BIOS 785 Project Report: Exploring Human Breast Cancer Atlas scRNA-Seq Data
Team members: Seowoo Kim, Taeim Kwon, Arthi Hariharan

Introduction

Cancer is a group of diseases which is marked by the hallmark features of Uncontrolled proliferation, angiogenesis, metastasis and evading cell death. Recent research has shown that the tumor microenvironment plays an important role in determining the fate of the cancer cell.

Breast cancer has a heterogenous microenvironment. Immune cells such as T-cells, B-cells, NK-cells, Neutrophils, Macrophages, Fibroblasts, Mesenchymal cells and other stem cells are present along with normal and cancerous epithelial cells which comprise the Breast cancer Tumor Microenvironment (Anderson & Simon,2020). Breast cancers are of different subtypes based on the presence or absence of certain gene signatures. Mutations in the BRCA gene are frequently associated with hereditary breast cancer. Breast cancer is clinically classified based on the presence of estrogen receptor (ER), progesterone receptor (PR) and expression of HER2. It is broadly classified as the following subtypes: Luminal (ER+,PR+/-), HER2+ and Triple negative (ER-,PR-,HER2-) (Wu et. al,2021). The different subtypes of cancer are known to exhibit different prognostic markers and respond to treatment differently. Treatments for breast cancer target metastasis, angiogenesis and the tumor microenvironment. It is essential to understand the underlying differences in the gene expression across the different subtypes and the interactions among the cells in the tumor microenvironment to develop effective treatment strategies.

Metastasis and Angiogenesis are commonly targeted hallmarks for cancer treatment. Angiogenesis can be VEGF-mediated or non-VEGF mediated (such as PDGF family, FGF family) and monoclonal antibodies such as Bevacizumab target VEGF. Chemokine family genes are known to play a role in metastasis and inflammation and also influence angiogenesis (Bruno et al., 2014). Chemokines are emerging drug targets for different types of cancers (Poeta et.al , 2019). Wu et al. report that chemokines play a role in tumor promotion and immune regulation and are also markers of metastasis in cancer cells. Therefore, we look at *CXCR* family genes, *VEGF* family genes and *PDGF* family genes to see their differential expression across subtypes.

In this project, we aim to identify the differential expression of genes across breast cancer subtypes, focusing on cancer hallmarks – metastasis and inflammation markers (*CXCR family*) and angiogenesis markers (*VEGF*, *PDGF family*). We also aim to show a gene co-expression network in the cells involved with tumor microenvironment, particularly for mesenchymal cells.

Here, we adapted the sc-RNA Seq dataset from Wu et al. (2021). They generated a breast cancer atlas with sc-RNA Seq and Bulk RNA Seq data from 26 primary untreated tumors collected from patients (11 ER+, 5 HER2+ and 10 TNBC). Single Cell RNA Sequencing was performed using 10X Chromium with a total of 5000 to 7000 cells targeted per well. Libraries were sequenced on NextSeq 500 Platform (Illumina) with pair-ended sequencing and dual indexing. A total of 26 cycles for Read 1, 8 for i7 index and 98 cycles for Read 2 were performed. The raw bcl files demultiplexed and mapped to reference genome GRCh38 using Cell Ranger Single Cell software. *EmptyDrops* method from the DropletUtils Package applied for cell filtering with additional cutoffs for cells with a gene and UMI count greater than 200 and 250 respectively and a mitochondrial percentage less than 20%. The authors used Seurat V3.0.0 for data normalisation, dimensionality reduction and clustering. The clusters were annotated using

Garnett Method (v 0.1.4). In the original paper, the authors provide an comprehensive assessment of the breast cancer tumor environment and spatially map tumor heterogeneity.

Methods

Data preprocessing

The dataset consisted of barcodes.tsv, genes.tsv, count.mtx, and metadata, which were read in RStudio and used to set up a Seurat object with Seurat v.4.0.0. From the raw dataset, the initial object of class Seurat has 29,733 features across 100,064 samples. First we created the Seurat object and conducted Quality Control. We subset the samples by giving conditions to samples that nFeature_RNA is bigger than 250, nCount_RNA is bigger than 500 and percent.mt is smaller than 20. 1,777 samples were filtered out after this pre-processing. We then normalized the object with the 'LogNoramlize' method to select the highly variable features and scaled the data. After pre-processing, we ran PCA. We used an 'elbow plot' to determine the dimensionality of the dataset to determine the dimension for UMAP. We used default parameters for normalization, scaling, feature selection, and dimensionality. To demonstrate the UMAP consistently, a seed was set. After determining the dimension, we performed the annotation based on the metadata. Then we reproduced the UMAP plot and Feature plots in the paper with manual annotation using FindAllMarkers().

Differential Expression Analysis

Differential Expression analysis was performed using the in-built functions in the seurat package. The idents were renamed to the subset given in the metadata to use for finding markers. The 'FindMarkers' function in the Seurat package was used to perform testing gene expression across the subtypes. Highly differential genes were identified comparing HER2+ vs TNBC, ER+ vs HER2+ and TNBC vs ER+. 'FindAllMarkers' function was used to find highly expressed genes across all three subtypes. The seurat function uses Wilcoxon Rank sum test to identify genes which are expressed differentially. The output includes the log fold-change of the average expression between the two groups (avg_log2FC), the percentage of cells expressing the marker in the two groups (pct.1 and pct.2) and adjusted p-value based on bonferroni correction.

The differential expression of genes involved in metastasis and angiogenesis, specifically *CXCR* family genes, *VEGF* family genes and *PDGF* family genes, were tested across subtypes using the 'features' option in the 'FindMarkers' function.

'DEEnrichRPlot' function of seurat to identify biological functions related to the differentially expressed genes. MSigDB_Hallmark_2020 EnrichR database is used for enrichment analysis in all three subtypes of breast cancer.

