

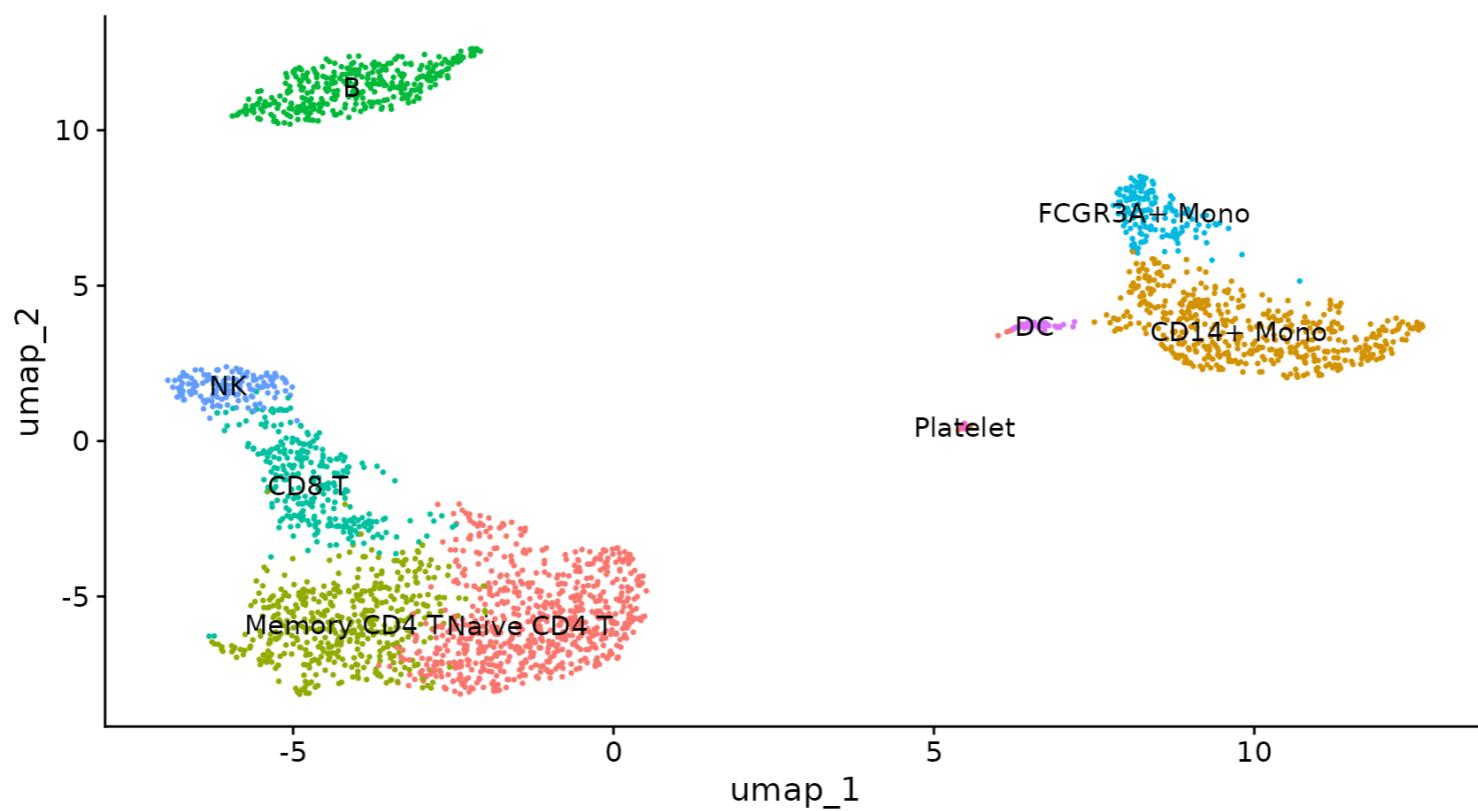
Guidelines for annotating single-cell transcriptomic datasets

Seattle Area Single Cell (SASC) user group meeting

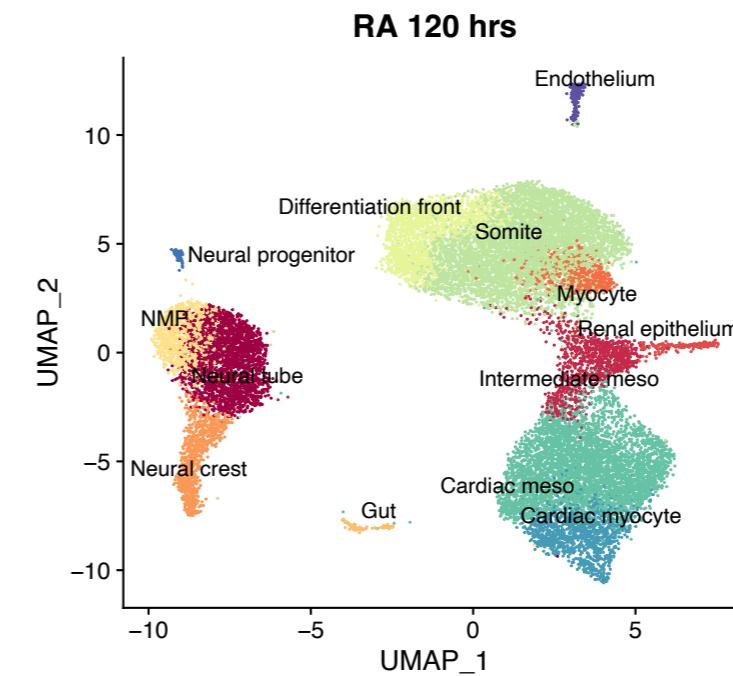
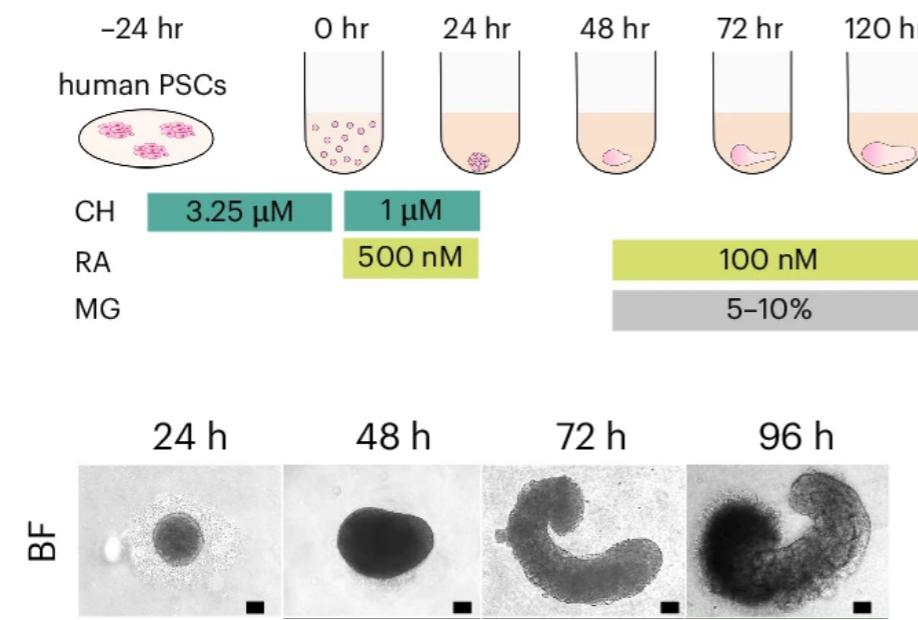
Wei Yang (UW GS/Shendure lab)

