Final Project OMICS
Chairetaki Antonia
December 2025

Seurat Guided Clustering Tutorial

Figure 1: The following plot displaying the distribution of key quality control metrics per cell (Peripheral Blood Monoluclear cell), specifically the number of unique genes detected, total RNA molecule counts, and the percentage of mitochondrial reads.

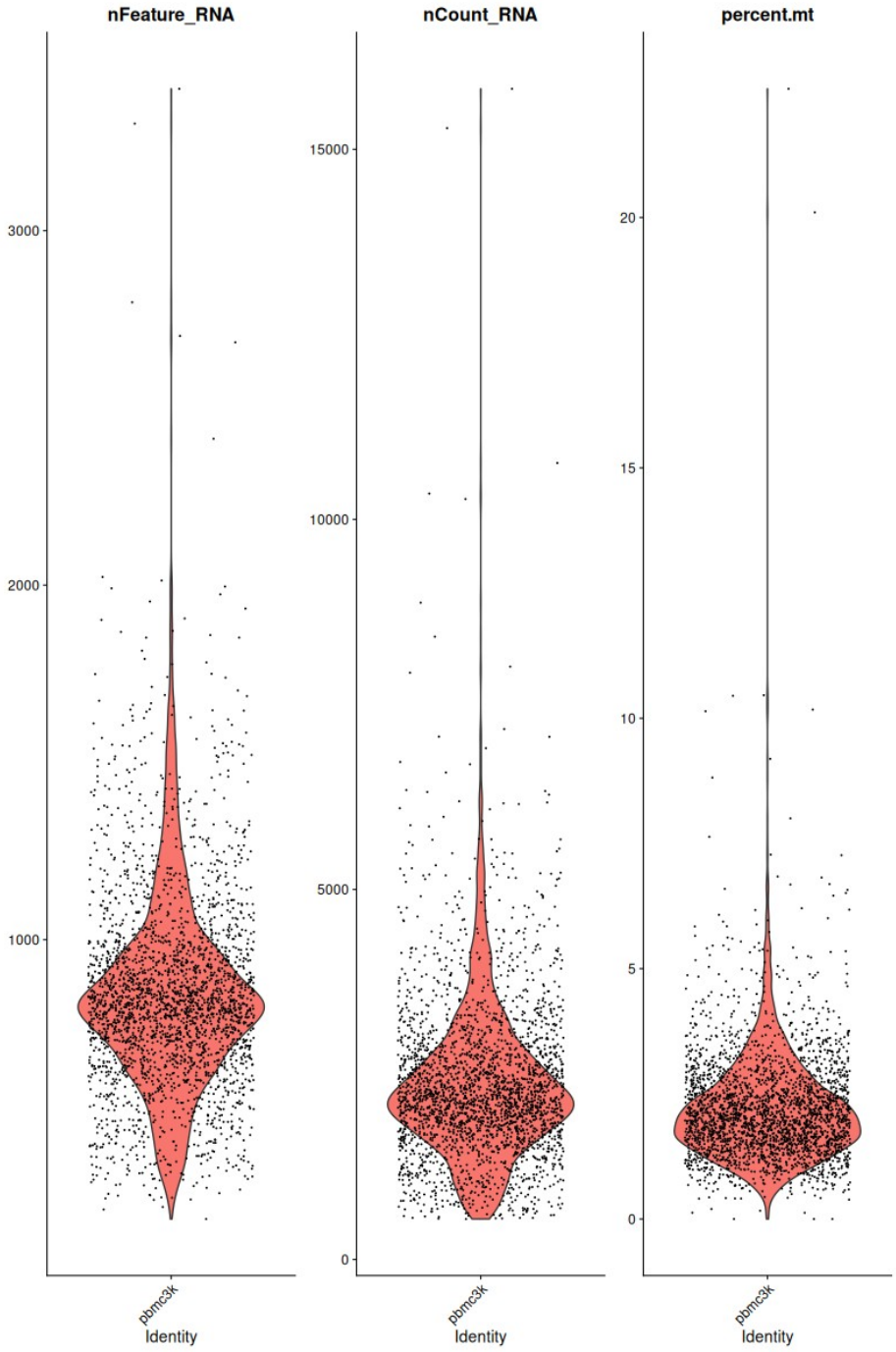


Figure 2: The scatter plots below illustrating the relationship between sequencing depth and mitochondrial percentage (left), and the correlation between sequencing depth and detected gene counts (right) used to assess cell quality prior to filtering.

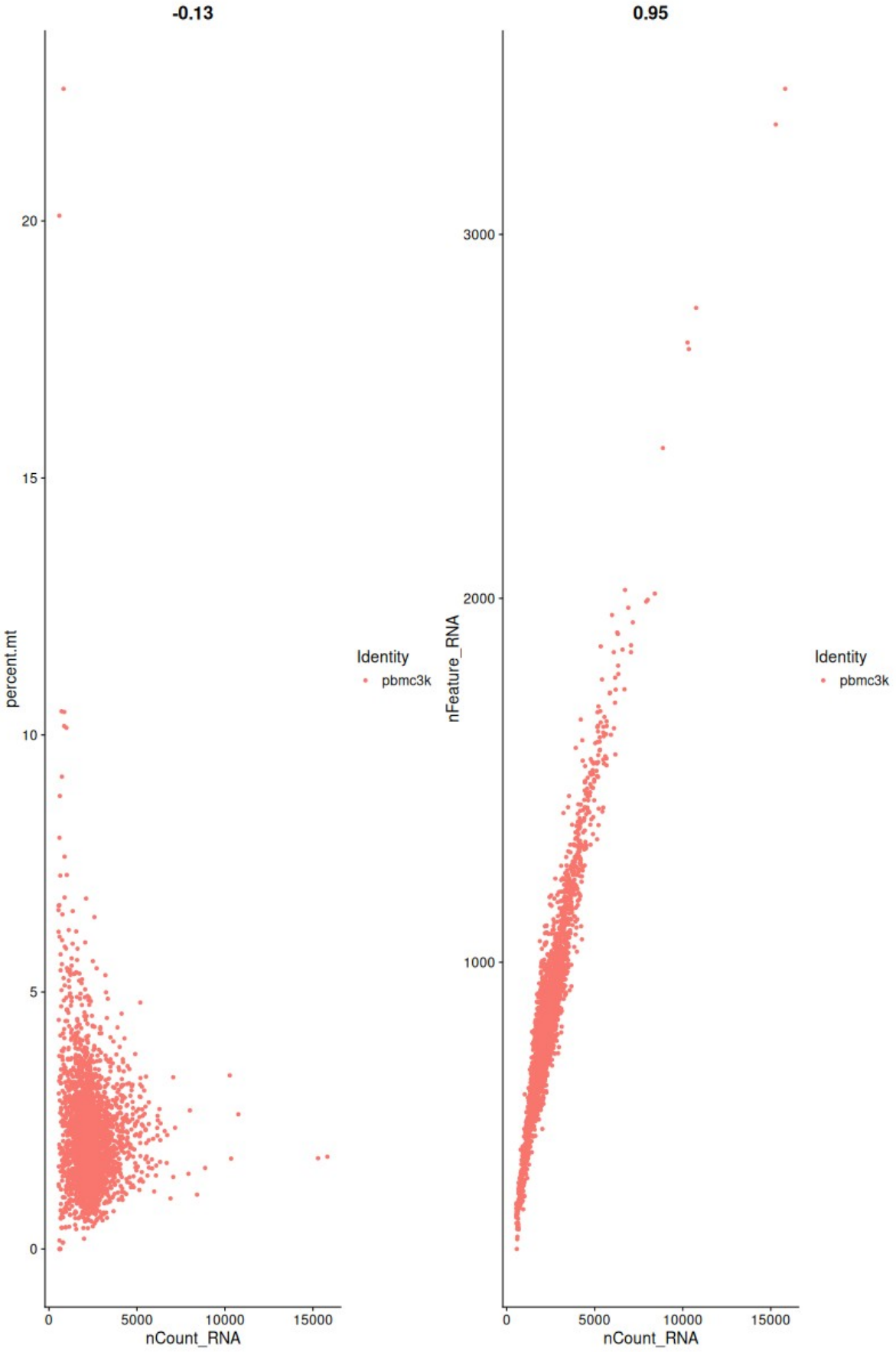


Figure 3: The scatter plots below displaying the relationship between average expression and standardized variance to identify highly variable features, with the top 10 most variable genes labeled in the right panel.

