Deciphering the immune cell dynamics in Breast Tumor Microenvironment
BMI 710 Final Project
Cheryl Gu
May 5th, 2024

Introduction

The diverse immune cells within the tumor microenvironment (TME) play critical roles in cancer progression and responses to immunotherapies. Recent research has highlighted that immune cells in non-lymphoid tissues would influence the outcomes in cancers such as melanoma, lung, and kidney cancers, but they have not yet fully understood in breast cancer (Topalian et al., 2015; Fan and Rudensky, 2016). Building on the previous study, we aimed to further uncover the immune cell phenotypes in breast cancer through a comprehensive single cell RNA sequencing profiling of 45,000 immune cells from eight breast tumors, with the matched normal breast tissue, blood, and lymph node samples. (Azizi et al., 2018)

The tumor tissue samples from the dataset were collected from women undertaking primary breast cancer surgery, while the normal tissues were gathered from same patients during contralateral prophylactic mastectomy surgeries, and the peripheral blood mononuclear cells (PBMCs) were collected before they had surgery. They dissociate and isolate the tissue by first cutting the specimens into 1 mm^3 pieces, and then used the tissue Liberase TL to enzymatically digest the tissue fragments. After that, the digested tissue was filtered washed, and stained with an anti-CD45 antibody and DAPI in order to differentiate live and dead cells. Once stained, cells are sorted using fluorescence-activated cell sorting (FACS) on a FACSARIA sorter, which allows for the high-precision sorting of viable immune single cells prepared for the downstream RNA sequencing. (Azizi et al., 2018)

Methodology

To ensure the integrity and reliability of our single-cell RNA sequencing data, we implemented a rigorous quality control (QC) process. First, we assessed the quality of the data by examining three key metrics: the number of detected genes per cell (nFeature_RNA), the total count of transcripts per cell (nCount_RNA), and the percentage of mitochondrial transcripts (percent_mito). *Using violin plots for visualizing these three metrics, we filtered out the cells expressing too high number of genes (<500), as low values indicating dying cells or insufficient capture, and the cells with too high percentage of mitochondrial gene (>10%) as high levels of mitochondrial indicating apoptotic or damaged cells. (Figure 1)*

After QC, we addressed for the potential batch effect variables by plotting them on a Uniform Manifold Approximation and Projection (UMAP) to visualize how each cluster are differentiate across samples. The variable we chose is "patient", since individual patients would give intrinsic biological differences that affect gene expression, such as genetic background or age or other health conditions other than breast cancer. *From the pre-harmony UMAP, we observed that the distinct clusters form very likely due to the batches by “patient”, as the clusters are not mixing very well, indicating the batch effects caused by “patient” variable. After applying harmony integration to the data, the confounding effects due to batches are removed and the clusters seem more integrated and mixed by diverse colors, representing each cluster are not separated by “patient” variable anymore. (Figure 2)*

The data processing pipeline for the scRNA-seq data analysis includes normalization, feature selection and scaling, dimensionality reduction, as well as clustering. First, for data normalization, we performed LogNormalize method, which scaled the gene expression measurements by total expression, multiplies this by a scale factor of 10,000 (which is default) and log-transformed the result. Following normalization, we applied feature selection to identify top 2000 most highly variable genes across the dataset, using the Variance Stabilizing Transformation (VST) method, which is particularly effective at identifying genes that exhibit the most significant biological variation across different cell states and conditions.

For dimensionality, we scaled the normalized data and initially employed principal component analysis (PCA) to manage the complexity of the data, as it can reduce the dataset's dimensionality by transforming the original variables into first few principal components that captures most of the variation present in the data. *To determine the optimal number of principal components to retain, we generated an elbow plot from the PCA results. This plot graphically displayed the percentage of variance explained by each principal component. We identified the 18th principal component as the flexion point, indicating that the first 18 components captured the bulk of the variance in the data while minimizing information loss. (Figure 3)* Additionally, considering the potential for batch effects that could skew the analysis, we utilized batch-corrected principal components for further steps. These batch corrections were performed using Harmony, as mentioned before.

