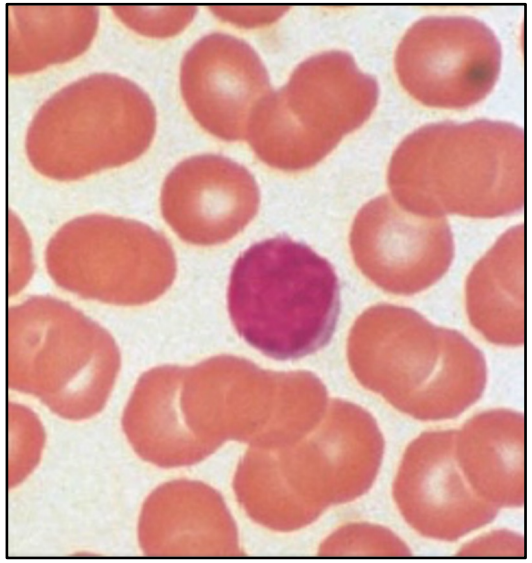
Differential expression and biological interpretation

Single-Cell RNA-Seq Computational Workshop

October 24, 2023



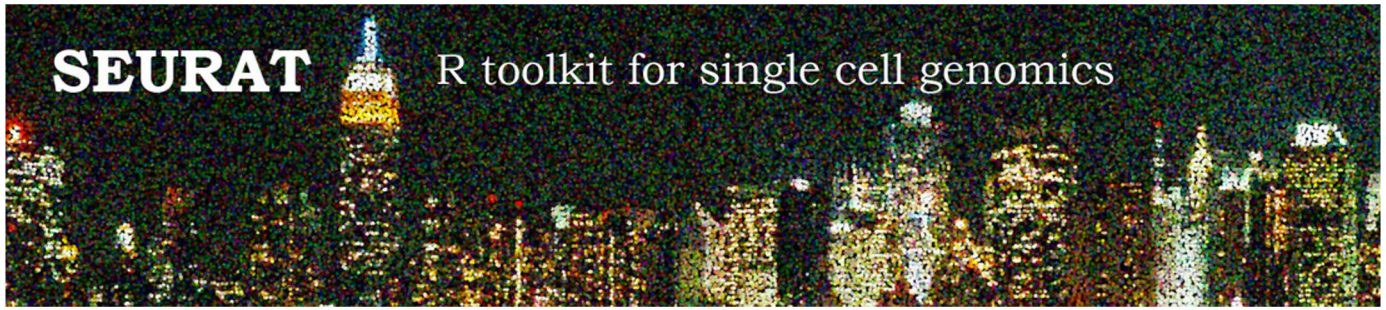
BROAD INSTITUTE

KLARMAN CELL OBSERVATORY AT BROAD INSTITUTE

HUMAN CELL ATLAS

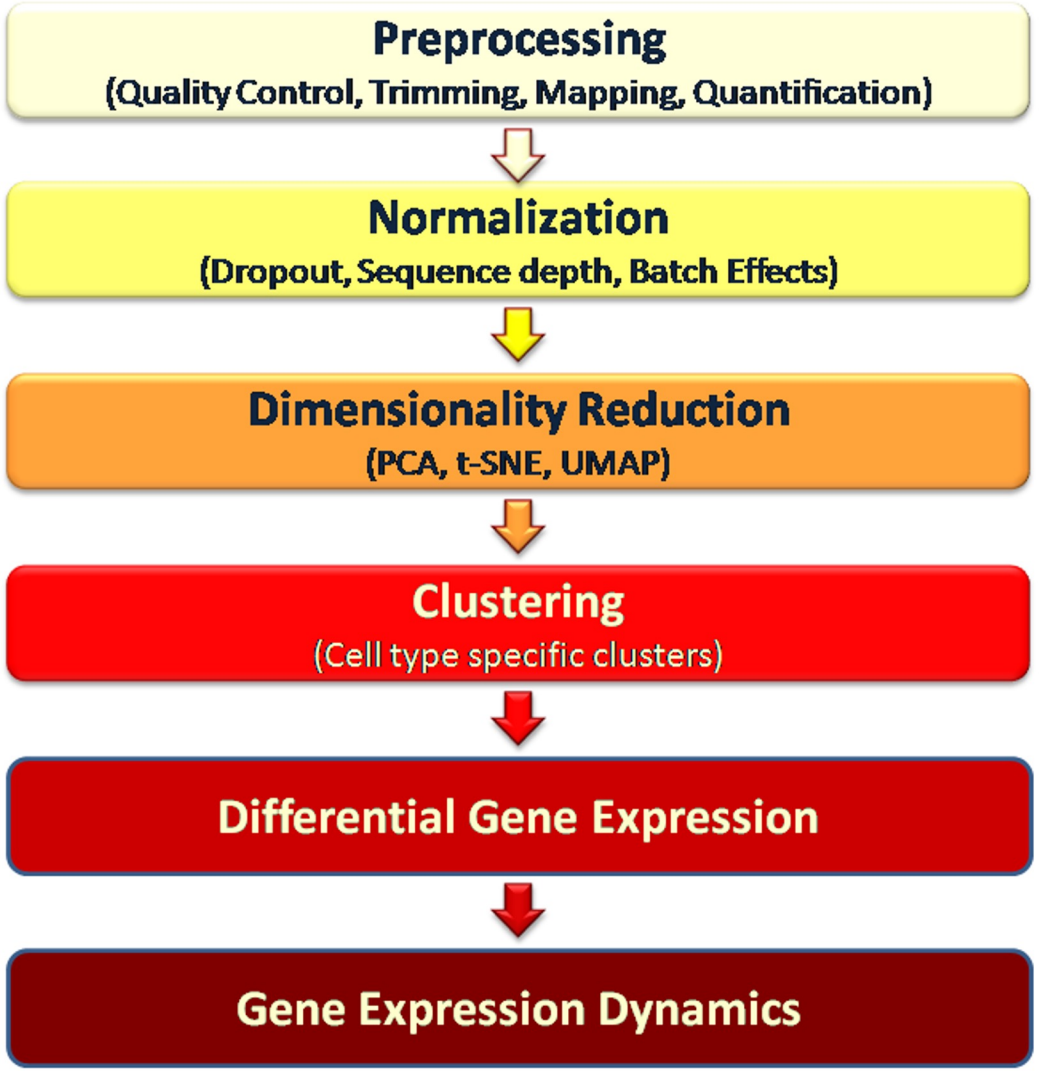
humancellatlas.org

Main tools for single-cell data analysis



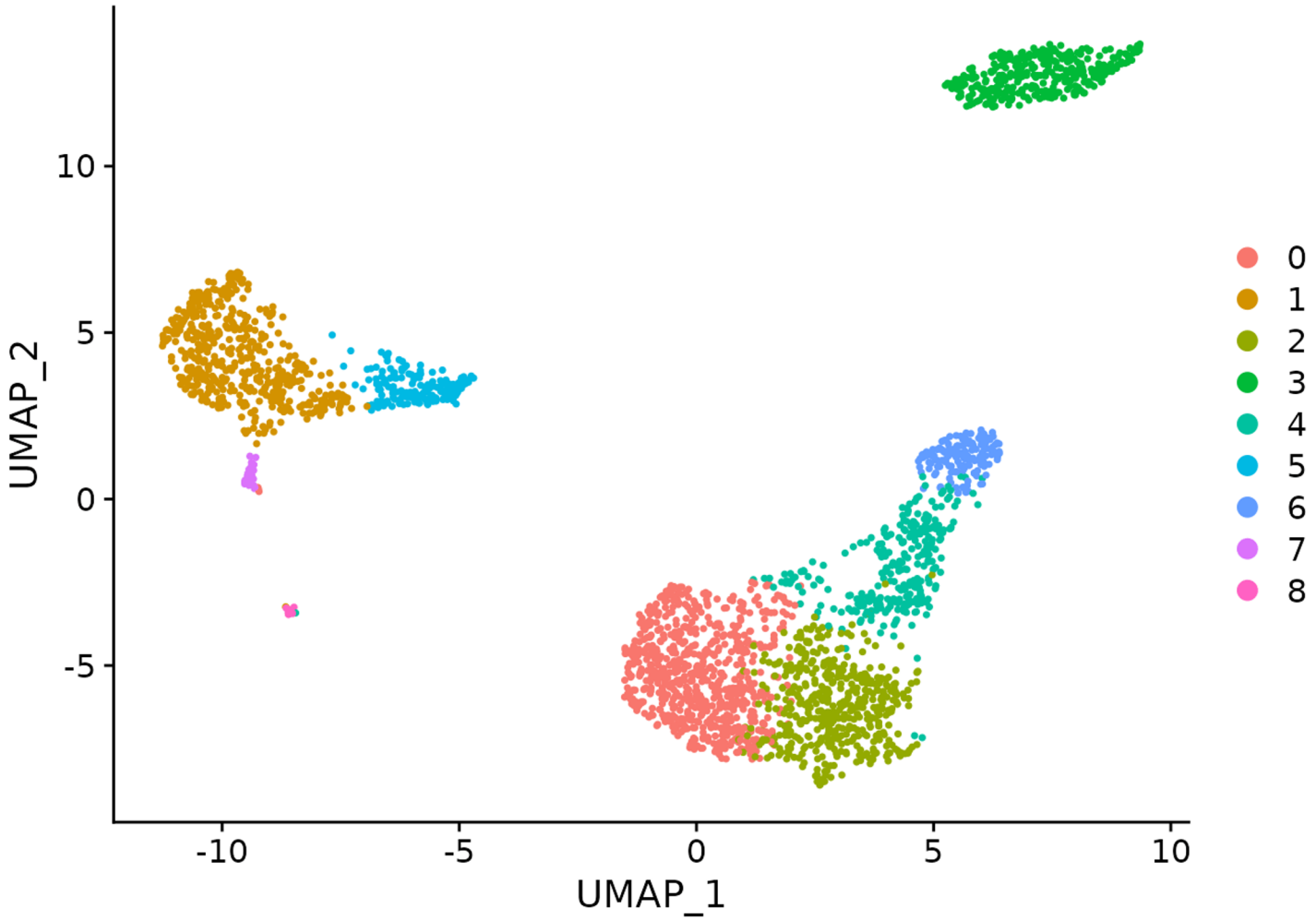
Methods

- Data import
- Normalization
- Dimension reductior
  - PCA
  - tSNE
  - UMAP
- Data visualization
- Differential expressi
- Geneset enrichmen



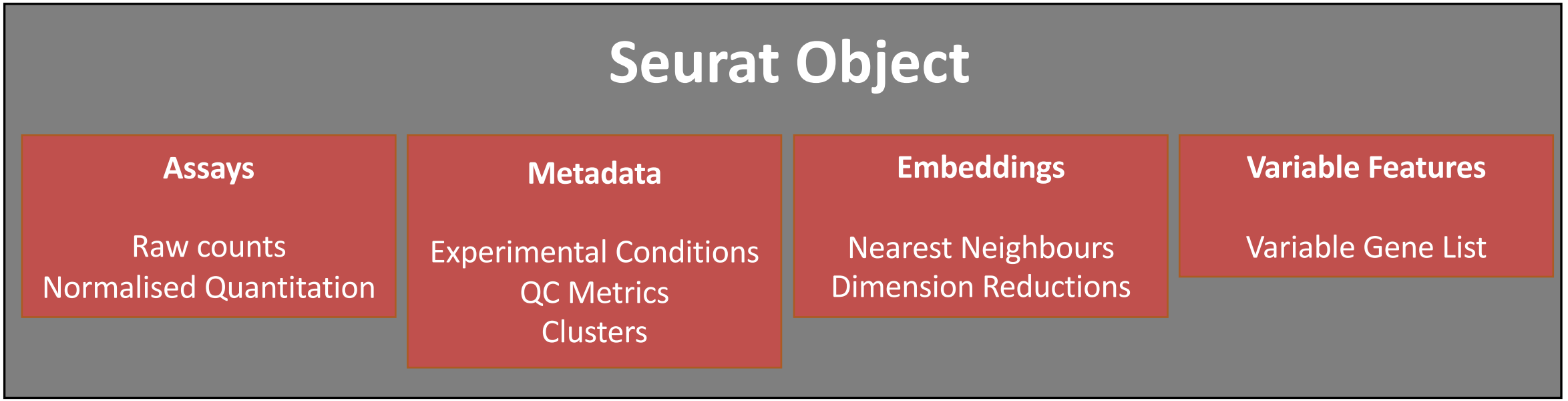
Seurat

- The most used tool
- Good documentatior
- Several tutorials
- Many methods
- Extendable



Seurat Data Structure

- All data in object
- Created from raw matrix or h5 file



Seurat Metadata

- Data Frame of QC metrics (cols = metrics, rows = cells)
  - Imported classifications (not automatically carried over from cellranger, add from aggr csv file)
  - Derived clusters
  - Some automatically defined – can add your own

- Access directly or indirectly
  - `data$my.qc.metric`
  - `data@meta.data$my.qc.metric`
  - `data[[]]`

