



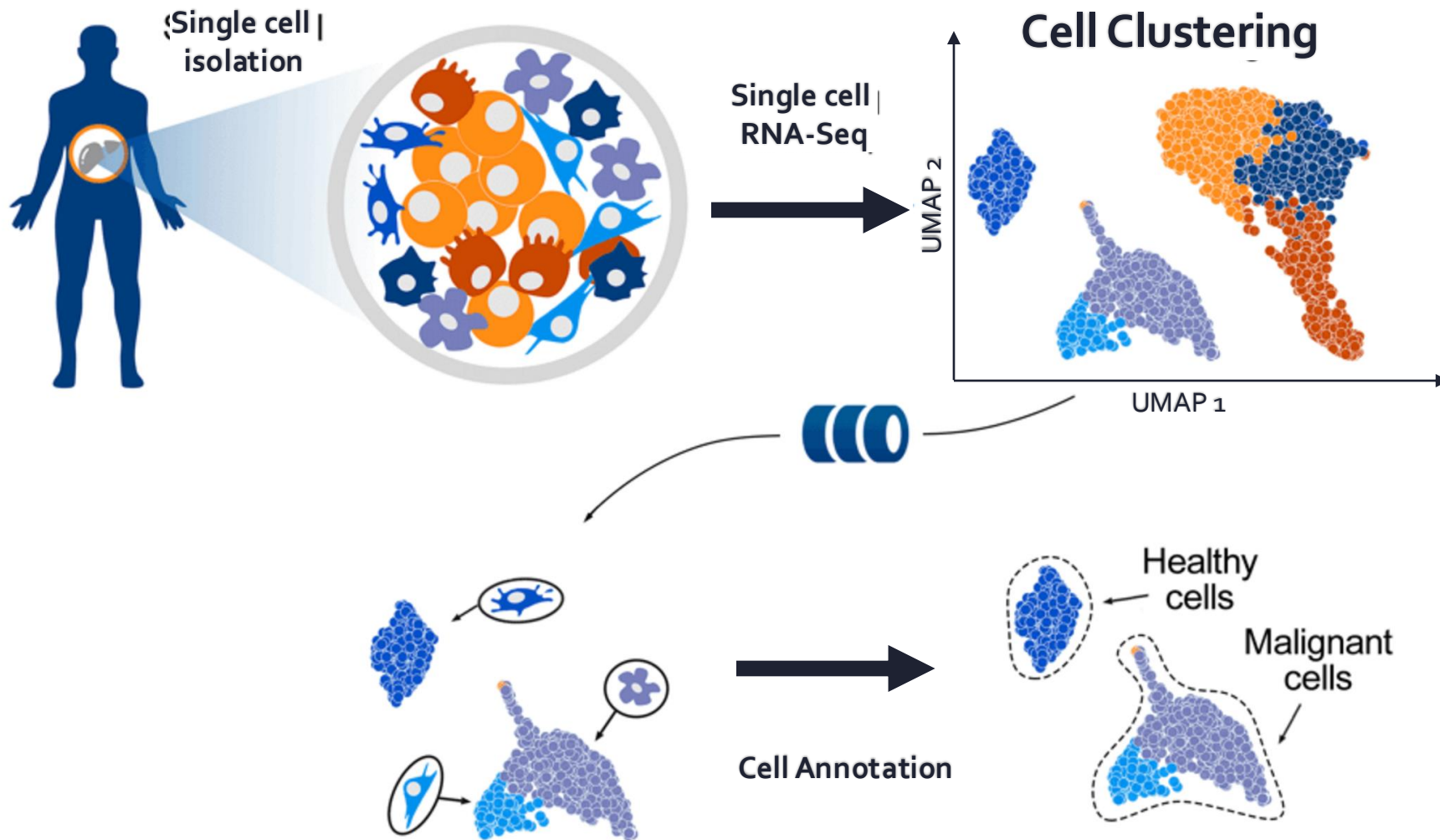
Swiss Institute of
Bioinformatics

Cells Annotation

Luciano Cascione, PhD
Bioinformatics Core Unit

LUCIANO CASCIONE, PHD
BELLINZONA, Nov. 14TH 2025

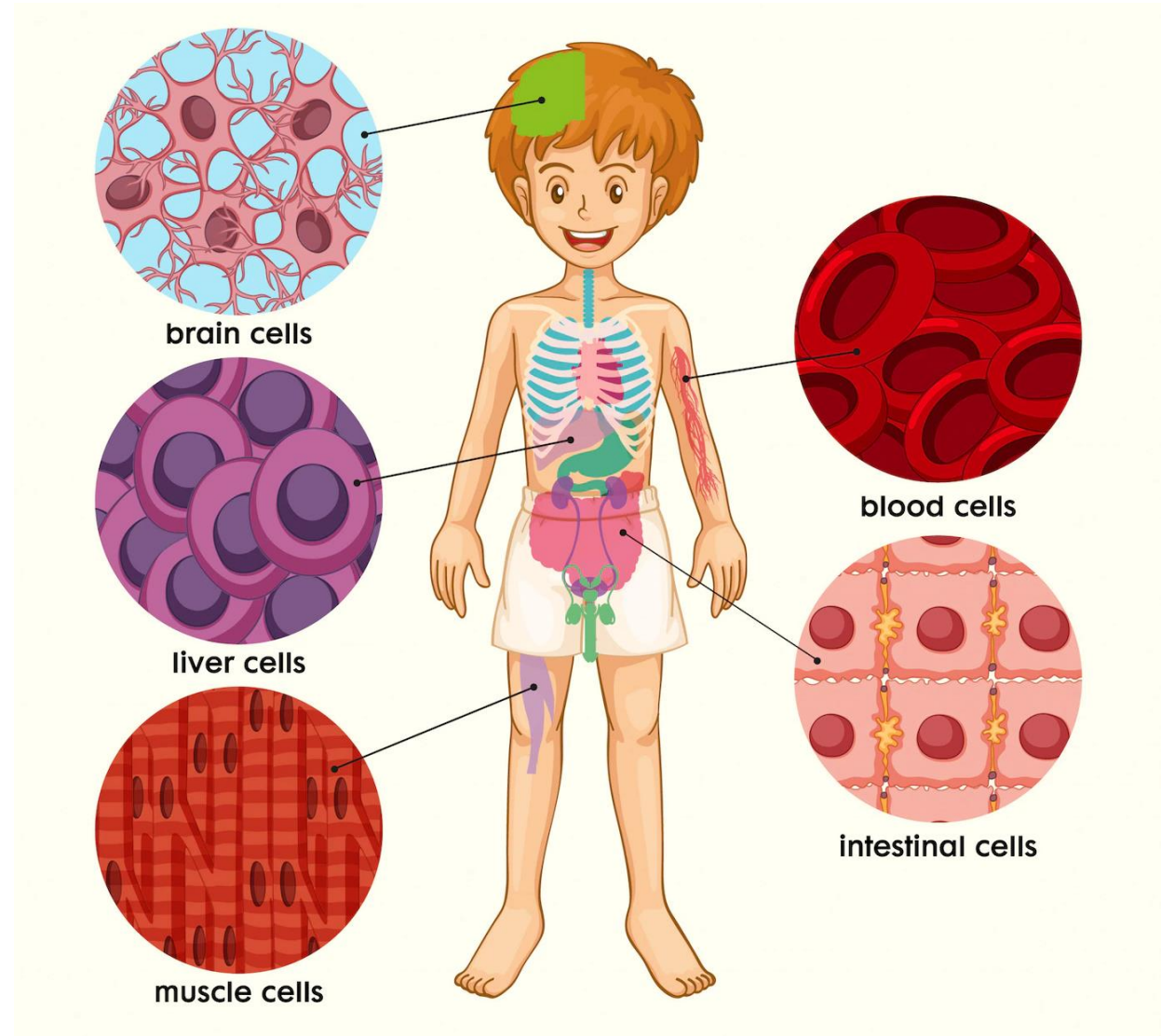
scRNA-Seq workflow



The fundamental unit of life

Cells could be defined in terms of:

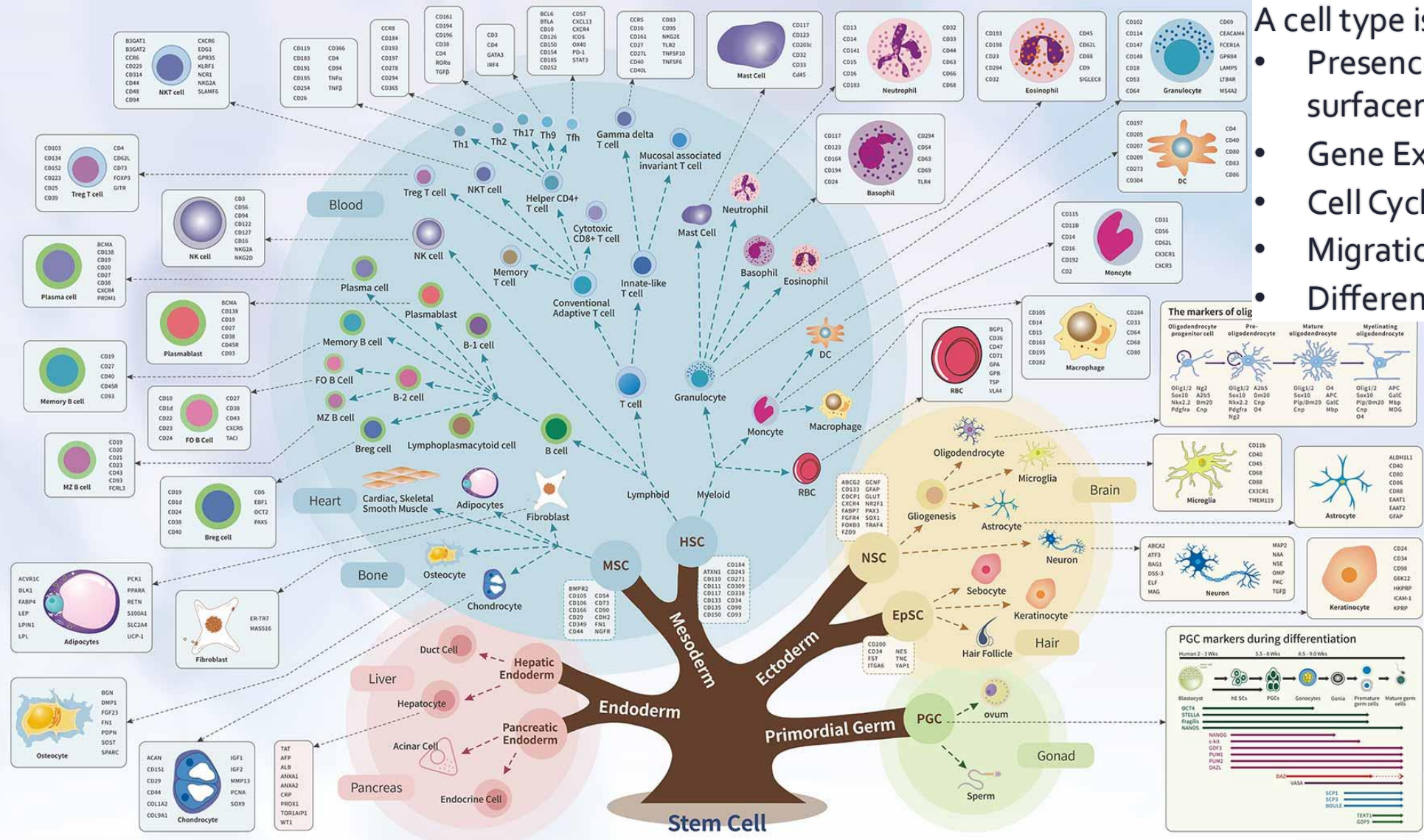
- Function
- Location
- Tissue type
- Cell morphology



What is a "cell type"?

A cell type is characterized by:

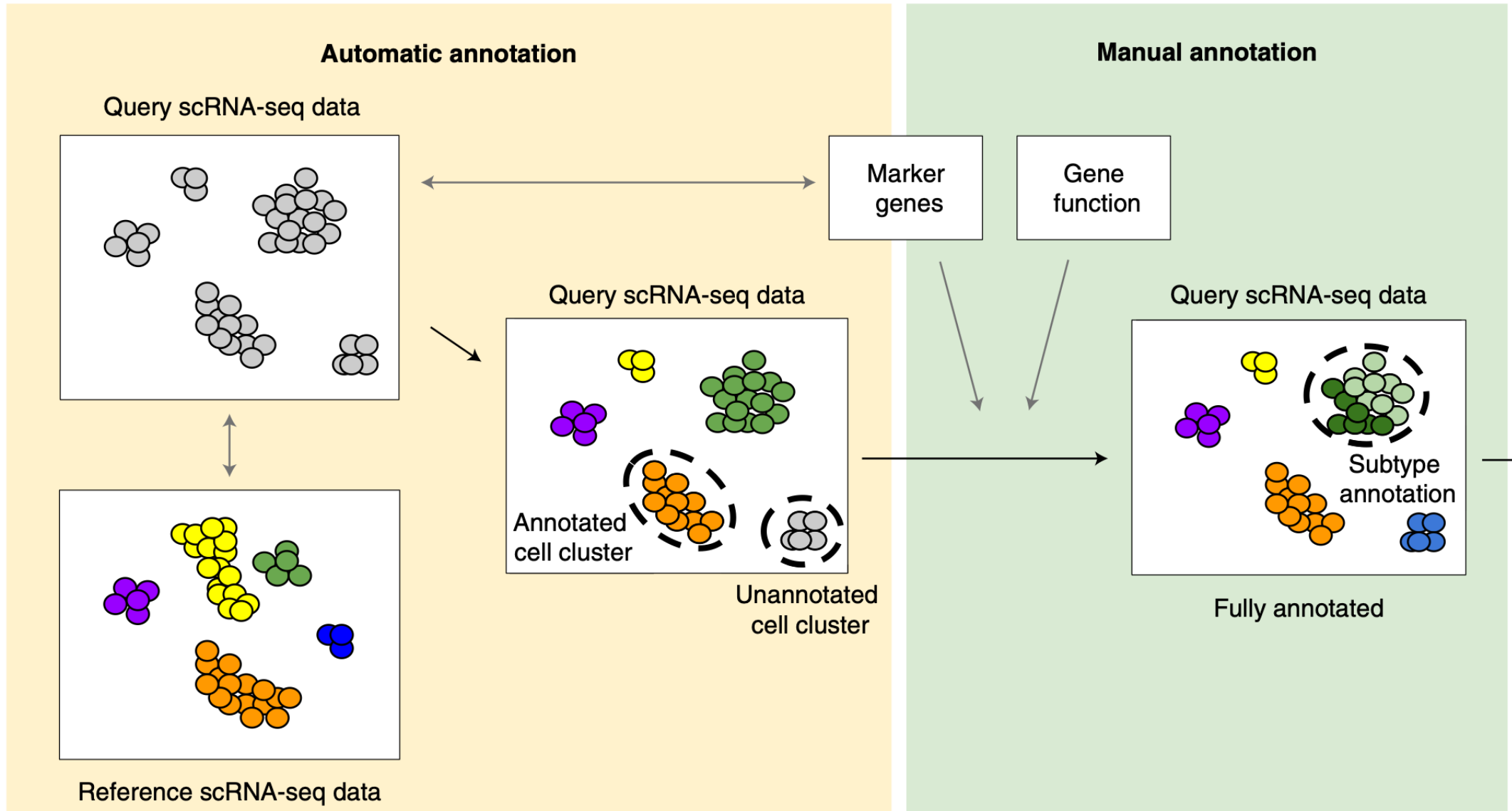
- Presence/absence of cell surfacemarkers
- Gene Expression
- Cell Cycle phase
- Migration State
- Differentiation: cell state



Why should we identify cell types?