After that, we used UMAP for further reduction and visualization, which allows to explore data structure in a low-dimensional space, enhancing our ability to identify distinct clusters of cells. Finally, for clustering, we used the Louvain algorithm to construct a shared nearest neighbor (SNN) graph based on the Euclidean distance in PCA space. *We chose an appropriate resolution parameter as 0.1 to balance between overclustering and underclustering, ensuring that the clusters were neither too granular nor too coarse. (Figure 4)*

To identify marker genes for each cluster, we conducted differential expression (DE) analysis using the Wilcoxon Rank Sum test by the function "FindAllMarkers" in Seurat, and we conserved the top 10 expressed marker genes for each cluster, and ordered them by their positive log 2-fold change (log2FC) values since positive log2FC correspond to more highly expressed genes, which can allow us to pinpoint genes that were significantly upregulated in specific clusters. The identified markers were then used to annotate the clusters, helping in the interpretation of cell types based on known gene expression profiles. *According to the heatmap of top 10 marker genes in each cluster, we focused on at least two gene signatures per cluster to identify the corresponding cell types (Figure 5).* When annotating cell types for each cluster, we referenced from the original literature like Table S2 and Table S3 from the paper where we got the data from (Azizi et al., 2018), as well as other cell type calling tools like CellMarker 2.0 (http://bio-bigdata.hrbmu.edu.cn/CellMarker/). We then validated by coloring the UMAP by each marker gene expression and plotting violin plots of for each marker gene expression across clusters. *The example pipeline for annotating cluster 3 is that we found "IGHD" and "PAX5"are two top-expressed DE genes as B cell gene signatures, (Table1) we confirmed from both original literature and CellMarker2.0, and then plot UMAP and violin plot by "IGHD" feature. From both UMAP and violin plots, we observed that "IGHD" only expressed very high and densely in cluster 3, which validates our annotation that only cluster 3 represents the B cells in this data. (Figure 6)*

Further Analyses

In this study, we would like to address the biological question: "How does the abundance of lymphocytes, such as T cells and B cells, differ between normal and tumor tissues in breast cancer?" Therefore, we conducted differential abundance of T cells and B cells and the corresponding statistical testing to understand the insights of lymphocyte abundance differences within normal and tumor tissues.

To analyze differential abundance, we first annotated the identified cell clusters with their respective cell types. We then calculated the proportion of each major immune cell type, focusing on T cells and B cells, across different tissue types (tumor vs. normal). Initial findings revealed that there was no significant difference in the proportion of T cells between normal and tumor tissues, suggesting a stable presence of T cells across these conditions. However, *when we extended our analysis to B cells, the results indicated a statistically significant difference in their abundance between normal and tumor tissues as shown in the bar plot. (Figure 7) This was confirmed through a Chi-squared test, which returned a p-va*lue < 2.2e-16, *highlighting a marked differential abundance of B cells.*

Since the original literature also identified cell heterogeneity within the tumor microenvironment, particularly among myeloid cells, by utilizing diffusion maps. The diffusion maps revealed four primary branches, indicating varied cell states among these cells. This motivates the idea for conducting a trajectory analysis using Monocle 3 with the same data. In specific, since we do not have a time-specific variable to do that, we will use "tissue" and "cell type" instead since it contains lymph nodes region location, which is often where cell starts to differentiate. Then, to conduct trajectory inference, we employed Monocle 3 that calculated pseudotime time to order cells and to construct a map demonstrating the cell trajectories. *We chose monocytes from lymph site as the precursor cell, and then calculated the inferred developmental trajectory as shown in the trajectory plot. We also plot the monocyte and dendritic cell, pDC gene signature on the trajectory plot to further observe the path that monocytes differentiated from lymph region to tumor region and developed into DC and pDC. (Figure 8)*

References

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti, R. C., Kadaveru, K., Mazutis, L., & Rudensky, A. Y. (2018). Single-cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell, 174(5), 1293. https://doi.org/10.1016/j.cell.2018.05.060

Fan, X., & Rudensky, A. Y. (2016). Hallmarks of Tissue-Resident Lymphocytes. Cell, 164(6), 1198–1211. https://doi.org/10.1016/j.cell.2016.02.048

Topalian, S. L., Drake, C. G., & Pardoll, D. M. (2015). Immune checkpoint blockade: A common denominator approach to cancer therapy. Cancer Cell, 27(4), 450. https://doi.org/10.1016/j.ccell.2015.03.001