2024-09-27

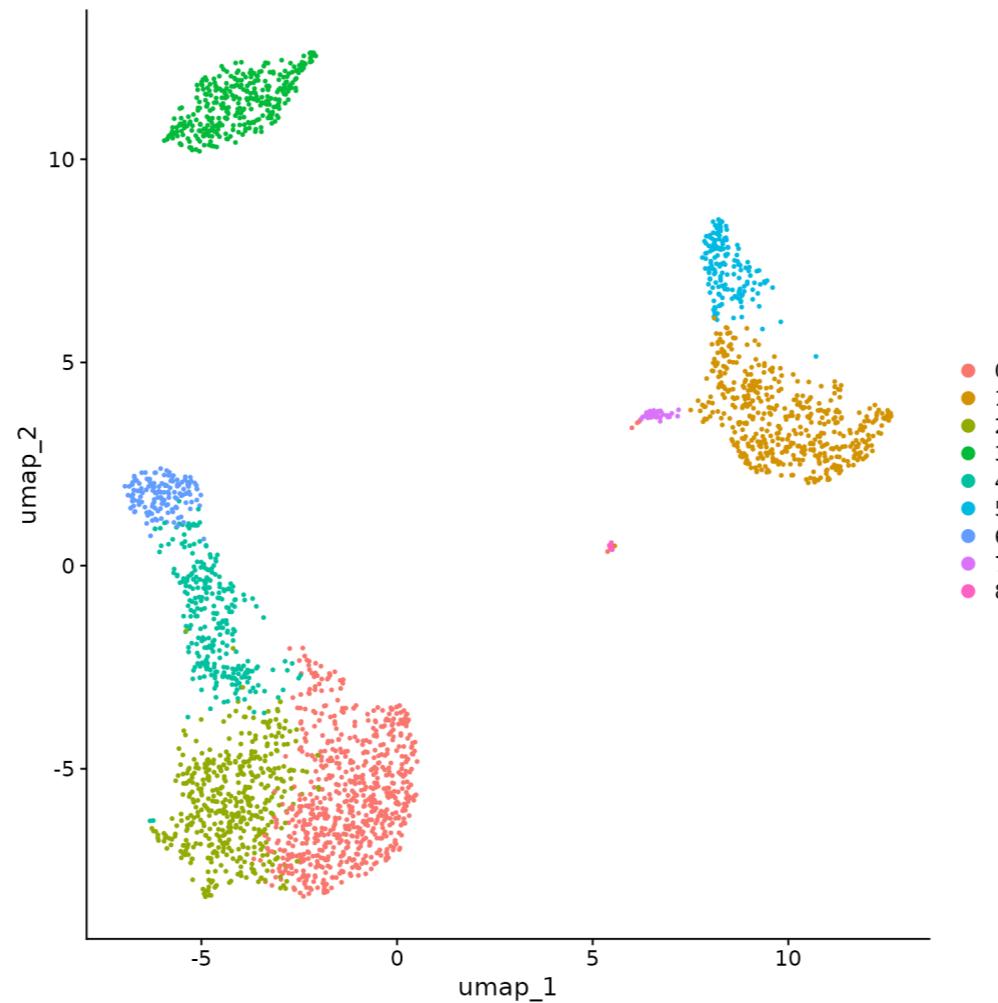
Dataset 1 - PBMC3k



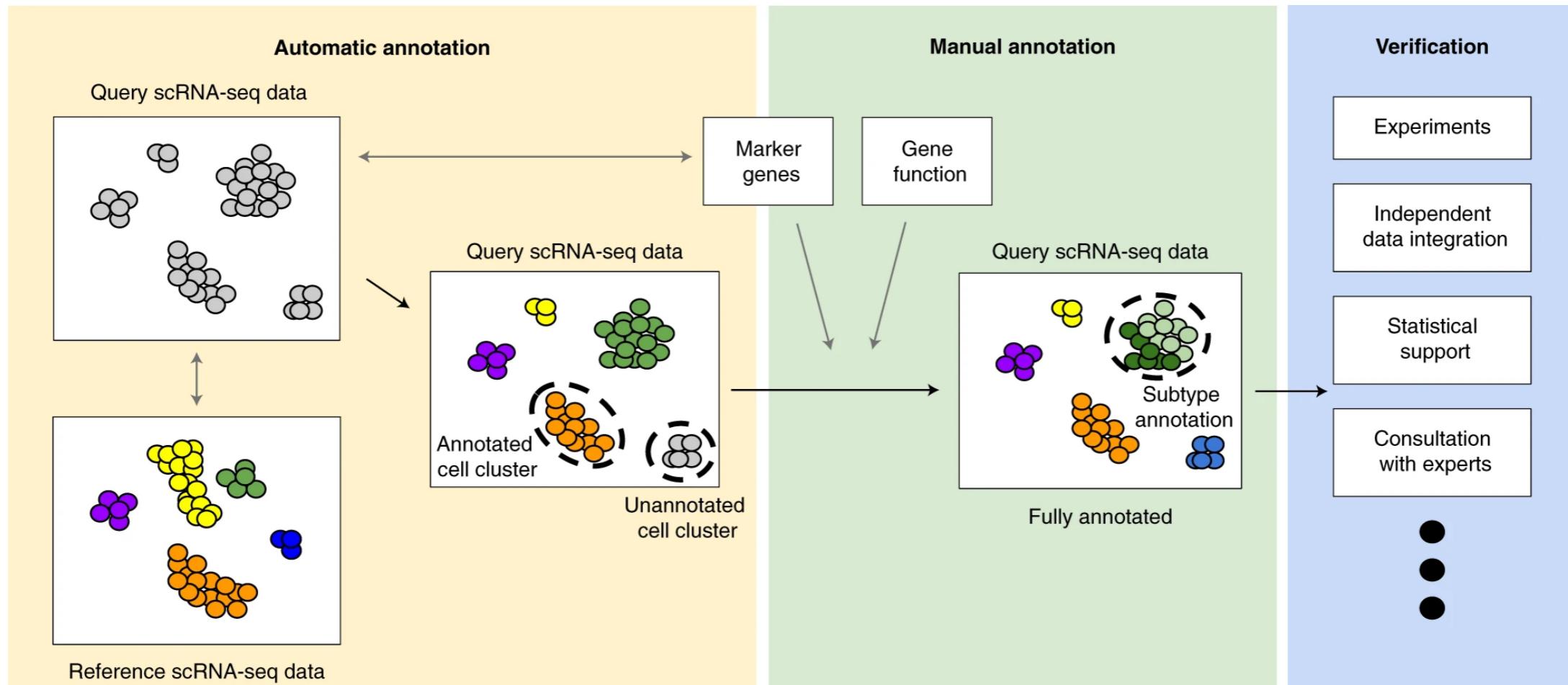
Dataset 2 - Human RA gastruloids



To start with cell type annotation



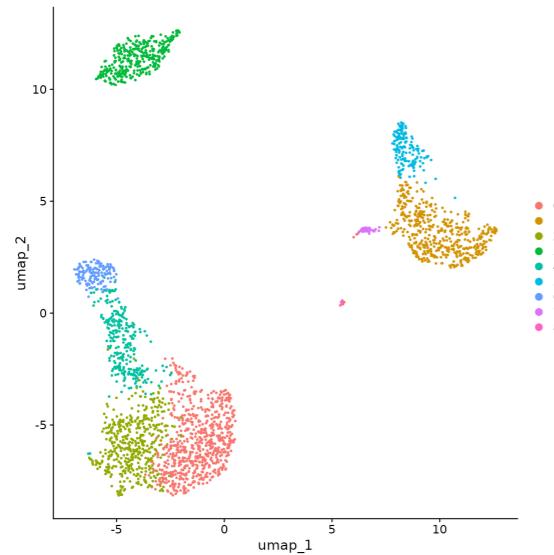
Workflow of cell type annotation



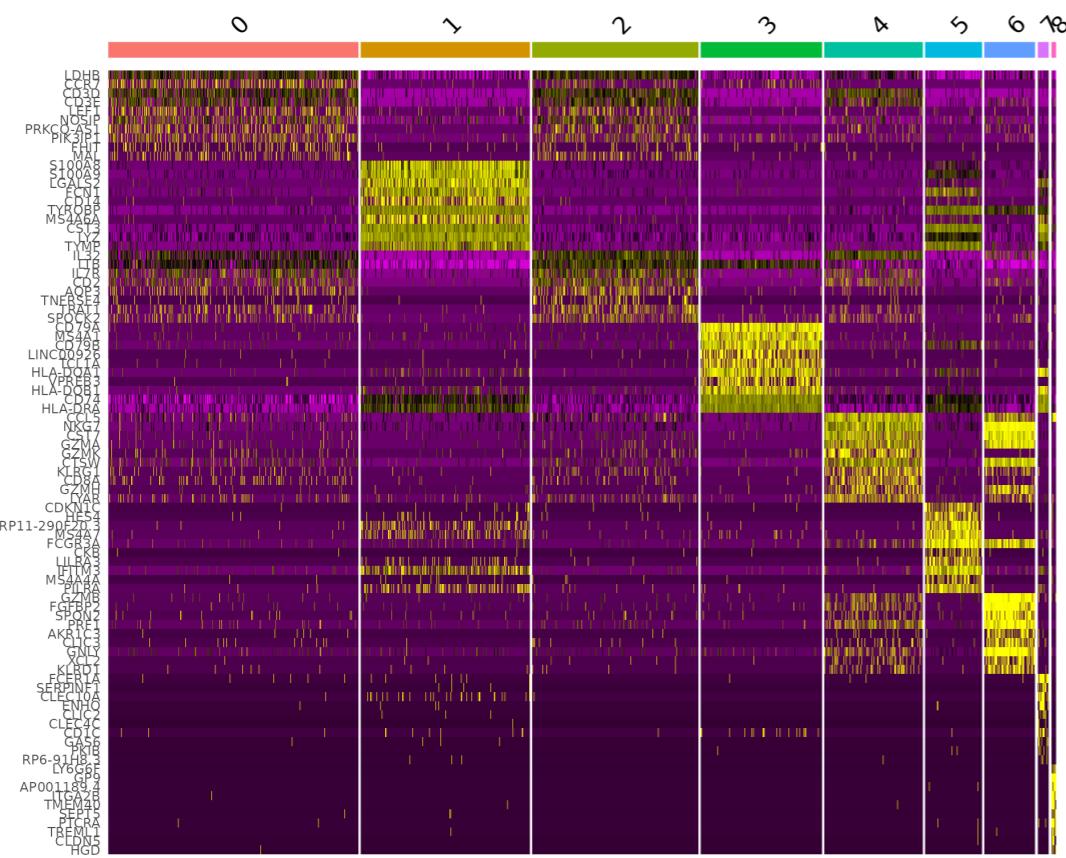
- Integration and label transfer
- NNLS
- Web tools
- Marker genes
- Gene modules
- Functional enrichment

