An Improved Algorithm for Rare Cell-Type Detection Based on GiniClust3

George Saab, Christia Victoriano, Chenghao Zhang
Columbia University

Abstract **-** The role of rare cells in determining the pathogenesis of cancer, mediating immune responses and other diseases has been discovered, thus detecting rare cells could help us to have a new and deeper understanding of their role in biological functions and disease progression. Yet how to efficiently detect rare genes remains a pressing problem. Methods have been created to identify common cell populations, but lack the sensitivity to find rare populations. Alternative methods designed specifically for rare cell cluster identification, such as GiniClust3, were proposed to alleviate the issues with traditional common clustering schemes by integrating both common and rare cell identification metrics, but fail to identify potential dependencies between gene-gene interactions due to their assumption on genes being independent of each other. Our DESC-integrated method aims to improve upon the existing GiniClust3 framework by utilizing a more sophisticated deep learning approach. Our intention is to improve the clustering step to allow for more sensitive identification of rare cell populations within scRNA-seq datasets.

A. Introduction

The rise of single-cell RNA-sequencing (scRNA-seq) technology has made highly parallel, robust genome-wide transcriptional analysis of individual cells among complex tissue a reality. Studying gene expression at single-cell resolution will allow us to further characterize heterogeneous tissues, providing the potential for identifying biologically-significant phenomena previously undetected in bulk tissue analysis. Importantly, scRNA-seq has transformed the ability of scientists to characterize, classify, and identify cell types and states. Proper cell type classification enables robust downstream analysis of gene expression of these cell types, allowing us to elucidate the functions of these cells within biological systems, which might enable the identification of new biomarkers or more specific therapeutic targets.

In particular, the identification of rare or minor cell types within bulk heterogeneous tissue can lead us to novel insights into their roles in biological function and disease progression. Examples of rare cell types include circulating tumor cells, cancer stem cells, progenitor cells, antigen-specific T cells, and invariant natural killer T cells. They can also include minor subpopulations or subtypes of more common cell types, such as endothelial cells or hepatocytes, with distinct gene expression profiles that have more specialized functions or responses [1]. Many of these rare cell populations play an important role in determining the pathogenesis of cancer, mediating immune responses, and angiogenesis in cancer and other diseases [2].

Many methods have been developed for characterization of cell types among complex tissues from single-cell transcriptomic data [3-5]. However, most of these methods are designed to be effective at identifying common cell populations but are not sensitive enough to detect rare or minor cell types. Identification of rare cell types from scRNA-seq data presents a major computational challenge, as these cell populations can represent as small as a few cells in a dataset of thousands, and thus the distinctive features of these cell types can be mistaken for noise or technical variation when it is actually biological variation.

Methods specifically developed for detecting the transcriptomes of rare cell types are scarce. Among the most prominent is RaceID [6]. Following a k-means clustering step, RaceID identifies outlier cells in each cluster, defining an outlier cell as expressing a certain number of outlier genes at levels significantly exceeding the modeled noise. Thus, cell type identification is not reliant on global differences between cells– only on a few outlier genes. However, the algorithm is limited by its use of k-means as the unsupervised clustering algorithm prior to outlier identification, which is reliant upon selecting the correct number of clusters. Additionally, its outlier detection method involves computationally expensive parametric modeling for the detection of outlier expression profiles, which reduces its scalability to larger datasets.

A promising alternative to RaceID is GiniClust3, which uses a more straightforward two-pronged algorithm to identify both rare and common cell types [7]. In GiniClust3, two gene selection methods are used on the same dataset in parallel: the Gini index, which is used to systematically identify genes within rare cell types, and the Fano factor, which is a valuable metric for capturing differentially expressed genes specific to common cell types [8]. Following gene selection, GiniClust3 performs Leiden clustering on each resulting expression matrix, and the result is two differing clustering results which are then consolidated into a simplified consensus matrix via a cluster-aware, weighted consensus clustering approach. A final k-means clustering step is then performed on the consensus matrix to yield the final cell type clusters. GiniClust3 is less computationally expensive than RaceID and can classify both rare and common cell types within the same algorithm. However, like RaceID, GiniClust3 is heavily reliant on unsupervised clustering following feature selection to detect communities

of outlier cells. Furthermore, it fails to alleviate the technical issues stemming from batch effects.

