

Swiss Institute of  
Bioinformatics

Single cell transcriptomics data analysis

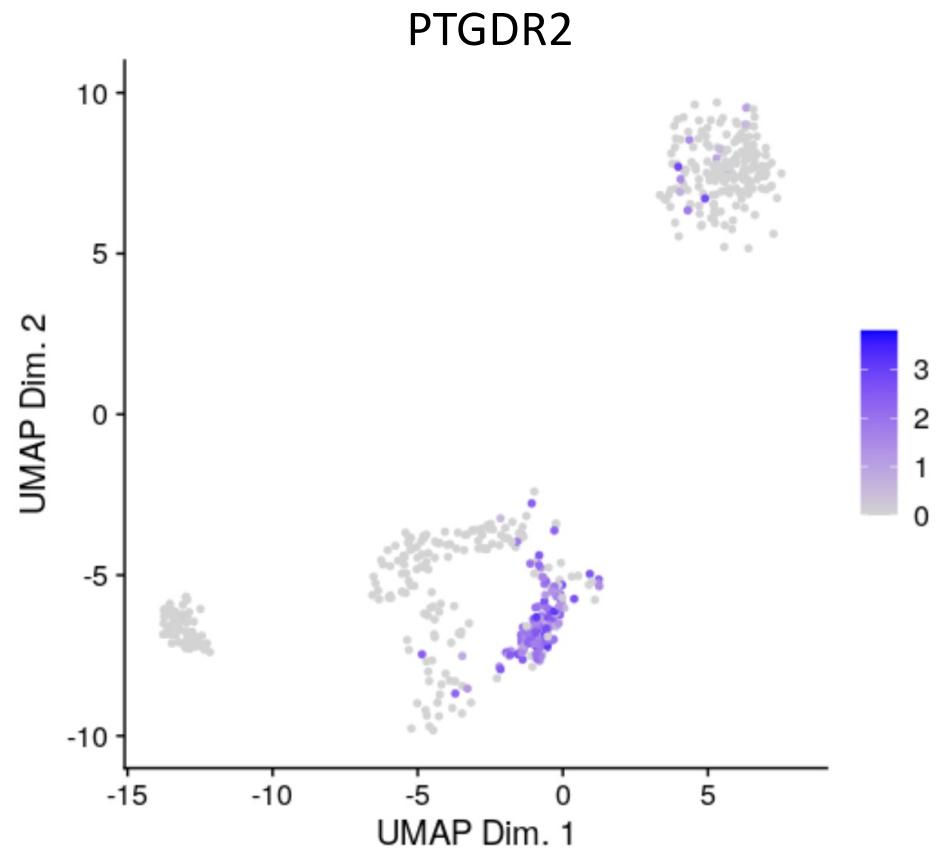
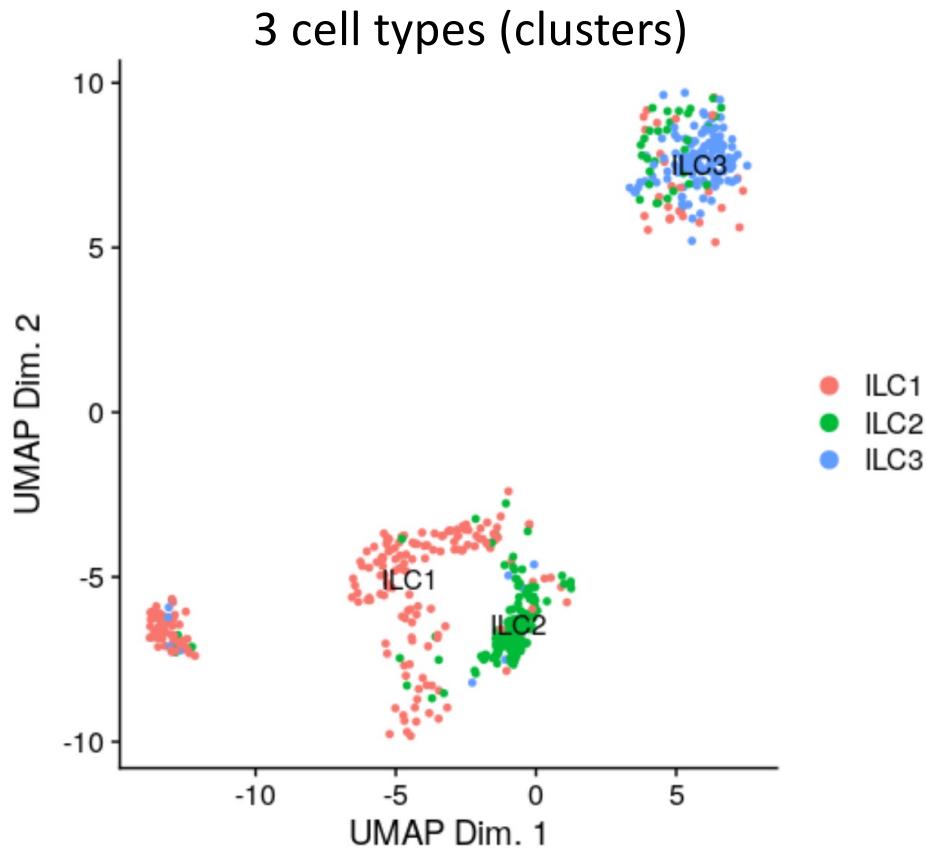
Marker gene identification  
&  
Differential gene expression analysis

# Two types of gene expression analysis:

- **Marker gene identification:**  
genes overexpressed by each cell type, cell cluster, ..., within the dataset => can *help with cell type annotation*
- **Differential gene expression analysis:**  
genes impacted by experimental conditions within a cell type, cell cluster, ..., etc

# Marker gene identification

Which genes are more (or less) expressed in one cell type than in the others?



# Marker gene identification

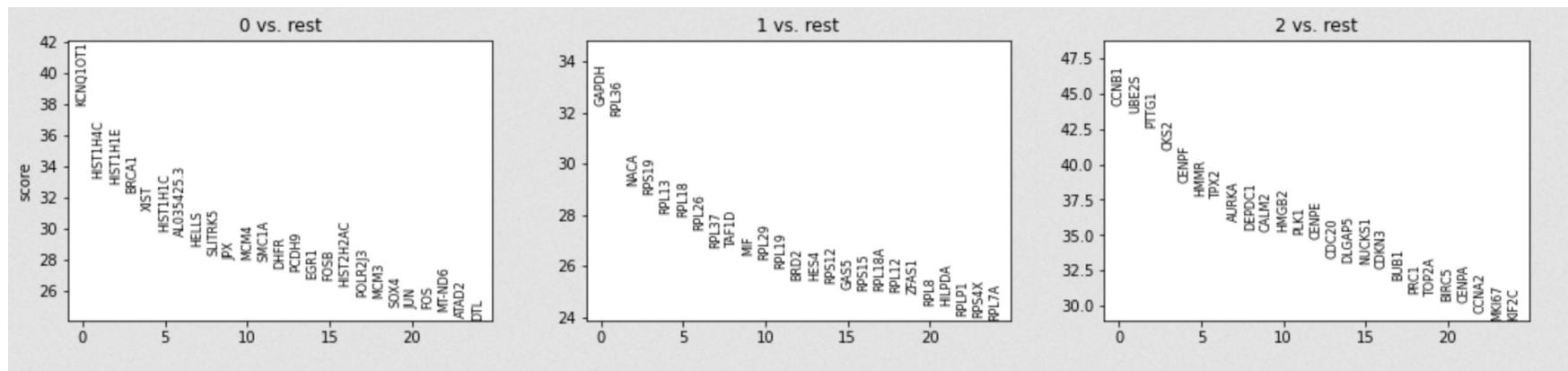
- Compare each cluster of cells against all other cells = 2 group-comparison
- **Wilcoxon Rank Sum test (Mann-Whitney U test)**

```
sc.tl.rank_genes_groups(adata,  
groupby="leiden", method="wilcoxon")
```

