MONASH University

MONASH
BIOINFORMATICS
PLATFORM

Celaref: Annotating single-cell RNAseq clusters by similarity to reference datasets

Sarah Williams

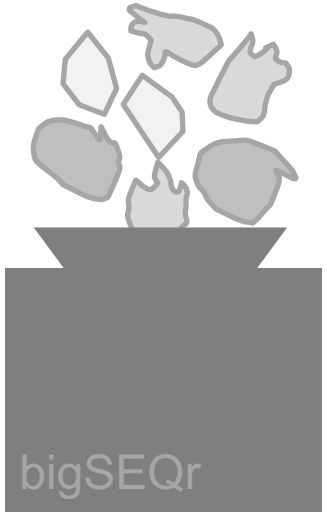
Bioconductor Asia 2018

GROUP OF EIGHT AUSTRALIA

@MonashBioinfo
bioinformatics.platform@monash.edu
https://platforms.monash.edu/bioinformatics/

scRNAseq outline

Cells

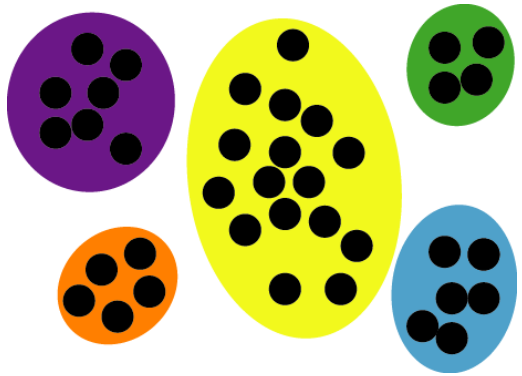


Fastq sequence

Counts table

	cell001	cell002	cell003	cell004	cell005	cell006	cell007	cell008	...
SYN1									
SHANK3									
KCNC1									
RPS6KA2									
KCNA4									
KLF4									
LRRC2									
KCNC2									
P53									
...									

cell-cluster assignment



Sequence

Count gene expression per cell

Cluster cells

Clustering – then what?

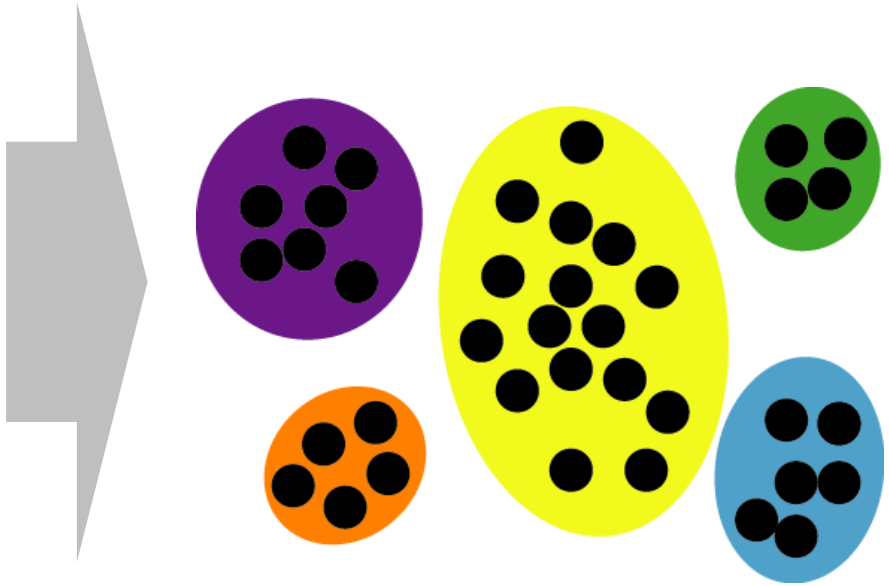
Many

different

clustering

tools

exist



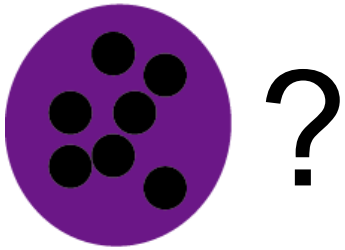
What type of cells are in each cluster?

Celaref is not one of them!

What's in that cluster?

How to annotate my cell clusters?

- **Don't.**
  - Let reference data inform the clustering.
- **Manually**
  - Look at marker genes / expression patterns
- **Cell cluster labelling tools**
  - Scmap/scfind
  - MUDAN
  - **Celaref**



Why another tool? - celaref

Celaref: Per-cluster **Ce**ll **La**belling by **Ref**erence

Bioconductor : https://bioconductor.org/packages/release/bioc/html/celaref.html
Github : https://github.com/MonashBioinformaticsPlatform/celaref

- Suggests *cluster* labels on similarity
  - By previously annotated experiment
  - Quantified by p-value.
- Tractably based on differential expression

Celaref package – Input and output

Experiment:
- Counts table (genes per cell)
- Cell clusters

+

Reference Dataset
- Counts table (genes per cell)
- ***Annotated*** Cell clusters

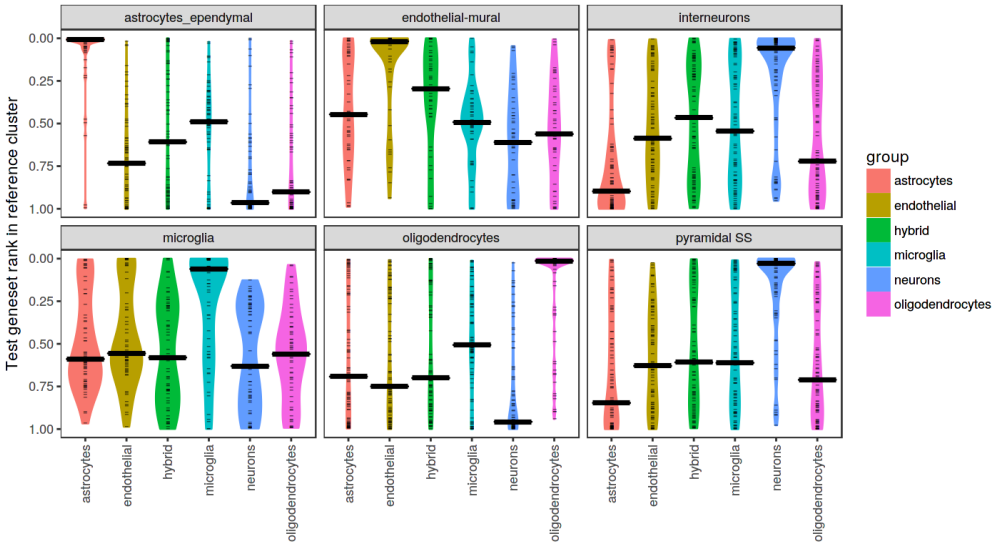
```
# Differential expression
de_table.ref   <- contrast_each_group_to_the_rest(toy_ref_se, dataset_name="ref")
de_table.query <- contrast_each_group_to_the_rest(toy_query_se, dataset_name="query")

# Plot
make_ranking_violin_plot(de_table.test=de_table.query, de_table.ref=de_table.ref)

# And get group labels
make_ref_similarity_names(de_table.query, de_table.ref)
```

Suggested cluster names:
- C5:pyramidal SS|interneurons, p=3.5^-10
- C3:no_similarity, p=NA
- C4:microglia, p=2.7^-10

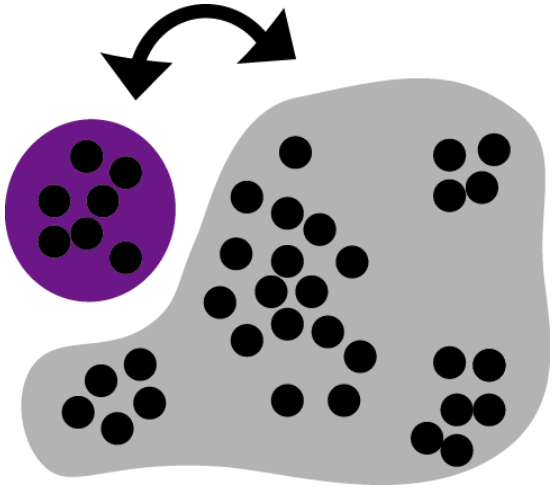
Visualization of similarity:

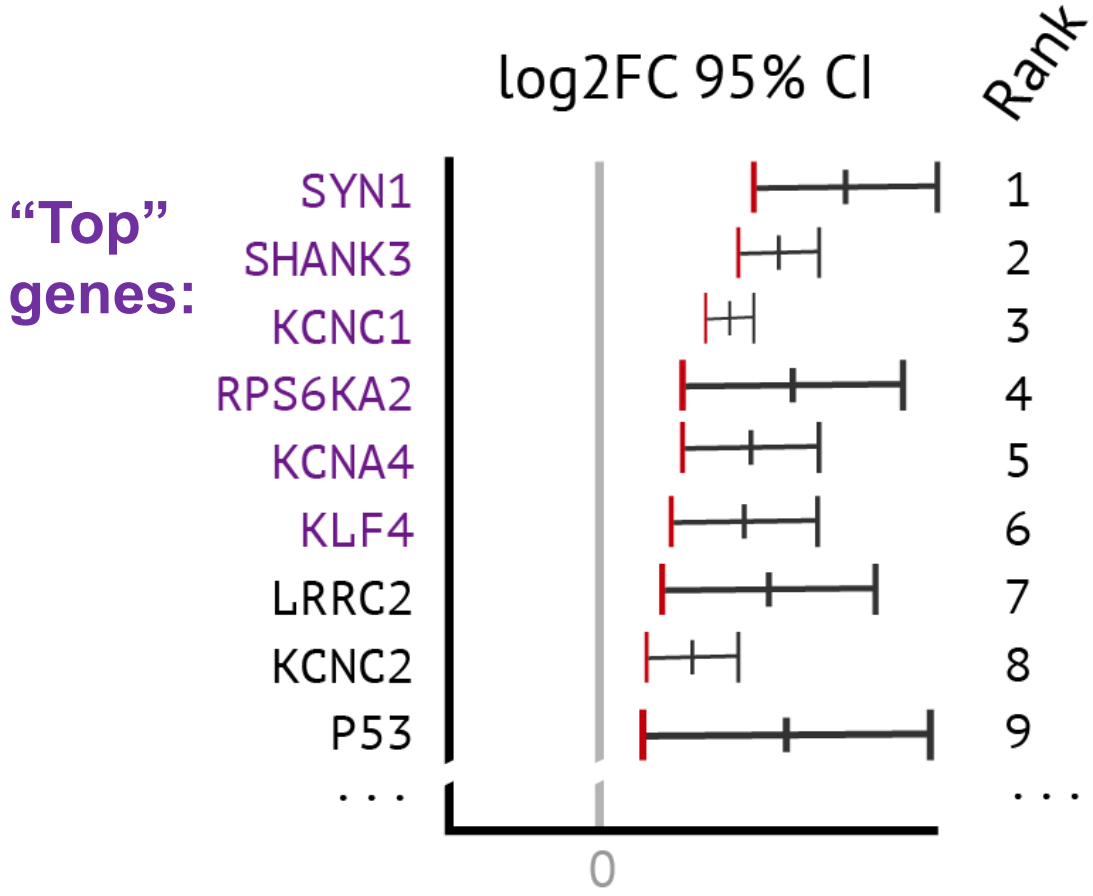


Method – Differential Expression

Get the highly expressed genes for the query cluster

=> are they highly expressed in a reference cell type?





Differential expression with MAST

"Top" genes = significantly, 2-fold **over**represnted , max 100

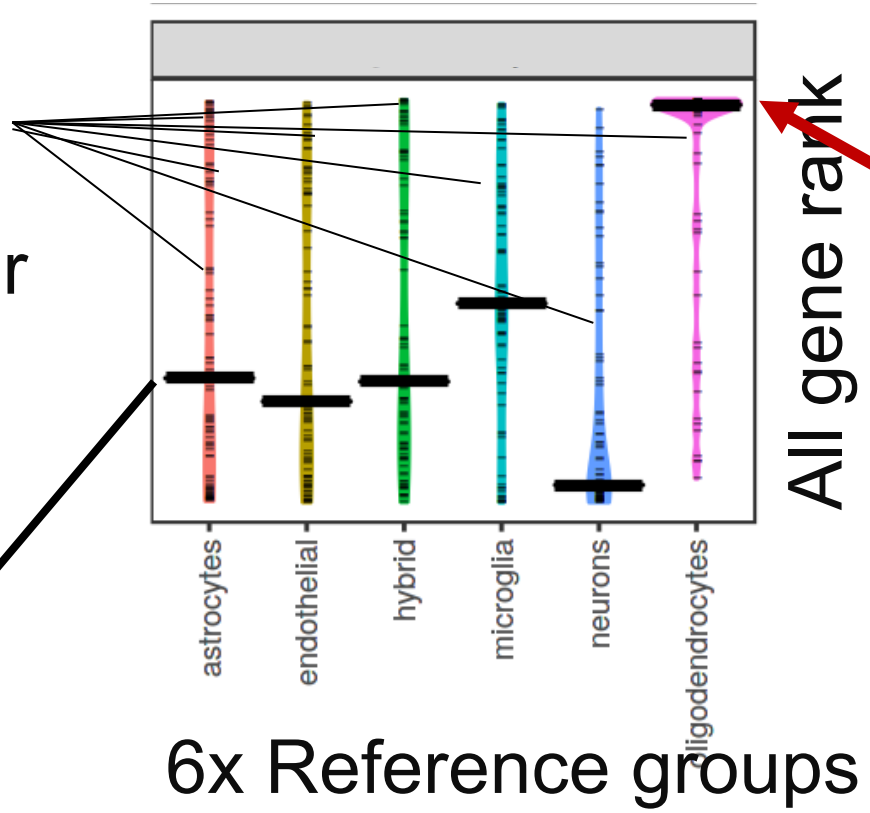
Method - ‘Top’ gene distributions

Get the highly expressed genes for the query cluster

=> are they highly expressed in a reference cell type?

‘Top’ genes from one query cluster

median ‘top’ gene rank

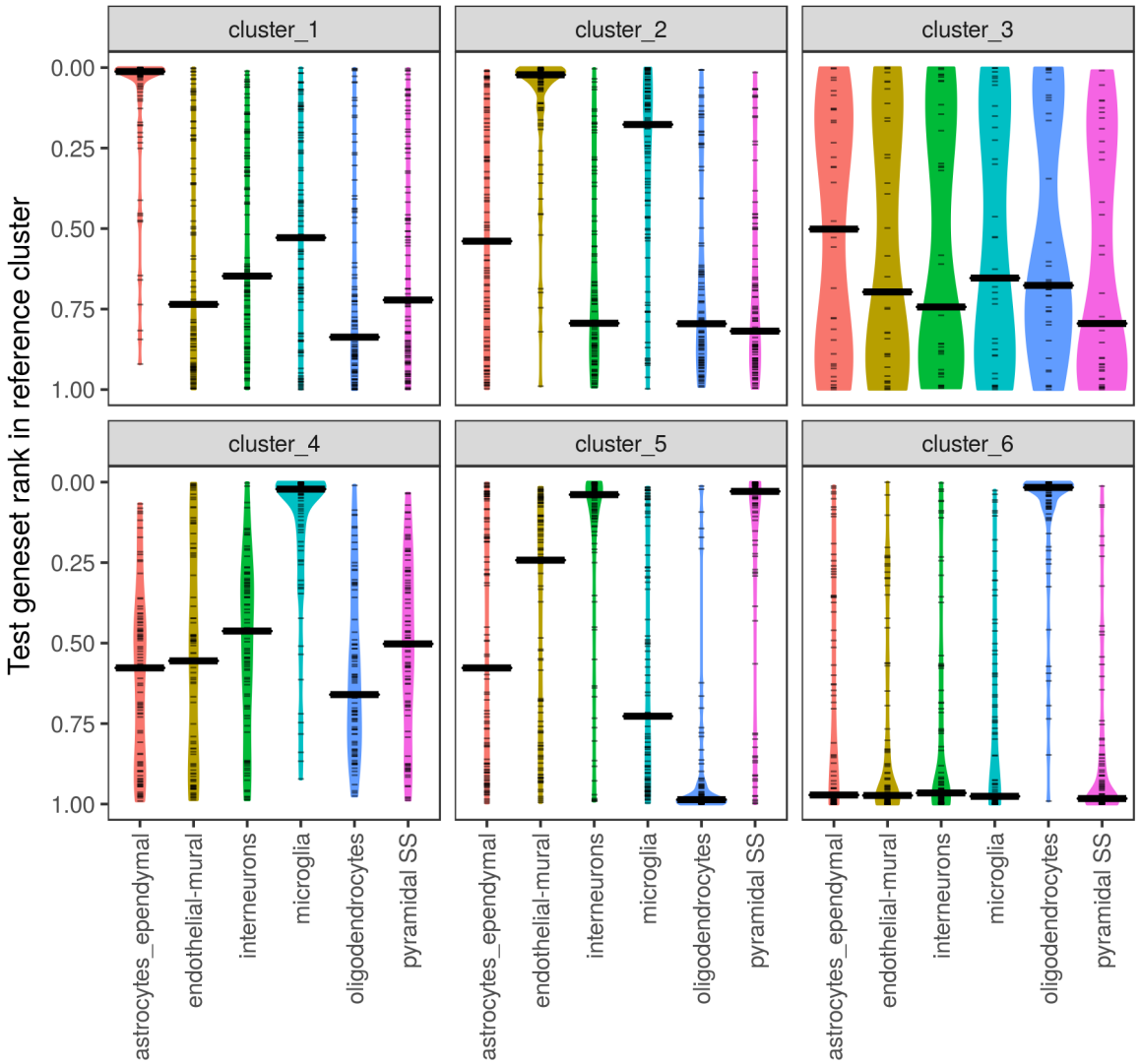


6x Reference groups

Yes – this one!

(They are Oligodendrocytes!)

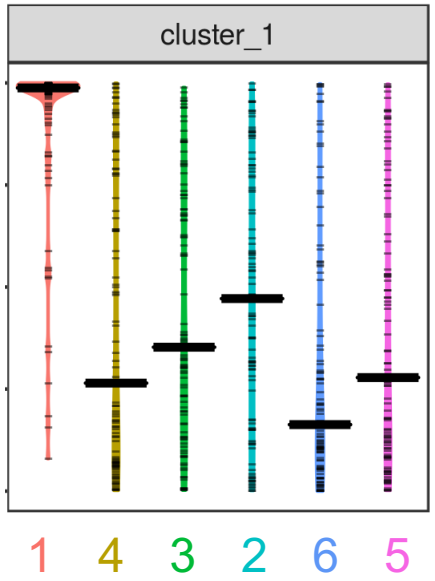
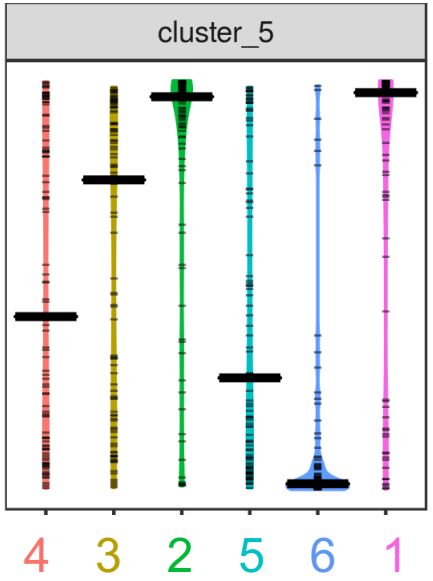
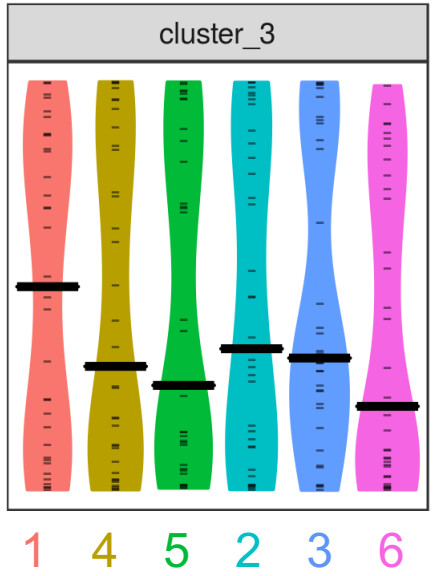
Example – Clusters 1-6 vs a brain reference



Darmarnis 2014 Human brain
vs
Zeisel 2015 Mouse brain

Query Group	Short Label	pval
cluster_1	cluster_1:astrocytes_ependymal	2.98e-23
cluster_2	cluster_2:endothelial-mural	8.44e-10
cluster_3	cluster_3:no_similarity	NA
cluster_4	cluster_4:microglia	2.71e-19
cluster_5	cluster_5:pyramidal SS|interneurons	3.49e-10
cluster_6	cluster_6:oligodendrocytes	2.15e-28

1:1, 1:many or 1:none?

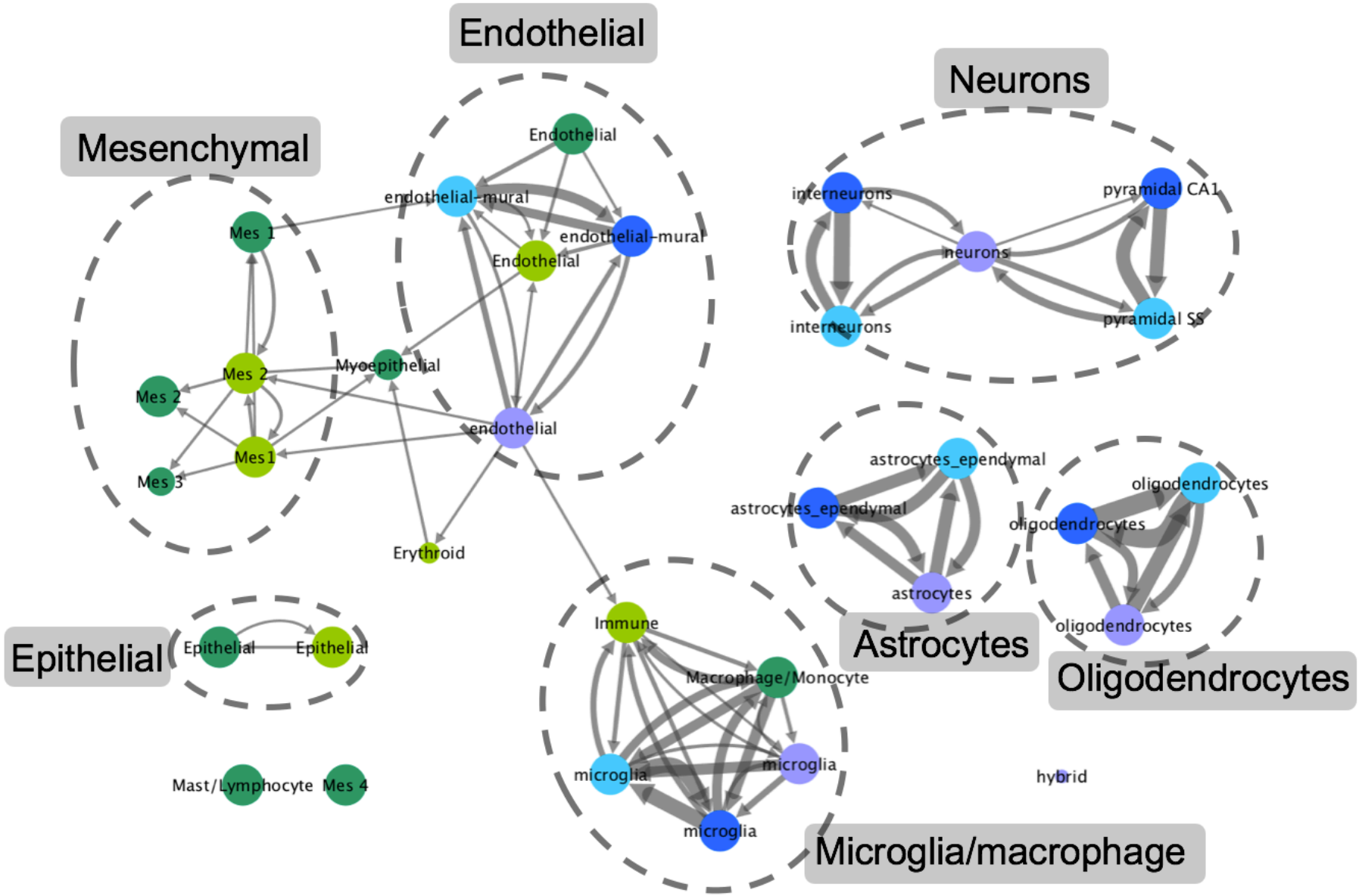
	"astrocytes"	*"neurons"*	*"hybrid"*
'True' cell types : (Darmanis 2014)			
Order by median rank:			
Mann–Whitney U Tests:	* p = 2.98e-23	* p=3.49e-10	
Label:	*astrocytes_ependymal*	*pyramidal SS*|*interneurons*	*no_similarity*

Cross-tissue Comparisons

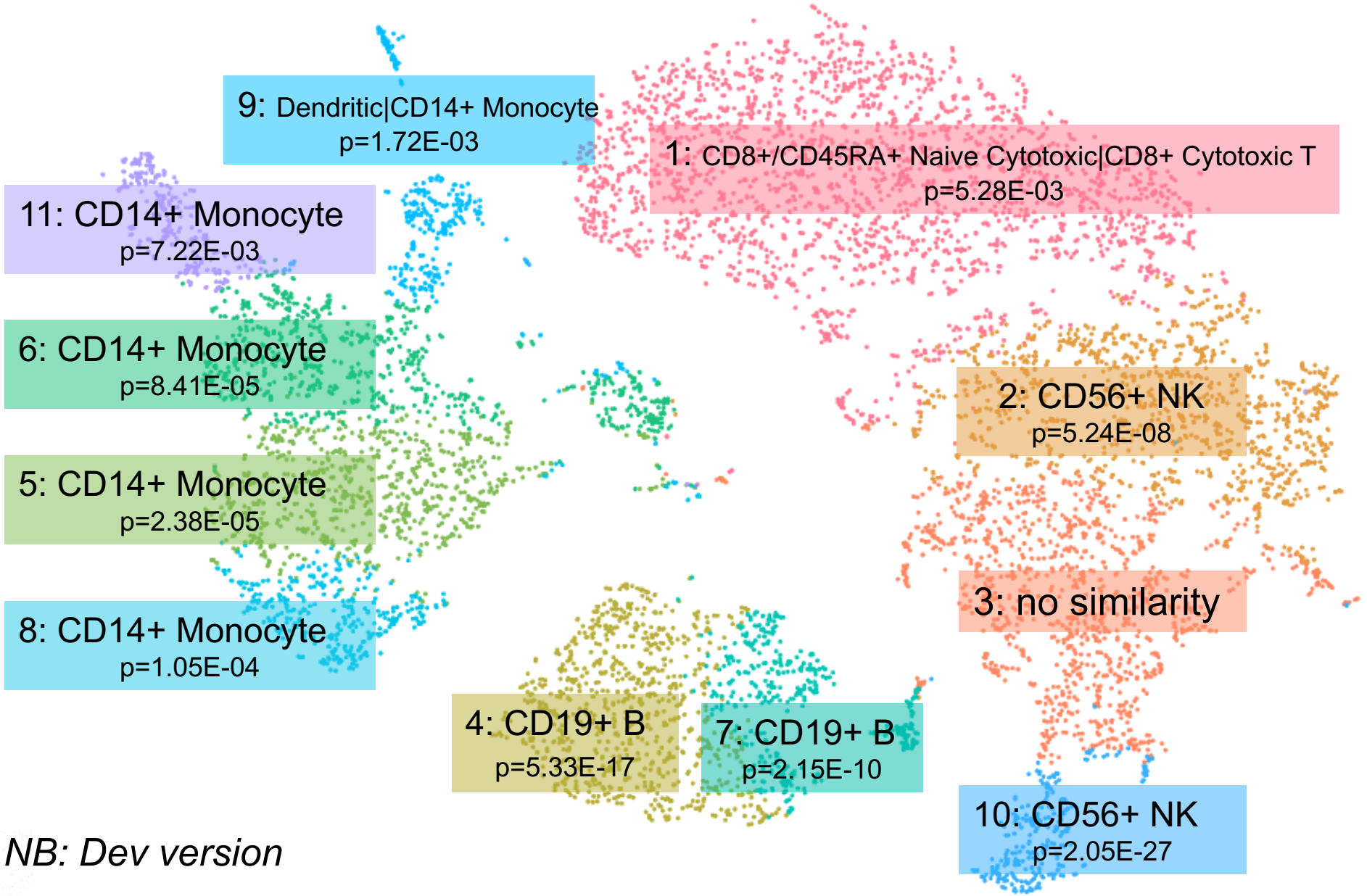
Darmanis 2014:
Human Brain

Zeisel 2015:
Mouse Brain
Cortex
Hippocampus

Farmer 2017:
Mouse
Lacrimal Gland
E16 P4



PBMCs (Blood cells) – Hard



10X PBMC8k (graphclust)

Vs

Purified PBMCs FACs sorted. Zheng et al (2017)

NB: Dev version

Contrast with scmap

Celaref
cluster labels + p-value

(p= 3.49e-10) pyramidal SS|interneurons

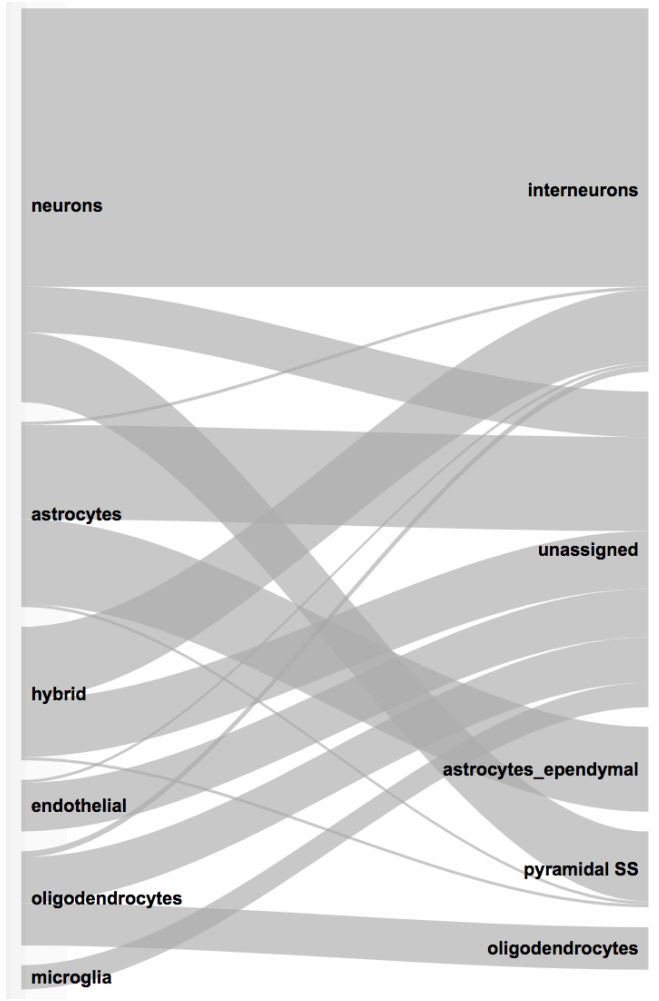
(p=2.98e-23) astrocytes_ependymal

(p= 8.44e-10) endothelial-mural

(p= 2.15e-28) oligodendrocytes

(p= 2.71e-19) Microglia

Scmap
Per-cell assignments



Darmarnis 2014
Human brain
vs
Zeisel 2015
Mouse brain

Conclusions

Celaref package – match clusters to reference data.

- Similar or different tissues, technologies
- P-values indicate confidence
- Tractable labels from differential expression
- Performance scales by number of datasets

Applications:

- Use previous or public experiments to label your new one
- Evaluation of different clustering on an experiment

V1.0 now in bioconductor 3.8 (*with vignette!)*

Bioconductor : https://bioconductor.org/packages/release/bioc/html/celaref.html

Github (dev version) : https://github.com/MonashBioinformaticsPlatform/celaref

References

Scmap: Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, *15*(5), 359–362. http://doi.org/10.1038/nmeth.4644

MAST: Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., … Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*(1), 1–13. http://doi.org/10.1186/s13059-015-0844-5

Human brain dataset: Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., … Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, *112*(23), 201507125. http://doi.org/10.1073/pnas.1507125112

Mouse brain datasets: Zeisel, A., Manchado, A. B. M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., … Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, *347*(6226), 1138–42. http://doi.org/10.1126/science.aaa1934

Mouse lacrimal gland datasets: Farmer, D. T., Nathan, S., Finley, J. K., Shengyang Yu, K., Emmerson, E., Byrnes, L. E., … Knox, S. M. (2017). Defining epithelial cell dynamics and lineage relationships in the developing lacrimal gland. *Development*, *144*(13), 2517–2528. http://doi.org/10.1242/dev.150789

Purified PBMC data: Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 1–12. http://doi.org/10.1038/ncomms14049

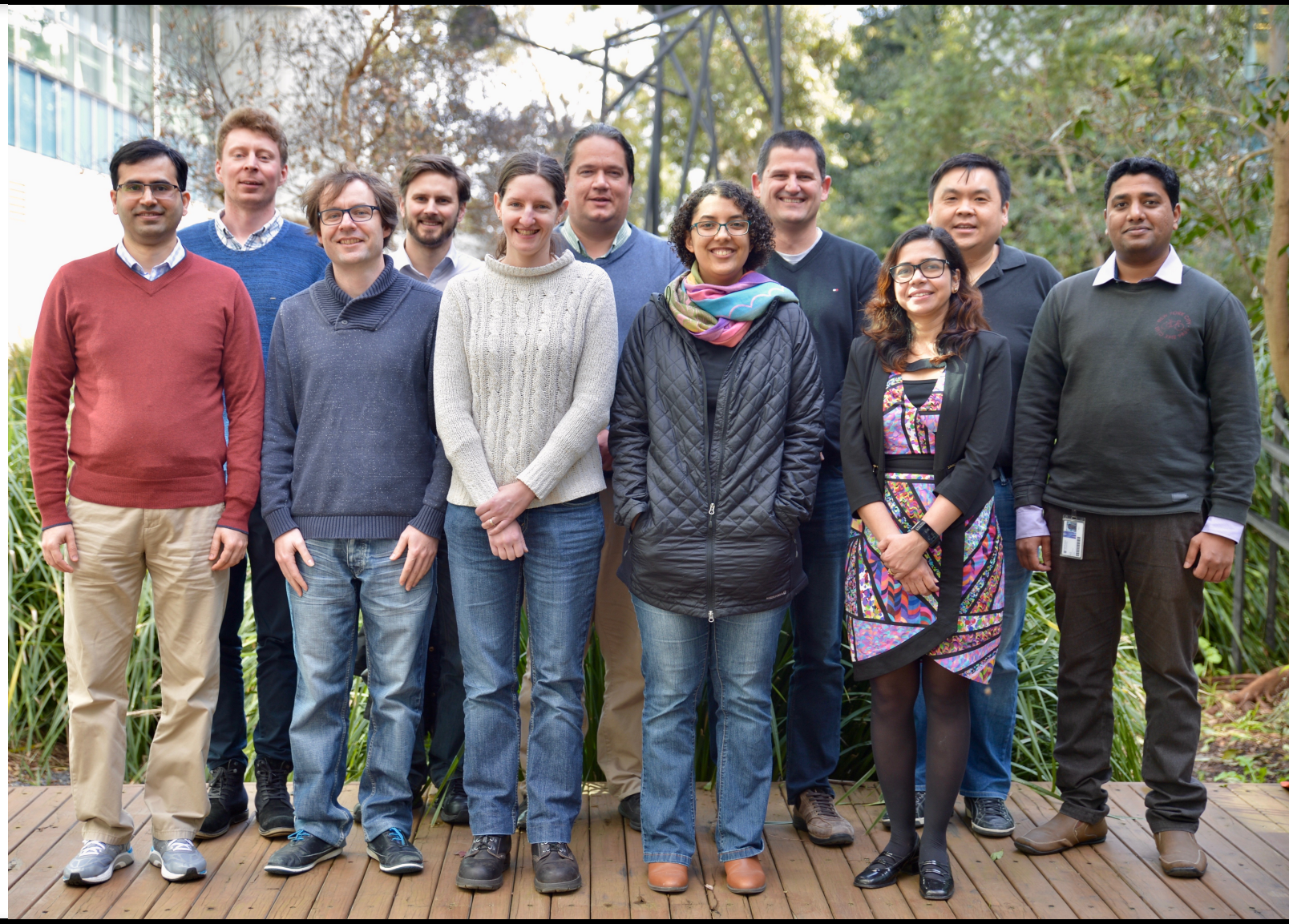
Thanks

Monash Bioinformatics Platform:

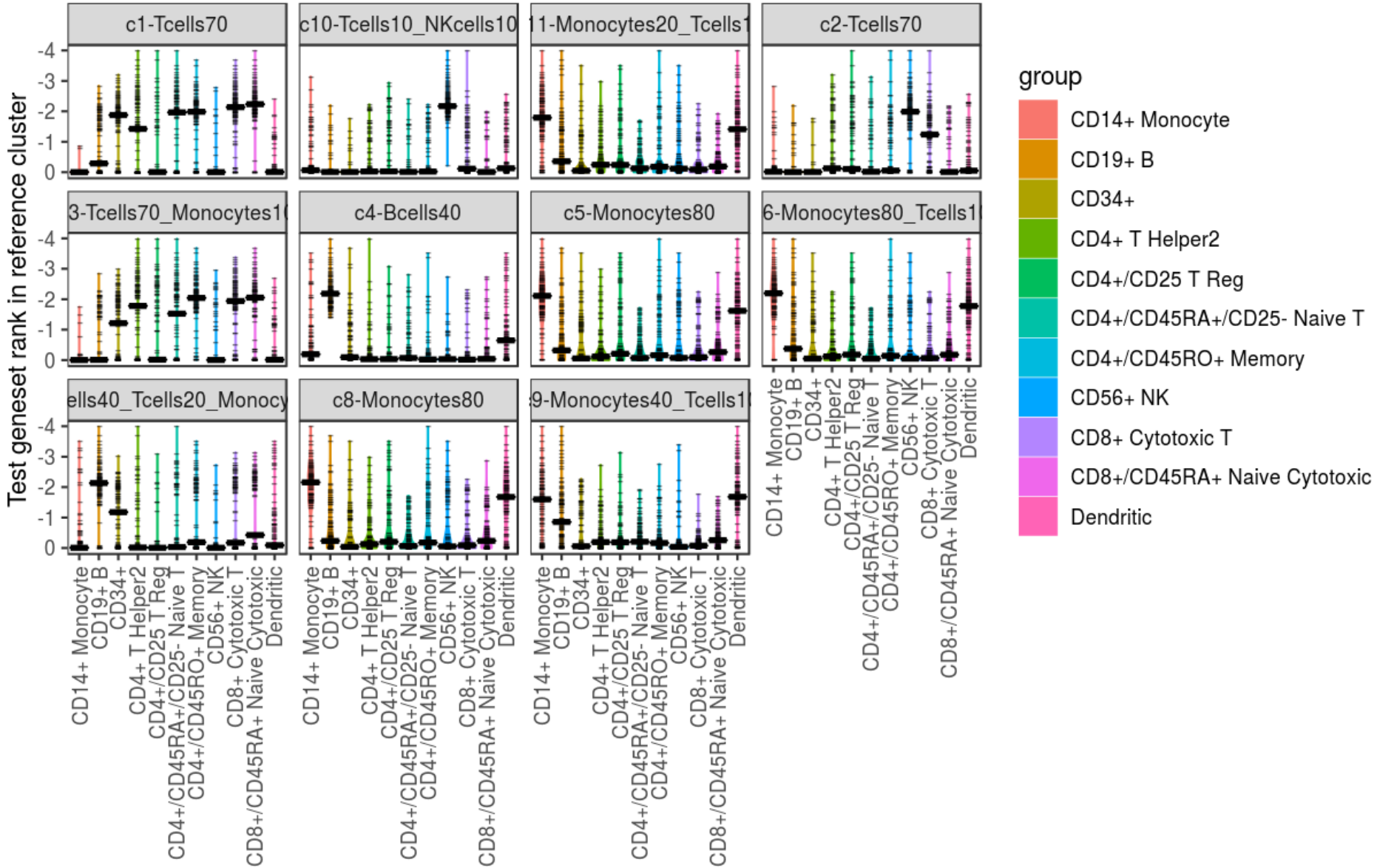
- **David Powell**
- Stuart Archer
- Adele Barugahare
- **Paul Harrison**
- Andrew Perry
- Michael See
- Anup Shah
- Kirill Tsyganov
- Sarah Williams
- Nick Wong

Monash Bioinformatics Platform/
Monash School of Biological sciences

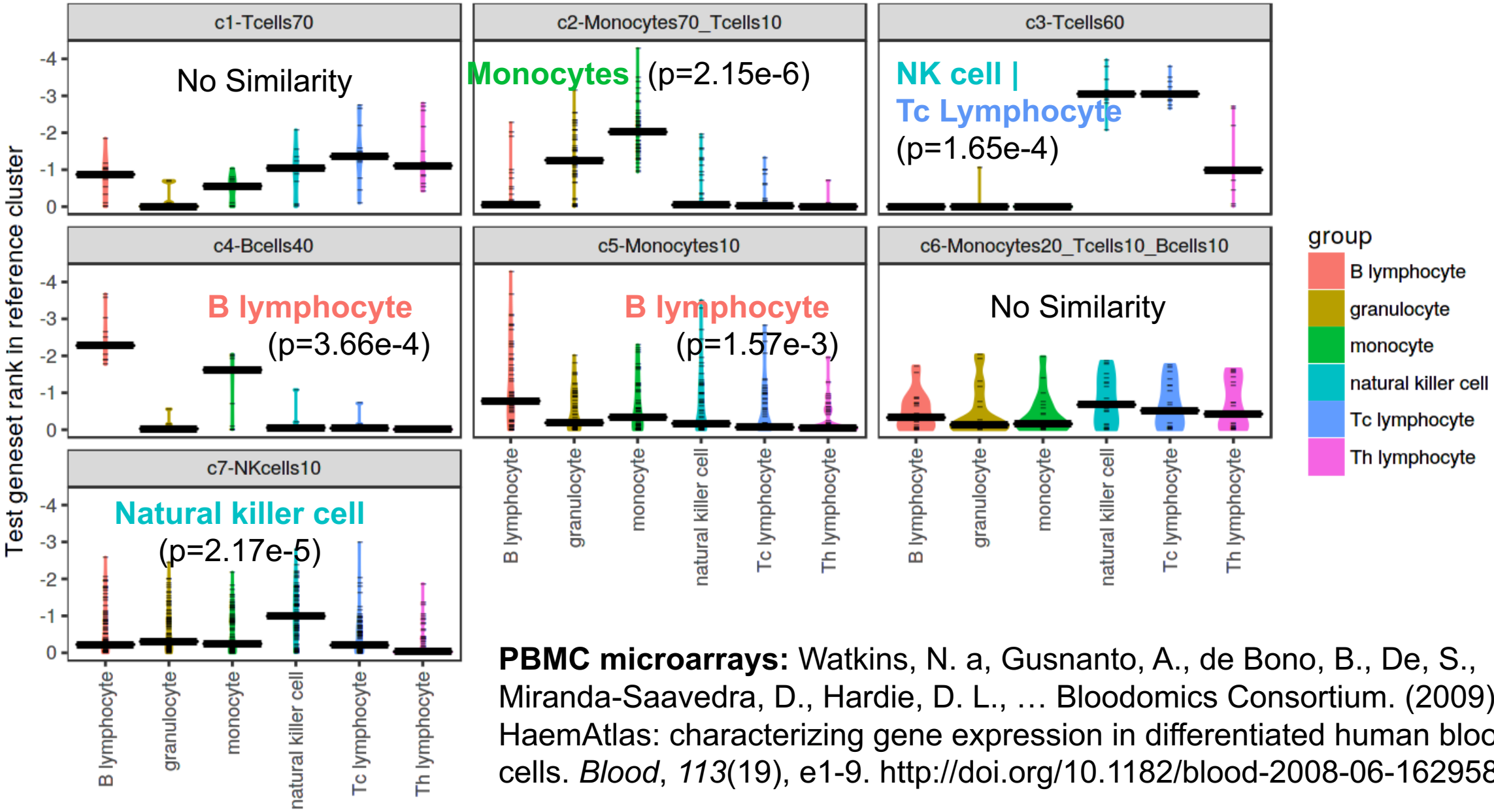
- **Sonika Tyagi**



PBMCs (Blood cells) – 10X example data

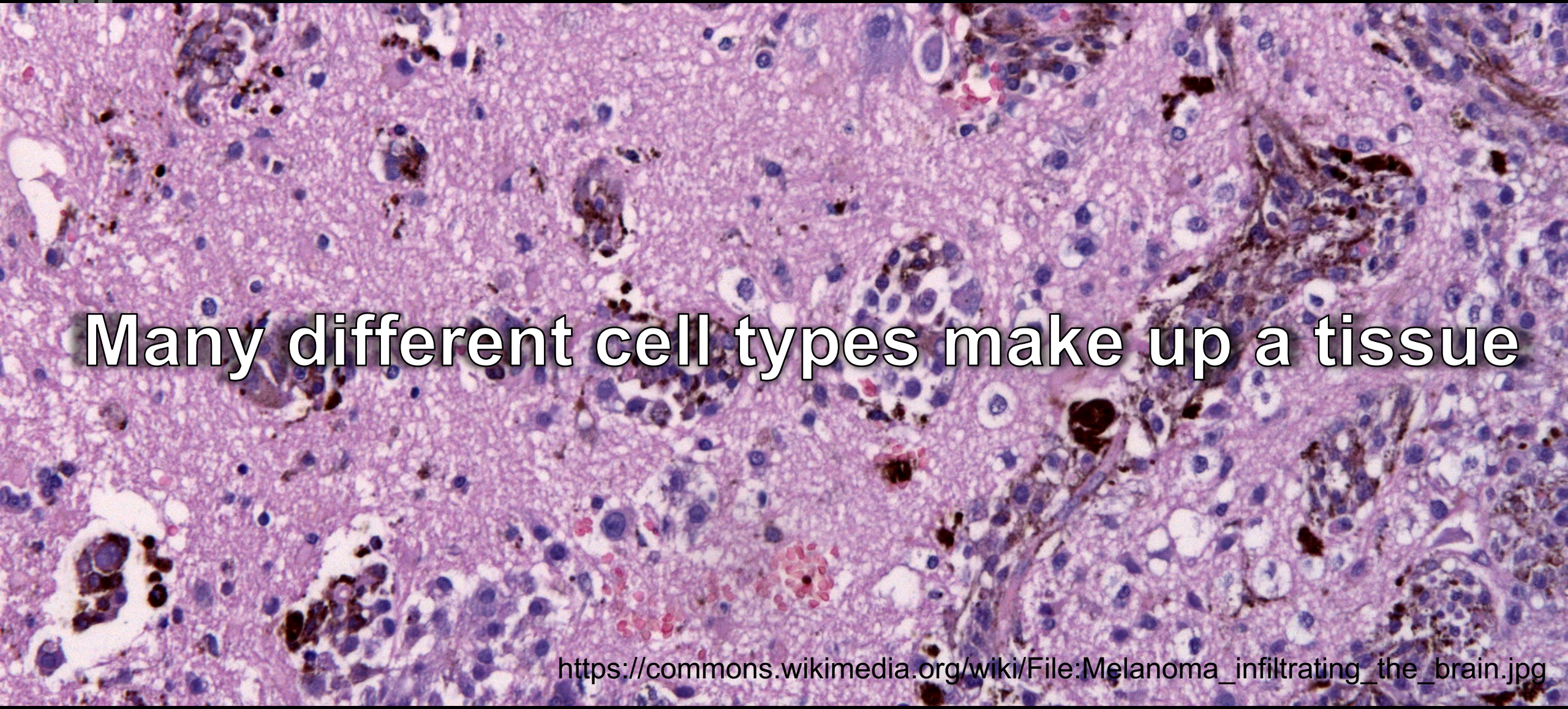


PBMCs (Blood cells) – 10X example data



PBMC microarrays: Watkins, N. a, Gusnanto, A., de Bono, B., De, S., Miranda-Saavedra, D., Hardie, D. L., … Bloodomics Consortium. (2009). A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, *113*(19), e1-9. http://doi.org/10.1182/blood-2008-06-162958

Single Cell sequencing



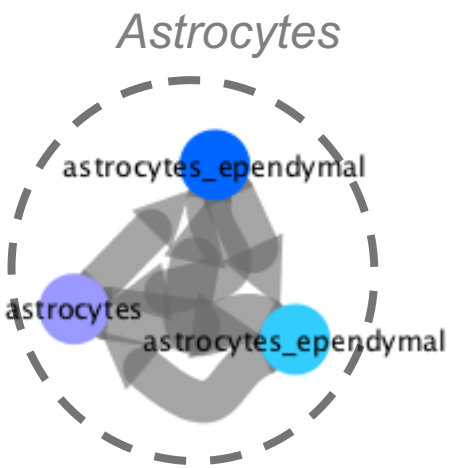
https://commons.wikimedia.org/wiki/File:Melanoma_infiltrating_the_brain.jpg

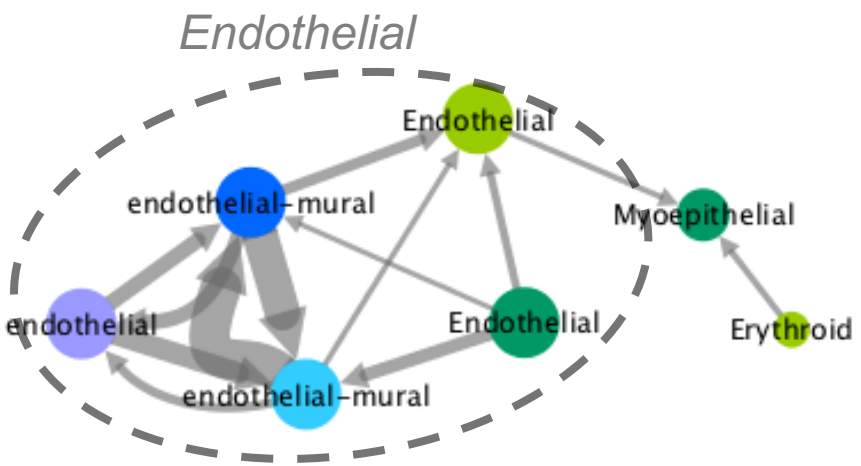
Cross-tissue Comparisons

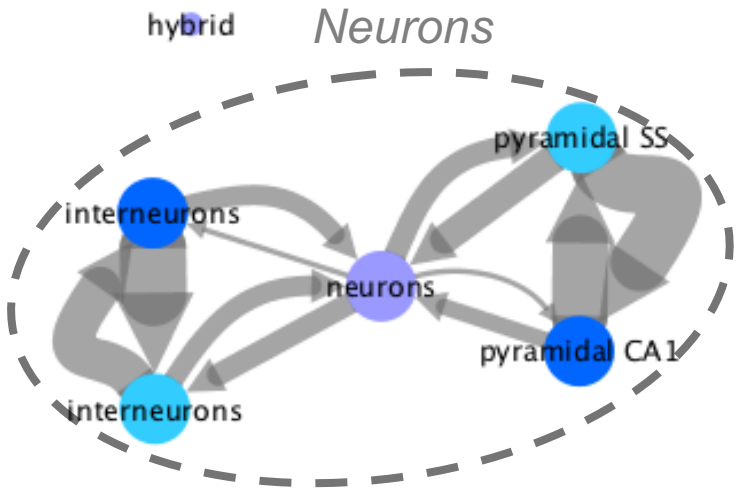
Darmanis 2014:
Human Brain

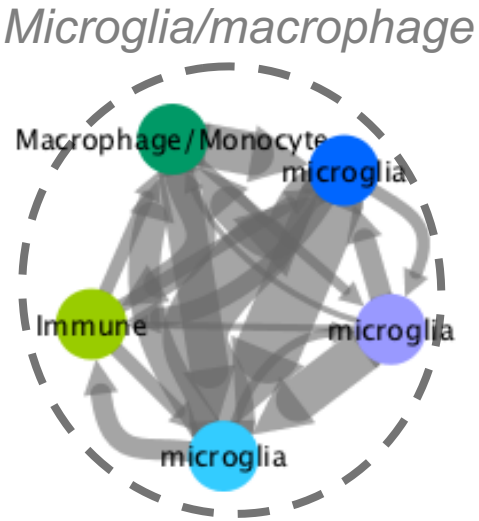
Zeisel 2015:
Mouse Brain
Cortex
Hippocampus

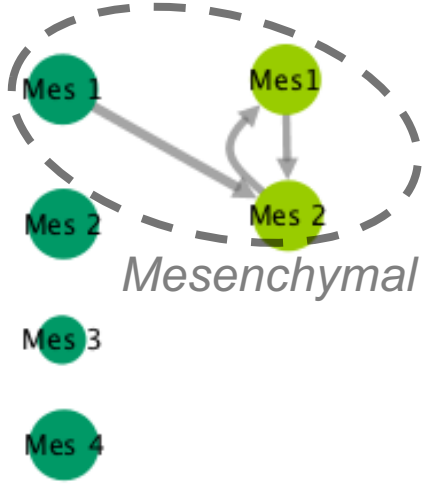
Farmer 2017:
Mouse
Lacrimal Gland
E16 **P4**

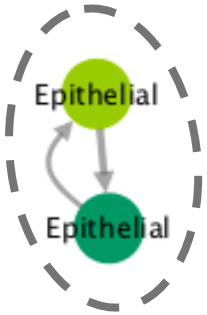












Mast/Lymphocyte