Manual annotation - marker genes

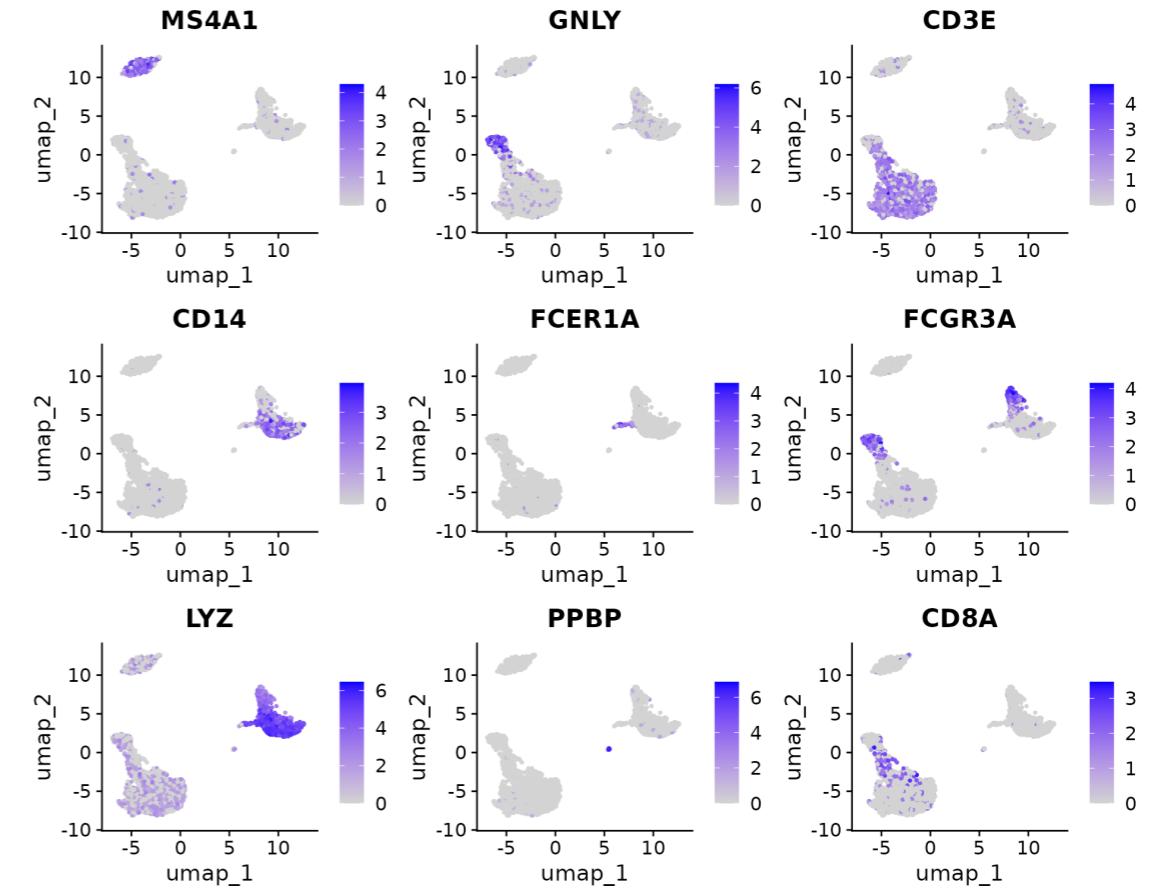
PBMC



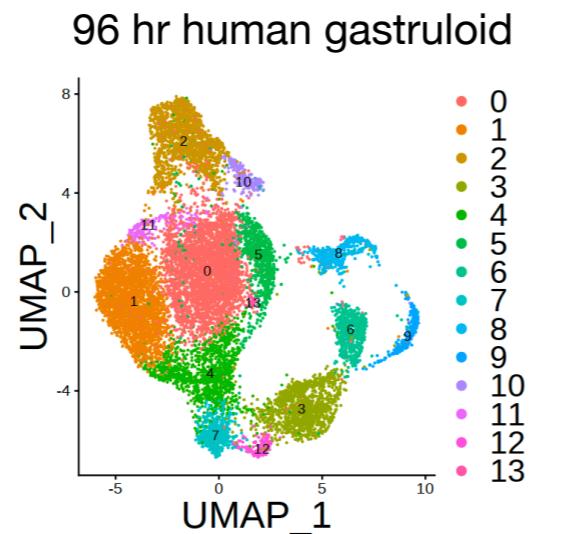
Data-driven



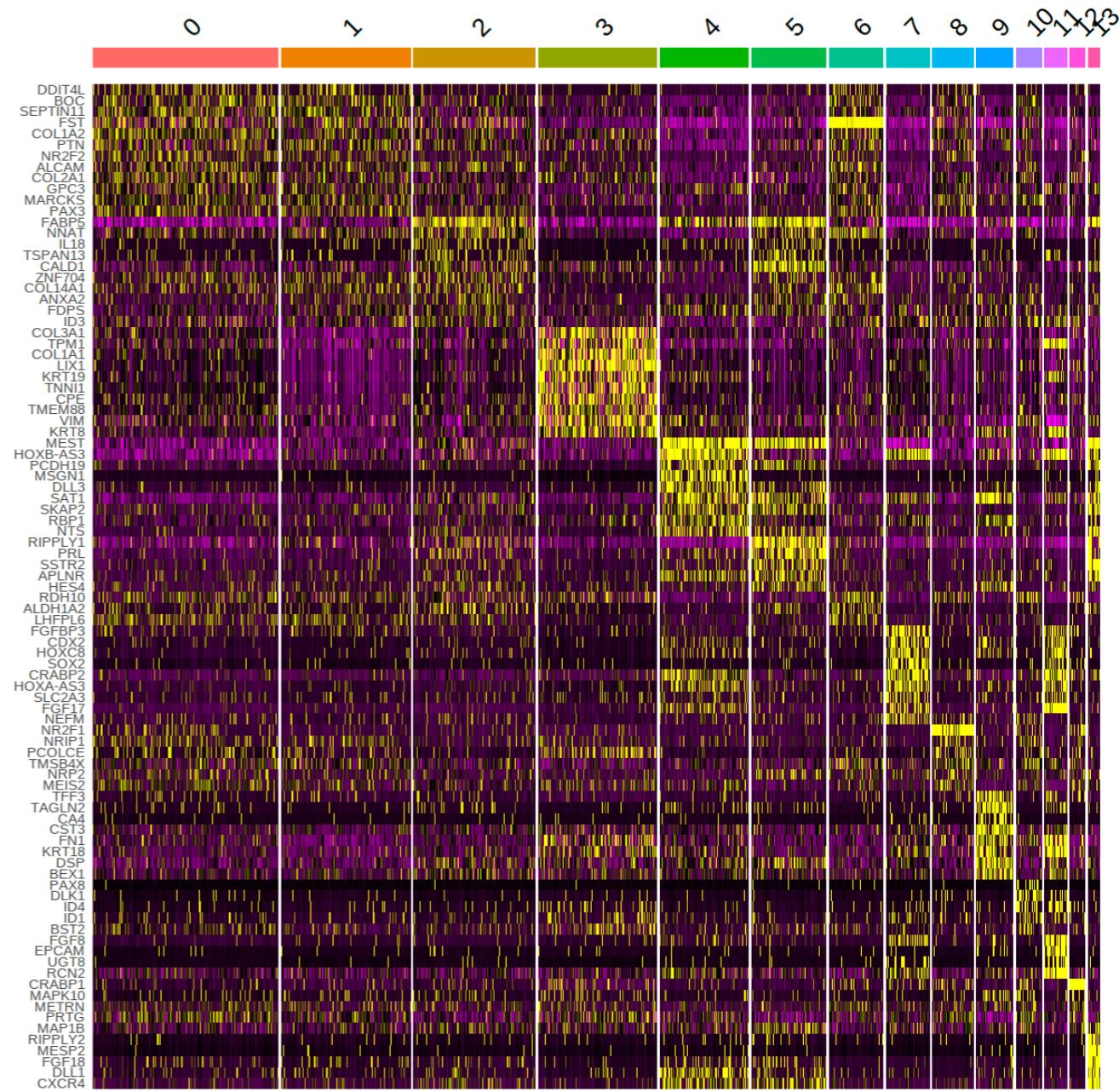
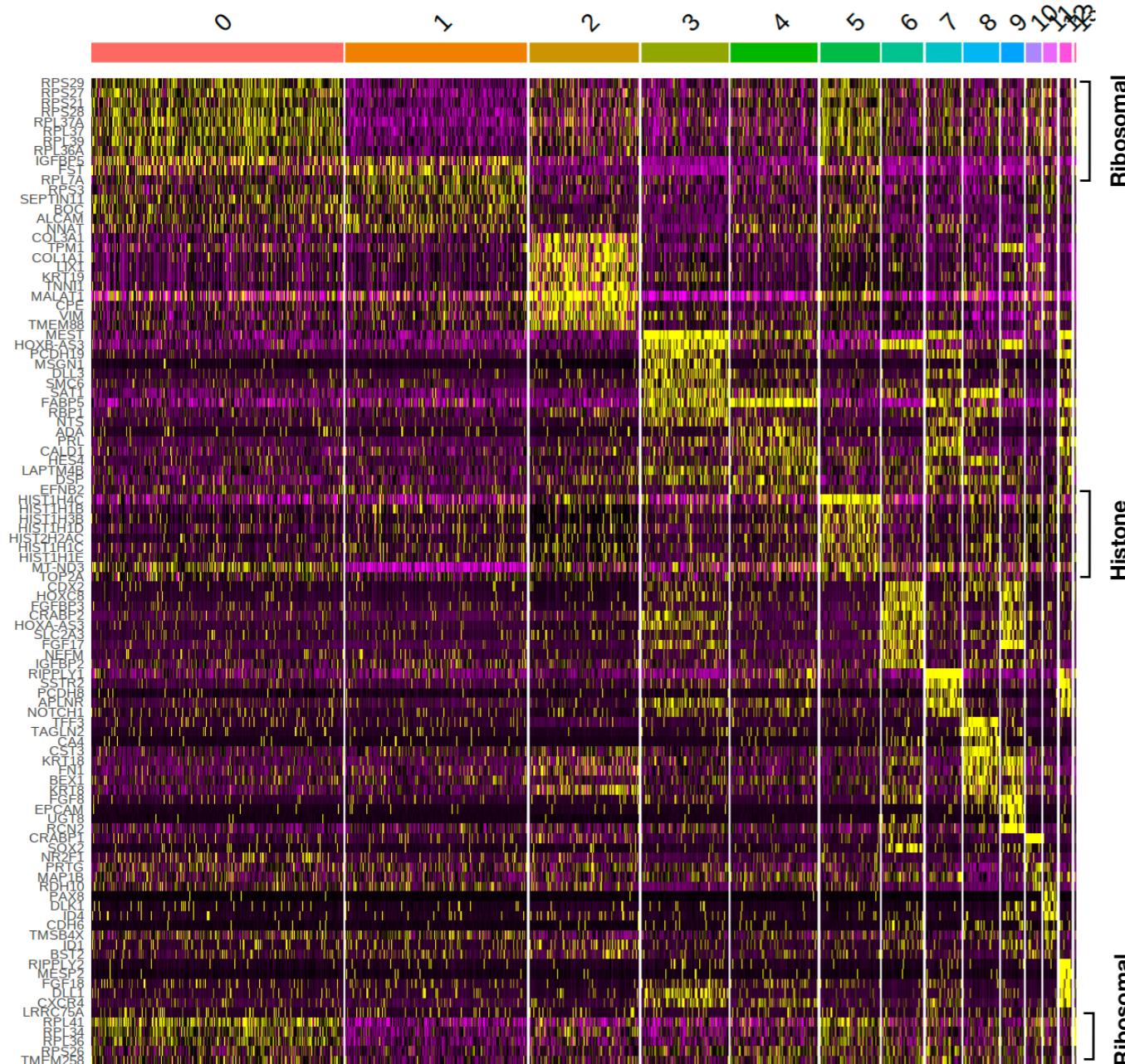
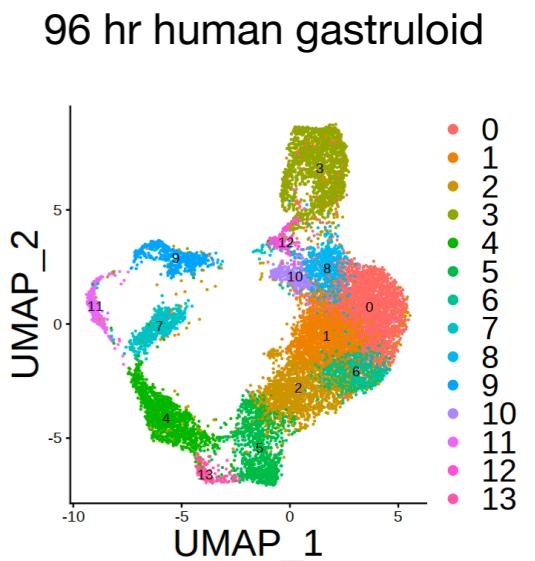
Literature



