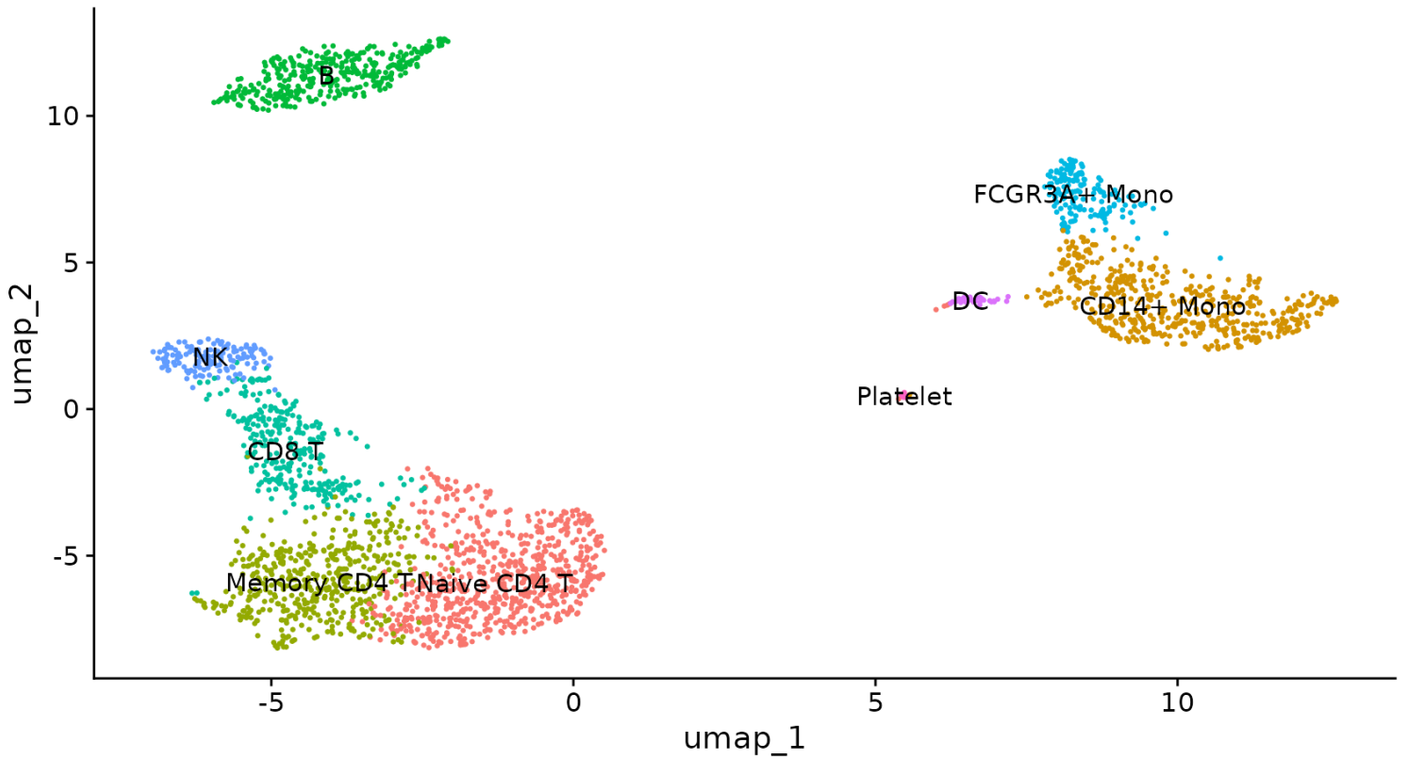
Guidelines for annotating single-cell transcriptiomic datasets

Seattle Area Single Cell (SASC) user group meeting

Wei Yang (UW GS/Shendure lab)

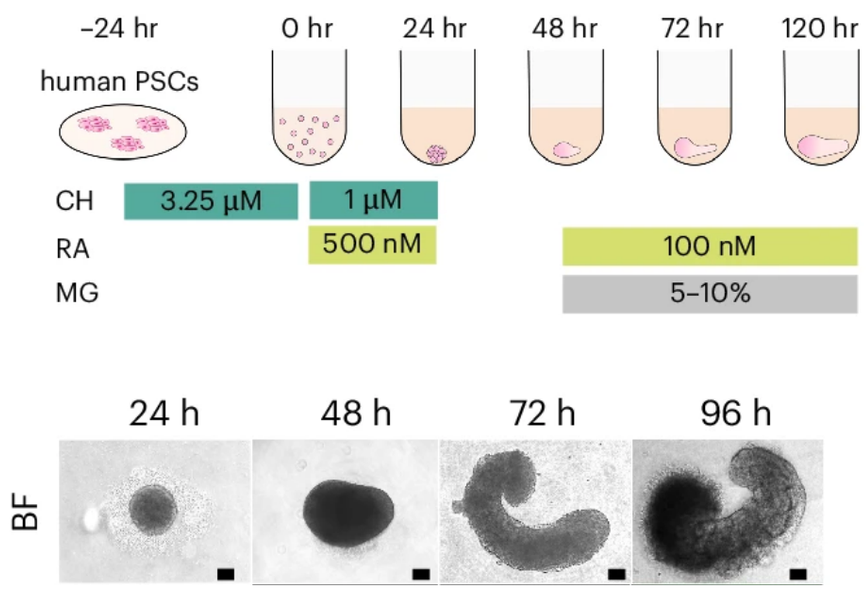
2024-09-27

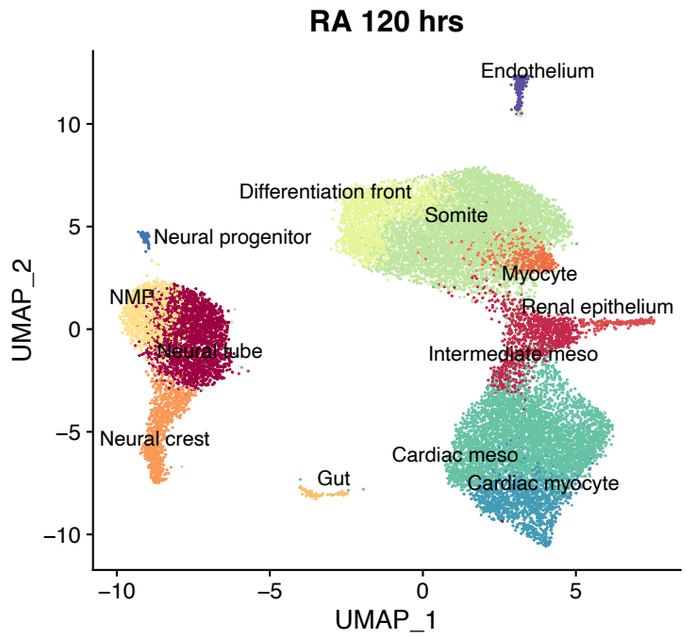
Dataset 1 - PBMC3k



Zheng et al., 2017

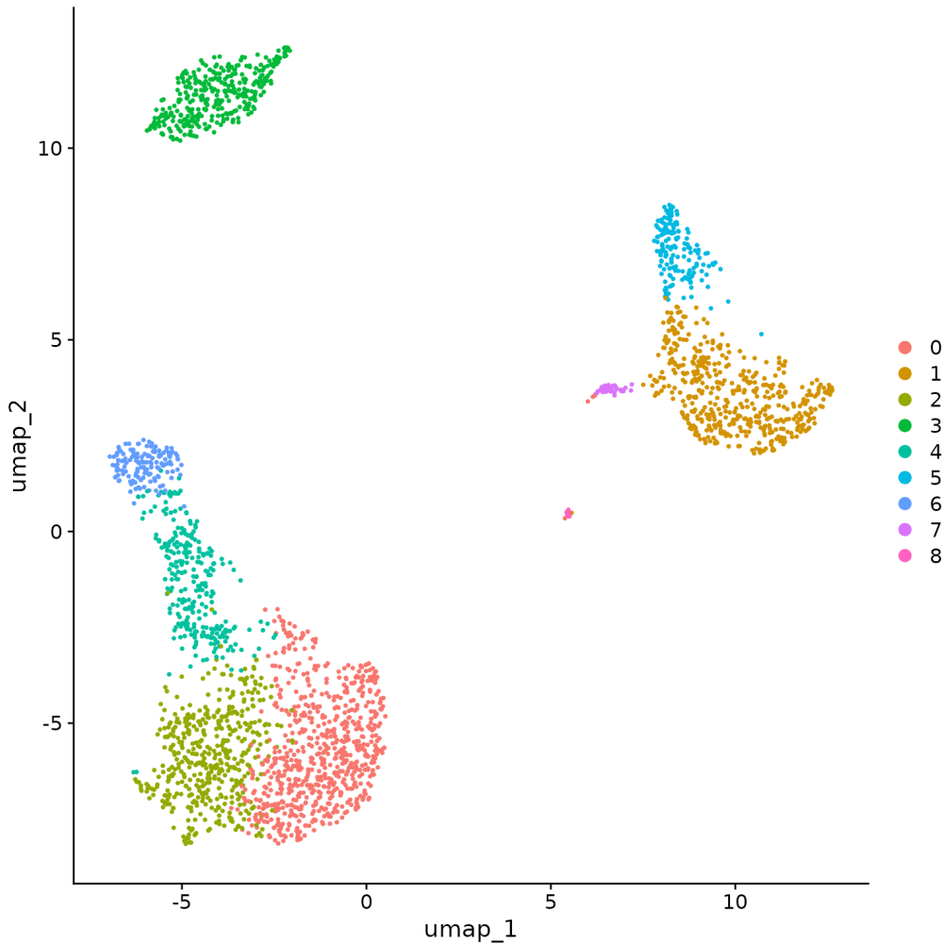
Dataset 2 - Human RA gastruloids





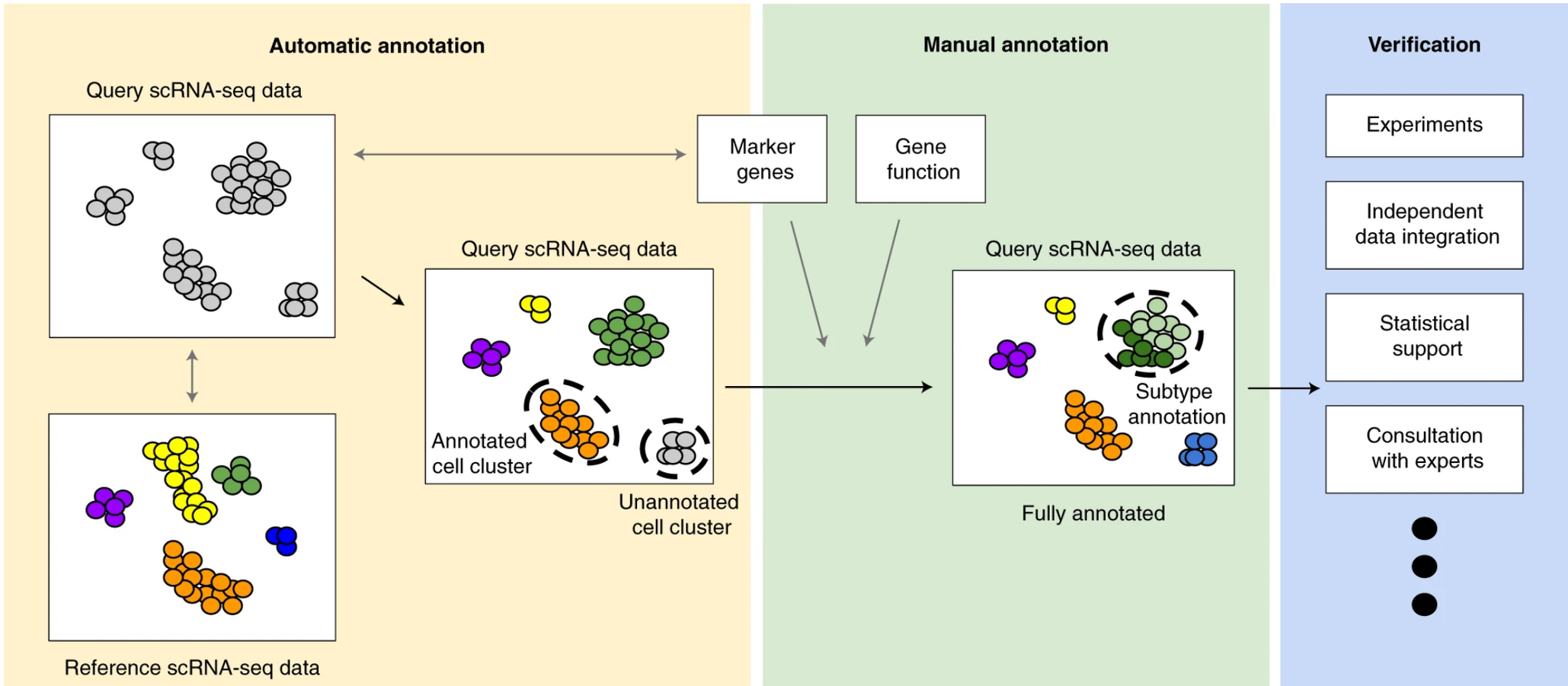
Hamazaki, Yang et al., 2024

To start with cell type annotation



Zheng et al., 2017

Workflow of cell type annotation

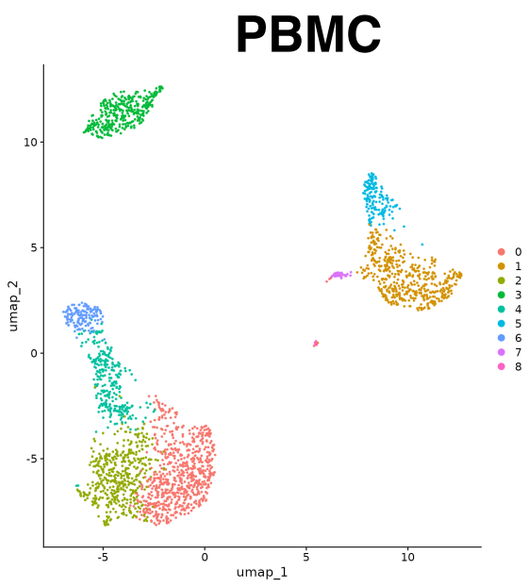


- Integration and label transfer
- NNLS
- Web tools

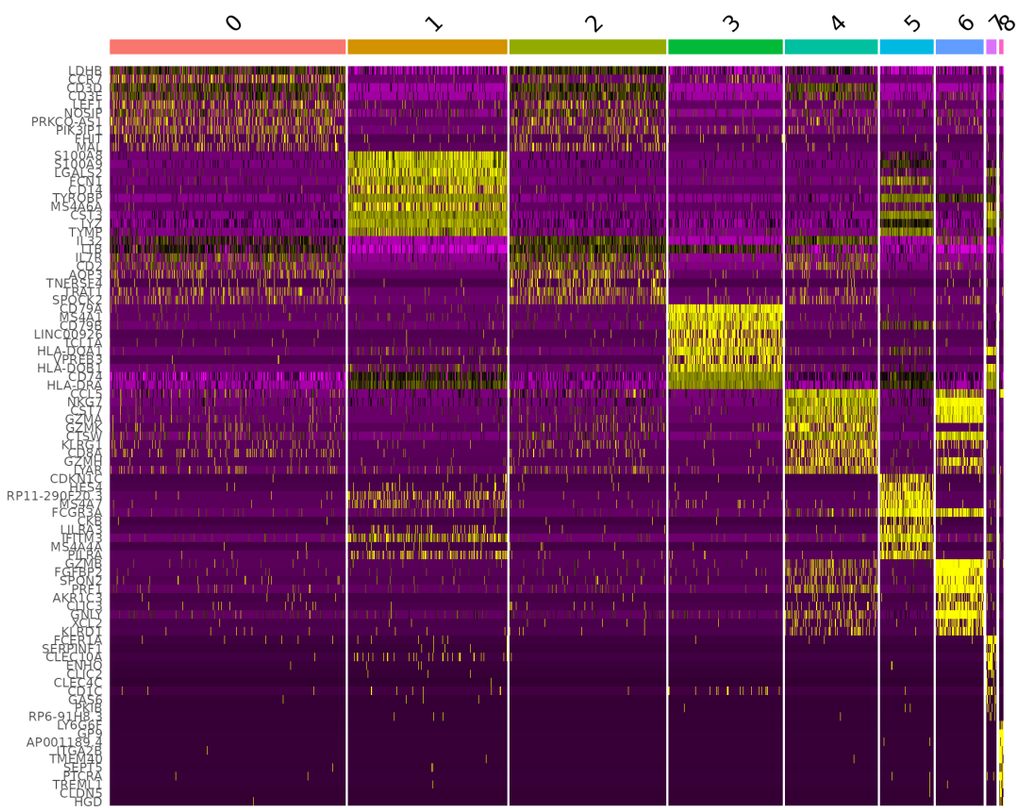
- Marker genes
- Gene modules
- Functional enrichment

Clark et al., 2021

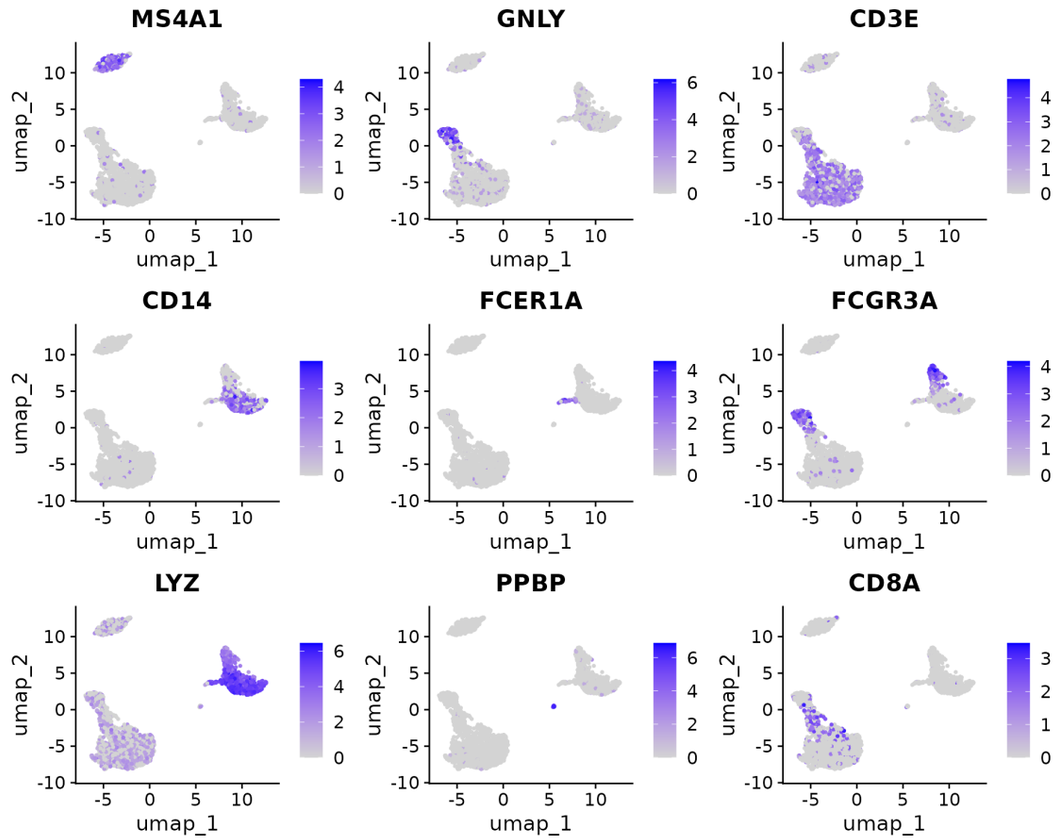
Manual annotation - marker genes



Data-driven



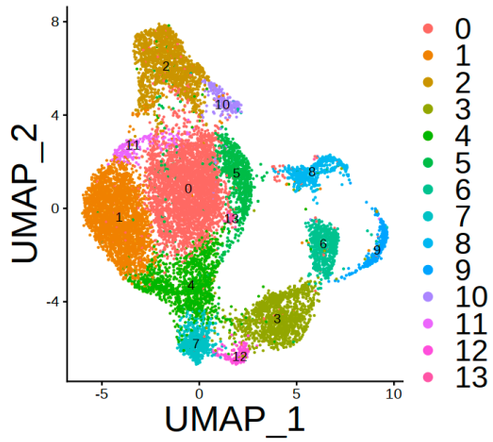
Literature



Seurat tutorial

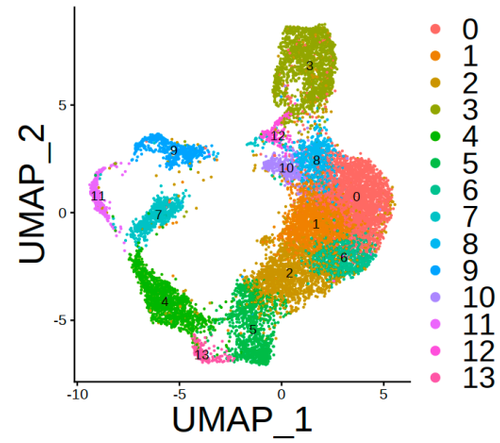
Before filtration of technical genes

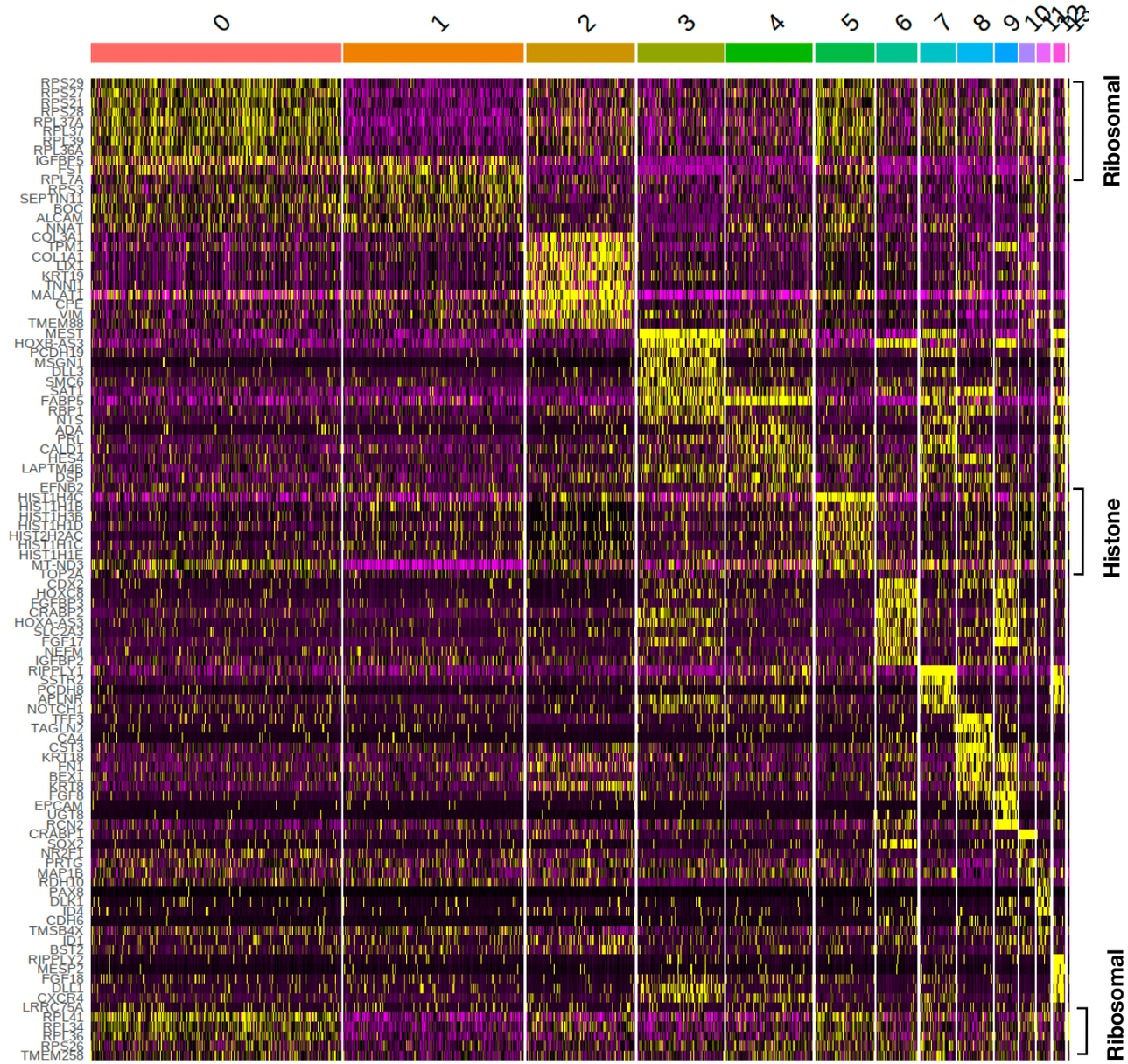
96 hr human gastruloid

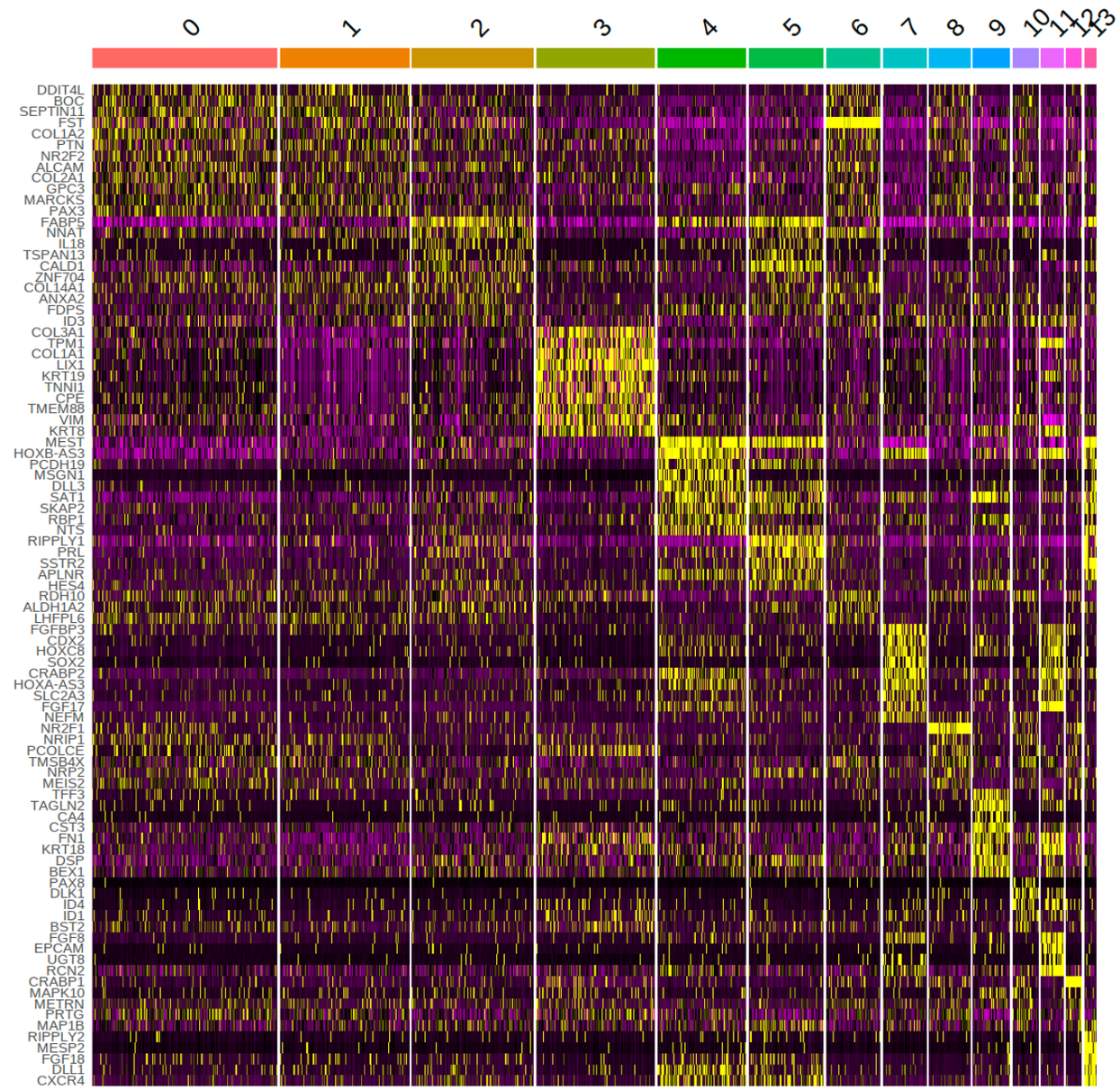


After filtration of batch technical genes

96 hr human gastruloid

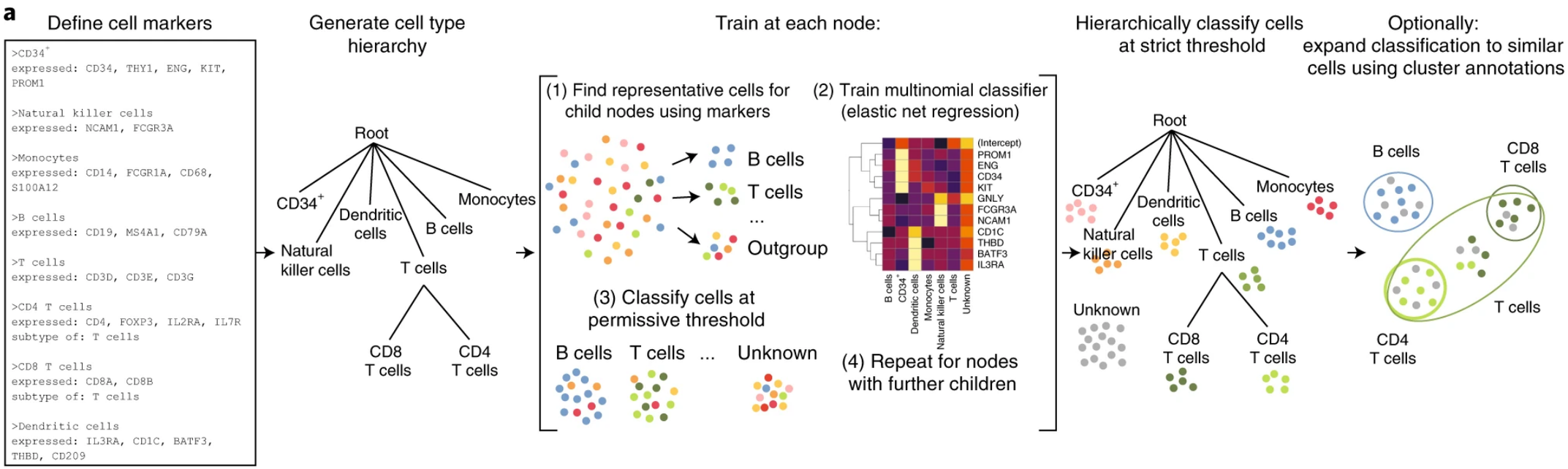






Manual annotation - marker genes

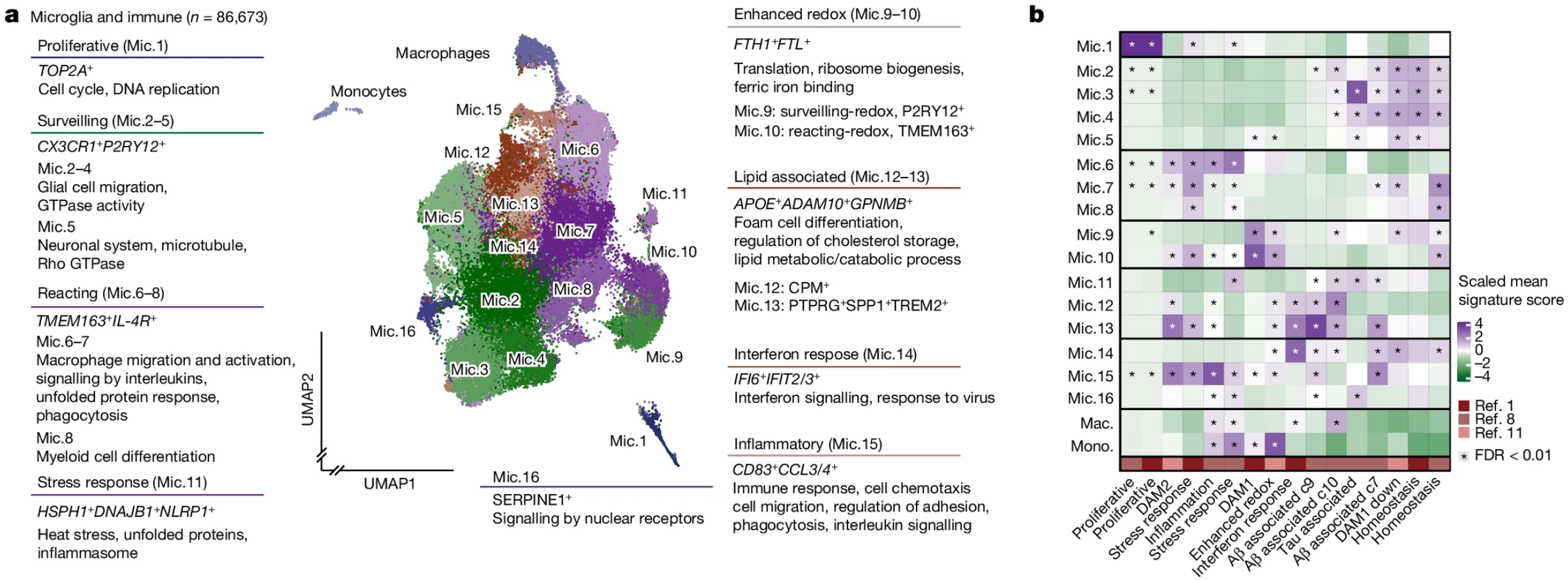
- Garnett trained a classifier with a set of marker genes for each cell type for fast annotation based on marker gene expression.



Pliner et al., 2019

Manual annotation - gene modules and functional enrichment

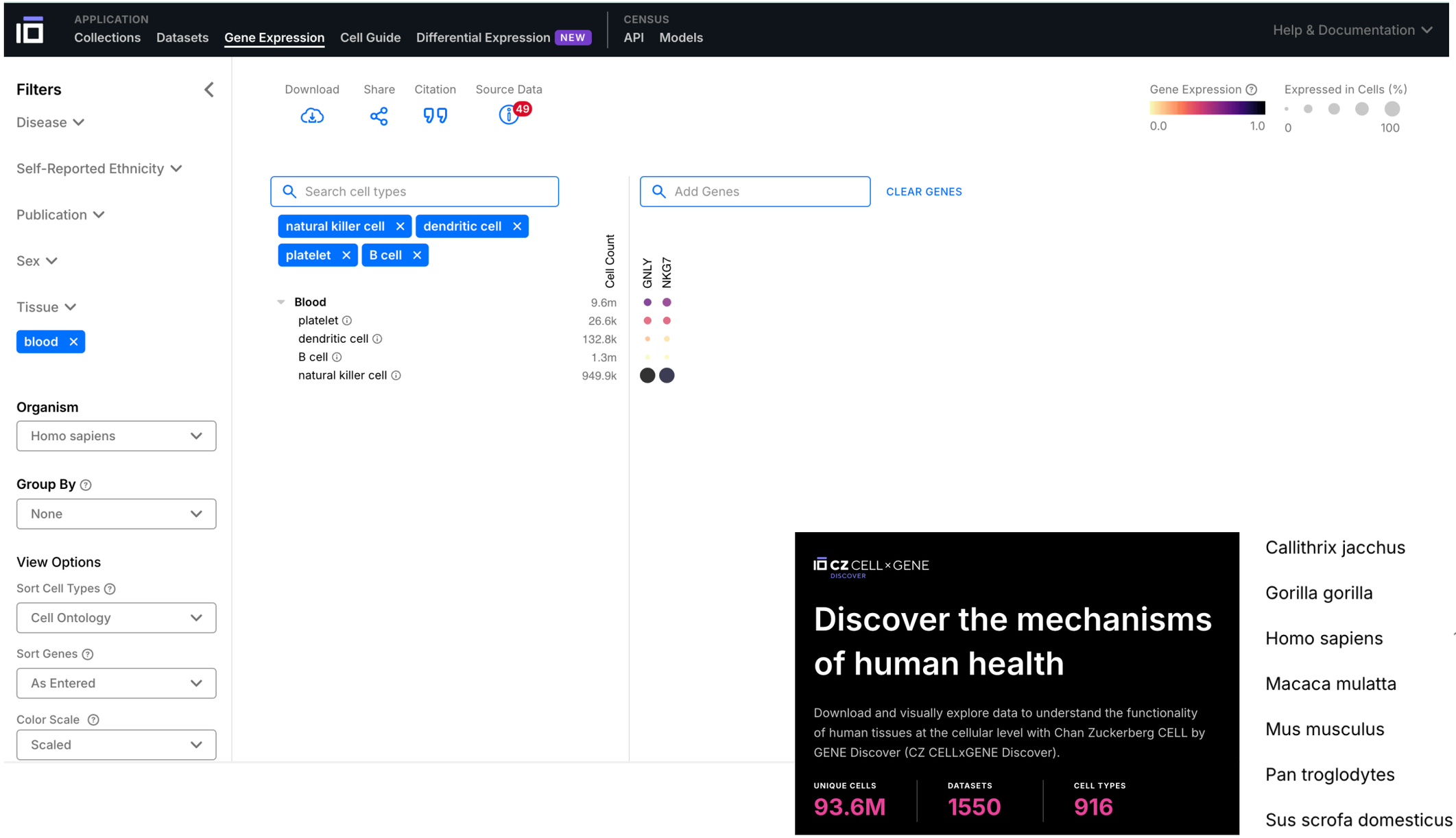
- In de novo annotation, gene modules and GO terms are useful to define functions of a new cell subtype.



Green et al., 2024

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

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

Filter

Show cell type: NK cells (98) get tsv file add marker

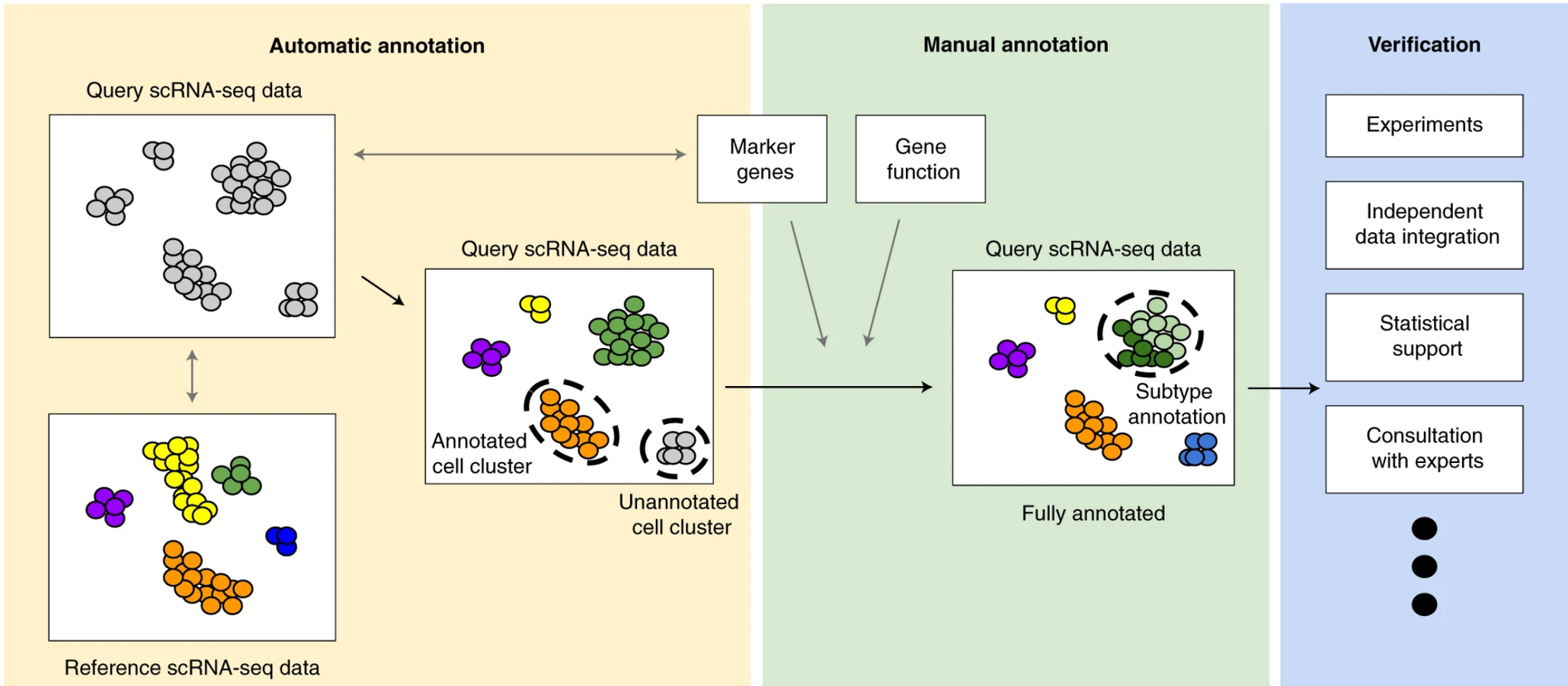
Gene expression markers for NK cells

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	NKG5,LAG-2,D2S69E,TLA519,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).
- Maker 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

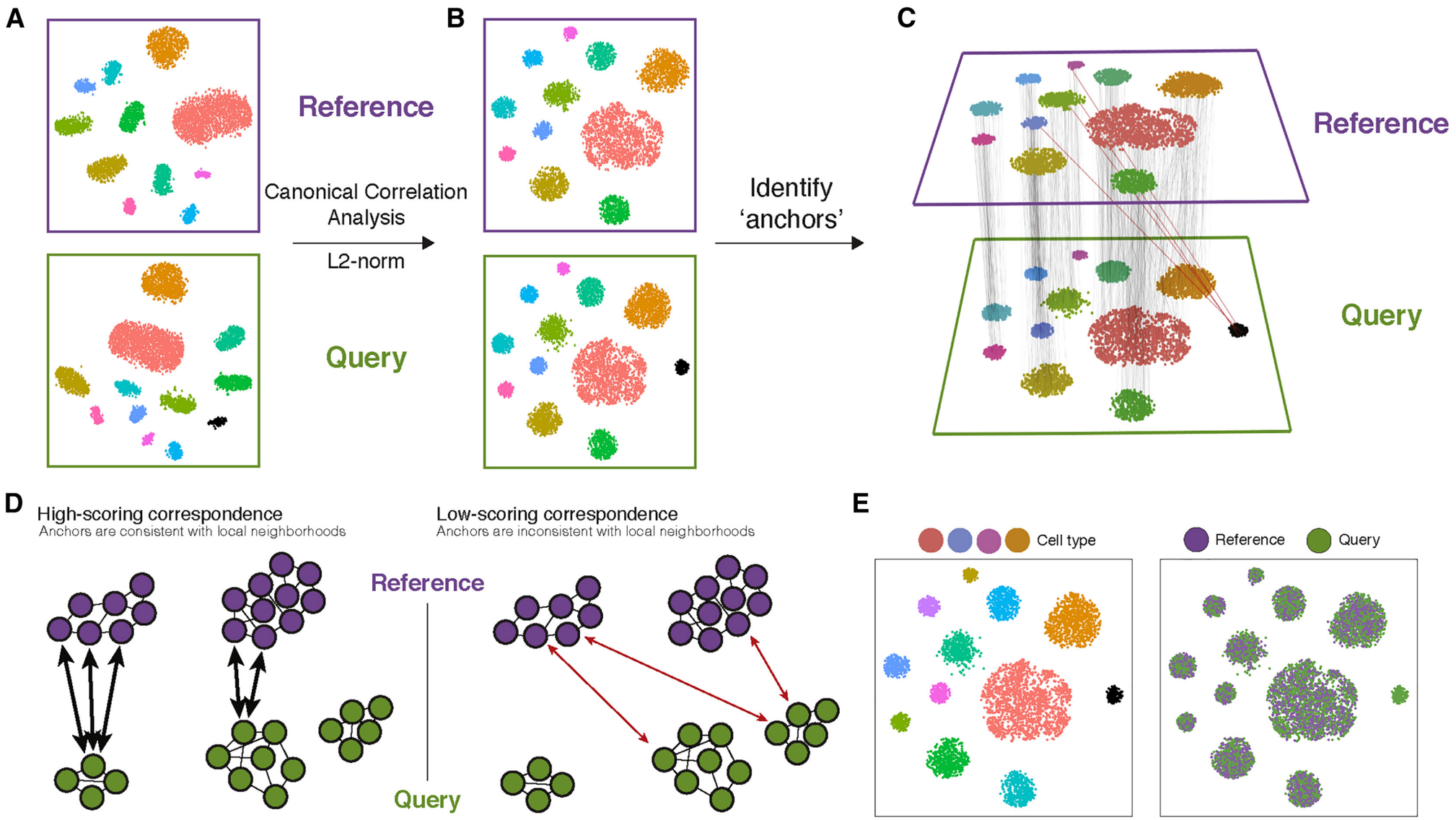


- Integration and label transfer
- NNLS
- Web tools

- Marker genes
- Gene modules
- GO terms

Clark et al., 2021

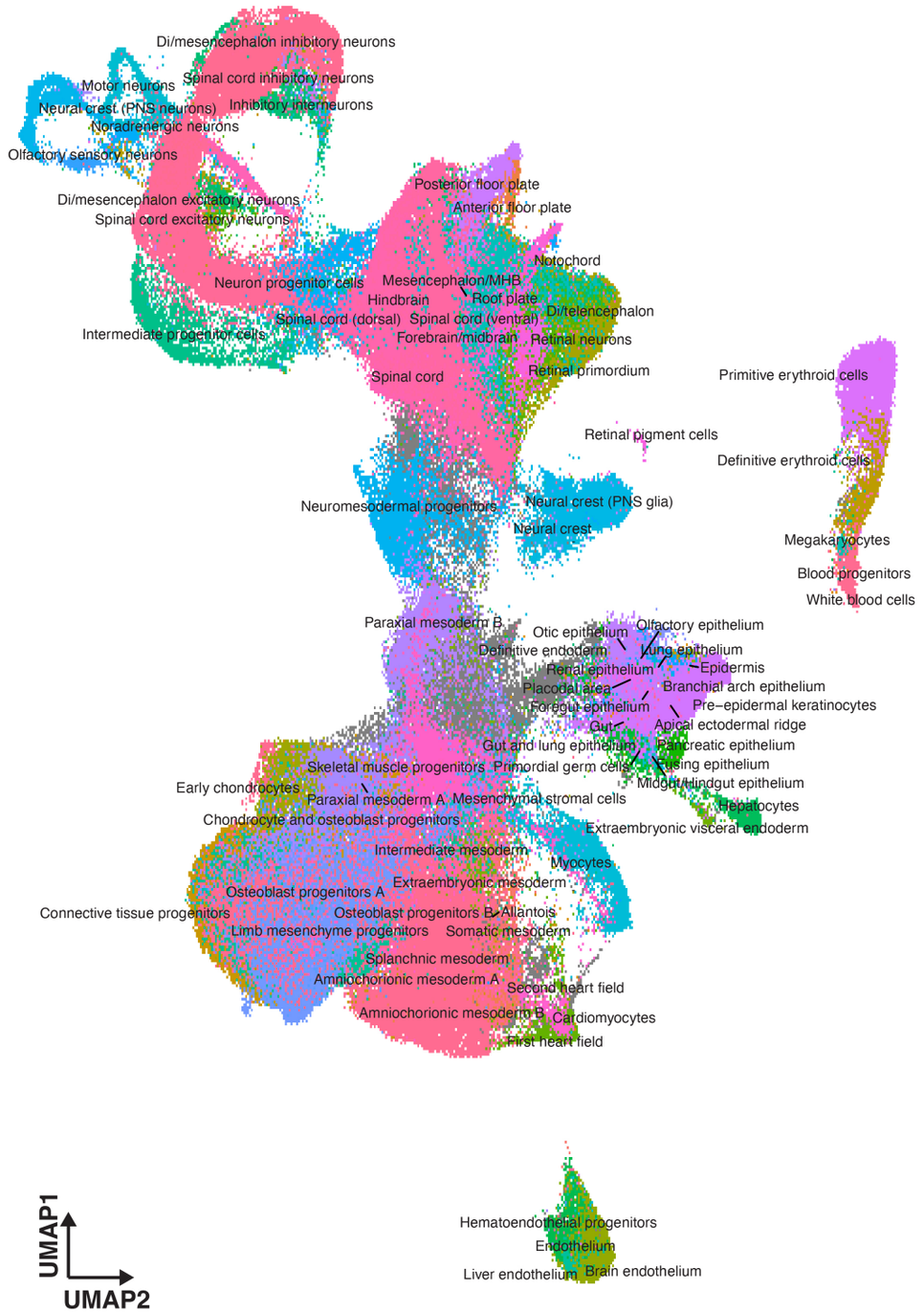
Automatic annotation - integration and label transfer



