Single-cell Bioinformatics

Assignment Project 1– Submission until 17.11.24

Zain Ziad (Matr. Nr.: 7068669)

Task 3: Add Meta-data

GSM4138872_scRNA_BMMC_D1T1

Cells: 6270
Genes: 20287

GSM4138873_scRNA_BMMC_D1T2

Cells: 6332
Genes: 20287

GSM4138874_scRNA_CD34_D2T1

Cells: 2424
Genes: 20287

GSM4138875_scRNA_CD34_D3T1

Cells: 5752
Genes: 20287

Meta-data:

orig.ident: Sample name
donor: Individual source
replicate: Technical repeat
sex: Gender/sex
nCount_RNA: Number of unique genes detected in cell.
nFeature_RNA: Total RNA molecule count in cell.

Task 4.1: Preprocessing

Q. Which steps do you perform before and after merging (task: 4.2) and why?

Filtering:

Filtering is important for ensuring data quality and biological relevance while reducing technical noise in the data. The specific metrics and parameters used to perform filtering are detailed below.

Normalization:

Normalization adjusts for technical variations, such as differences in sequencing depth, capture efficiency, and library size between cells.

Feature Selection:

Detailed below.

Doublet removal:

Detailed below.

Q. Name the parameters that have been used for filtering. Argue why you have used them and how you have chosen the cut-off parameters.

The parameters that have been used for filtering are:

nFeature_RNA (750-2500): This represents the number of unique genes detected per cell. The lower bound of 750 helps remove empty droplets, damaged cells, or background noise. The upper bound of 2500 helps eliminate potential doublets, as an unusually high number of genes often indicates multiple cells were captured together.

nCount_RNA (1500-6000): This represents the total RNA molecule count per cell. The lower threshold of 1500 removes low-quality cells with insufficient RNA content, while the upper limit of 6000 helps exclude potential doublets and cells with abnormally high RNA content that might indicate stress or technical artifacts.

Percent.mt: Represents mitochondrial percentage, which is typically filtered with a threshold around 5-10% in scRNA-seq data. A high percentage of mitochondrial reads (>10%) often indicates cell stress, damage, or death, as dying cells tend to lose cytoplasmic RNA while maintaining mitochondrial RNA. No mitochondrial genes were found in all 4 samples, thus filtering based on mitochondrial percentage was not performed.

Plots before and after QC:

BMMC_D1T1

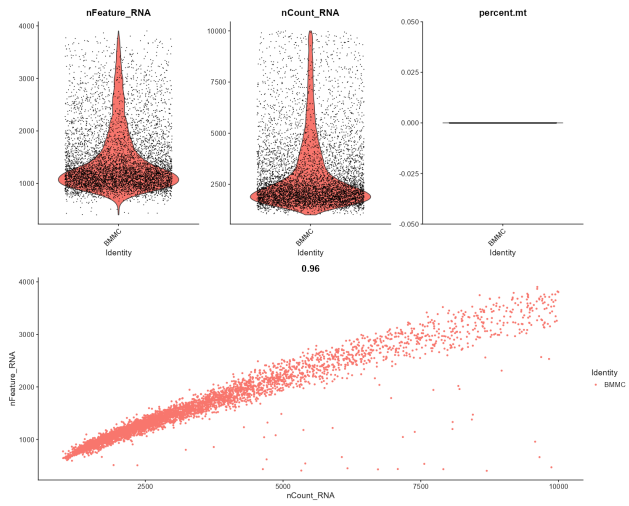

Figure 1: Before

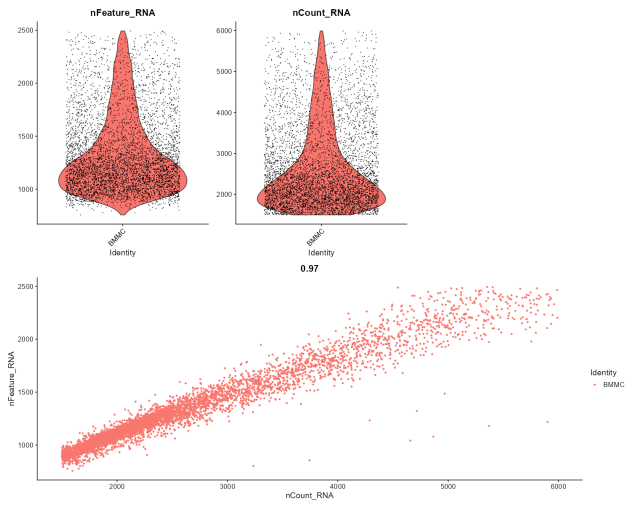

Figure 2: After

BMMC_D1T2

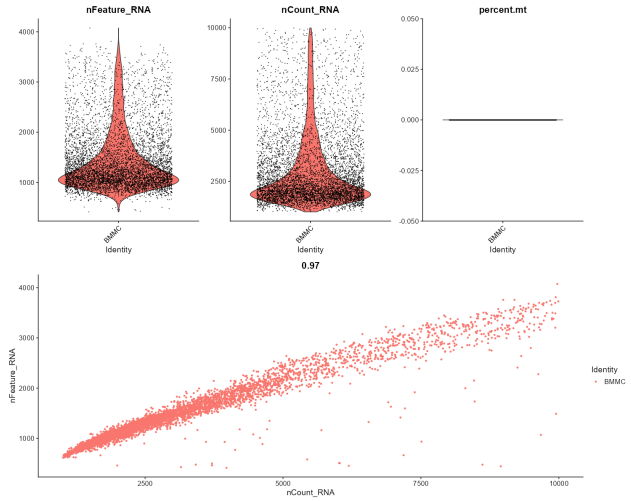

Figure 3: Before

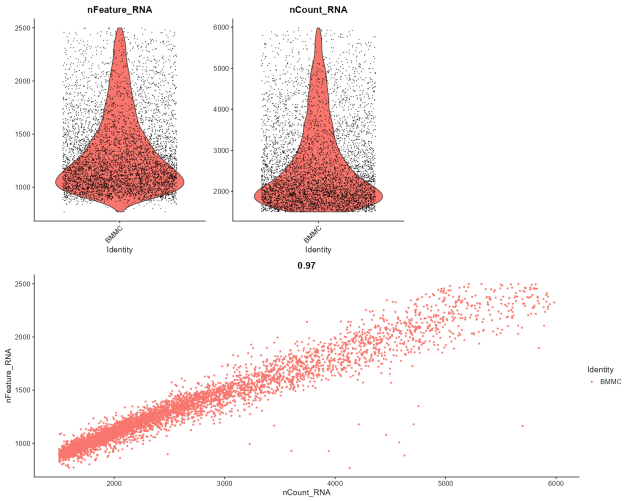

Figure 4: After

CD34_D2T1

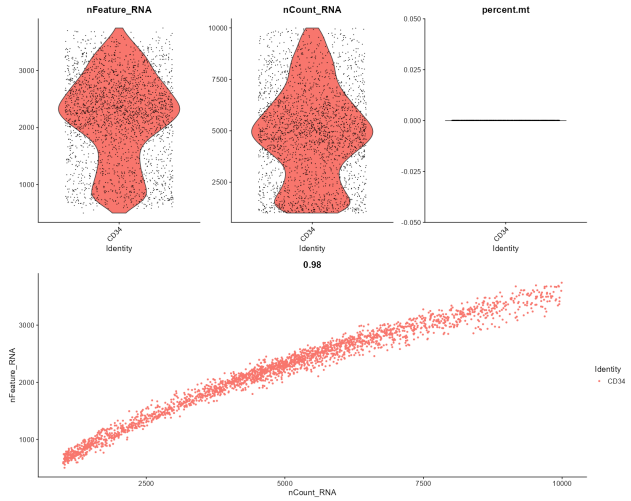

Figure 5: Before

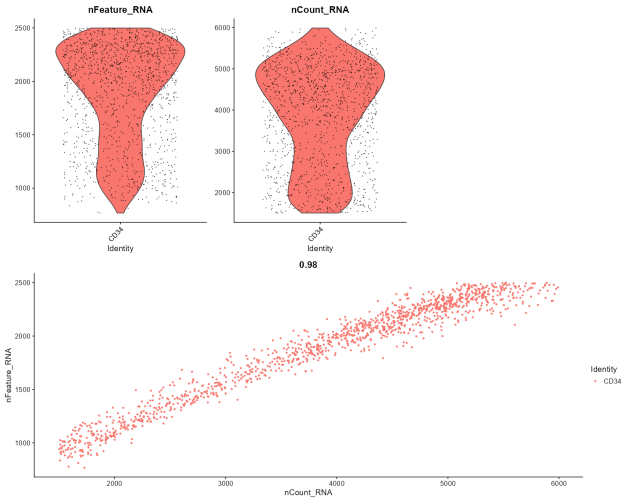

Figure 6: After

CD34_D3T1

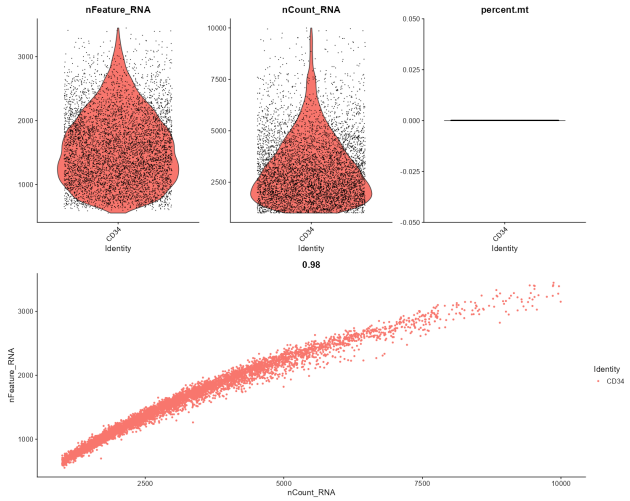

Figure 7: Before

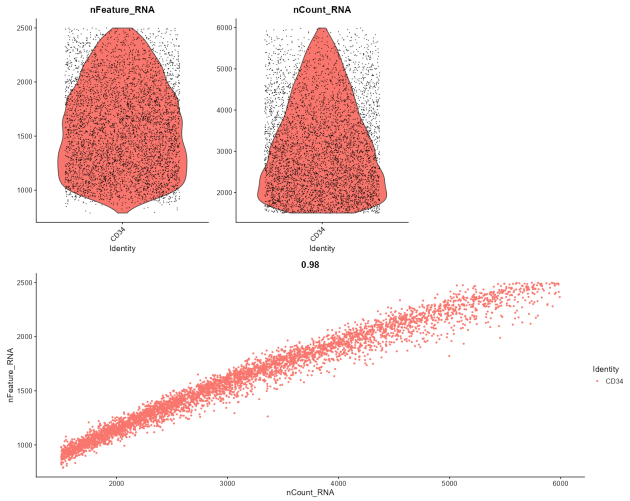

Figure 8: After

Q. Explain why we perform doublet removal.

Doublets are technical artifacts that arise in scRNA-seq data when two or more cells are mistakenly considered as a single cell, leading to mixed gene expression profiles that don't represent true biological cell states, potentially misleading downstream analyses like clustering and cell type identification.