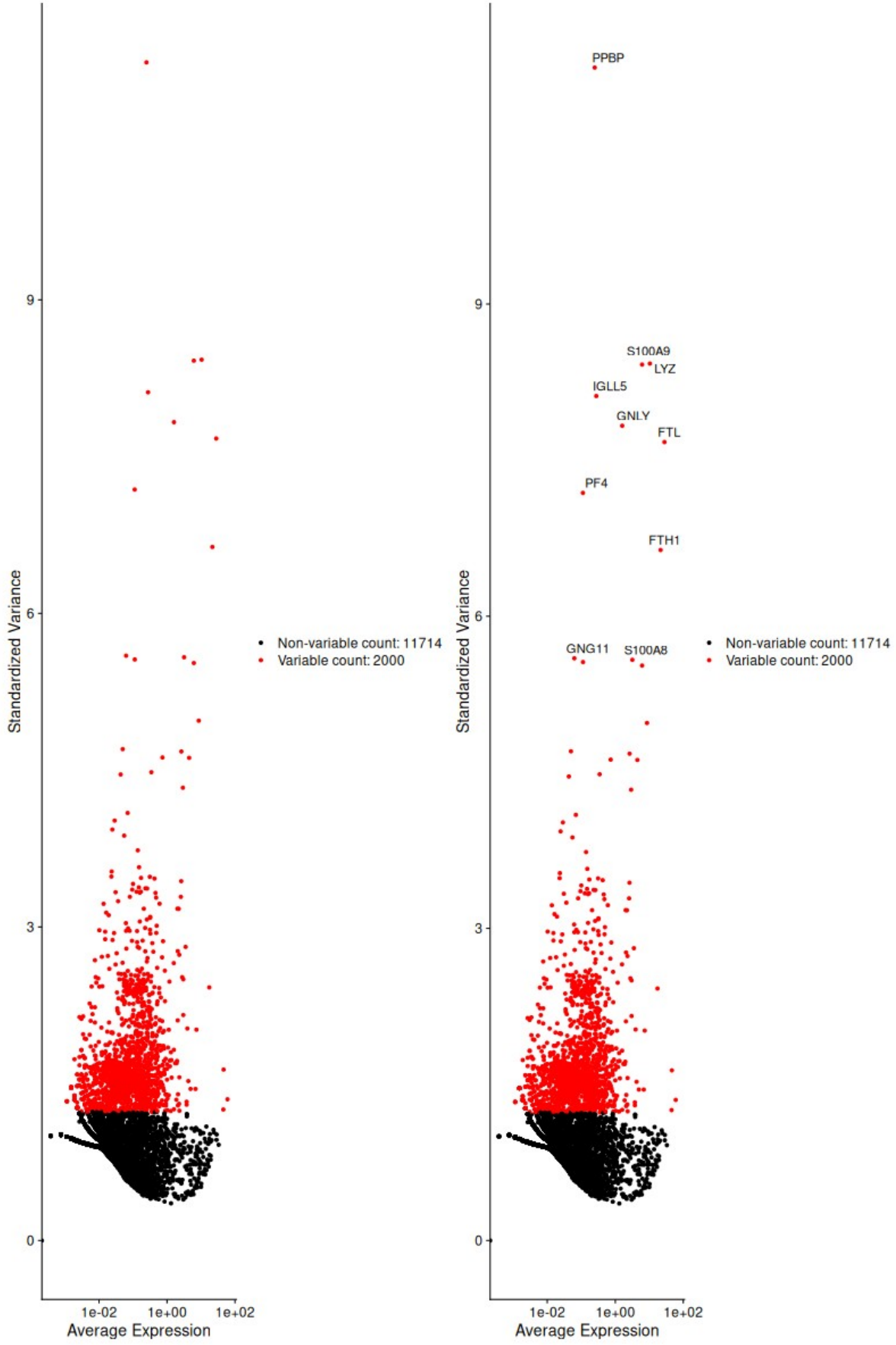


Figure 4: The plot displaying the top contributing genes for the first two principal components, ranked by their loading scores to illustrate the specific features responsible for the greatest variance in the dataset.

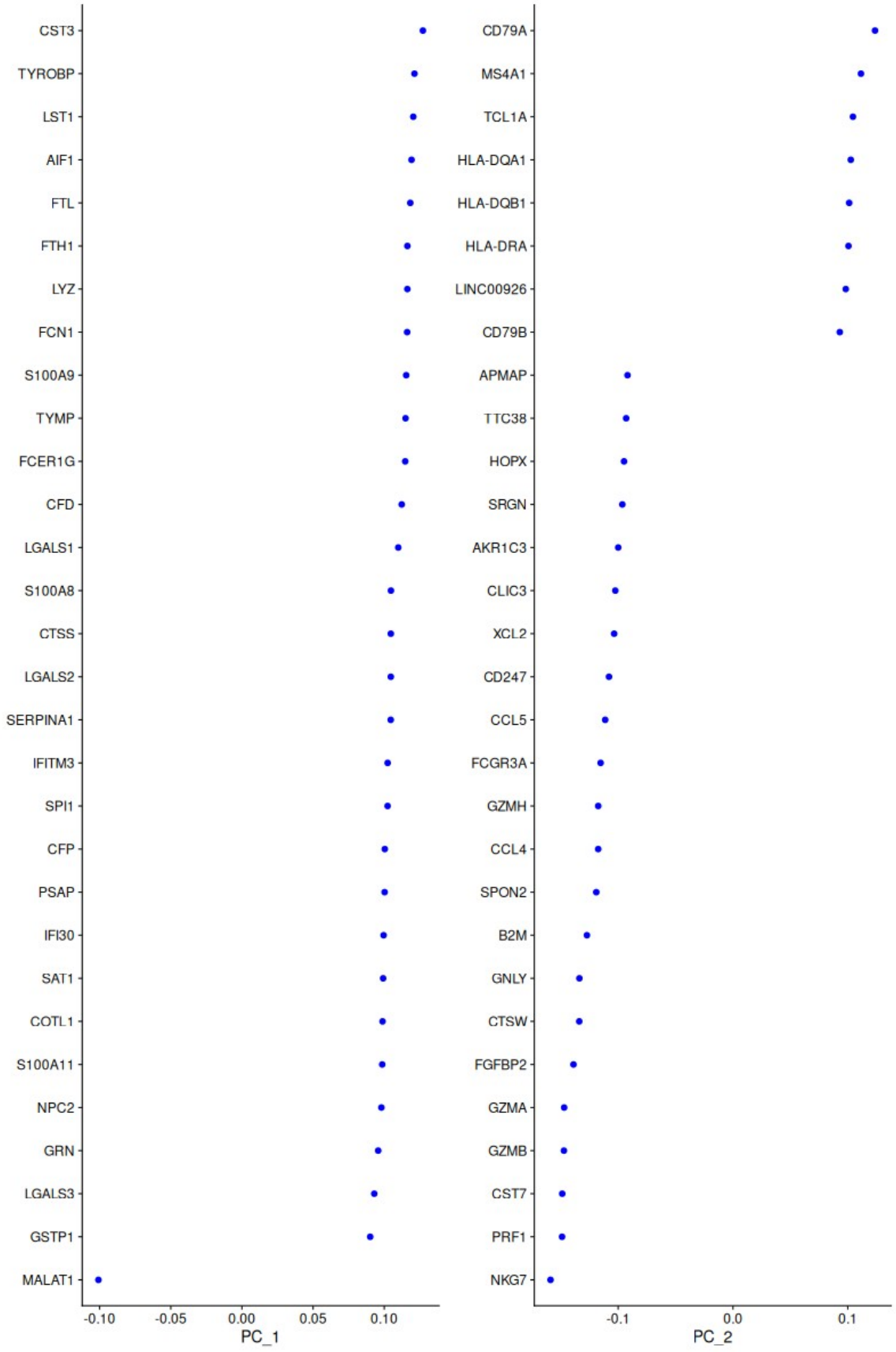


Figure 5: The plot visualizing the distribution of individual cells along the first two principal components (PC1 vs. PC2) of a Principal Component Analysis (PCA), to reveal global structure of our data distribution.

