Amy Gill
Interim report
4/19/21

Background

I am replicating and extending analysis from Karaayvaz et al., 2018, a single-cell RNA-seq study analyzing 1189 cells from 6 patients with triple negative breast cancer (TNBC). The goal of this replication is to classify cells according to their lineage (epithelial, stromal, endothelial, macrophage, T cell, B cell).

Results

First, I used a literature-based gene expression panel of 49 genes. As in the original paper, cells were labeled using the following expression rules (expression is defined as a normalized gene expression value greater than or equal to 1):

- Epithelial
    - Case 1: expressing 2 or more breast epithelial markers (EPCAM, EGFR, CDH1, KRT14, ITGA6, KRT5, TP63, KRT17, MME, KRT8, KRT18, KRT19, FOXA1, GATA3, MUC1, CD24, KIT, GABRP)
    - Case 2: expressing a single well-defined breast epithelial marker (EPCAM, KRT8, KRT18, KRT19) above the median level of gene expression for that gene in that patient
- Stromal
    - Case 1: expressing 2 or more stromal markers (FAP, COL1A1, COL3A1, COL5A1, ACTA2, TAGLN, LUM, FBLN1, COL6A3, COL1A2, COL6A1, COL6A2) and no other markers
    - Case 2: expressing 3 or more stromal markers and no more than 1 endothelial marker
- Endothelial
    - Case 1: expressing 2 or more endothelial markers and no other markers
    - Case 2: expressing 3 or more endothelial markers and no more than 1 stromal marker
- T cell
    - Case 1: expressing 2 or more T cell markers (CD2, CD3D, CD3E, CD8A, CD8B) and no B cell or macrophage markers
    - Case 2: expressing the pan-immune marker PTPRC, 1 T cell marker, and no B cell or macrophage markers
    - Case 3: expressing 3 or more T cell markers and no more than 1 B cell or macrophage marker
- B cell
    - Case 1: expressing 2 or more B cell markers (MS4A1, CD79A, CD79B, BLNK) and no T cell or macrophage markers
    - Case 2: expressing the pan-immune marker PTPRC, 1 B cell marker, and no T cell or macrophage markers
    - Case 3: expressing 3 or more B cell markers and no more than 1 T cell or macrophage marker
- Macrophage
    - Case 1: expressing 2 or more macrophage markers (CD14, CD68, CD163, CSF1R) and no B cell or T cell markers

    - Case 2: expressing the pan-immune marker PTPRC, 1 macrophage marker, and no B cell or T cell markers
    - Case 3: expressing 3 or more macrophage markers and no more than 1 B cell or T cell marker
- Undecided
    - Belonging to 2 or more cell types according to previous rules
    - *Exception:* cells classified as epithelial and stromal, and no other cell types, were interpreted as epithelial cells undergoing epithelial-mesenchymal transition (EMT), a common phenotype in cancerous epithelial cells
- Unknown
    - Not classified as any cell type above

I plotted the cells using tSNE and UMAP and colored by cell type (Figure 1). The majority of cells are epithelial, as expected in breast cancer, with distinct populations of stromal cells and endothelial cells (right of Figure 1A, top of Figure 1B) as well as immune cells (left of Figure 1A, bottom of figure 1B). The observed cell type numbers and clusters correspond well, though not perfectly, with the reported values and patterns in Karaayvaz et al., 2018 (data not shown). Although I applied the classification rules as written, my analysis produced fewer cells of unknown types and more T cells. However, as the T cells are all part of a single well-defined cluster, it is likely they all are true T cells.