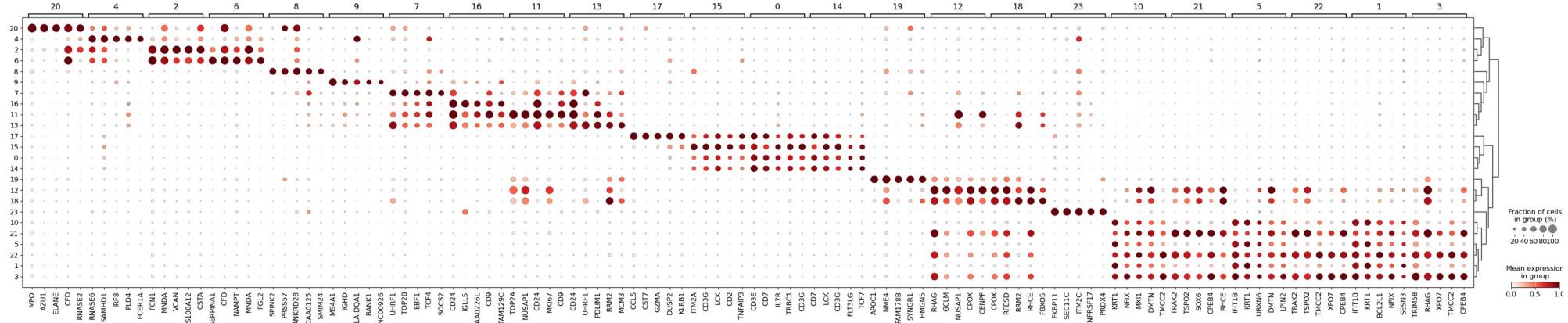
-> Returns an AnnData

- Plot the MW U statistic - the higher, the more significant:

```
sc.pl.rank_genes_groups(adata, n_genes=25)
```



# Marker gene identification



```
sc.pl.rank_genes_groups_dotplot(adata, groupby="leiden", n_genes=5)
```

Table of marker genes in each cluster versus all other cells - obtain  $\log_2$ (fold change), p-value and adjusted p-value

	Gene	scores	pval_adj	lfc	cluster
<b>87796</b>	zika	23.704487	0.0	8.062958	4
<b>87797</b>	SNHG15	16.631556	0.0	1.818314	4
<b>87798</b>	SDF2L1	16.453564	0.0	1.920571	4
<b>87799</b>	XBP1	16.216284	0.0	1.708001	4
<b>87800</b>	EIF4EBP1	15.350019	0.0	1.237626	4
...	...	...	...	...	...
<b>109739</b>	SOX4	-13.782868	0.0	-1.743386	4
<b>109740</b>	PGK1	-13.823959	0.0	-1.340903	4
<b>109741</b>	MSMO1	-13.977442	0.0	-2.380198	4
<b>109743</b>	SLC25A3	-16.256285	0.0	-1.478398	4
<b>109744</b>	XIST	-16.642595	0.0	-1.296944	4

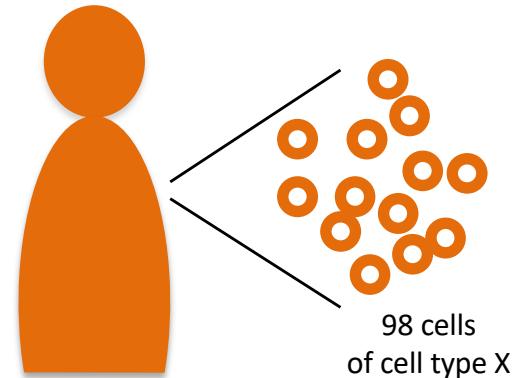
[https://github.com/mousepixels/sanbomics\\_scripts/blob/main/Sca-npy\\_intro\\_pp\\_clustering\\_markers.ipynb](https://github.com/mousepixels/sanbomics_scripts/blob/main/Sca-npy_intro_pp_clustering_markers.ipynb)

[https://www.youtube.com/watch?v=5HuOGZEu2HY&t=21s&ab\\_channel=Sanbomics](https://www.youtube.com/watch?v=5HuOGZEu2HY&t=21s&ab_channel=Sanbomics)

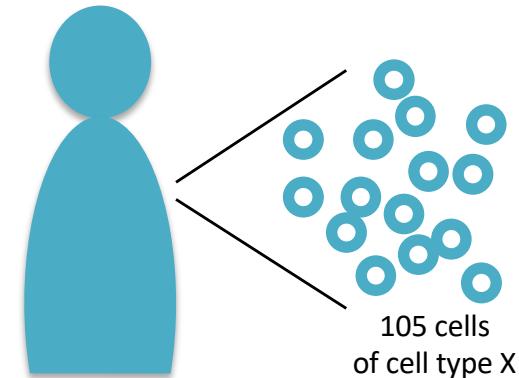
396 rows x 5 columns

# DGE analysis between 2 conditions :

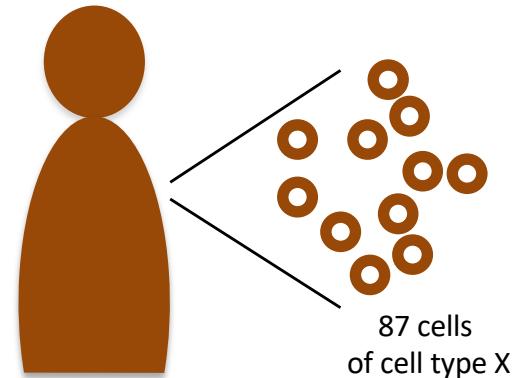
Healthy donor A



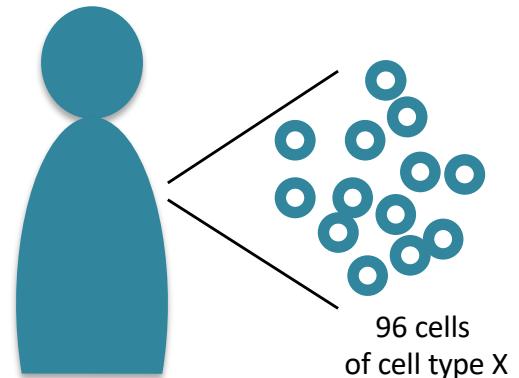
Patient A



Healthy donor B



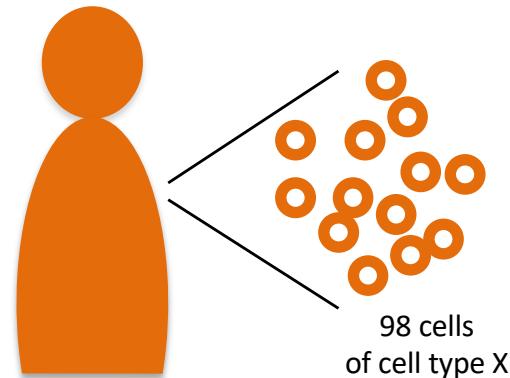
Patient B



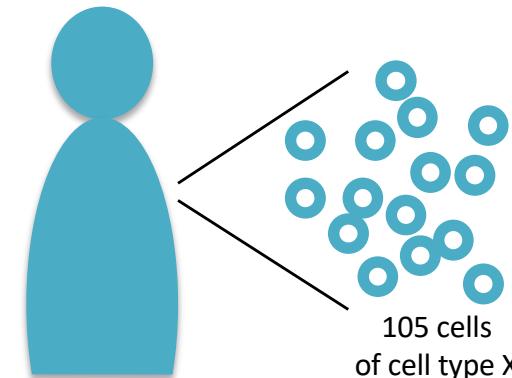
# DGE analysis between 2 conditions : Problem of pseudo-replication?

How many independent replicates do we have,  
~200 or 2 replicates per condition?

