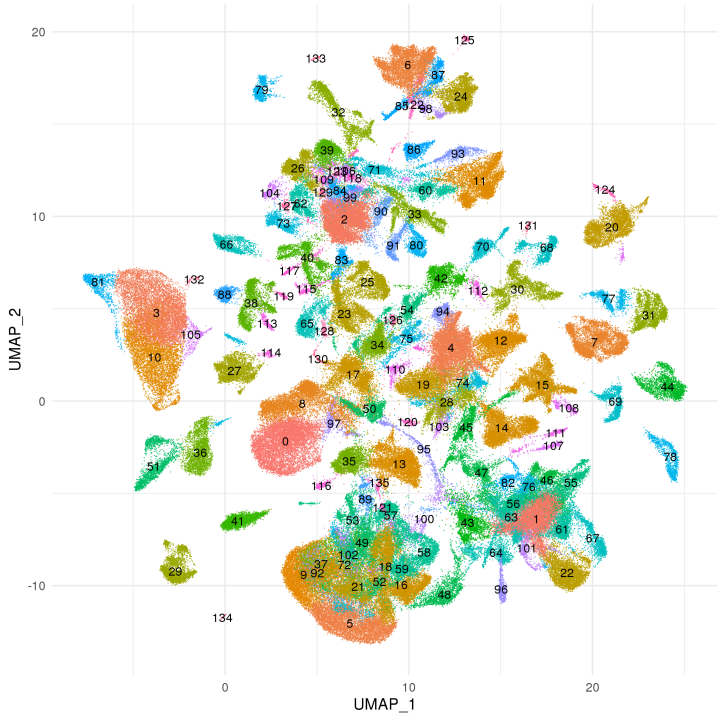
pipeline_seurat.py: summary report

Microwell-seq Mouse Cell Atlas (240k cells)

Sansom group

November 5, 2020



Sample: mca

Run specs: no. components: 75, cluster resolution: 3, cluster algorithm: leiden, de test: wilcox

Code: https://github.com/sansomlab/tenx

Contents

1 Introduction **3**
1.1 Optional tasks . . . 3

2 Data quality control **4**
2.1 Quality assessment and removal of low-quality cells . . . 4

3 Removal of unwanted variation **6**
3.1 Removal of unwanted variation (data normalisation) . . . 6
3.2 Summary statistics . . . 6

4 Dimension reduction **8**
4.1 Scree (elbow) plot . . . 8
4.2 Component heatmaps . . . 9

5 Visualisation of clusters and factors of interest **10**
5.1 umap.mindist_0 plot colored by cluster_id . . . 10
5.2 umap.mindist_0.1 plot colored by cluster_id . . . 12
5.3 umap.mindist_0.3 plot colored by cluster_id . . . 14
5.4 umap.mindist_0.5 plot colored by cluster_id . . . 16
5.5 umap.mindist_0.7 plot colored by cluster_id . . . 18
5.6 umap plot colored by nCount_RNA . . . 20
5.7 umap plot colored by percent.mito . . . 21
5.8 umap plot colored by Tissue . . . 22

6 Plots of summary statistics **24**
6.1 Cells by cluster . . . 24
6.2 Number of genes per cell per cluster . . . 25
6.3 Number of umi per cell per cluster . . . 26

7 Cluster dissimilarity **27**
7.1 Dissimilarity by gene expression . . . 27

8 Identification of cluster marker genes **29**

9 Top cluster marker genes **30**

1 Introduction

The core of the data analysis was performed using Seurat and scanpy:

- The construction of the nearest neighbor graph, clustering and UMAP computation were performed using scanpy (or scvelo for use of hnswlib).
- The differential expression analysis was performed using Seurat.
- The geneset analysis was performed using gsfisher
- Please see https://github/sansomlab/tenx for more details.

The key parameter choices used for this analysis were:

- The number of pca components: 75
- The number of nearest neighbors: 20
- The distance metric used for the nearest neighbor graph: euclidean
- The method used for construction of the nearest neighbor graph: hnsw
- The resolution of the clustering: 3
- The clustering algorithm: leiden
- The differential expression test: wilcox

1.1 Optional tasks

This table summarises the status of the optional tasks. Tasks set to "True" were run.

task	run
explore_hvg_and_cell_cycle	False
singleR	False
jackstraw	False
compare_clusters	True
characterise_markers	False
top_marker_heatmap	True
extra_cluster_marker_plots	True
diffusionmap	False
phate	False
paga	False
velocity	False
knownmarkers	False
marker_report	False
exprsreport	False
genesets	False
cellbrowser	False

2 Data quality control

2.1 Quality assessment and removal of low-quality cells

Figure 1: Basic QC plots

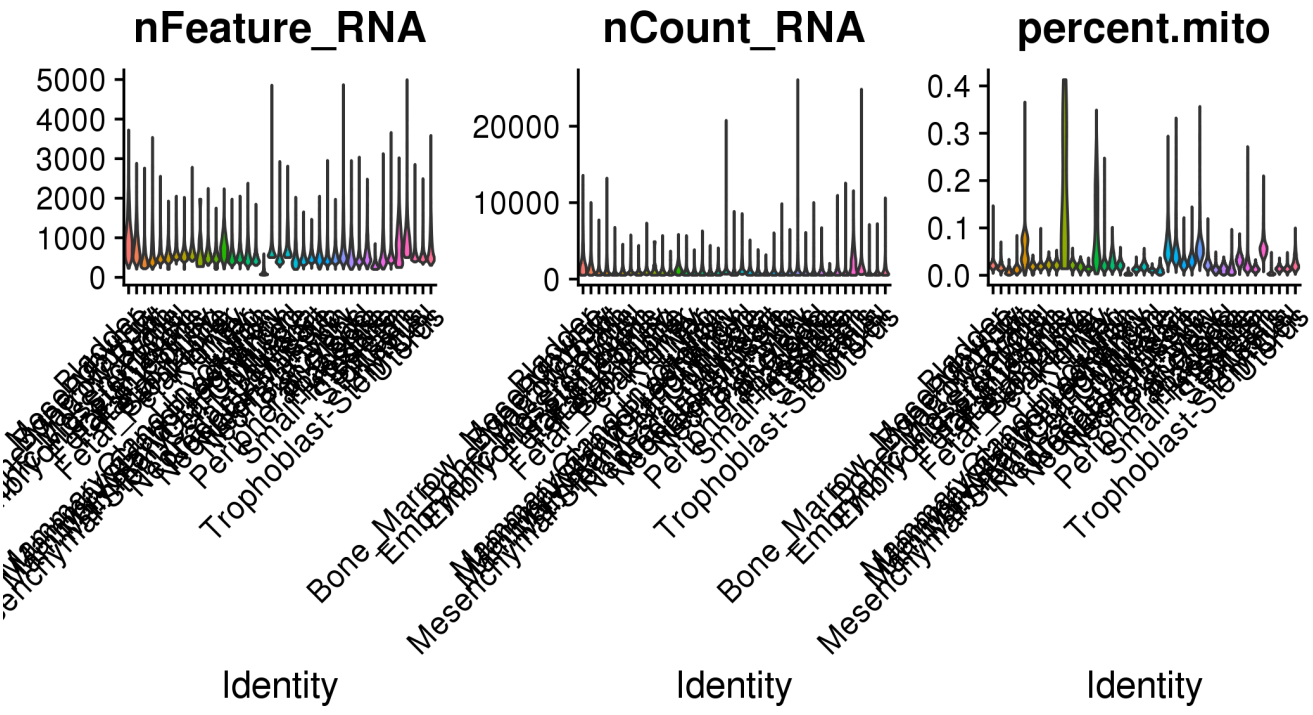


Figure 2: QC: violin plots

The dataset was filtered to remove (1) cells expressing fewer than 0 genes per cell and (2) cells with a fraction of mitochondrial reads greater than 1. Genes expressed in less than 3 cells were removed from the analysis.

3 Removal of unwanted variation

3.1 Removal of unwanted variation (data normalisation)

- The type of normalization applied was: log-normalization.
- A linear model was used to regress out the latent variables [percent.mito] before further analysis.
- The type of cell cycle regression applied was: none.

3.2 Summary statistics

	x
no_cells	242533.00
qc_min_gene_threshold	0.00
qc_min_percent_mito_threshold	0.00
qc_max_percent_mito_threshold	1.00
no_cells_after_qc	239347.00

Table 1: Run statistics

	input	after_qc_filters
Bladder	2746	2746
Bone_Marrow_Mesenchyme	7365	7364
Bone-Marrow	9049	8796
Bone-Marrow_c-kit	26483	26406
Brain	4038	4033
Embryonic-Mesenchyme	2771	2768
Embryonic-Stem-Cell	9991	9991
Fetal_Brain	4369	4368
Fetal_Intestine	6076	6074
Fetal_Kidney	11	11
Fetal_Lung	6453	6450
Fetal_Stomache	6192	6191
Fetal-Liver	2699	2696
Kidney	4673	4673
Liver	4685	4655
Lung	6940	6940
MammaryGland.Involution	4821	4820
MammaryGland.Lactation	13538	11831
MammaryGland.Pregnancy	4909	4908
MammaryGland.Virgin	5380	5379
Mesenchymal-Stem-Cell-Cultured	7319	7318
Muscle	1102	1078
Neonatal-Calvaria	7964	7964
Neonatal-Heart	3948	3948
Neonatal-Muscle	4873	4872
Neonatal-Rib	6262	6261
Neonatal-Skin	3392	3392
Ovary	4363	4362
Pancreas	3610	3583
Peripheral_Blood	7095	7029
Placenta	4346	4257
Prostate	2505	2505
Small-Intestine	6684	6683
Spleen	1970	1968
Stomach	2389	2389
Testis	14005	13125
Thymus	4289	4289
Trophoblast-Stem-Cell	19489	19485
Uterus	3739	3739

