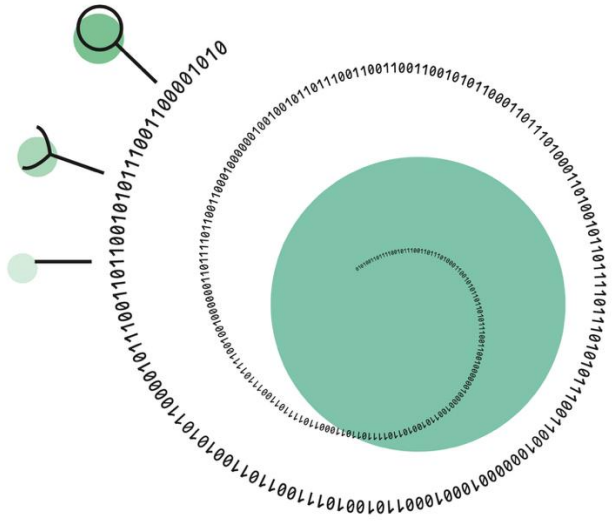
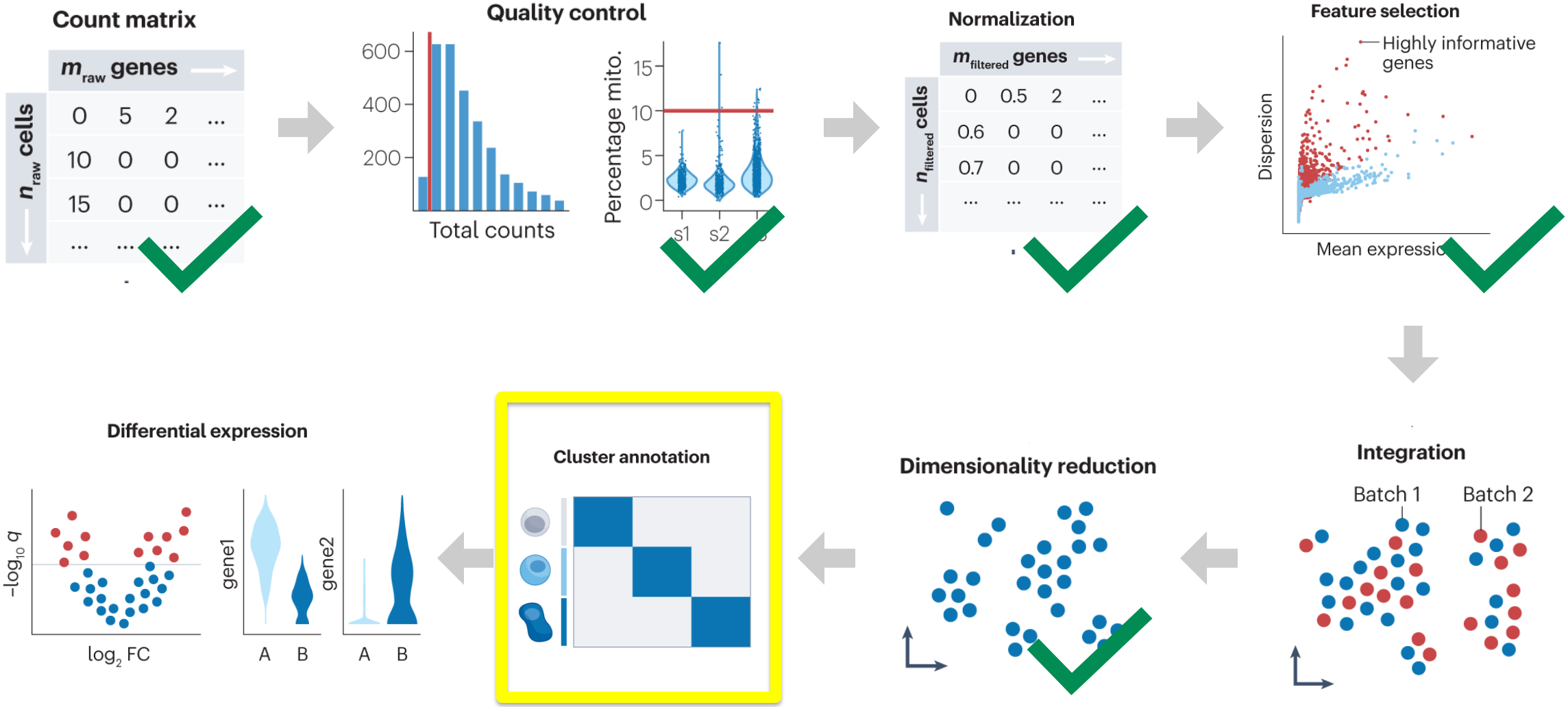


Single Dell Data Analysis Course

Cluster-based cell type annotation

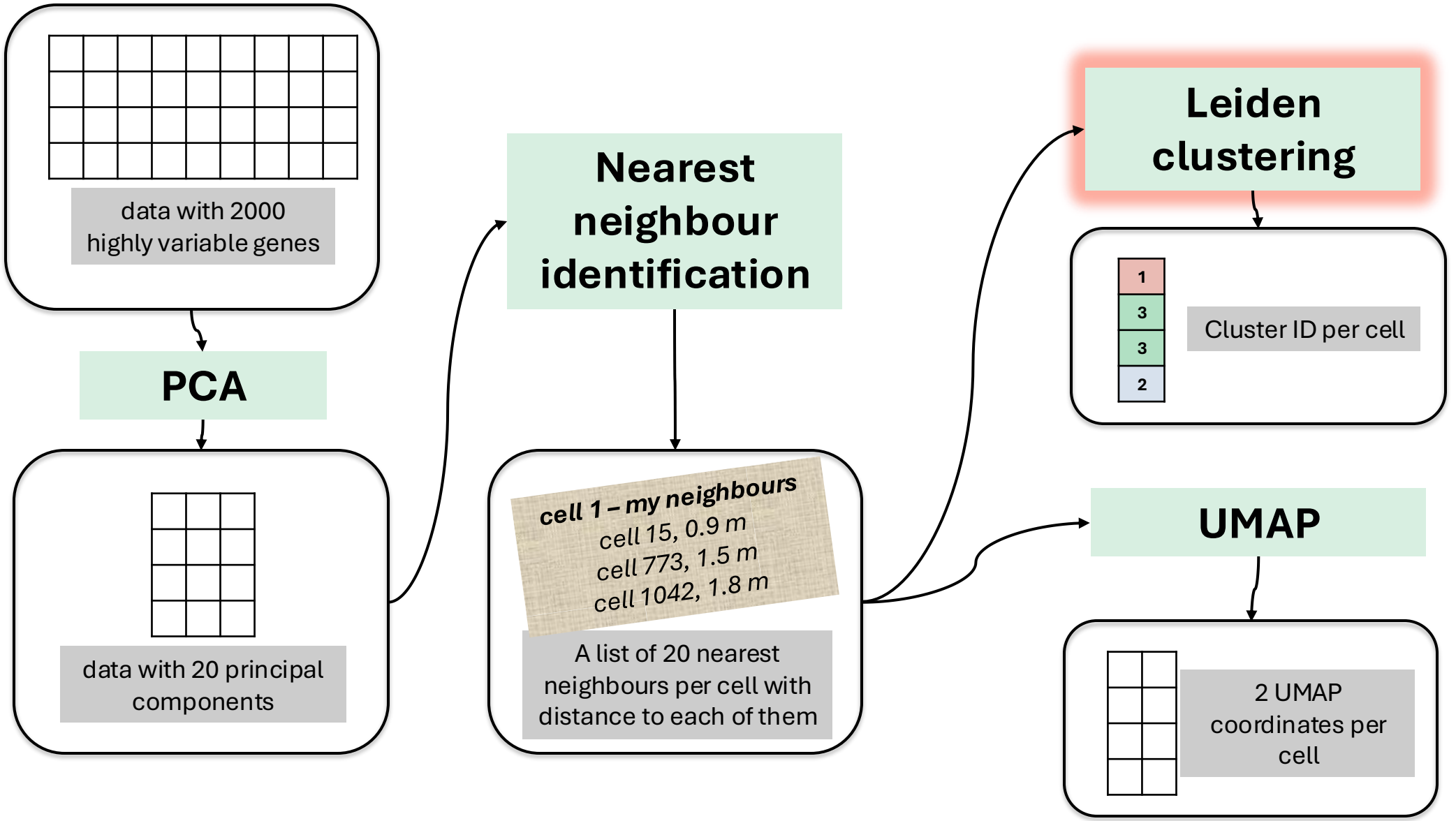
Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Today