Following were the distributions of doublets vs singlets in each sample:

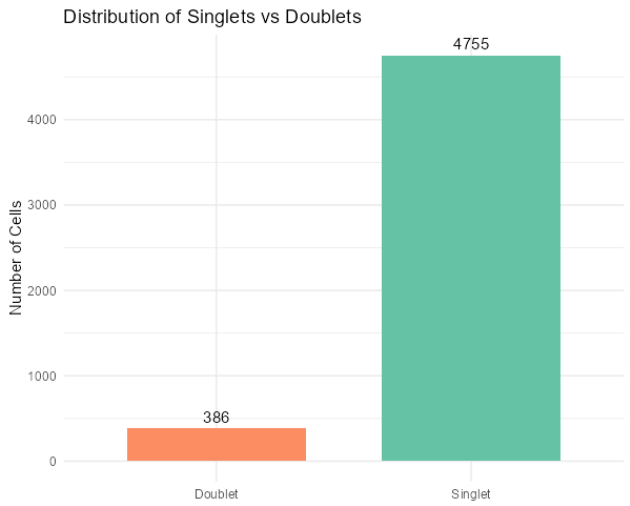

Figure 9: BMMC_D1T1

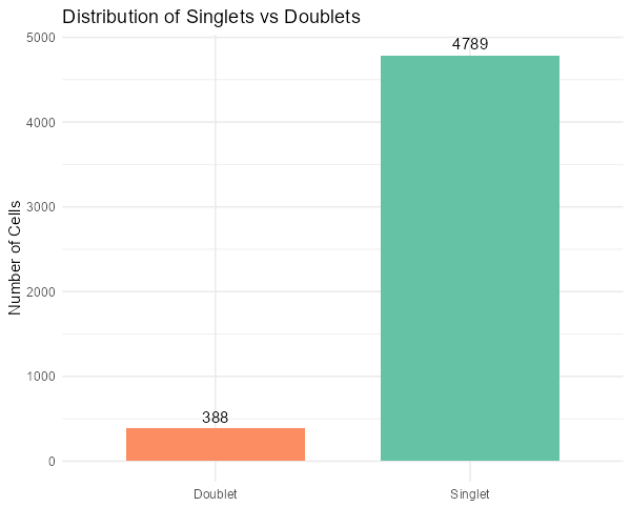

Figure 10: BMMC_D1T2

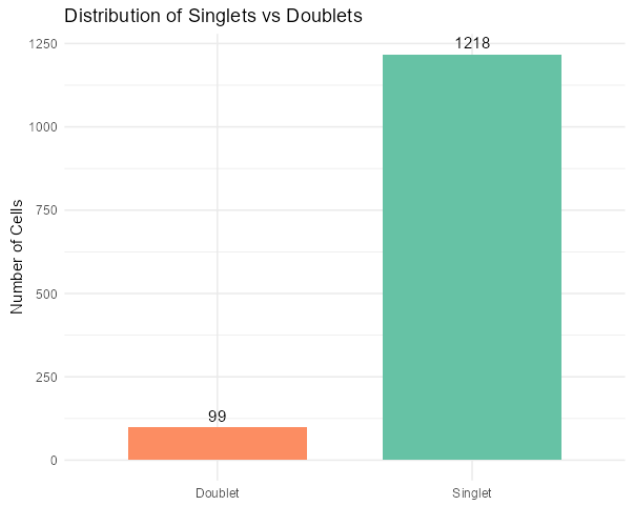

Figure 11: CD34_D3T1

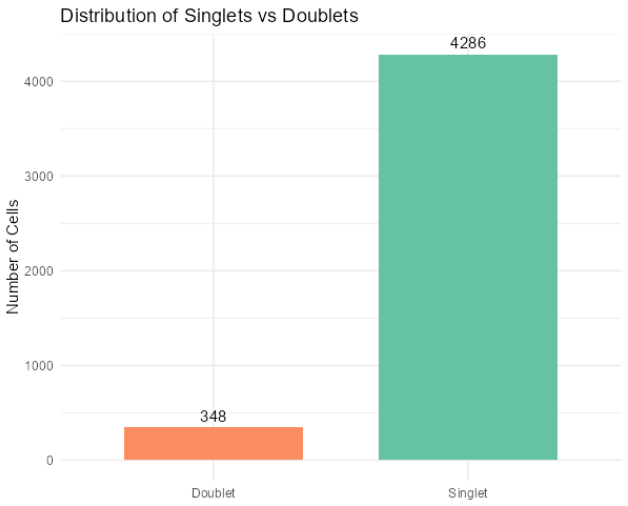

Figure 12: CD34_D3T1

The percentage of predicted doublets in each sample (around 7.5%) was within the acceptable range, so it was safe to remove them without risking the loss of true single cells.

Q. Which Normalization method is used by the Seurat Normalization function by default?

The default "LogNormalize" method was used here. Which for each cell, divides the expression values by the cell's total expression, then multiplies it by a scaling factor of 10,000, and finally applies a natural log transformation.

Q. What is the purpose of Feature Selection? How are they selected?

Feature selection is performed to focus on the most biologically relevant genes while reducing noise from the other genes. Seurat's default "vst" method was used here, which first models the relationship between gene variance and mean expression by fitting a curve using local regression. It then standardizes gene expression values based on this relationship and calculates the final variance scores, highlighting genes that show more variation than expected given their expression level.

Task 4.2: Batch-Correction

Q. Is Batch-Correction necessary? If yes, name the parameters and explain (with the necessary plots) why a correction for this parameter may be necessary.

Clustering with and without Integration (Batch-Correction):

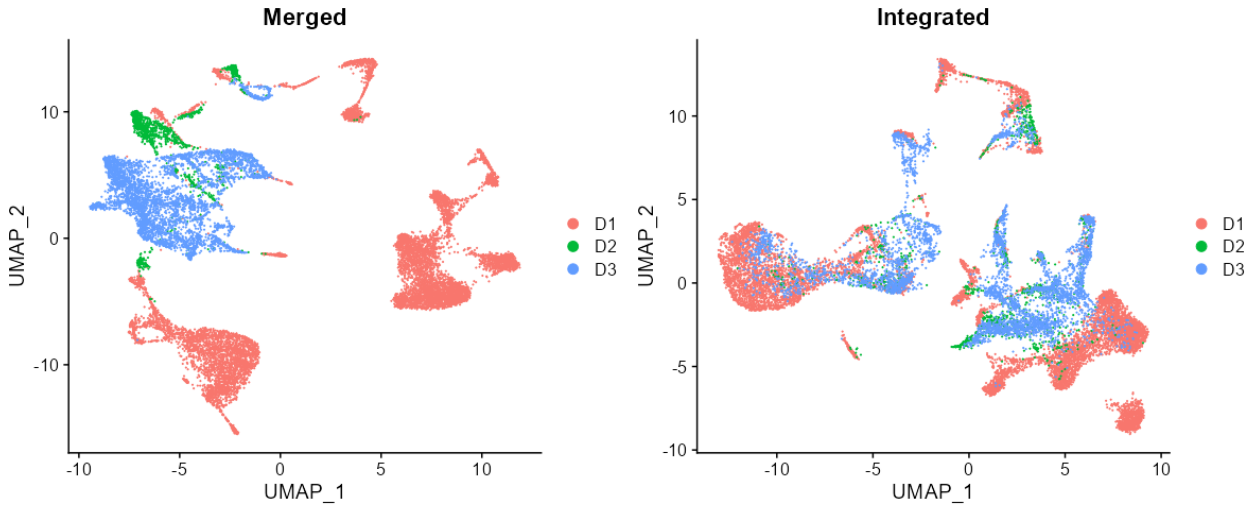

Figure 13: (Left) Clustering without integration. (Right) Clustering with integration.

As evident from the left plot, the clustering was performed entirely based on donors (samples), and does not actually represent the underlying biological differences between cells (like cell types or states). Batch correction helps remove these unwanted technical effects while preserving the actual biological variations, so we can properly integrate data across different samples and perform downstream analysis (right plot).

Task 5.1: Dimensionality Reduction

Q. How did you choose the number of dimensions? Use a plot to explain.

The optimal number of principal components (PCs) for analysis is determined using an elbow plot, which displays the percentage of variance explained by each PC. The "elbow" point in this plot represents where additional PCs begin to contribute minimally to the overall variance, which suggests that subsequent dimensions mainly capture noise rather than meaningful biological information.

This point was found to be approximately near **20 PCs**. We will select **30 PCs** for further analysis. This is to capture additional subtle biological variation in the data. Given that the curve still shows a gradual decline at this point rather than complete flattening, these additional PCs likely contain meaningful biological signal rather than just technical noise.

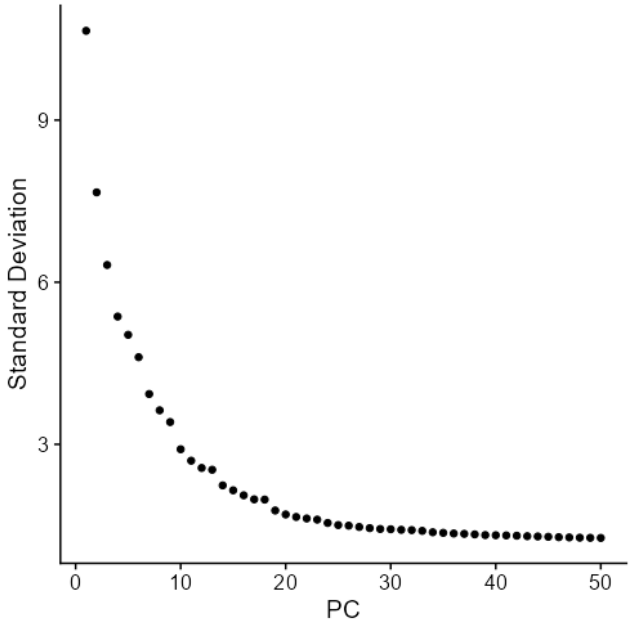

Figure 14: The standard deviation of each PC

Explain why we use a combination of PCA with UMAP for clustering and not only one of the methods?
We use a combination of PCA with UMAP because its more reliable than using either method alone. PCA first performs linear dimensionality reduction on high-dimensional data, removing technical noise and making analysis computationally manageable. This preprocessed data is then passed onto UMAP, which captures non-linear relationships and preserves local structure, resulting in more distinct clusters. While UMAP may not perfectly preserve density and can create artificial cluster separations, this combination leverages both methods' strengths to achieve better visualization and biologically meaningful clustering results than using either method alone.

