EPiC Series in Computing

Volume 83, 2022, Pages 155–164

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

Comparative Single-cell RNA-sequencing Cluster Analysis for Traumatic Brain Injury Marker Genes Detection

Audra Addison and Tayo Obafemi-Ajayi

Missouri State University, Springfield, Missouri
aba97s@missouristate.edu, tayoobafemiajayi@missouristate.edu

Single-cell RNA-sequencing (scRNA-seq) is a high-resolution transcriptomic approach used to discover gene expression patterns among cell types to study precise biological functions. Unsupervised machine learning (clustering) is of central importance for the analysis of scRNA-seq data. It can identify putative cell types, uncover regulatory relationships, and track cell lineages and trajectories. A key issue in clustering scRNA-seq data is determining which clustering method is appropriate to use, since varied methods can yield diverse results. Current approaches usually focus on a one method and manually select a seemingly meaningful result. From a biological relevance perspective, it is vital to distinguish between normal and pathogenic cell types using marker genes. We present a learning framework for comparing outcomes of multiple scRNA-seq clustering methods to determine the most optimal results. We address the challenges of model selection and validation metrics in the context of traumatic brain injury (TBI) applications. We compare clustering performance of five clustering algorithms and two dimensionality reduction techniques implemented in both Seurat and Scanpy packages.

1 Introduction

One of the most significant frontiers for computational scientists is the engineering of human healthcare delivery based on intelligent analysis of health data [15]. In diseases with high heterogeneity (such as cancer, autism spectrum disorders, autoimmune diseases, traumatic brain injury, etc.), there is a need to identify more homogeneous, clinically meaningful subgroups of patients to enable individualized care [16]. Advances in intelligent learning algorithms along with concurrent increase in biomedical data availability have yielded significant data-driven solutions with high potential to improve opportunities for personalized medicine [2]. In particular, computational analysis of single-cell RNA-sequencing (scRNA-seq) has amassed traction over the past few years as advances in high throughput sequencing technologies have allowed researchers to collect large catalogues detailing the transcriptomes of individual cells [5].

ScRNA-seq is a high-resolution transcriptomic approach used to discover gene expression patterns among cell types to study precise biological functions. It is an improvement over other sequencing technologies such as bulk RNA-sequencing where gene expression is taken as the total across the sample. Gene expression is measured as the number of times a gene has been "read," or transcribed to RNA. Since the genome of each cell within the same organism is identical, cell types are distinguished from each other by the expression level of each gene.

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB2022 (EPiC Series in Computing, vol. 83), pp. 155–164

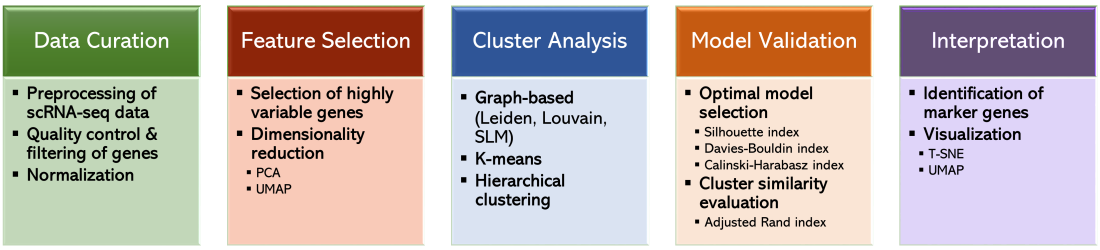


Figure 1: Overview of learning framework for analysis of scRNA-seq data.

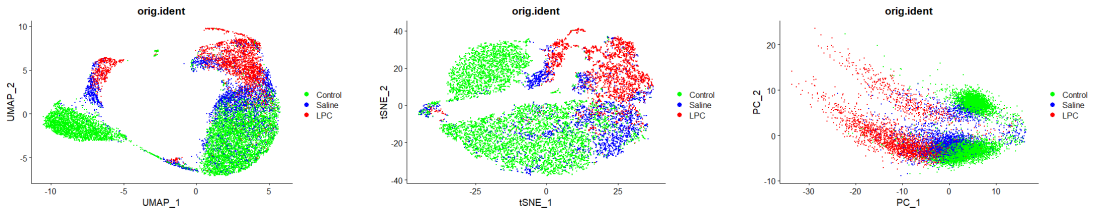
Beyond simple changes in average gene expression between cell types, scRNA-seq enables a high granularity of changes in expression to be unraveled. It reveals interesting and informative changes in expression patterns as well as cell type-specific changes in cell state across samples [7]. The ability to define cell types via unsupervised machine learning (clustering) on the basis of transcriptome similarity has emerged as one of the most powerful applications of scRNA-seq [5].

