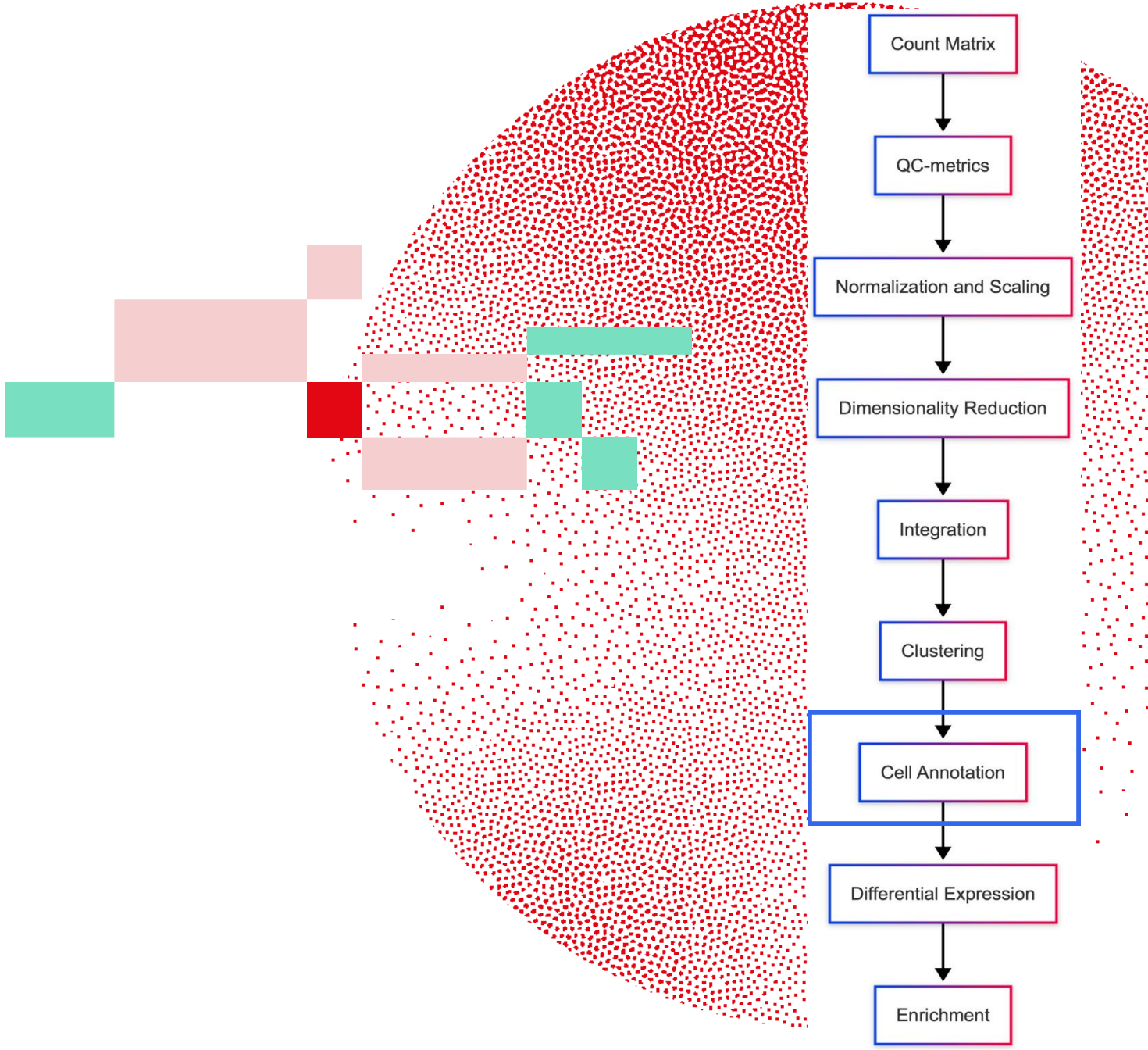
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Cell Annotation



Joana Carlevaro Fita

July 2-4, 2025

Adapted from previous year courses

elixir SWITZERLAND

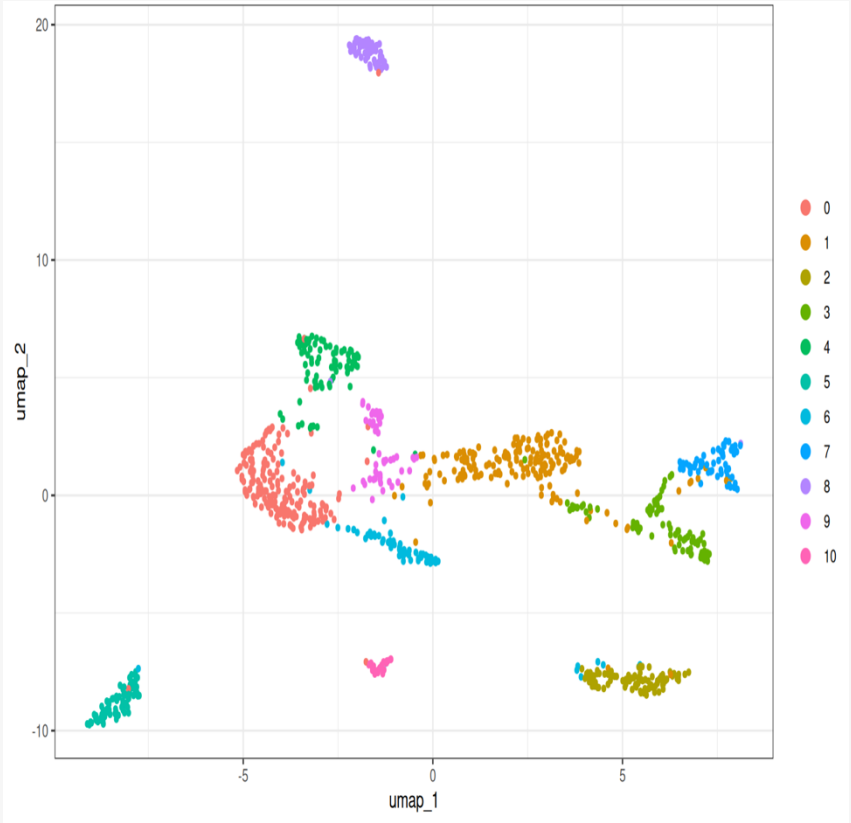
Learning objectives

Understand the relevance and challenges of cell annotation

Get an introduction to different cell annotation approaches and applications

Perform cell annotation using singleR package

Cell Annotation



Example of research questions

Identify cell heterogeneity

(different cell abundances)

Identify new cell type subsets

(by function, molecular signature..)

