Amy Gill
EN.580.743
5/13/21

Replicability of single-cell RNA-seq cell type identification

Introduction

Tissues and tumors are ecosystems of functionally distinct cell types in various states, and these differences are obscured in traditional bulk analysis methods like Western blotting or bulk RNA sequencing. Single-cell RNA sequencing (scRNA-seq) allows transcriptome-scale investigation of tissue heterogeneity on a cellular level. Downstream applications of scRNA-seq data can elucidate which cell types are present in a tissue, their relative proportions, molecular differences between those cell types, heterogeneity within cell types, signaling and transcriptional networks active in each cell type, and which cell types have expression features that allow them to functionally interact with each other. Many downstream applications of scRNA-seq first require individual cells in a dataset to be classified into cell types to analyze clusters of similar cells, but there is no single best approach for cell type classification.

The three main categories of scRNA-seq cell typing methods – marker gene approaches, supervised machine learning, and unsupervised clustering – have distinct benefits and drawbacks. Supervised approaches compare expression profiles of unknown cells to cells with known gold-standard identities, but this approach requires reliable training data on representative cell types, which must first be typed by one of the other approaches in this list. Marker gene approaches use existing biological knowledge of gene expression patterns to classify cells based on their expression of known cell-type-specific molecular features. However, this approach requires extensive prior knowledge, and selected markers may not be specifically or consistently expressed in real data. Unsupervised clustering algorithms require no prior knowledge or gold-standard reference data for clustering, but different algorithms produce different cell groupings, and identifying the biological identity and relevance of resulting clusters is non-trivial. Because different cell typing methods use different assumptions and approaches, identified clusters vary based on the method, and using multiple methods for cell typing on the same data can help determine whether identified populations represent high confidence groupings (Andrews and Hemberg, 2018).

The goal of this analysis is to compare cell typing methods for scRNA-seq data from triple negative breast cancer (TNBC). As no reference dataset yet exists for human breast tissue, supervised methods cannot be applied. Here, a literature-based marker gene approach for cell typing is compared to two different unsupervised clustering methods: Louvain community detection and consensus k-means clustering. The clusters produced by each approach are compared to clusters produced by the other approaches to analyze the degree to which different algorithms produce the same groups of cells or generate clusters with similar expression profiles.

Data

scRNA-seq data from 6 fresh TNBC tumors from human patients (Karaayvaz et al., 2018) were downloaded from GEO (GSE118389) in both transcript per million (TPM) count format and quality-control filtered, normalized format. Briefly, the authors removed low quality cells (> 4 median absolute deviations (MADs) below the median for library size, number of expressed genes, or mRNA amount) and unexpressed genes ($\log_2(TPM+1) < 0.1$ in 95% of cells for each patient). Normalization steps included transformation of TPM values into relative counts with the Census algorithm (Qiu et al., 2017), normalization of Census counts using deconvolution in scran (Lun et al., 2016), and removing additional sources of variation with RUVSeq (Risso et al., 2014). Most analysis in this report was performed with the normalized data; TPM data were only used to calculate dropout rate in SC3 clustering (see Methods). All analysis was performed in R 4.0.5.

Methods

Pre-processing

To better identify clusters of biologically meaningful cell types across cancers, rather than identifying patient-specific effects, the patient ID variable was removed by linear regression using `Seurat::ScaleData`. The 2000

most variable genes were identified using the "vst" selection method. Dimension reduction by principal component analysis of the 2000 most variable genes was performed with `Seurat::RunPCA`. tSNE and UMAP projections were found using the top 50 principal components using `Seurat::RunTSNE` and `Seurat::RunUMAP`. The method for `Seurat::RunUMAP` was the R-native UWOT using the cosine metric.

Cell typing based on marker genes

Cell typing was performed as described in Karaayvaz et al. (2018). The table of 49 marker genes and their corresponding cell types (epithelial, stromal, endothelial, macrophage, T cell, B cell) was imported from the PDF of the supplement using the pdftools library (Table 1). A marker gene was considered to be expressed in a cell if its normalized expression level was greater than or equal to 1. Cells were classified according to the decision rules shown in Table 1.

Louvain clustering with Seurat

Louvain community detection was performed with the Seurat package (Satija et al., 2015). Cell-level shared nearest neighbors were computed with `Seurat::FindNeighbors`. Louvain clustering was performed with `Seurat::FindClusters`. Resolutions from 0.4-1.2 were tested based on the suggested range in the Seurat documentation for datasets of 3000 cells or fewer. At resolutions greater than or equal to 1, the number of clusters leveled out at 18, so a resolution of 1 was chosen (data not shown).

Consensus k-means clustering

Single cell consensus clustering was performed with the SC3 package (Kiselev et al., 2017). The optimal value of k=29 for k-means clustering was estimated based on Tracy-Widom theory of random matrices using `SC3::sc3_estimate_k`. Distances between cells (Euclidean, Spearman and Pearson) were calculated with `SC3::sc3_calc_dists` and the distance matrices were transformed with PCA and the graph Laplacian using `SC3::sc3_calc_transfs`. k-means clustering was performed with the estimated optimal k=29 using `SC3::sc3_kmeans`, which performs clustering independently with each combination of distance metric and transformation for a variety of random starting seeds. Consensus clustering of these diverse k-means outputs was performed with `SC3::sc3_calc_consens`.

Cluster visualization and differential gene expression

For all clustering methods, clusters were visualized on tSNE and UMAP projections using `Seurat::DimPlot`. Genes differentially expressed across clusters were identified using `Seurat::FindAllMarkers`. Genes were declared to be differentially expressed if they had adjusted p-value < .05 and log fold change > 0.5. The top 5 genes or top 10 genes differentially expressed across clusters were visualized using `Seurat::DoHeatmap`.

Quantification of cluster consistency across methods with multimodal intersection analysis (MIA)

MIA (Moncada et al., 2020) was used for two purposes: first, to quantify the degree to which different clustering approaches generate clusters composed of the same cells, and second, to quantify the degree to which different clustering approaches generate clusters with the same upregulated differentially expressed genes. Given two clustering approaches, each cluster from one approach is compared to each cluster from the other approach and a hypergeometric test for enrichment is performed on either the list of cells belonging to those clusters or the list of differentially expressed genes associated with each cluster. p-values from the hypergeometric test are plotted on a heatmap to highlight which clusters from one approach best match clusters from the other approach.

Results

Marker-based cell typing

Cell types determined by marker gene expression tend to be grouped together on tSNE and UMAP projections (Fig. 1A-B). Most cells are epithelial, and epithelial cells form several distinct clusters that may represent epithelial cell subtypes or heterogeneous tumor clones. Non-epithelial cell types tend to form compact and

consistent clusters, particularly T cells and macrophages. One cluster of epithelial cells is closer to stromal cells than other epithelial clusters; these cells may reflect true epithelial cells undergoing epithelial-mesenchymal transition (EMT), or they may reflect misclassified stromal cells that were labeled epithelial due to more permissive decision rules for epithelial classification. Most undecided cells are within immune clusters, suggesting that the decision rules for immune types may be too vague given the similarity of these cell types. In contrast, most unknown cells cluster with epithelial cells, suggesting that the decision rules for epithelial cells may not be flexible enough to capture all normal or cancerous breast epithelium.

The cell type classifications in this replication study broadly agree with the numbers reported by the original authors, though some slight to moderate discrepancies exist in each category (Fig. 1C). Importantly, although the replication yielded a much higher number of T cells than the original study, all of these T cells are in a single well-defined cluster on tSNE and UMAP, supporting a likely shared identity of these cells.

Individual patients have between 75 and 300 cells contributed to the dataset (Fig. 1D). For each patient, the largest category of cells is epithelial cells, as expected from a cancer of epithelial origin (Fig. 1E). Not all cell types are present in each tumor; in particular, all immune cells (B cells, T cells, macrophages) are absent from patients PT081 and PT126, and T cells are only observed in patients PT058 and PT084, but when T cells are observed they are the second most common cell type after epithelial cells. The absence of immune cells in some patients is supported by the original paper's description that some patient samples were intentionally immune-depleted by flow sorting for CD45- cells before sequencing, although the paper does not note the number and identity of immune-depleted samples.

The expression profiles of the top 5 differentially expressed genes in each cell type are fairly consistent across cells of a type, though epithelial cells are more heterogeneous than other cell types (Fig. 1F). This internal consistency suggests marker-based cell typing creates cell groups with true biological differences. Differentially expressed genes include several marker genes used to assign cell type, but also include known markers not used by decision rules (i.e. LCK for T cells, CD34 for endothelial cells). However, several cells labeled undecided have expression profiles that clearly match only one tested cell type, suggesting that these cells could be conclusively assigned to an identity with a different approach.