Co-expression Network Analysis

The package called hdWGCNA was employed for the co-expression network analysis on the Seurat object after initial processing. The object was run to create a hdWGCNA experiment within the Seurat object such that genes were selected if they were expressed in at least 5% of cells. Then metacells, collections of similar cells coming from the identical biological sample in smaller groups, were constructed with the k-Nearest Neighbors (kNN) algorithm. We used k = 25, producing a metacell gene expression matrix based on each cell type.

The experiment for mesenchymal cells that were of the cell type we specifically used as a reference was created based on the metacell expression matrix for downstream network analysis. The testing of soft-power threshold was plotted to determine the optimal soft-power threshold, a key to construct an adjacency matrix of gene to gene correlation. Then the co-expression network based on the soft-power threshold was constructed and used to plot the WGCNA dendrogram for the mesenchymal hdWGCNA experiment. The PCA was performed for harmonized module eigengenes, and eigen-based connectivity was computed to find the hub genes, which were the ones largely connected with other genes in each module. Then the percentage of expression of each module was plotted for the cell types. The hdWGCNA modules underwent enrichment tests so our modules can be visualized more meaningfully in terms of biology, using one of the Enrichr databases, MSigDB_Hallmark_2020.

Results

Reproduction of results

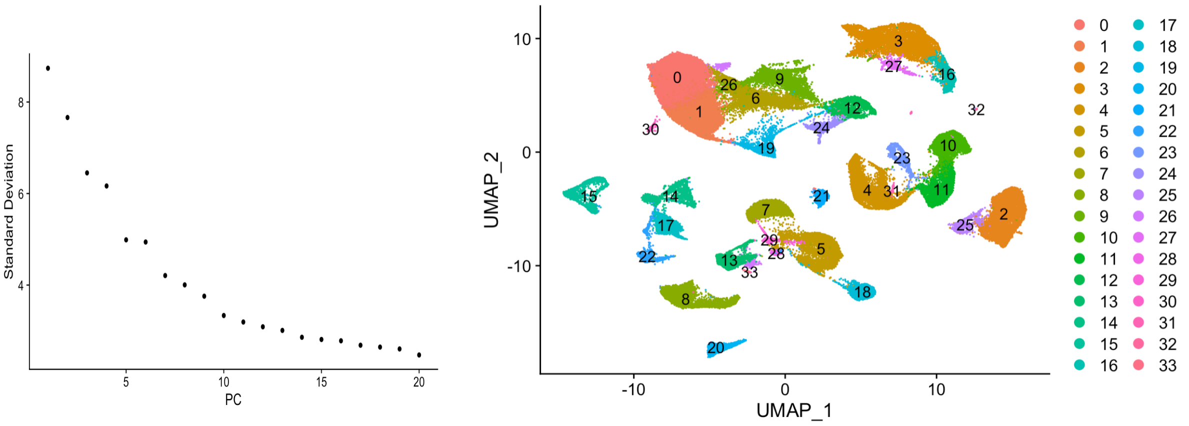

Figure 1. Elbow plot(left) and Reproduced UMAP after cell type annotation(right). Total of 34 clusters, numbered from 0 to 33.

Since the original paper did not provide information about the parameters chosen for dimensionality, Elbow plot was used to determine the dimensionality. Based on the elbow plot we determined the dimensionality as 16. The resulting UMAP had 34 clusters.

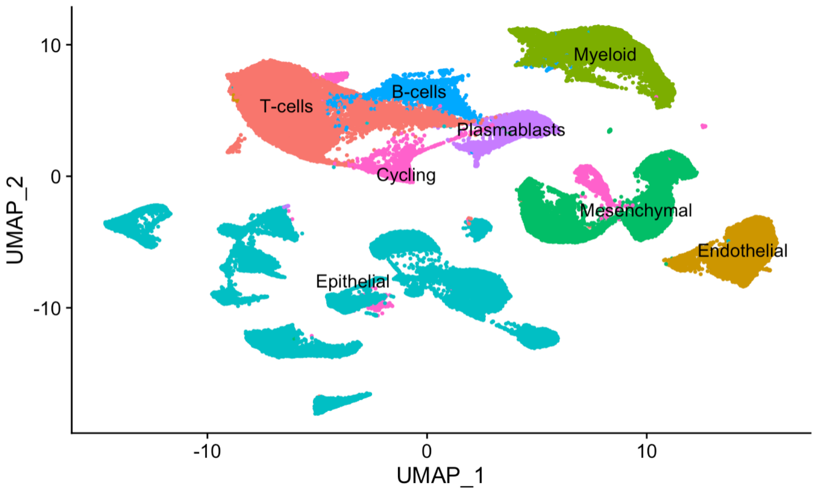

Figure 2. Annotated UMAP plots of major cell types.

Annotation based on metadata did not map all the clusters to a cell type. Therefore, we looked for marker genes for each cluster that were not annotated with information from the metadata to identify the clusters. Comparing the marker genes expressed in each cell, we made additional notes for cell type allocation. For example, if *KRT18* and *KRT8*, which are expressed in epithelial cells but not found in the other cells, were noted from a cluster, the cluster would be renamed as epithelial cells. Through this process, we put allocations on every cell type in the plot. (Figure 2).