Table 2: Numbers of cells

4 Dimension reduction

4.1 Scree (elbow) plot

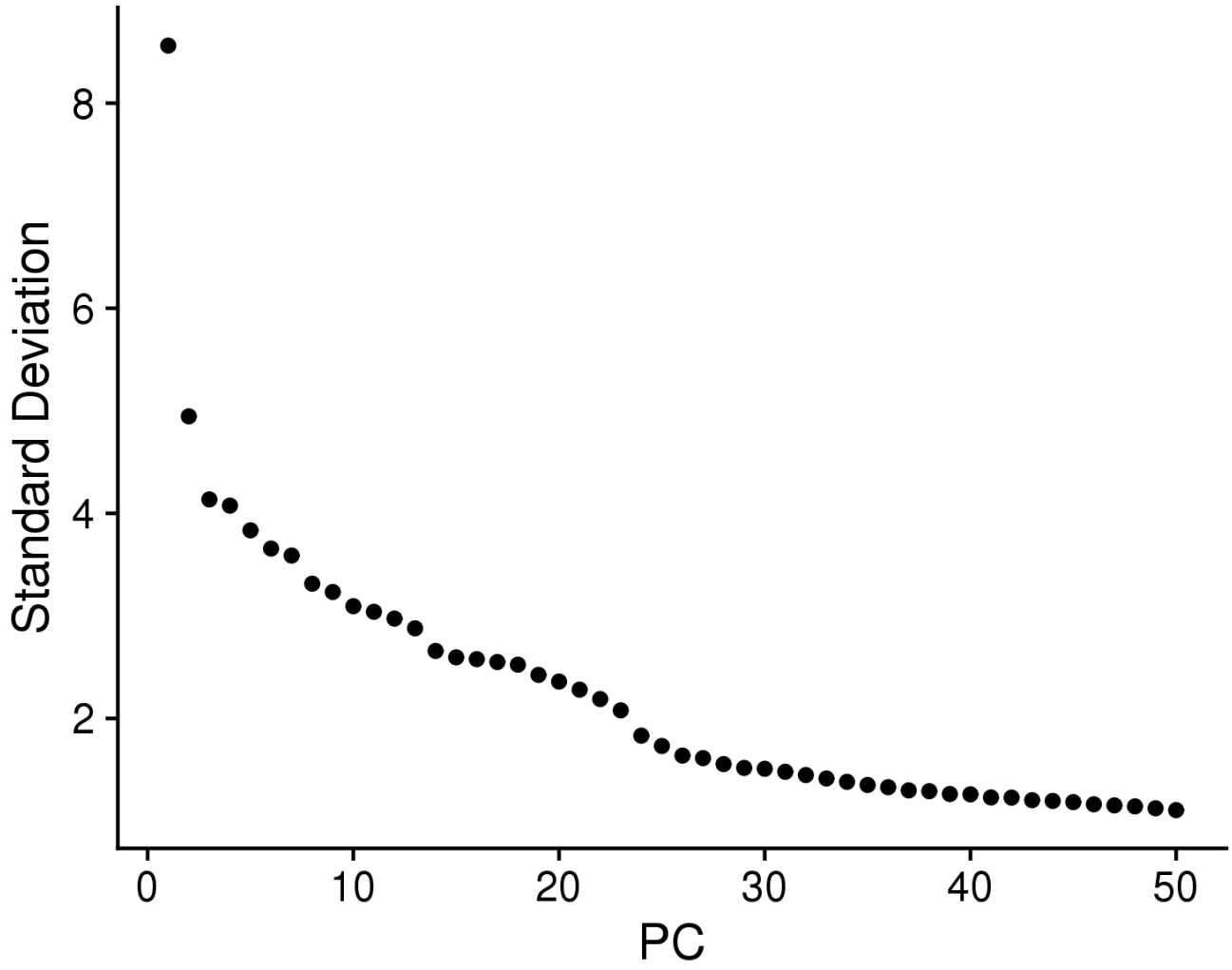


Figure 3: Scree plot showing proportion of variance explained by each PCA component

4.2 Component heatmaps

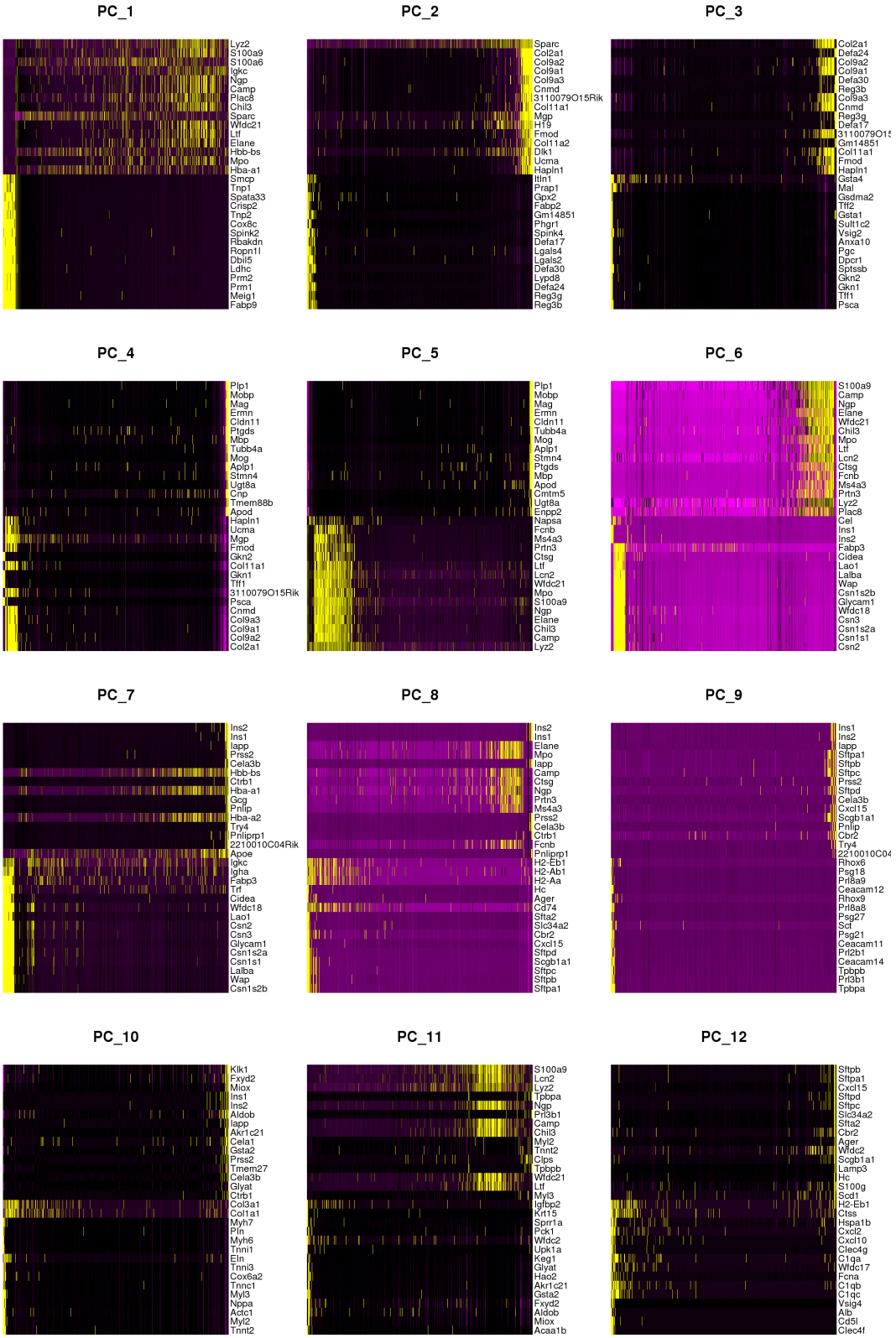


Figure 4: Heatmaps of the top genes for each PCA component

5 Visualisation of clusters and factors of interest

5.1 umap.mindist_0 plot colored by cluster_id

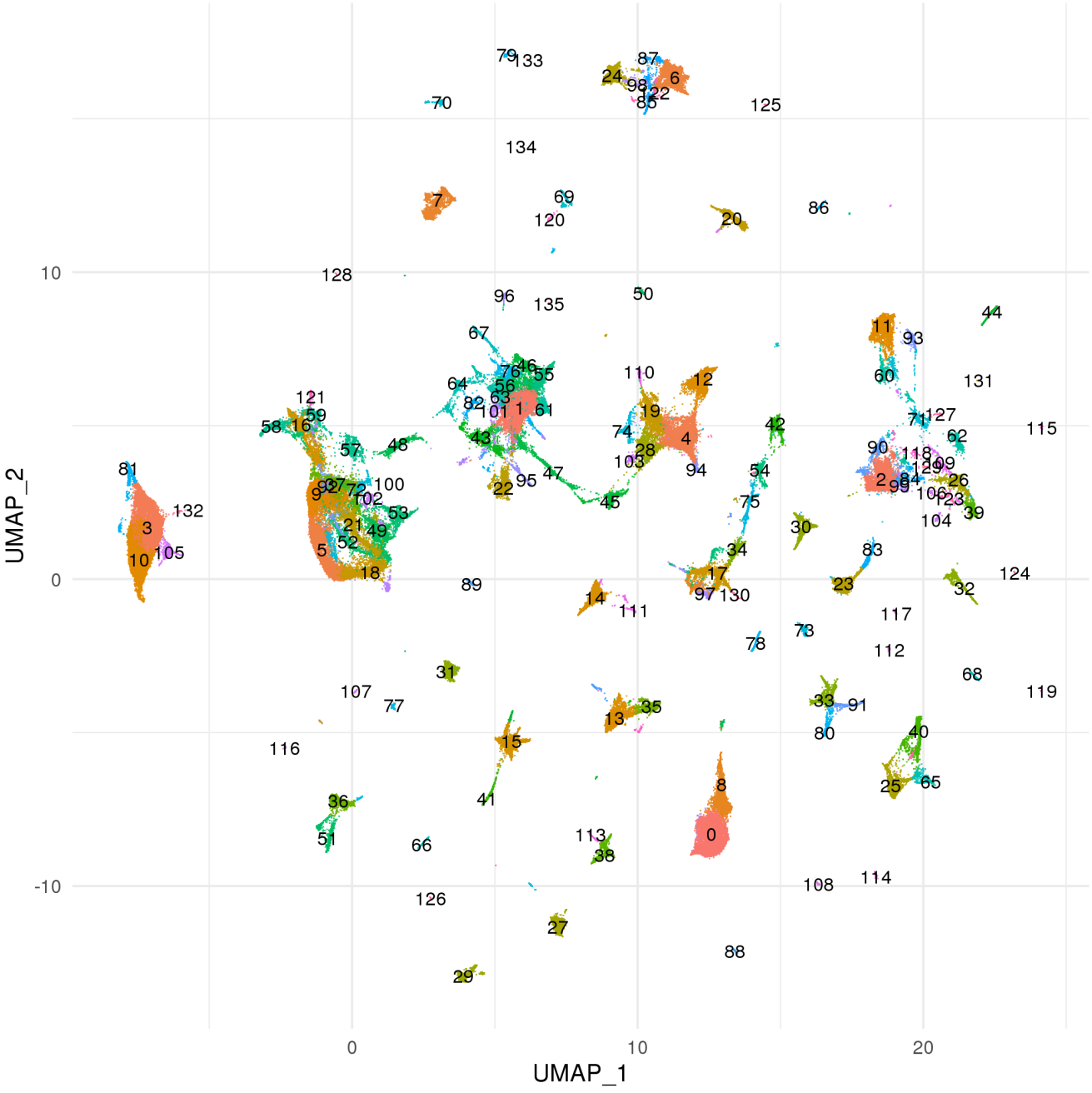


Figure 5: umap.mindist_0 plot colored by cluster_id

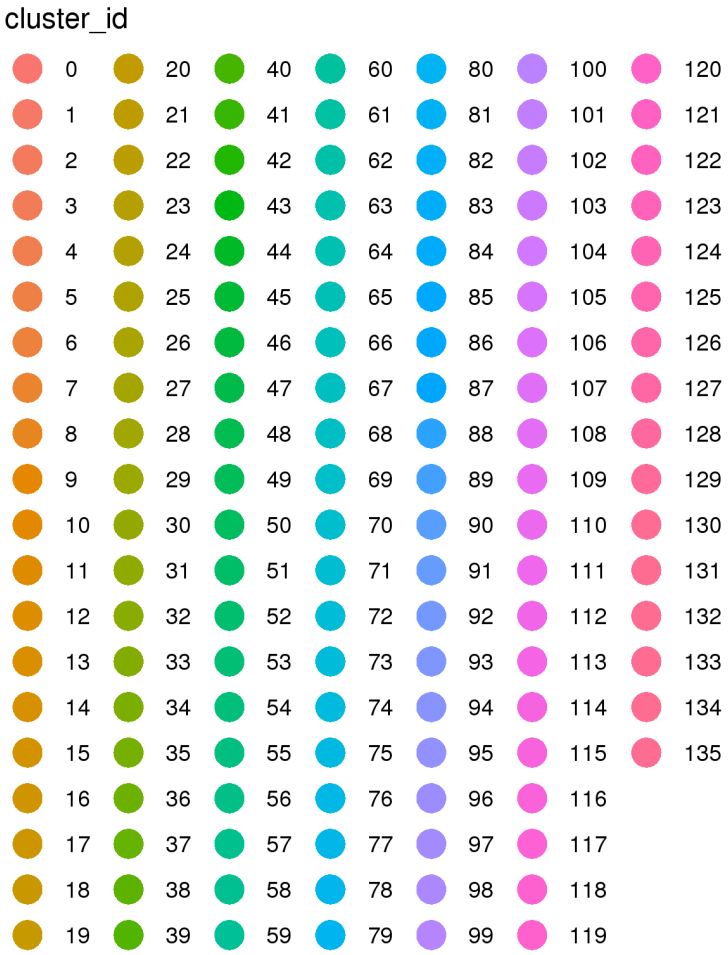


Figure 6: umap.mindist_0 plot colored by cluster_id plot legend

5.2 umap.mindist_0.1 plot colored by cluster_id

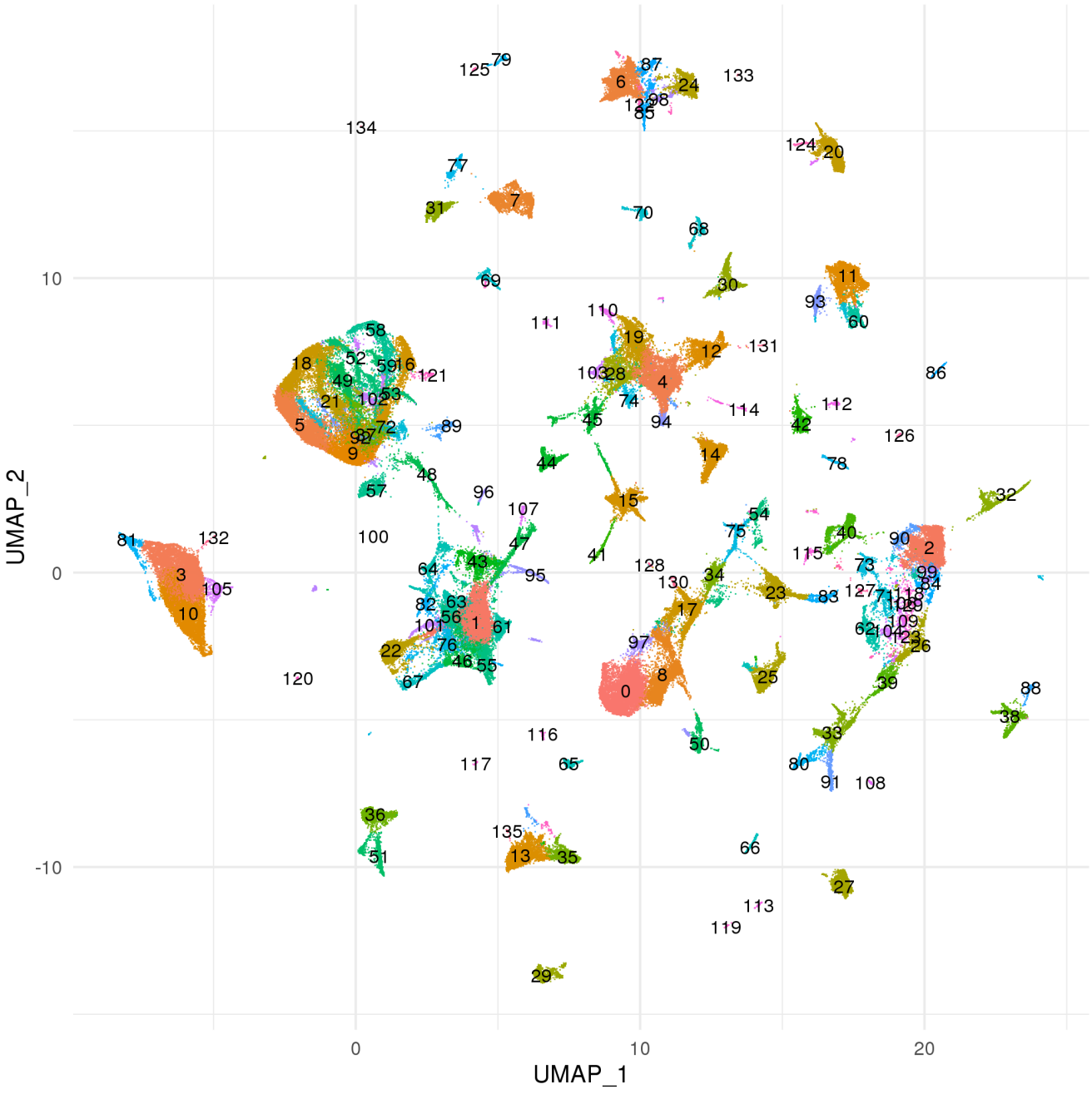


Figure 7: umap.mindist_0.1 plot colored by cluster_id

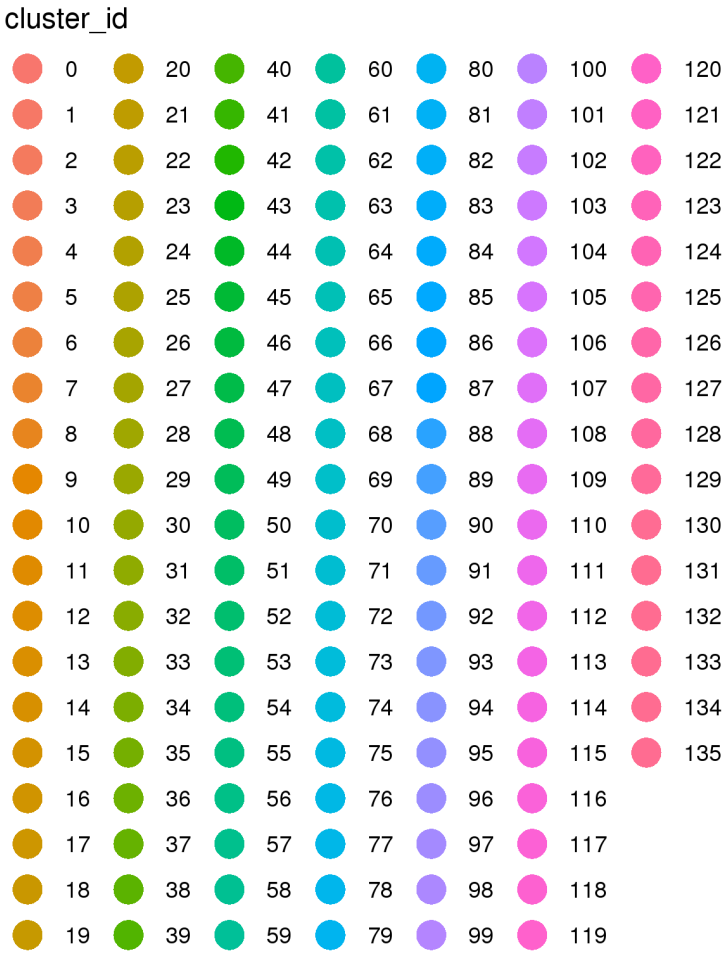


Figure 8: umap.mindist_0.1 plot colored by cluster_id plot legend

5.3 umap.mindist_0.3 plot colored by cluster_id

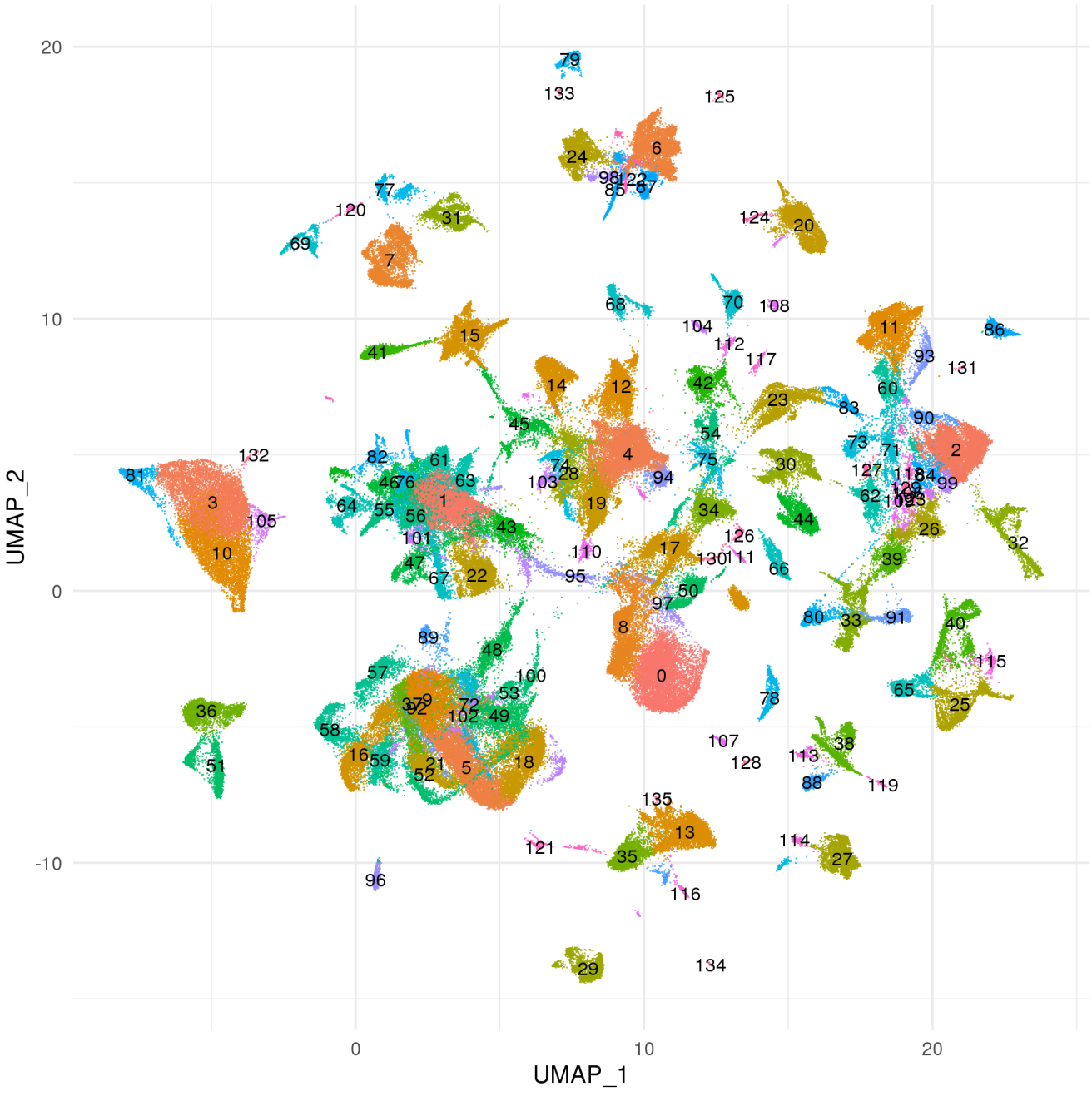


Figure 9: umap.mindist_0.3 plot colored by cluster_id

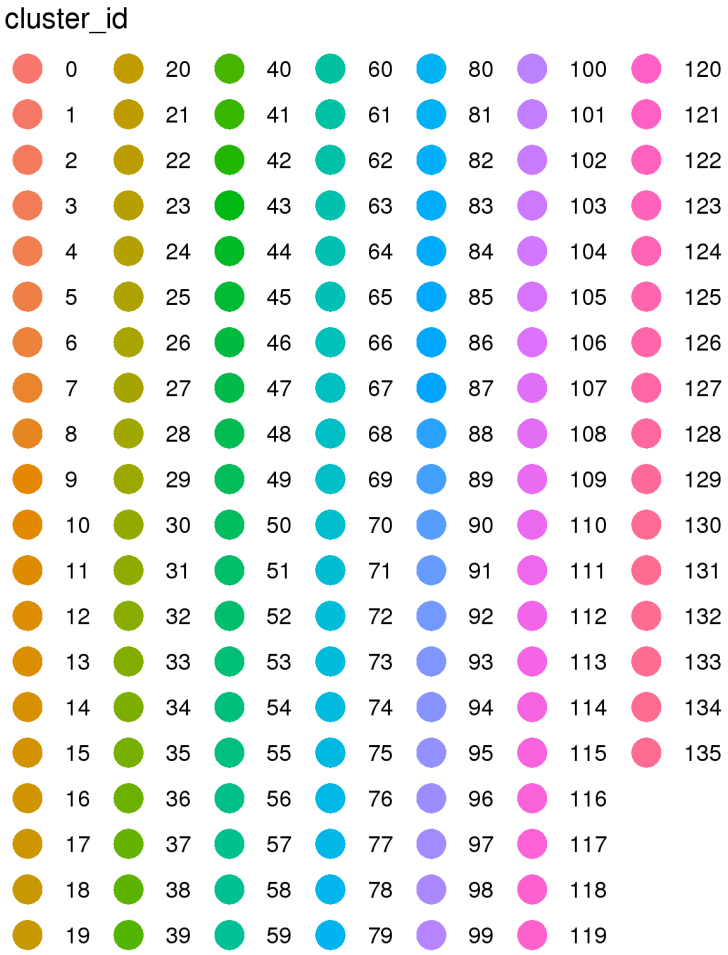


Figure 10: umap.mindist_0.3 plot colored by cluster_id plot legend

5.4 umap.mindist_0.5 plot colored by cluster_id

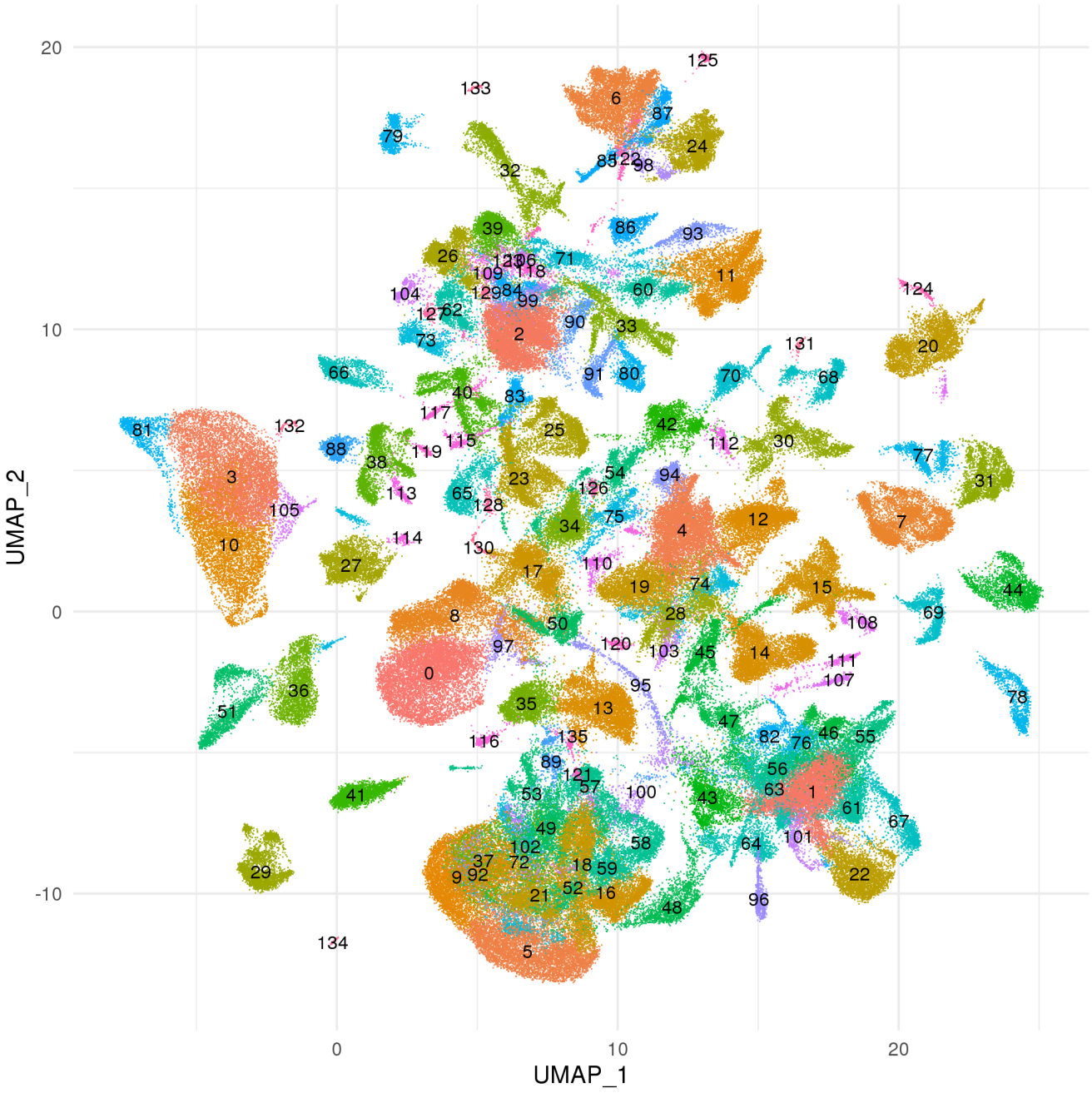


Figure 11: umap.mindist_0.5 plot colored by cluster_id

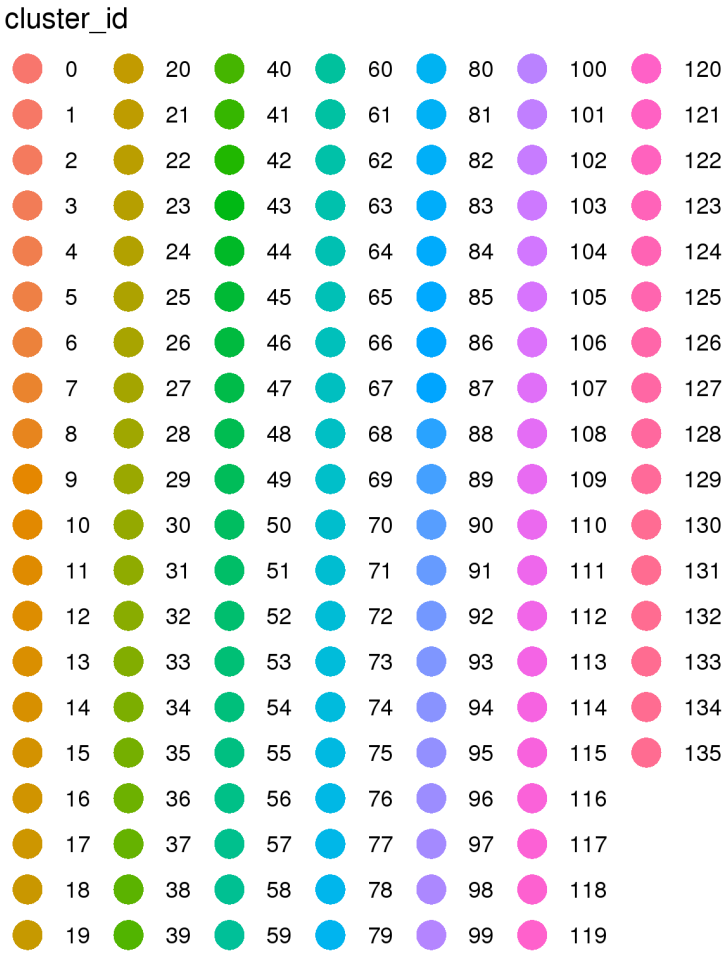


Figure 12: umap.mindist_0.5 plot colored by cluster_id plot legend

5.5 umap.mindist_0.7 plot colored by cluster_id

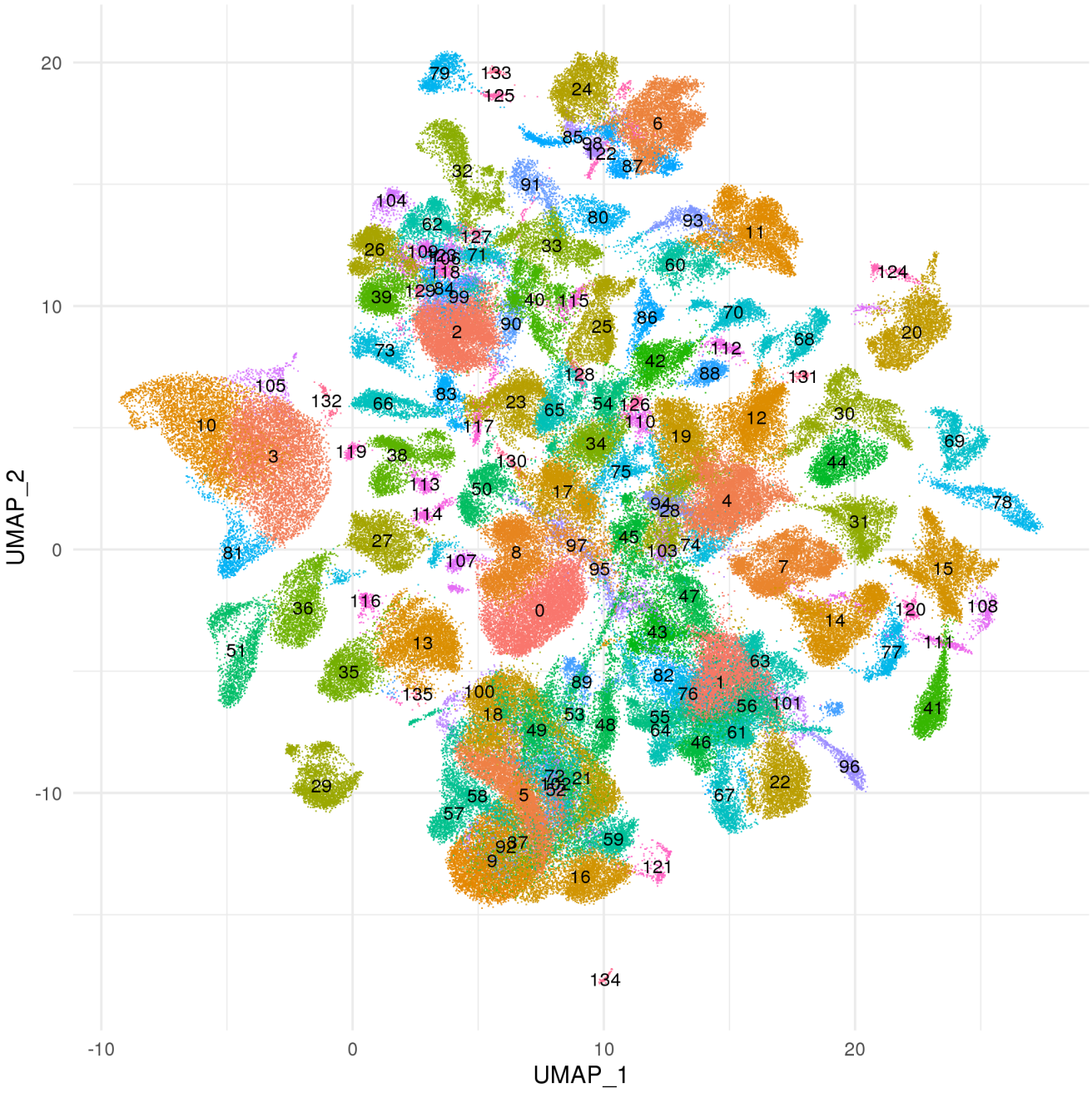


Figure 13: umap.mindist_0.7 plot colored by cluster_id

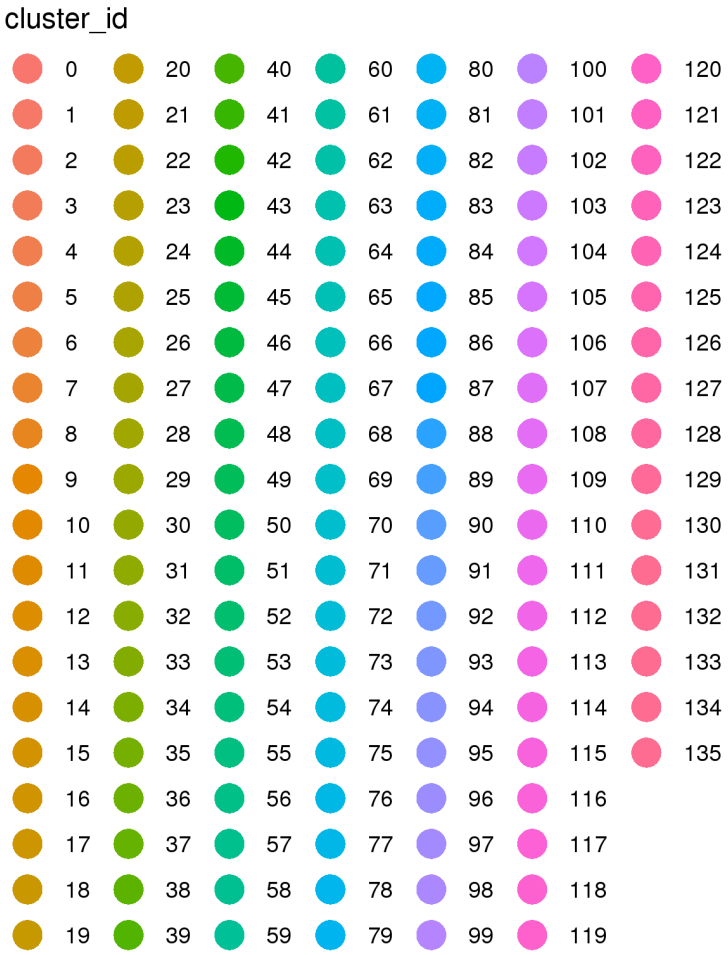


Figure 14: umap.mindist_0.7 plot colored by cluster_id plot legend

5.6 umap plot colored by nCount_RNA

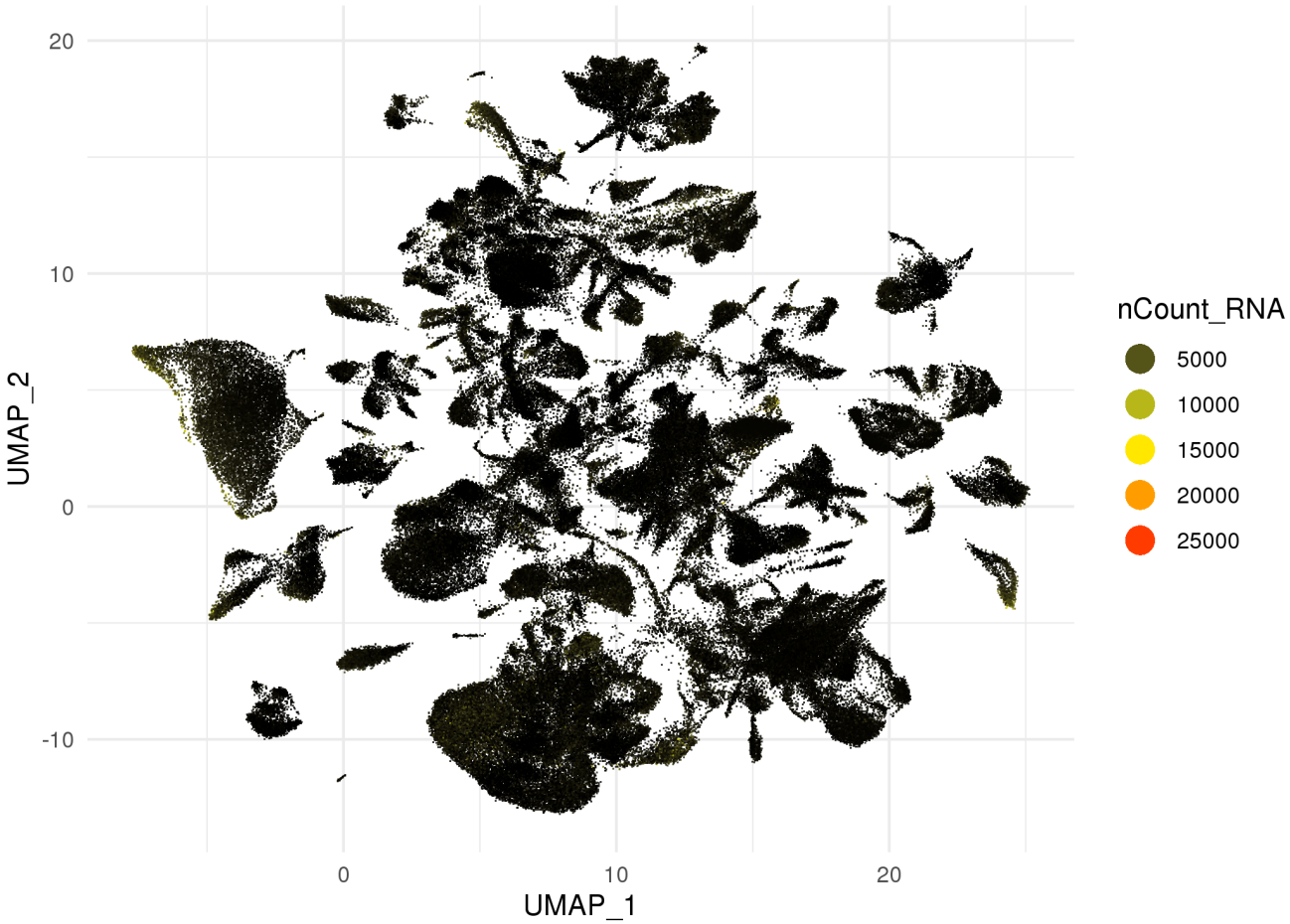


Figure 15: umap plot colored by nCount_RNA

5.7 umap plot colored by percent.mito

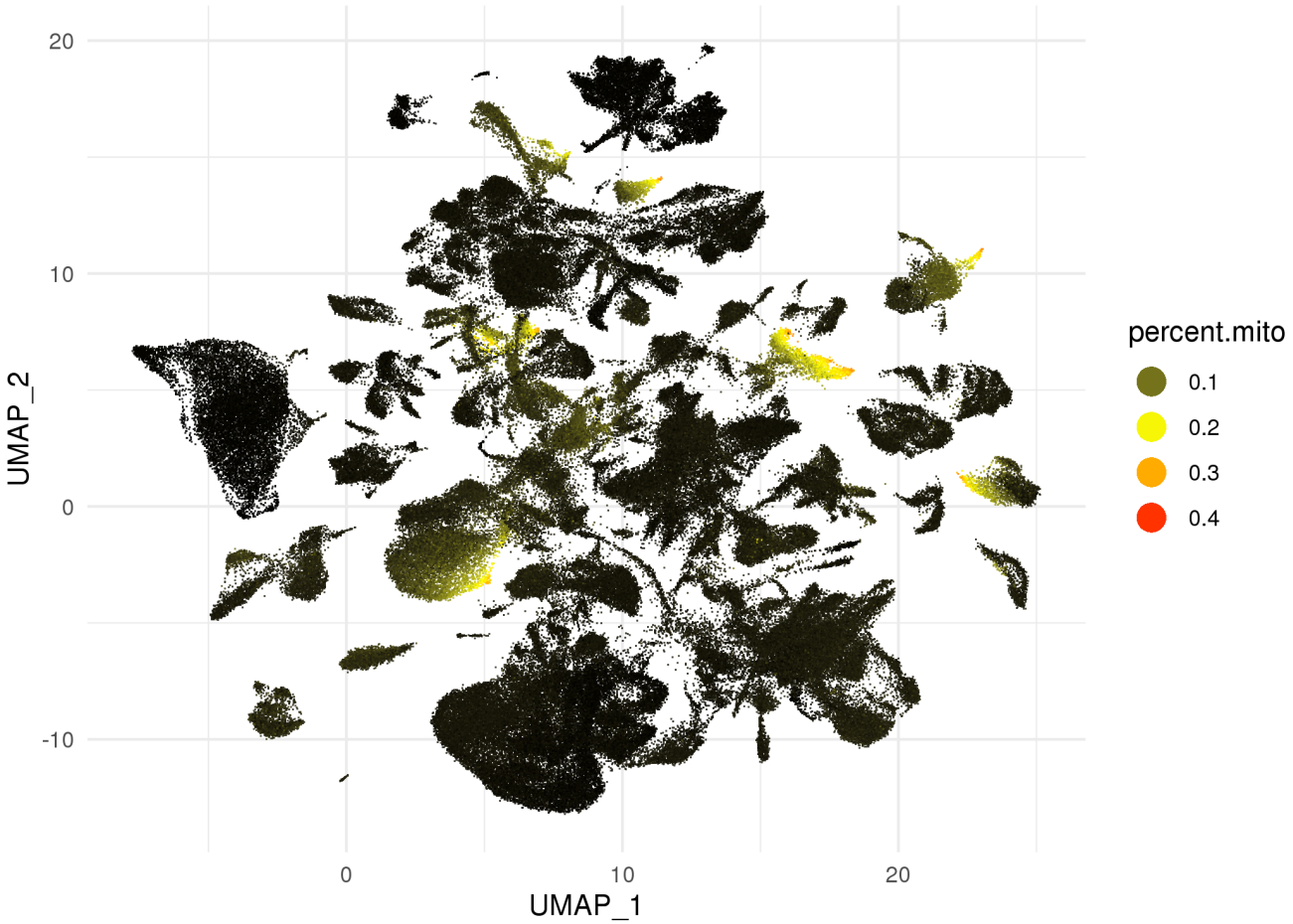


Figure 16: umap plot colored by percent.mito

5.8 umap plot colored by Tissue

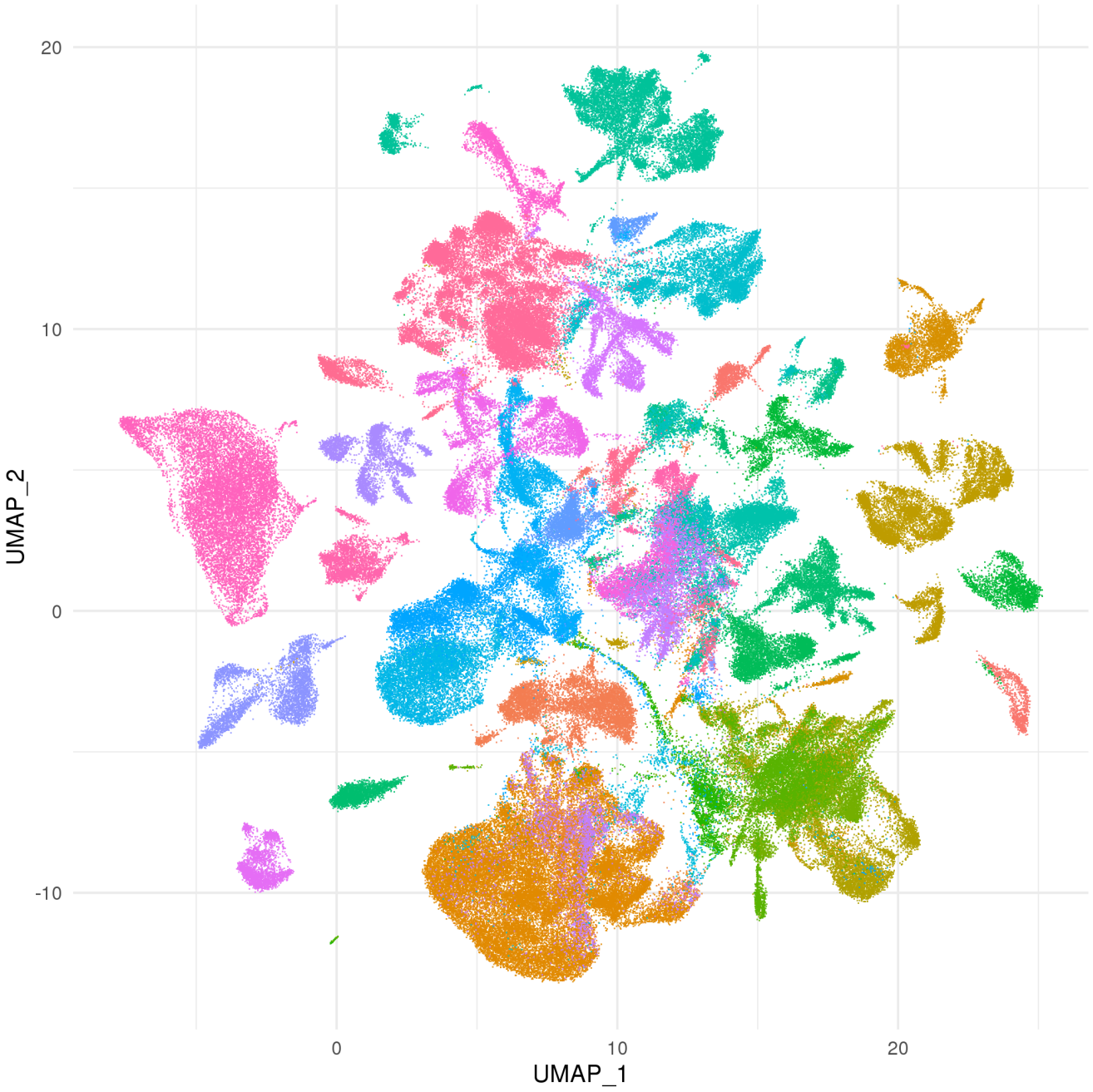


Figure 17: umap plot colored by Tissue

Figure 18: umap plot colored by Tissue plot legend

6 Plots of summary statistics

Plots of summary statistics (e.g. cell number) by factor of interest (e.g. cluster)

6.1 Cells by cluster

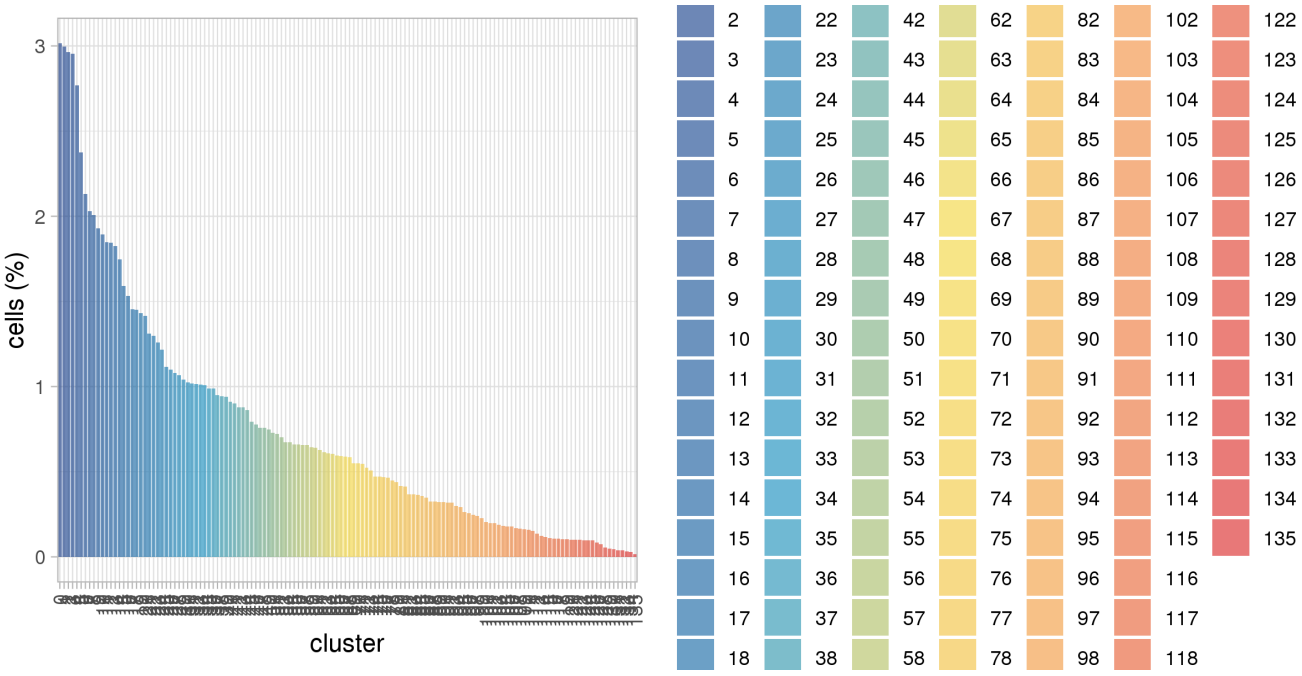


Figure 19: Cells by cluster

6.2 Number of genes per cell per cluster

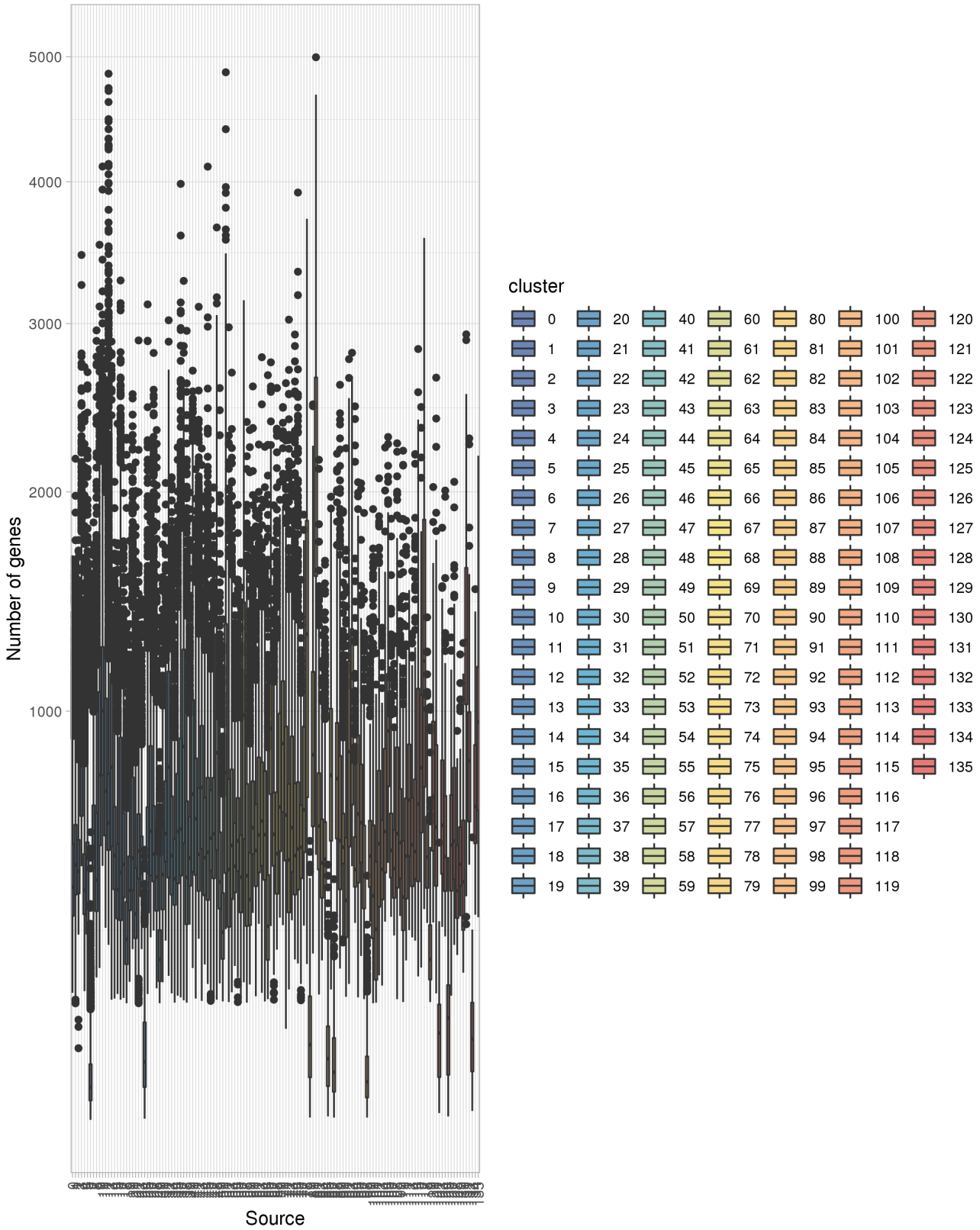


Figure 20: Number of genes per cell per cluster

6.3 Number of umi per cell per cluster

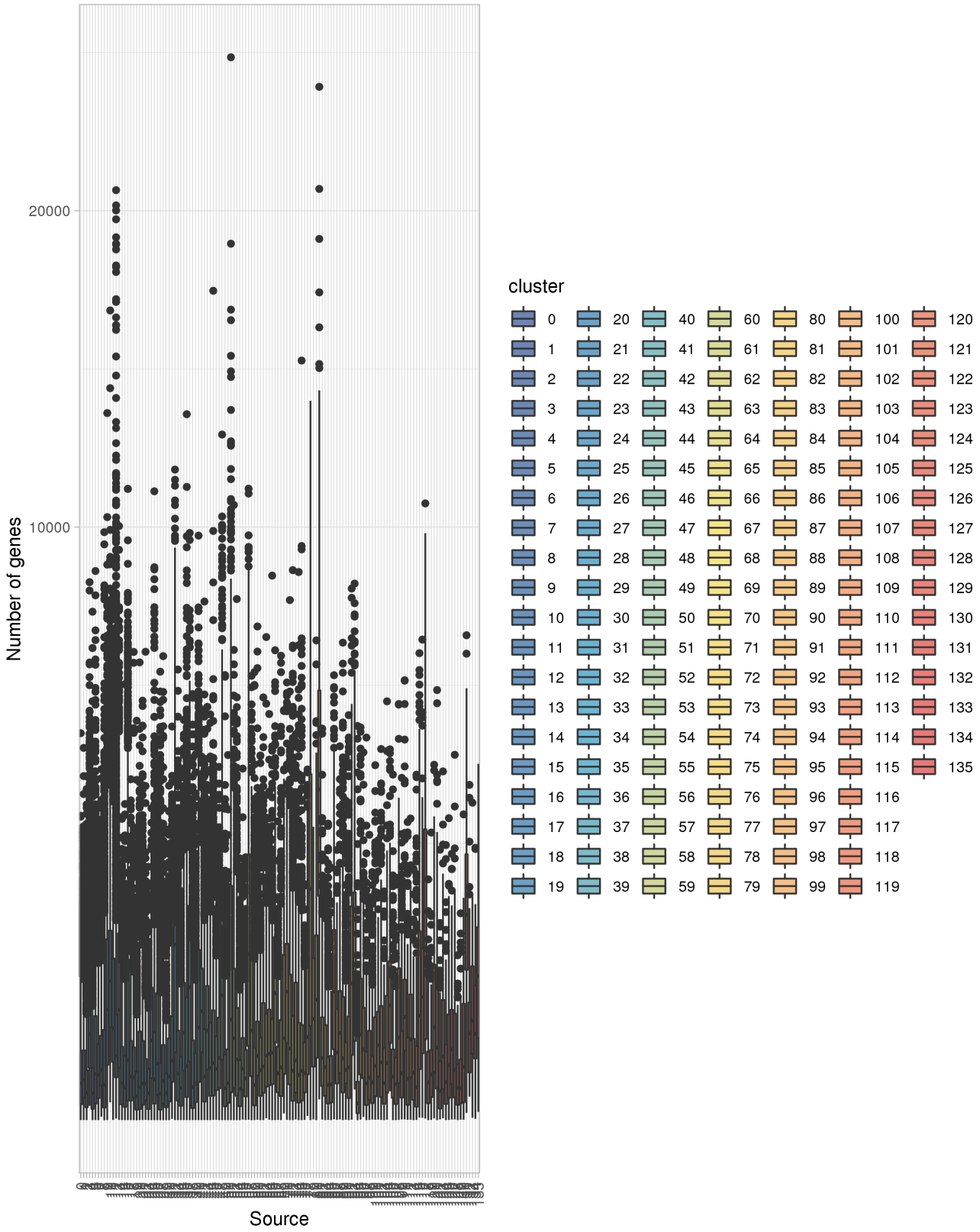


Figure 21: Number of umi per cell per cluster

7 Cluster dissimilarity

7.1 Dissimilarity by gene expression

The distances between the clusters was assessed using the "BuildClusterTree" function in the Seurat package, which "constructs a phylogenetic tree relating the "average" cell from each identity class".

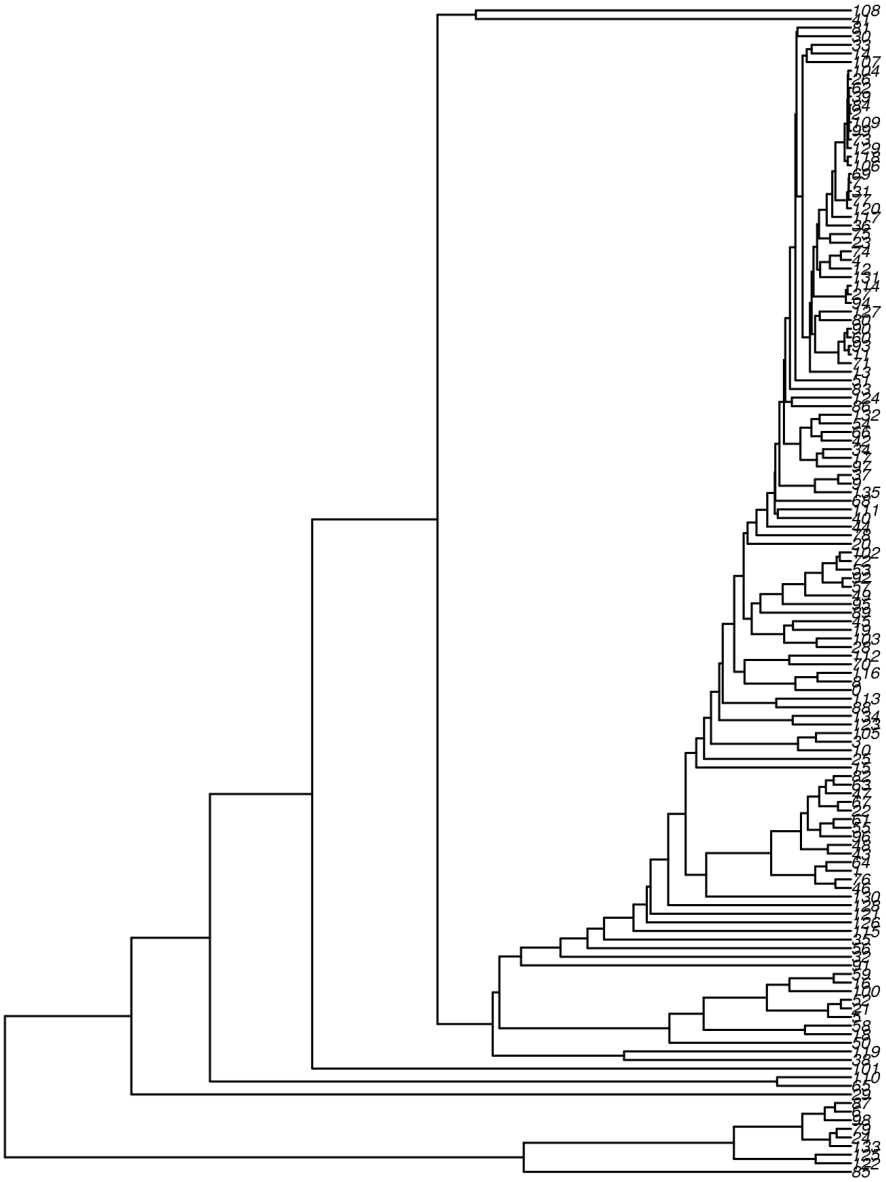


Figure 22: Visualisation of inter-cluster distances (cluster average, gene-based)

8 Identification of cluster marker genes

Cluster marker genes were identified using the Seurat FindMarkers routine and the wilcox test. A summary table containing all of the significant markers for all of the clusters (based on BH adjusted p value) is available separately.

Key parameters are:

- Differential expression methods: wilcox
- Testing limited to genes with a log fold change of > 0.25
- Testing limited to genes detected in a minimum fraction of 0.1 of cells
- Conservation factor applied: None

9 Top cluster marker genes

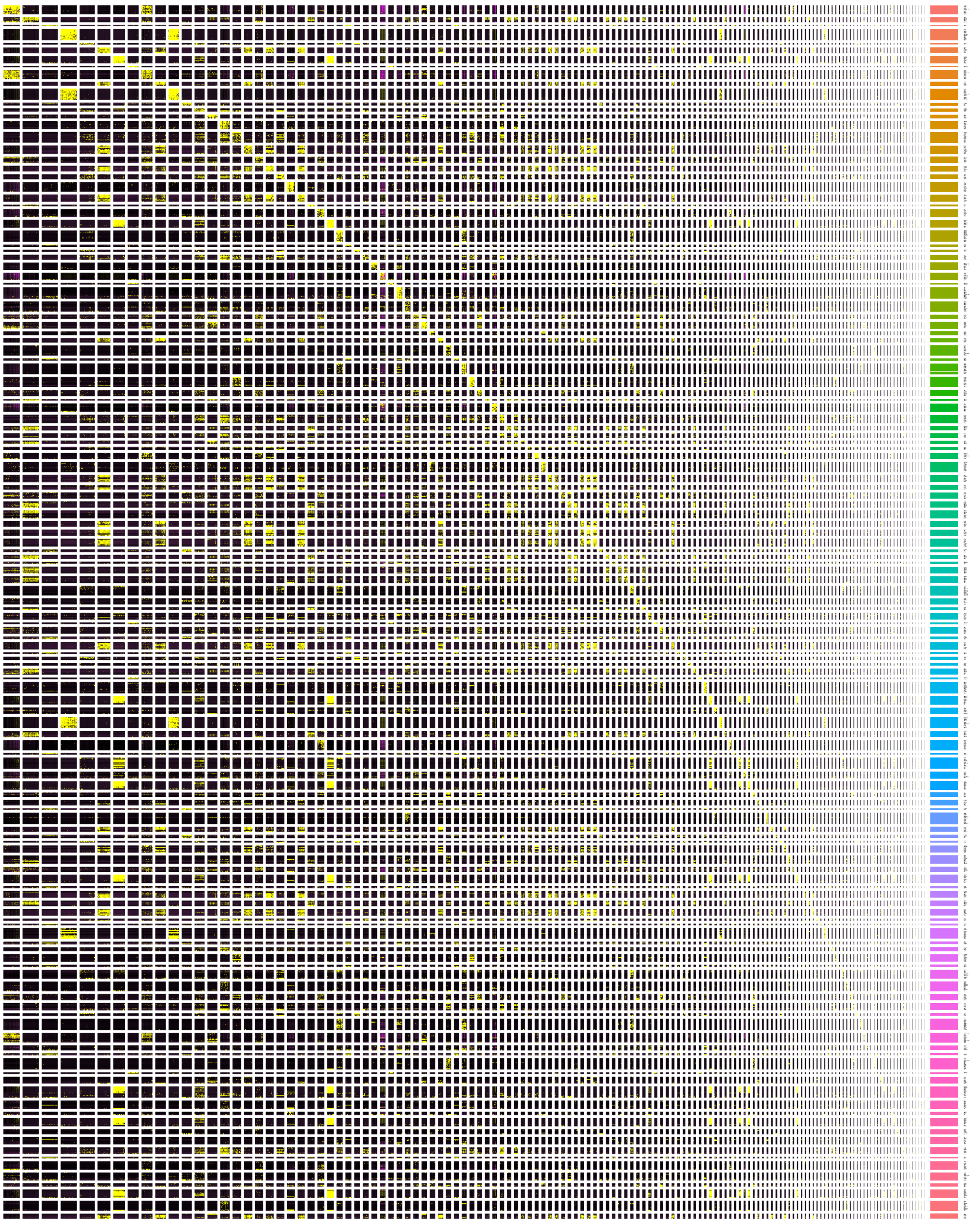


Figure 23: Heatmap of the top cluster-specific genes (based on differential expression analysis)