Task 5.2: Clustering

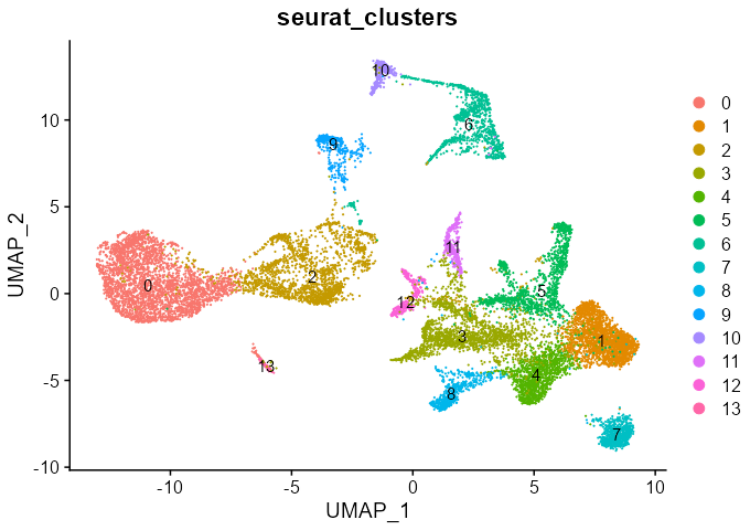

Figure 15: Clustering results

Task 6.1: Automatic Annotation

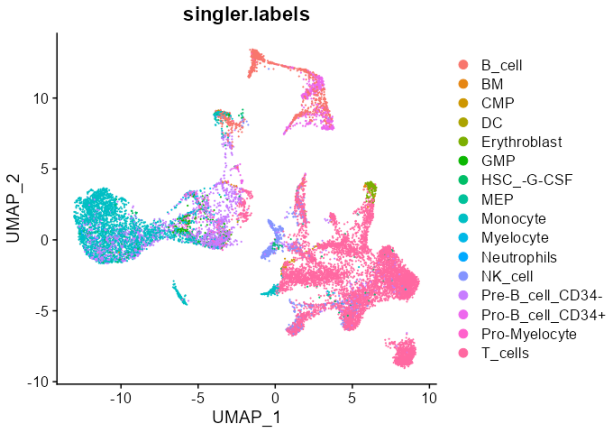

Figure 16: SingleR annotation results

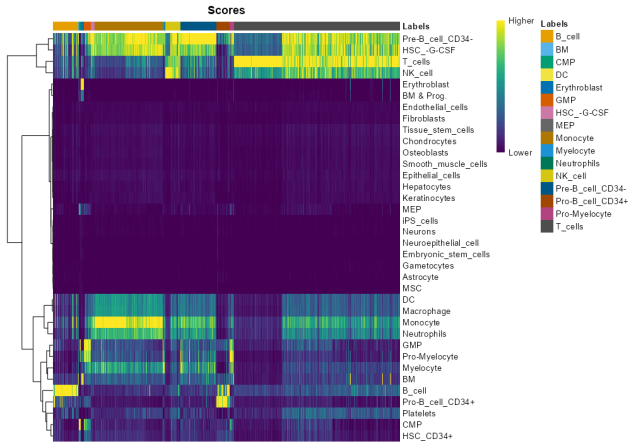

Figure 17: SingleR scores

Task 6.2: Manual Annotation

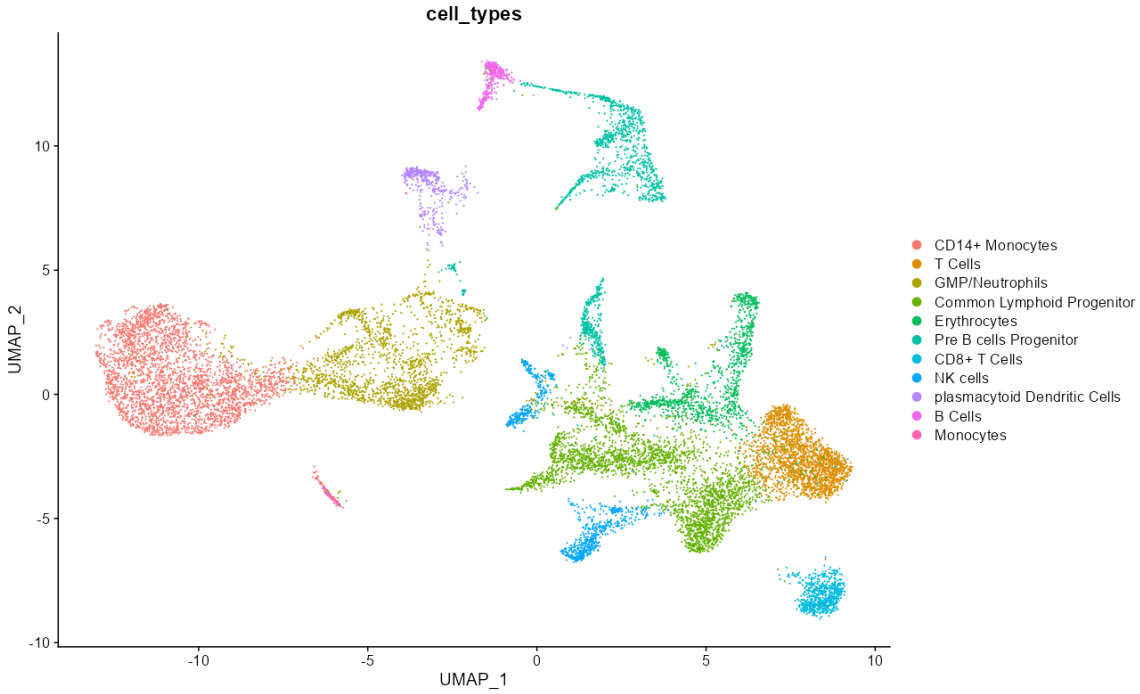

Figure 18: Manual annotation using the provided Table 2

Comparision between Automatic vs Manual Annotation

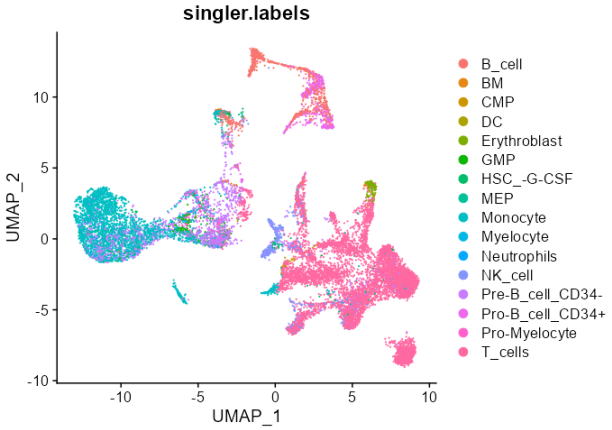

Figure 19: Automatic Annotation

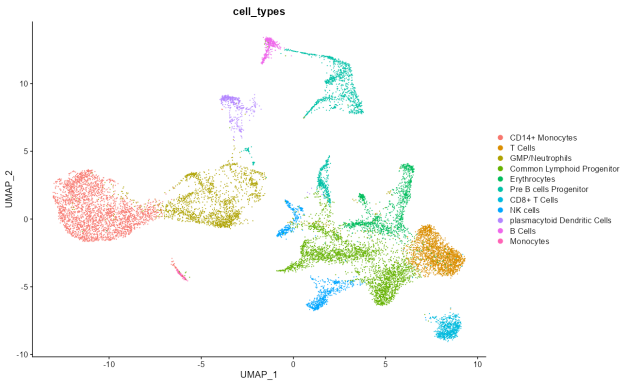

Figure 20: Manual Annotation

Both annotation methods produced similar results, though SingleR provided more specific cell type labels, making the classifications somewhat more complex/detailed.

Q. Show the gene expression of 3 marker genes in the different clusters using a Violin plot and in the different cells as a UMAP Plot.