Seurat Metadata

```
> head(data[[]])
  nCount_RNA nFeature_RNA gem_id     orig.ident
1       6538         2108      1 Influenza_day1
2       6742         1579      1 Influenza_day1
3       1420          810      1 Influenza_day1
4       1077          681      1 Influenza_day1
5       6303         2342      1 Influenza_day1
6       8947         2793      1 Influenza_day1
```

Seurat Quantitative Data

- Counts
  - Top level is a sparse matrix (rows = genes, cols = cells)
  - Can also use `GetAssayData(data, slot="counts")`
  - Can also use `data@assays$RNA@counts`

- Normalised data
  - A second independent matrix
  - `GetAssayData(data, slot="data")`
  - `data@assays$RNA@data`

- Can filter by subsetting the top level matrix

Seurat Quantitative Data

```
> GetAssayData(data, slot="counts")
11 x 2 sparse Matrix
          TGAGA-1 AAGGG-1
Rgs20           .       .
Atp6v1h         1       2
Alkal1          .       .
St18            .       .
Pcmtd1          .       1
Gm26901         .       .
Sntg1           .       .
```

```
> data[,]
11 x 2 sparse Matrix
          TGAGA-1 AAGGG-1
Rgs20           .       .
Atp6v1h     0.928   1.377
Alkal1          .       .
St18            .       .
Pcmtd1          .   0.909
Gm26901         .       .
Sntg1           .       .
```

Seurat Embeddings

- Reductions
  - `Embeddings(data,reduction = "umap")`
  - `data@reductions$umap@cell.embeddings`

- Rows = cells, Cols = Projection axes
  - PCA
  - tSNE
  - UMAP

Seurat Embeddings

```
> Embeddings(data,reduction = "pca")
                PC_1          PC_2         PC_3          PC_4         PC_5
Reg3g     0.02209522  0.008870765 -0.01791399 -0.042812700 -0.18932551
Scgb3a1   0.01829751  0.005249075 -0.02640954 -0.044957754 -0.18863055
Retnla    0.02616540  0.005942941 -0.03967109 -0.049142055 -0.16423270
Bpifb1    0.01966987  0.005381136 -0.02367615 -0.045729821 -0.19999561
Cxcl13   -0.01369808 -0.010579552  0.08571013  0.004081651 -0.02472911
```

```
> Embeddings(data,reduction = "tsne")
                       tSNE_1     tSNE_2
AAACCTGAGAGTGAGA-1  14.182877 -12.592458
AAACCTGAGCGAAGGG-1 -24.612876   5.739248
AAACCTGAGCGTCTAT-1  21.212219   4.717356
AAACCTGAGCTACCTA-1   5.228508  21.568443
AAACCTGAGCTCCCAG-1  18.923168   6.299075
```

Seurat Methods

- Data Parsing
  - `Read10X`
  - `Read10X_h5*`
  - `CreateSeuratObject`
- Data Normalisation
  - `NormalizeData`
  - `ScaleData`
- Graphics
  - Violin Plot – metadata or expression (`VlnPlot`)
  - Feature plot (`FeatureScatter`)
  - Projection Plot (`DimPlot`, `DimHeatmap`)
- Dimension reduction
  - `RunPCA`
  - `RunTSNE`
  - `RunUMAP`
- Statistics
  - Select Variable Genes `FindVariableFeatures`
  - Build nearest neighbour graph `FindNeighbors`
  - Build graph based cell clusters `FindClusters`
  - Find genes to classify clusters (multiple tests) `FindMarkers`

*Requires installing the `hdf5r`

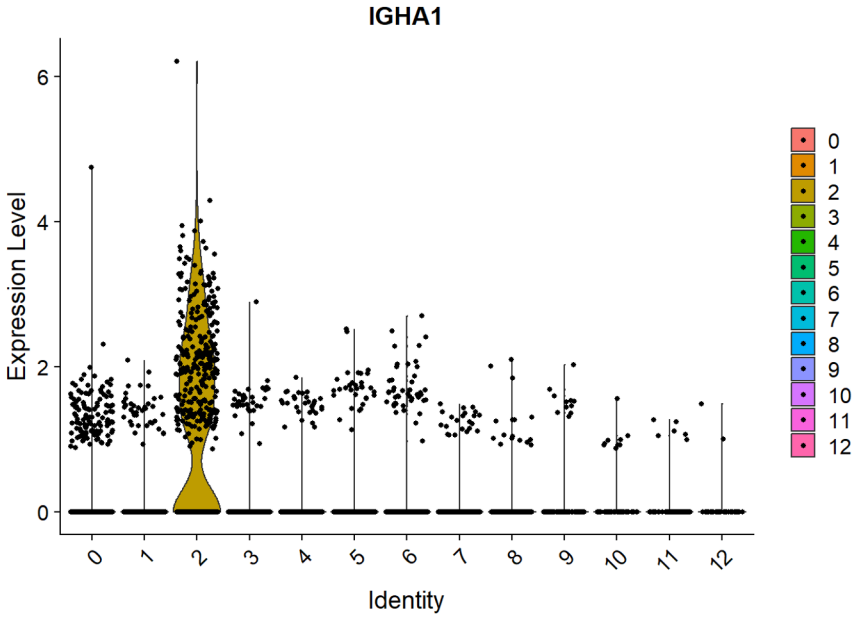
Statistical analysis of differences between clusters

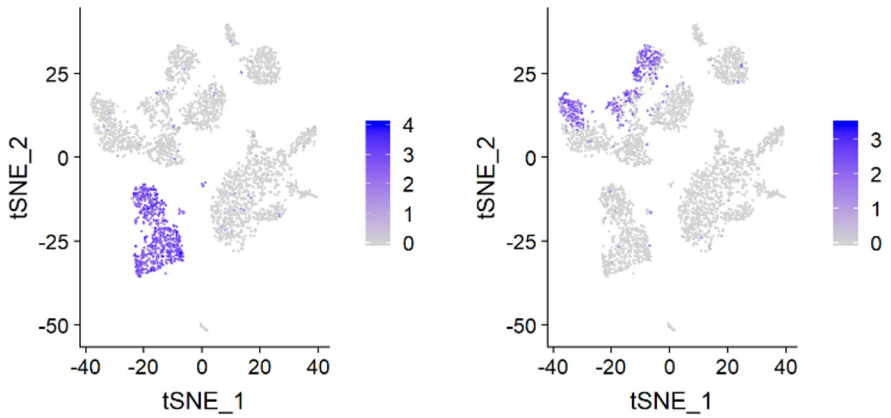
- Different types of hits
  - Quantitatively significant between clusters
  - Qualitatively different (predictive) of cluster membership
- Different type of markers
  - Global: Distinguish one cluster from all of the rest of the data
  - Local: Distinguish one cluster from another defined set of clusters
- Often filter genes based on coverage in the set or the size of groups
- Several choices of method to identify genes

Statistical analysis of differences between clusters

- Non-parametric
  - Wilcox rank sum test
- Parametric
  - T-test
  - Negative binomial (eg DESeq)

- Classification
  - ROC analysis
- Specialised
  - MAST






```
FindMarkers(
        data,
        ident.1 = 2,
        ident.2 = 6,
        test.use = "roc",
        only.pos = TRUE
)
```

Wang *et al. BMC Bioinformatics* (2019) 20:40
https://doi.org/10.1186/s12859-019-2599-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

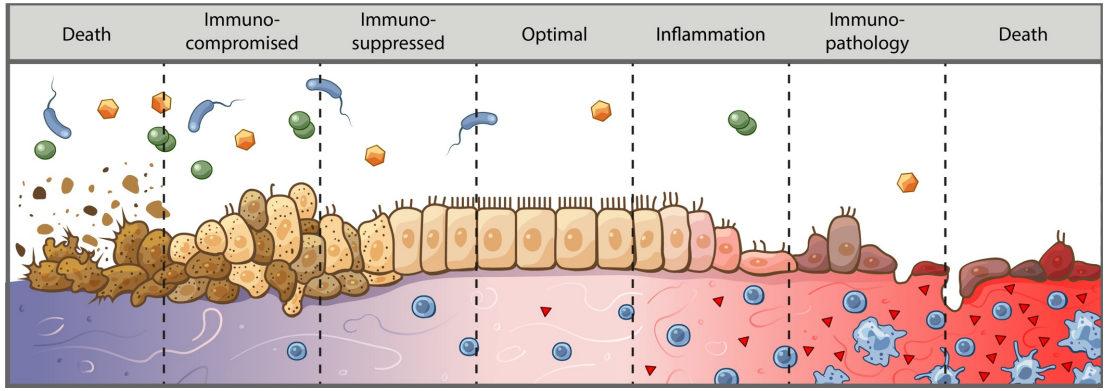
Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data

Tianyu Wang[1], Boyang Li[2], Craig E. Nelson[3] and Sheida Nabavi[4*]

Conclusions: In general, agreement among the tools in calling DE genes is not high. There is a trade-off between true-positive rates and the precision of calling DE genes. Methods with higher true positive rates tend to show low precision due to their introducing false positives, whereas methods with high precision show low true positive rates due to identifying few DE genes. We observed that current methods designed for scRNAseq data do not tend to show better performance compared to methods designed for bulk RNAseq

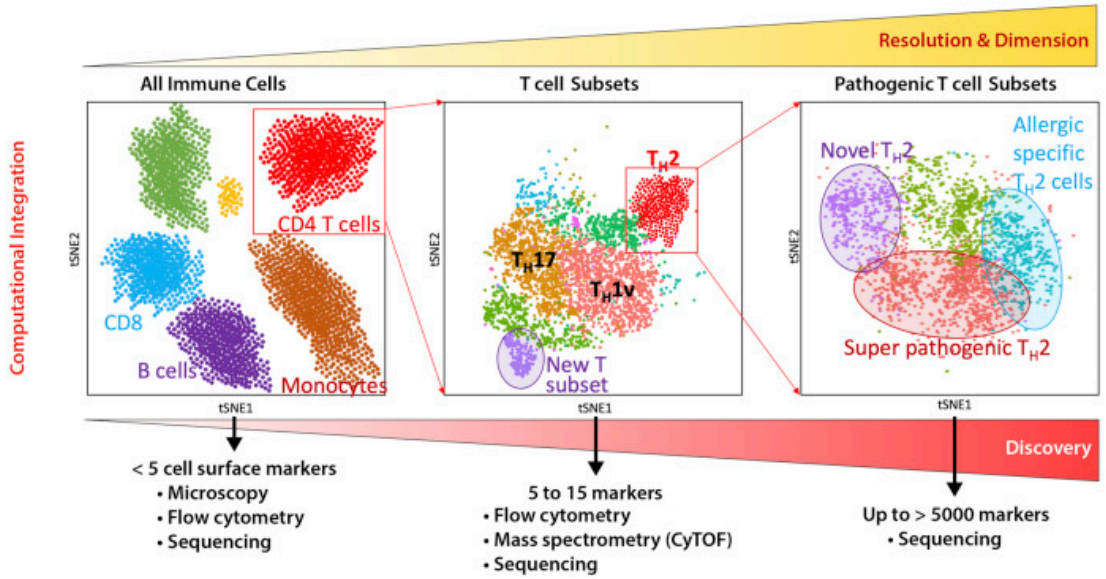
What are we actually annotating?

- 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



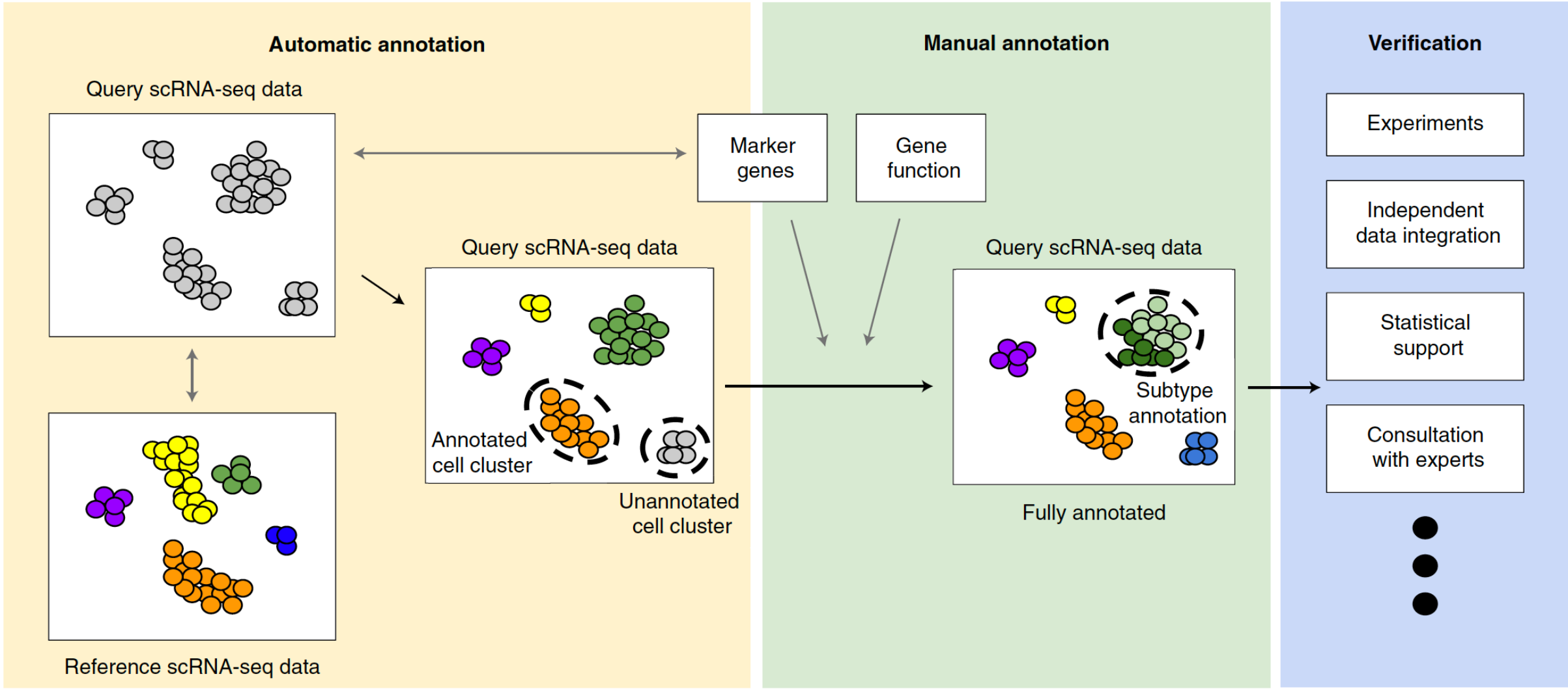
Biology to your data!

- 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.*

How to identify the cell types of clustered cells?



Clarke, Z. A. at *al.* (2021) *Nature Protocols*. 16: 2749-2764.
Abdelaal, T. et al., (2019) *Genome Biology*. 20:194.

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

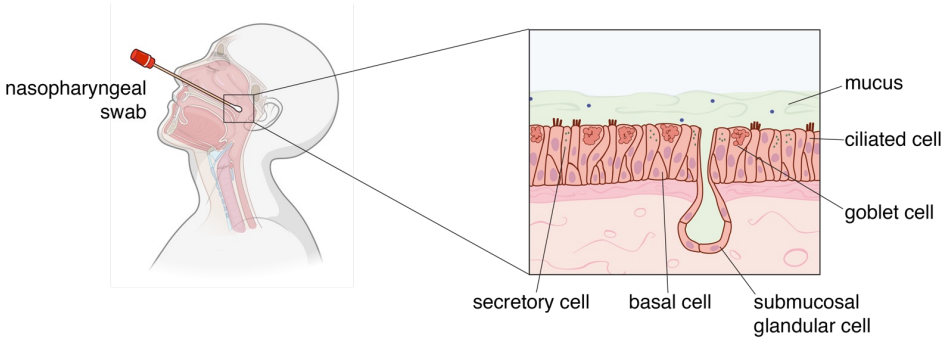
Host-pathogen annotations

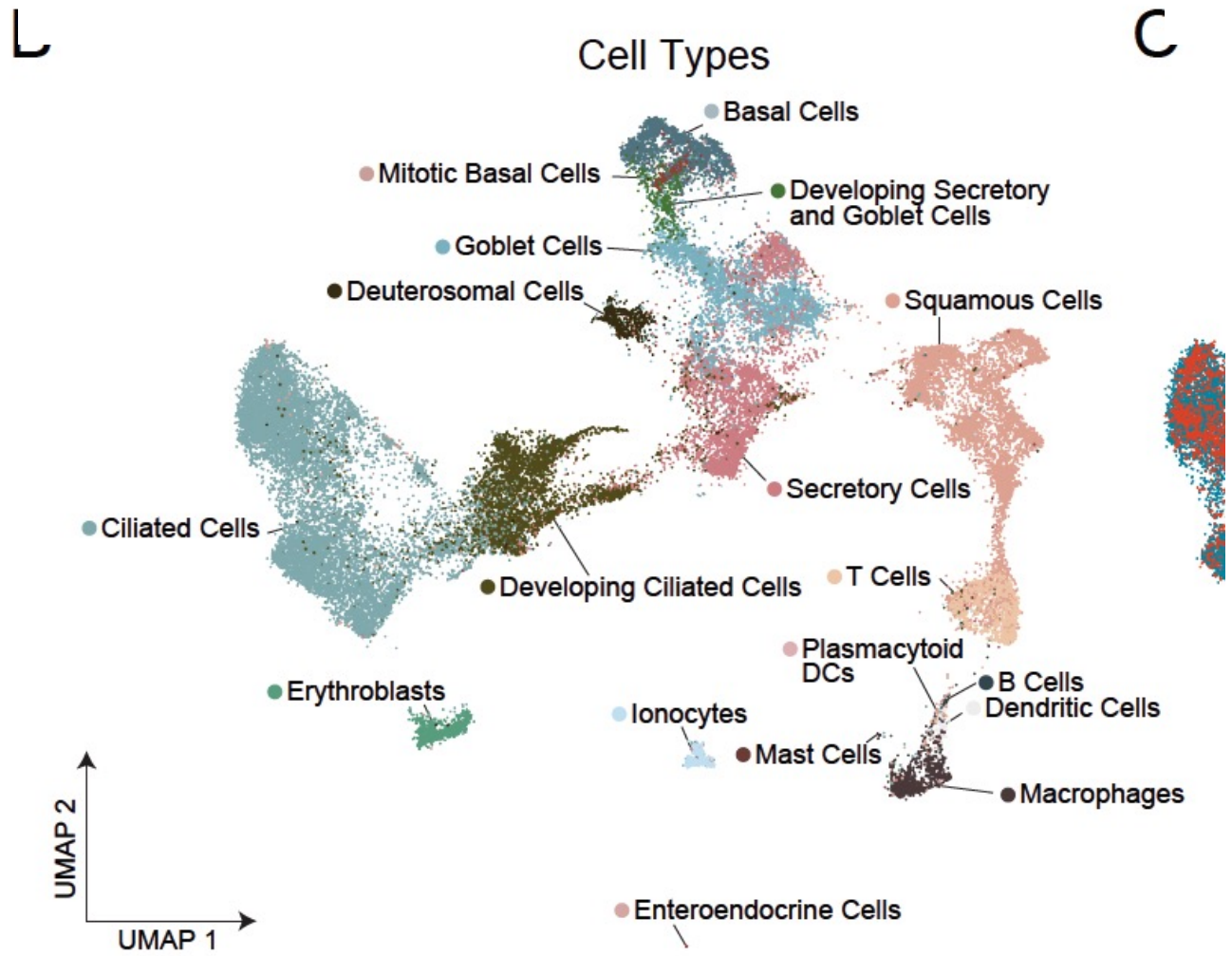
- ScRNA-seq will capture poly-adenylated pathogen reads for intra-cellular pathogens
- Required alignment against custom, combined host-pathogen genome
- Pathogen RNA != intra-cellular pathogen presence

  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

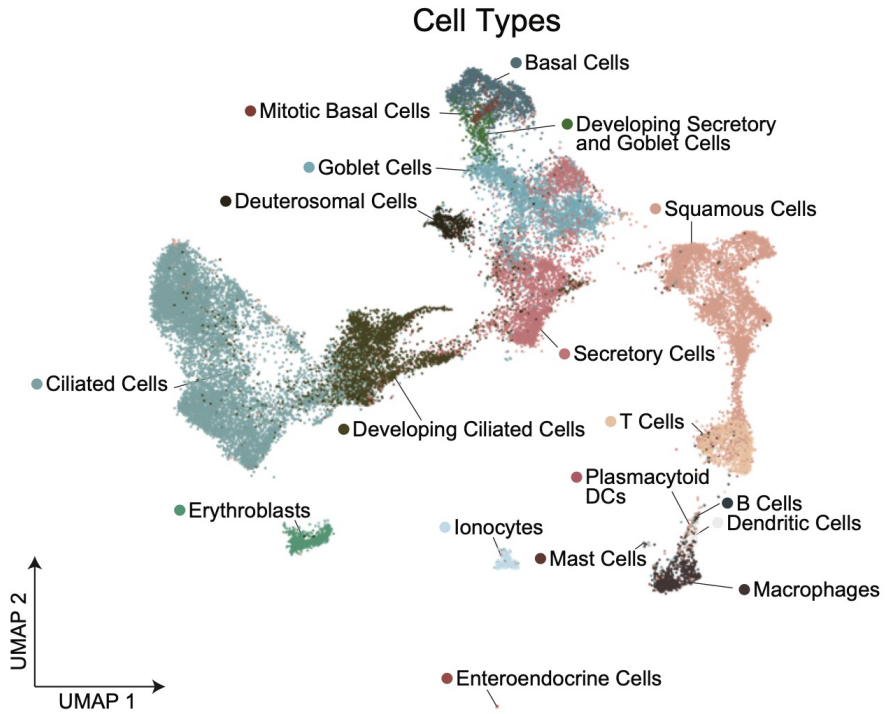




32588 cells across 58 individuals

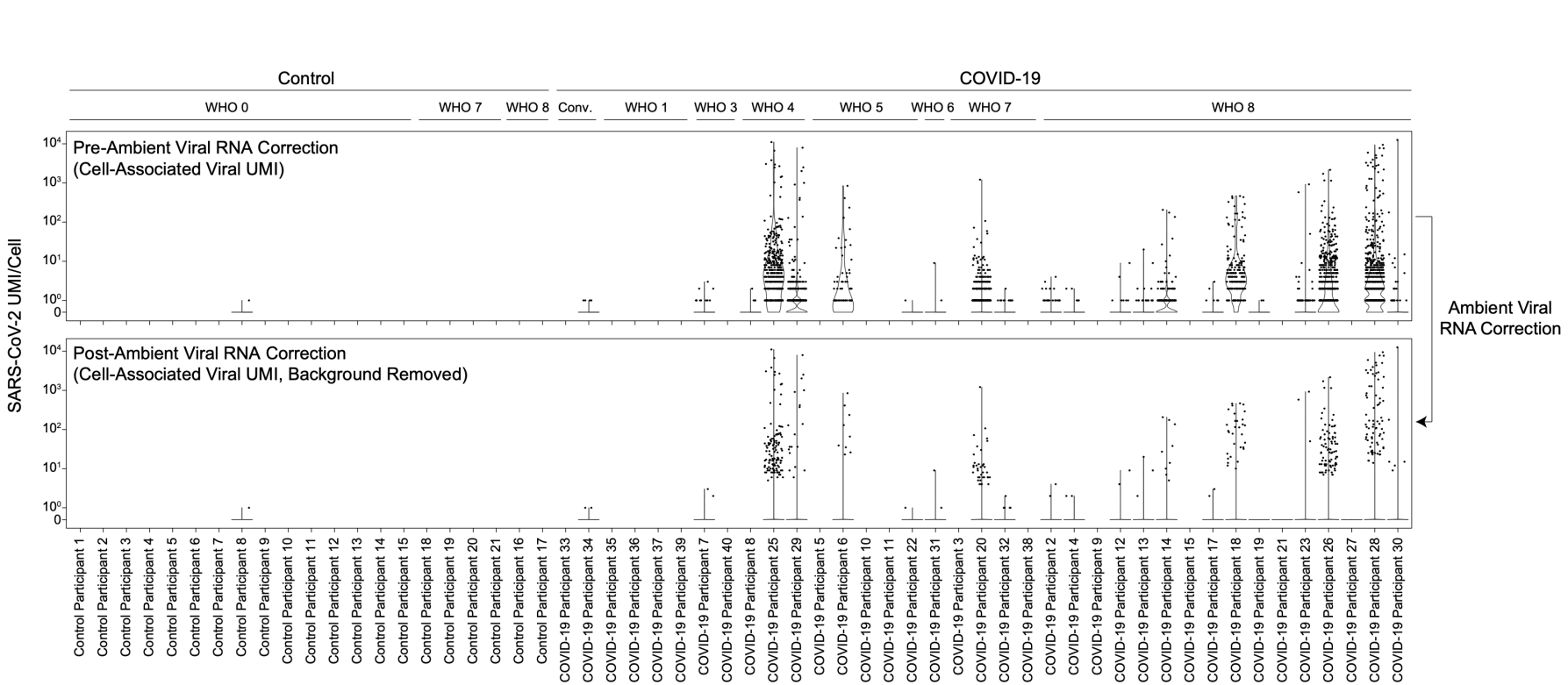
Ziegler et al., *Cell*, 2021.

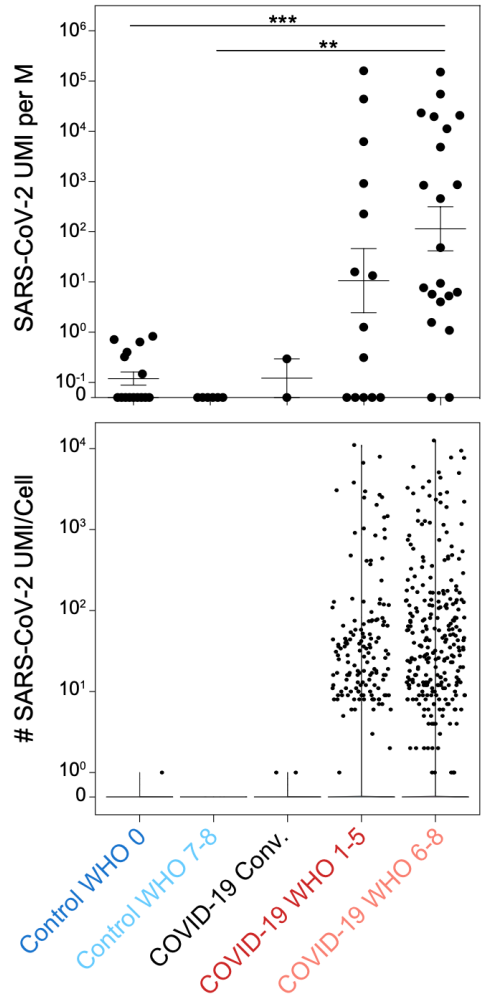
Host-pathogen annotations (cont'd): A COVID-19 example



Ziegler et al., *Cell*, 2021.

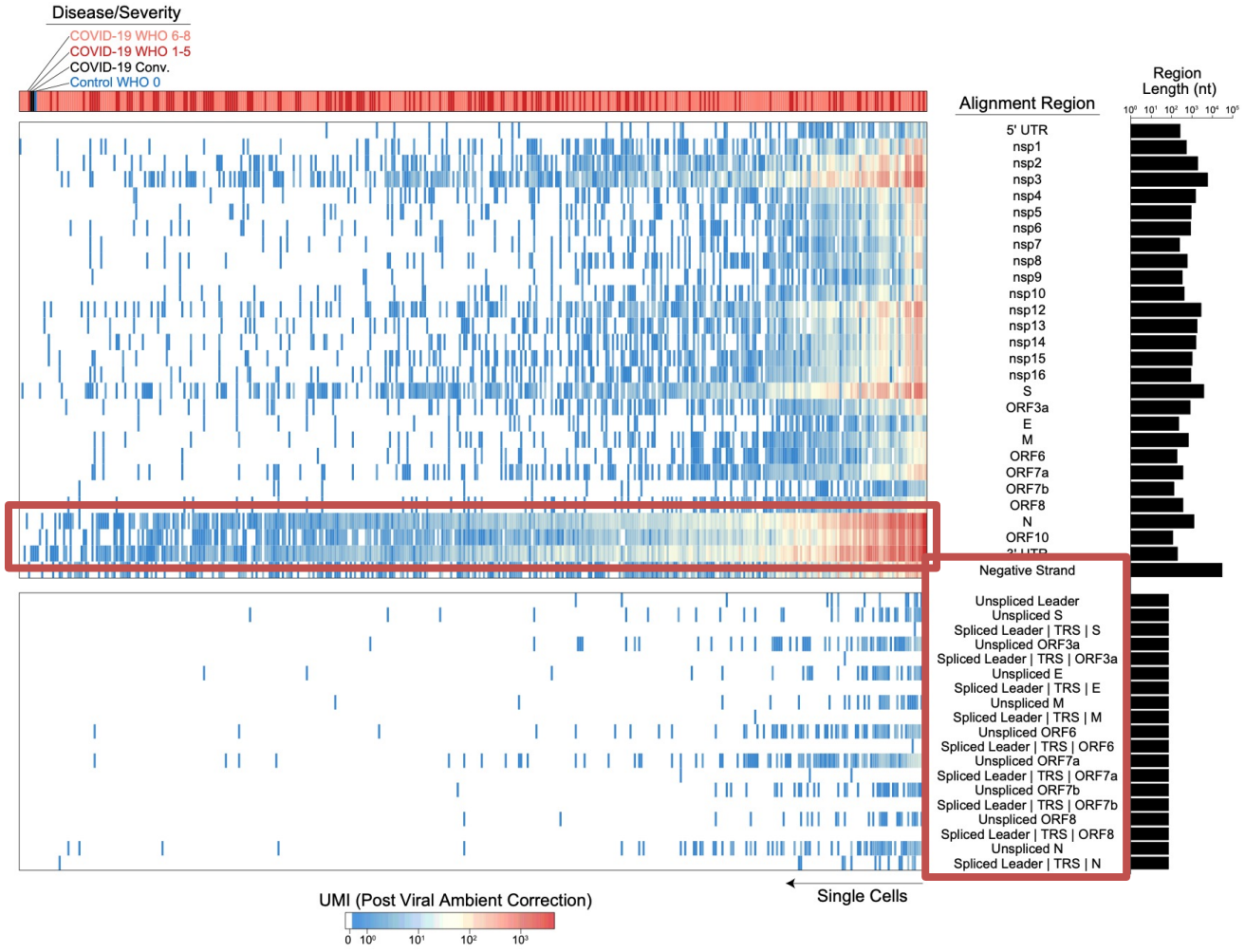
Host-pathogen annotations (cont'd): A COVID-19 example

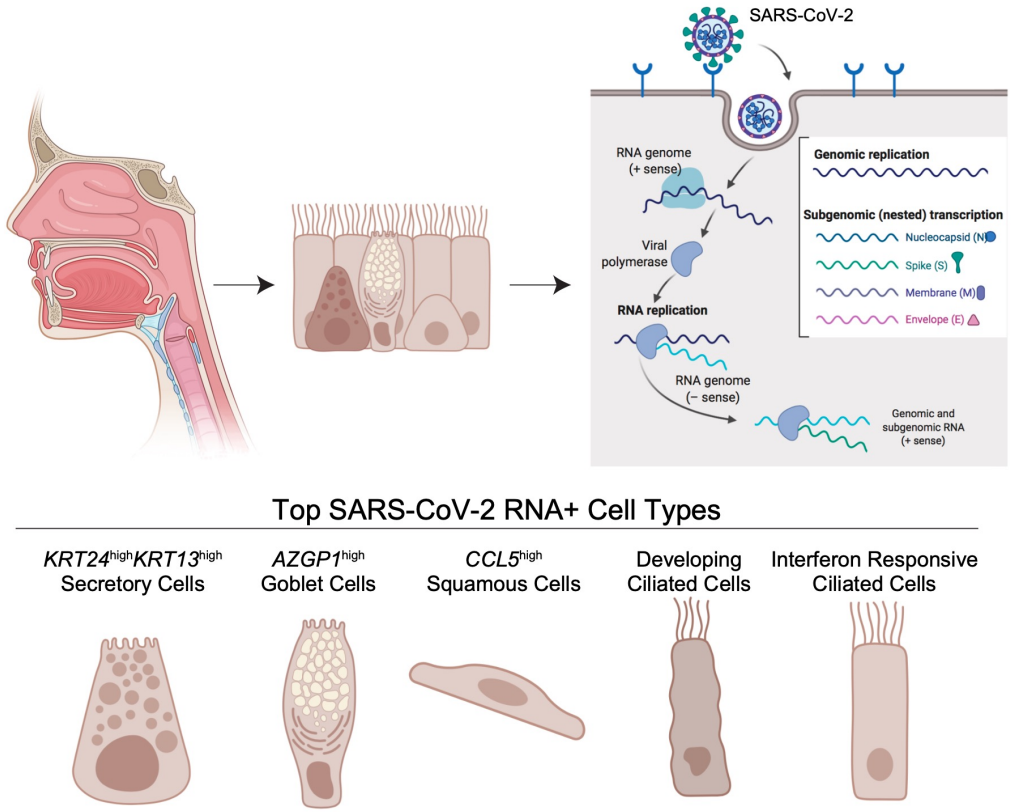




Fleming et al., *bioRxiv*, 2019; Kotliar et al., *Cell*, 2020; Ziegler et al., *Cell*, 2021.

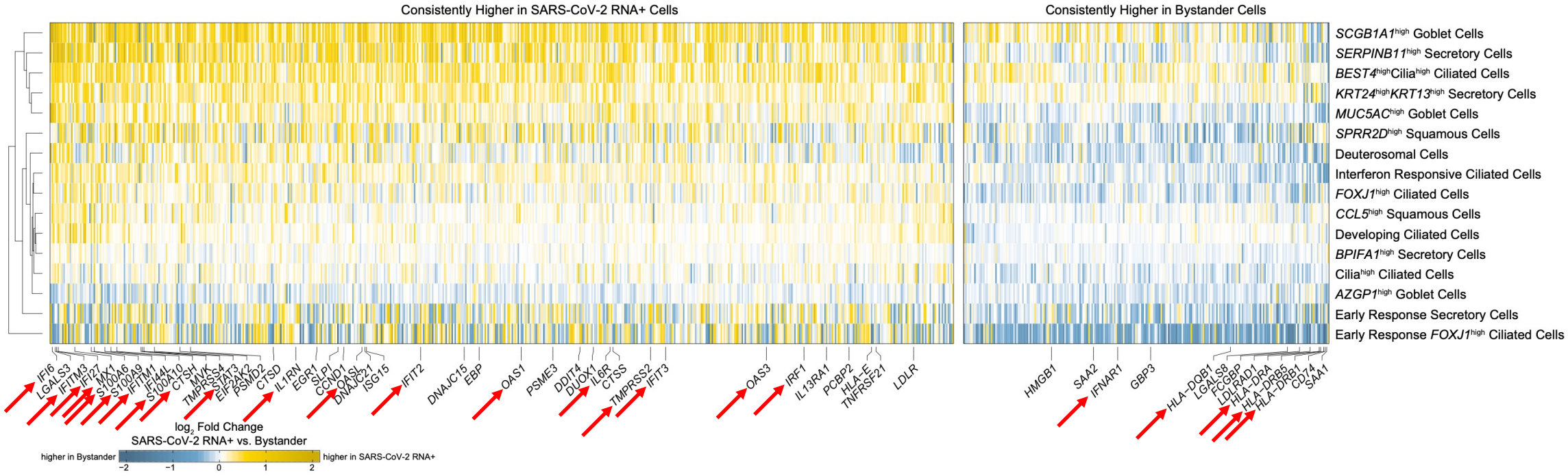
Host-pathogen annotations (cont'd): A COVID-19 example





Ziegler et al., *Cell*, 2021.

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



Ziegler et al., *Cell*, 2021.

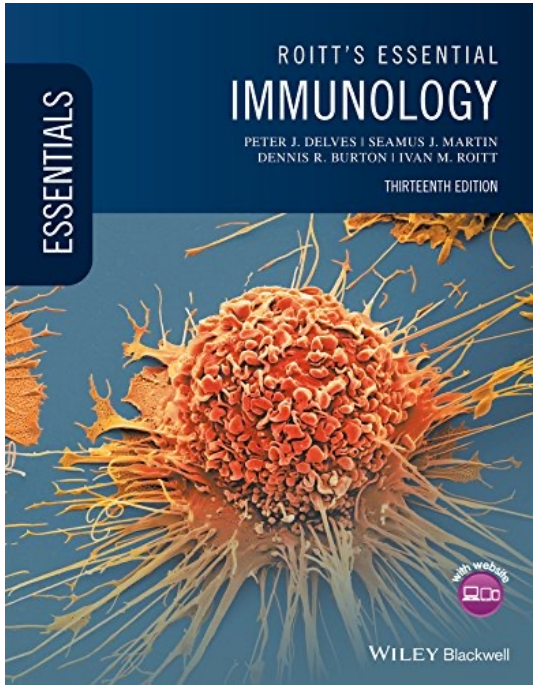
Approach 2: Databases and Reference atlases

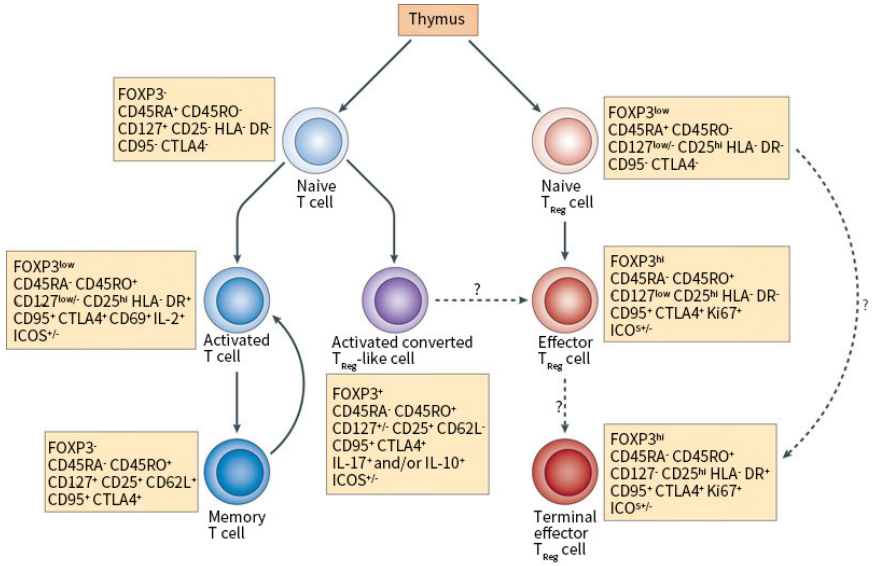
- Marker genes and databases
  - Literature
  - Tissue and protein DB
  - Single-cell atlases
- Automated tools
  - Correlation-based
  - Supervised classification

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

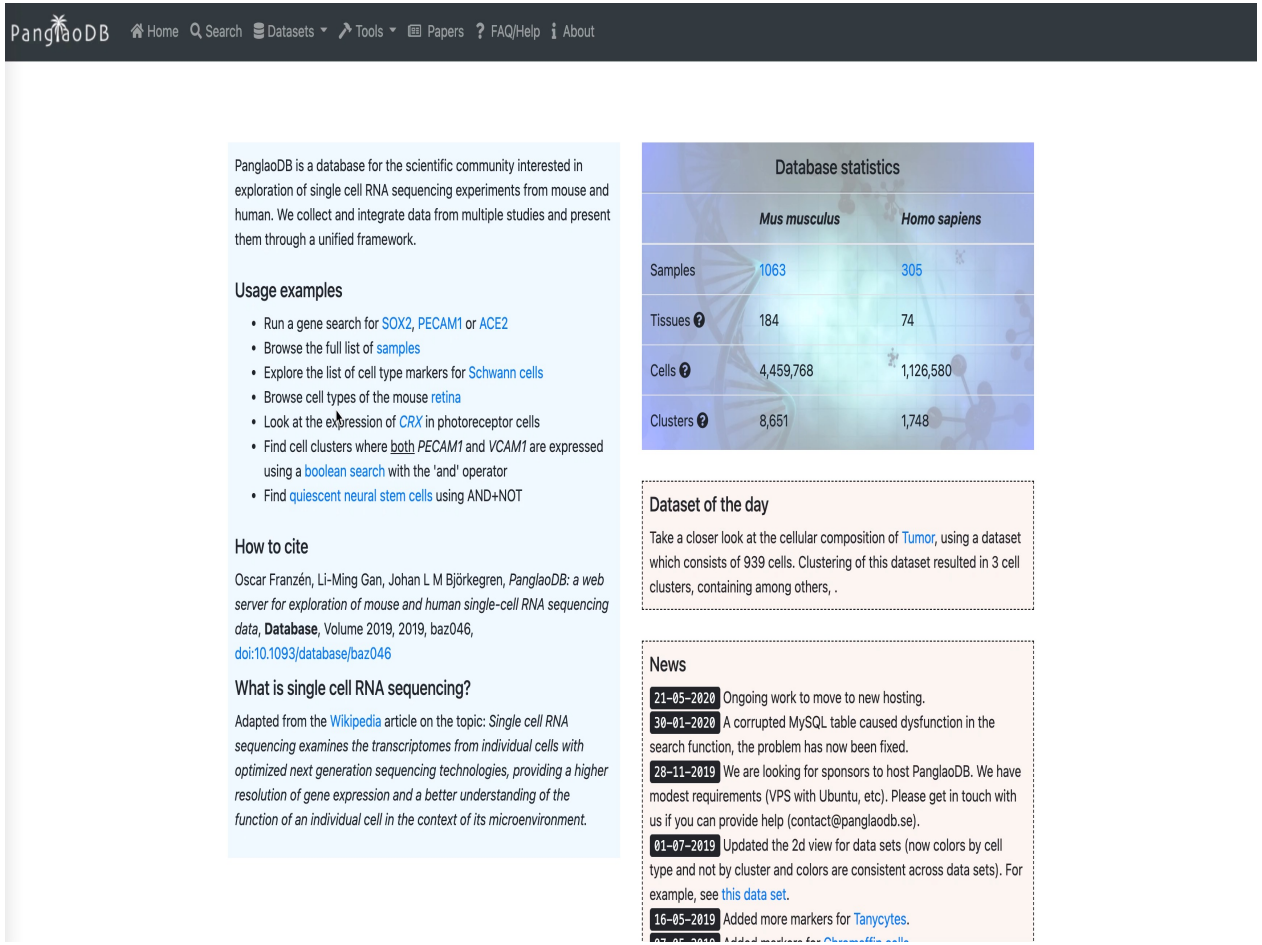
in Meaning you have your set of genes and search for them one by one these DB

- 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/



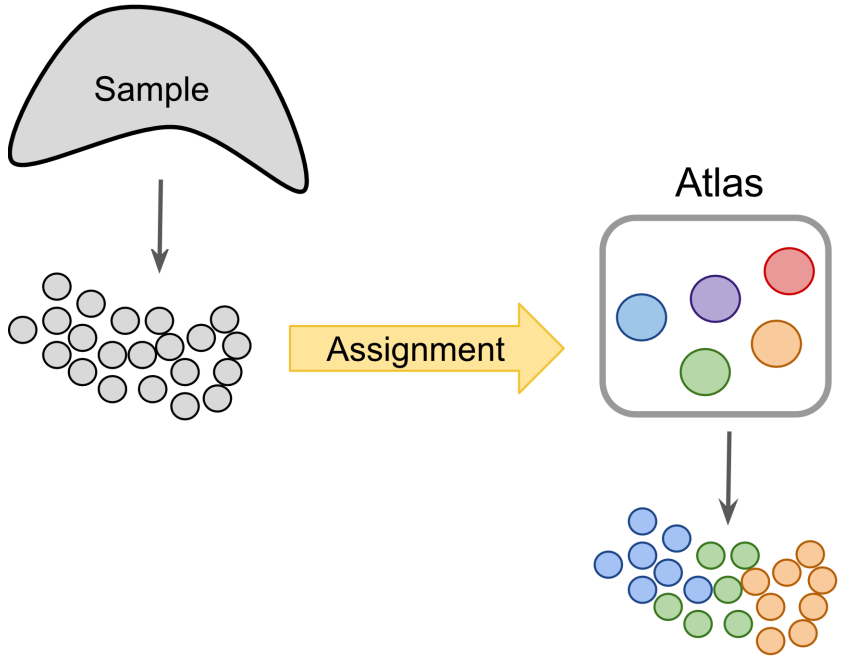
Approach 2: Automated tools

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

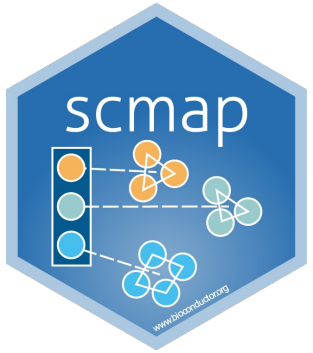
- Correlation-based
- Supervised classification

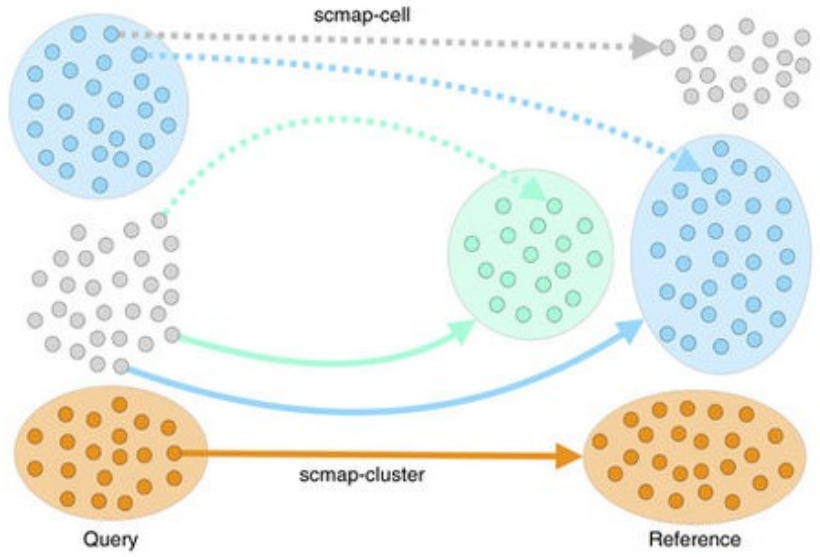


Approach 2: Automated tools

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



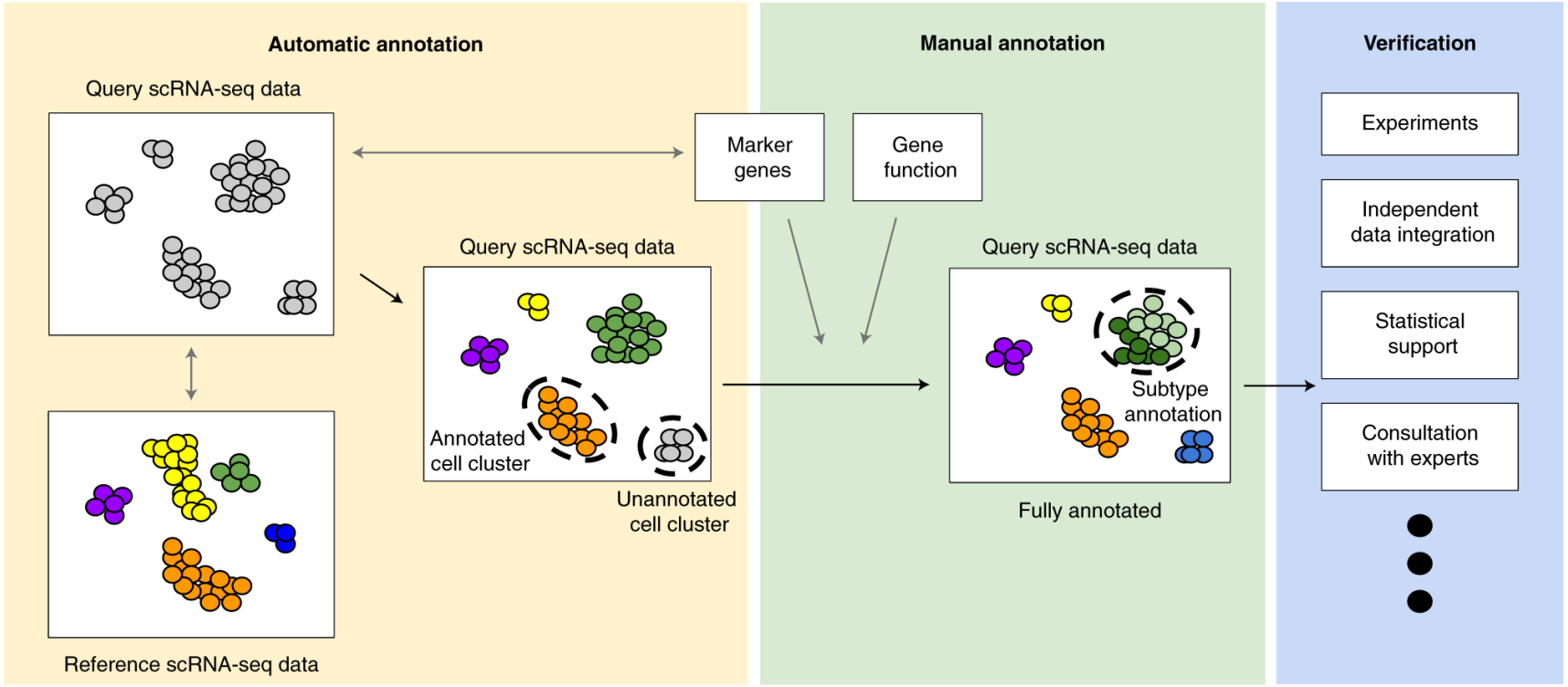


The search can be done by cluster or by cell , in each cluster/cell by its centroid (a vector of the median value of the expression of each gene) and measure the similarity between c and each cluster centroid or cell in the refence is calculated and ranked

Pasquini et al, 2021, Computational and Structural Biotechnology Journal

Recent helpful articles on data integration methods

- 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).

Differentially expressed genes across biological conditions

When detecting gene differentially expressed across different biological conditions (e.g. case vs control, treated vs untreated, etc…), what is the replicate? A single cell? A sample?

Treating a single cell as a replicate will lead to false discoveries. Strongly consider using pseudobulk models or mixed effect models when comparing samples across conditions.

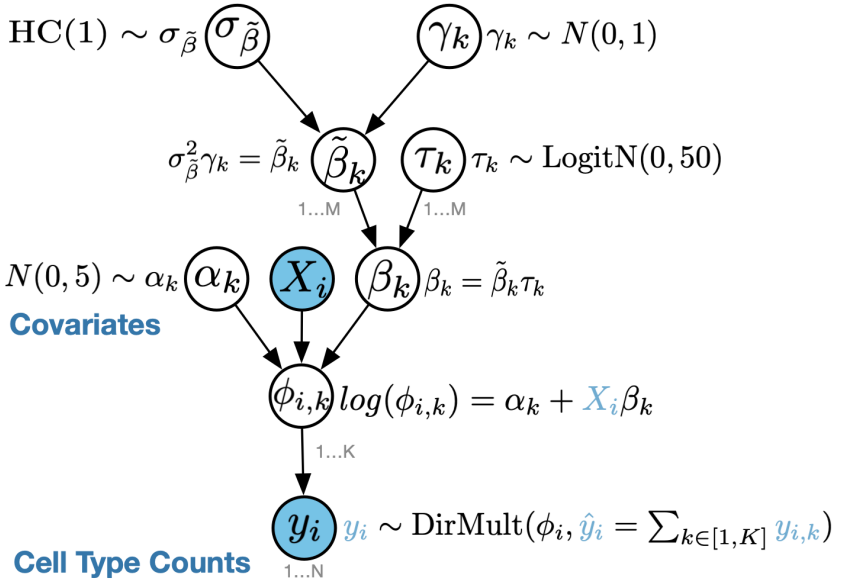
- Crowell, Helena L., et al. "Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data." *Nature communications* (2020).
- Squair, Jordan W., et al. "Confronting false discoveries in single-cell differential expression." *Nature communications* (2021).
- https://www.nxn.se/valent/2019/2/15/handling-confounded-samples-for-differential-expression-in-scrna-seq-experiments

Testing for significant changes in cell type abundance across biological conditions

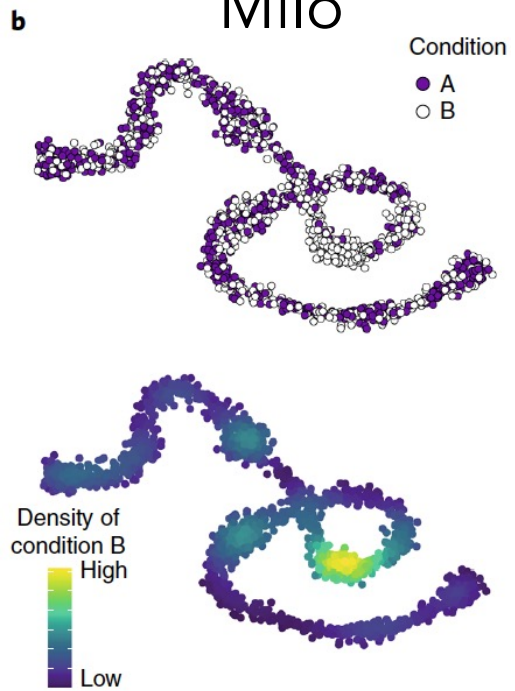
Assume we have 3 cell types (A, B, C) in a treated and untreated sample. If the frequency of cell type A increases in the treated sample due to the treatment, what will happen to the frequencies of cell types B and C?

The compositional nature of cell type abundance data means the abundances of cell type A, cell type B, and cell type C are not independent of one other. Their frequencies must always sum to 1.

scCODA

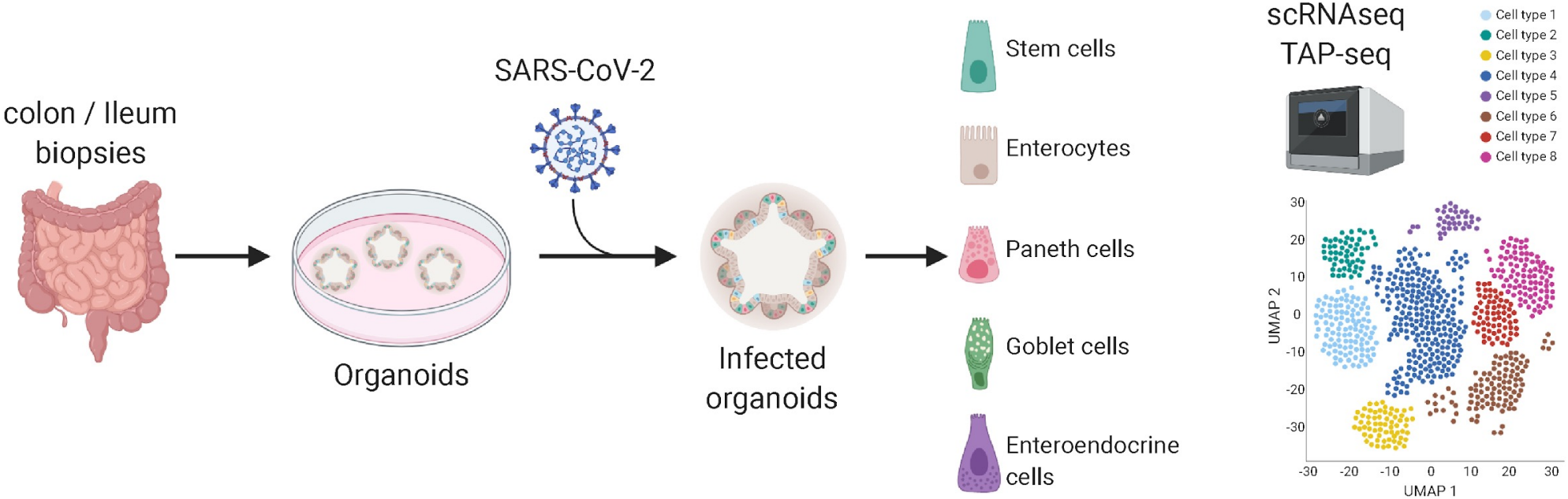


Milo

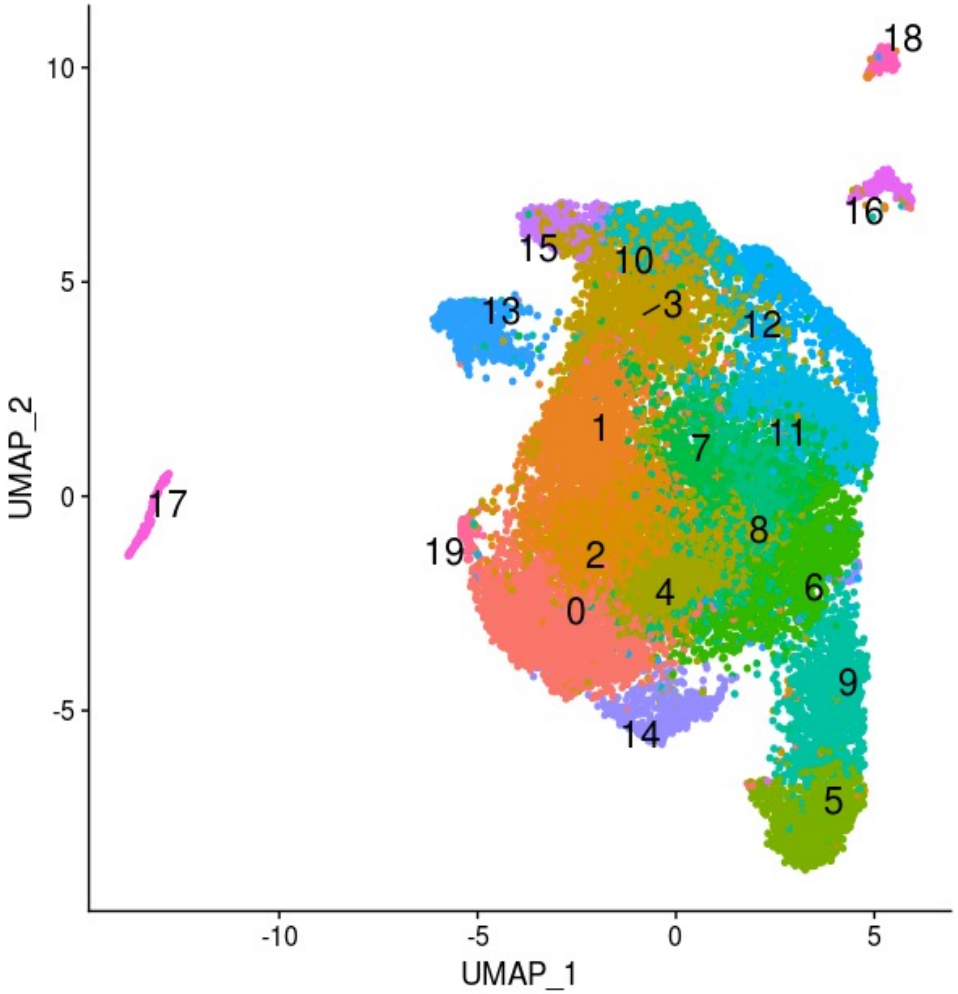


Büttner et al., *Nature Communications*. (2021).
Dann et al., *Nature Biotech*. (2022).

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

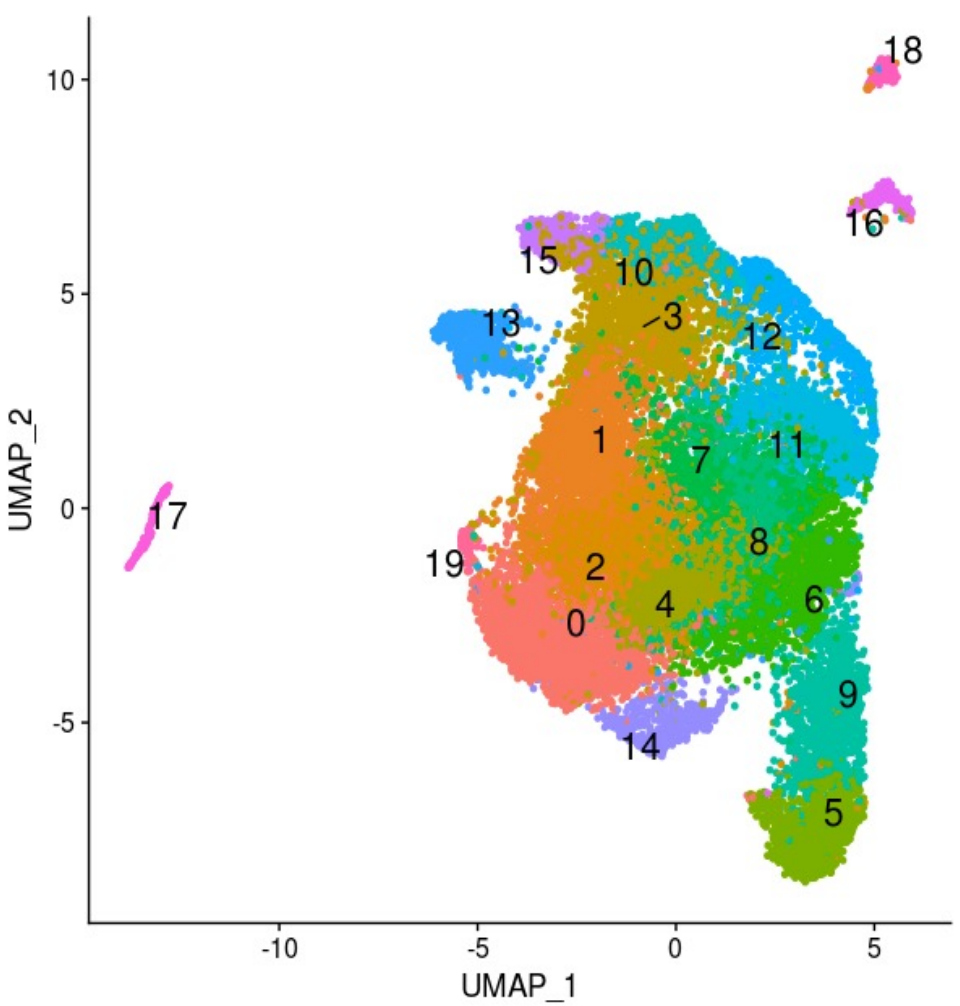
Resource

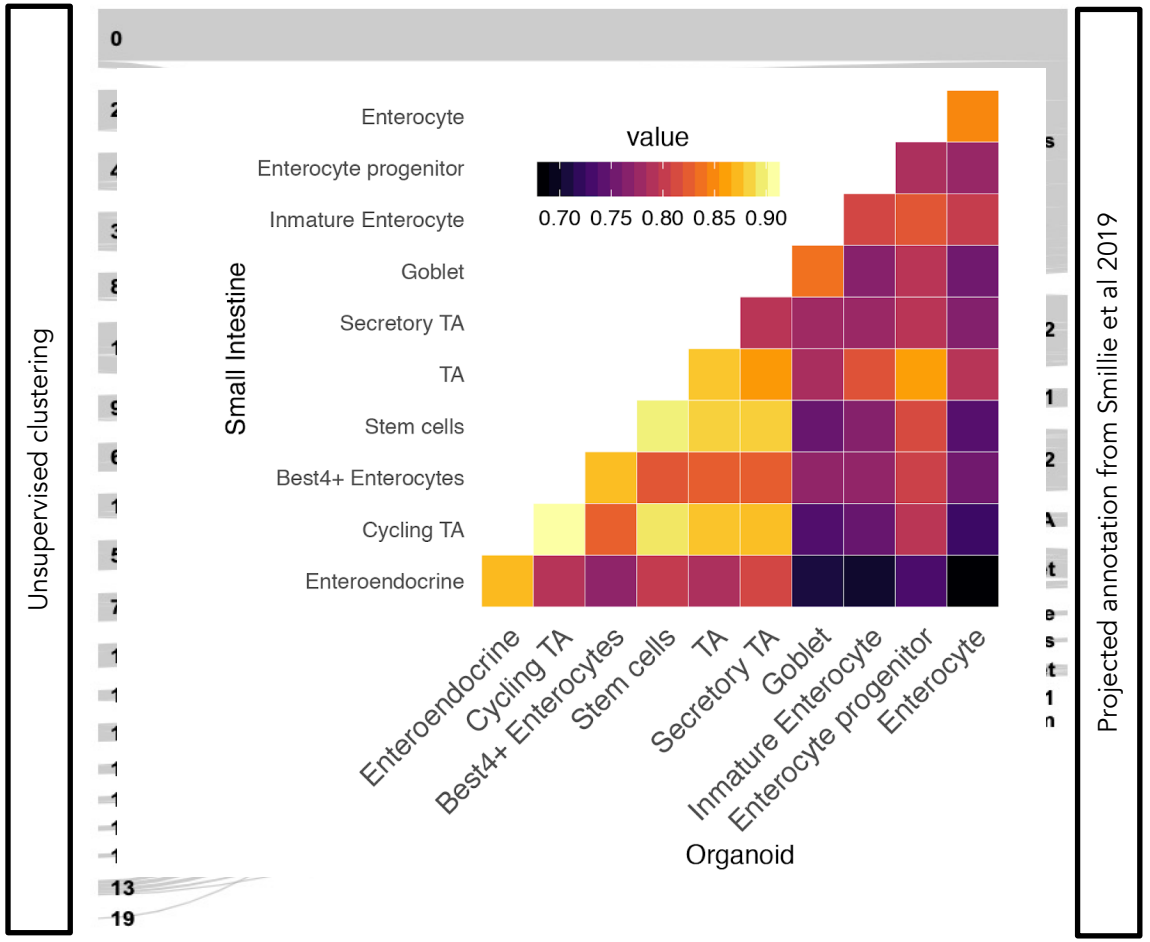
Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis

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

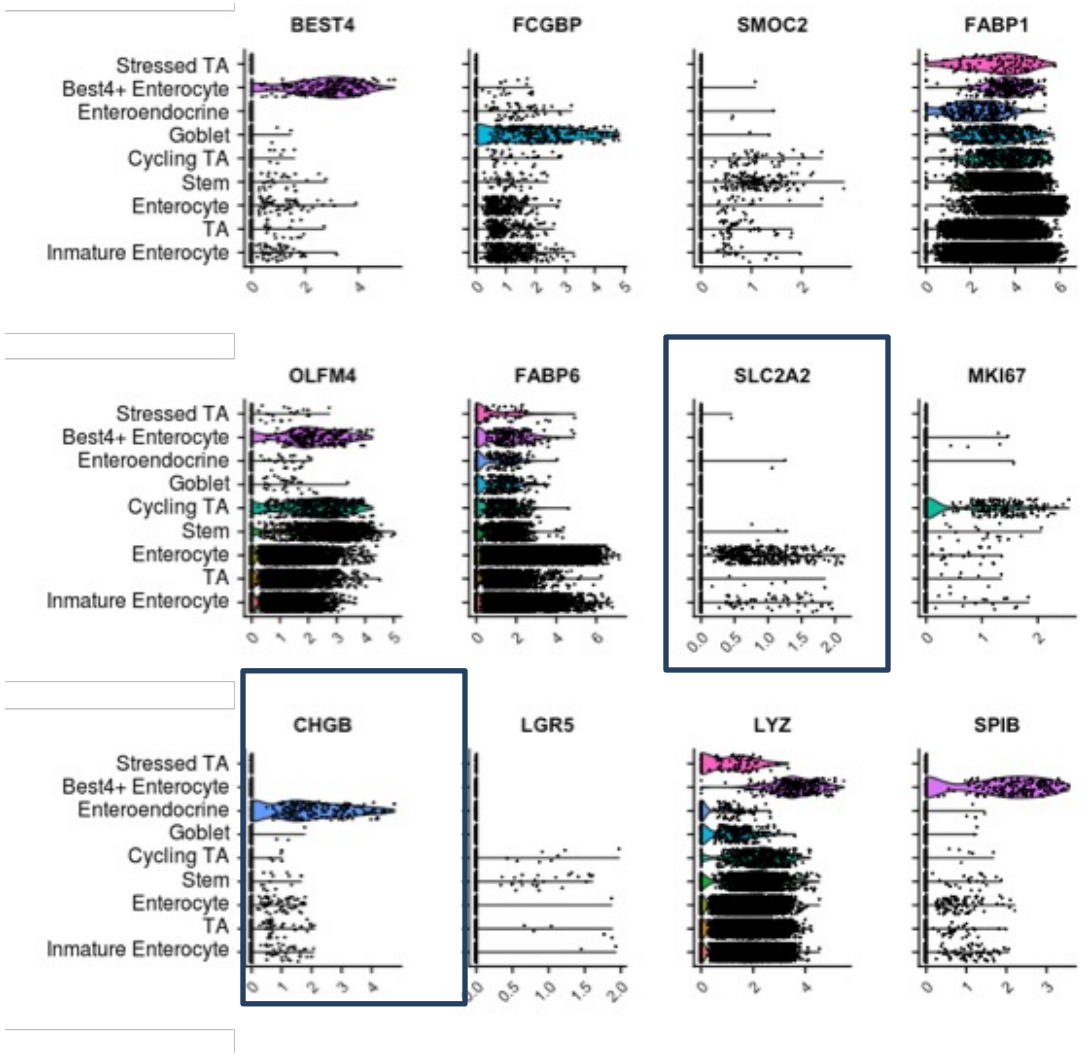


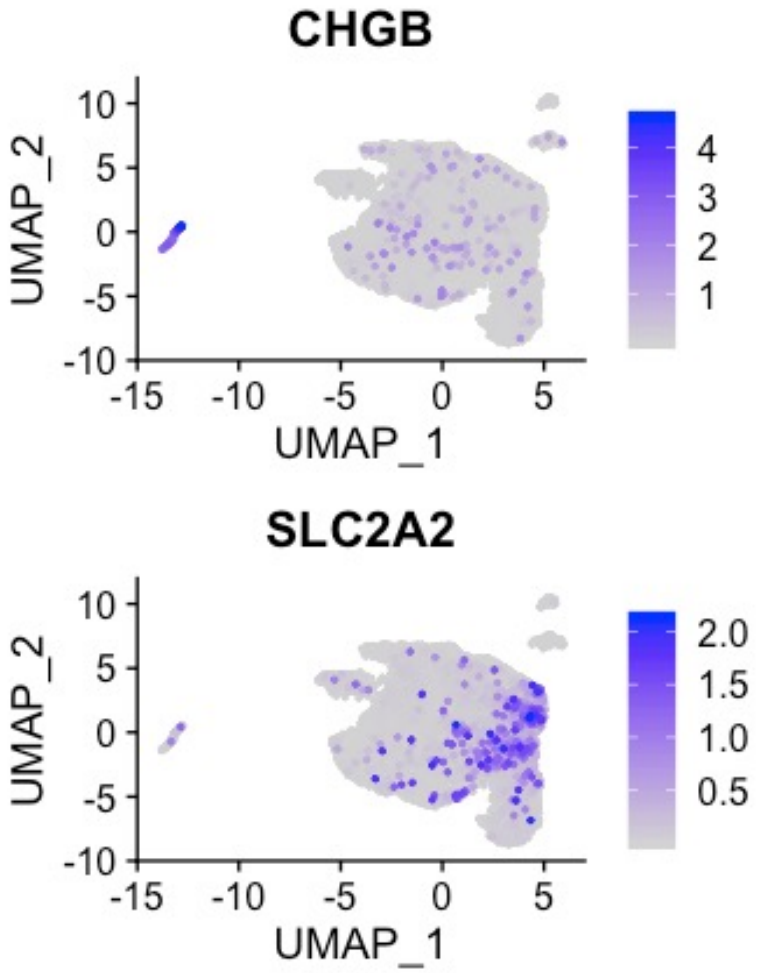


scMAP

B.