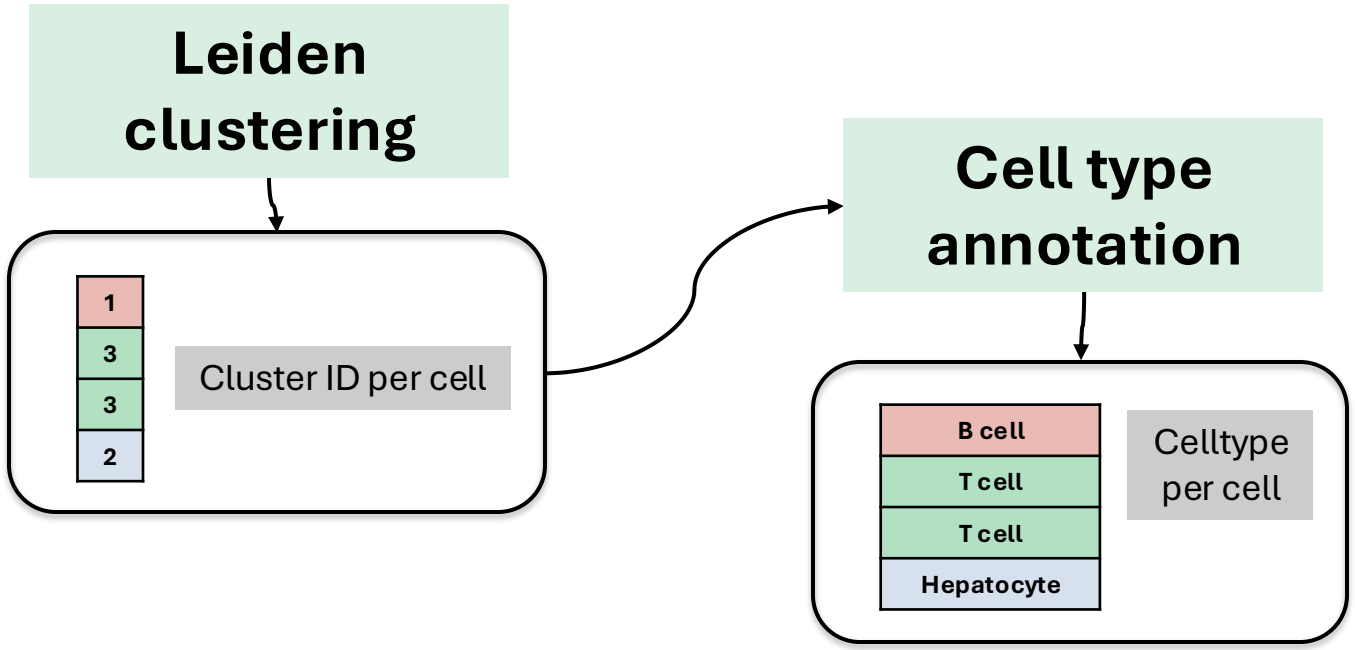


Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

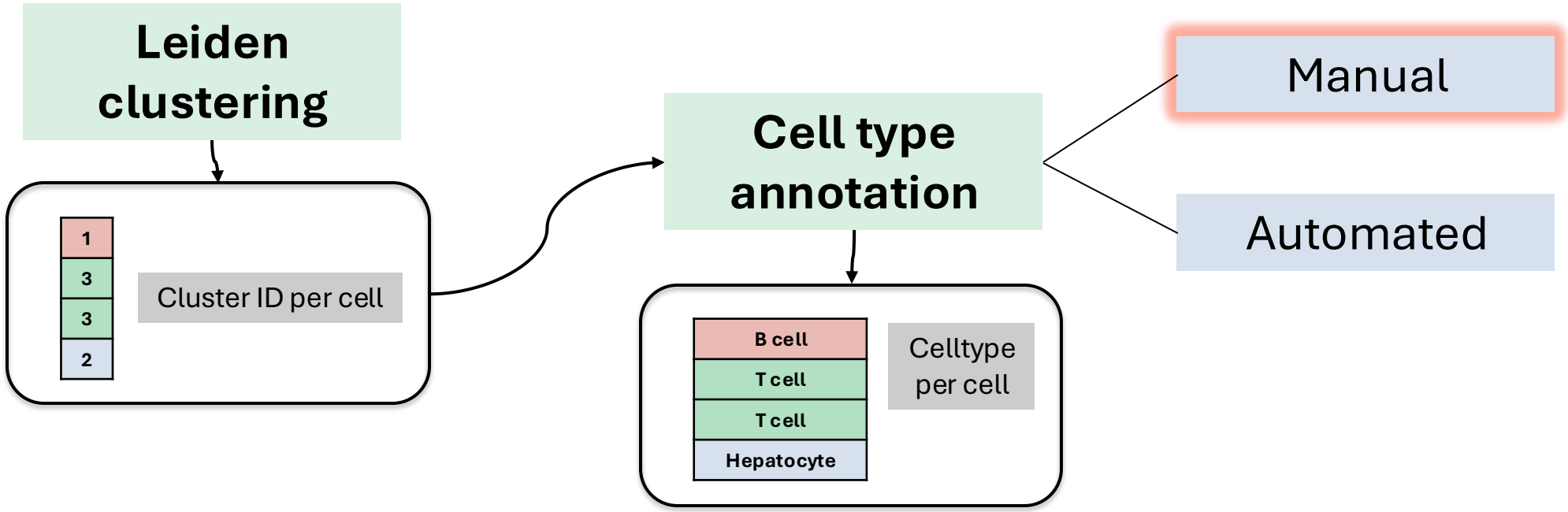
Recap: Where we stand after a whole lot of processing



Turning clusters into knowledge with annotation and differential expression analysis



Turning clusters into knowledge with annotation and differential expression analysis



Turning clusters into knowledge with annotation and differential expression analysis

Leiden clustering

1
3
3
2

Cluster ID per cell

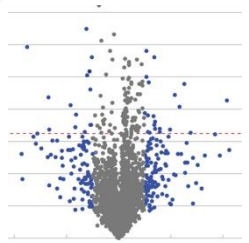
Cell type annotation

B cell
T cell
T cell
Hepatocyte

Celltype per cell

Differential cell type abundance

Differentially expressed genes



DEGs per cell type / cluster

DEGs per condition

contrast	gene	log(fold change)	P value
B cells vs rest	A	5	0.001
B cells vs rest	B	0.1	0.04
T cells vs rest	C	3	0.002

Cell type annotation ≈ giving names to clusters

What is a cell type?

Hongkui Zeng (2022) What is a cell type and how to define it?, Cell, Volume 185, Issue 15, https://doi.org/10.1016/j.cell.2022.06.031.

Cell type annotation ≈ giving names to clusters

What is a cell type?

A cellular phenotype that is

- robust across datasets
- Linked to specific functions
- Identifiable based on specific markers (most often proteins or transcripts)

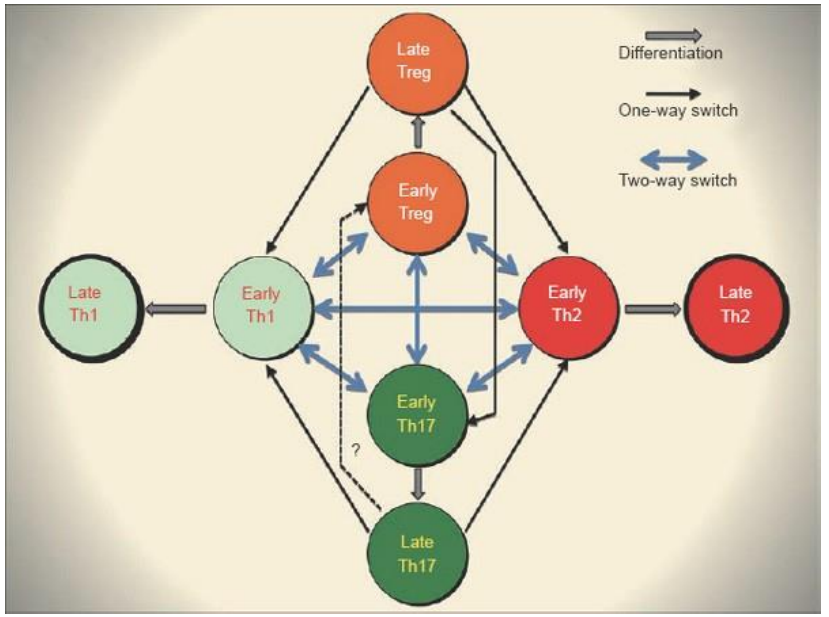
Hongkui Zeng (2022) What is a cell type and how to define it?, Cell, Volume 185, Issue 15, https://doi.org/10.1016/j.cell.2022.06.031.

Cell type annotation ≈ giving names to clusters

What is a cell type?

A cellular phenotype that is

- robust across datasets
- Linked to specific functions
- Identifiable based on specific markers (most often proteins or transcripts)



cell state ⟷ cell type

Hongkui Zeng (2022) What is a cell type and how to define it?, Cell, Volume 185, Issue 15, https://doi.org/10.1016/j.cell.2022.06.031.

Cell type annotation ≈ giving names to clusters

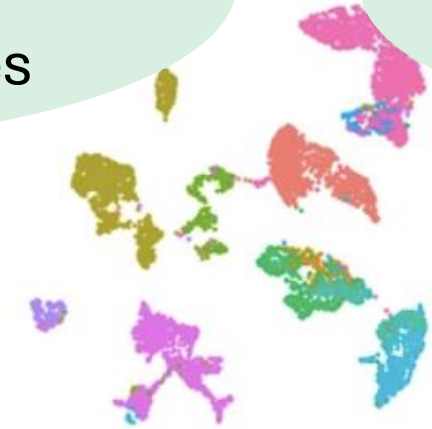
Why do we need to give names to clusters?

...could we not analyse everything based on our computed embeddings, trajectories and clusters?

Cell type annotation ≈ giving names to clusters

Why do we need to give names to clusters?

To speak about biology with others in established categories

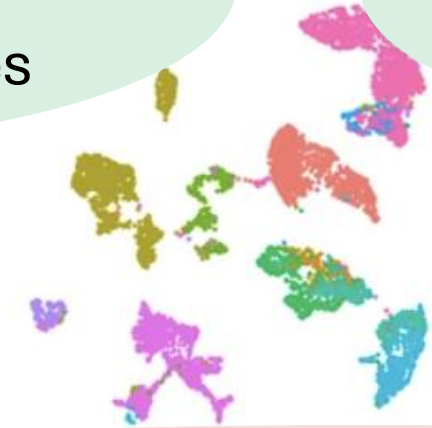


To link sc omics based results to decades of immunological research

Cell type annotation ≈ giving names to clusters

Why do we need to give names to clusters?