In this work, we are interested in scRNA-seq cluster analysis with respect to traumatic brain injury (TBI). TBI is a heterogeneous neurological disorder that is a disruption of brain function caused by a blow to the head which could result in lasting physical, emotional and cognitive impairments [24]. It is a global health problem affecting over 10 million people worldwide and is a leading cause of death, neurologic and neuropsychiatric disability in the United States [23]. Recent studies reveal that secondary damage of TBI occurs as apoptotic cascades are triggered in neurons after an initial impact which could result in permanent brain damage, disability, or death [23]. In some patients, inflammatory and apoptotic pathways are still active years after the trauma and cause neurodegenerative mechanisms that are also characteristic of other neurological chronic disorders such as Alzheimer's disease (AD). ScRNA-seq has the potential to discover differentially expressed genes responsible for promoting these pathways, as revealed by the transcriptional changes. Cluster analysis of these changes can lead to new insights in identifying complex differentially expressed genes that respond to TBI as well as determine cell populations that are altered post-injury. ScRNA-seq cluster analysis of TBI could enhance our understanding of persistent inflammation in the subacute and chronic time points after injury that affects surrounding neurons, key cells, and/or genes. This can also aid medical researchers in targeting recovery pathways for therapeutic intervention and predict TBI patient outcome.

Though significant progress has been made in clustering of scRNA-seq data, some questions remain unanswered. A key issue is how to determine which clustering method is appropriate for a given single cell data set and what parameters to utilize in order to achieve the most optimal clustering result, especially since different methods produce results with little overlap [3]. Validating the clustering results also remains a significant issue for scRNA-seq data as no ground truth is available. Our objective is to develop a learning framework for comparing outcomes of multiple scRNA-seq clustering methods to determine the most optimal results. We will address the challenges of model selection and validation metrics in the context of identifying possibly relevant TBI marker genes.

2 Methods

The overall framework, as illustrated in Figure 1, consists of five key steps. Data curation and feature selection are conducted prior to cluster analysis to ensure that the filtered representative data adheres to the single cell quality control standards. We are interested in comparing



(a) UMAP (b) t-SNE (c) PCA

Figure 2: Seurat plots comparing the gene expression for each condition (LPC injury = red, saline injury = blue, control = green) [4]: UMAP vs t-SNE vs PCA.

two dimensionality reduction techniques in the feature selection phase. Varied graph based clustering methods (Leiden, Louvain, smart local moving algorithm), as well as K-means and hierarchical clustering are performed on the reduced data with varying resolution and number of k clusters. Model validation involves optimal model selection utilizing certain internal validation metrics as well as evaluating similarity of clusters obtained across varied parameter settings. The final step, model interpretation, focuses on visualization and discussion of the clustering results as well as the process of determining the relevant marker genes.

2.1 Data Curation

We analyze two publicly available TBI data sets in this work. The first data set, GSE101901 [1] (single cell sequencing of hippocampus tissues in TBI), is derived from the hippocampus region of six samples of 10-week-old C57BL/6 male mice. A mild fluid percussion injury is imposed on three of the mice to reflect TBI while the remainder three are the sham (control) mice. 8397 hippocampal cells were collected and sequenced for 17621 genes, after 24 hours, for at least 500 genes and 900 transcriptomes per cell. The second data set is a subset of the GSE121654 [4] (complex cell-state changes revealed by single cell RNA sequencing of 76,149 microglia throughout the mouse lifespan and in the injured brain) data. The specific TBI samples derived from the data set consists of 13,359 microglial cells of the white matter brain region from nine 100-day old male mice (6 TBI, 3 control). The TBI condition is simulated via a demyelination injury (3 lysolecithin (LPC), 3 saline injection) that depicts the damage that occurs to the myelin sheath of the axons after a TBI. The total sample consists of 13755 genes across 13359 cells.