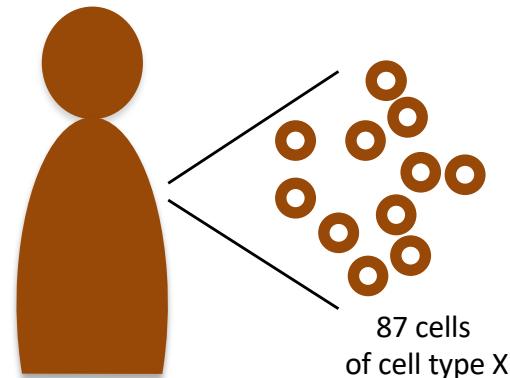
Healthy donor A



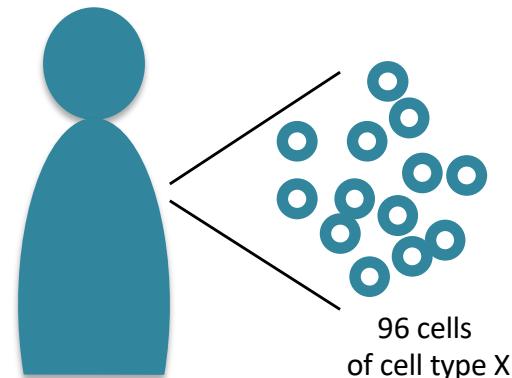
Patient A



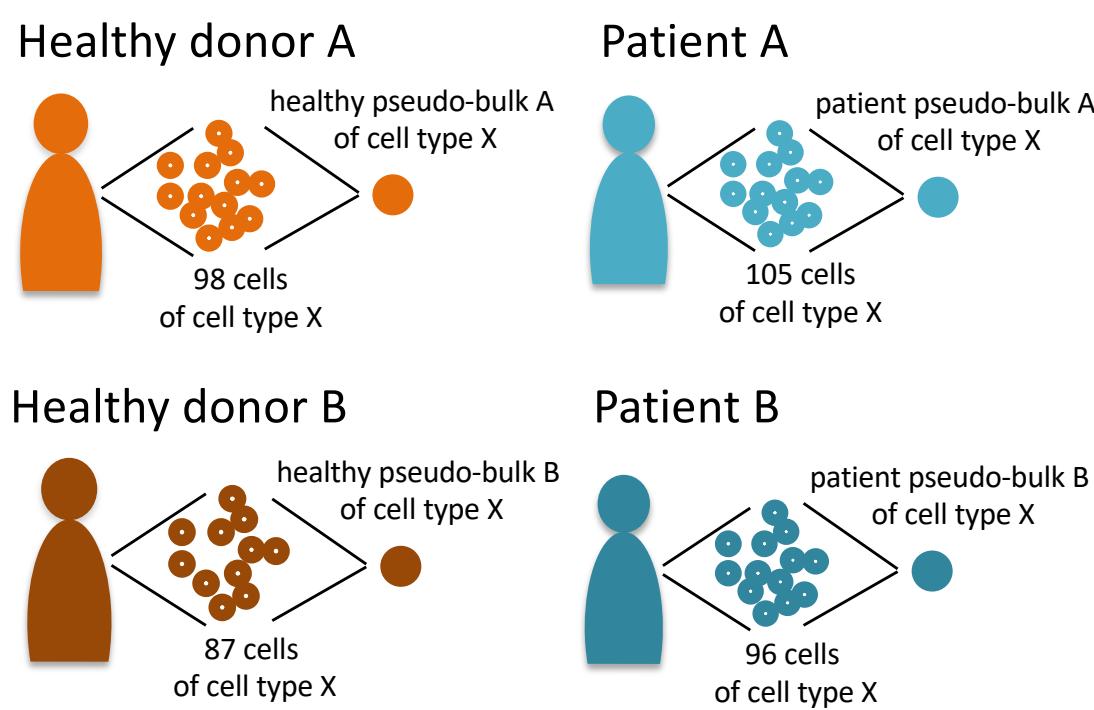
Healthy donor B



Patient B



# Pseudo-bulk DE analysis



**For cell type/cluster X:**  
From a matrix of  
386 cells  $\times$  33694 genes

To a matrix of  
4 pseudo-bulk samples  $\times$  33694 genes

Perform a DGE analysis of  
patient vs healthy with n=2 per condition

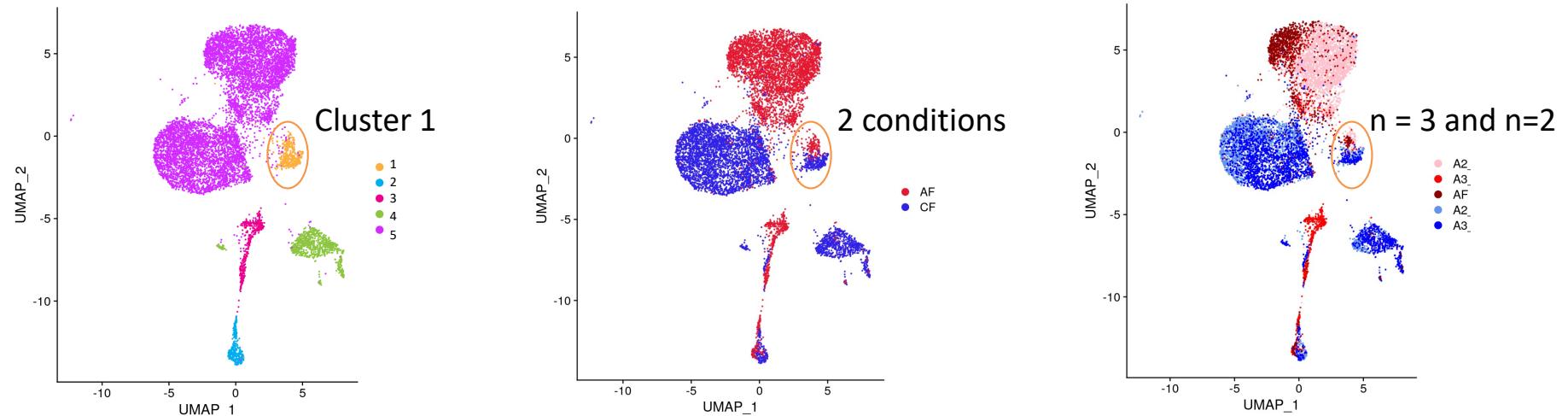
Repeat for every individual  
cell type/cluster

# PyDESeq2

[https://pydeseq2.readthedocs.io/en/latest/auto\\_examples/index.html](https://pydeseq2.readthedocs.io/en/latest/auto_examples/index.html)

- Implementation of R's DESeq2 package: Method designed for bulk RNA seq analysis
- **Wald test:** DESeq2 models count data using a negative binomial distribution. It fits a GLM to the counts of each gene. The Wald test is used to determine the significance of individual coefficients in the GLM fitted to each gene.  
 $\text{counts}_i \sim \text{condition} + \text{other covariates}$
- It can incorporate factors such as experimental conditions (multi-factor, factorial, ...), batch effects and other covariates.

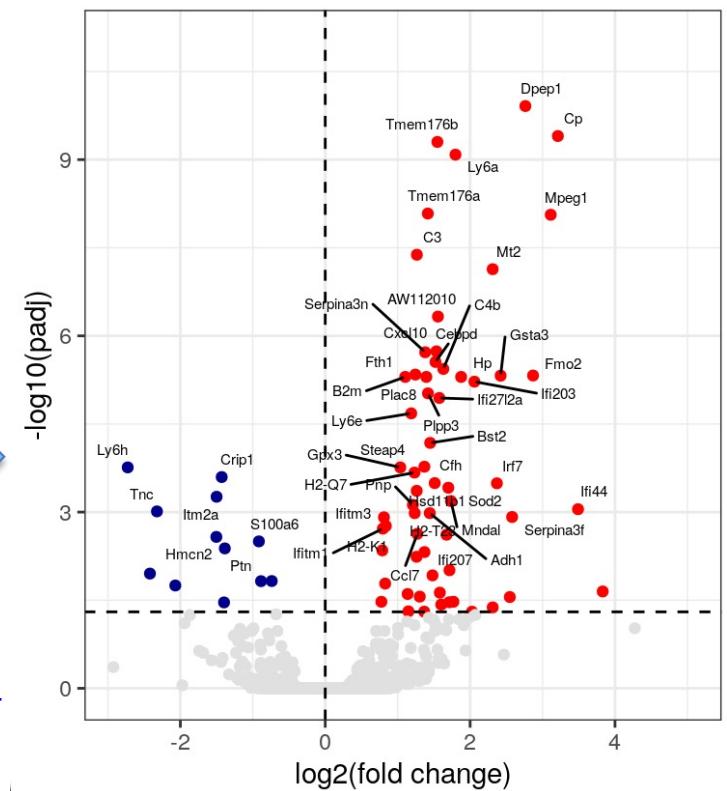
# DGE analysis between 2 conditions :



Sum counts of all cells of cluster 1 per sample:

	Xkr4	Gm1992	Gm3738	Rp1	Sox17	Gm3732	..
A2_AF	18	1	240	0	4	68	..
A3_AF	25	0	489	0	5	45	..
AF	40	2	500	0	1	32	..
A2_CF	70	0	407	0	0	45	..
A3_CF	36	2	230	0	3	23	..