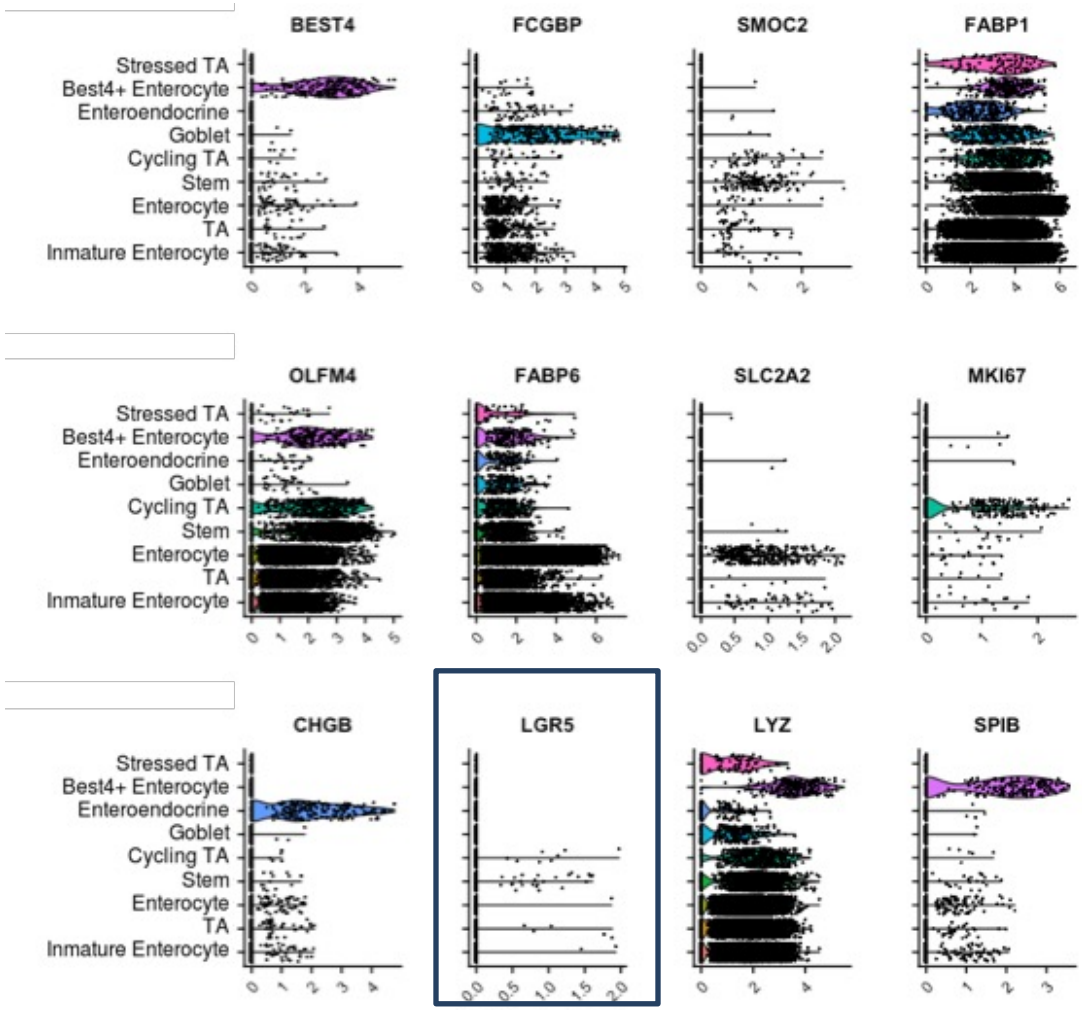
Markers for validation – cell type

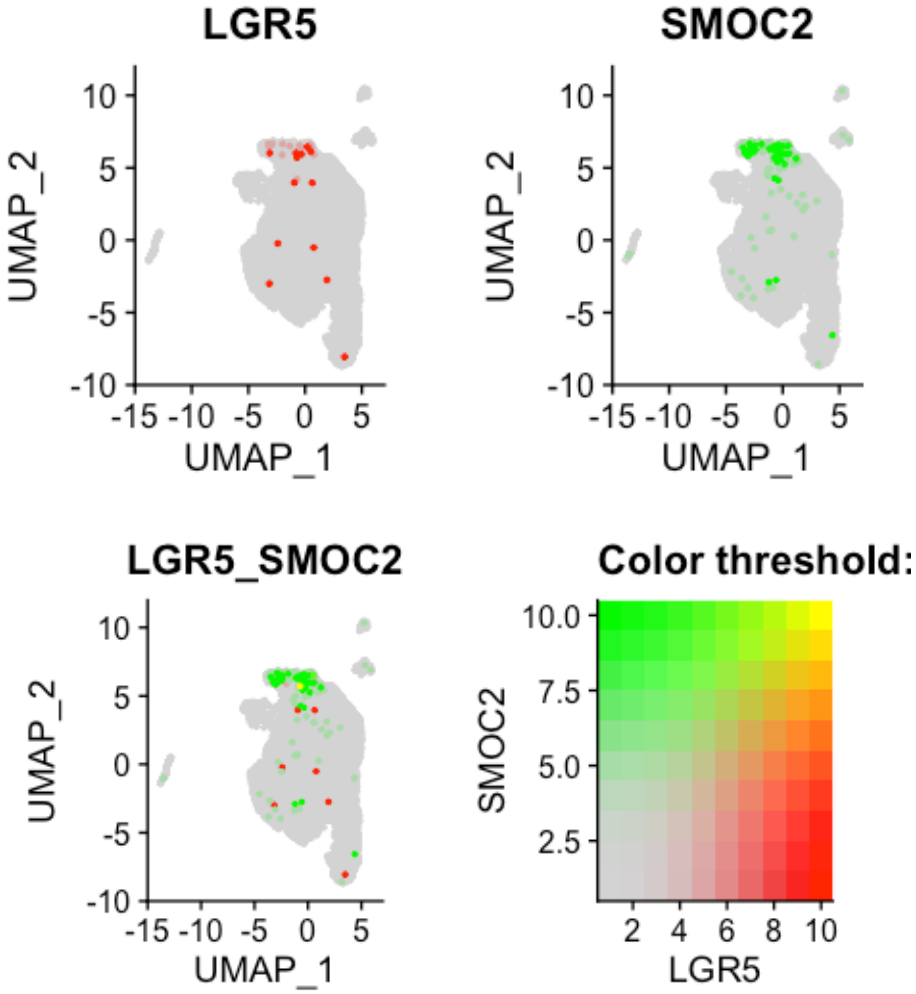




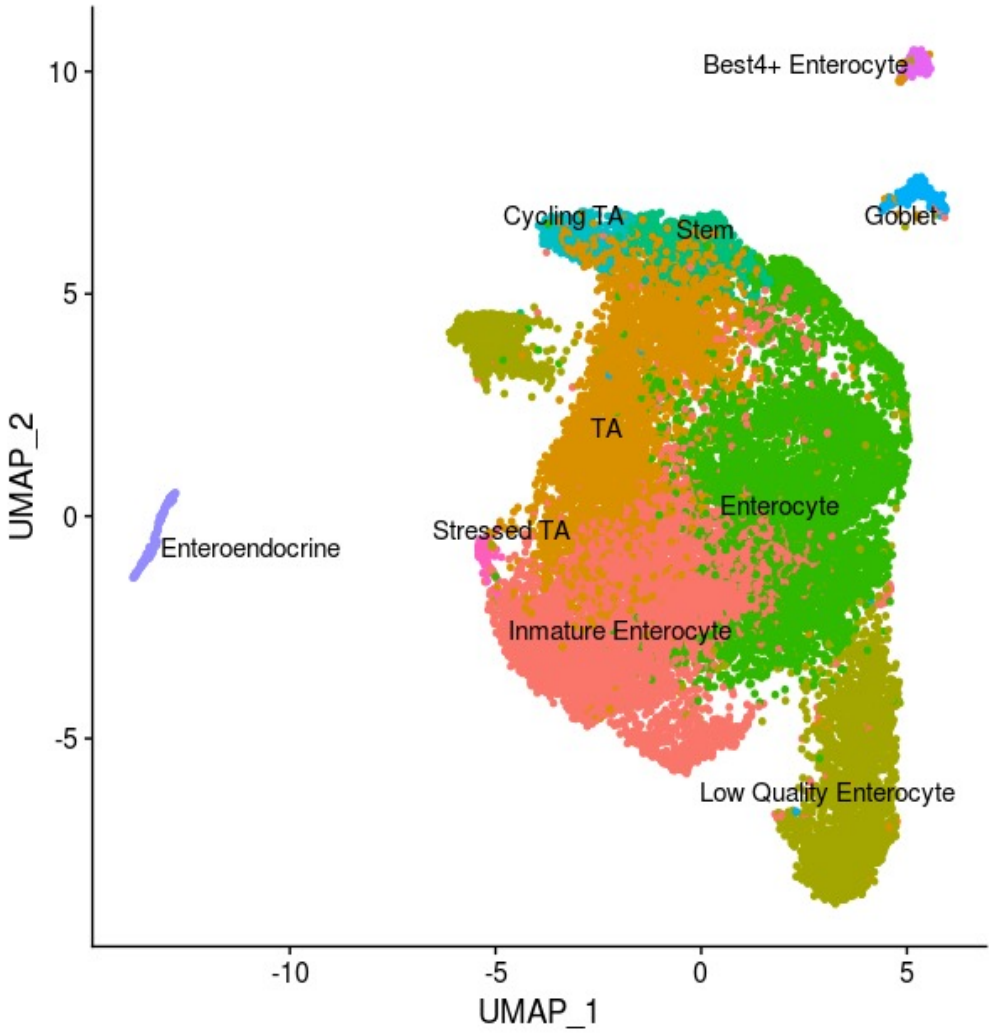
B.

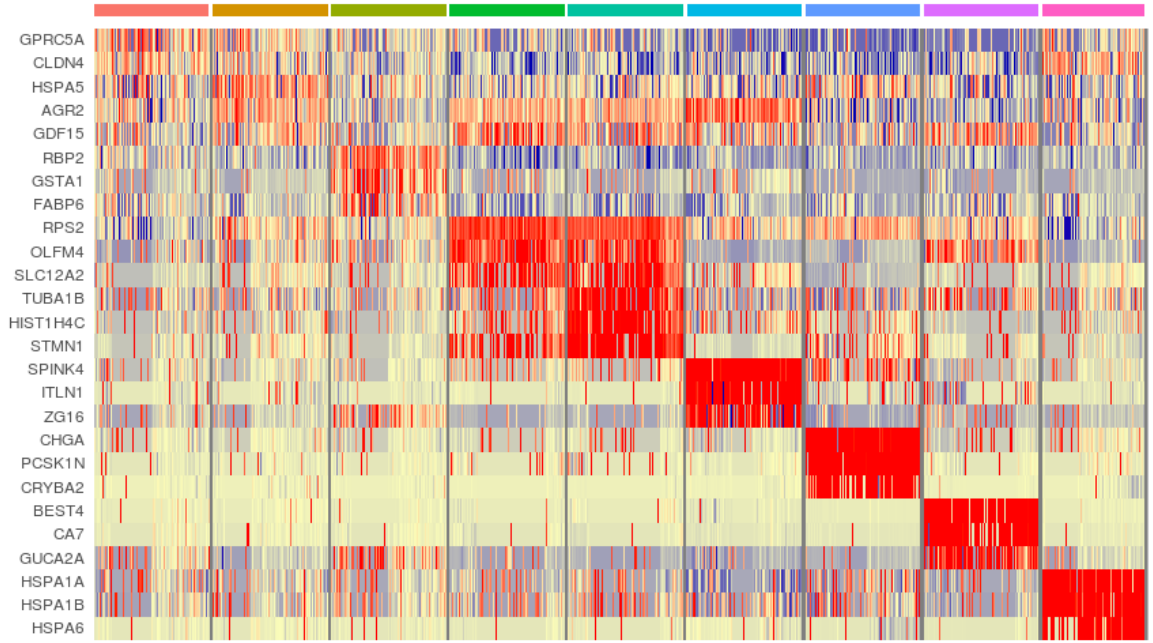
Markers for validation – cell type



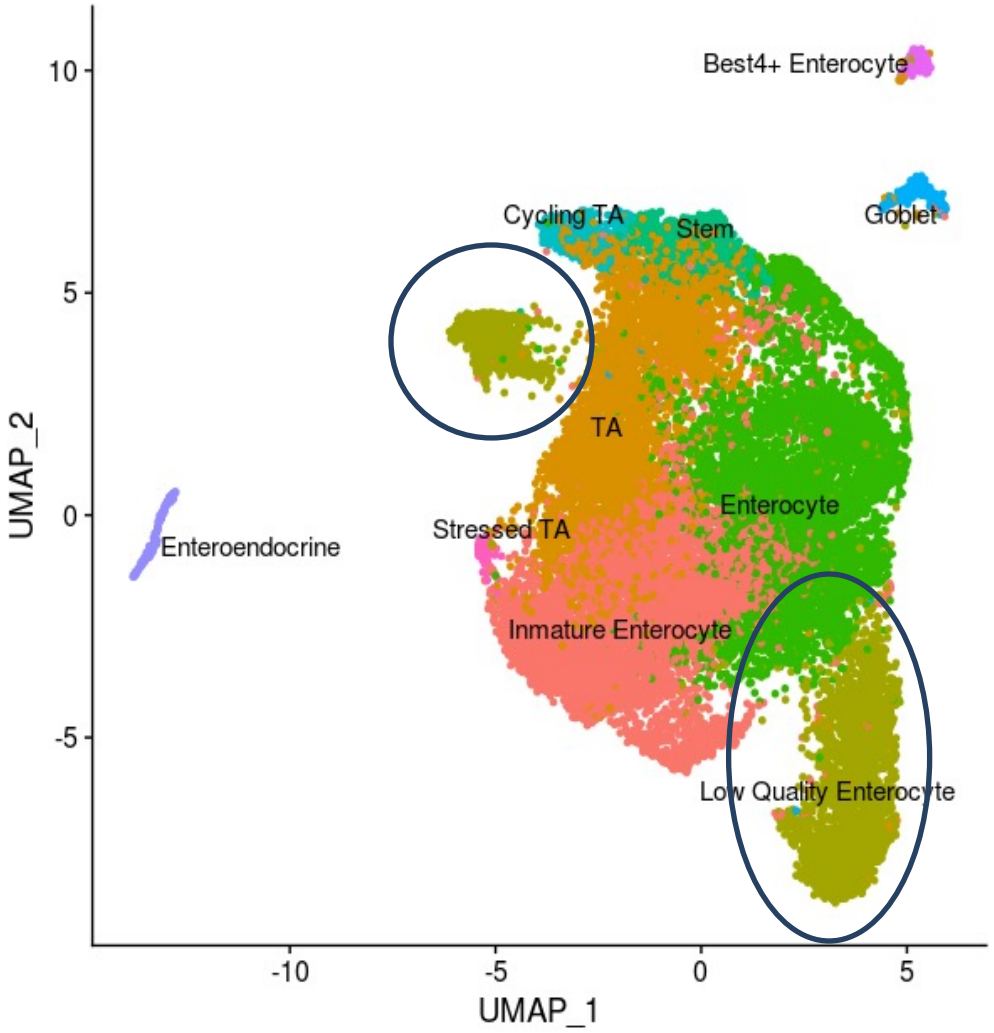


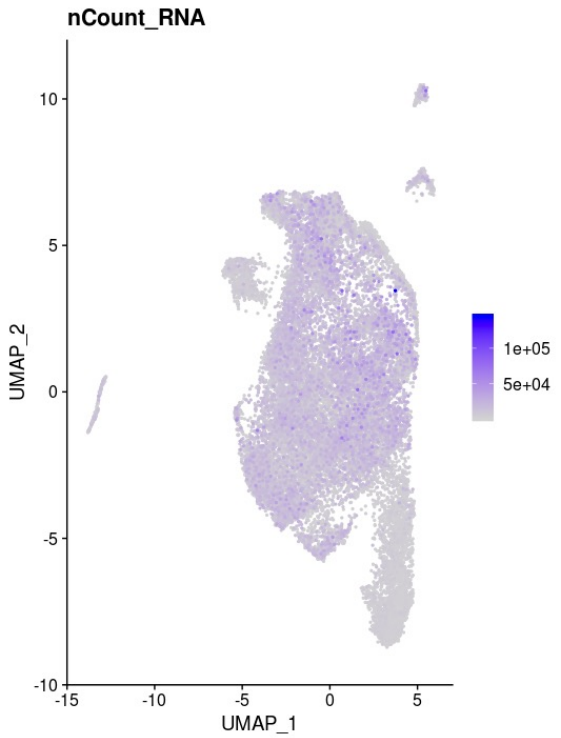
Cluster Annotation

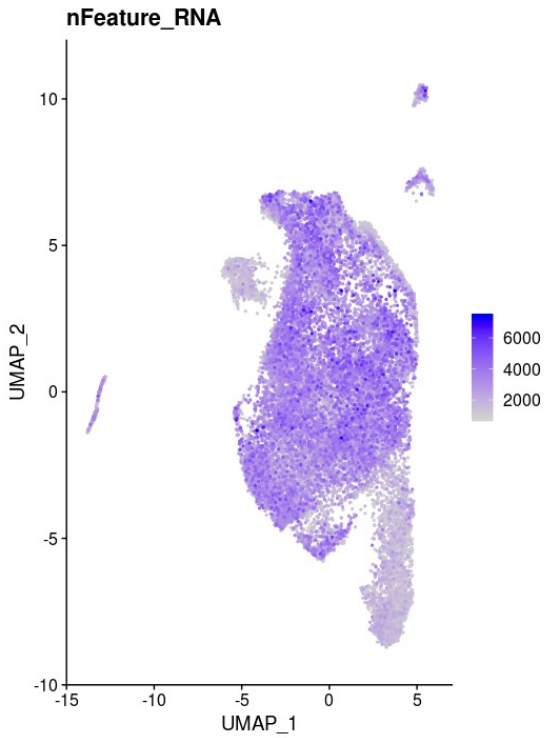




Cluster Annotation







Cluster Annotation