To ensure consistency in analysis, the preprocessing, filtering, and normalization steps were all implemented using the Seurat R package [17]. Cells were filtered out if they met at least one of the following conditions. 1) Cells with unique feature (gene) counts over 2500, as this implies duplicated reads. 2) Cells with less than 200 unique feature counts, which indicates low quality or resulting from empty droplets. 3) Cells with excessive mitochondrial contamination (denoted by over 5% of their feature counts), also considered low quality or dying cells. This resulted in 17621 genes across 7678 cells for the GSE101901 data set, and 13755 genes across 13246 cells for the GSE121654 data set. No cells were filtered due to mitochondrial contamination in either data set. The feature expression data was subsequently normalized using the aggregate sum of the expression measurements, a scale factor of 10,000, and natural log transformation.

2.2 Feature Selection

Feature selection using dimensionality reduction is an essential step, prior to clustering, for single cell analysis [5]. Usually these data sets are very large, as there are potentially thousands of genes that could define a cell, presenting both challenges and opportunities. A large data set ensures that analyses will have high power and enhances ability to detect rare cell types. However, visualizing and interpreting clustering results can be computationally difficult. Too many genes contribute to the 'curse of dimensionality' problem as too much information could bias the results due to noise, and hide significant patterns within the data. The goal of the feature selection phase is to identify the most informative genes for distinguishing between the cell types and combine them so that the gene expression patterns among the cells can be detected. This also speeds up the computational efficiency of the cluster analysis. The subset of most informative genes is based on those that exhibit high variability across the cells, and thus denote heterogeneous features to prioritize for subsequent analysis. This is determined by modeling the mean-variance relationship inherent in data [20].

We investigated three commonly used dimensionality reduction techniques: Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) [5]. PCA utilizes a linear approach to select the components that explain the most variation in the data with relatively few parameters to tune. It also preserves the global structure of the data, however it is heavily influenced by outliers and performs poorly at preserving local structure and detecting nonlinear relationships among cells. T-SNE and UMAP are both nonlinear methods that attempt to preserve the local structure of the data. Though t-SNE is more robust to possible outliers, it has a relatively longer computational time and does not preserve global structure well. T-SNE can only embed 2-3 dimensions, so it cannot directly work with high-dimensional data (sometimes an initial round of PCA is applied). UMAP is an efficient technique that relies on distance metrics to find similarities between neighbors. It is capable of preserving both the global and local structure of the data depending on the parameters.

These dimensionality reduction methods are also useful for visualization of the data, before and after clustering. This is helpful for illustrating and interpreting the cluster structures within the heterogeneous, high-dimensional scRNA-seq data. Figure 2 depicts the visualization of the GSE121654 data based on their known condition (control, saline, and LPC) using all three methods. As can be observed, UMAP yielded the most defined visual structures. Hence, for the remainder of this paper, we compare the performance of the clustering methods on PCA and UMAP transformed data and utilize UMAP for all the visualization post clustering.

2.3 Cluster Analysis

Graph-based community detection (clustering) methods are commonly used in single cell cluster analysis due in part to their scalability. The Louvain algorithm is a well-known algorithm based on optimizing modularity [21]. It has low time complexity and is frequently employed for clustering large data sets. The Leiden algorithm [21] represents an improvement over Louvain. It converges to a partition in which all subsets of all communities are locally optimally assigned. It is also faster than the Louvain algorithm and uncovers better partitions. In addition, it has been proven that the Leiden algorithm yields communities that are guaranteed to be connected. The SLM algorithm is another modularity-based community detection algorithm [22]. In contrast to the other modularity based methods, it splits the communities by moving sets of nodes between the communities to search for more possibilities to increase modularity. We apply these graph based methods for varying values of the resolution parameter (influences the resulting

number of clusters). In addition, we also apply k-means algorithm as well as hierarchical (agglomerative) clustering method for varying values of number of clusters (k). These clustering methods were implemented in both Seurat and Scanpy [18] packages.

2.4 Model Validation

We evaluated the varied clustering results to determine the most optimal results using inter- and intra-cluster similarity. Three different internal validation metrics were applied: Silhouette index (SI), Davies-Bouldin (DB) index, and Calinski-Harabansz (CH) index [11]. SI relies on the compactness and separation of the clusters in its evaluation based on the pairwise difference of between-and within-cluster distances. The scores range from -1 to 1, with negative scores representing samples with wrong assignments and scores of 0 indicating overlapping clusters while 1 denotes the most optimal result. The DB index measures the average value of similarity between each cluster and its most similar cluster. Similar to SI, it also computes the compactness and separation of the clusters, but in contrast, it compares each cluster to every other cluster using the sum of the average distances from the center of each cluster. Thus, clusters that are far from other clusters and are densely packed result in a better score. Values closest to 0 indicate more optimal performance. The CH index measures between-cluster isolation and within-cluster coherence. A higher score indicates better cluster performance. To compare similarity of clustering results, we utilized the adjusted rand index (ARI). The similarity score ranges from -1.0 to 1.0. A score of 1.0 means that the clustering results are identical, while a score of 0 means the clustering results are random or completely independent of each other. A negative score suggests that the clustering results have an orthogonal relationship.

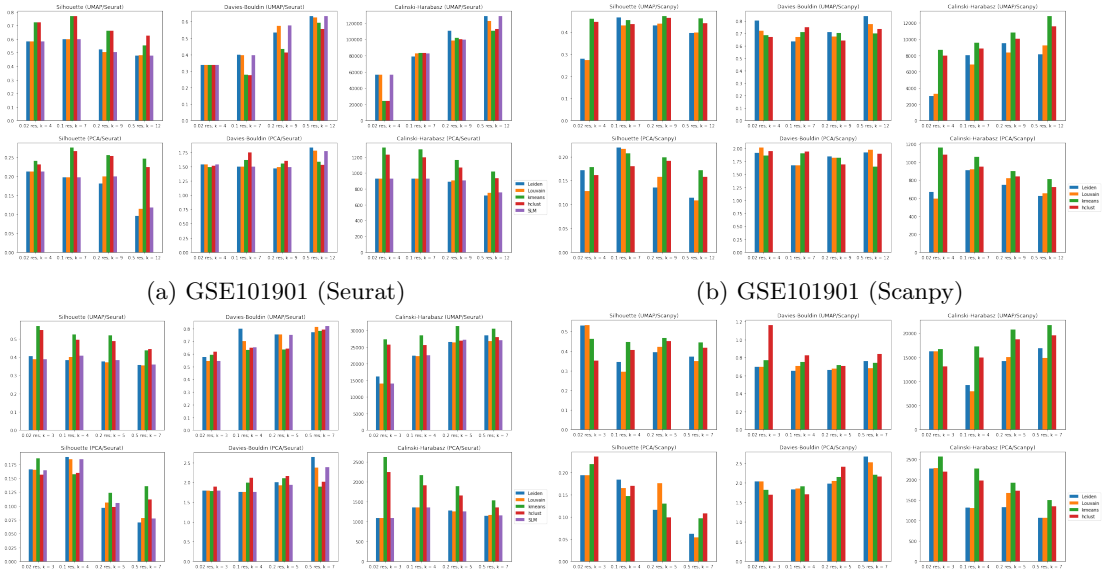
2.5 Model Interpretation

For scRNA-seq cluster analysis, visualization of the resulting clusters is very key in model interpretation. It allows the human users to effectively extract meaningful biological information and identify novel cell subtypes based on the discovered patterns and relationships within the data [8]. As mentioned in Section 2.2, to project high-dimensional data into a 2D or 3D space for visualization, we utilize the dimension reduction techniques that were employed for feature selection. Different visualization techniques exploit the nonlinear structures and patterns in the same result in multiple ways. One technique may reveal a pattern more clearly than the other (Figure 2), hence we utilized UMAP for all the visualization plots. Another component of model interpretation is the identification and analysis of the marker genes from the clusters. Marker genes are defined as the genes within a cluster or group of cells that are highly differentiated from other clusters or groups of cells. These genes can be used to define cell types or identify genes responsible for promoting certain pathways.