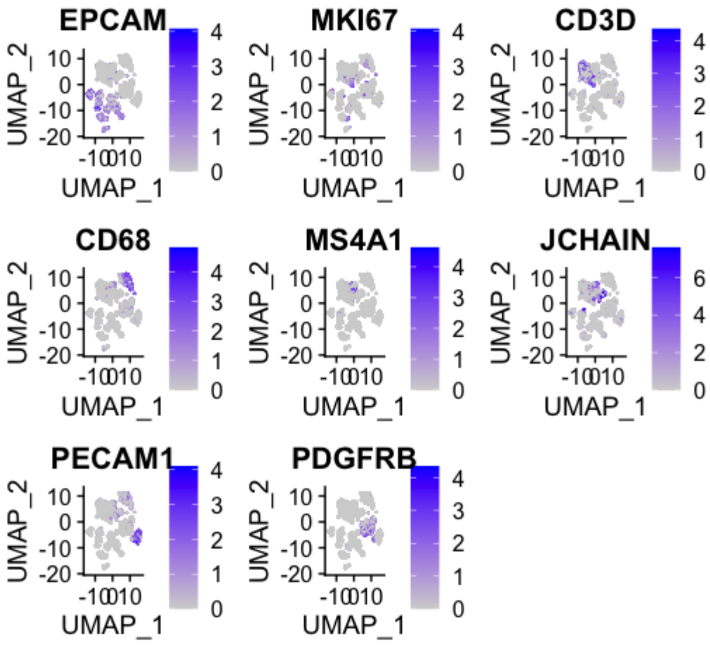

Figure 3. Reproduced feature plots of the eight marker genes used in Wu et al.: *EPCAM, MKI67, CD3D, CD68, MS4A1, JCHAIN, PECAM1, and PDGFRB*. It is observed that expression corresponds to the cell type clusters annotated in Figure 2.

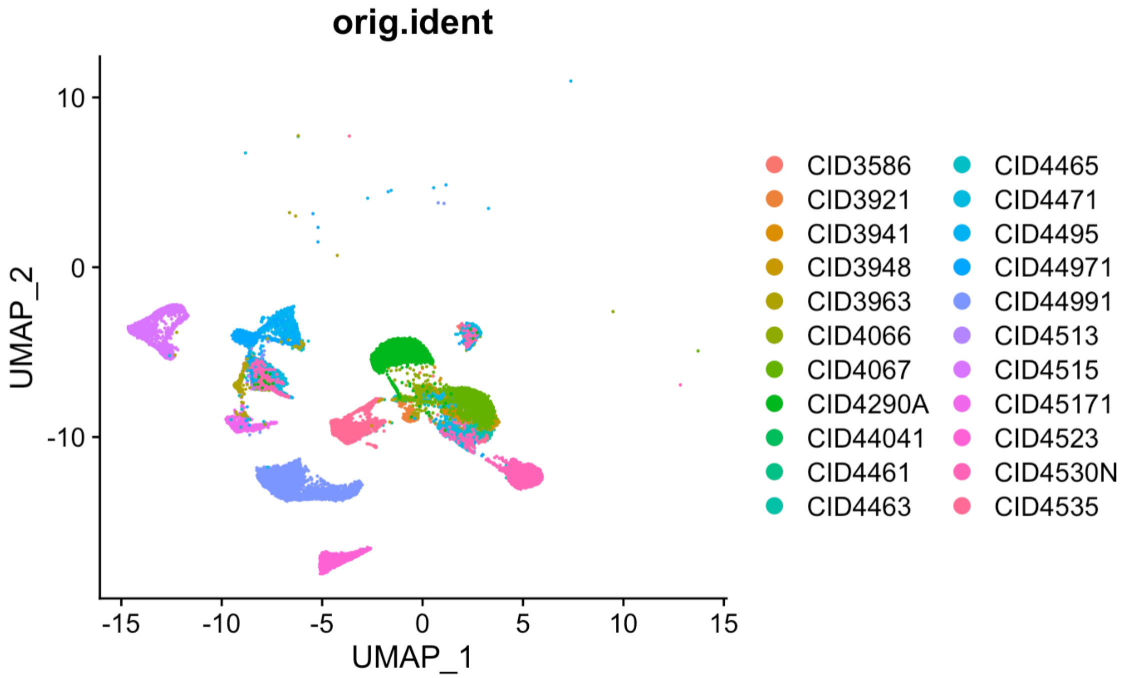

Figure 4. Reproduced UMAP output grouped by sample IDs.

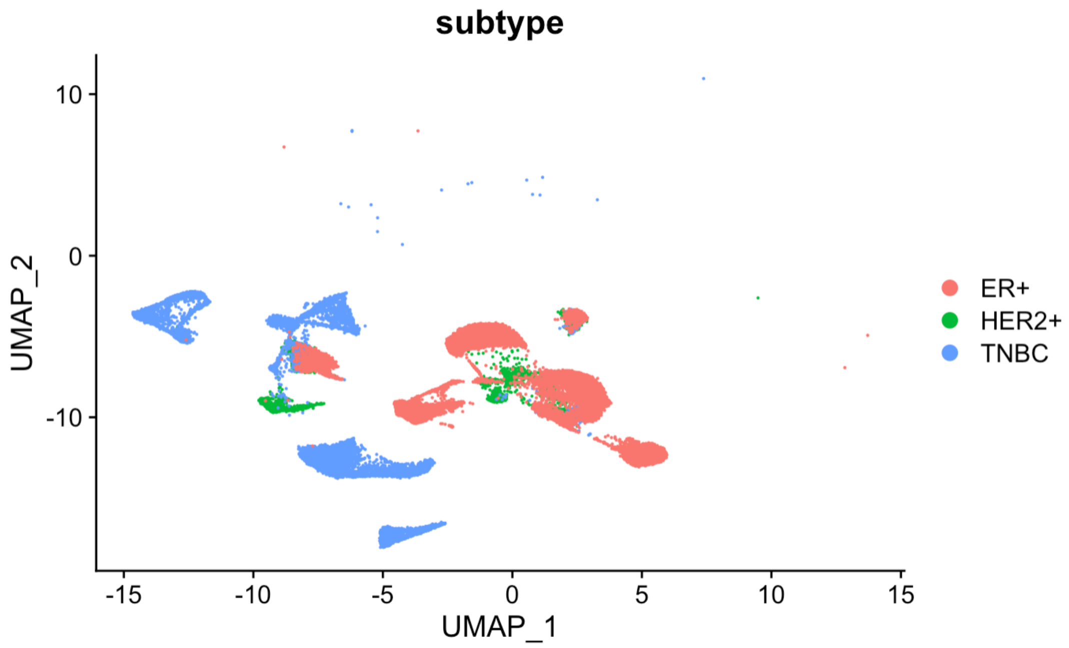

Figure 5. Reproduced UMAP output grouped by cancer subtypes.