Before filtration of technical genes

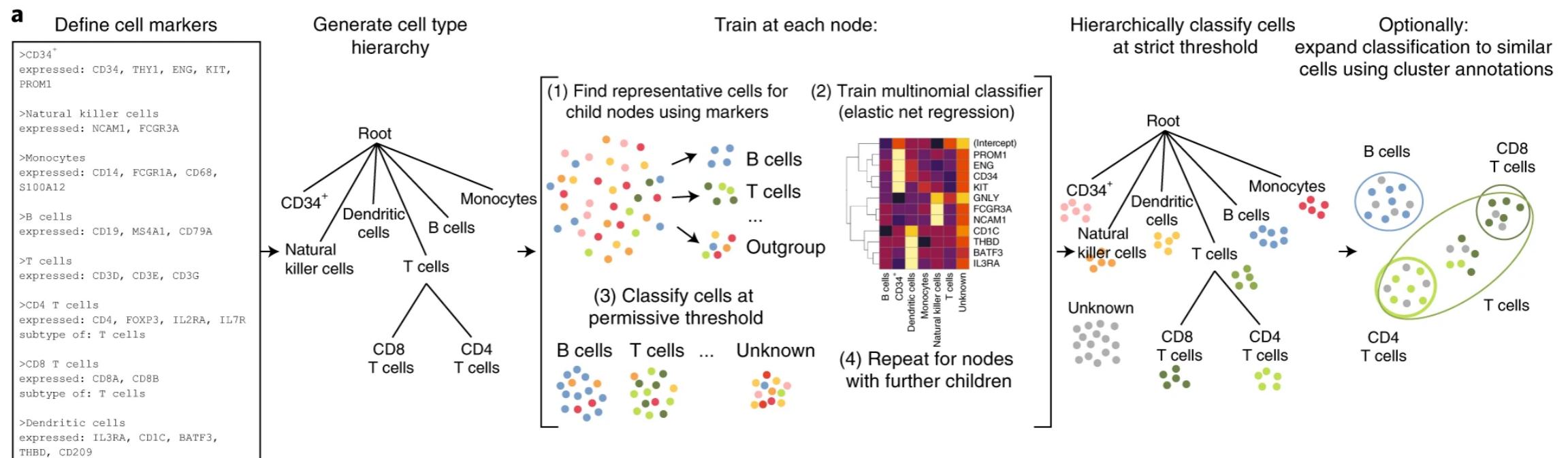


After filtration of batch technical genes



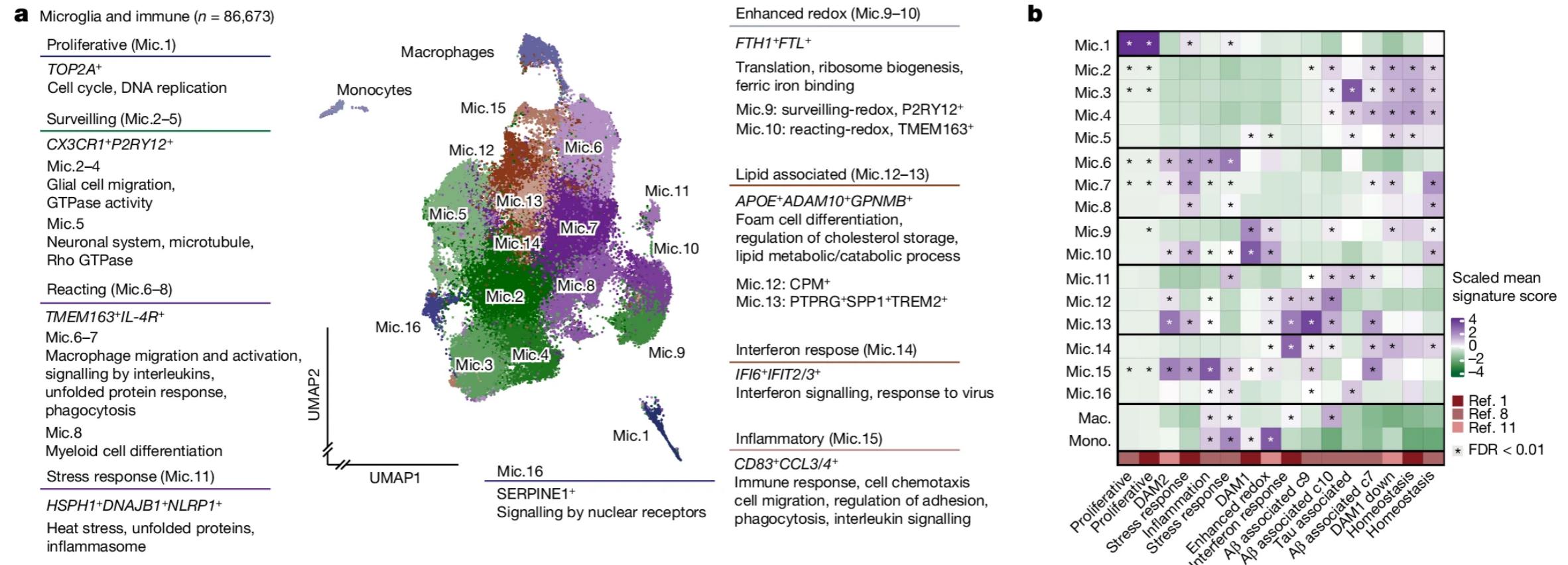
Manual annotation - marker genes

- Garnett trained a classifier with a set of marker genes for each cell type for fast annotation based on marker gene expression.



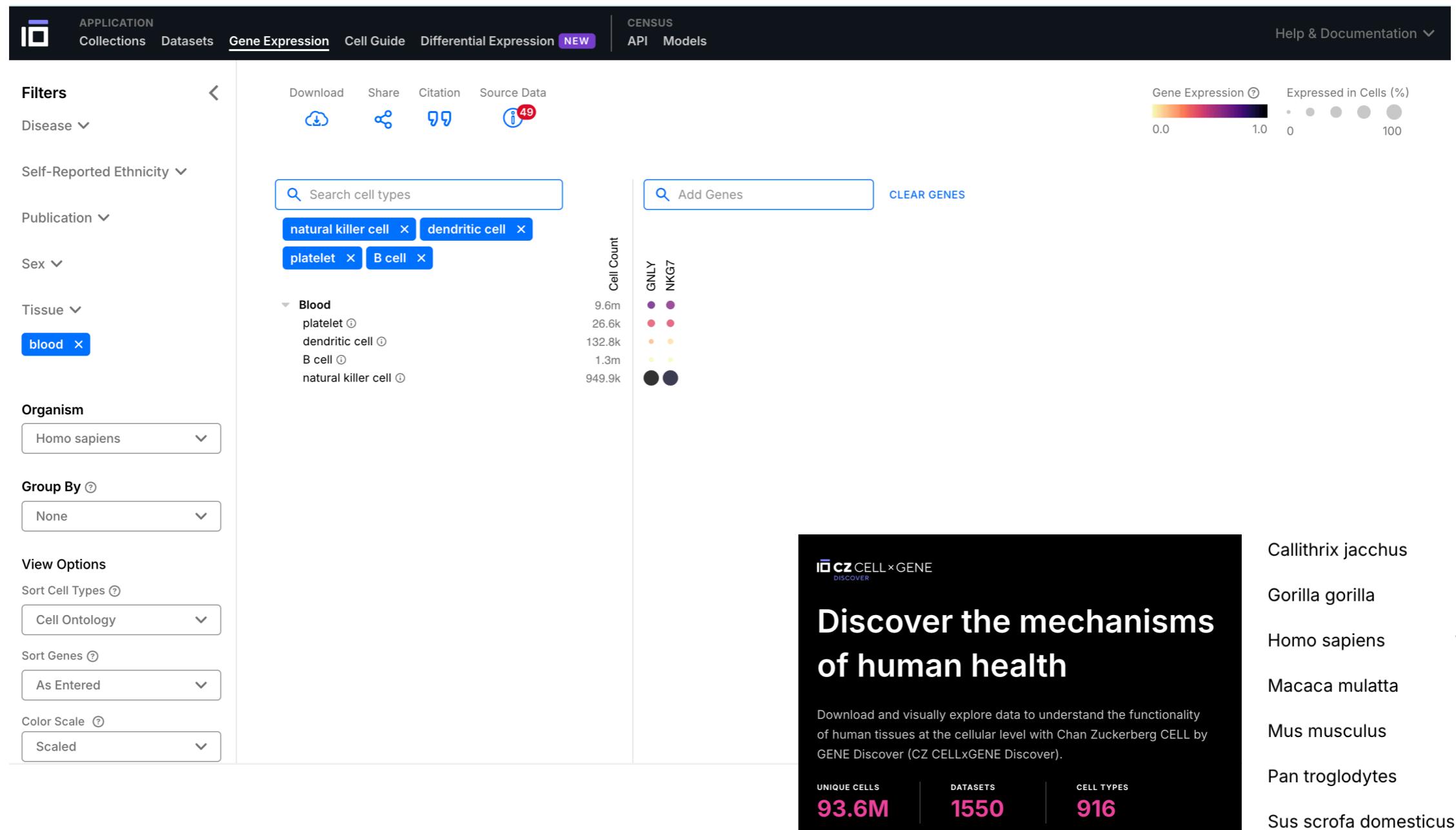
Manual annotation - gene modules and functional enrichment