3 Results and Analysis

3.1 Experiment Setup

The implementation of the data curation and feature selection steps were all done in Seurat for consistency. Prior to dimensionality reduction, the top 10 highly variable genes (features) were determined using the mean-variance method in Seurat. The data matrix is scaled using a linear transformation such that the mean expression across all cells was 0 and the variance across all cells was 1. PCA and UMAP dimensionality reduction techniques were then compared using



(a) GSE101901 (Seurat) (b) GSE101901 (Scanpy)

(c) GSE121654 (Seurat) (d) GSE121654 Scanpy)

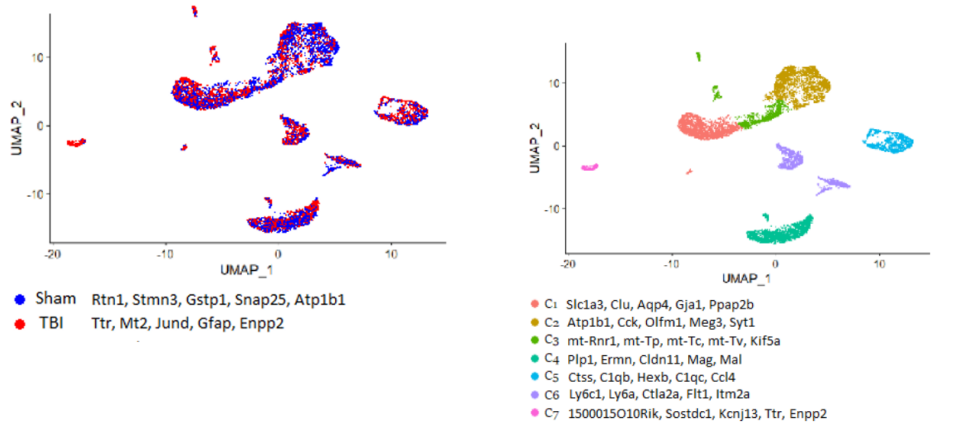
Figure 3: Comparison of clustering performance using validation metrics per data set by each clustering algorithm and dimensionality reduction technique.

the top 10 components generated from each method to compute the k-nearest neighbors (kNN) graph using euclidean distance for the community detection clustering algorithms. Kmeans and hierarchical clustering algorithms were also performed using both PCA and UMAP coordinates as well as the euclidean distance metric.

The graph based clustering techniques (Leiden, Louvain, and SLM) were implemented both in Seurat and Scanpy to compare their performance using varying resolutions (0.05, 0.02, 0.1, 0.2, 0.5). For ease of comparison with the graph based clustering results, the k-means hierarchical clustering methods were implemented with $k = [4, 7, 9, 12]$ for the GSE101901 data set; and $k = [3, 4, 5, 7]$ for the GSE121654 data set to match the range of clusters found using the community detection methods. We extracted the top 5 marker genes within each cluster for each experiment for the optimal clustering results (as determined by the validation metrics) to identify potential populations of cells susceptible to TBI.

3.2 Cluster Validation

The SI, DB, and CH validation results for both GSE101901 and GSE121654 are shown in Figure 3 by clustering algorithm and dimensionality reduction technique and by package environment implementation. As can be observed, the results obtained vary by implementation package, even though similar user-defined parameters and preprocessing were utilized. More experiments are needed to fully investigate the underlying reasons for the differences. Given that the validation metrics take into account differing properties of the cluster structure, it isn't surprising that the results differ. The SI and DB indices tended to either be worse or stay the same as the value of k or resolution increased. Interestingly, the CH index results for Scanpy and Seurat were oppositely correlated for both data sets. Increasing the value of k or resolution always resulted in better CH scores in Seurat but worse CH scores in Scanpy. The similarity between



(a) Visualization of GSE101901 by TBI (red) and Sham/control (blue) cells.