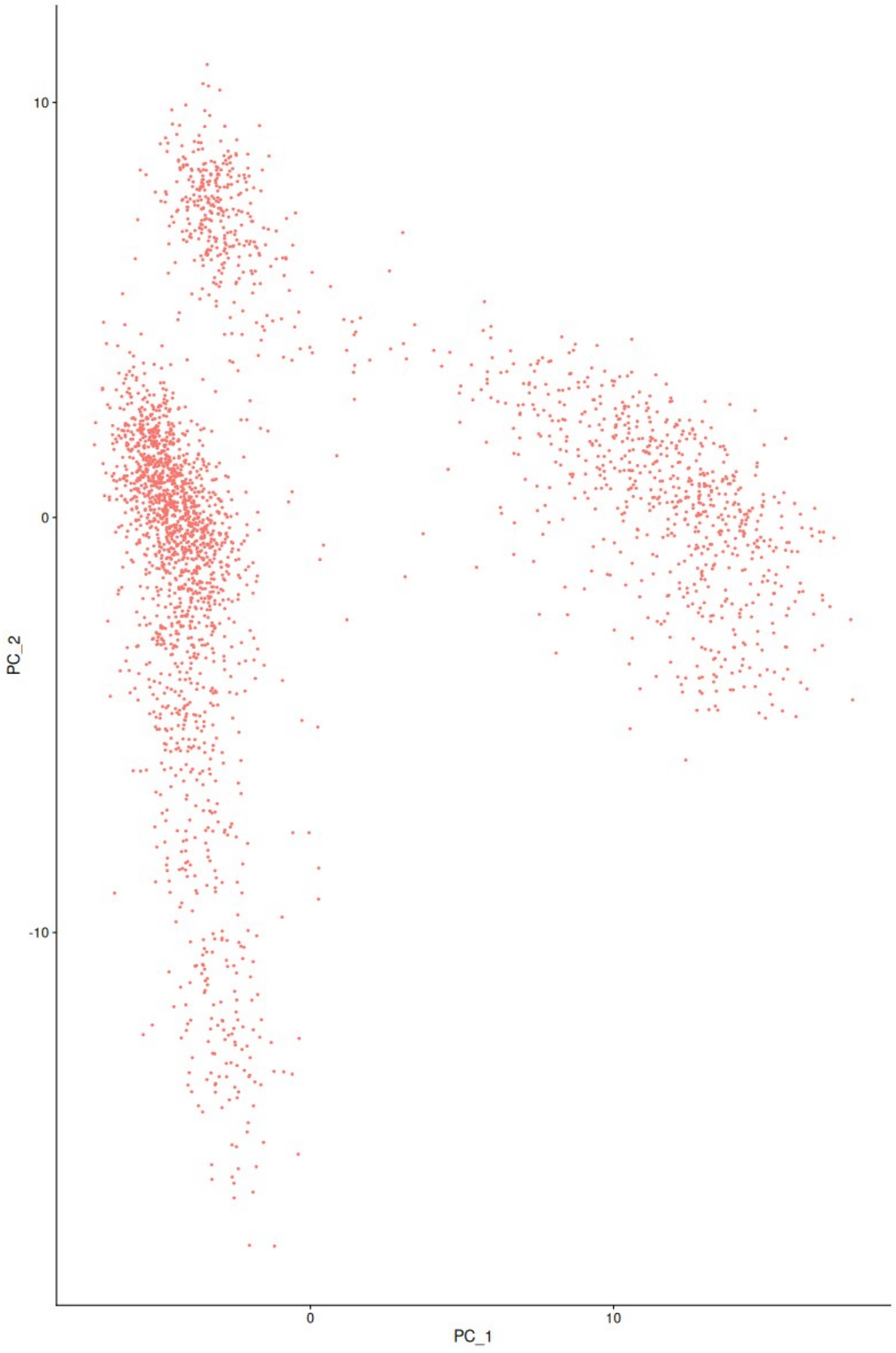


Figure 6: The following heatmap displaying the expression levels of the top contributing genes for the first principal component across 500 cells, highlighting the primary sources of heterogeneity in that dimension.

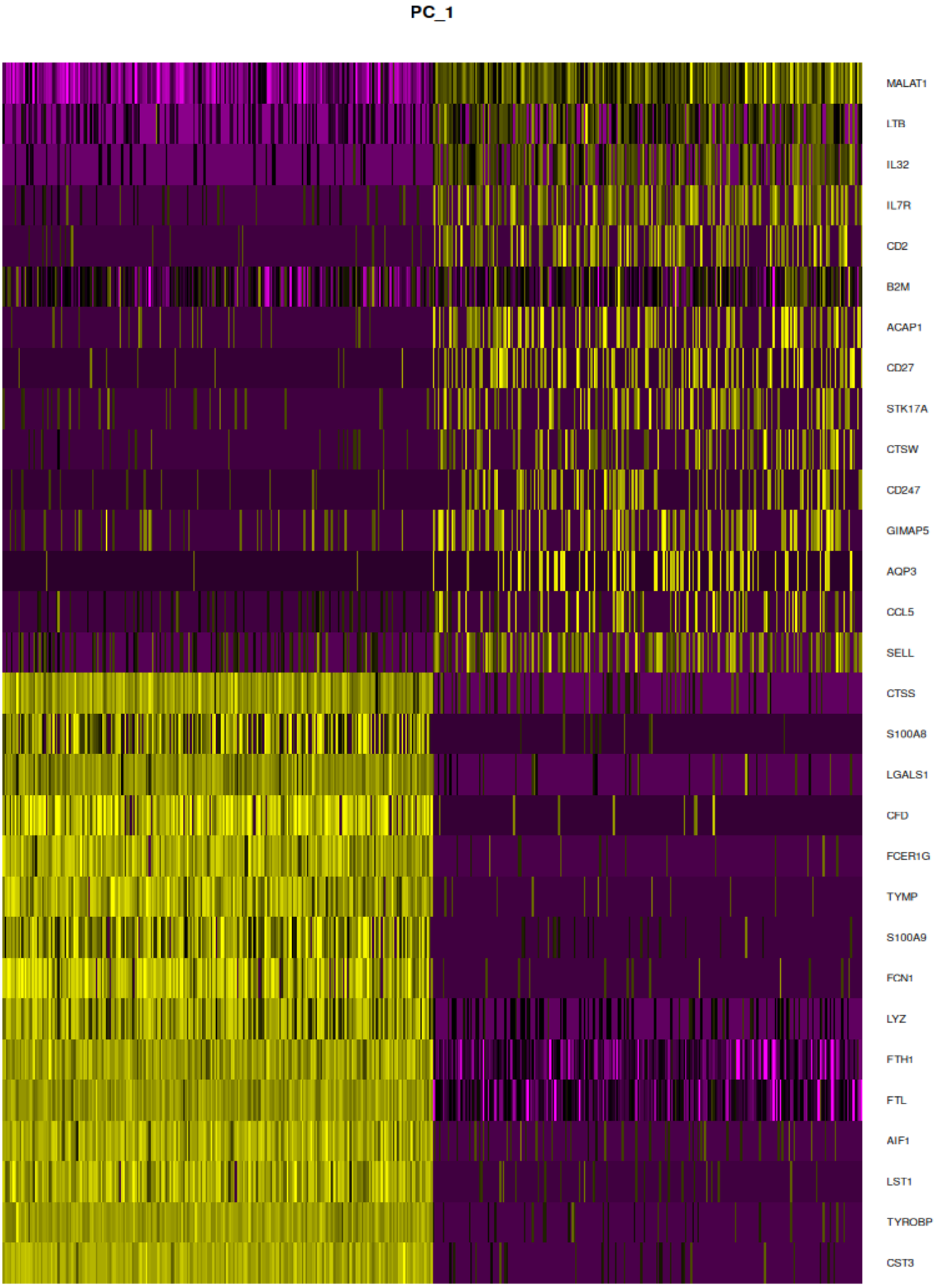


Figure 7: This figure presents 15 separate heatmaps corresponding to the first 15 principal components, where each panel displays the expression levels of the most strongly correlated genes across 500 individual cells. The distinct bands of high and low expression intensity within the earlier panels indicate that those dimensions capture clear biological signals, helping to determine where the structural variation ends and random noise begins.

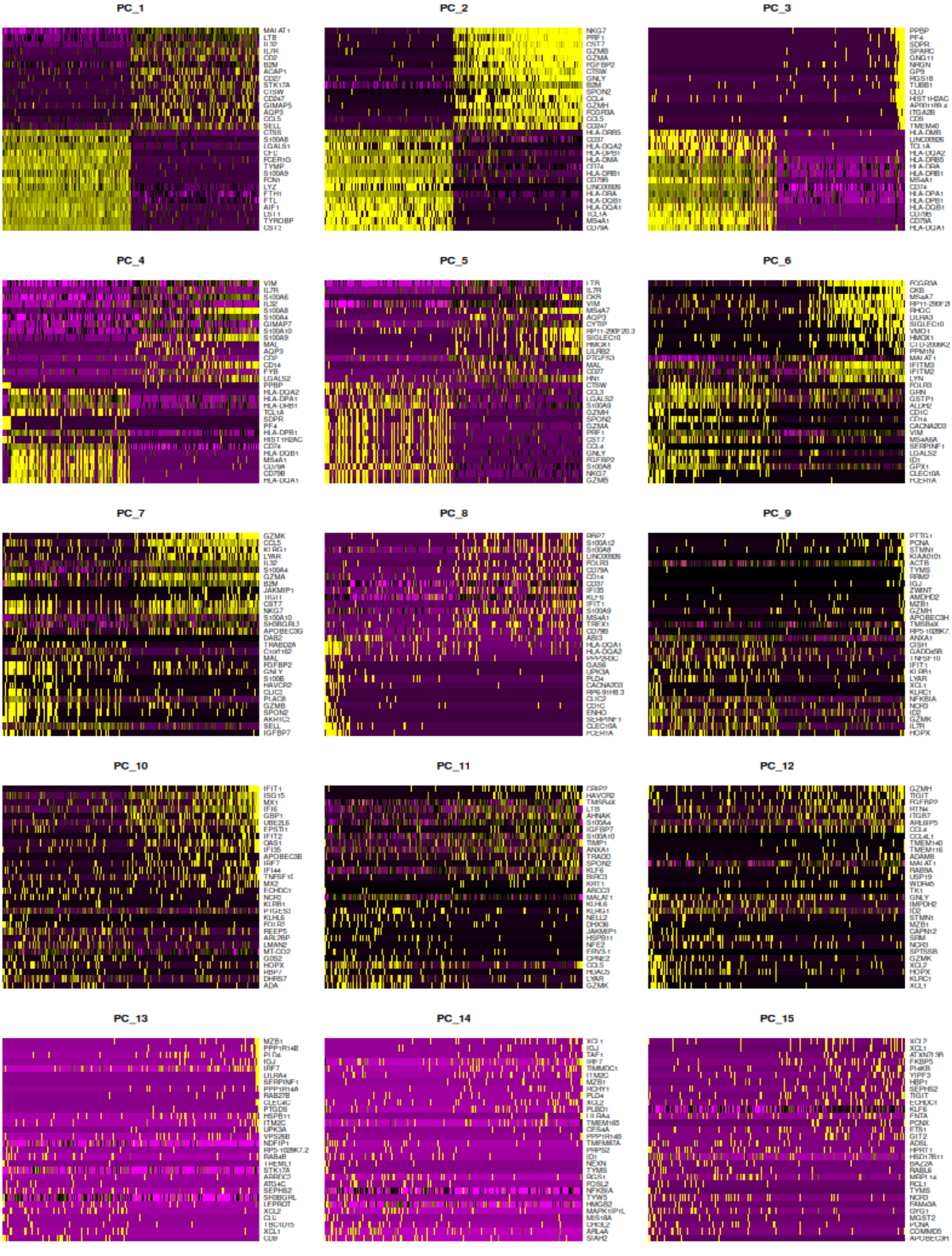


Figure 8: This elbow plot displaying the standard deviation of each principal component, used to visually identify the inflection point where true biological signal diminishes into random noise.