- In de novo annotation, gene modules and GO terms are useful to define functions of a new cell subtype.



Manual annotation - web tools

<https://cellxgene.cziscience.com/gene-expression>



Manual annotation - web tools

<https://panglaodb.se/index.html>

PanglaoDB Home Search Datasets Tools Papers FAQ/Help About

Filter

Show cell type: NK cells (98)

get tsv file add marker

Database statistics

	Mus musculus	Homo sapiens
Samples	1063	305
Tissues	184	74
Cells	4,459,768	1,126,580
Clusters	8,651	1,748

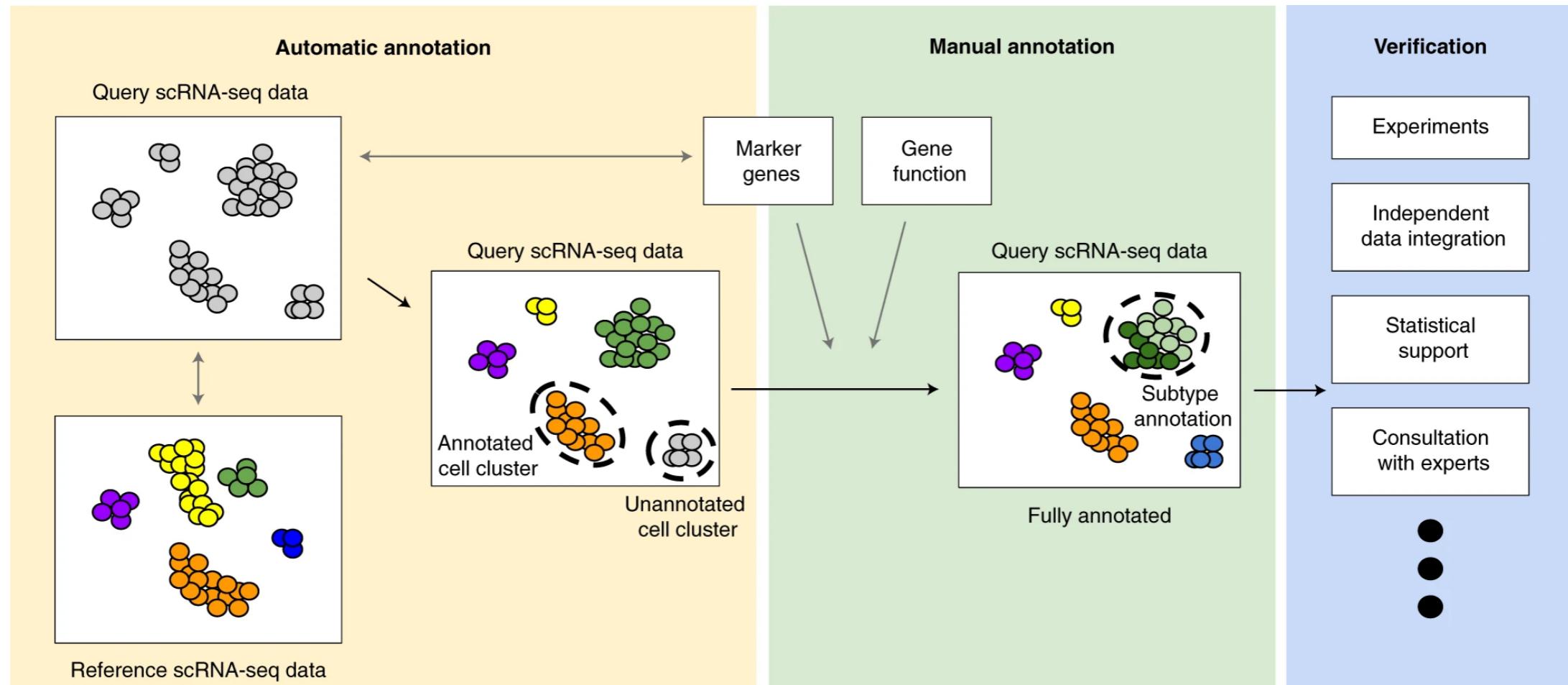
Gene expression markers for NK cells

Vote(s)	Species	Official gene symbol	UI	Sensitivity (human)	Sensitivity (mouse)	Specificity (human)	Specificity (mouse)	Marker count	Cell type	Germ layer	Organ	Aliases	Product description	Disease	Action
19	Mm Hs	TRDC	0.011	0.351	NA	0.018	NA	4	NK cells	Mesoderm	Immune system		T cell receptor delta constant		flag
8	Mm Hs	NKG7	0.056	0.946	1	0.088	0.04	3	NK cells	Mesoderm	Immune system	GMP-17	natural killer cell granule protein 7		flag
5	Hs	KLRF1	0.003	0.257	NA	0.015	NA	2	NK cells	Mesoderm	Immune system	CLEC5C,NKp80	killer cell lectin like receptor F1		flag
3	Mm Hs	KLRD1	0.02	0.595	0.915	0.026	0.011	3	NK cells	Mesoderm	Immune system	CD94	killer cell lectin like receptor D1		flag
3	Hs	GNLY	0.013	0.851	NA	0.064	NA	2	NK cells	Mesoderm	Immune system	LAG-2,D2S69E,TLA51,LAG9,LAG2	granulysin		flag
3	Mm Hs	NCR1	0.004	NA	0.798	NA	0	1	NK cells	Mesoderm	Immune system	NK-p46,NKP46,CD335,LY94	natural cytotoxicity triggering receptor 1		flag
2	Mm Hs	GZMA	0.026	0.905	0.787	0.046	0.014	3	NK cells	Mesoderm	Immune system	HFSP,CTLA3	granzyme A		flag
1	Mm Hs	HOPX	0.033	0.514	NA	0.082	0.022	10	NK cells	Mesoderm	Immune system	LAGY,OB1,NECC1,SMAP31	HOP homeobox		flag
1	Mm Hs	ITGAM	0.025	NA	0.011	0.003	0.03	8	NK cells	Mesoderm	Immune system	CD11b,CR3A,CD11B	integrin subunit alpha M	Y	flag
1	Mm Hs	TGFB1	0.11	0.122	NA	0.043	NA	3	NK cells	Mesoderm	Immune system	CED,TGFbeta,TGFb,DPD1	transforming growth factor beta 1	Y	flag
1	Mm Hs	GZMB	0.017	0.514	0.67	0.05	0.004	7	NK cells	Mesoderm	Immune system	CCPI,CGL-1,CSP-B,CGL1,CTSGL1,SECT,CSPB	granzyme B		flag
1	Mm	KLRE1	0.005	NA	0.862	NA	0.001	1	NK cells	Mesoderm	Immune system		killer cell lectin-like receptor family E member 1		flag