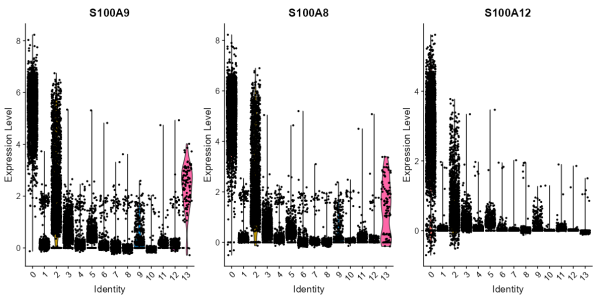

Figure 21: Violin plot for Cluster 0 marker genes

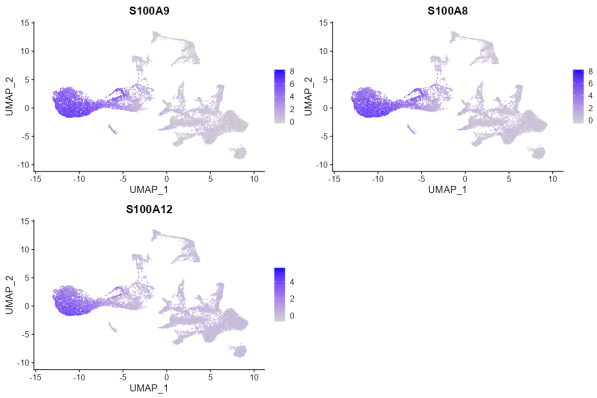

Figure 22: UMAP plot for Cluster 0 marker genes

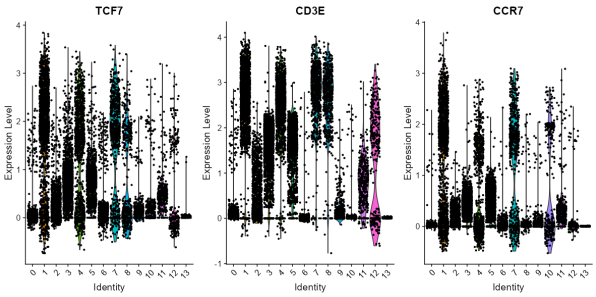

Figure 23: Violin plot for Cluster 1 marker genes

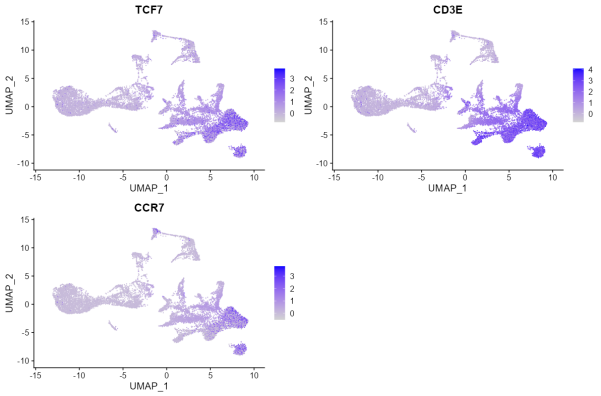

Figure 24: UMAP plot for Cluster 1 marker genes

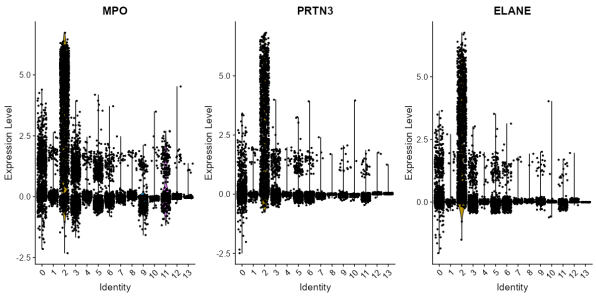

Figure 25: Violin plot for Cluster 2 marker genes

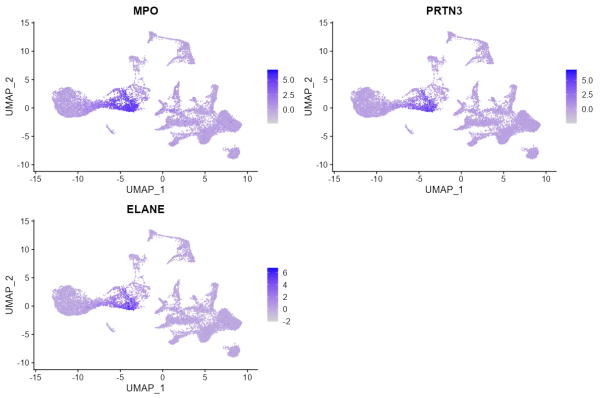

Figure 26: UMAP plot for Cluster 2 marker genes

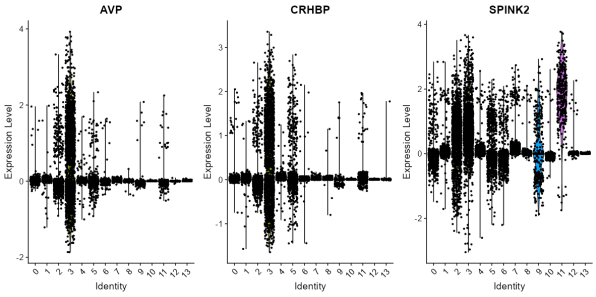

Figure 27: Violin plot for Cluster 3 marker genes

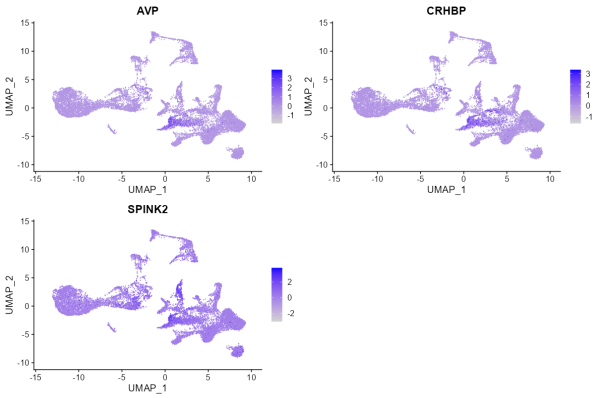

Figure 28: UMAP plot for Cluster 3 marker genes

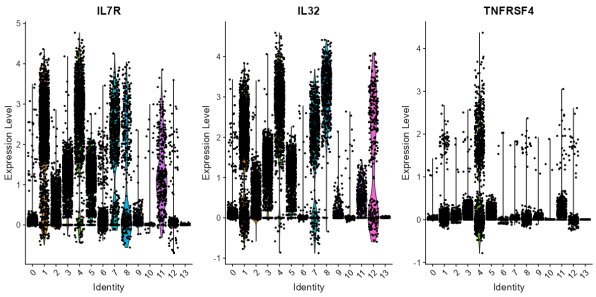

Figure 29: Violin plot for Cluster 4 marker genes

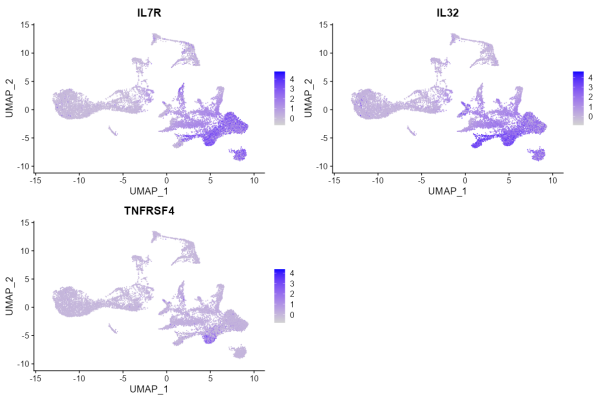

Figure 30: UMAP plot for Cluster 4 marker genes

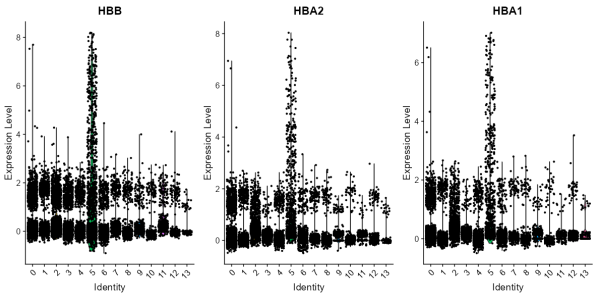

Figure 31: Violin plot for Cluster 5 marker genes

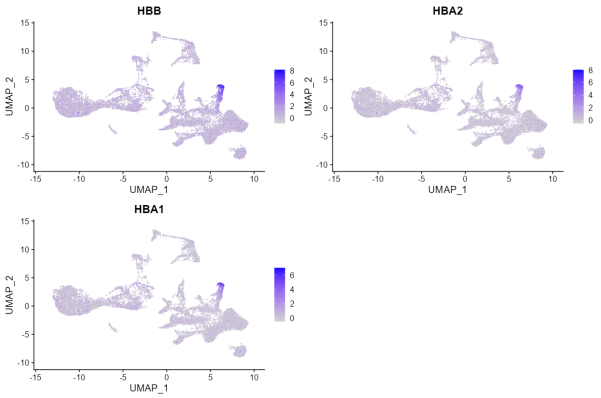

Figure 32: UMAP plot for Cluster 5 marker genes

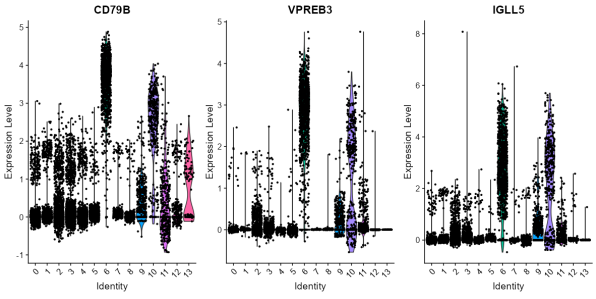

Figure 33: Violin plot for Cluster 6 marker genes

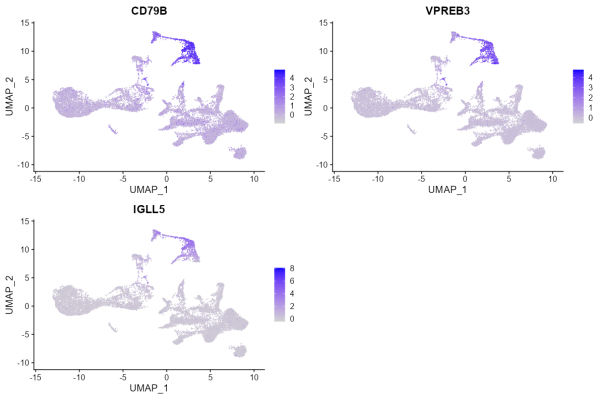

Figure 34: UMAP plot for Cluster 6 marker genes

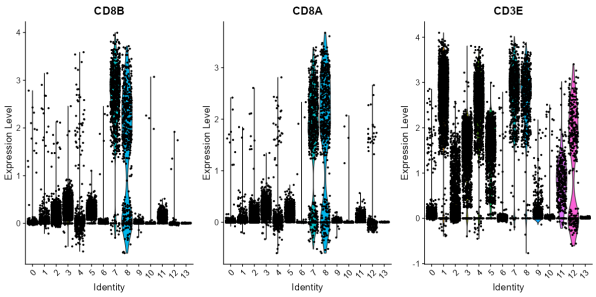

Figure 35: Violin plot for Cluster 7 marker genes

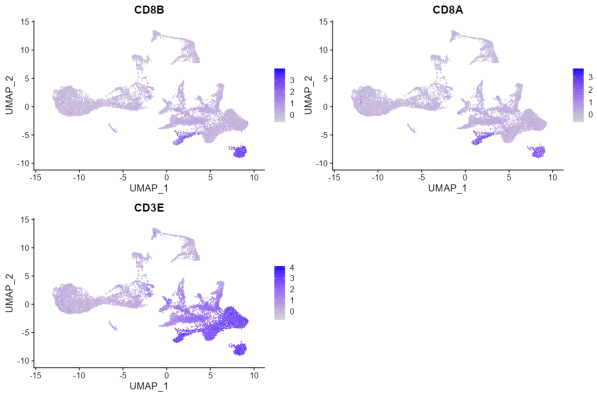

Figure 36: UMAP plot for Cluster 7 marker genes

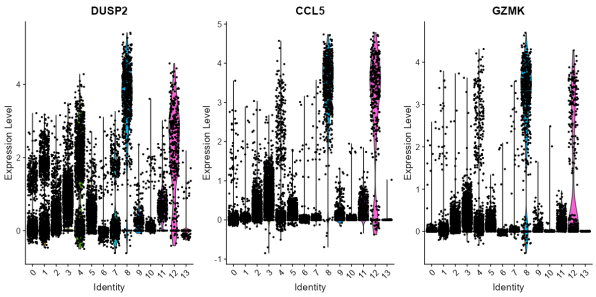

Figure 37: Violin plot for Cluster 8 marker genes

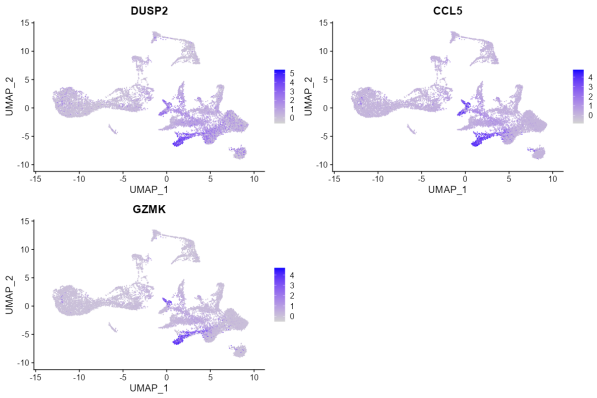

Figure 38: UMAP plot for Cluster 8 marker genes

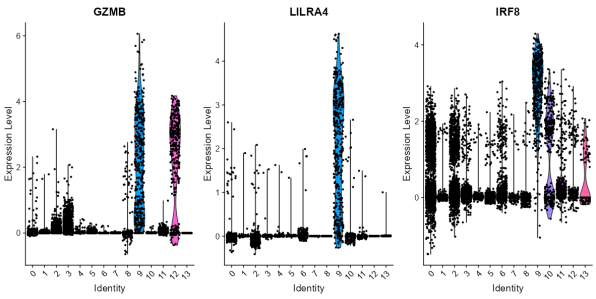

