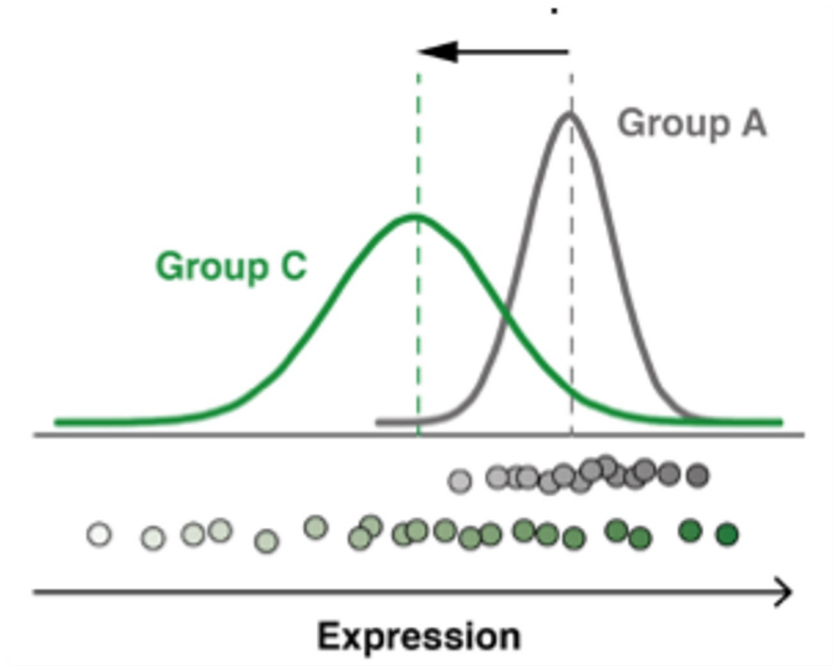


Single-cell RNA-Seq analysis: Clustering principal components

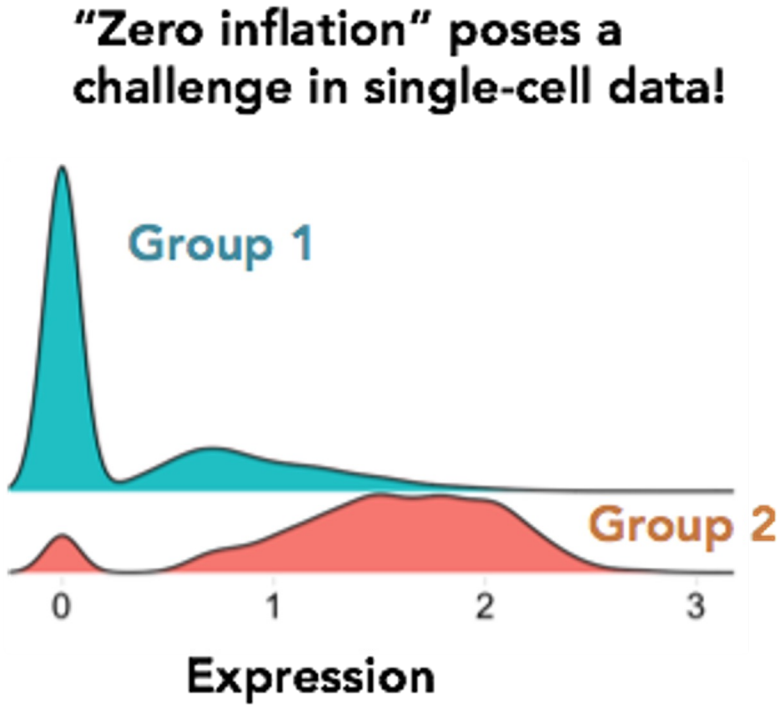
- Graph-based clustering is based on **community detection**.
- Many different algorithms for community detection:
  - Louvain (heuristic), Infomap, Walktrap
- Most of them are based on modularity maximization