- Find new cell types which have been missed by using “standard” surface markers
- To compare the abundance of cell types in different conditions
- Follow cell fate and determine cell differentiation mechanisms
- To determine which cell types might communicate with each other

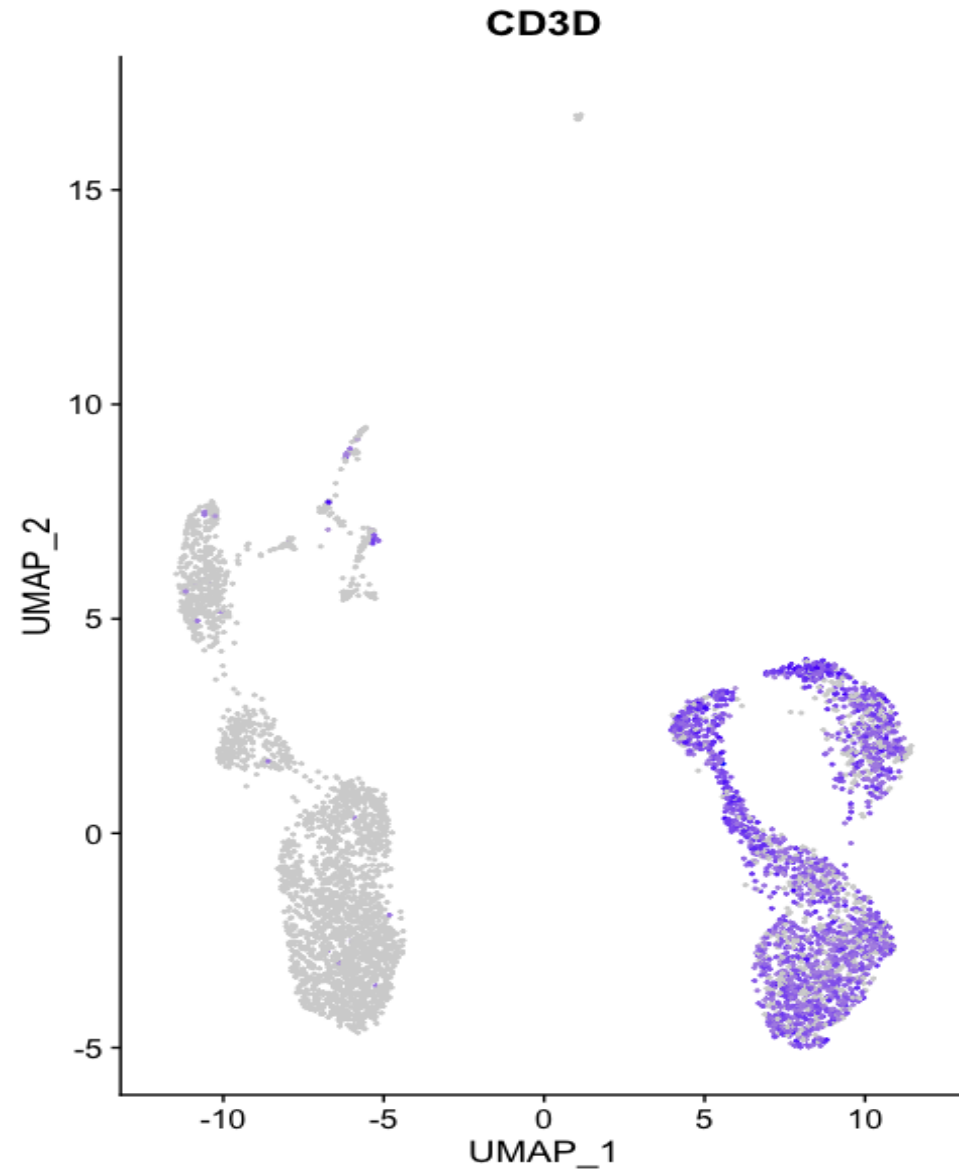
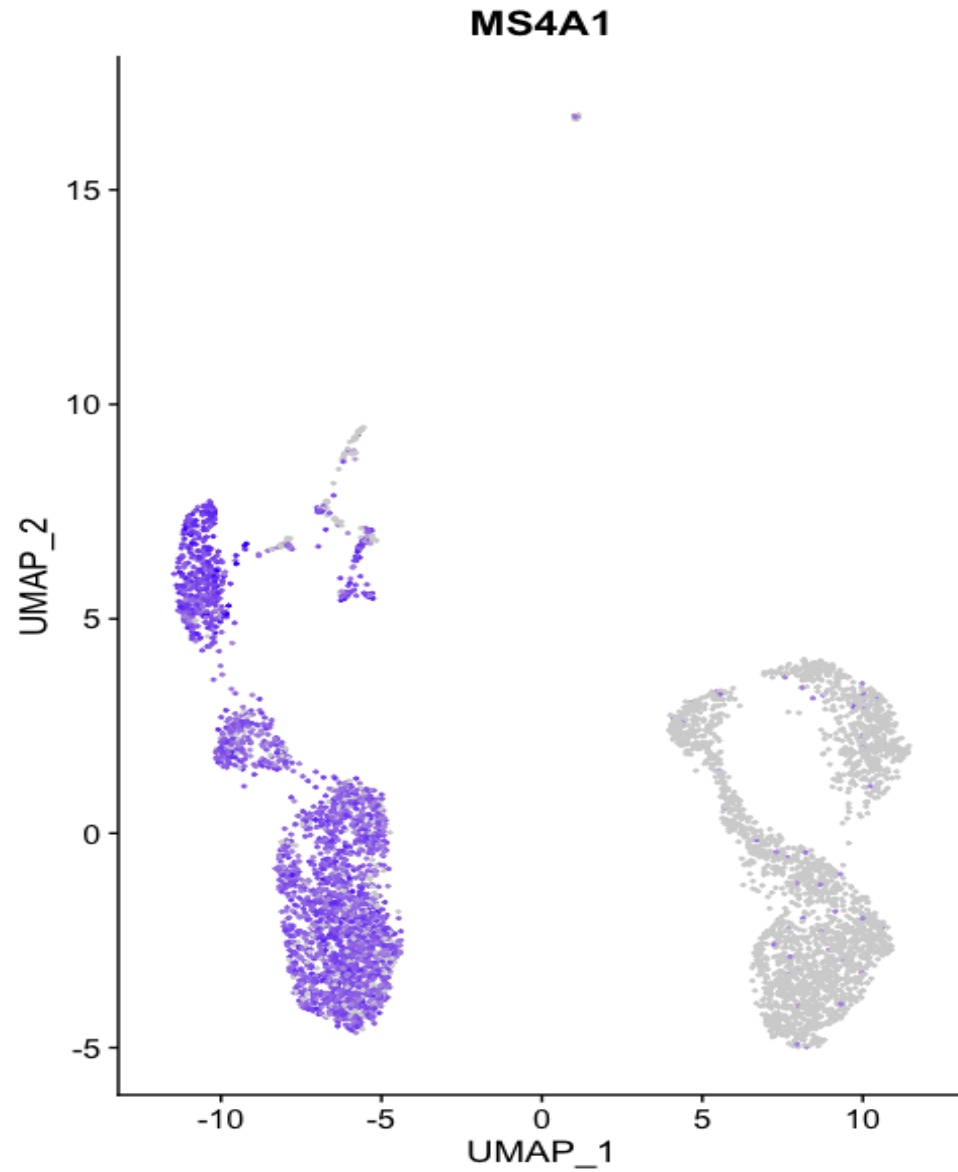
Cell Annotation workflow