Figure 39: Violin plot for Cluster 9 marker genes

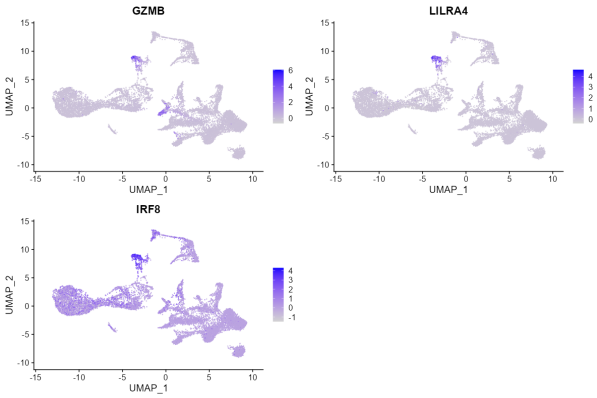

Figure 40: UMAP plot for Cluster 9 marker genes

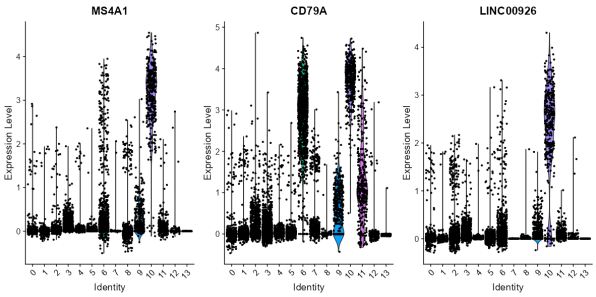

Figure 41: Violin plot for Cluster 10 marker genes

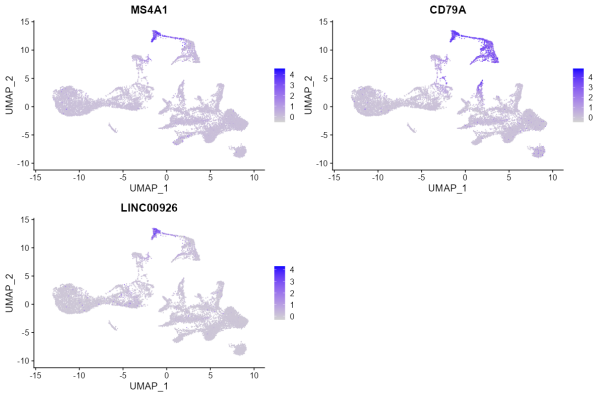

Figure 42: UMAP plot for Cluster 10 marker genes

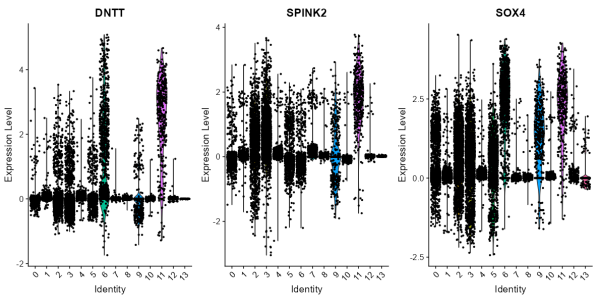

Figure 43: Violin plot for Cluster 11 marker genes

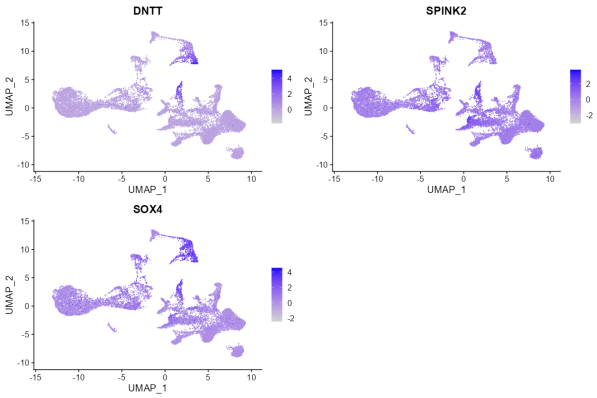

Figure 44: UMAP plot for Cluster 11 marker genes

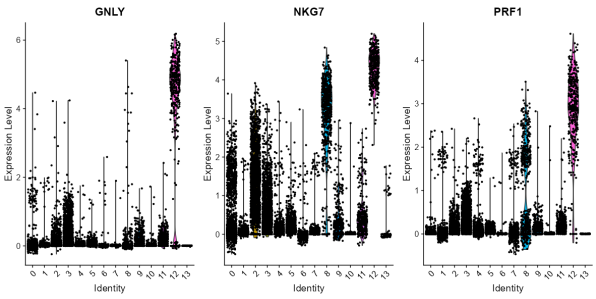

Figure 45: Violin plot for Cluster 12 marker genes

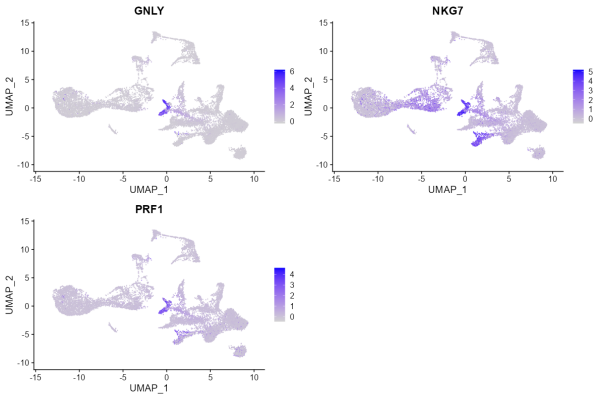

Figure 46: UMAP plot for Cluster 12 marker genes

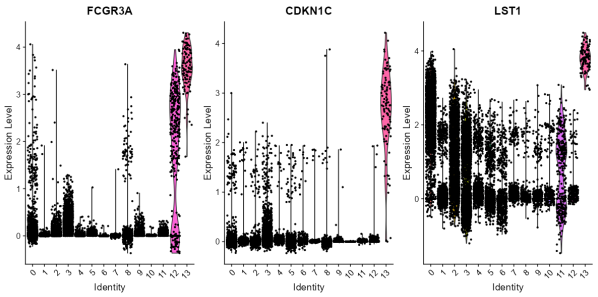

Figure 47: Violin plot for Cluster 13 marker genes

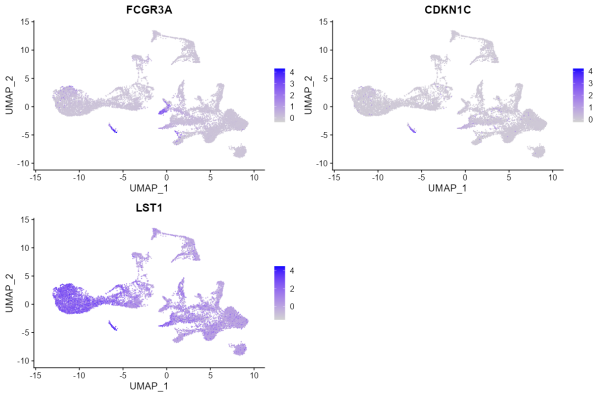

Figure 48: UMAP plot for Cluster 13 marker genes

Task 6.3: Cell-Type Proportions

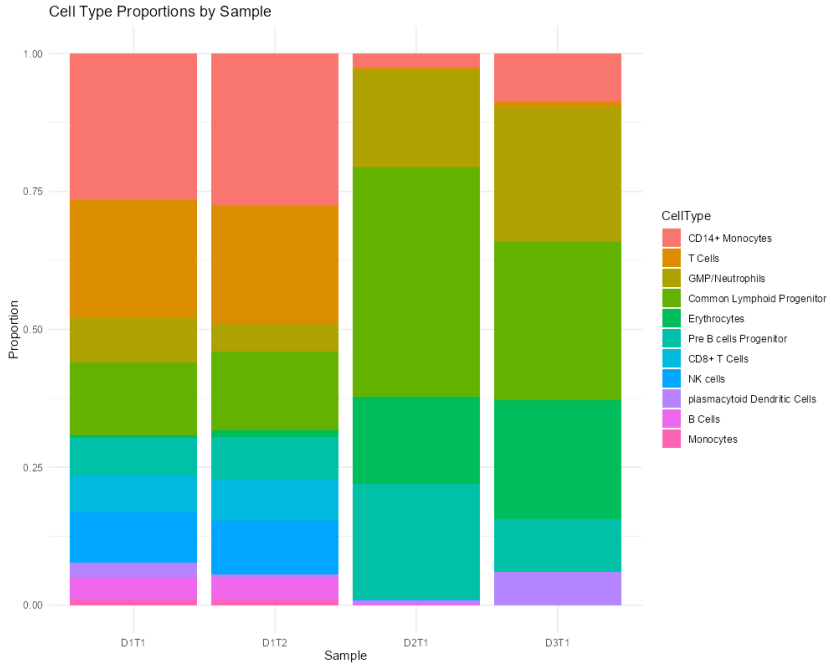

Figure 49: Stacked barplot of celltype proportions in each sample

D1T1 and D1T2 are from regular bone marrow samples, which naturally contain a mix of mature and immature cells, including B cells and CD14+ Monocytes. These samples show similar profiles because they're from the same donor and same collection method. D2T1 and D3T1 show more common lymphoid progenitor cells and fewer mature cells (like CD14+ Monocytes) because they're CD34+ selected samples. CD34 is a marker for hematopoietic stem and progenitor cells, so this selection specifically enriches for immature/progenitor cells while excluding mature cell types. This explains why these samples have higher proportions of progenitor cells compared to the normal bone marrow samples (D1T1/D1T2).