Figures and tables

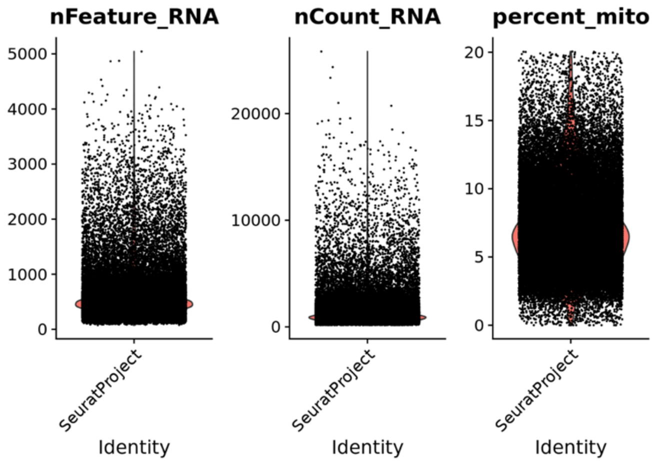
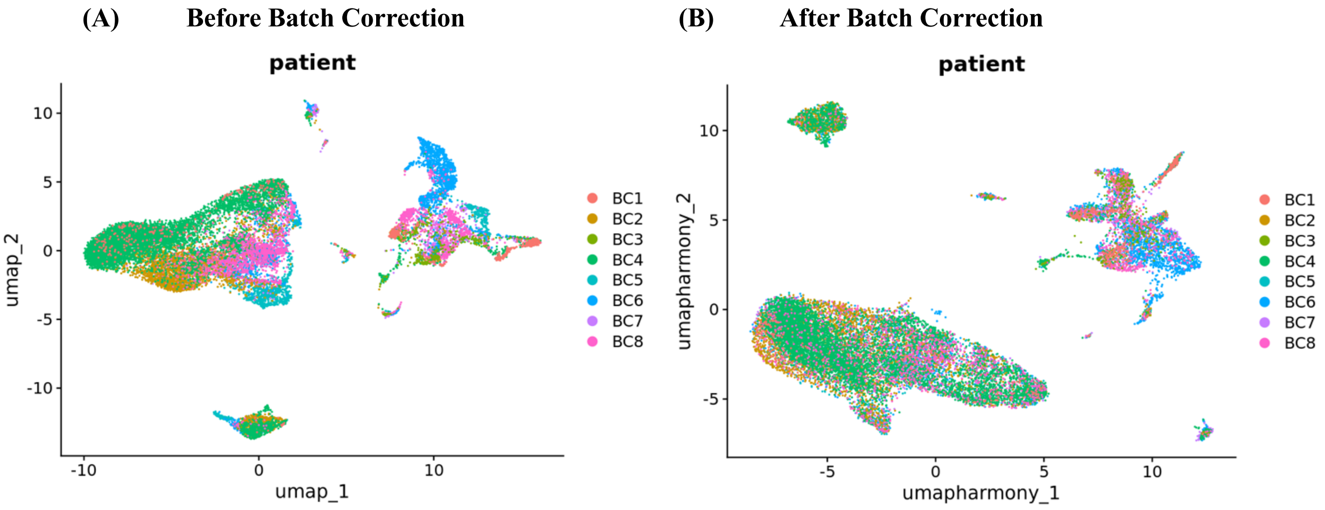


Figure 1: Quality Control Metrics for Single-Cell RNA Sequencing Data. Quality control analysis of single-cell RNA-seq data visualizing key metrics in violin plots used to assess cell viability and data integrity. Cells with extreme values in these metrics were excluded from subsequent analyses. (Left: nFeature_RNA, displays the number of genes detected per cell across the dataset. Middle: nCount_RNA: displays the total mRNA transcripts counted per cell. Right: percent_mito, displays the percentage of mitochondrial versus nuclear genes.)


(A) Before Batch Correction

(B) After Batch Correction

Figure 2: UMAP Visualization of Single-Cell RNA-seq Data Before vs. After Batch Correction. (A) UMAP plot demonstrates the clustering of single-cell RNA-seq data from eight breast cancer patients (BC1 to BC8) before batch correction. Each color represents cells from a different patient, highlighting distinct clusters formed due to batch effects, which may obscure true biological variation. **(B)** Post-Harmony batch correction, the UMAP plot illustrates a more integrated clustering of cells from the same eight breast cancer patients (BC1 to BC8). Batch correction has

minimized technical disparities, allowing for the visualization of clusters based more closely on biological similarities rather than technical differences.