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 cell type annotation

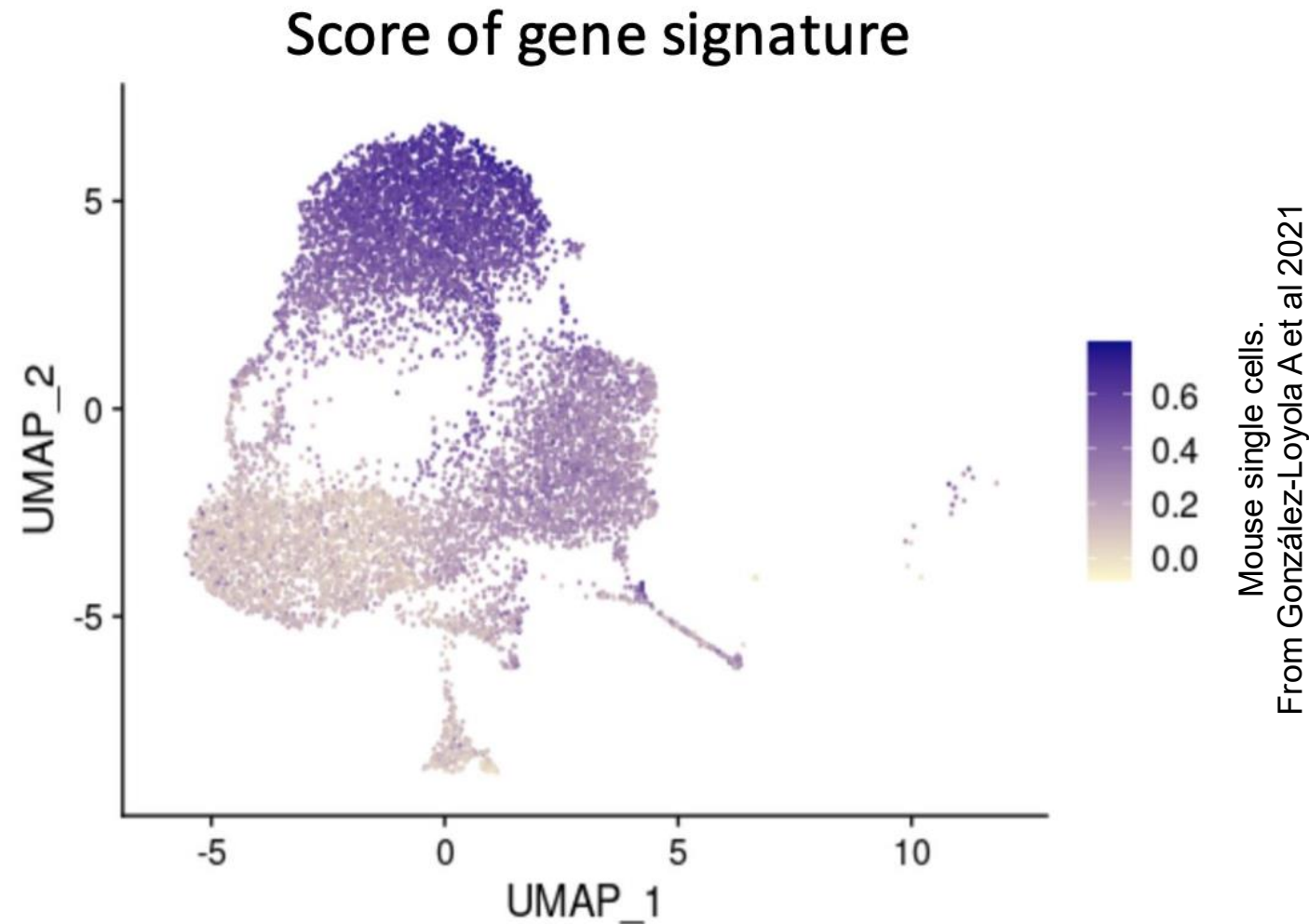


Databases with cell type markers genes

- 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

Module Score

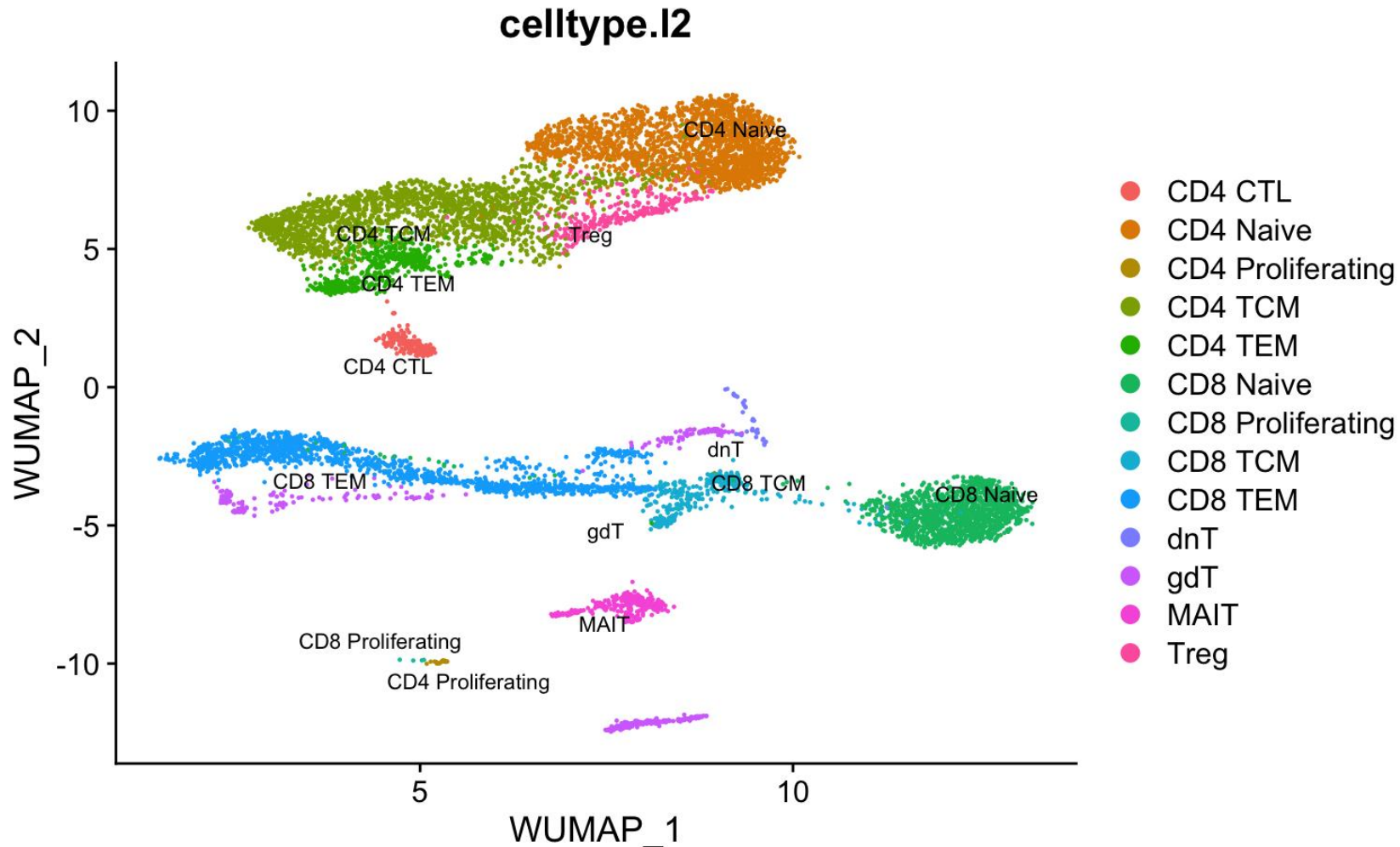
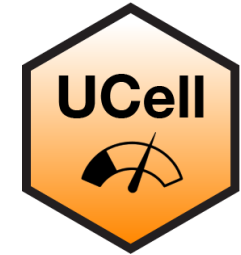
Compare expression level of genes belonging to the signature to “control” genes with similar expression level to signature genes (Tirosh et al 2016, Science 352:6282)