Compare conditions

(ie. tumor vs normal)

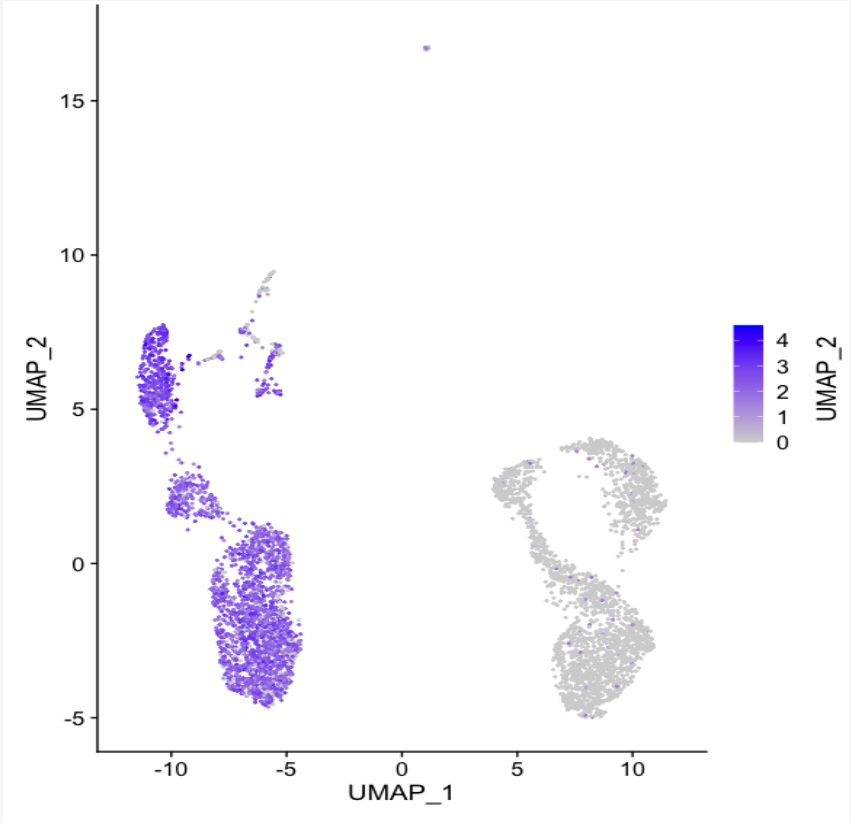
Follow cell fate

(ie. Development/differentiation)

Annotate cell cycle

(for QC or as biological question)

Cell Annotation



Example of research questions

Identify cell heterogeneity

(different cell abundances)

Identify new cell type subsets

(by function, molecular signature..)

Compare conditions

(ie. tumor vs normal)

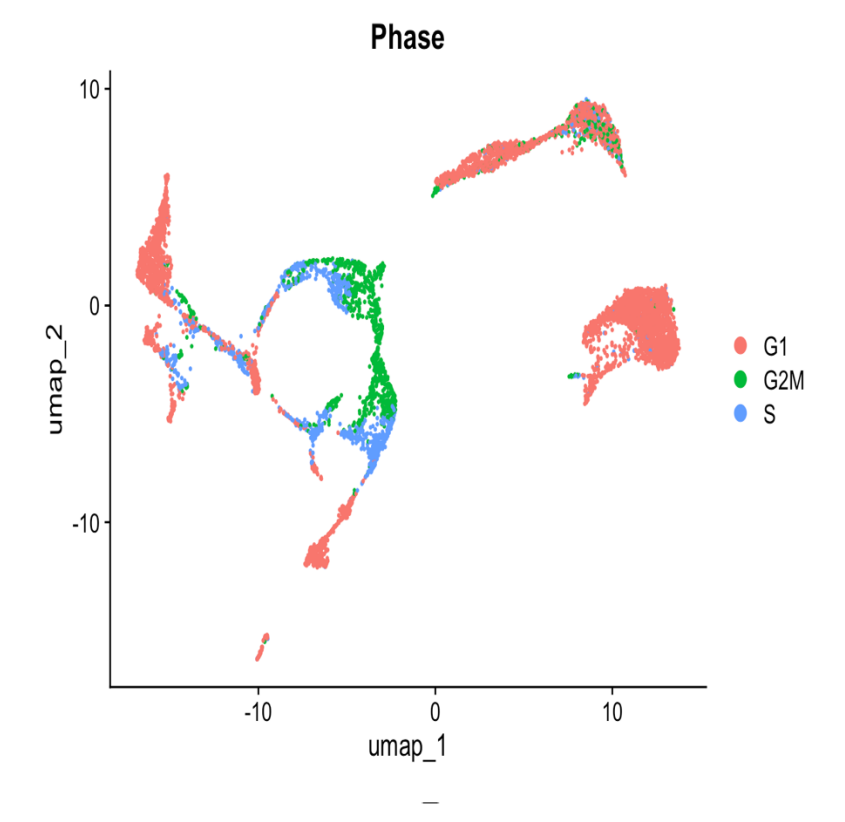
Follow cell fate

(ie. Development/differentiation)

Annotate cell cycle

(for QC or as biological question)

Cell Annotation



Example of research questions

Identify cell heterogeneity

(different cell abundances)

Identify new cell type subsets

(by function, molecular signature..)

Compare conditions

(ie. tumor vs normal)