Differential Expression Analysis

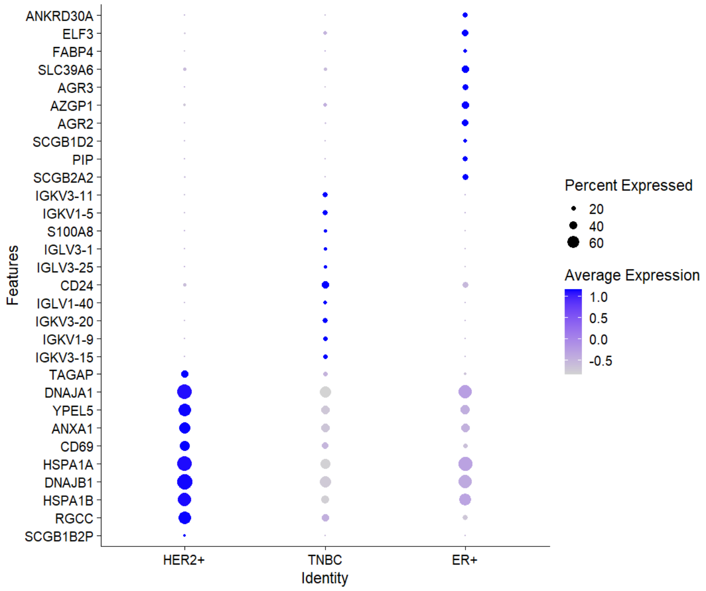

Figure 6. Dot plot of Top 10 Differentially expressed genes across the breast cancer subtype.

The highly expressed genes in each subtype were identified using the 'FindAllMarkers' function. The top 10 genes in each subtype were identified and visualized as dotplot. It is observed that the genes highly expressed in HER2+ had very low expression in the TNBC subtype and weak expression in ER+ subtype. Similarly, the top genes expressed in TNBC and ER+ had very low expression in other subtypes (Figure 6).

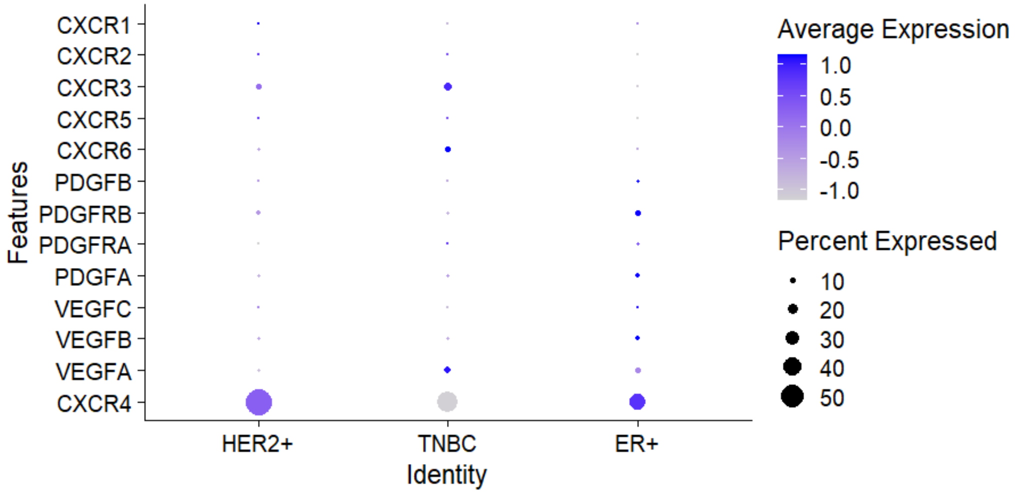

Figure 7. Dot Plot of differential expression of marker genes involved in metastasis and angiogenesis. (Metastasis - *CXCR1, CXCR2, CXCR3, CXCR4, CXCR5, CXCR6. Angiogenesis - VEGFA, VEGFB, VEGFC, PDGFA,PDGFB,PDGFRA,PDGFRB*)

Analyzing the differential expression of the markers, we observed that *CXCR4* had higher expression in ER+ compared to HER2+ and TNBC had negative differential expression. Other chemokines of the *CXCR* family had very low expression. The angiogenesis markers did not have significant expression (below 0.25 avg_log2FC threshold). (Figure 7).

