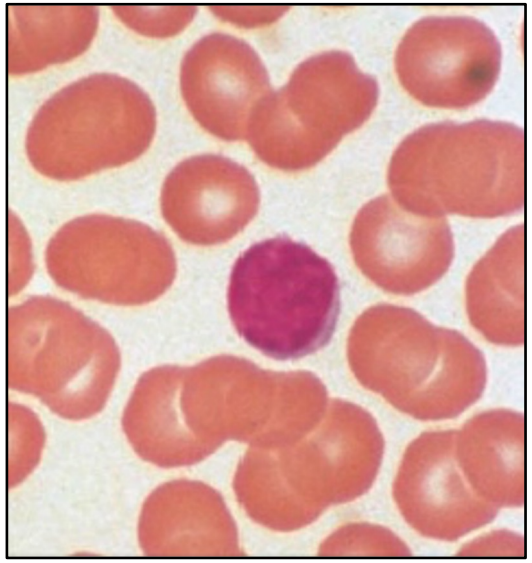
Identification and biological interpretation of cell subsets

Orr Ashenberg and Sergio Triana

October 19, 2022

HCA Africa: Single-Cell RNA-Seq Analysis Workshop



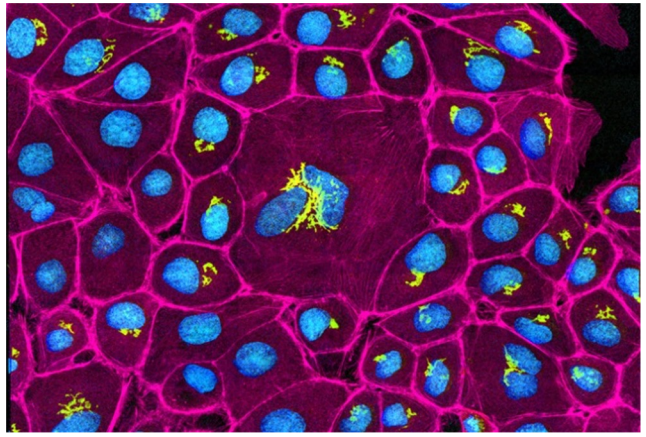
BROAD INSTITUTE

KLARMAN CELL OBSERVATORY AT BROAD INSTITUTE

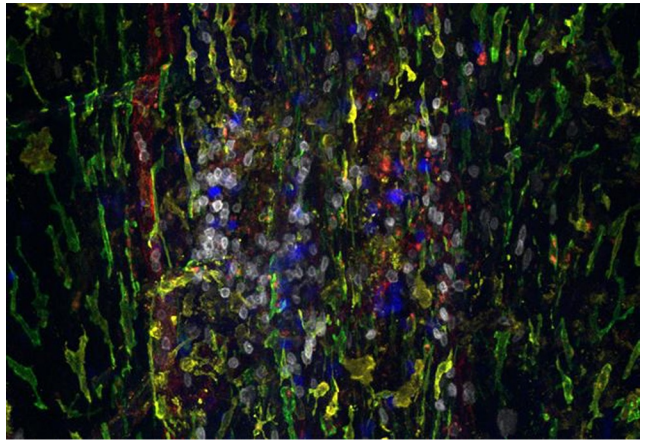
humancellatlas.org

Incredible diversity in cell types, states, and interactions across human tissues

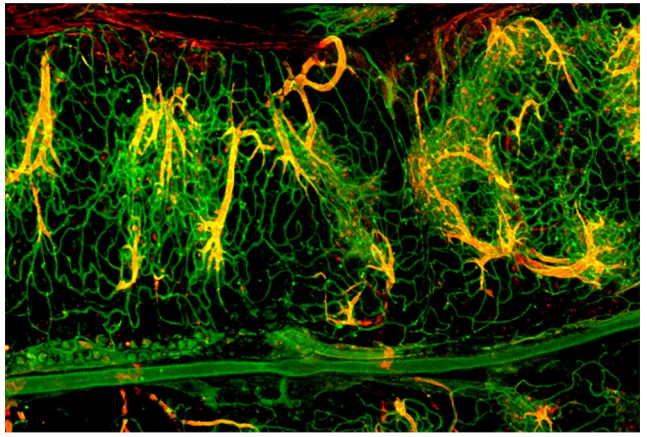
Skin epithelium



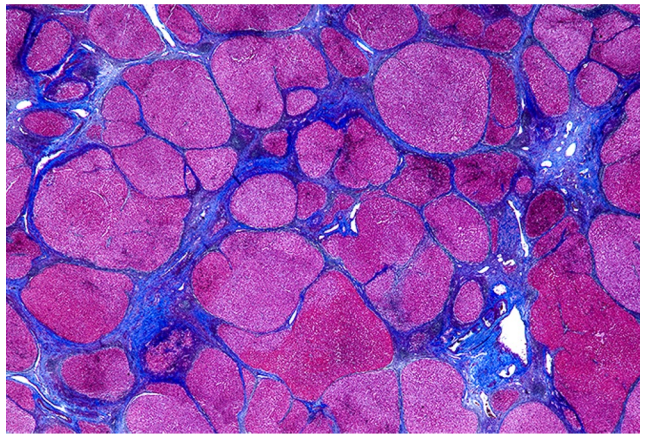
Brain meninges



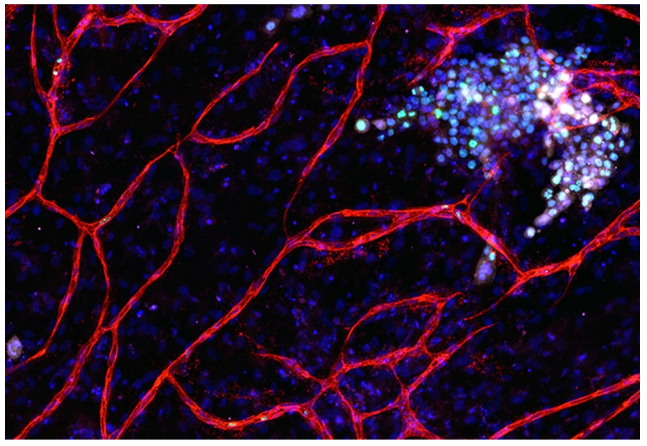
Blood vessels



Small intestine



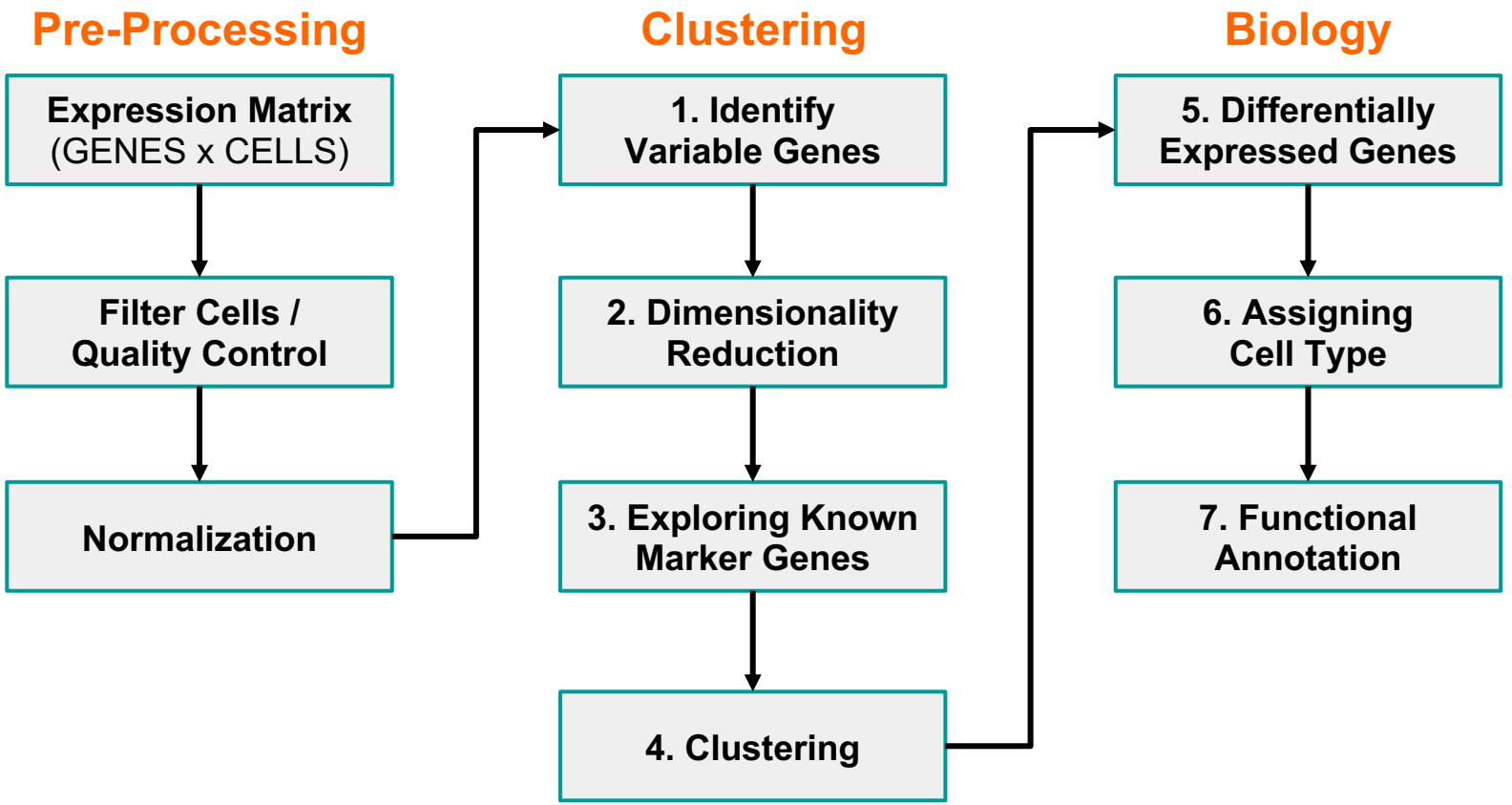
Liver cirrhosis



Breast cancer

http://www.cell.com/pictureshow/skin
https://library.med.utah.edu/WebPath/webpath.html

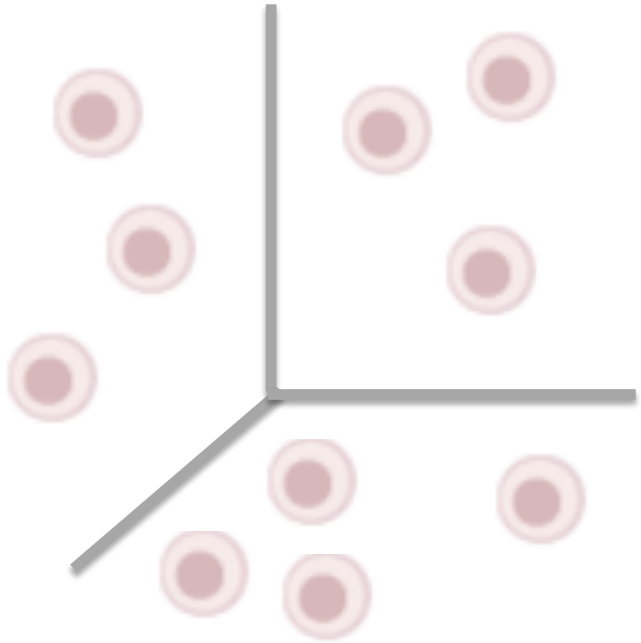
Single-cell RNA-Seq analysis pipeline
Analyzing the expression data



Determining cell type, state, and/or function: 1: Identifying highly variable genes

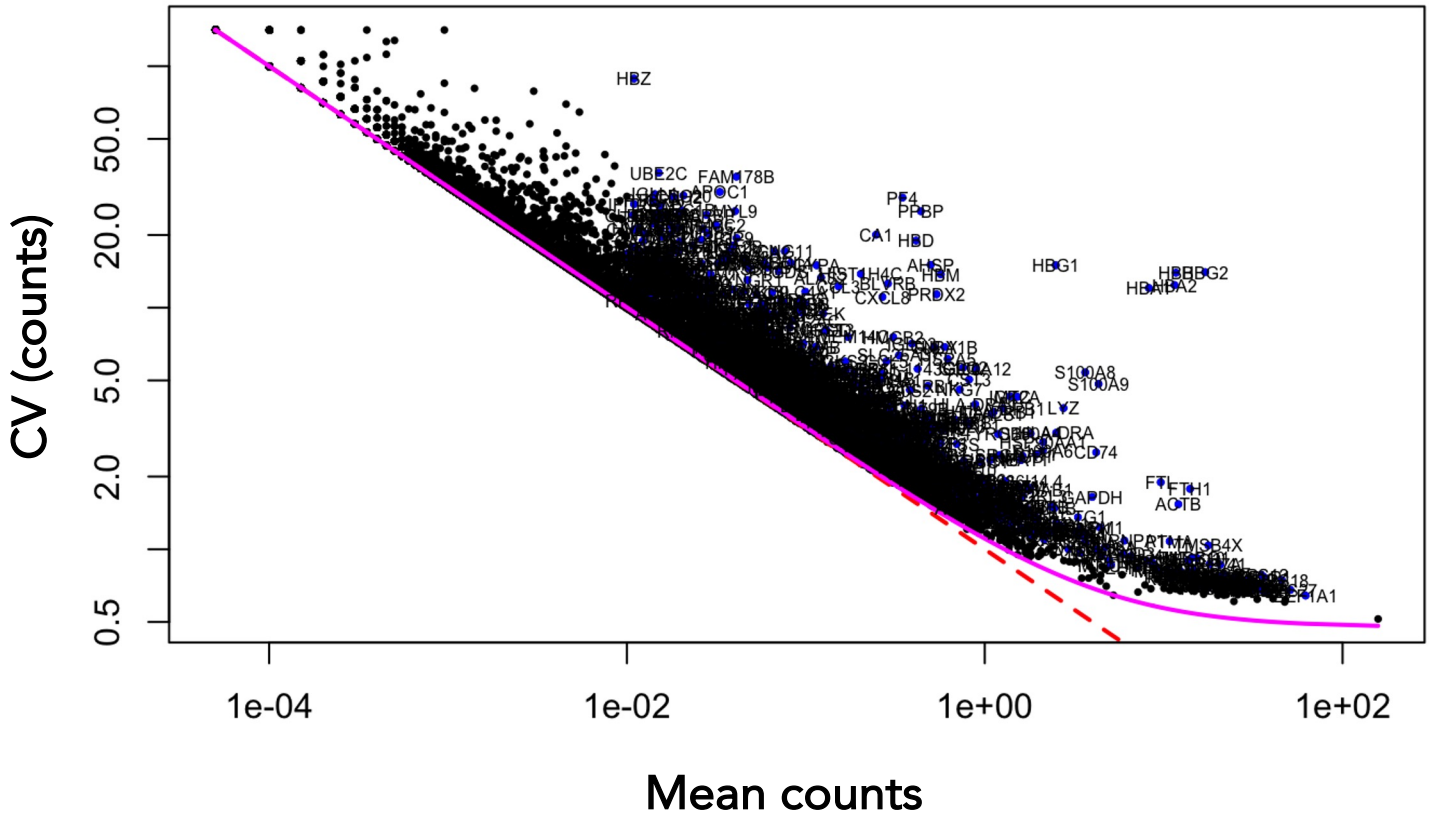
Cells are in ~20,000 dimensional space (one dimension for each gene)

- many genes are lowly detected or noisy measurements



- variable genes contain the biological signal we are interested in

Determining cell type, state, and/or function: 1: Identifying highly variable genes



Determining cell type, state, and/or function: 2: 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

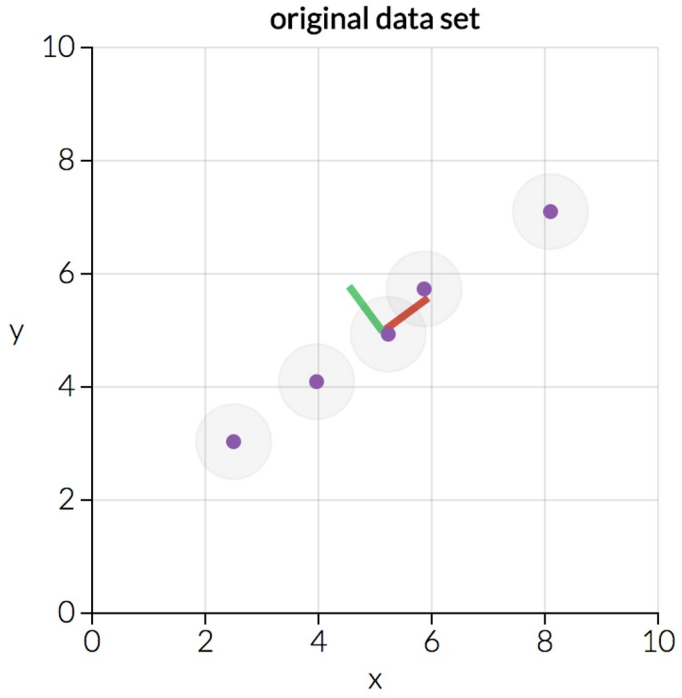
Determining cell type, state, and/or function: 2: Dimensionality reduction

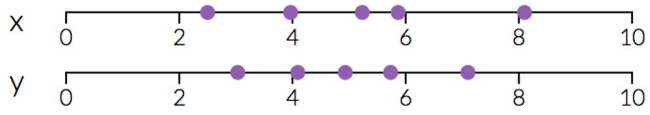
One approach to simplification is to **assume that the data of interest lies within lower-dimensional space**. If the data of interest is of low enough dimension, the data can be visualised in the low-dimensional space.

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)