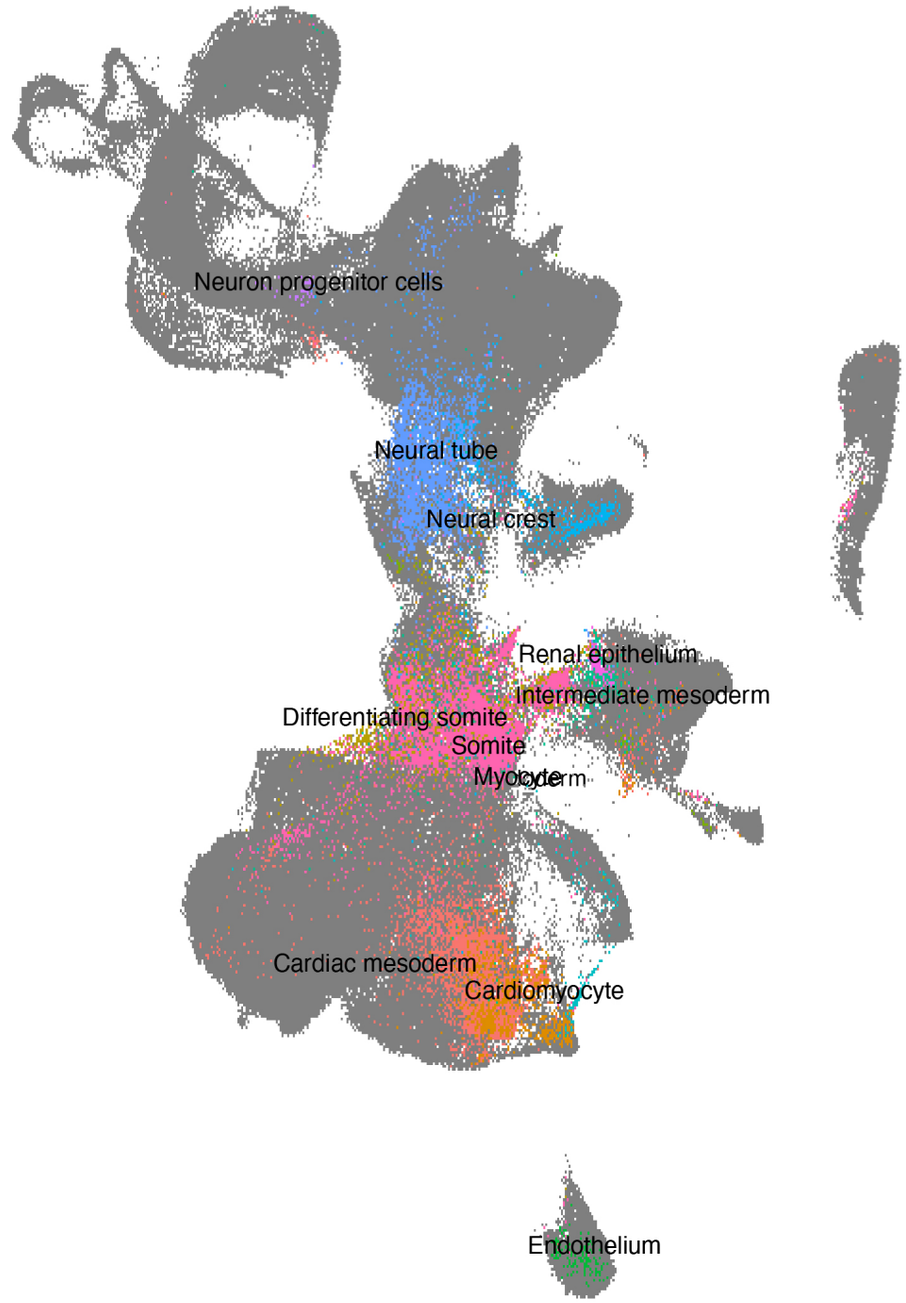
Seurat V3

Automatic annotation - integration and label transfer

Mouse embryo E8.5-13.5



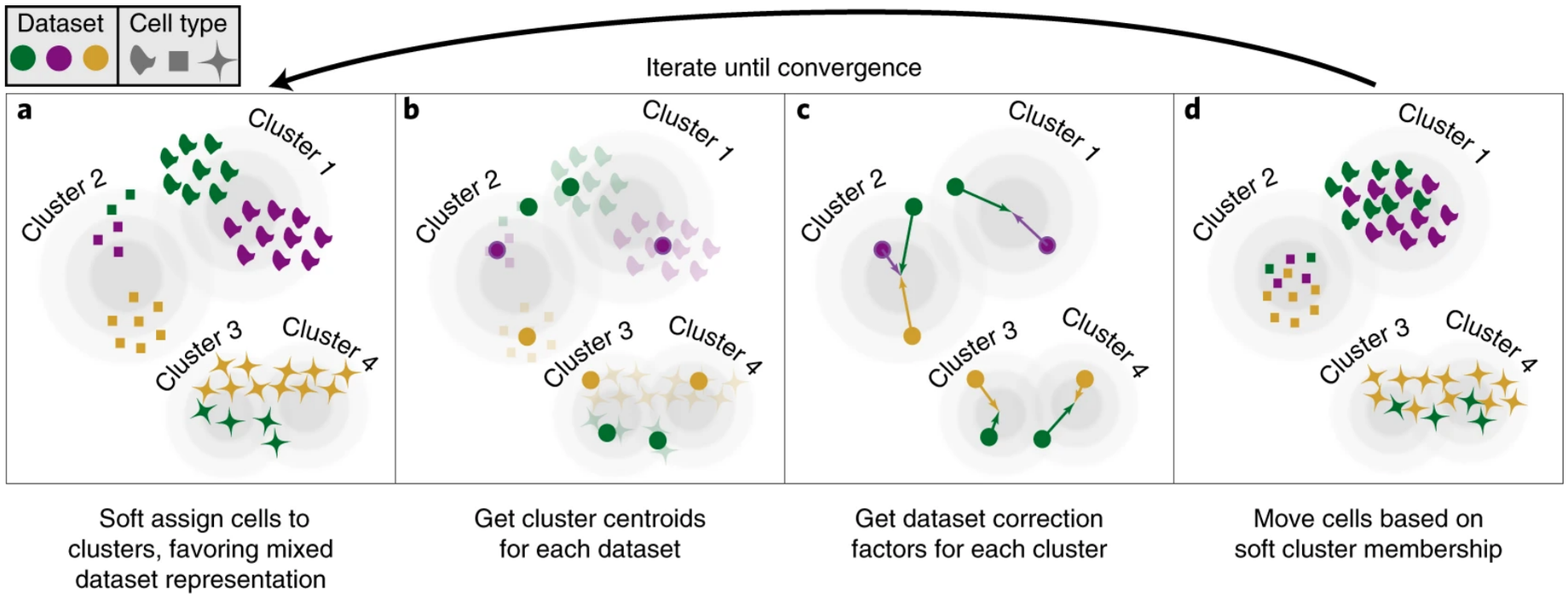
Human RA gastruloids (120 hrs)



* Integration by Seurat V3 (Hamazaki, Yang et al., 2024)

Automatic annotation - integration and label transfer

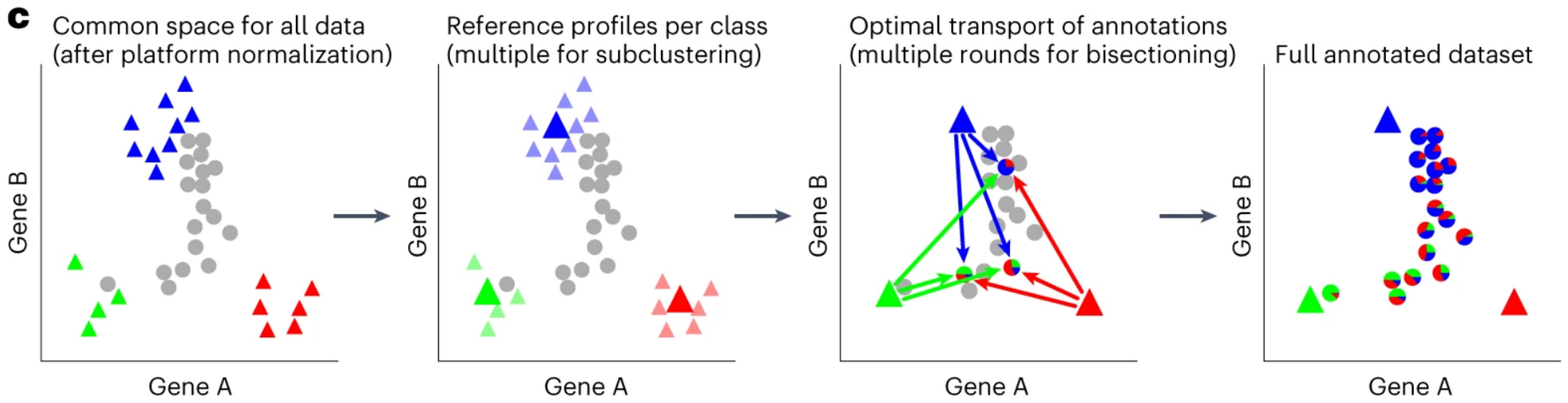
- Harmony: much faster than MNN approach but work poorly for cross-species integration.



Korsunsky et al., 2019

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

Mages et al., 2023

Automatic annotation - integration and label transfer

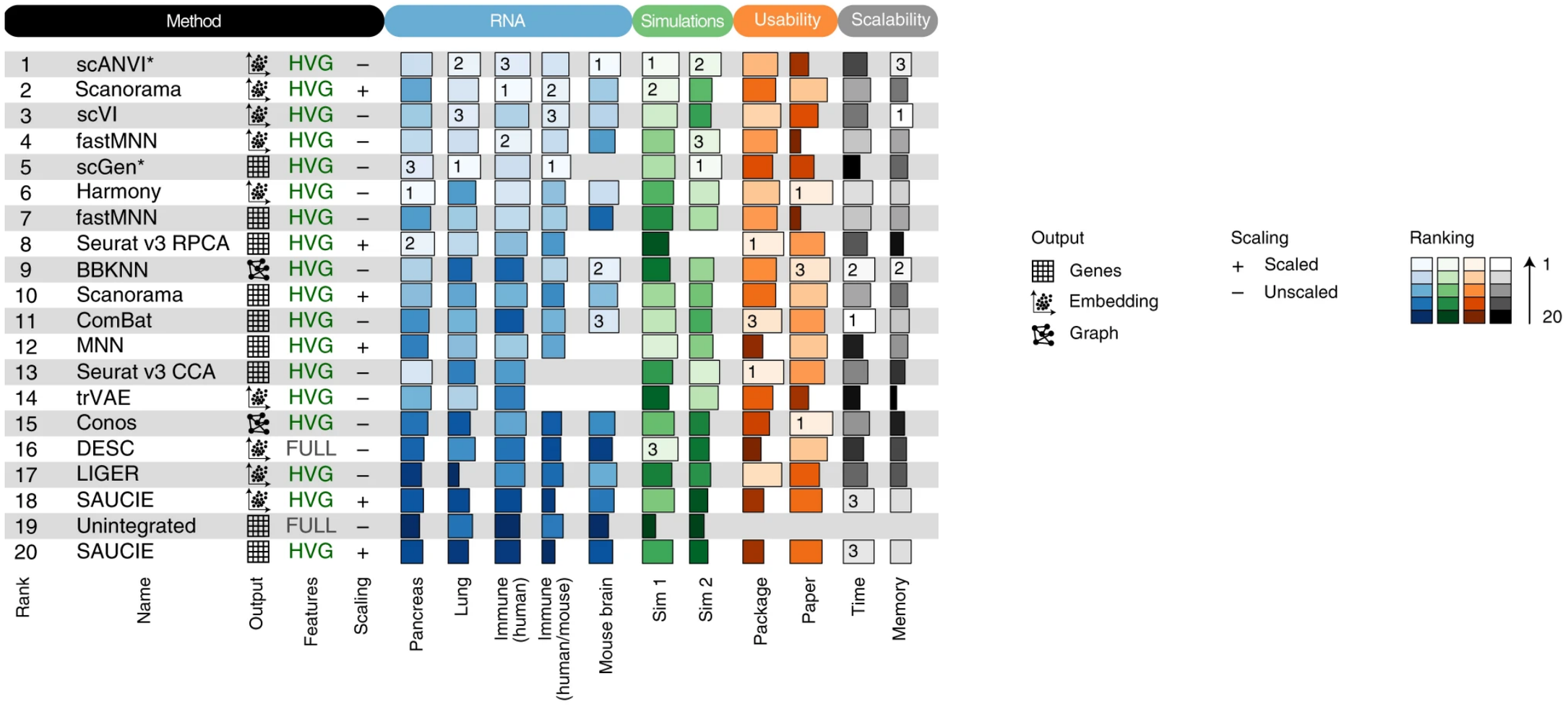
- Other algorithms

Analysis | Open access | Published: 23 December 2021

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



Luecken et al., 2021

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.

$$Ta = \beta 0a + \beta 1a Mb$$

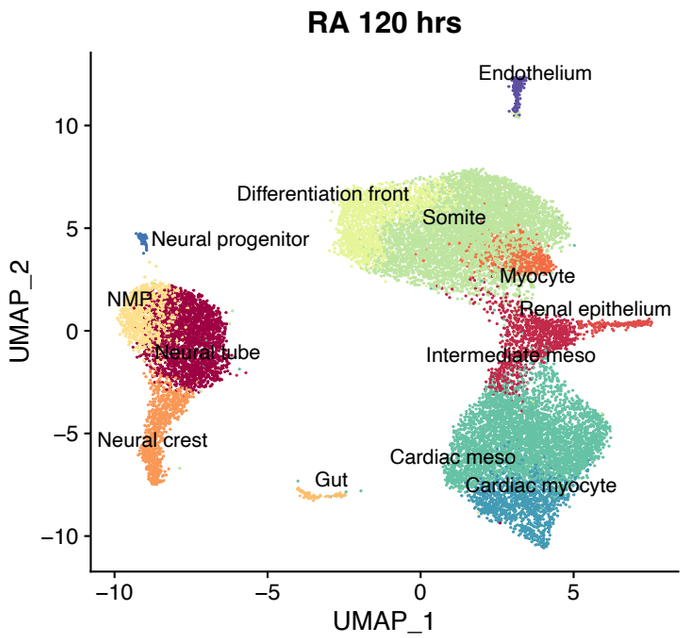
$$Tb = \beta 0b + \beta 1b Ma$$

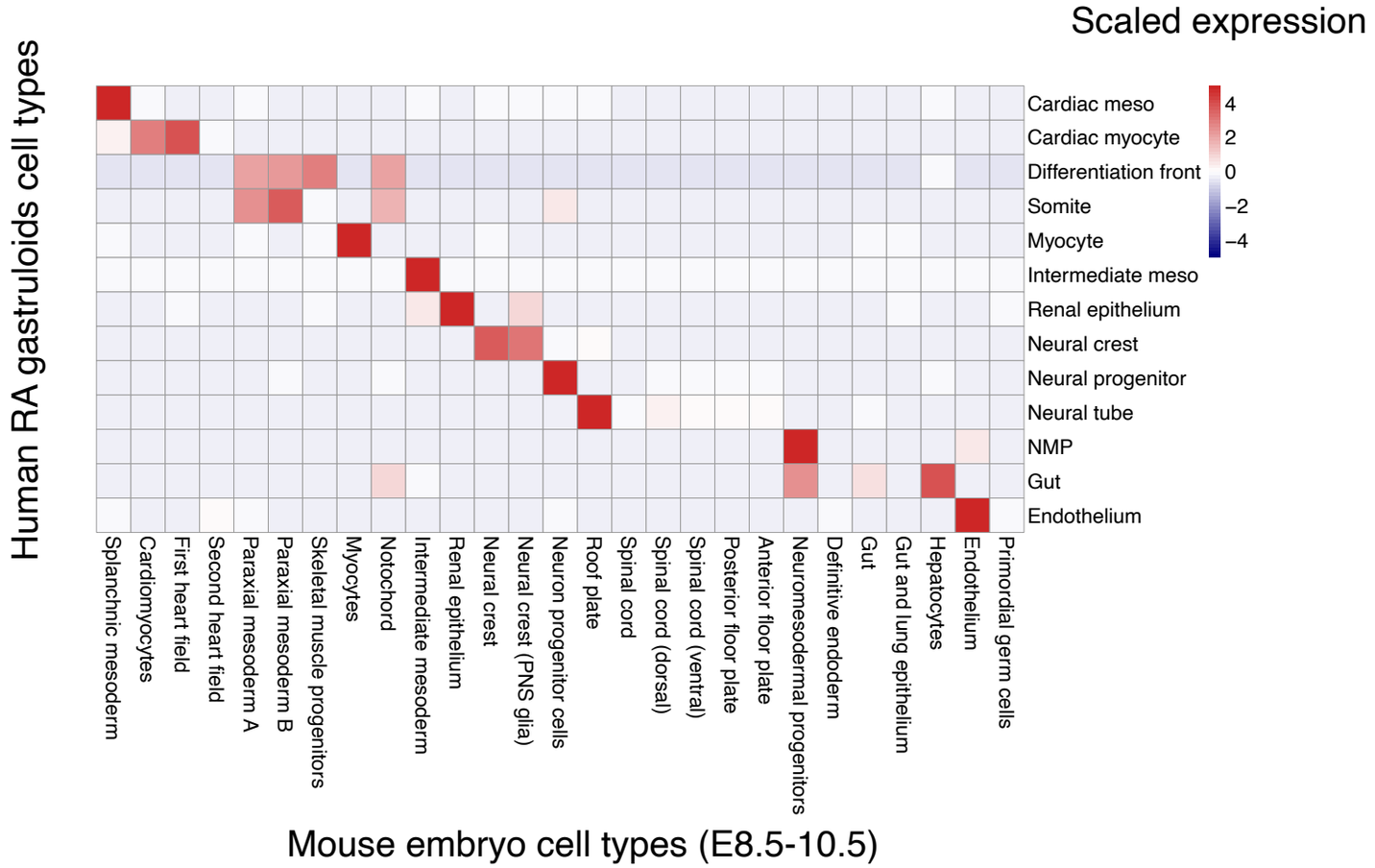
$$\beta = 2(\beta 1ab + 0.001)(\beta 1ba + 0.001)$$

Ta/b: gene expression in target cell type
Ma/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.





Hamazaki, Yang et al., 2024

Automatic annotation - web tools

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

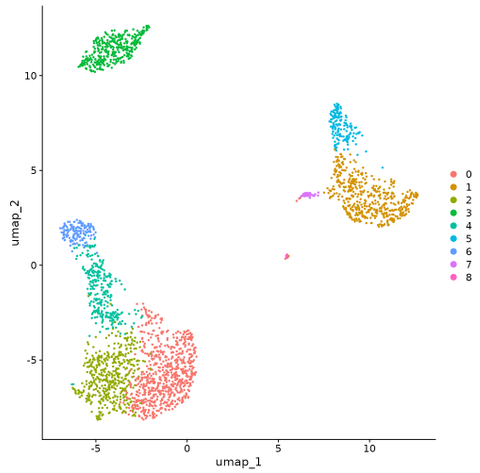
	Azimuth	**Tabula Sapiens**	**MapmyCells**
Species	Human(11) and mouse (1)	Human	Human and mouse
Organs	PBMC, Motor cortex, pancreas, fetal, kidney, bone marrow, lung, adipose, tonsil, heart, liver	24 organs	Whole brain and middle temporal gyrus
Assay	RNA(12) and ATAC(2)	RNA	RNA
Interface	Web portal, 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