However both RaceID and GiniClust3 perform gene expression analysis on a gene-by-gene basis. These two approaches assume that the expression of each gene is independent of each other, thus ignoring the possible dependence between the expression of different genes. These methods are therefore computationally efficient but are likely to miss interesting patterns when a set of genes forms a specific expression altogether and may ignore certain patterns when one gene expression is influenced by other genes.

In this study, we aim to improve upon GiniClust3 by leveraging more sophisticated deep learning methods to improve the clustering step in order to allow for more sensitive identification of rare cell populations within large scRNA-seq datasets. Specifically, we have integrated DESC [9] into the GiniClust3 pipeline in place of simple Leiden clustering. DESC is an unsupervised deep embedding algorithm that clusters scRNA-seq data by iteratively optimizing a clustering objective function, gradually removing technical variation through iterative self-learning. This iterative procedure balances biological and technical differences between clusters. DESC also assigns cluster-specific probabilities to each cell, providing valuable, interpretable information for identification of novel cell subpopulations or transitive states by showing degrees of similarities to other clusters.

To determine the validity of our algorithm as well a scompare the performance of GiniClust3 and our deep-learning-based approach, we have applied both methods to a previously annotated scRNA-seq dataset of sinus node and atrial muscle tissue biopsies from 4 mice hearts in which we artificially generated “rare cell types” by altering the number of cells in certain clusters.

B. Computational Methodology

A. *Data pre-processing*

Since rare cell types are observed in only less than one percent of the total cell population, many conventional pre-processing pipelines must be avoided. Some of these techniques, such as left-truncated total Unique Molecular Identifier (UMI) histograms, result in the removal of subset populations that may only express a few key genes. We opted to remove cells expressing less than 2000 genes, and genes expressed in fewer than 3 cells as suggested by Tsoucas et. al. [9] The resulting cell group was then normalized to a UMI count per cell of 10,000.

B. *Parallel feature selection using Gini index and Fano factor analysis*

In GiniClust3, after data pre-processing, genes specific to rare cell types are identified using the Gini index. The Gini index was originally developed to study social inequality and has been used to identify countries whose wealth is concentrated within a small number of individuals [10]. Thus, it is suitable for identifying rare cell type-specific genes. For each gene X, the cells are sorted based on its expression levels from lowest to highest, and then the cumulated expression levels of X as more and more cells are included from the ranked list are evaluated. This functional relationship is plotted and is called a Lorenz curve. The Gini index is defined as two times the area between the Lorenz curve and the diagonal. The Gini index ranges from 0 (most uniform) to 1 (most extreme). The Gini index values are then normalized using a two-step locally estimated scatterplot smoothing (LOESS) regression procedure. A smooth curve is fit through all data points by LOESS regression, outliers for which the residues are above the 75th percentile are removed, and then LOESS is used to refit another smooth curve to the remaining data points. For each gene, Gini index value was calculated by subtracting the original value by the fitted trend. P values were further estimated based on a normal distribution approximation. Genes with Gini index value ≥ 0.6 and p value < 0.0001 are labeled as high Gini genes and selected for downstream analysis.

To make the algorithm robust for detecting genes differentially expressed in common cell types, highly variable genes are selected on the same dataset using the Fano factor as a metric for variability in parallel.

The Fano factor is defined as the variance over the mean expression value for each gene.

$$F = \frac{\sigma^2}{\mu}$$

The top 1000 highly variable genes based on Fano factor were chosen for further analysis. The result of this parallel feature selection step is two separate expression matrices using different sets of genes.

C. *Clustering using DESC*

The resulting expression matrices from the previous step are both clustered using DESC. Using a deep neural network, DESC initializes parameters obtained from an autoencoder and learns a nonlinear mapping function from the original scRNA-seq data space to a low-dimensional feature space by iteratively optimizing a clustering objective function.