Louvain clustering

Louvain clusters reflect consistent visual groupings in the projected data (Fig. 2A-B). In comparison to the marker-based cell types, most cell types are well-represented by 1 or 2 clusters (T cells and cluster 3, B cells and cluster 16, macrophages and clusters 10/15, stroma and clusters 5/9. Endothelial cells do not have an independent cluster: they are grouped in cluster 5 with stromal cells, although they form a visually distinct subcluster that could likely be revealed by higher resolution clustering. Numerous clusters have an epithelial identity. Undecided and unknown cells are spread throughout existing clusters, suggesting they likely reflect cells of one of the 6 original types that were not properly classified in the marker-based cell typing rather than reflecting additional untested populations of cells. Interestingly, most of the clusters are uniquely defined by their top 5 differentially expressed genes (Fig. 2C), but some are not, and all clusters without clear defining gene expression profiles are epithelial in the marker-based cell typing.

Consensus k-means clustering

Consensus k-means clustering produces small clusters that often subdivide apparently well-defined groupings in the projected data (Fig. 3A-B). For example, in comparison to the marker-based cell types, the distinct T cell group is now divided into 2 different clusters with no obvious demarcation line (9-10). Some of this enhanced resolution may indeed reflect biological divisions – for example, endothelial cells are now within their own cluster (21), while in the Louvain clustering, endothelial cells were grouped with stroma – but the implication of other cluster subdivisions is unclear. Inspection of the top 10 differentially expressed genes per cluster suggests that the consensus k-means clustering approach subdivided clusters too finely (Fig. 3C). While some clusters show unique expression patterns (for example, 6 and 29), other groups of distinct clusters have highly similar patterns (for example, 9-10 and 8-19-20), suggesting this approach may not be returning the most parsimonious group of biologically relevant clusters. Additional clusters had an extremely small number of cells (cluster 2 = 1 cell, cluster 5 = 5 cells) and may not reflect biologically meaningful subsets or may represent populations that are

too rare to characterize with statistical confidence. These results suggest that the estimated optimal value of k from SC3 may produce overly fine resolution and that the true number of clusters is less than 29.

Comparison between identified clusters

Multimodal intersection analysis (MIA) identified statistically significant overlap between cell groups formed by marker-based cell typing, Louvain clustering, and consensus k-means clustering (Fig. 4). When comparing cell membership in clusters produced by each algorithm, clusters associated with T cell, stromal, or macrophage identities tend to have stronger enrichment than clusters associated with epithelial cells, likely because epithelial cells are subdivided into a larger number of clusters. Clusters with a high degree of overlap in cell membership also tend to have a high degree of overlap in upregulated gene sets. However, the upregulated gene set overlap was not fully redundant with cell overlap, as some clusters with low cell overlap showed strong gene set enrichment (for example, Louvain clusters 6-7 with epithelial cells) and some clusters with significant cell overlap showed weak gene set enrichment (Louvain clusters 16-17 with B cells; k-means cluster 21 with endothelial cells, numerous k-means clusters with epithelial cells). Louvain clusters and k-means clusters associated with the same cell type also tended to be associated with each other at the cell level, gene level, or both. In general, Louvain clusters and k-means clusters tended to share more enrichment in upregulated genes than in cell identity, likely reflecting their differing subdivisions of cells of the 6 investigated cell types into clusters.

Conclusions

Replicability of tSNE visualization and original marker-based cell typing

Replicating the regression of the patient ID variable, dimensional reduction, and tSNE plotting was a straightforward process using functions from Seurat. However, the original paper's tSNE clustering looks slightly different, although the general structure of numerous epithelial clusters and some smaller outlying clusters of other cell types is conserved (data not shown). One immediate explanation is that tSNE projection is a stochastic process with variable results. Another important factor is that the original paper selected the most variable genes for dimensionality reduction using Monocle (although the exact method was unclear), while this analysis selected variable genes using Seurat. Use of different lists of variable genes will affect visualizations and would also affect the results of any cell typing algorithm like clustering that relies on expression patterns of variable genes.

Marker-based cell typing required some work to extract the list of marker genes, perform expression tests on each gene, and program the decision rules. Despite following the decision rules as written in the original paper, the cell type identities returned by marker-based cell typing differed (Figure 1C). The largest discrepancies in the replication set with respect to original definitions were a decreased number of unknown cells and an increased number of T cells. The reason for these discrepancies is not immediately clear. In the supplement, the authors describe that they refined their classifications of undecided and unknown cells by using insights from unsupervised clustering, but (1) the basis of their methods and insights is not completely described and (2) the authors still ended up with significantly more unknown cells than in the replication. The marker gene approach should not have been affected by the difference in variable gene identification, and the author's original normalized data were used. The only clear preprocessing difference that could explain this discrepancy is that I regressed out the patient variable using a Seurat function that includes linear regression, while the authors do not state which function they used to regress out the patient variable, and it is possible there is an algorithmic difference in the two approaches we used.

Strengths and weaknesses of tested methods for cell typing

Marker-based cell typing generated biologically interpretable groups of cells that both grouped together on tSNE/UMAP and had consistent gene expression patterns. However, this approach involved implementing fairly arbitrary decision rules based on prior knowledge and resulted in many unclassified cells, including cells whose gene expression clearly suggested they belonged to a certain type tested in the study. Louvain clustering required no prior knowledge, classified every cell, and generated compact clusters with similar gene expression profiles. However, the resulting Louvain clusters are not immediately interpretable, and additional biological knowledge must be applied to assign cell types to the resulting clusters. Consensus k-means clustering is heavily dependent on the choice of cluster number k. In this case, the estimated optimal k generated by SC3 produced overly

resolved clusters that may not be biologically relevant. Experimentation with different values of k may produce more parsimonious clusters but was beyond the scope of this study. As with Louvain clustering, the k-means clusters are not immediately interpretable and require supplementation with further biological knowledge to assign cell types. As these approaches provide complementary information about relationships between cells, opportunities exist to supplement individual approaches by combining them in an ensemble method. Additional research would be required to devise rational ways to integrate information from diverse clustering and cell typing methods in an ensemble.

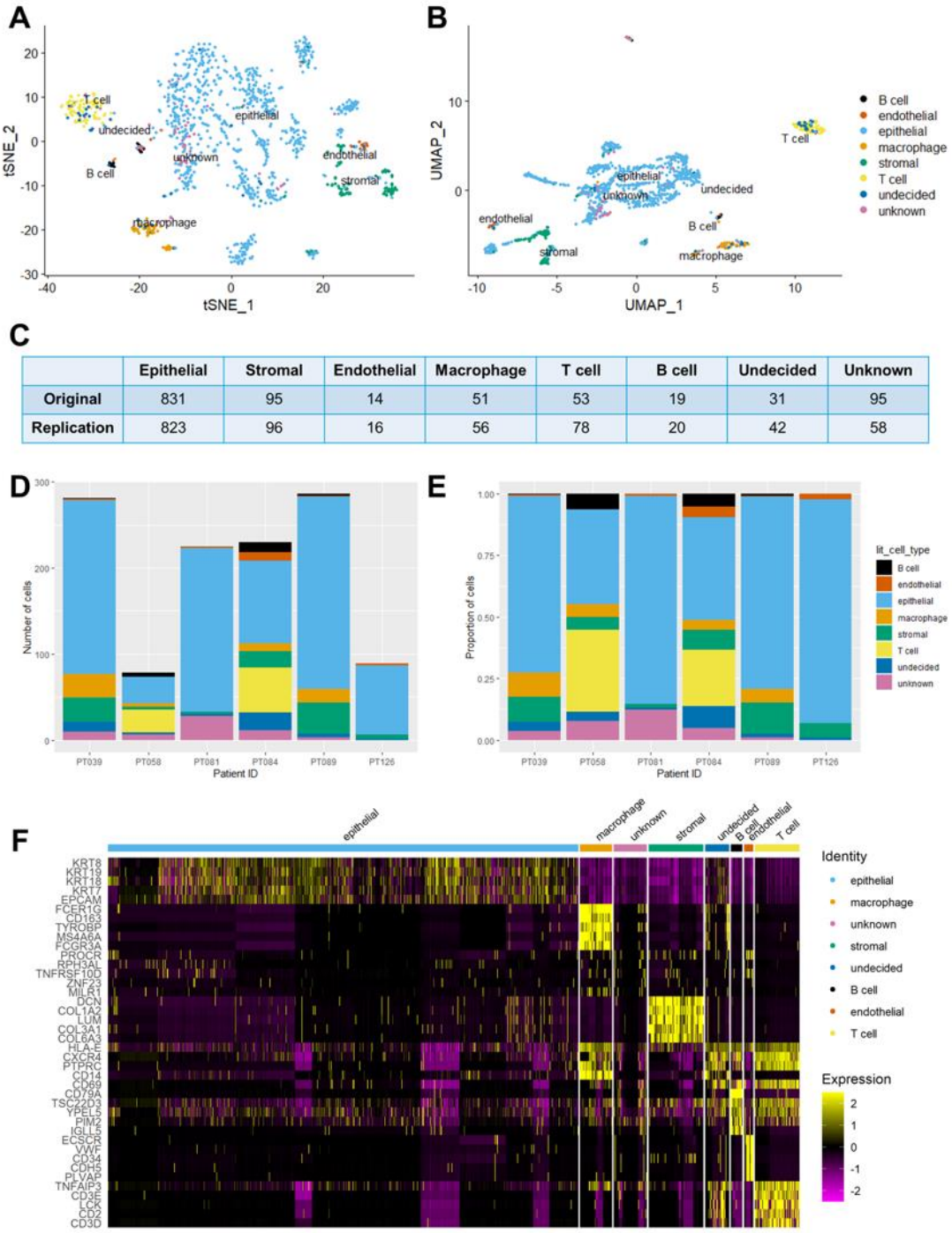
Biological implications of findings

One essential conclusion is that the results of any scRNA-seq analysis involving an early cell typing step will depend on the cell typing algorithm used. This implies that findings from distinct studies may not be directly comparable if different approaches were used for cell typing. Importantly, some features of clusters (such as upregulated genes) were robustly retained across various approaches. These features reproducible across multiple cell typing approaches could be important landmarks for researchers comparing results of multiple studies and could be investigated as novel or important markers of biologically relevant cell types. It is notable that non-epithelial and non-tumor cell types form more distinct and easily identifiable clusters than various epithelial populations in this study, while the epithelial cells were highly heterogeneous. While clustering can show the existence of this heterogeneity, approaches beyond naïve clustering (such as analysis of mutations, copy number variations, and cell cycle state) will be useful for distinguishing normal epithelium from cancer and identifying subpopulations and heterogeneity in the tumor. If the goal of a study is to investigate intratumoral heterogeneity, it may also be useful to perform additional rounds of clustering and differential expression analysis on only the epithelial cells or suspected tumor cells in the absence of non-epithelial types.

Figures

Table 1. Marker genes and decision rules for classification of cells as 6 different cell types.

Cell type	Markers	Classification rules (1 or more cases met)
Breast epithelial	**EPCAM, KRT8, KRT18, KRT19,** EGFR, CDH1, KRT14, ITGA6, KRT5, TP63, KRT17, MME, FOXA1, GATA3, MUC1, CD24, KIT, GABRP	1. 2+ breast epithelial markers 2. 1 **well-defined (bold)** breast epithelial marker expressed above median level for that marker for that patient 3. Cell labeled as epithelial and stromal only
Stromal	FAP, COL1A1, COL3A1, COL5A1, ACTA2, TAGLN, LUM, FBLN1, COL6A3, COL1A2, COL6A1, COL6A2	1. 2+ stromal markers and no other markers 2. 3+ stromal markers and 0-1 endothelial marker
Endothelial	PECAM1, VWF, CDH5, SELE	1. 2+ endothelial markers and no other markers 2. 3+ endothelial markers and 0-1 stromal marker
Immune types (T cell, B cell, macrophage)	*Pan-immune:* PTPRC *T cell:* CD2, CD3D, CD3E, CD8A, CD8B *B cell:* MS4A1, CD79A, CD79B, BLNK *Macrophage:* CD14, CD68, CD163, CSF1R	1. 2+ type-specific immune markers (T, B, macrophage) and no other immune markers 2. Pan-immune marker PTPRC, 1 type-specific immune marker, no other immune markers 3. 3+ type-specific immune markers and 0-1 other immune markers
Undecided		Assigned to 2+ cell types above *(Exception: Epithelial/stromal only labeled as epithelial)*
Unknown		Not assigned to any cell type above



	Epithelial	Stromal	Endothelial	Macrophage	T cell	B cell	Undecided	Unknown
Original	831	95	14	51	53	19	31	95
Replication	823	96	16	56	78	20	42	58

Figure 1. *Marker-based cell typing of TNBC cells.* ***A)*** *tSNE and* ***B)*** *UMAP projections of cell types based on marker gene expression.* ***C)*** *Comparison of cell type counts reported in the original analysis (Karaayvaz et al., 2018) and this replication study.* ***D)*** *Cell type numbers and* ***E)*** *proportions in each patient.* ***F)*** *Top 5 differentially expressed genes for each cell type.*

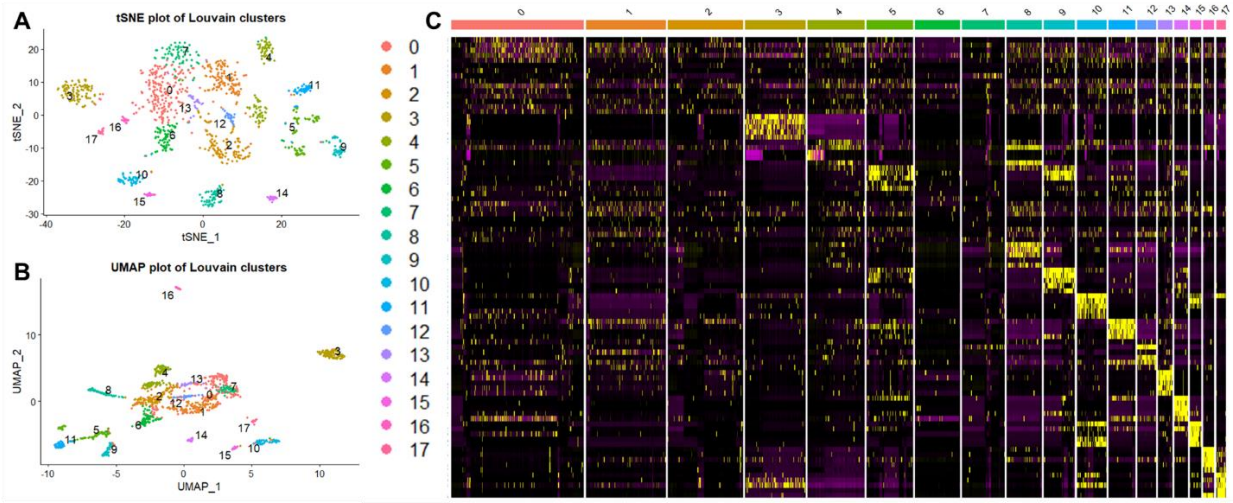


Figure 2. *Louvain clustering of TNBC cells with Seurat.* ***A)*** *tSNE and* ***B)*** *UMAP projections of Louvain clusters.* ***C)*** *Top 5 differentially expressed genes for each cluster.*

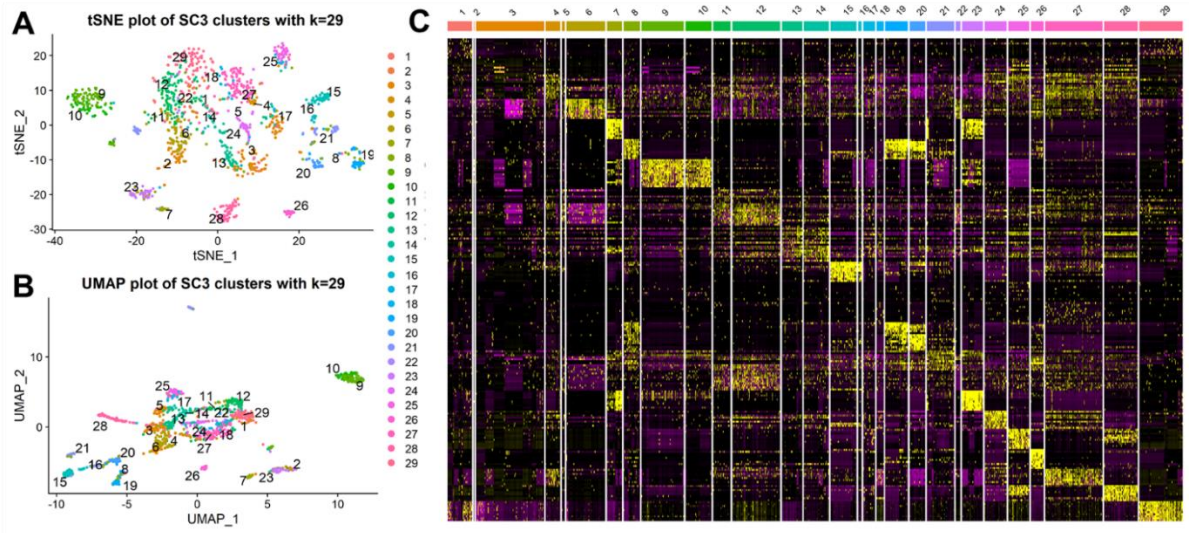


Figure 3. *Consensus k-means clustering of TNBC cells with k=29 using SC3.* ***A)*** *tSNE and* ***B)*** *UMAP projections of k-means clusters.* ***C)*** *Top 10 differentially expressed genes for each cluster.*

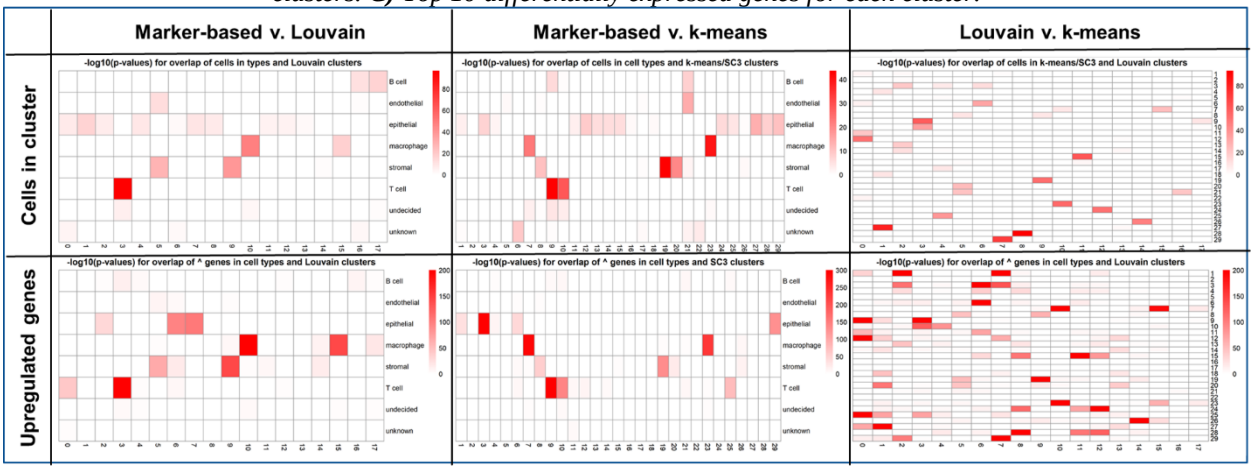


Figure 4. *MIA comparison of cell assignment to clusters (top) and genes upregulated in clusters (bottom) in marker-based types versus Louvain clusters (left), marker-based types versus k-means clusters (center), and Louvain clusters versus k-means clusters (right). Values are -log$_{10}$(p-values).*

Course project statement

This work is not currently related to any of my previous research or outside projects. All analysis shown was performed specifically for this course.

References

Andrews, T. S., & Hemberg, M. (2018). Identifying cell populations with scRNASeq. Molecular Aspects of Medicine, 59, 114–122. https://doi.org/10.1016/j.mam.2017.07.002

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

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., & Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. Nature Methods, 14(5), 483–486. https://doi.org/10.1038/nmeth.4236

Lun A. T. L., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology, 17(1), 75. https://doi.org/10.1186/s13059-016-0947-7

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., & Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nature Biotechnology, 38(3), 333–342. https://doi.org/10.1038/s41587-019-0392-8

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., & Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nature Methods, 14(3), 309–315. https://doi.org/10.1038/nmeth.4150

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology, 32(9), 896–902. https://doi.org/10.1038/nbt.2931

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression. Nature Biotechnology, 33(5), 495–502. https://doi.org/10.1038/nbt.3192