Single-cell RNA-Seq analysis: Differentially expressed genes



Bulk



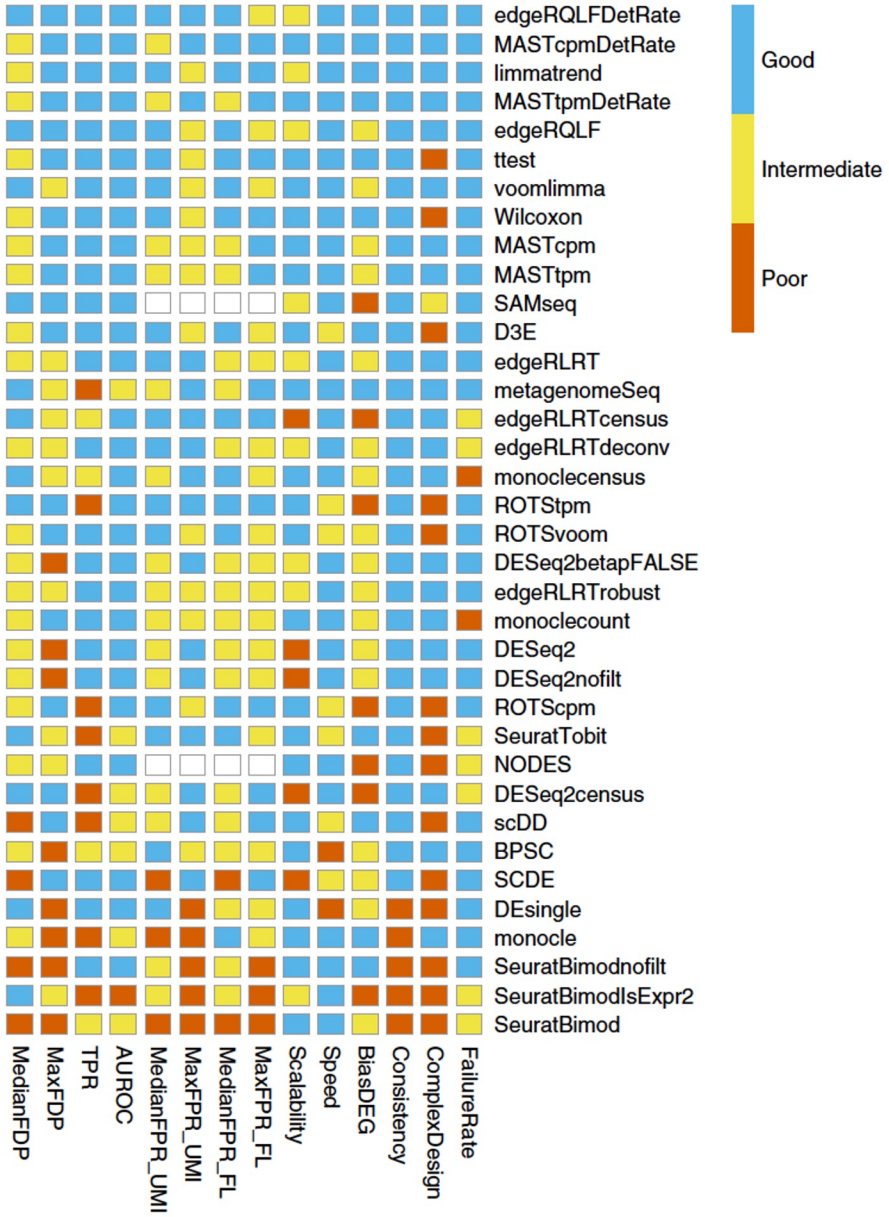
Single cell

Single-cell RNA-Seq analysis: Differentially expressed genes

To annotate cell subsets, we typically perform **pairwise comparisons** of gene expression, **between pairs of cell clusters**, using some of the following tests:

- "wilcox" : Wilcoxon rank sum test (default)
- "t" : Student's t-test
- "poisson" : Likelihood ratio test assuming an underlying poisson distribution. Use only for UMI-based datasets
- "negbinom" : Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets
- Others…