Determining cell type, state, and/or function: 2: Dimensionality reduction



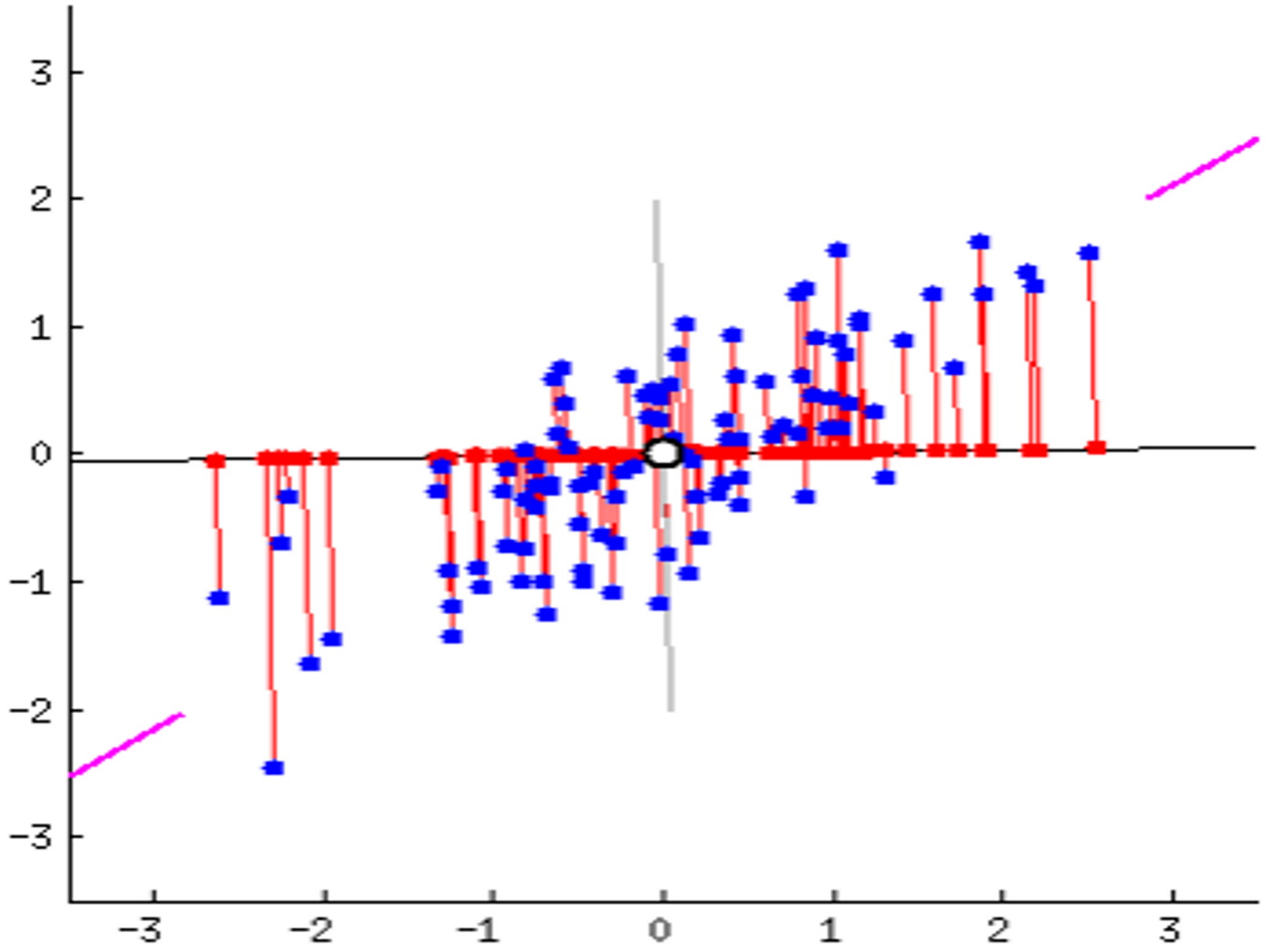


http://setosa.io/ev/principal-component-analysis/

Determining cell type, state, and/or function: 2: Dimensionality reduction

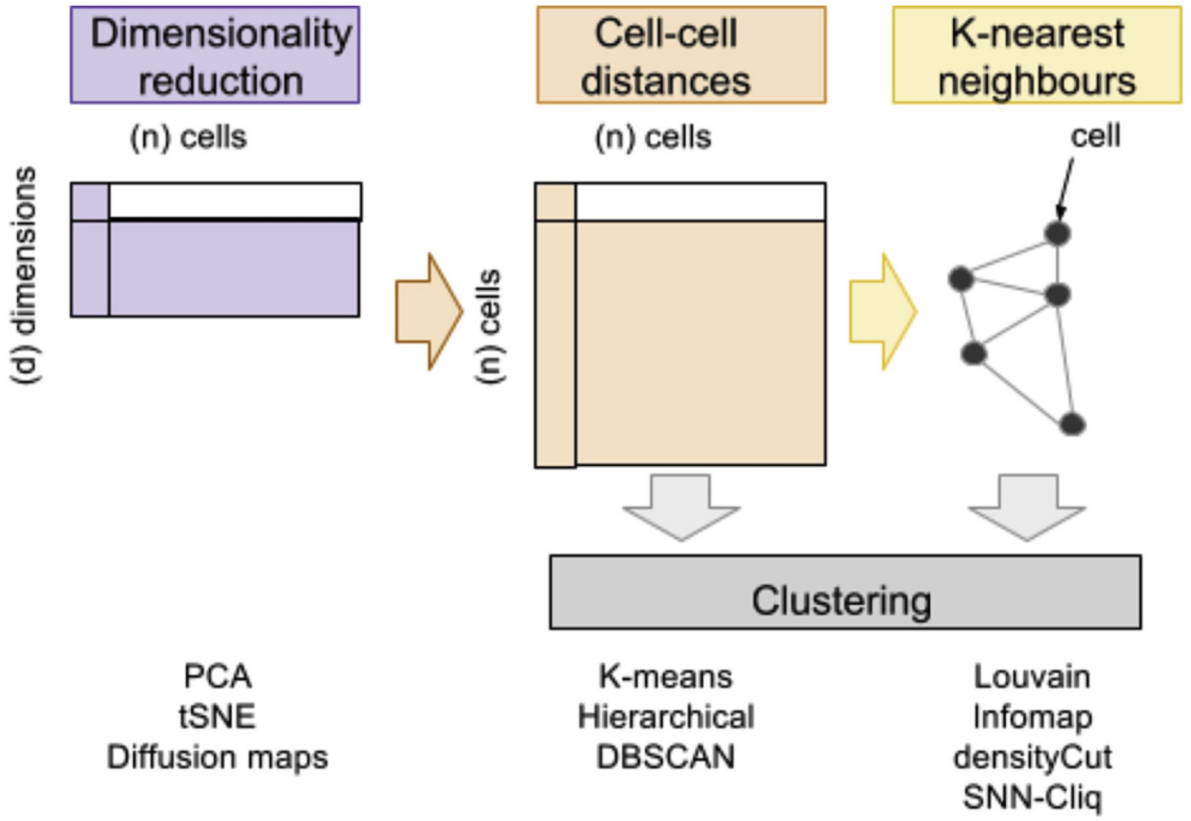
- PCA is a dimensionality reduction method that transforms a **set of features** into a set of **linearly uncorrelated variables** called principal components

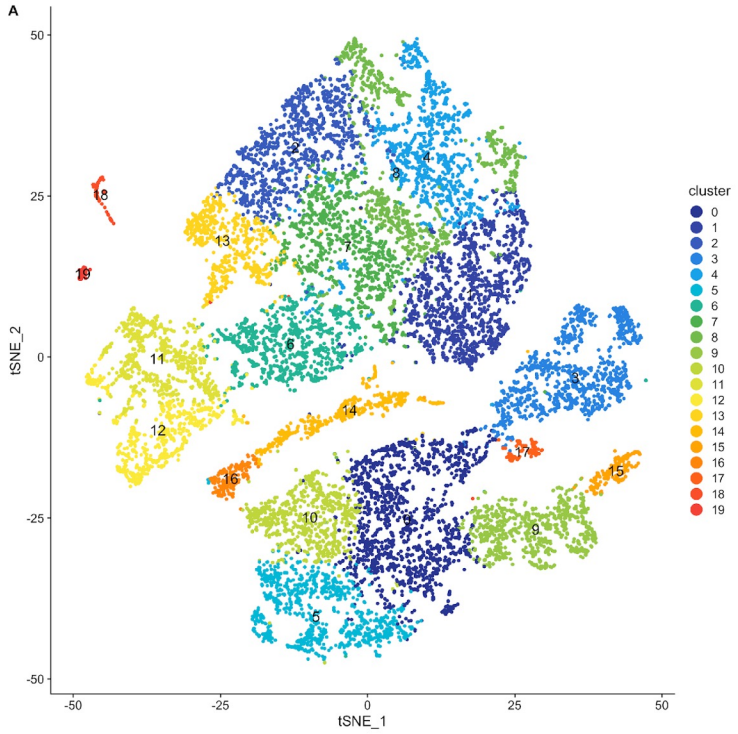
- The first principal component contains the most variance, and each component after contains as much variance while still being orthogonal to other components



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

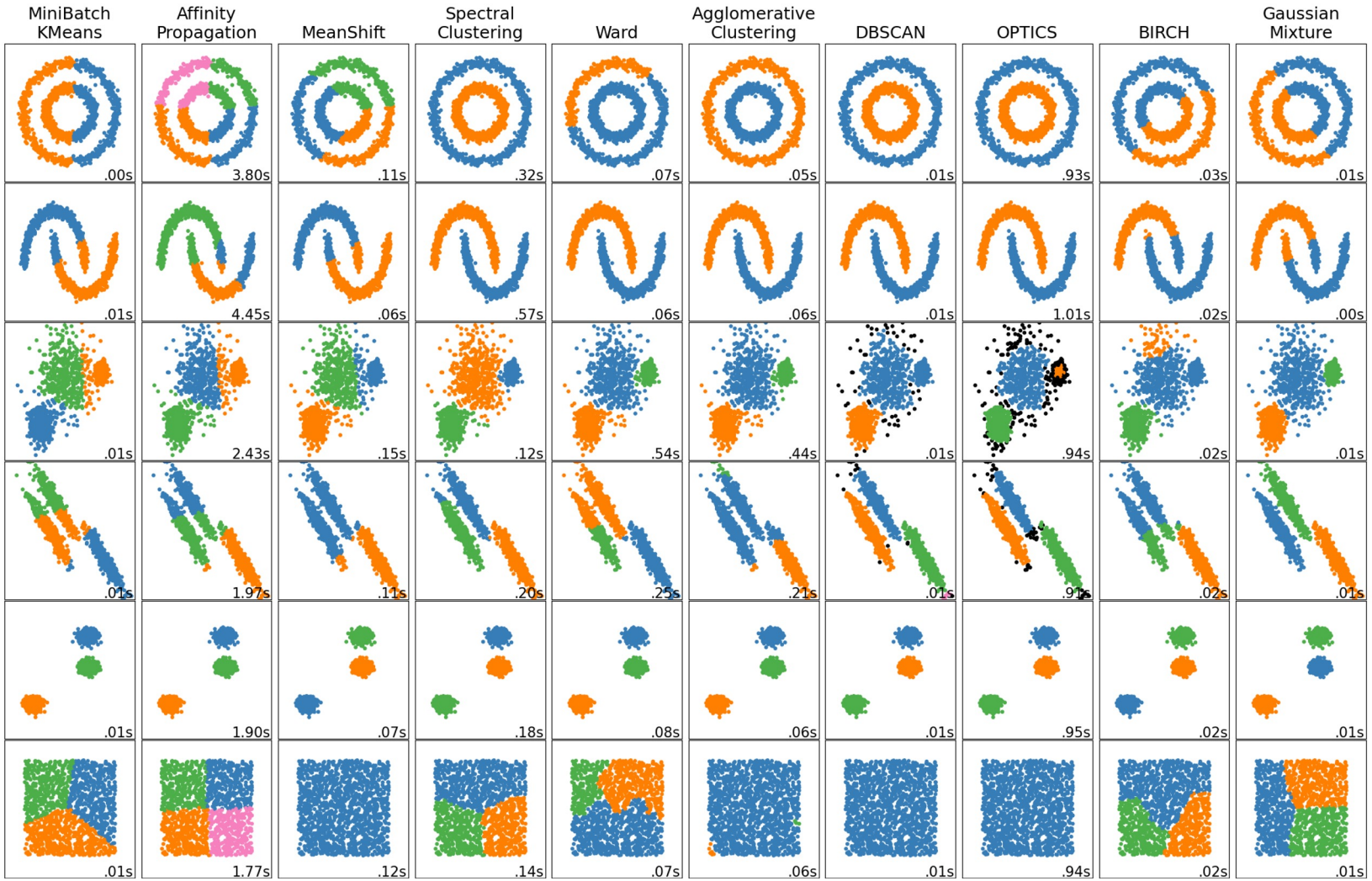
Determining cell type, state, and/or function: 3: Clustering principal components





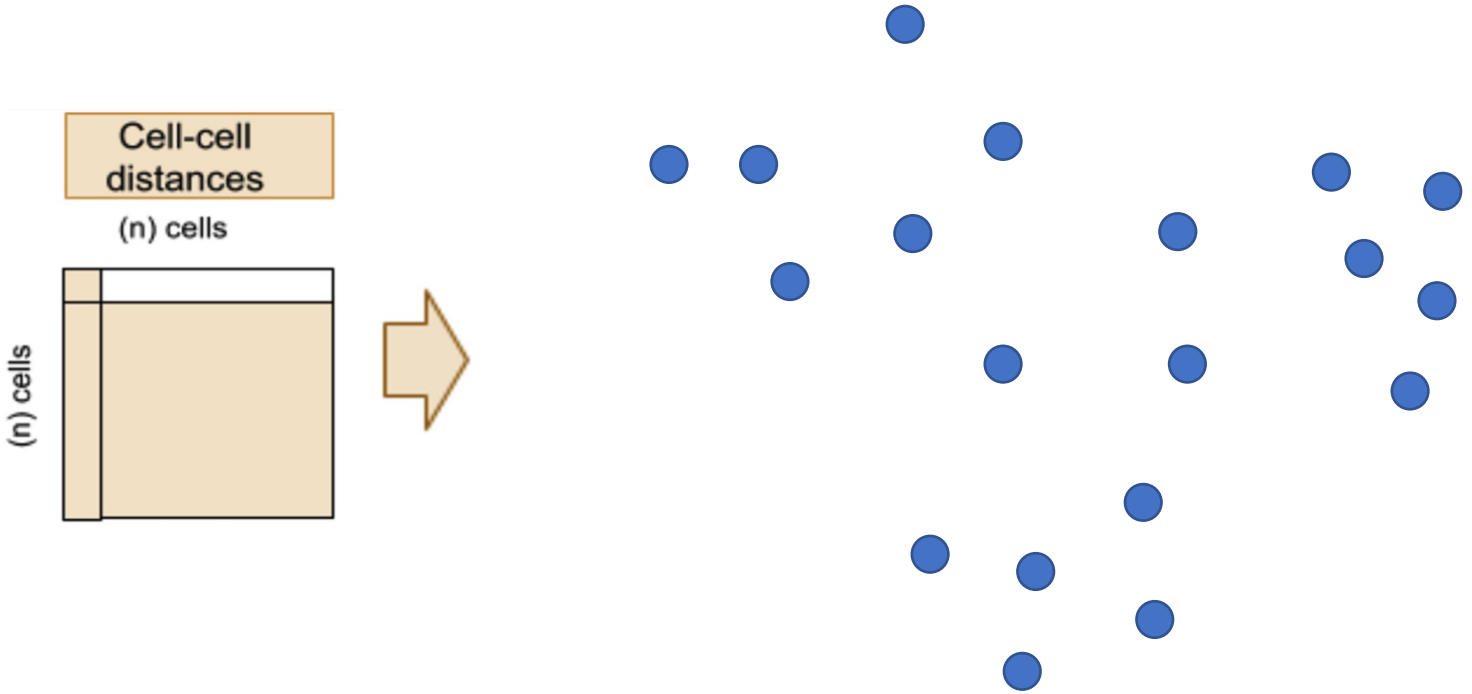
Andrews, Tallulah S. at *al.* (2017) *Molecular aspects of medicine.*

Determining cell type, state, and/or function: 3: Clustering principal components

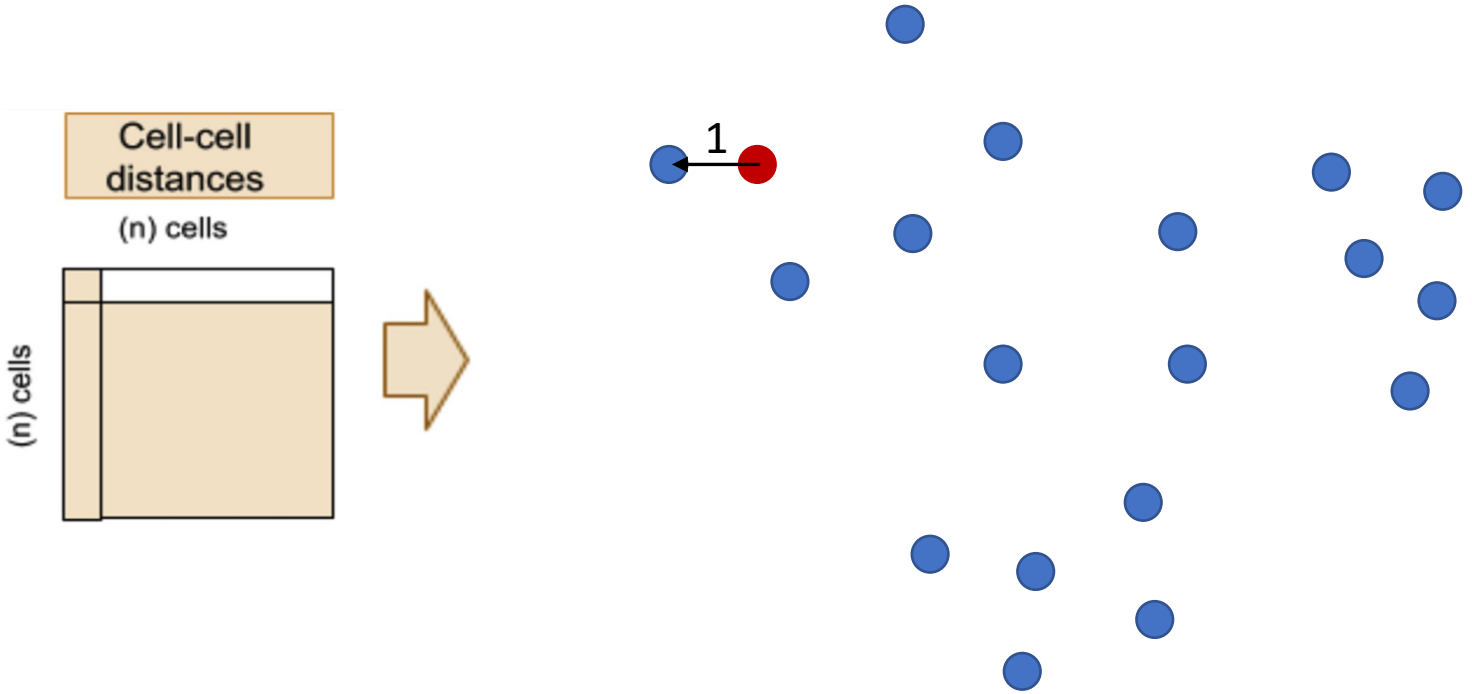


https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

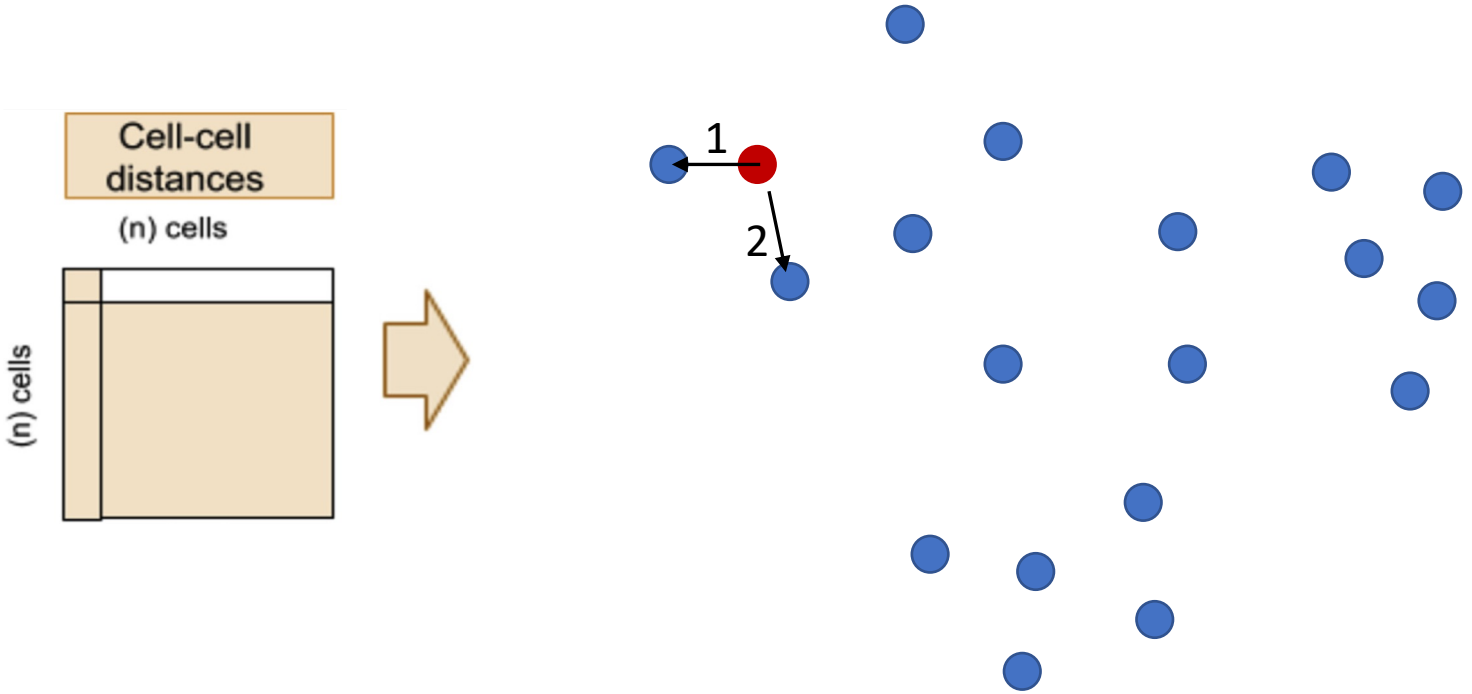
Determining cell type, state, and/or function: 3: Clustering principal components



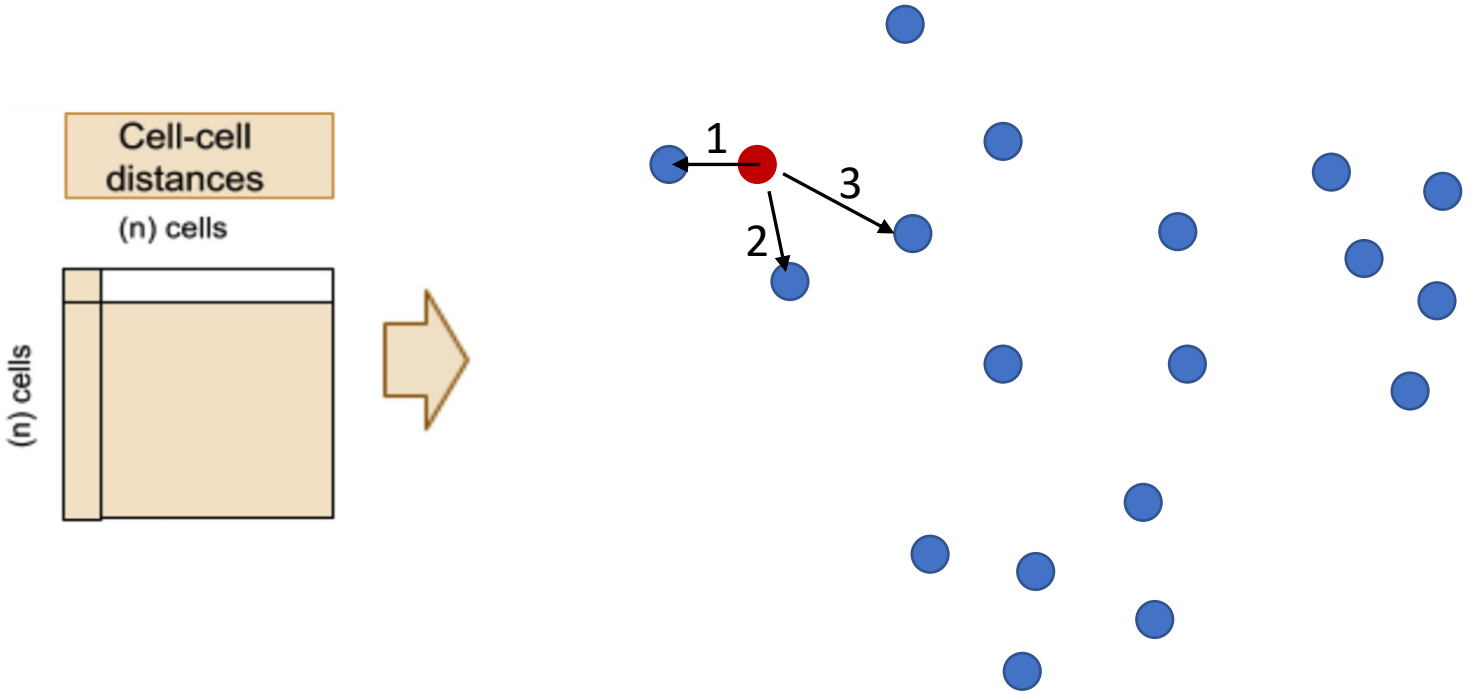
Determining cell type, state, and/or function: 3: Clustering principal components