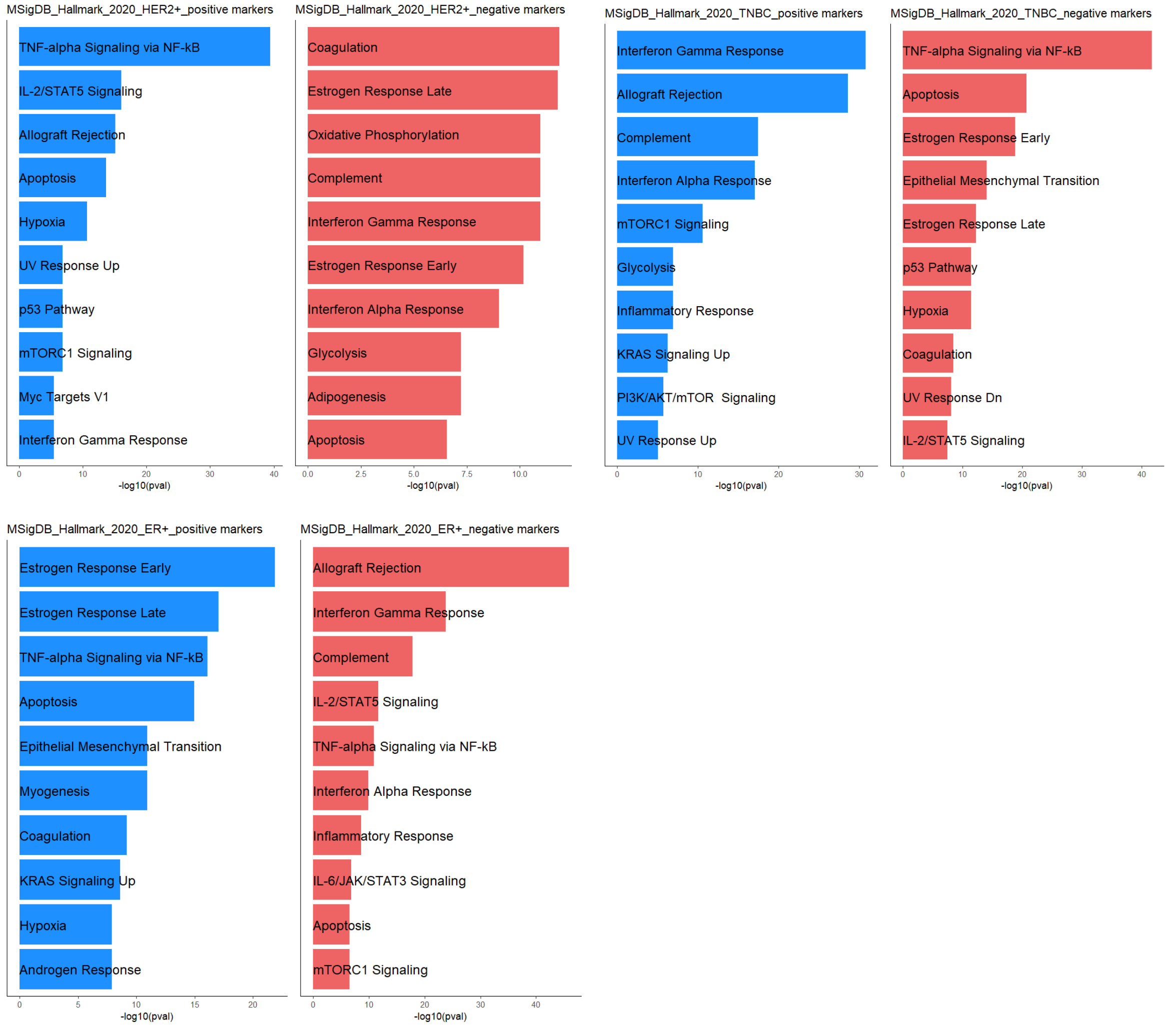

Figure 8. EnrichR Barplot of differential expressed genes across subtypes.

Pathway enrichment for the differentially expressed genes for all three subtypes was done using the MSigDB_Hallmark_2020 database using EnrichR. The “DEEnrichRPlot " outputs enriched and depleted terms from the EnrichR database as bar plots. It can be observed that some GO terms show up in both enriched and depleted plots. TNF-alpha signaling via NF-kB, apoptosis are some GO terms seen in both enriched and depleted bar plots for the different subtypes. In the ER+ enriched set, we can note that Estrogen Response genes are enriched, possibly due to the presence of the estrogen receptor and depleted in the TNBC set (Figure 8).

Co-expression analysis

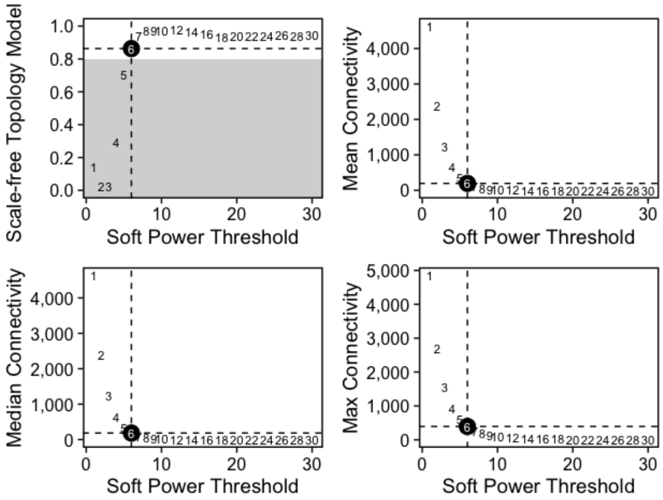

Figure 9. Plot of testing soft-power threshold.

The optimal soft power threshold turned out to be 6 based on the plot of testing soft-power threshold. That is, soft power = 6 seemed to be the most effective power that would affect the co-expression network to have more strong connections but less weak connections (Figure 9).

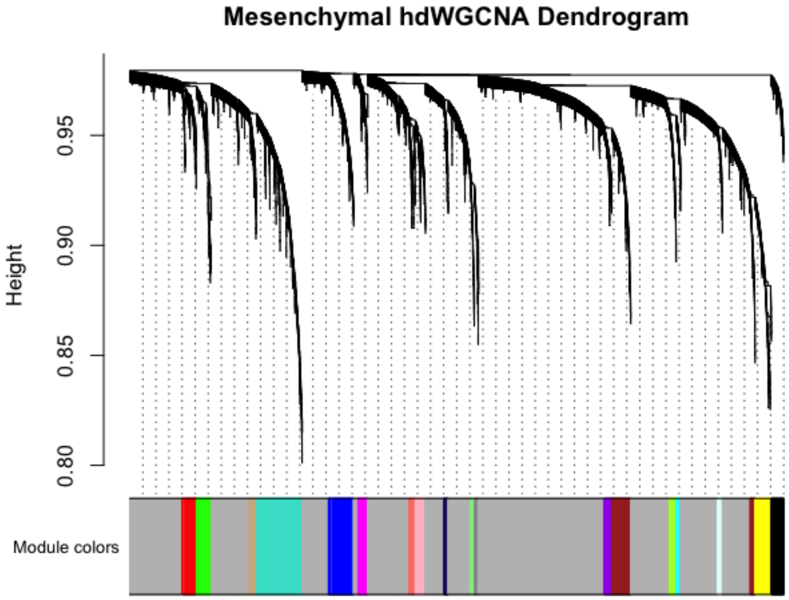

Figure 10. Mesenchymal cell hdWGCNA dendrogram.