To speak about biology with others in established categories



To link sc omics based results to decades of immunological research

The “pure ML” view: We should not let ourselves be biased by old papers based on few markers, instead reanalyse biology from scratch

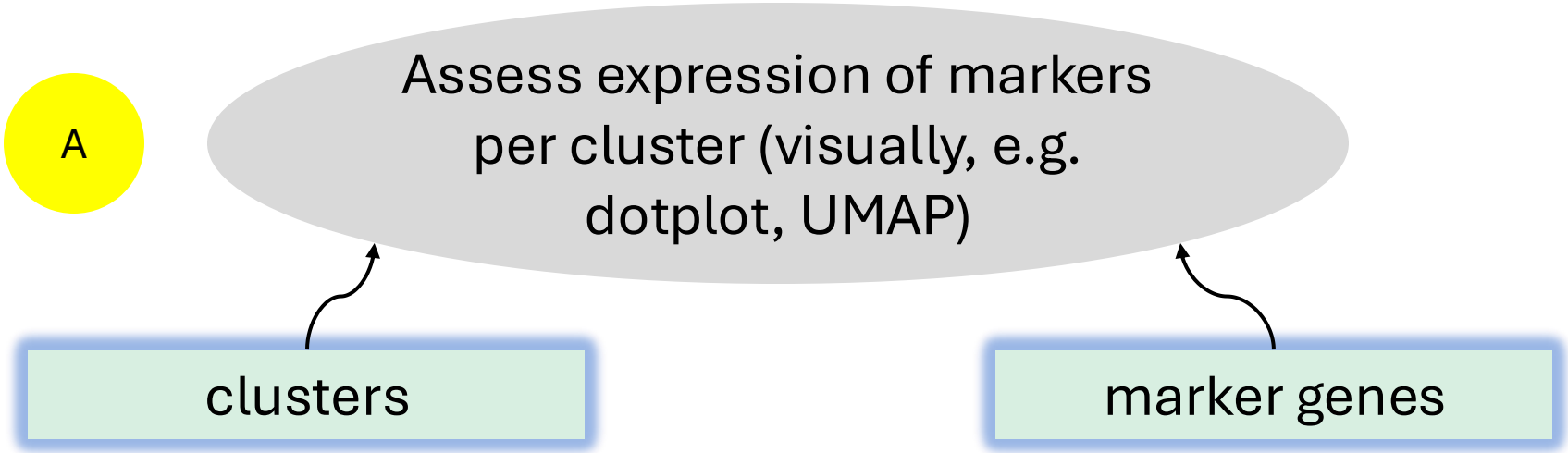
Manual cell type annotation

clusters

marker genes

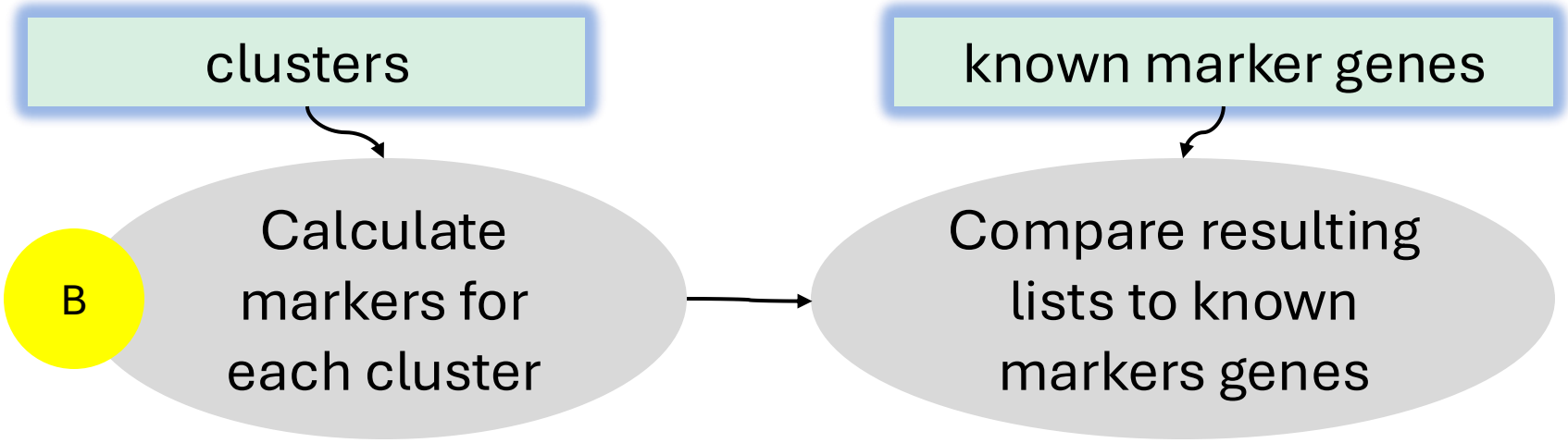
https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

Manual cell type annotation



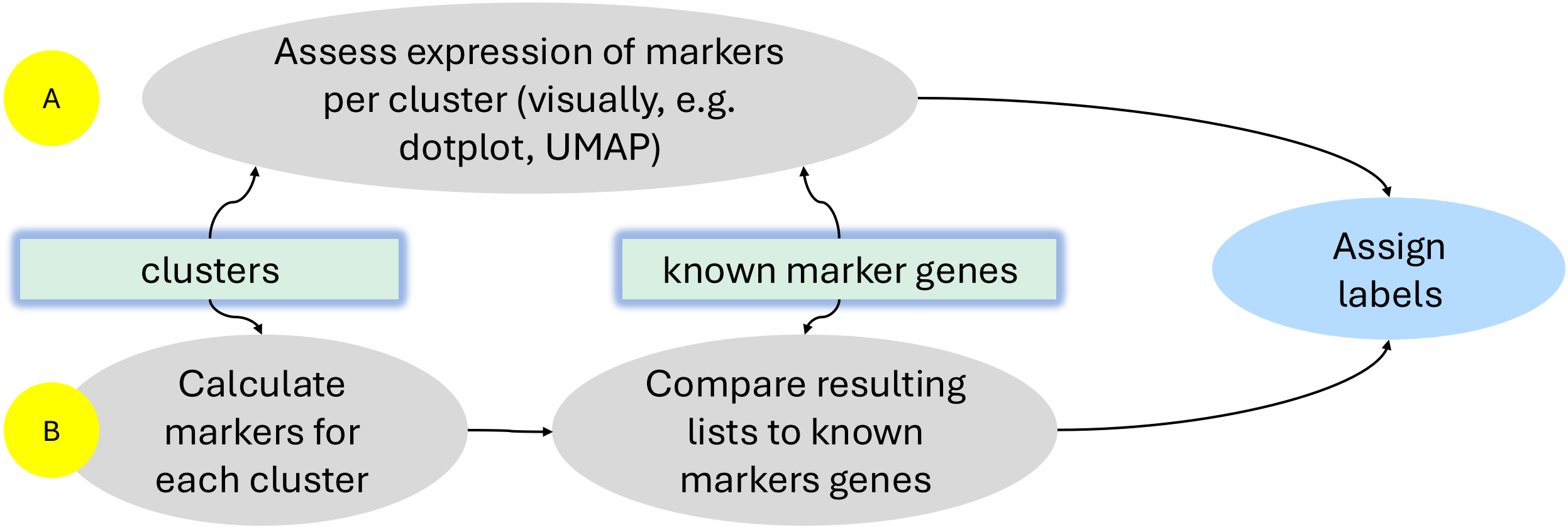
https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