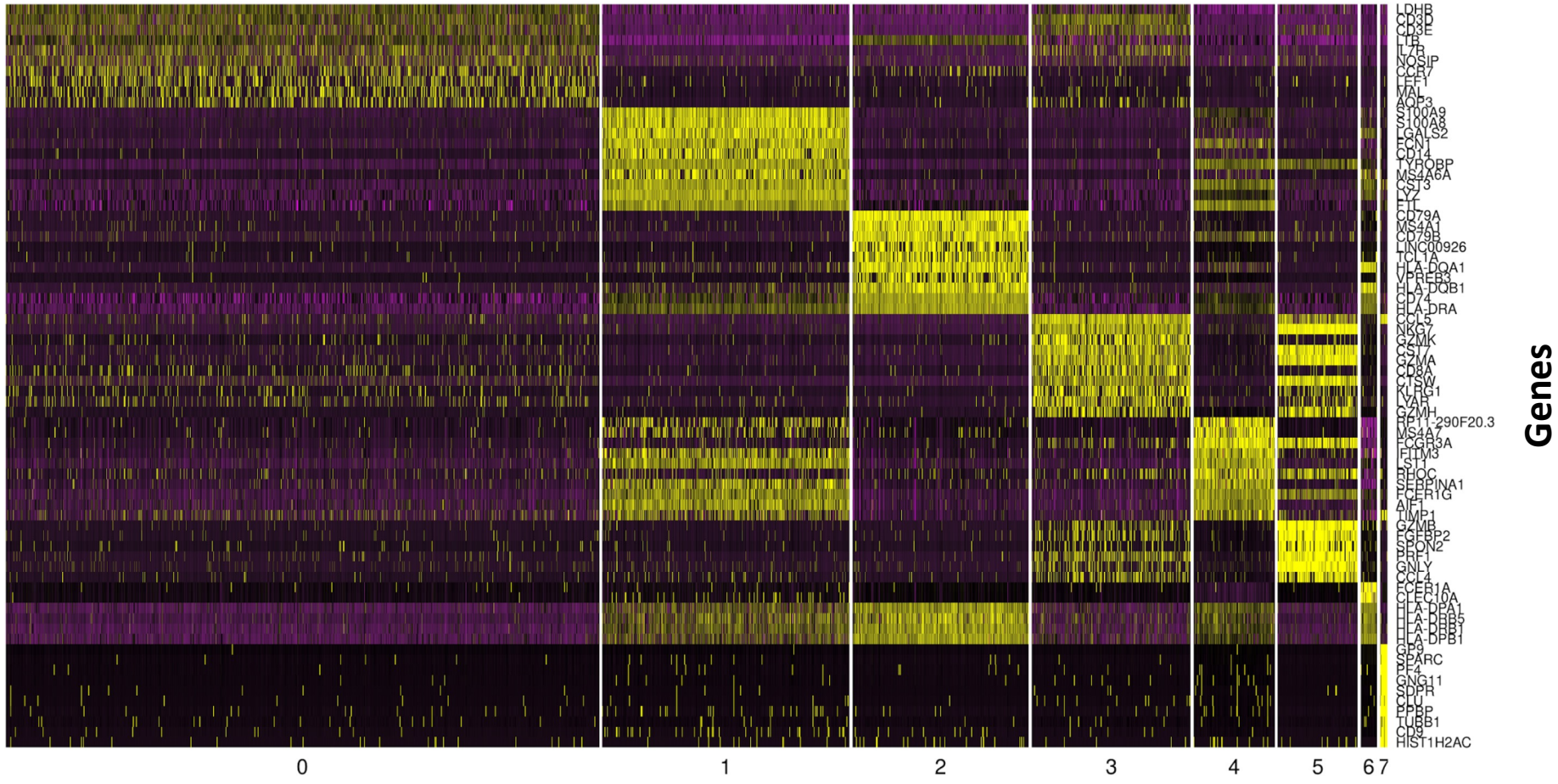
Single-cell RNA-Seq analysis: Differentially expressed genes



There are many ways to test for differential expression!

Soneson C. at *al.* (2018) *Nature Methods*. 15: 255-261.

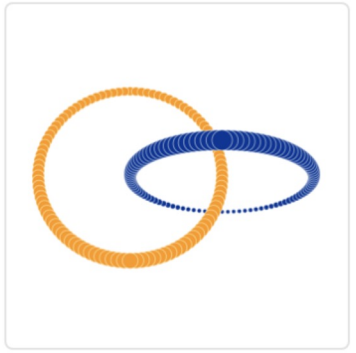
Single-cell RNA-Seq analysis: Differentially expressed genes



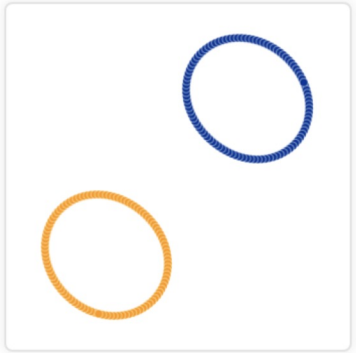
Single-cell RNA-Seq analysis: Visualizing cells in lower dimensions

High dimension manifold

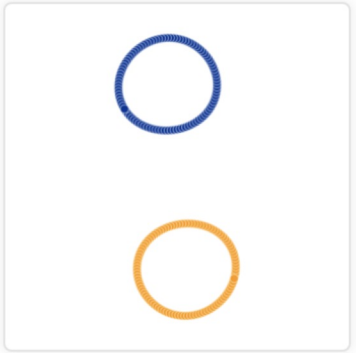
Low dimension representation



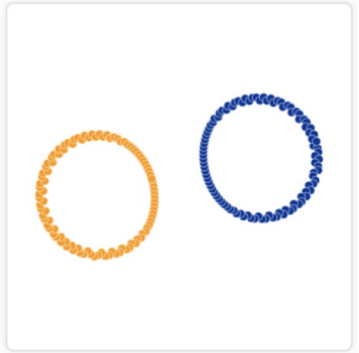
Original



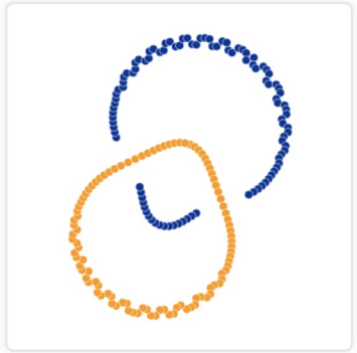
Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000

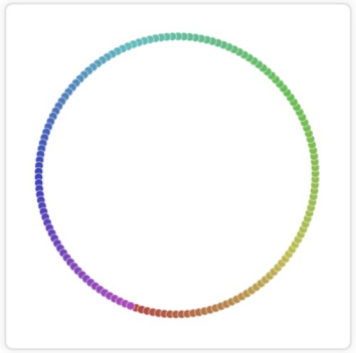


Perplexity: 30
Step: 5,000



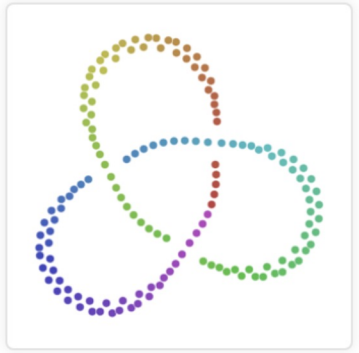
Perplexity: 50
Step: 5,000

Original

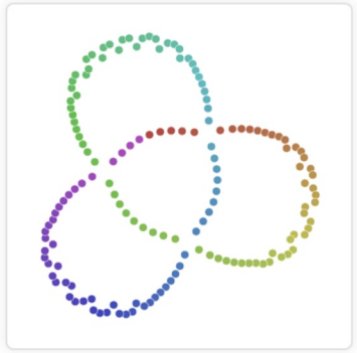


Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000

Place cells with similar local neighborhoods in high-dimension space together in low-dimension space.

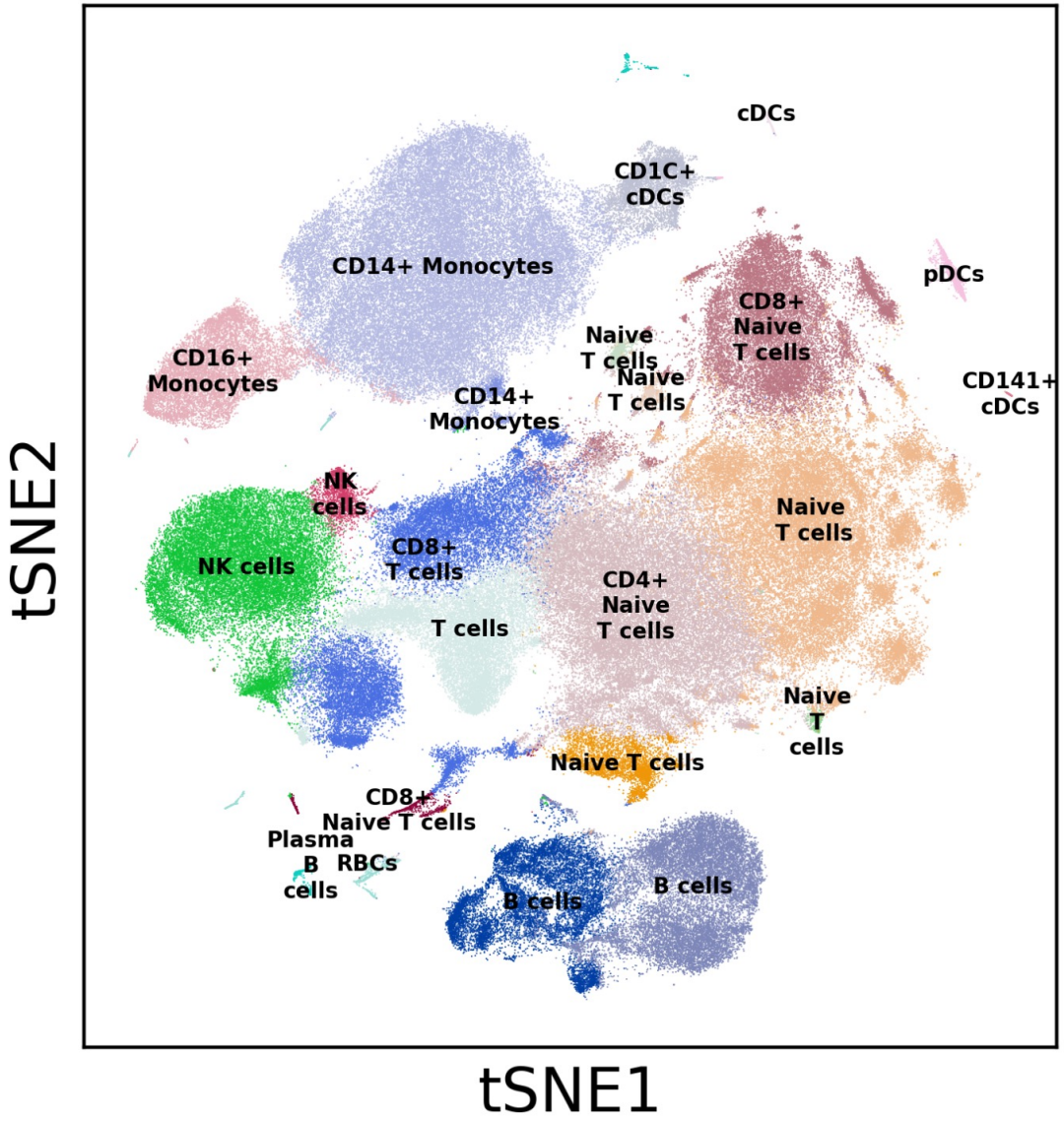
t-SNE and UMAP are popular visualizations.

https://distill.pub/2016/misread-tsne
https://pair-code.github.io/understanding-umap

Single-cell RNA-Seq analysis: Cell state annotation

500K immune cells collected from healthy humans.

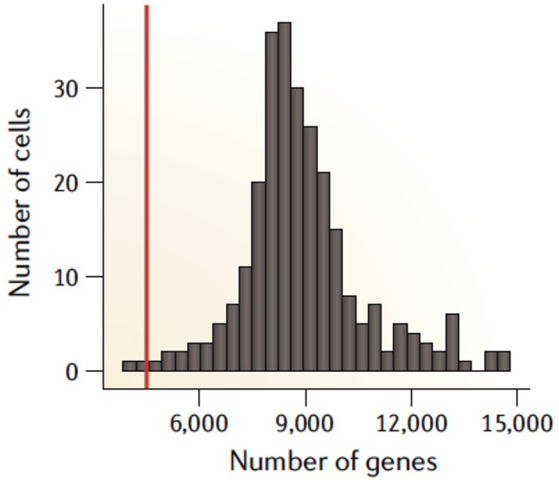
- Immune cells are colored by their cluster identity and visualized in a tSNE embedding.



Cell types are identified by inspecting genes differentially expressed in one cluster relative to other clusters.

Determining cell type, state, and function: Recap

Quality control



Normalization

Feature selection

Dimensional reduction

Cell-cell distances

Unsupervised clustering

Kiselev C. at *al.* (2019) *Nature Reviews Genetics*. 20: 273-282.