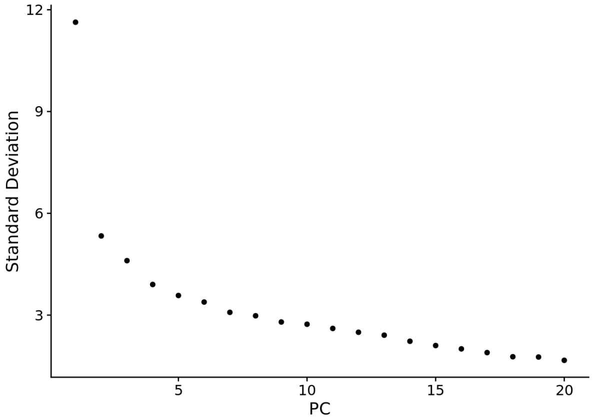


Figure 3: Elbow Plot of Principal Component Analysis. Elbow plot depicts the standard deviation of the first 20 principal components (PCs) derived from the PCA of single-cell RNA sequencing data. The plot is used to determine the optimal number of principal components to retain for downstream analyses. A noticeable drop in standard deviation after the first few components suggests diminishing returns on additional component. This visualization helps selecting a suitable number of PCs to capture significant biological variation without overfitting the model.

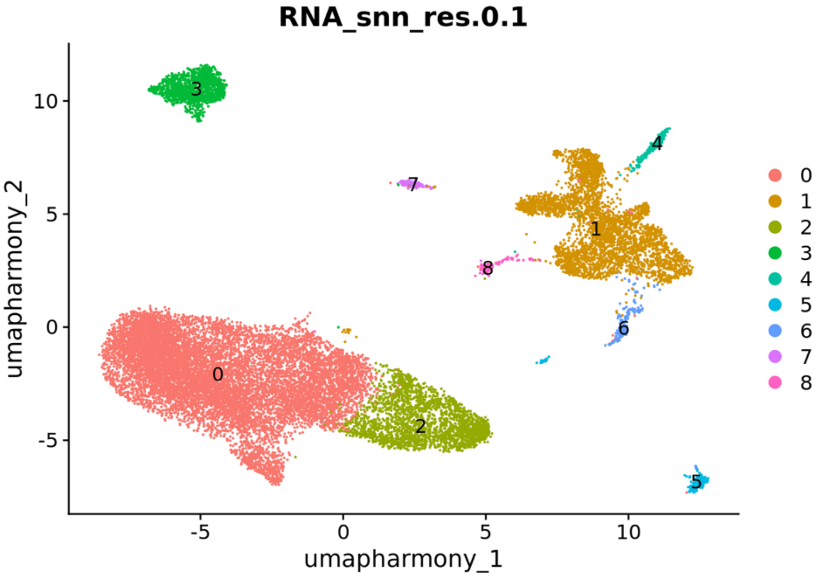


Figure 4: UMAP Visualization of Clustered Single-Cell RNA Data. This UMAP plot presents the clustering results of single-cell RNA sequencing data, processed using batch-corrected principal components and visualized at a Louvain resolution of 0.1. Each color represents one of nine distinct clusters (0 to 8).

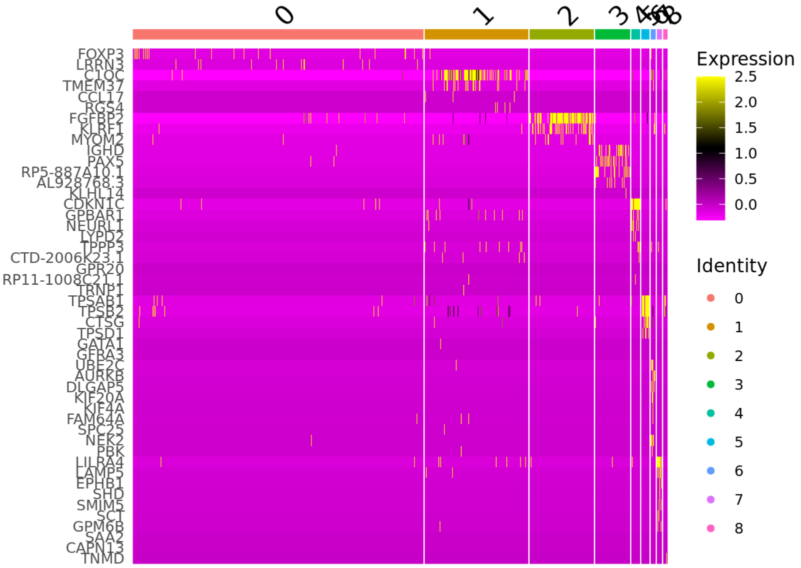


Figure 5: Heatmap of Differential Gene Expression Across Clusters. This heatmap visualizes the differential expression of selected genes across nine identified cell clusters (0 through 8) in single-cell RNA sequencing data. Each row represents a gene, and each column represents a cluster, with the color intensity indicating the level of gene expression (from low in purple to high in yellow).

(A) UMAP visualization **(B). Violin plot visualization**

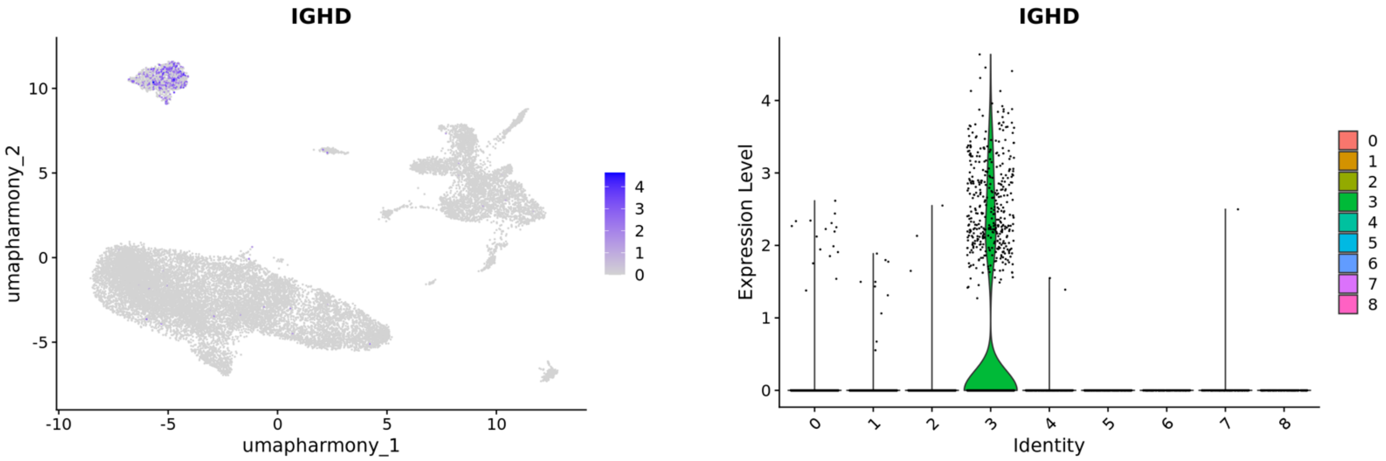


Figure 6: Expression Analysis of IGHD in Single-Cell RNA-seq Data. (A) Cells expressing IGHD are highlighted on the UMAP plot, showing spatial distribution and expression level of marker gene IGHD. **(B)** Violin plot depicting IGHD expression levels across different clusters (0 through 8), with each cluster represented by a distinct color. The green fill in the plot highlights the expression profile within cluster 3, designated as B cells, where IGHD is notably expressed, marking it as a specific marker for this cell type.

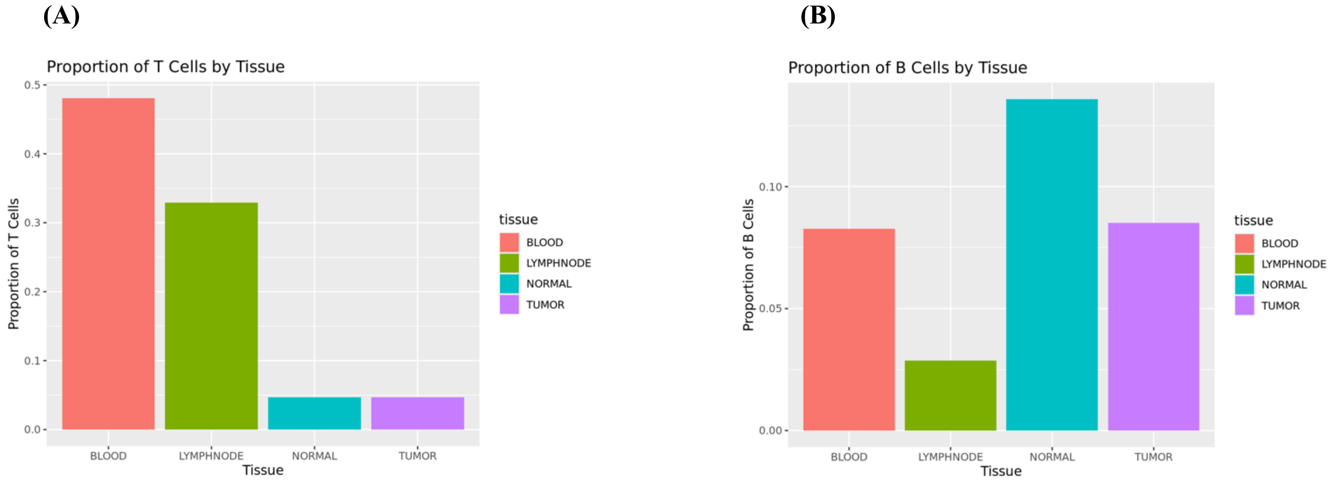
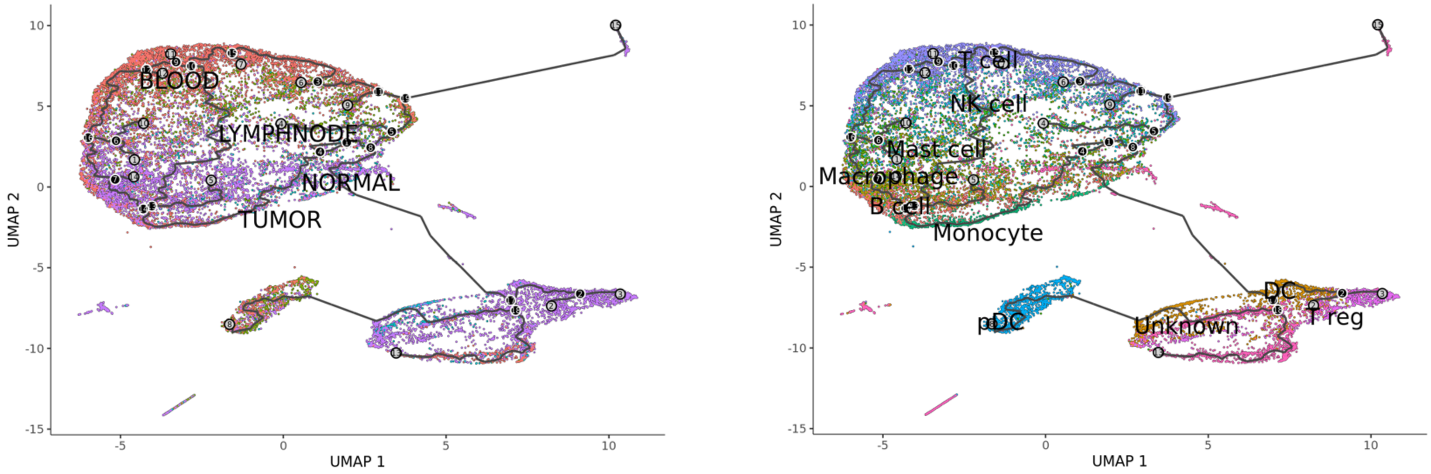
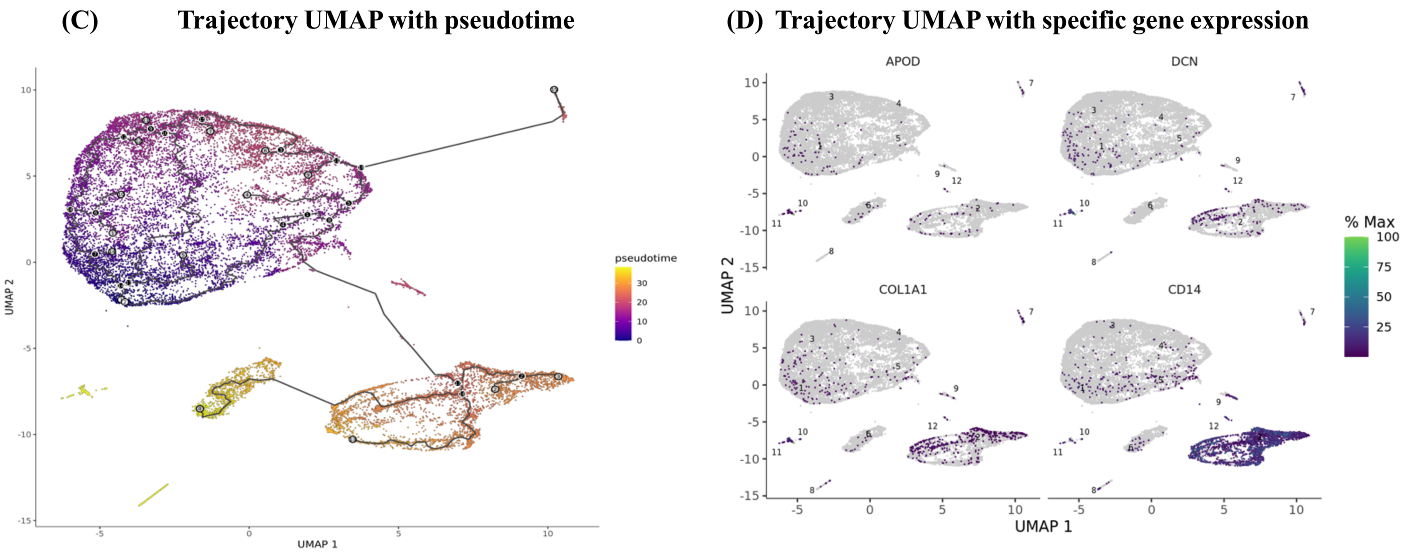


Figure 7: Differential Abundance of T Cells and B Cells Across Tissue Types. (A) Bar plot shows the distribution of T cells, which are notably higher in blood and lymph nodes compared to normal and tumor tissues, suggesting a no significant differences of T cell abundance in both tumor vs. normal states. **(B**) Bar plots shows that B cells are predominantly abundant in lymph nodes, followed by a significant presence in normal and tumor tissues. The marked increase in B cells within tumor tissues compared to normal suggests potential involvement in the local immune response or possibly in the immunopathology of breast cancer.

(A) Trajectory UMAP by tissue **(B) Trajectory UMAP by cell type**





(C) Trajectory UMAP with pseudotime

(D) Trajectory UMAP with specific gene expression

Figure 8*: Comprehensive UMAP Visualizations of Cell Trajectories and Gene Expression in Breast Cancer. (A)** *UMAP plot annotated with tissue types, illustrating the distribution and clustering of cells from different tissue sources: blood, lymph node, normal, and tumor.* ***(B) *UMAP plots annotated with cell types, including T cells, NK cells, B cells, monocytes, macrophages, mast cells, dendritic cells (DCs), plasmacytoid dendritic cells (pDCs), and T regulatory cells (T regs).* ***(C)*** *A trajectory analysis overlaid on a UMAP visualization, where cells are color-graded based on their pseudotime values, indicating the inferred developmental progression from a starting point (presumably in a lymphoid site) toward differentiated states.* ***(D)*** *Focused gene expression plots for selected marker genes APOD, DCN, COL1A1, and CD14 across the UMAP landscape. Each subplot represents the expression level of one marker, highlighting areas of high expression (darker shades) related to specific cell states or clusters.*

Cluster	Cell Type	Marker Genes
0	T cell	MAL, FOXP3, RTKN2
1	Macrophage	C1QC, TMEM37
2	NK cell	FGFBP2, KLRF1, MYOM2
3	B cell	IGHD, PAX5
4	Dendritic Cell (DC)	CDKN1C, GPBAR1, NEURL1
5	Mast cell	TPSAB1, TPSB2, CTSG
6	Plasmacytoid DC (pDC)	LRRC26, CLEC4C
7	T regulatory (T reg)	UBE2C, ASPM
8	Monocyte	APOD, DCN

***Table 1*: Differentially Expressed Gene Markers by Cluster.** *This table lists the primary marker genes identified for each cluster derived from differential expression analysis. The genes listed are those with notably high expression in their respective clusters, serving as distinctive molecular signatures for cell type annotation.*