Manual cell type annotation



https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

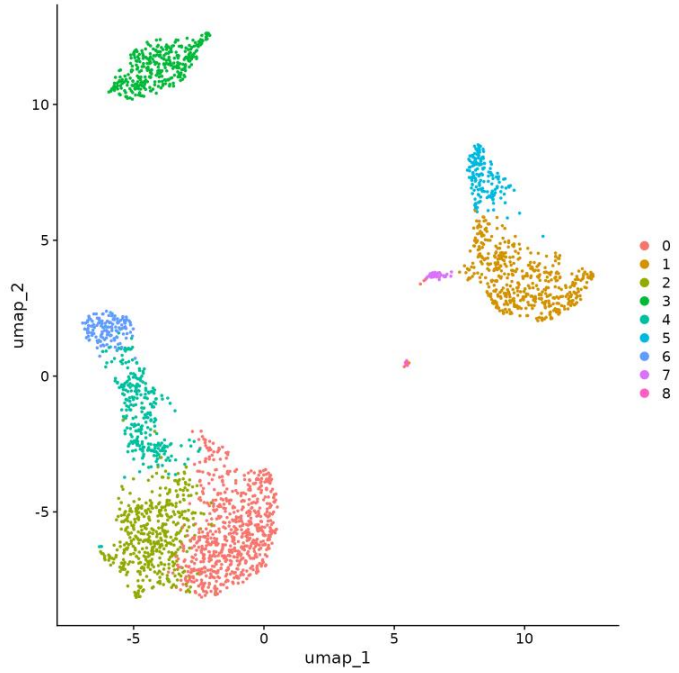
Manual cell type annotation



https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

Method A - PBMC annotation / ingredients

clusters



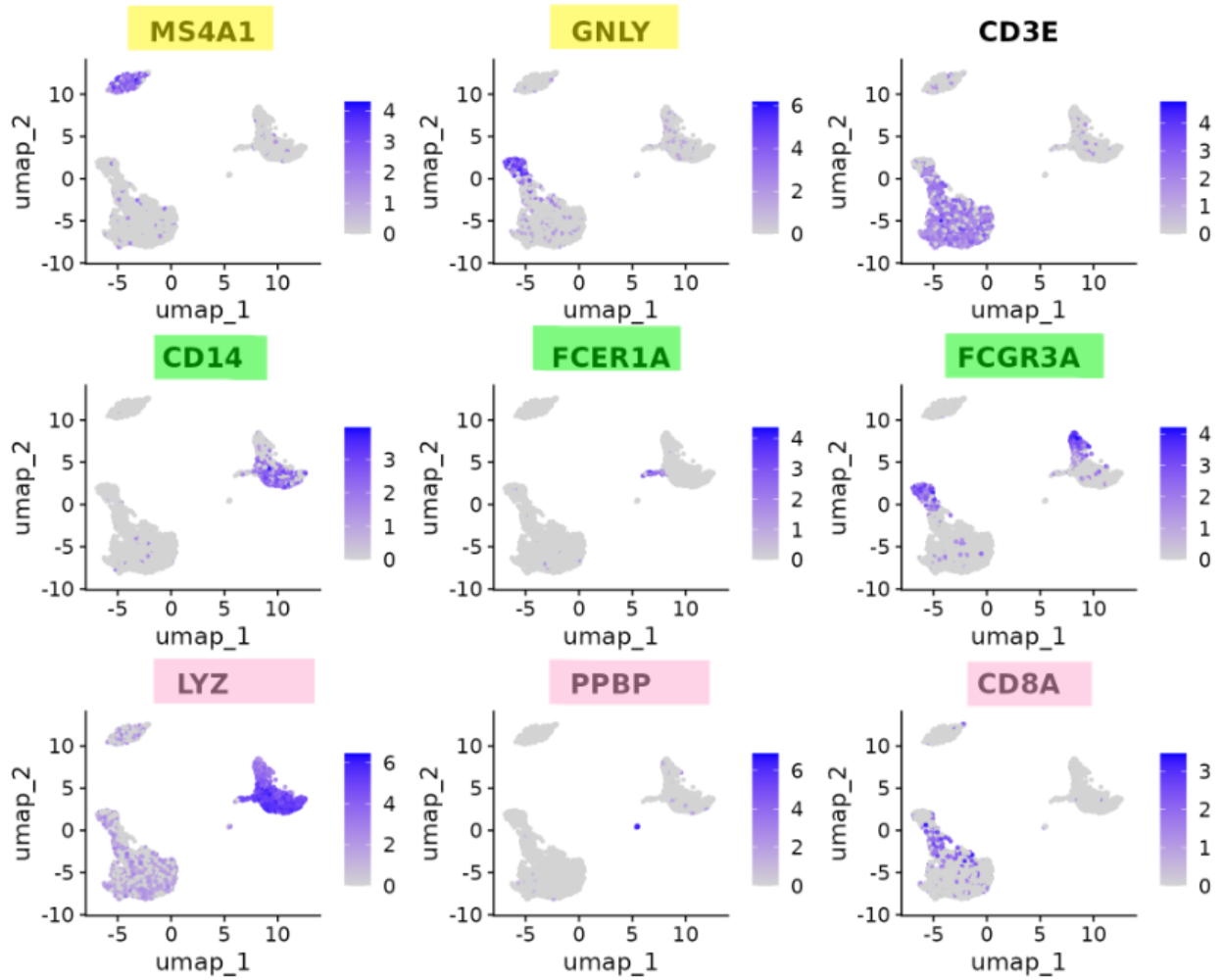
known marker genes

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Method A - PBMC annotation / UMAP inspection

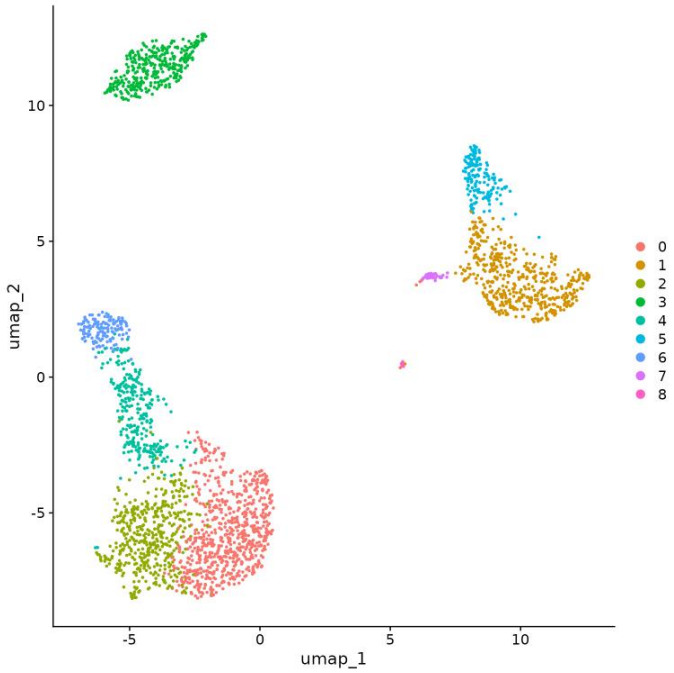
Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

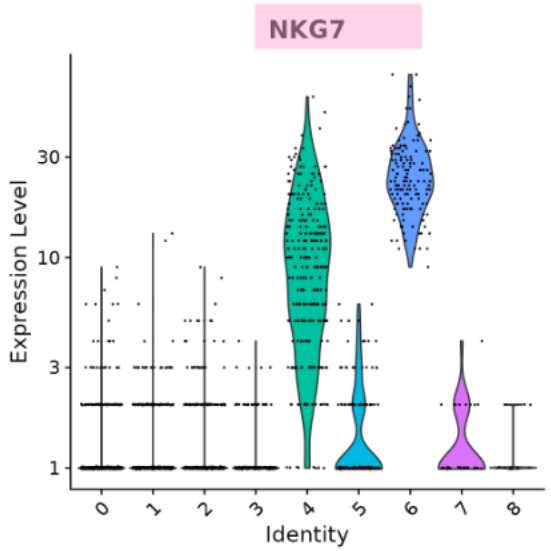


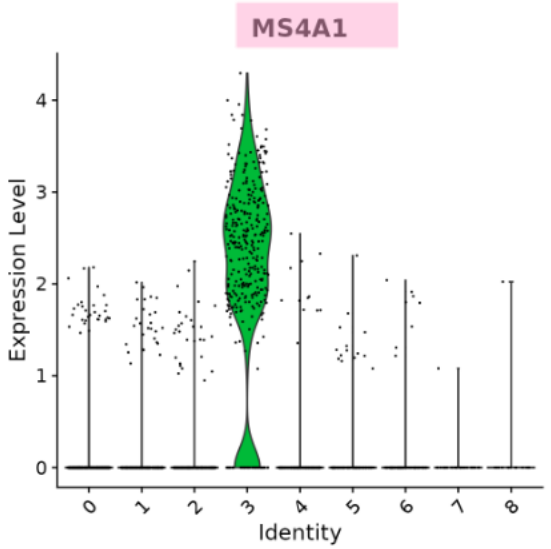
https://satijalab.org/seurat/articles/pbmc3k_tutorial