Task 7.1: Differential Expression Analysis on cell-types

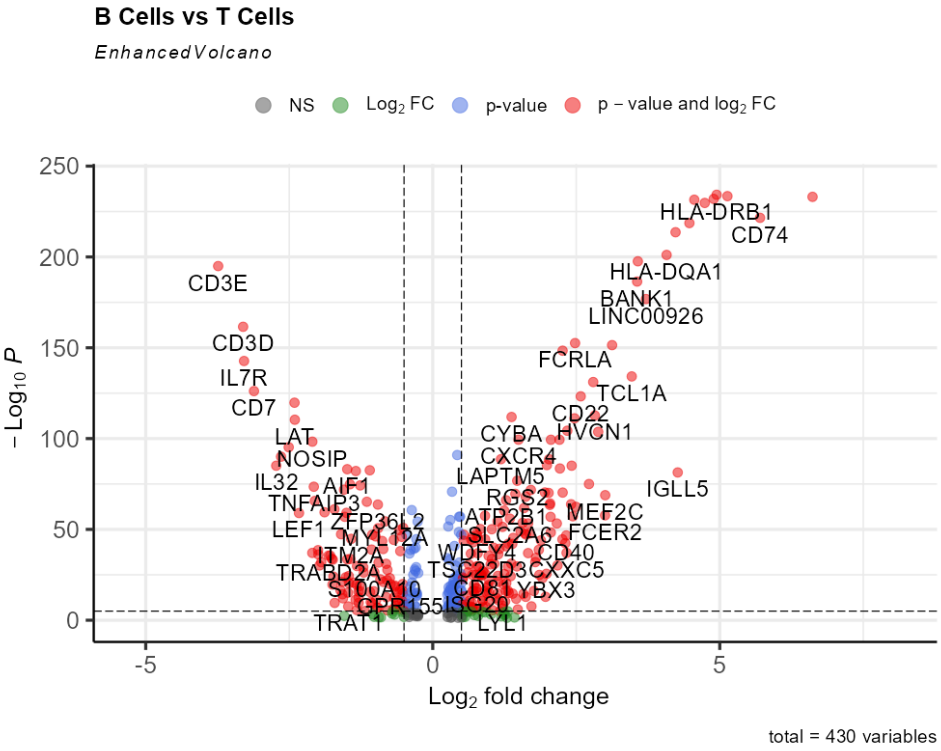

Figure 50: Bcells vs T cells

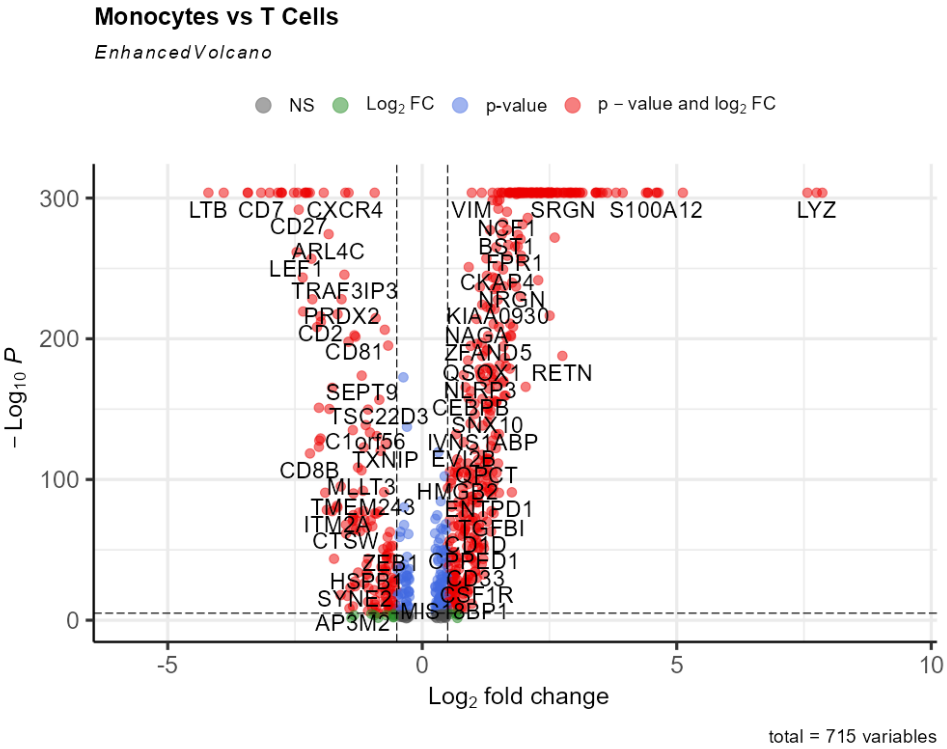

Figure 51: T-cells vs Monocytes

Q. How do naive Tcells differ from memory T cells?

Memory T cells have previously encountered their specific antigen and can mount faster, thus stronger immune responses upon re-exposure to the same pathogen, while naive T cells have never encountered their antigen and respond more slowly. When naive T cells first encounter their antigen, some become effector cells for immediate response, while others differentiate into long-lived memory cells that provide fast protection against future encounters with the same pathogen.

Task 7.2: Plot Differentially expressed genes

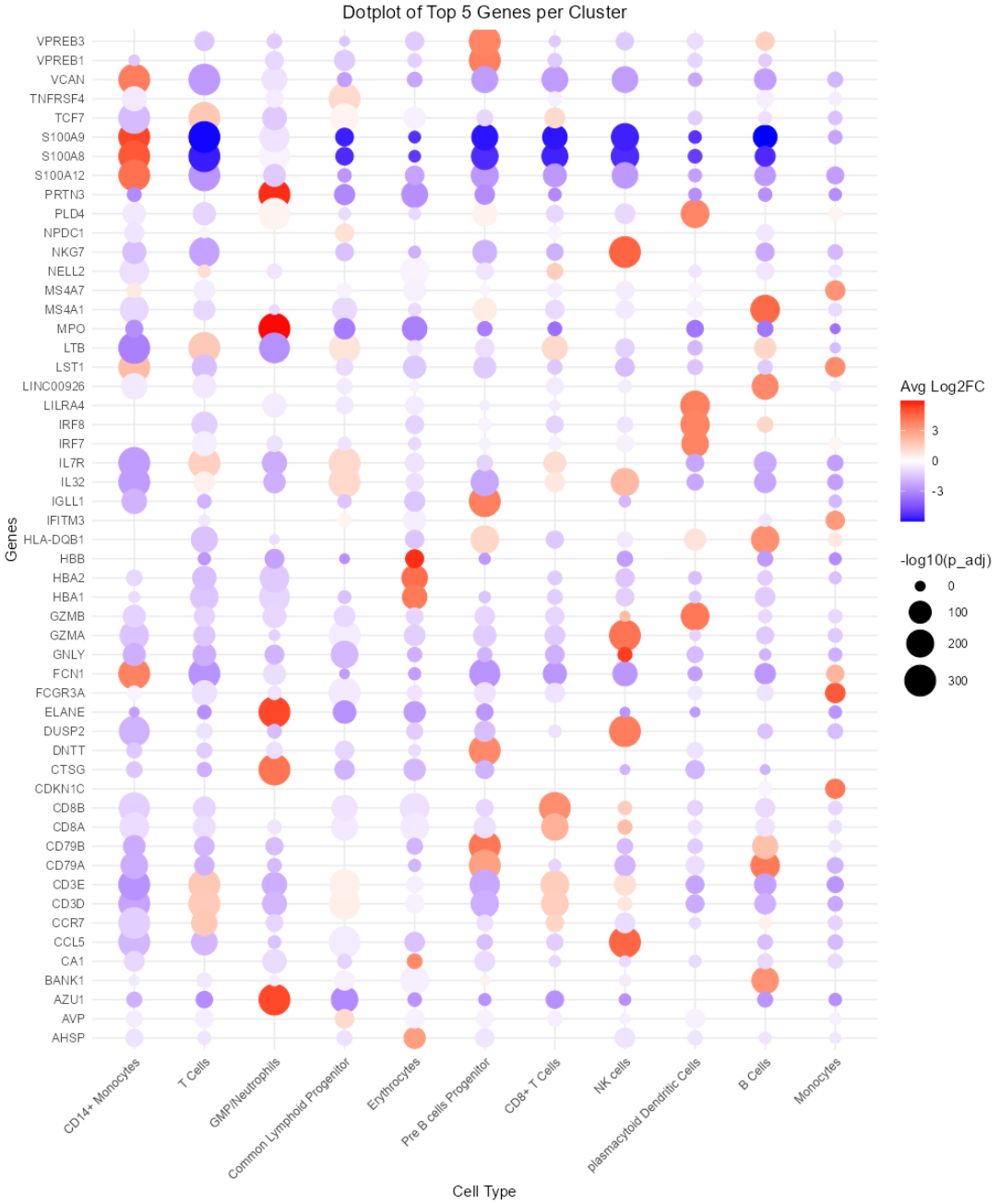

Figure 52: T-cells vs Monocytes

Task 8.1: Differential Expression Analysis on groups

Gene	avg_log2FC	p_val_adj
GNLY	2.16	0.00
S100A12	1.99	0.00
S100A9	1.92	2.36E-150
S100A8	1.75	3.43E-168
CCL5	1.61	0.00

Table 1: BMMC vs CD34+ Differentially expressed genes

Gene	avg_log2FC	p_val_adj
FCGR3A	1.19	3.88E-101
S100A12	1.14	1.07E-32
NCF1	1.02	3.15E-25
CD14	0.95	1.77E-40
HBB	0.92	2.19E-12

Table 2: BMMC vs CD34+ Differentially expressed genes (Monocyte cells only)

Task 8.2: Pathway analysis on groups

Pathway with the lowest p-value in BMMC: Neutrophil Degranulation
Pathway with the lowest p-value in CD34+: DNA metabolic process

Biological meaning

BMMC samples contain mature immune cells, including neutrophils, which are equipped for immediate immune responses - so the prominence of neutrophil degranulation pathways. These cells are fully differentiated and ready to perform immune functions.

CD34+ samples are enriched for stem and progenitor cells that are actively dividing, self-renewing, and differentiating into various blood cell types. This explains why DNA metabolic processes are the top pathway, as these cells require DNA replication and repair activity for their developmental processes rather than immune response functions.

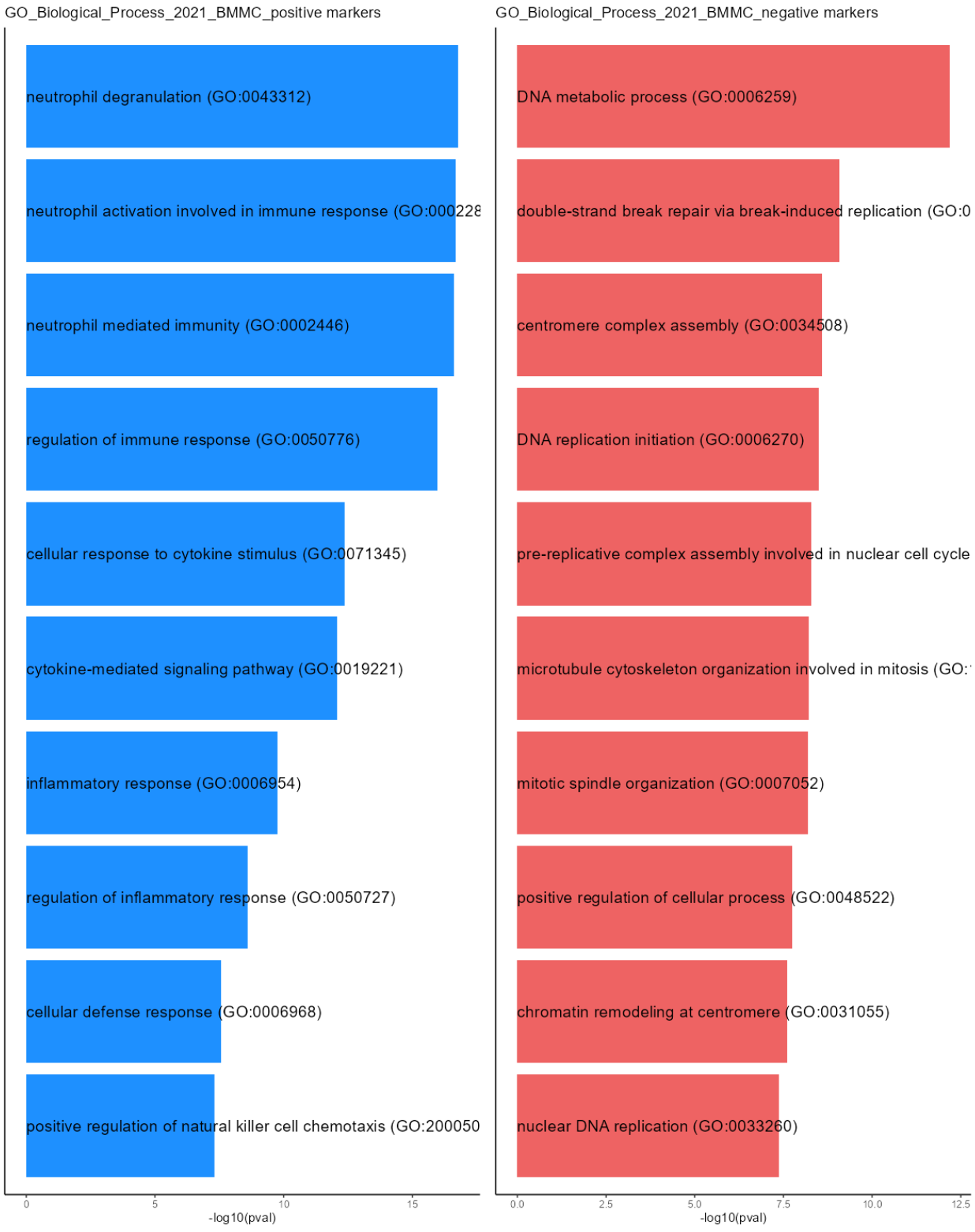

Figure 53: Pathway analysis using EnrichR

Task 9: Trajectory Analysis

Why is this a good group to do trajectory analysis? Which other group do you think may be a good choice?