Manual annotation - summary

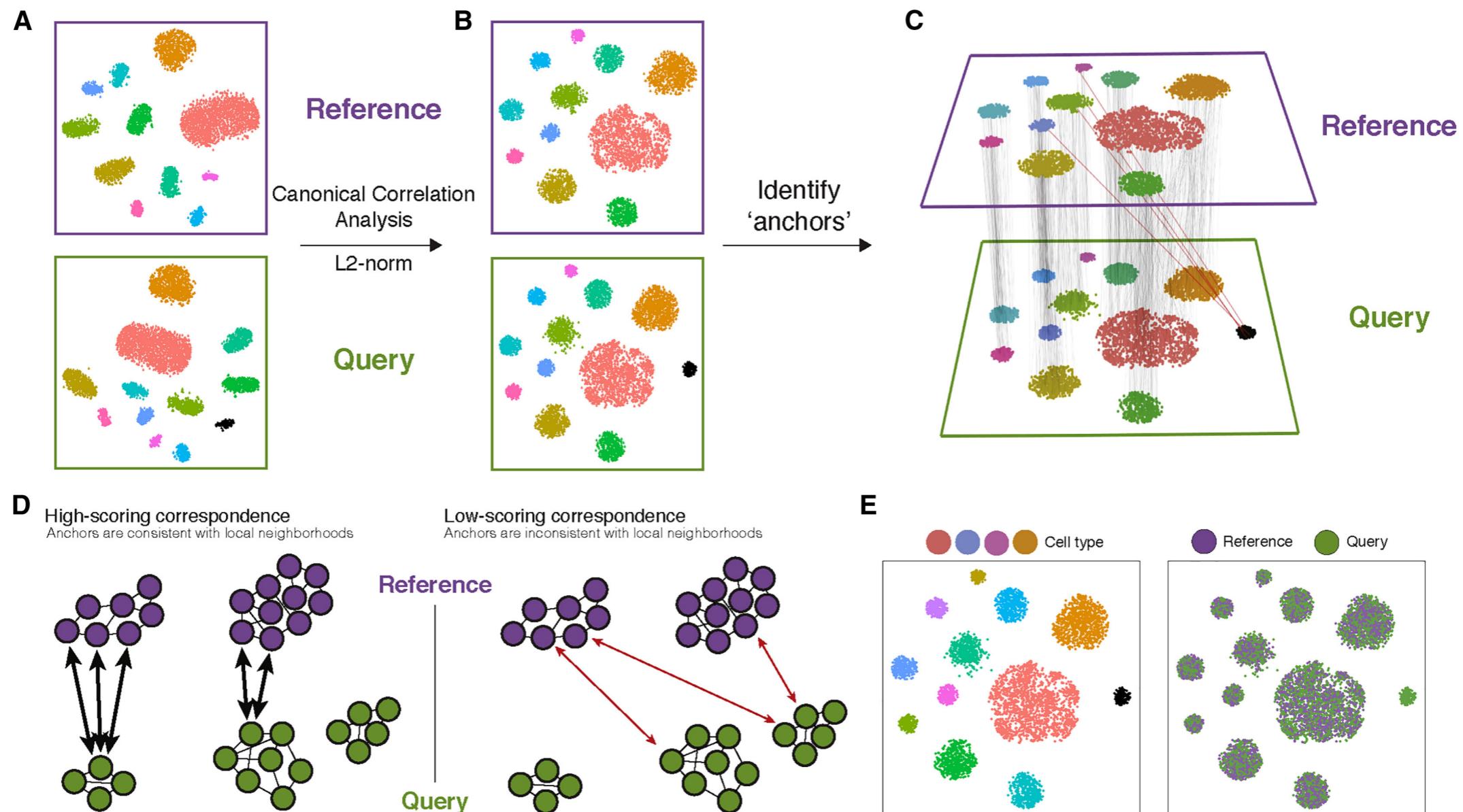
- In manual annotation, we are looking at marker genes that could distinguish a cluster from the rest.
- Marker genes can be identified in a data-driven way. This is very useful for de novo discovery of de novo annotation of cell types/states. However, some marker genes could be introduced by technical factors (e.g. ribosomal and histone).
- Marker genes can also be identified from literature. However, the expression level of such marker genes might be different in different studies (i.e. profiling methods, organisms, locations).
- Web tools are available to explore gene expression in a large number of cell types. The annotation of these webtools stay at a broad level.

Workflow of cell annotation



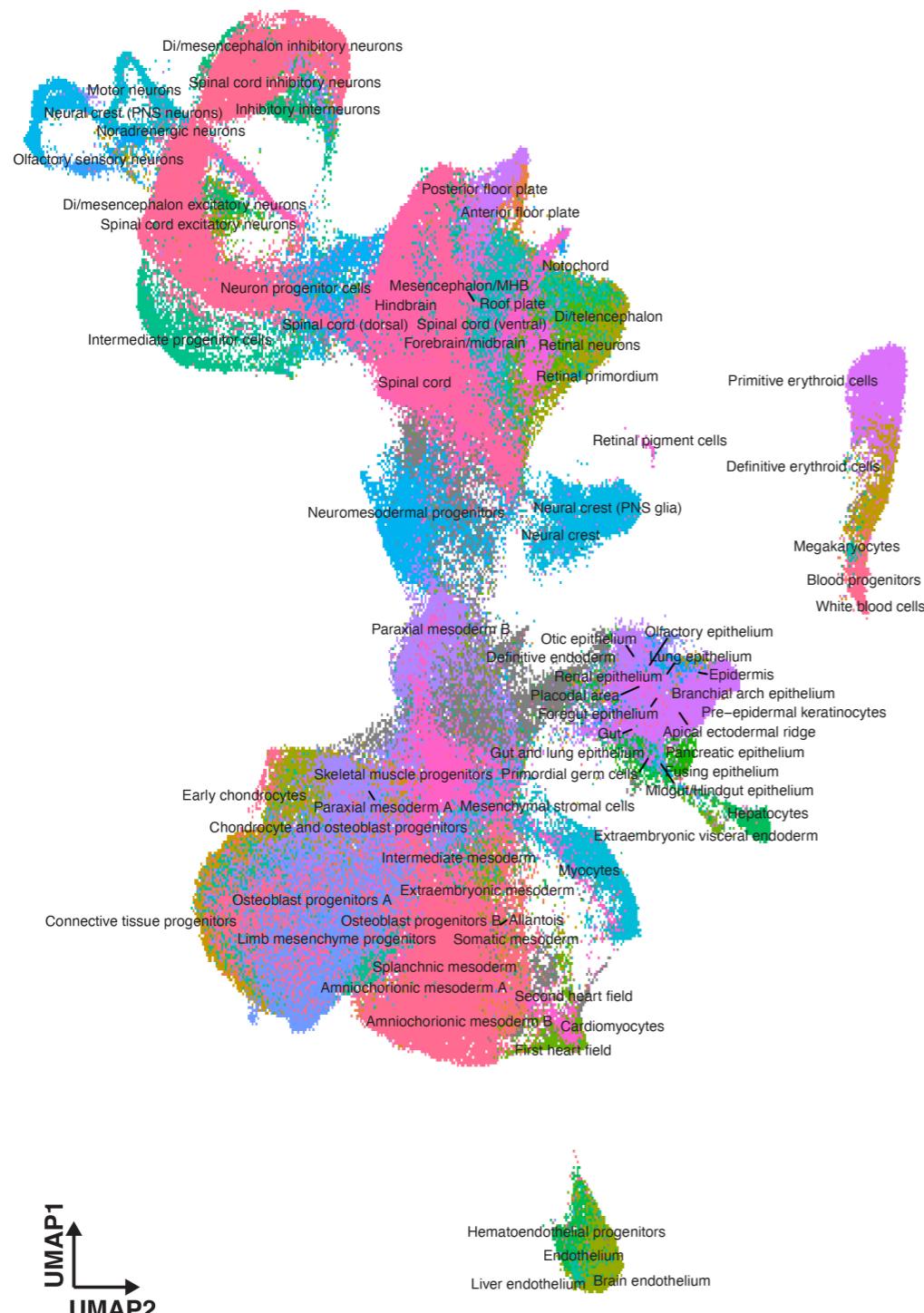
- Integration and label transfer
- NNLS
- Web tools
- Marker genes
- Gene modules
- GO terms