Method A - PBMC annotation / violin plots of marker genes

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

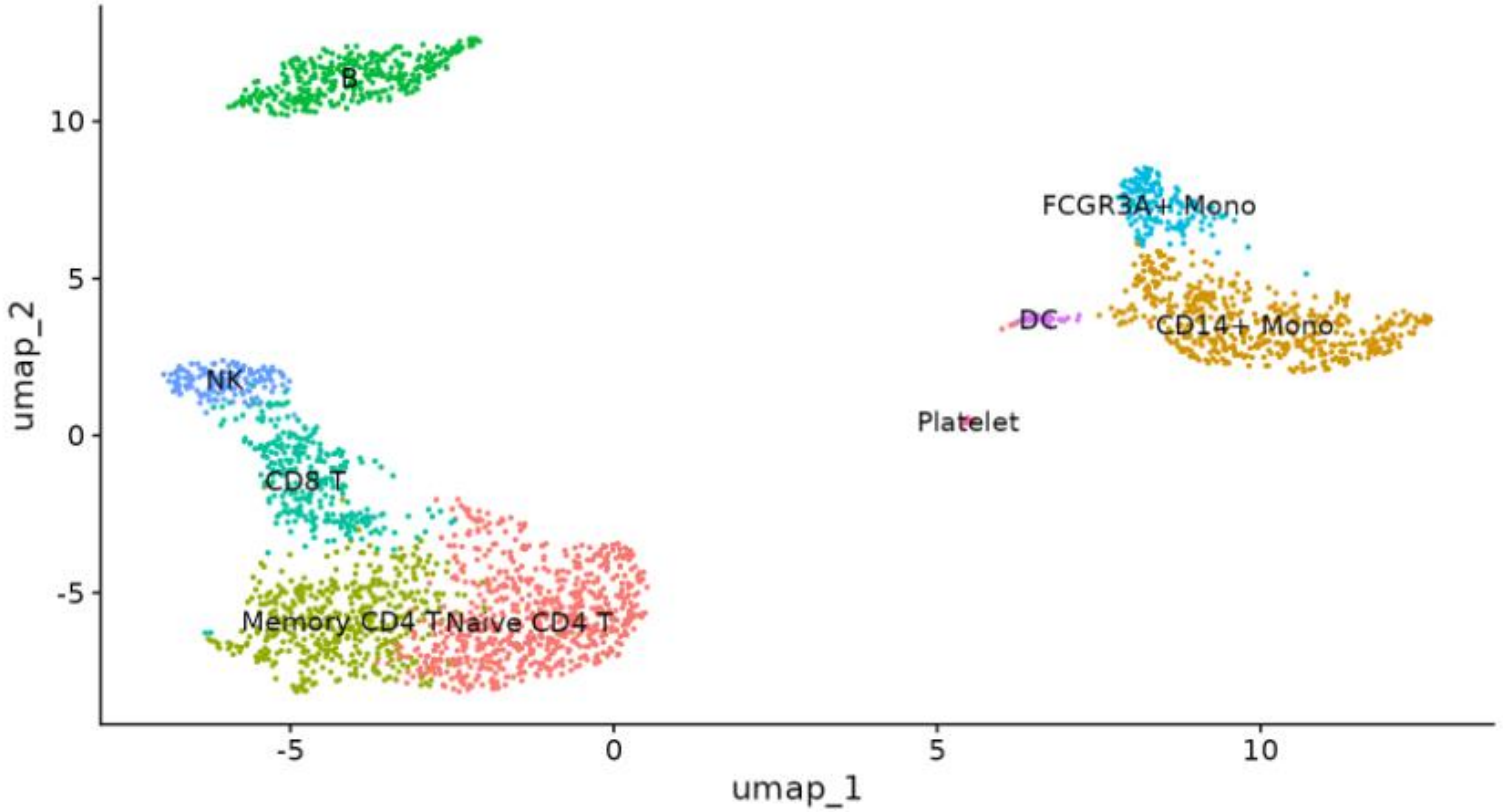






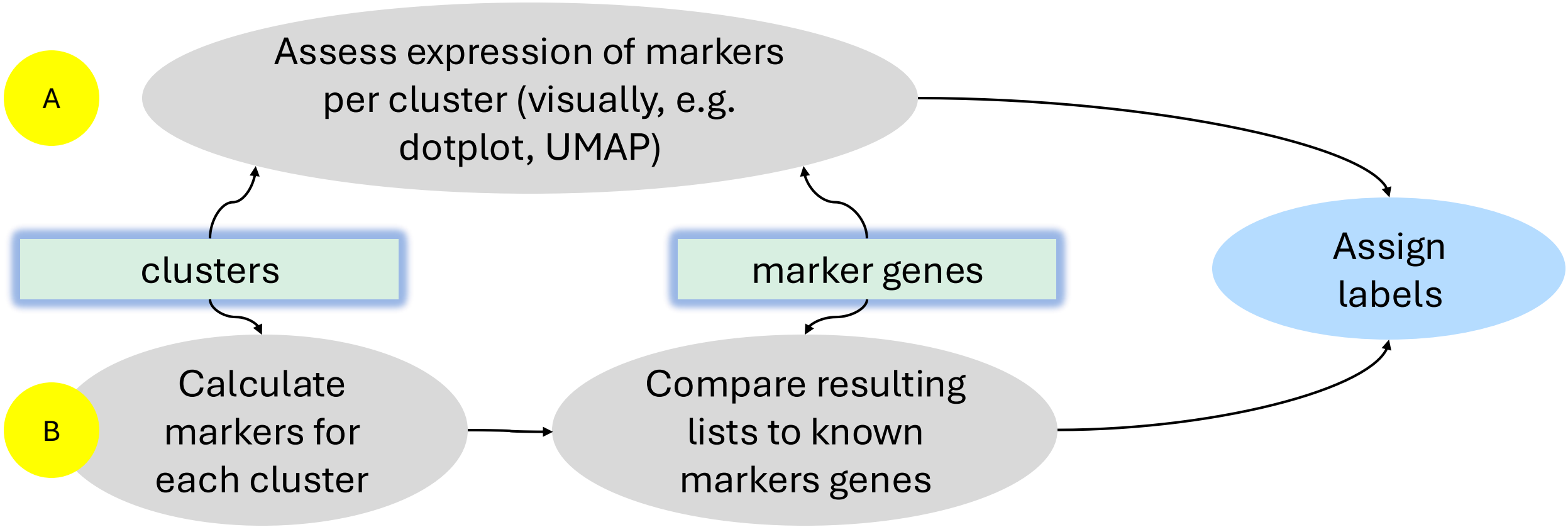
https://satijalab.org/seurat/articles/pbmc3k_tutorial

Method A - PBMC annotation / results



https://satijalab.org/seurat/articles/pbmc3k_tutorial

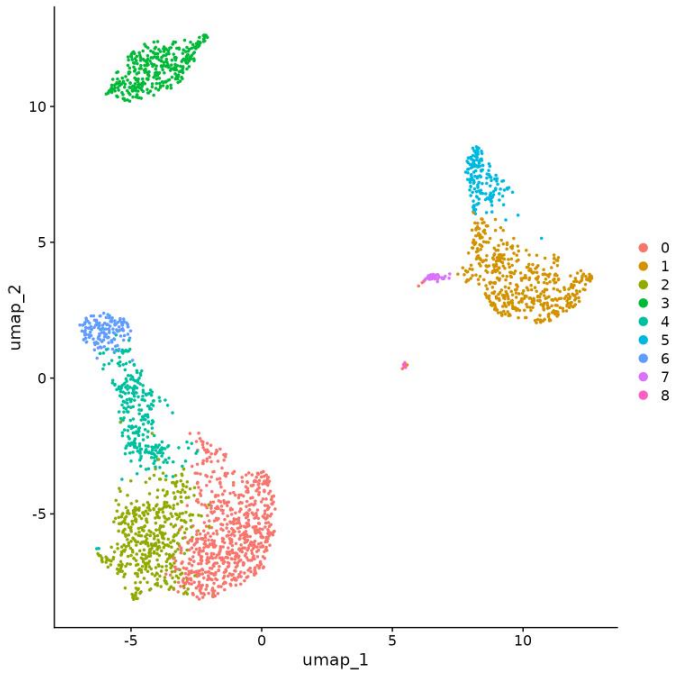
Manual cell type annotation



https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

Method B - PBMC annotation / ingredients

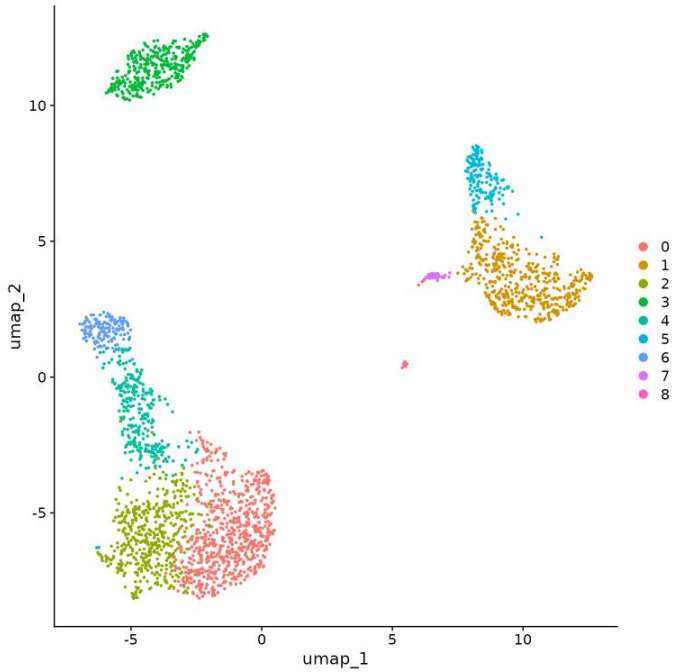
clusters



https://satijalab.org/seurat/articles/pbmc3k_tutorial

Method B - PBMC annotation / ingredients

clusters → Calculate markers for each cluster



Seurat

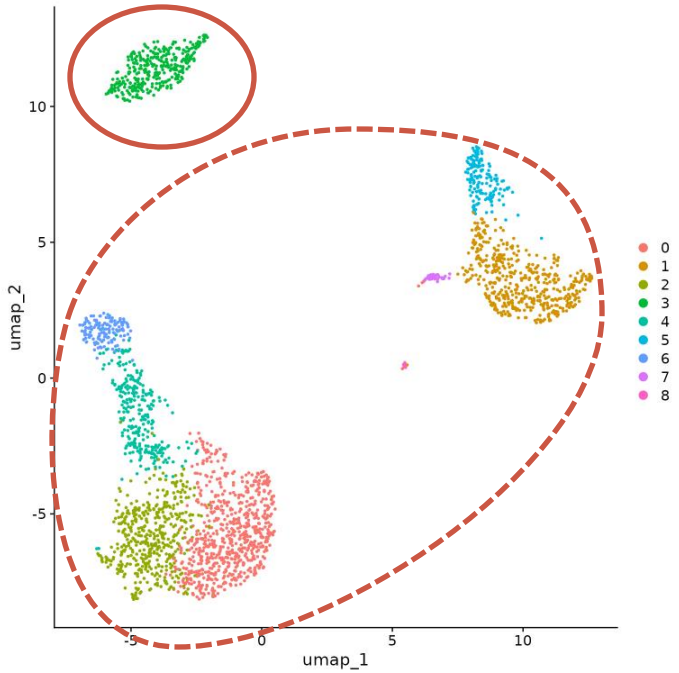
```
cluster0.markers <- FindMarkers(pbmc, ident.1 = 0, logfc.threshold = 0.25, test.use = "roc", onl
y.pos = TRUE)
```

scanpy

```
sc.tl.rank_genes_groups(
    adata, groupby="leiden_2", method="wilcoxon", key_added="dea_leiden_2"
)
```

https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

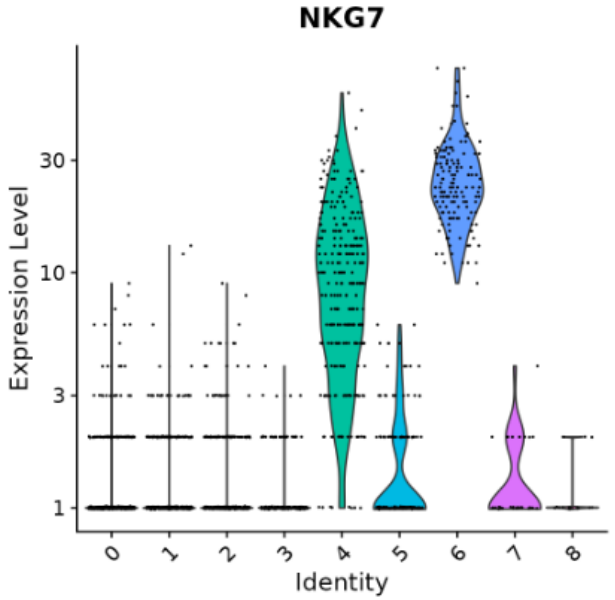
Calculating marker genes per cluster



1. Select cluster for comparison
(for marker gene detection we typically compare "cluster X" vs "all the rest")

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Calculating marker genes per cluster