The specific trajectory that was chosen for analysis was:
Common Lymphoid Progenitor → Pre B cells Progenitor → B Cells

Primarily because it represents one of the most well-defined and crucial developmental pathways in hematopoiesis. This trajectory captures the progression of cells from early lymphoid progenitors to fully mature B cells.
Common Lymphoid Progenitor serve as the starting point, representing the earliest stage of lymphoid development, while Pre B cells mark a critical intermediate stage where cells become increasingly committed to the B cell lineage. The final stage, mature B cells, represents the endpoint of this development. Also, looking at the cell type UMAP again we can see a clear spatial connectivity, suggesting a developmental relationship in our dataset as well.

Another group that maybe a good choice:
GMP/Neutrophils → CD14+ Monocytes
This also shows good spatial connectivity in the UMAP visualization, and would represent the myeloid lineage development, which is another major branch of hematopoiesis.

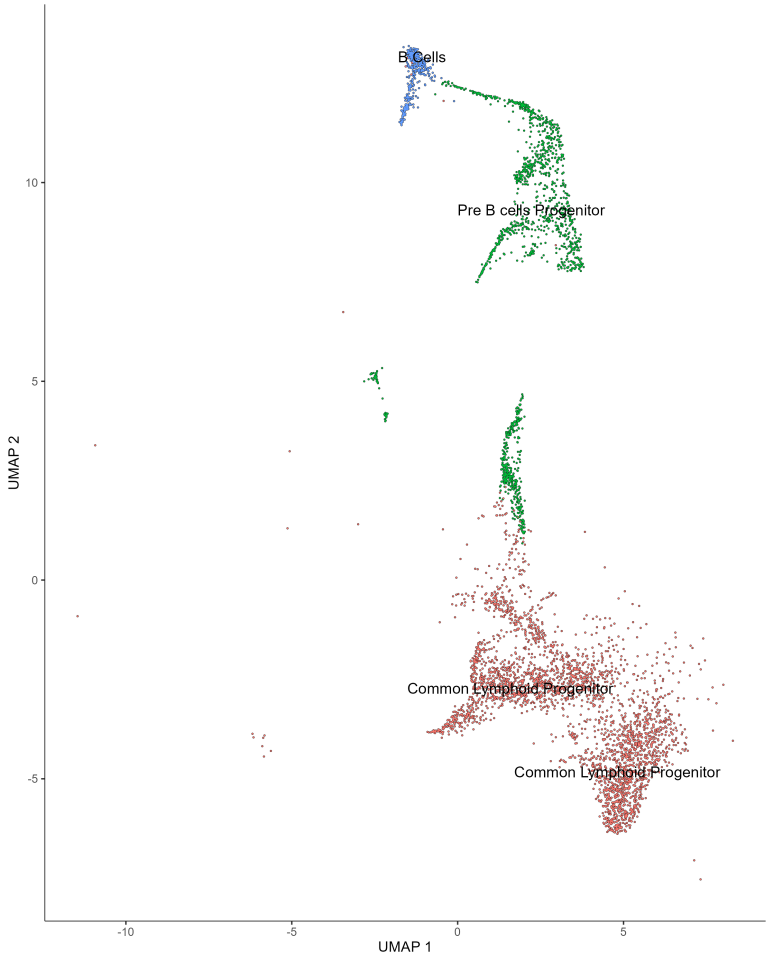

Figure 54: Selected clusters for trajectory analysis

Task 9.2: Select root-nodes manually

Why is the selection of the root nodes important for the algorithm?

The selection of root nodes in Monocle3 is important because it determines the starting point and direction of cell development in the trajectory analysis. This choice affects how pseudotime is calculated and how gene expression changes are interpreted. If we select an incorrect root node, it can mislead the entire developmental timeline, which would be a misinterpretation of the actual biological processes and cellular relationships.

Which points are a good choice for root nodes of the analysis and why?

When manually selecting root nodes for trajectory analysis, we should look for cells that are positioned at the extreme ends or "tips" of the Common Lymphoid Progenitor (CLP) cluster in the UMAP. These specific points are ideal choices because they represent the earliest cells in the developmental progression. We should focus on cells located at the outermost region of the CLP cluster that connects towards the Pre B cells and other lymphoid populations, as these points likely represent the true starting population of cells.

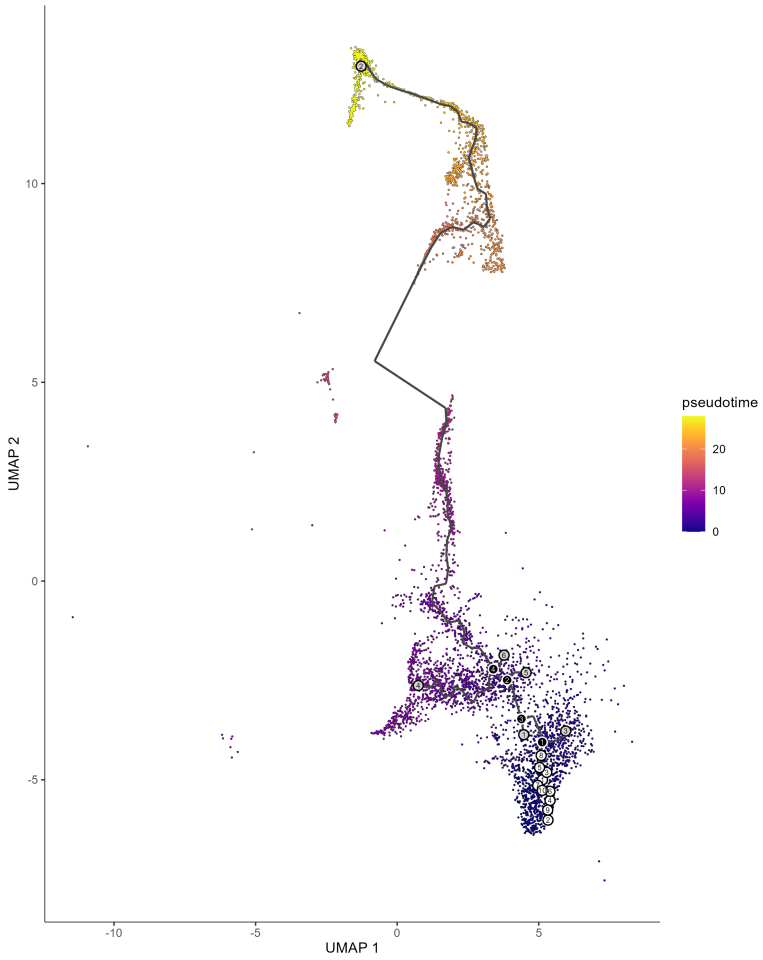

Figure 55: Pseudotime trajectory (Manual nodes selection)

Task 9.3: Select root-nodes automatically

Try to Select root nodes automatically. Did it improve the results? Explain why.

Automatic root selection was performed by using Monocle3's get_earliest_principal_node function, as defined in it's documentation. Since we don't have a time-based column in our dataset, the function was modified to use **cell_ids** of the earliest/progenitor state. Which in our case was the CLP cells. The automatic and manual root node selections produced essentially identical results. The trajectory structure is the same in both plots and the pseudotime coloring is also consistent between both. So the algorithm successfully identified the biologically relevant starting points.

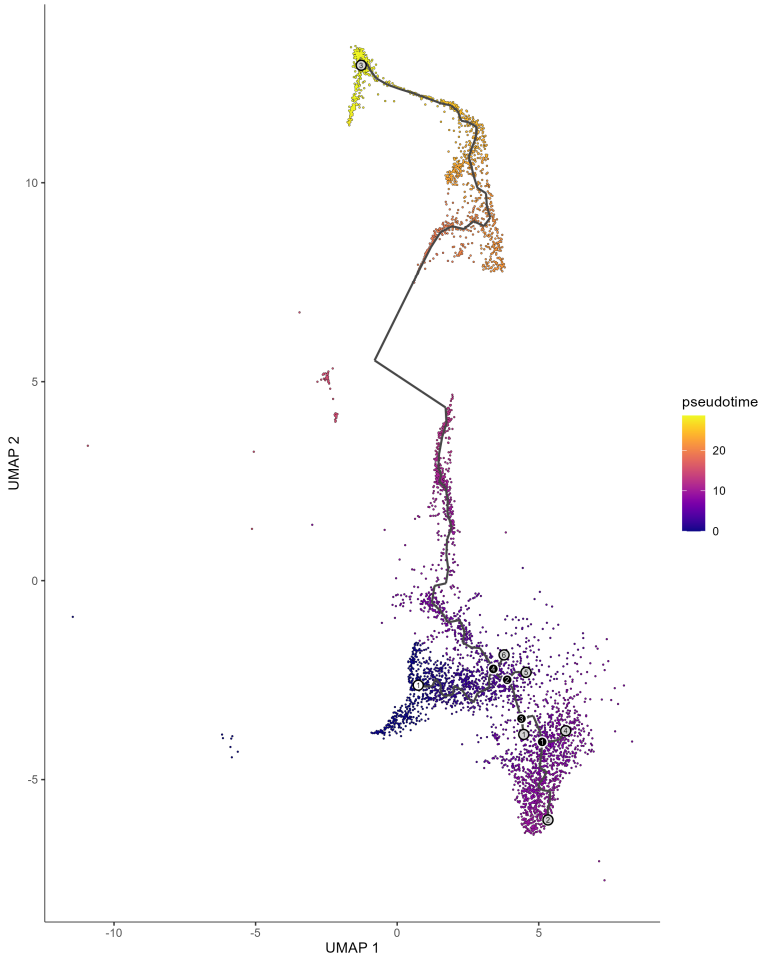

Figure 56: Pseudotime trajectory (Automatic nodes selection)

Task 10: Cell-Cell Communication

Find the signaling pathways that can be found in both groups.

- MIF Pathway
- CD99 Pathway
- APP (Amyloid Processing) Pathway
- GALECTIN Pathway
- ITGB2 (Integrin) Pathway

Show the number of interactions and the interaction strength for each group.