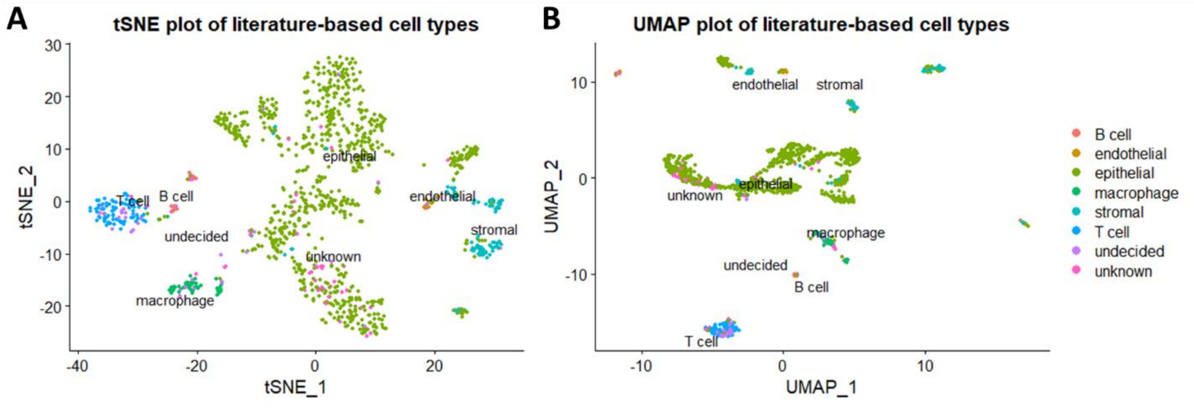

Figure 1. (A) tSNE and (B) UMAP plots of cell types determined by a literature-based panel of marker genes.

Then, I clustered the cells using Louvain clustering as implemented in Seurat (using a resolution of 1.2 – other resolutions will be explored). Louvain clustering identified 17 distinct clusters. I plotted the cells using tSNE and UMAP and colored by Seurat cluster (Figure 2). Many of these clusters corresponded well to the literature-based cell types: cluster 3 corresponds to T cells; cluster 13 corresponds to B cells; clusters 11, 15 and 16 correspond to macrophages; cluster 10 corresponds mainly to endothelial cells (with some stromal cells), and other clusters correspond to subsets of epithelial cells.

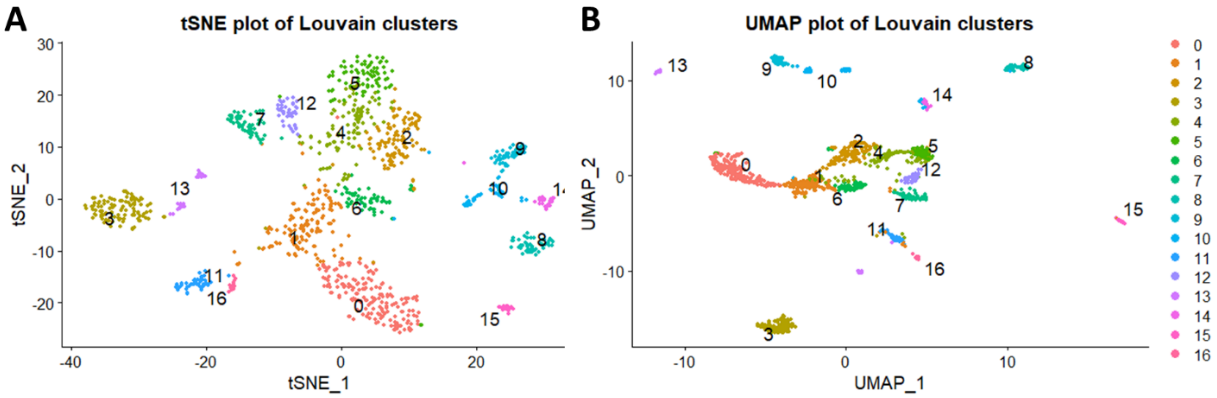

Figure 2. (A) tSNE and (B) UMAP plots of Louvain clusters.

I also investigated genes that were differentially expressed across the various Louvain clusters, as determined using Seurat (Figure 3). Most clusters were well-defined by the top 5 differentially expressed genes, although clusters 1, 2, 4 and 5 were less clearly distinguished from other clusters by their top 5 genes. Notably, these poorly resolved clusters are composed of epithelial cells in Figure 1. Several differentially expressed genes are known markers of the cell types in matching clusters; for example, cluster 3 shows overexpression of CD3D, a canonical T cell marker included in the literature-based panel, as well as overexpression of LCK, a canonical T cell marker not included in the panel.