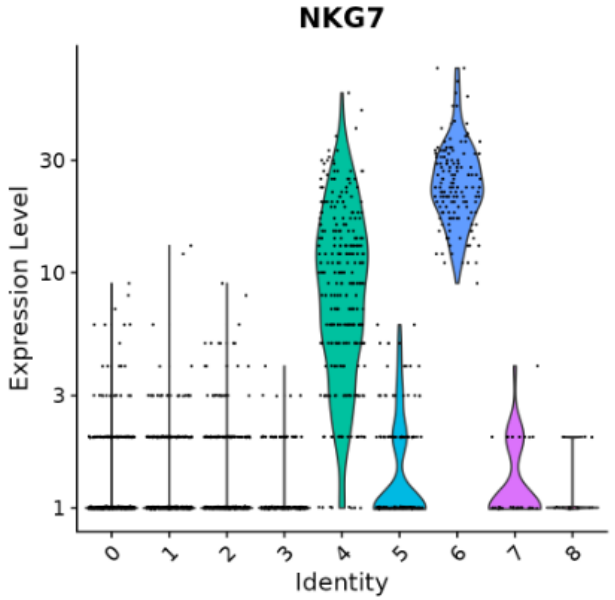


1. Select cluster for comparison (X)

2. For each gene, perform statistical test that compares the expression distribution between X and rest

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Calculating marker genes per cluster



1. Select cluster for comparison (X)

2. For each gene, perform statistical test that compares the expression distribution between X and rest

Classical testing

- t-test
- Wilcoxon's test

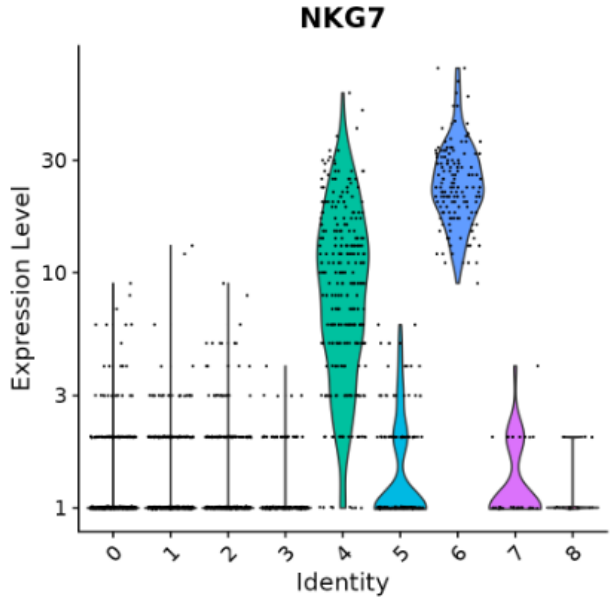
...

Classifier-based

- Logistic regression
- ROC analysis

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Calculating marker genes per cluster



1. Select cluster for comparison (X)

2. For each gene, perform statistical test that compares the expression distribution between X and rest

3. Apply multiple testing correction to resulting p-values

https://satijalab.org/seurat/articles/pbmc3k_tutorial

Calculating marker genes per cluster

p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
1.783378e-11	0.6126976	0.327	0.441	2.966472e-07	0	CAP1
1.111275e-05	0.5448506	0.201	0.277	1.848495e-01	0	NDUFA2
5.572391e-69	0.9152596	0.705	0.325	9.269115e-65	1	IL7R
1.834643e-44	0.7485499	0.308	0.086	3.051746e-40	1	MAL
0.000000e+00	4.4769283	0.968	0.207	0.000000e+00	2	S100A9
0.000000e+00	4.1251610	0.932	0.114	0.000000e+00	2	S100A8
1.124982e-304	4.1155861	0.992	0.192	1.871295e-300	3	NKG7
9.974215e-141	4.0851525	0.596	0.113	1.659111e-136	3	GNLY
0.000000e+00	2.7777043	0.934	0.043	0.000000e+00	4	CD79A
1.689832e-184	2.3696654	1.000	0.818	2.810866e-180	4	CD74
6.129817e-167	2.2891146	0.954	0.138	1.019634e-162	5	FCGR3A
1.497900e-118	2.3893744	1.000	0.316	2.491607e-114	5	LST1
8.448866e-05	0.6377758	1.000	0.956	1.000000e+00	6	RPL36
1.450393e-03	0.5964958	0.459	0.288	1.000000e+00	6	GYPC
2.002848e-181	7.1596296	1.000	0.011	3.331538e-177	7	PF4
6.022917e-101	8.5050281	1.000	0.025	1.001852e-96	7	PPBP

1. Select cluster for comparison (X)

2. For each gene, perform statistical test that compares the expression distribution between X and rest

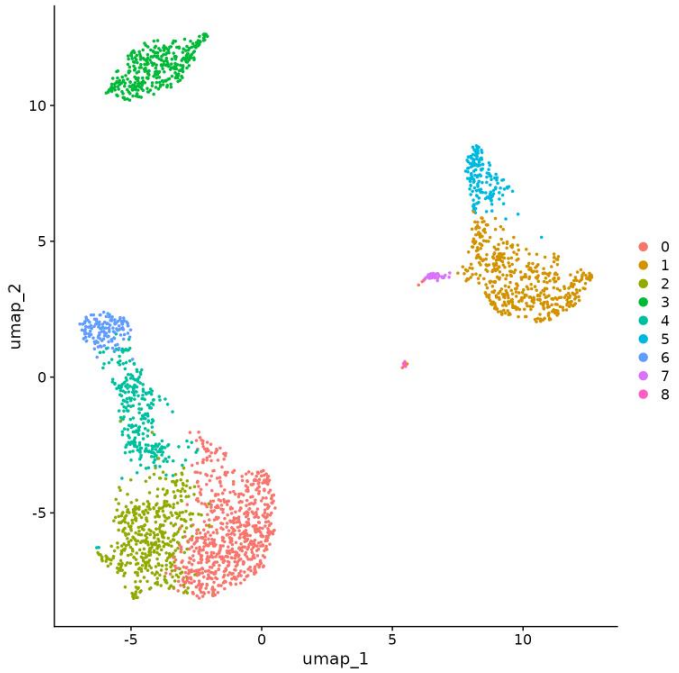
3. Apply multiple testing correction to resulting p-values

4. Identify significant / most important marker genes by sorting + thresholding **fold changes** and **p-values**

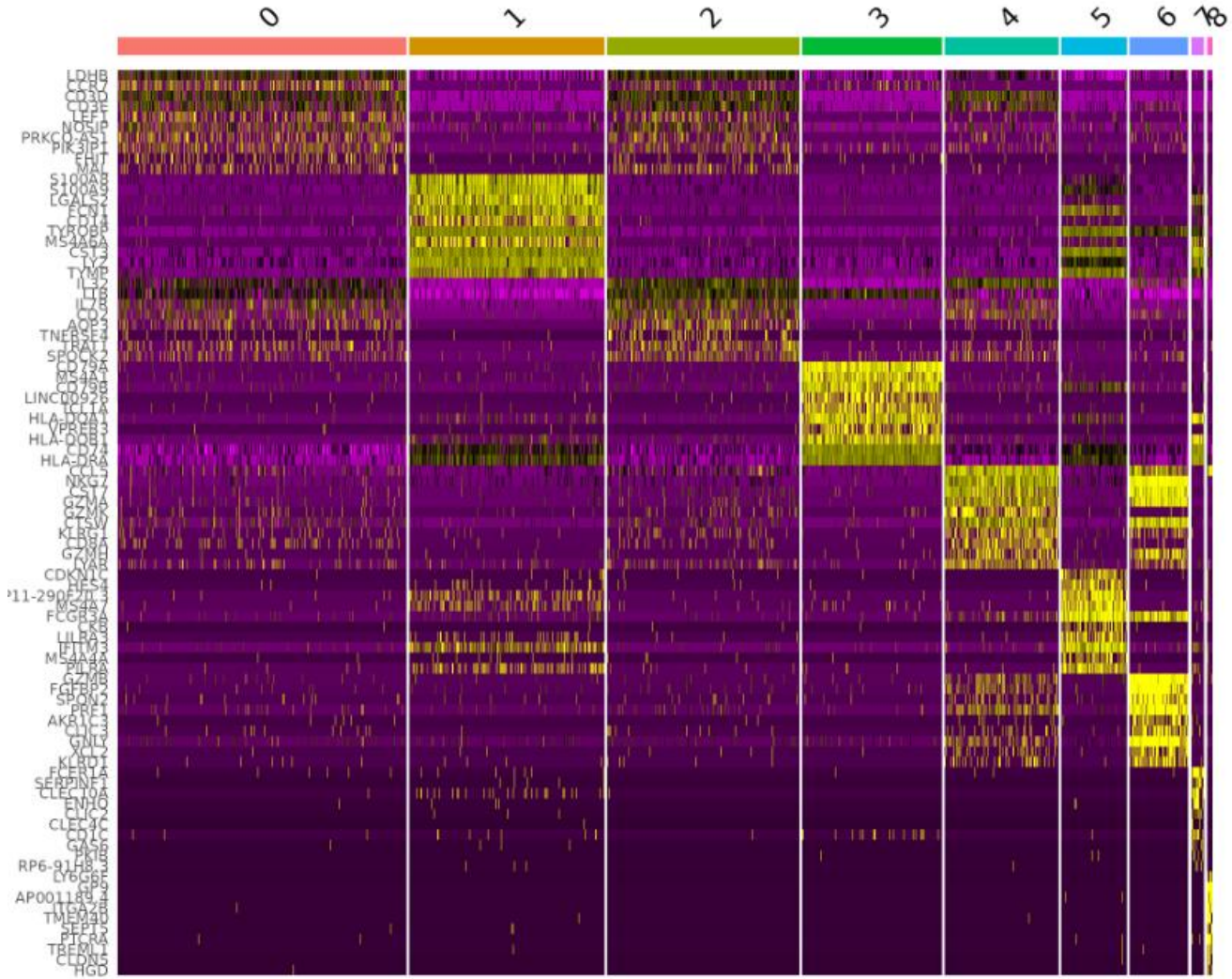
https://satijalab.org/seurat/articles/pbmc3k_tutorial

Method B - PBMC annotation / marker heatmap

clusters



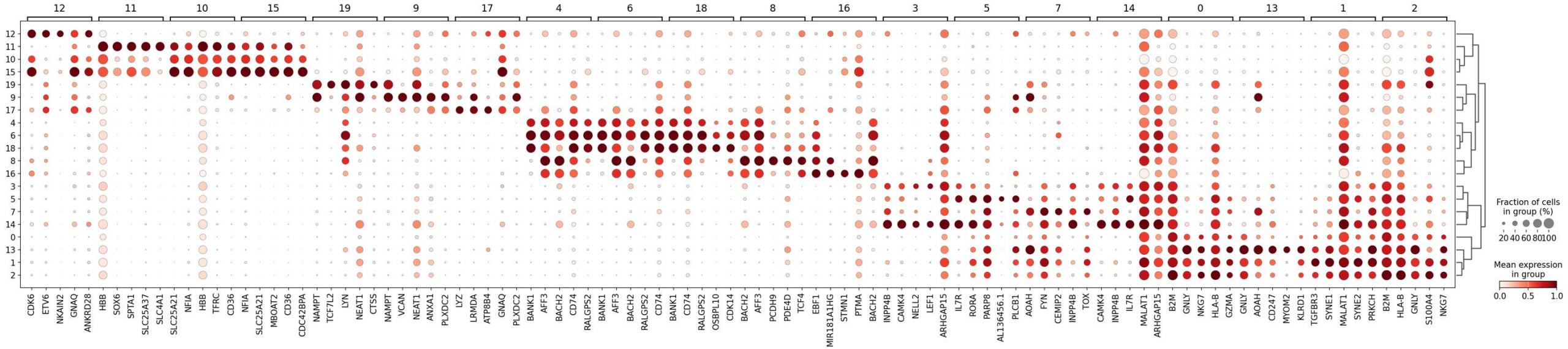
calculated marker genes



https://satijalab.org/seurat/articles/pbmc3k_tutorial
https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

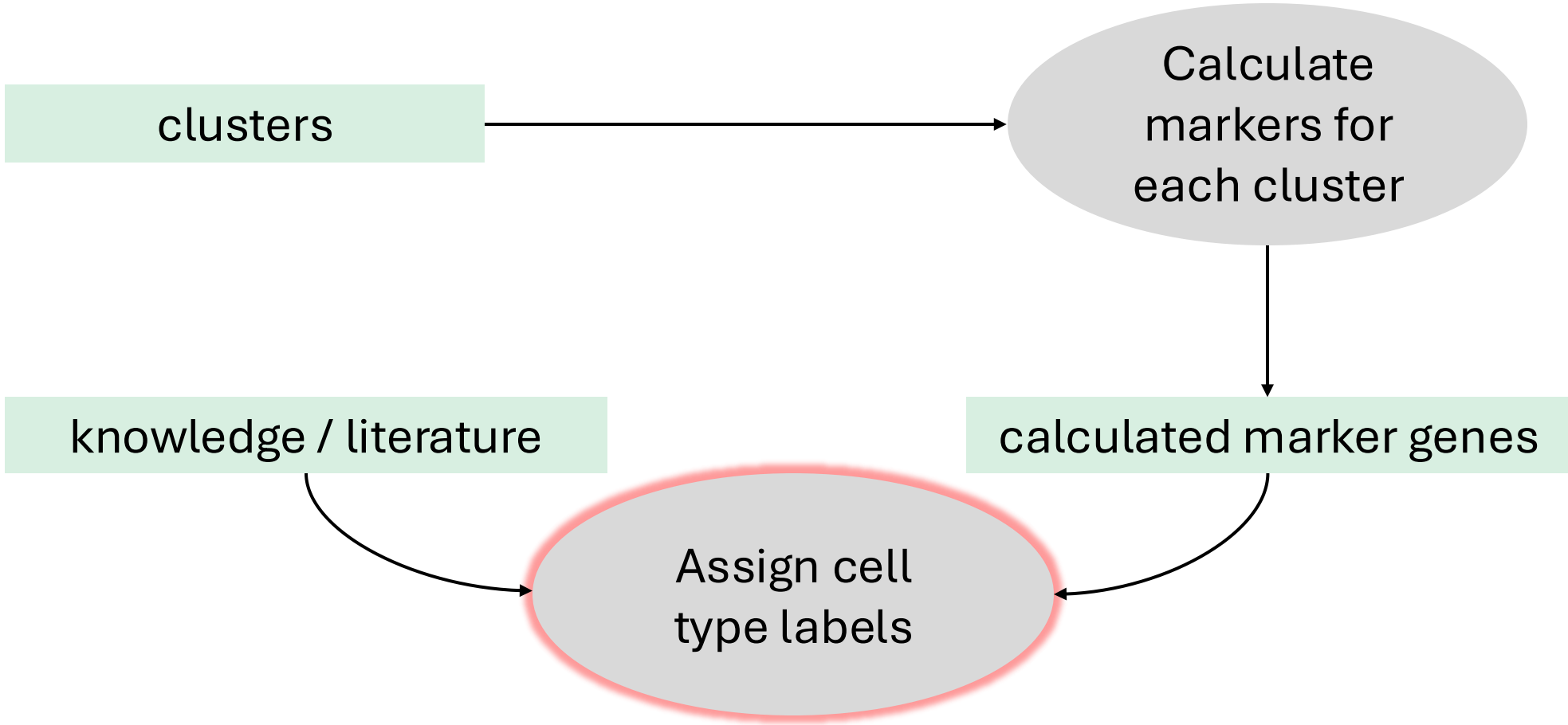
Method B - PBMC annotation / marker dotplot

calculated marker genes

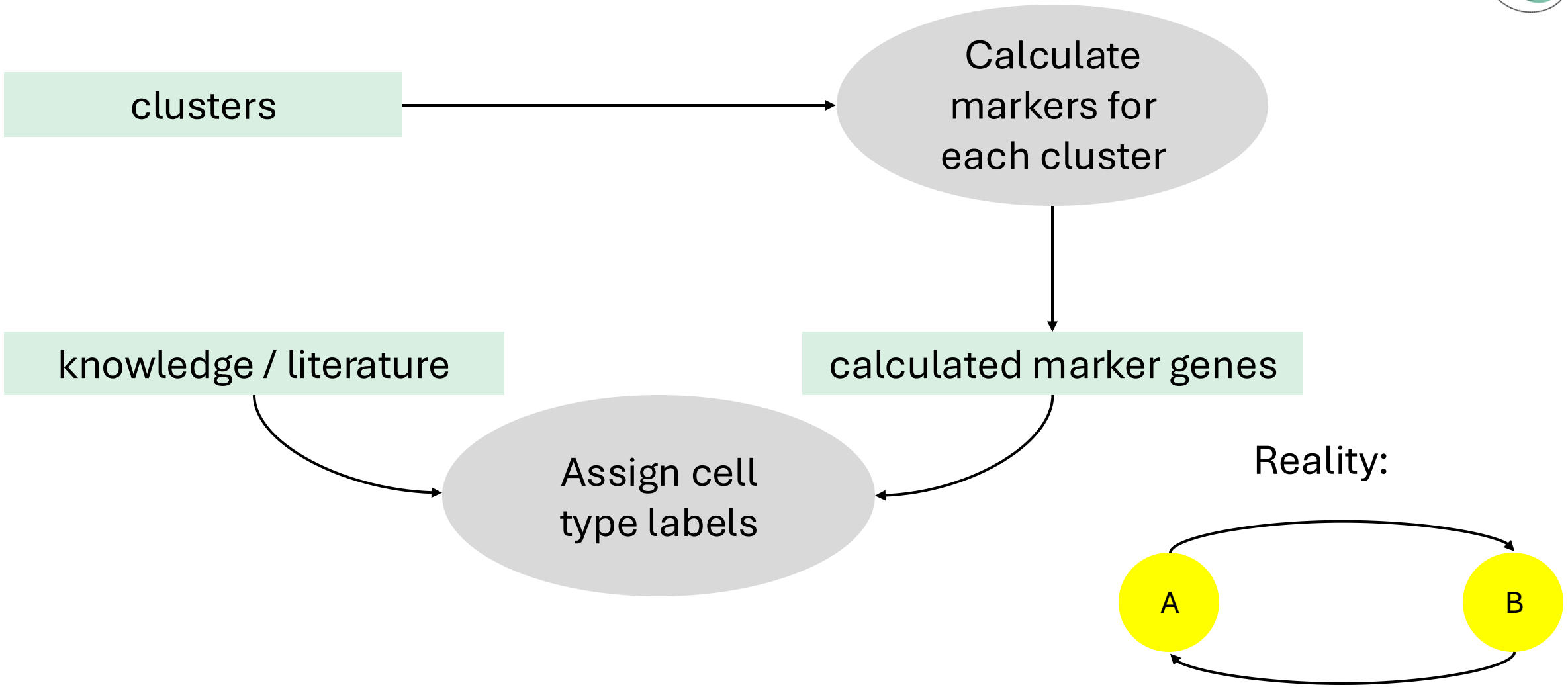


https://www.sc-best-practices.org/cellular_structure/annotation.html#manual-annotation