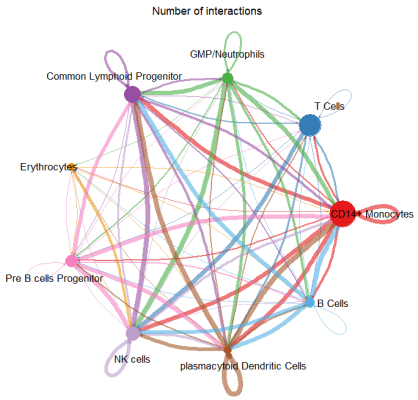

Figure 57: Number of Interactions (BMMC Group)

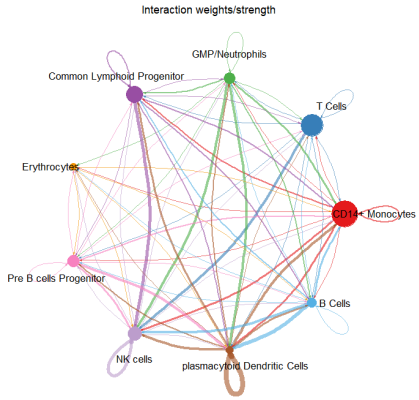

Figure 58: Interaction weight/strength (BMMC Group)

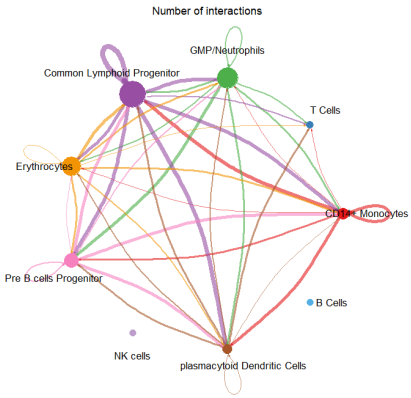

Figure 59: Number of Interactions (CD34 Group)

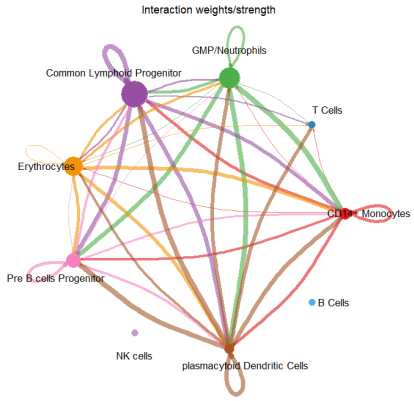

Figure 60: Interaction weight/strength (CD34 Group)

Choose one pathway, and display the results in a circle plot for each group (BMMC).

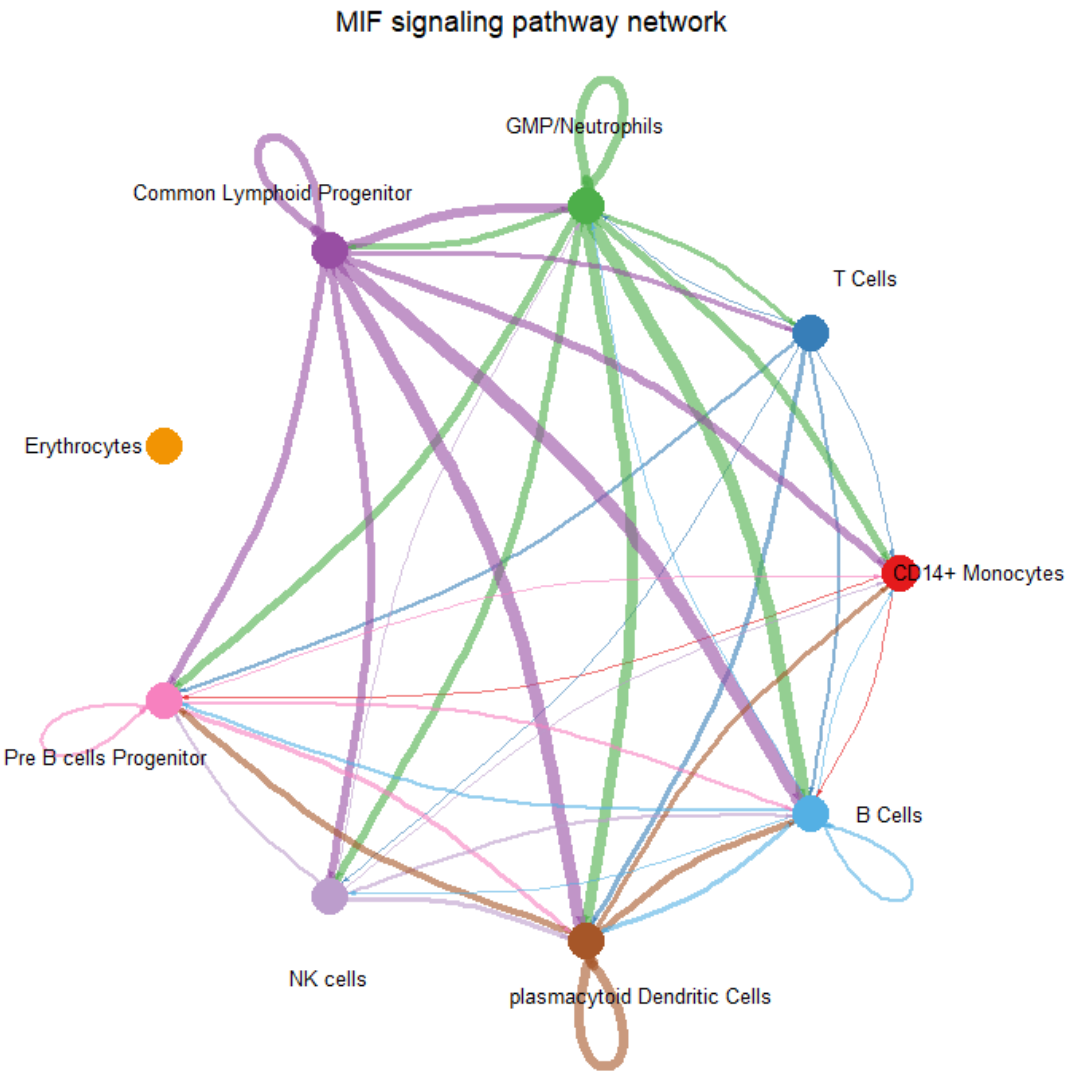

Figure 61: MIF Pathway Circleplot

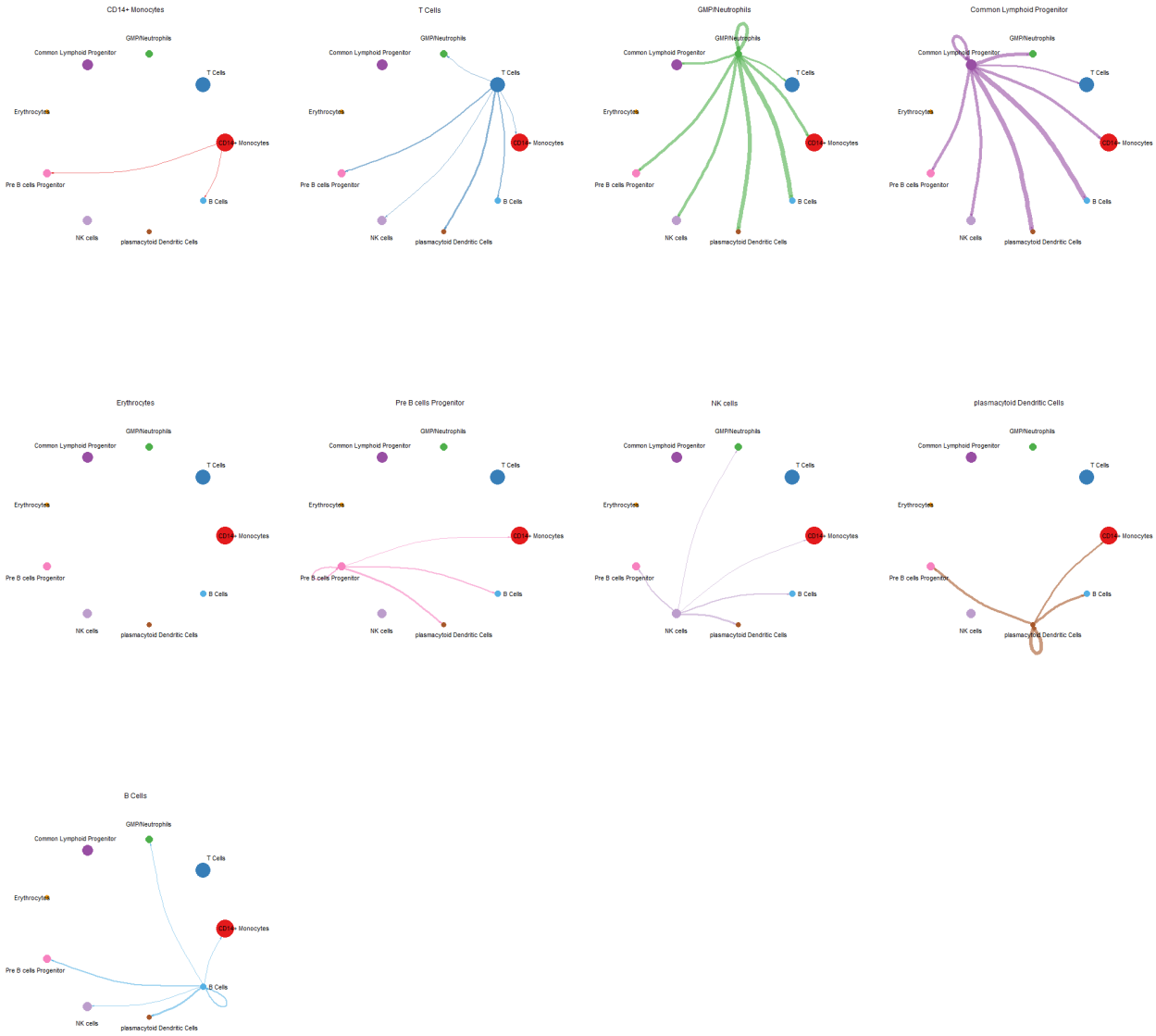

Figure 62: MIF Pathway Circleplot (per cell group)

Task 11: Summary

The integrated analysis of bone marrow samples revealed distinct cellular compositions and developmental states between regular bone marrow (BMMC) and CD34+ selected populations. The methods used for quality control and batch correction successfully mitigated technical variations while preserving biological signals, which made the data reliable for downstream analysis. The clustering analysis identified 14 distinct cell populations, with BMMC samples showing a broader spectrum of mature immune cells, while CD34+ samples were enriched for progenitor populations, confirming successful CD34+ selection. Differential expression analysis highlighted key marker genes distinguishing cell types, particularly between B cells, T cells, and monocytes. Notably, pathway analysis revealed that BMMC samples were enriched for immune response pathways (particularly neutrophil degranulation), while CD34+ samples showed enrichment for DNA metabolic processes, reflecting their distinct biological roles. The trajectory analysis of B cell development (CLP $\rightarrow$ Pre-B $\rightarrow$ B cells) provided insights into lymphoid differentiation paths. And lastly, cell-cell communication analysis identified several active signaling pathways, particularly MIF signaling.

Alternative approaches could include:

- Using Seurat's SCTransform function for normalization.
- Using different integration methods like Harmony or BBKNN.
- Applying velocity analysis (RNA velocity).
- Replicating the analysis in Python specific tools, including Scanpy and SCVI.
- Using deep learning approaches for cell type annotation. (scANVI, scMMT)