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
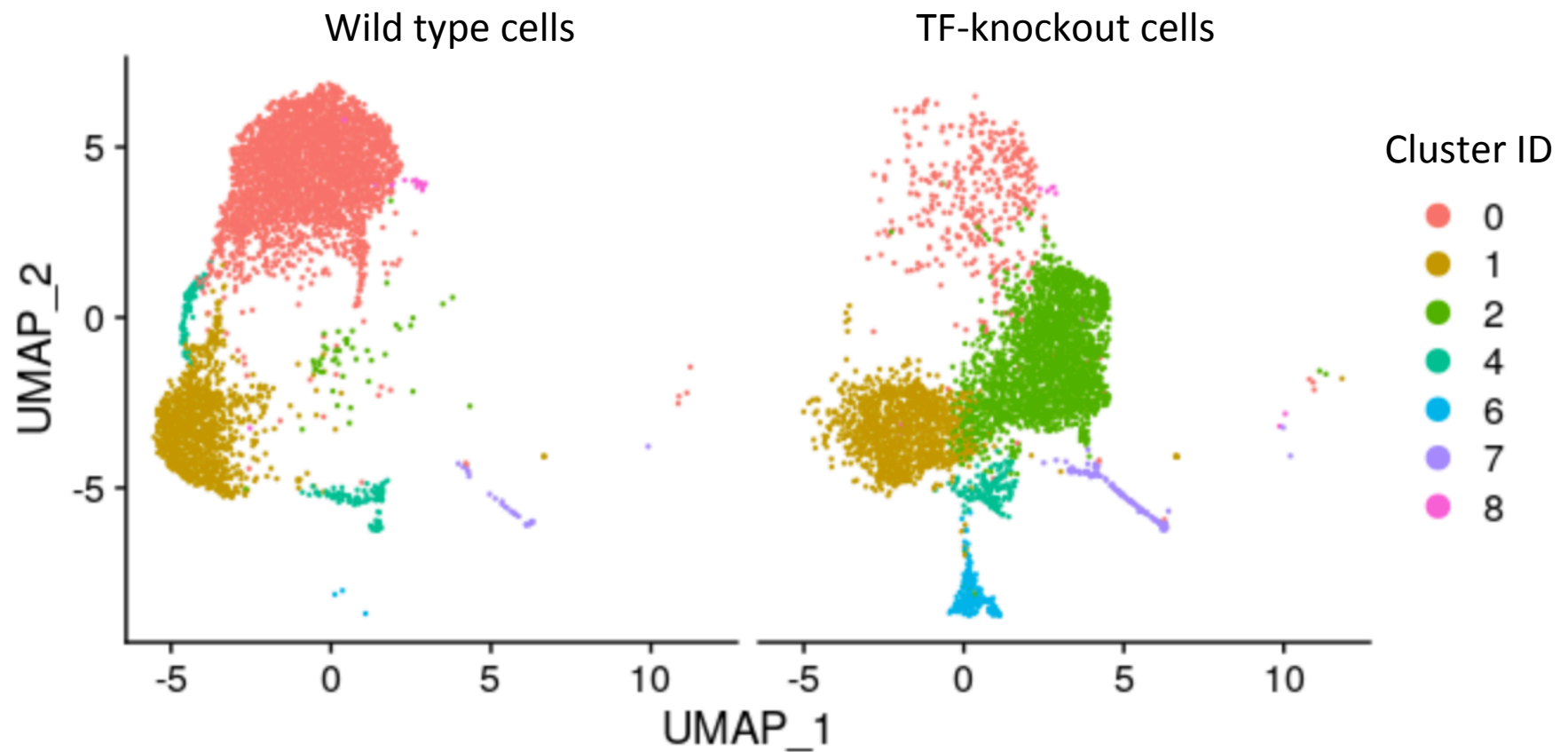 - migration state
  - differentiation: cell state

Wild-type *C. elegans* hermaphrodite stained to highlight the nuclei of all cells
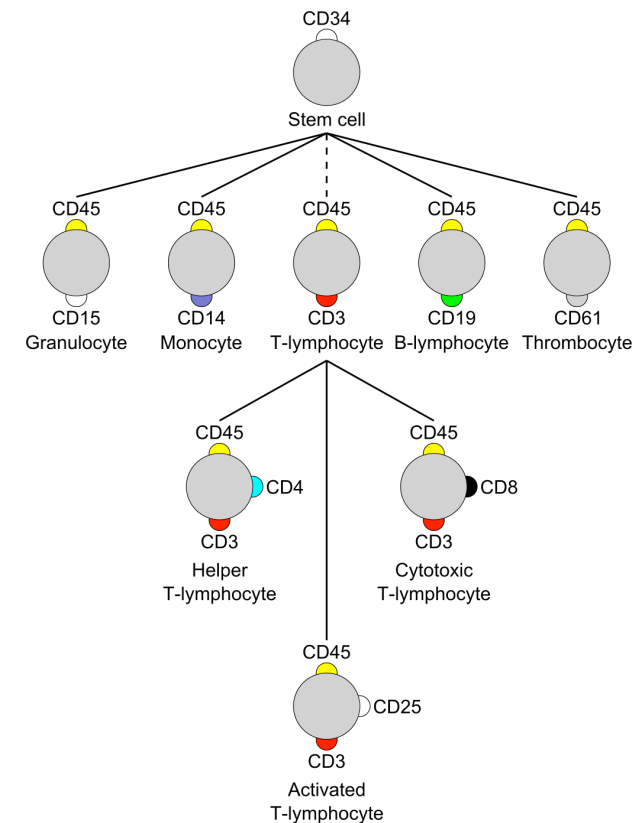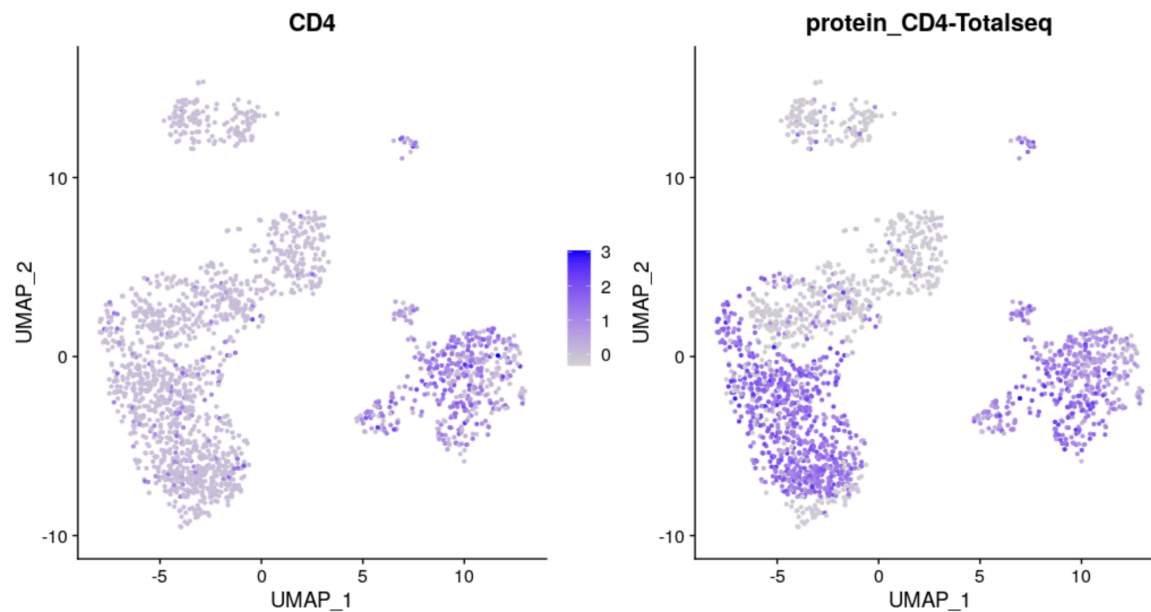
Why should we identify cell types?

- 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?

Wild type cells

TF-knockout cells
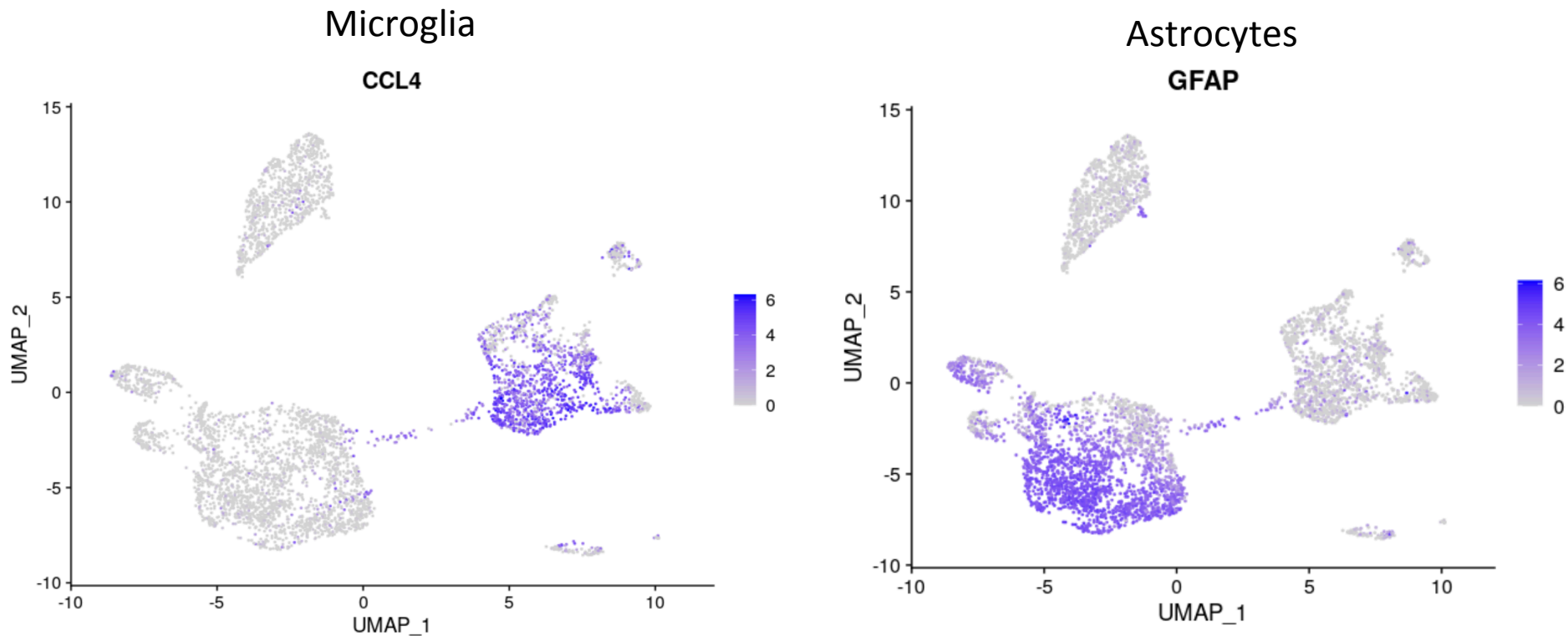
Cluster ID

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)

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
- Methods:
  - Assign a cell type per individual cell or per cluster of cells (better per cell)
  - Assignment of cell type via correlation of each cell/cluster to the "reference"
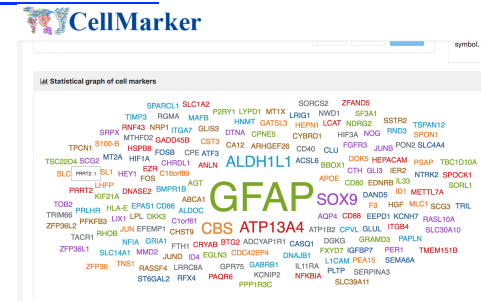
Manual annotation using known marker genes

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

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)
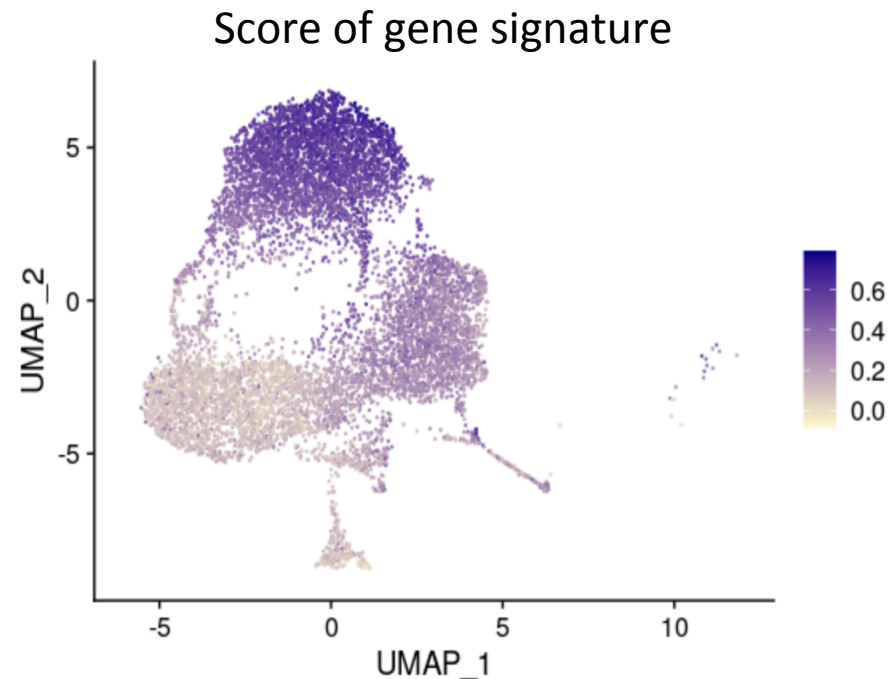 http://bio-bigdata.hrbmu.edu.cn/CellMarker/

- SingleR https://github.com/dviraran/SingleR (Aran et al. 2019), access via celldex package, eg 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
  (source of data to play with)

Module score

Tirosh et al 2016, Science 352:6282

Compare expression level of genes
belonging to the signature to "control"
genes with similar expression level to
signature genes

Score of gene signature


```
Seurat::AddModuleScore(object, features=list() )
```

UCell: Robust and scalable single-cell gene signature scoring

`UCell` is an R package for scoring gene signatures in single-cell datasets. UCell scores, based on the Mann-Whitney U statistic, are robust to dataset size and heterogeneity, and their calculation demands relatively less computing time and memory than other robust methods, enabling the processing of large datasets (>10^5 cells). UCell can be applied to any cell vs. gene data matrix, and includes functions to directly interact with Seurat and Bioconductor's SingleCellExperiment objects.

https://github.com/carmonalab/UCell
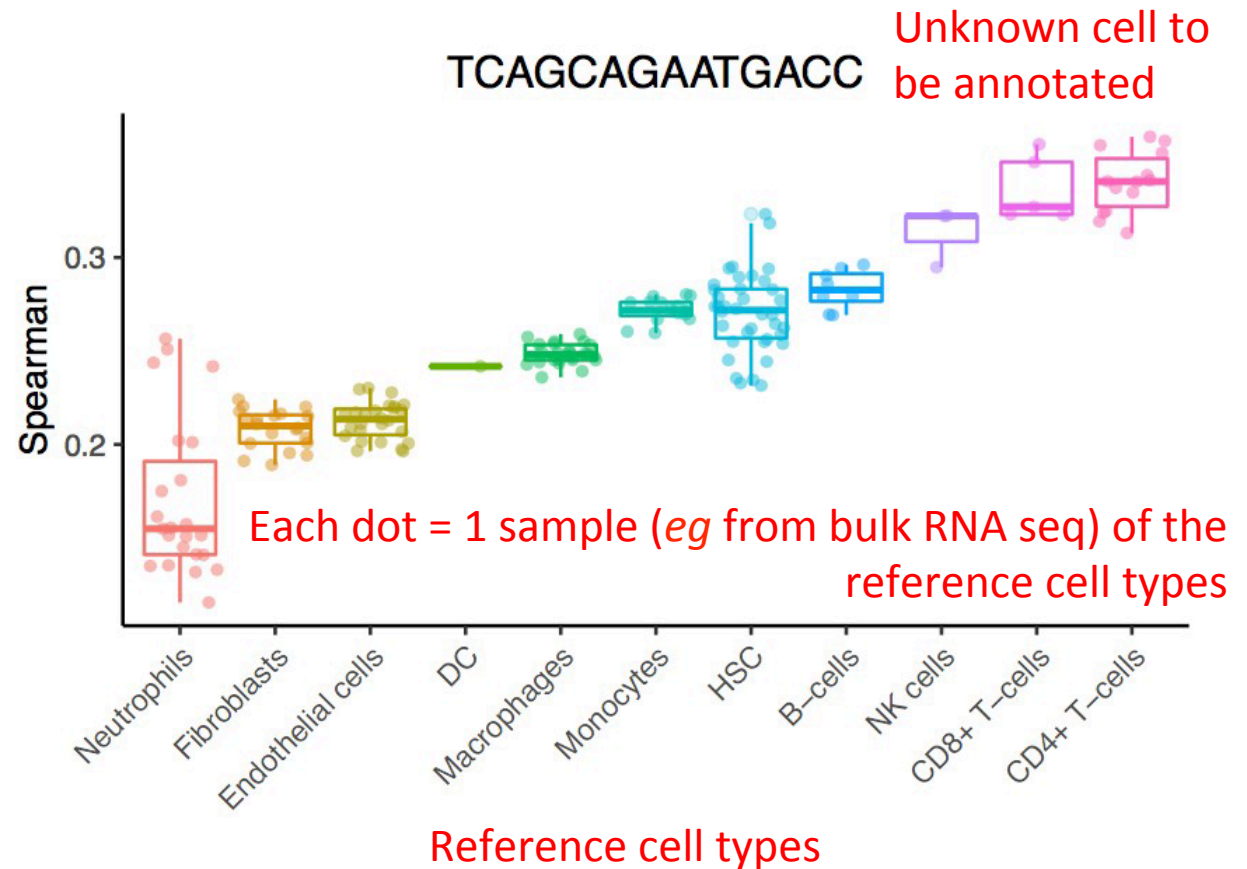
SingleR

SingleR

Easy access to rich reference data:

    - HPCA: manually-annotated Human Primary Cell Atlas
    37 main types, 157 subtypes, 713 samples
    - BluePrint +ENCODE
    24 main types, 43 subtypes, 259 bulk RNAseq
    samples
- Mouse: ImmGen and 'mouse.rnaseq' (brain-specific)

Classifies cells to both main types and subtypes, performs both single cell-wise and cluster-wise annotation

SingleR

Correlate each cell (or average of cluster) to each reference cell type

TCAGCAGAATGACC

Unknown cell to be annotated

Each dot = 1 sample (*eg* from bulk RNA seq) of the reference cell types
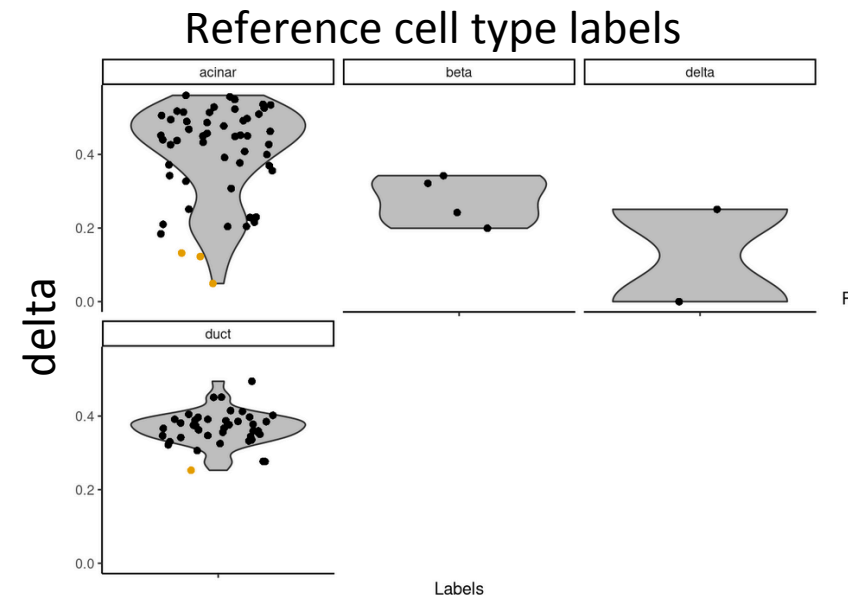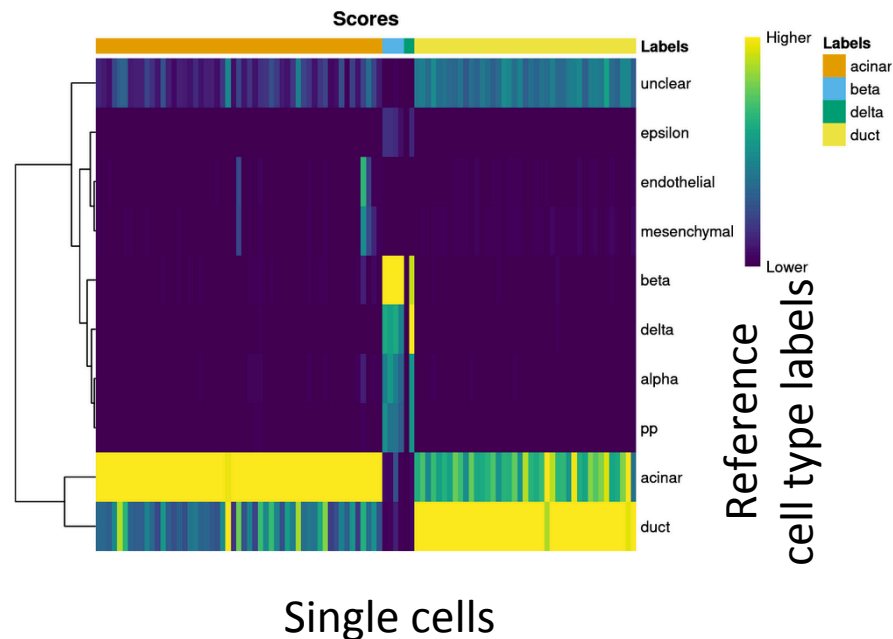
Reference cell types

Per reference cell type, correlate to the topN genes that have median(expression) > median(expression) in all other cell types

Fine Tuning step...

SingleR − annotation diagnostics
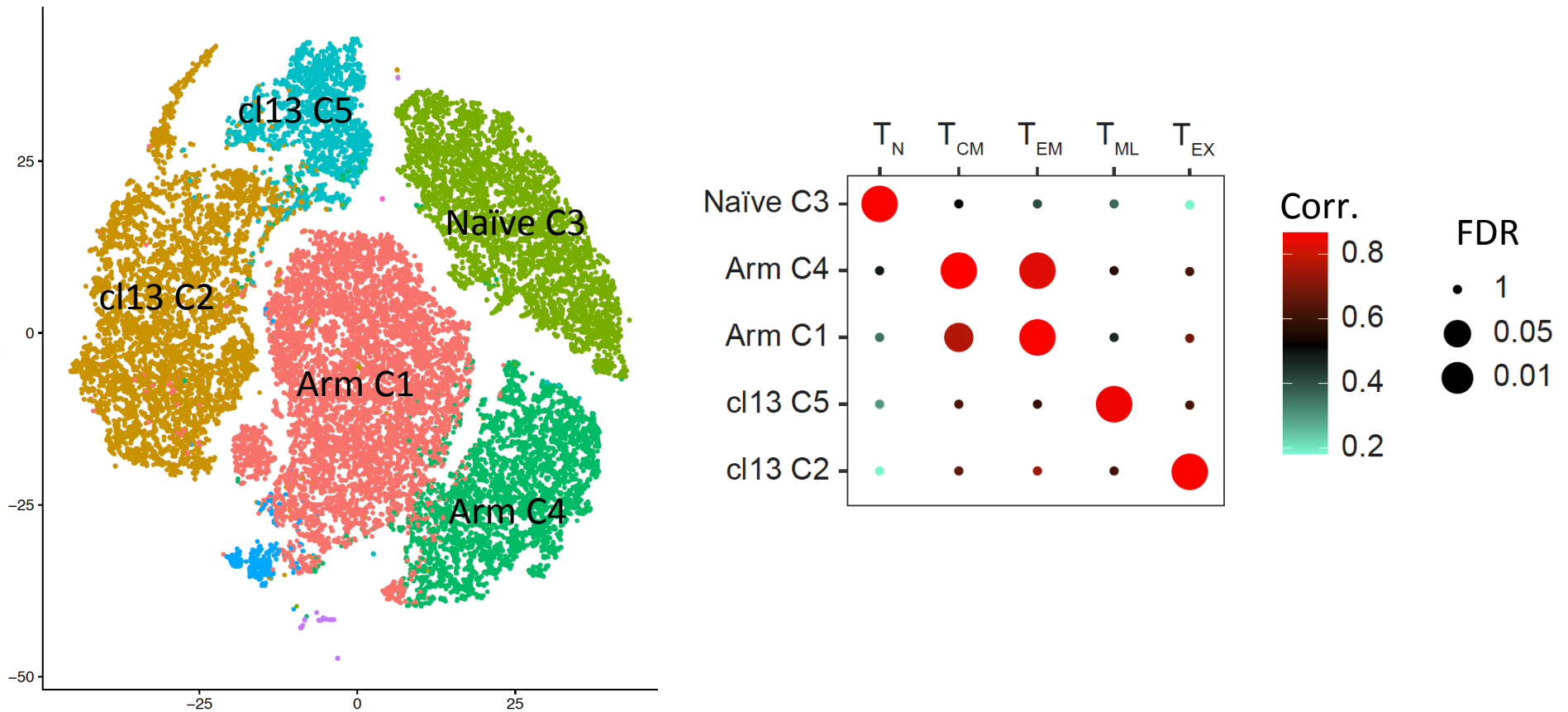
Heatmap of scores per cell for each label:

Single cells

Reference cell type labels

plotScoreHeatmap(singleR.object)

plotDeltaDistribution(singleR.object)

Try them during the practical exercise

https://bioconductor.org/packages/release/bioc/vignettes/SingleR/inst/doc/SingleR.html#4_Annotation_diagnostics

SingleR can also be used to evaluate similarity to a custom reference

Arm and cl13 = 2 different strains of LCM virus
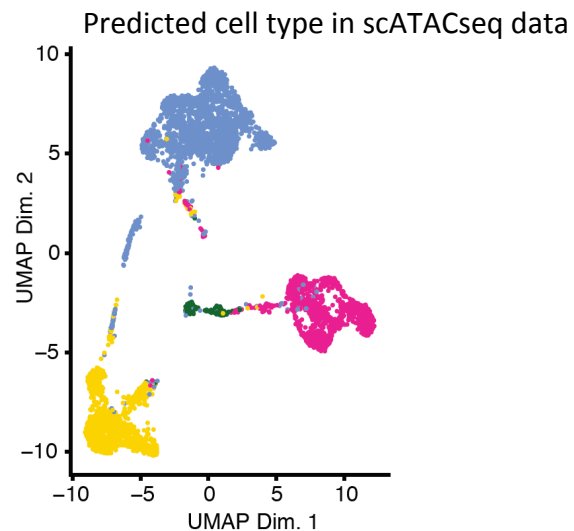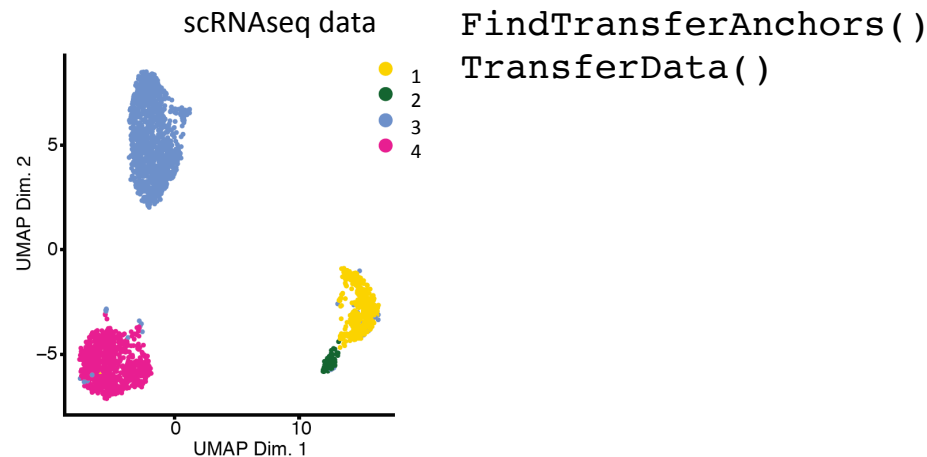
Several methods are available

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$_{rejection}$	0.19.2	Python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	Python	kNN ($k = 9$)	No	No	[29]

Abdelaal et al 2019 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z

Seurat, Azimuth (R. Satija lab)

Cell type label transfer

scRNAseq data


```
FindTransferAnchors()
TransferData()
```

Predicted cell type in scATACseq data

No rejection

https://satijalab.org/seurat/articles/integration_mapping.html

Reference-based mapping

11 human references (Sept 2022)

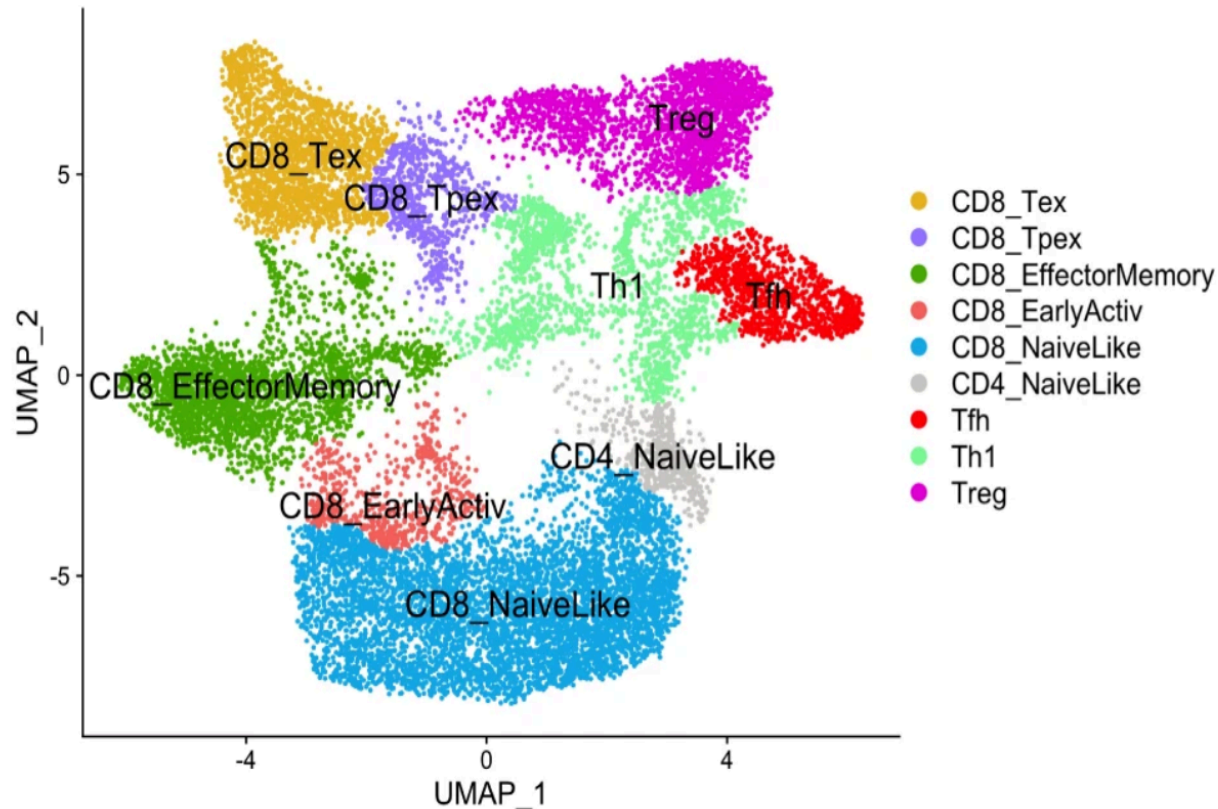https://azimuth.hubmapconsortium.org/

R vignette to run locally:
https://satijalab.github.io/azimuth/articles/run_azimuth_tutorial.html

```
bm <- RunAzimuth(query = "x",
reference = "bonemarrowref")
```

Combine published scRNAseq datasets to create an atlas that can be used as a reference

ProjecTILs, an algorithm for reference atlas projection
Andreatta et al 2021 Nat. Comm. https://www.nature.com/articles/s41467-021-23324-4

https://github.com/carmonalab/ProjecTILs

Additional links

https://bioconductor.org/books/release/OSCA/cell-type-annotation.html#assigning-cell-labels-from-reference-data

Dealing with multimodal single cell data:

    Argelaguet et al 2021

https://www.nature.com/articles/s41587-021-00895-7

Review on automated cell annotation, Pasquini et al 2021

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

Question on cell type annotation