Method B - PBMC annotation / ingredients



Method B - PBMC annotation / ingredients

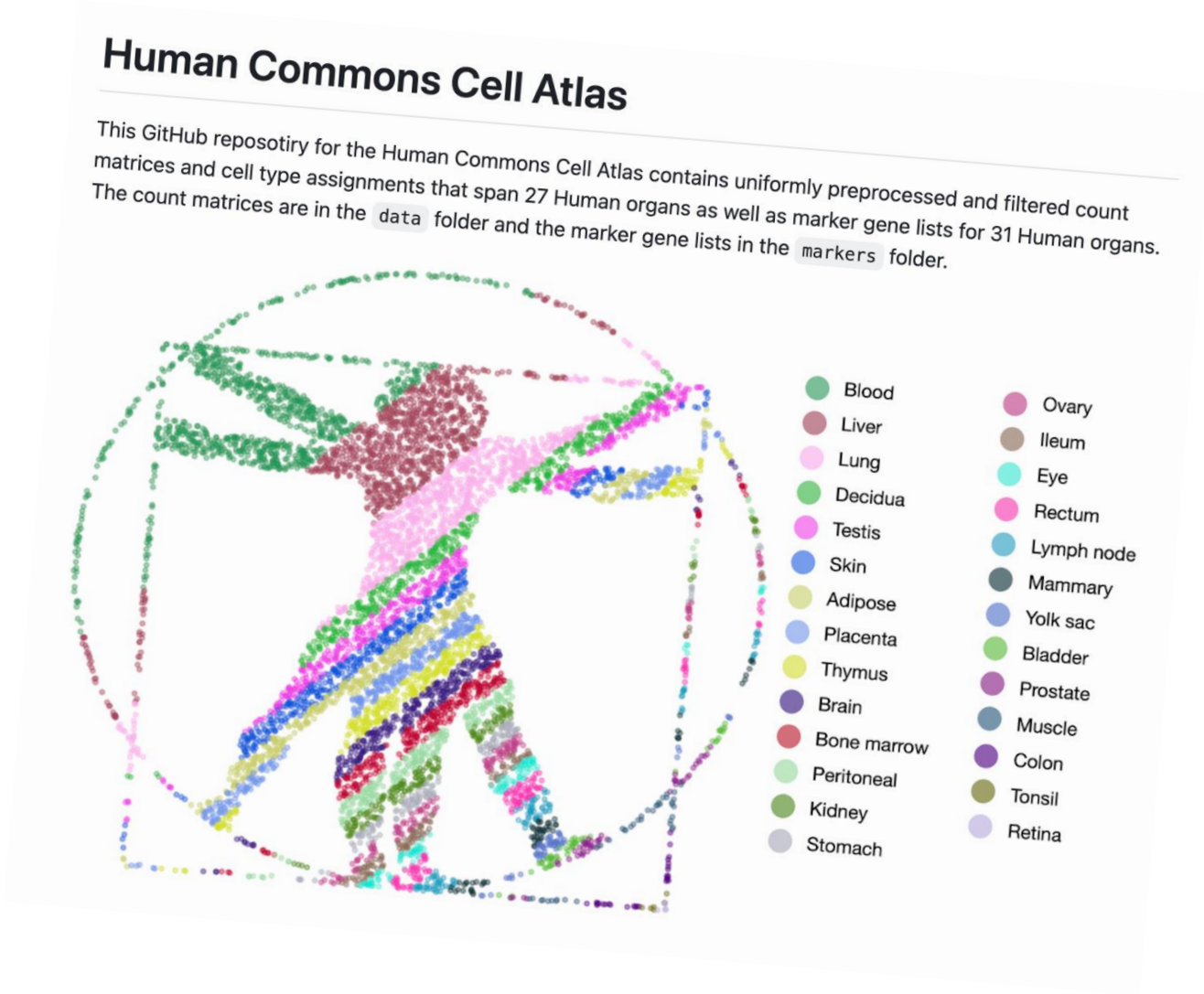


Where to find marker genes?

General literature

Annotated single-cell datasets

Dedicated marker gene resources



https://github.com/cellatlas/human

Outlook: Automated annotation methods

! Automated annotation methods currently serve as a starting point for annotation, not as an endpoint. Quality depends on:

Annotation method

Quality of training data

https://www.sc-best-practices.org/cellular_structure/annotation.html#general-remarks

Outlook: Automated annotation methods

! Automated annotation methods currently serve as a starting point for annotation, not as an endpoint. Quality depends on:

Annotation method

Quality of training data

Similarity between training and query data

https://www.sc-best-practices.org/cellular_structure/annotation.html#general-remarks

Outlook: Automated annotation methods

! Automated annotation methods currently serve as a starting point for annotation, not as an endpoint. Quality depends on:

Annotation method

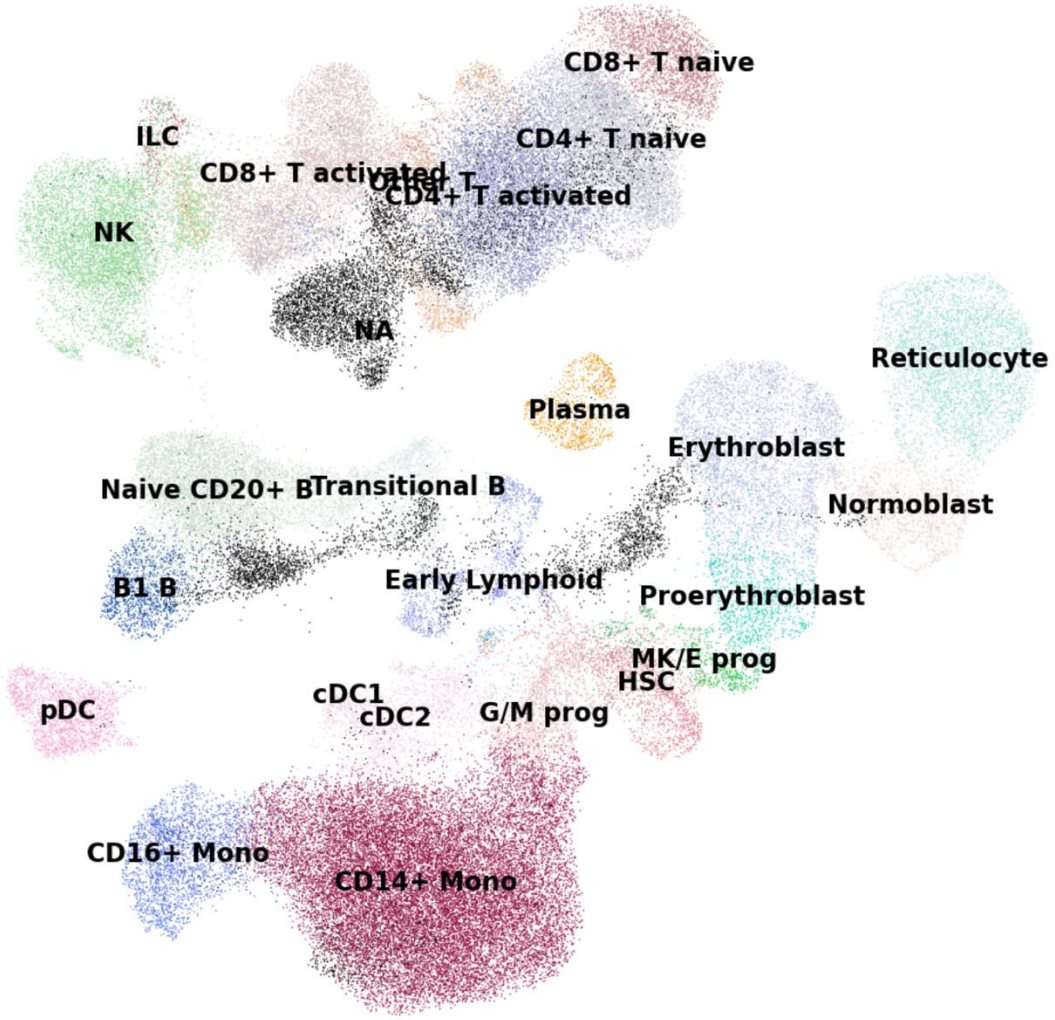
Quality of training data

Similarity between training and query data

- Marker gene-based methods (e.g. Garnett, CellAssign)
- Classifier trained on previous datasets/atlases (e.g. CellTypist, Clustifyr)
- Reference mapping (e.g. scArches, Symphony, Azimuth)

https://www.sc-best-practices.org/cellular_structure/annotation.html#general-remarks

Finally, an annotated dataset.



https://www.sc-best-practices.org/cellular_structure/annotation.html