Automatic annotation - integration and label transfer

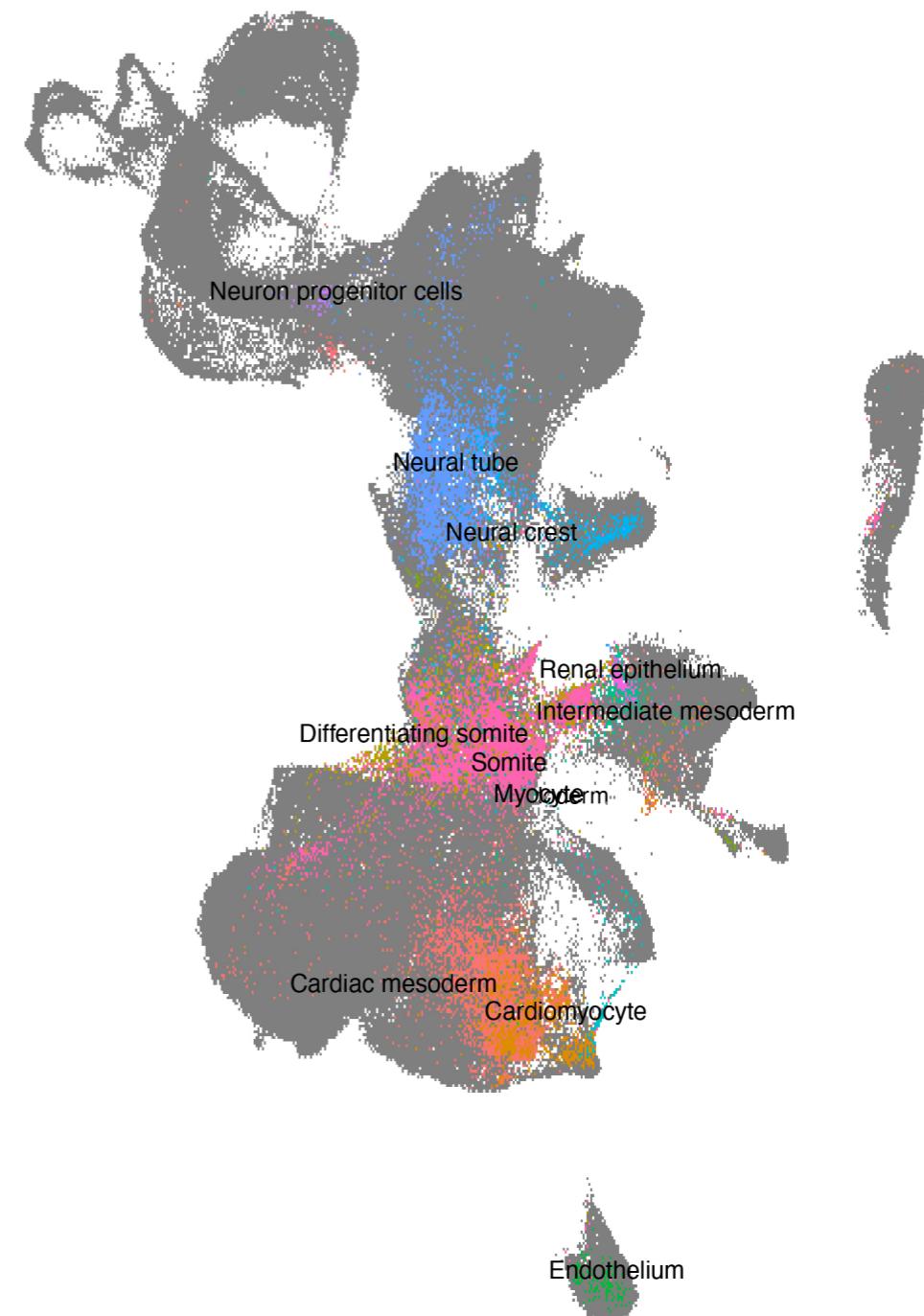


Automatic annotation - integration and label transfer

Mouse embryo E8.5-13.5

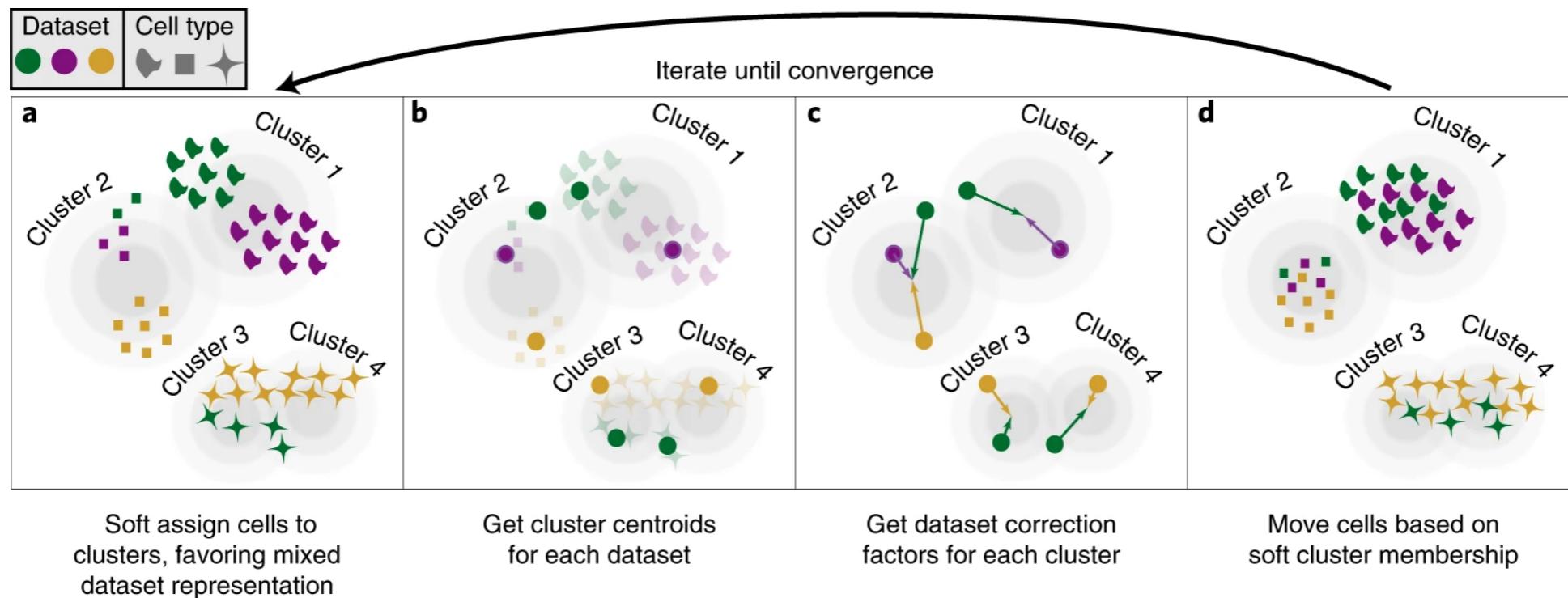


Human RA gastruloids (120 hrs)



Automatic annotation - integration and label transfer

- Harmony: much faster than MNN approach but work poorly for cross-species integration.



Soft assign cells to clusters, favoring mixed dataset representation

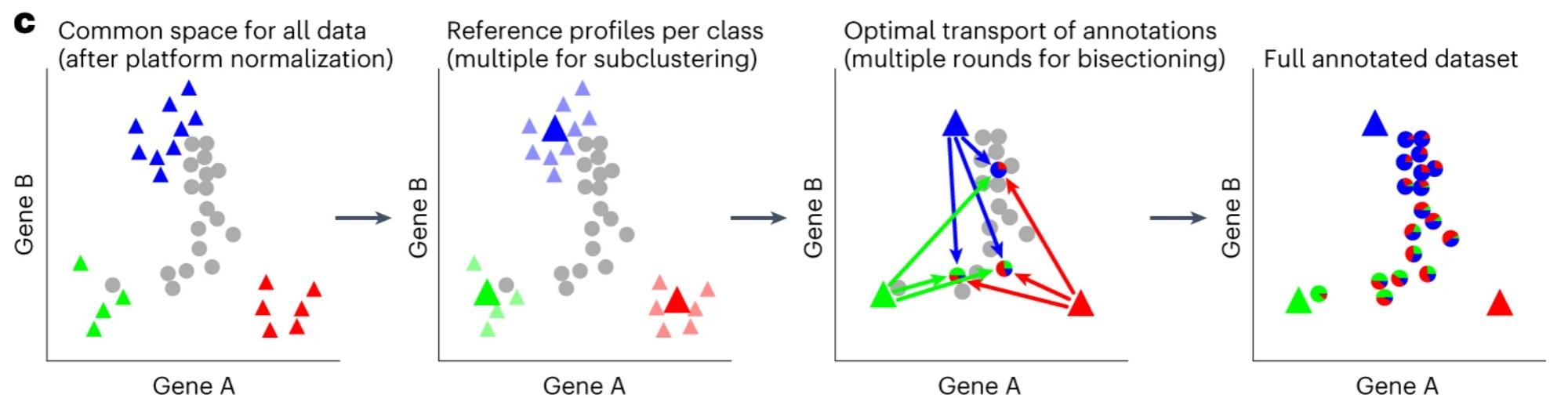
Get cluster centroids for each dataset

Get dataset correction factors for each cluster

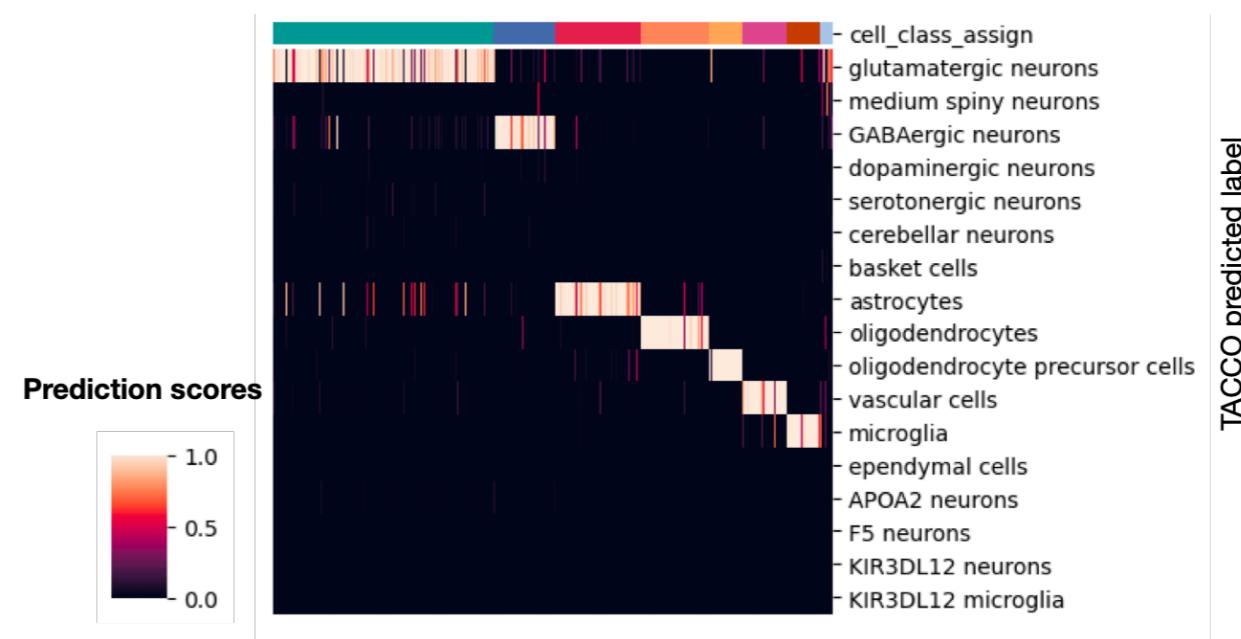
Move cells based on soft cluster membership

Automatic annotation - integration and label transfer

- TACCO: implemented in Python, much faster than MNN approach, works well for both same-species and cross-species integration.