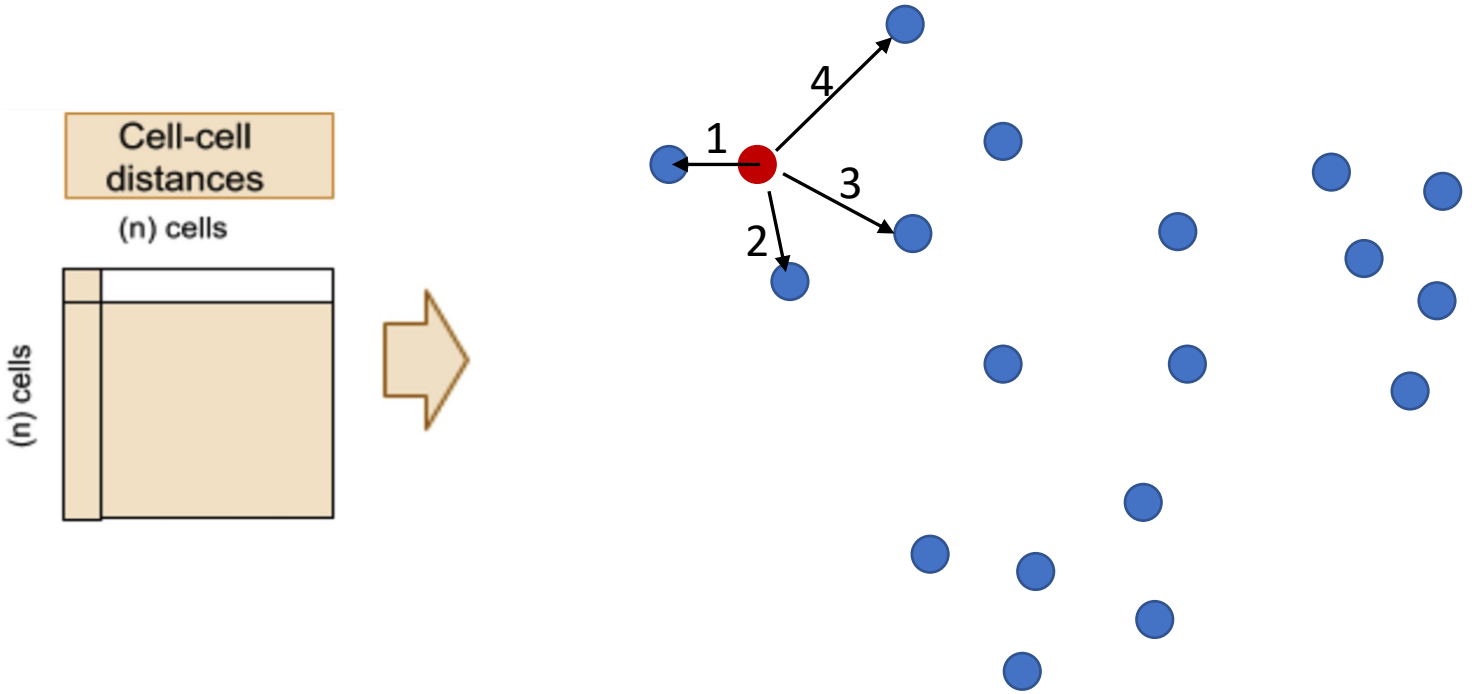
Determining cell type, state, and/or function: 3: Clustering principal components



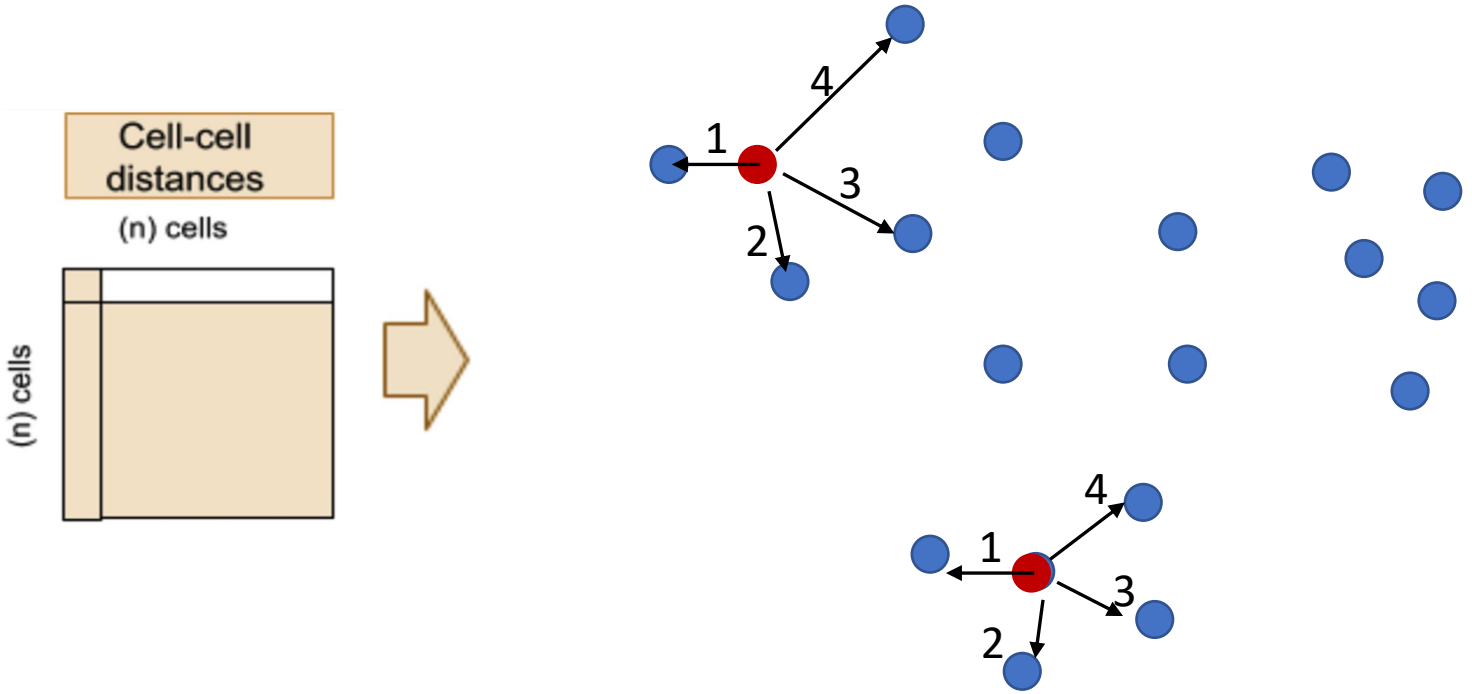
Determining cell type, state, and/or function: 3: Clustering principal components



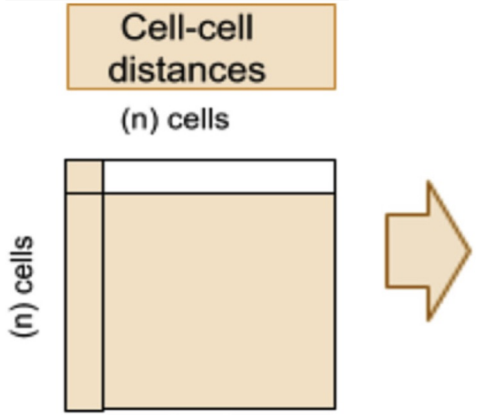
Determining cell type, state, and/or function: 3: Clustering principal components

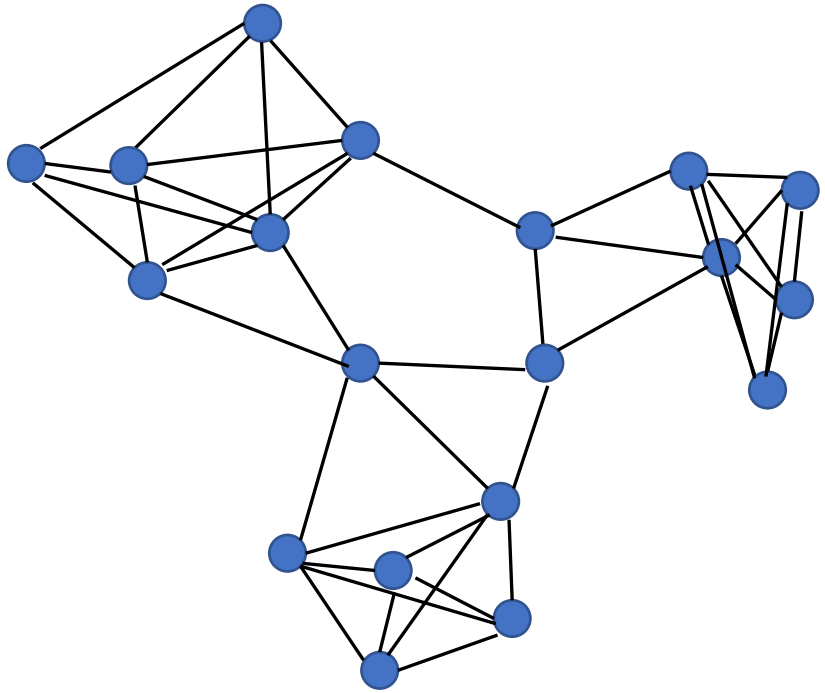


Determining cell type, state, and/or function: 3: Clustering principal components

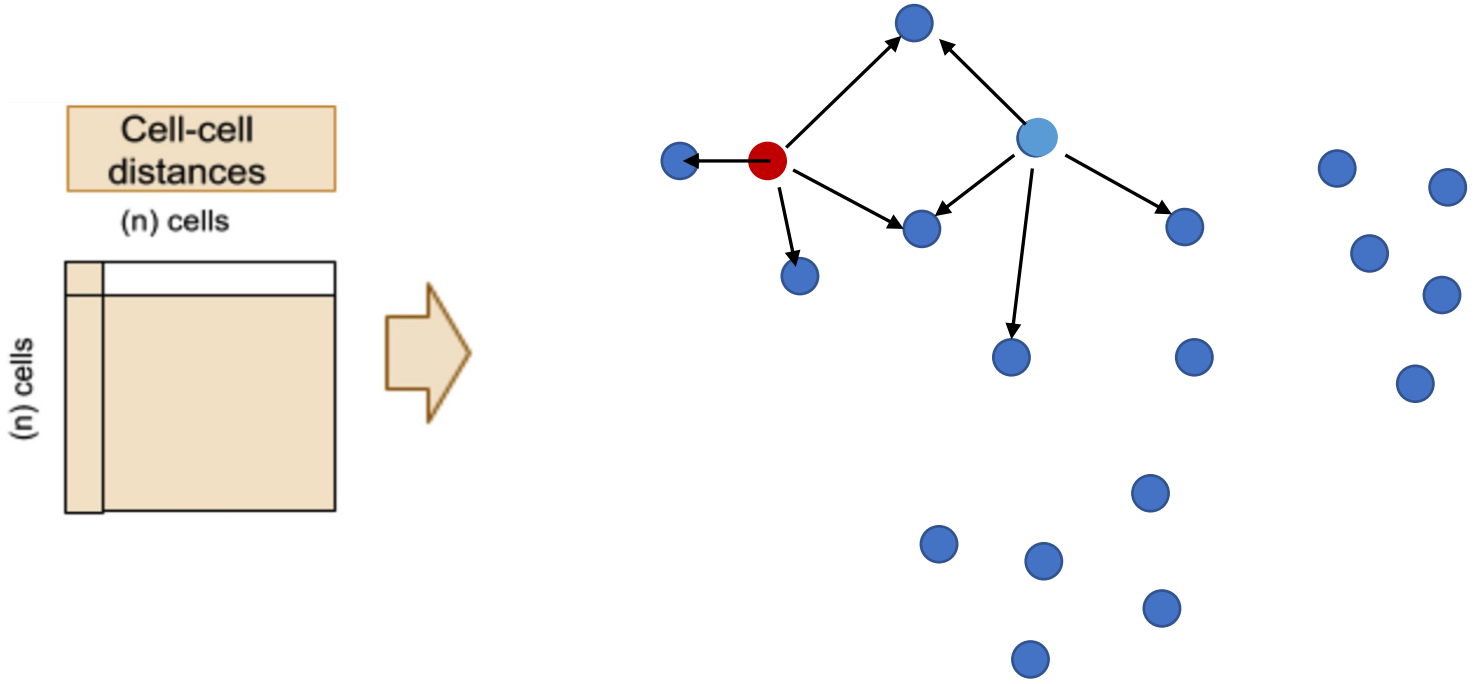


Determining cell type, state, and/or function: 3: Clustering principal components



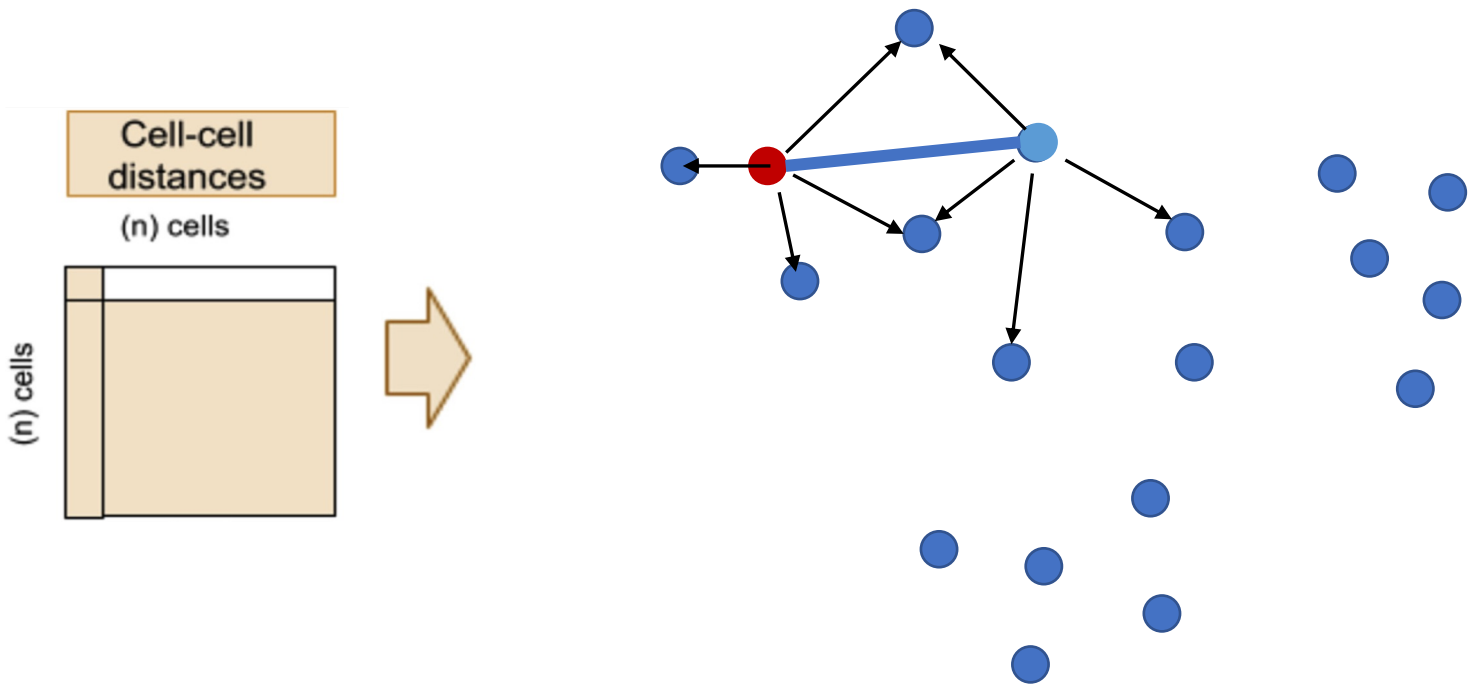


Determining cell type, state, and/or function: 3: Clustering principal components

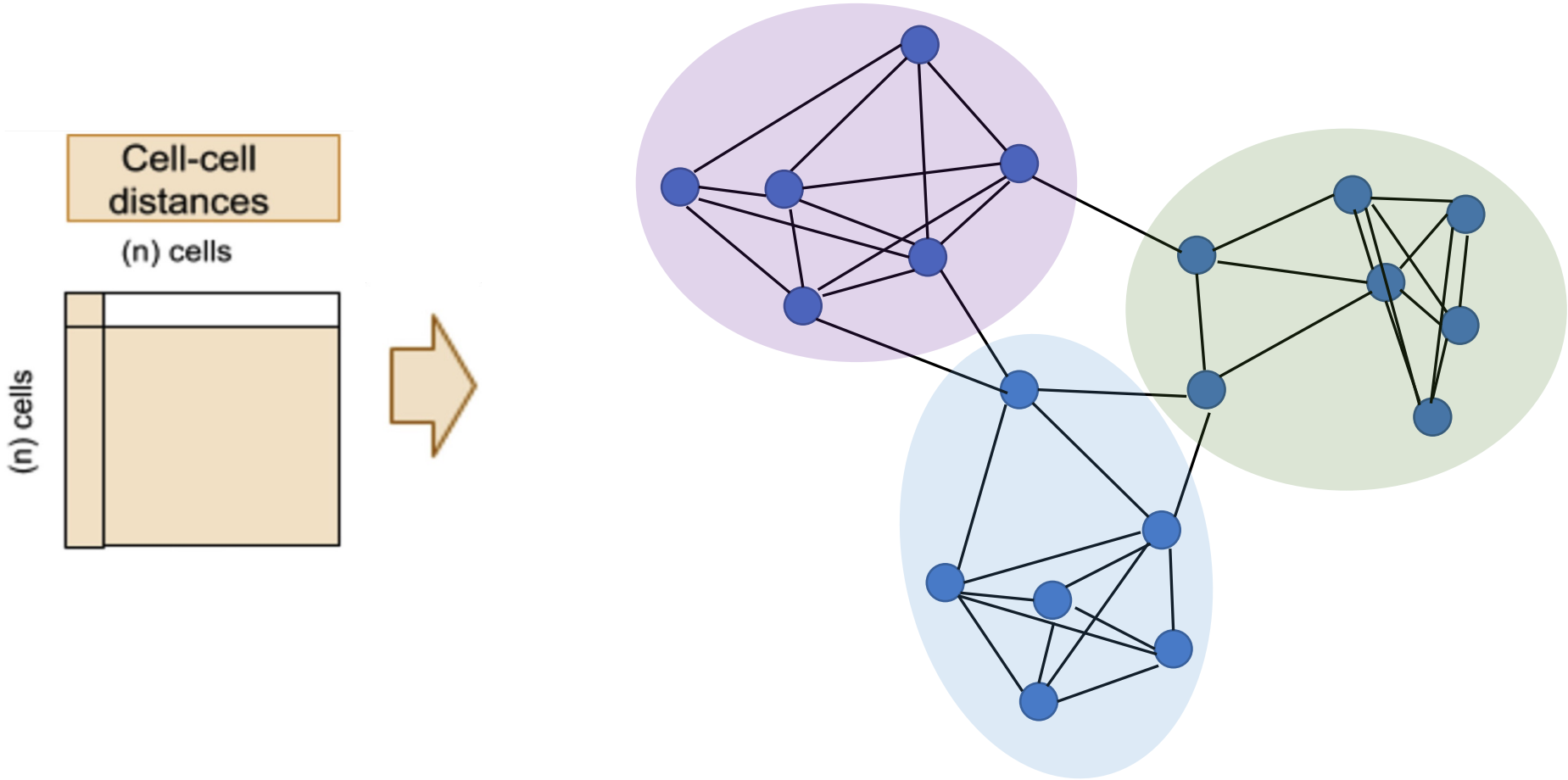


Determining cell type, state, and/or function: 3: Clustering principal components

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

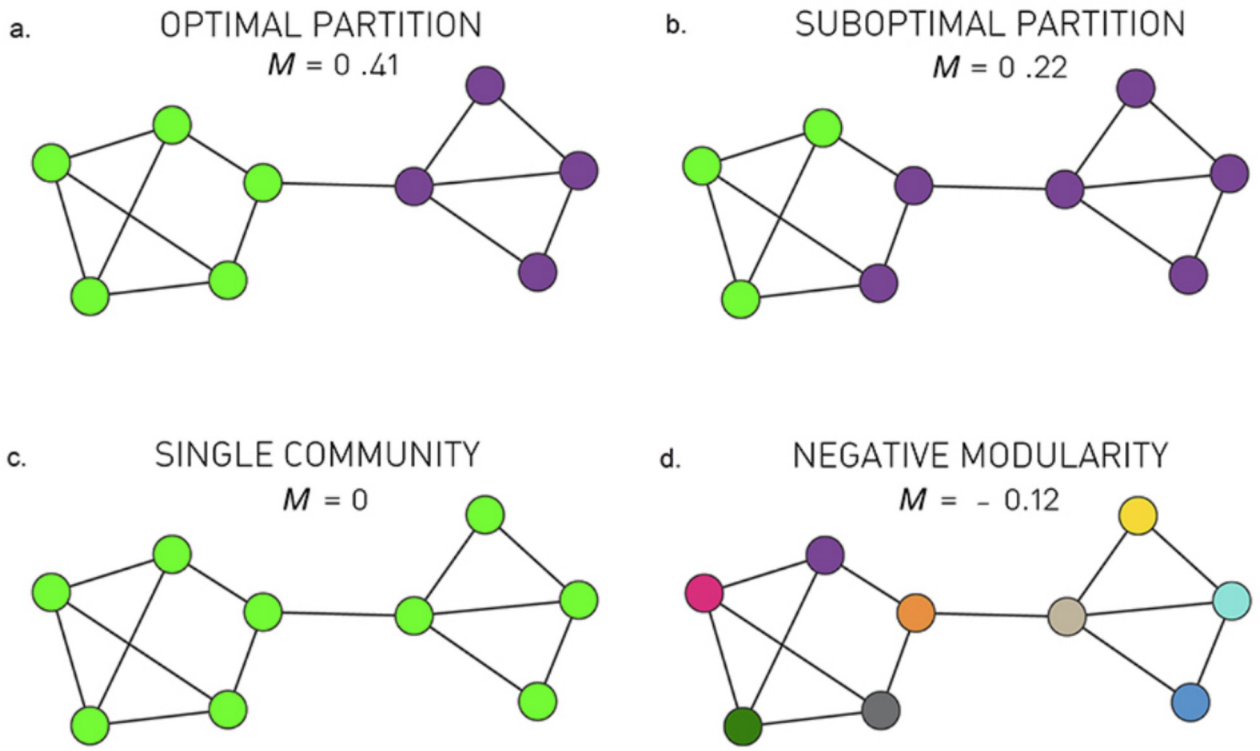


Determining cell type, state, and/or function: 3: Clustering principal components

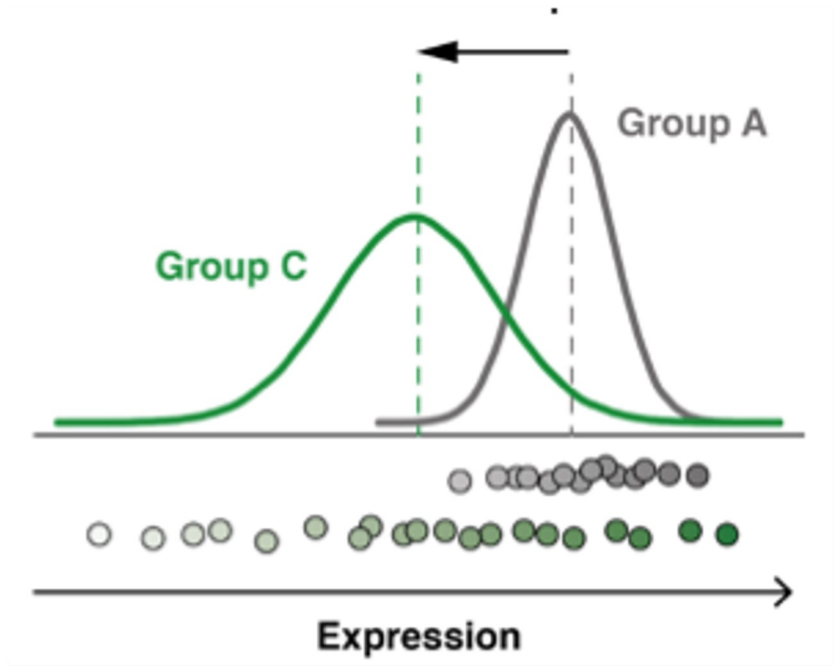


Determining cell type, state, and/or function: 3: Clustering principal components

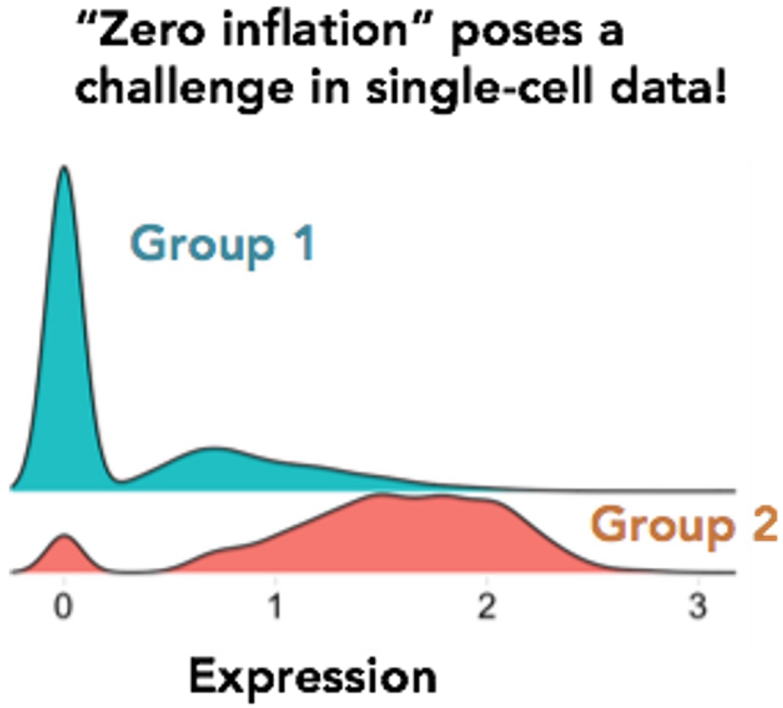
- Graph-based clustering is based on **community detection**.
- Many different algorithms for community detection:
  - Louvain (heuristic), Infomap, Walktrap
- Most of them are based on modularity maximization



Determining cell type, state, and/or function: 4: Identifying differentially expressed genes



Bulk



Single cell

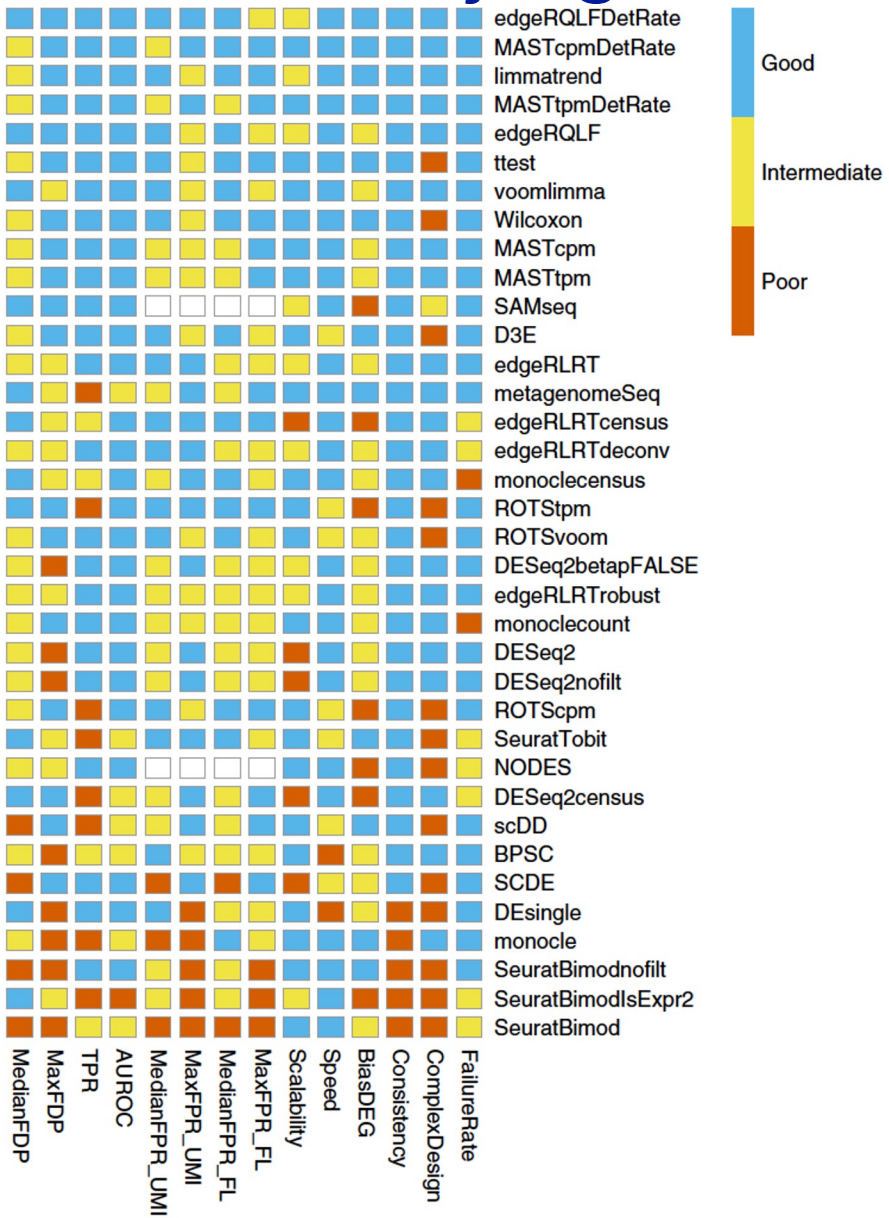
Determining cell type, state, and/or function: 4: Identifying differentially expressed genes

In scRNA-seq we often do not have a defined set of experimental conditions.

Instead, we can 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…

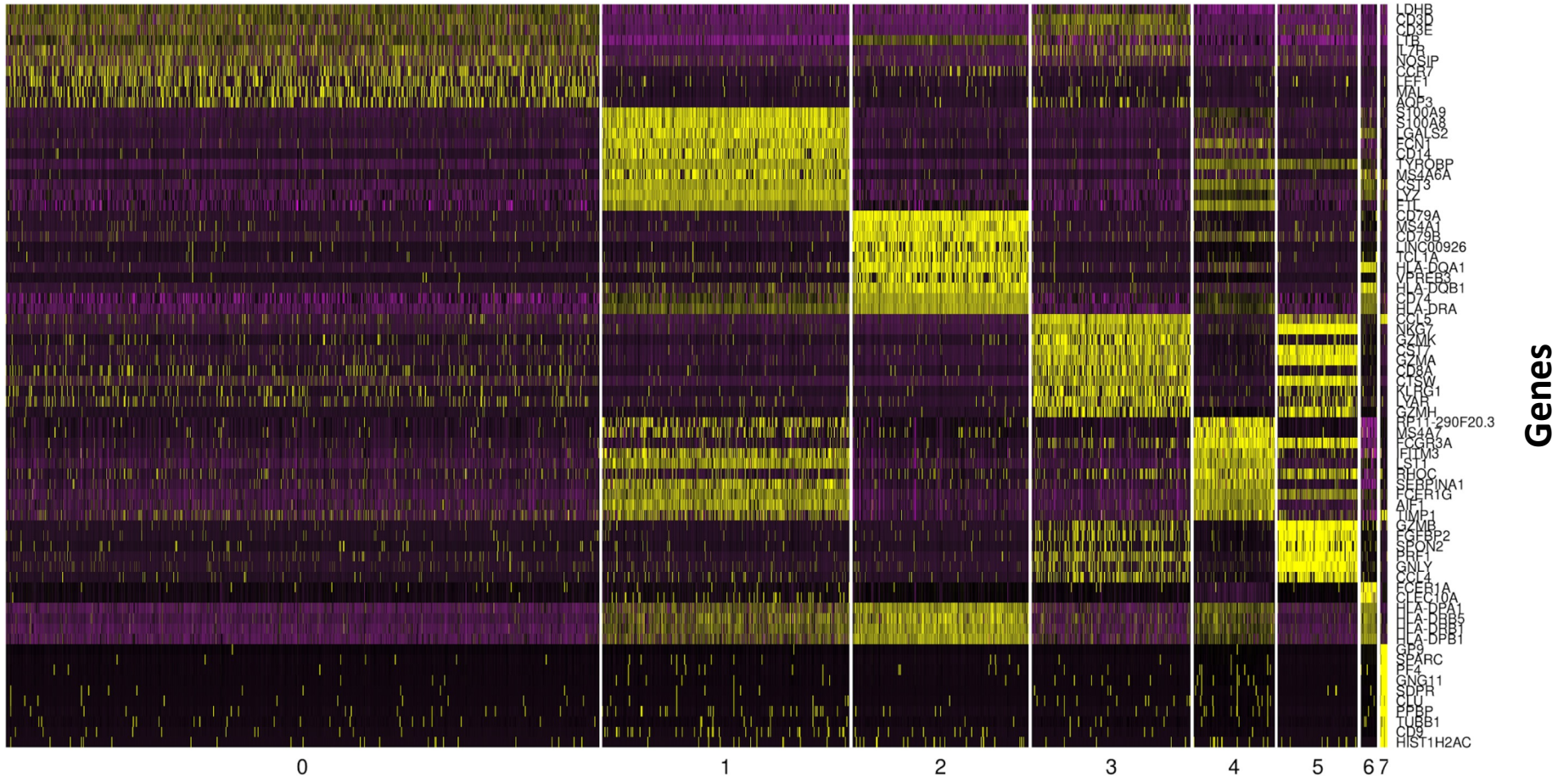
Determining cell type, state, and/or function: 4: Identifying differentially expressed genes



There are many ways to test for differential expression!

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

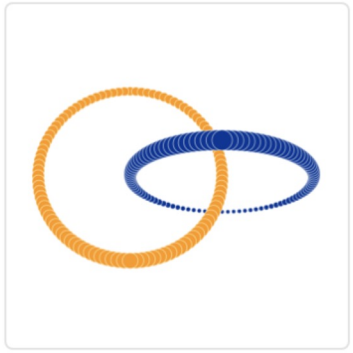
Determining cell type, state, and/or function: 4: Identifying differentially expressed genes



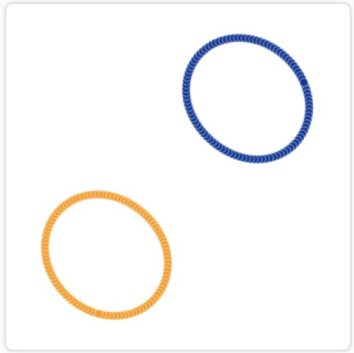
Determining cell type, state, and/or function: 4: Visualizing cells in lower dimension

High dimension manifold

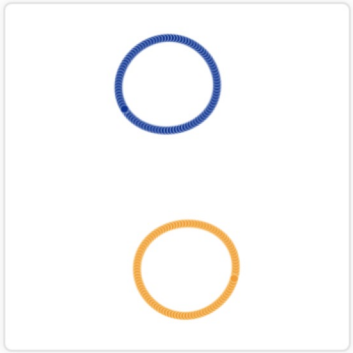
Low dimension representation



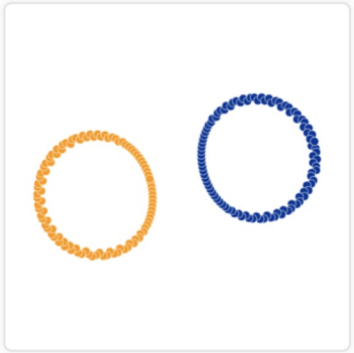
Original



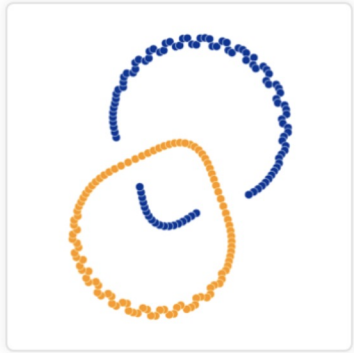
Perplexity: 2
Step: 5,000



Perplexity: 5
Step: 5,000

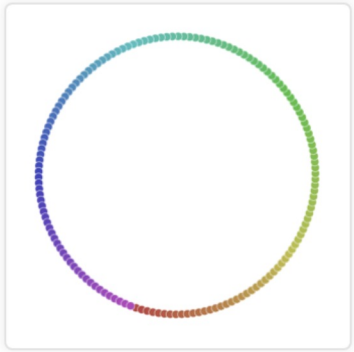


Perplexity: 30
Step: 5,000



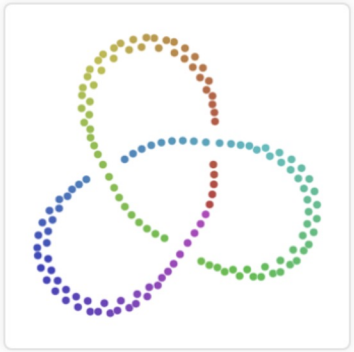
Perplexity: 50
Step: 5,000

Original

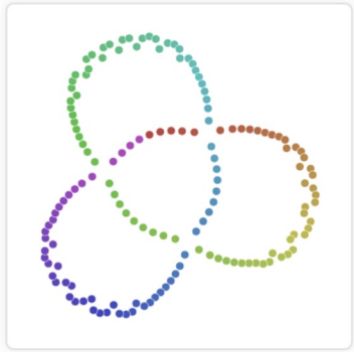


Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



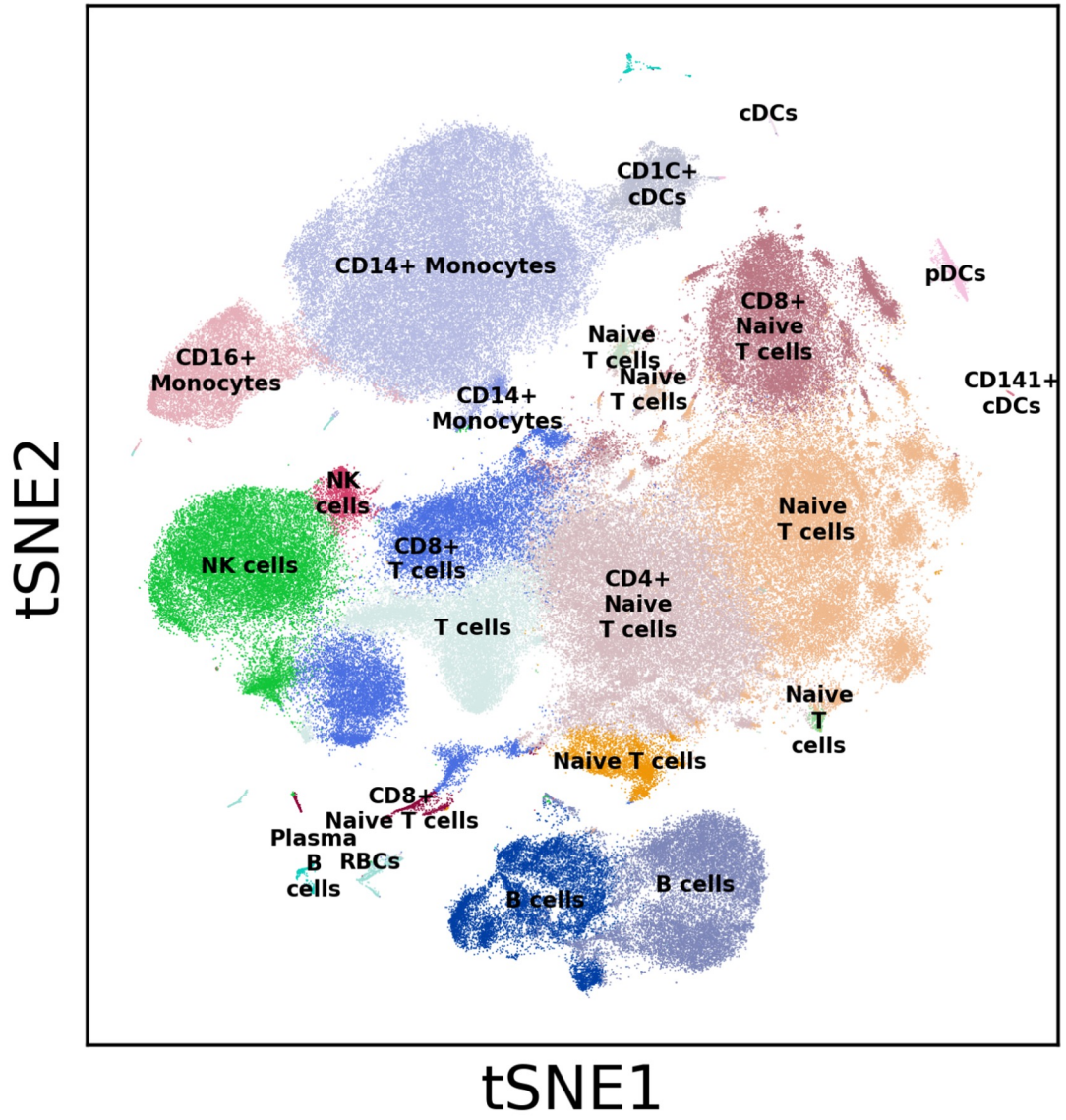
Perplexity: 50
Step: 5,000

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

t-SNE and UMAP are popular visualizations.

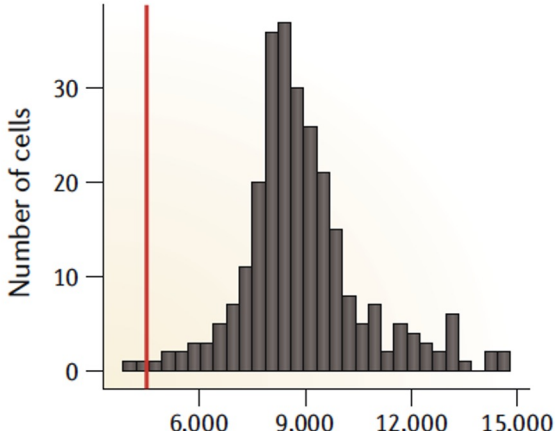
https://distill.pub/2016/misread-tsne
https://pair-code.github.io/understanding-umap/

Determining cell type, state, and/or function: 4: Assigning cell type



Determining cell type, state, and function: Recap

Quality control



Normalization

Feature selection

Dimensional reduction

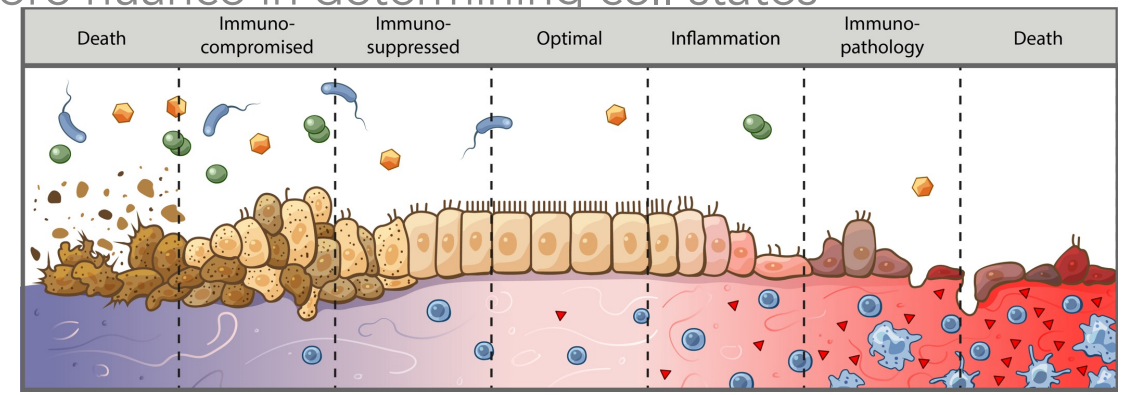
Cell-cell distances

Unsupervised clustering

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

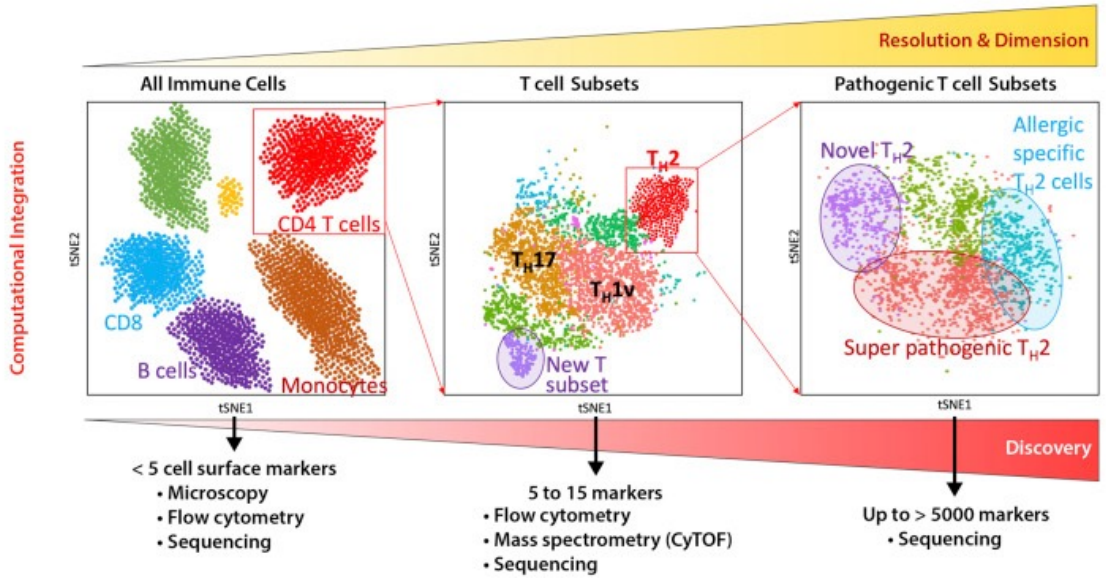
What are we actually annotating?

- Cells exist along multiple continuums
  - Ex. Differentiation, inflammatory response, anti-viral response, EMT transition
- Conventional profiling techniques puts cells into “buckets”
- Unbiased approaches allows us to appreciate how cell types/states relate to each other along a spectrum
- We have a hard time conceptualizing this spectrum—so we need to annotate cells similarly to conventional techniques
  - Usually with a bit more nuance in determining cell states



Biology to your data!

- Annotating cells in single-cell gene expression data is a challenge!
  - -What's a cell type?
    - Gene expression is not discrete but mostly a continuum
    - Gene expression ≠ cellular function
  - -Resolution
    - Subsets
    - Phenotypes
    - Differentiation



Seumois G, Vijayanand P 2019 . *J Allergy Clin Immunol.*

Approach 1: *De novo* cluster annotation

- Three main approaches to assigning clusters
  - Annotating based on known marker genes
  - Annotating on differential gene expression
  - Annotating on previously reported module scoring
- Edge cases
  - Doublets
  - Low-quality cells
  - Host-pathogen annotations
- Clustering accuracy
  - Over-clustering
  - Under-clustering
  - Iterative clustering

Annotating based on known marker genes

- Requires understanding of expected cell types and their markers
  - Biological expertise here is critical
- Many protein and RNA markers overlap
  - However, not all protein markers translate well into RNA space
- Experience with cell types in sample and scRNA-seq will help identify which protein markers do or do not translate to RNA space
- Most useful for identifying coarse cell types and some cell states, but difficult to annotate pathological cell states in disease

Annotating on differential gene expression

- Begin with one-against-all differential gene expression
  - Easier to annotate based on upregulated rather than downregulated genes
- Known marker genes that are highly expressed should show up here
  - May not depending on differential gene expression cutoffs
- Top 20 to 30 genes should give an indication of the cell states as well as cell type
  - Ex. Inflammatory, anti-viral, low-quality, etc.

Annotating on differential gene expression (cont'd)

- This is an iterative process
  - You should not expect to be able to confidently identify all your cell clusters on your first pass
- There will likely be clusters that are very similar to each other
  - Run differential gene expression between those similar clusters ONLY to figure out what's different between them
  - If differential gene expression doesn't make sense biologically, combine clusters into one
- Continue to iterate and re-check work as you go

Annotating on previously reported module scoring

- Supplementary tables will usually list differential gene expression used for assigning clusters in other studies
  - Use these to your advantage
  - Great for checking work
- Make sure species and system are correct
- Technology used may influence results

These basic techniques should not be used in a vacuum—they build on each other to convince you that you have annotated your clusters correctly

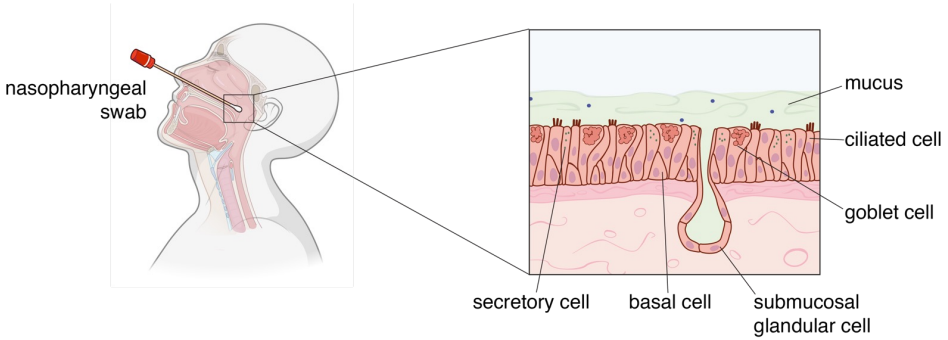
Host-pathogen annotations

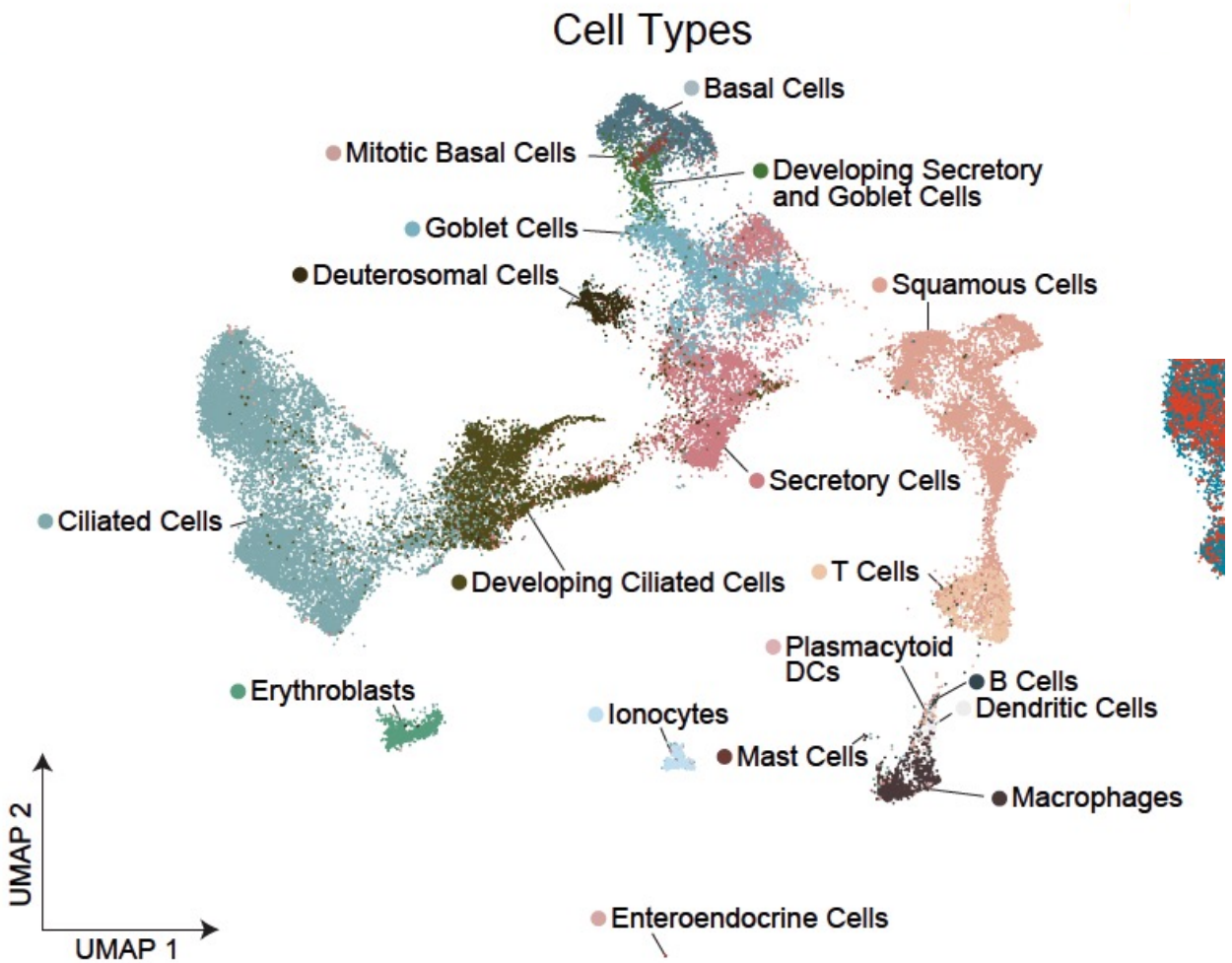
- ScRNA-seq will capture poly-adenylated pathogen reads for intra-cellular pathogens
- Required alignment against custom, combined host-pathogen genome
- Pathogen RNA != intra-cellular pathogen presence

  Especially in low quantities, pathogen RNA may be from ambient background contamination
- Intra-cellular pathogen RNA != active infection

  Phagocytes sample their environment and other cells continuously and may readily capture pathogen RNA

Host-pathogen annotations (cont'd): A COVID-19 example

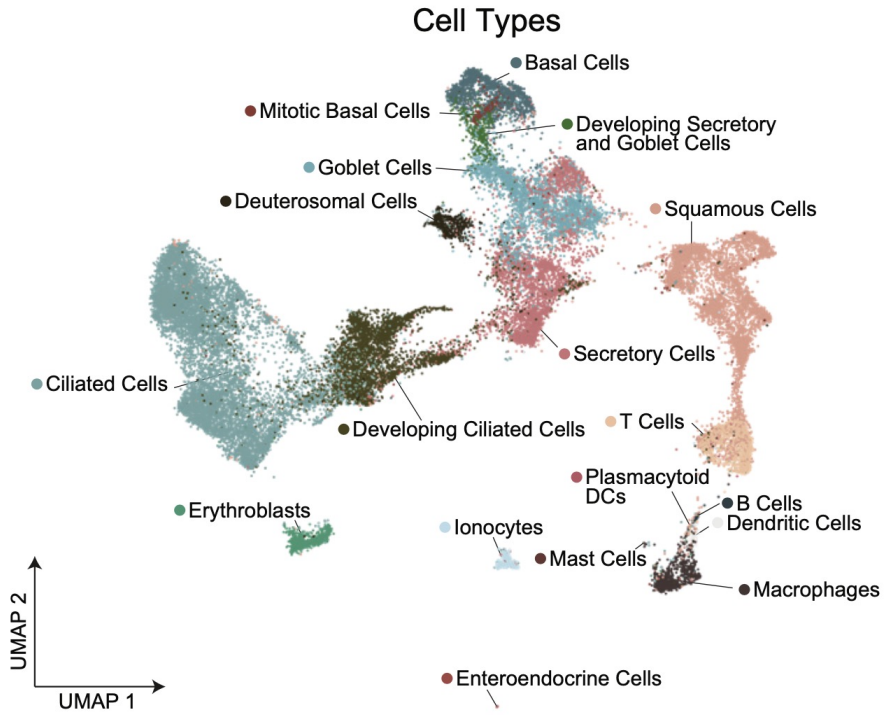




32588 cells across 58 individuals

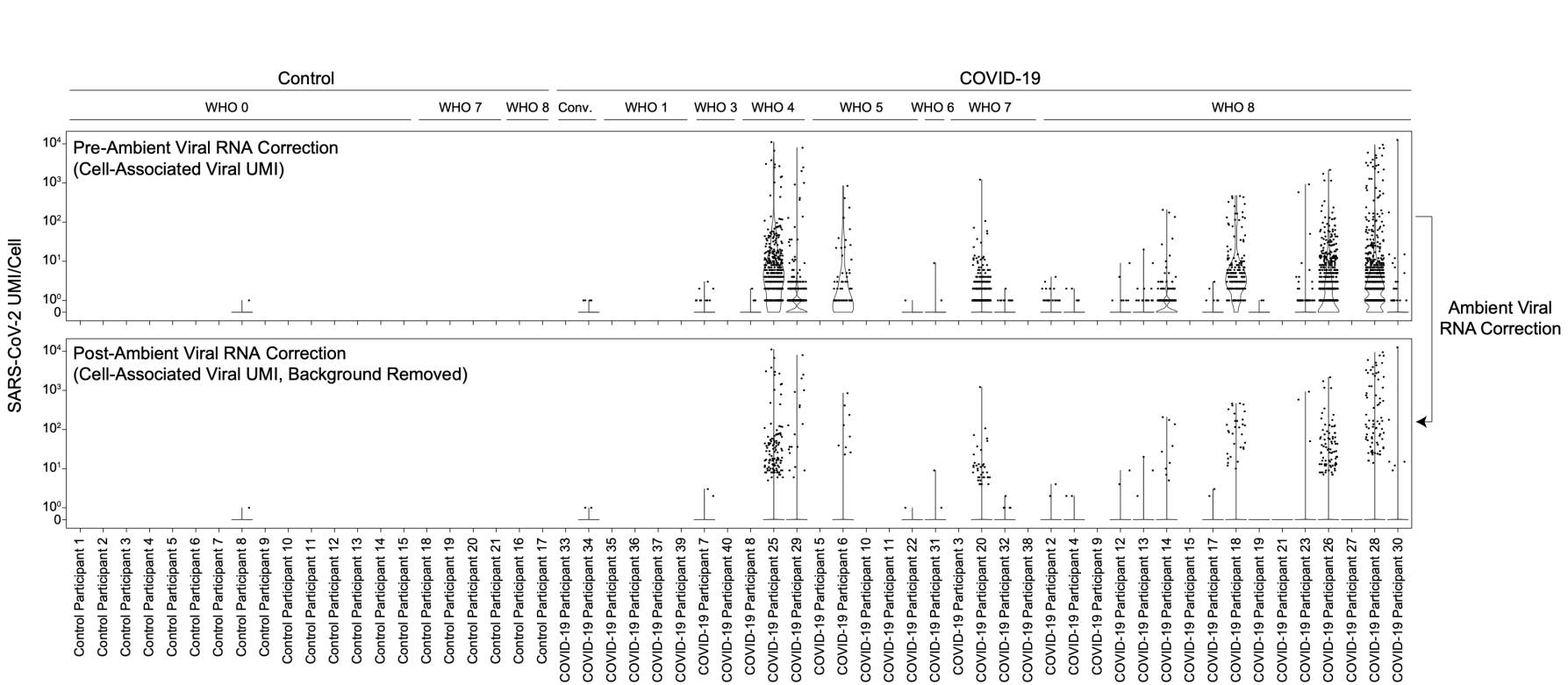
Ziegler et al., *Cell*, 2021.

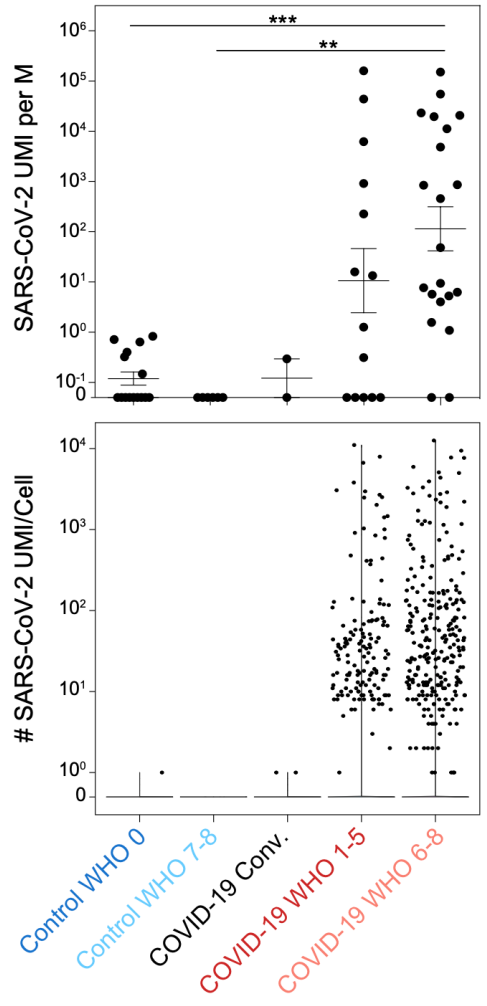
Host-pathogen annotations (cont'd): A COVID-19 example



Ziegler et al., *Cell*, 2021.

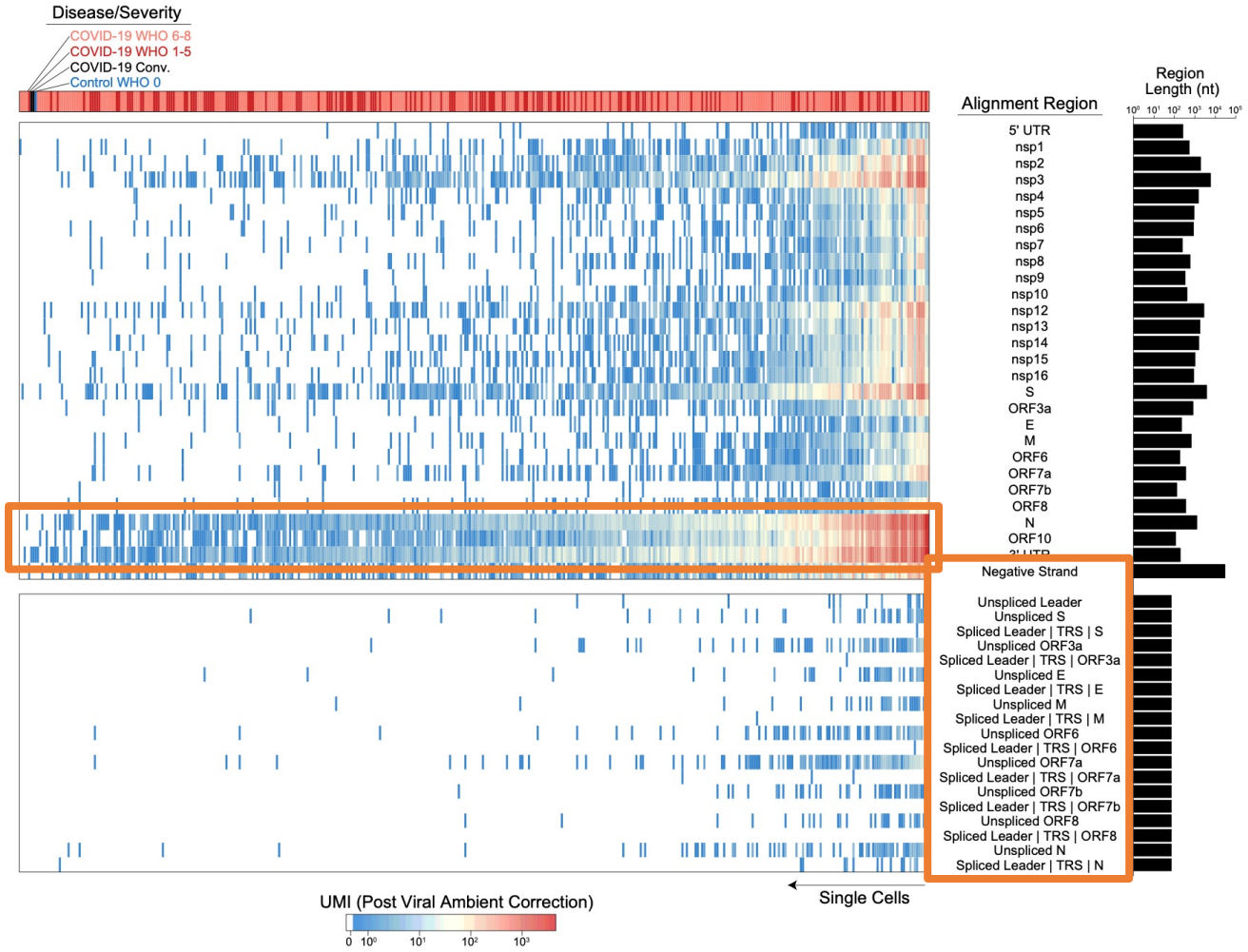
Host-pathogen annotations (cont'd): A COVID-19 example

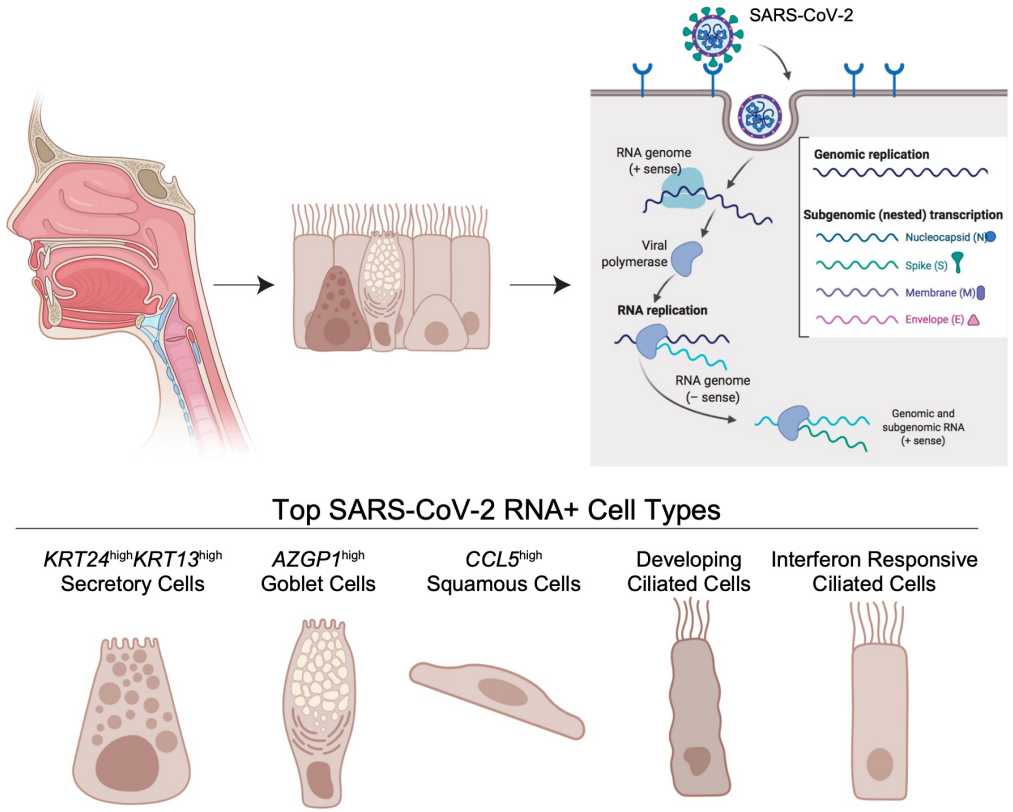




Fleming et al., *bioRxiv*, 2019; Kotliar et al., *Cell*, 2020; Ziegler et al., *Cell*, 2021.

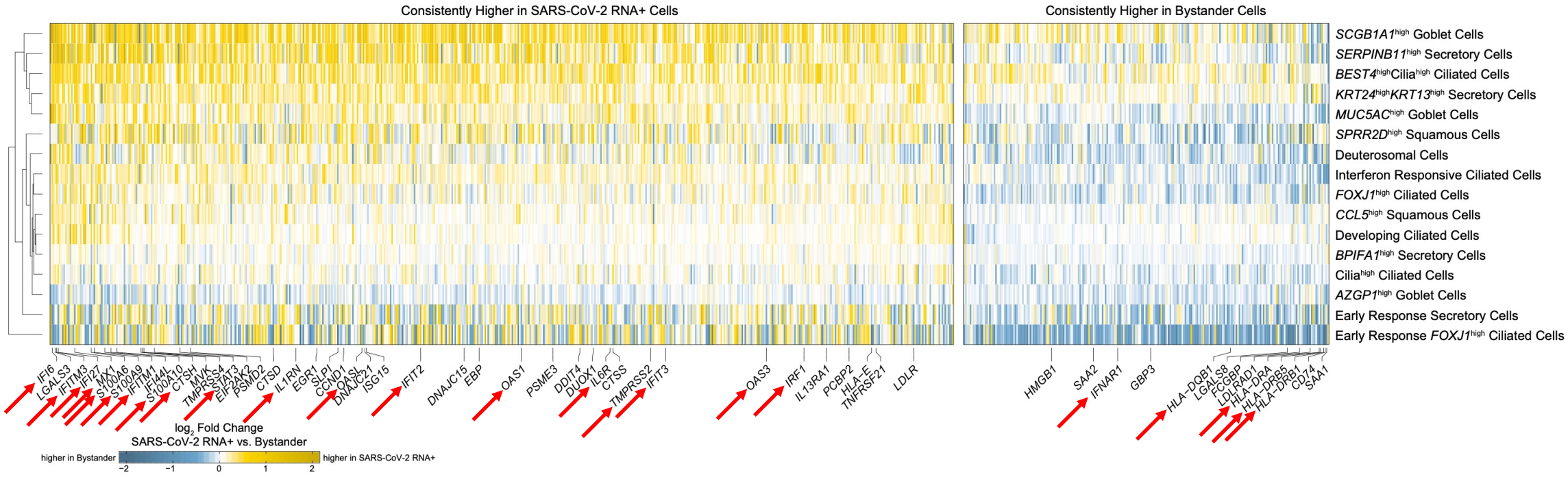
Host-pathogen annotations (cont'd): A COVID-19 example





Ziegler et al., *Cell*, 2021.

Host-pathogen annotations (cont'd): A COVID-19 example



Ziegler et al., *Cell*, 2021.

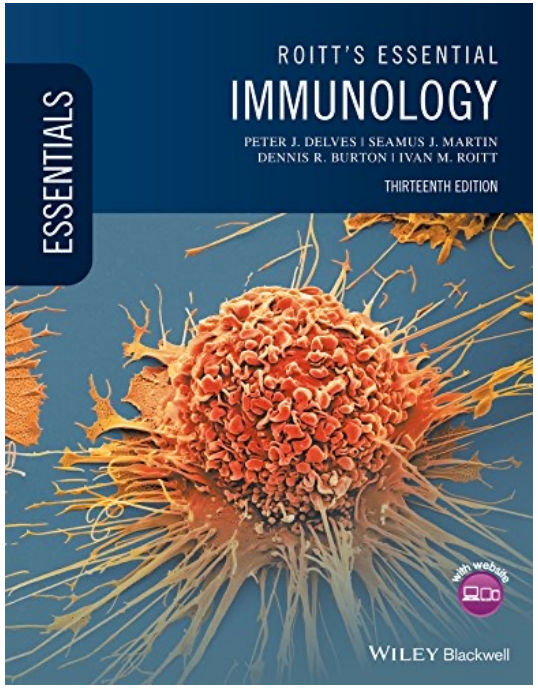
Approach 2: Databases and Reference atlases

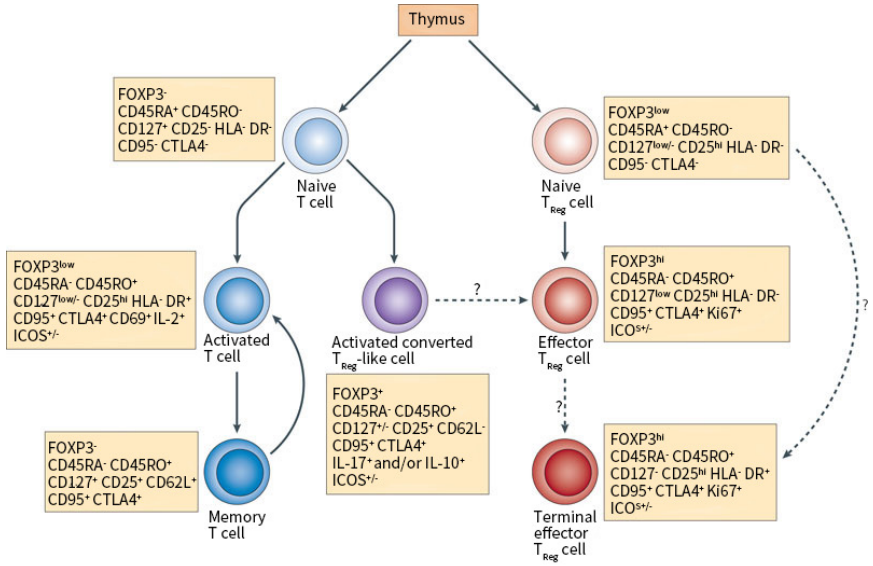
- Marker genes and databases
  - Literature
  - Tissue and protein DB
  - Single-cell atlases
- Automated tools
  - Correlation-based
  - Supervised classification

Approach 2: Marker genes and databases

Literature!

- Single-cell RNAseq is a new field but cell types are not!
- Book knowledge is important, know your model!
- Rely on experts and collaborators





Approach 2: Marker genes and databases

Tissue and protein DB

- Protein cell atlas - > https://www.proteinatlas.org/
- Immunological Genome Project - > https://www.immgen.org/
- GTEx (Genotype-Tissue Expression) -> https://gtexportal.org/home/

Approach 2: Marker genes and databases

Other organism

- Mouse https://www.emouseatlas.org
- Fly https://flybase.org/
- *C. elegans* https://www.wormatlas.org/
- Zebra fish https://bio-atlas.psu.edu/zf/

NCBI
https://www.ncbi.nlm.nih.gov/
The European Bioinformatics Institute (EMBL- EBI)
https://www.ebi.ac.uk/services

Approach 2: Single-cell atlases

Single-cell DB for manual annotation

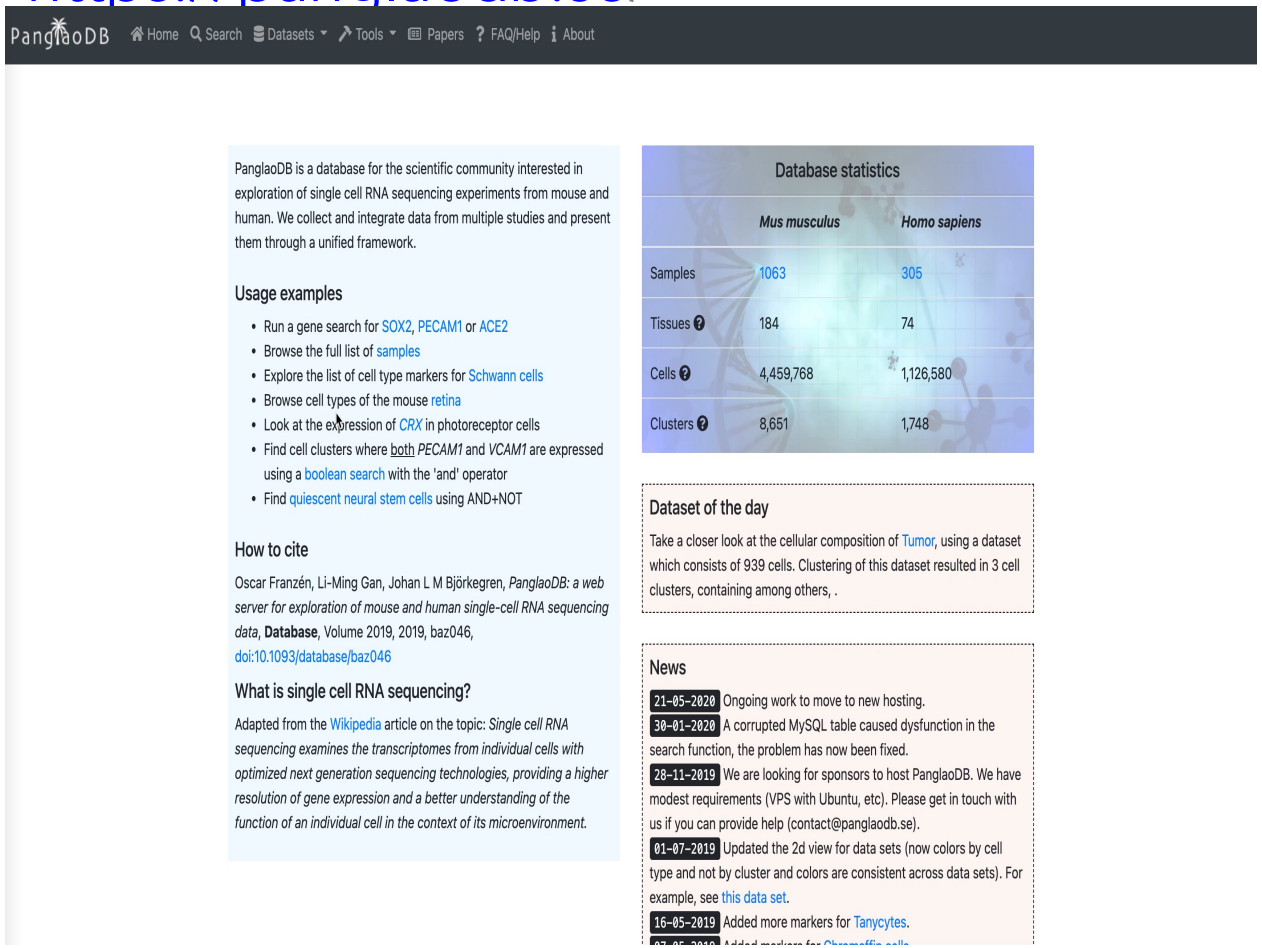
Meaning you have your set of genes and search for them one by one in these DB

- Human cell atlas (HCA) https://www.ebi.ac.uk/humancellatlas/project-catalogue/
- Tabula murris (Mouse atlas)
- Broad institute Single-cell portal https://singlecell.broadinstitute.org/
- PanglaoDB https://panglaodb.se/
- CellMarker http://biocc.hrbmu.edu.cn/CellMarker/index.jsp

Approach 2: Single-cell atlases

Single-cell DB for manual annotation

PanglaoDB https://panglaodb.se/



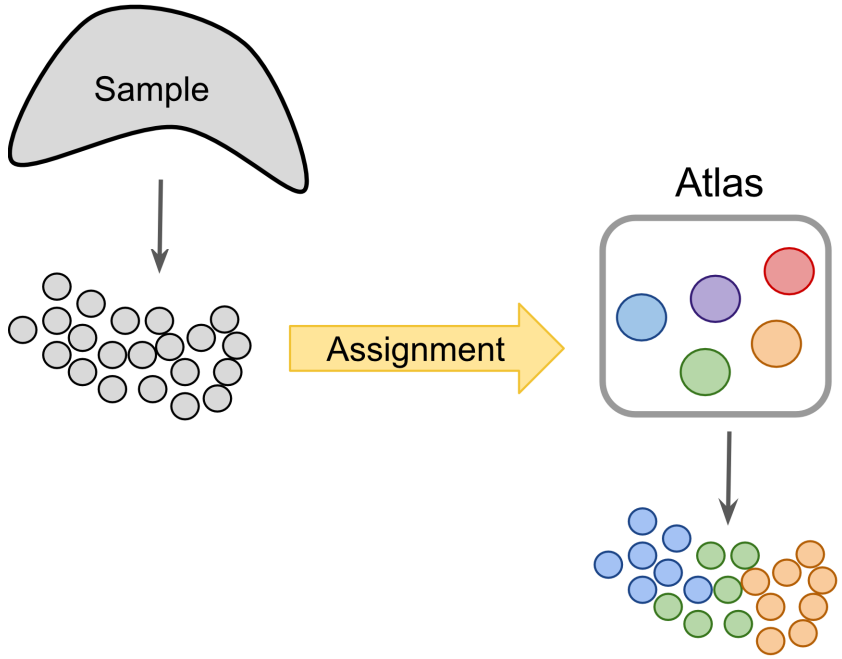
Approach 2: Automated tools

Project cells or clusters from a new sample onto a reference (Cell Atlas or previous study) to identify cells with novel/unknown identity ----> Cell type like BLAST method

- *IMPORTANT: A reference cell type information is needed*
- *Can be done intra-species*
- *Different-omics!*

Different approaches

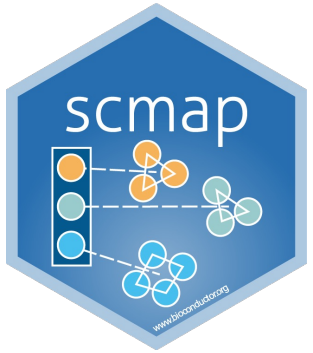
- Correlation-based
- Supervised classification

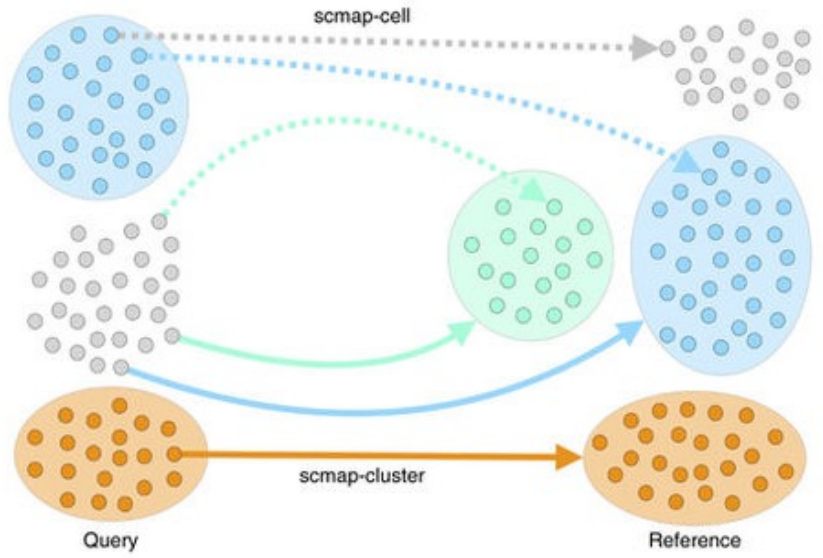


Approach 2: Automated tools

Approach	Tool name	Language	Computational approach
Correlation-based	scmap-cluster	R, web app	Cosine, Spearman, Pearson
	scmap-cell	R, web app	Cosine distance based kNN
	SingleR	R	Spearman
	CHETAH	R, Shiny app	Spearman + confidence
	scMatch	Python	Spearman, Pearson
	ClustifyR	R	Spearman, Pearson, Kendall, cosine
	CIPR	R, Shiny app	Dot product, Spearman, Pearson
Supervised classification-based	CaSTLe	R	XGBoost classifier
	Moana	Python	kNN-smoothing + SVM
	LAmbDA	Python	Multiple ML techniques
	superCT	Web app	Artificial Neural Network
	SingleCellNet	R	Random Forest
	Garnett	R	Elastic net regression
	scPred	R	SVM
	ACTINN	Python	Artificial Neural Network
	OnClass	Python	kNN and Bilinear Neural Network
	scClassify	R, Shiny app	Weighted kNN classifier
	scArches	Python	Autoencoder

Approach 2: Automated tools



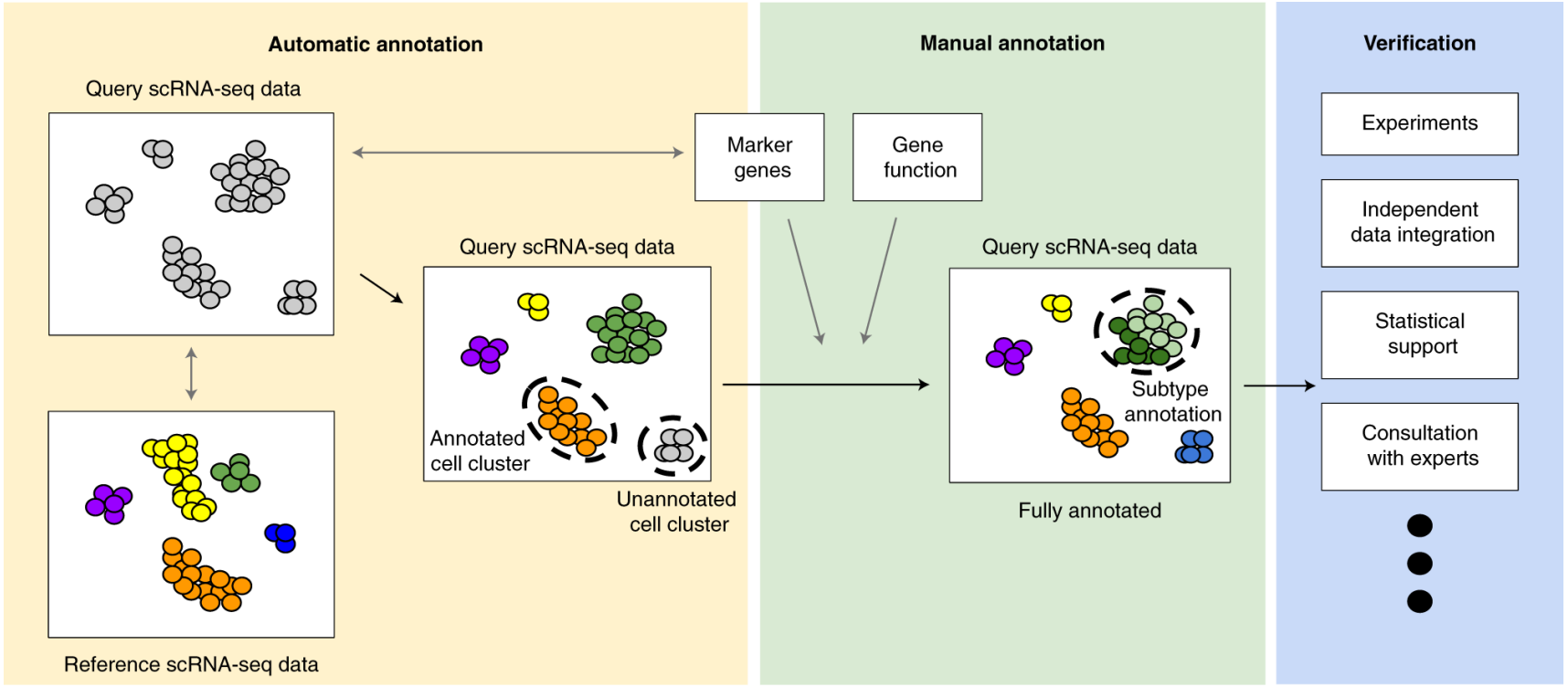


The search can be done by cluster or by cell , in each cluster/cell by its centroid (a vector of the median value of the expression of each gene) and measure the similarity between c and each cluster centroid or cell in the refence is calculated and ranked

Pasquini et al, 2021, Computational and Structural Biotechnology Journal

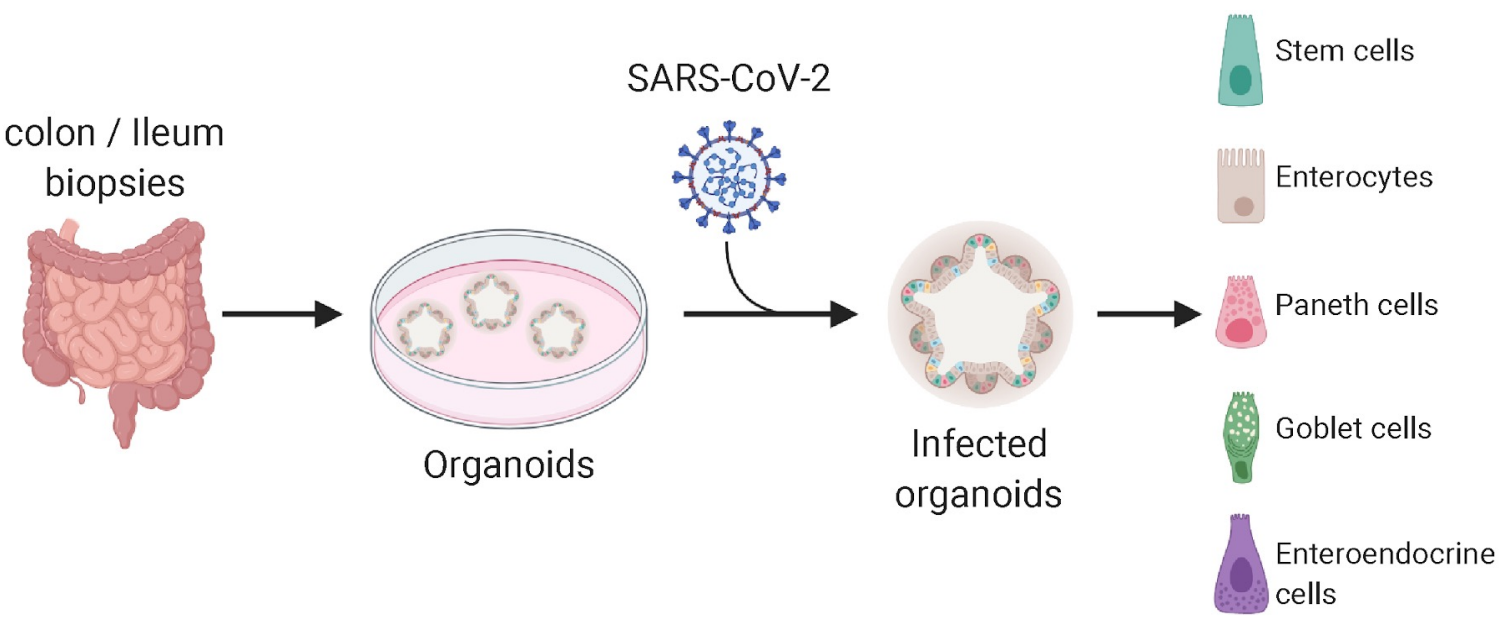
Overview

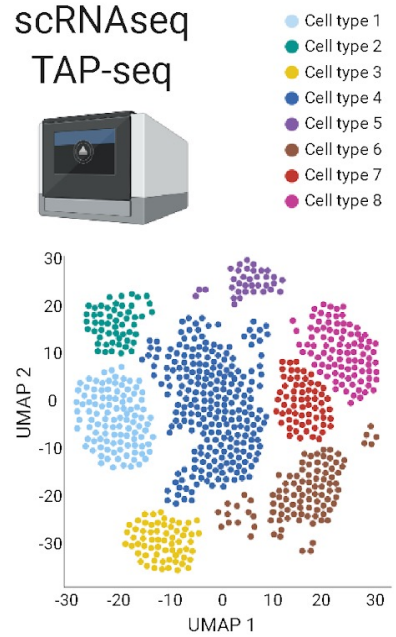
- An ideal cluster/cell type annotation process is composed of three major steps: automatic cell annotation, manual cell annotation and verification



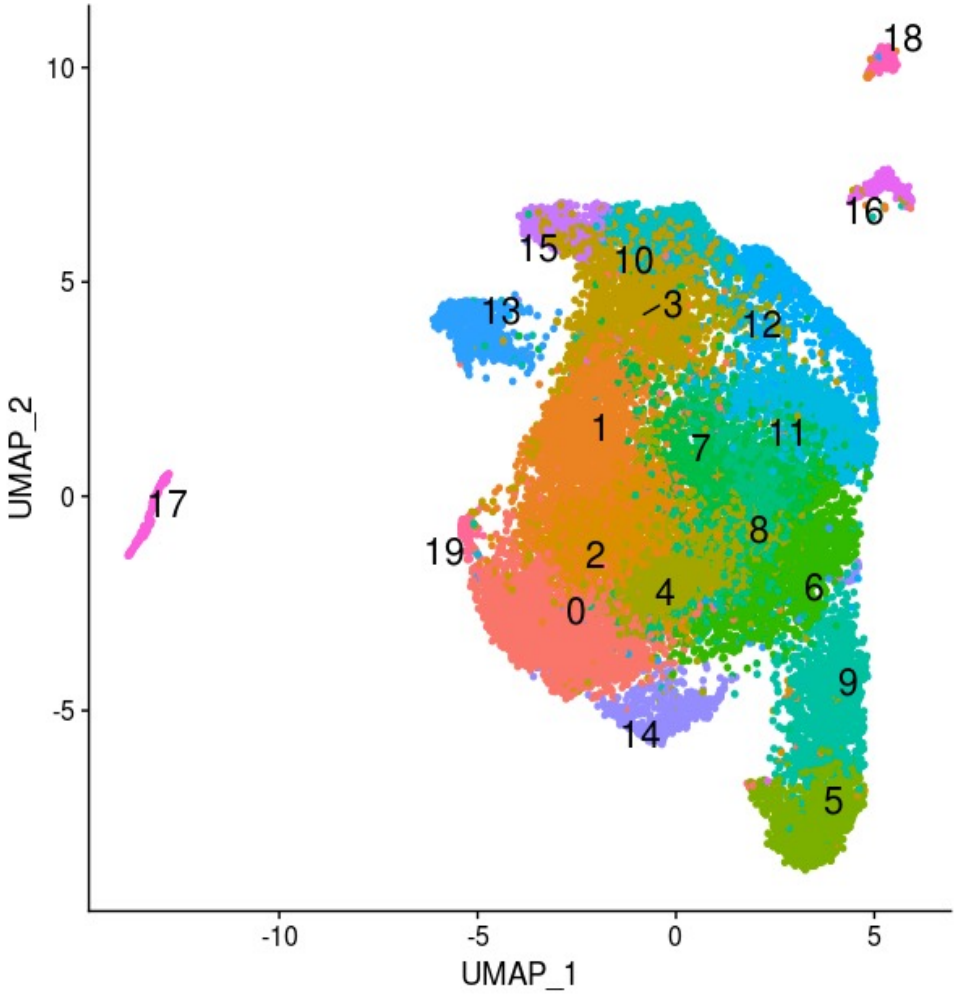
Clarke, Z.A., Andrews, T.S., Atif, J. *et al Nat Protoc* **16,** 2749–2764 (2021).

Example


colon / Ileum
biopsies
SARS-CoV-2
Organoids
Infected
organoids
Stem cells
Enterocytes
Paneth cells
Goblet cells
Enteroendocrine
cells


scRNAseq
TAP-seq
Cell type 1
Cell type 2
Cell type 3
Cell type 4
Cell type 5
Cell type 6
Cell type 7
Cell type 8
UMAP 2
UMAP 1
30
20
10
0
-10
-20
-30

Example



Cell

Volume 178, Issue 3, 25 July 2019, Pages 714-730.e22

CellPress

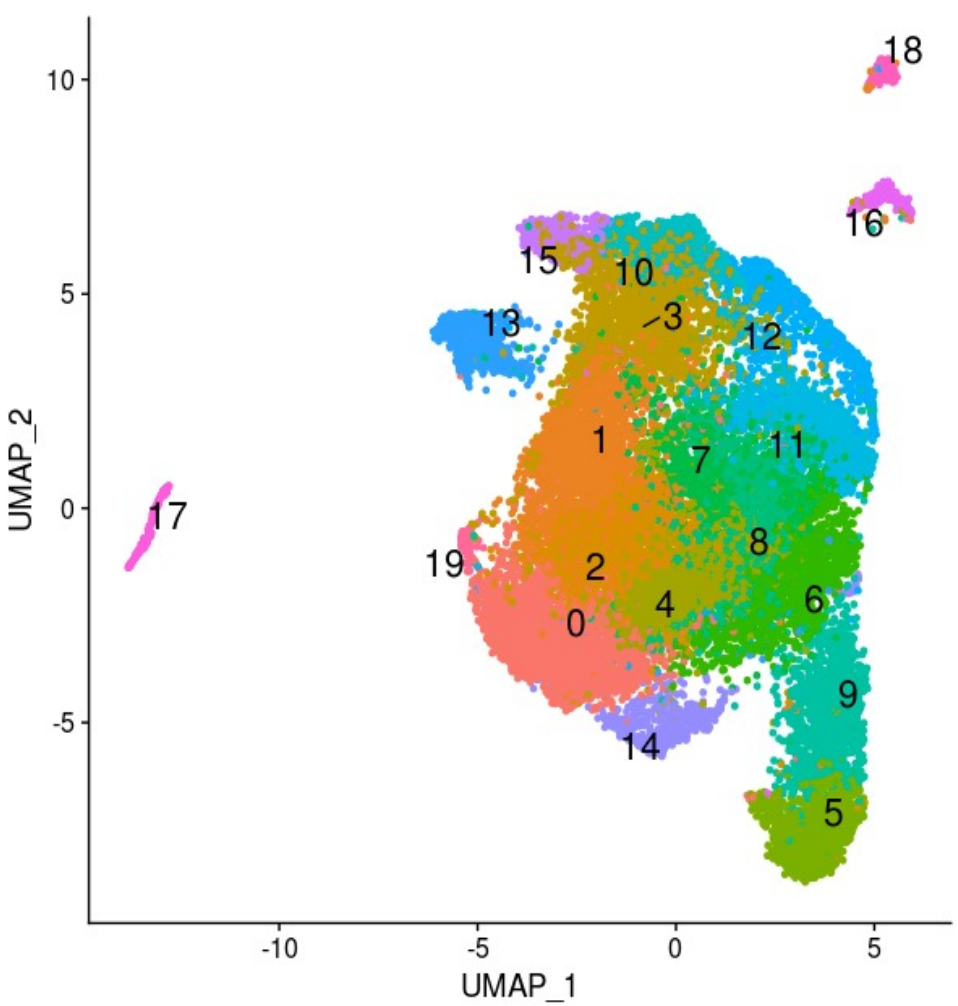
Resource

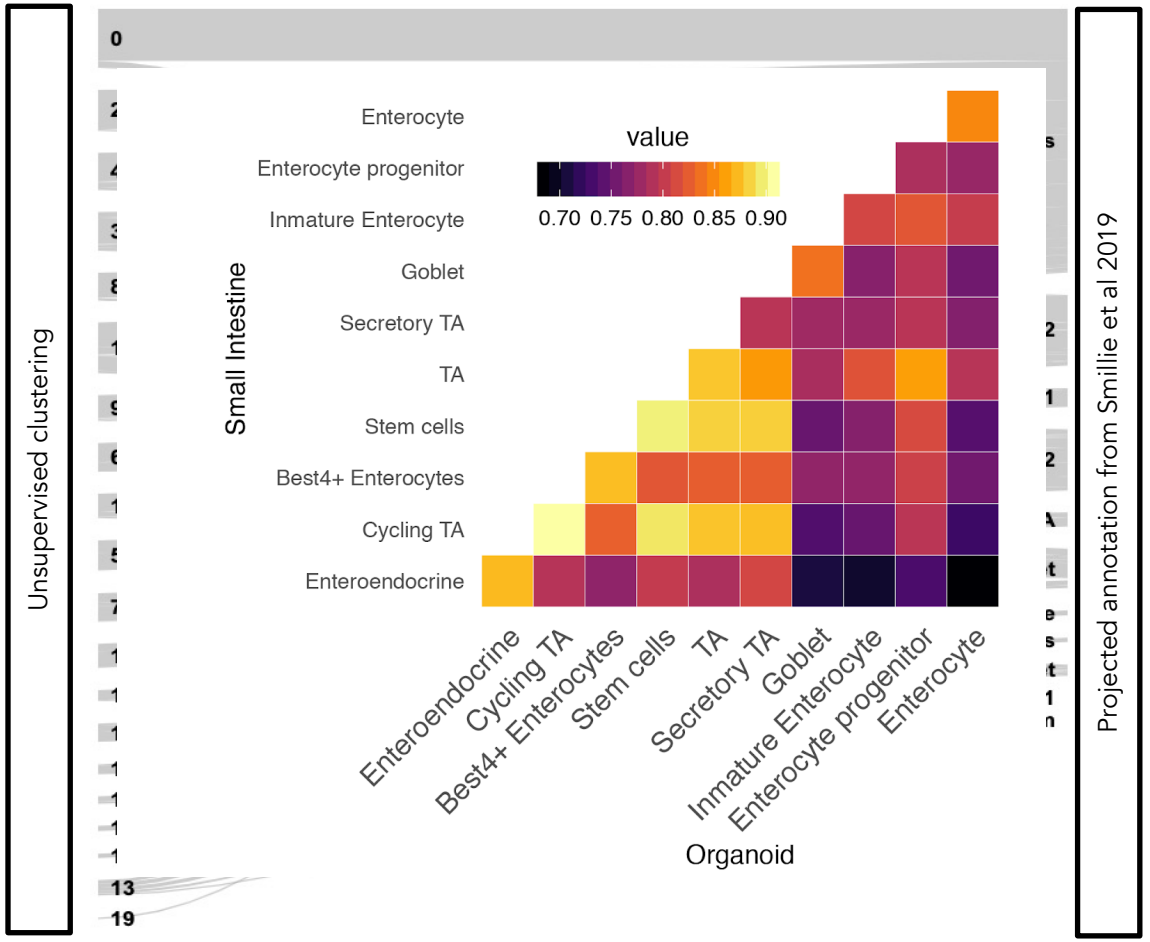
Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis

Christopher S. Smillie [1, 19], Moshe Biton [1, 2, 19], Jose Ordovas-Montanes [1, 3, 4, 5, 6, 7, 19], Keri M. Sullivan [8], Grace Burgin [1], Daniel B. Graham [2, 8, 9, 10, 11], Rebecca H. Herbst [1, 12], Noga Rogel [1], Michal Slyper [1], Julia Waldman [1], Malika Sud [1], Elizabeth Andrews [8], Gabriella Velonias [8], Adam L. Haber [1], Karthik Jagadeesh [1], Sanja Vickovic [1], Junmei Yao [14], Christine Stevens [9] ... Aviv Regev [1, 18, 20]