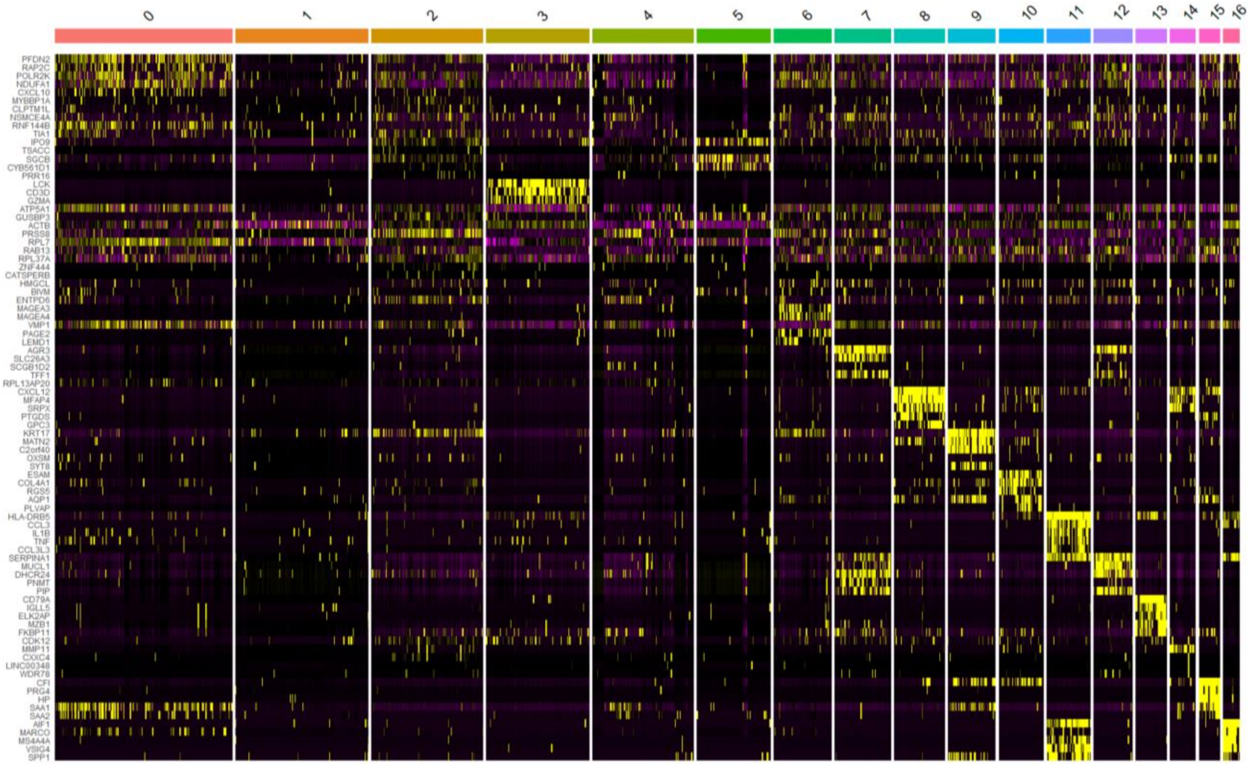

Figure 3. Heatmap of the top 5 differentially expressed genes from each cluster. Yellow indicates high expression and purple indicates low expression.

Future directions

In the original paper, tSNE and clustering were performed after regressing out the patient variable. I did not remove the patient variable first, but will explore the impact of doing so. To extend this study, I intend to explore how cell types are classified using additional clustering algorithms, such as hierarchical clustering, k-nearest neighbors, k-means, and infomap. I plan to characterize the degree of overlap between clusters produced by different algorithms using multimodal integration analysis (MIA) (Cabello-Aguilar et al., 2020). In essence, MIA is a hypergeometric test for the extent of overlap between gene sets that are differentially expressed according to one analysis method relative to another analysis method. This will allow quantitative evaluation of concordance or discordance of clusters produced by different methods.

References

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., & Colinge, J. (2020). SingleCellSignalR: Inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research, 48*(10), e55–e55. https://doi.org/10.1093/nar/gkaa183

Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F., & Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications, 9*(1), 3588. https://doi.org/10.1038/s41467-018-06052-0