Each single gene is represented with a leaf, and each color on the module colors band represents a co-expression module that the genes were assigned to. The gray region is not related to the modules such that genes in the gray region were not assigned to any of the co-expression modules. The height was used to determine if the genes that were differentially expressed from the other genes in a module could be allocated to another module. We identified 18 different modules (Figure 10).

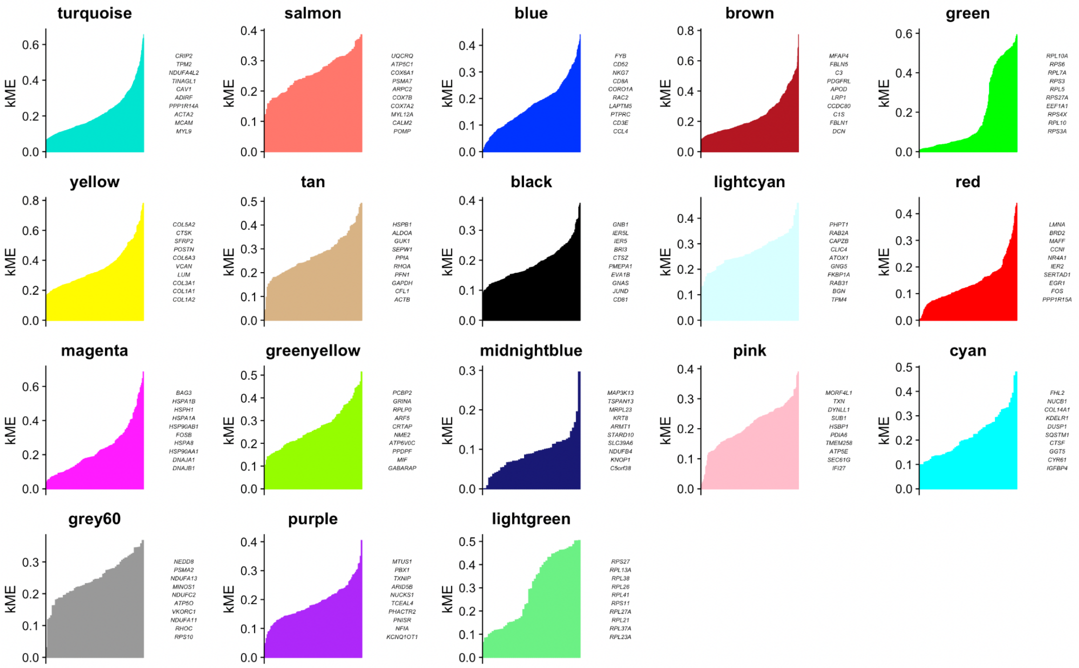

Figure 11. Visualization of the genes ranked by kME in each module.

The colors of the modules identified from the dendrogram correspond to the colors of the plot. Each plot showed the genes ranked by kME, eigen-based connectivity, in a module. The genes involved with collagen, *COL5A2, COL6A3, COL3A1, COL1A1,* and *COL1A2*, appeared in the yellow module. Also, the *KRT8* gene, which was used as a marker gene for epithelial cells for manual annotation in our pre-processing step, was found from the midnight blue module (Figure 11).

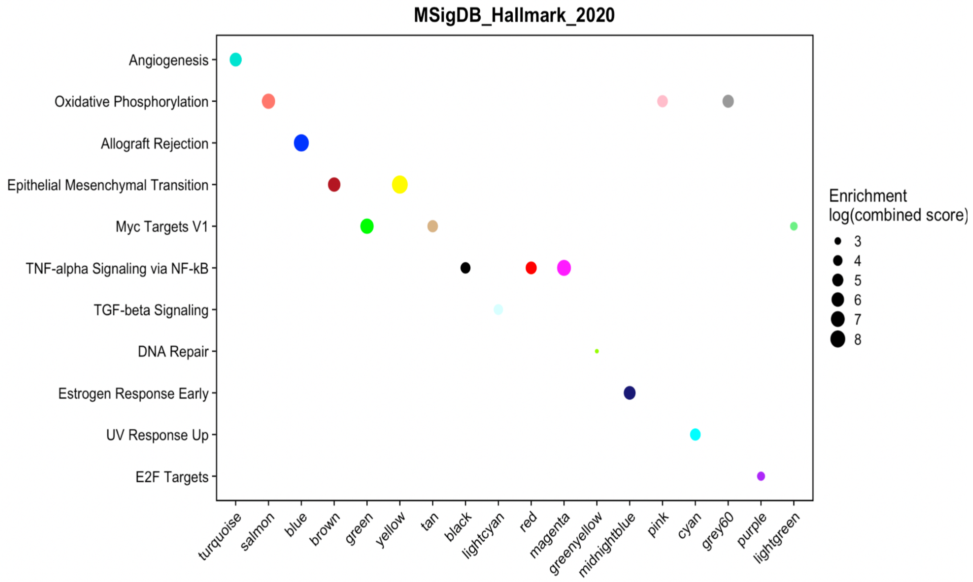

Figure 12. EnrichR Dot Plot based on “MSigDB_Hallmark_2020” EnrichR database.

The EnrichR Dot Plot (Figure 12) was made to see our results in a biological sense with the “MSigDB_Hallmark_2020” EnrichR database used. The descriptions on the y-axis were the top ranked results from the database, and the modules were enriched across the different descriptions. We can see that the descriptions that the modules were enriched and cell types that the modules were highly expressed in on average from Figure 13 were somewhat consistent. We would like to note that the brown and yellow modules that were enriched in the description of epithelial mesenchymal transition had a high average expression in the mesenchymal cells but low average expression in the epithelial cells. The process of epithelial mesenchymal transition would be done from epithelial cells to mesenchymal cells, so it would be sufficient to suggest that the brown and yellow modules were in transition from epithelial cells to mesenchymal cells based on Figure 13.

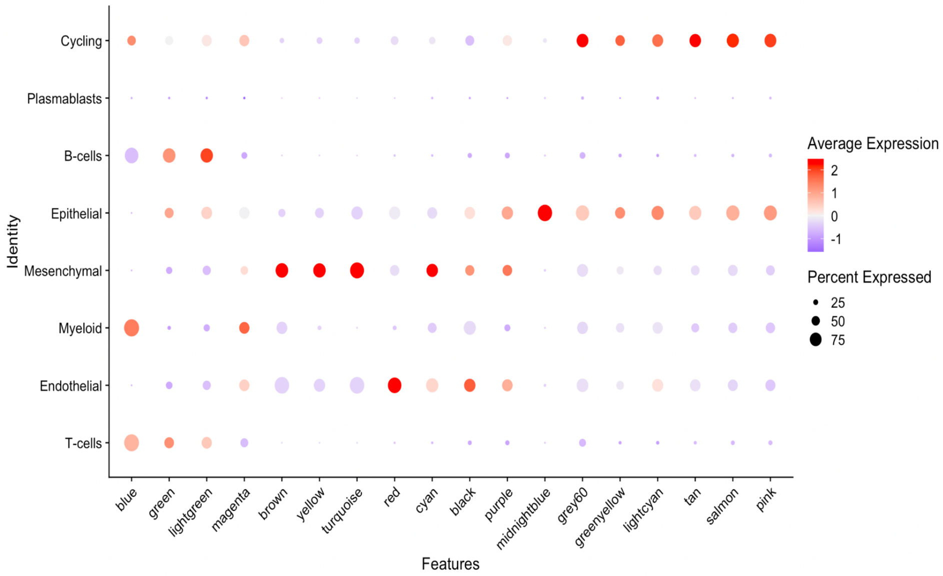

Figure 13. Dot plot of co-expressed gene modules by cell types.

The dot plot of co-expressed gene modules by cell types visualizes how each module is expressed across different cell types. The brown, yellow, turquoise, and cyan modules had a higher expression level in the mesenchymal cells than the other modules did. In addition, we could see the plot corresponded to the graph from Figure 11 that the midnight blue module which had KRT8 genes, one of the marker genes for epithelial cells, were highly expressed in epithelial cells on average.

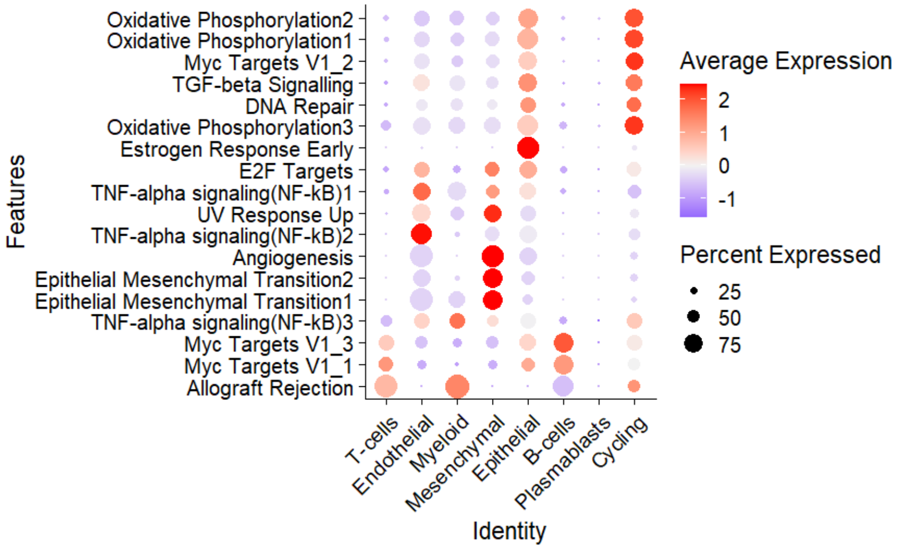

Figure 14. Dot plots of renamed co-expressed gene modules by cell types after comparing the figure 12 and figure 13.