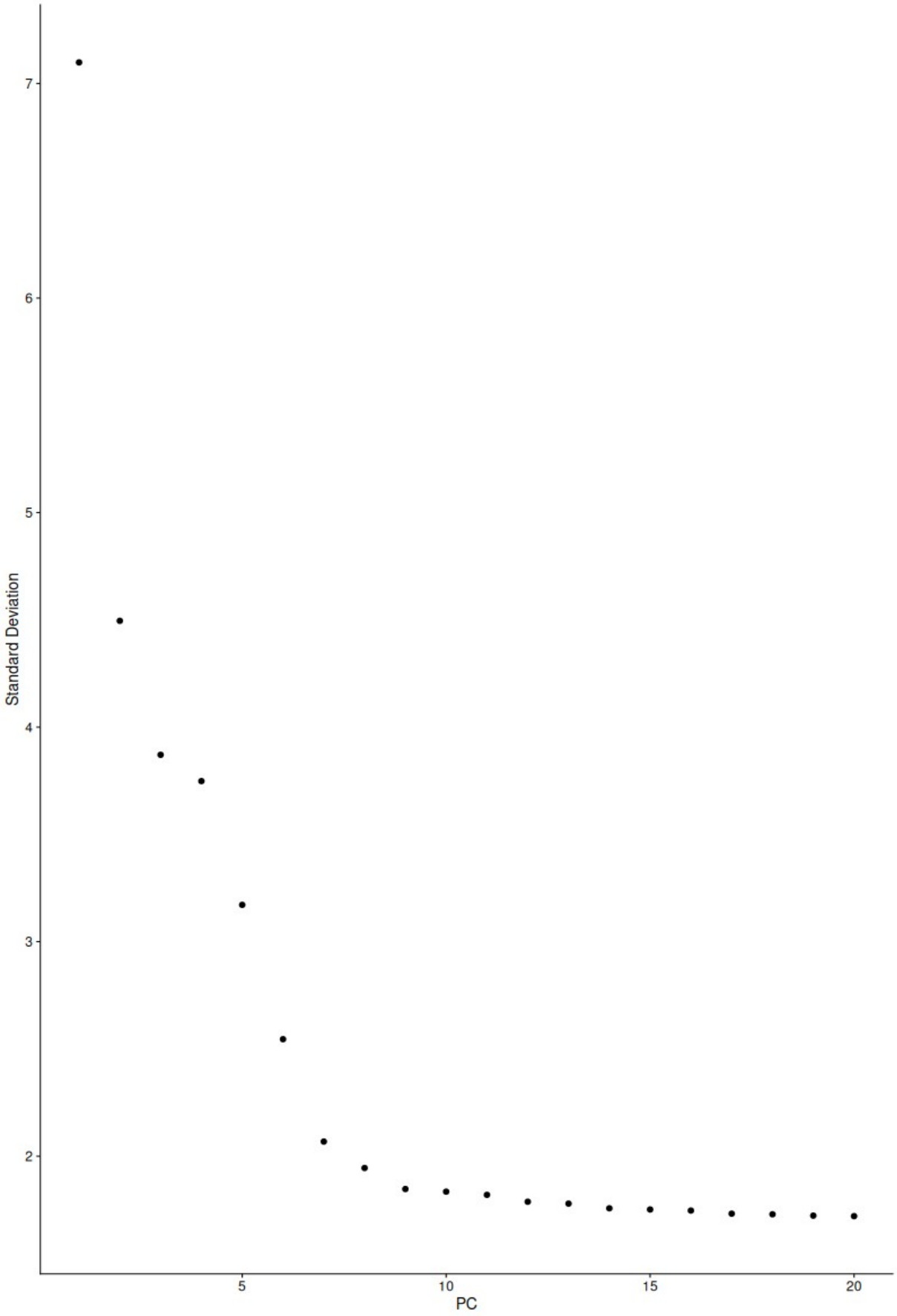


Figure 9: This shows a UMAP projection visualizing the single-cell clustering results, where each point represents an individual cell and colors indicate distinct clusters identified based on the first 10 principal components, with resolution 0.5.

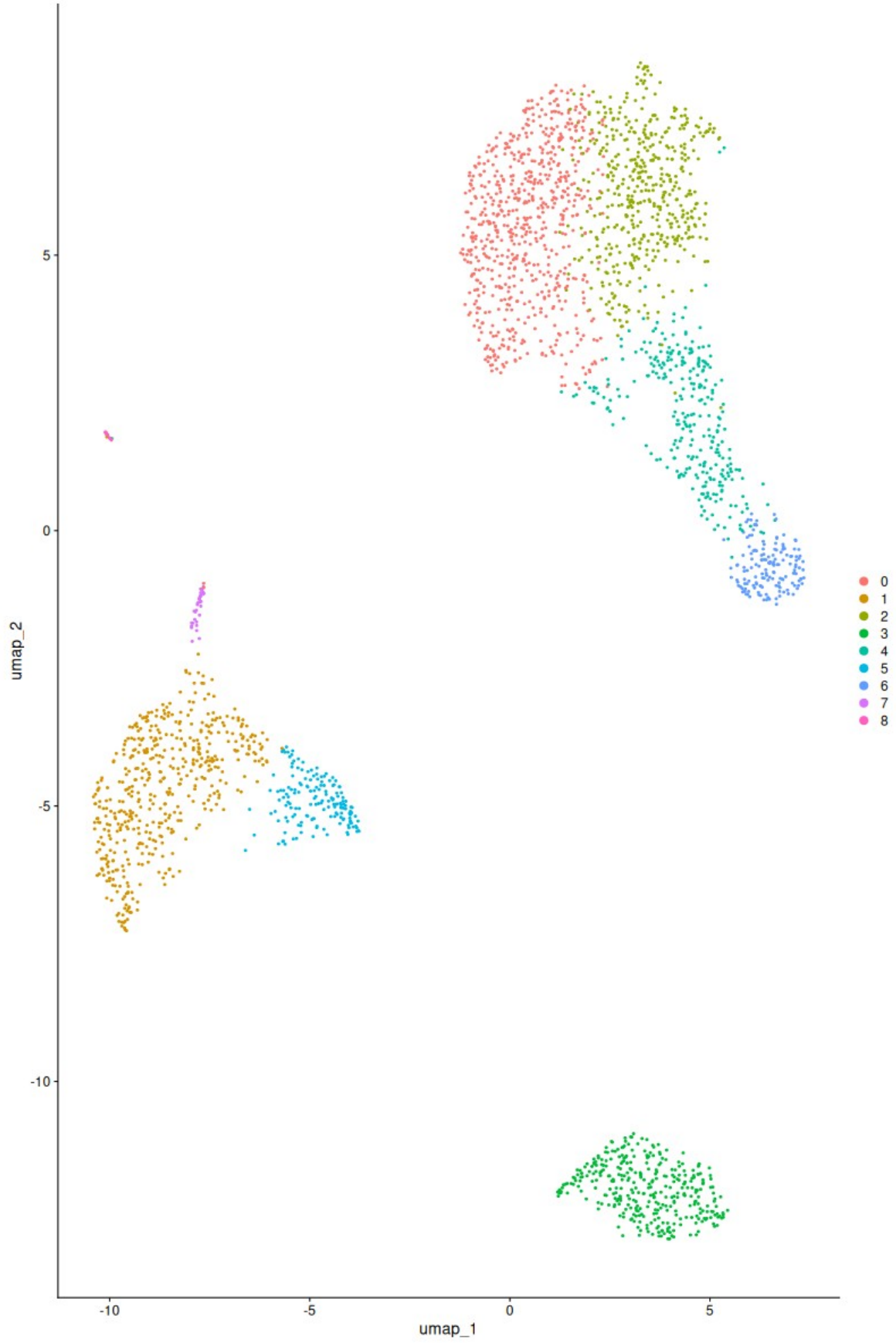


Figure 10: This shows a UMAP projection with the distinct clusters identified based on the first 5 principal components, with resolution 0.5. We observe that this affect our results and giving us a less clear clustering, loosing some clusters and the samples shows a broader distribution in the space compare with the figure 9.

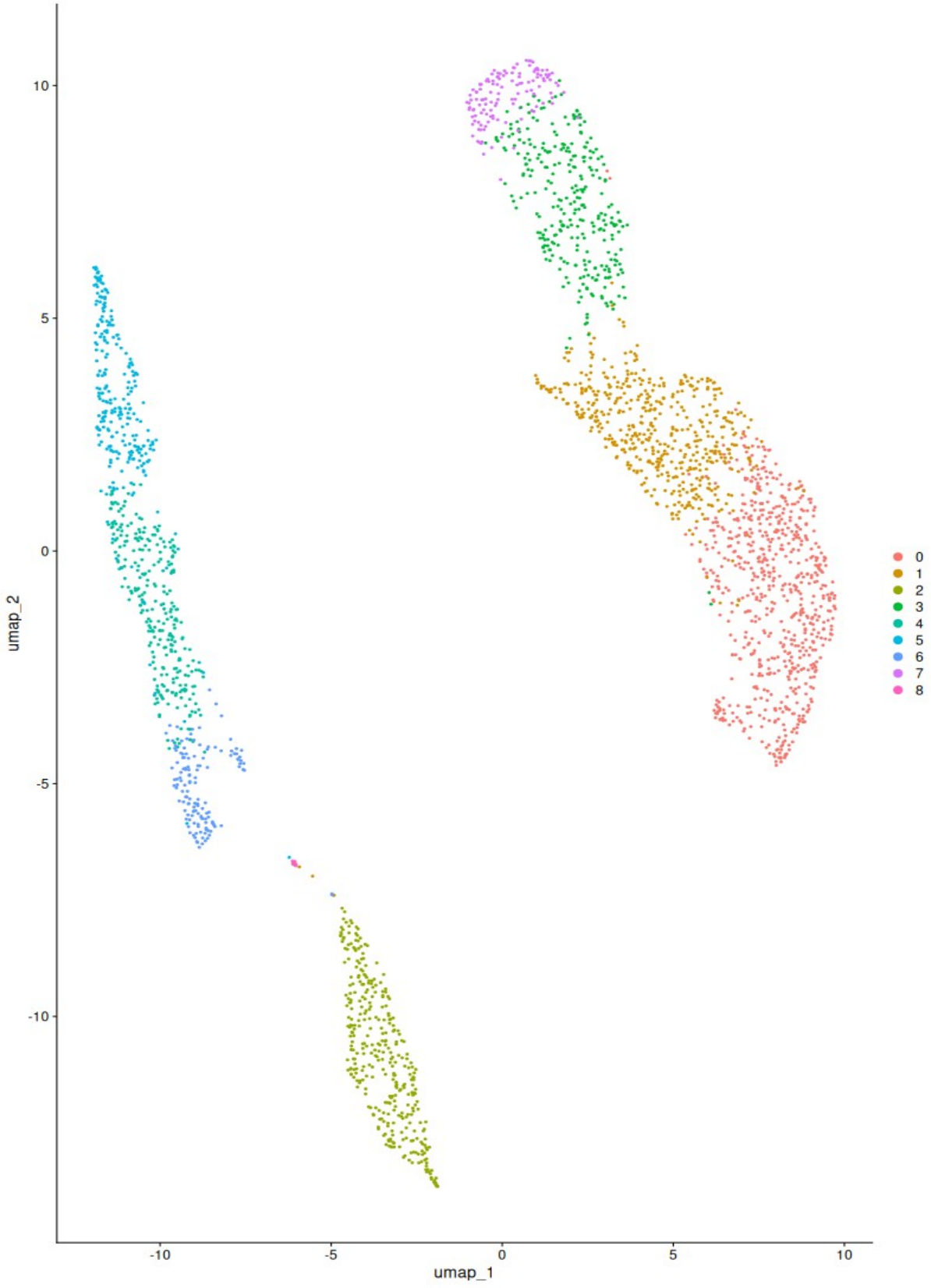


Figure 11: In this UMAP projection, we have used the same resolution (0.5), but we keek 15 principal components in the PCA. We observe that this affect the plot significant compare to figure 9 and 10 and show as a better separation among the clusters.

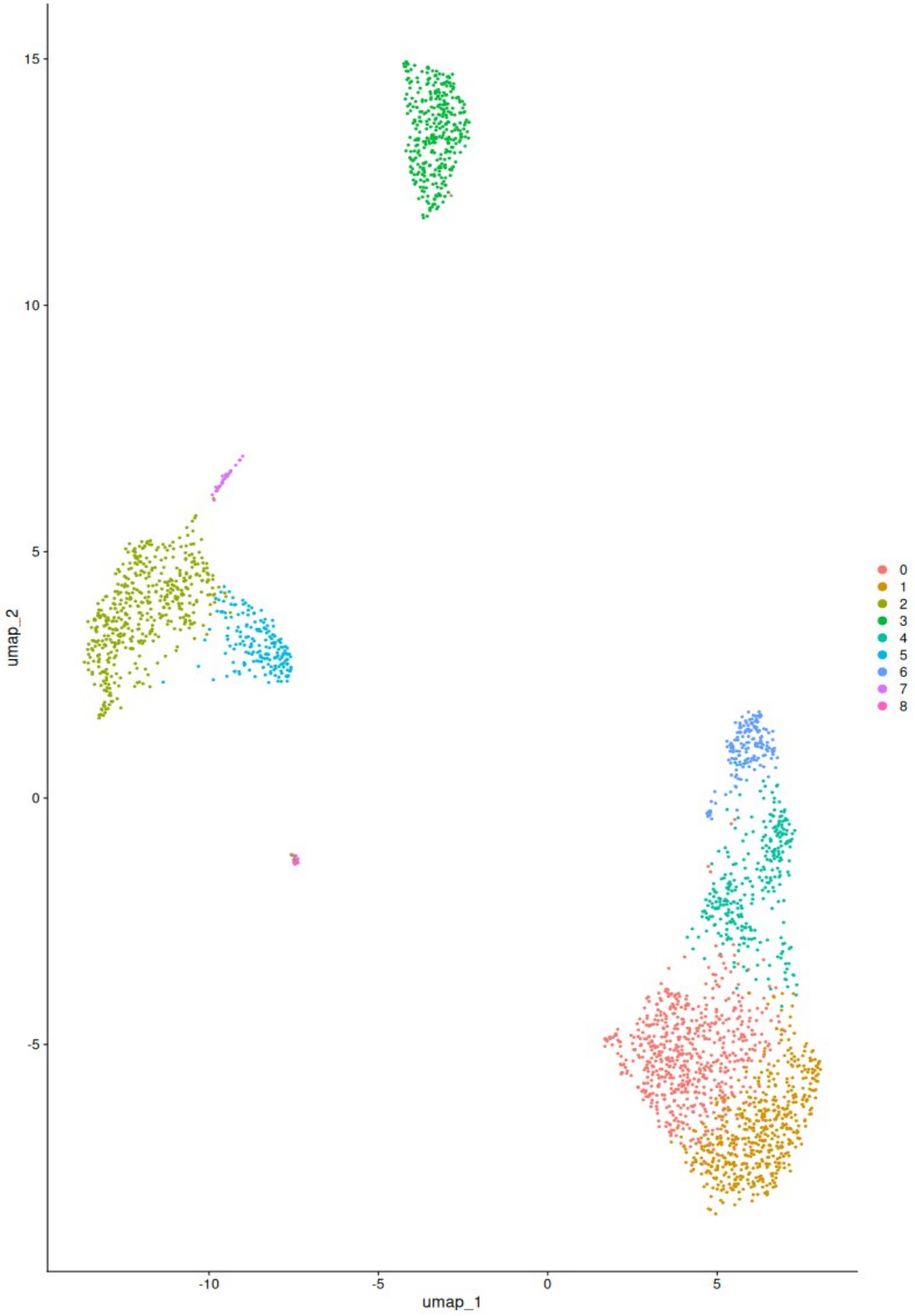


Figure 12: In this UMAP projection, we have maintain 10 principal components and change the resolution to 0.3. The position of the samples is the same as figure 9, but the number of the clusters looks decline like we zoom out and separate the samples to bigger subgroups, by merging the clusters that have a lot of similarities.

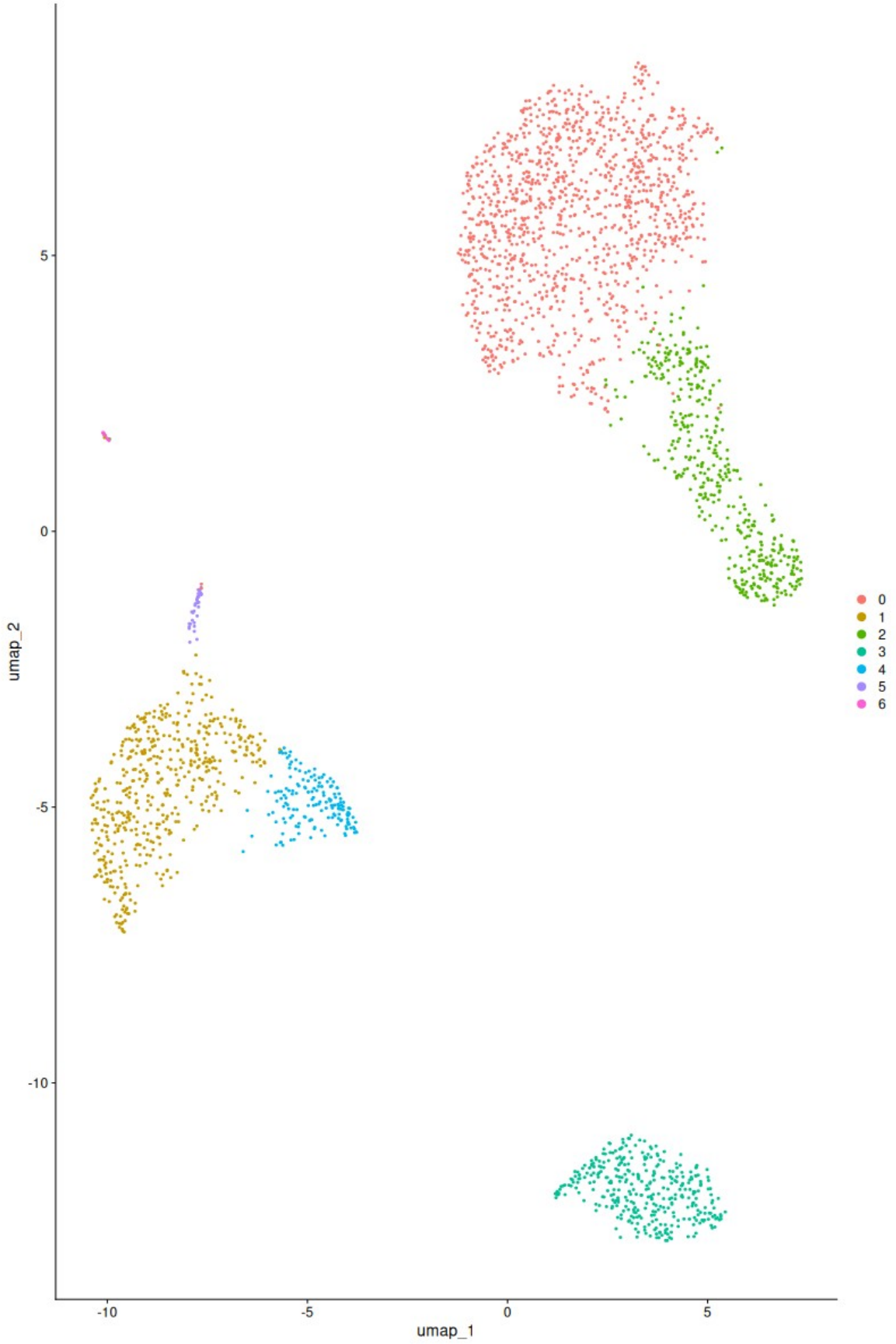
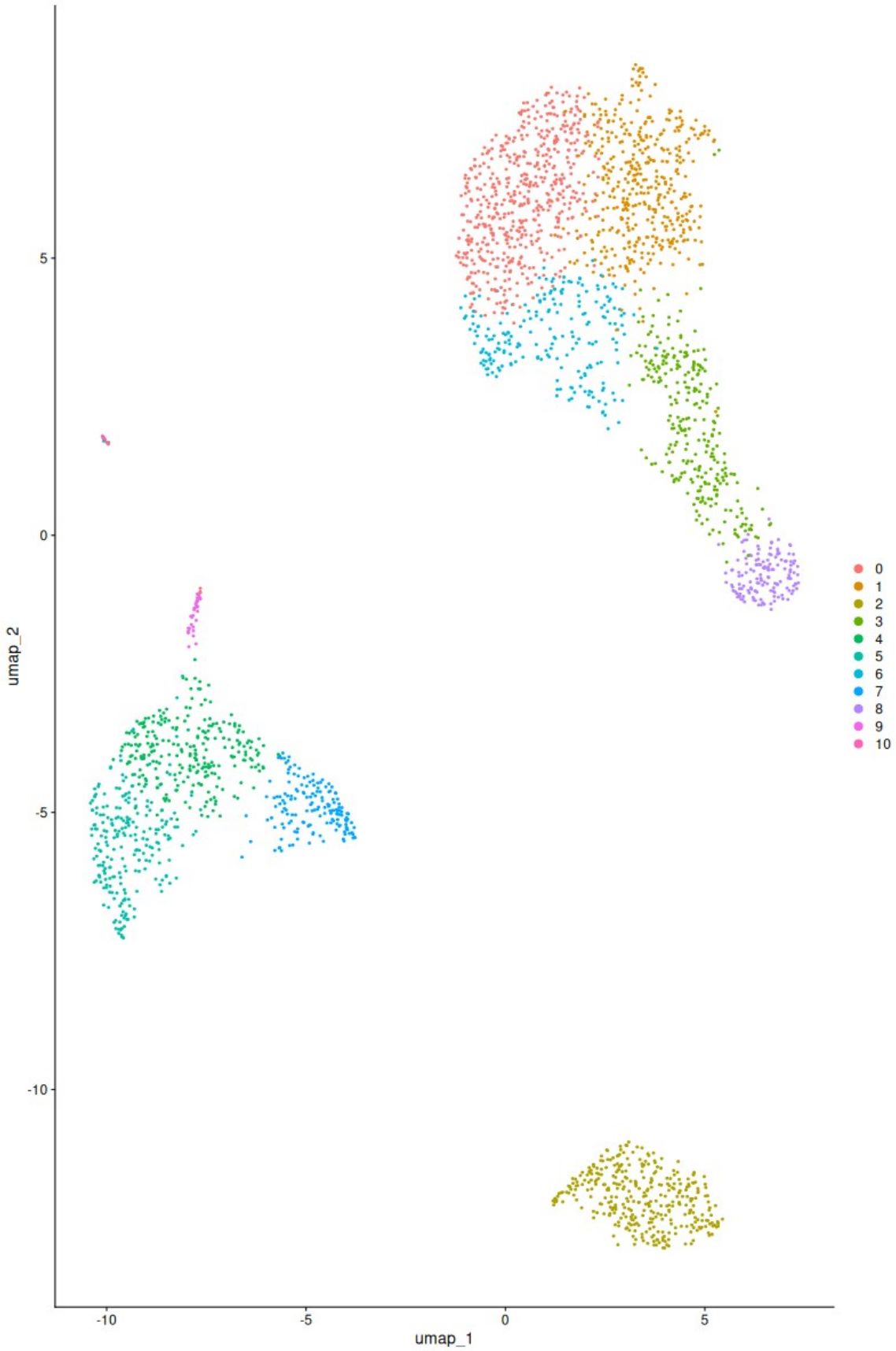


Figure 13: In this UMAP projection, we have maintain 10 principal components and change the resolution to 0.8. Here we observe that the position of the sample is maintain from figure 9, but contrary to figure 12, we are doing a zoom in and increasing the number of clusters, which can mean that we separate different subtypes of the same cell types.



<u>Figure 14:</u> In this Violin plots is displayed the expression probability distributions of the B-cell lineage markers *MS4A1* and *CD79A* across all clusters, demonstrating their specific enrichment within a distinct cell population, which we expect to correspond to the B-cells population.

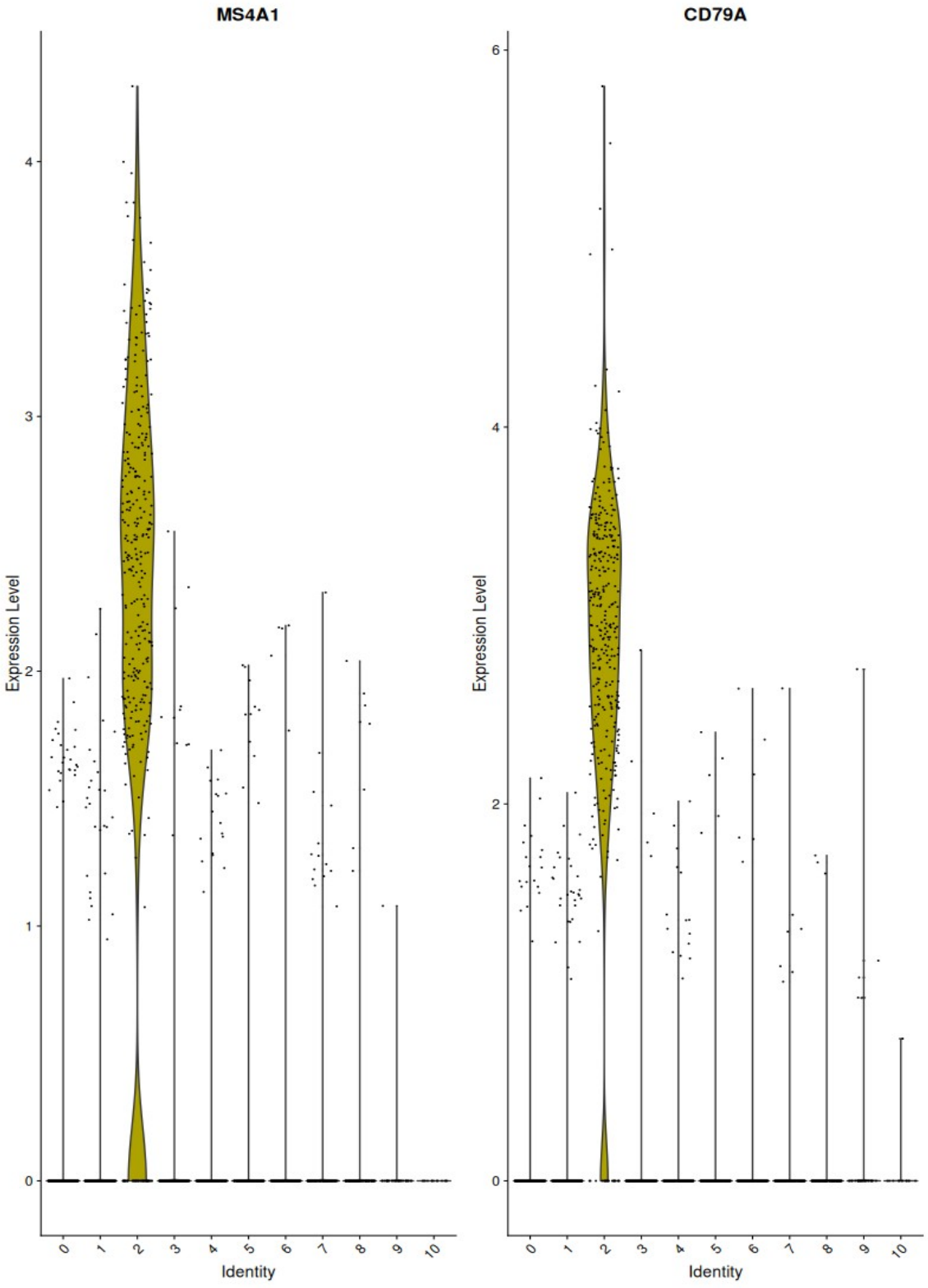


Figure 15: This violin plot displaying the distribution of log-transformed raw read counts for the markers *NKG7* and *PF4*, allowing for the assessment of absolute expression levels across clusters independent of normalization scaling.

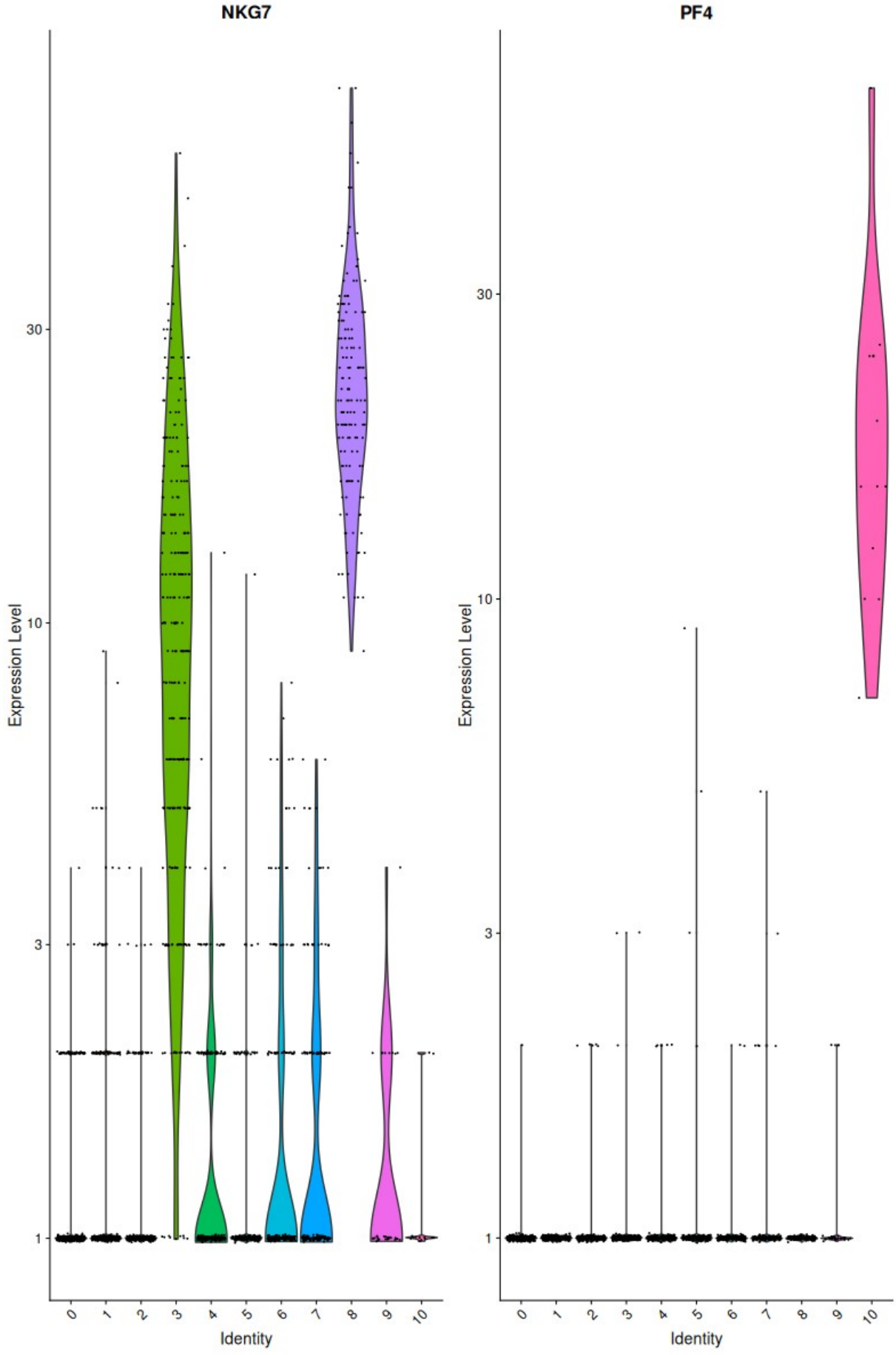
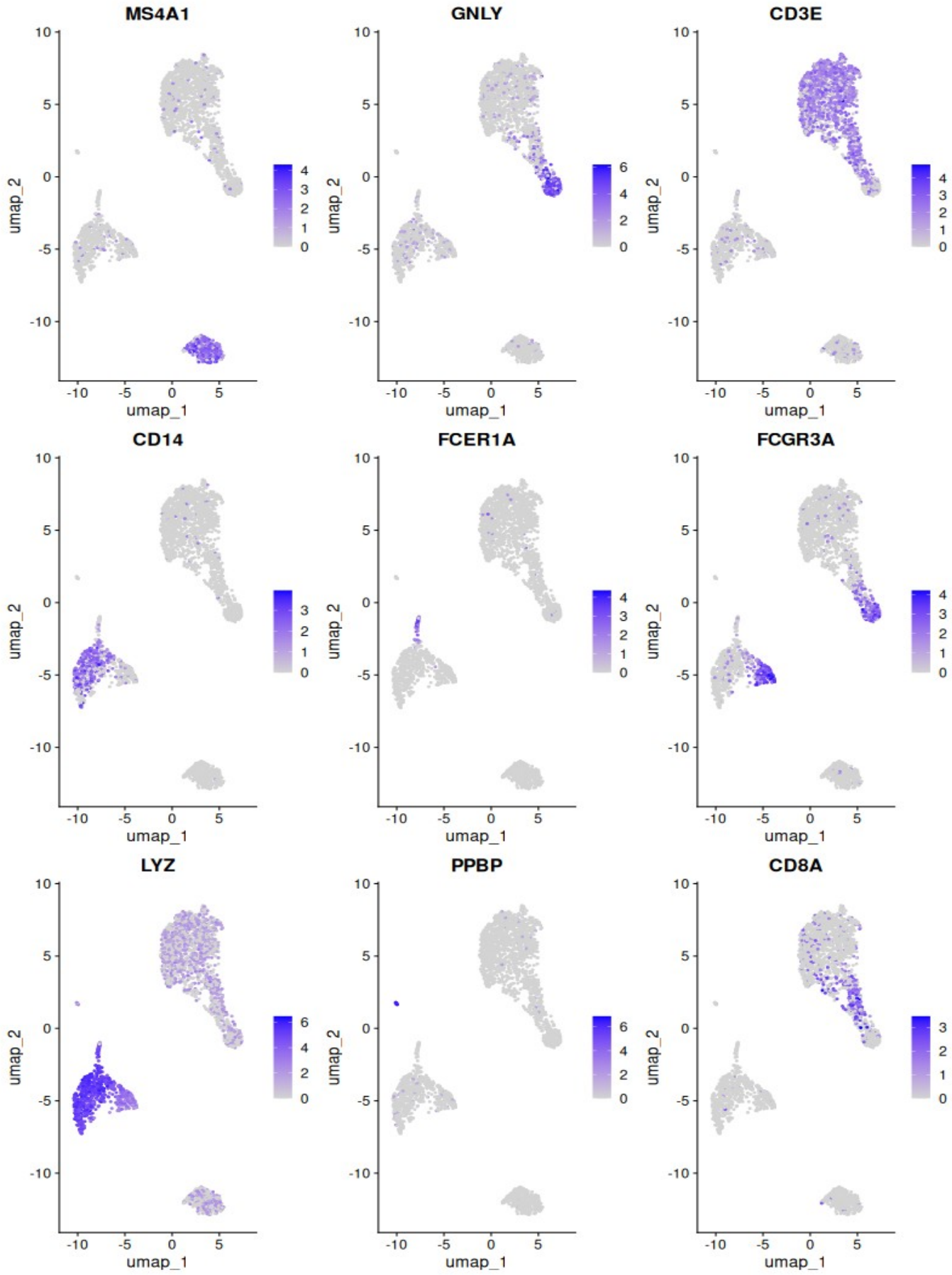


Figure 16: This feature plots visualizing the expression distribution of key canonical marker genes on the UMAP projection, used to map specific biological identities to the defined cell clusters.



<u>Figure 17:</u> This feature plots illustrating the differential expression of *CCL5*, *MS4A1*, and *SEPT5* across the UMAP projection. We can observe that there is a distinct localization of Cytotoxic T/NK cells (SEPT5) and B cells (MS4A1), respectively, while Platelets (SEPT5) are not appear at all, which is consistent with our analysis, given the fact that we select only nuclear cell, during this experiment.

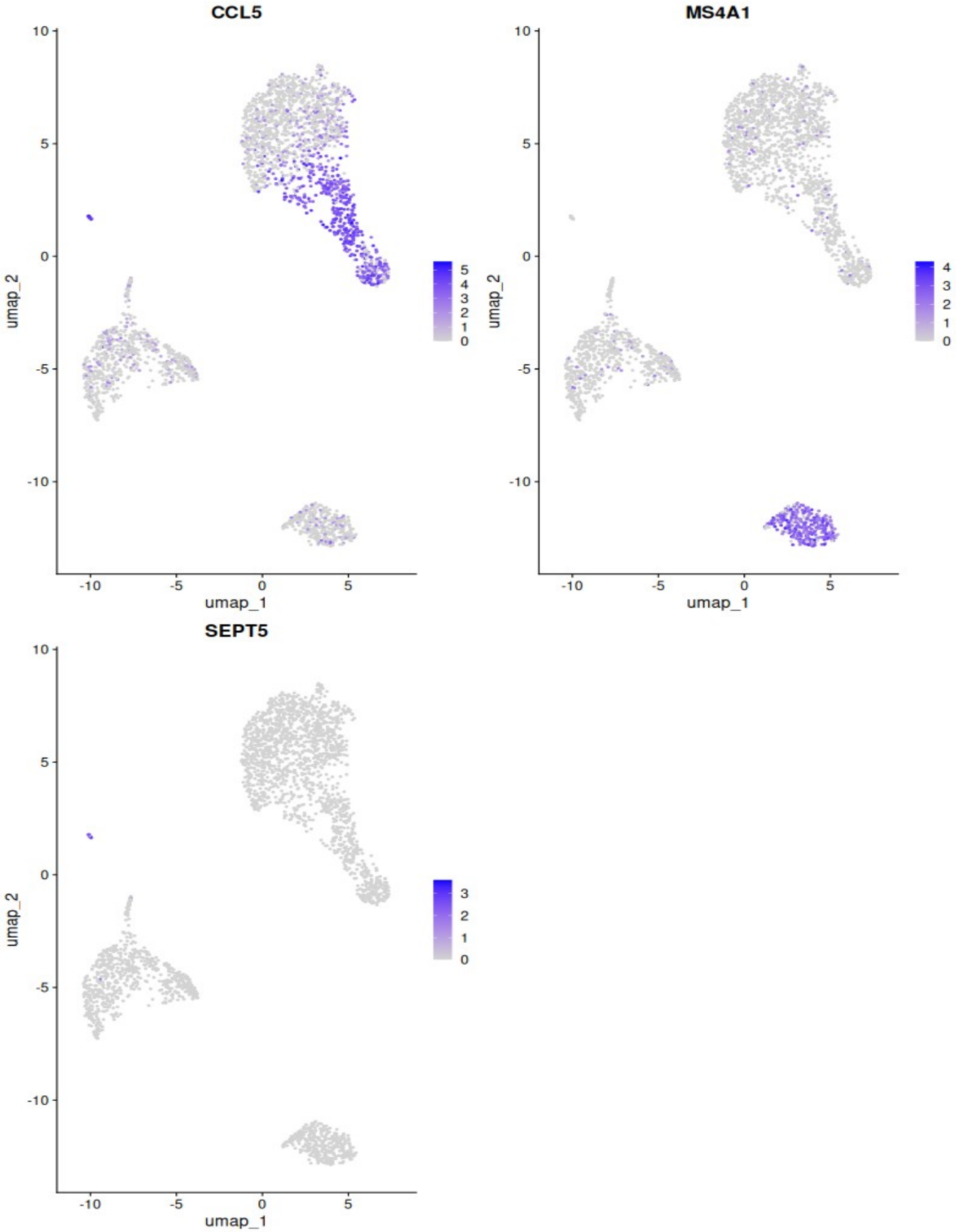


Figure 18: This heatmap visualizing the scaled expression of the top 10 upregulated marker genes for each cluster, displaying the distinct transcriptional signatures that define each cell population.

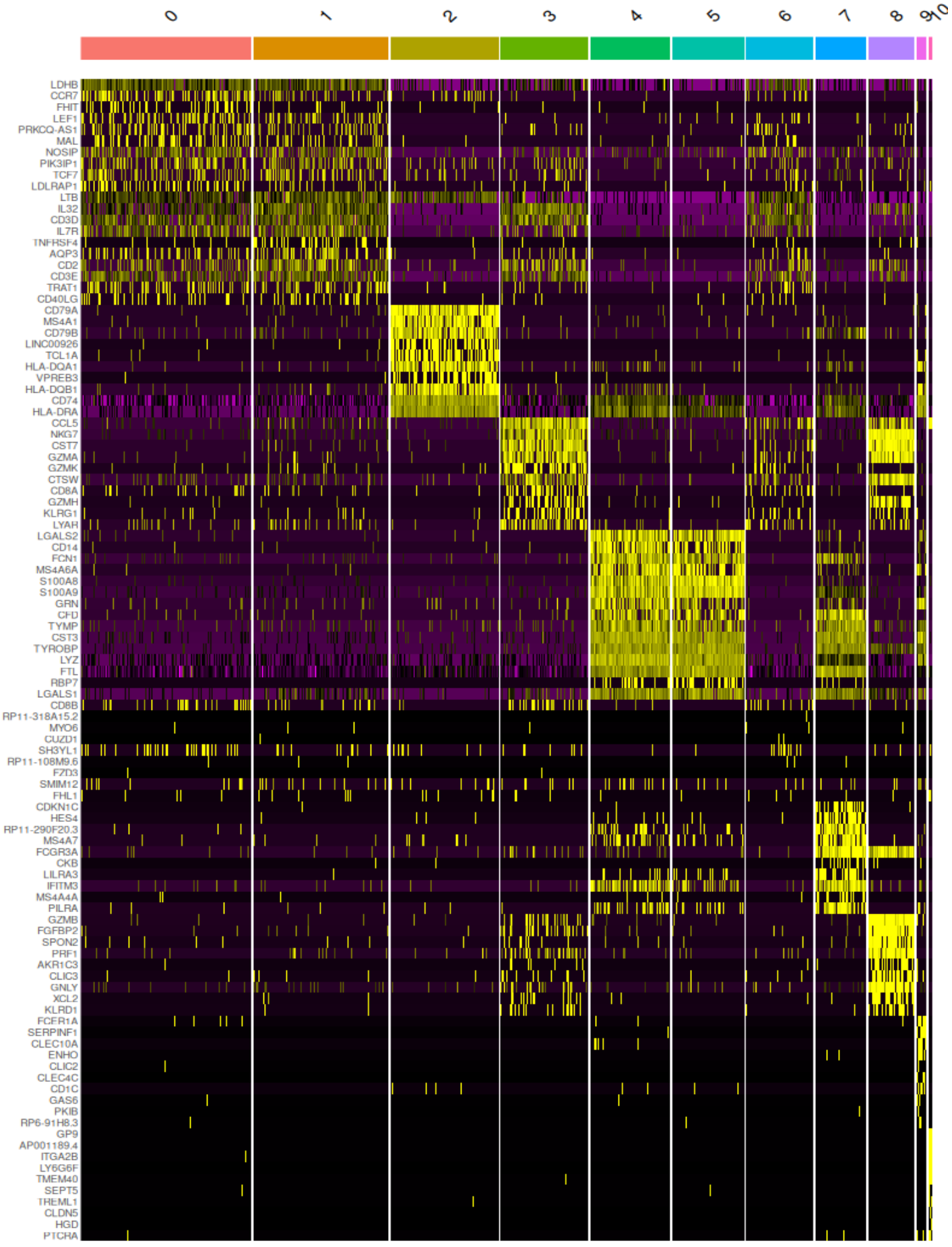


Figure 19: This UMAP projection showing the final cell type annotations, where the numerical cluster IDs have been replaced with biological labels derived from the canonical marker gene analysis.