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

UCell

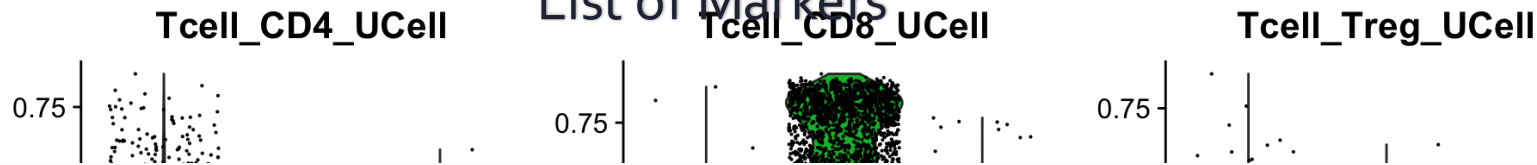
A robust and scalable single-cell gene signature scoring, uses *positive* and *negative* genes and mitigates data sparsity by nearest neighbours smoothing



UCell

A robust and scalable single-cell gene signature scoring, uses *positive* and *negative* genes and mitigates data sparsity by nearest neighbours smoothing

List of Markers



```
markers <- list()
markers$Tcell_CD4 <- c("CD4", "CD40LG")
markers$Tcell_CD8 <- c("CD8A", "CD8B")
markers$Tcell_Treg <- c("FOXP3", "IL2RA")
markers$Tcell_MAIT <- c("KLRB1", "SLC4A10", "NCR3")
markers$Tcell_gd <- c("TRDC", "TRGC1", "TRGC2", "TRDV1", "TRAC-", "TRBC1-", "TRBC2-")
markers$Tcell_NK <- c("FGFBP2", "SPON2", "KLRF1", "FCGR3A", "KLRD1", "TRDC", "CD3E-",
  "CD3G-")
```