(b) Top clustering result for GSE101901 by SI & DB indices: hierarchical clustering (k = 7)

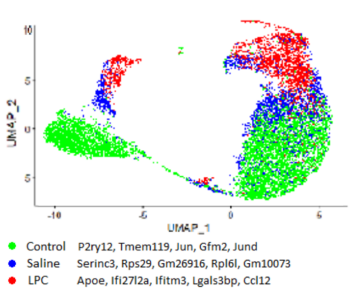
Figure 4: Top clustering result of GSE101901 compared to distribution of control and TBI cells. The top five marker genes per cluster are listed in the legend. C7 cluster is mainly TBI cells.

the clustering results were also compared to each other using their ARI scores. Leiden and Louvain methods resulted in the most similar clusters according to their average ARI of 0.880, followed by comparing two different resolutions of the same clustering method (average ARI of 0.721). Comparing Seurat and Scanpy clusters resulted in the lowest average ARI of 0.444.

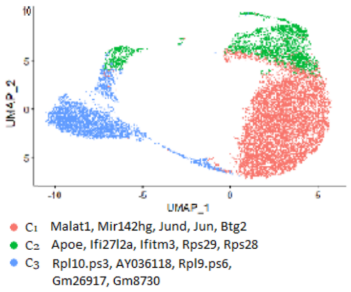
Generally, clustering results obtained using the Seurat package and UMAP coordinates generated better validation metrics than those using the Scanpy package or PCA coordinates. For the GSE101901 data set, the worst validation scores (SI of 0.05, DB of 2.66, CH of 1063.31) resulted from using Louvain at 0.5 or 0.02 resolution on PCA coordinates using Scanpy. The best clustering results (SI of 0.72, DB of 0.28, and CH of 128,633.94) was obtained by clustering using UMAP coordinates with Seurat's hierarchical clustering technique ($k = 7$). For the GSE121654 data set, the worst validation scores (SI of 0.11, DB of 2.01, CH of 594.73) resulted from using Leiden or Louvain at 0.5 resolution. The best clustering results (SI of 0.57, DB of 0.55, CH of 31,365.61) was obtained with UMAP coordinates using k-means ($k = 3$ and $k = 5$) and Louvain (resolution of 0.02) algorithms in R.

3.3 Analysis of TBI Marker Genes Findings

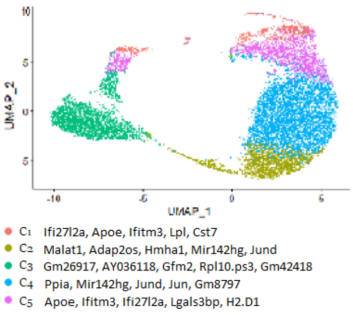
The best clustering result of the GSE101901 data set according to two validation metrics (SI and DB) is shown in Figure 4b. When compared to the distribution of TBI and control cells in Figure 4a, C7 cluster stands out as a group of cells defined by TBI. C7 cluster and TBI cells share two marker genes: Ttr and Enpp2. In contrast, the other six clusters do not share any marker genes with those associated with TBI. For the GSE121654 data set, the marker genes of the top three clustering results according to each validation metric are compared to the marker genes for each condition (LPC injury, saline injury, and control) in Figure 5. C2 cluster from the best SI score (Figure 5b), C1 and C5 clusters from the best CH score (Figure 5c), and C2 cluster from the best DB score (Figure 5d) seem to match up well with the LPC injury cells as shown in Figure 5a. All top five marker genes (Apoe, Ifi27la, Ifitm3, Lgals3bp, and Ccl12) expressed in LPC cells were also marker genes in C2 cluster with the best DB score, while the



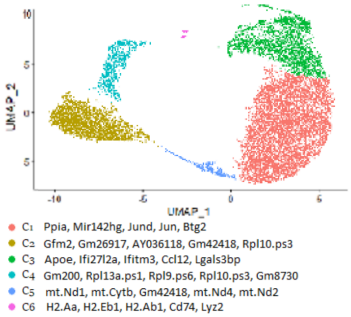
(a) Visualization by conditions: LPC (red), saline (blue), and control (green) cells.