366.650 cells from human colon mucosa

Cluster Annotation

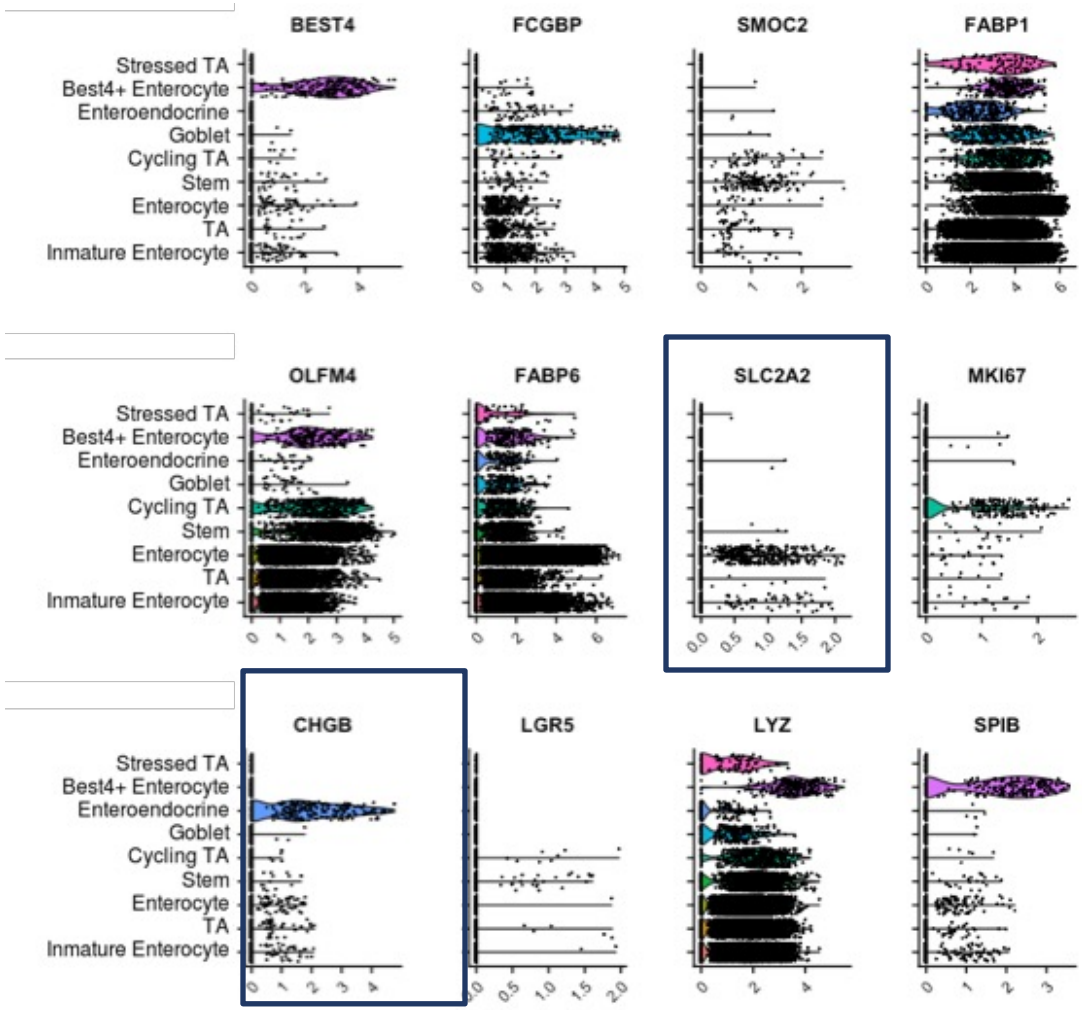


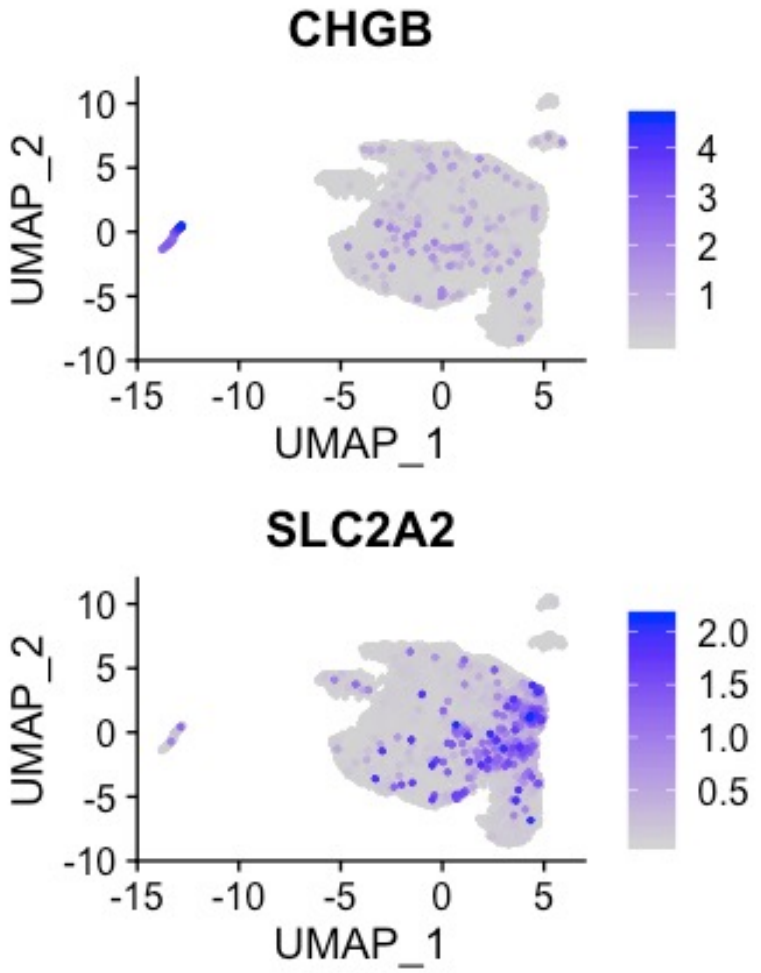


scMAP

B.

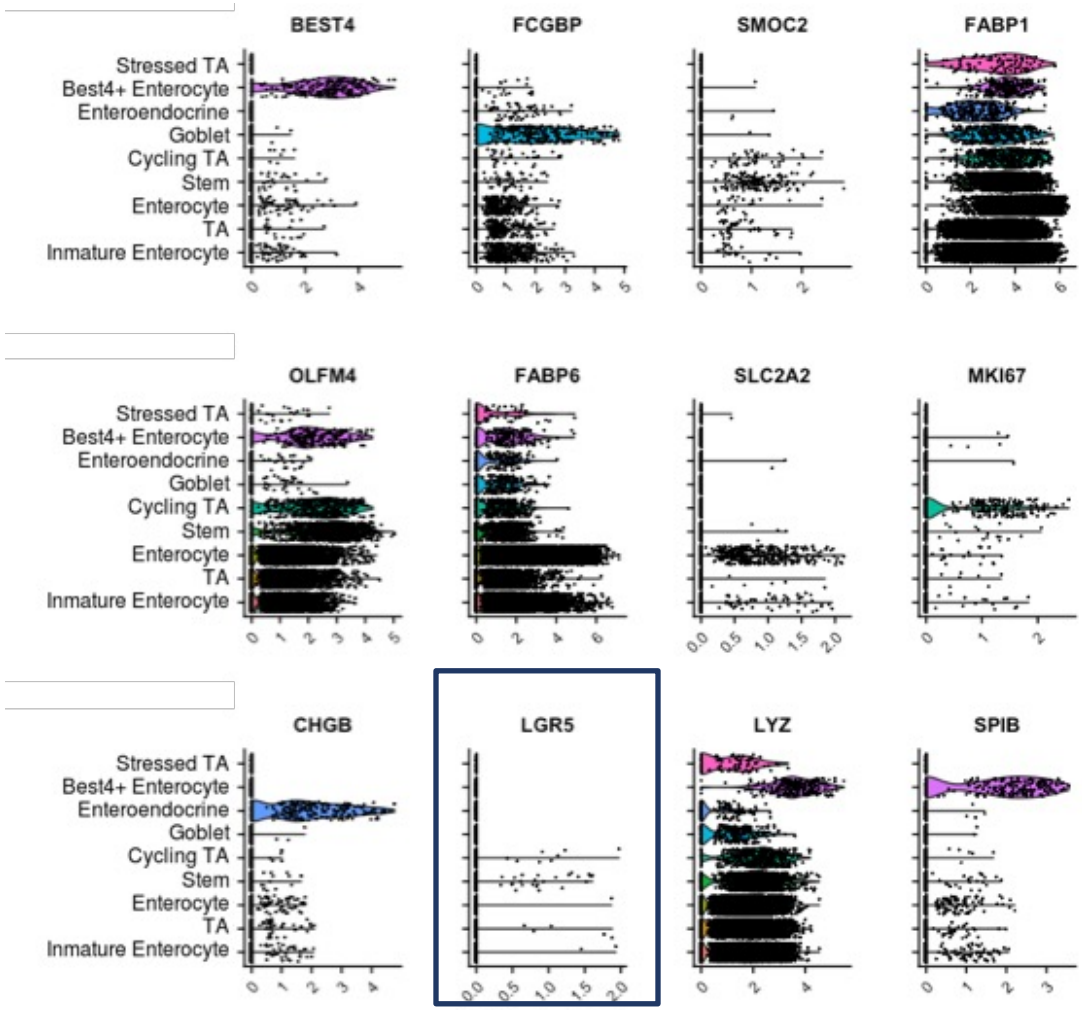
Markers for validation – cell type

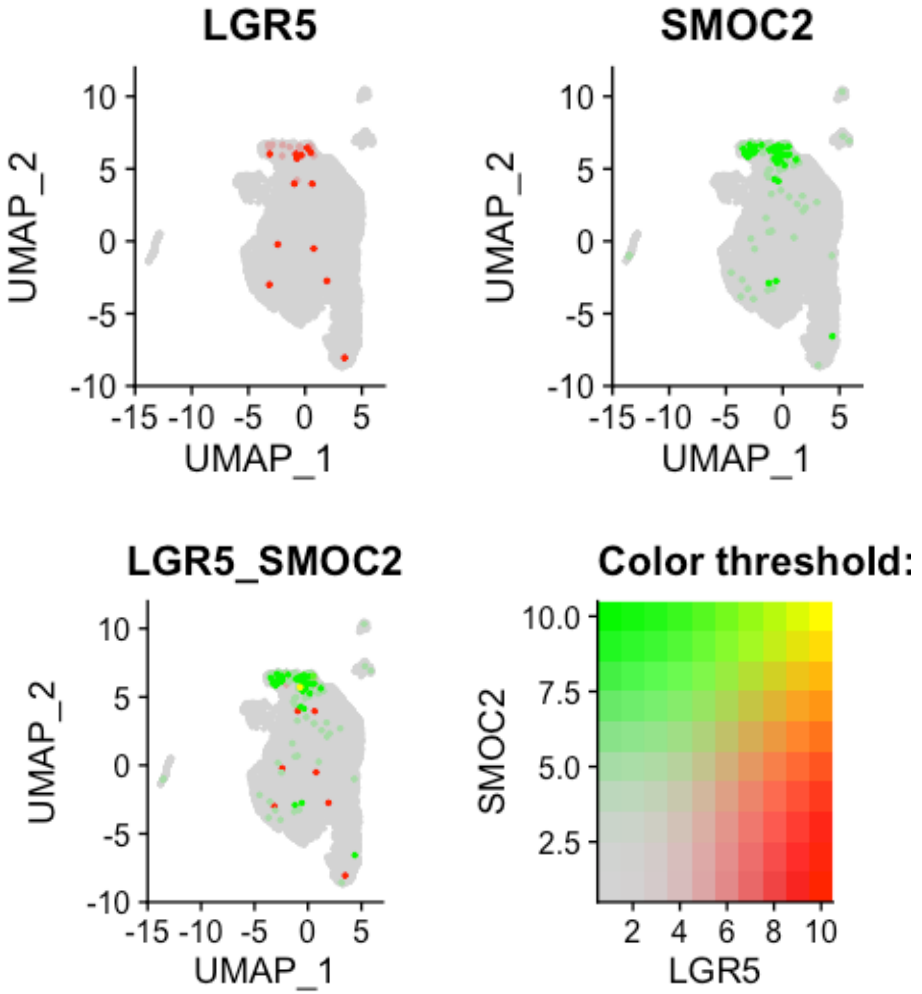




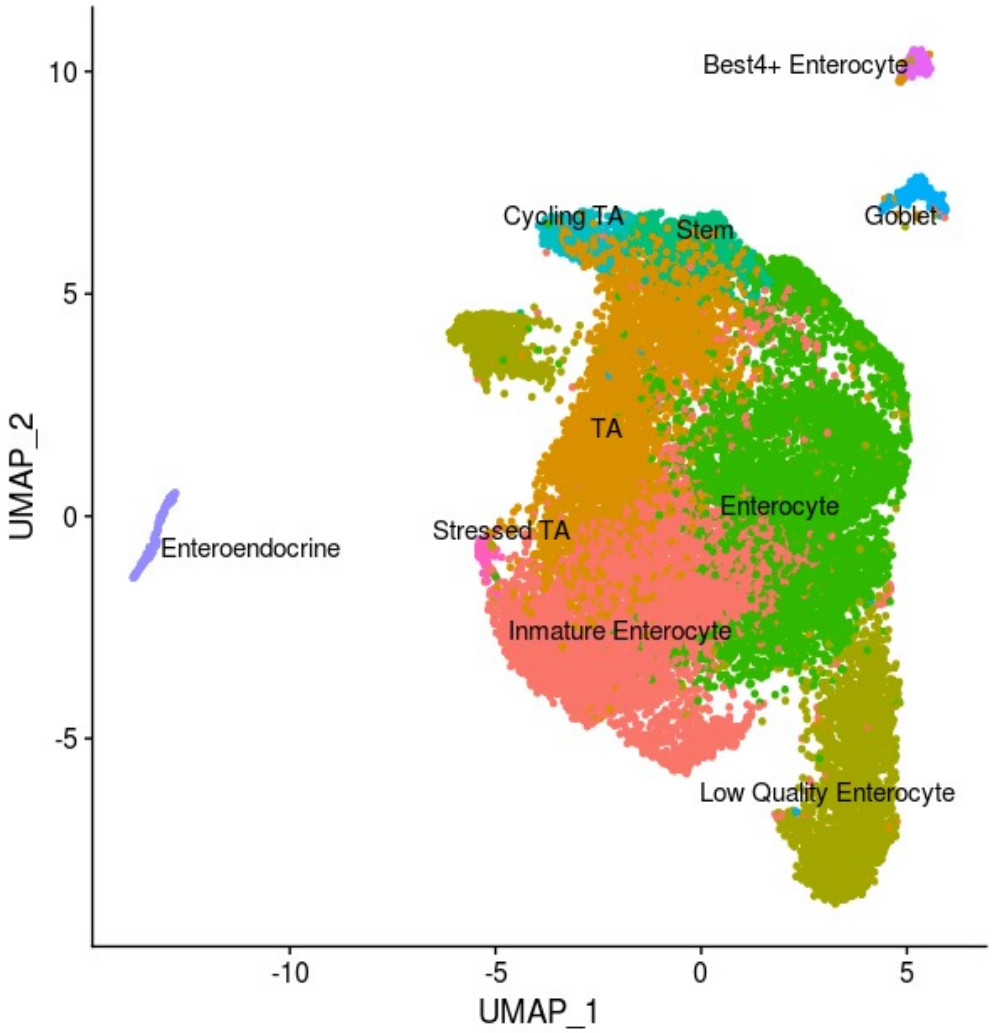
B.

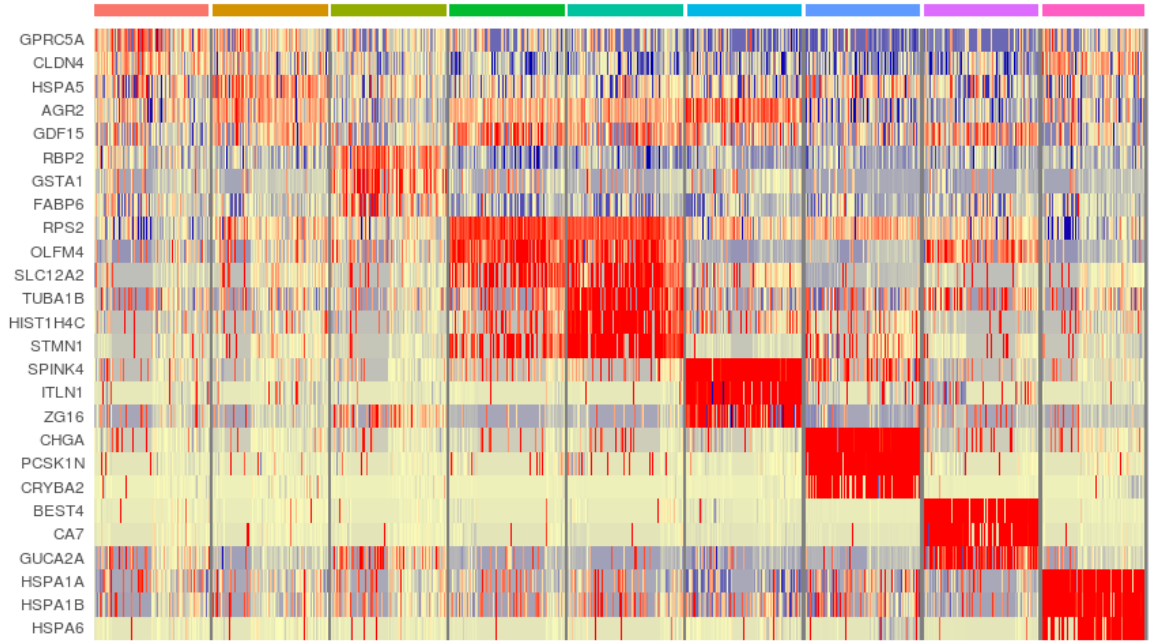
Markers for validation – cell type



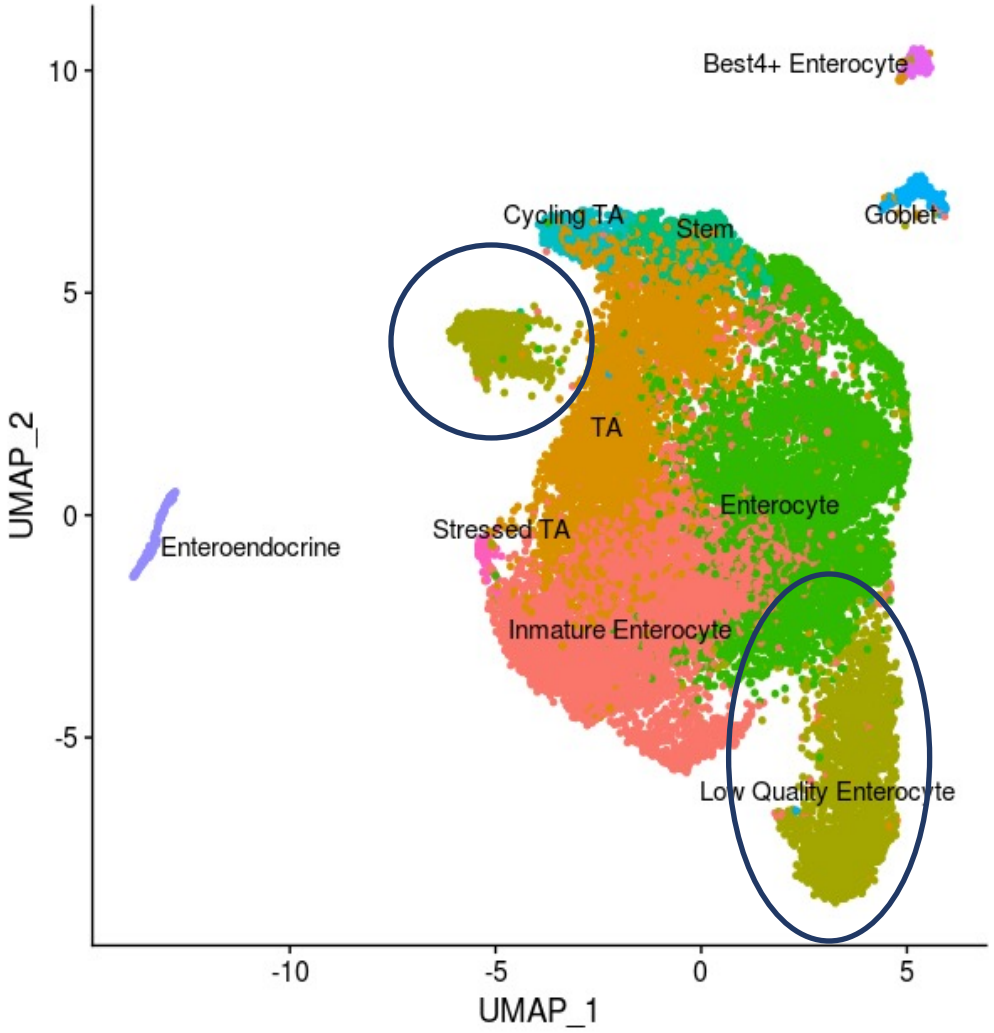


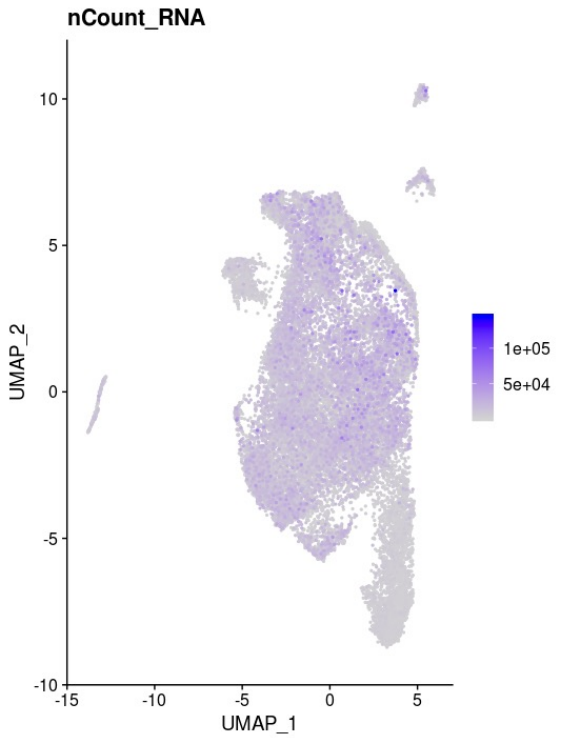
Cluster Annotation

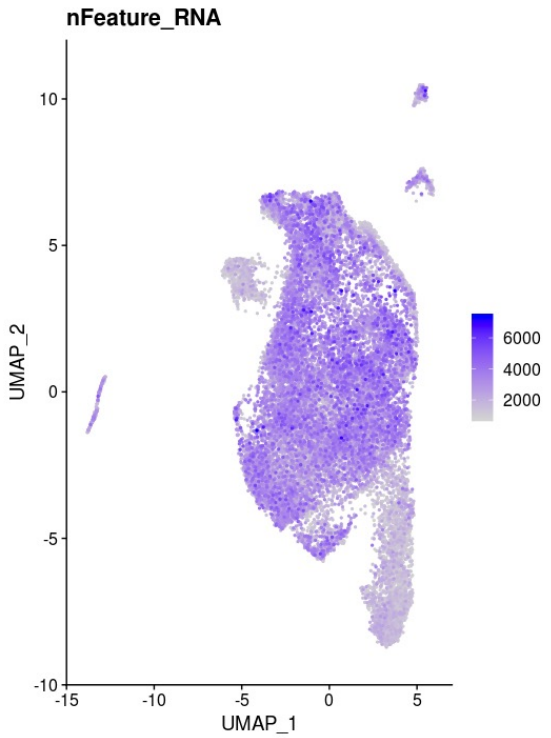




Cluster Annotation







Cluster Annotation