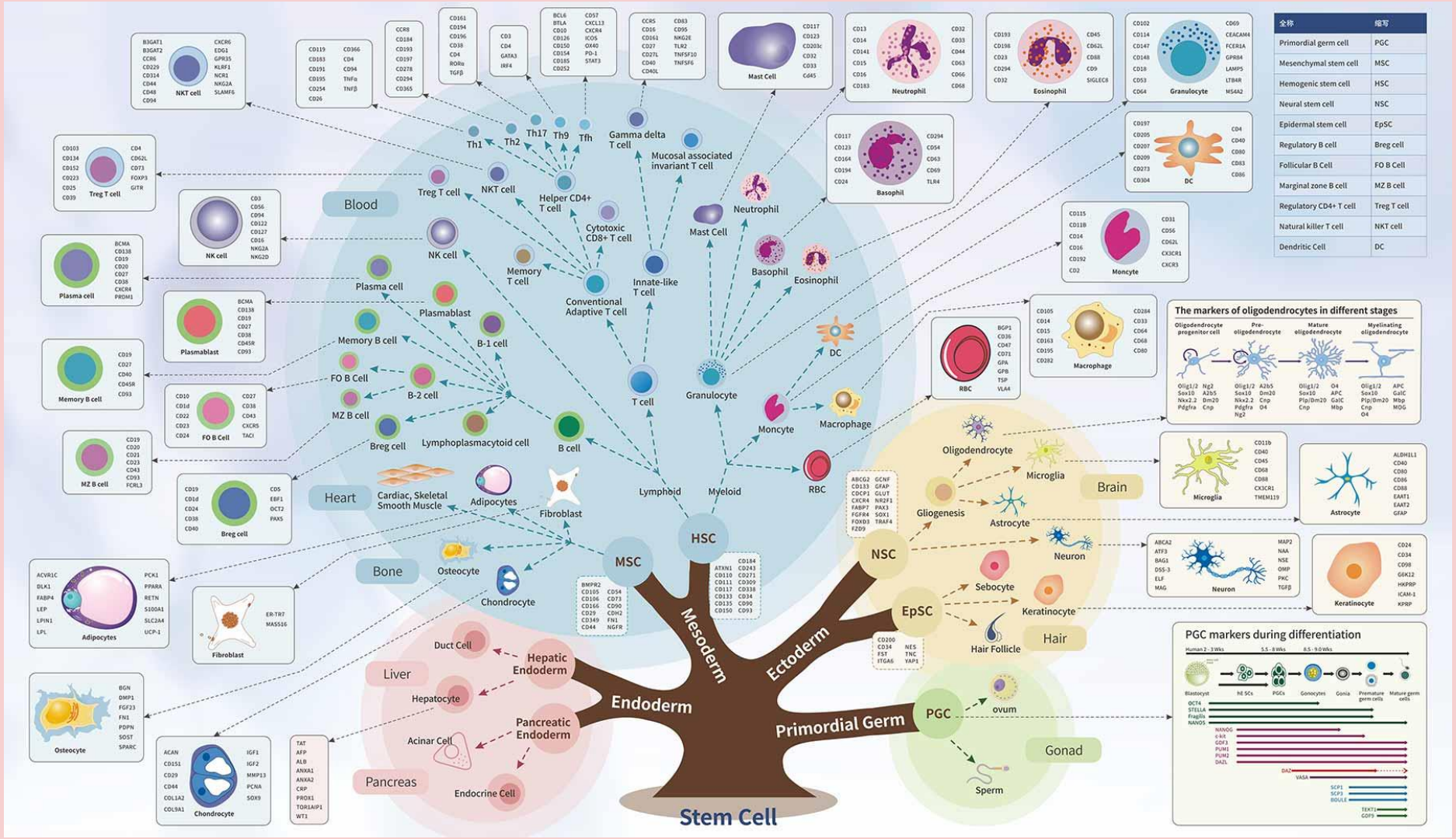
Follow cell fate

(ie. Development/differentiation)

Annotate cell cycle

(for QC or as biological question)

What is a cell type?



全称	缩写
Primordial germ cell	PGC
Mesenchymal stem cell	MSC
Hemogenic stem cell	HSC
Neural stem cell	NSC
Epidermal stem cell	EpSC
Regulatory B cell	Breg cell
Follicular B Cell	FO B Cell
Marginal zone B cell	MZ B cell
Regulatory CD4+ T cell	Treg T cell
Natural killer T cell	NKT cell
Dendritic Cell	DC

Adapted from: https://www.cusabio.com/statics/images/Cell-Marker.jpg

Cell Annotation

Cells could be defined in terms of:

- Function
- Location
- Tissue type
- Cell morphology

Identify and label cell clusters using relevant biological terms

Example of research questions

Identify cell heterogeneity
(different cell abundances)

Identify new cell type subsets
(by function, molecular signature..)

Compare conditions
(ie. tumor vs normal)

Follow cell fate
(ie. Development/differentiation)

Annotate cell cycle
(for QC or as biological question)

Cell Annotation Approaches

Manual →

↓

Automatic

Gene Markers

Identify marker genes from the literature that give us a proxy to explain cell identity

Cell Annotation Approaches

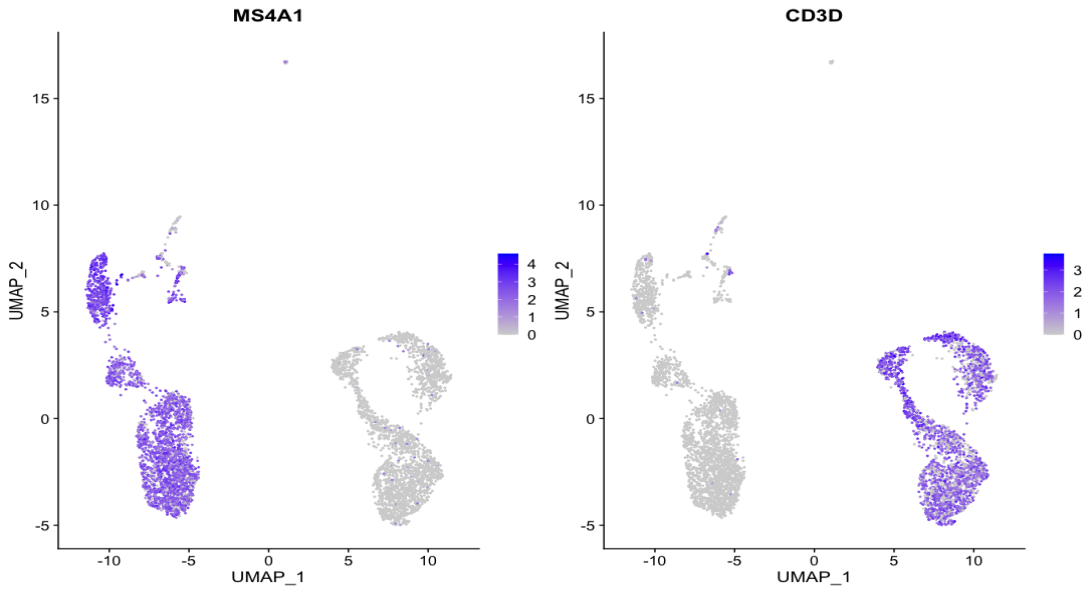
Manual →

↓

Automatic

Gene Markers

Identify marker genes from the literature that give us a proxy to explain cell identity