(b) SI; k-means (k = 3)



(c) CH; k-means (k = 5)



(d) DB; Louvain (0.02 res)

Figure 5: Top clustering result of GSE121654, using SI, DB, and CH indices, compared to the control and TBI (Saline & LPC) cells. We identify the top 5 marker genes per cluster.

top CH score had four of the five (Apoe, Ifitm3, Ifi27l2a, and Lgals3bp) in C5 cluster 5 and the top SI score had three of the five (Apoe, Ifi27l2a, and Ifitm3) in C2 cluster. The clustering result with the best SI score resulted in three clusters, distributed similarly to the conditions in Figure 5a. The best clustering results according to CH and DB metrics resulted in more clusters, but also more accurate marker genes.

For the GSE101901 data set, there appears to be a clear optimal clustering result out of the experiments performed. C7 cluster from this result seems to be specific to TBI, meaning the marker genes from this cluster are likely involved in promoting pathways that are triggered after a TBI. Two genes, Ttr and Enpp2, were highly differentiated in both TBI cells and in cluster 7. The transthyretin protein encoded by the Ttr gene is responsible for transporting the thyroid hormone thyroxine across the blood brain barrier [14] and binding to beta-amyloid deposition; those with reduced expression of this gene are at a greater risk of AD [19]. Enpp2 has previously been found to be differentially expressed in metastatic cancers, including brain cancer [10]. Other sources that utilized the same data set [1, 25] both found significant results regarding the Ttr gene. Ttr was found to be the most differentially expressed gene after TBI among all hippocampal cell types; using this information, modulating the Ttr pathway was found to mitigate adverse effects of TBI [1]. The transthyretin protein encoded by Ttr has also been significantly increased in endothelial and mural cells including pericytes after TBI [25].

The optimal clustering results for the GSE121654 data set were able to successfully uncover most, if not all, of the marker genes associated with LPC injury. The Apoe gene was the most differential gene associated with LPC injury; certain alleles of this gene promote the

development of Alzheimer's disease (AD) [6], providing an interesting link between AD and TBI. IFN-stimulating genes Ifi27l2a and Ifitm3 have previously been described as preventing viruses from invading the central nervous system and spreading in the brain [9]. Lgals3bp is down-regulated in the presence of beta-amyloid, a biomarker for Alzheimer's [12]. Ccl12 is a gene that has been confirmed to remain persistently upregulated following TBI [13]. Several of these marker genes specific towards LPC injury agreed with the results obtained in [4], where a majority of the cells within the LPC-specific cluster (named IR2) were found to express the genes Fcrls, Apoe, Ifl27l2a, Cxcl10, Ccl4, and Birc5. The cluster IR2-specific genes were found to be variably upregulated, so the IR2 cluster was subclustered to determine four specific populations of microglia that respond to LPC injury.

4 Conclusion

We presented a learning framework for comparing outcomes of multiple scRNA-seq clustering methods to determine the most optimal results. We tackled the challenges of model selection and validation metrics in the context of identifying possibly relevant TBI marker genes. ScRNA-seq has the potential to discover differentially expressed genes responsible for promoting these pathways, as revealed by the transcriptional changes. The marker genes specific to TBI tend to agree across clustering methods, but not across data sets. Only one marker gene (Jund) appears in both data sets, and it is a marker gene for TBI cells in GSE101901 and for control cells in GSE121654. These differences are probably due to the different brain regions (hippocampus vs subcortical white matter), cell types (heterogenous vs microglia), or time after the TBI (1 day vs 7 days). There were no shared marker genes between LPC and saline injury cells in the GSE121654 data set, suggesting that the type of injury may play a significant role in the expression of certain genes.

References

[1] Douglas Arneson, Guanglin Zhang, Zhe Ying, Yumei Zhuang, Hyae Ran Byun, In Sook Ahn, Fernando Gomez-Pinilla, and Xia Yang. Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nature Communications*, 9(1):1–18, 2018.

[2] Lien A Bui, Dacosta Yeboah, Louis Steinmeister, Sima Azizi, Daniel Hier, Donald Wunsch, Gayla R Olbricht, and Tayo Obafemi-Ajayi. Heterogeneity in blood biomarker trajectories after mild TBI revealed by unsupervised learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[3] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7, 2018.

[4] Timothy R Hammond, Connor Dufort, Lasse Dissing-Olesen, Stefanie Giera, Adam Young, Alec Wysoker, Alec J Walker, Frederick Gergits, Michael Segel, James Nemesh, et al. Single-cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. *Immunity*, 50(1):253–271, 2019.

[5] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[6] Olga N Kokiko-Cochran and Jonathan P Godbout. The inflammatory continuum of traumatic brain injury and Alzheimer's disease. *Frontiers in Immunology*, 9:672, 2018.

[7] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

[8] Zhenqiu Liu. Visualizing single-cell RNA-seq data with semisupervised principal component analysis. *International Journal of Molecular Sciences*, 21(16):5797, 2020.

[9] Huanle Luo and Tian Wang. Recent advances in understanding West Nile virus host immunity and viral pathogenesis. *F1000Research*, 7, 2018.

[10] Shahan Mamoor. Enpp2 is a differentially expressed gene in human metastatic breast cancer, in the brain and in the lymph nodes. *OSF Preprints*, 2021.

[11] Thy Nguyen, Jason Viehman, Dacosta Yeboah, Gayla R Olbricht, and Tayo Obafemi-Ajayi. Statistical comparative analysis and evaluation of validation indices for clustering optimization. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 3081–3090. IEEE, 2020.

[12] Jong-Chan Park, Sun-Ho Han, Hangyeore Lee, Hyobin Jeong, Min Soo Byun, Jingi Bae, Hokeun Kim, Dong Young Lee, Dahyun Yi, Seong A Shin, et al. Prognostic plasma protein panel for aβ deposition in the brain in Alzheimer's disease. *Progress in Neurobiology*, 183:101690, 2019.

[13] Katarzyna Popiolek-Barczyk, Agata Ciechanowska, Katarzyna Ciapała, Katarzyna Pawlik, Marco Oggioni, Domenico Mercurio, Maria-Grazia De Simoni, and Joanna Mika. The ccl2/ccl7/ccl12/ccr2 pathway is substantially and persistently upregulated in mice after traumatic brain injury, and ccl2 modulates the complement system in microglia. *Molecular and Cellular Probes*, 54:101671, 2020.

[14] Samantha J Richardson, Roshen C Wijayagunaratne, Damian G D'Souza, Veerle M Darras, and Stijn LJ Van Herck. Transport of thyroid hormones via the choroid plexus into the brain: the roles of transthyretin and thyroid hormone transmembrane transporters. *Frontiers in Neuroscience*, 9:66, 2015.

[15] Suchi Saria. A $3 trillion challenge to computational scientists: Transforming healthcare delivery. *IEEE Intelligent Systems*, 29(4):82–87, 2014.

[16] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.

[17] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[18] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.

[19] João Carlos Sousa, Isabel Cardoso, Fernanda Marques, Maria João Saraiva, and Joana Almeida Palha. Transthyretin and Alzheimer's disease: where in the brain? *Neurobiology of Aging*, 28(5):713–718, 2007.

[20] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[21] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):1–12, 2019.

[22] Ludo Waltman and Nees Jan Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, 86(11):1–14, 2013.

[23] Kristina G Witcher, Chelsea E Bray, Titikorn Chunchai, Fangli Zhao, Shane M O'Neil, Alan J Gordillo, Warren A Campbell, Daniel B McKim, Xiaoyu Liu, Julia E Dziabis, et al. Traumatic brain injury causes chronic cortical inflammation and neuronal dysfunction mediated by microglia. *Journal of Neuroscience*, 41(7):1597–1616, 2021.

[24] Dacosta Yeboah, Louis Steinmeister, Daniel B Hier, Bassam Hadi, Donald C Wunsch, Gayla R Olbricht, and Tayo Obafemi-Ajayi. An explainable and statistically validated ensemble clustering model applied to the identification of traumatic brain injury subgroups. *IEEE Access*, 8:180690–180705, 2020.

[25] Bangyan Zhang, Xinying Guo, and Zhen Zhao. Secondary single-cell transcriptomic analysis reveals common molecular signatures of cerebrovascular injury between traumatic brain injury and aging. In *AAIC Neuroscience Next*. ALZ, 2020.