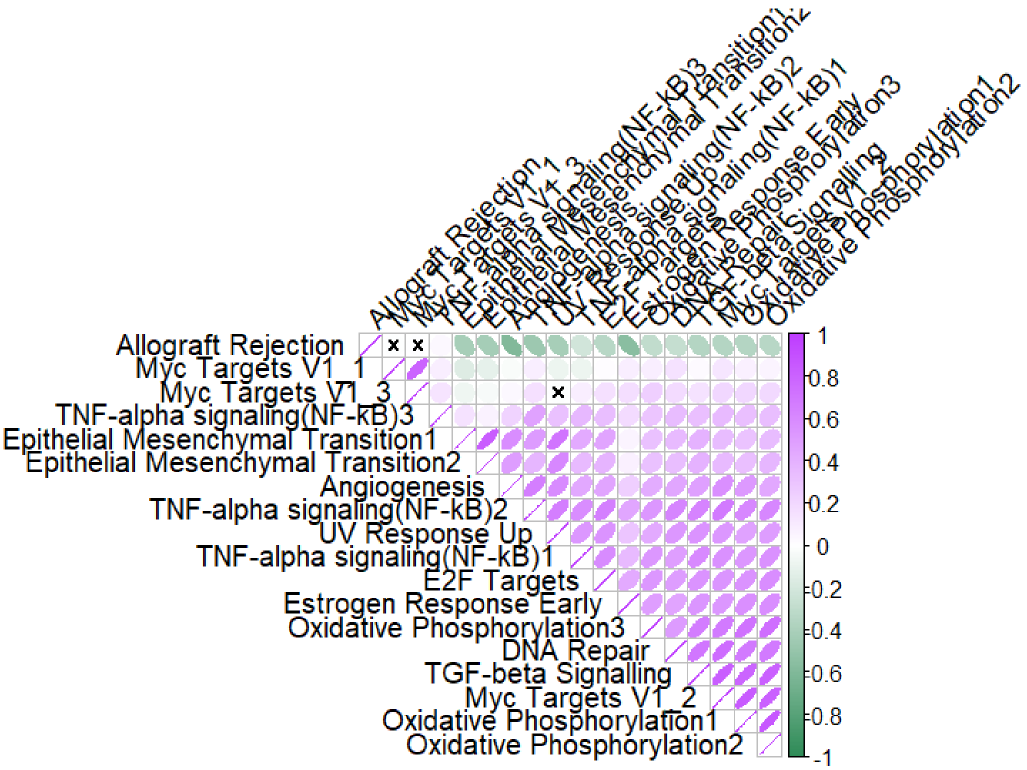

Figure 15. Module correlations with the modules renamed after comparing the figure 12 and figure 13.

Discussion

Differential expression analysis

The marker genes for angiogenesis did not have significant expression across all the subtypes. Out of all *CXCR* family genes in the dataset, only *CXCR4* showed significant differential expression. This may depend on the stage of the tumor. Information about whether the patient had a metastatic or benign tumor is not provided. This information could be helpful in making further conclusions. From differential expression results we can look for genes which can help with targeted treatment of different subtypes of breast cancer. For example, we can find that some features are highly expressed by HER2+ but not that highly expressed in other subtypes and vice versa.

Co-expression network analysis

The yellow module included *COL5A2, COL6A3, COL3A1, COL1A1,* and *COL1A2* as the top ranked genes (Figure 11) and was one of the modules that were highly expressed in the mesenchymal cells (Figure 12). Finding *COL1A1, COL1A2,* and *COL3A1* in the module related to the mesenchymal cells was consistent with one study that used the breast cancer datasets (Gordon 2023). The author explained that studying such genes was helpful to understand the mesenchymal cells involved with tumor microenvironment (Gordon 2023). Another study found that the epithelial mesenchymal transition process was significantly associated with *COL1A1, COL1A2, COL3A1,* and *COL5A2* (Yin 2021). This was also consistent with our results as the process that the yellow module including those genes were enriched in was the epithelial mesenchymal transition. Finally, another paper published by Papanicolaou et al. profiled breast cancer tumor microenvironment and found that metastasis was impacted by collagen genes. This would be consistent with our result as some collagen genes were identified from the module involved with the epithelial mesenchymal transition, which can be influenced by tumor microenvironment (Jing et al. 2011).

Limitation of this project

The next topic we want to discuss is contradiction in the EnrichR Barplot of differential expressed genes across subtypes as mentioned above. This problem might have been caused by the dimension of the dataset. This is because the result of the test might be different depending on how we subset the spatial dataset. Therefore, this is our limitation as our main topic was not related to spatial dimension in this project. However, this limitation could be solved if we divide the dataset into specific dimensions such as single cell type subsets or incorporate spatial analysis in future analysis.

References

1. Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C. L., Gnanasambandapillai, V., Dutertre, C. A., … Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, *53*(9), 1334–1347.
2. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., & Swarup, V. (2022). High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems. *bioRxiv*, 2022-09.
3. Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A. C., Head, E., ... & Swarup, V. (2021). Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature genetics*, *53*(8), 1143-1155
4. Gaiteri C, Ding Y, French B, Tseng GC, Sibille E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 2014 Jan;13(1):13-24. doi: 10.1111/gbb.12106. Epub 2013 Dec 10. PMID: 24320616; PMCID: PMC3896950.
5. Gordon, J. A., Evans, M. F., Ghule, P. N., Lee, K., Vacek, P., Sprague, B. L., ... & Stein, J. L. (2023). Identification of molecularly unique tumor-associated mesenchymal stromal cells in breast cancer patients. *Plos one*, *18*(3), e0282473.
6. Mollica Poeta, V., Massara, M., Capucetti, A., & Bonecchi, R. (2019). Chemokines and Chemokine Receptors: New Targets for Cancer Immunotherapy. Frontiers in immunology, 10, 379. https://doi.org/10.3389/fimmu.2019.00379
7. Yin, W., Zhu, H., Tan, J. *et al.* Identification of collagen genes related to immune infiltration and epithelial-mesenchymal transition in glioma. *Cancer Cell Int* 21, 276 (2021). https://doi.org/10.1186/s12935-021-01982-0
8. Valdés-Mora, F., Salomon, R., Gloss, B. S., Law, A. M. K., Venhuizen, J., Castillo, L., Murphy, K. J., Magenau, A., Papanicolaou, M., Rodriguez de la Fuente, L., Roden, D. L., Colino-Sanguino, Y., Kikhtyak, Z., Farbehi, N., Conway, J. R. W., Sikta, N., Oakes, S. R., Cox, T.

R., O'Donoghue, S. I., Timpson, P., … Gallego-Ortega, D. (2021). Single-cell transcriptomics reveals involution mimicry during the specification of the basal breast cancer subtype. *Cell reports*, *35*(2), 108945. https://doi.org/10.1016/j.celrep.2021.108945

9. Jing, Y., Han, Z., Zhang, S., Liu, Y., & Wei, L. (2011). Epithelial-Mesenchymal Transition in tumor microenvironment. *Cell & bioscience*, *1*, 29. https://doi.org/10.1186/2045-3701-1-29

BIOS 785 Project Report: Exploring Human Breast Cancer Atlas scRNA-Seq Data
Team members: Seowoo Kim, Taeim Kwon, Arthi Hariharan