Example data



Automatic annotation - integration and label transfer

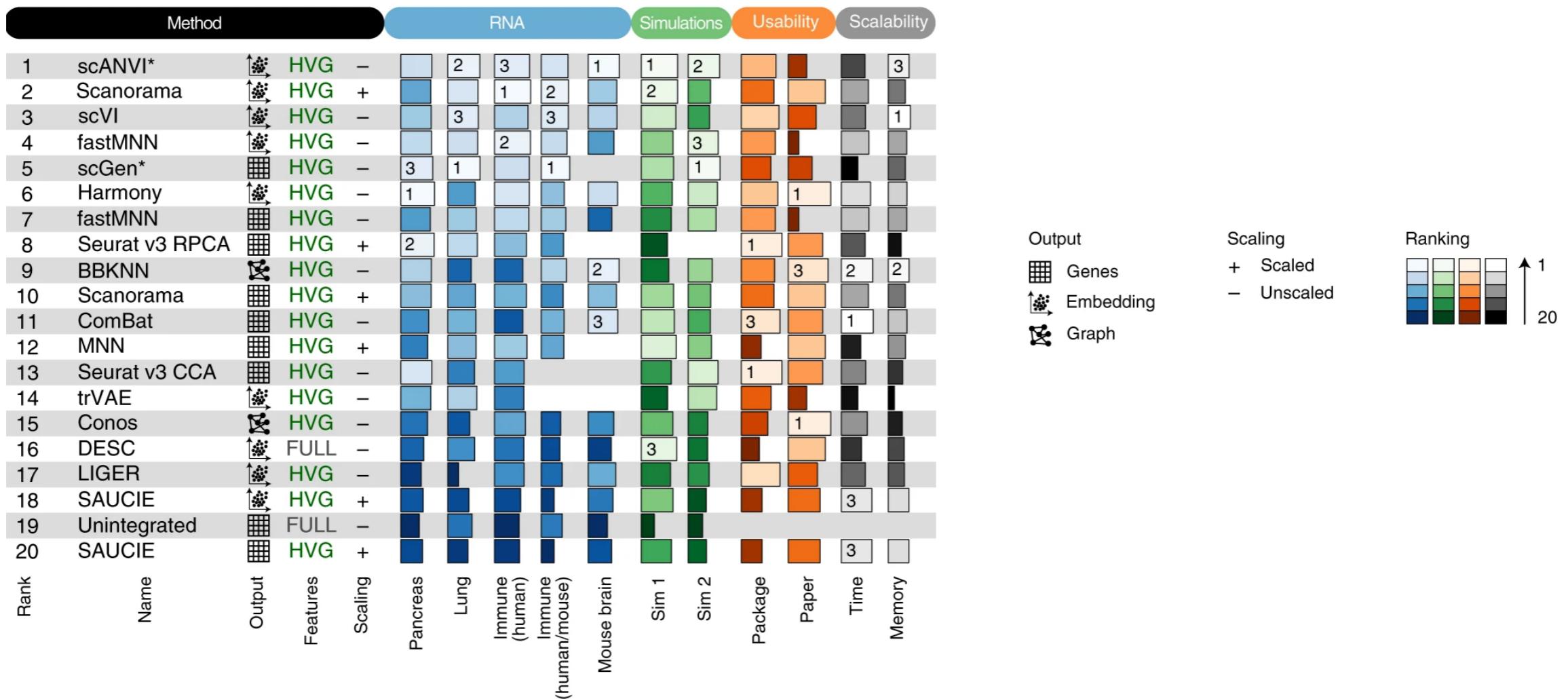
Analysis | [Open access](#) | Published: 23 December 2021

- Other algorithms

Benchmarking atlas-level data integration in single-cell genomics

[Malte D. Luecken](#), [M. Büttner](#), [K. Chaichoompu](#), [A. Danese](#), [M. Interlandi](#), [M. F. Mueller](#), [D. C. Strobl](#), [L. Zappia](#), [M. Dugas](#), [M. Colomé-Tatché](#) & [Fabian J. Theis](#)

[Nature Methods](#) **19**, 41–50 (2022) | [Cite this article](#)



Automatic annotation - NNLS

- NNLS is an optimization problem of a linear model with non-negative constraints on the coefficients.
- The coefficient in the linear model indicates how similar gene expression of a given cell type in dataset b to the target cell type in dataset a.

$$T_a = \beta_0 a + \beta_1 a M_b$$

$$T_b = \beta_0 b + \beta_1 b M_a$$



$$\beta = 2(\beta_1 ab + 0.001)(\beta_1 ba + 0.001)$$

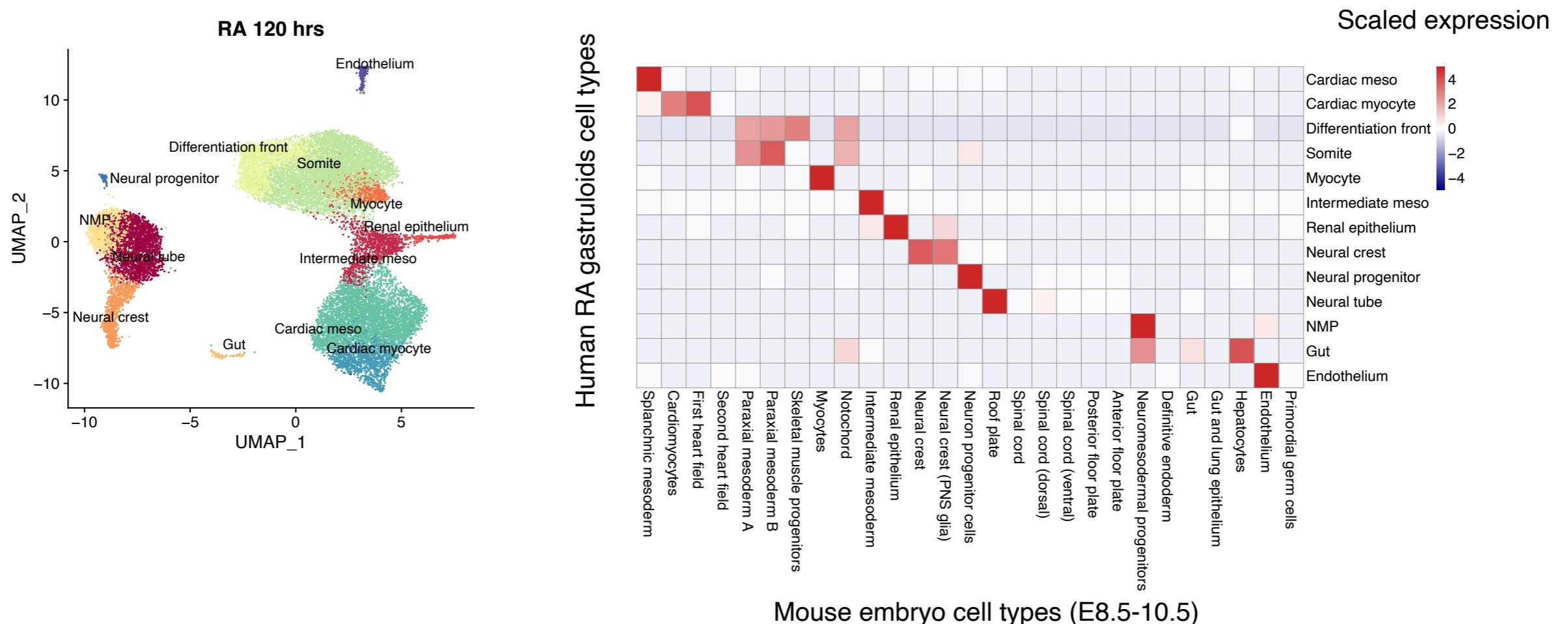
T_a/b : gene expression in target cell type

M_a/b : matrix of gene expression of all cell types

$\beta_0/1$: coefficient

Automatic annotation - NNLS

- Requires less computational resources because it works on pseudobulk data.



Automatic annotation - web tools

- Reference-based web tools.
- Minimal requirement for programming.
- Available to a limited set of organisms and organs.

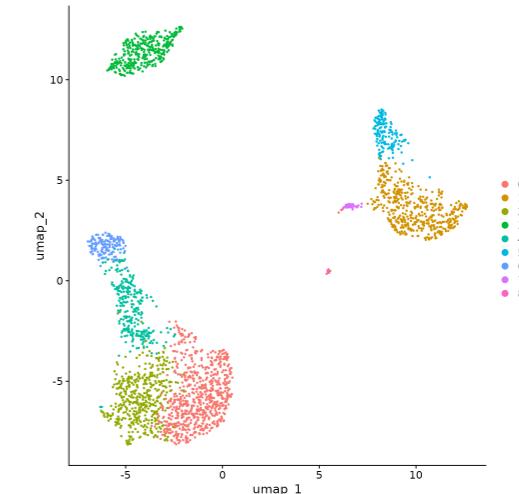
	Azimuth	Tabula Sapiens	MapmyCells
Species	Human(11) and mouse (1)	Human	Human and mouse
Organs	PBMC, Motor cortex, pancreas, fetal, kidney, bone marrow, lung, adipose, tonsil, heart, liver	24 organs	Whole brain and middle temporal gyrus
Assay	RNA(12) and ATAC(2)	RNA	RNA
Interface	Web portal, Seurat	Google collab	Web portal
Link	https://azimuth.hubmapconsortium.org/	https://tabula-sapiens.sf.czbiohub.org/annotateuserdata	https://knowledge.brain-map.org/mapmycells/process/

Automatic annotation - summary

- A well-annotated reference is required for automatic annotation.
- Many algorithms are available for integration and label transfer. The main consideration is computational resources (e.g. run time, memory).
- The quality of integration may be different between same-species and cross-species integration.
- Web-tools are less computational demanding options but have limited sets of reference for now.

Workshop

- PBMC
 - Performing normalization, dimensionality reduction and clustering.
 - Identifying data-driven marker genes for each pbmc cluster.
 - Visualizing enrichment of marker gene expression from literatures individually or as a gene module.
 - Exploring functional meaning of marker genes.
 - Integrating query to reference in Azimuth.



- Human RA gastruloids
 - Mapping cell types in gastruloids to mouse embryo with NNLS