Cell Annotation Approaches

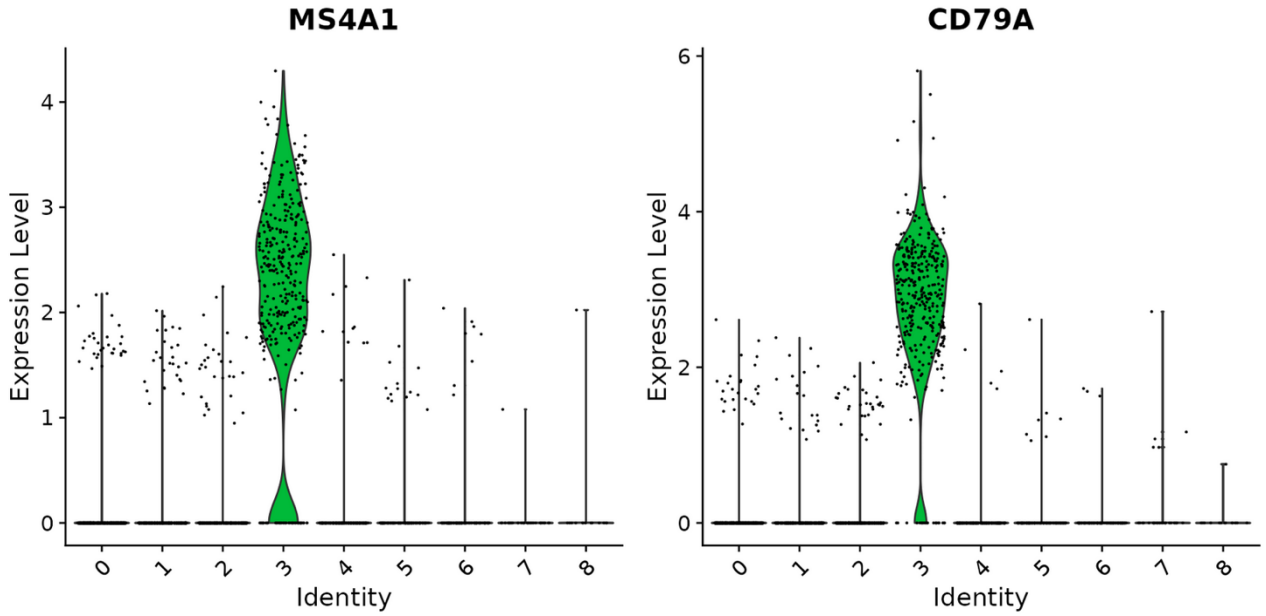
Manual →

↓

Automatic

Gene Markers

Identify marker genes from the literature that give us a proxy to explain cell identity

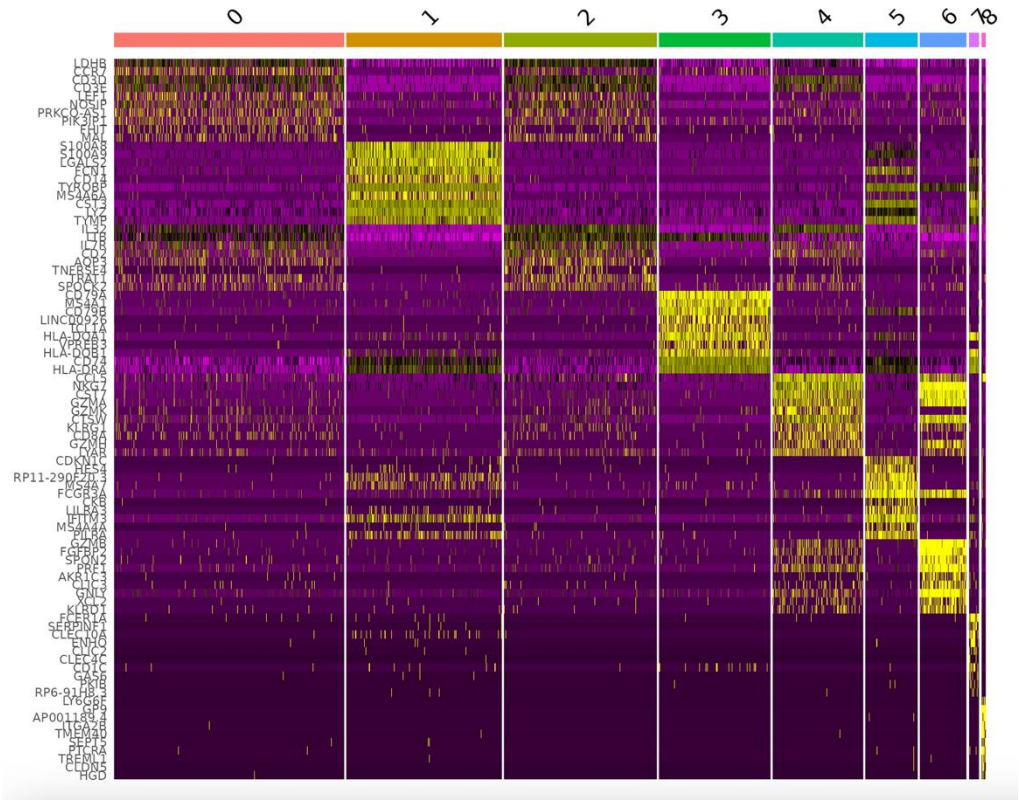


Cell Annotation Approaches

Manual →

Gene Markers

Finding differentially expressed features (cluster biomarkers)



Automatic

Cell Annotation Approaches

Manual →

Gene Markers

↓

Automatic

It requires expert knowledge

Can be time consuming, subjective

Good markers are not always known or specific to one cell type

Cell surface proteins may not always correlate with mRNA levels

Master transcription factors that drive cell fate are often better markers than cell surface proteins

You may need a combination of genes, one may not be enough

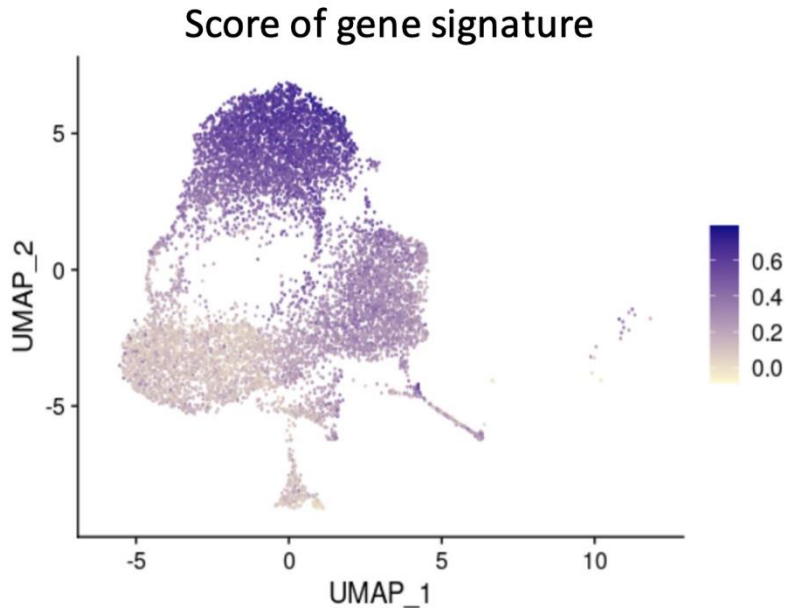
Cell Annotation Approaches

Manual

Module Score

Automatic

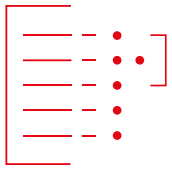
Compare expression level of genes belonging to the signature to "control genes" with similar expression level (Tirosh et al 2016, Science 352:6282)




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

Cell Annotation Approaches

Manual



Automatic

- Cell level
- Cluster level

It reduces the chance of missing cell differences / cell types

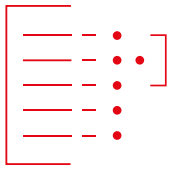
If read counts are low, more difficult to correctly predict cell type

It is faster and could be more accurate (expression levels are based on several cells)

May be challenging for datasets with gradients

Cell Annotation Approaches

Manual



Automatic

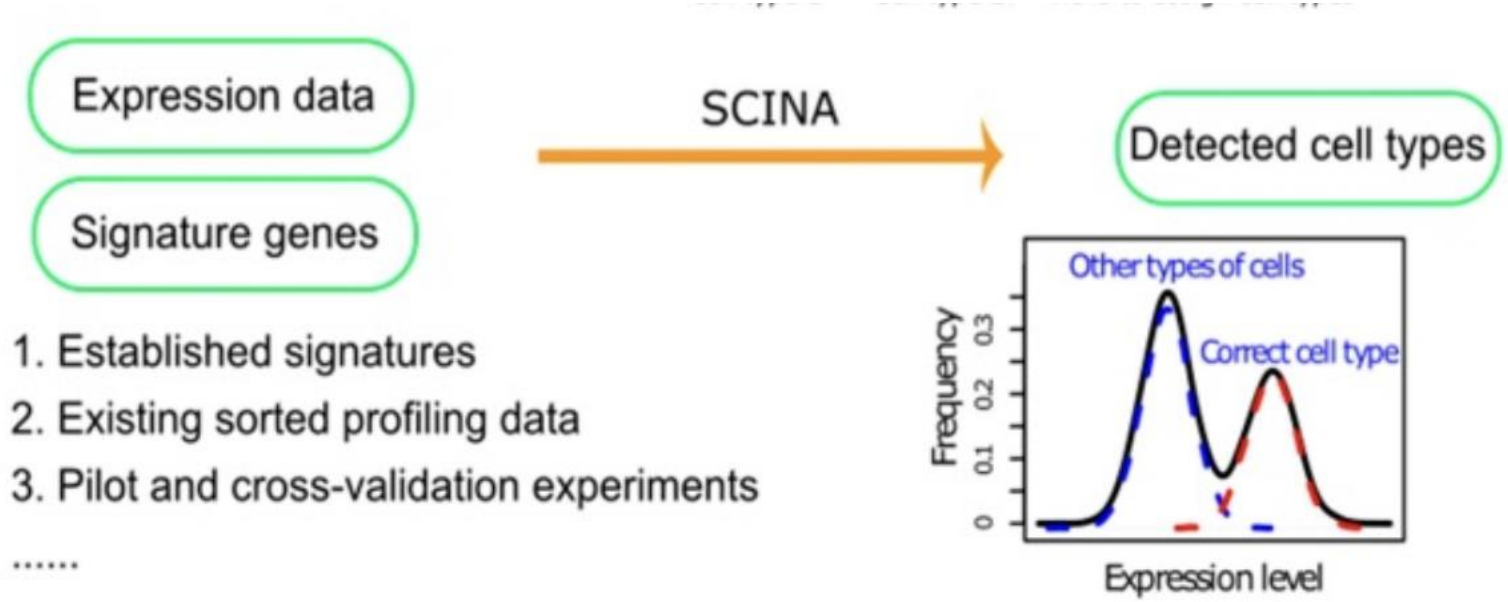
- Gene markers
- Reference dataset

Automatic Annotation

- ***Marker based automatic annotation***

Known relationships between gene markers and cell types are obtained from db or literature

- ***SCINA****:* supervised and automated approach for assigning cell types based on prior knowledge of signature genes and can directly arrive at detected cell types. The gene signatures could come from a variety of sources.



Automatic Annotation

- ***Marker based automatic annotation***

Known relationships between gene markers and cell types are obtained from db or literature

- ***SCINA****:* supervised and automated approach for assigning cell types based on prior knowledge of signature genes and can directly arrive at detected cell types. The gene signatures could come from a variety of sources.
- **AUCell**: uses the "Area Under the Curve" (AUC) to calculate whether a critical subset of the input gene set is enriched within the expressed genes for each cell.
- **GSVA:** Given a db of markers gene set, it identifies enriched sets in within the highly expressed genes in a cluster

Automatic Annotation

- ***Marker based automatic annotation***

Known relationships between gene markers and cell types are obtained from db or literature

Works well specially if several genes are used

Reliable markers are not available for all cell types

Automatic Annotation

- ***Reference based automatic annotation***

  ScRNAseq data to be annotated (query) is compared to a similar and expertly annotated reference dataset (reference). Reference label transferred to query

Automatic Annotation

- ***Reference based automatic annotation***

ScRNAseq data to be annotated (query) is compared to a similar and expertly annotated reference dataset (reference). Reference label transferred to query

- ***Scmap****: one of the best performing tools. High accuracy and good avoidance of incorrect labelling when there are novel cell types*
- ***SingleCellNet****: Good accuracy if all cell types are present, low if data is incomplete or is a poor match*
- ***SingleR****: based on spearman rank correlation. An advantage is that it includes a general database for reference datasets.*

Automatic Annotation

- ***Reference based automatic annotation***

ScRNAseq data to be annotated (query) is compared to a similar and expertly annotated reference dataset (reference). Reference label transferred to query

Relies on available datasets and resources may be incomplete.

Depends on quality of reference dataset

It needs further curation

Automatic annotation

- Easy access to rich reference data:

  - HPCA: hand-annotated Human Primary Cell Atlas 37 main types, 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 can also be used to evaluate similarity to a custom reference

SIB

SingleR - Steps

Step1

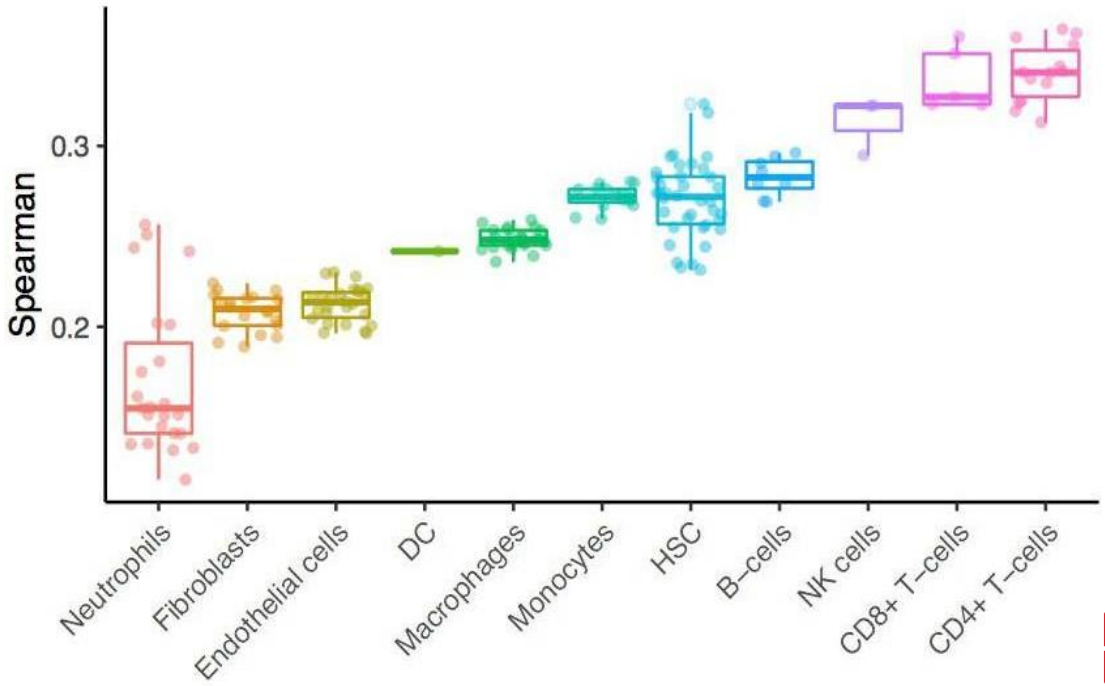
- Identifying variable genes among cell types in the reference set

Step2

- Correlating each single-cell transcriptome with each sample in the reference set

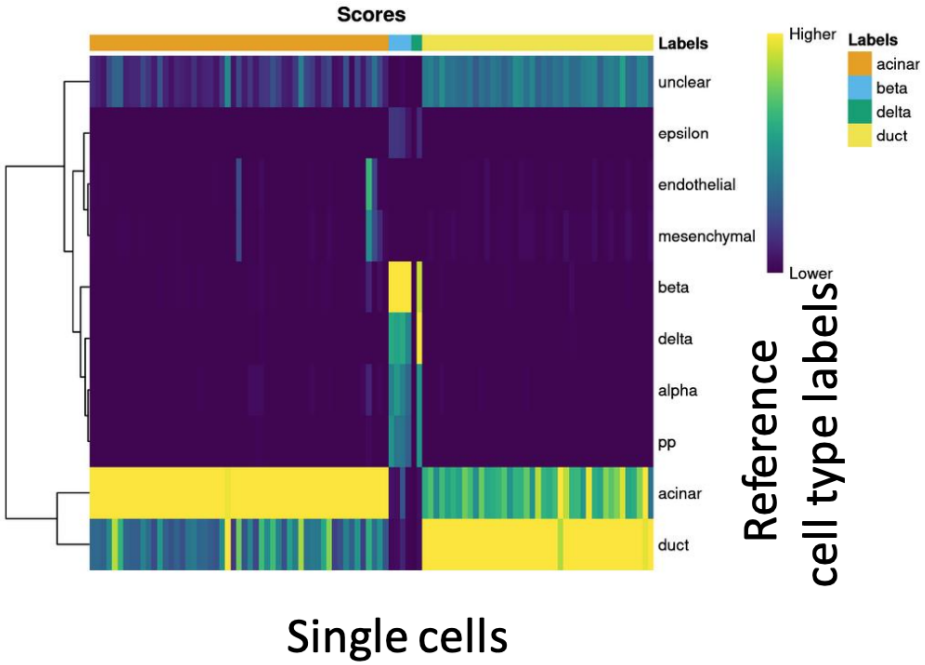
Step3

- Iterative fine-tuning: reducing the reference set to only top cell types



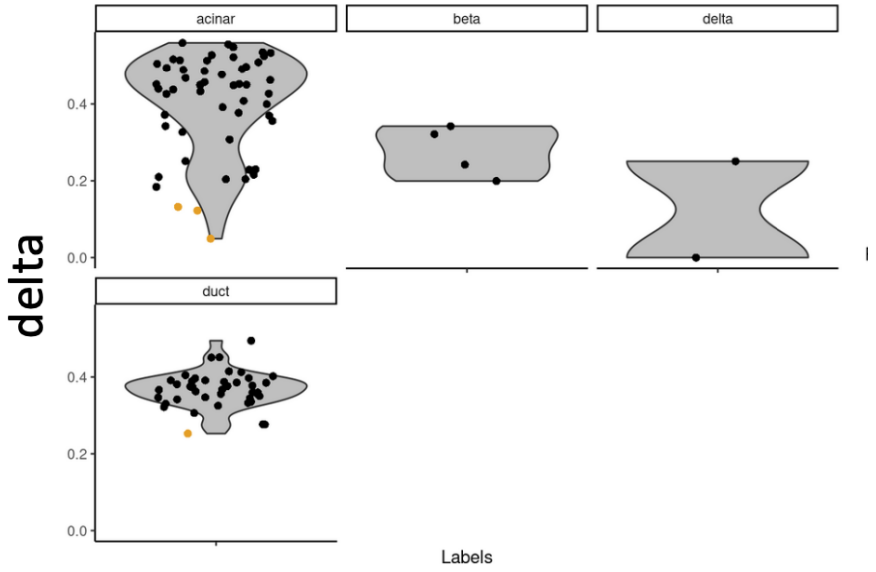
SingleR – annotation diagnostics

Heatmap of scores per cell for each label:




```
plotScoreHeatmap(singleR.object)
```

Reference cell type labels



"delta": the difference between the score for the assigned label and the median across all labels for each cell

```
plotDeltaDistribution(singleR.object)
```

Try them during the practical exercise

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

Methods avaible

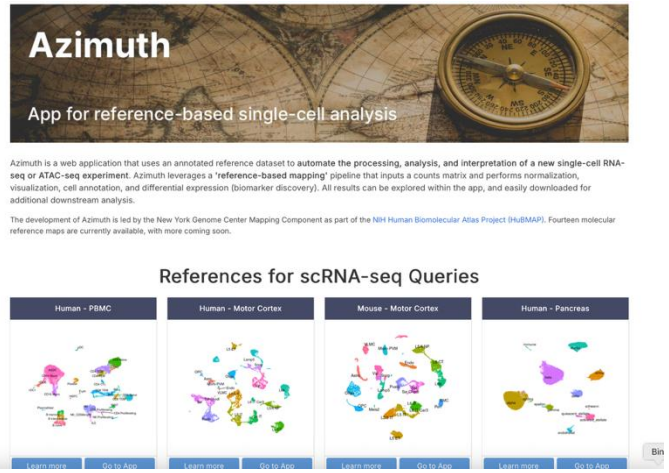
Tool	Type	Language	Resolution	Approach	Allows 'None'	Notes
singleCell Net[42]	Reference based	R	Single cells	Relative-expression gene pairs + random forest	Yes, but rarely does so even when it should[33]	10–100× slower than other methods; high accuracy
scmap-cluster[41]	Reference based	R	Single cells	Consistent correlations	Yes	Fastest method available; balances false-positives and false-negatives; includes web interface for use with a large pre-built reference or custom reference set
scmap-cell[41]	Reference based	R	Single cells	Approximate nearest neighbors	Yes	Assigns individual cells to nearest neighbor cells in reference; allows mapping of cell trajectories; fast and scalable
singleR[43]	Reference based	R	Single cells	Hierarchical clustering and Spearman correlations	No	Includes a large marker reference; does not scale to data sets of ≥10,000 cells; includes web interface with marker database
Scikit-learn[102]	Reference based	Python	Multiple possible	k-nearest neighbors, support vector machine, random forest, nearest mean classifier and linear discriminant analysis	(Optional)	Expertise required for correct design and appropriate training of classifier while avoiding overtraining
AUCell[103]	Marker based	R	Single cells	Area under the curve to estimate marker gene set enrichment	Yes	Because of low detection rates at the level of single cells, it requires many markers for every cell type
SCINA[34]	Marker based	R	Single cells	Expectation maximization, Gaussian mixture model	(Optional)	Simultaneously clusters and annotates cells; robust to the inclusion of incorrect marker genes
GSEA/GSVA[36,104]	Marker based	R/Java	Clusters of cells	Enrichment test	Yes	Marker gene lists must be reformatted in GMT format. Markers must all be differentially expressed in the same direction in the cluster
Harmony[105]	Integration (Box 2)	R	Single cells	Iterative clustering and adjustment	Yes	Integrates only lower-dimensional projection of the data; seamlessly integrated into Seurat pipeline; may overcorrect data
Seurat-canonical correlation analysis[106]	Integration (Box 2)	R	Single cells	MNN anchors + canonical correlation analysis	Yes	Accuracy depends on the accuracy of MNN anchors, which are automatically-identified corresponding cells across data sets
mnnCorrect[107]	Integration (Box 2)	R	Single cells	MNN pairs + singular value decomposition	Yes	Accuracy depends on the accuracy of MNN pairs (cells matched between data sets). Referred to in Box 2
Linked inference of genomic experimental relationships (LIGER)[108]	Integration (Box 2)	R	Single cells	Non-negative matrix factorization	Yes	Allows interpretation of data set–specific and shared factors of variation. Referred to in Box 2

Databases with cell type markers genes

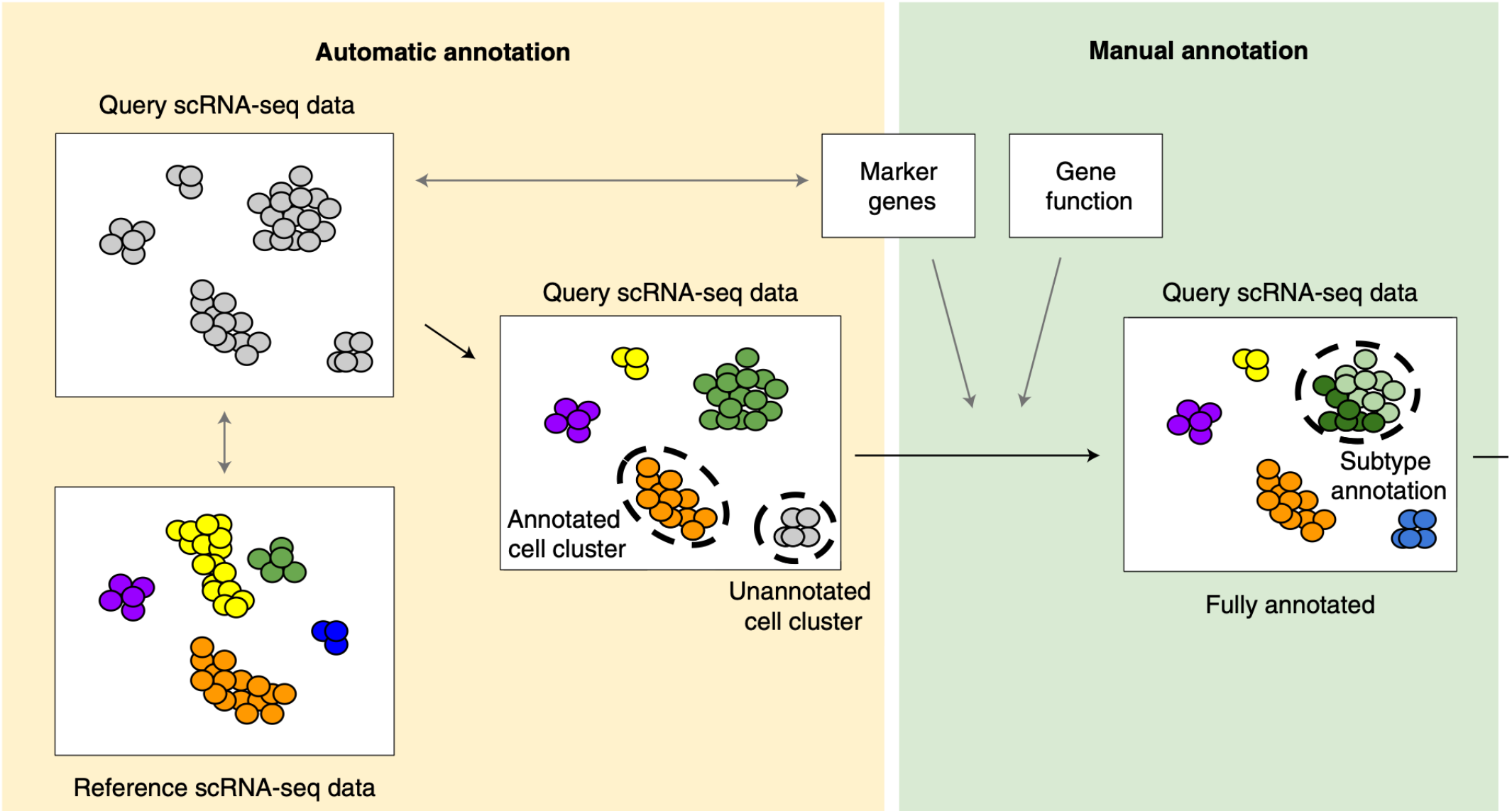
- PanglaoDB https://panglaodb.se/ (mouse and human)
  R: https://cran.r-project.org/web/packages/rPanglaoDB/index.html
- CellMarker (mouse and human): http://bio-bigdata.hrbmu.edu.cn/CellMarker/
- SingleR https://github.com/dviraran/SingleR (Aran et al. 2019), access via celldex package, e.g. human primary cell atlas (microarrays, bulk)
- 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

Reference datasets available

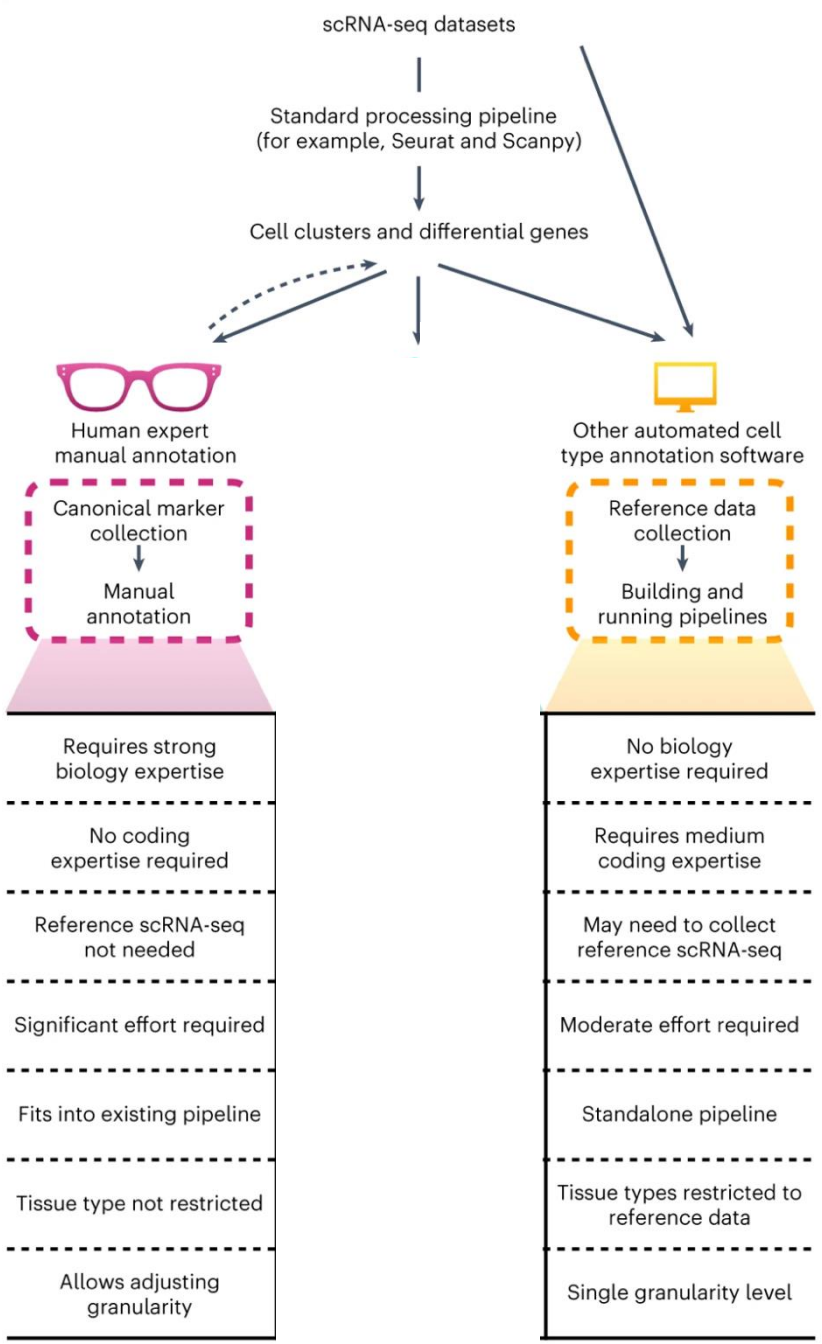
- Single cell portal: https://singlecell.broadinstitute.org/single_cell (10 species)
- Single Cell Expression Atlas: https://www.ebi.ac.uk/gxa/sc/experiments/E-MTAB-10662/results/cell-plots (single cell expression across species)
- Gene Expression Omnibus (GEO)
- Azimuth https://azimuth.hubmapconsortium.org/ (app for reference-based single-cell analysis)



Summary


Automatic annotation
Manual annotation
Query scRNA-seq data
Marker genes
Gene function
Query scRNA-seq data
Query scRNA-seq data
Annotated cell cluster
Unannotated cell cluster
Subtype annotation
Fully annotated
Reference scRNA-seq data

Summary

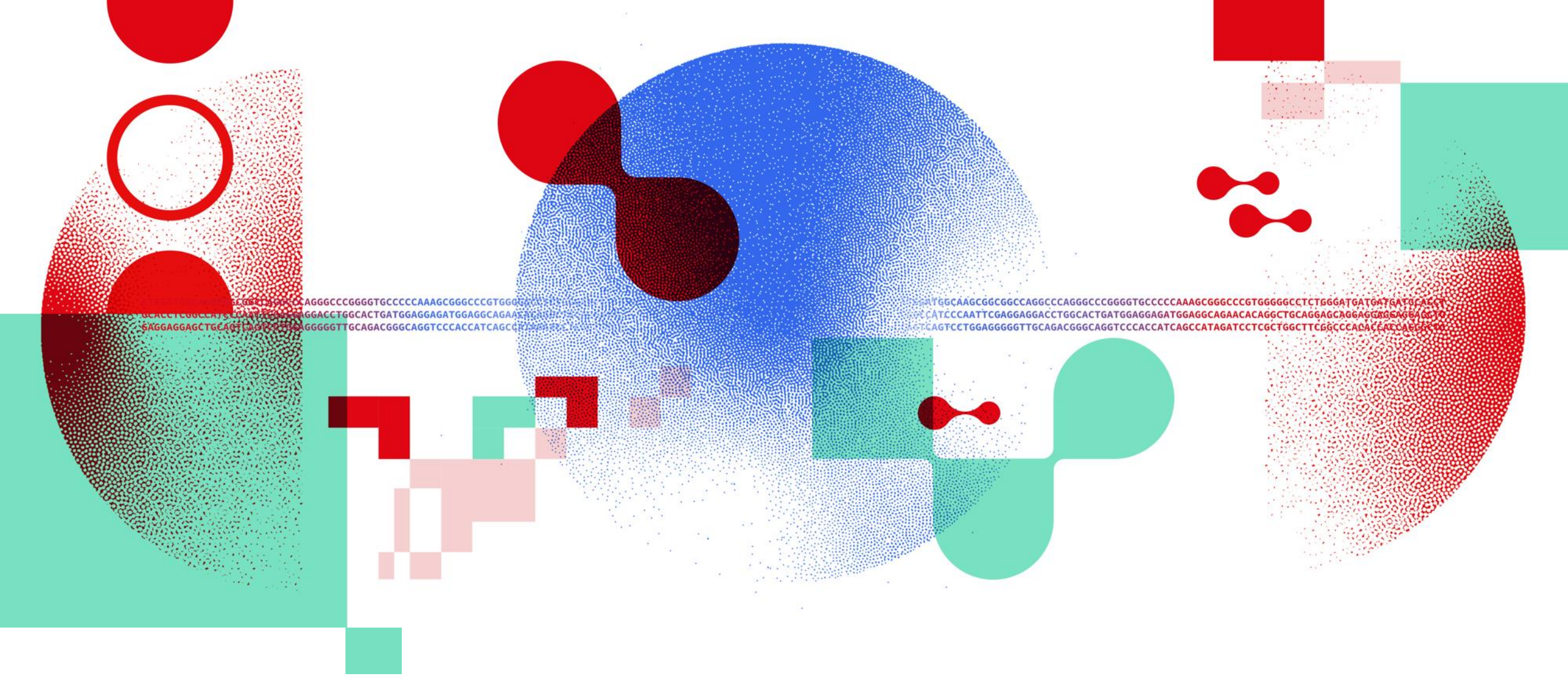


Try different methods

! Needs manual curation

Depends on quality and completeness of reference

Adapted from: Hou and Ji, Nature Methods 2024



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss