A cluster of blue and white cells against a dark blue background.

Analysis of a single-cell RNA-seq dataset

Ines Guix
December 2021

GENERAL INFORMATION

- N = **28,700** cells from **4** murine mammary tumors.
- Information on the expression of > 20,000 genes per cell.
- Single cells were dissociated from mouse mammary gland tumors, immuno-stained and **sorted on the CD45+ marker**.
- In 2 samples, single cells were dissociated with **enzymatic digestion** and in 2 samples with the **SimpleFlowTM System**.
- Single cell RNA-seq was performed with 10xGenomics.



THE GOALS



Prepare a pipeline to analyze single-cell RNA-seq data for current and future use in our lab



Test whether there are differences between samples dissociated with distinct methodologies
(enzymatic digestion versus SimpleFlow)

INDEX

1. Quality control & cell filtering
2. Normalization & dimensionality reduction
3. Clustering & artifact control
4. Marker selection
5. Cell type annotation
6. Coloring by genes
7. Comparison of the dissociation methods
8. The take aways



01

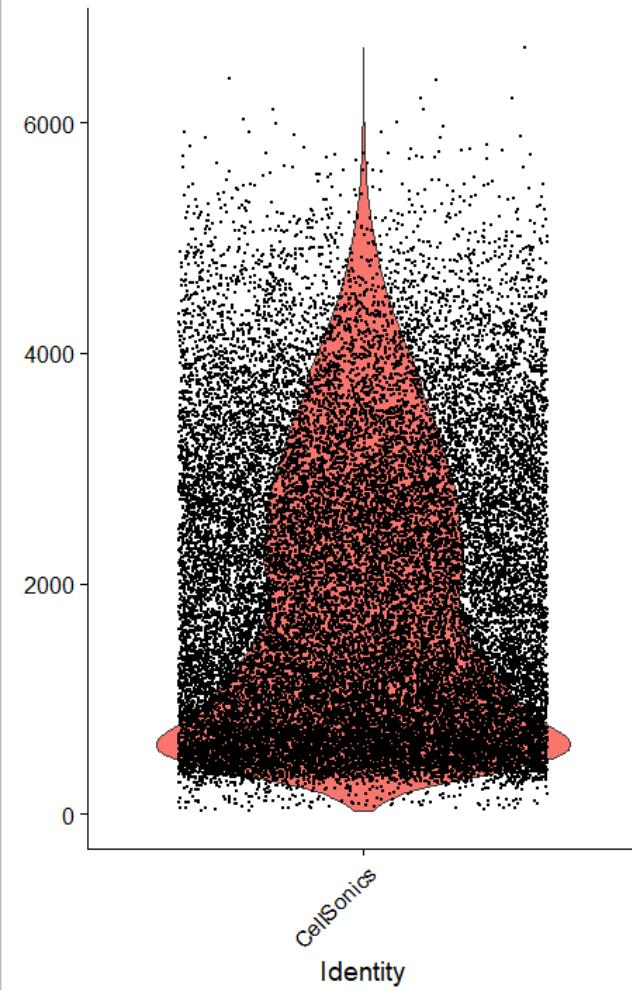
Quality control & cell filtering



1. Import to R the output of 10X cell ranger (filtered_feature_bc_matrix)
2. Create a **Seurat** object
3. Calculate quality control metrics:
 - nCount RNA (UMI reads/cell)
 - nFeature RNA (number of detected genes/cell)
 - % mitochondrial genes
 - % ribosomal genes

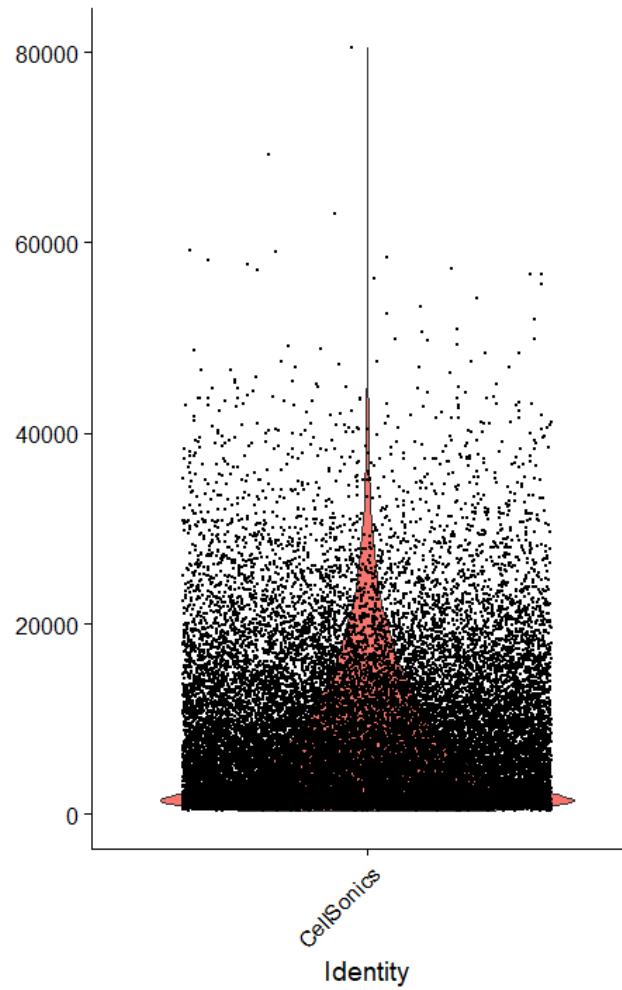
(# detected genes/cell)

nFeature_RNA

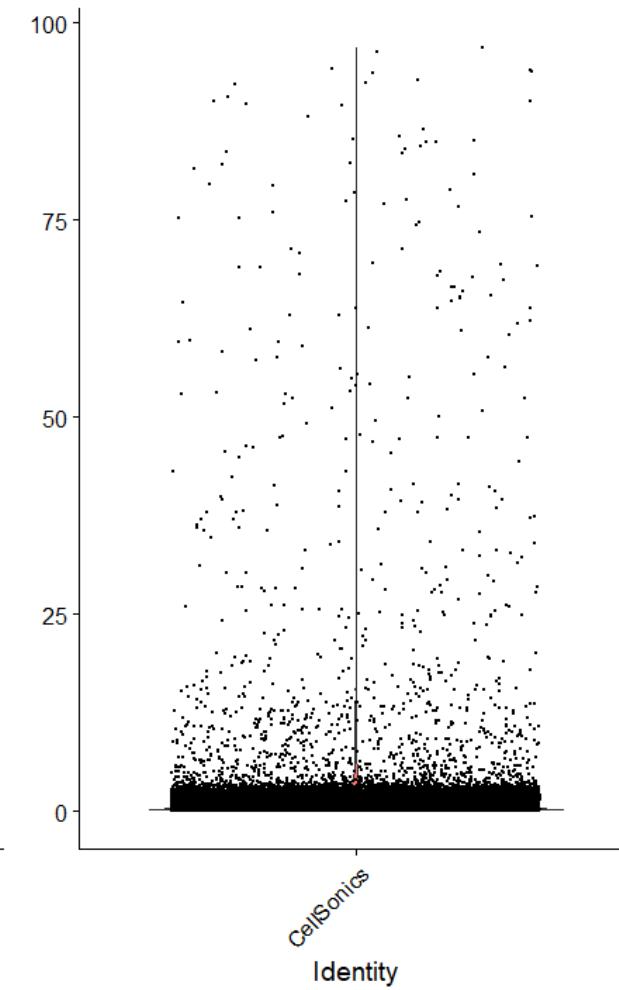


(UMI reads/cell)

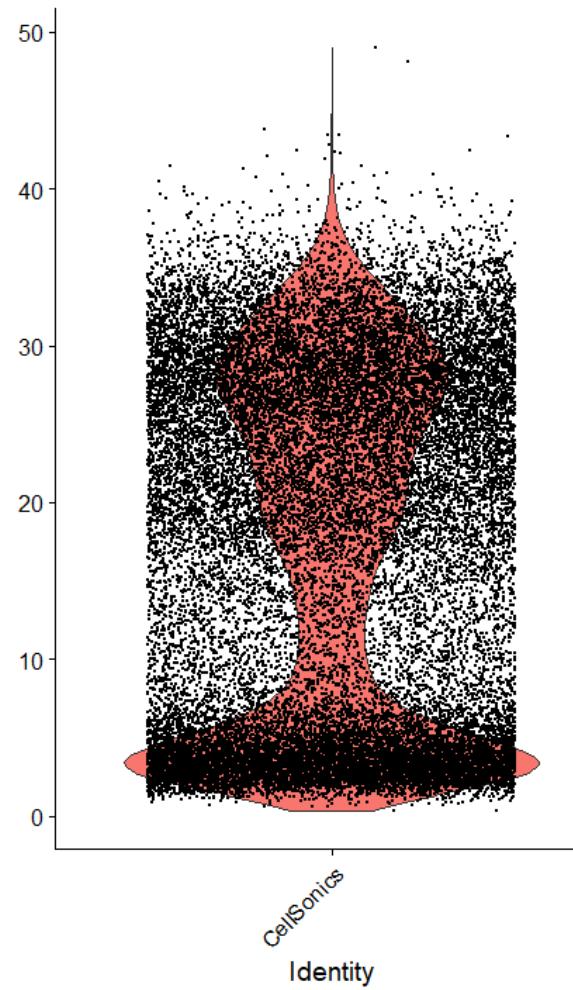
nCount_RNA



percent.mt



percent.rb



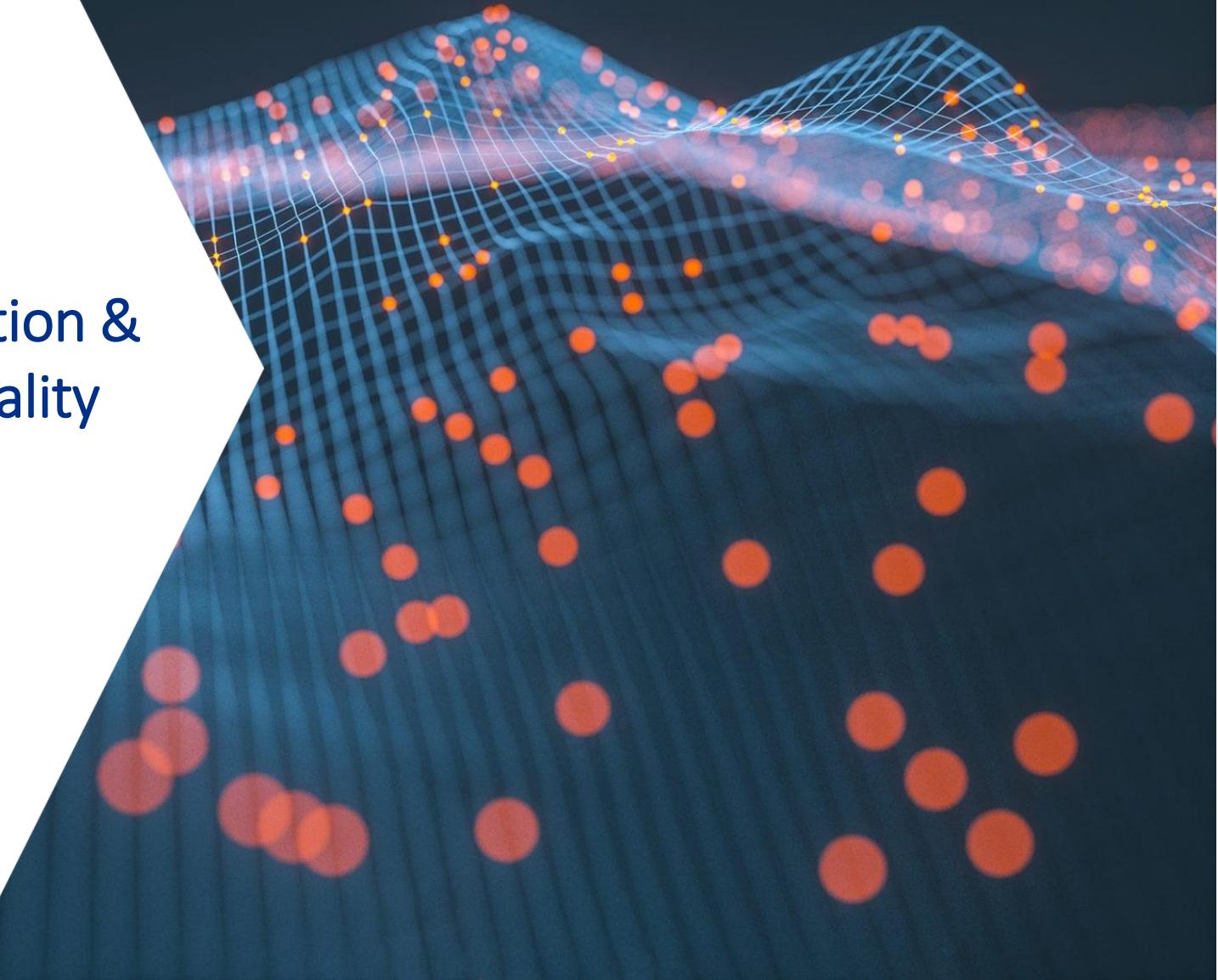
4. Based on that, filter cells to remove dead cells / empty droplets / doublets

```
> table(srat[['QC']]) #Most cells (26.0007) pass the quality control.
```

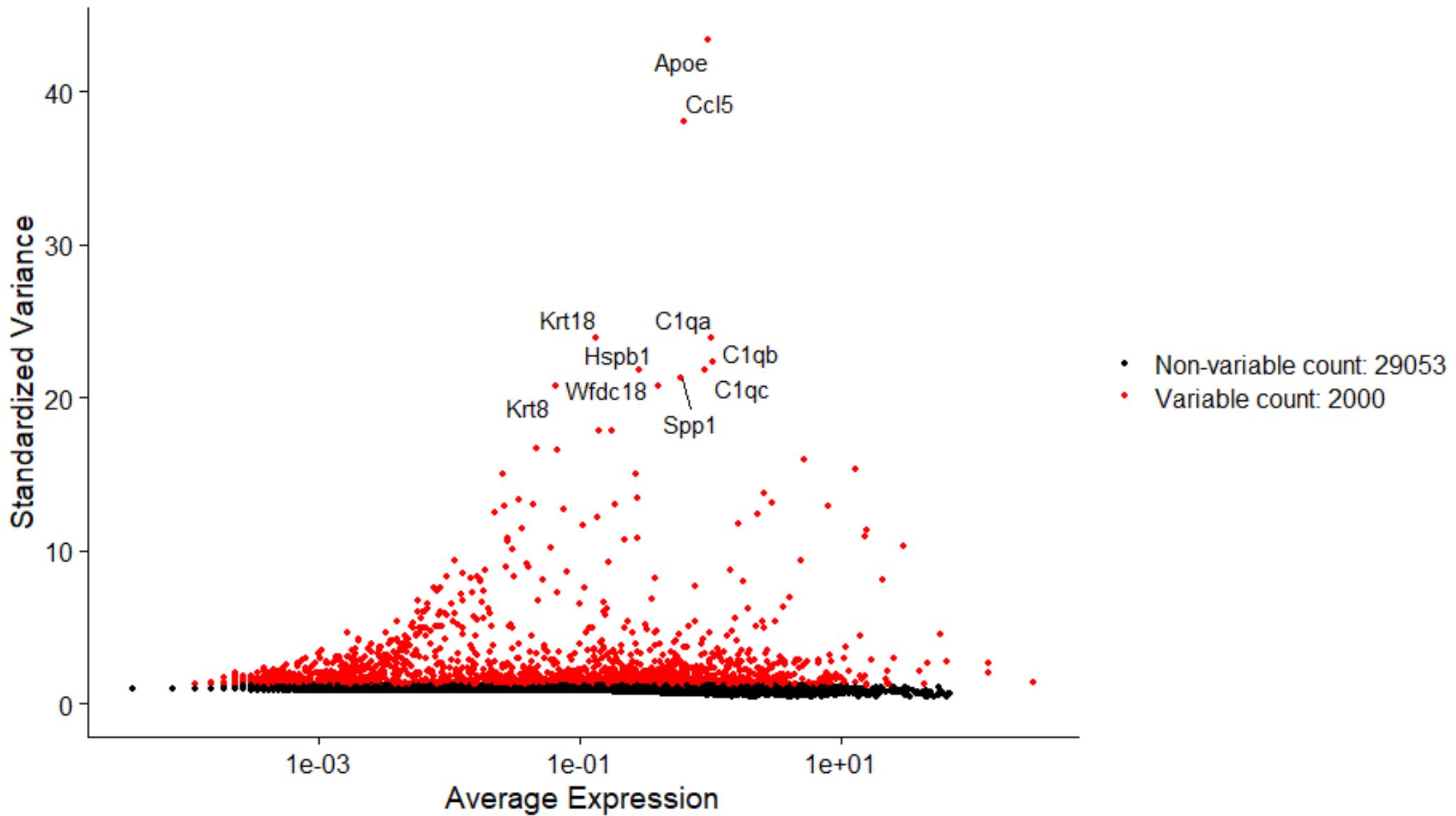
High_MT	High_MT,Low_nFeature	Pass
78	2615	26007

02

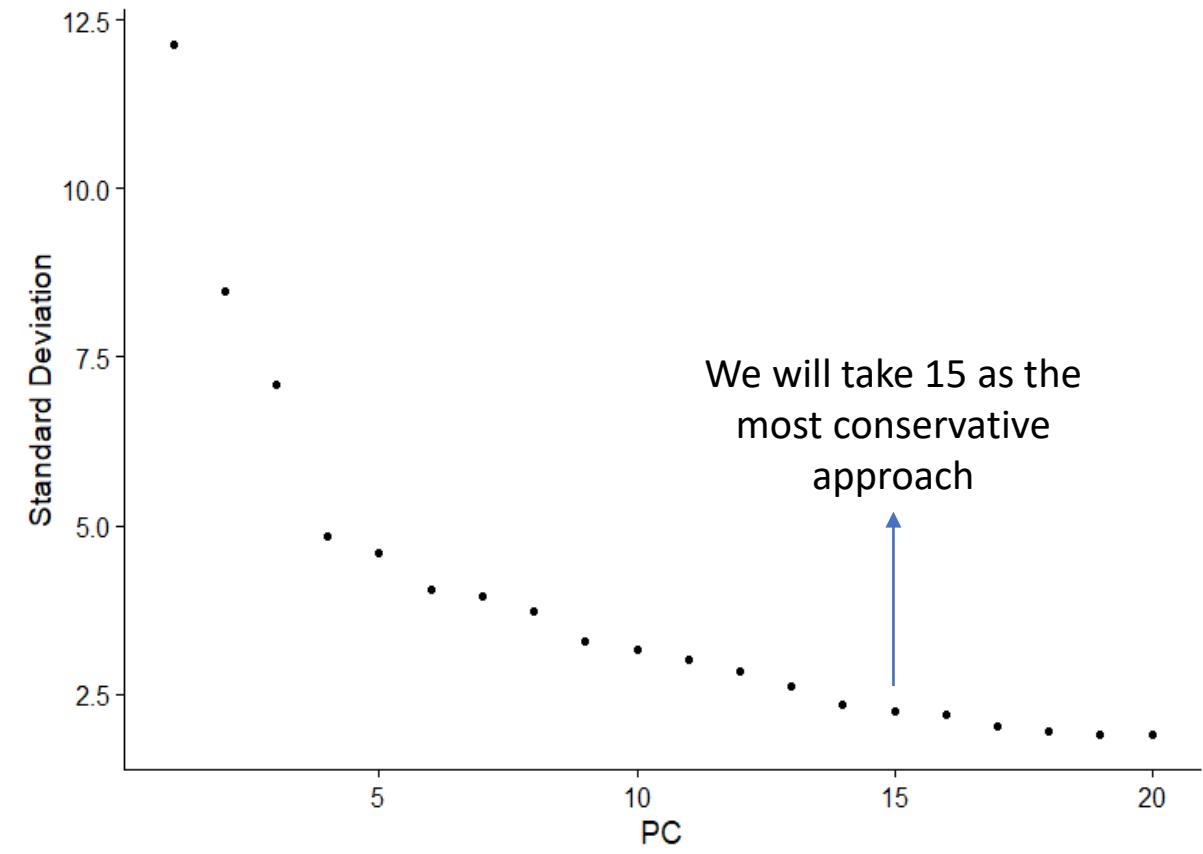
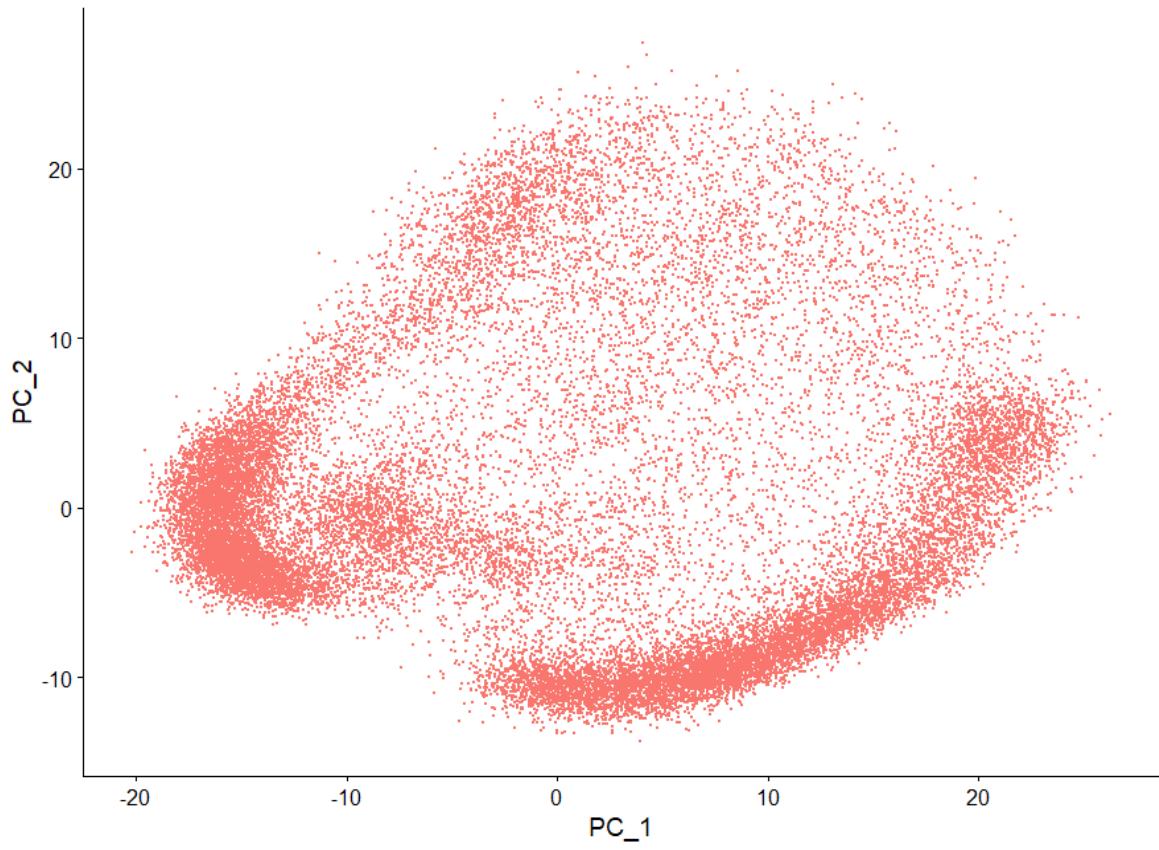
Normalization &
dimensionality
reduction



5. Normalize the data to account for sequencing depth using the `NormalizeData()` R function
 - Method: Feature counts for each cell are divided by the total counts for that cell and multiplied by the `scale.factor` (10.000 default). This is then log transformed.
6. Identify the 2.000 genes with the highest variance between samples using the `FindVariableFeatures()` R function



7. Scale the data by Z-score transforming the gene expression
8. Perform a Principal Component Analysis (PCA) dimensionality reduction using the 2.000 genes with the highest variance
9. Identify how many Principal Components (PCs) we should use in downstream analyses by ranking the % of variance explained by each of them using the ElbowPlot() R function

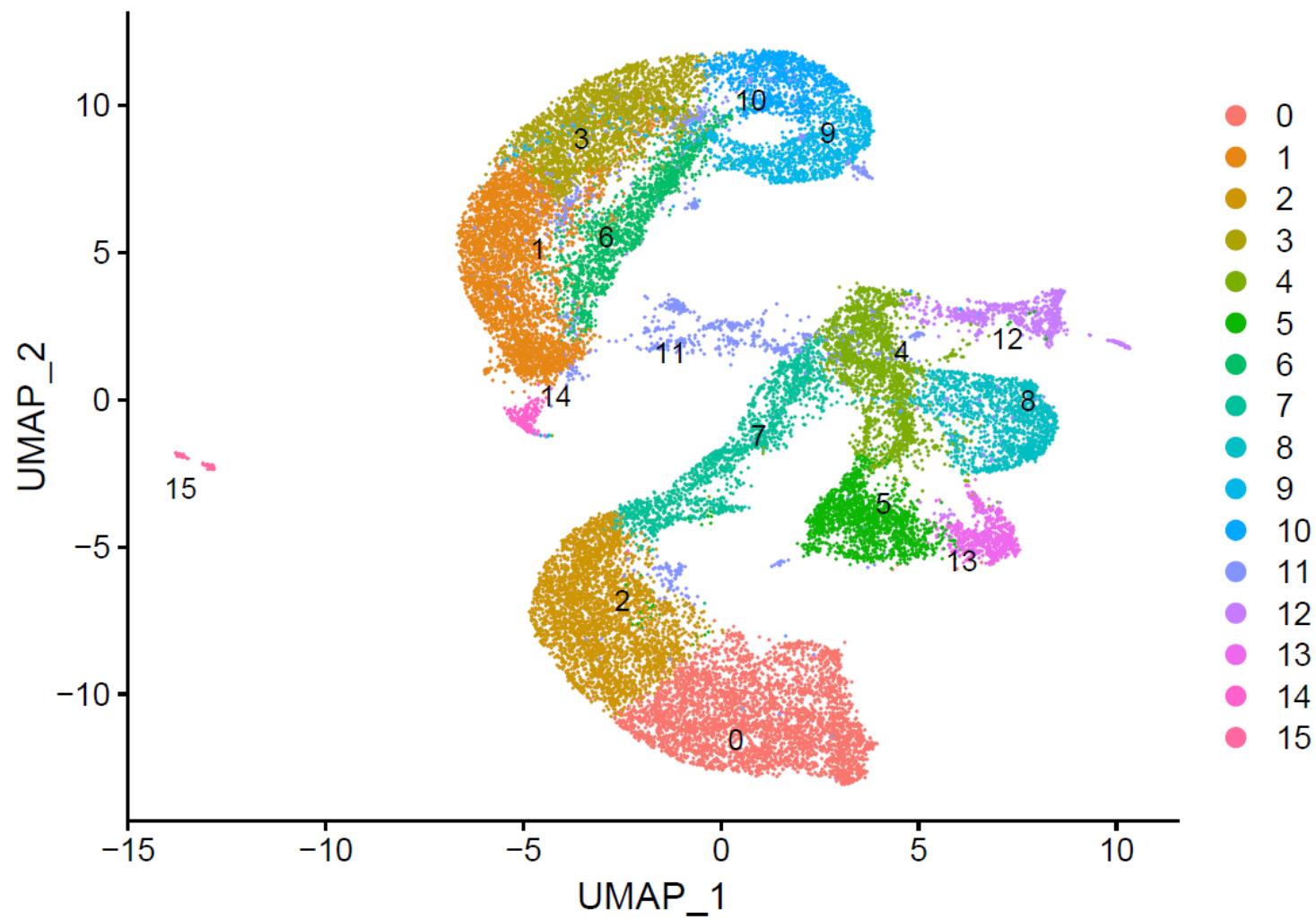


03

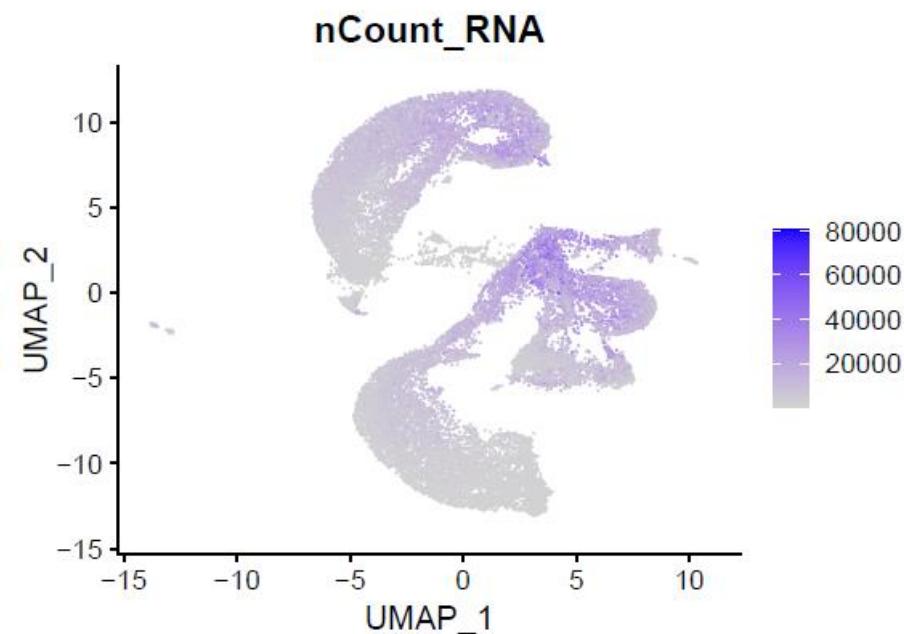
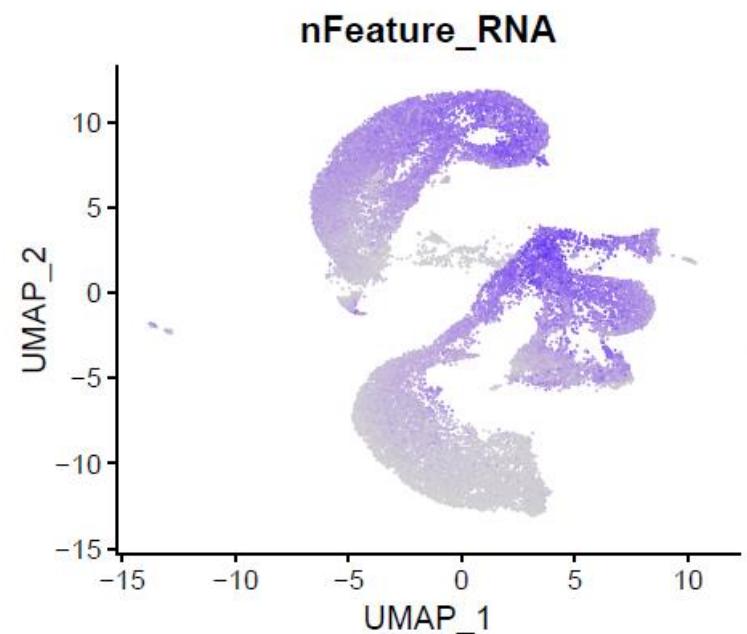
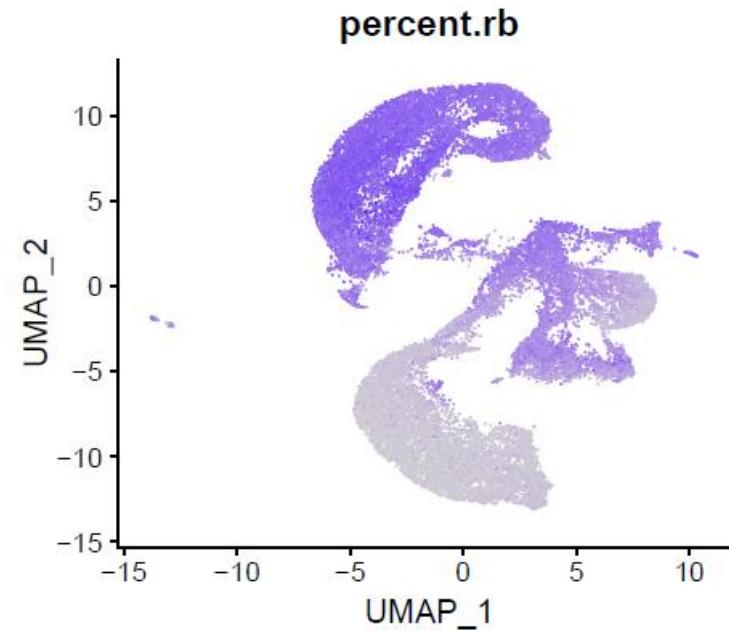
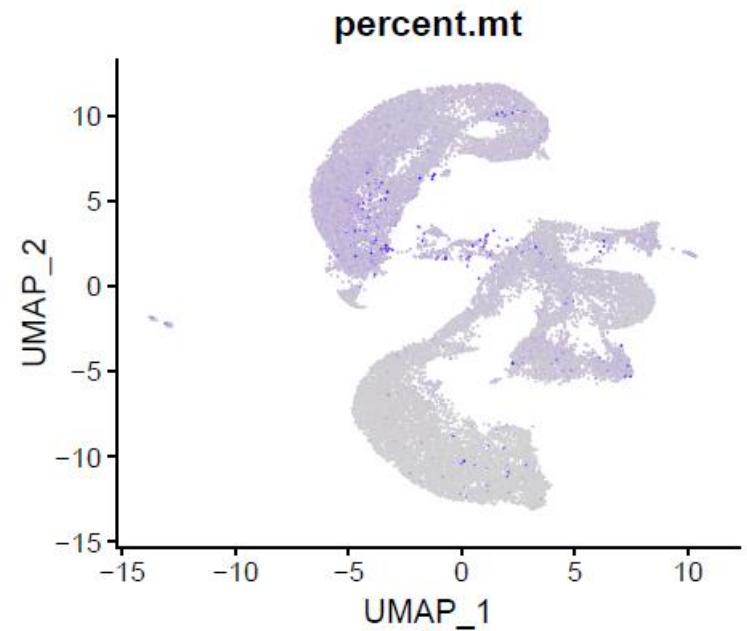
Clustering & artifact control



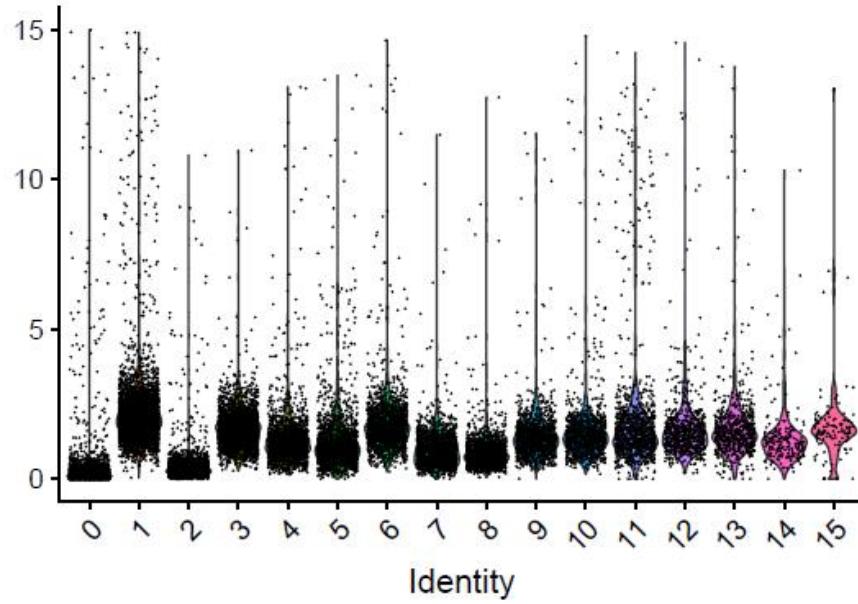
10. Using these number of PCs, perform clustering of the cells using the K-Nearest Neighbors (KNN) algorithm based on the euclidean distance in PCA space with the FindNeighbors() and FindClusters() R functions
11. For visualization purposes, generate UMAP reduced dimensionality representation of the results



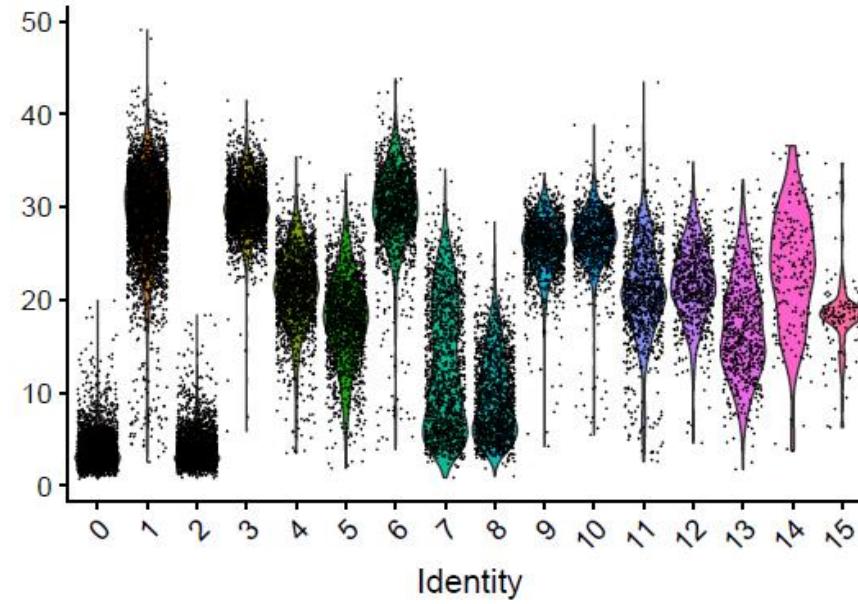
12. In order to control for possible artifacts, compare the distribution within clusters of the quality control (QC) metrics (nCount RNA, nFeature RNA, % mitochondrial genes, % ribosomal genes, +/- cell cycle score)



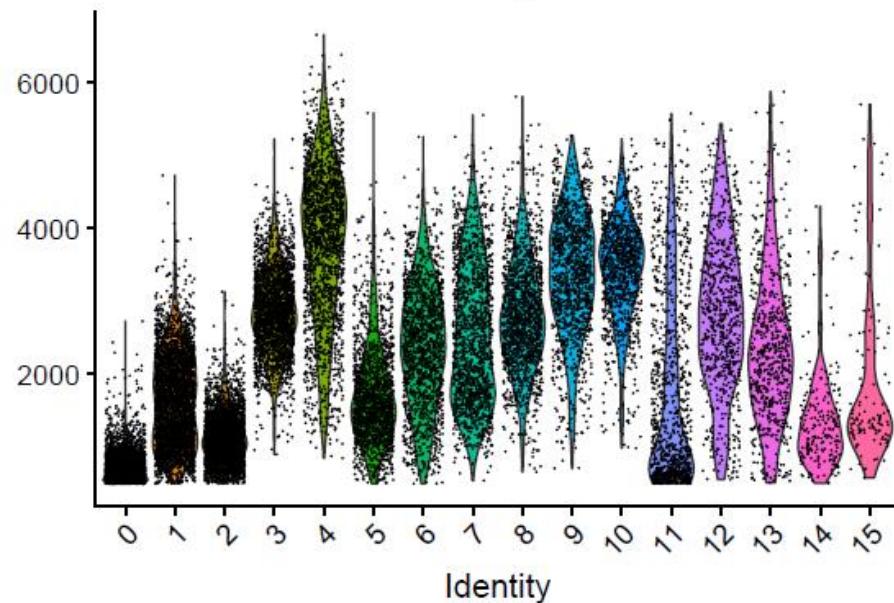
percent.mt



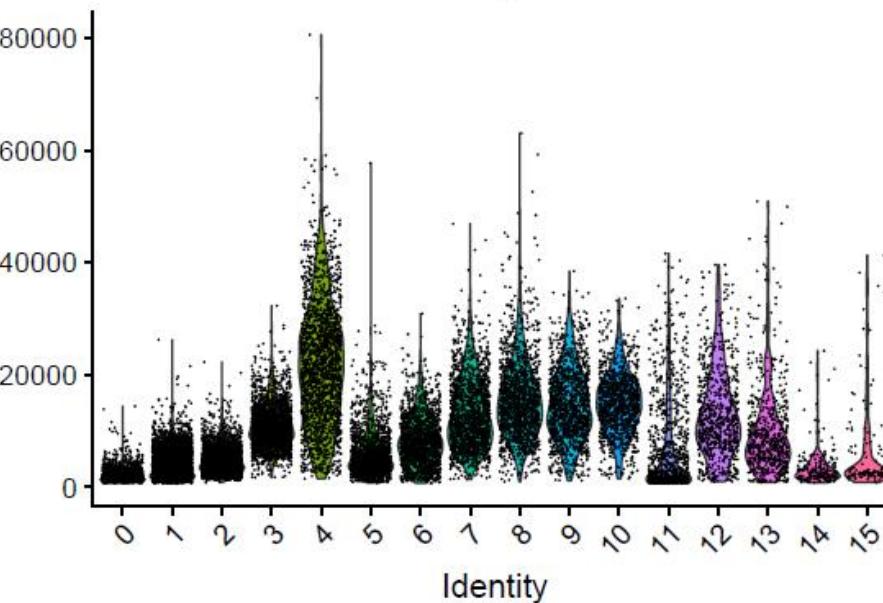
percent.rb



nFeature_RNA



nCount_RNA



13. Correct for the identified confounding factors that change dramatically between clusters using using vars.to.regres from the SCTransform() R function

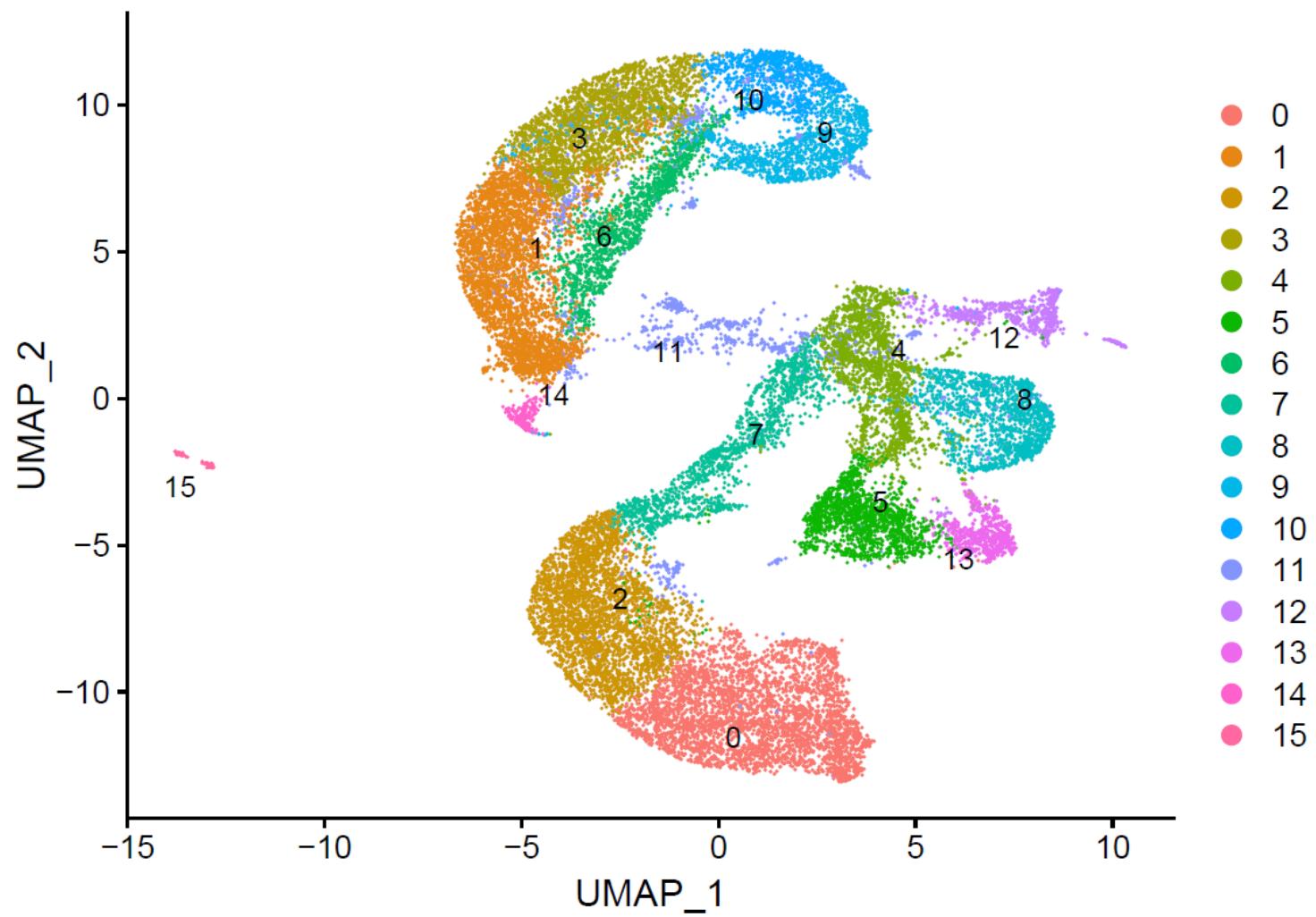
In this case, however, the distribution of none of the metrics seems to be an artifact after consultation with the experts (Ye lab from UCSF), so no correction is needed