identity

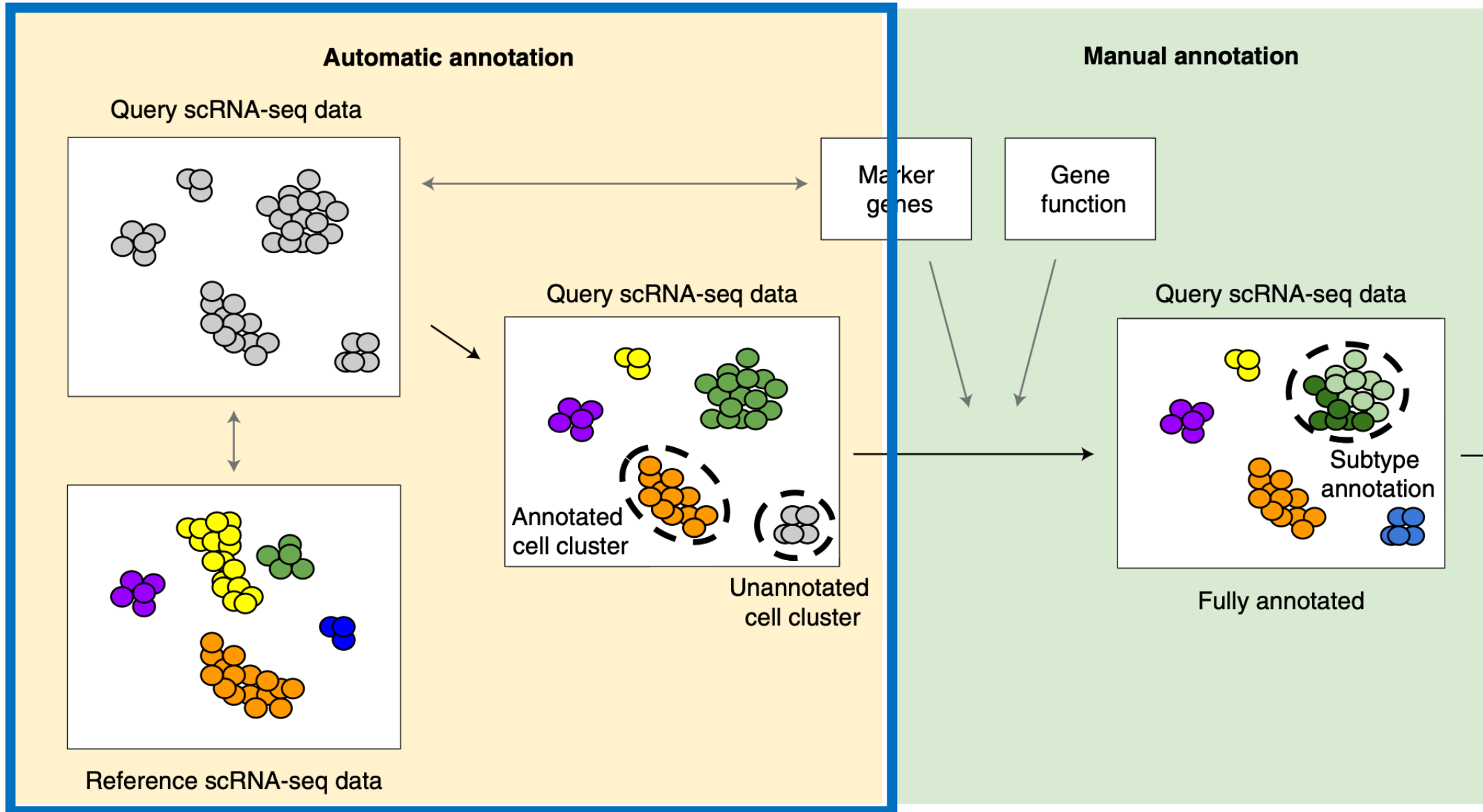
identity

identity

```
seu.obj <- AddModuleScore_UCell(seu.obj, features = markers)
```

<https://github.com/carmonalab/UCell>

Cell Annotation workflow



Automatic annotation

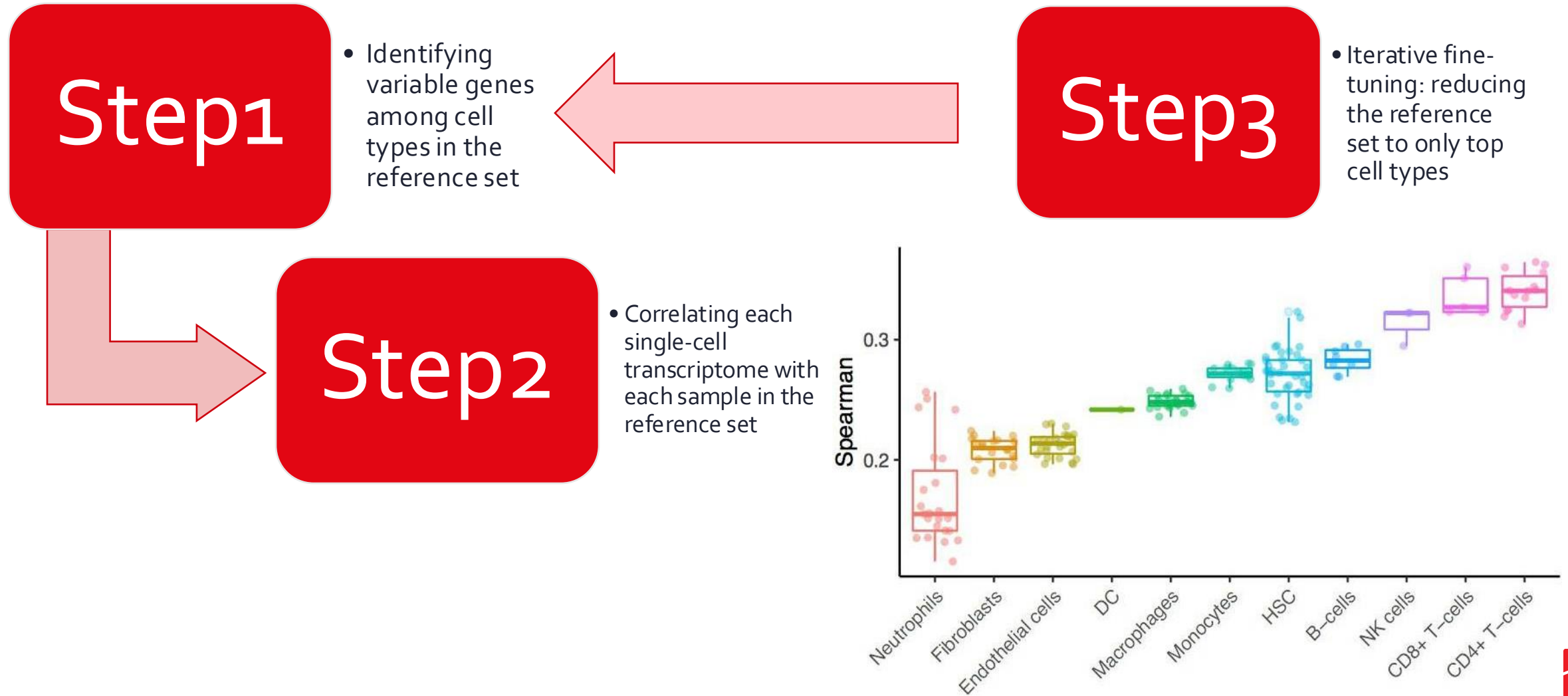
Easy access to rich reference data:

- HPCA: hand-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

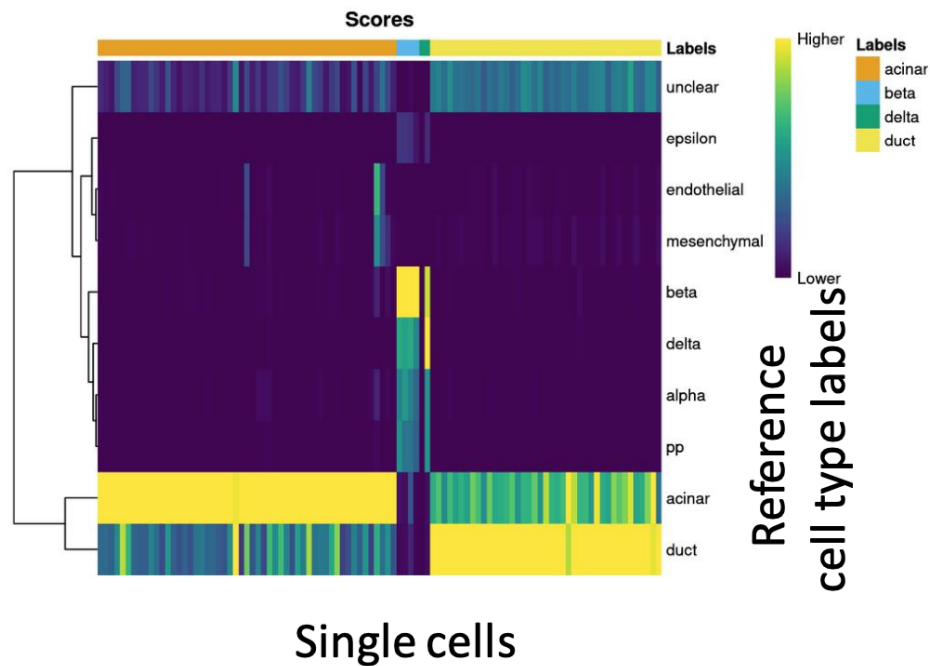


SingleR - Steps

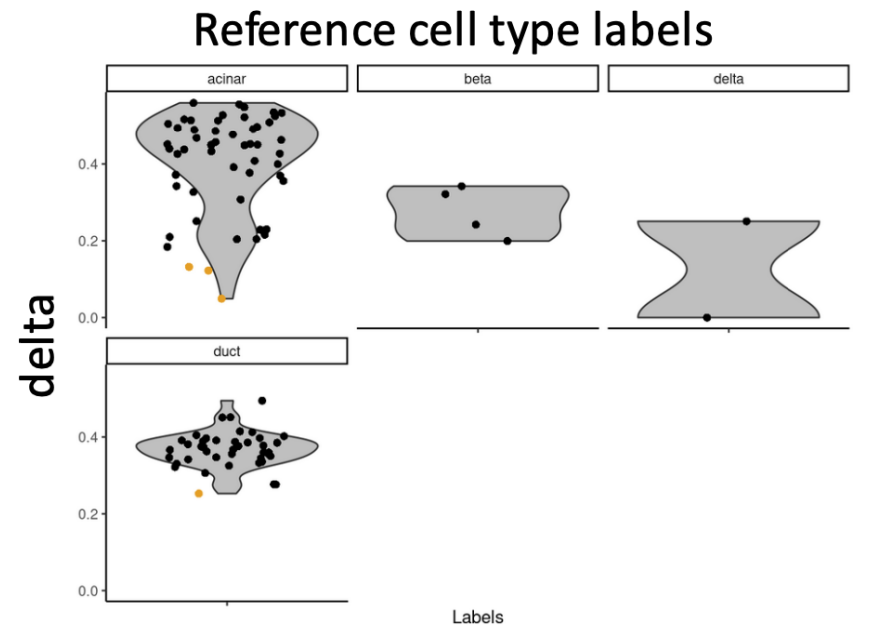


SingleR – annotation diagnostics

Heatmap of scores per cell for each label:



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



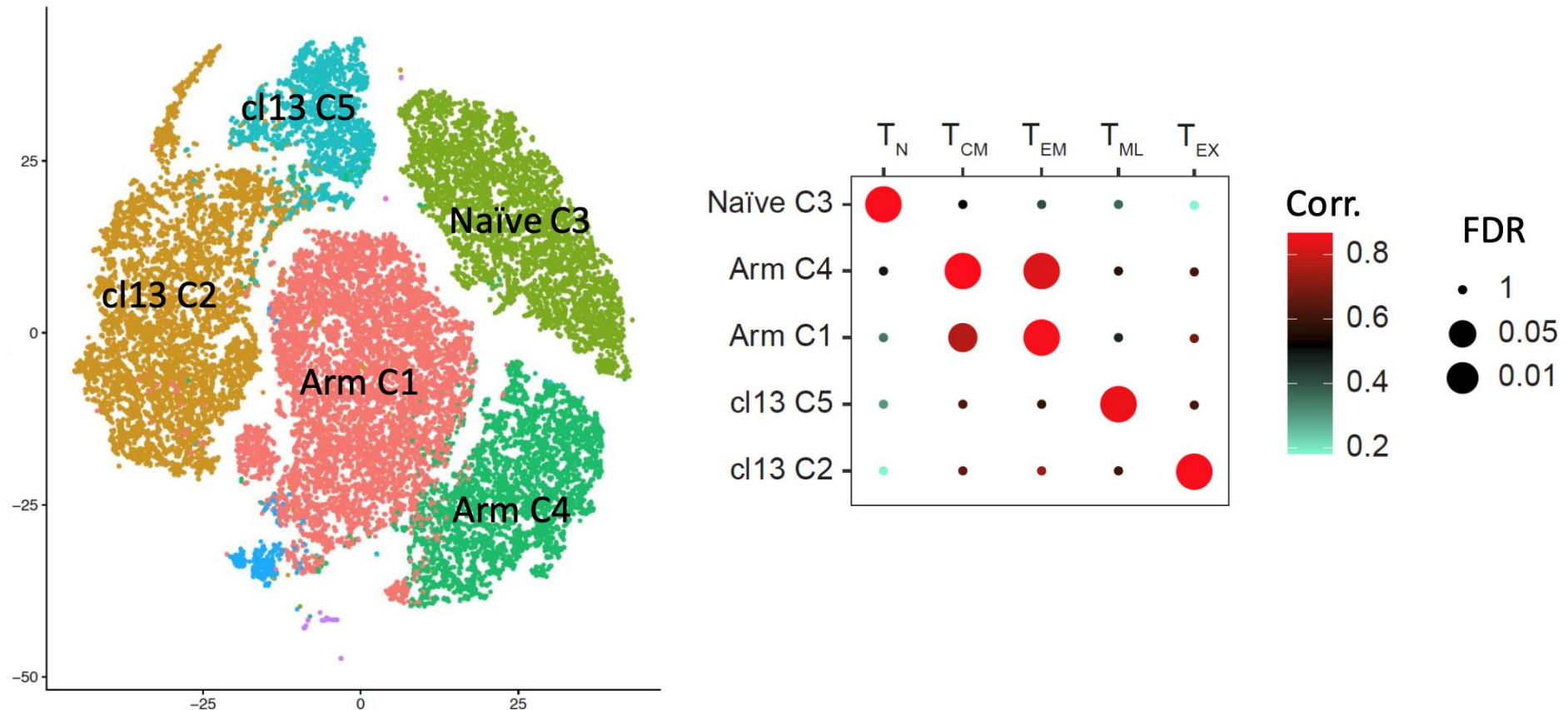
“delta”: the difference between the score for the assigned label and the median across all labels for each cell

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

Try them during the practical exercise

SingleR – Customization

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



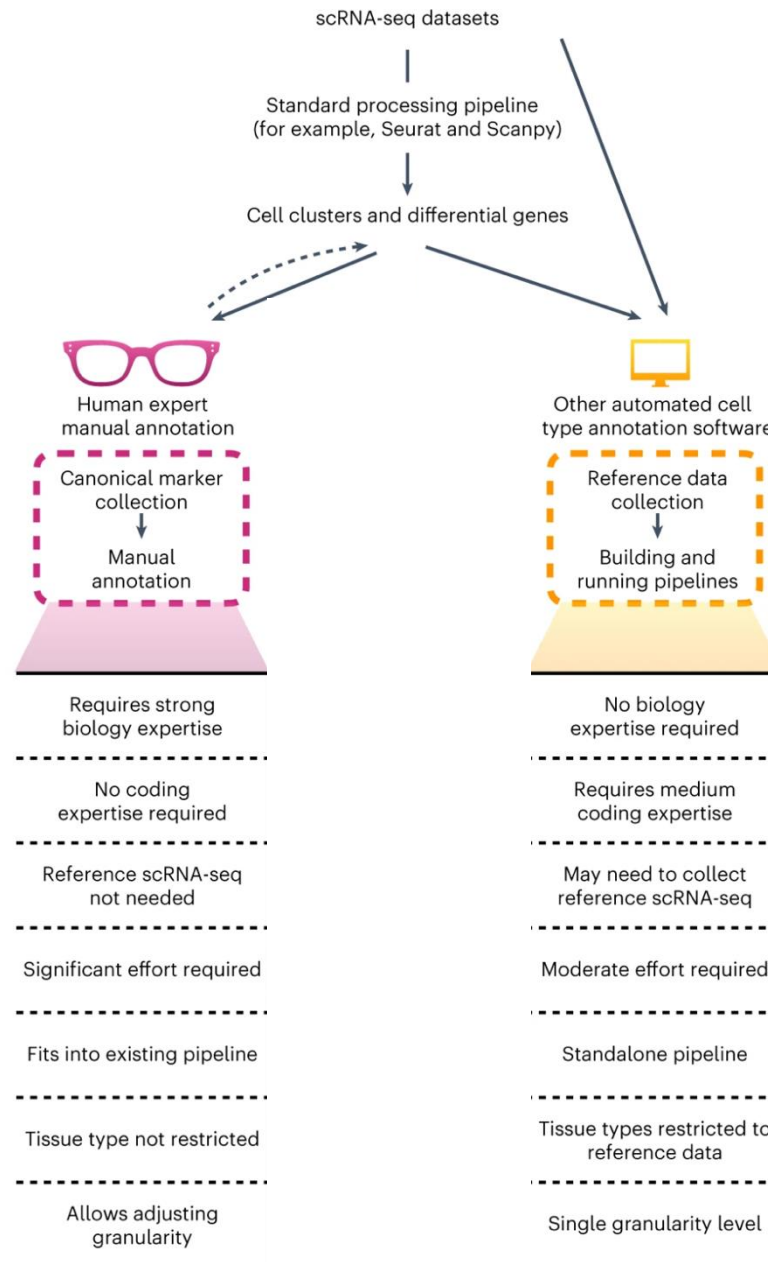
Arm and cl13 = 2 different strains of LCM virus

Charmoy et al 2021 <https://doi.org/10.1016/j.celrep.2021.109672>

Methods available

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 $\geq 10,000$ 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

Manual vs Automated cell type annotation



Will AI replace bioinformaticians?



Will AI replace bioinformaticians?



No, but a bioinformatician who uses AI will

AI application in Cell type identification

nature methods



Brief Communication

<https://doi.org/10.1038/s41592-024-02235-4>

Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis

Received: 16 April 2023

Accepted: 5 March 2024

Published online: 25 March 2024

 Check for updates

Wenpin Hou ¹✉ & Zhicheng Ji ²✉

Here we demonstrate that the large language model GPT-4 can accurately annotate cell types using marker gene information in single-cell RNA sequencing analysis. When evaluated across hundreds of tissue and cell types, GPT-4 generates cell type annotations exhibiting strong concordance with manual annotations. This capability can considerably reduce the effort and expertise required for cell type annotation. Additionally, we have developed an R software package GPTCelltype for GPT-4's automated cell type annotation.

Examples

Chat 1



Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name.

CR2, CD24, FAS, CXCR3, CD1c
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROX1, LYVE1
TPSAB1, FCER1A, TPSB2, KIT, CD69, HDC
ACTA2, MYO1B, ACTA2, ANPEP, DES, MCAM, PDGFRB, CSPG4



1. Dendritic cells
2. Luminal epithelial cells
3. Lymphatic endothelial cells
4. Mast cells
5. Pericytes



Be more specific



1. Myeloid dendritic cells (mDCs)
2. Prostate secretory luminal epithelial cells
3. Prostate lymphatic endothelial cells
- 4 . Prostate tissue-resident mast cells
5. Prostate vascular pericytes

Chat 2



Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name. Some could be a mixture of multiple cell types. Some could be unknown cell types.

KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROX1, LYVE1
CD69, IL7R,, CD3D, CD3E, CD3G,ACTA2, MYO1B, ACTA2, ANPEP, PDGFRB, CSPG4
DDX49,LOC105371196,MTND1P30,LOC105373682,TAGLN2,ZNF836,ZNF677,COILP1



1. Prostate epithelial cells
2. Lymphatic endothelial cells
3. T cell and smooth muscle cell mixture
4. Unknown cell type

All that glitters is not gold



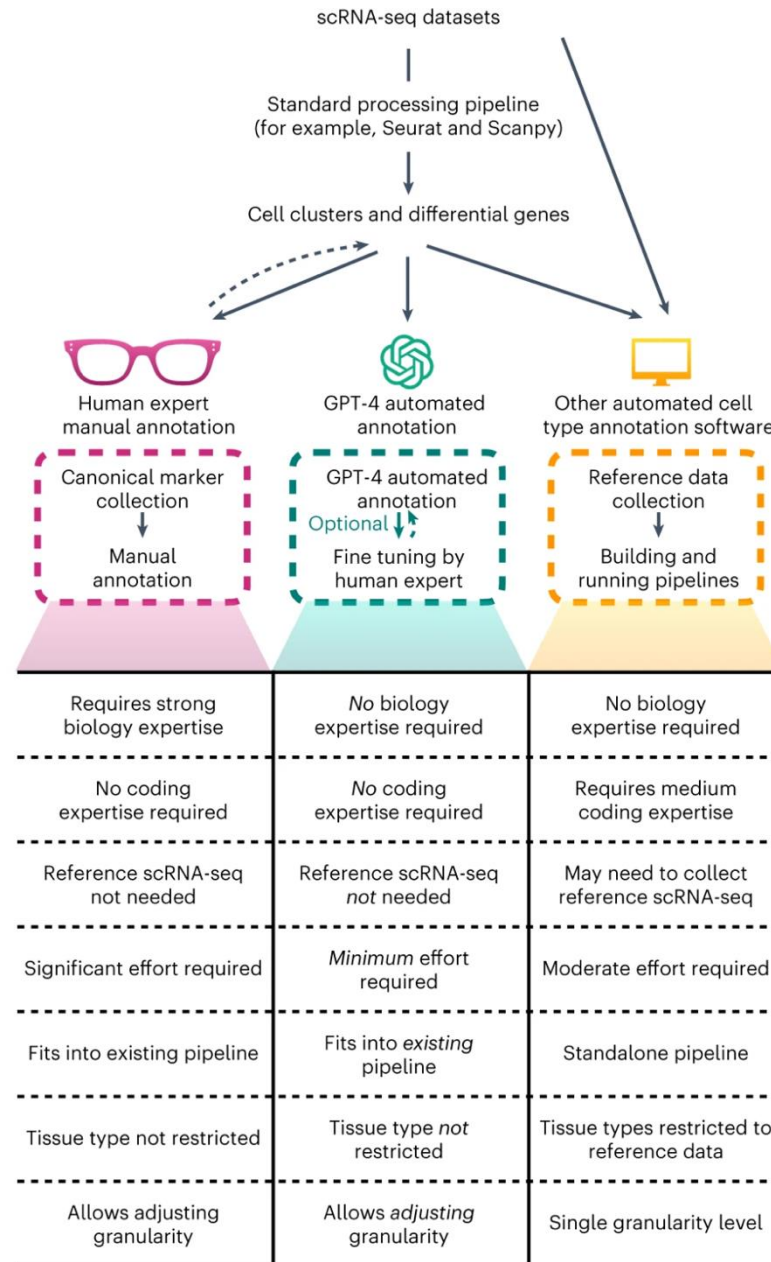
AI-generated cell annotation requiring human evaluation to ensure annotation quality and reliability

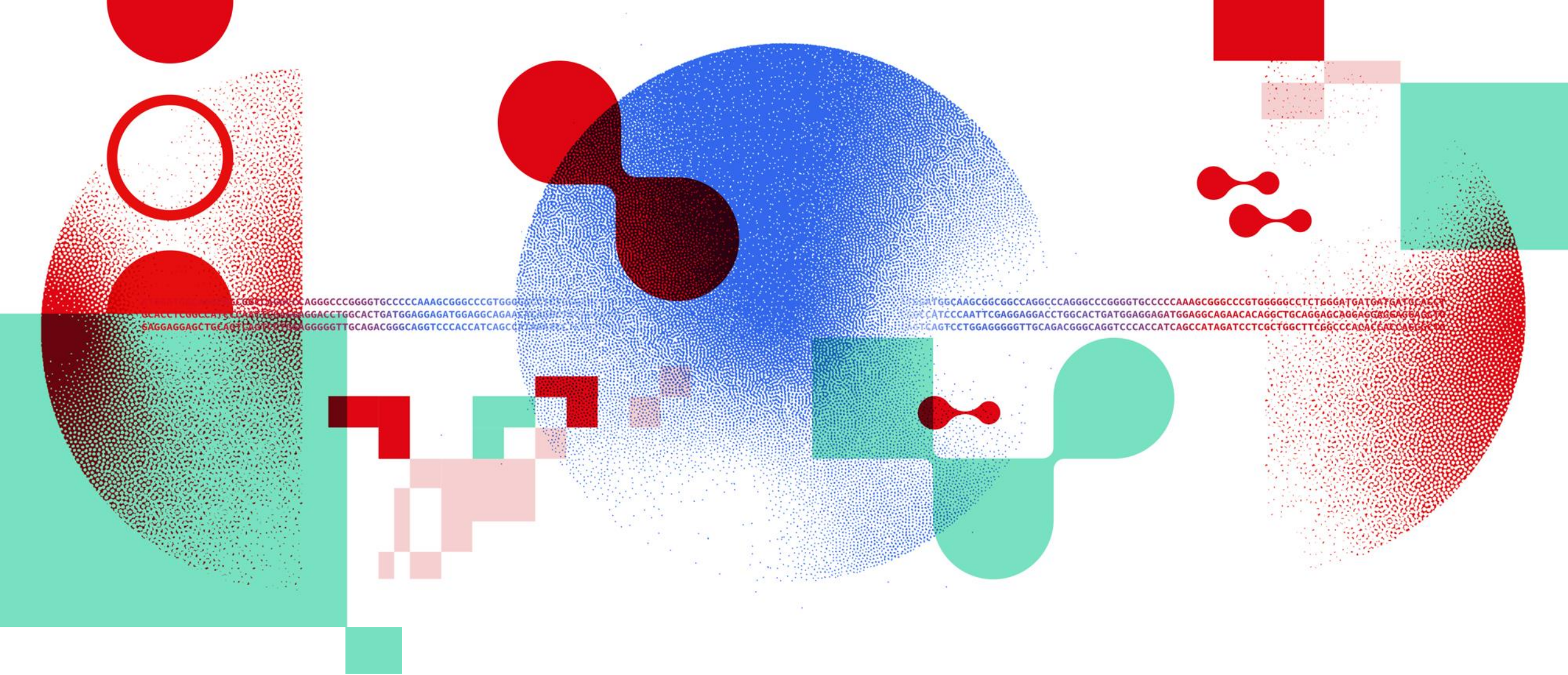
Human involvement in the optional fine-tuning of the model may affect reproducibility due to subjectivity and could limit the scalability of the model in large datasets

High noise levels in scRNA-seq data and unreliable differential genes can adversely affect GPT-4's annotations

GPT-4 risks artificial intelligence hallucination.

AI vs Manual vs Automated cell type annotation





Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss