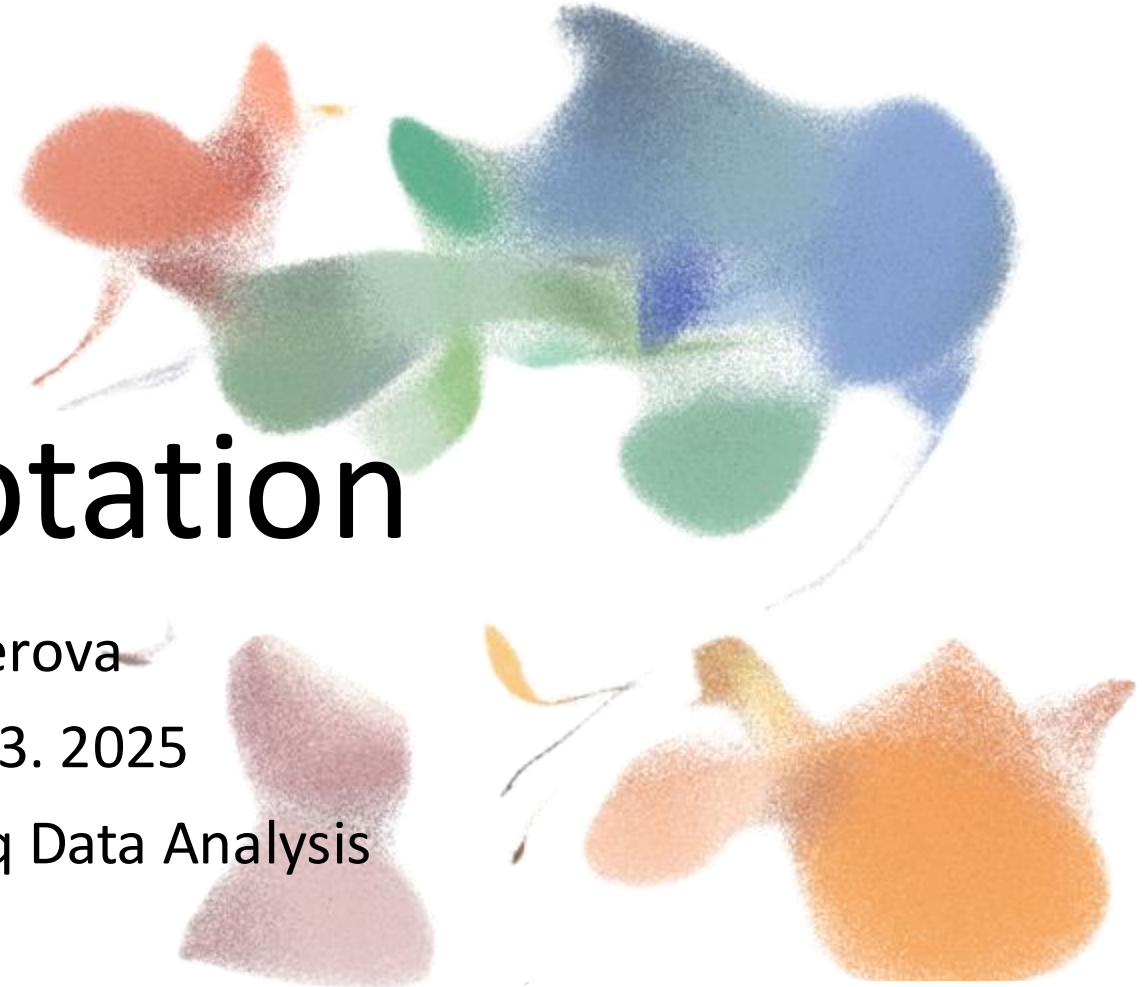


Cell annotation

Lucie Pfeiferova

December 1.-3. 2025

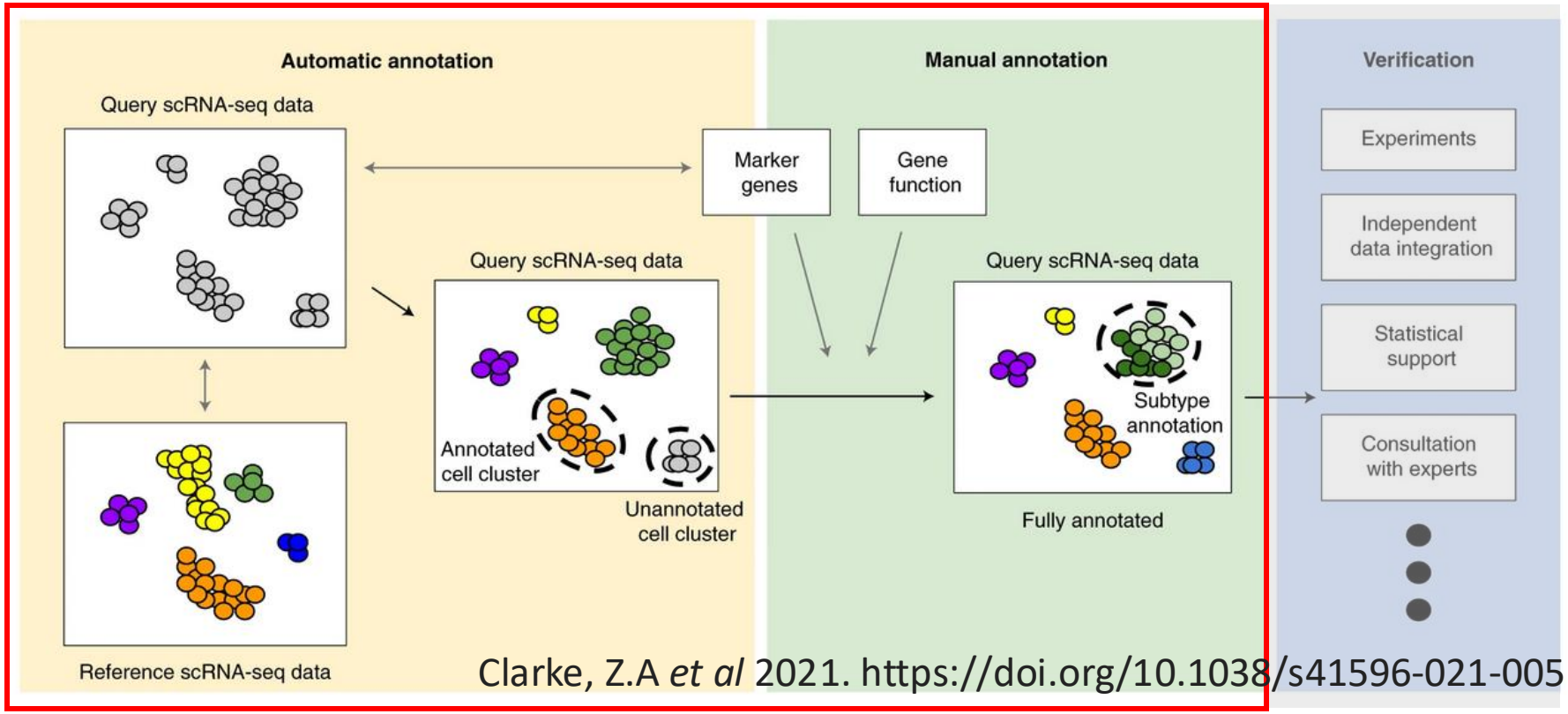
Course on scRNA-seq Data Analysis

Learning outcome

- Understand the challenges of cell annotation in scRNA-seq
- Understand different cell annotation approaches

Motivation

exploit prior information I) using marker genes for expected cell types, II) curated gene set associated with specific biological processes III) directly comparison of expression profiles between published reference datasets



Clarke, Z.A *et al* 2021. https://doi.org/10.1038/s41596-021-00534-0

Manual vs automatic annotation

Manual: using marker genes

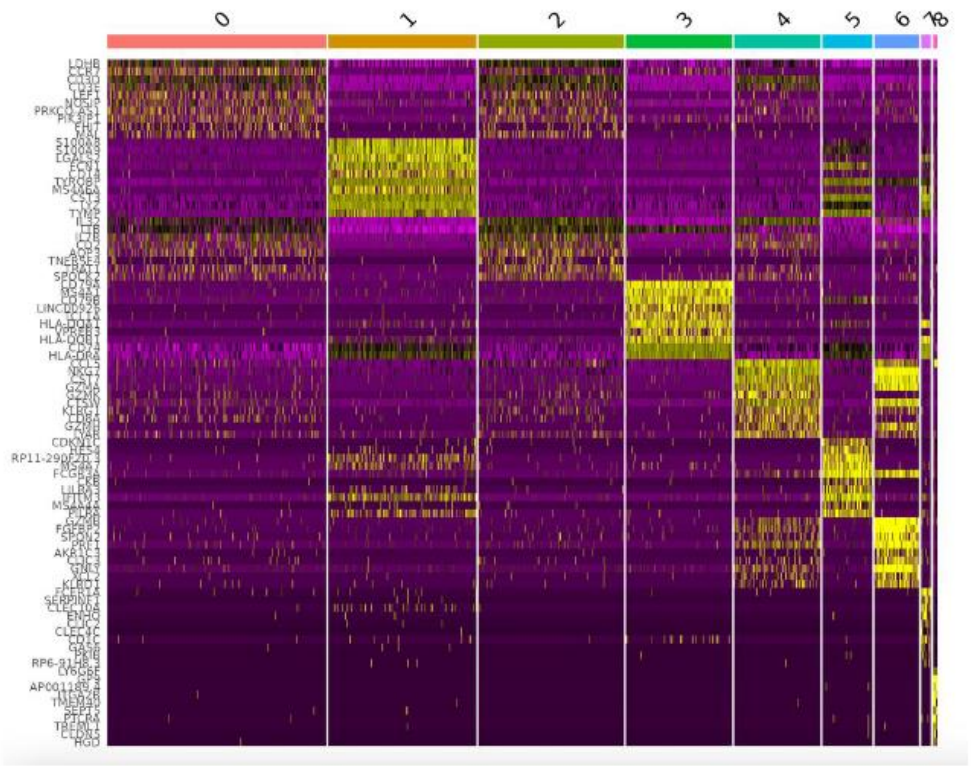
- Time consuming
- Requires expert (or at least prior knowledge
- Might be subjective

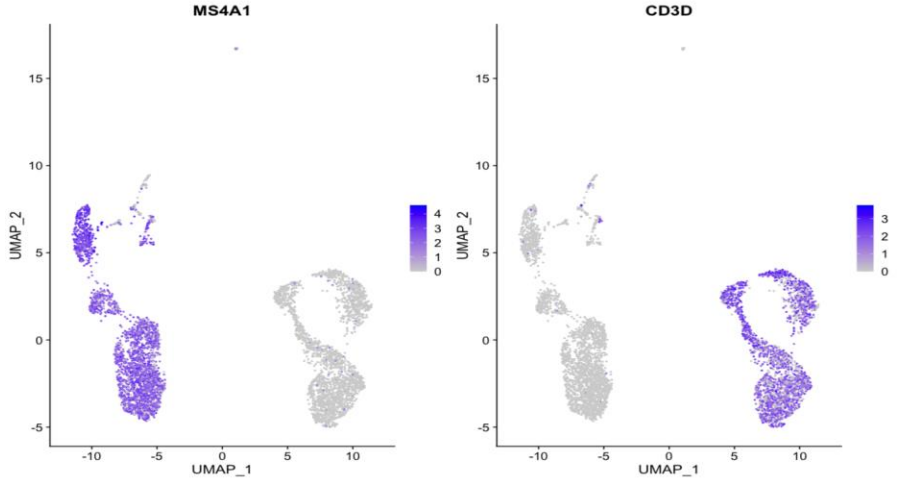
Automatic: requires a reference

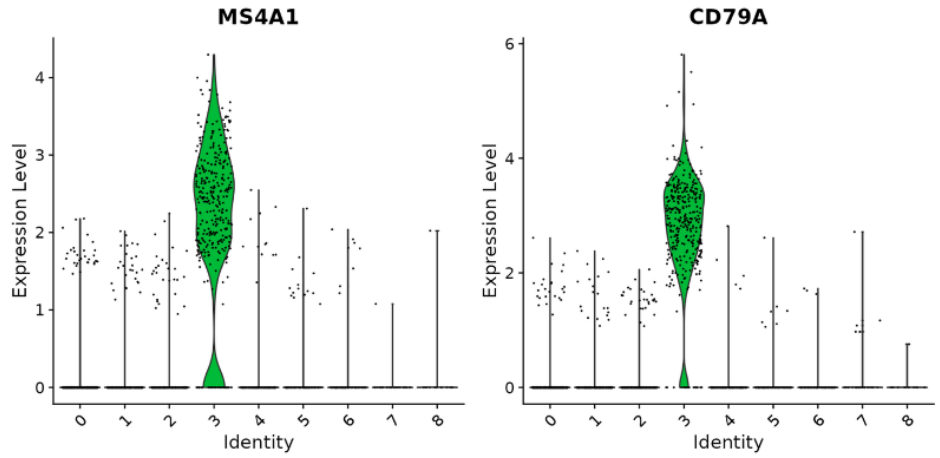
- Either use complete cell type-specific mRNA expression profiles based on bulk RNA-seq from FACS-sorted or use of a reference of manually curated cells picked from scRNA-seq data sets
- Can miss cell types if they are not included in the reference

Manual annotation – Gene markers

Finding differentially expressed features (cluster biomarkers)

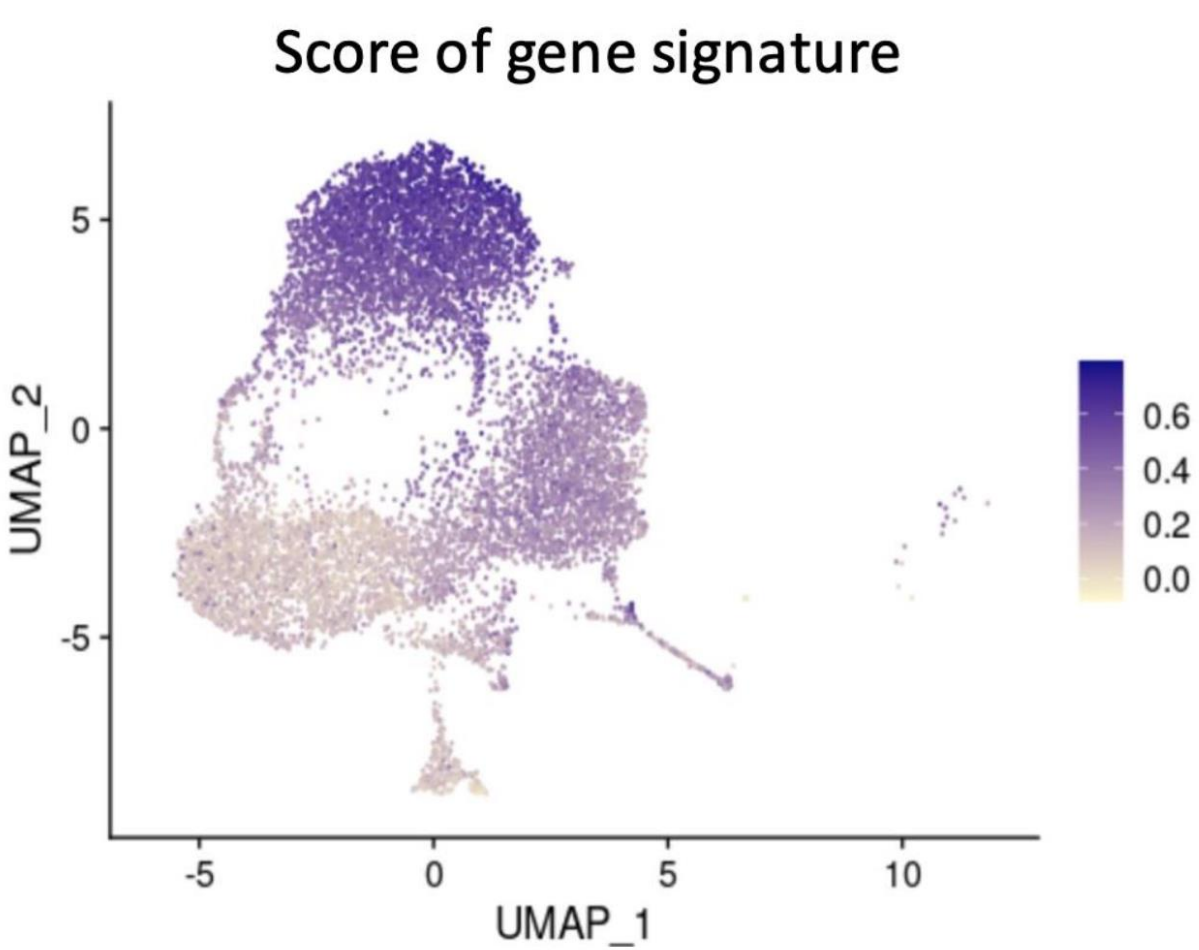






Seurat AddModuleScore

	orig.ident	nCount_RNA	nFeature_RNA	Tcell	Myeloid	NK	Plasma_cell
cHTNA1	Zilionis_immune	7516	2613	0	0.5227121	0	0.00000000
cHNVA2	Zilionis_immune	5684	1981	0	0.5112892	0	0.00000000
cALZN3	Zilionis_immune	4558	1867	0	0.3584502	0	0.07540874
cFWBP4	Zilionis_immune	2915	1308	0	0.1546426	0	0.00000000
							0.00000000
							0.00000000

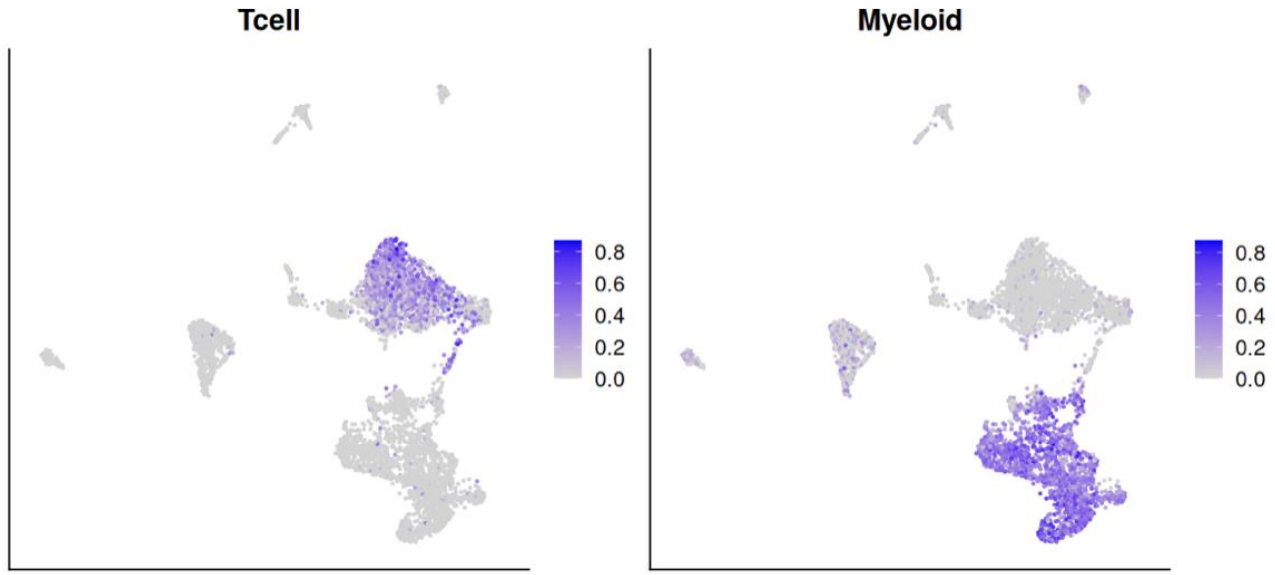


UCell (In Seurat AddModuleScore_Ucell)

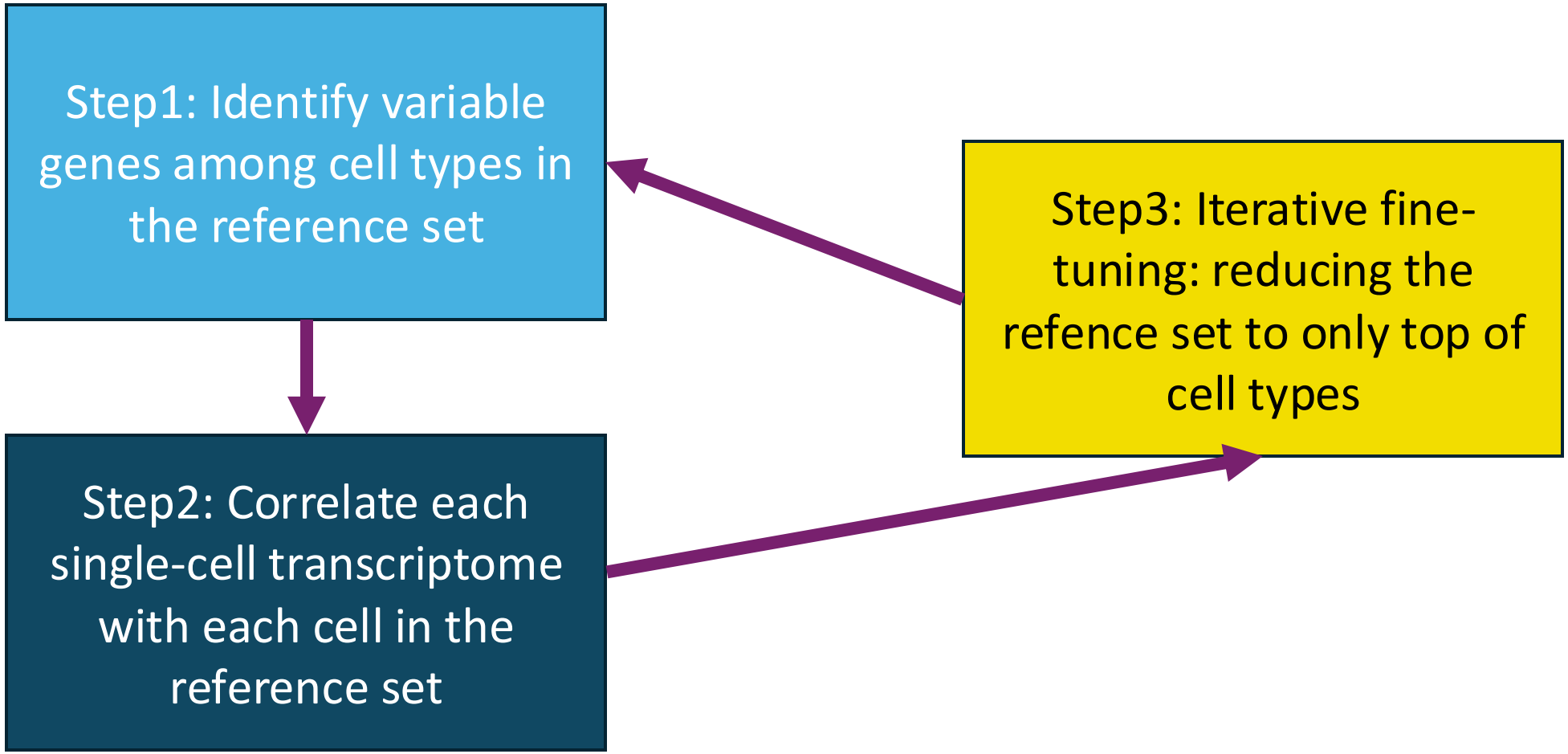
UCell supports positive and negative gene sets within a signature. Simply append + or - signs to the genes to include them in positive and negative sets, respectively. For example:

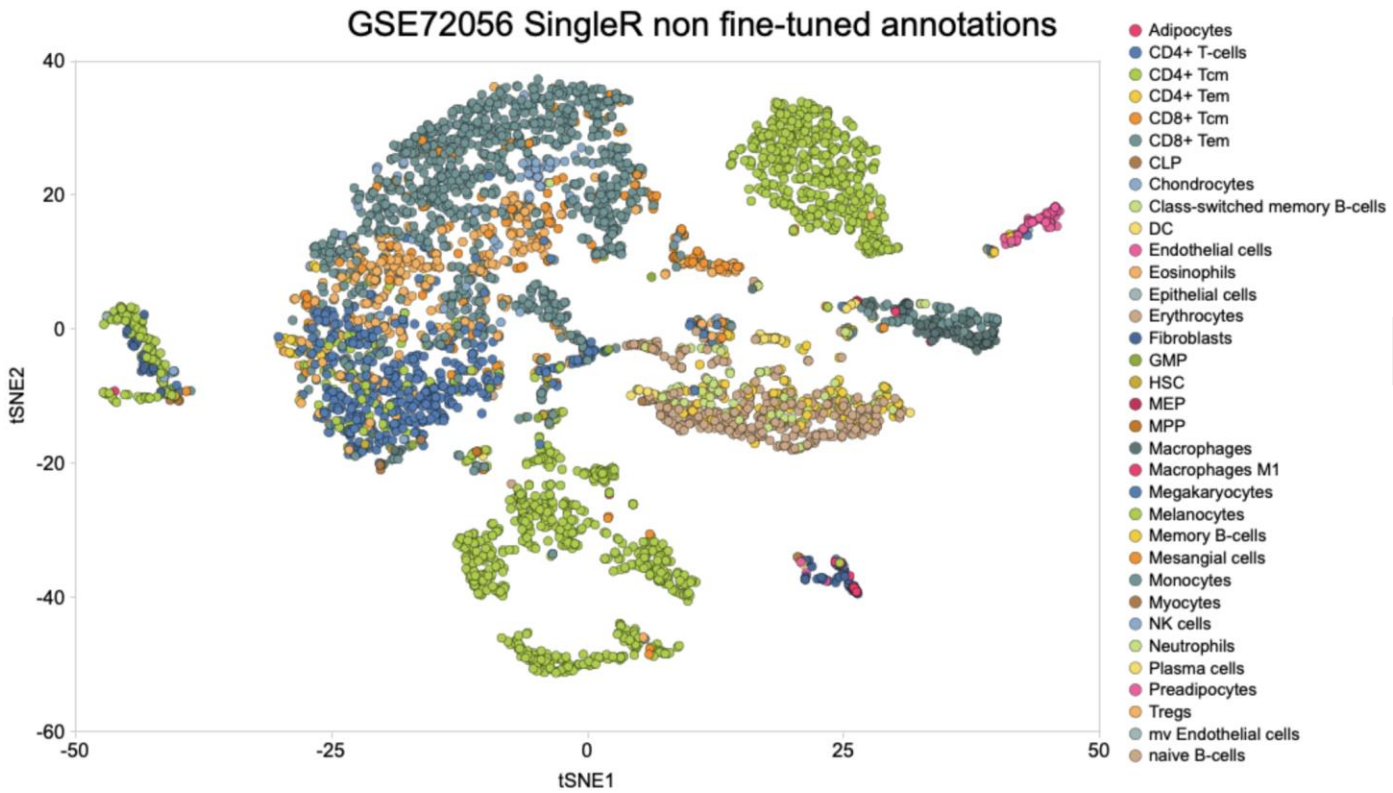
```
signatures <- list(
    CD8T = c("CD8A+","CD8B+","CD4-"),
    CD4 = c("TRAC+","CD4+","CD40LG+","CD8A-","CD8B-"),
    NK = c("KLRD1+","NCR1+","NKG7+","CD3D-","CD3E-")
)
```

```
##              CD8T        CD4       NK
## bcHTNA1  0.000000 0.14975523 0.000000
## bcHNVA2  0.000000 0.02503338 0.000000
## bcALZN3  0.000000 0.00000000 0.000000
## bcFWBP4  0.000000 0.00000000 0.000000
## bcBJYE5  0.000000 0.28627058 0.000000
## bcGSBJ6  0.000000 0.00000000 0.000000
## bcHQGJ7  0.000000 0.00000000 0.000000
## bcHKKM8  0.000000 0.21161549 0.000000
## bcIGQU9  0.000000 0.28649310 0.000000
```

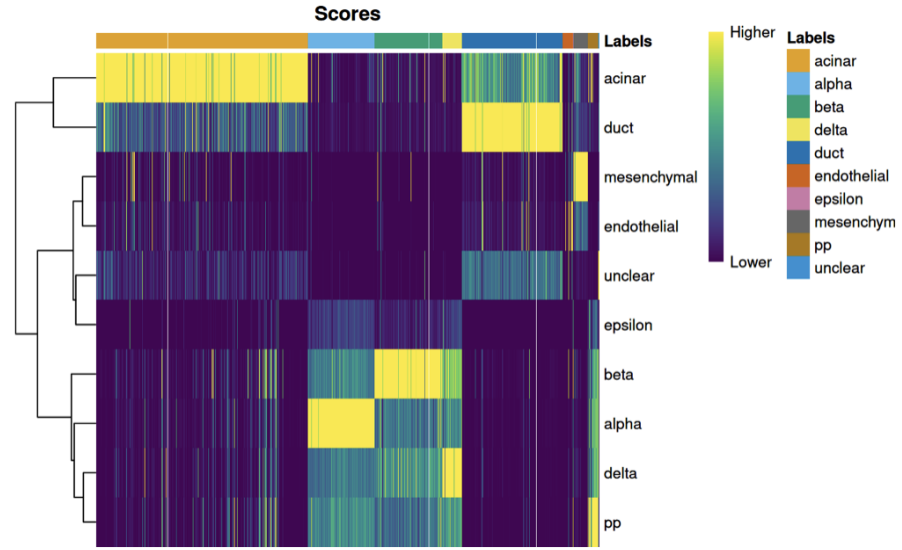


Automatic annotation - SingleR

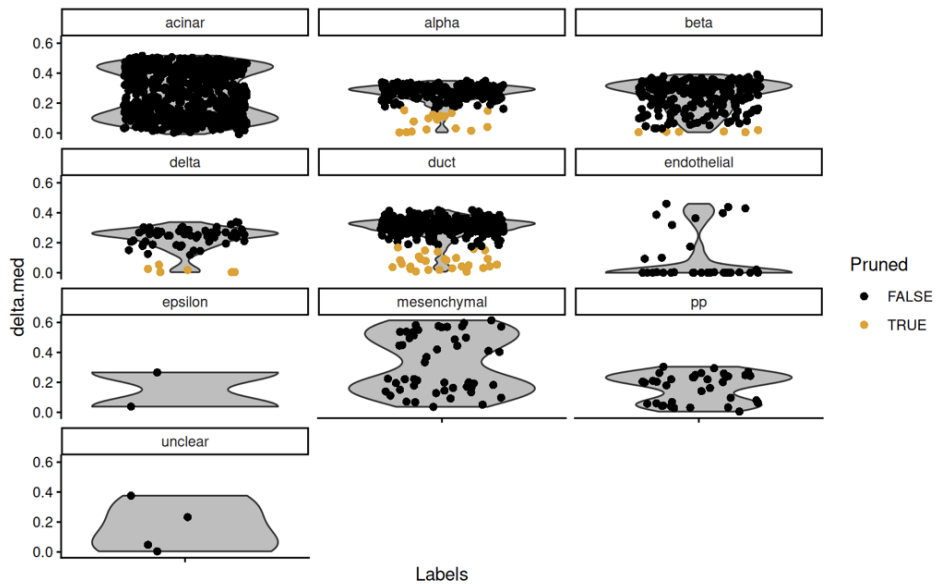





```
plotScoreHeatmap(pred.grun)
```




```
plotDeltaDistribution(pred.grun, ncol = 3)
```



AI in scRNA-seq annotation

- Traditional cell-type annotation in scRNA-seq depends heavily on expert knowledge or reference datasets that may be incomplete or biased
- AI-based foundation models (***might***) bring automation, scalability, and improved generalization across tissues, species, and experimental platforms.

scGPT

- Pretrained on millions of single cells across diverse tissues
- Learns contextual relationships between genes and cell states
- Performs cell-type annotation, batch correction, data integration, and perturbation prediction
- Works in a zero-shot or few-shot setting
- Reduces need for manual feature engineering

scFoundation

- Unified multimodal embedding space
- Supports cross-modality annotation and integration
- Provides standardized cell-type labels across datasets
- More robust to dataset size, batch effects, and modality noise
- Can transfer knowledge from one omic type to another

Bonus: Database with reference genes/sets

- Databases with cell type markers genes
- • PanglaoDB https://panglaodb.se/ (mouse and human)
- R: https://cran.r-project.org/web/packages/rPanglaoDB/index.html
- • CellMarker (mouse and human): http://bio-bigdata.hrbmu.edu.cn/CellMarker/
- • SingleR https://github.com/dviraran/SingleR (Aran et al.), access via celldex package,
- e.g. human primary cell atlas (microarrays)
- • Human Cell Atlas https://www.humancellatlas.org (Regev et al.) single cell RNA seq
- atlas, also some mouse data
- • Single cell portal: https://singlecell.broadinstitute.org/single_cell

- Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019). “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.” Nat. Immunol., 20, 163-172. doi:10.1038/s41590-018-0276-y.
- Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R (2023). “Dictionary learning for integrative, multimodal and scalable single-cell analysis.” Nature Biotechnology. doi:10.1038/s41587-023-01767-y.
- Clarke, Z.A., Andrews, T.S., Atif, J. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc 16, 2749–2764 (2021). https://doi.org/10.1038/s41596-021-00534-0
- Andreatta M, Carmona SJ. UCell: Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796-3798. Published 2021 Jun 30. doi:10.1016/j.csbj.2021.06.043
- https://sib-swiss.github.io/single-cell-r-training/