04

Marker selection



14. Find markers for every cluster by comparing it to all remaining cells using the Wilcoxon Rank Sum test with the `FindAllMarkers()` R function



Top 5 markers of each cluster

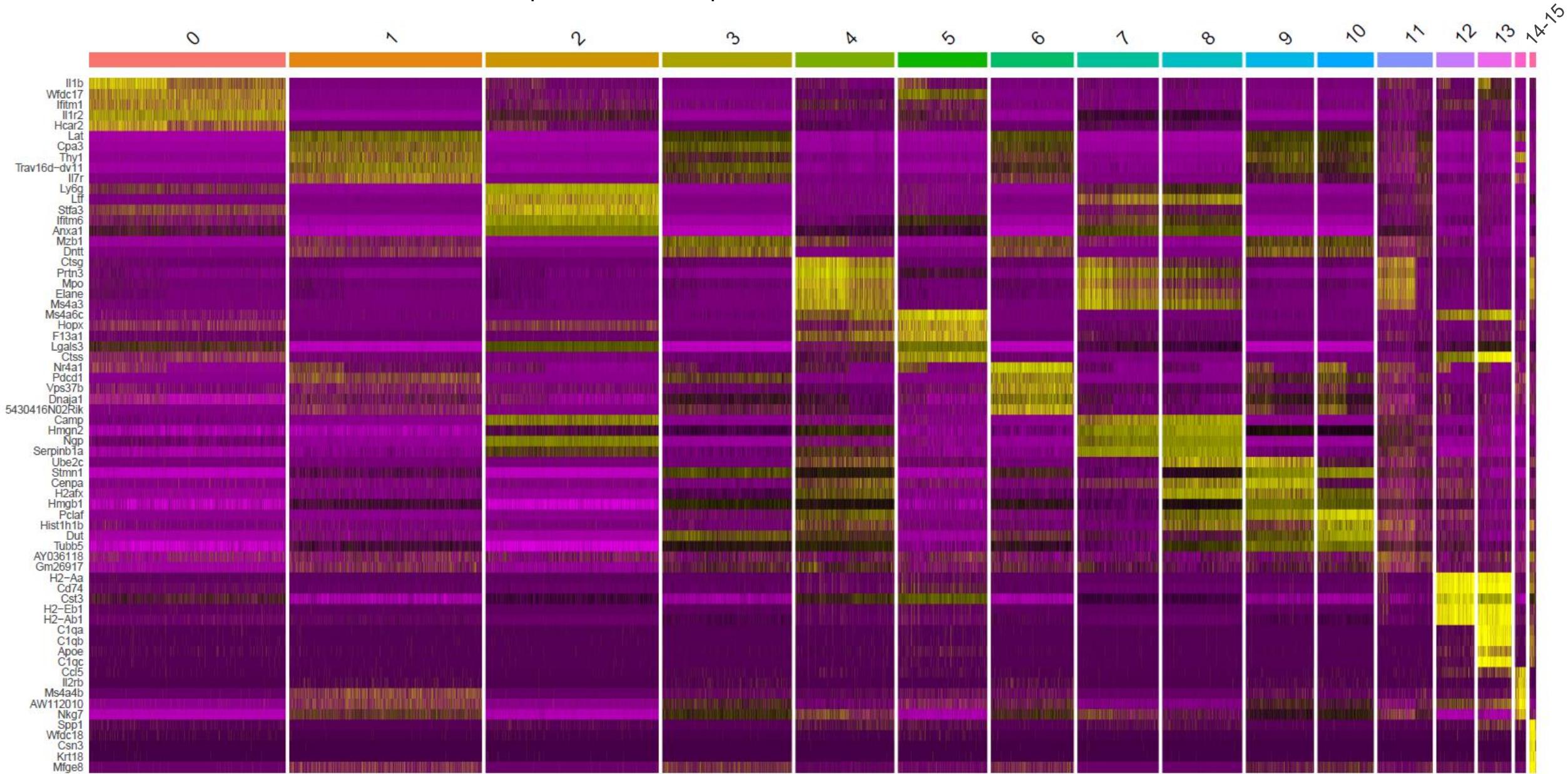
cluster	gene
0	Il1b
0	Wfdc17
0	Ifitm1
0	Il1r2
0	Hcar2
1	Lat
1	Cpa3
1	Thy1
1	Trav16d-dv11
1	Il7r
2	Ly6g
2	Ltf
2	Stfa3
2	Ifitm6
2	Anxa1
3	Lgals1
3	Cpa3
3	Mzb1
3	Dntt
3	Pebp1

cluster	gene
4	Ctsg
4	Prtn3
4	Mpo
4	Elane
4	Ms4a3
5	Ms4a6c
5	Hopx
5	F13a1
5	Lgals3
5	Ctss
6	Nr4a1
6	Pdcd1
6	Vps37b
6	Dnaja1
6	5430416N02Rik
7	Elane
7	Mpo
7	Prtn3
7	Camp
7	Hmgn2

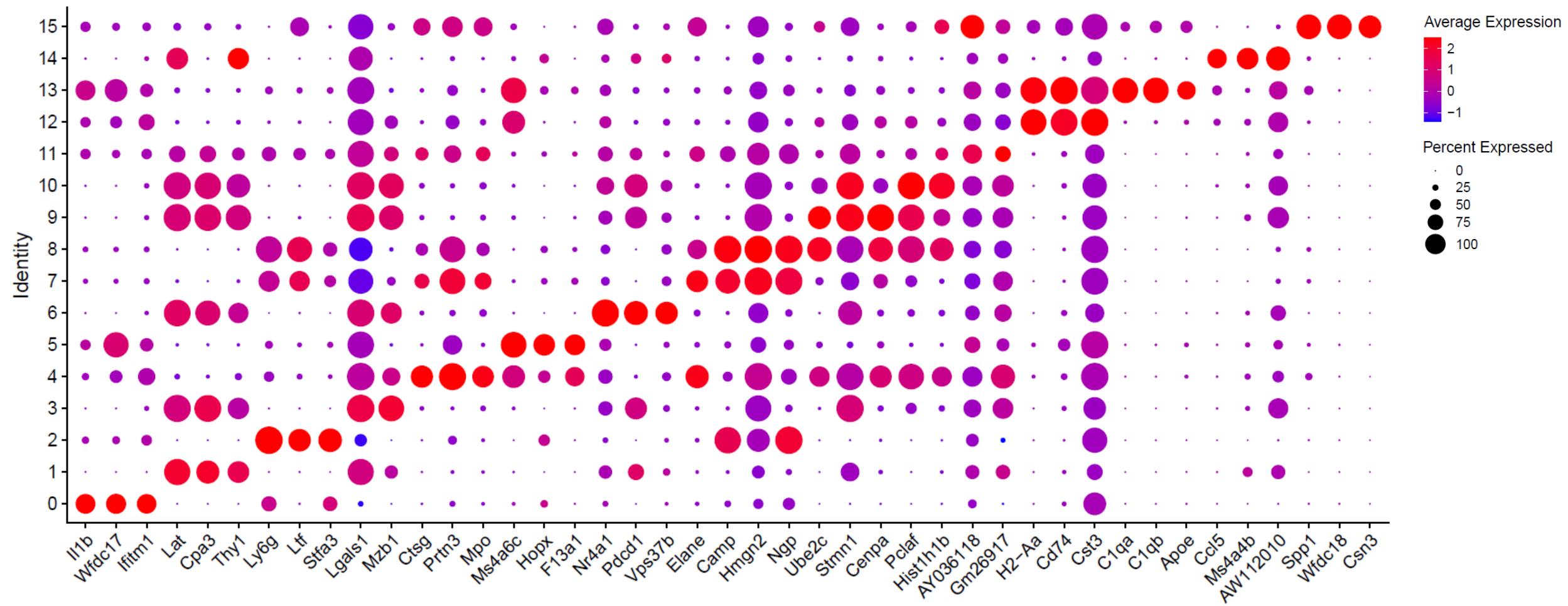
cluster	gene
8	Camp
8	Hmgn2
8	Ngp
8	Serpina1a
8	Ube2c
9	Ube2c
9	Stmn1
9	Cenpa
9	H2afx
9	Hmgb1
10	Pclf
10	Stmn1
10	Hist1h1b
10	Dut
10	Tubb5
11	mt-Cytb
11	Elane
11	Prtn3
11	AY036118
11	Gm26917

cluster	gene
12	H2-Aa
12	Cd74
12	Cst3
12	H2-Eb1
12	H2-Ab1
13	C1qa
13	C1qb
13	Apoe
13	C1qc
13	Cd74
14	Ccl5
14	Il2rb
14	Ms4a4b
14	AW112010
14	Nkg7
15	Spp1
15	Wfdc18
15	Csn3
15	Krt18
15	Mfge8

Gene expression heatmap of the main markers across clusters



Gene expression bubble chart of the main markers across clusters



05

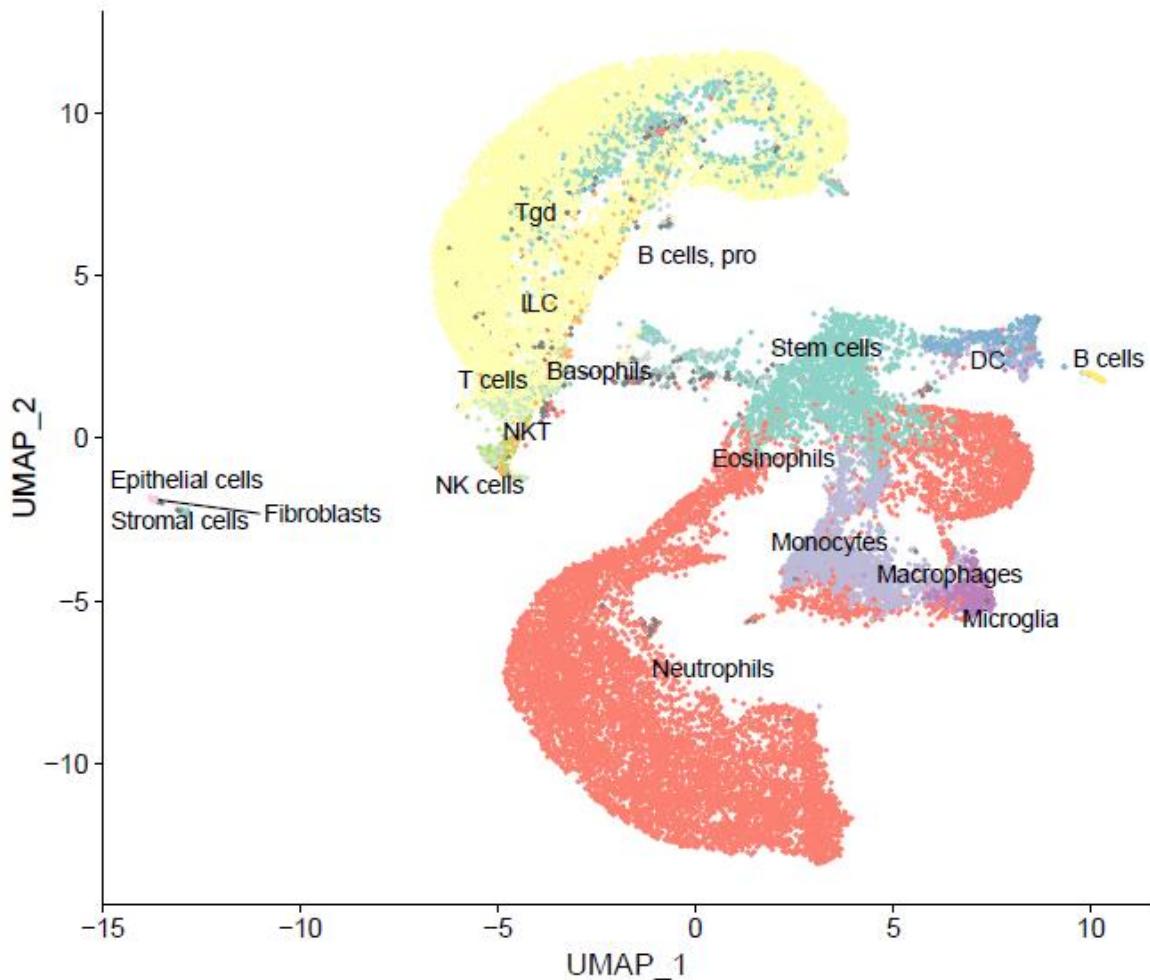
Cell type annotation



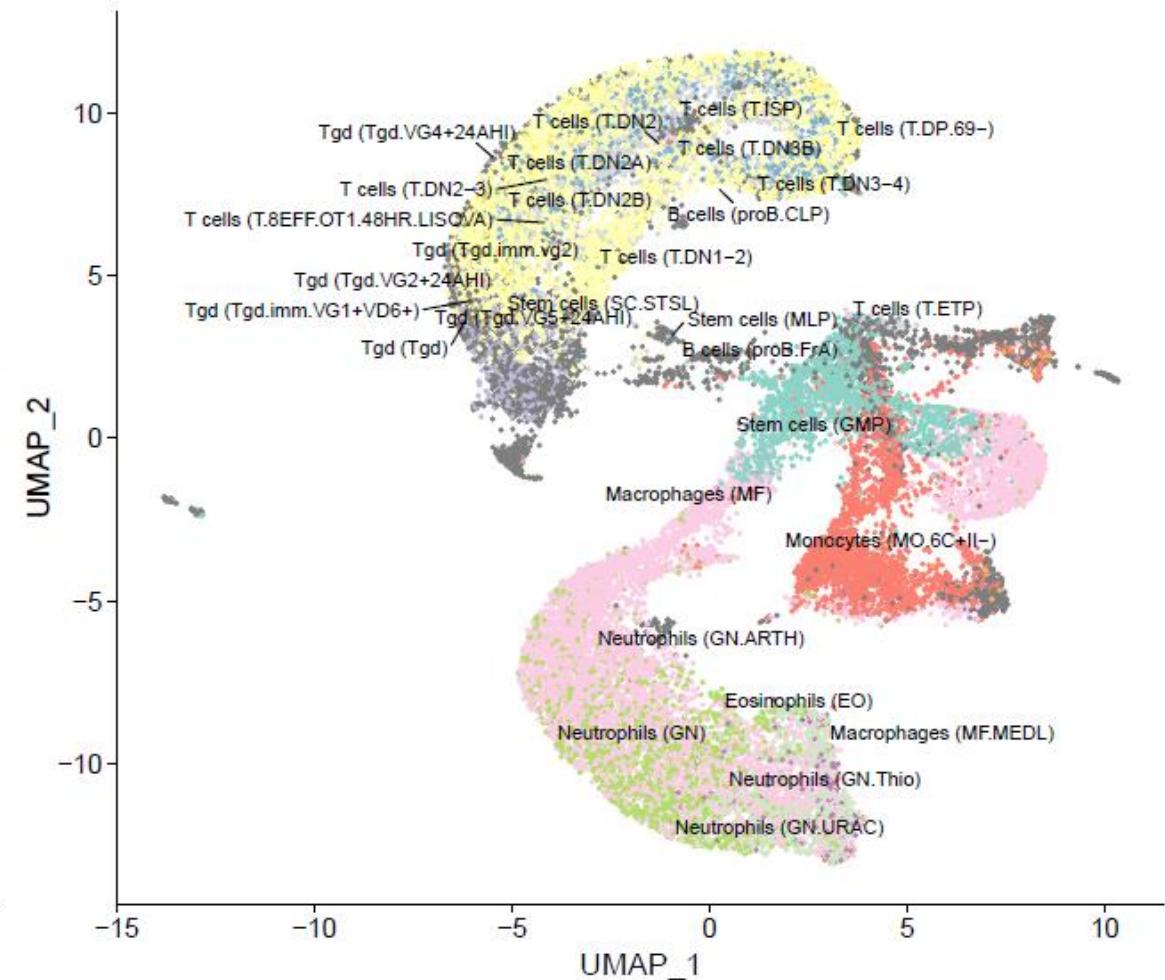
15. Assign cell types to the clusters using the ImmGen database as reference

- Convert the Seurat object into a single cell experiment object
- Get the ImmGen reference database from celldex package
- Assign cell types to the clusters using SingleR

Cell types: ImmGen (main labels)



Cell types: ImmGen (fine labels)



16. Check whether the cell types assigned with the main labels (general) and the fine labels (detailed subgroups) match well
17. Check what are the cell types from each cluster

ImmGen (main labels)

DC

Macrophages

Monocytes

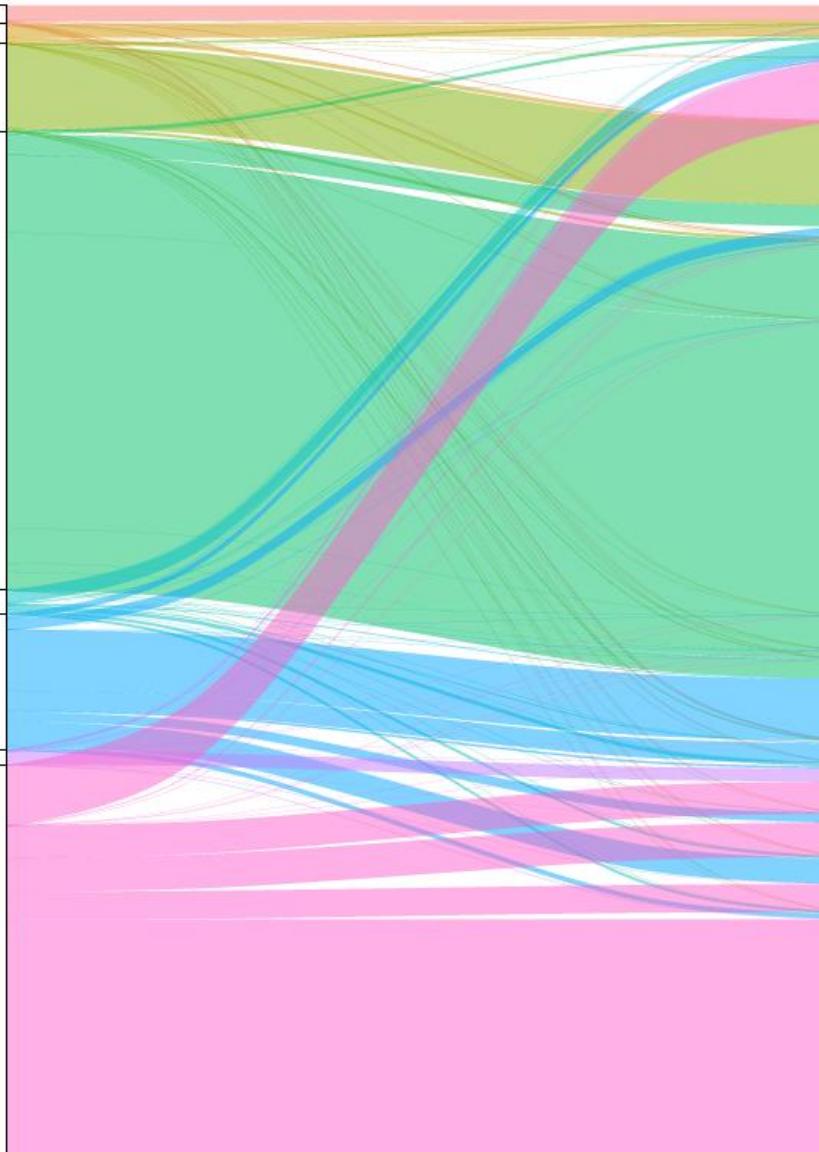
Neutrophils

Other

Stem cells

T cells

Tgd



ImmGen (fine labels)

DC

Other

Tgd

Monocytes (MO.6C+II-)

Monocytes (other)

Neutrophils (GN)

Neutrophils (GN.ARTH)

Neutrophils (GN.URAC)

Neutrophils (other)

Stem cells (GMP)

Stem cells (other)

T cells (other)

T cells (T.DN2)

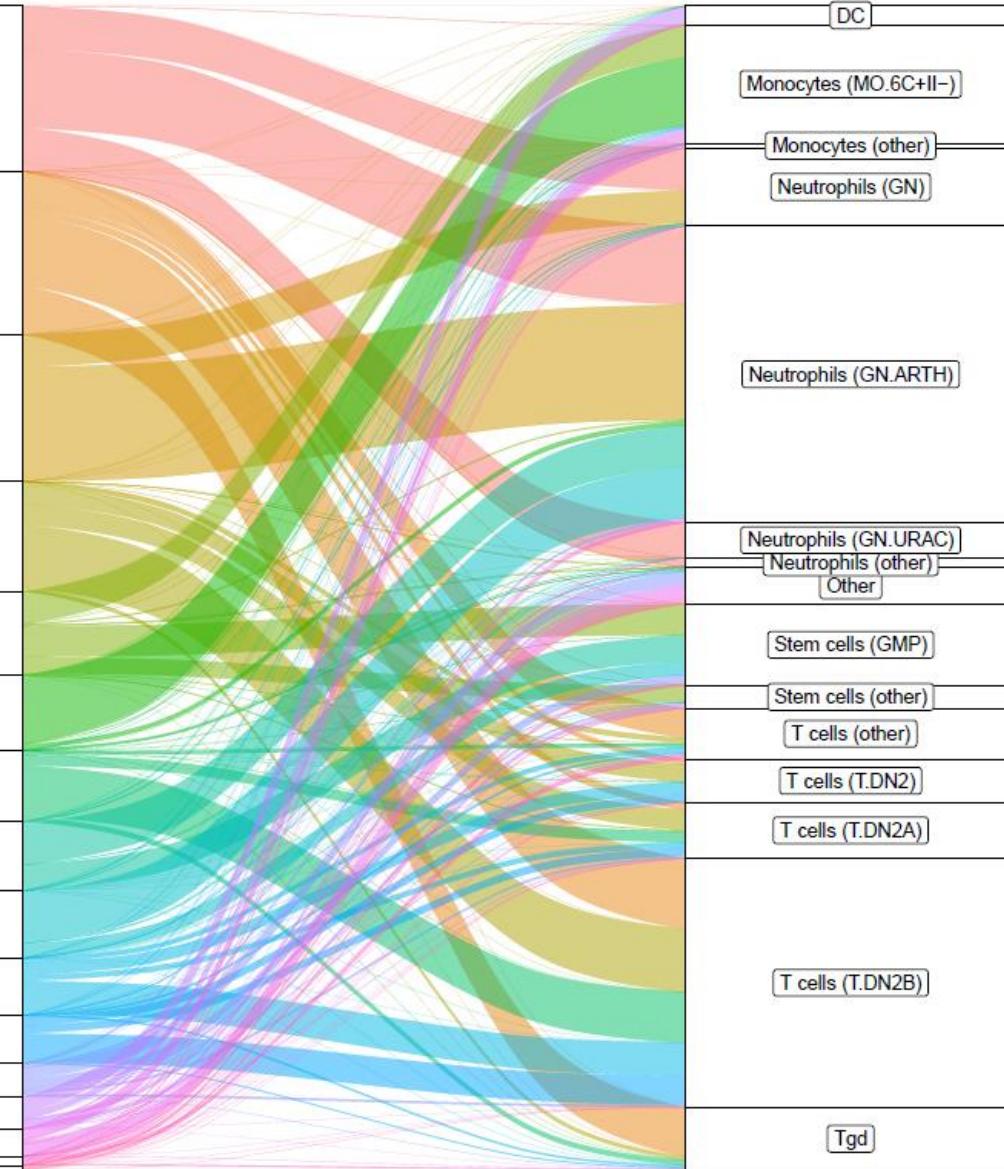
T cells (T.DN2A)

T cells (T.DN2B)

Clusters

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

ImmGen (fine labels)



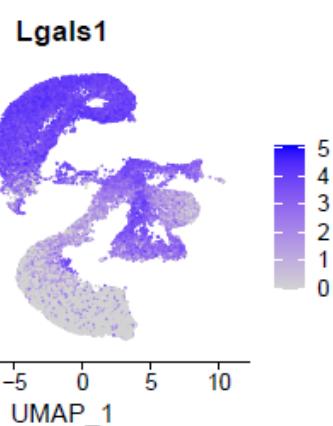
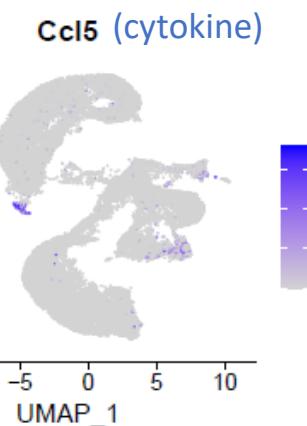
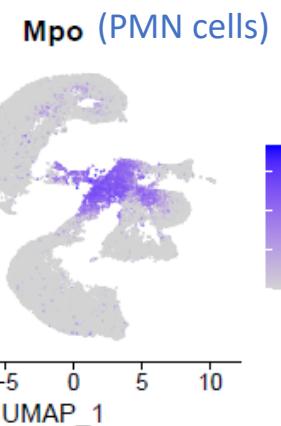
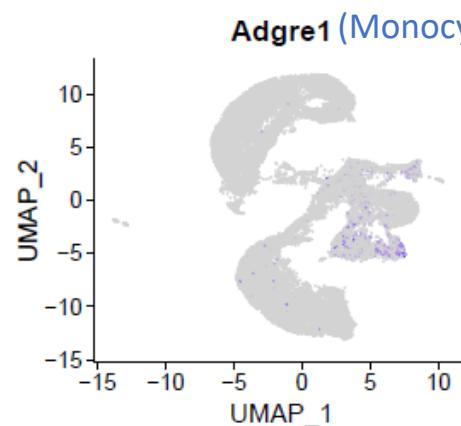
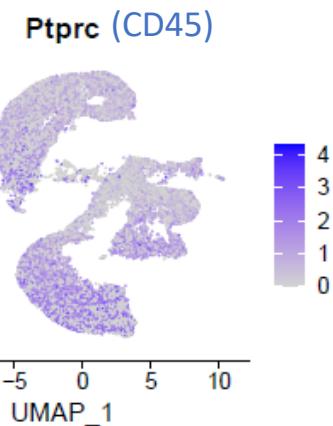
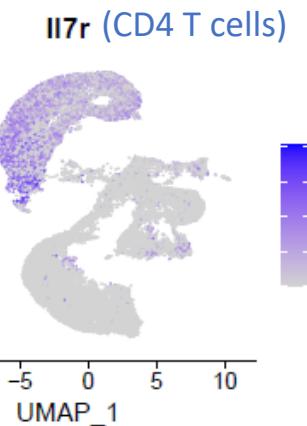
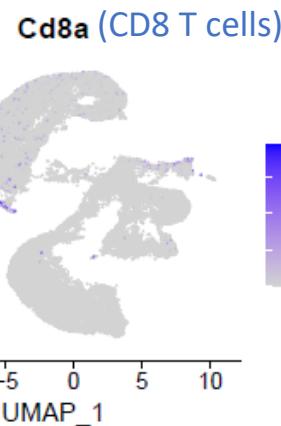
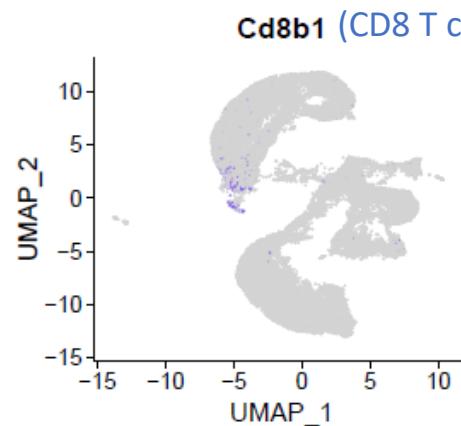
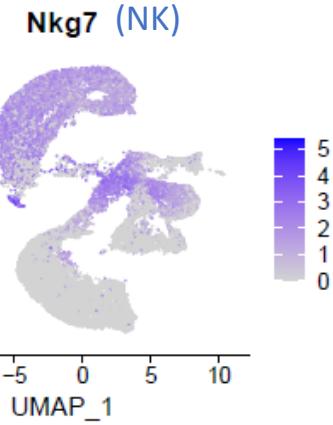
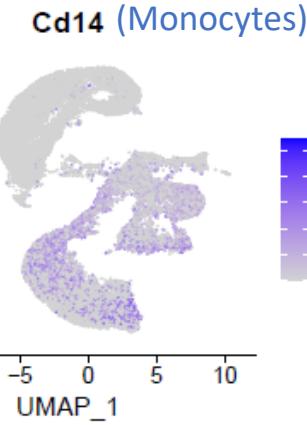
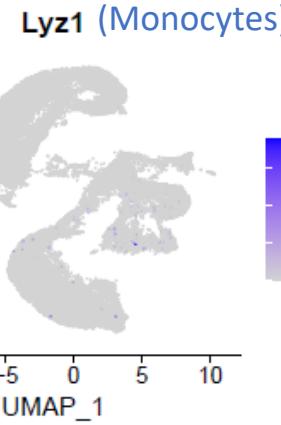
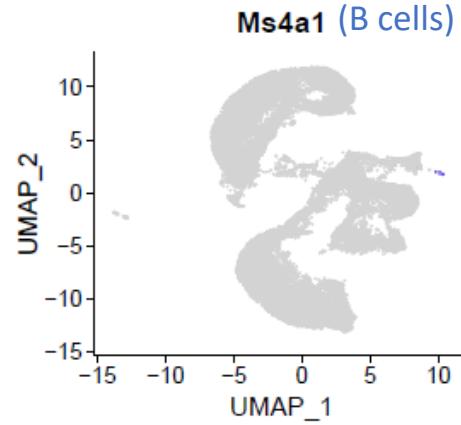
06

Coloring by genes

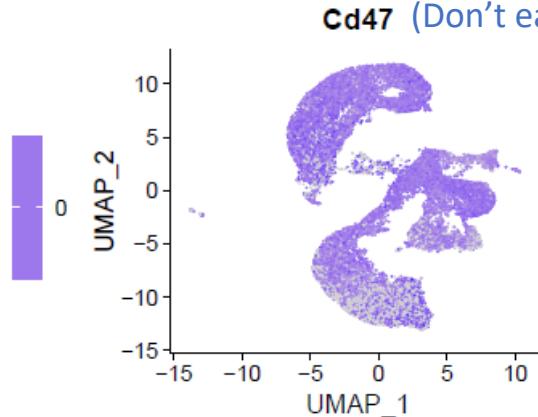
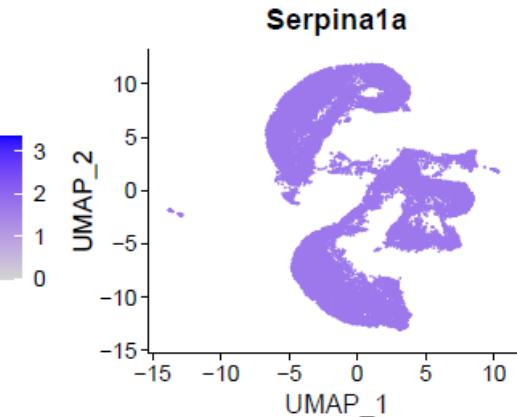
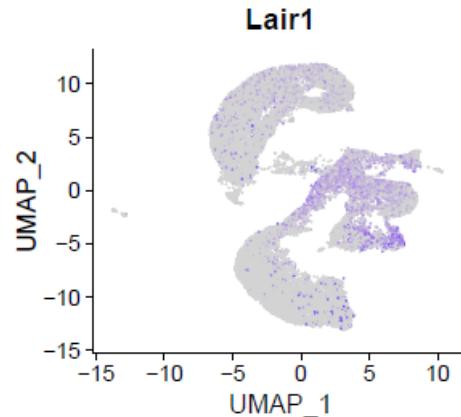
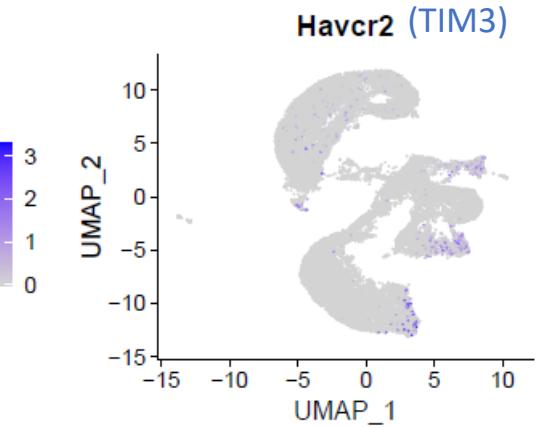
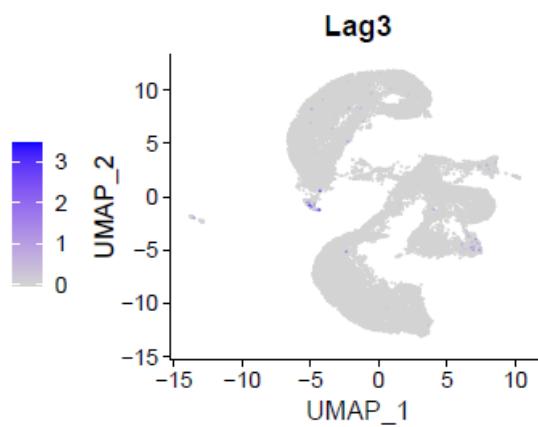
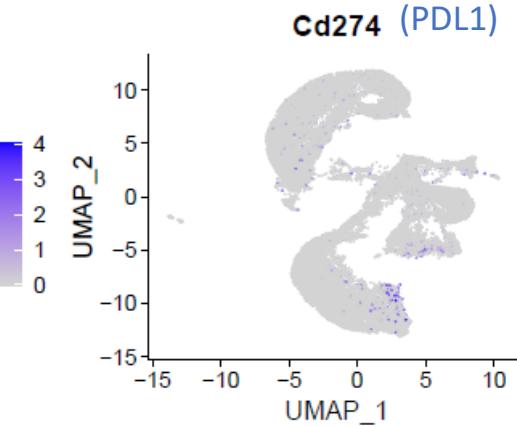
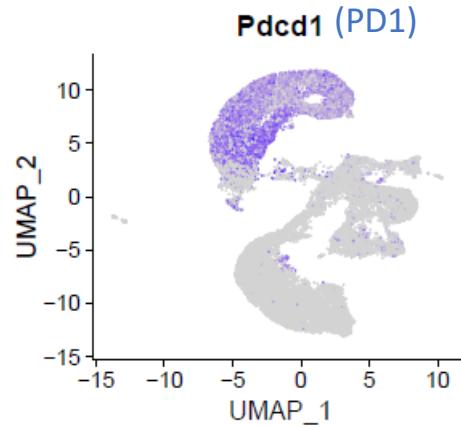
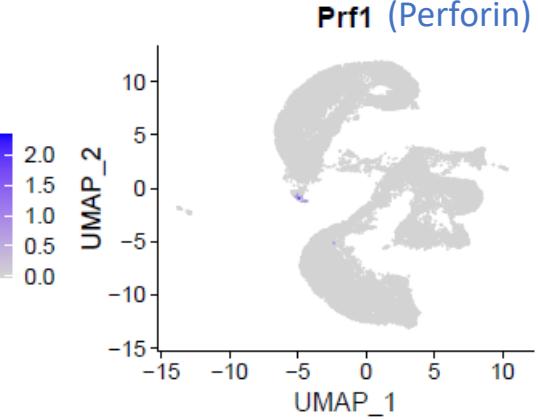
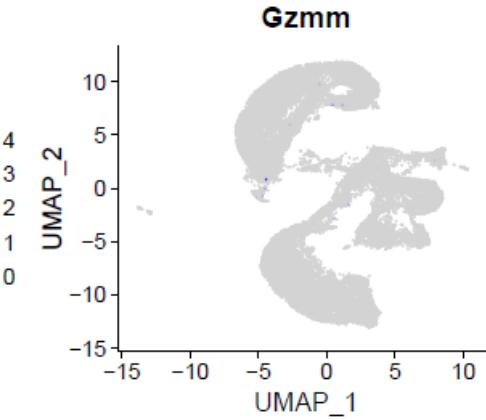
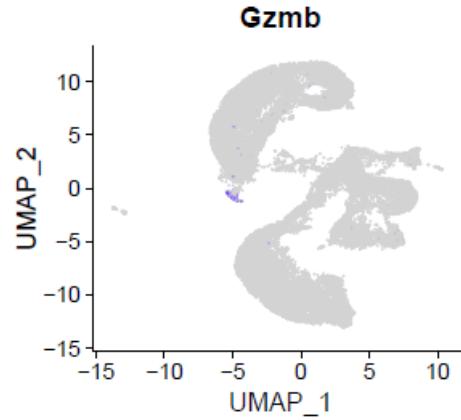


18. Visualize the distribution of the expression of some interesting genes across cells

Distribution of some famous markers

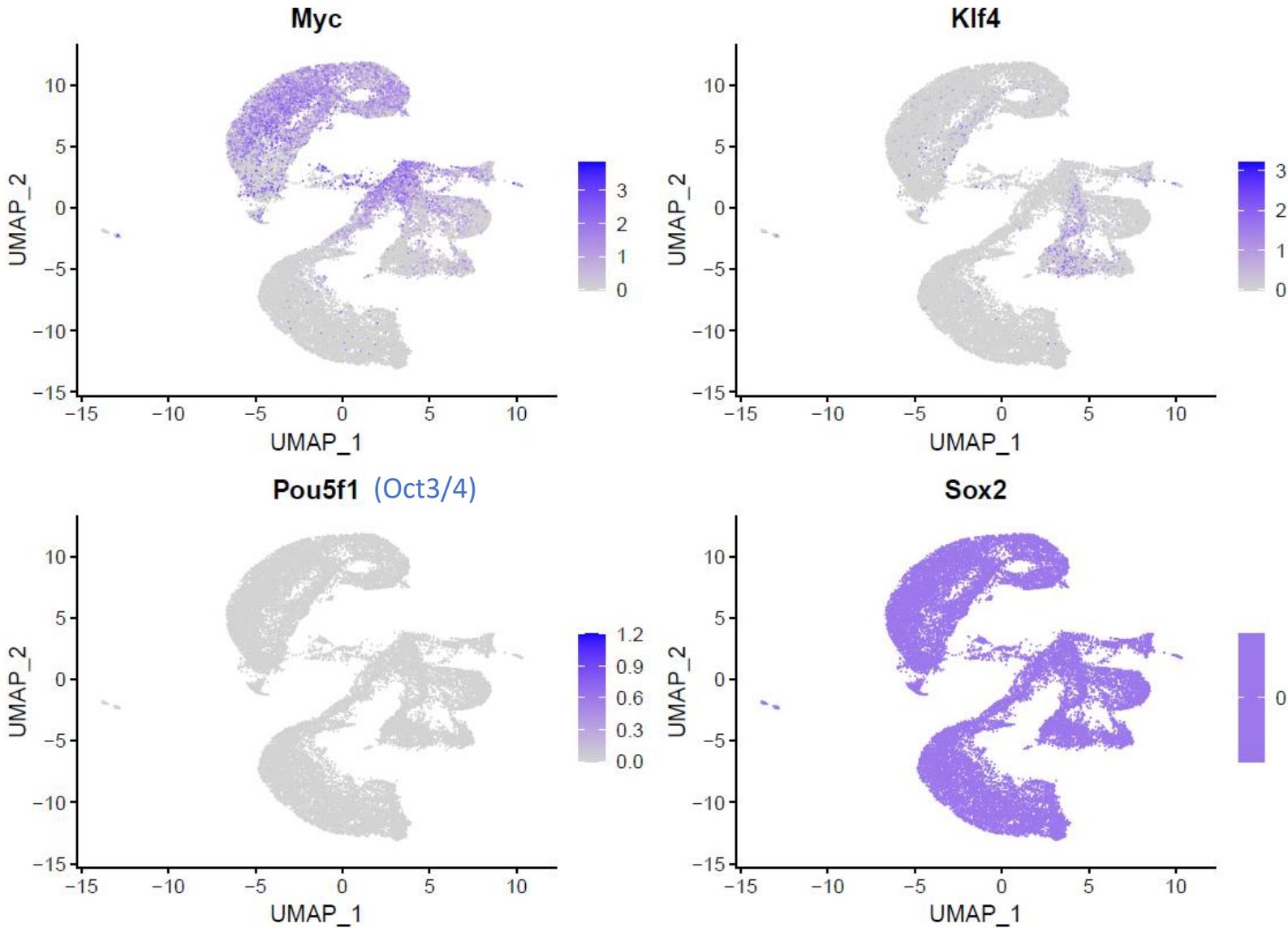


Distribution of some markers we were curious about



Granzyme h,
Lair2 and
Serpine A
not available

Distribution of the Yamanaka factors

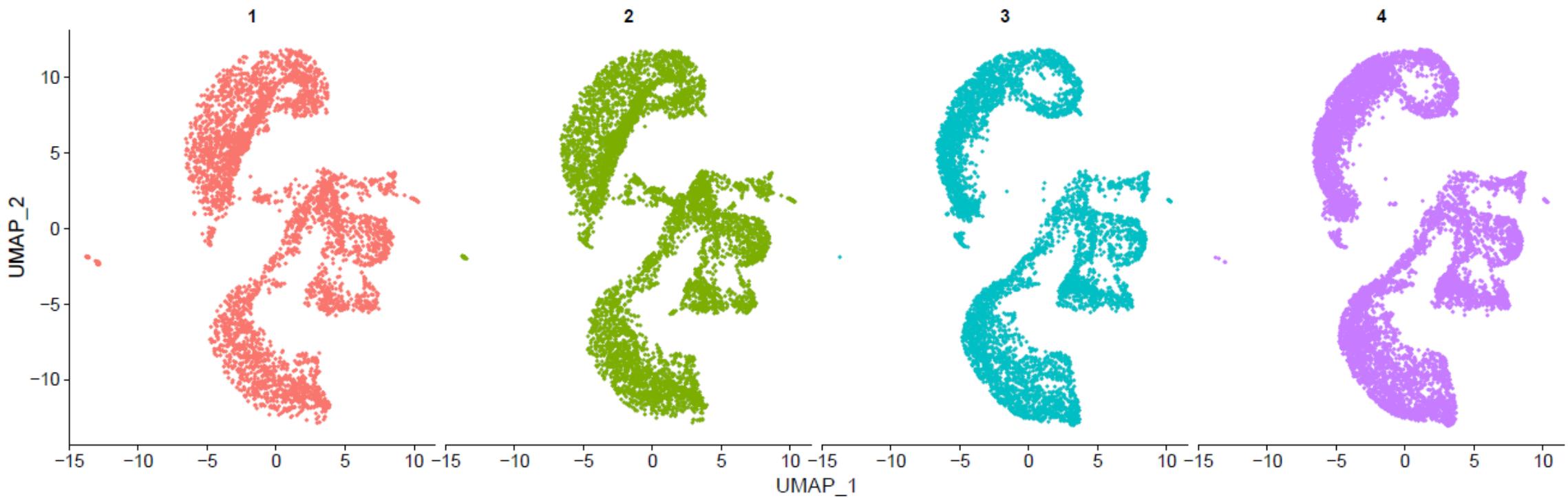


07

Comparison of
the dissociation
methods

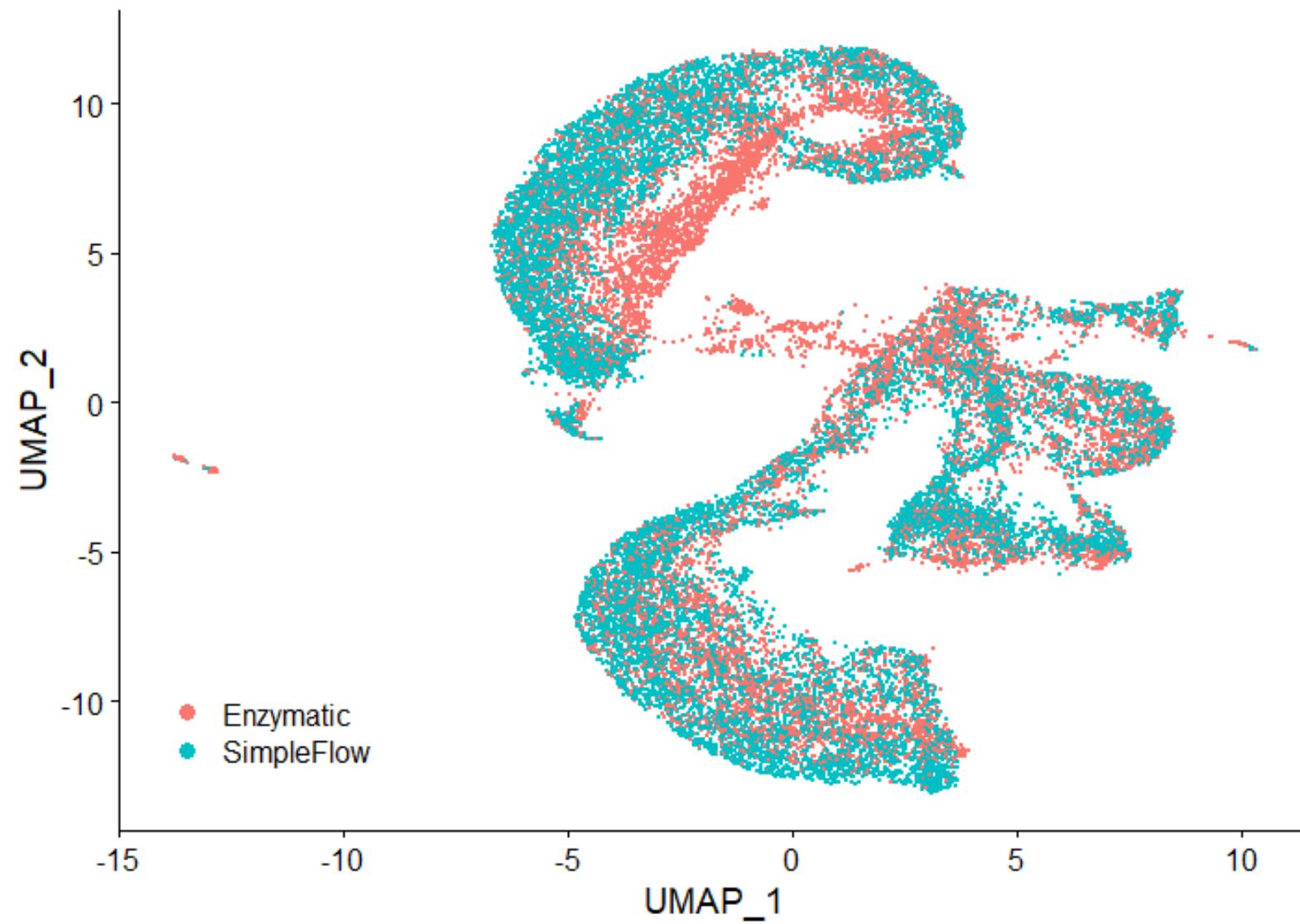


1. Visualize the distribution of cells from each murine mammary tumor and from the two dissociation groups



Samples 1 & 2: Single cells were dissociated with Enzymatic digestion

Samples 3 & 4: Single cells were dissociated with SimpleFlow™ System



2. Perform a differential gene expression (DGE) analysis between the two dissociation groups to identify the expression of what genes changes for cells of the same type under different dissociation techniques

For that, we will use the `FindAllMarkers()` R function that uses the Wilcoxon Rank Sum test

For simplicity, we will focus on the largest cell types

Top 15 differentially expressed genes between dissociation methods by cell type

```
> head(Monocytes, n = 15)
```

		p_val	avg_log2FC	pct.1	pct.2	p_val_adj
•	Hspa1b	1.040215e-155	1.5454332	0.538	0.031	3.230179e-151
•	Nr4a1	5.232288e-130	2.4541140	0.743	0.241	1.624782e-125
•	Hspa1a	4.230262e-110	1.5763296	0.497	0.068	1.313623e-105
•	Dnajb1	1.186188e-75	1.6965480	0.659	0.305	3.683470e-71
Atf3		9.978361e-59	2.0161393	0.790	0.593	3.098580e-54
Ccl4		4.264070e-51	2.0004175	0.622	0.319	1.324122e-46
Hsp90aa1		5.711892e-46	1.0097817	0.786	0.552	1.773714e-41
Cdkn1a		7.336598e-39	1.0691153	0.688	0.447	2.278234e-34
H3f3b		1.700543e-37	0.5771917	1.000	0.999	5.280697e-33
Btg2		4.847031e-35	0.8355704	0.874	0.785	1.505149e-30
Srgn		8.298998e-34	0.8085828	0.903	0.772	2.577088e-29
Osm		4.413170e-32	1.0151987	0.428	0.199	1.370422e-27
Ier3		1.714669e-31	1.1314937	0.852	0.734	5.324563e-27
Ccl3		2.459093e-31	0.8946007	0.772	0.589	7.636220e-27
Il1b		1.501253e-30	2.0214160	0.450	0.237	4.661841e-26

```
> head(Neutrophils, n = 15)
```

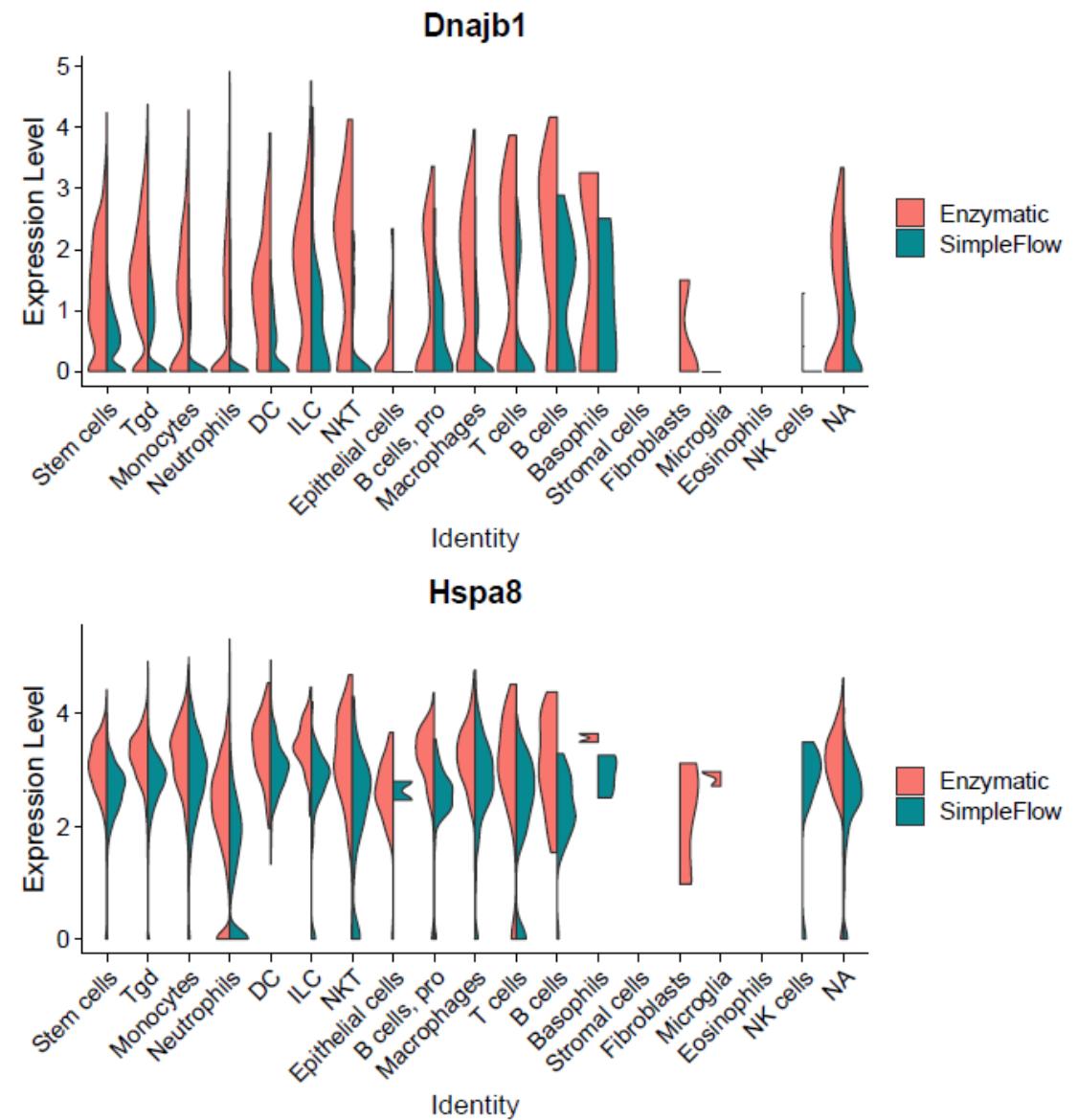
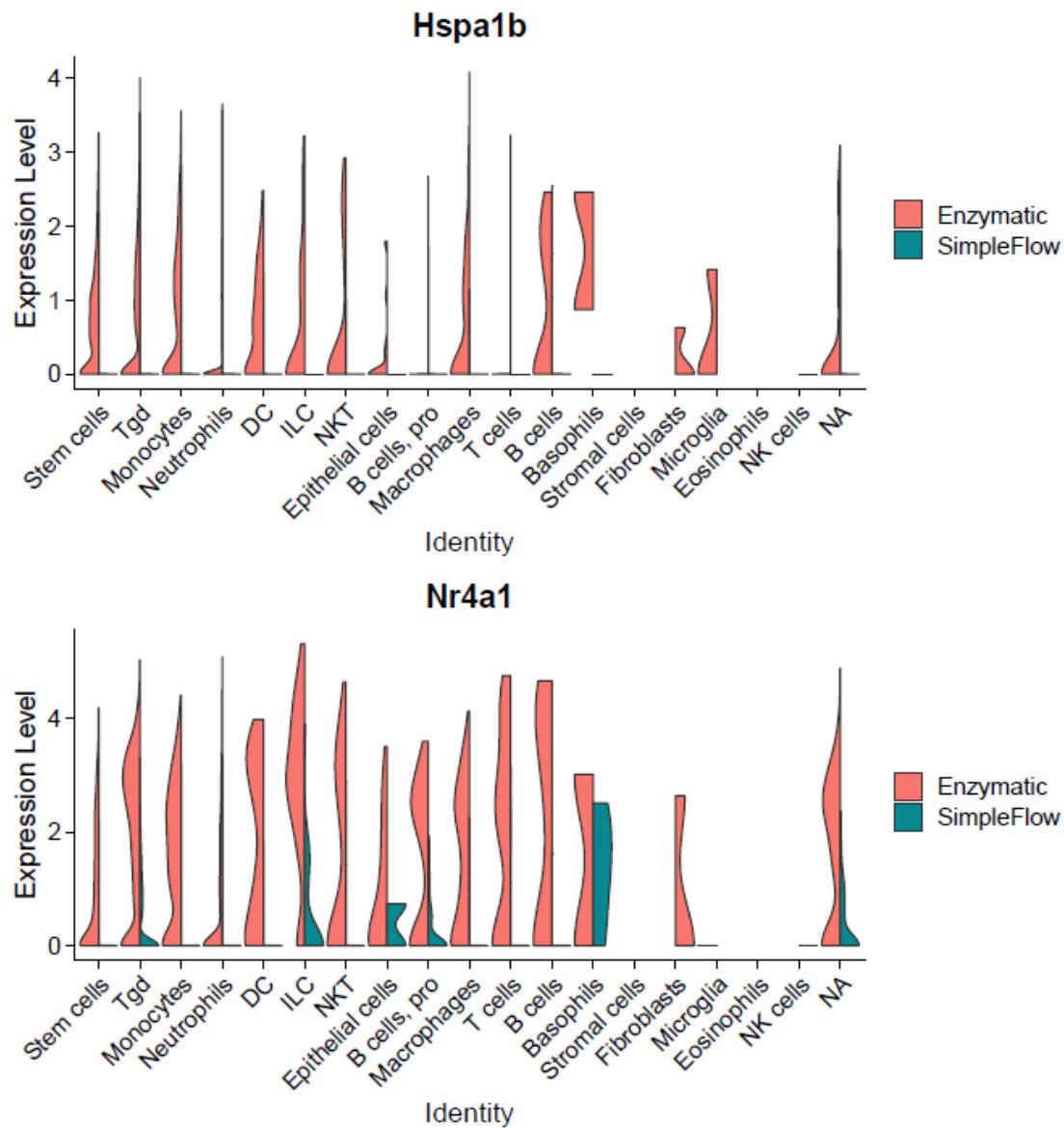
		p_val	avg_log2FC	pct.1	pct.2	p_val_adj
	G0s2	0.000000e+00	2.2647916	0.762	0.531	0.000000e+00
•	Hspa8	0.000000e+00	1.1134147	0.879	0.711	0.000000e+00
•	Nr4a1	0.000000e+00	1.9639366	0.435	0.071	0.000000e+00
•	Hspa1b	0.000000e+00	0.7527038	0.256	0.011	0.000000e+00
	Csrnp1	1.446207e-240	1.0648900	0.448	0.153	4.490907e-236
•	Dnajb1	2.659682e-220	1.1938296	0.607	0.334	8.259110e-216
	Klf6	1.515432e-217	0.8794987	0.892	0.740	4.705870e-213
	Trem1	6.158587e-211	1.0356728	0.686	0.403	1.912426e-206
	Plaur	1.202167e-208	1.0528873	0.825	0.656	3.733089e-204
	Hsp90aa1	1.559493e-204	0.9325669	0.580	0.320	4.842694e-200
	Fosl1	1.995955e-200	0.8110221	0.302	0.076	6.198039e-196
	Nfkbid	6.757802e-194	0.8075579	0.298	0.078	2.098500e-189
	Tnfaip6	5.856903e-184	0.6785843	0.175	0.016	1.818744e-179
•	Hspa1a	1.105735e-183	0.5400613	0.163	0.011	3.433638e-179
	Hsp90ab1	2.126125e-175	0.8851453	0.778	0.622	6.602257e-171

```
> head(Tgd, n = 15)
```

		p_val	avg_log2FC	pct.1	pct.2	p_val_adj
•	Dnajb1	0.000000e+00	1.4599982	0.854	0.613	0.000000e+00
•	Hspa8	0.000000e+00	0.6472637	0.996	0.985	0.000000e+00
•	Nr4a1	0.000000e+00	2.6383861	0.785	0.420	0.000000e+00
•	Hspa1b	0.000000e+00	1.5541028	0.560	0.016	0.000000e+00
•	Hspa1a	0.000000e+00	1.6776488	0.446	0.014	0.000000e+00
	Hsp90aa1	2.583884e-275	0.6212419	0.969	0.909	8.023735e-271
	Hsp90ab1	3.865601e-230	0.3527013	0.999	0.997	1.200385e-225
	Cdkn1a	2.824298e-223	1.1446326	0.448	0.169	8.770294e-219
	Ppp1r15a	3.299994e-209	0.8235387	0.765	0.499	1.024747e-204
	Plaur	9.011602e-207	1.0946360	0.645	0.361	2.798373e-202
	Atf3	4.832174e-199	0.9064369	0.207	0.015	1.500535e-194
	Tnfsf11	2.023240e-198	0.6370298	0.195	0.011	6.282768e-194
	Csrnp1	5.380092e-192	0.8110122	0.475	0.203	1.670680e-187
	Nr4a2	5.067422e-191	0.8392308	0.403	0.147	1.573587e-186
	Jun	1.030369e-186	0.8721499	0.897	0.748	3.199606e-182

```
> head(STEM.CELLS, n = 15)
```

		p_val	avg_log2FC	pct.1	pct.2	p_val_adj
•	Hspa1b	2.711773e-247	1.2752226	0.660	0.047	8.420869e-243
•	Hspa1a	1.161467e-155	1.2118726	0.489	0.044	3.606703e-151
•	Dnajb1	7.792747e-132	1.5208896	0.839	0.710	2.419882e-127
•	Nr4a1	3.926794e-123	1.9038705	0.587	0.220	1.219387e-118
•	Hspa8	2.028165e-112	0.6117943	0.987	0.995	6.298060e-108
	Hsp90aa1	2.567353e-91	0.5895800	0.962	0.991	7.972401e-87
	Jun	7.146602e-67	0.8336549	0.907	0.888	2.219234e-62
	Hspf1	7.508380e-61	0.4787178	0.433	0.167	2.331577e-56
	Hnrnpab	8.344769e-60	-0.4266460	0.874	0.980	2.591301e-55
	Rps18	2.037699e-56	0.3285766	0.948	0.988	6.327668e-52
	Bag3	2.962035e-56	0.5430598	0.403	0.160	9.198009e-52
	Srsf2	4.995195e-56	-0.3655303	0.882	0.973	1.551158e-51
	Gnai2	9.259230e-50	-0.4248351	0.819	0.947	2.875269e-45
	Ppp1cc	1.249587e-49	-0.3403015	0.826	0.946	3.880342e-45
	Gm10260	1.705010e-48	-0.2793919	0.980	1.000	5.294567e-44



3. Perform pathway analysis of the top 15 differentially expressed genes per cell type between dissociation groups using the DAVID platform ([link](#))

Main differentially expressed pathways between dissociation methods by cell type

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment	Benjamini
□	UP_KEYWORDS	Stress response	RT		6	40.0	6.8E-10	121.0	3.0E-8
□	UP_KEYWORDS	Chaperone	RT		6	40.0	7.6E-8	47.5	1.7E-6
□	GOTERM_MF_DIRECT	unfolded protein binding	RT		5	33.3	2.2E-7	83.1	1.6E-5
□	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		6	40.0	2.5E-7	30.5	5.2E-6
□	KEGG_PATHWAY	Antigen processing and presentation	RT		5	33.3	8.1E-7	52.1	8.5E-6

Neutrophils (n=10,202):

1) **Stress response**

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment	Benjamini
□	UP_KEYWORDS	Stress response	RT		7	46.7	3.1E-12	141.1	1.6E-10
□	UP_KEYWORDS	Chaperone	RT		6	40.0	7.6E-8	47.5	1.9E-6
□	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		7	46.7	1.5E-7	22.9	9.2E-6
□	GOTERM_MF_DIRECT	unfolded protein binding	RT		5	33.3	3.1E-7	77.5	2.9E-5
□	KEGG_PATHWAY	Estrogen signaling pathway	RT		6	40.0	3.6E-7	33.6	1.1E-5

Tgd (n=9,706):

1) **Stress response**

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment	Benjamini
□	UP_KEYWORDS	Stress response	RT		6	40.0	6.8E-10	121.0	3.8E-8
□	INTERPRO	Heat shock protein 70 family	RT		4	26.7	7.1E-8	422.4	1.4E-6
□	INTERPRO	Heat shock protein 70, conserved site	RT		4	26.7	7.1E-8	422.4	1.4E-6
□	UP_KEYWORDS	Chaperone	RT		6	40.0	7.6E-8	47.5	1.5E-6
□	UP_KEYWORDS	Acetylation	RT		12	80.0	8.1E-8	5.8	1.5E-6

Stem cells (n=3,109):

1) **Stress response**

2) **Heat shock protein 70 family**

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment	Benjamini
□	GOTERM_BP_DIRECT	response to heat	RT		4	26.7	1.1E-5	3.2E-3	
□	UP_KEYWORDS	Stress response	RT		4	26.7	1.2E-5	6.2E-4	
□	GOTERM_BP_DIRECT	negative regulation of cell proliferation	RT		5	33.3	1.7E-4	2.5E-2	
□	UP_KEYWORDS	Chaperone	RT		4	26.7	2.0E-4	3.8E-3	
□	UP_KEYWORDS	Cytokine	RT		4	26.7	2.3E-4	3.8E-3	

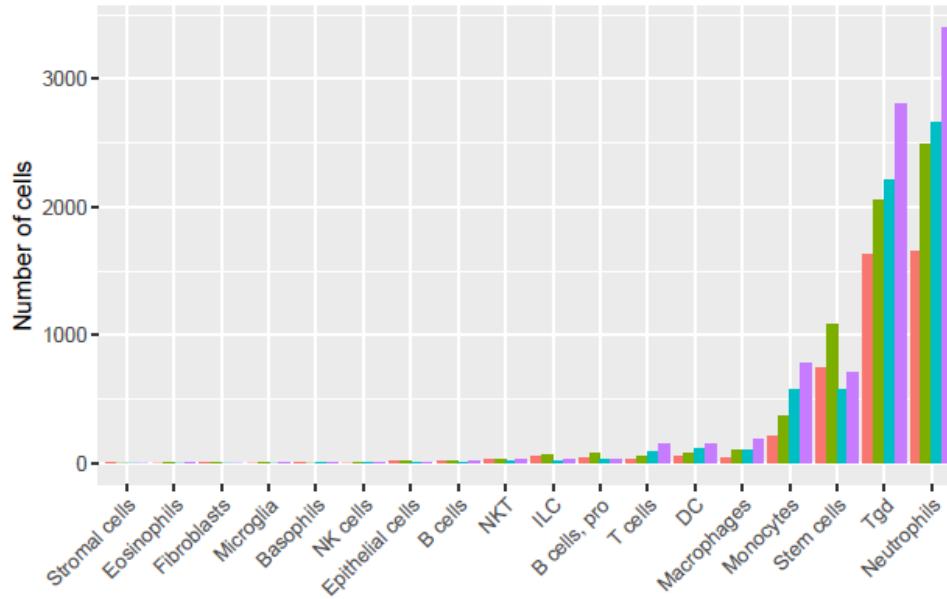
Monocytes (n=1,939):

1) **Response to heat**

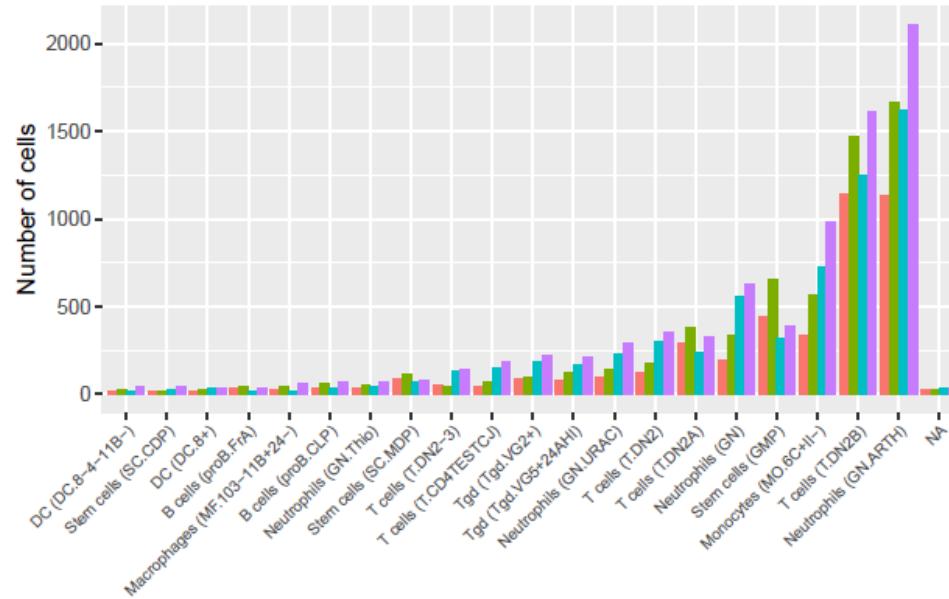
2) **Stress response**

4. Compare the cell type composition of samples from the two dissociation groups

Cell types: ImmGen (main labels)

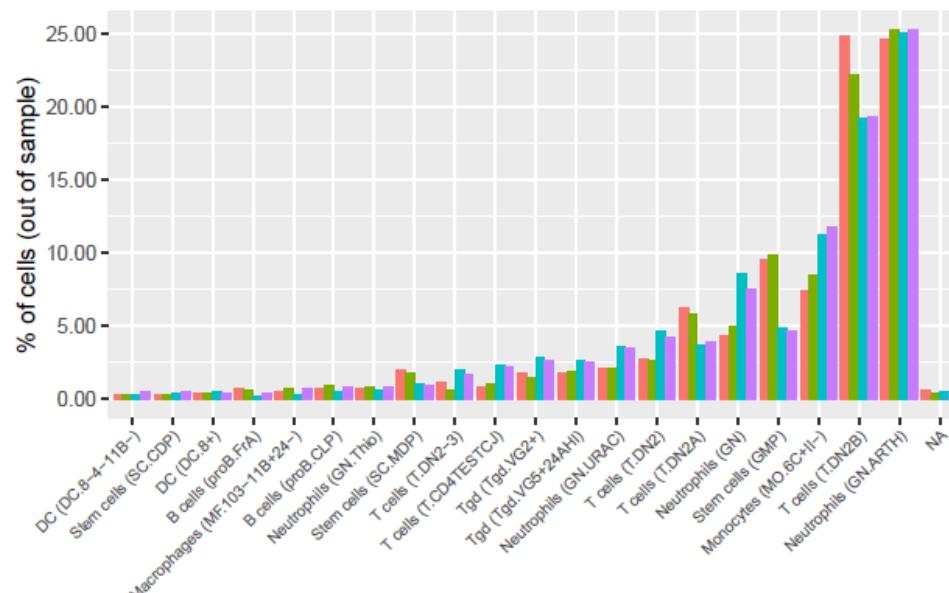
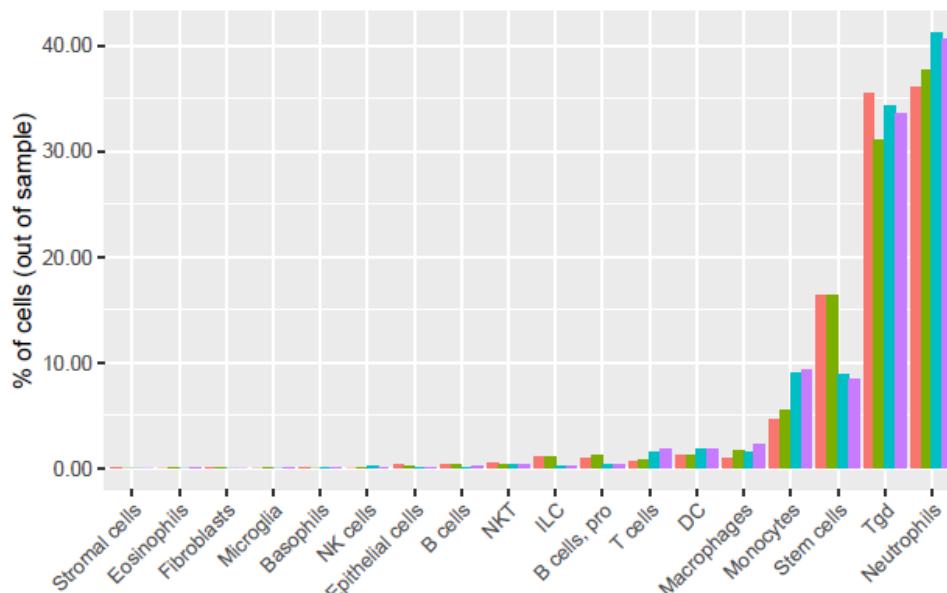


Cell types: ImmGen (fine labels)



Sample ID
1
2
3
4

Samples 1 & 2:
Single cells were
dissociated with
enzymatic digestion

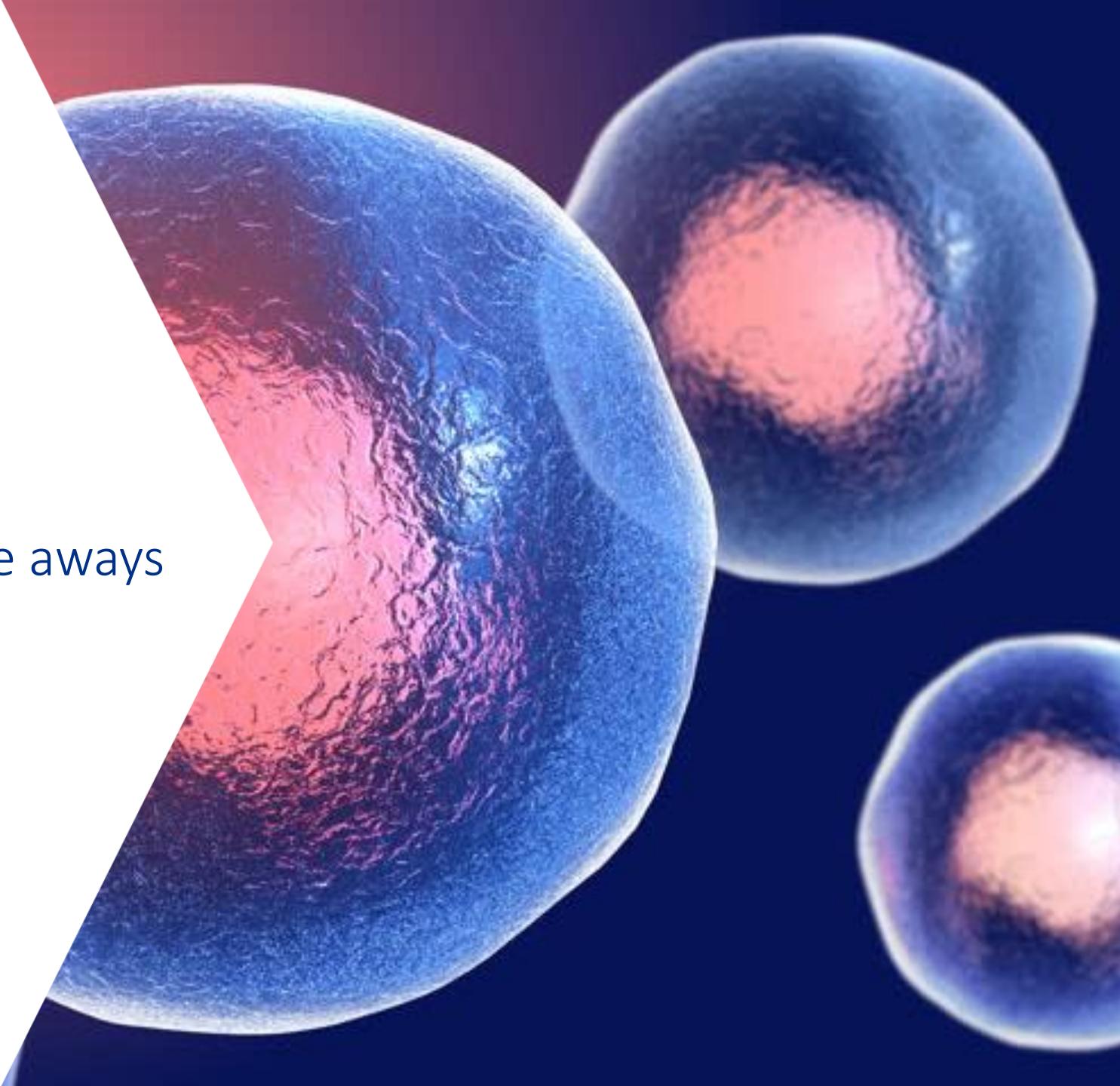


Total # cells per sample:

1 6596
2 6610
3 6458
4 8343

08

The take aways



Summary of key insights:



The most frequent cell types in the samples are neutrophils (GN.ARTH), T cells (T.DN2B), and monocytes (MO.6C+II-).



The cell type compositions of samples dissociated with enzymatic digestion and the SimpleFlowTM System are very similar, though samples dissociated with enzymatic digestion contain a slightly higher proportion of T cells (T.DN2B).



Cells dissociated with enzymatic digestion have a substantial increase in expression of genes from the “stress response” and the “response to heat” pathways. Overexpression of these pathways is indicative of cellular stress conditions imposed by the circumstances of tissue digestion and the high temperature of enzymatic digestion (it is performed at 37°, whereas SimpleFlowTM is performed at 4°).



Overall, the SimpleFlowTM system has demonstrated to be an excellent dissociation methodology of single cells because: 1) it generates similar results to those obtained with the conventional methodology of enzymatic digestion, and 2) reduces the cellular stress imposed by the procedures.



THANKS

Does anyone have any questions?