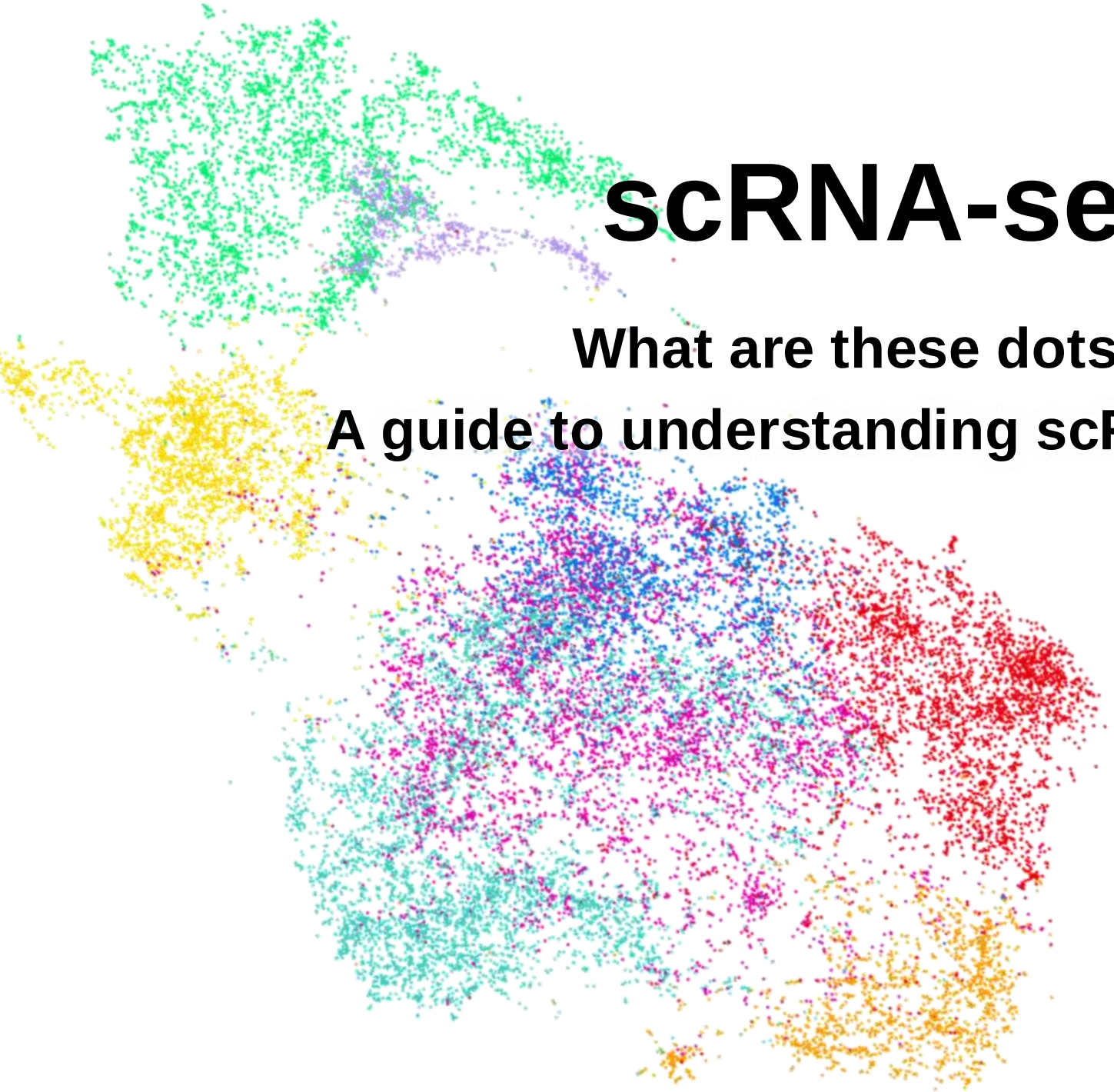


scRNA-seq 101

What are these dots in my plots?

A guide to understanding scRNA-seq data analysis

Computational Biology in Cancer

Marta Bica
Nuno L. Barbosa-Morais

Champalimaud Foundation
09/11/2022

FACULDADE DE
MEDICINA
LISBOA

Instituto
de Medicina
Molecular

João
Lobo
Antunes

Marta

ana.bica@medicina.ulisboa.pt

2021-**present**
PhD candidate (FCT scholar), iMM/FMUL
"Single-cell transcriptomics in unravelling the complexity of the breast tumour microenvironment"

2019-2021
GenomePT Research Fellow, Disease Transcriptomics lab, iMM

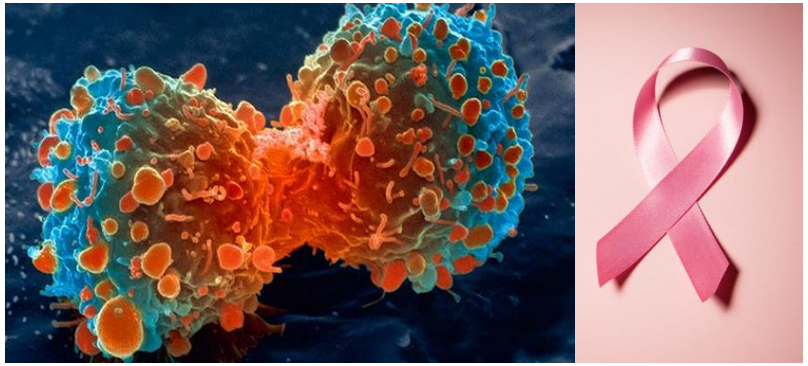
2017-2019
MSc in Bioinformatics and Computational Biology, Fac. Sciences, U. Lisbon
"Single-cell transcriptomics in unravelling the molecular complexity of immunity in human disease"

2013-2017
BSc in Health Sciences, U. Lisbon

What we do: disease transcriptomics

Goal *Understand how ageing-associated molecular (RNA) changes in human tissues increase proneness to disease*

(Disease) models

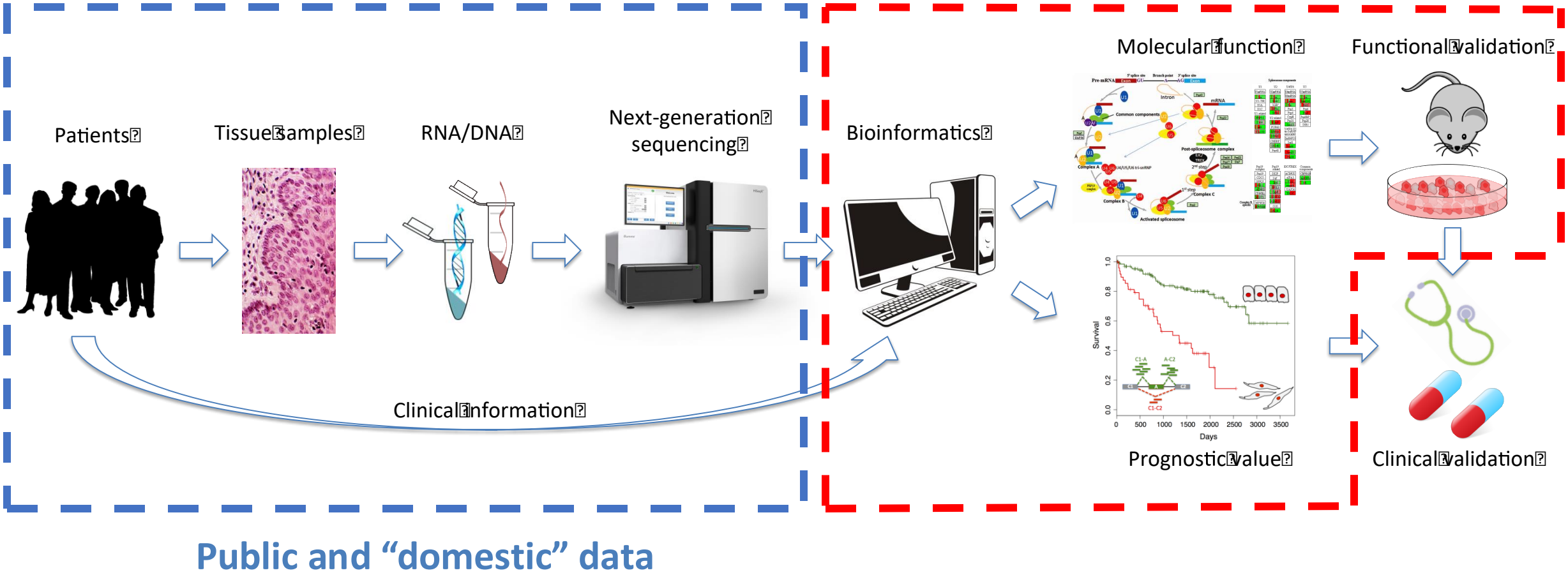


Cancer

Neurodegeneration

Ageing

What we do: disease transcriptomics

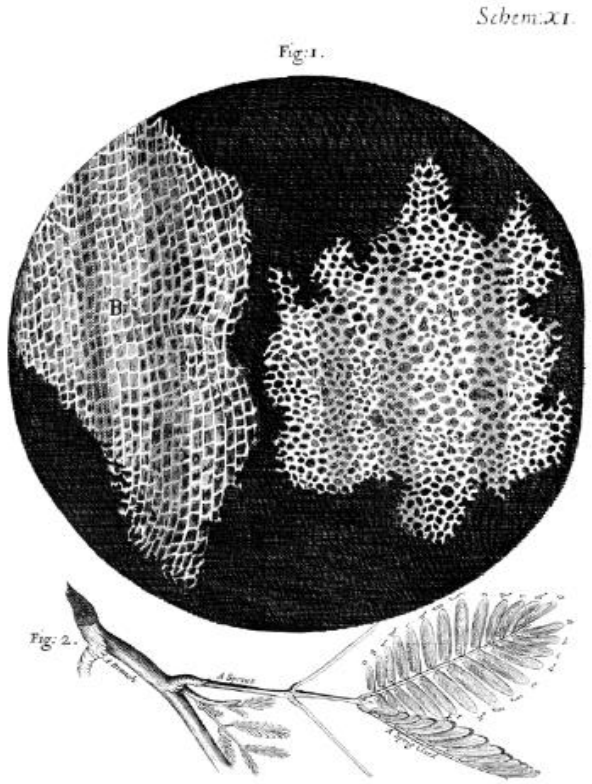


Public and “domestic” data

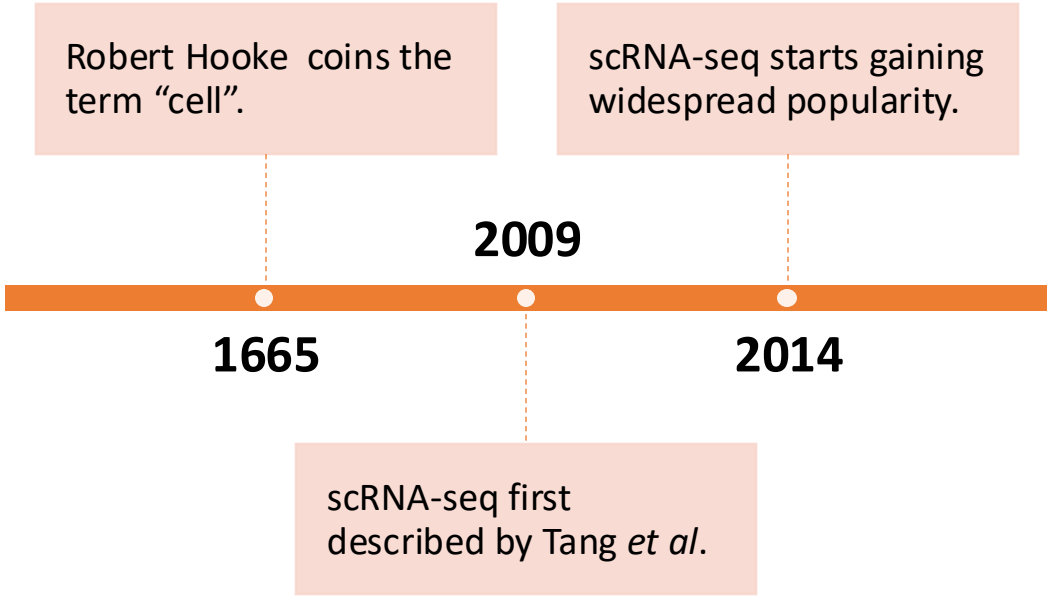
Outline

- What analysing big biological data (i.e., being a computational biologist) entails
- Understanding all steps of scRNA-seq data analysis
- Case study: *Single-cell transcriptomics in unravelling the therapeutic potential of myeloid cells in breast cancer*
- DIY with ***scStudio***

350 years studying the cell



Drawing of cork cells seen through the microscope.



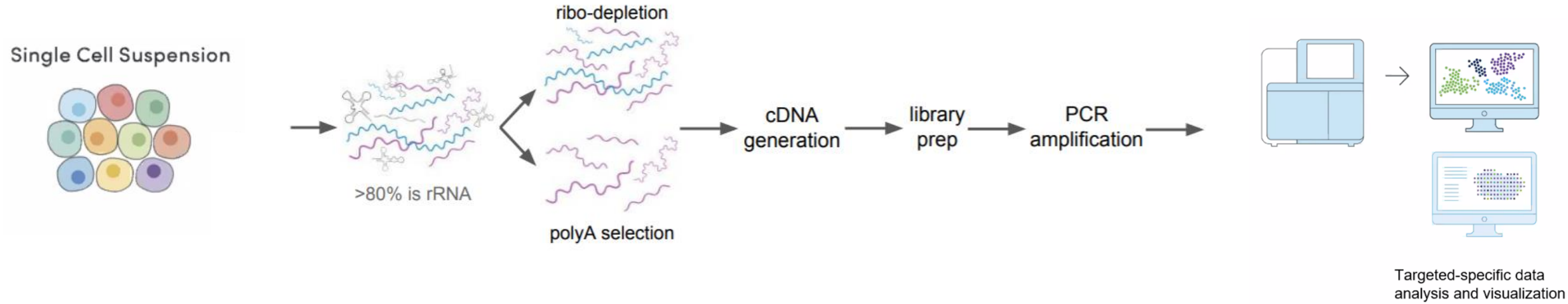
What is single-cell transcriptomics?

omics (ō′mĭks, ŏm′ĭks)

n.

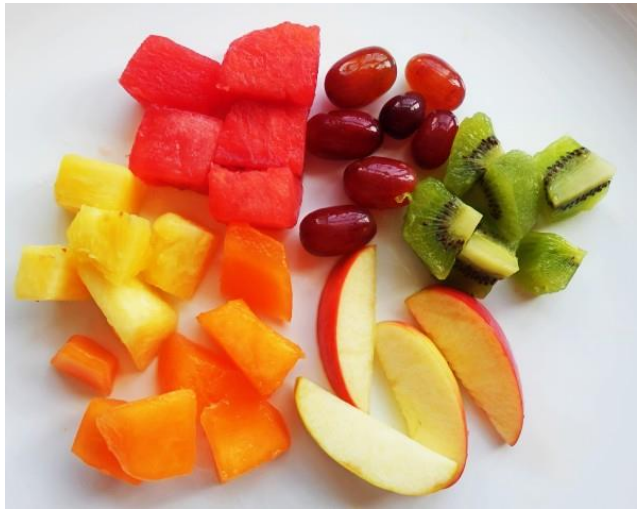
1. *(used with a sing. or pl. verb)* Analysis of large amounts of data representing an entire set of some kind, especially the entire set of molecules, such as proteins, lipids, or metabolites, in a cell, organ, or organism.

2. *(used with a sing. verb)* Any of the fields employing this approach, as proteomics or metabolomics.



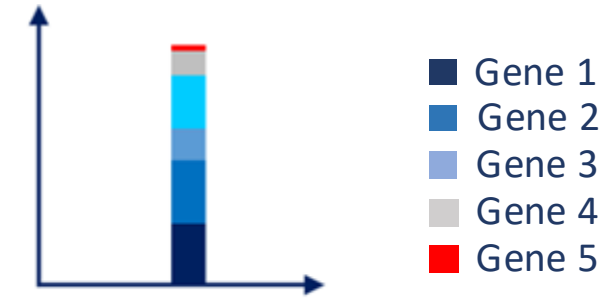
What is the advantage of single-cell?

Bulk RNA-seq

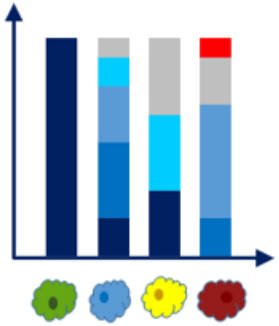


Single-cell RNA-seq

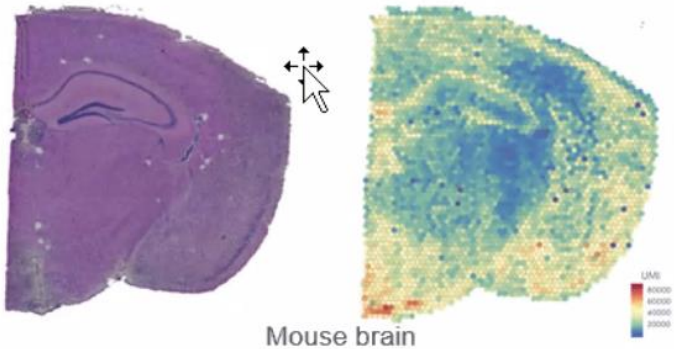
Spatial transcriptomics



Average gene expression from all cells



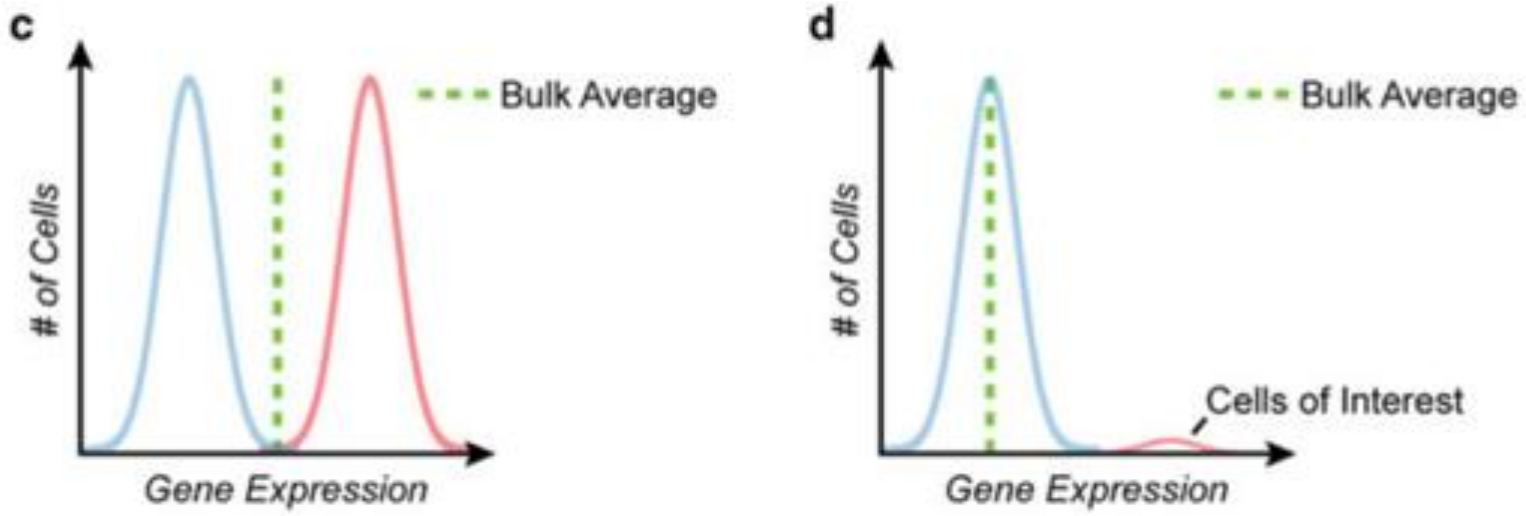
Each cell type has distinct expression profile



Mouse brain

Single-cell vs bulk RNA-seq

Bulk RNA-seq is not sufficient for studying cell diversity in an heterogenous tissue



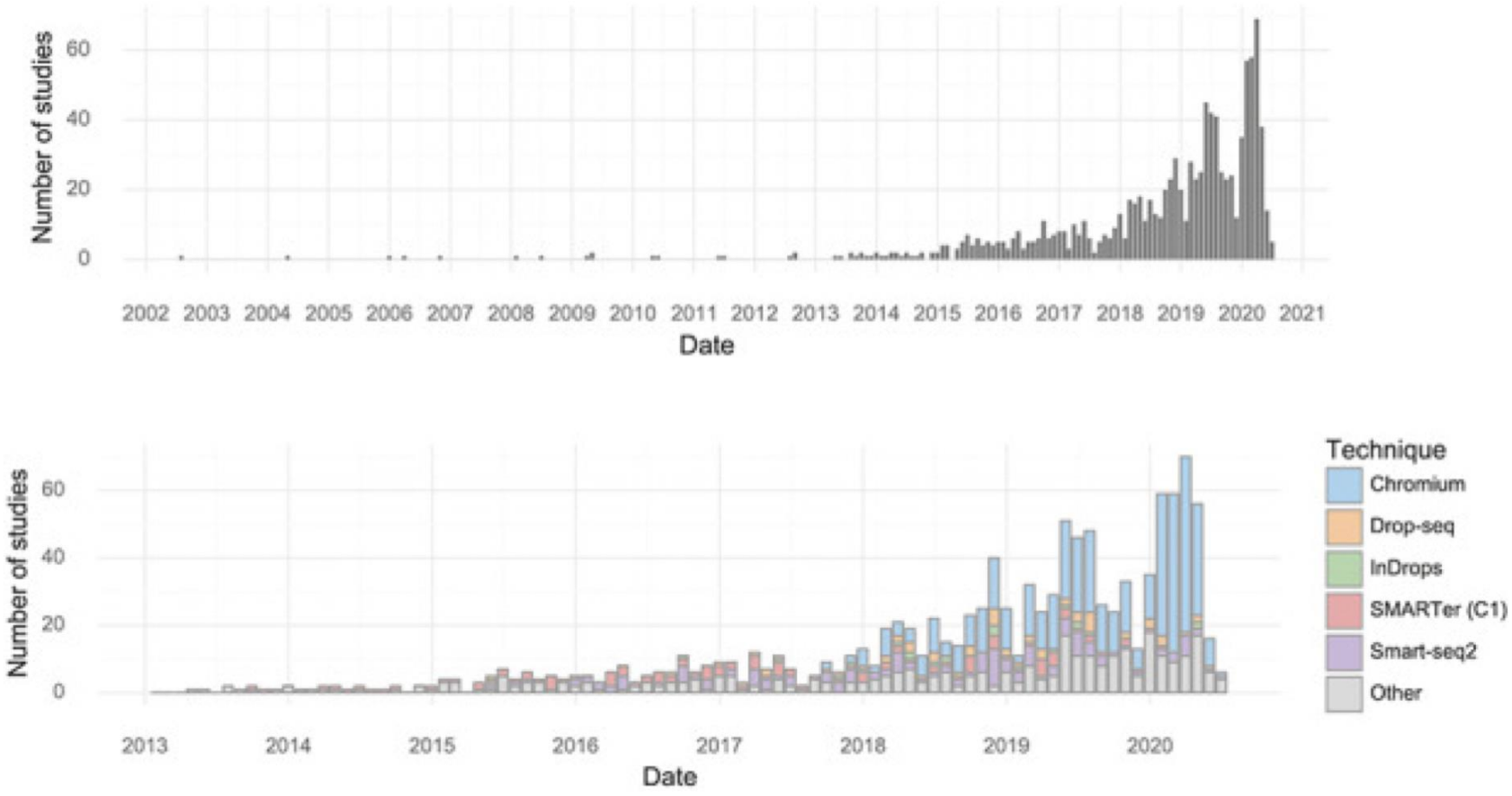
Methods in Molecular Biology, 2012. Single-Cell Analysis.

Types of questions

- Descriptive Analysis (one condition):
  - *How many cell-types are there? What are they?*
  - *Is Gene X expressed in a particular cell-type?*
  - *Are there developmental trajectories/gradients?*
- Relational (one condition, multiple cell-types):
  - *Which cell-types are similar to other cell-types?*
  - *Does cell-type A communicate with cell-type B?*
  - *Is cell-type A a progenitor of cell-type B?*
- Comparative Analyses (multiple conditions):
  - *How does this cell-type change during disease?*
  - *In response to a drug/mutation?*
  - *When in contact with a different cell-type?*

- What does this high-resolution view of cellular transitions tell us about switches in cell state?

The growth of the number of scRNA-seq studies is exponential



Application areas

Developmental biology

Neurology

Infectious disease

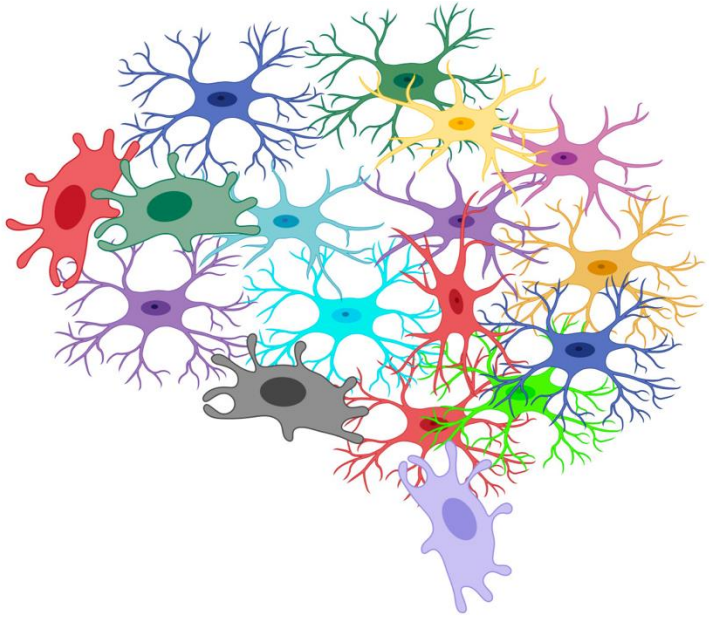
Cardiovascular research

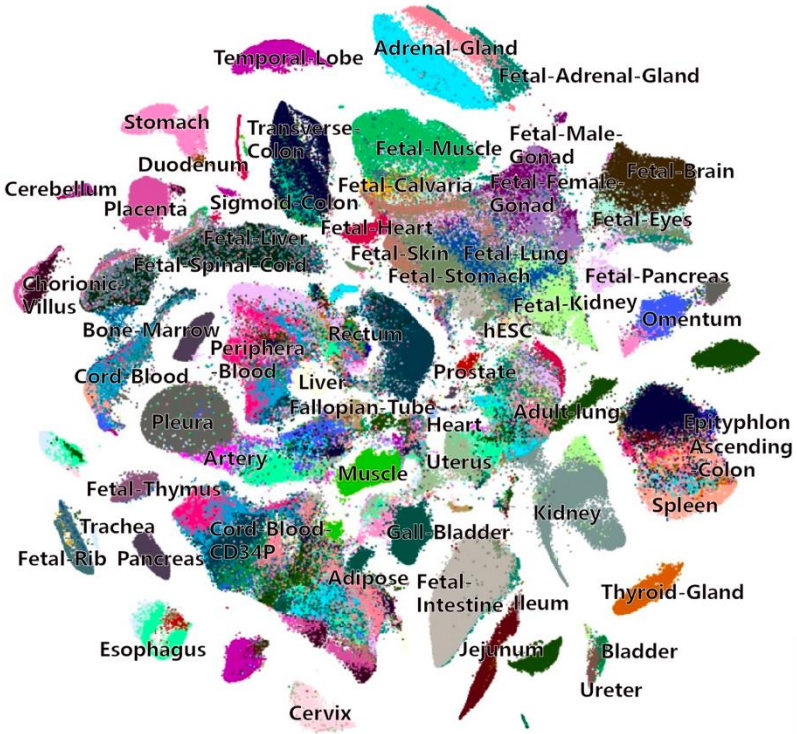
Oncology

Immunology

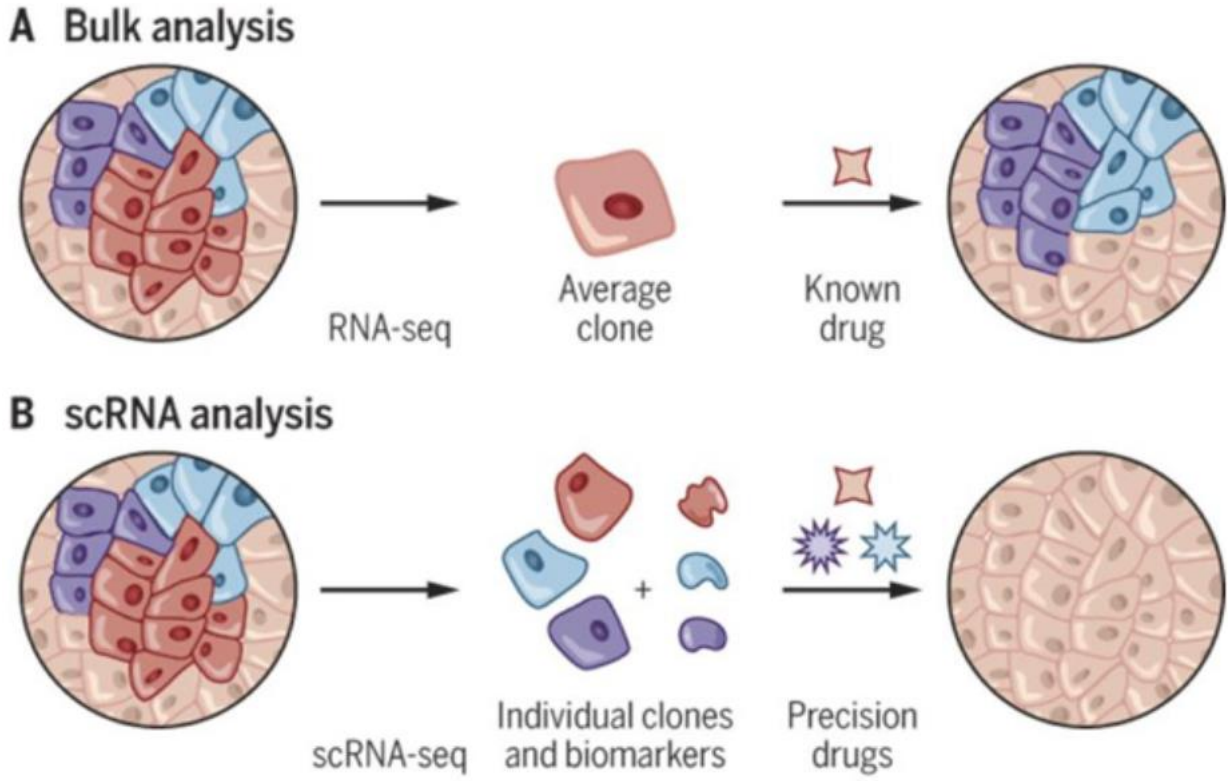
Ageing

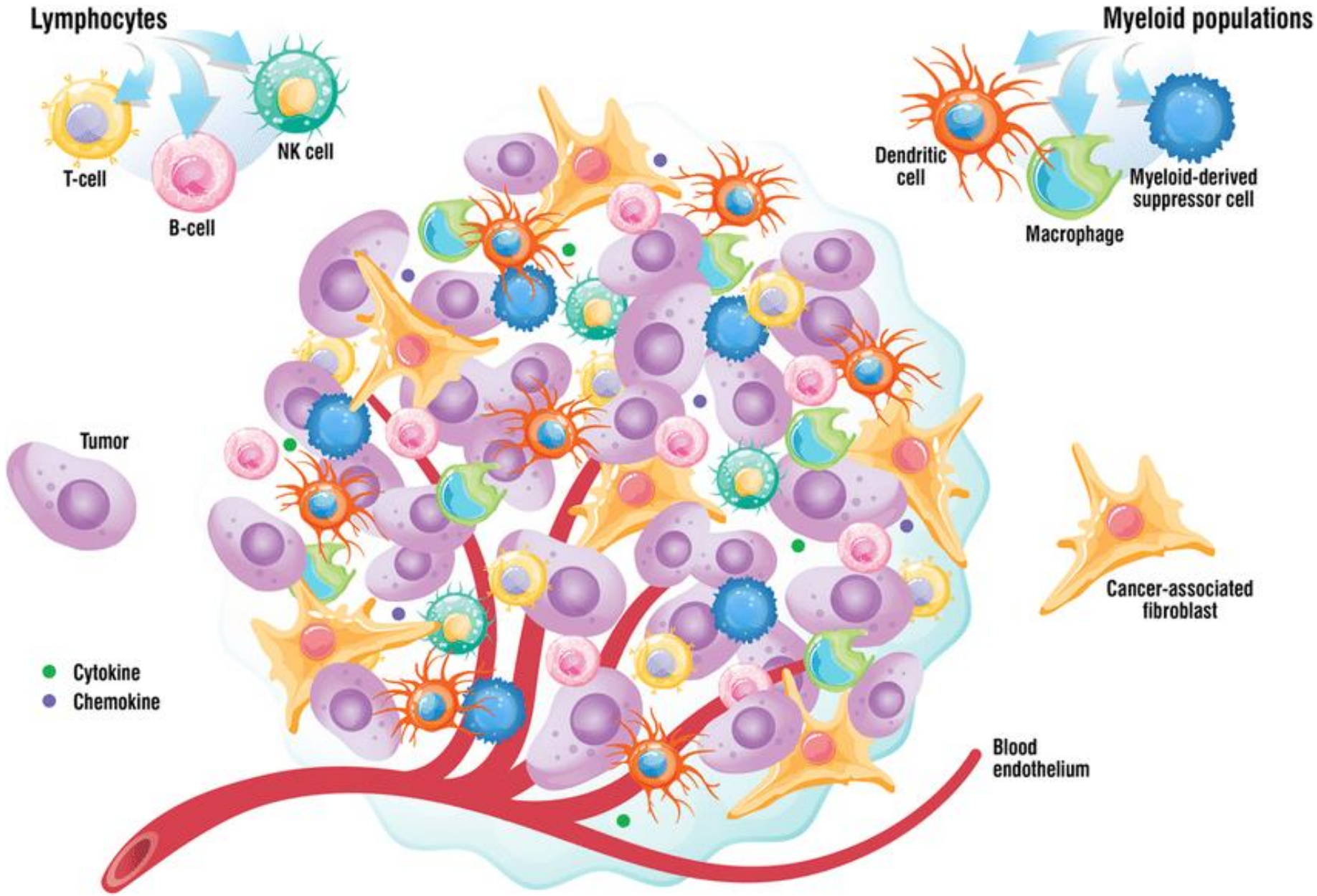
Stem cell research





Single-cell analyses to tailor treatments




Lymphocytes
T-cell
B-cell
NK cell
Myeloid populations
Dendritic cell
Macrophage
Myeloid-derived suppressor cell
Tumor
Cancer-associated fibroblast
Cytokine
Chemokine
Blood endothelium

Human Cell Atlas – reference of physiology

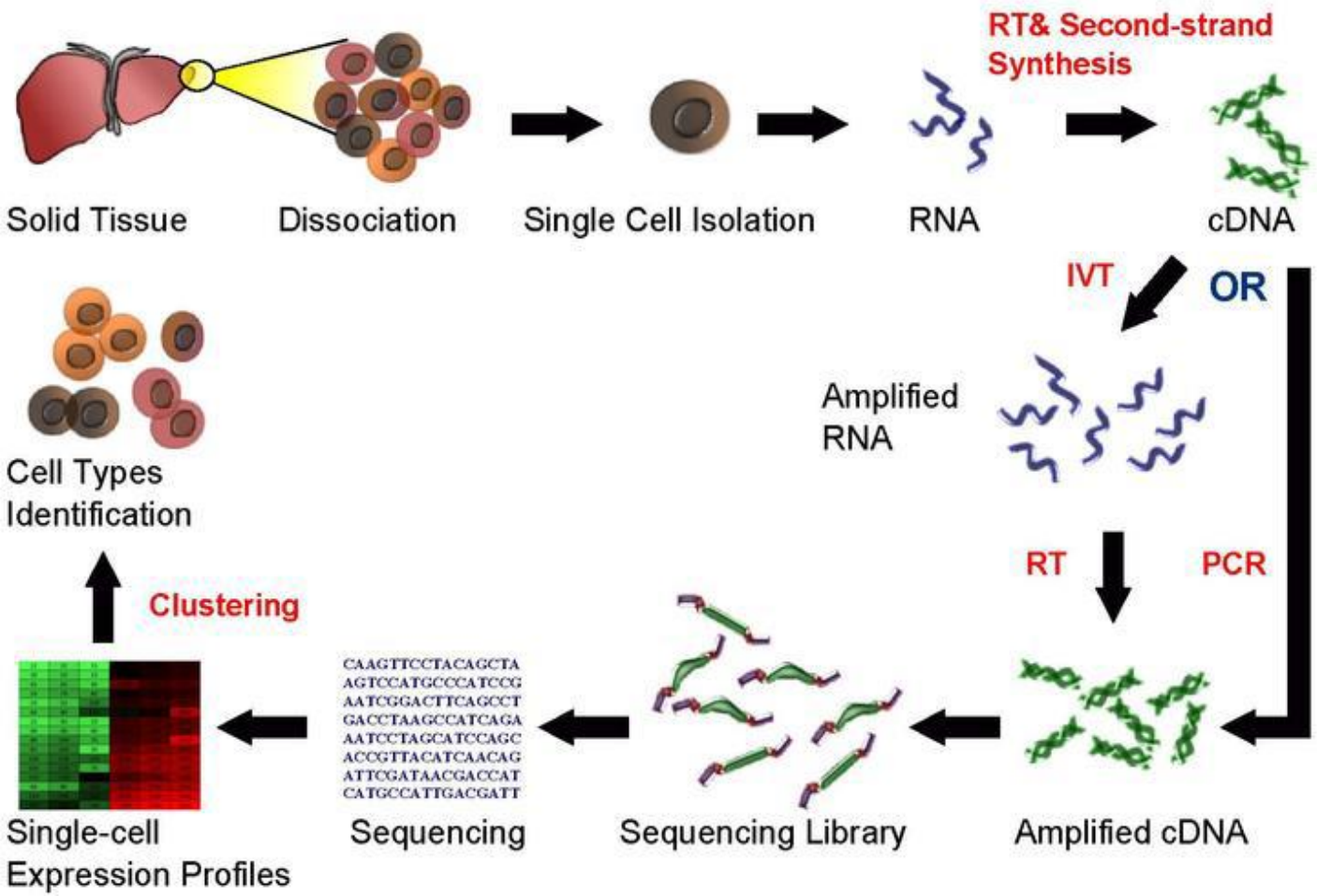
HUMAN CELL ATLAS

"Google maps" for the human body

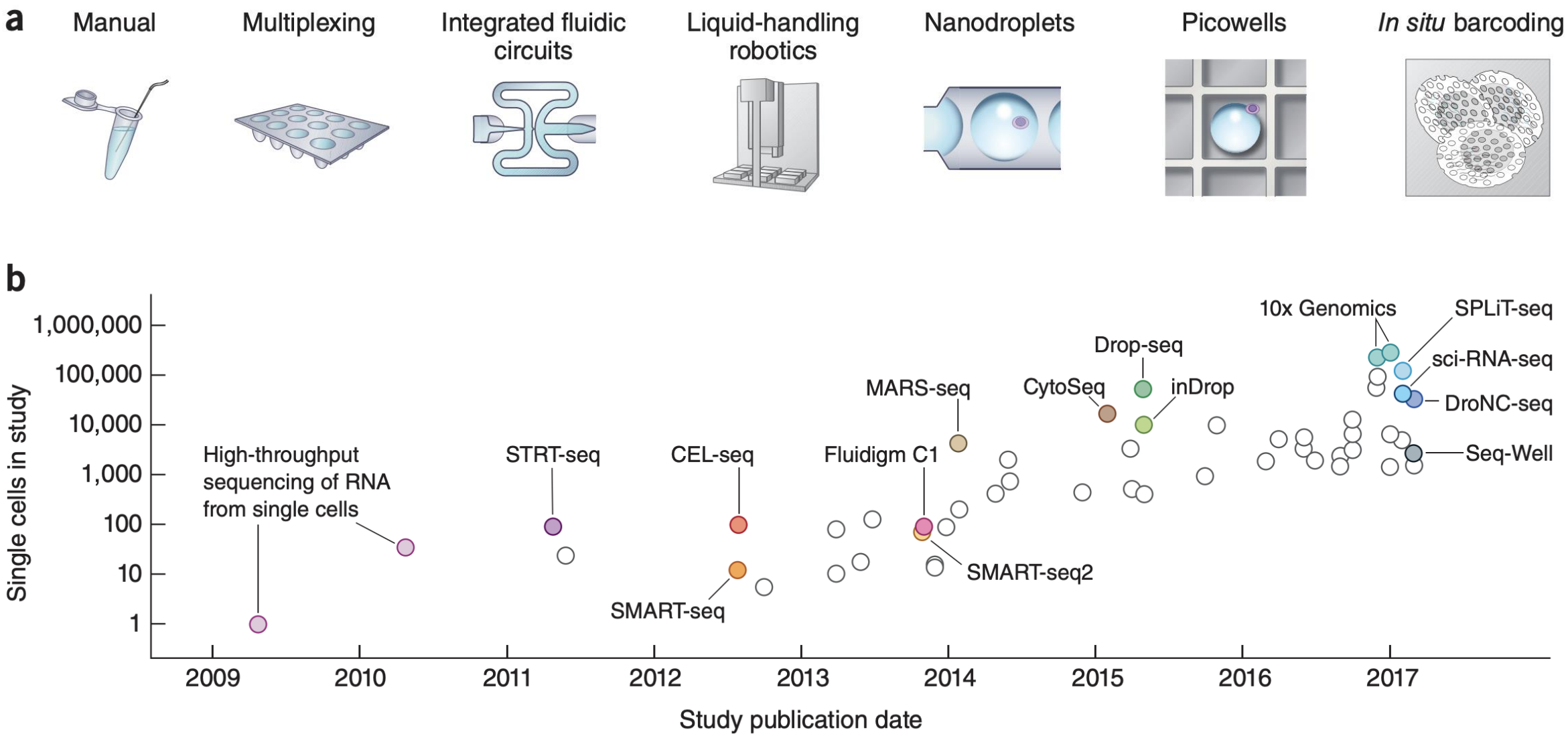
versus

scRNAseq + spatial methods
to define cells in tissues

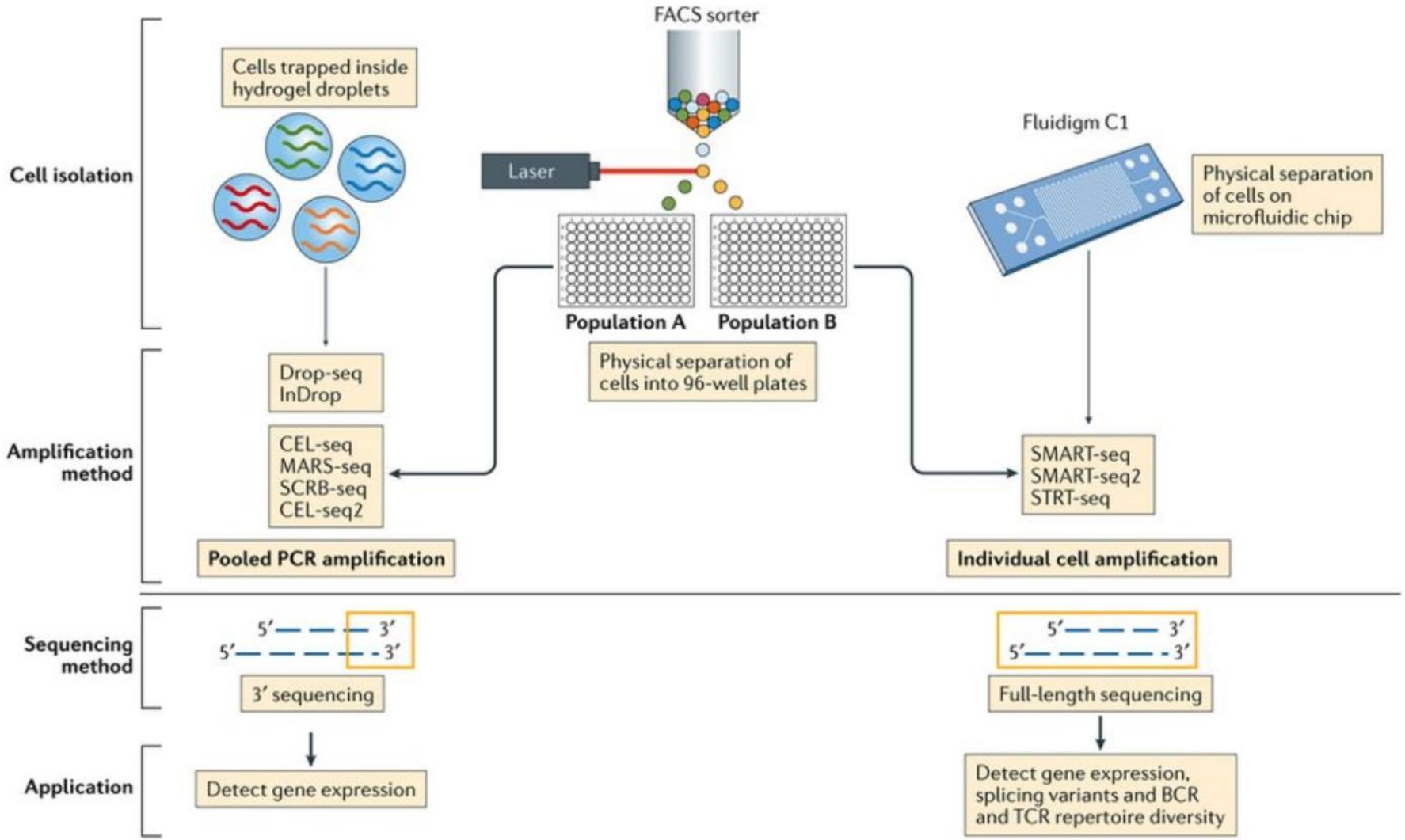
Workflow


RT& Second-strand Synthesis
Solid Tissue
Dissociation
Single Cell Isolation
RNA
cDNA
IVT
OR
Amplified RNA
Cell Types Identification
RT
PCR
Clustering
CAAGTTCCTACAGCTA
AGTCCATGCCCATCCG
AATCGGACTTCAGCCT
GACCTAAGCCATCAGA
AATCCTAGCATCCAGC
ACCGTTACATCAACAG
ATTCGATAACGACCAT
CATGCCATTGACGATT
Single-cell Expression Profiles
Sequencing
Sequencing Library
Amplified cDNA

Many protocols are available



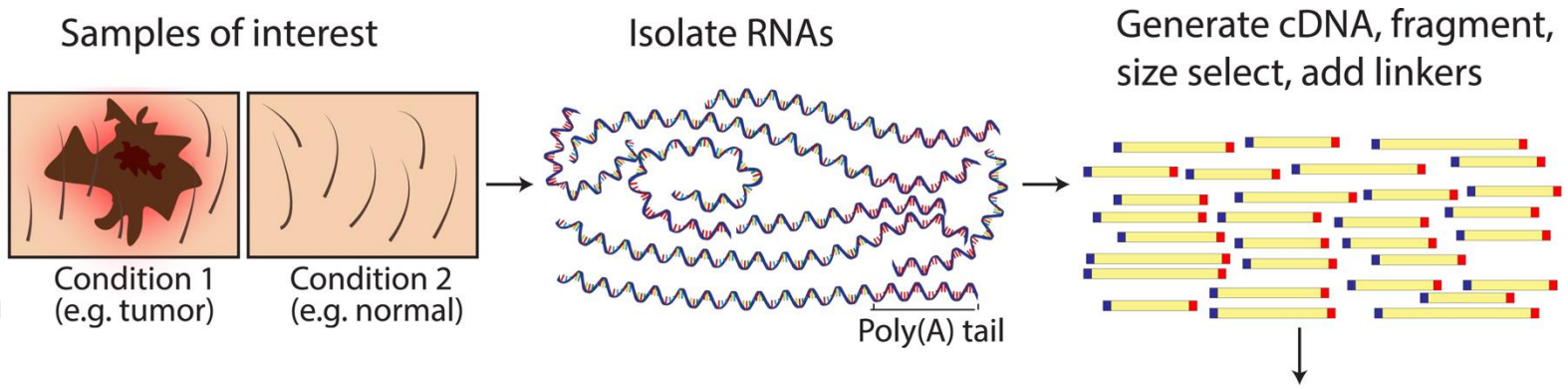
Svensson et al 2018

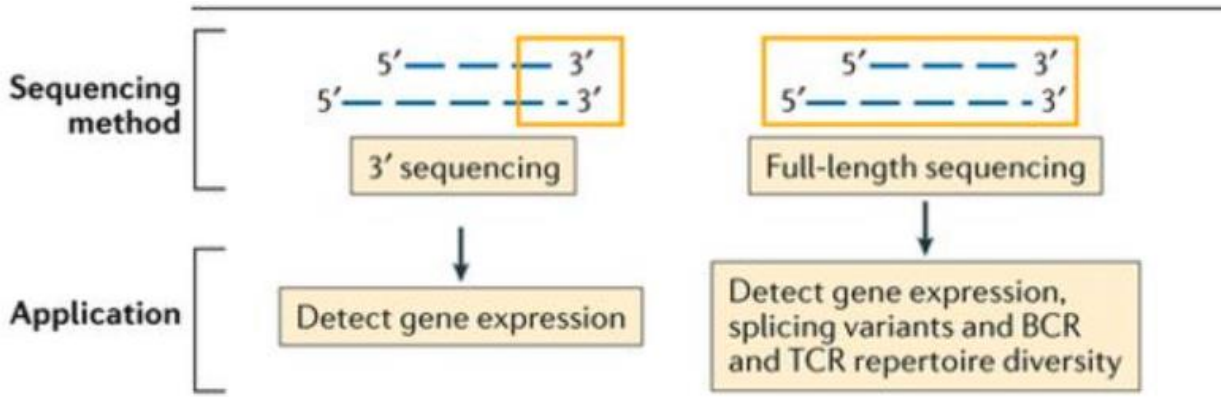


Nature Reviews | Immunology

Important concepts

- Transcript vs read
- Full-length vs 3’/5’ ends
- Library size/sequencing depth
- Library complexity





Library complexity

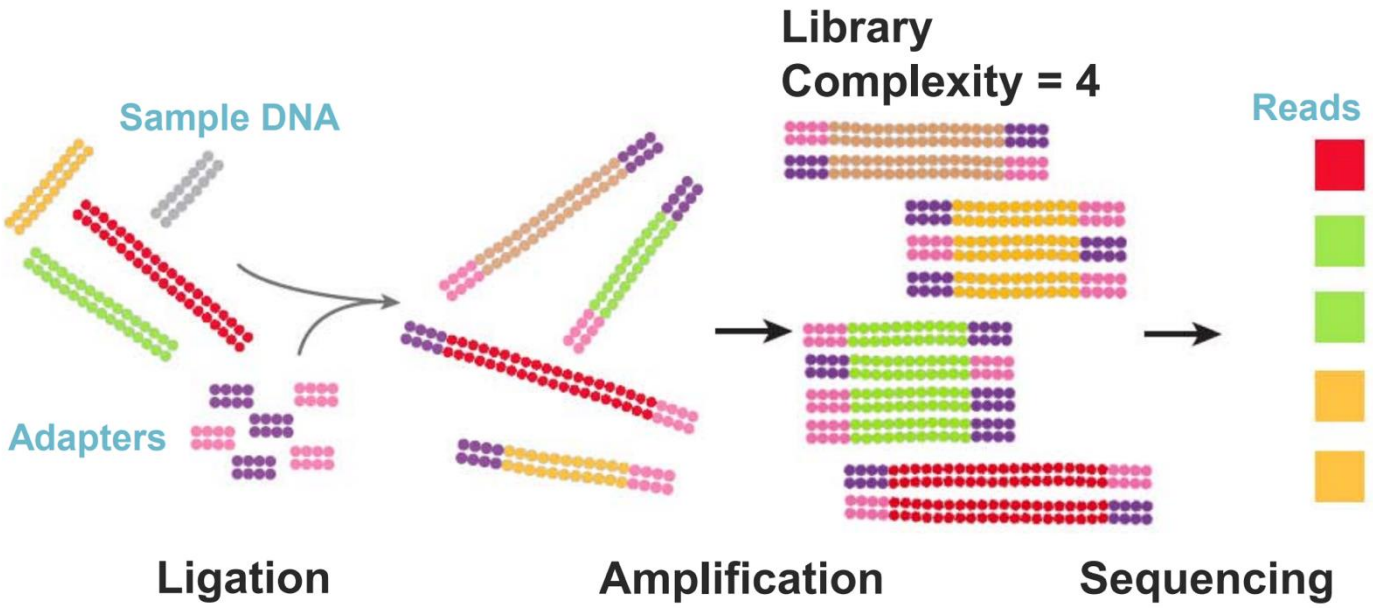
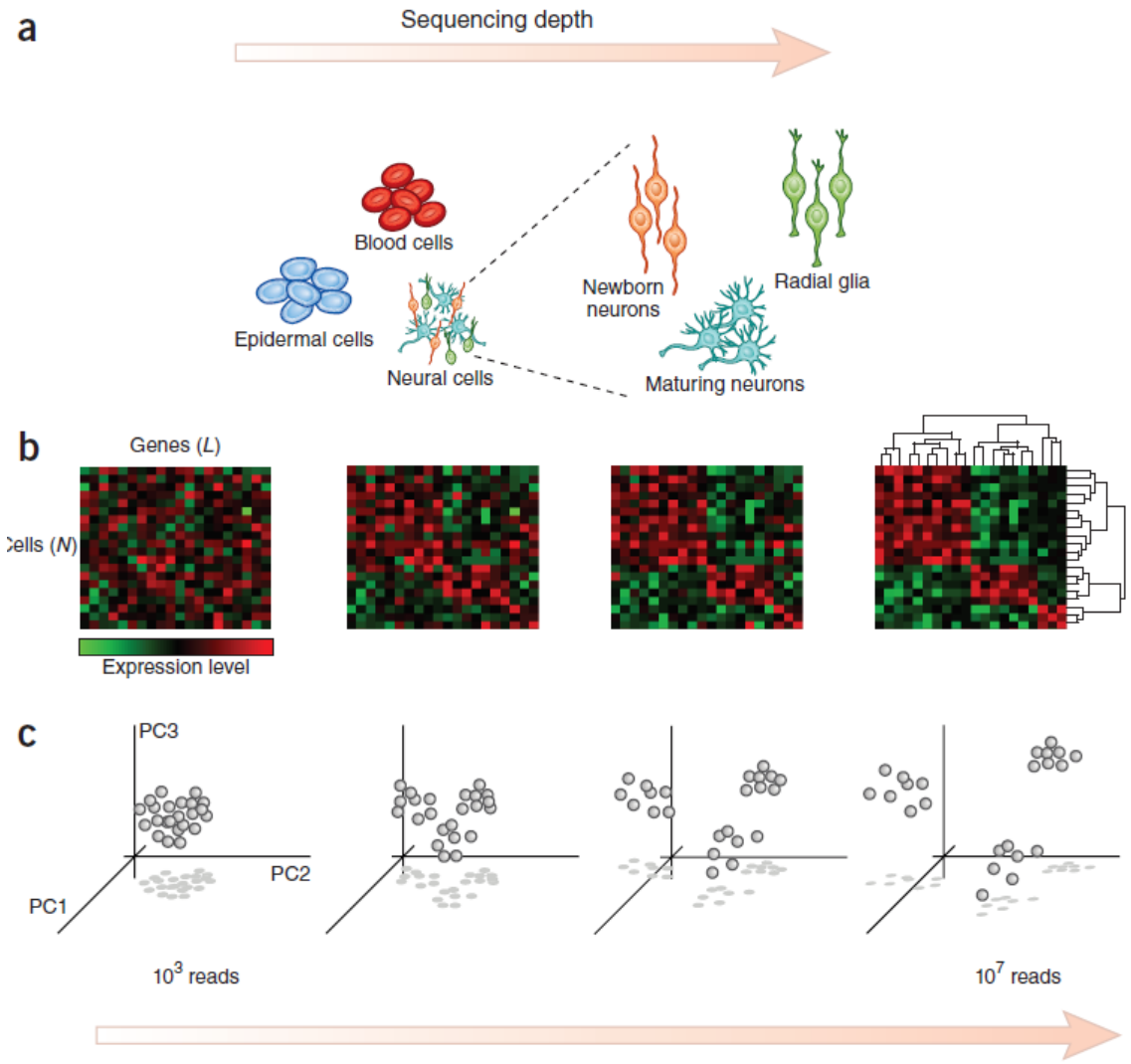


Image adapted from Mardis, *ARGHG* (2008)

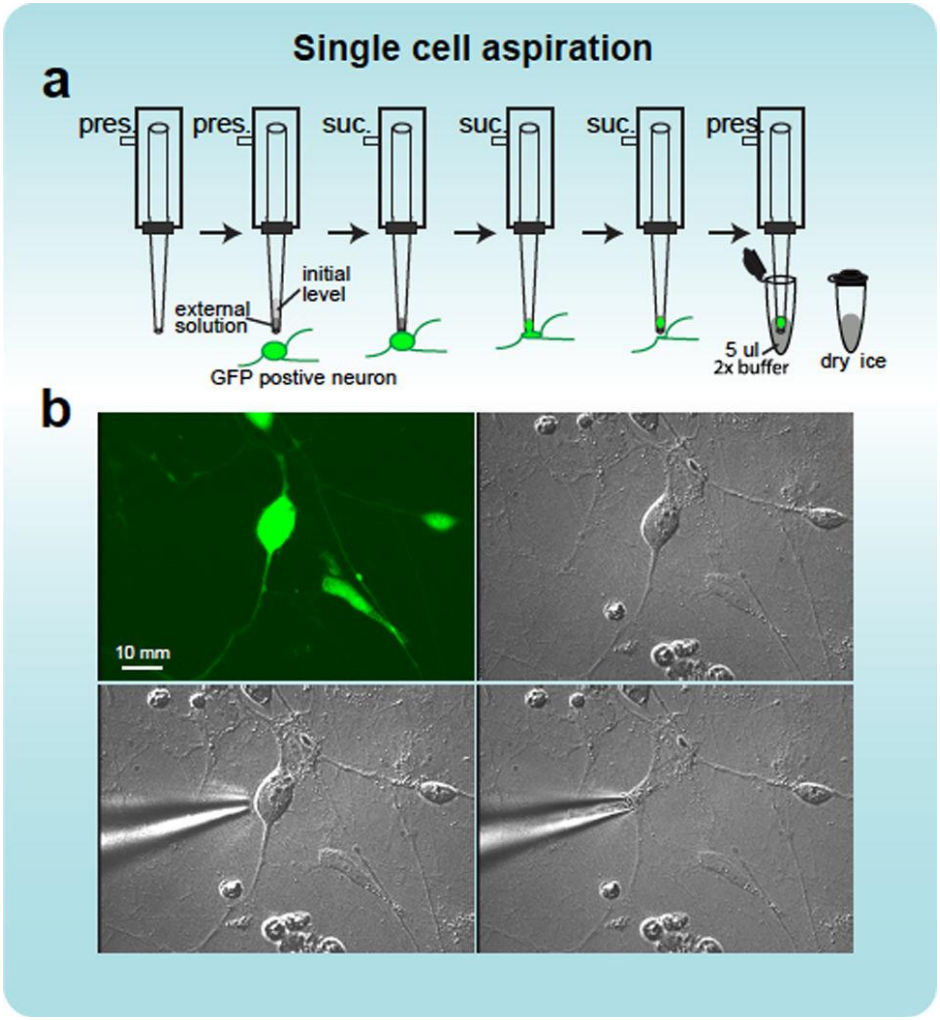
Sequencing depth vs number of cells



Different scRNA-seq protocols with different cell number and sequencing depth capacities can be used in an integrated approach.

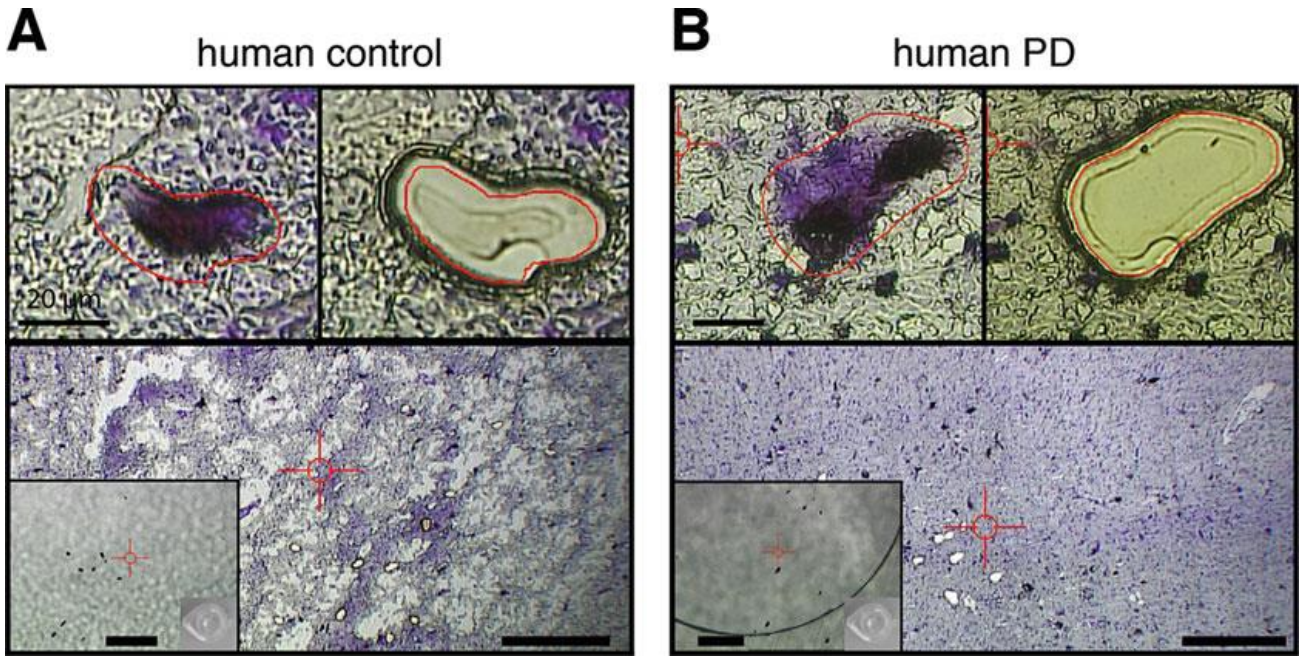
Streets, A. and Huang, Y. (2014). How deep is enough in single-cell RNA-seq?. *Nature Biotechnology*, 32(10), pp.1005-1006

Micromanipulation



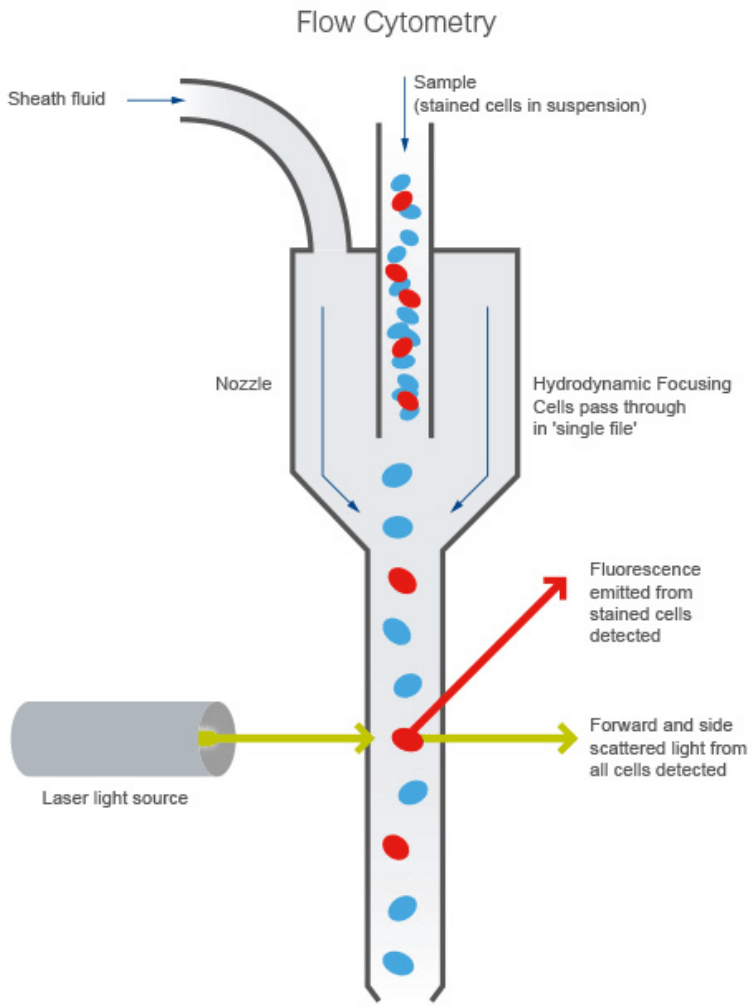
Citri, A., Pang, Z., Südhof, T., Wernig, M. and Malenka, R. (2011). Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nature Protocols*, 7(1), pp.118-127

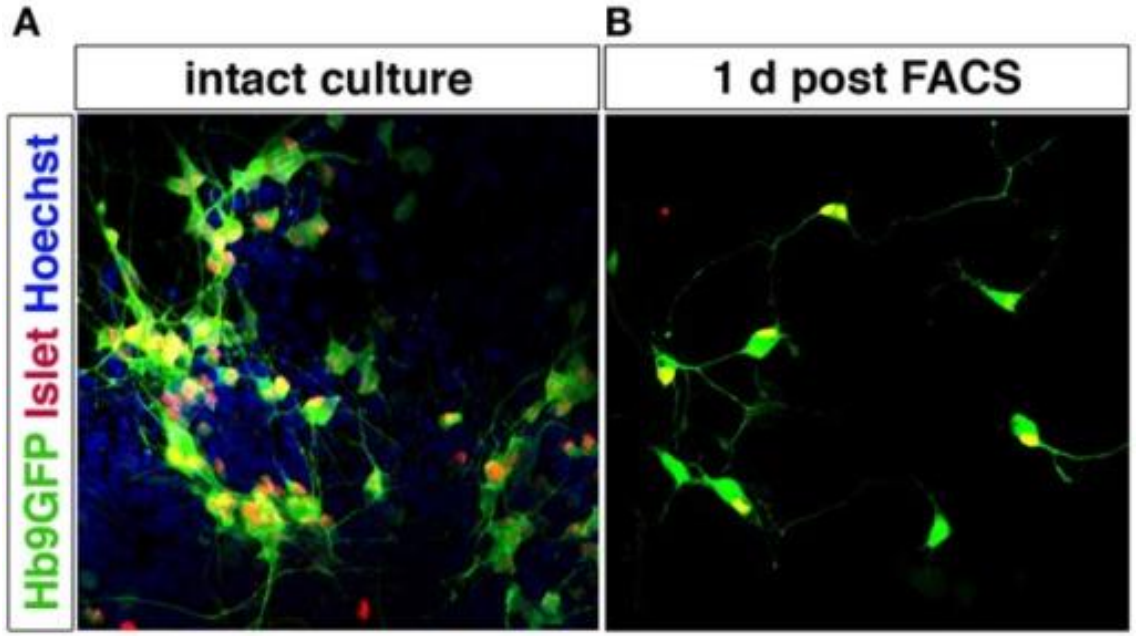
Microdissection



Duda, J., Fauler, M., Gründemann, J. and Liss, B. (2018). Cell-Specific RNA Quantification in Human SN DA Neurons from Heterogeneous Post-mortem Midbrain Samples by UV-Laser Microdissection and RT-qPCR. *Methods in Molecular Biology*, pp.335-360.

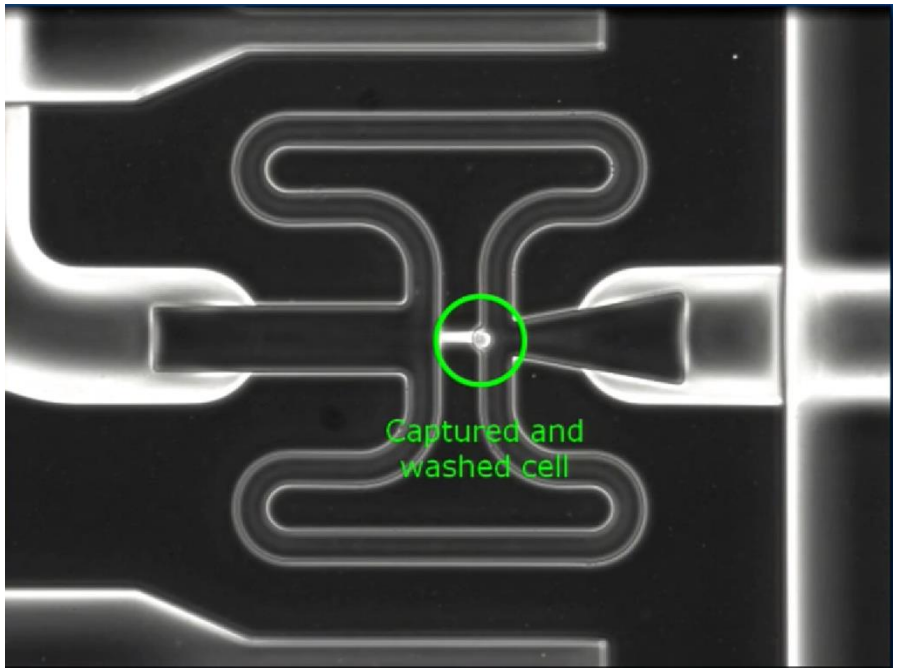
FACS - Fluorescence-activated cell sorting

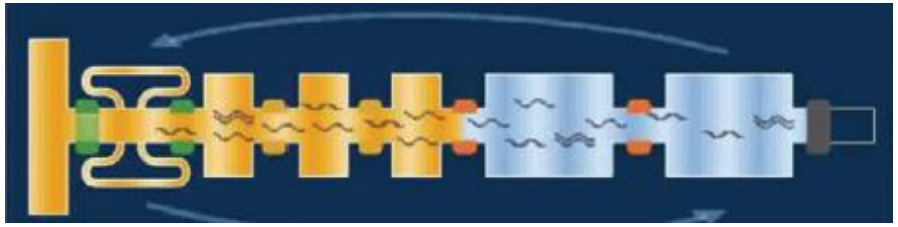


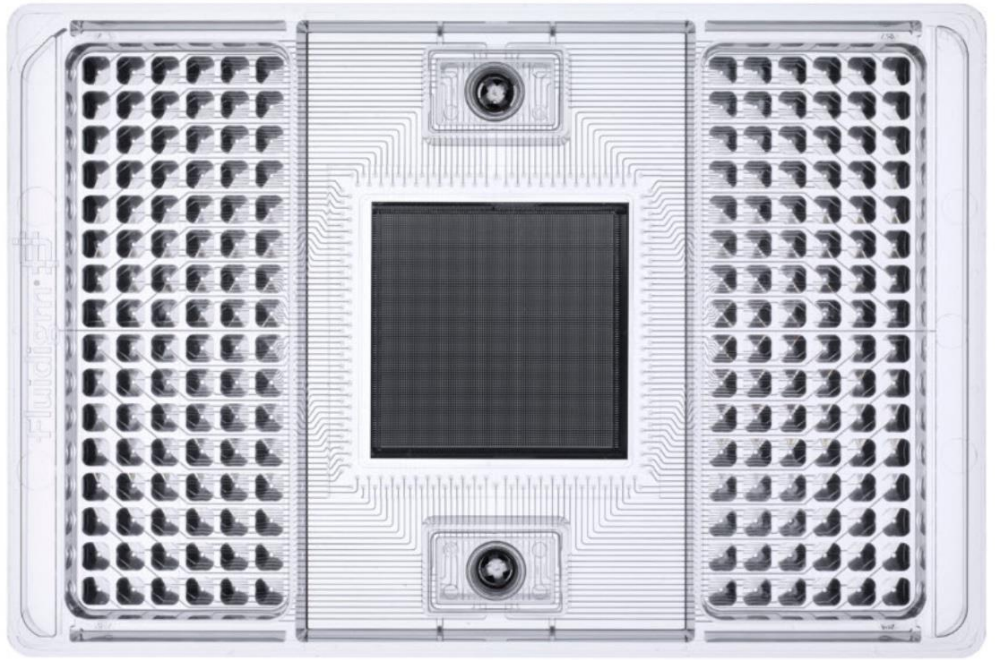


Allodi, I. and Hedlund, E. (2014). Directed midbrain and spinal cord neurogenesis from pluripotent stem cells to model development and disease in a dish. *Frontiers in Neuroscience*, 8.

Microfluidics

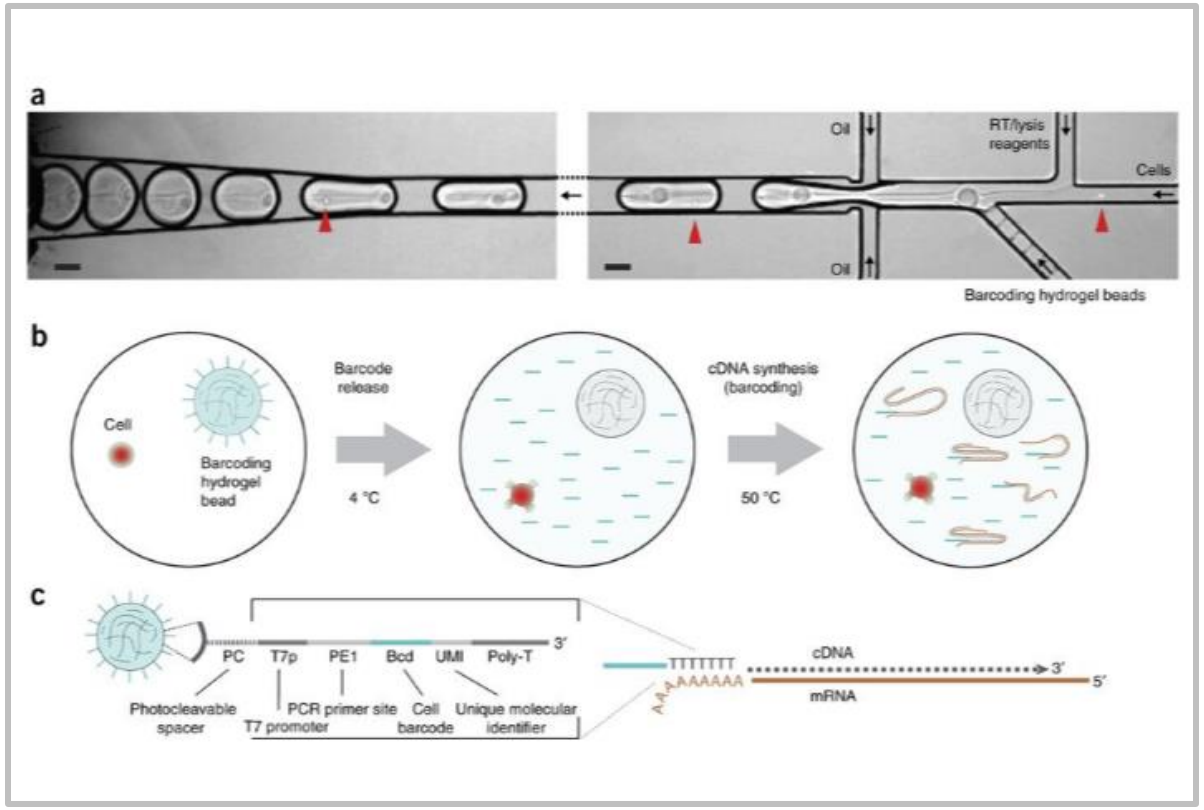




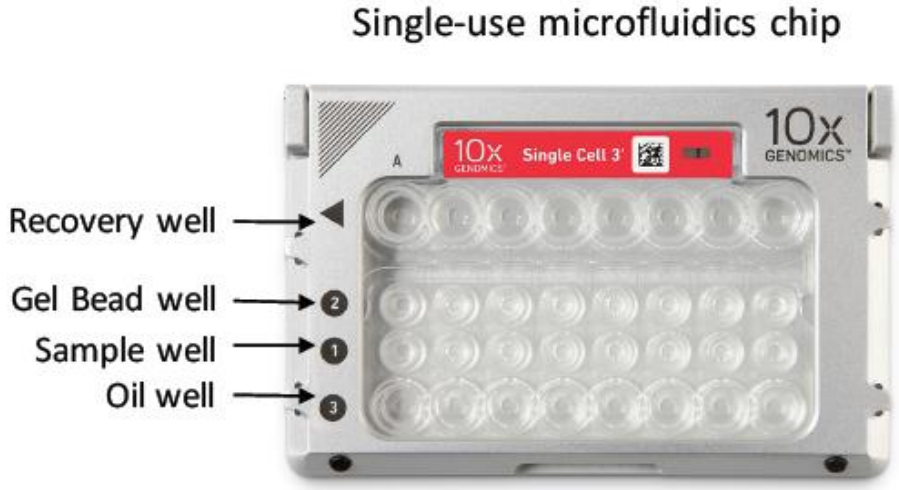


Single-Cell Gene Expression Profiling using TaqMan® Gene Expression Assays with the C1™ Single-Cell Auto Prep System Technical Note

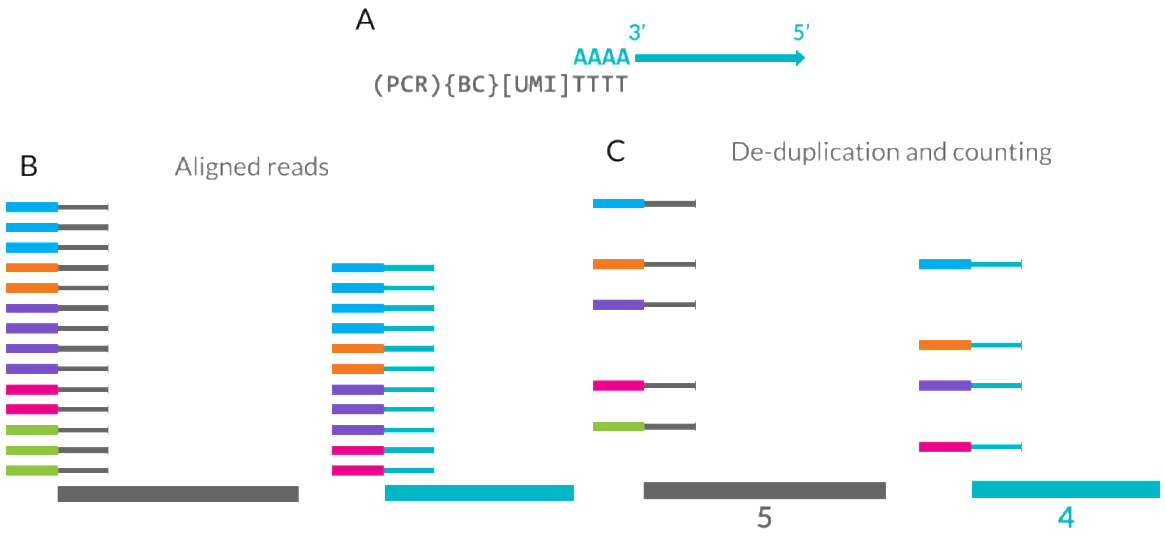
Microdroplet



Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. and Mazutis, L. (2016). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1), pp.44-73.

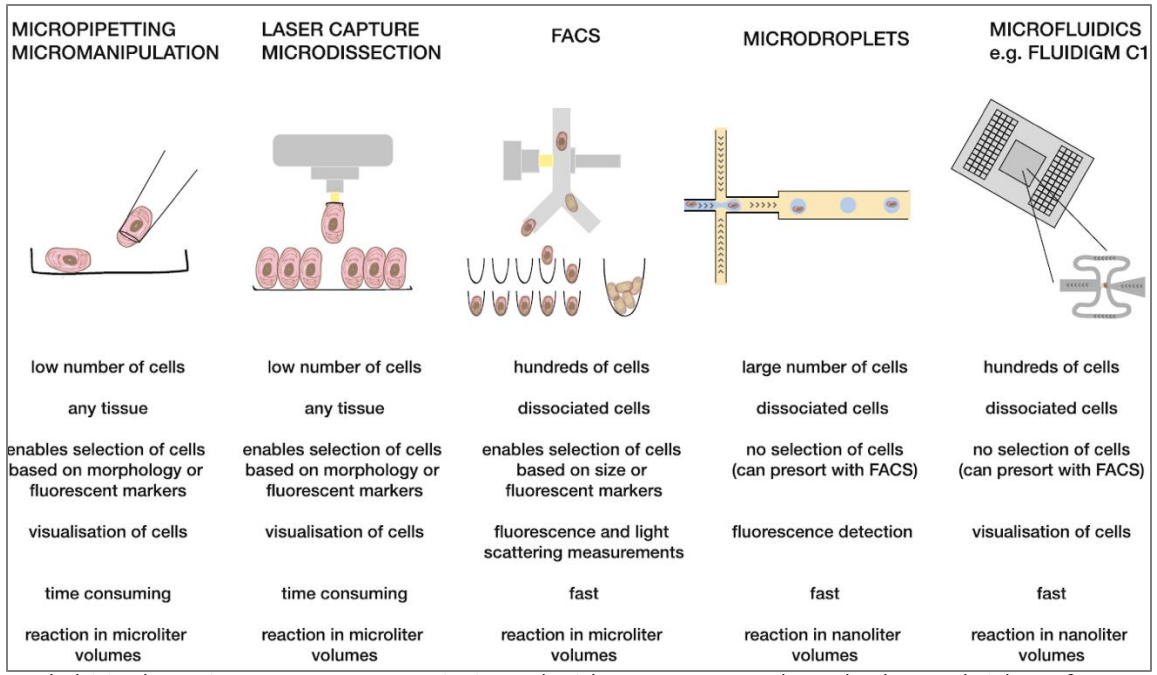


UMIs – unique molecular identifiers



- Small barcodes added to each cDNA during reverse transcription
- Enable removal of amplification noise

Summary

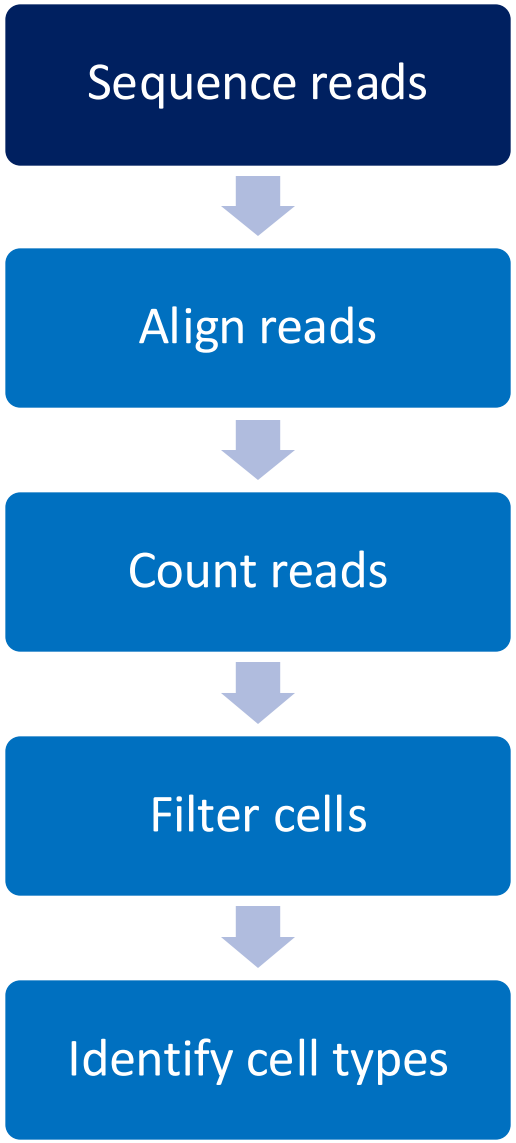


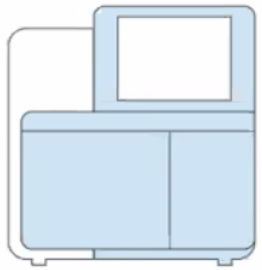
Kolodziejczyk, A., Kim, J., Svensson, V., Marioni, J. and Teichmann, S. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4), pp.610-620.

	FACS	CyTOF	qPCR	Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2)	Fluidigm C1	Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2)	Massively parallel approaches (Drop-seq, InDrop)
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500	48–96	500–2,000	5,000–10,000
Cost	$0.05 per cell	$35 per cell	$1 per cell	$3–6 per well	$35 per cell	$3–6 per well	$0.05 per cell
Sensitivity	Up to 17 markers	Up to 40 markers	10–30 genes per cell	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	6,000–9,00 genes per cell for cell lines; 1,000–5,000 genes per cell for primary cells	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells

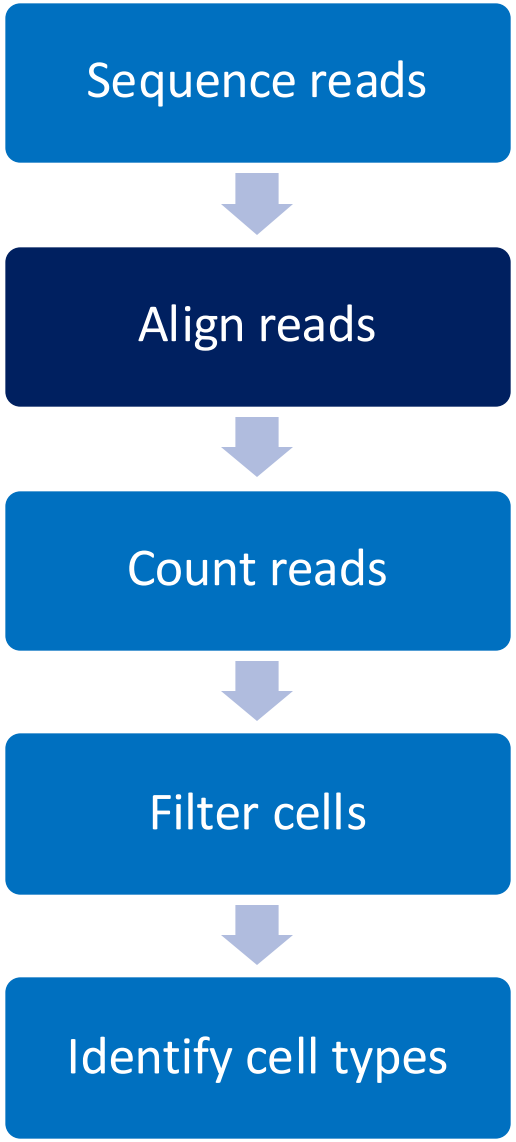
Papalexi, E. and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1), pp.35-45.

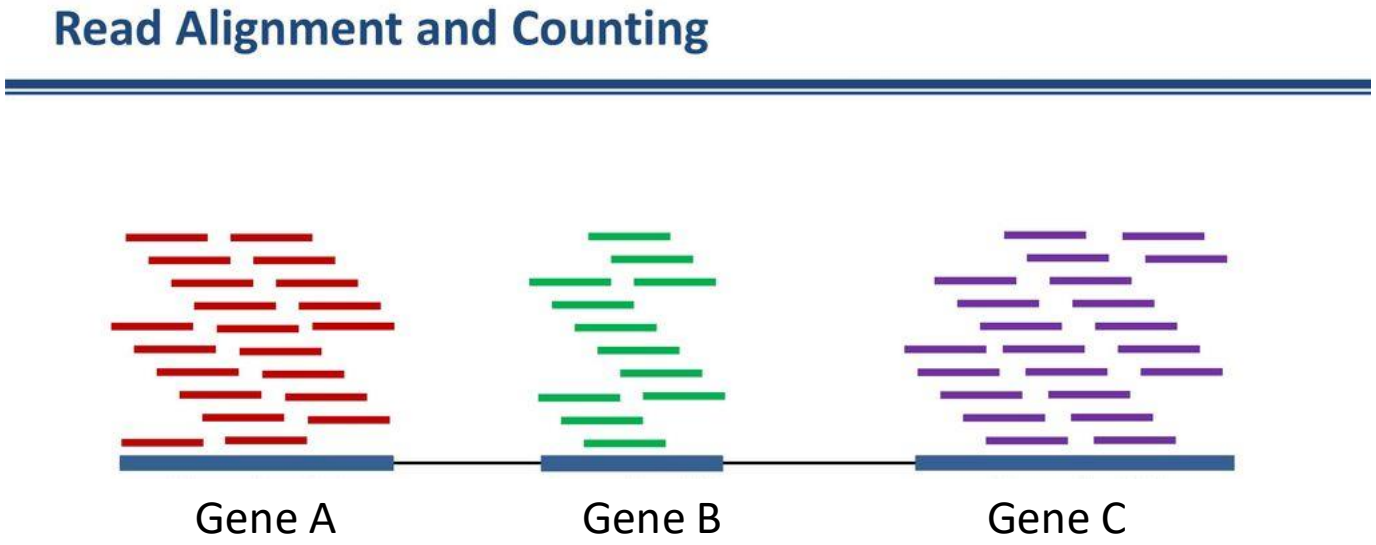
Computational analysis



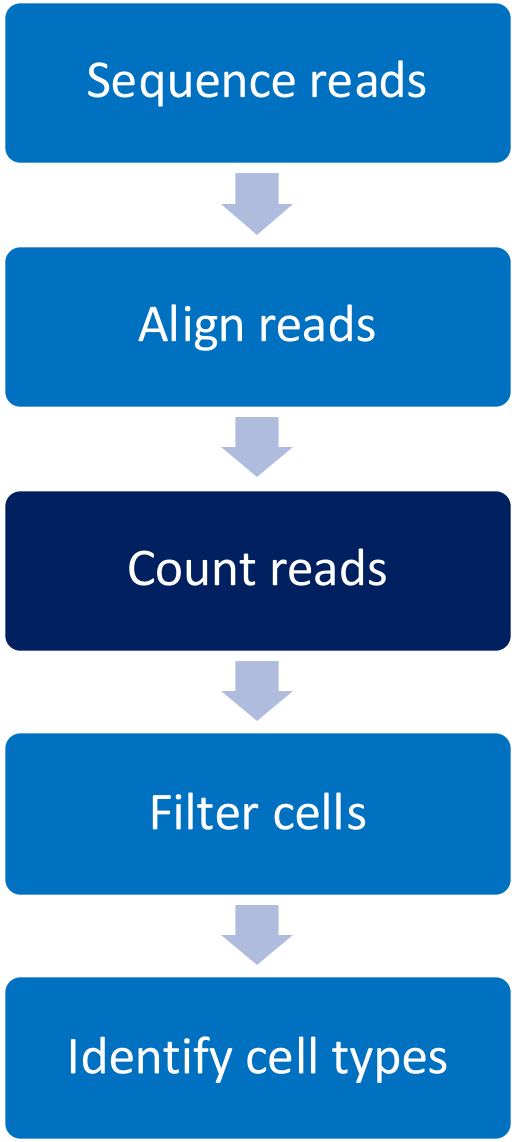


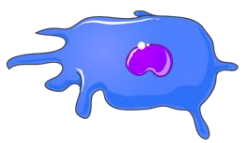
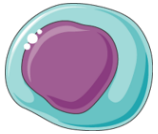
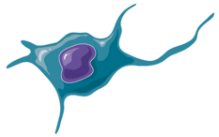
Computational analysis





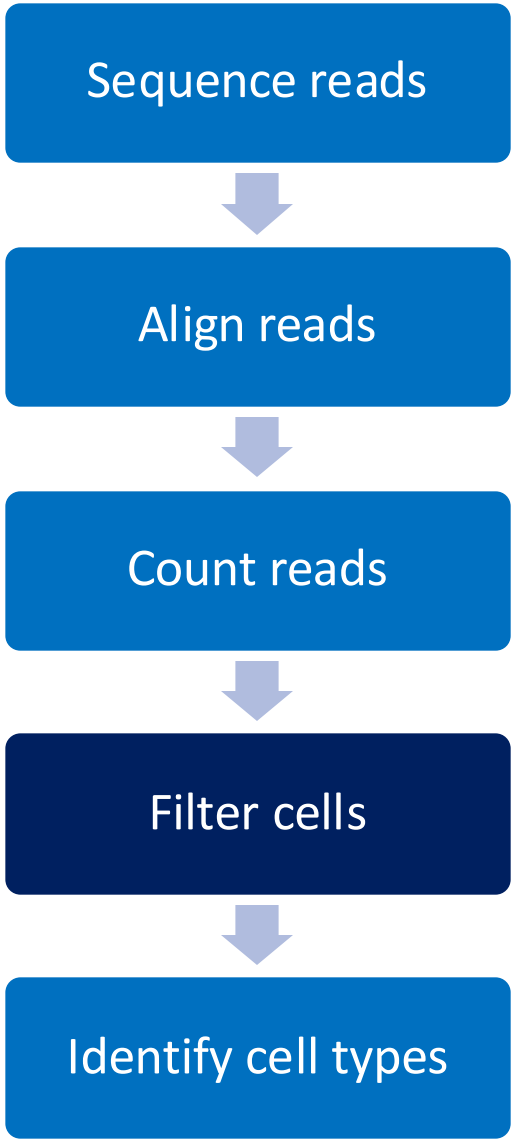
Computational analysis

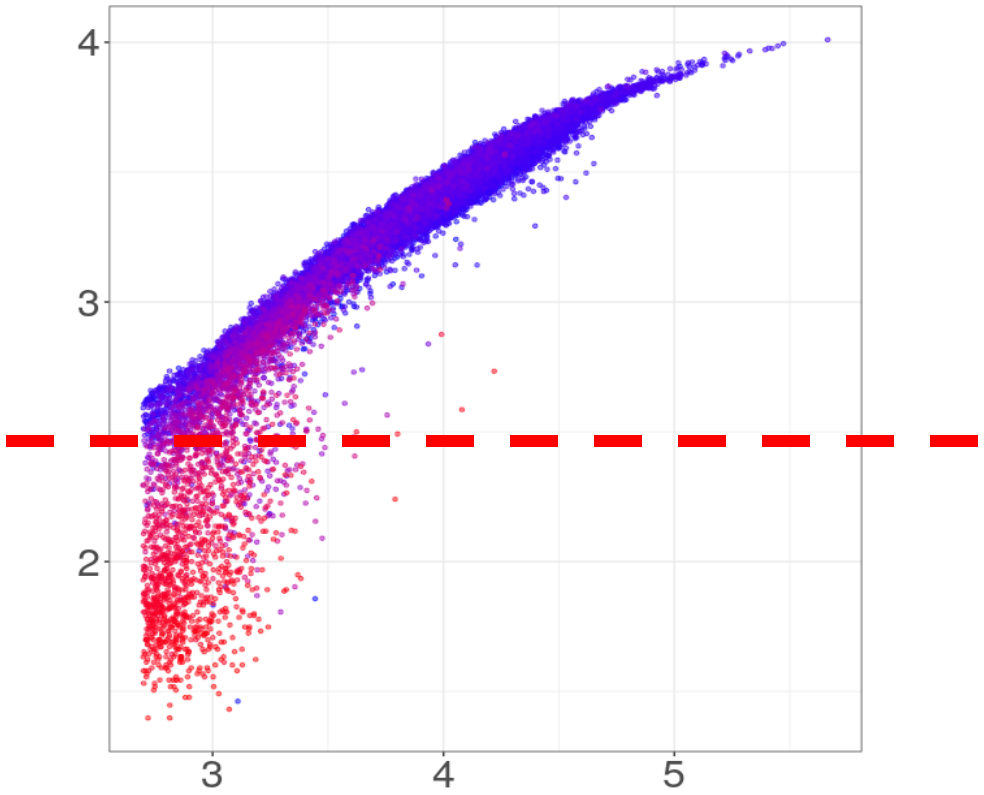


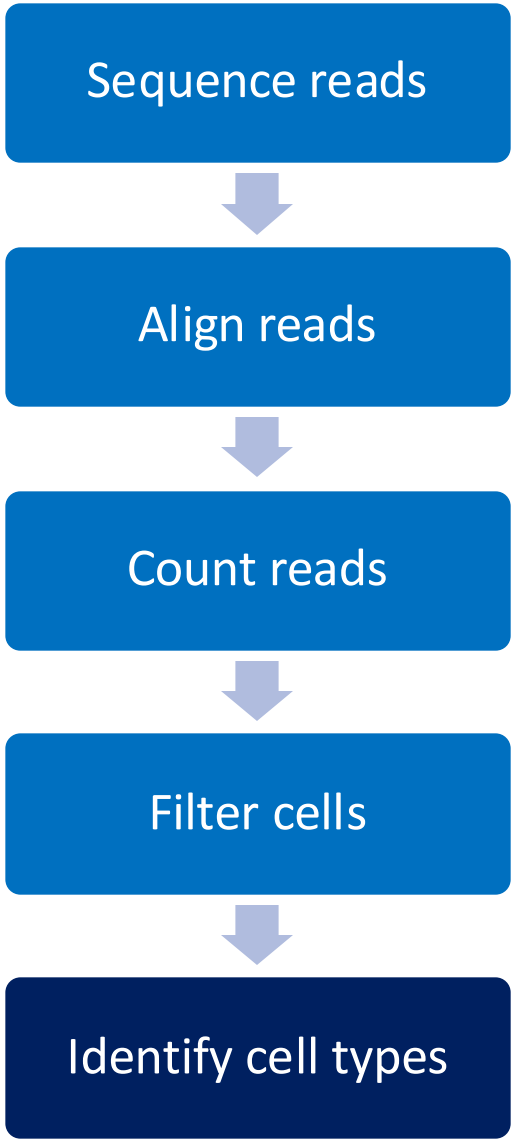
Genes	Cell 1	Cell 2	Cell 3	…
A	0	0	0	…
B	0	0	1	…
C	0	10	2	…
D	5	0	0	…
E	1	3	0	…
F	0	0	0	…
G	2	1	1	…
H	0	0	0	…
I	0	22	5	…
…	…	…	…	…

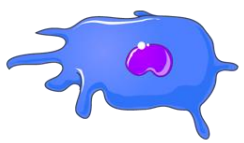
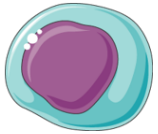
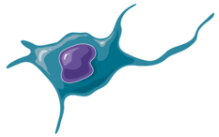
Computational analysis





Computational analysis



Genes	Macrophage	T cell	Dendritic cell	…
A	0	0	0	…
B	0	0	1	…
C	0	10	2	…
D	0	0	0	…
E	1	3	0	…
F	0	0	0	…
G	2	1	1	…
H	0	0	0	…
I	0	22	5	…
…	…	…	…	…

The complexity of analyzing scRNA-seq data

- Large volume of data
- Gene dropouts
  - a gene is observed at a low or moderate expression level in one cell but is not detected in another cell of the same cell type

The complexity of analyzing scRNA-seq data

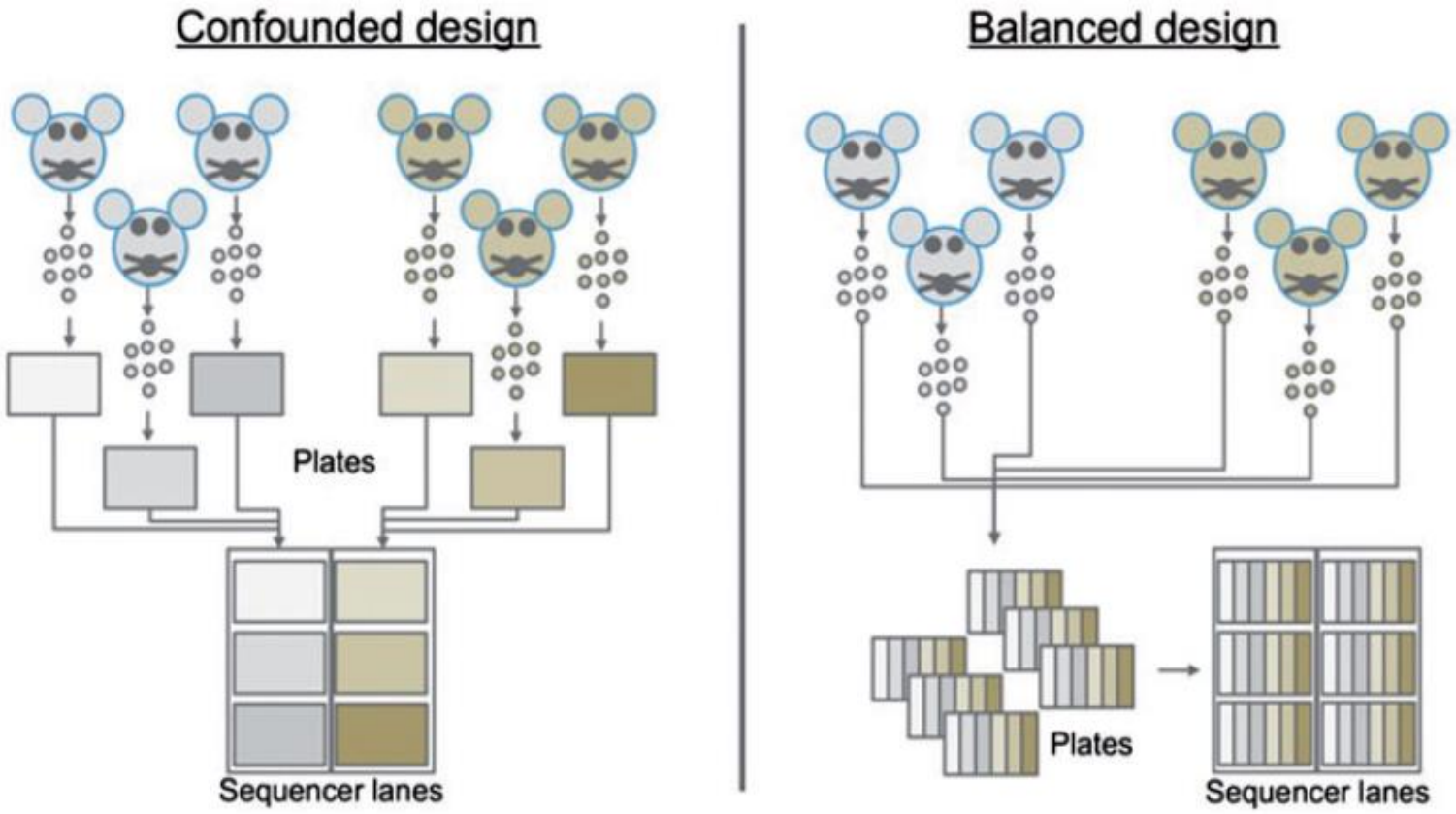
- Large volume of data
- Gene dropouts
  - a gene is observed at a low or moderate expression level in one cell but is not detected in another cell of the same cell type
- Technical variability across cells/samples (batch effect)

What is batch effect?

Batch Effect

Random technical artefacts which occur during handling/processing

- Samples were prepared in different labs
- Samples were prepared in the same lab but at different times
- Samples were processed on different plates/runs

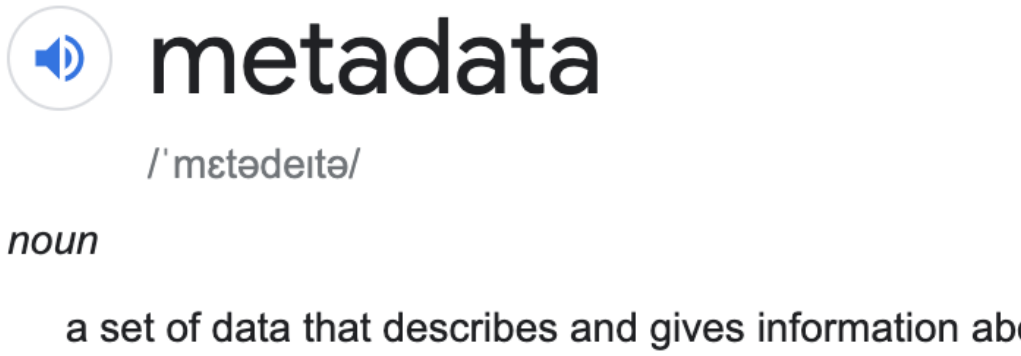


Sort cells from different biological conditions into different wells of the same plate.

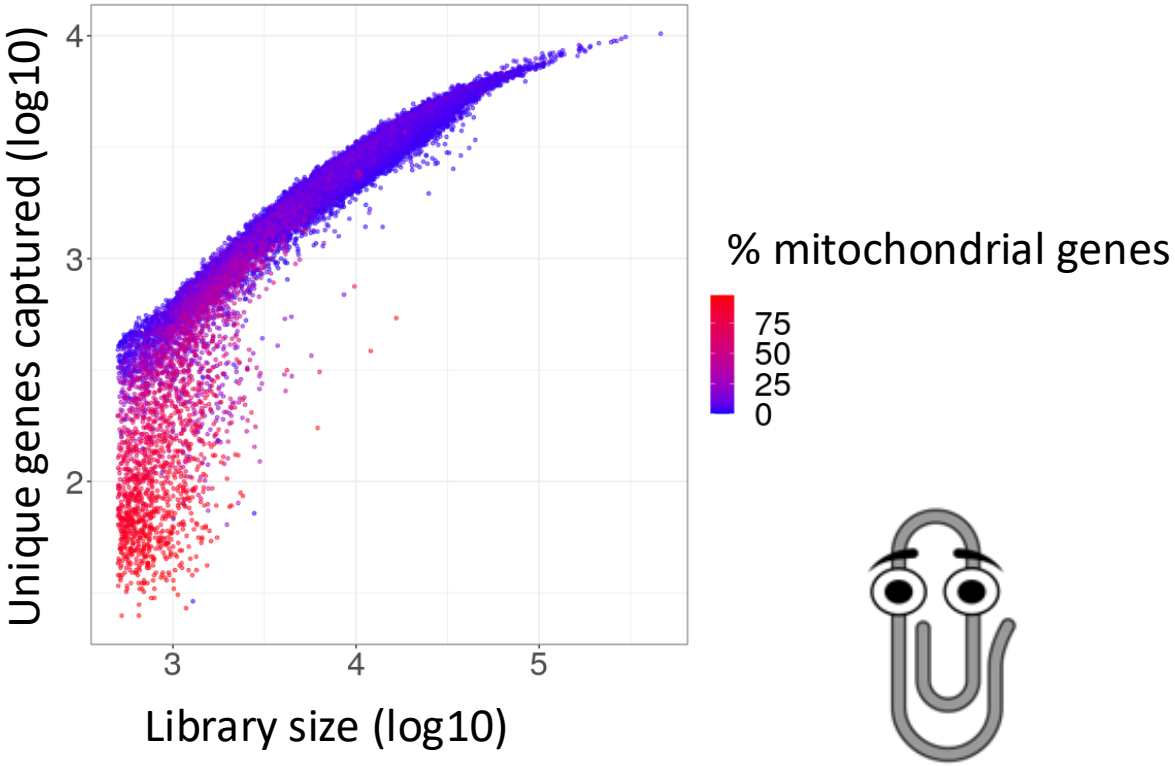
Importance of Metadata (Bioinformaticians aren't psychic)

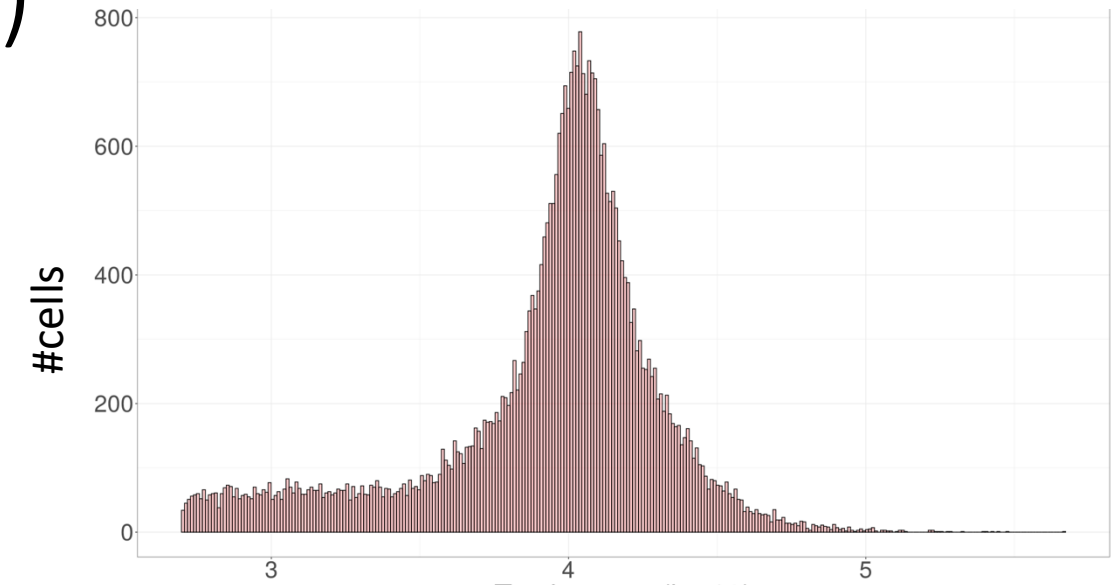
Any metadata that can be associated to an individual cell or a sample/batch can be used to check and/or correct the final data.

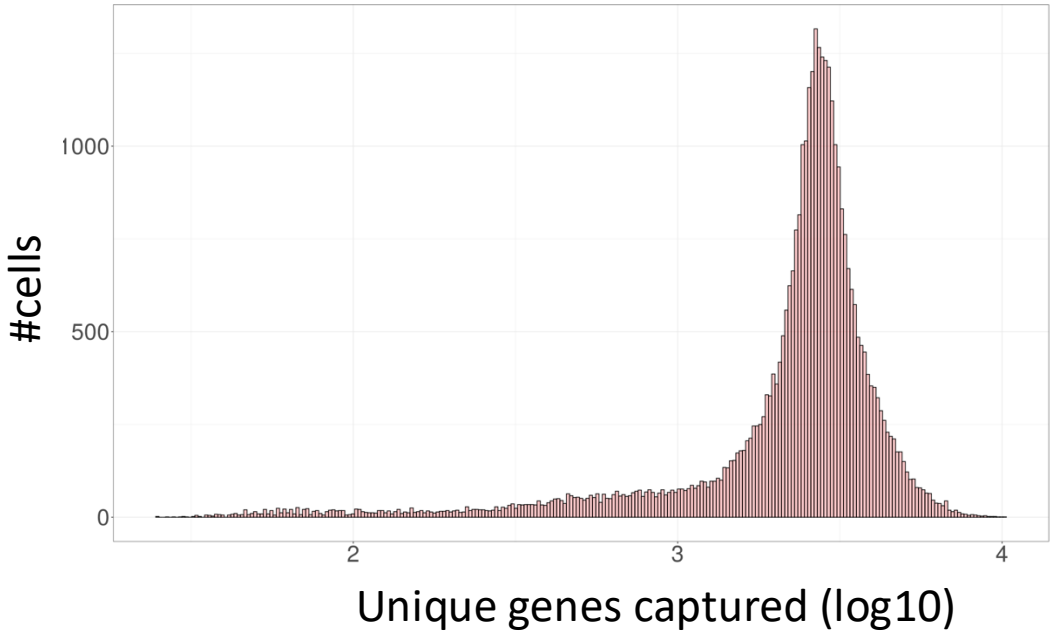
Computational biologists cannot tell you what went wrong just from the sequencing data!



Quality control (cell filtering)







Percentage of mitochondrial genes

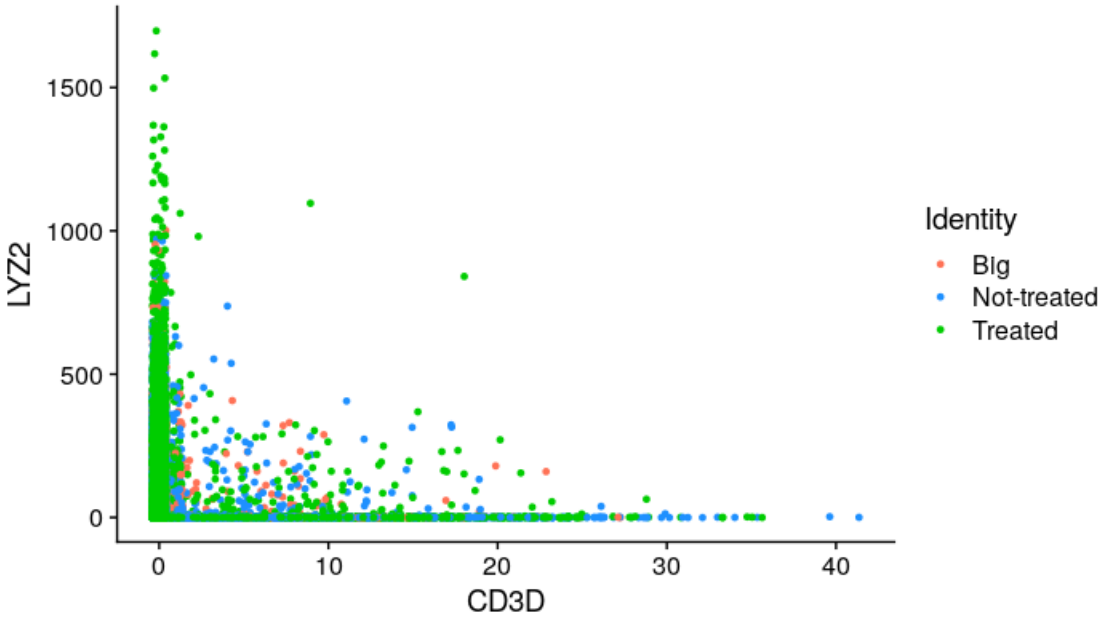
- Lysed cells lose cytoplasmic RNA, while the mitochondrial transcripts remain within the intact mitochondria.
- Apoptotic cells express mitochondrial genes and export these transcripts to the cytoplasm.

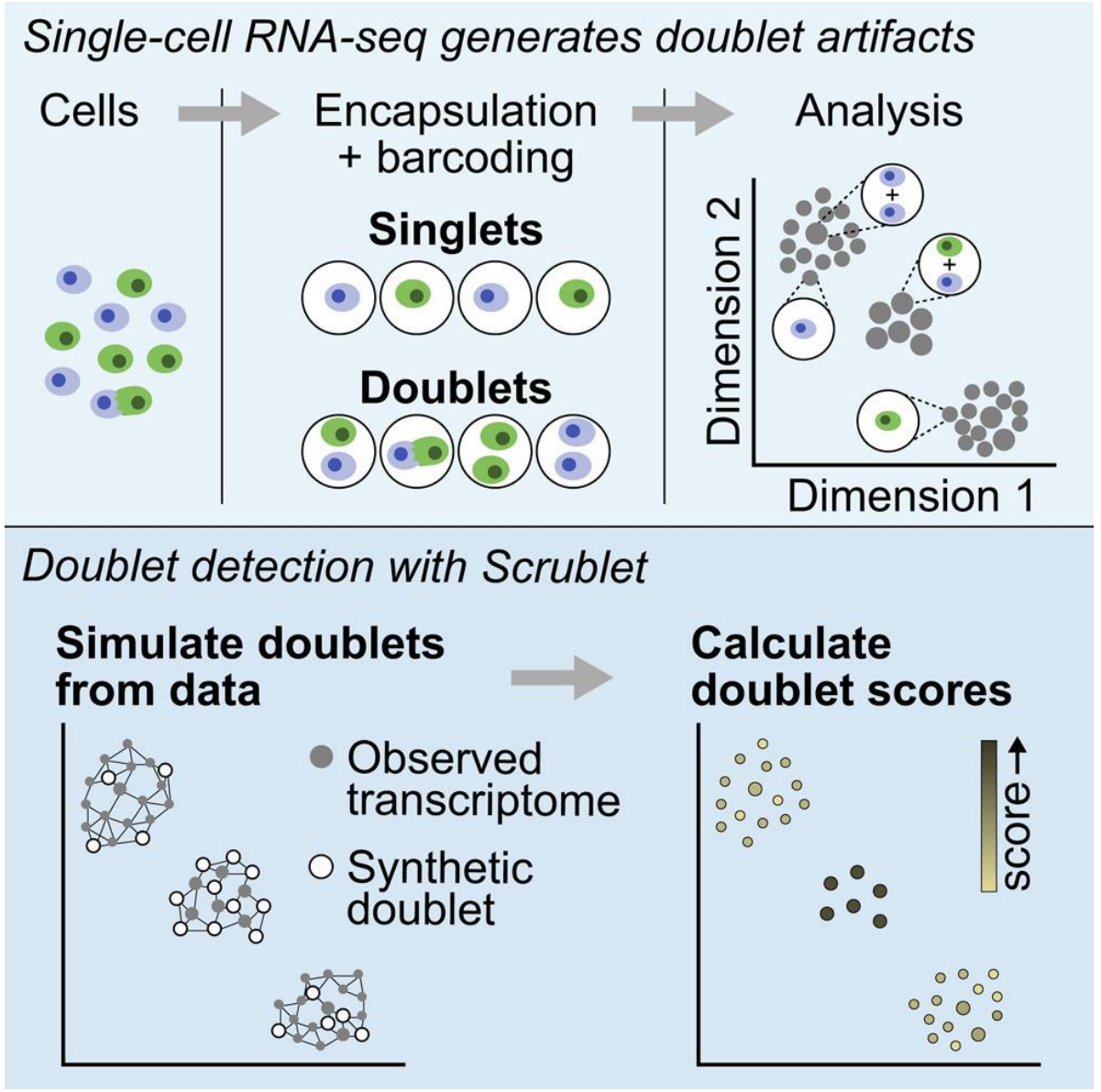
Cells that show an increased expression of mitochondrial genes likely represent a group of dying cells.

Looking for doublets

- Marker genes
- Synthetic doublets

E.g., cells with T cell and macrophage markers?

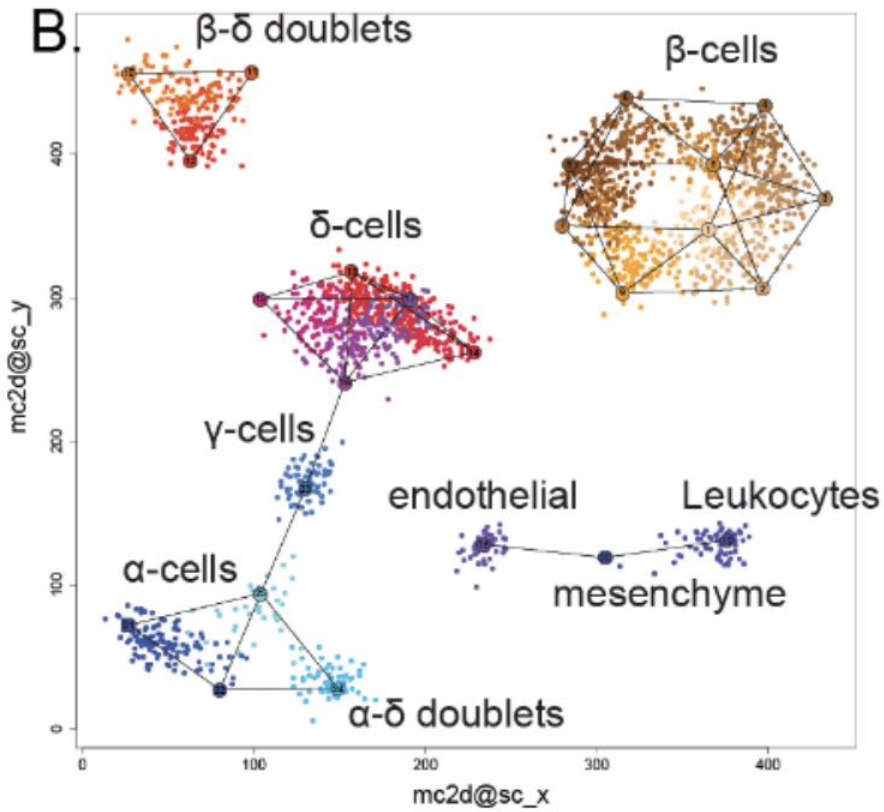




Physically interacting beta-delta pairs in the regenerating pancreas revealed by single-cell sequencing

Eran Yanowski, Nancy-Sarah Yacovzada, Eyal David, Amir Giladi, Diego Jaitin, Lydia Farack, Adi Egozi, Danny Ben-Zvi, Shalev Itzkovitz, Ido Amit, Eran Hornstein

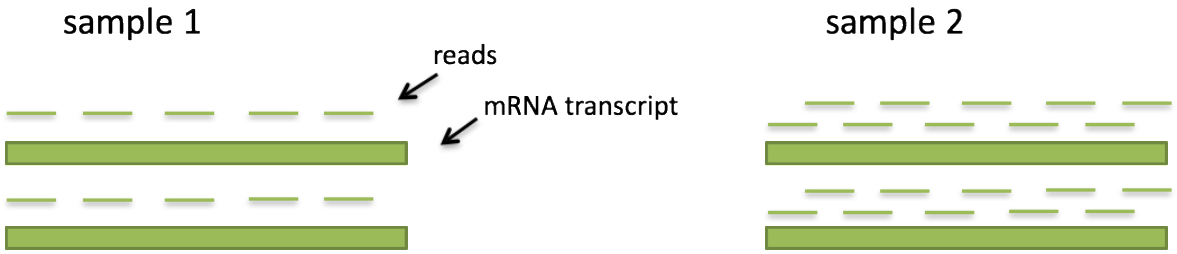
doi: https://doi.org/10.1101/2021.02.22.432216

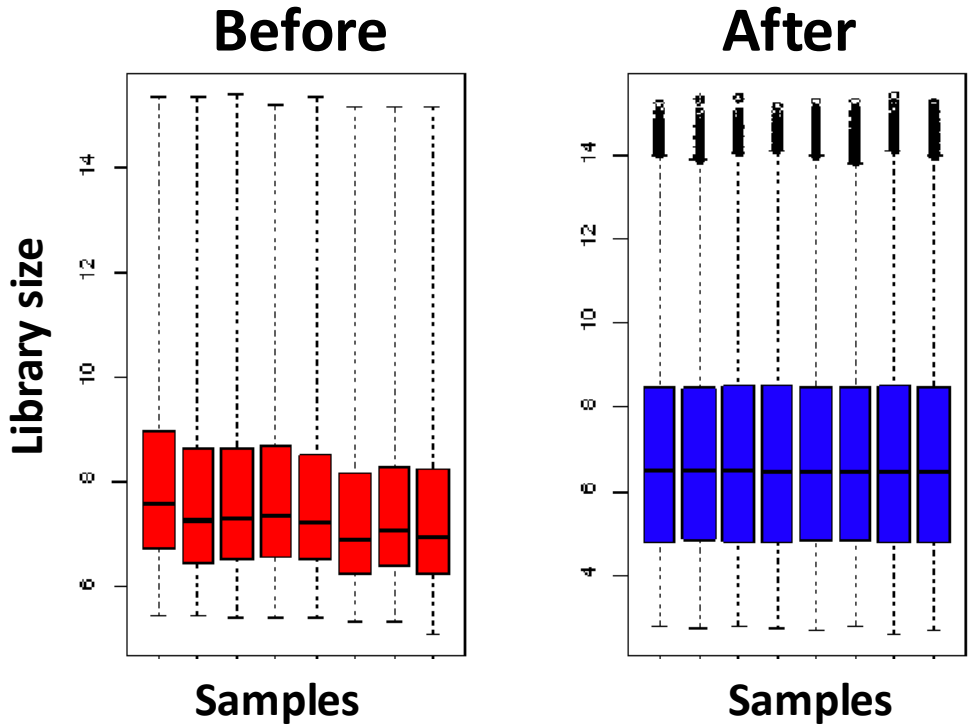


- Unexpectedly identified a relatively large number of beta-delta pairs and alpha-delta pairs (doublets) that is untypical for that protocol;
- Demonstrated that the pairing of delta to beta cells is specific and is significantly enriched, relative to other endocrine cell couples
- Overall, the study reveals genuine pairs of beta and delta cells in the endocrine pancreas.

Why do we need normalization?

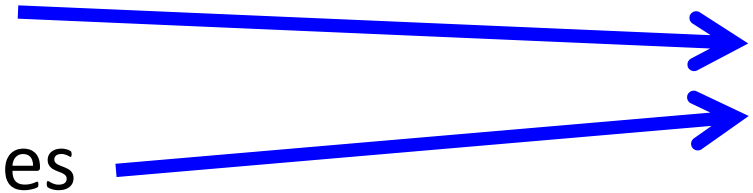
Different sequencing depth





How do we summarize/make sense of big data?

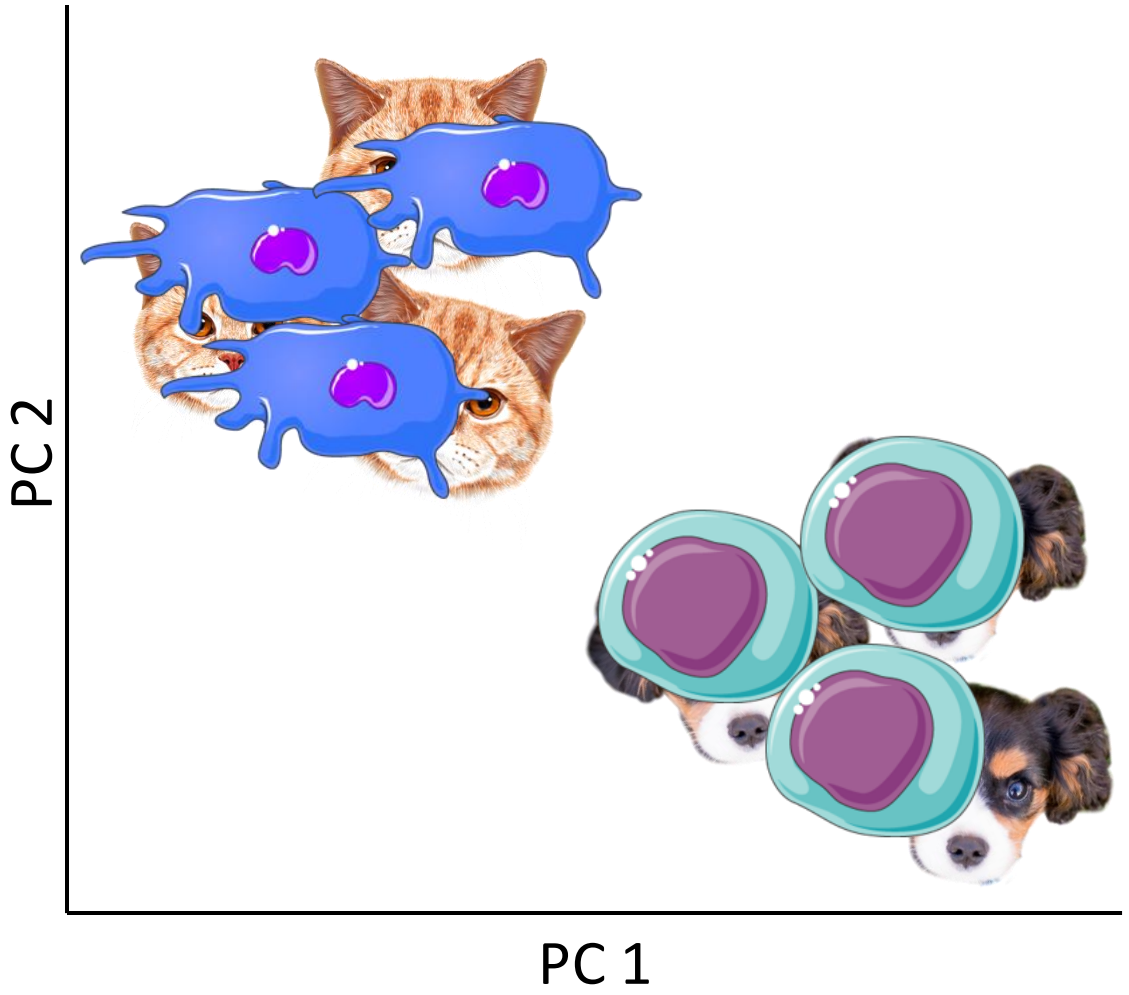
- **Dimension reduction**: removing redundancy from many often correlated variables to give a smaller and more manageable set
- **Clustering**: discovering classes
- **Gene sets**: meaningfully grouping genes to reduce the number of analyzed features, add biological significance and additively increasing sensitivity for detecting changes



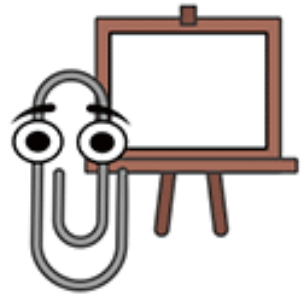
Visual Analytics
(i.e. plots)

Visualization: dimensionality reduction

Principal Component Analysis

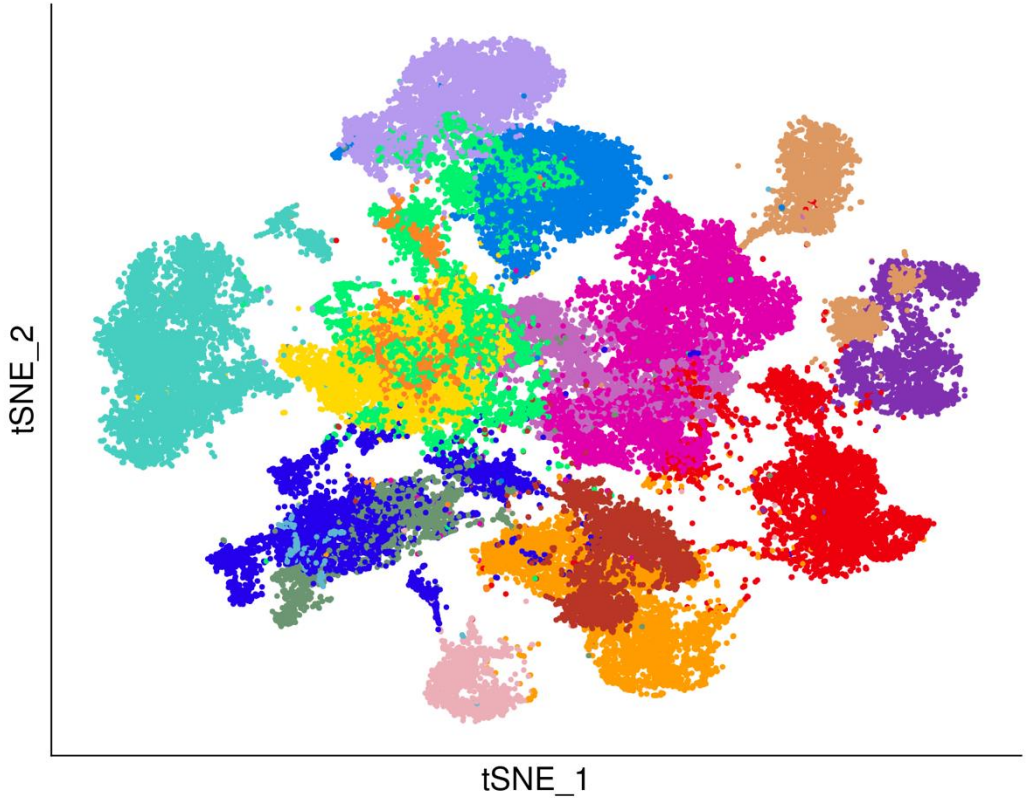


Emphasize **variation** and bring out strong patterns in a dataset.

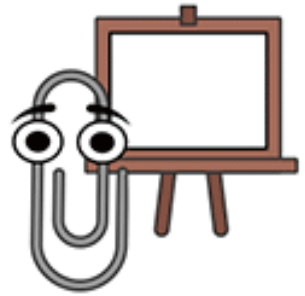


Visualization: dimensionality reduction & clustering

t-SNE: t-distributed stochastic neighbor embedding



Finds patterns in the data based on **similarity** of data points with genes (distinct groups).



Visualization: dimensionality reduction & clustering

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

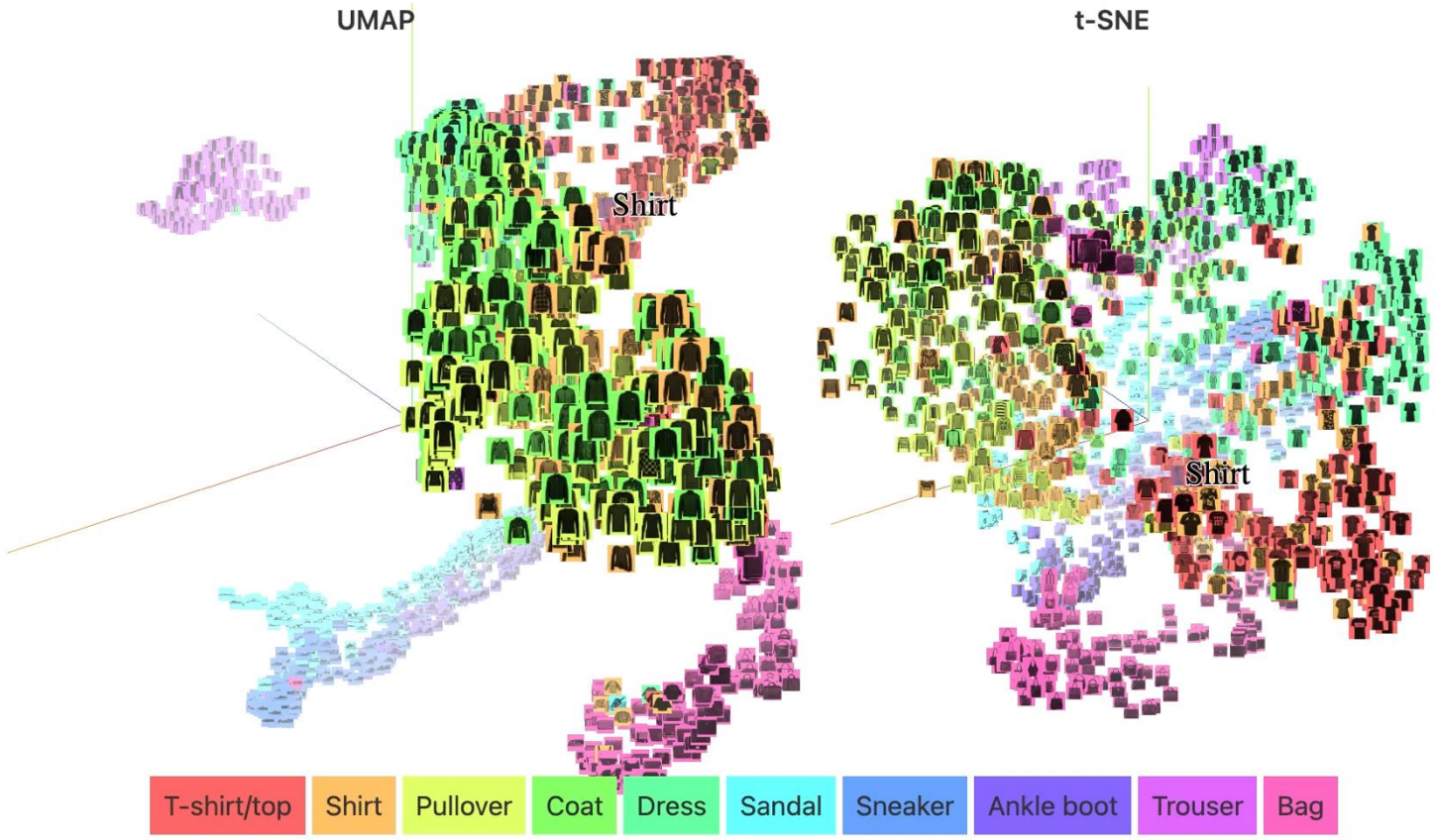
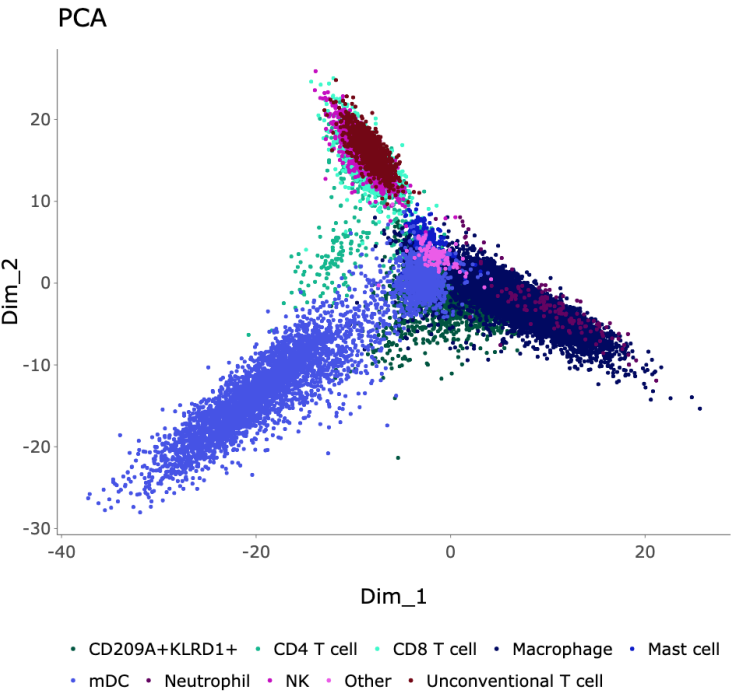
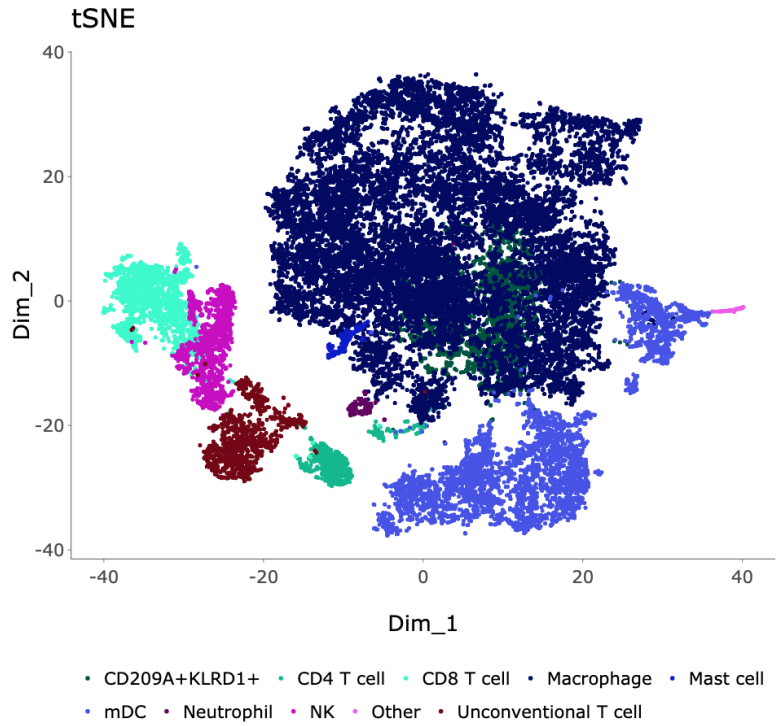
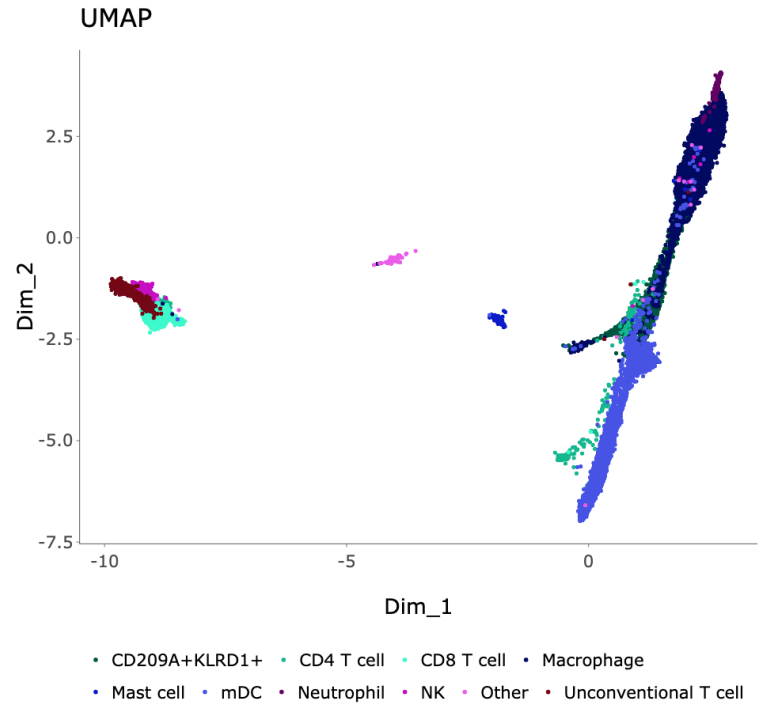


Figure 2: *Dimensionality reduction applied to the Fashion MNIST dataset. 28x28 images of clothing items in 10 categories are encoded as 784-dimensional vectors and then projected to 3 using UMAP and t-SNE.*





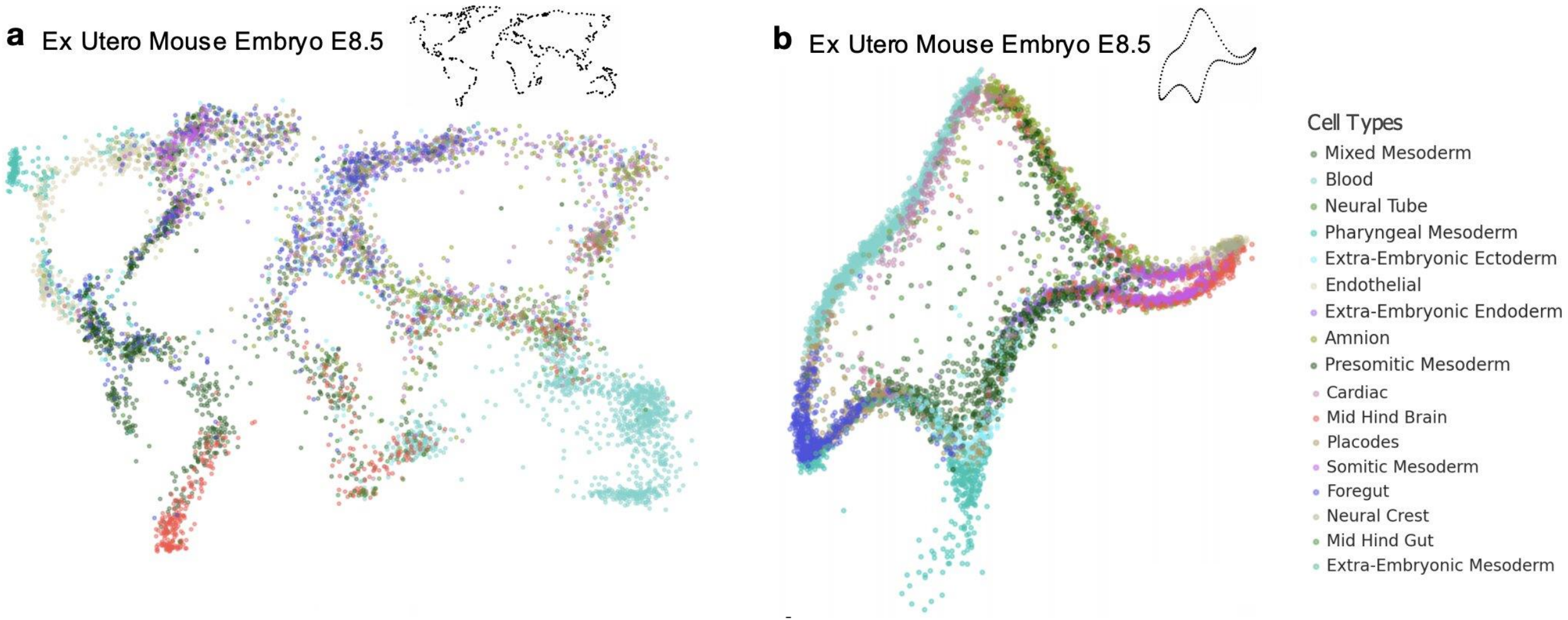


← **Sequência**

Lior Pachter
@lpachter

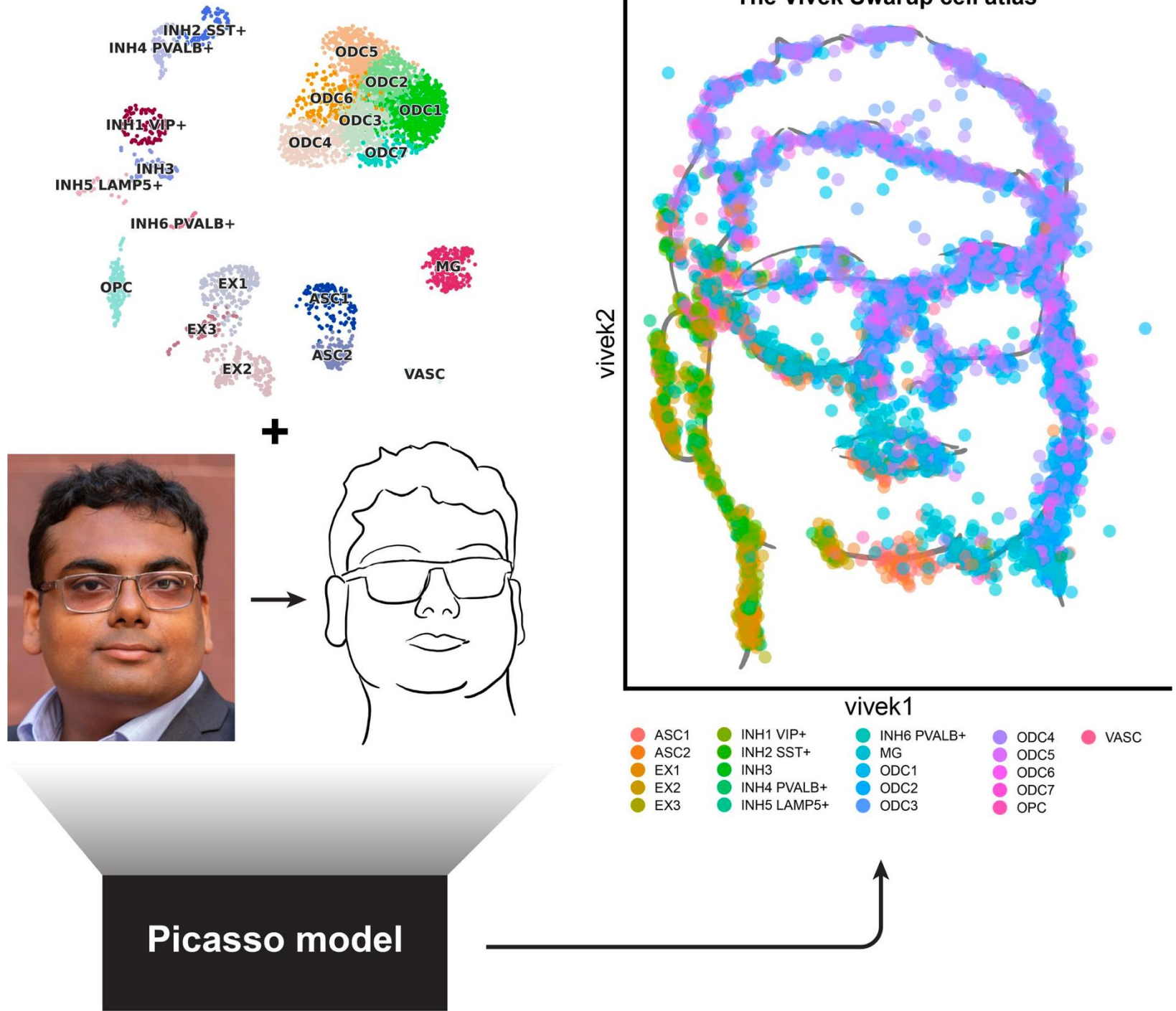
It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary.🧵
biorxiv.org/content/10.110...

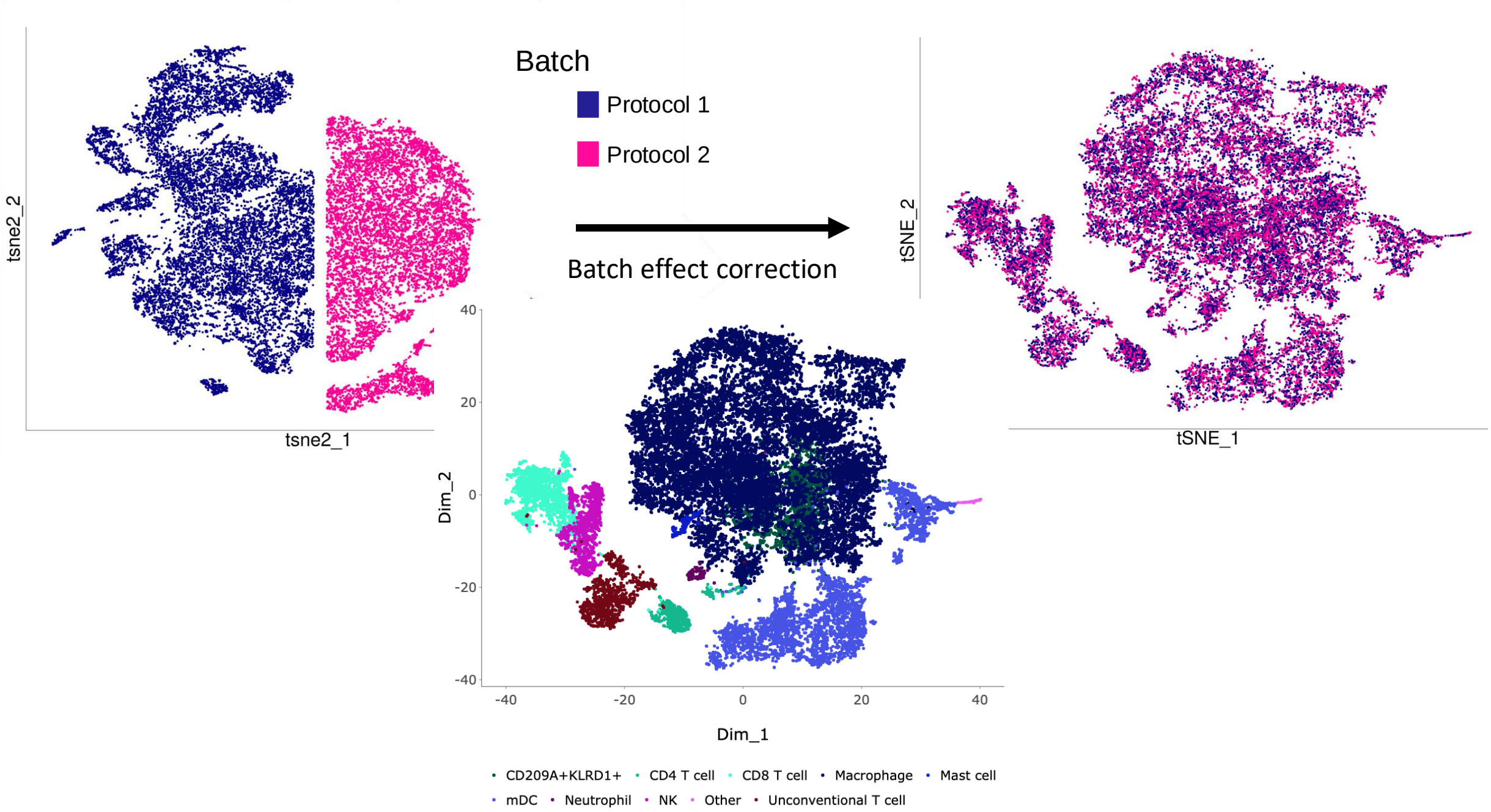
Traduzir Tweet



Lior Pachter @lpachter · 27 de ago de 2021

Picasso can produce quantitatively similar plots that are qualitatively very different- here is the same dataset as a world map and as von Neumann's elephant:

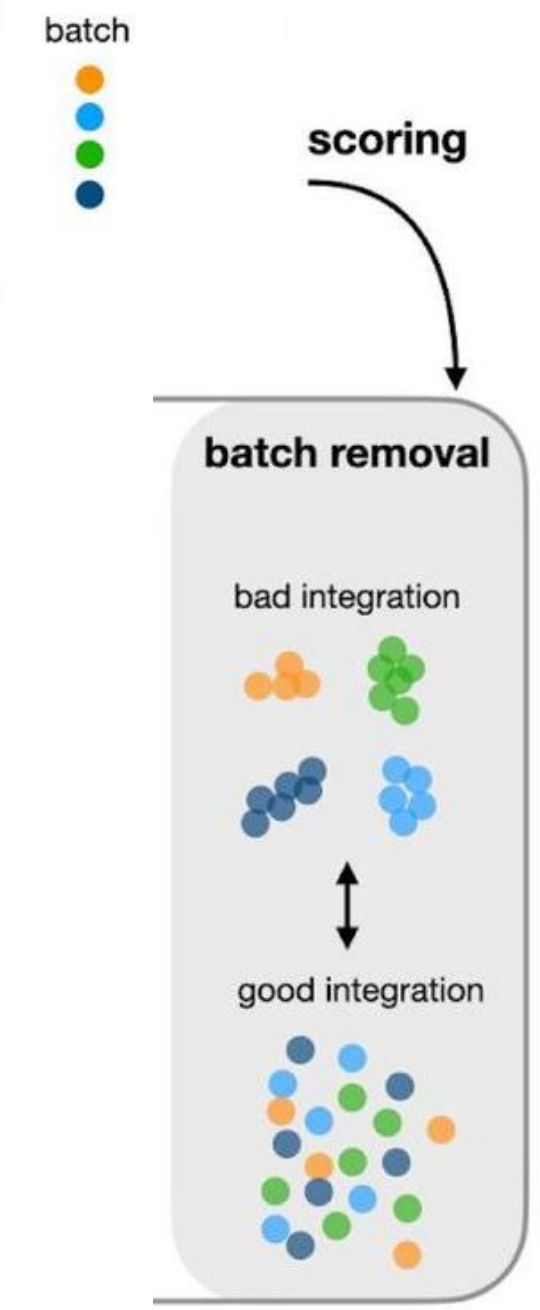

INH2 SST+
INH4 PVALB+
ODC5
ODC2
ODC6
ODC1
ODC3
INH1 VIP+
ODC4
ODC7
INH3
INH5 LAMP5+
INH6 PVALB+
MG
OPC
EX1
ASC1
EX3
ASC2
EX2
VASC
+
Picasso model
The Vivek Swarup cell atlas
vivek2
vivek1
ASC1
ASC2
EX1
EX2
EX3
INH1 VIP+
INH2 SST+
INH3
INH4 PVALB+
INH5 LAMP5+
INH6 PVALB+
MG
ODC1
ODC2
ODC3
ODC4
ODC5
ODC6
ODC7
OPC
VASC


Batch
Protocol 1
Protocol 2
Batch effect correction
tsne2_2
tsne2_1
tSNE_2
tSNE_1
Dim_2
Dim_1
40
20
0
-20
-40
-40
-20
0
20
40
CD209A+KLRD1+
CD4 T cell
CD8 T cell
Macrophage
Mast cell
mDC
Neutrophil
NK
Other
Unconventional T cell

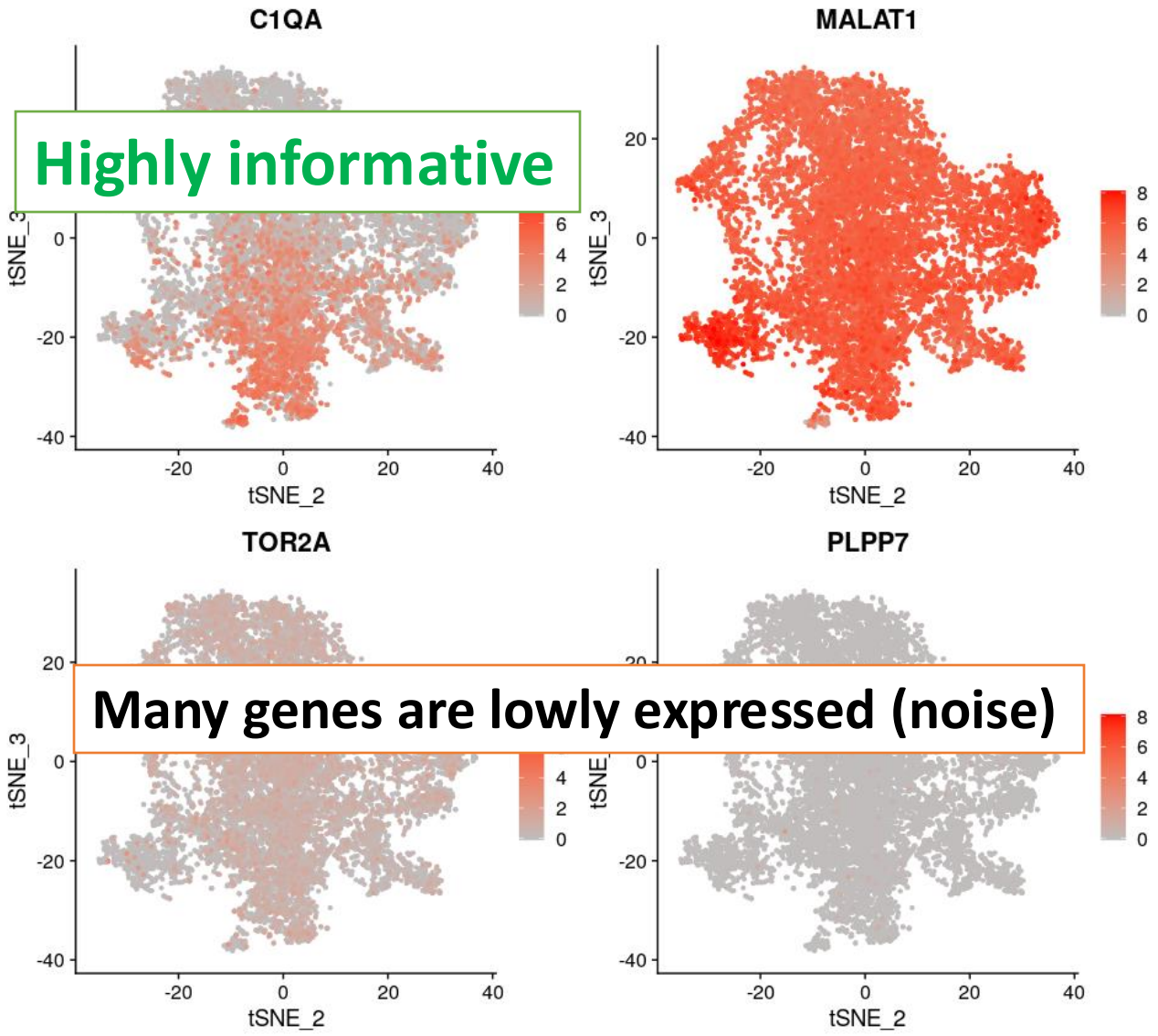
Batch effect correction

Technical and biological differences cannot be confounded:

- At least one cell type in both batches
- We cannot have a perfect overlap between cell type and batch

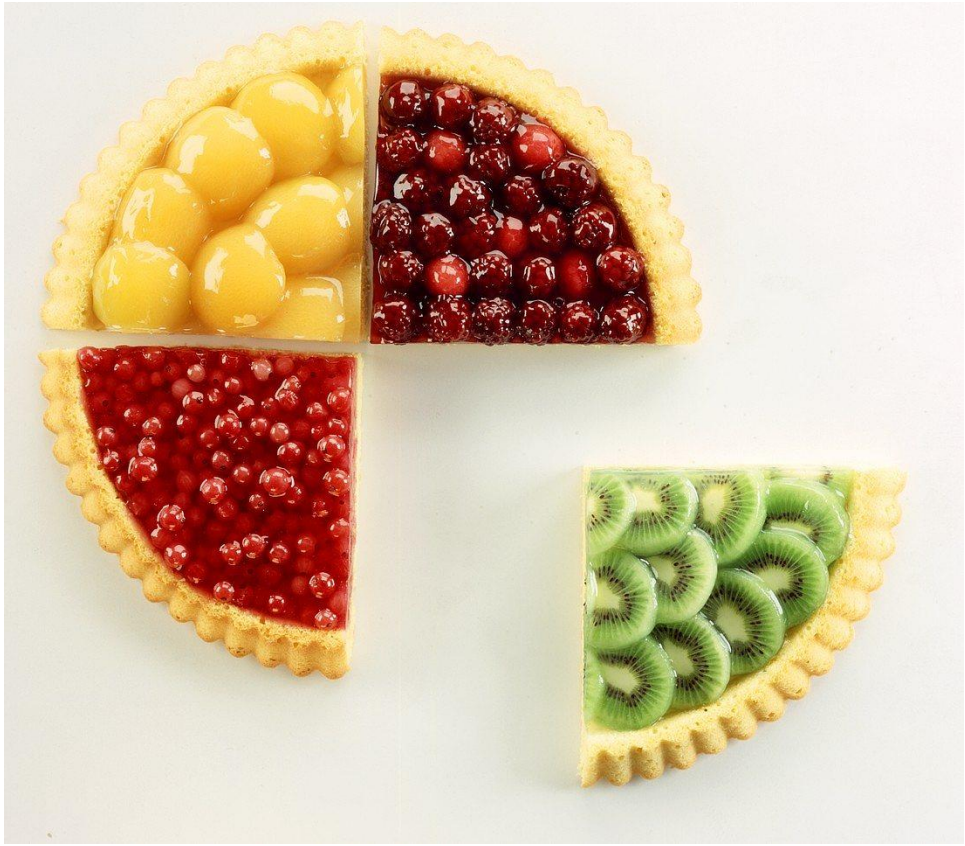


Which are the informative genes?

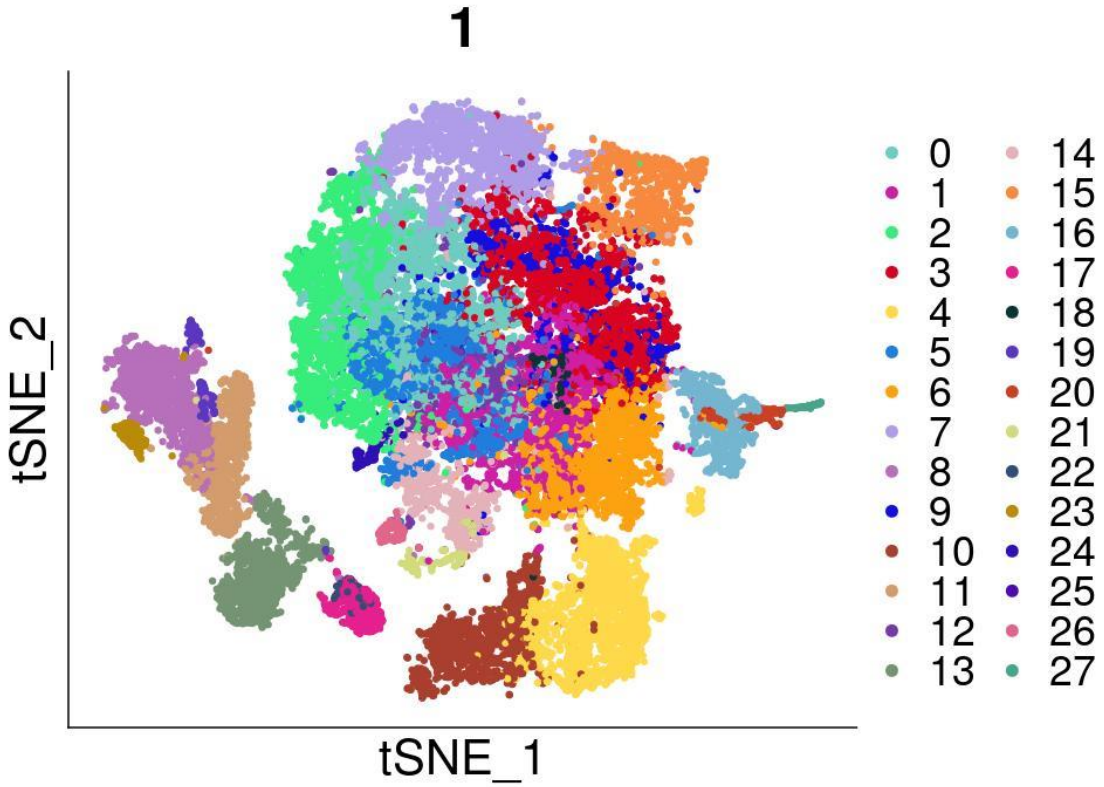


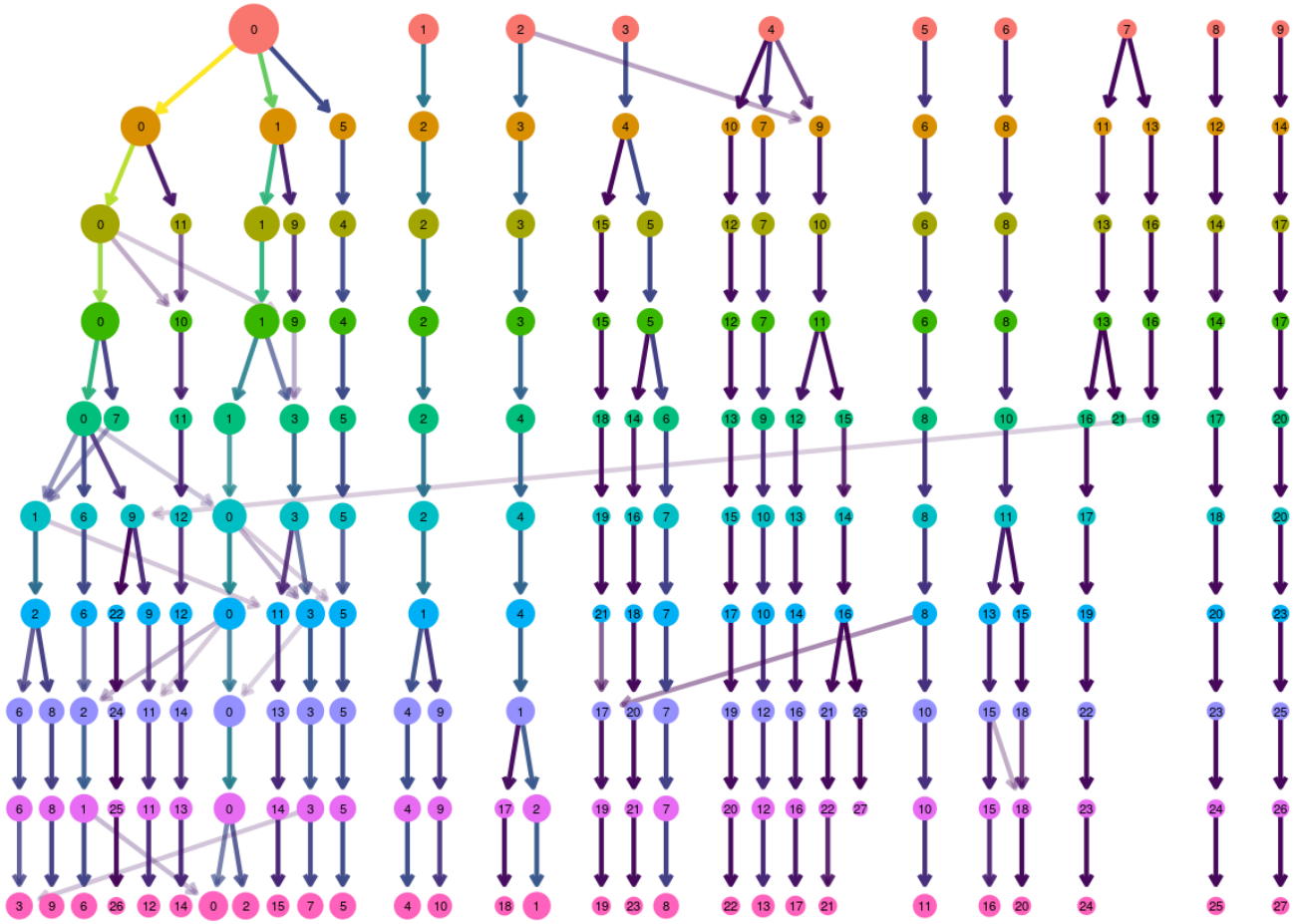
What is clustering?

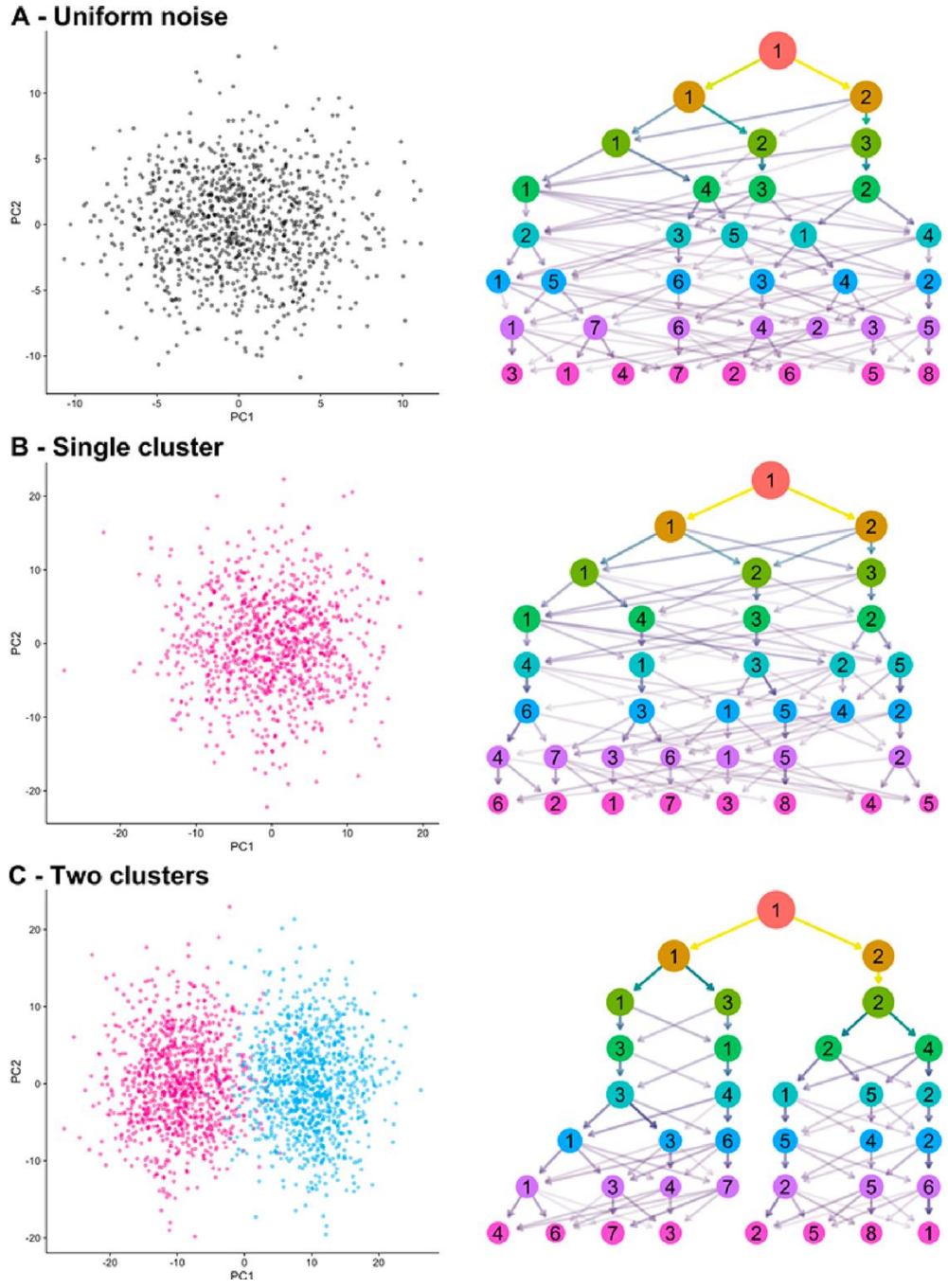
Group cells based on similarities, without knowing *a priori* their cell type.



How many clusters?







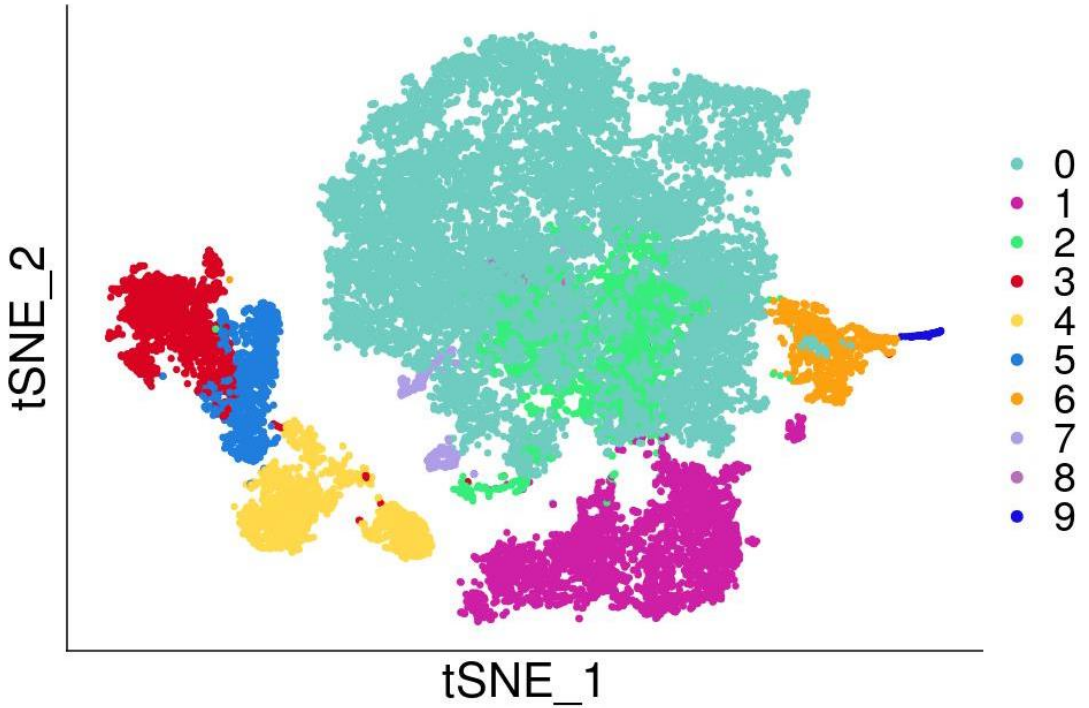
Clustering trees: a visualization for evaluating clusterings at multiple resolutions

Luke Zappia, Alicia Oshlack

GigaScience, Volume 7, Issue 7, July 2018, giy083,
https://doi.org/10.1093/gigascience/giy083
Published: 11 July 2018 **Article history**

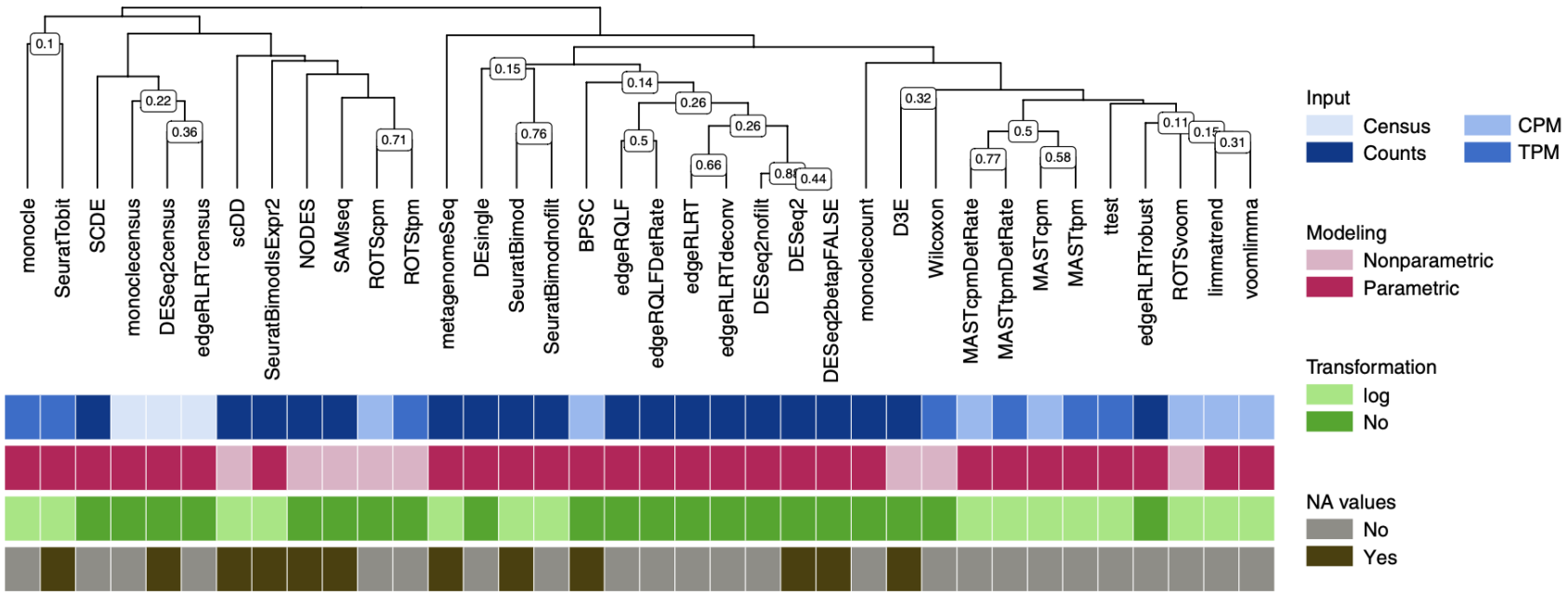
How do you know if a cluster is “real”?

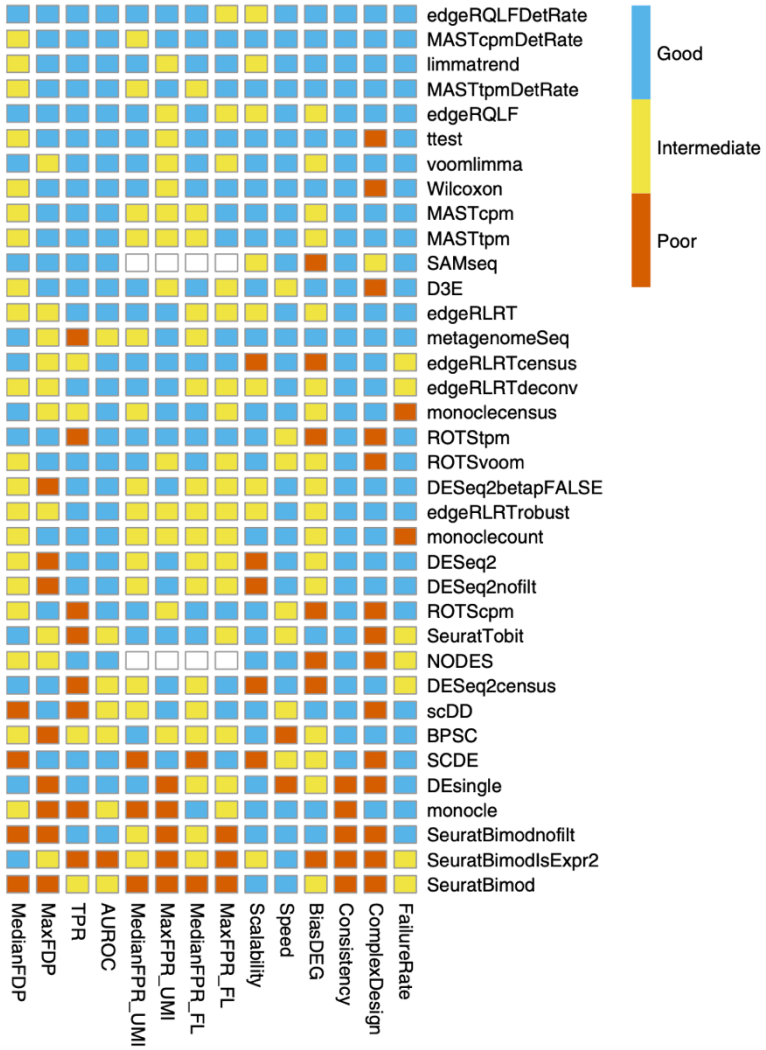
- **Significant marker genes / differentially expressed genes**
- Consistency across experimental replicates



Differential expression analysis

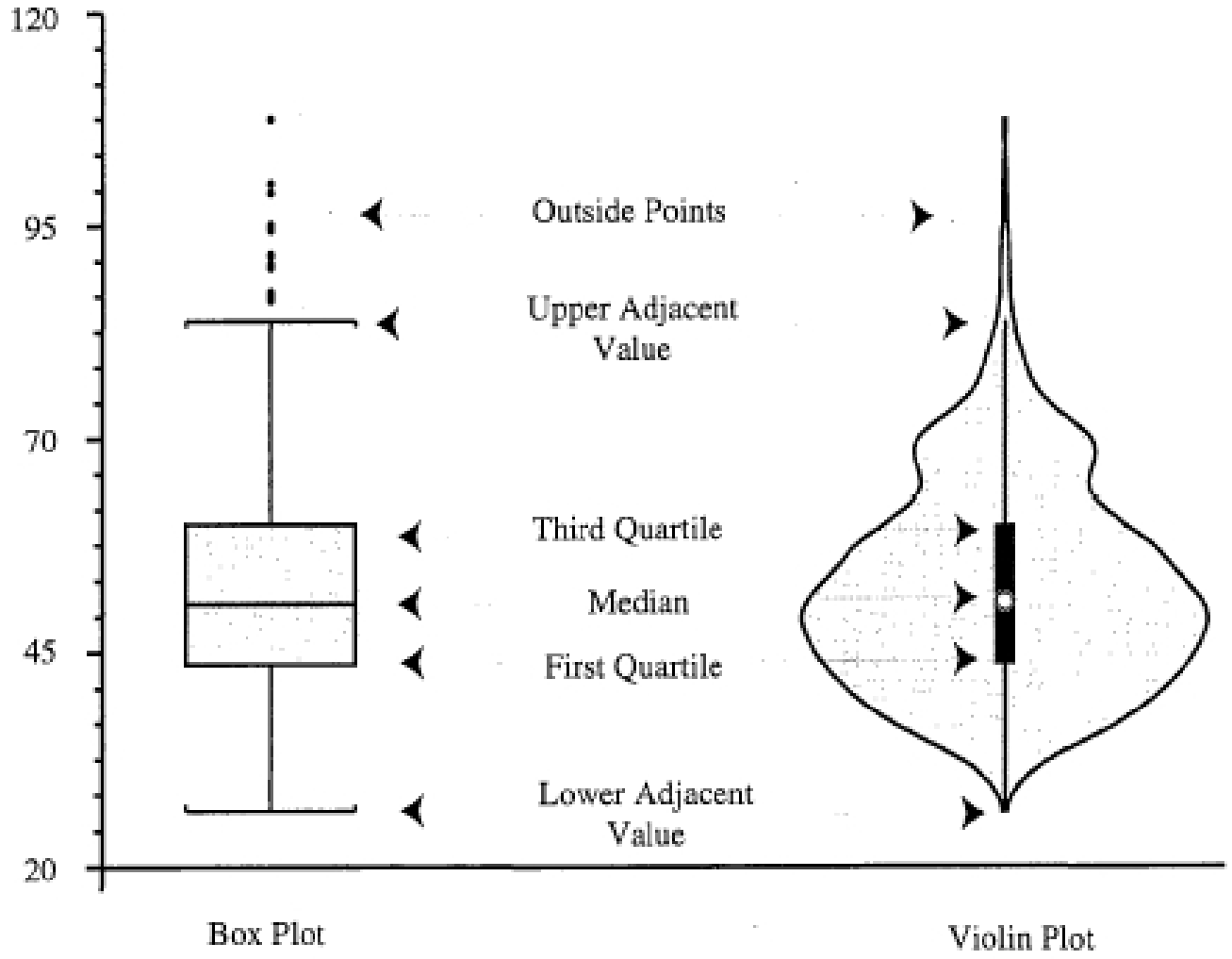
Methods are ranked by their average performance across the criteria





Soneson, C., Robinson, M. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 15, 255–261 (2018). https://doi.org/10.1038/nmeth.4612

Violin-plot


120
95
70
45
20
Outside Points
Upper Adjacent Value
Third Quartile
Median
First Quartile
Lower Adjacent Value
Box Plot
Violin Plot

Information Systems Research

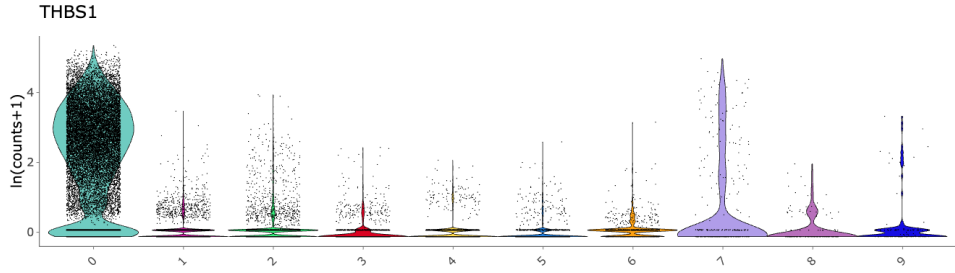
Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem

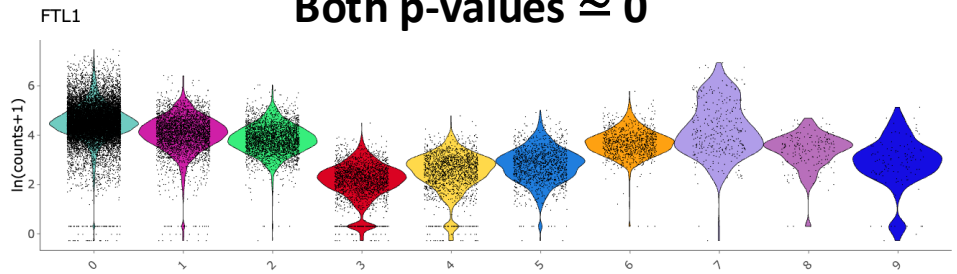
Mingfeng Lin, Henry C. Lucas Jr, Galit Shmueli

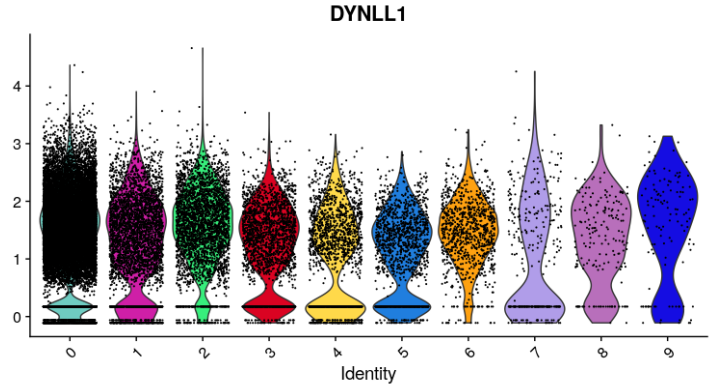
To cite this article:
Mingfeng Lin, Henry C. Lucas Jr, Galit Shmueli (2013) Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research 24(4):906-917. http://dx.doi.org/10.1287/isre.2013.0480

Cluster 0 vs others



Both p-values ≃ 0



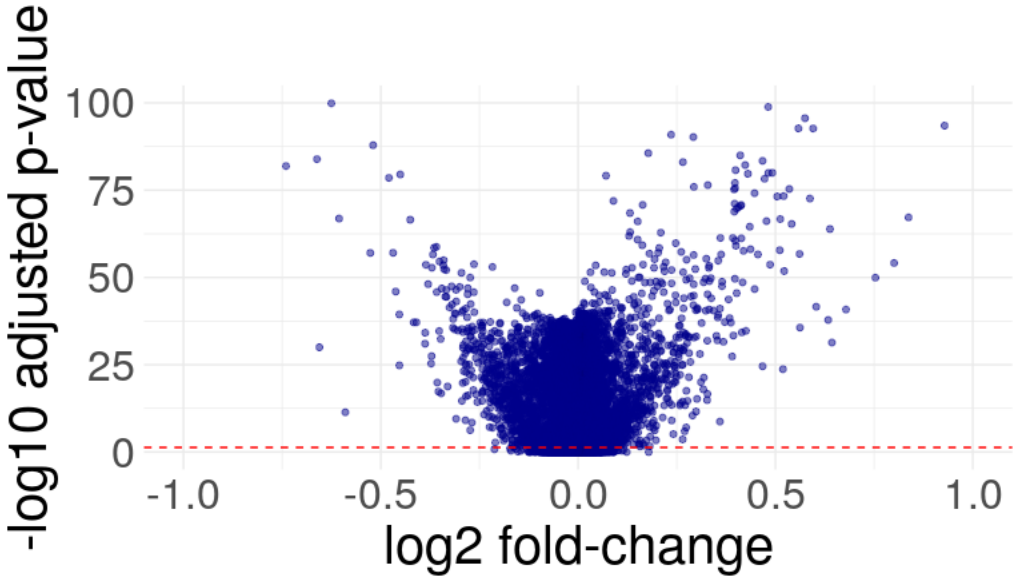


P-value = 5.68e-218

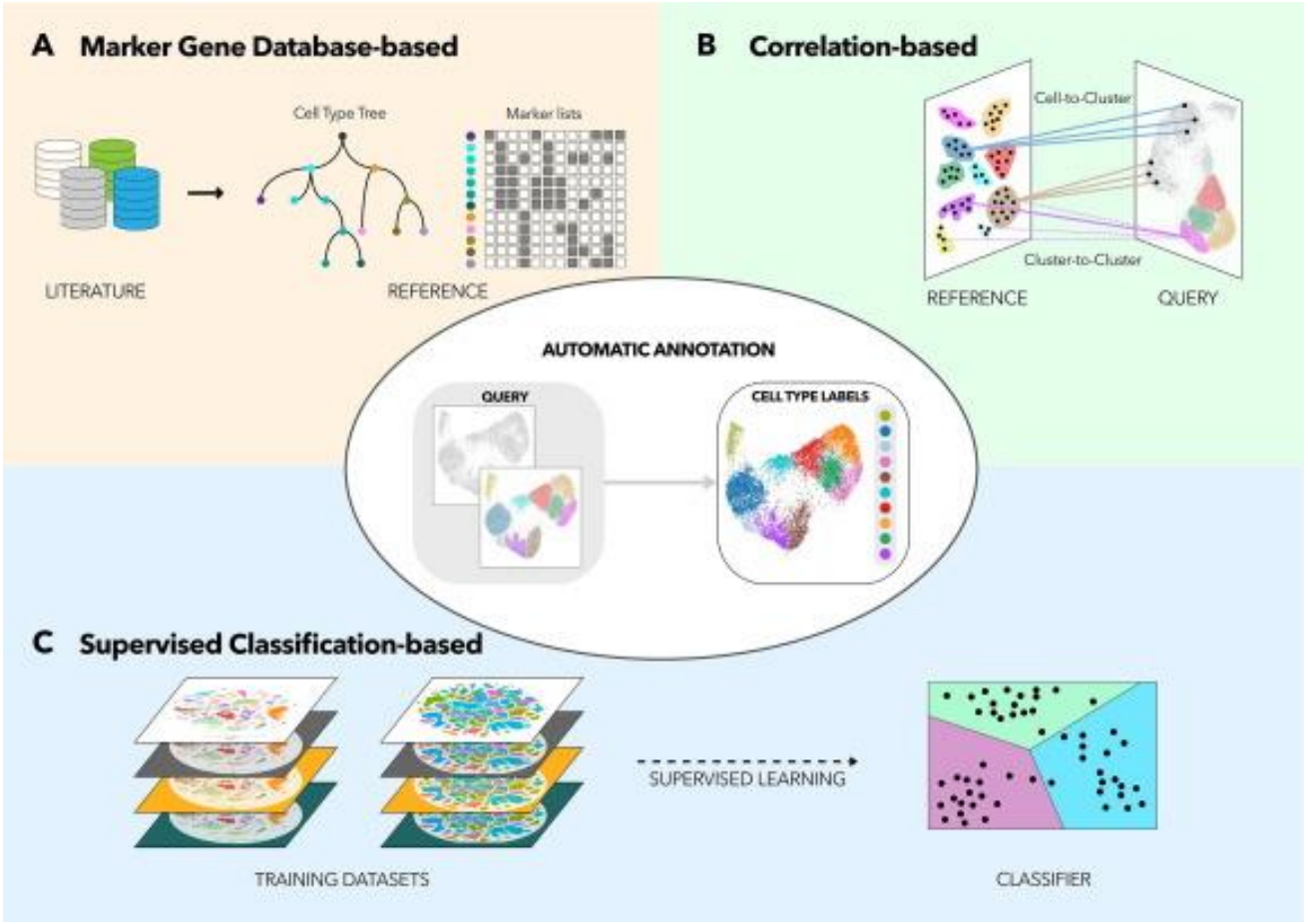
P-value interpretation

- In very large samples, p-values go quickly to zero, and solely relying on p-values can lead the researcher to claim support for results of no practical significance. Even minuscule effects can become statistically significant.

- Conclusions based on significance alone, claiming that the null hypothesis is rejected, are meaningless unless interpreted in light of the actual magnitude of the effect size.



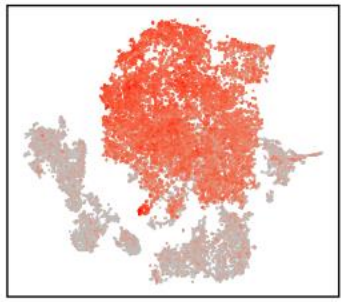
Cell type annotation: different methods



Pasquini G. et al, Computational and Structural Biotechnology Journal 2021

Macrophages

CD14

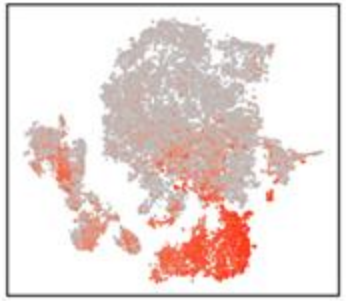


APC

H2-AA

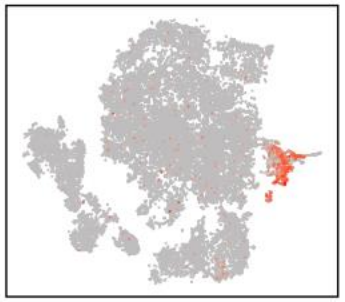
mDC2

CCR7



mDC1

XCR1

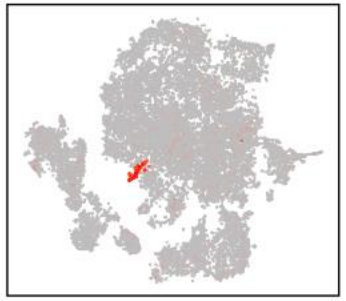


pDC

SIGLECH

Mast cells

MCPT4

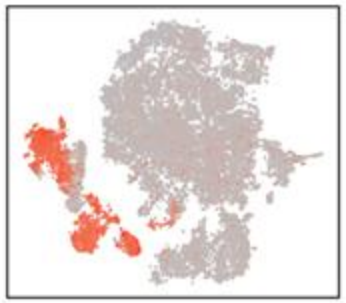


Lymphocytes

CD2

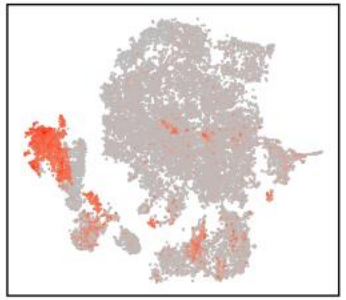
T cells

CD3D



Cytotoxic T cell

CD8A

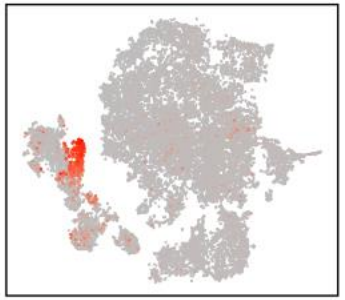


Tregs

FOXP3

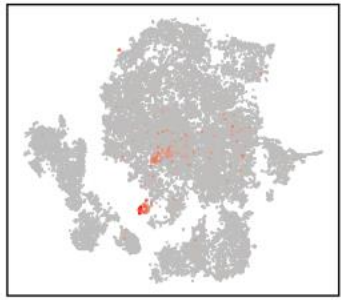
NK cells

GZMA



Neutrophils

RETNLG



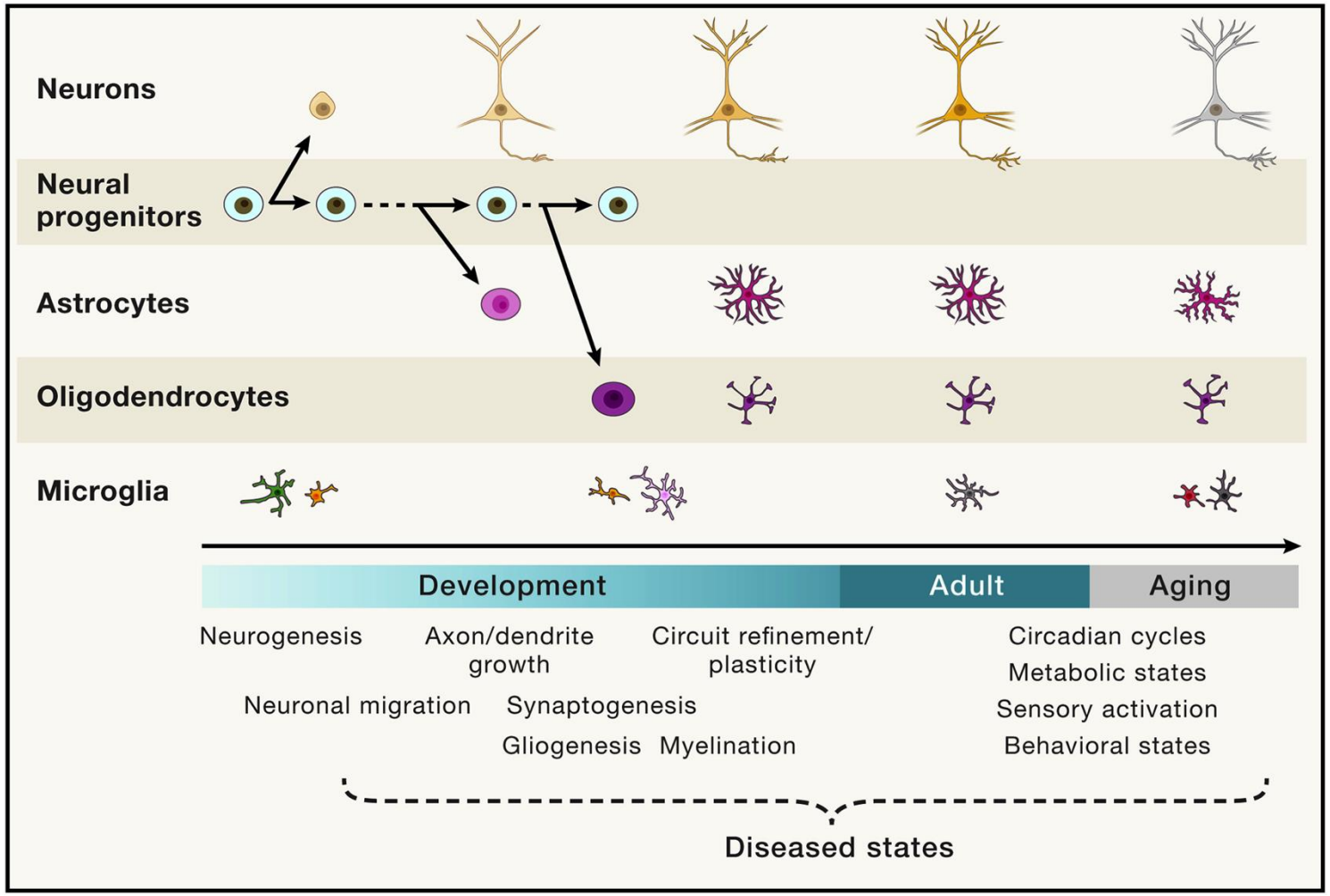
Careful with point overlapping

Same plots, different ordering of the cells

Clusters	scAnnotatR	GSEA	Canonical
0	T/NKT	NK/NKT	CD2;CD3D; CD8A (T/NKT)
1	NK/T/unknown	NK/NKT	CD2;NCAM1 (NK)
2	T/NKT/NK	T	CD2;CD3D; CD8A; CD8B (T/NKT)
3	NK	T/NK/NKT	CD2; FCGR3A(NK);
4	Mono/DC	Kupffer	CD68;FCGR2A (Kupffer); HLA-DRA; FCGR3A; CLEC7A (classical mono)
5	B	B	HLA-DRA;CD79A (B)
6	NK	Progenitor	CD2; FCGR3A; NCAM1;
7	pDC	DC	CD68;HLA-DRA; CLEC4C (pdC)
8	DC	DC	HLA-DRA; CD83 (DC); CLEC7A; CLEC9A(DC); XCR1
9	Plasma	Plasma	CD79A;CD27 (Plasma)

Clusters	scAnnotatR	GSEA	Canonical
0	T/NKT	NK /NKT	CD2;CD3D; CD8A (T/NKT)
1			CD2;NCAM1 (NK)
2			CD2;CD3D; CD8A; CD8B (T/NKT)
3			CD2; FCGR3A(NK);
4			CD68;FCGR2A (Kupffer); HLA-DRA; FCGR3A; CLEC7A (classical mono)
5			HLA-DRA;CD79A (B)
6			CD2; FCGR3A; NCAM1;
7			CD68;HLA-DRA; CLEC4C (pdC)
8	DC	DC	HLA-DRA; CD83 (DC); CLEC7A; CLEC9A(DC); XCR1
9	Plasma	Plasma	CD79A;CD27 (Plasma)

What is a cell type and how to define it?



Zeng, H. (2022) "What is a cell type and how to define it?," Cell, 185(15), pp. 2739–2755. Available at: https://doi.org/10.1016/j.cell.2022.06.031.

Farewell, “Cell Type.”

Allon Klein
Harvard Medical School

Cell Systems
Voices

What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism?

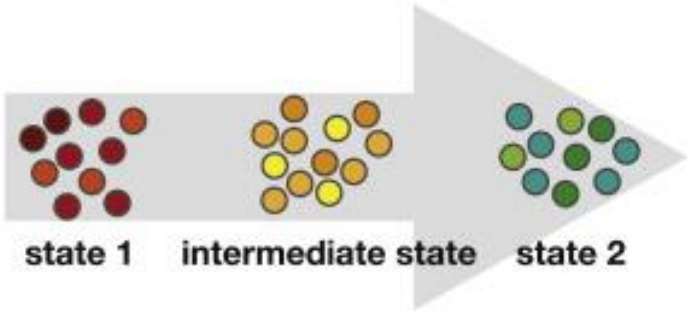
- Cell types have historically been defined by morphology, ontogeny, function, or molecular composition. **No single attribute has served for cell type classification.**
- Differences within cell types can be as large as differences between cell types, as seen in comparing fibroblasts between tissues.
- Are novel cell states in fact distinct “cell types”? Are different cell types points in a continuum of states?
- Some cells associate with two or more classical ”cell types”: **we find that there are no clean dividing lines.**

Cell type divisions are useful, but we must remember that they are artificial, imposed for our convenience and not because biology needs them.

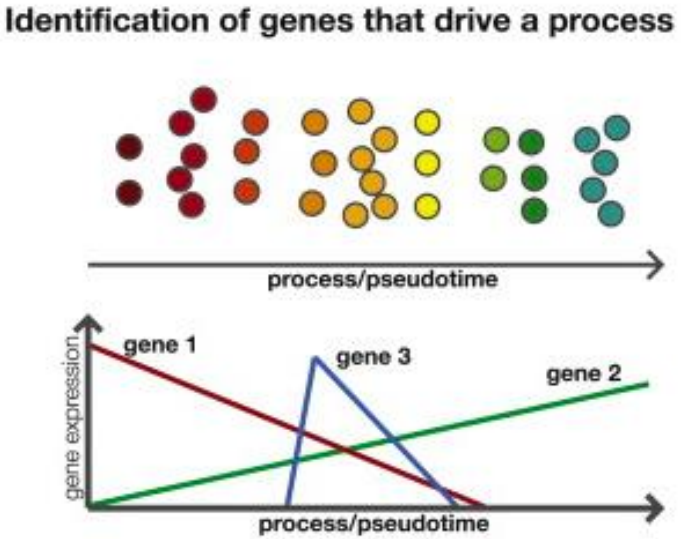
Trajectory Analysis

e.g. blood cells

e.g. intestinal crypt,
developing cerebral cortex



e.g. dendritic cells response to LPS



https://www.gsea-msigdb.org/gsea/index.jsp

UC San Diego

BROAD INSTITUTE

Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.
- View guidelines for **using RNA-seq datasets with GSEA**.
- Use the **GenePattern** platform to run analyses, including classical GSEA and a variation designed for single-sample analysis (**ssGSEA**).

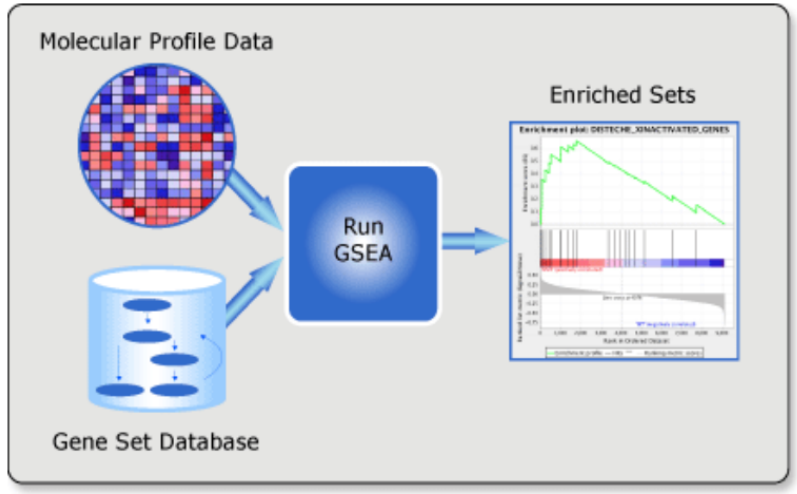
What's New

2-Oct-2022: GSEA 4.3.2 released. This is a minor release to fix a bug on the species consistency check. See the release notes for details.

7-Sep-2022: Announcing the **first release of Mouse MSigDB** (v2022.1.Mm) with ~16,000 gene sets that can be used directly for GSEA analysis of mouse datasets without the need for orthology conversion. A **new release of Human MSigDB** (v2022.1.Hs) includes updates to Reactome, GO, HPO, and WikiPathways. See the release notes for details.

7-Sep-2022: GSEA 4.3.0 released. Adds support for convenient access to the new Mouse MSigDB database and makes adjustments for our new versioning scheme. See the release notes for details.

29-Mar-2022: MSigDB gene sets are now available as JSON bundles. Check our Downloads page for the full collections. Smaller bundles as well as extended metadata in TSV format are available from our Search page.



License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Citing GSEA

To cite your use of the GSEA software, a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et al. (2005, PNAS) and Mootha, Lindgren, et al. (2003, Nature Genetics).

Funding

GSEA and MSigDB are currently funded by a grant from NCI's Informatics Technology for Cancer Research (ITCR)

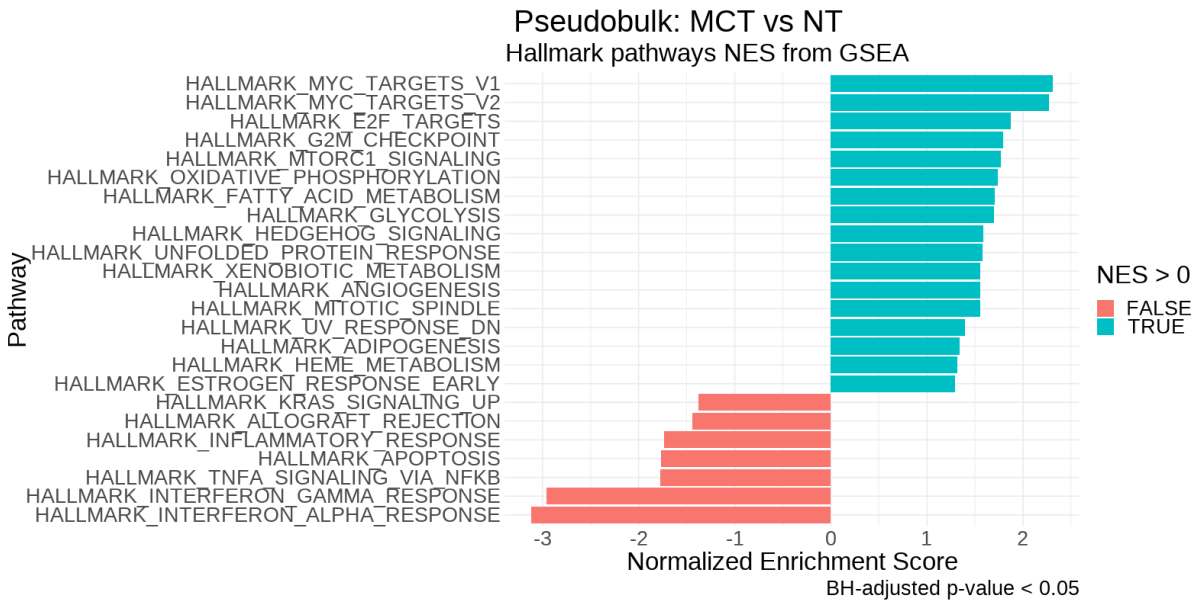
NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

Functional Enrichment analysis

clusterPseudobulk: MCT vs NT

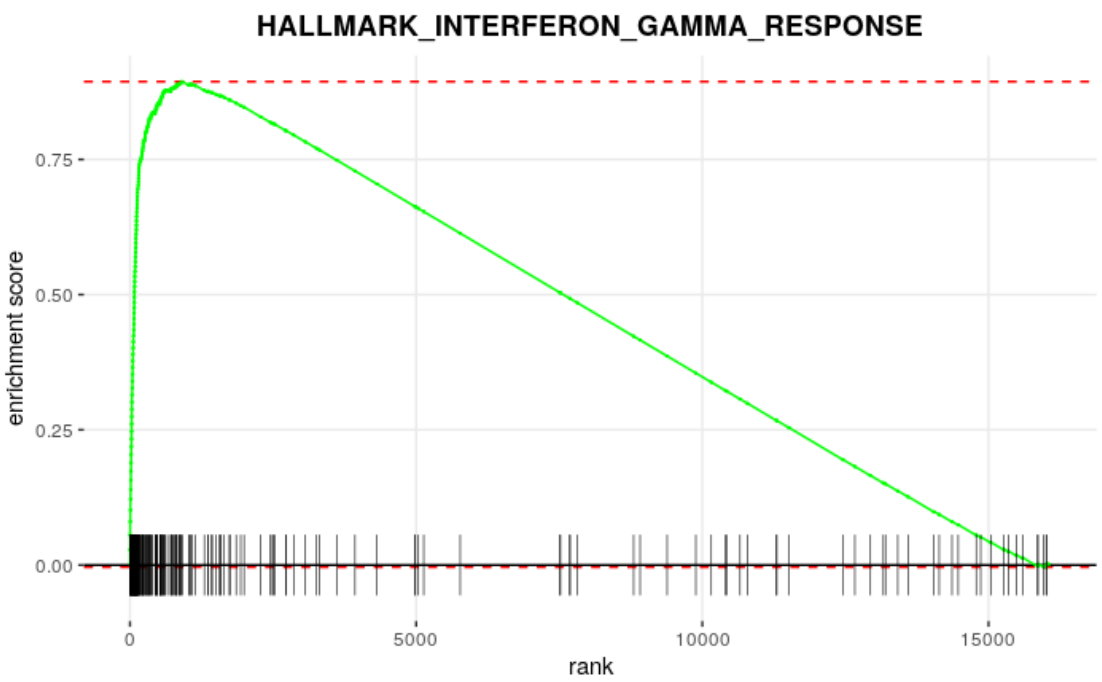
ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
MT-CYTB	0.862815876317793	10.9421227921704	8.14309165903894	3.22573676083731e-06	0.00264886894533225	4.95029458754272
GM4951	1.85932253418593	2.69706227722278	6.98101940987948	1.50751497386754e-05	0.00554718569874005	3.44609870293504
CD1D1	1.1666235688033	2.51356621960229	6.80705792888968	1.92631345846731e-05	0.00606002156992019	3.19252842174441
MT-ND4	0.856160469094567	10.5160708747382	6.57780419430834	2.67701740547538e-05	0.00663924751224761	2.84869586534704
PTGR1	1.18267227937947	4.70230825967235	6.54196572707112	2.82011714868461e-05	0.00667085892807032	2.85014904988757
PCDH7	1.8047375155905	1.84649878944963	6.50636314542766	2.9703527792315e-05	0.00674434017062514	2.72760087989876
NRP2	1.42466712221994	5.28009904686116	6.09411649072013	5.48578847542285e-05	0.00956843249280884	2.17402924504152
LAMC1	1.51676307375369	2.49841955893912	6.04407342847072	5.91929381633442e-05	0.00993764163931071	2.1488822888903
LRP8	1.08773416313284	3.18401129406438	5.92811950930207	7.06943944170921e-05	0.0111482917741378	1.97698999027814
ABCC1	0.983806926350957	5.87561123302048	5.67061876911585	0.000105572825280483	0.0142921421027092	1.5089888349843
PLXNA1	1.14127721356301	4.31066430944664	5.66903293400059	0.000105836941833256	0.0142921421027092	1.54997119133808
MT-ND1	0.825461181754333	10.4703814437754	5.64125785807246	0.00011057776955824	0.0142921421027092	1.42376539697407
MMP12	1.70115887375092	7.12750555870784	5.64015541745733	0.00011077050950729	0.0142921421027092	1.45186811837878
UPP1	1.64594327974492	7.01611145308235	5.56109118874663	0.000125562457017118	0.015083695087547	1.32630081040849
AMPD3	1.1847208504036	2.0530388258453	5.48589845618451	0.00014157438869651	0.0159011464484166	1.326596249798
DDB1	0.713610406338776	5.0677072362233	5.39774178111415	0.000163131161300952	0.0175818552656519	1.0933017460562
SFXN1	1.40176558745072	3.88048157545922	5.32827553355182	0.000182546693133569	0.0184688276750931	1.0305545381789
XPOT	0.811367489255488	3.51056686919799	5.30796301106499	0.000188673118219302	0.0186493149792124	1.01501862835349
F830016B08RIK	2.198148928988	0.533341532270071	5.26580038542811	0.00020209132580372	0.0194756159163437	0.959888606416744
ADHFE1	3.31229739051408	-3.36975515404078	5.20781492303338	0.000222208010298211	0.0201473909812992	-0.456033583926329
BBS10	1.07723757518818	1.6415237113372	5.15416126438037	0.000242702954511527	0.0215288552604202	0.821740836616566
KCNQ1OT1	0.736176593302776	4.28948242606249	5.1333510447068	0.000251178887088611	0.0217855821401522	0.682027358392974
PGM2	0.727285371506859	4.31810594064081	5.10048322155165	0.000265205452015435	0.0224227306013294	0.649024122340703
MT-ND2	0.78480354554568	10.0852517584475	5.0305765328914	0.000297846679949458	0.022906317080645	0.427833050220357
IDE	0.883494892740096	3.72277492558127	4.92259538903845	0.000356808703837938	0.0260303620762518	0.372604548672169
GYS1	1.21225038968511	3.74839721287721	4.8549099231961	0.000399907472531751	0.0283788884598441	0.282926011655873

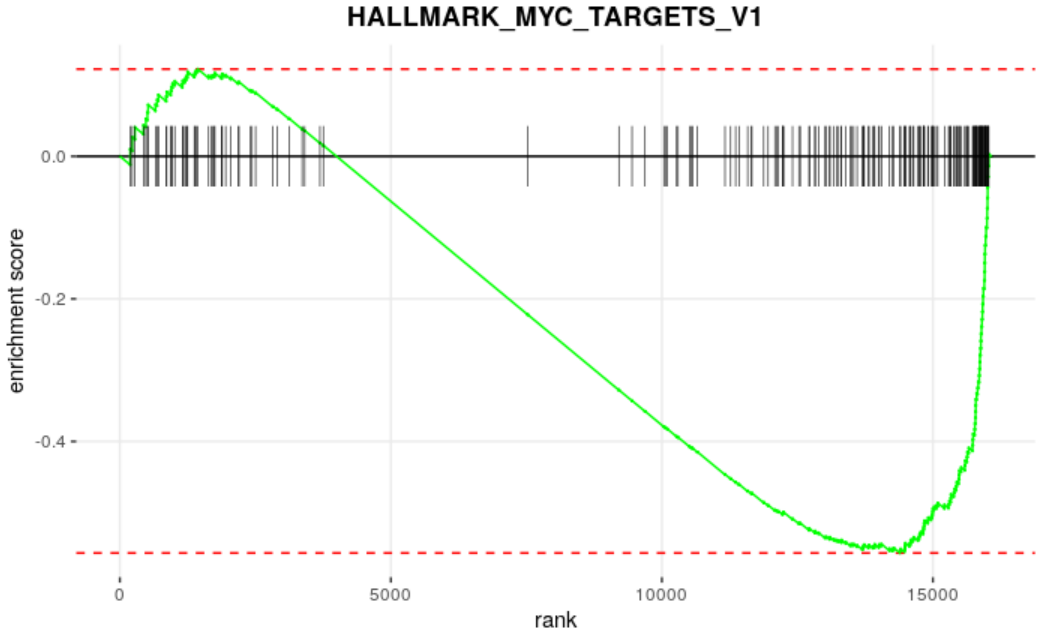
GSEA



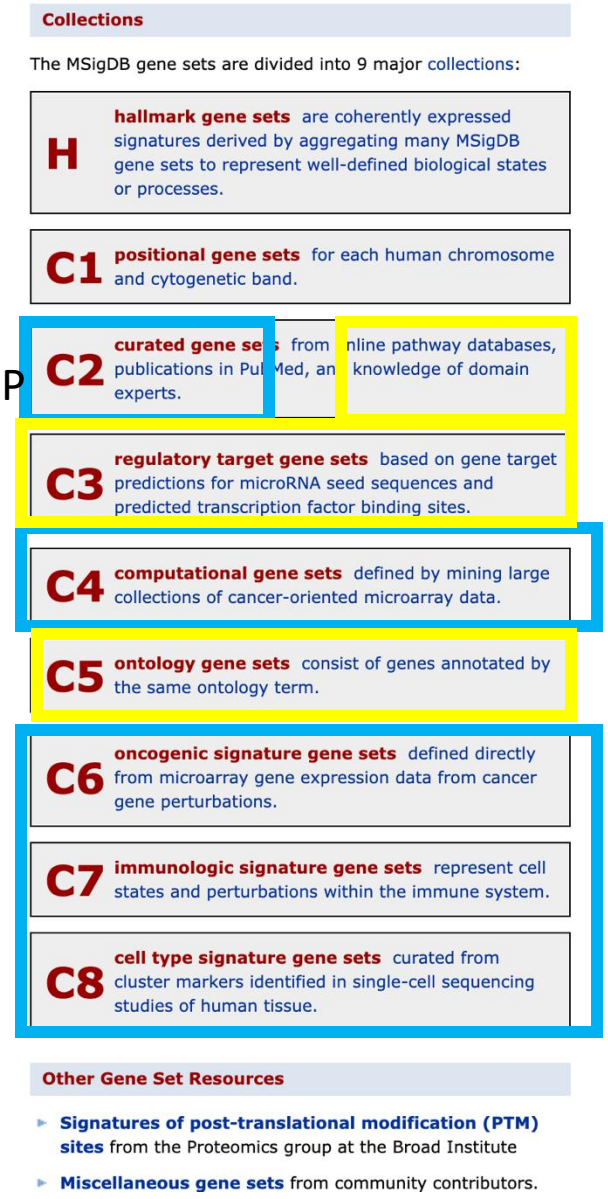
Functional enrichment analysis

Gene Set Enrichment Analysis (GSEA)





Molecular Signatures Database



Collections

The MSigDB gene sets are divided into 9 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene se** from nline pathway databases, publications in Pu Med, an knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Other Gene Set Resources

- **Signatures of post-translational modification (PTM) sites** from the Proteomics group at the Broad Institute
- **Miscellaneous gene sets** from community contributors.

CGP

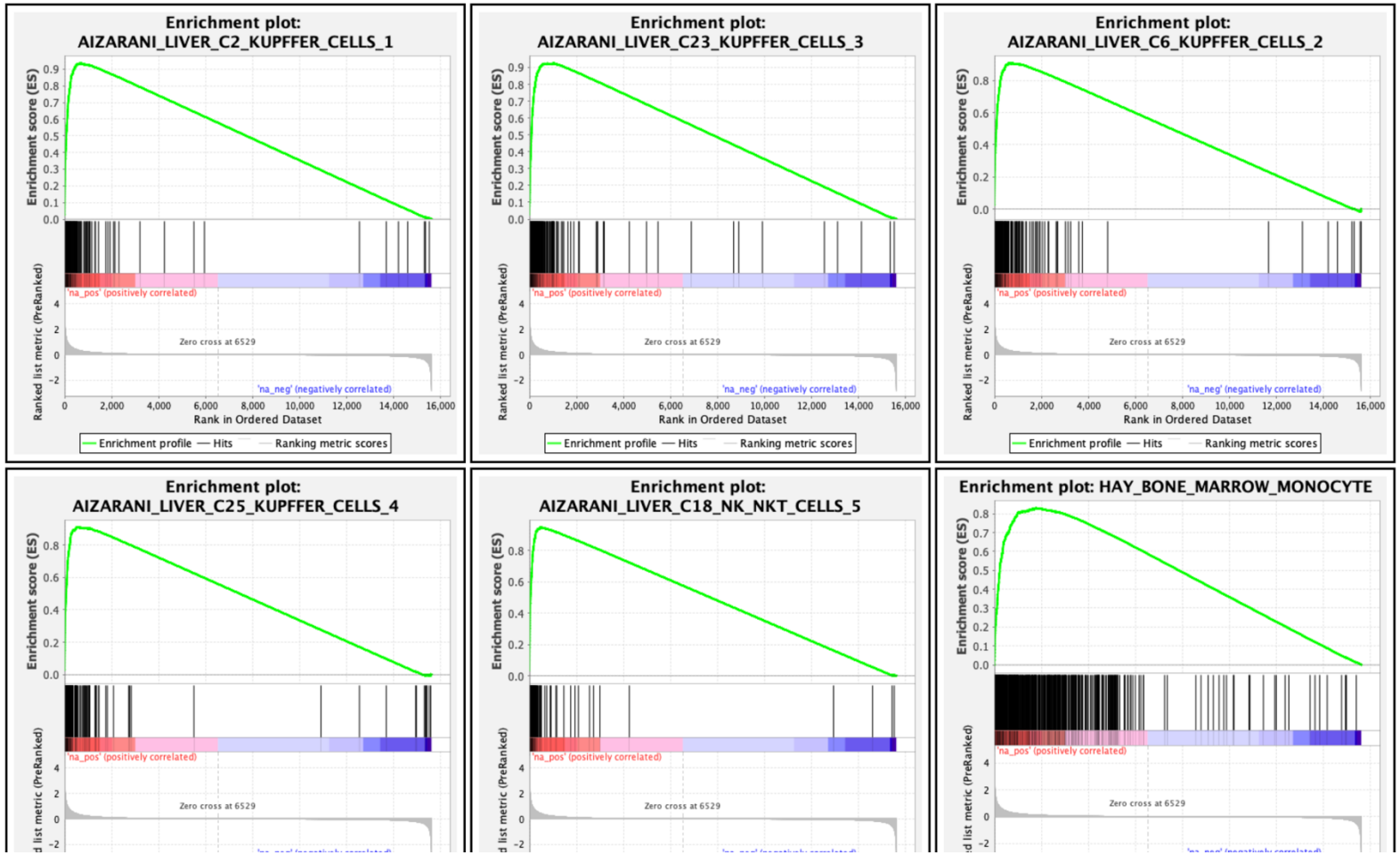
A priori knowledge-based vs data-driven

Biocarta, Kegg, reactome, wikipathways

"Nowadays, signature collections are getting larger, providing the benefit of a more complete coverage of the existing biological processes. However, the growth of these compendia is posing two main challenges related to the reliability and the redundancy of the collected gene sets."

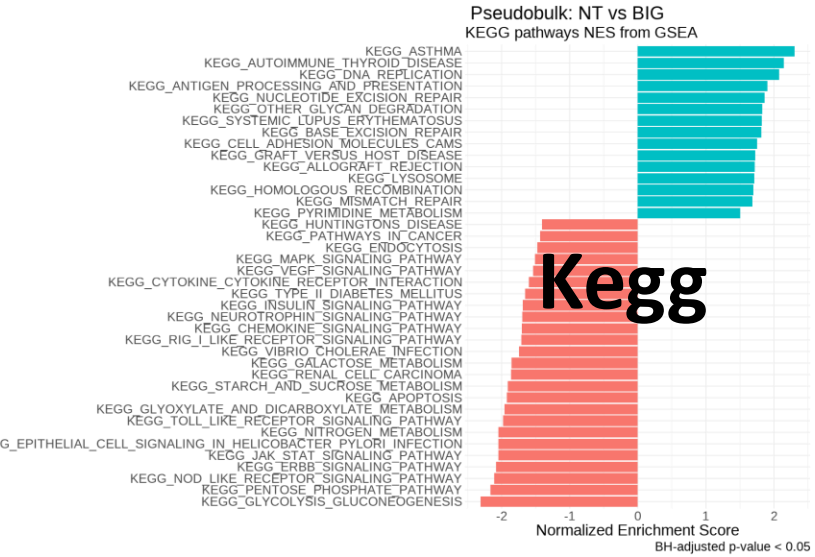
Gene signature redundancy

Table: Snapshot of enrichment results

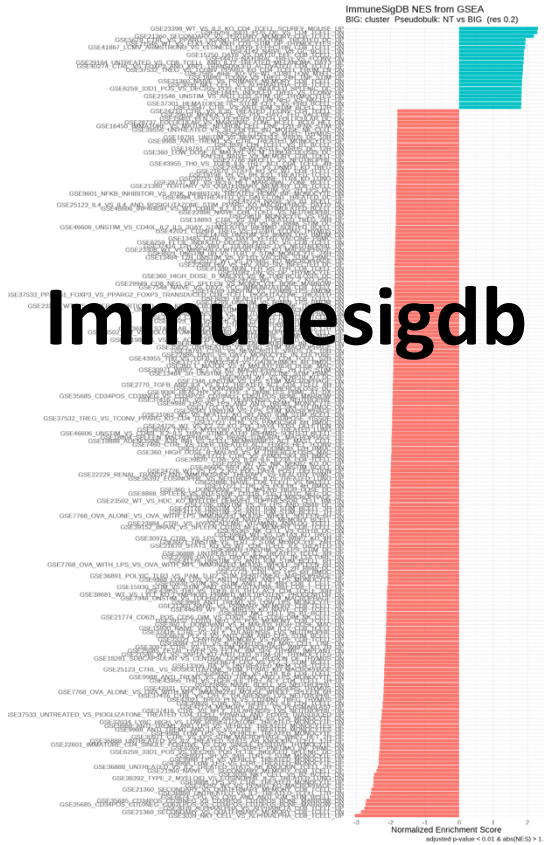


C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

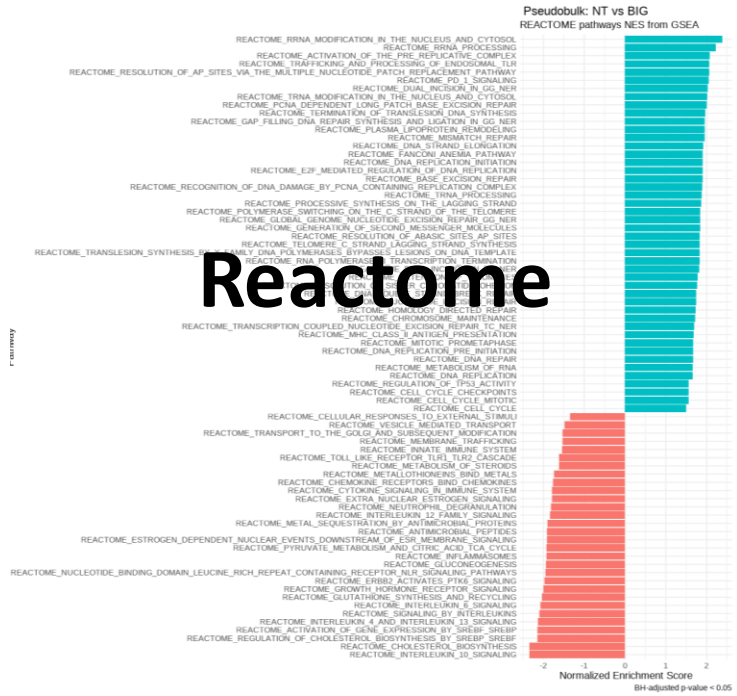
Problem: where to begin?



Kegg

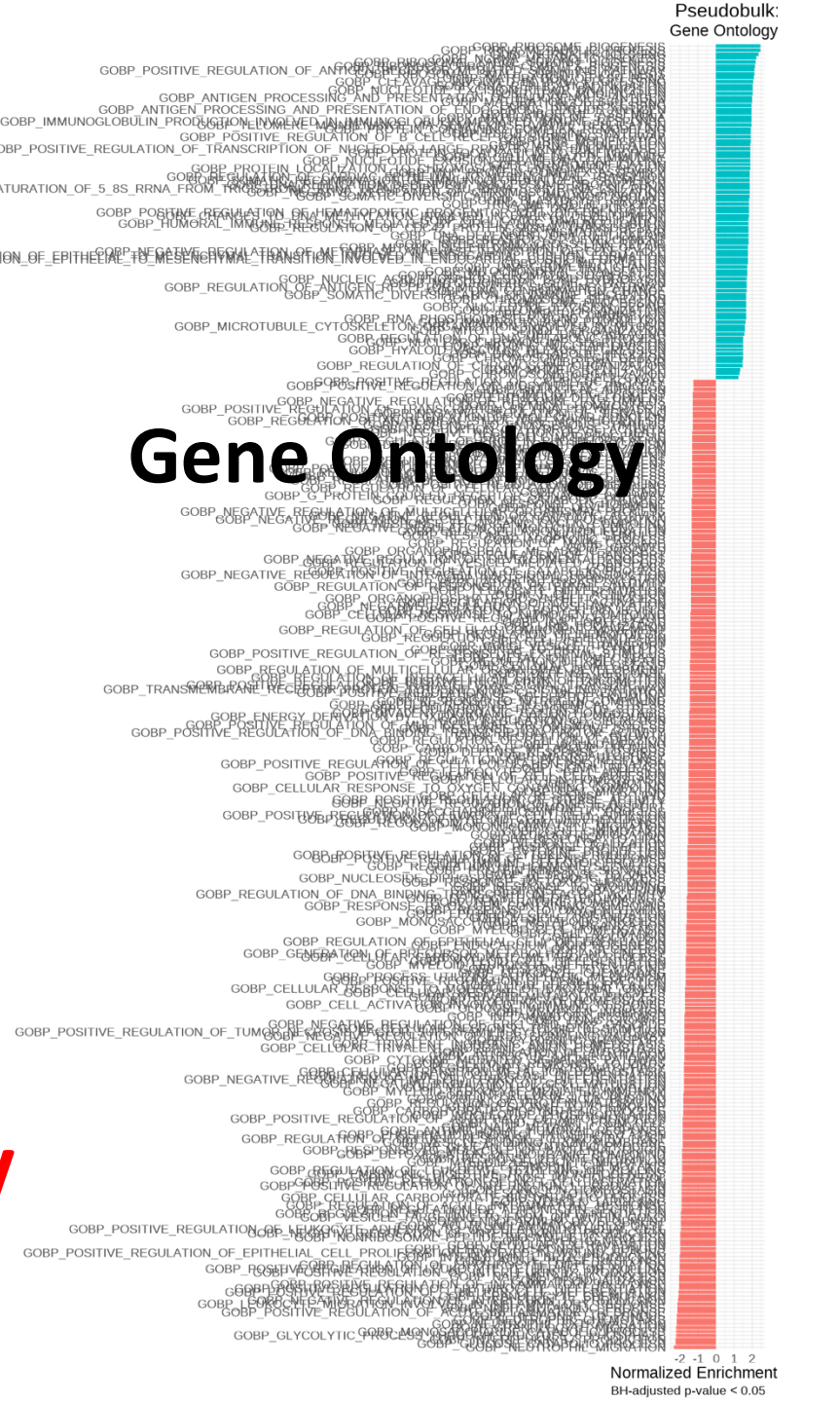


Immunesigdb



Reactome

Too much redundancy



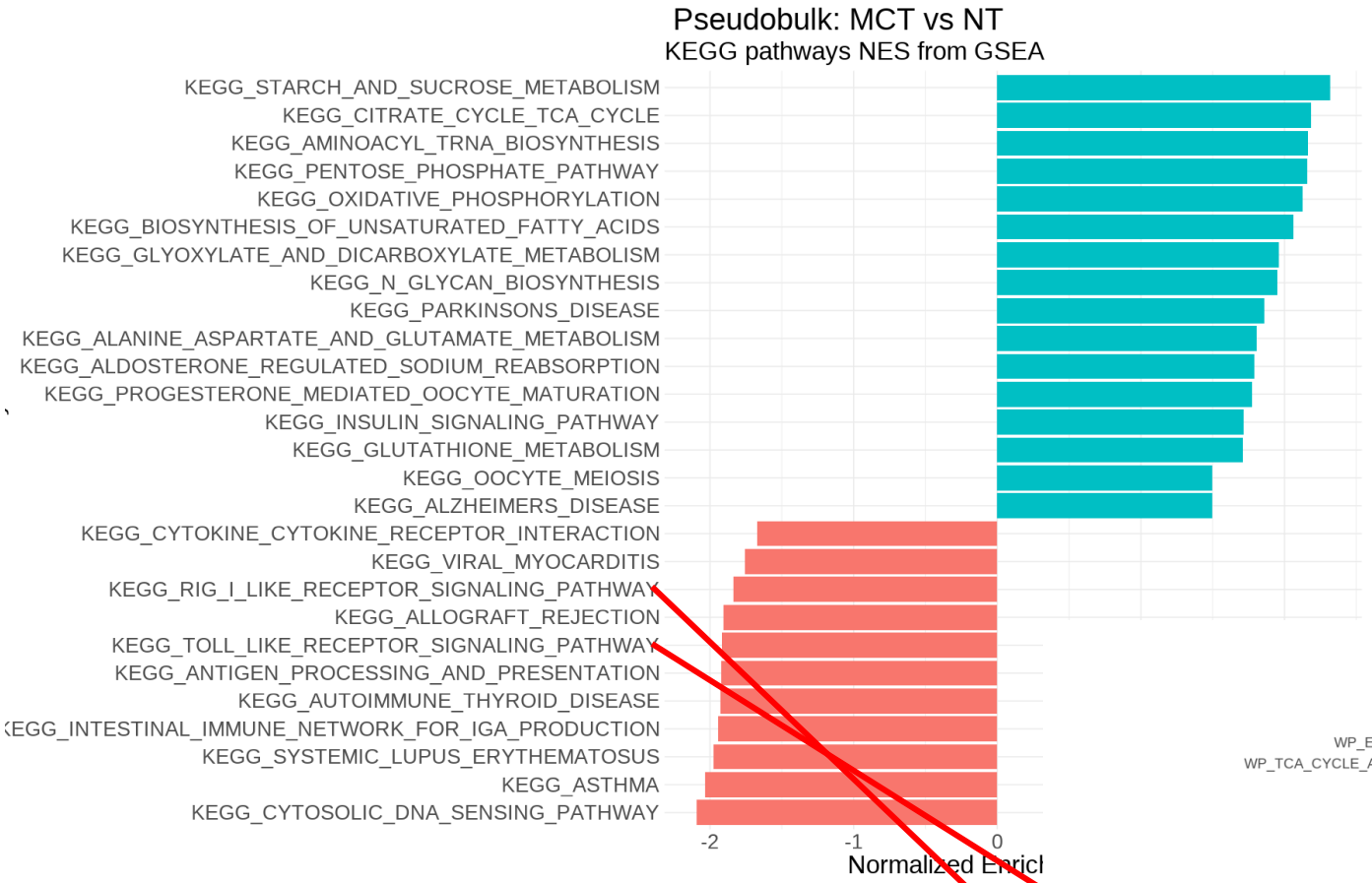
Gene Ontology

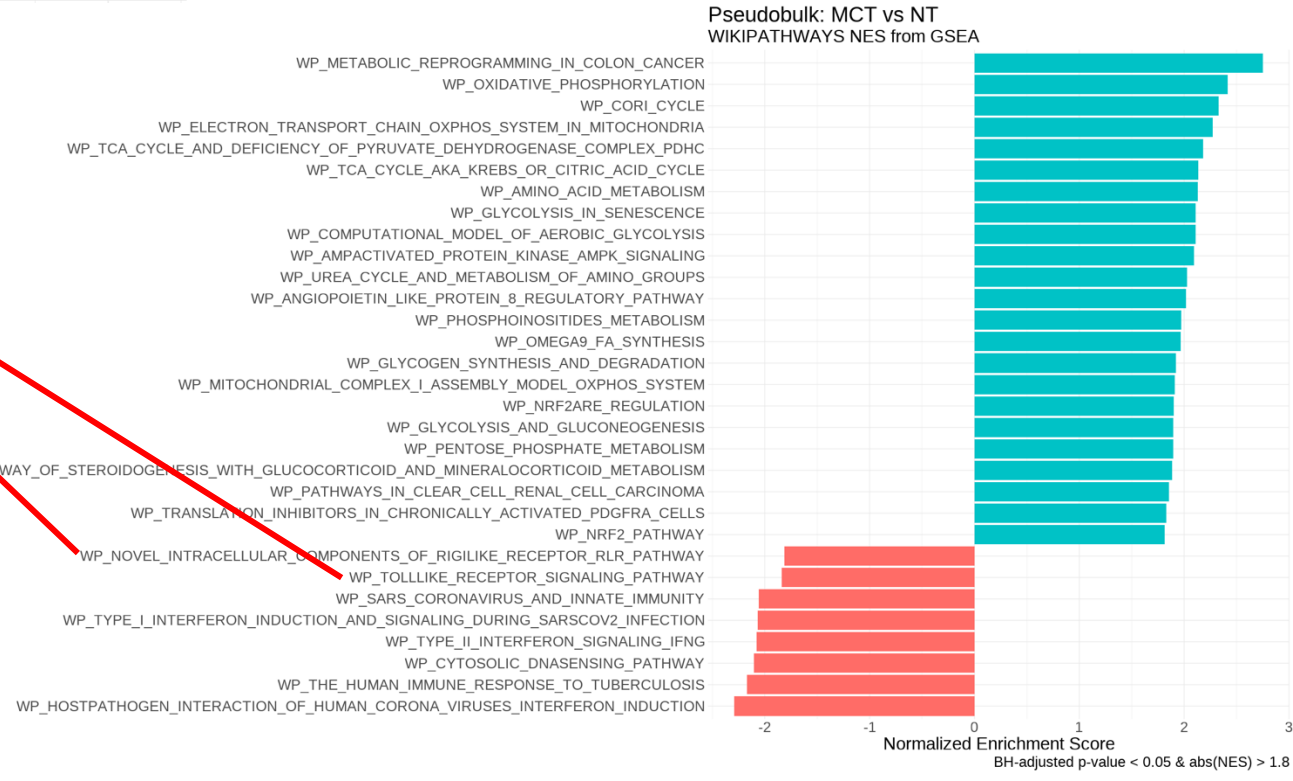
Types of gene set redundancy:

- **Compositional**: large intersection in terms of the genes composing them

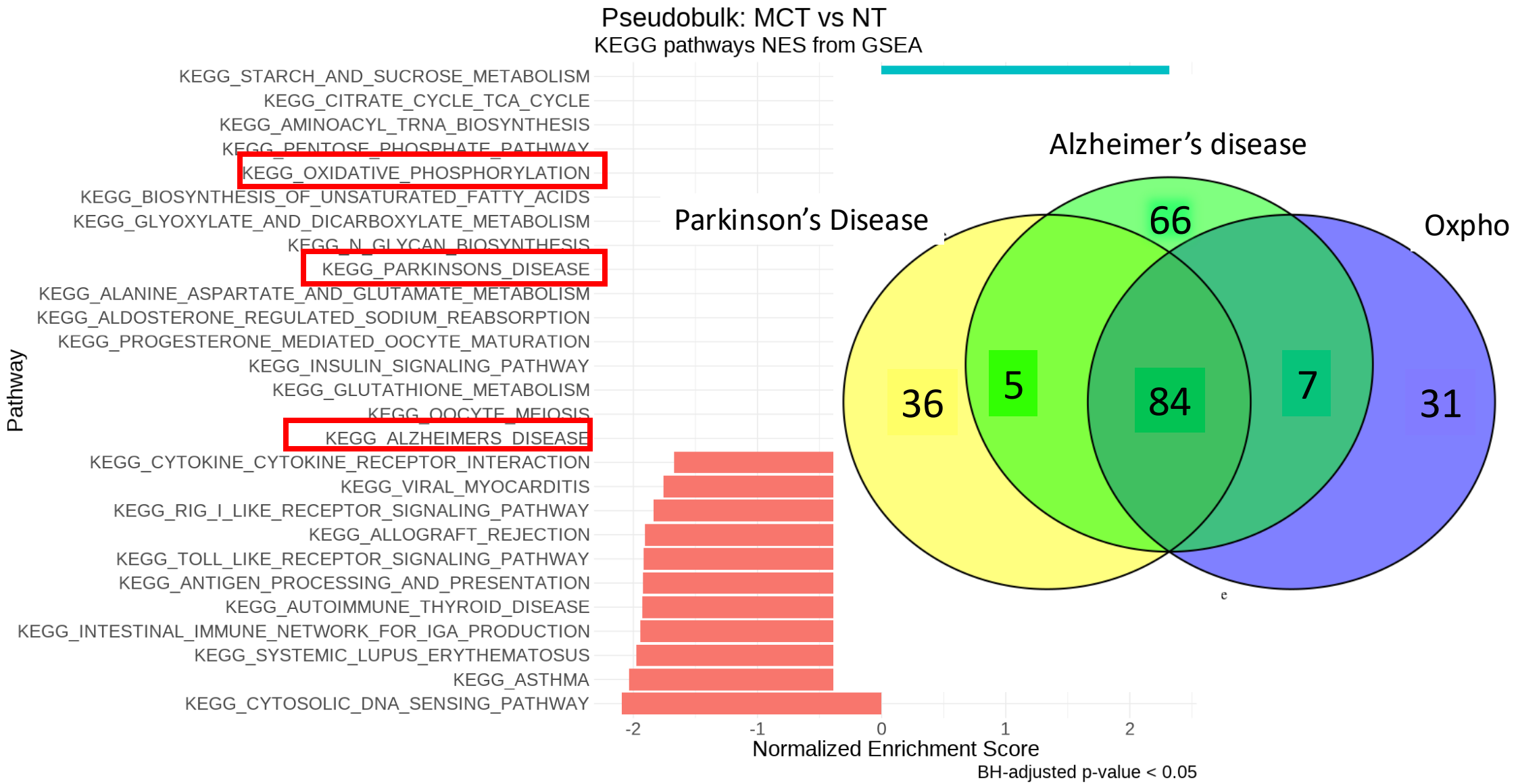
- **Functional**: highly scoring multiple gene sets belonging to analogous/related biological processes
  - two signatures can be functionally redundant even if having no overlap

Compositional redundancy: inter-collection

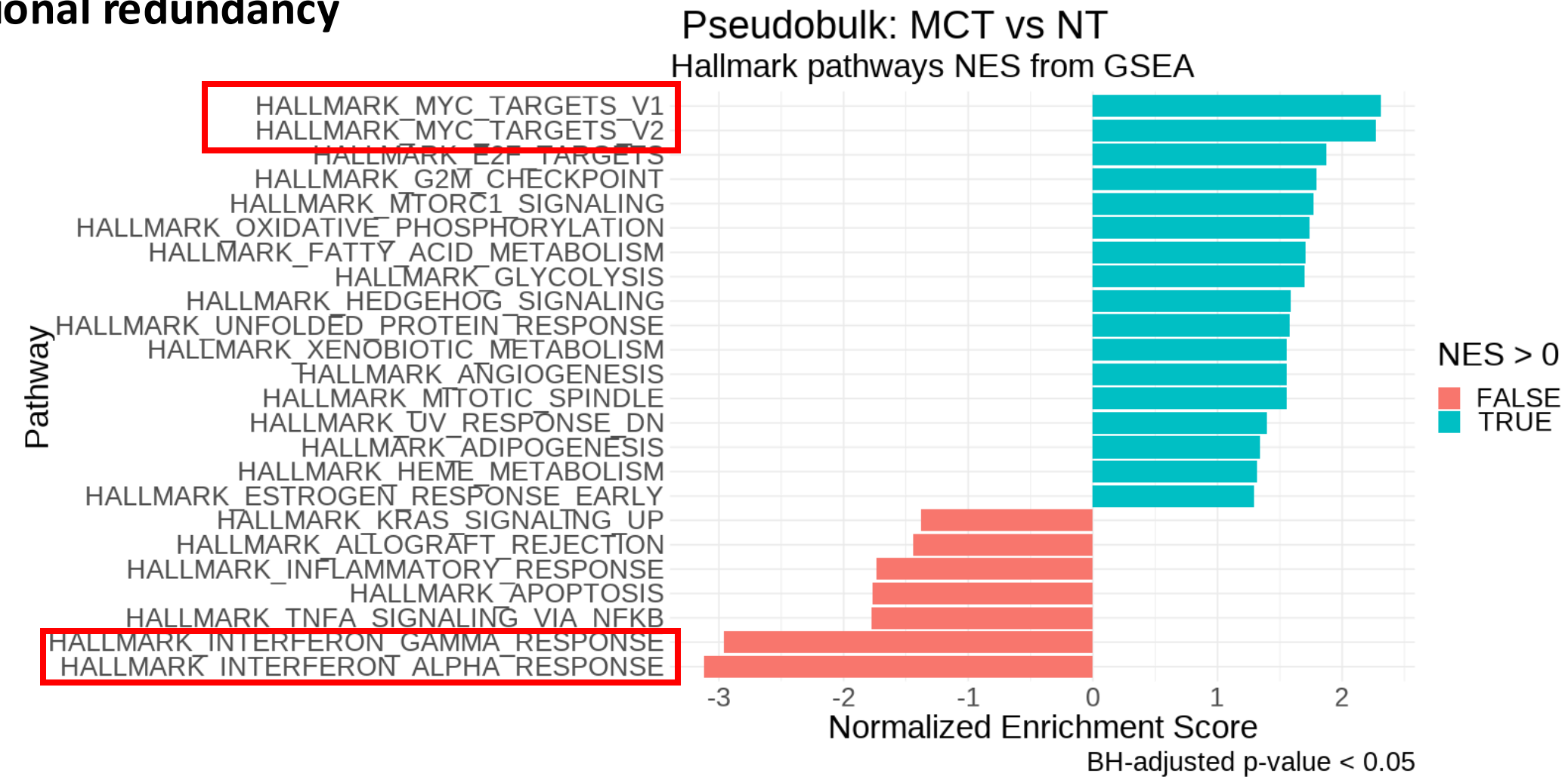




Compositional redundancy: intra-collection



Functional redundancy



The best attempt to define a robust and non-redundant collection of signatures is represented by MSigDB Hallmarks, by merging ***compositionally redundant signatures*** *and then refining the genes of the resulting signatures based on their ability to discriminate the associated phenotype. This methodology involves a* ***manual curation****, which might create a certain bias vis a vis an expert's opinion.*

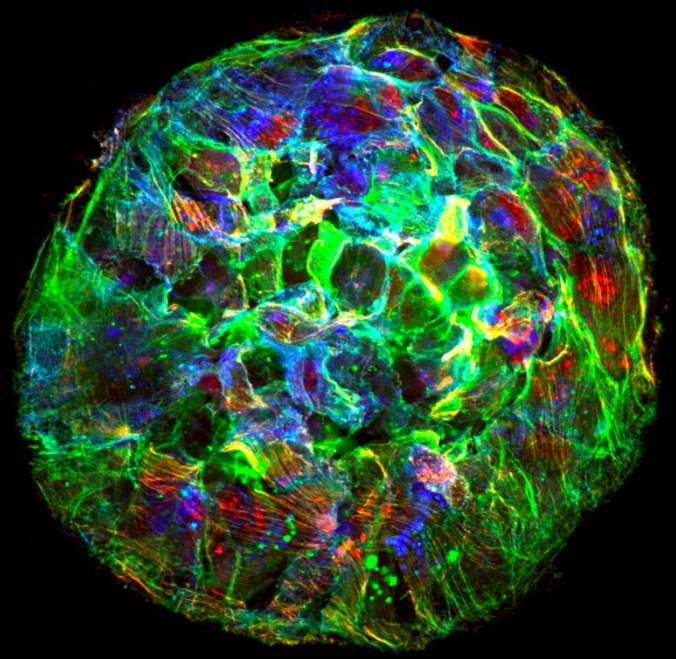
These multiple comparisons of redundant signatures can potentially hide relevant hits

pathway	pval	padj	NES	size
WP_ALLOGRAFT_REJECTION	2.27698545955137e-06	0.000466782019208031	2.01155594542176	73
WP_COMPLEMENT_AND_COAGULATION_CASCADES	0.000149279495411317	0.018361377935592	1.85741378762766	42
WP_COMPLEMENT_ACTIVATION	6.11890021363703e-05	0.00940780907846693	1.82776921063566	14
WP_PATHOGENESIS_OF_SARSCOV2_MEDIATED_BY_NSP9NSP10_COMPLEX	0.00259055226902698	0.177021071716844	1.70283554433149	16
WP_CANCER_IMMUNOTHERAPY_BY_PD1_BLOCKADE	0.00369893457037958	0.177678016524544	1.67382116176651	26
WP_HIPPOYAP_SIGNALING_PATHWAY	0.0265375343134456	0.544019453425634	1.59471718553067	23
WP_PPAR_SIGNALING_PATHWAY	0.0113808115783586	0.331598515092865	1.58178655683733	50
WP_NAD_BIOSYNTHETIC_PATHWAYS	0.0327831788528615	0.576047285557424	1.54500418186151	21
WP_EBOLA_VIRUS_PATHWAY_ON_HOST	0.00207441442193812	0.166228205354516	1.54313976096225	121
WP_PHOTODYNAMIC_THERAPYINDUCED_NFKB_SURVIVAL_SIGNALING	0.0303308517311893	0.565656594239853	1.52836409561111	33
WP_DISRUPTION_OF_POSTSYNAPTIC_SIGNALLING_BY_CNV	0.0216060405576167	0.459992354012789	1.51997914838898	24
WP_GDNFRET_SIGNALLING_AXIS	0.0298103883492606	0.565656594239853	1.51381041079167	15
WP_METABOLISM_OF_ALPHALINOLENIC_ACID	0.00216231811843273	0.166228205354516	1.50781924768374	5
WP_OXIDATIVE_DAMAGE	0.0416993893739801	0.633484162895928	1.49543828800141	37
WP_TYPE_II_INTERFERON_SIGNALING_IFNG	0.0303523050567726	0.565656594239853	1.47195477322025	42

Conclusion

- scRNA-seq is a **powerful and insightful method** for the analysis of gene expression with single-cell resolution;
- There are **many challenges and sources of variation** that can make the analysis of the data complex;
- Every dataset is unique and best practices tend to vary;
- Some tools will always give you an answer. That doesn't mean that answer is "true".

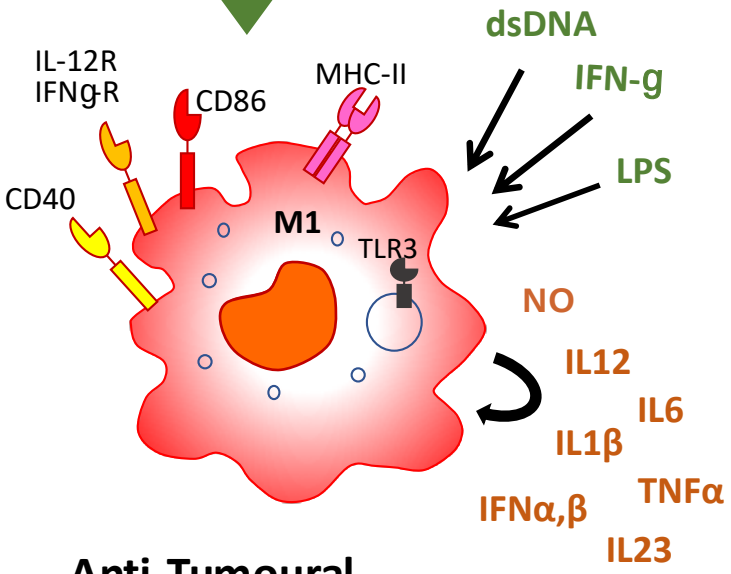
Single-cell transcriptomics in unravelling the therapeutic potential of myeloid cells in breast cancer



Tumour Associated Macrophages

Karine Serre

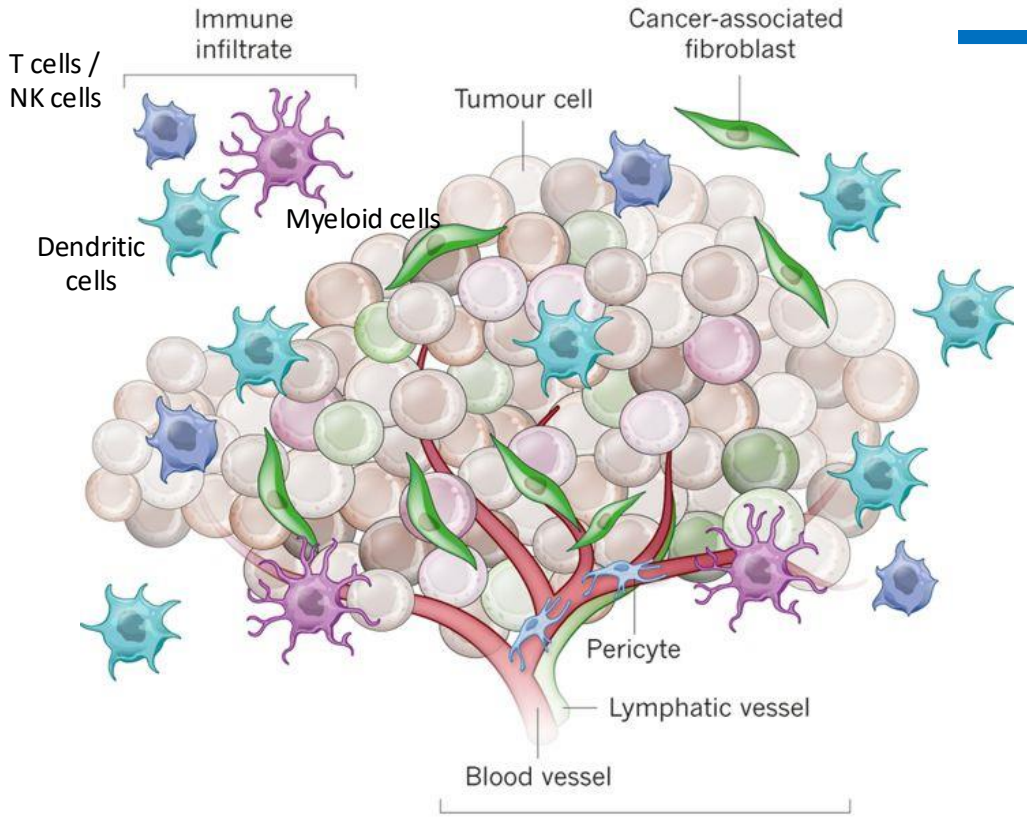
Carolina Jardim



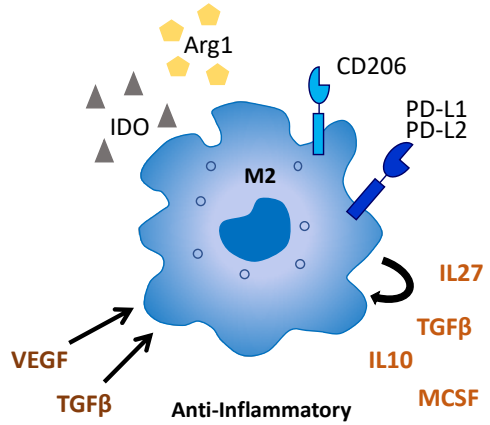
Anti-Tumoural

- Tumouricidal
- Immunostimulator
- Anti-angiogenic
- Th1 response

Primary Tumour Microenvironment (TME)



***Nature* volume 501, pages 346–354**



Pro-Tumoural

- Tumour survival
- Immunoregulator
- Hypoxia/Angiogenesis/Metastasis
- Therapeutic failure

Goals

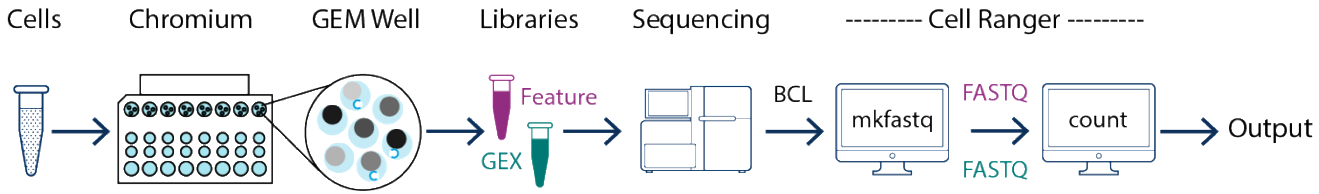
- ✓ Evaluate **the cellular diversity** and **molecular signatures** of myeloid cells with pro- and anti-tumour functions, focusing on **macrophages**
- ✓ Contribute to the creation of a **single-cell expression atlas** of myeloid lineage-specific anti-tumour molecular effectors
- ✓ Develop a **user-friendly open-source computational tool**, with an interactive graphical interface, to enable visual exploration, application of clustering algorithms and gene-centered analyses of single-cell transcriptomic data

Experimental assay

Karine Serre

Carolina Jardim

- Mouse syngeneic triple-negative breast cancer cell line model
- FACS (fluorescence-activated cell sorting) sorted myeloid cells coupled with 10X Genomics microdroplet technology



- Sequencing with an average of 50,000 reads/cell

Hypothesis

3 Treated tumours

9,209 cells

Experimentally-induced anti-tumour macrophages

3 Early/Untreated tumours

11,520 cells

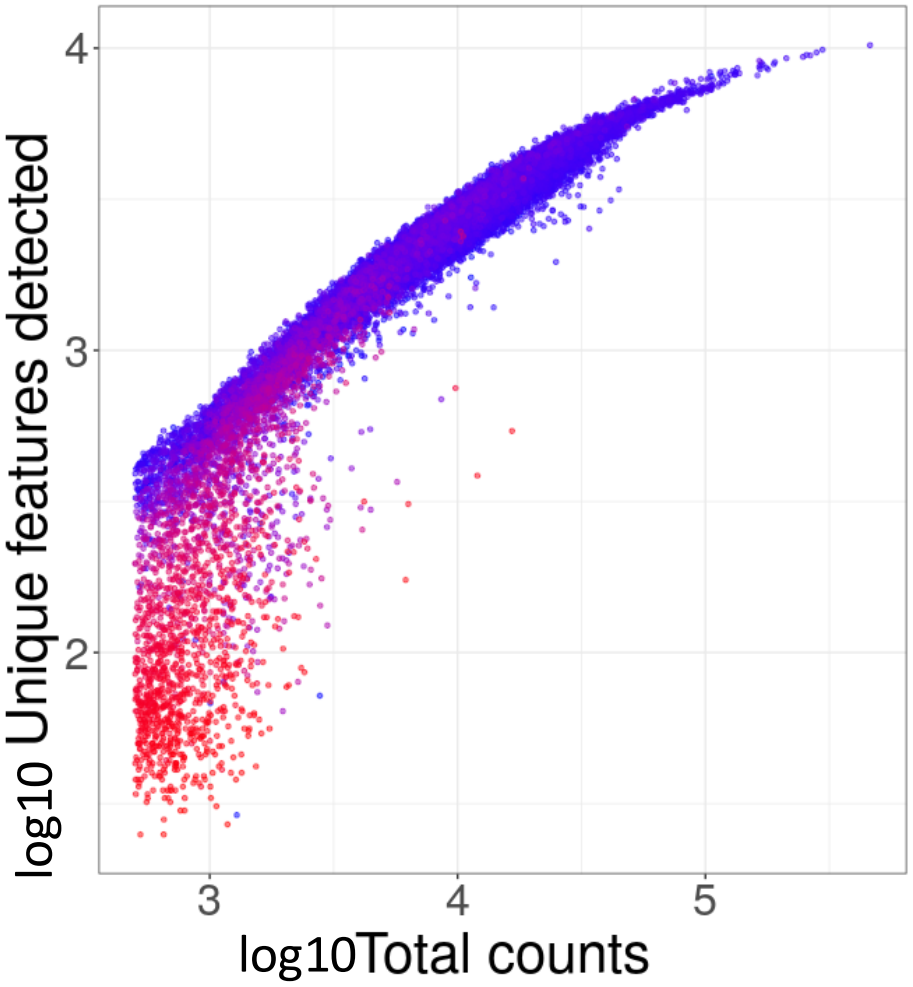
Ongoing battle: anti- and pro-tumour macrophages

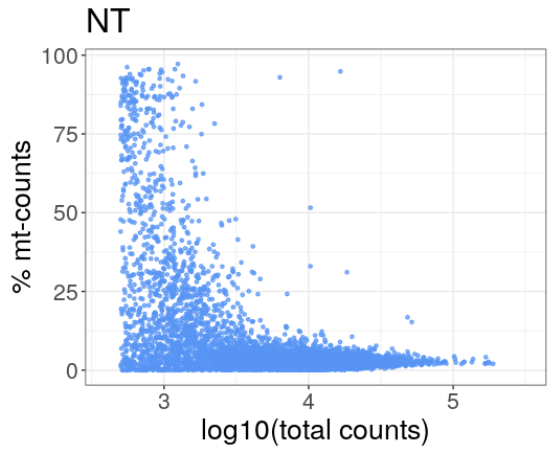
3 Late/Big tumours

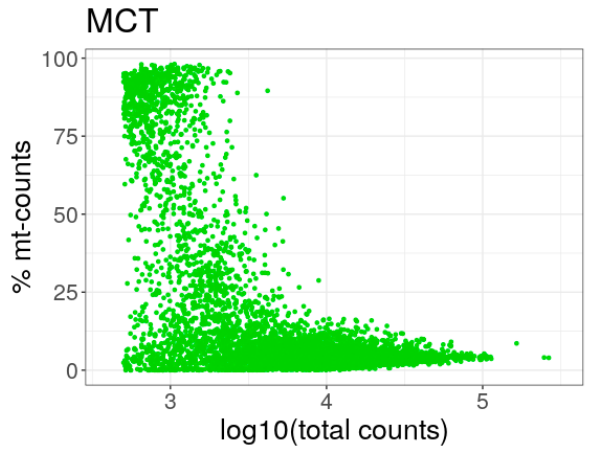
11,079 cells

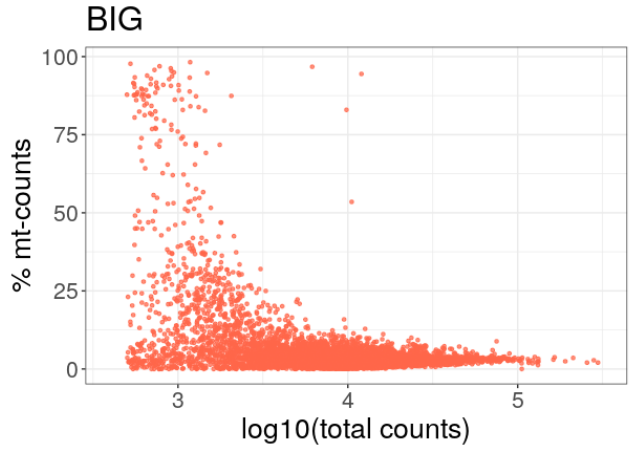
Battle is lost: pro-tumour macrophages

Quality Control

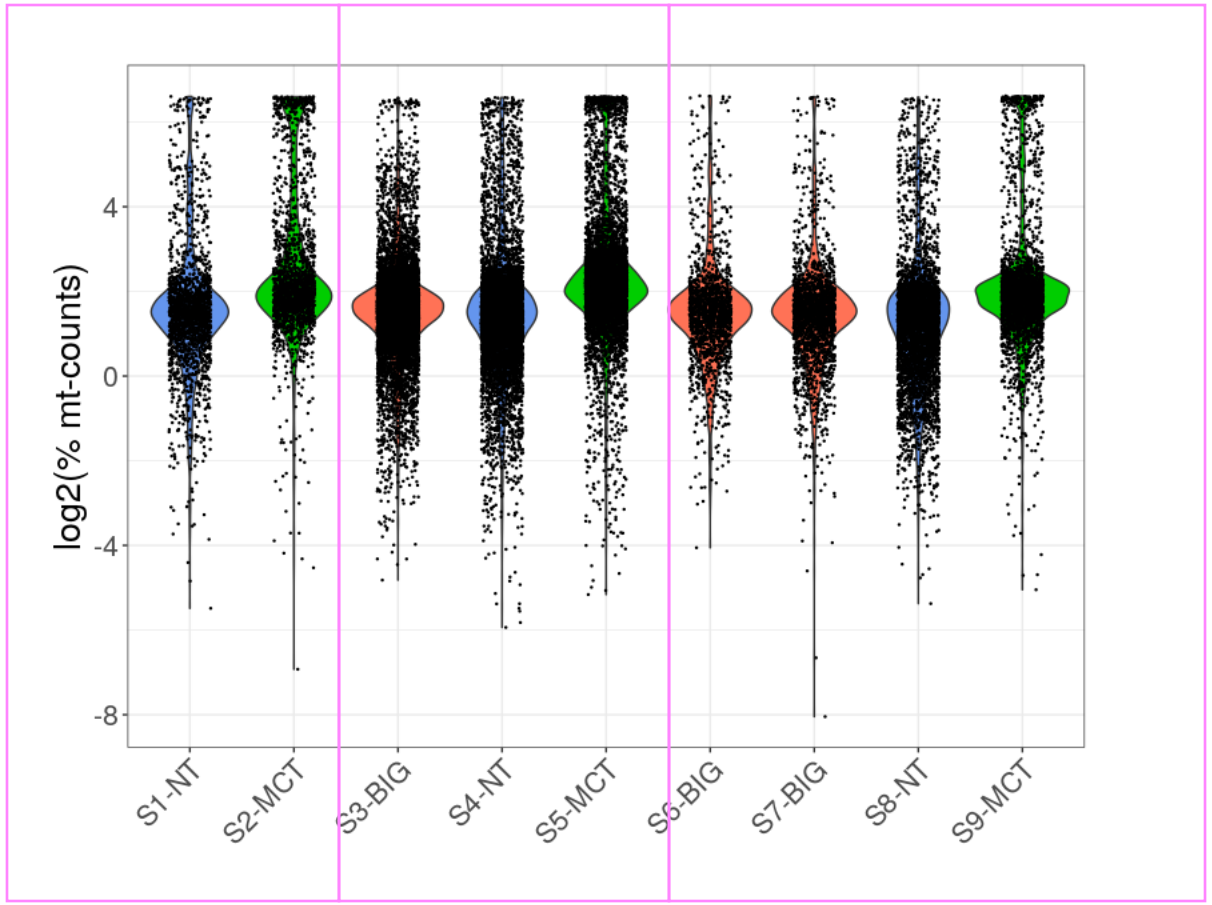

log10 Unique features detected
log10Total counts
% MT-gene counts
75
50
25
0


NT
% mt-counts
log10(total counts)

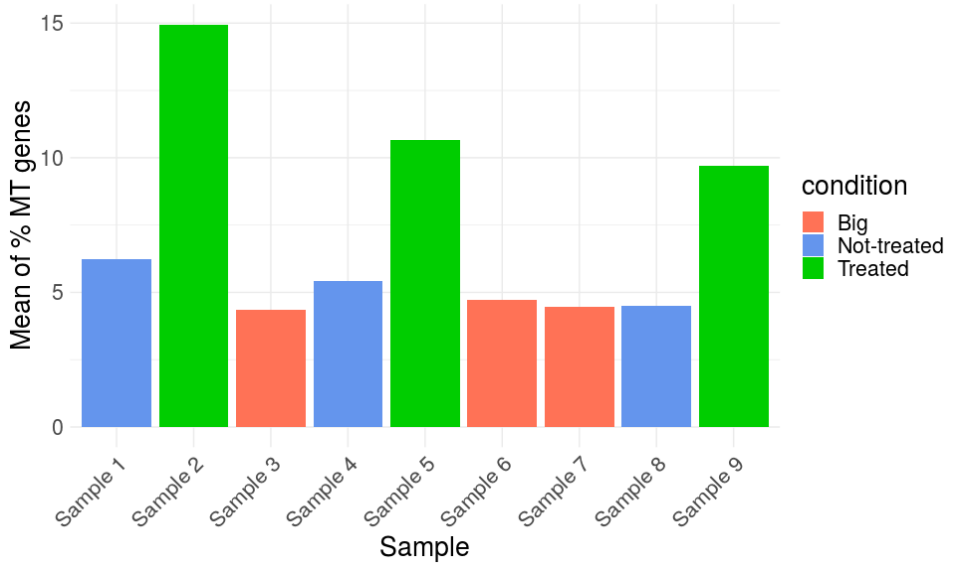

MCT
% mt-counts
log10(total counts)


BIG
% mt-counts
log10(total counts)

Biological vs technical



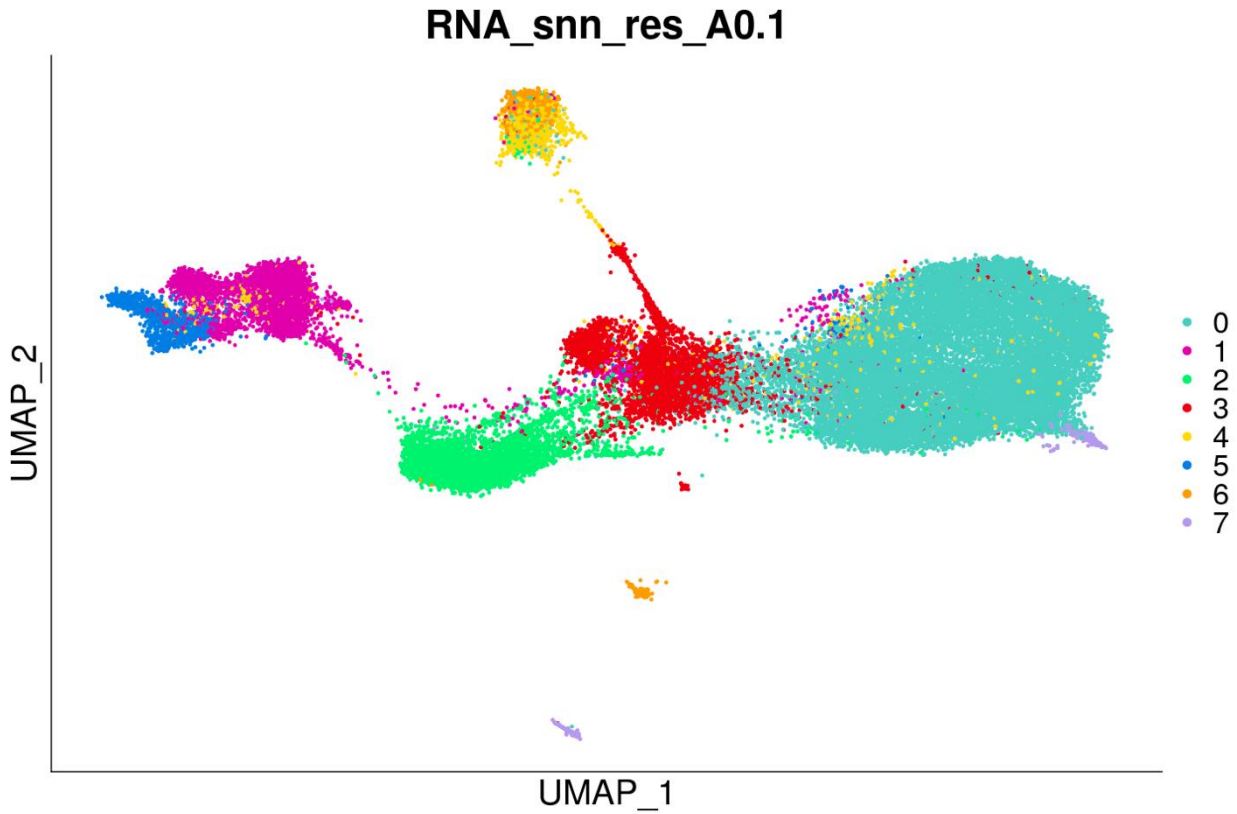
Day 1 Day 2 Day 3

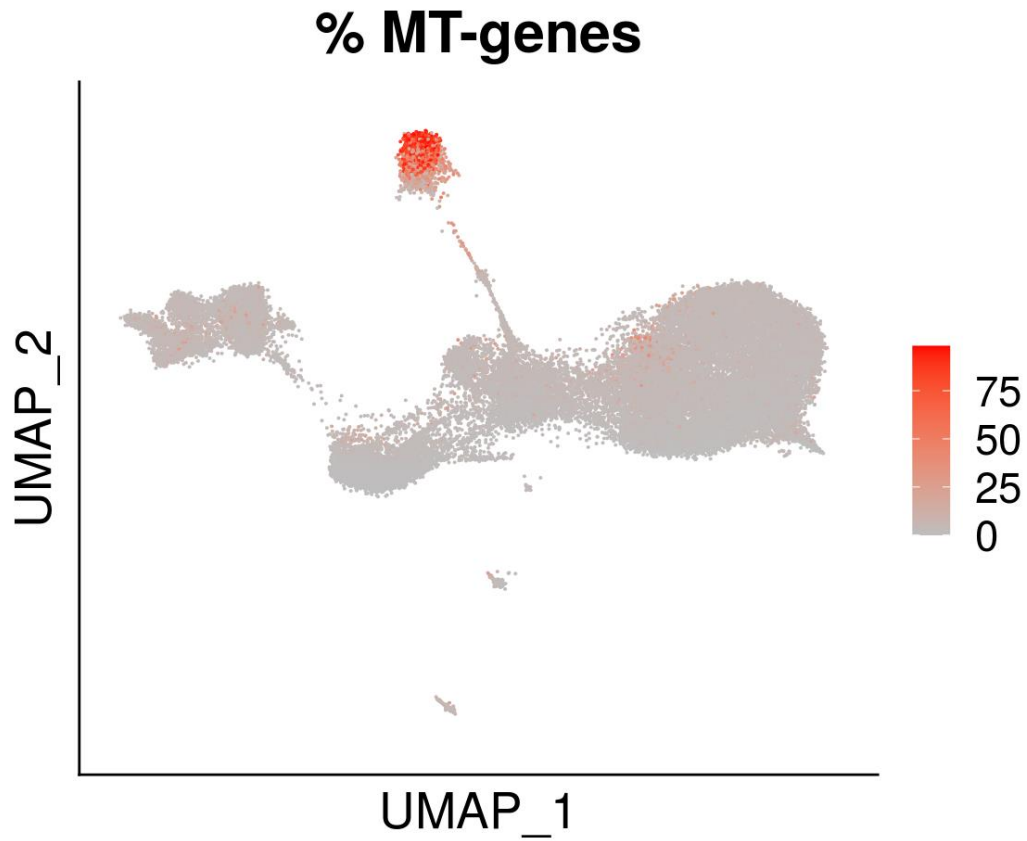


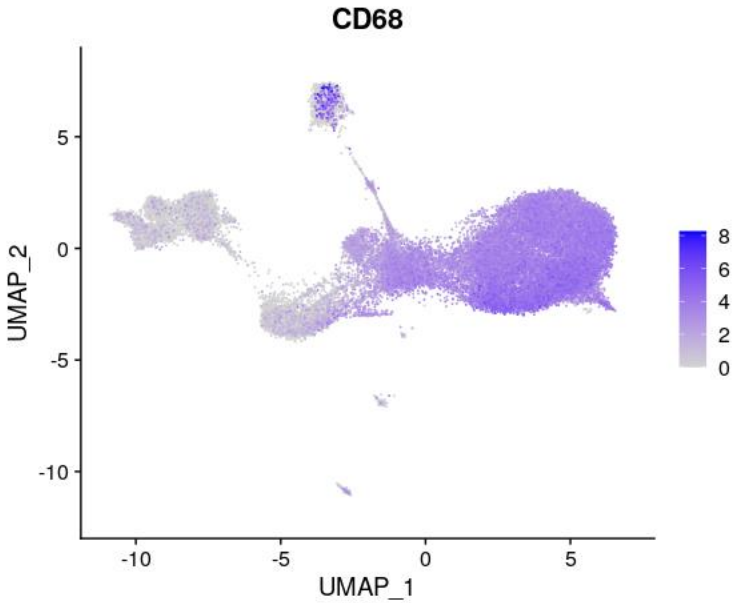

```
> summary(colData_df[sce$condition == "Treated",]$pct_counts_MT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.015   4.128  11.146   6.169  98.117
>
```

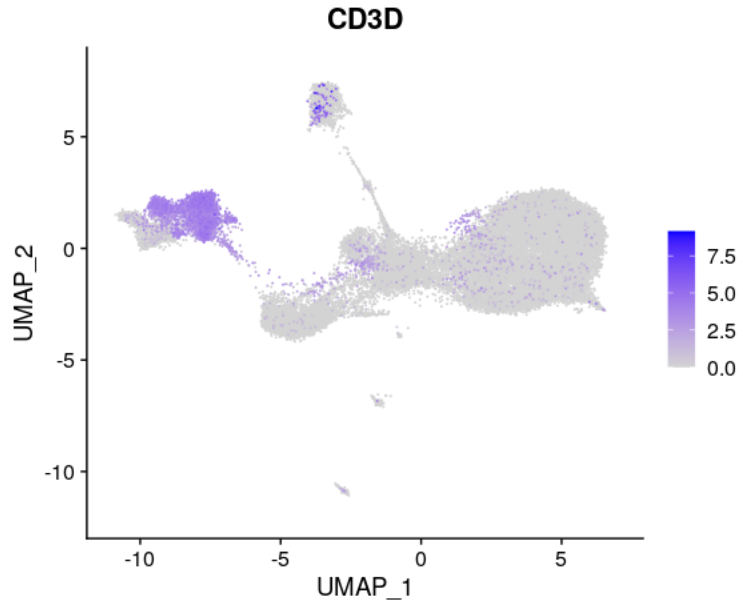
```
> summary(colData_df[sce$condition == "Not-treated",]$pct_counts_MT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.595   2.618   5.238   3.805  97.189
>
```

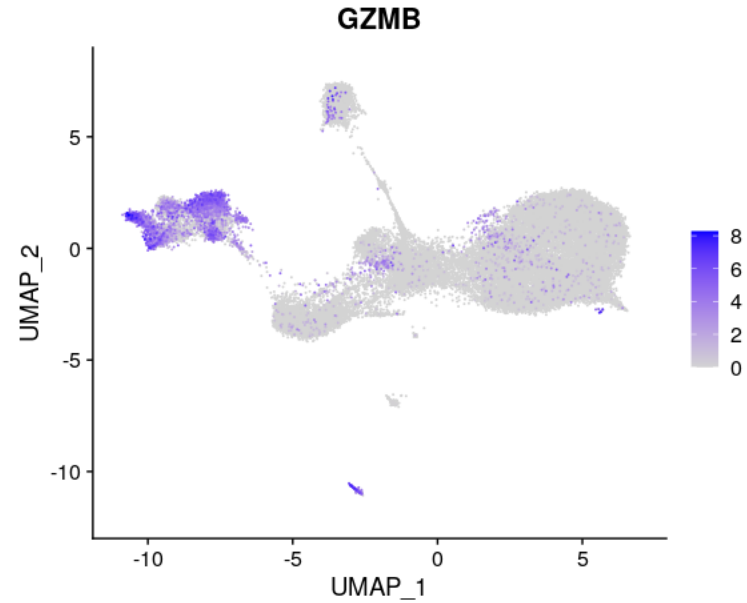
```
> summary(colData_df[sce$condition == "Big",]$pct_counts_MT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.035   2.899   4.421   3.893  98.217
>
```

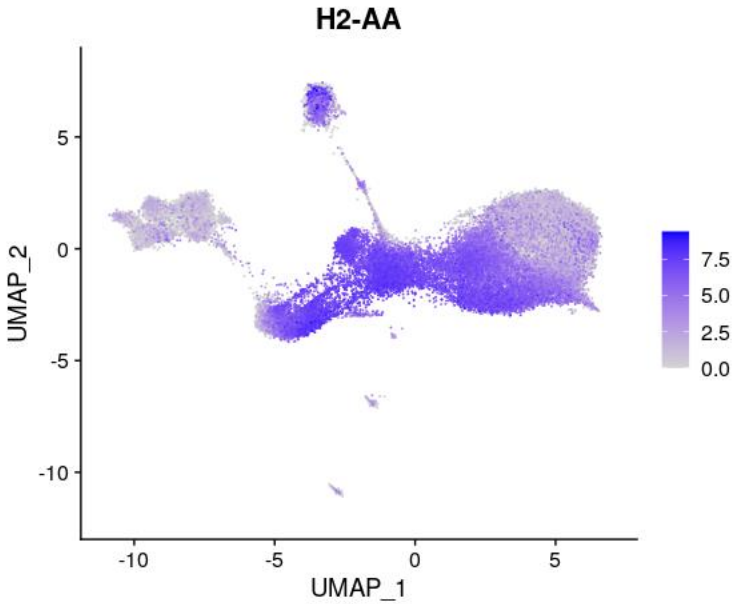

RNA_snn_res_A0.1
UMAP_2
UMAP_1
0
1
2
3
4
5
6
7

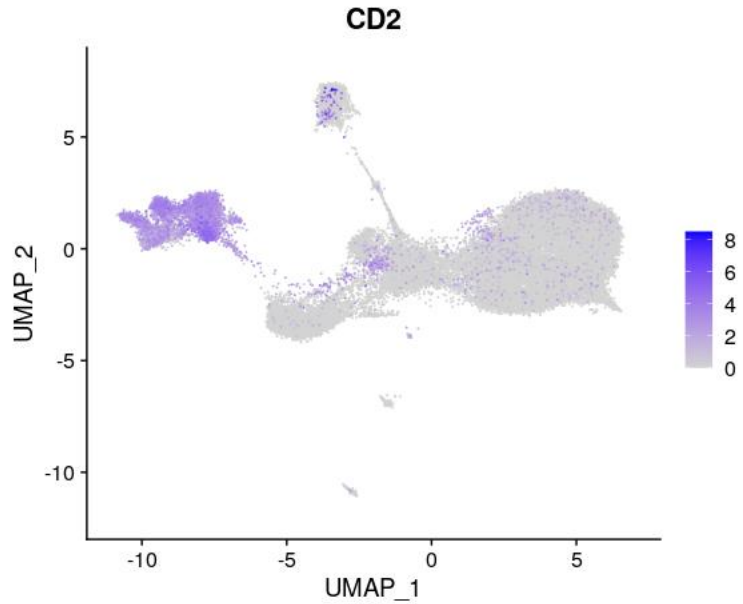

% MT-genes
UMAP_2
UMAP_1
75
50
25
0


CD68
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5
8
6
4
2
0


CD3D
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5
7.5
5.0
2.5
0.0

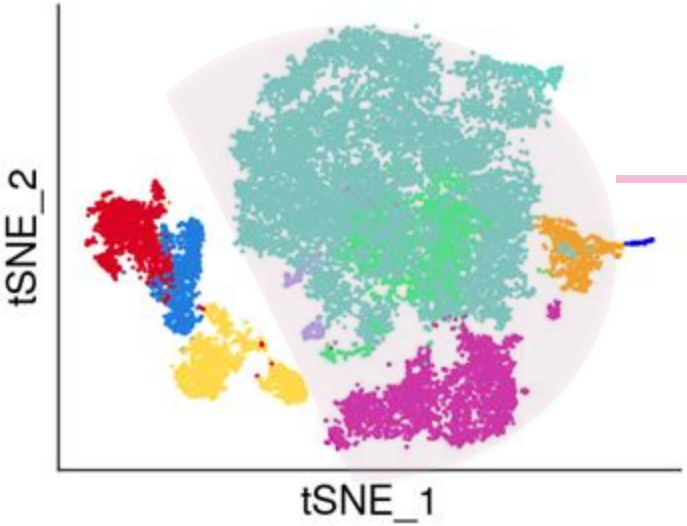

GZMB
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5
8
6
4
2
0


H2-AA
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5
7.5
5.0
2.5
0.0

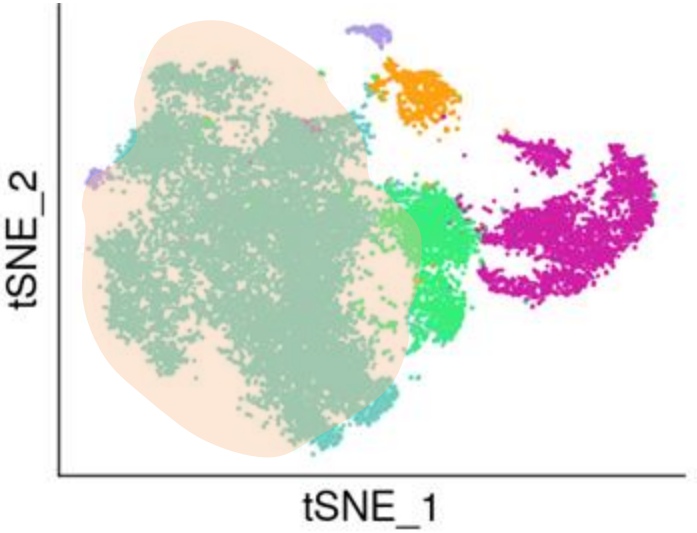

CD2
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5
8
6
4
2
0

Subset macrophages

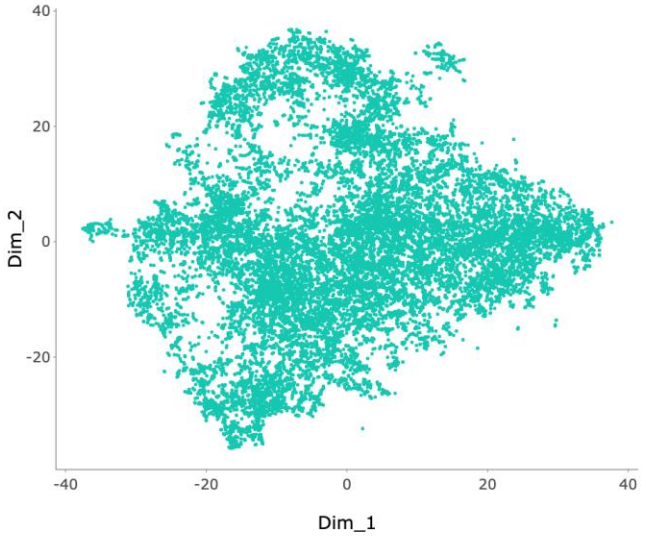
All cells



Myeloid cells



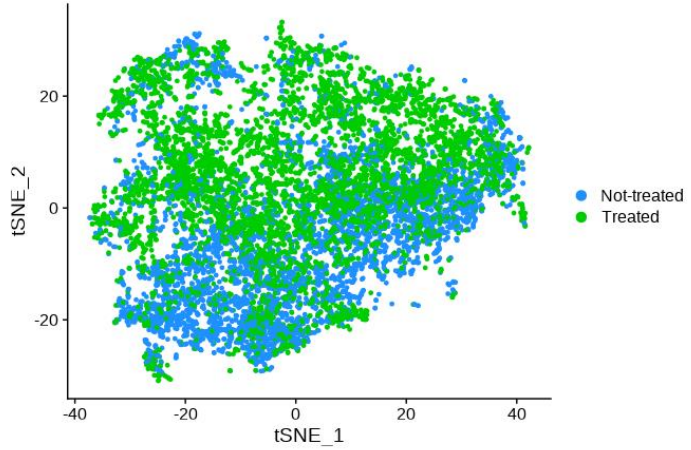
Macrophages

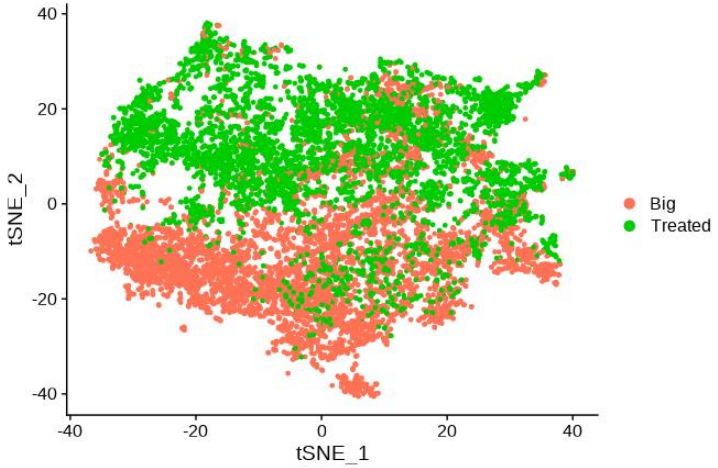


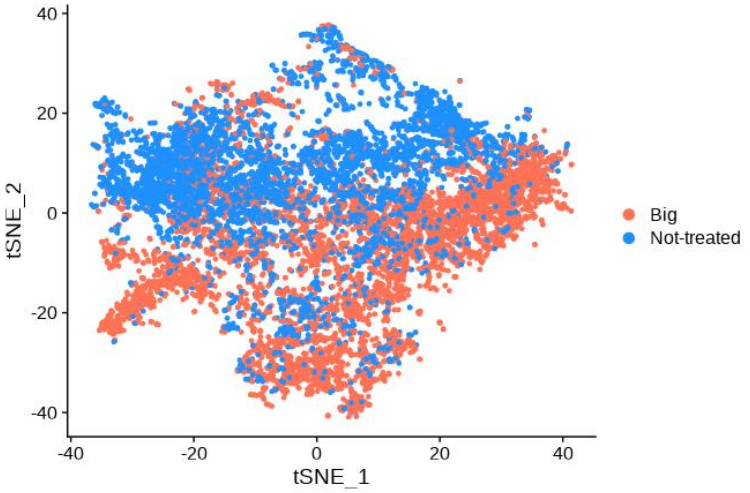
CD11b, CD14, Lyz2, Ly6c2, CD68, Fcgr3, Fcgr1, Adgre1, Csf1r, Mafb

Clustering analysis strategies

- Individual samples
- Condition
- <u>Pairs of conditions</u>
- All macrophages

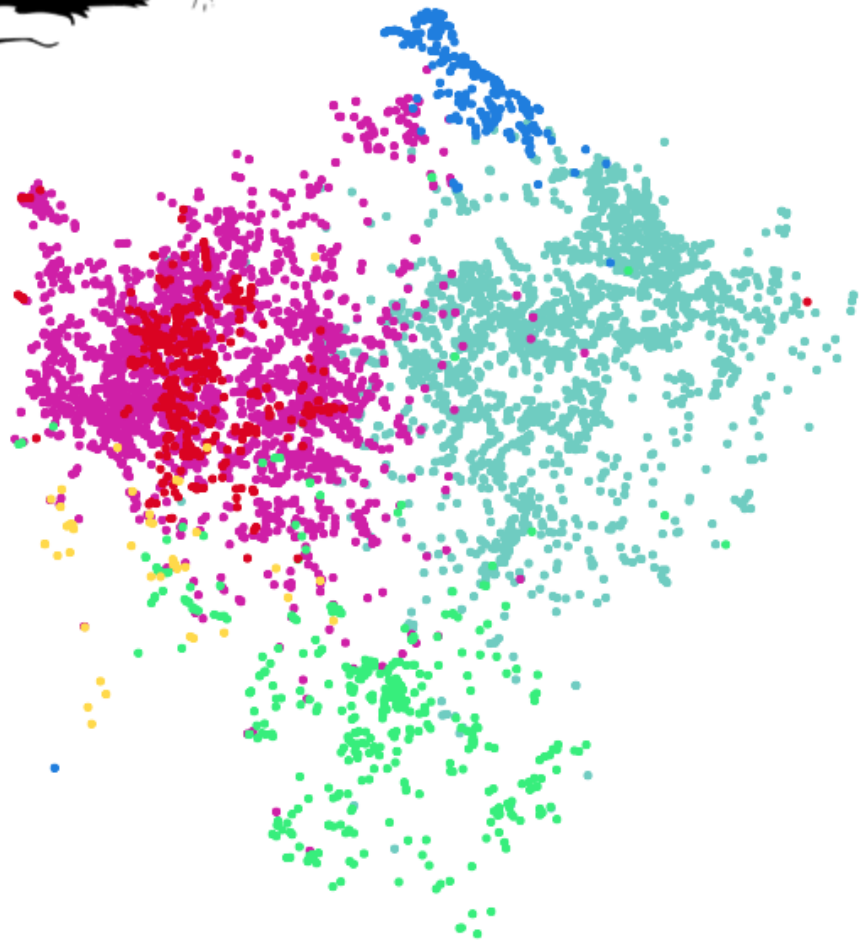






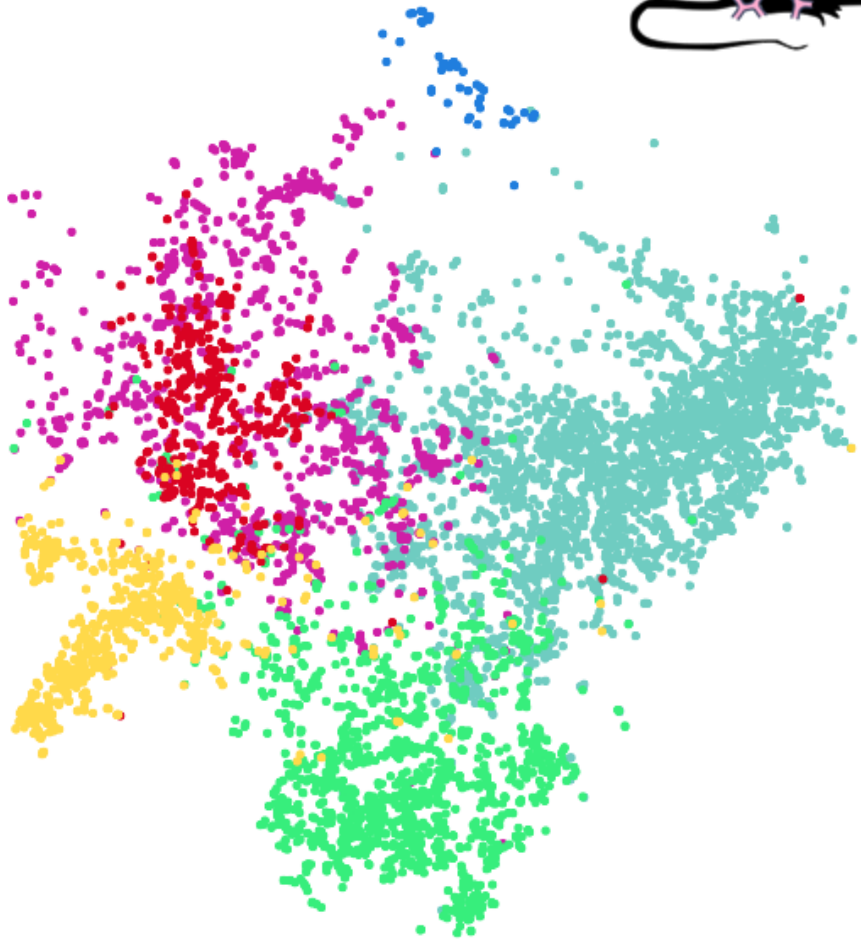
Early-stage (not-treated)

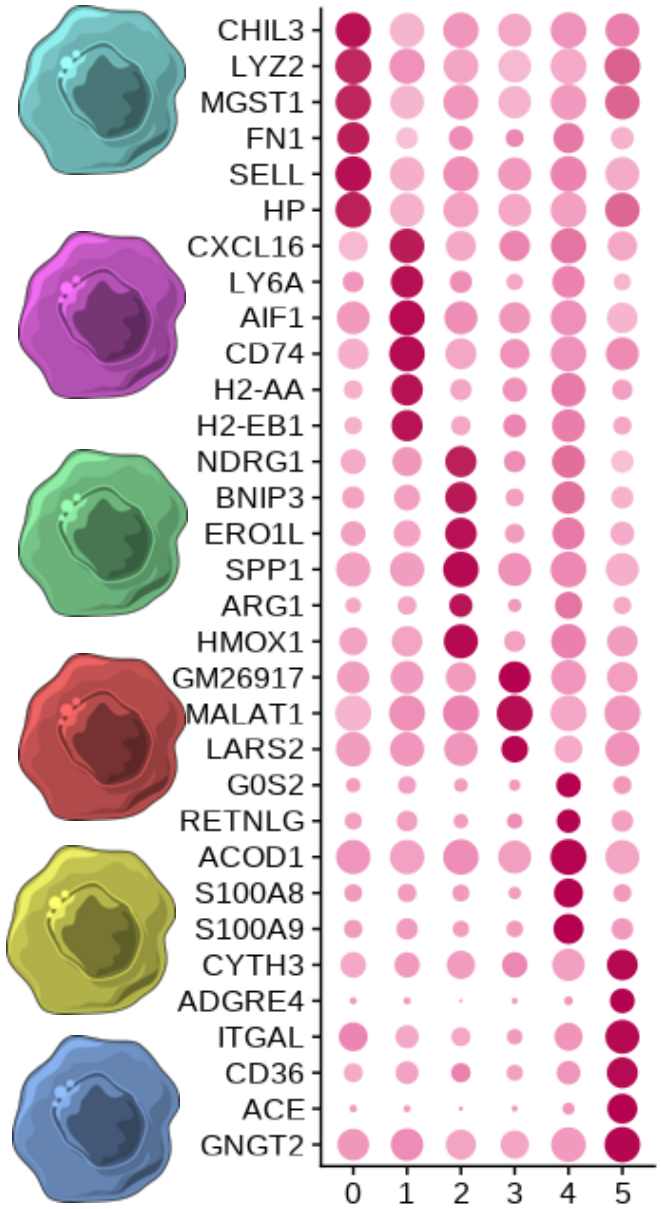
11,520 cells

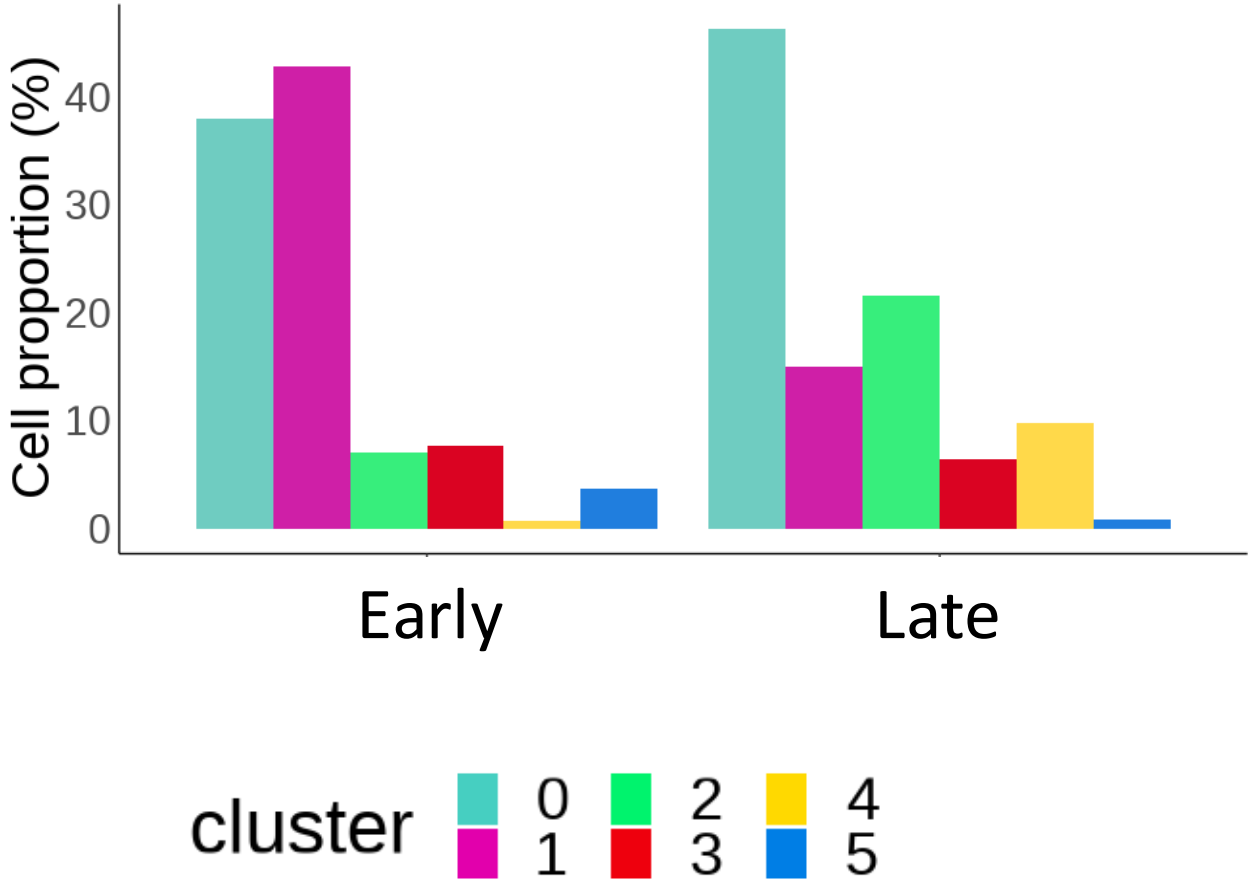


Late-stage (big)

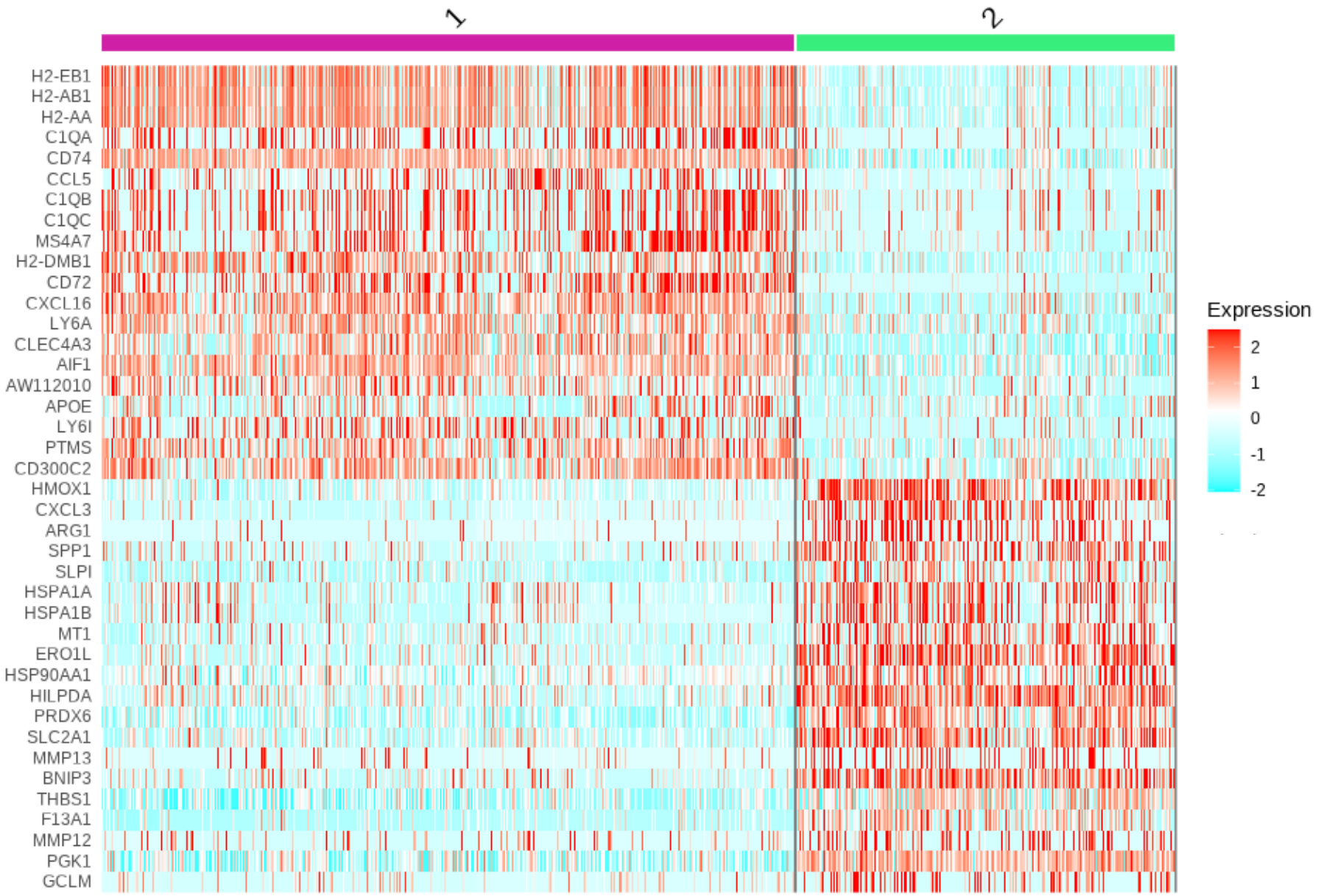
11,079 cells

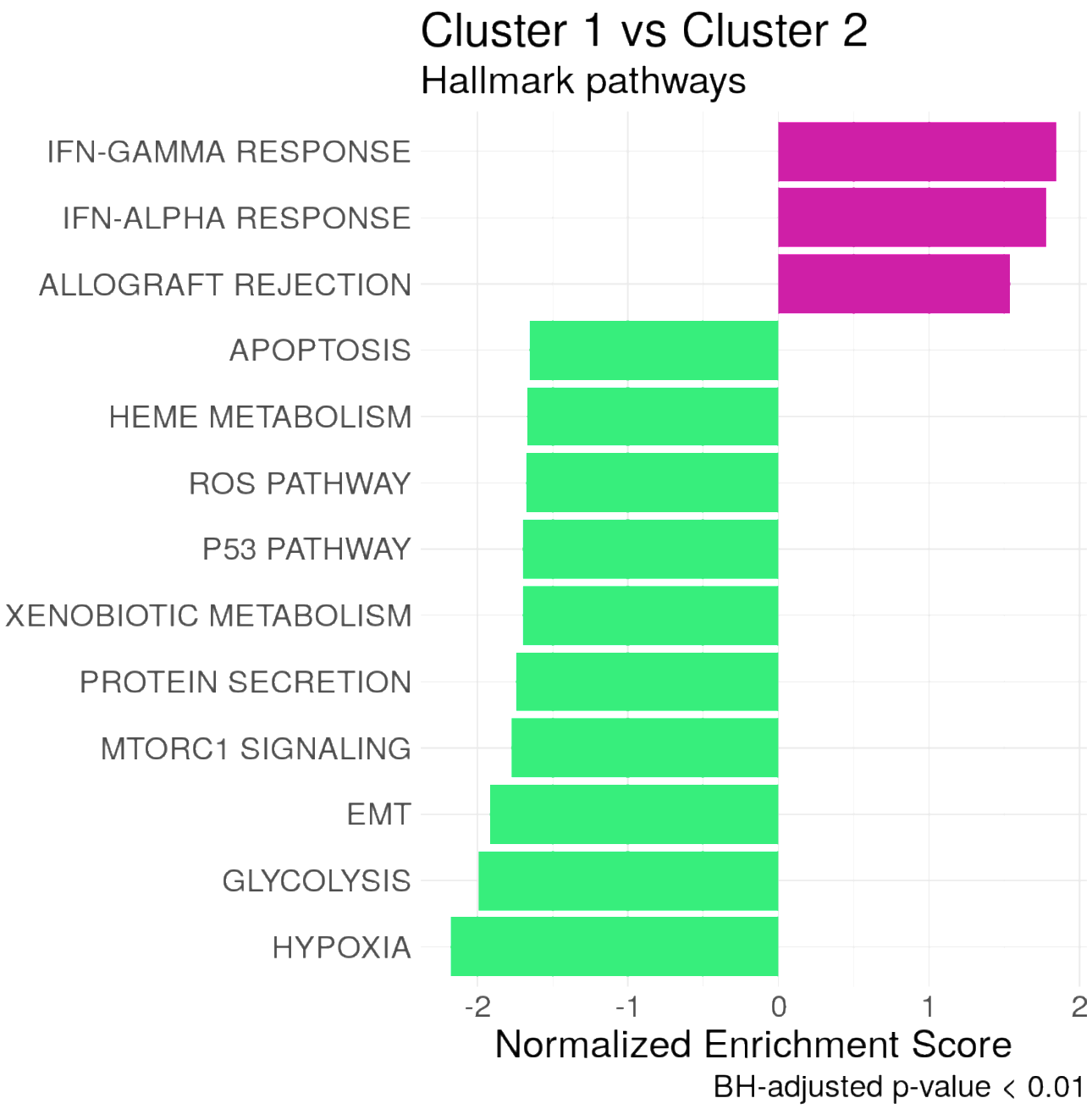


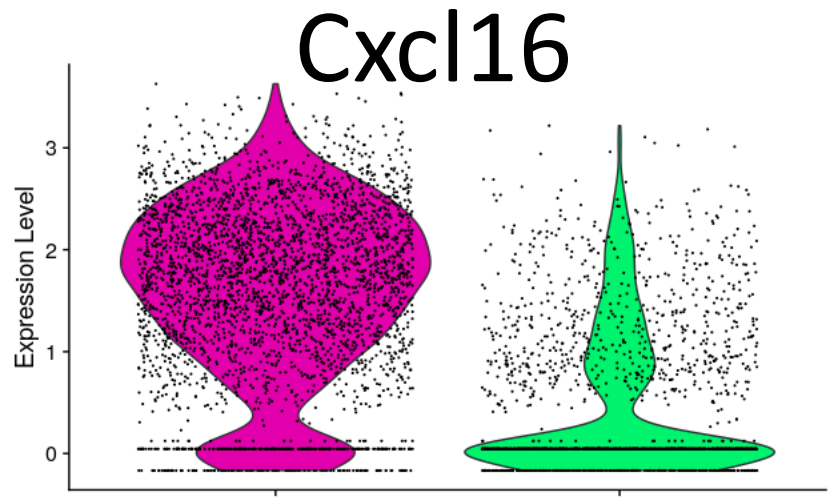

CHIL3
LYZ2
MGST1
FN1
SELL
HP
CXCL16
LY6A
AIF1
CD74
H2-AA
H2-EB1
NDRG1
BNIP3
ERO1L
SPP1
ARG1
HMOX1
GM26917
MALAT1
LARS2
G0S2
RETNLG
ACOD1
S100A8
S100A9
CYTH3
ADGRE4
ITGAL
CD36
ACE
GNGT2
0 1 2 3 4 5
Average expression
2
1
0
-1
Expressing cells (%)
25
50
75
100

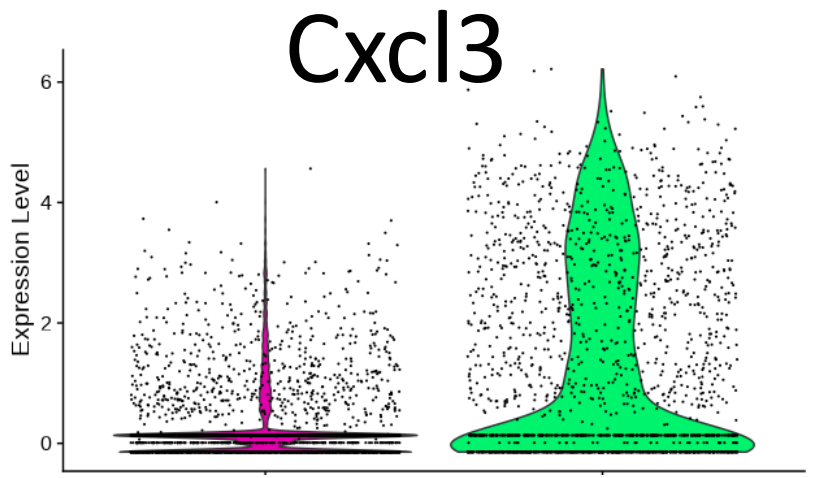

Cell proportion (%)
0
10
20
30
40
Early
Late
cluster
0
1
2
3
4
5

Cluster 1 vs Cluster 2

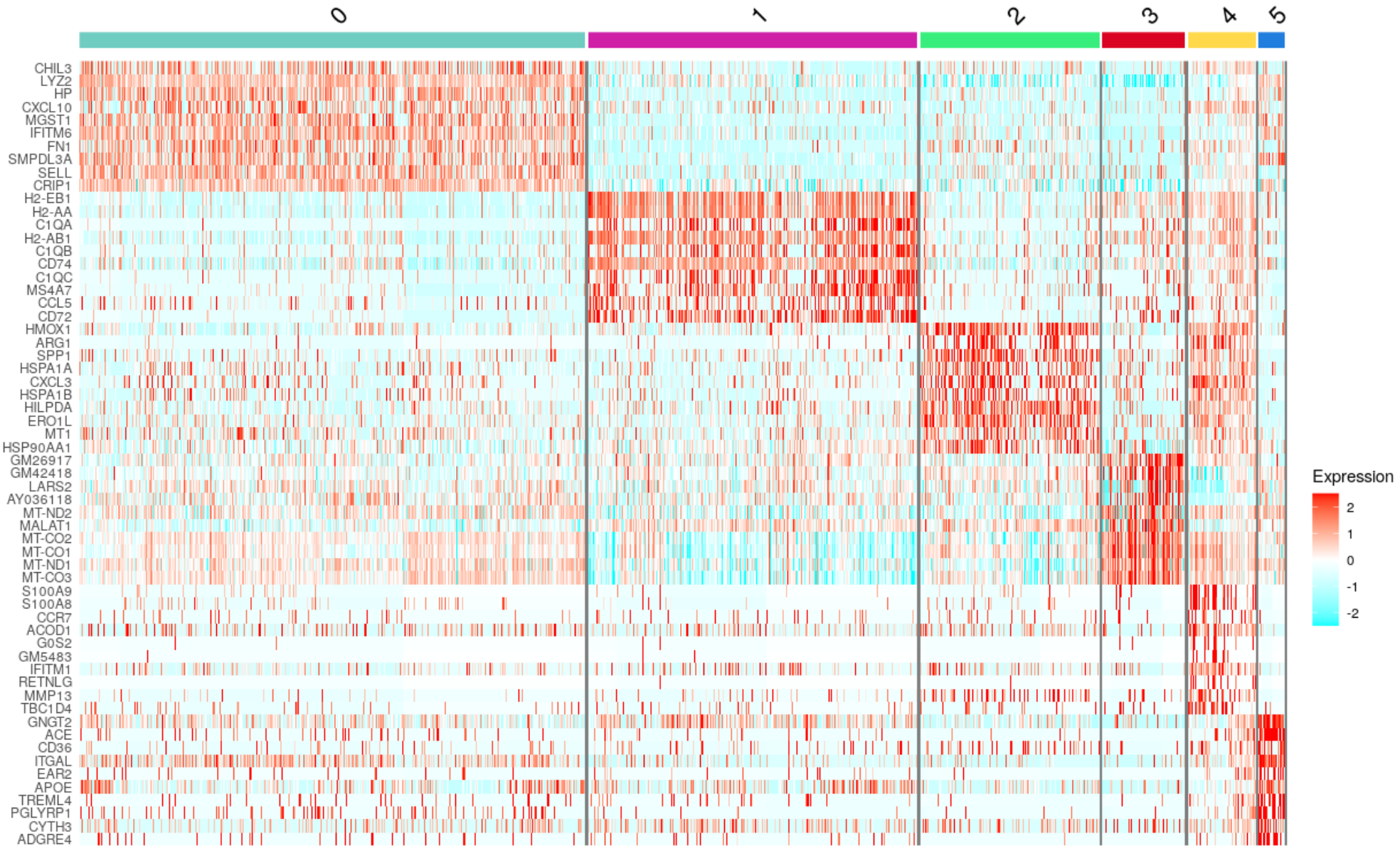


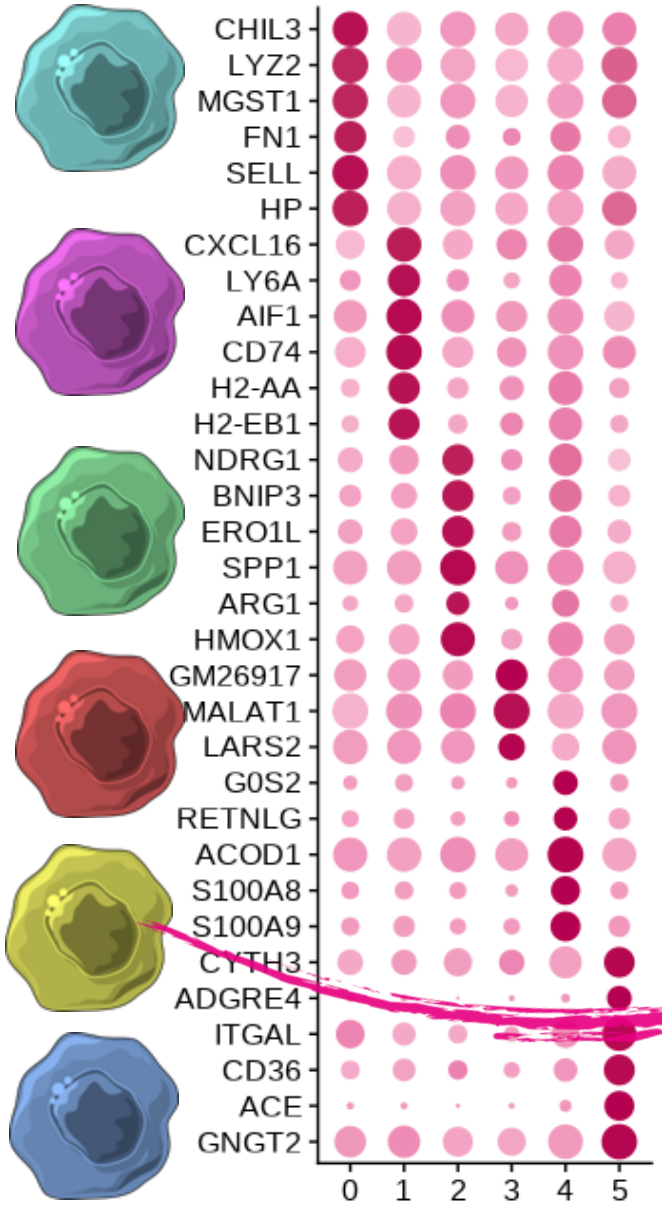

Cluster 1 vs Cluster 2
Hallmark pathways
IFN-GAMMA RESPONSE
IFN-ALPHA RESPONSE
ALLOGRAFT REJECTION
APOPTOSIS
HEME METABOLISM
ROS PATHWAY
P53 PATHWAY
XENOBIOTIC METABOLISM
PROTEIN SECRETION
MTORC1 SIGNALING
EMT
GLYCOLYSIS
HYPOXIA
-2
-1
0
1
2
Normalized Enrichment Score
BH-adjusted p-value < 0.01

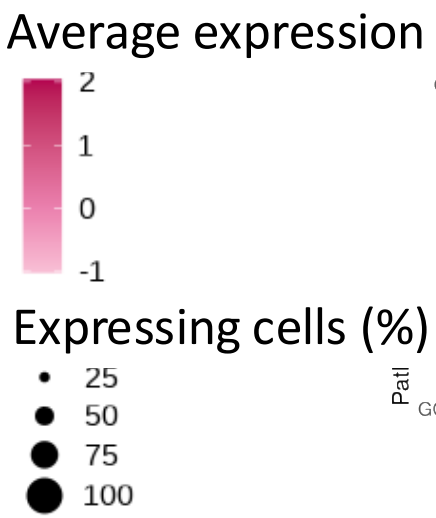

Cxcl16
Expression Level
0
1
2
3

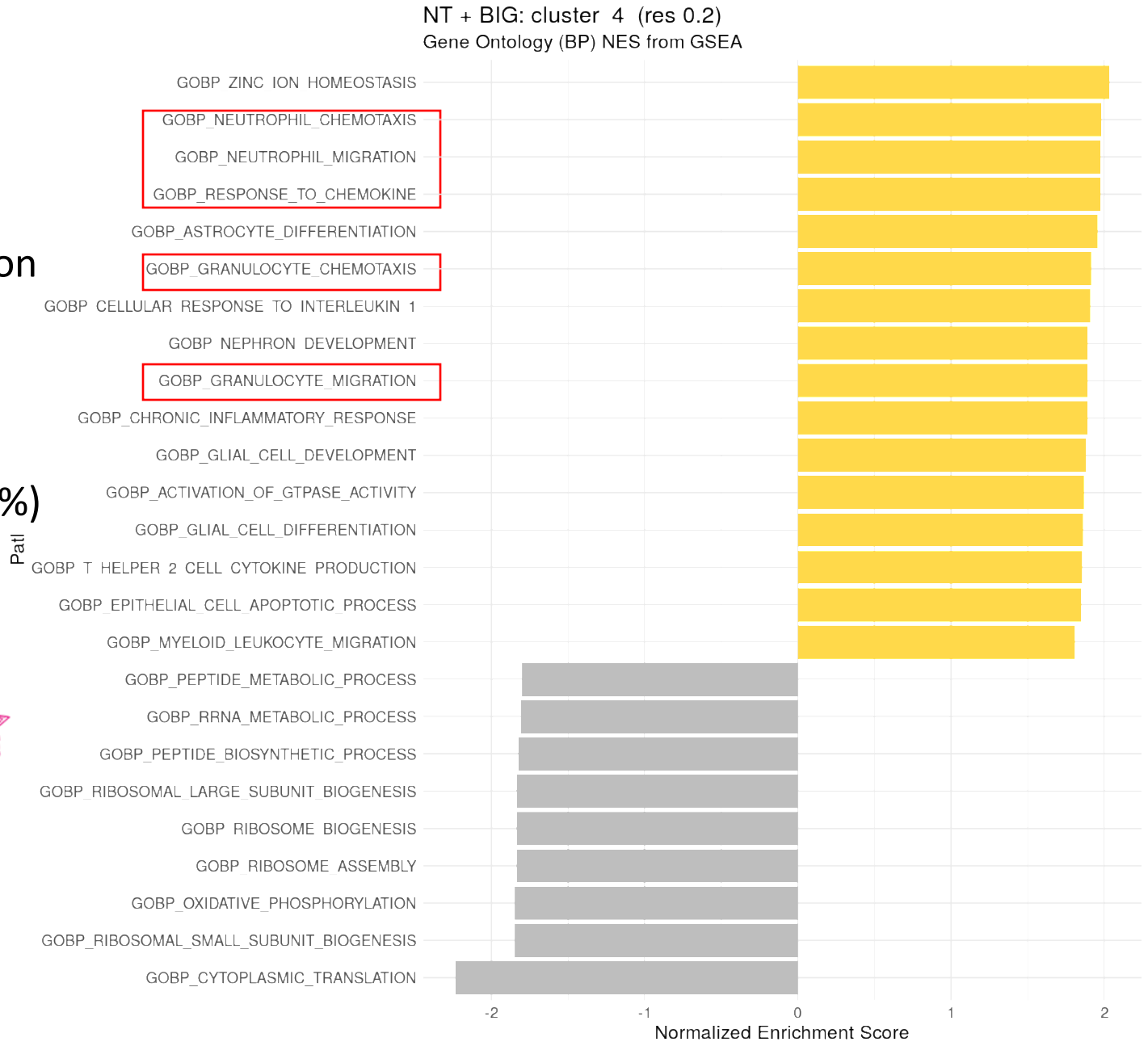

Cxcl3
Expression Level
0
2
4
6

One cluster vs others

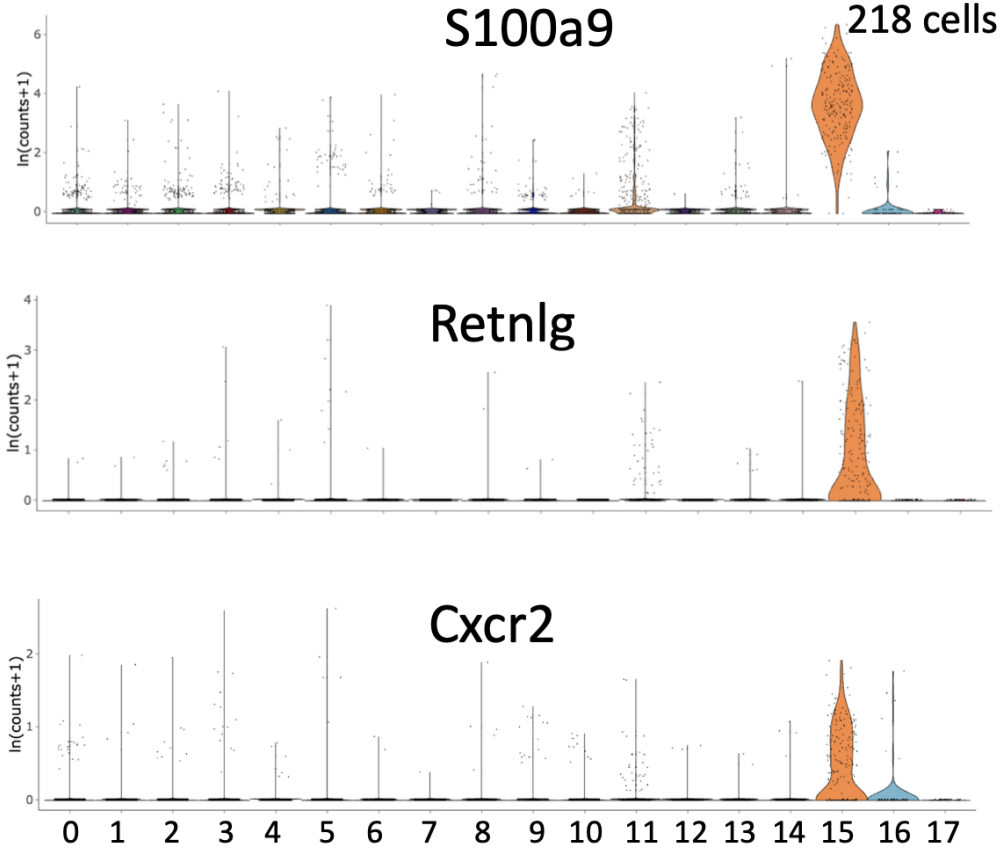


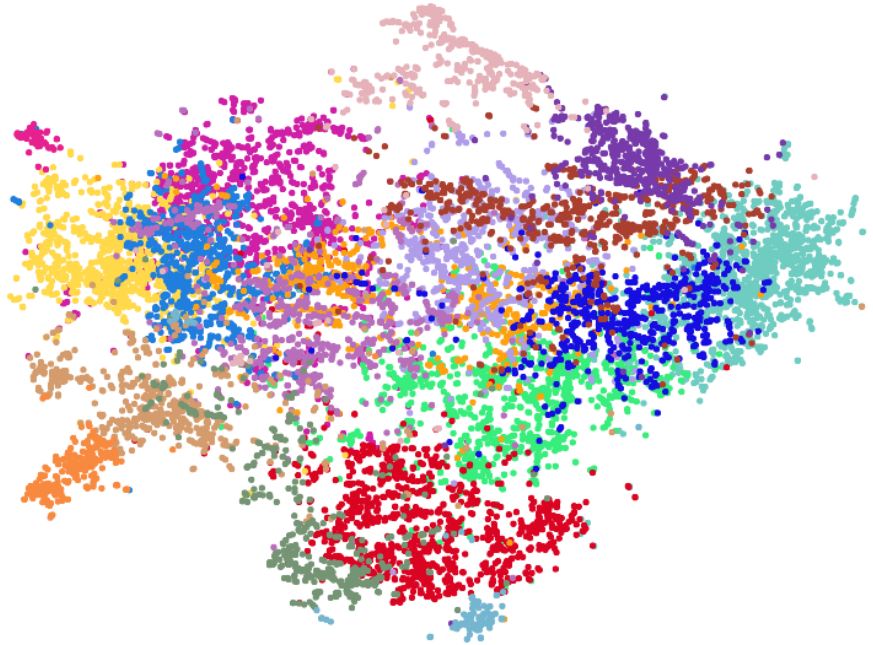

CHIL3
LYZ2
MGST1
FN1
SELL
HP
CXCL16
LY6A
AIF1
CD74
H2-AA
H2-EB1
NDRG1
BNIP3
ERO1L
SPP1
ARG1
HMOX1
GM26917
MALAT1
LARS2
G0S2
RETNLG
ACOD1
S100A8
S100A9
CYTH3
ADGRE4
ITGAL
CD36
ACE
GNGT2
0 1 2 3 4 5

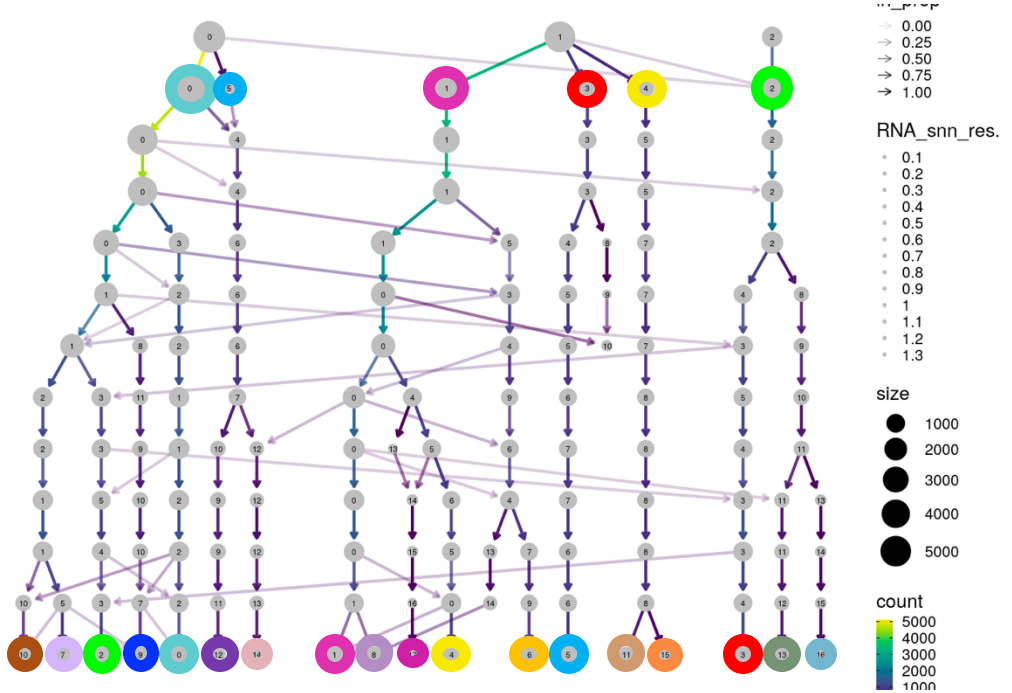

Average expression
2
1
0
-1
Expressing cells (%)
25
50
75
100

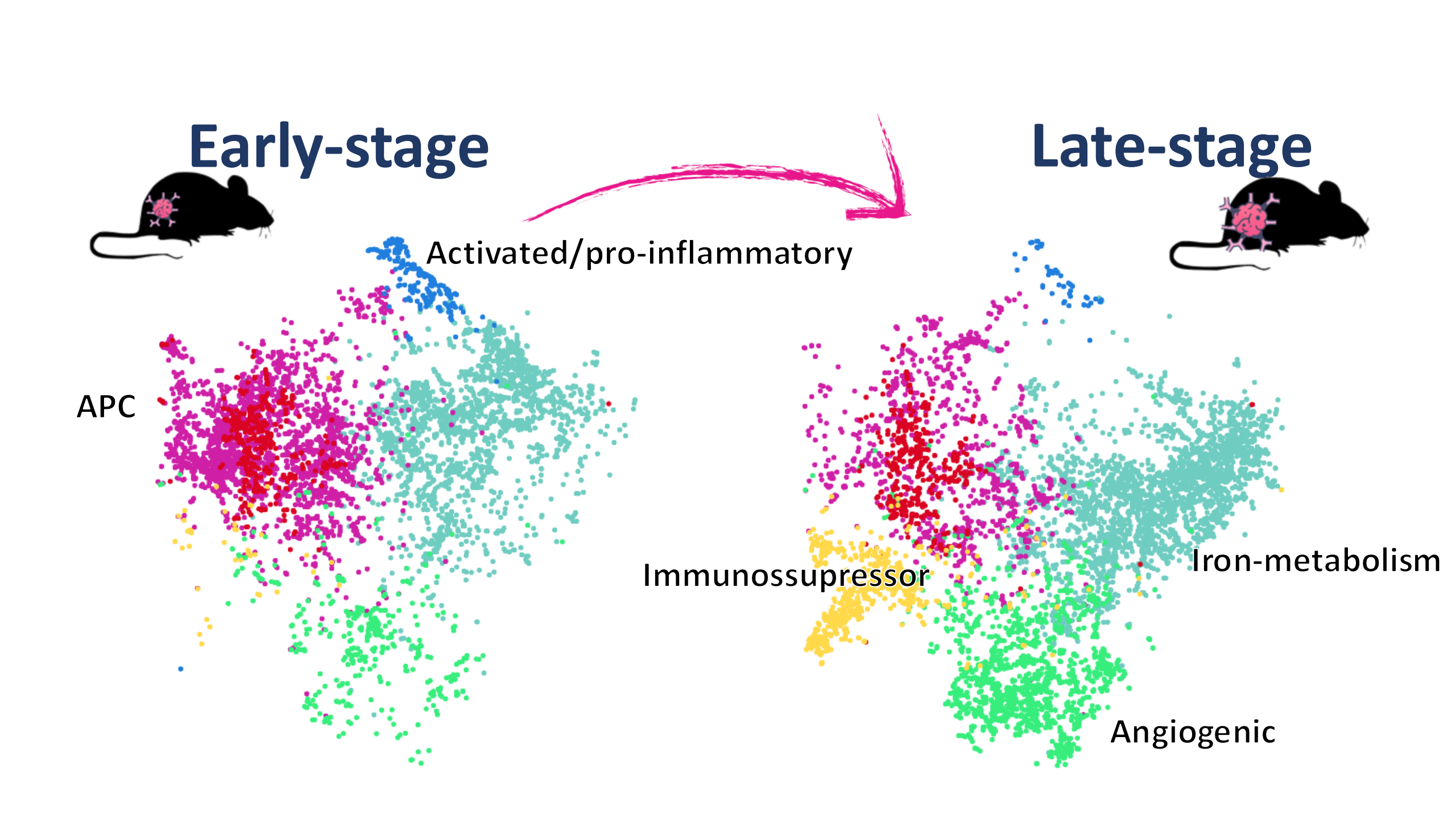

NT + BIG: cluster 4 (res 0.2)
Gene Ontology (BP) NES from GSEA
GOBP_ZINC_ION_HOMEOSTASIS
GOBP_NEUTROPHIL_CHEMOTAXIS
GOBP_NEUTROPHIL_MIGRATION
GOBP_RESPONSE_TO_CHEMOKINE
GOBP_ASTROCYTE_DIFFERENTIATION
GOBP_GRANULOCYTE_CHEMOTAXIS
GOBP_CELLULAR_RESPONSE_TO_INTERLEUKIN_1
GOBP_NEPHRON_DEVELOPMENT
GOBP_GRANULOCYTE_MIGRATION
GOBP_CHRONIC_INFLAMMATORY_RESPONSE
GOBP_GLIAL_CELL_DEVELOPMENT
GOBP_ACTIVATION_OF_GTPASE_ACTIVITY
GOBP_GLIAL_CELL_DIFFERENTIATION
GOBP_T_HELPER_2_CELL_CYTOKINE_PRODUCTION
GOBP_EPITHELIAL_CELL_APOPTOTIC_PROCESS
GOBP_MYELOID_LEUKOCYTE_MIGRATION
GOBP_PEPTIDE_METABOLIC_PROCESS
GOBP_RRNA_METABOLIC_PROCESS
GOBP_PEPTIDE_BIOSYNTHETIC_PROCESS
GOBP_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS
GOBP_RIBOSOME_BIOGENESIS
GOBP_RIBOSOME_ASSEMBLY
GOBP_OXIDATIVE_PHOSPHORYLATION
GOBP_RIBOSOMAL_SMALL_SUBUNIT_BIOGENESIS
GOBP_CYTOPLASMIC_TRANSLATION
Patl
-2 -1 0 1 2
Normalized Enrichment Score
BH-adjusted p-value < 0.05 & abs(NES) > 1.8

Identification of cell types using different clustering resolutions

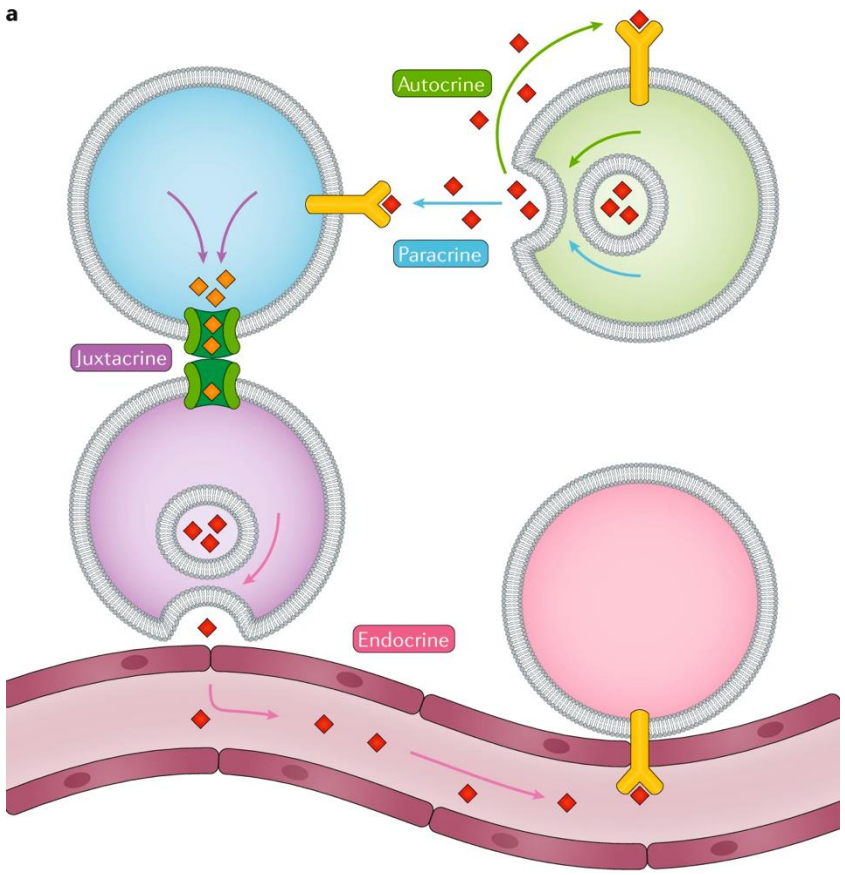


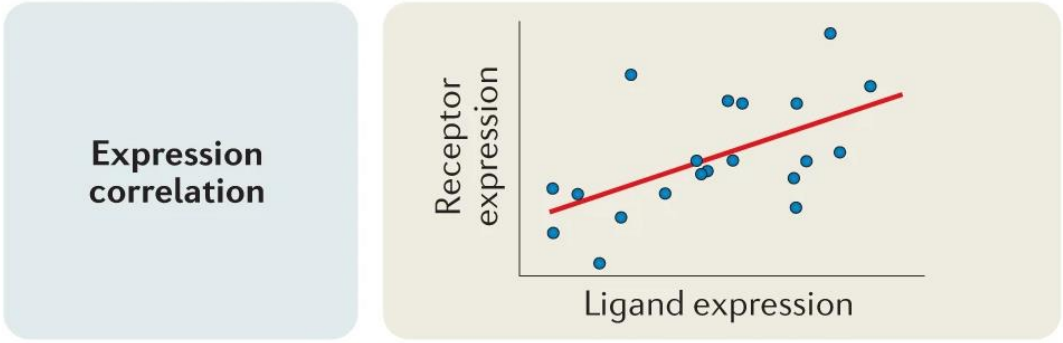


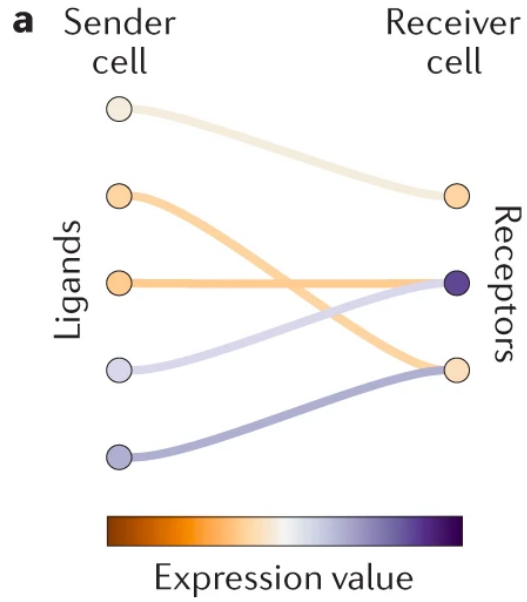



Early-stage
Late-stage
Activated/pro-inflammatory
APC
Immunossupressor
Iron-metabolism
Angiogenic

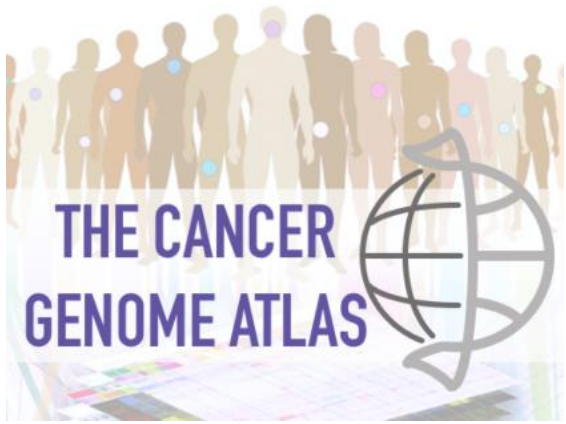
Future work: deciphering cell–cell interactions and communication from gene expression analysis

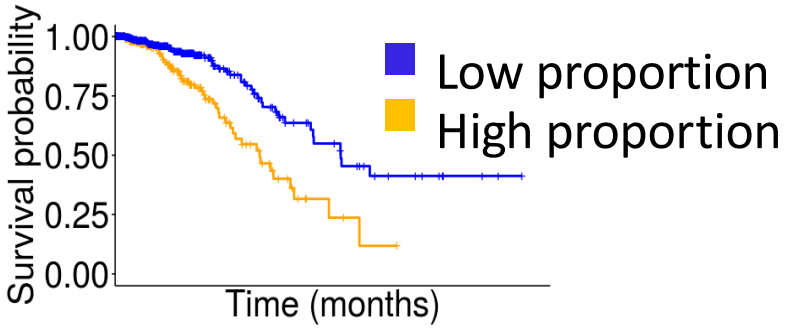


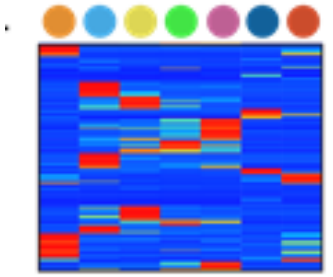




Future work: digital cytometry







Signature matrix

Decision support system: scStudio

- Real time user interface
- Shows graphical presentation
- Quick overview
- Supports interactions with data
- Drilling down into underlying details