DGE with  
PyDESeq2



[https://github.com/mousepixels/sanbomics\\_scripts/blob/main/pseudobulk\\_p\\_yDeseq2.ipynb](https://github.com/mousepixels/sanbomics_scripts/blob/main/pseudobulk_p_yDeseq2.ipynb)

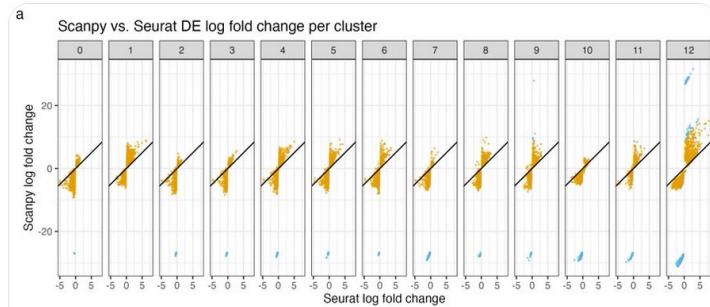
[https://www.youtube.com/watch?v=Ee0PQUwVH8Q&ab\\_channel=Sanbomics](https://www.youtube.com/watch?v=Ee0PQUwVH8Q&ab_channel=Sanbomics)

# $\log_2$ (fold change) discrepancy?

 **Lior Pachter**  
@lpachter

A  on why Seurat and Scanpy's log fold change calculations are discordant. 1/

(based on the Supplementary Notes from [biorxiv.org/content/10.1101/169438](https://www.biorxiv.org/content/10.1101/169438)).



## Seurat formula:

$$R_g = \log_2\left(\frac{1}{n_1} \sum_{i \in G_1} (\exp(Y_{ig}) - 1) + 1\right) - \log_2\left(\frac{1}{n_2} \sum_{i \in G_2} (\exp(Y_{ig}) - 1) + 1\right),$$

## Scanpy formula:

$$P_g = \log_2(\exp(\frac{1}{n_1} \sum_{i \in G_1} Y_{ig}) - 1 + \epsilon) - \log_2(\exp(\frac{1}{n_2} \sum_{i \in G_2} Y_{ig}) - 1 + \epsilon),$$

where  $\epsilon = 10^{-9}$ .

where  $Y_{ig}$  are the log-transformed expression values for cell  $i$  and gene  $g$ ,  $G_1$  and  $G_2$  are the indices for two groups of cells, and  $n_1$  and  $n_2$  are the numbers of cells in the respective groups.

<https://twitter.com/lpachter/status/1694387749967847874>

<https://divingintogeneticsandgenomics.com/post/do-you-really-understand-log2fold-change-in-single-cell-rnaseq-data/>

# Once we have identified marker or DE genes, what do we do?

scRNA sequencing pipeline

Differential expression analysis

Enrichment analysis

Several methods available, e.g.:

- over-representation analysis (ORA)
- gene set enrichment analysis (GSEA)

**Goal:** to gain biologically-meaningful insights from long gene lists

- Pathways from a collection like KEGG or Gene Ontology?
- Transcription factor targets ?
- Custom gene list from a publication?
- Genes associated with a disease ?
- Etc...

# Enrichment analysis

- GSEApY:

<https://gseapy.readthedocs.io/en/latest/introduction.html>

Tutorial:

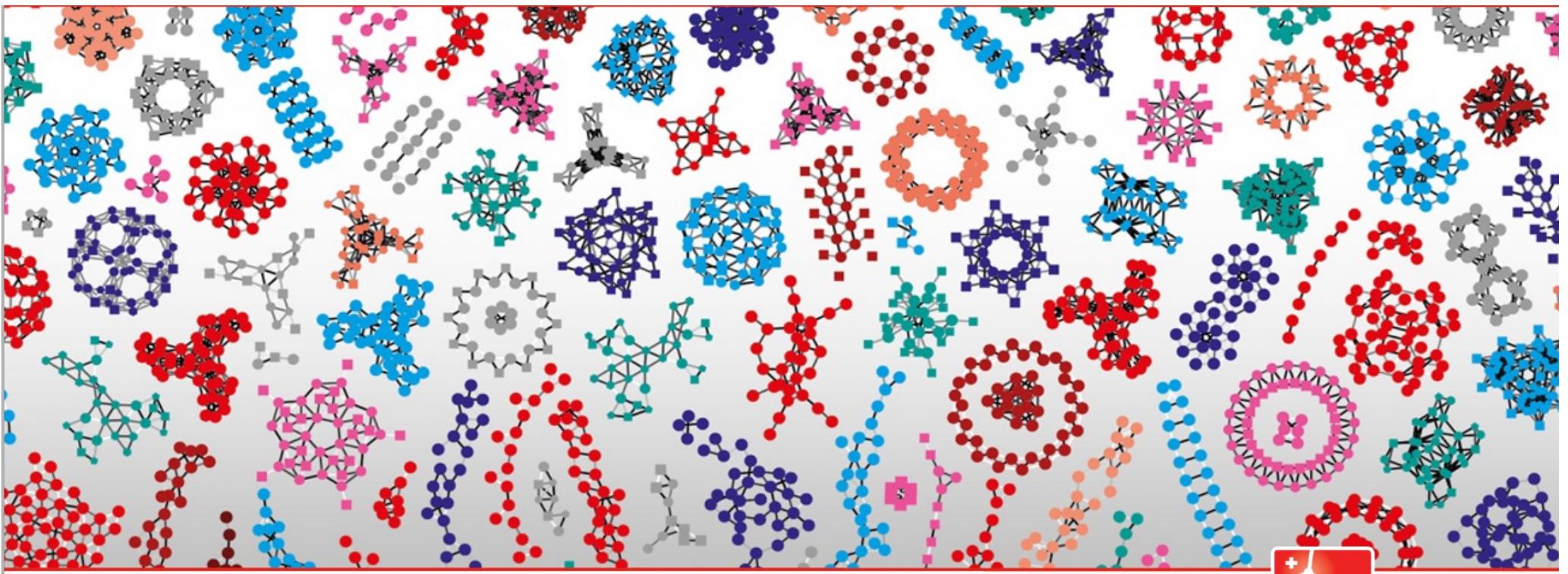
[https://github.com/mousepixels/sanbomics\\_scripts/blob/main/GSEA\\_in\\_python.ipynb](https://github.com/mousepixels/sanbomics_scripts/blob/main/GSEA_in_python.ipynb)

[https://www.youtube.com/watch?v=yOQcrUMCALw&t=302s&ab\\_channel=Sanbomics](https://www.youtube.com/watch?v=yOQcrUMCALw&t=302s&ab_channel=Sanbomics)

[https://nbisweden.github.io/workshop-scRNASeq/labs/scanpy/scanpy\\_05\\_dge.html#meta-dge\\_gsa\\_hyper](https://nbisweden.github.io/workshop-scRNASeq/labs/scanpy/scanpy_05_dge.html#meta-dge_gsa_hyper)

- (R: <https://sib-swiss.github.io/enrichment-analysis-training/>)

# Question on marker gene/DGE analysis



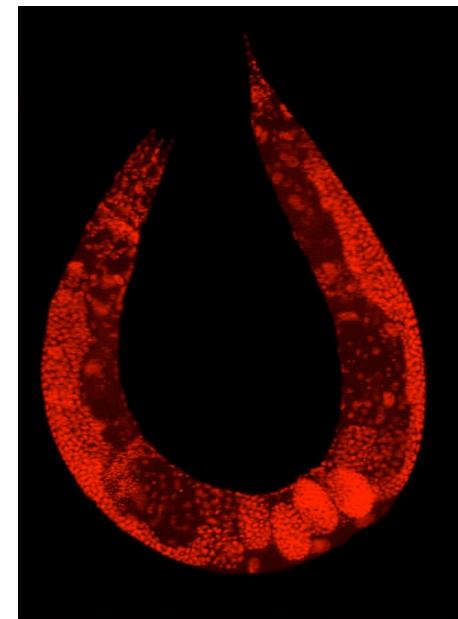
Swiss Institute of  
Bioinformatics

# Single cell transcriptomics data analysis

## Cell type annotation

# What is a “cell type”?

- Fundamental unit of life
- Originally defined in terms of function, location tissue type, cell morphology
- Later extended to
  - presence/absence of cell surface markers
  - gene expression (molecular profile)
- Currently very much less fixed
  - cell cycle phase
  - migration state
  - differentiation: cell state

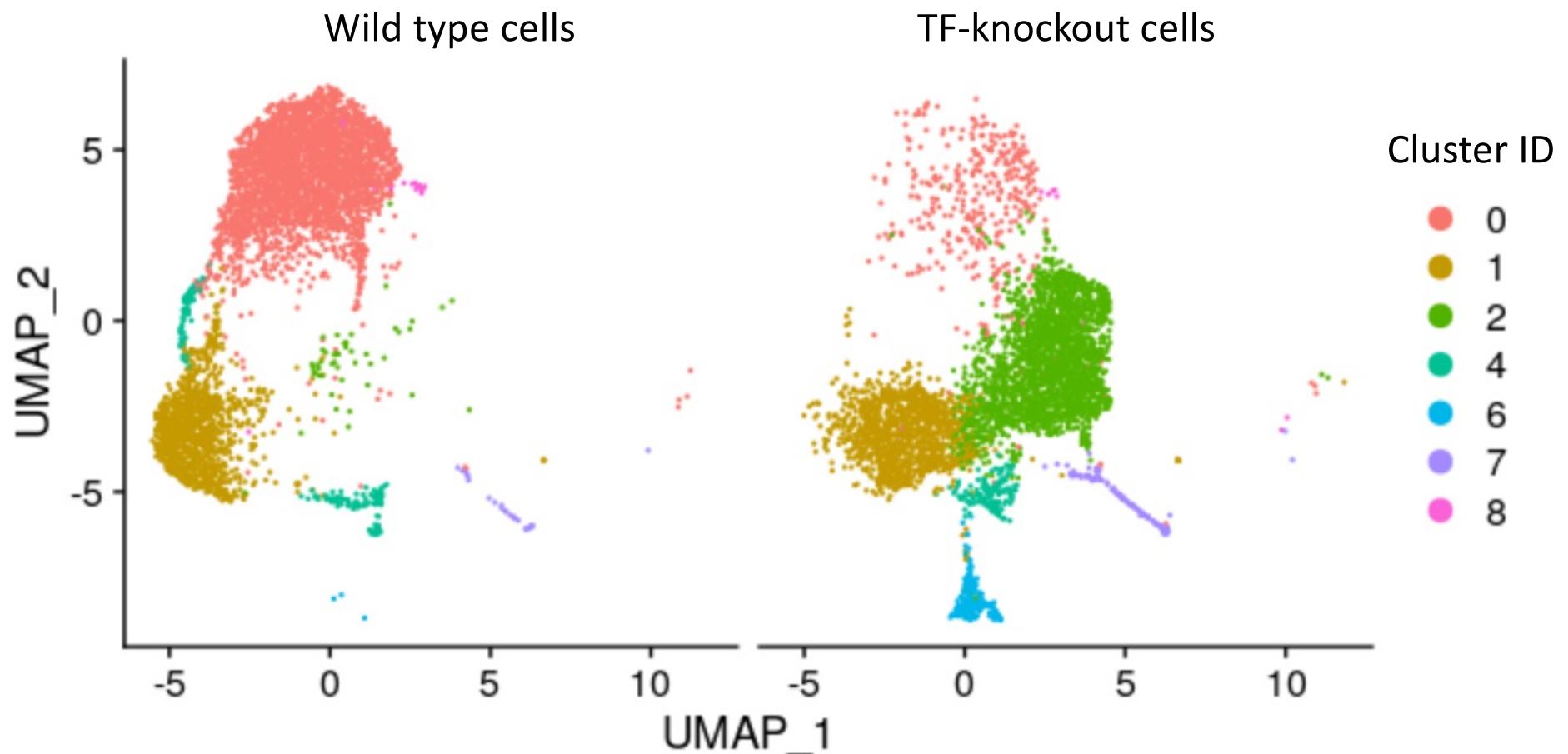


Wild-type *C. elegans* hermaphrodite  
stained to highlight the nuclei  
of all cells

# Why should we identify cell types?

- Samples are heterogeneous (in general)
- Tumor sample: how much do they differ from normal cell types?
- Find new cell types which have been missed by using “standard” surface markers
- Follow cell fate and determine cell differentiation mechanisms
- To determine which cell types might communicate with each other
- To compare the abundance of cell types in different conditions
- ...

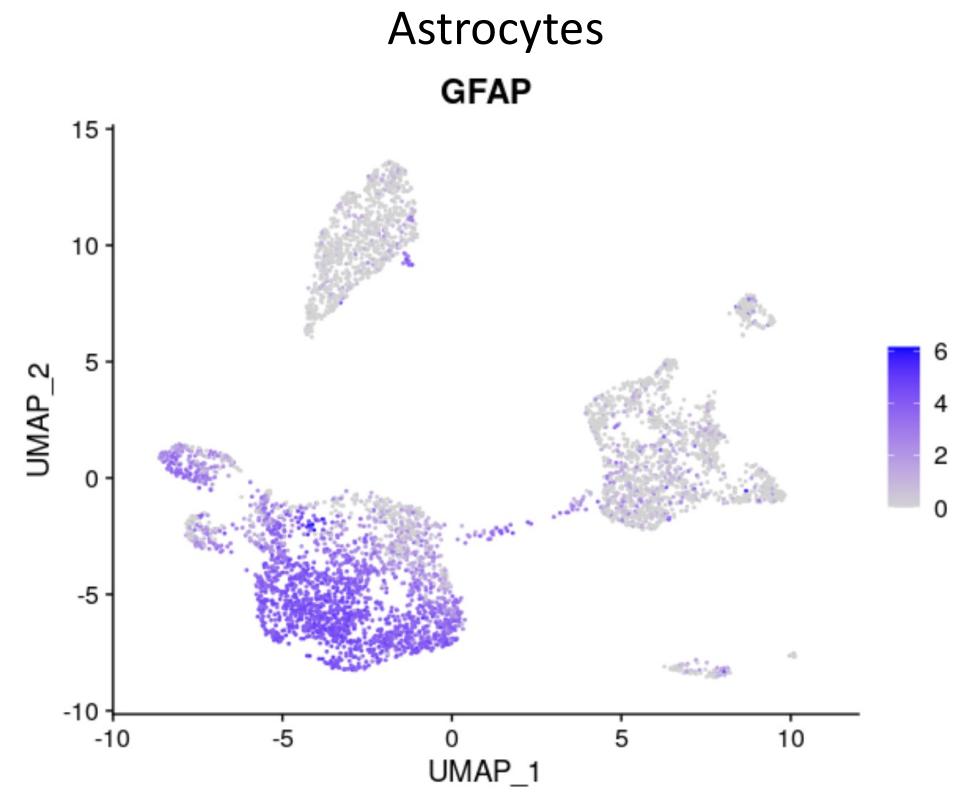
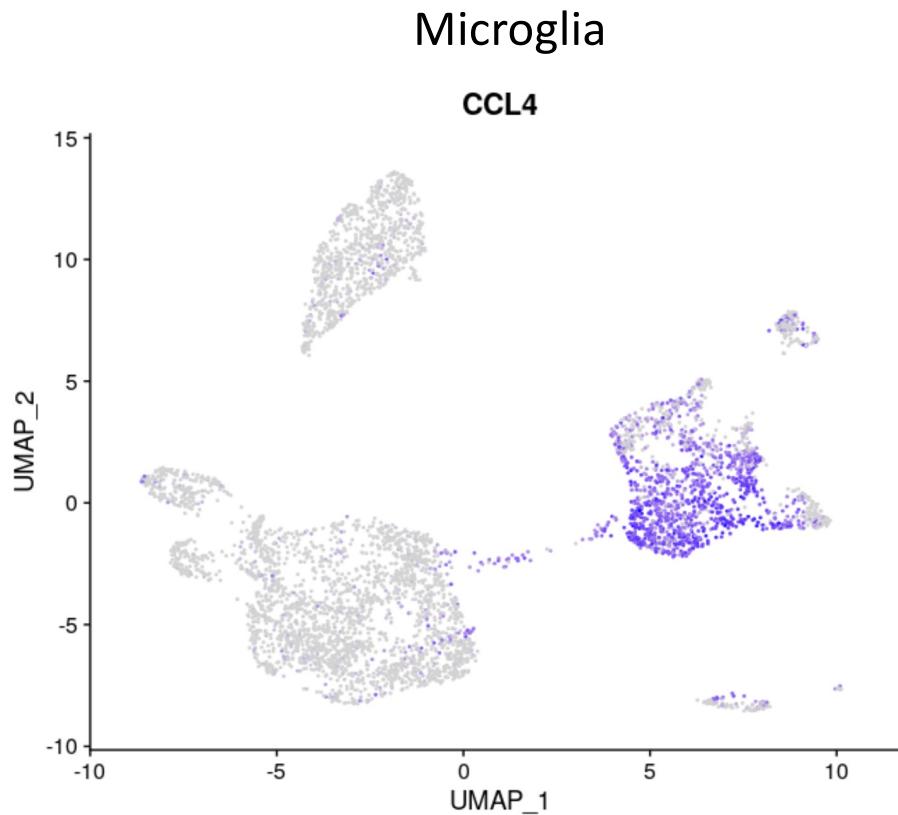
# Change in cell type abundances: what are the new cells?



# Manual vs automatic cell type annotation

- Manual: **using marker genes**
  - What most people do...
  - Time consuming
  - Requires expert knowledge
  - Sometimes subjective and inaccurate
- Automatic: **requires a reference**
  - Use complete cell type-specific mRNA expression profiles based on bulk RNAseq from FACS-sorted ‘pure’ populations
  - OR: Use “a reference” of manually curated cells picked from scRNA-seq data sets
  - Can miss cell types if they are not included in the reference

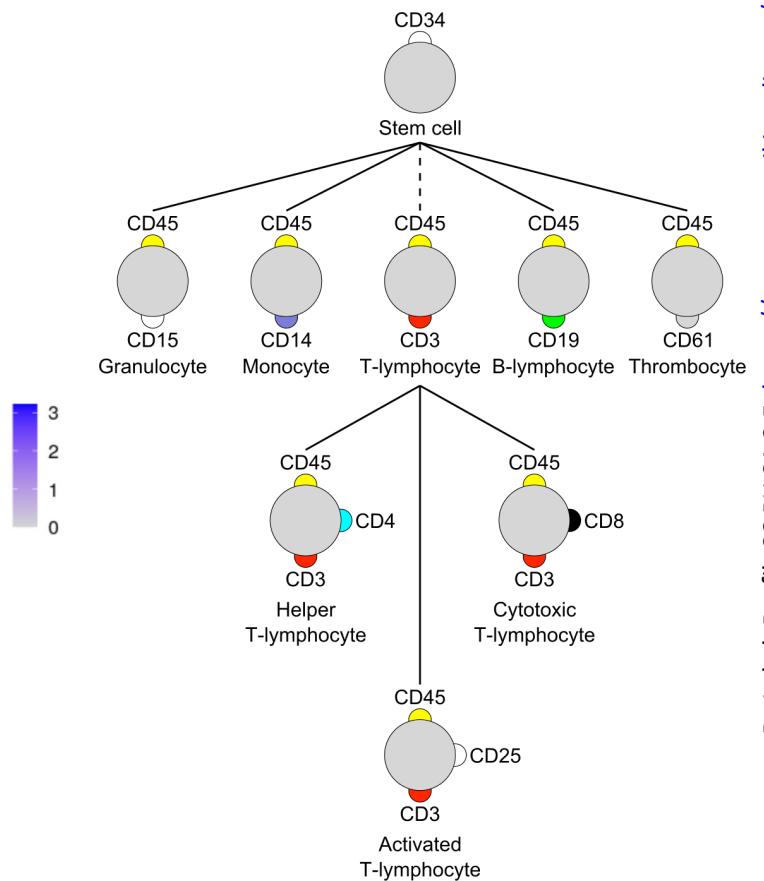
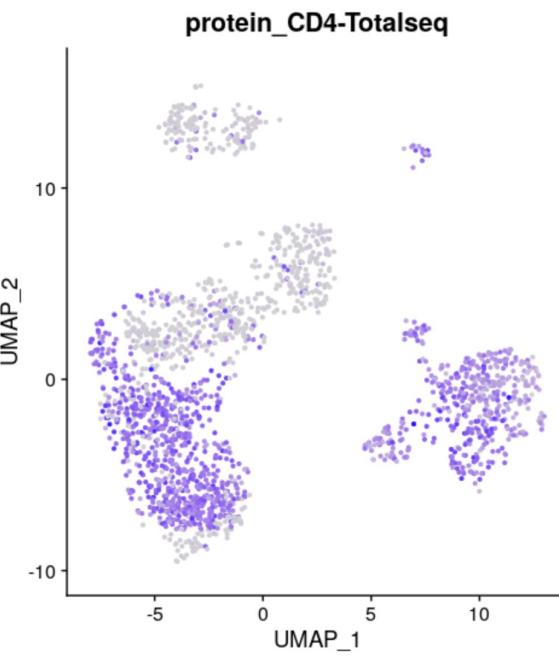
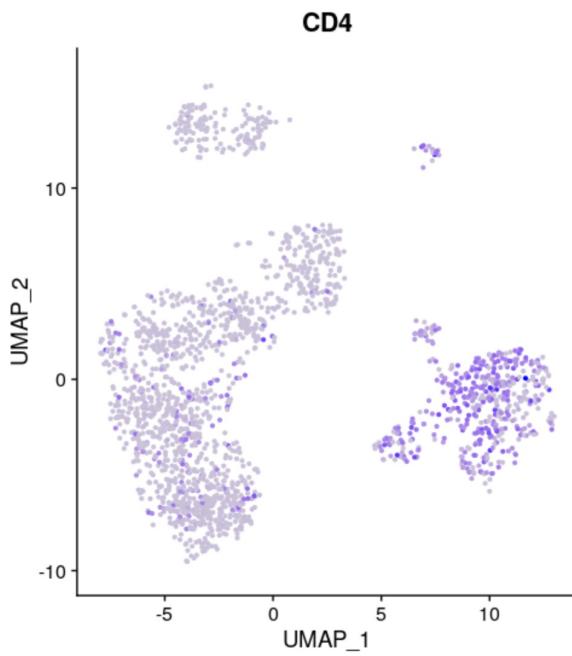
# Manual annotation using known marker genes



Human glioblastoma multiforme cells, 10x Genomics data (source of data to play with)  
[https://support.10xgenomics.com/single-cell-gene-expression/datasets/4.0.0/Parent\\_SC3v3\\_Human\\_Glioblastoma](https://support.10xgenomics.com/single-cell-gene-expression/datasets/4.0.0/Parent_SC3v3_Human_Glioblastoma)

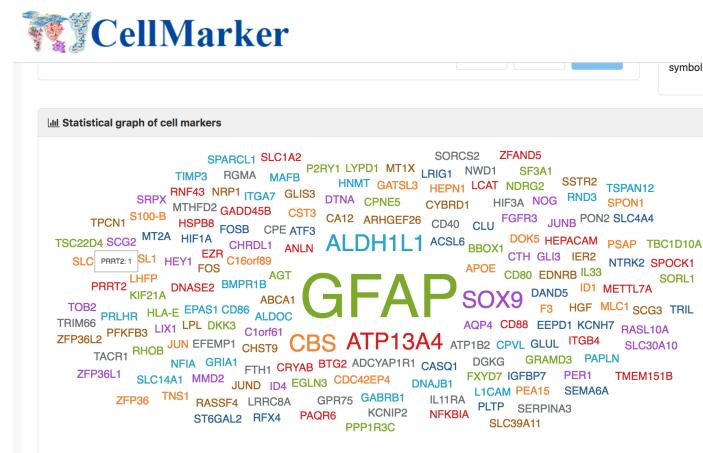
# Cell surface markers

- Often considered the gold standard esp. in immunology
- mRNA of cell surface markers sometimes lowly expressed or absent
- Use a combination of such marker genes, and also other genes like marker genes among clusters (eg secreted proteins or transcription factors)



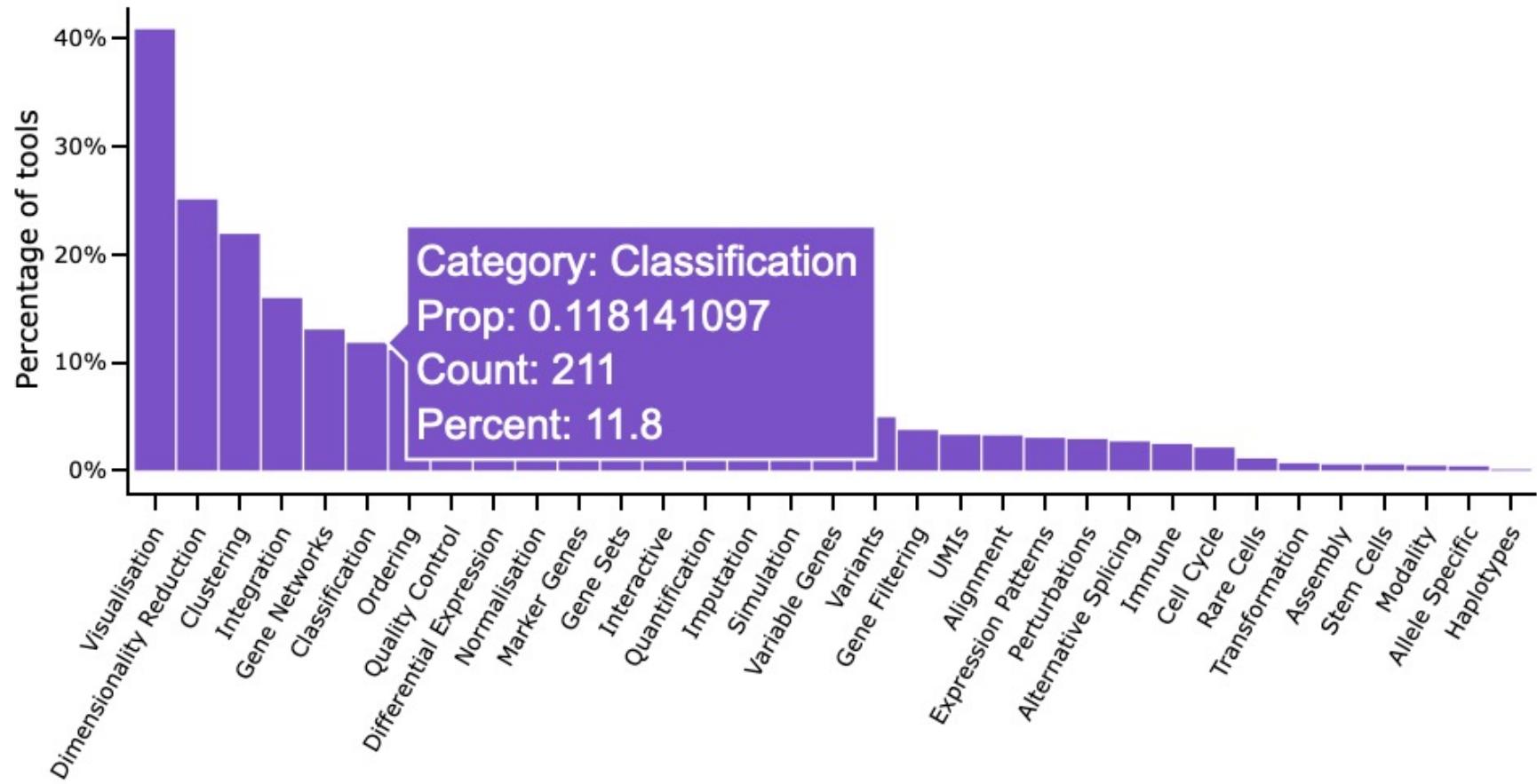
# Databases with cell type marker genes

- PanglaoDB <https://panglaodb.se/> (mouse and human)  
Check out <https://cran.r-project.org/web/packages/rPanglaoDB/index.html>
- CellMarker (mouse and human)  
<http://xteam.xbio.top/CellMarker/index.jsp>



-> Check whether your marker genes are part of any of the cluster marker genes

# Automated cell type annotation



# Several methods are available

**Table 1 Automatic cell identification methods included in this study**

From: [A comparison of automatic cell identification methods for single-cell RNA sequencing data](#)

Name	Version	Language	Underlying classifier	Prior knowledge	Rejection option	Reference
Garnett	0.1.4	R	Generalized linear model	Yes	Yes	[14]
Moana	0.1.1	Python	SVM with linear kernel	Yes	No	[15]
DigitalCellSorter	GitHub version: e369a34	Python	Voting based on cell type markers	Yes	No	[16]
SCINA	1.1.0	R	Bimodal distribution fitting for marker genes	Yes	No	[17]
scVI	0.3.0	Python	Neural network	No	No	[18]
Cell-BLAST	0.1.2	Python	Cell-to-cell similarity	No	Yes	[19]
ACTINN	GitHub version: 563bcc1	Python	Neural network	No	No	[20]
LAmbDA	GitHub version: 3891d72	Python	Random forest	No	No	[21]
scmapcluster	1.5.1	R	Nearest median classifier	No	Yes	[22]
scmapcell	1.5.1	R	kNN	No	Yes	[22]
scPred	0.0.0.9000	R	SVM with radial kernel	No	Yes	[23]
CHETAH	0.99.5	R	Correlation to training set	No	Yes	[24]
CaSTLe	GitHub version: 258b278	R	Random forest	No	No	[25]
SingleR	0.2.2	R	Correlation to training set	No	No	[26]
scID	0.0.0.9000	R	LDA	No	Yes	[27]
singleCellNet	0.1.0	R	Random forest	No	No	[28]
LDA	0.19.2	Python	LDA	No	No	[29]
NMC	0.19.2	Python	NMC	No	No	[29]
RF	0.19.2	Python	RF (50 trees)	No	No	[29]
SVM	0.19.2	Python	SVM (linear kernel)	No	No	[29]
SVM <sub>rejection</sub>	0.19.2	Python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	Python	kNN (k = 9)	No	No	[29]

# CellTypist

- Leverages machine learning models trained on large and diverse reference datasets to assign cell type labels in a query dataset.
- Classifiers are trained on the reference database: models learn to recognize patterns of gene expression characteristic of cell types.
- Compare the query gene expression profile to the reference profile, **predict** the most likely cell type
- CellTypist can assign broad or specific cell type label (user's choice)
- Obtain confidence score – QC your annotations

<https://www.celltypist.org/>

<https://www.science.org/doi/10.1126/science.abl5197>

# Sources of references

- **Celltypist** pre-trained models: `models.models_description()`

48 models: <https://www.celltypist.org/models>

Use your own reference: train the model with `celltypist.train()`

[https://colab.research.google.com/github/Teichlab/celltypist/blob/main/docs/notebook/celltypist\\_tutorial.ipynb#scrollTo=precise-bronze](https://colab.research.google.com/github/Teichlab/celltypist/blob/main/docs/notebook/celltypist_tutorial.ipynb#scrollTo=precise-bronze)

- **Tabula muris senis**: single-cell RNA-sequencing of different organs across the mouse lifespan, available as .h5ad files:

[https://figshare.com/articles/dataset/Tabula\\_Muris\\_Senis\\_Data\\_Objects/12654728/1](https://figshare.com/articles/dataset/Tabula_Muris_Senis_Data_Objects/12654728/1)

<https://github.com/czbiohub-sf/tabula-muris/blob/master/tabula-muris-on-aws.md>

- **Single-cell portal**: convert count matrix and cell labels to annData  
[https://singlecell.broadinstitute.org/single\\_cell](https://singlecell.broadinstitute.org/single_cell)

- Some **papers** provide link to .h5ad, eg human lymph node compartments:  
<https://www.nature.com/articles/s41587-021-01139-4>

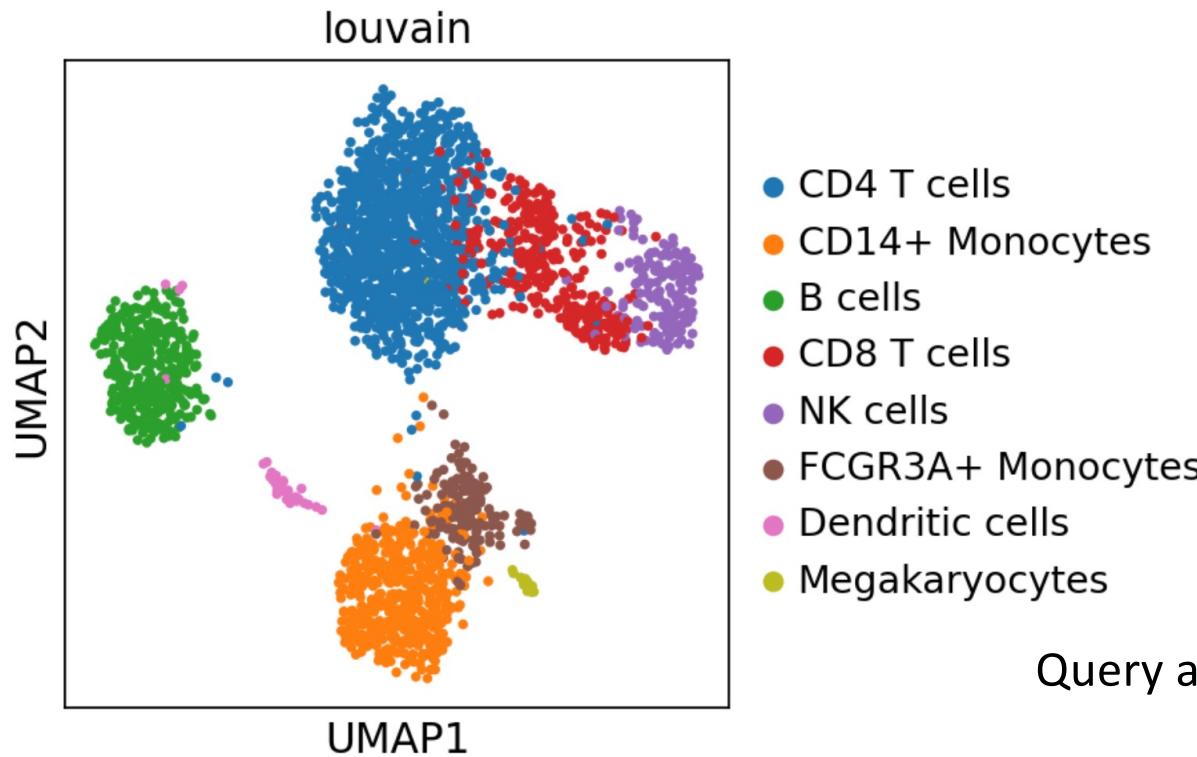
# ingest

<https://scanpy-tutorials.readthedocs.io/en/latest/integrating-data-using-ingest.html>

- Label transfer from an annotated annData to a query annData. Uses a kNN classifier for mapping labels and a UMAP for mapping the embeddings.
- Predict a cell type to each cell based on what it is most similar to.
- Fit a model on the reference data and use it to project new data.
- Look at what genes the different cell types express and use your biological knowledge to decide whether the annotation is good or should be improved.

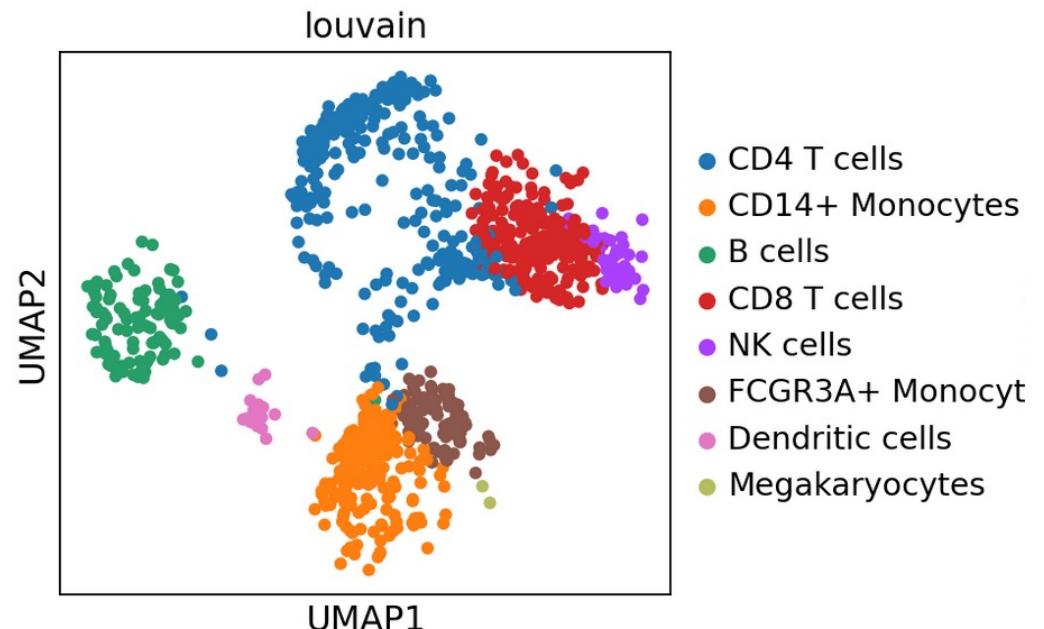
[https://nbisweden.github.io/workshop-scrNAseq/labs/scanpy/scanpy\\_06\\_celltyping.html#ingest](https://nbisweden.github.io/workshop-scrNAseq/labs/scanpy/scanpy_06_celltyping.html#ingest)

## Reference annotated annData



```
sc.tl.ingest(adata=adata,  
adata_ref=adata_ref,  
obs="celltype",  
embedding_method='pca')
```

## Query after ingest annotation



# Additional links

Review on automated cell annotation, Pasquini et al 2021

<https://www.sciencedirect.com/science/article/pii/S2001037021000192>

Question on cell type annotation