The DESC procedure begins with parameter initialization, in which a stacked autoencoder is used for pretraining and learning a low-dimensional representation of the input gene expression matrix. The stacked autoencoder network is initialized layer by layer; each layer is an autoencoder trained to reconstruct the output of the previous layer. Encoder layers are concatenated followed by decoder layers after layer-wise training, resulting in a multilayer autoencoder with a bottleneck layer in the middle. The decoder layers are then discarded, and the encoder layers are used as the initial mapping between the original data space and the dimension-reduced feature space. The Louvain clustering algorithm is then applied to the feature space obtained from the bottleneck layer, returning the number of

clusters and the corresponding centroids for each cluster. These will be used as the initial clusters.

Following cluster initialization, clustering is improved iteratively using an unsupervised algorithm that alternates between two steps until convergence. In the first step, a soft assignment of each cell is computed between the embedded points and the cluster centroids using a student's t-distribution as a kernel to measure the similarity between the embedded point and centroid

$$q_{ij} = \frac{(1+||z_i-\mu_j||^2/\alpha)^{-1}}{\sum_{j'} (1+||z_i-\mu_{j'}||^2/\alpha)^{-1}}.$$

In the second step, the clusters are refined by learning from cells with high-confidence cluster assignments. Specifically, the objective function is defined as KL divergence loss between the soft cell assignments qi and auxiliary distribution pi for cell i as

$$L = KL(P||Q) = \sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij} log \frac{p_{ij}}{q_{ij}}$$

where the auxiliary distribution P is defined as

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{n} q_{ij}}{\sum_{j=1}^{K} \left(q_{ij}^2 / \sum_{i=1}^{n} q_{ij}\right)}.$$

The encoder is fine-tuned by minimizing L iteratively. P also gives the probability that cell i belongs to cluster j, which can be used to measure the confidence of the cluster assignment for each cell. The cluster centroids and the deep neural network parameters are then optimized using stochastic gradient descent. During each iteration, the auxiliary distribution P is updated, and cluster centers and encoder parameters are optimized with the new P. This iterative procedure stops when the proportion of cells that changes cluster assignment between two consecutive steps is less than 0.005.

D. Integration into a single simplified consensus matrix

The output DESC matrices produced from the respective Gini index and Fano factor expression grids are then inputted into a weighted consensus clustering algorithm, as detailed by Tsoucas et. al. [9] and will be described again as follows. Cells are determined to be connected if they belong to the same cluster, as defined by the connectivity matrix equation

$$M_{ij}(P^{DG}) = 1, (i,j) \in C_k(P^{DG});\ 0\ otherwise$$

and

$$M_{ij}(P^{DF}) = 1, (i,j) \in C_k(P^{DF});\ 0\ otherwise,$$

where DG and DF are the partitions provided by DESC-GiniClust3 and the DESC-Fano resultant matrices, respectively, and C_k represents the set of k clusters. Cells that were identified to belong to a cluster among k clusters will be assigned a value of 1 while labeled 0 for un-assigned clusters. These binary values are then multiplied by their associated weight, producing the consensus matrix

$$CM_{ij} = w_{ij}^{DG} M_{ij}(P^{DG}) + w_{ij}^{DF} M_{ij}(P^{DF}),$$

Where CM is the formulated consensus matrix and w_{ij} being the associated standard weight for each cell connectivity in both DESC-Gini and DESC-Fano clustering results. In Tsoucas et. al., the weights produced by the GiniClust3 scheme (w_{ij}^{G}) is close to 1 for rare clusters and 0 for common clusters, specified by the function

$$w_{ij}^{G} = 1 - \frac{1}{1+e^{-f}}$$

where f is the detection sensitivity of rare cell clusters based on the proportion of cells present relative to a given cluster [9]. The implemented value of f determines the relative confidence towards the presence of a rare cluster and the associated weighted matrix is introduced to DESC to focus primarily on rare cluster identification. Afterwards, the sum of the weight connectivities provides a soft clustering, assigning a probability that a cell belongs to a specific cluster. Deterministic cluster assignment is given by the maximum function

$$C_f = max(w_{ij}^{DG}, w_{ij}^{DF}), (i,j) \in CM_{ij}(C_k)$$

where C_f is the final assigned cluster among $C_k$clusters in the consensus matrix. To transmit the relative probability of a given cell belonging to its cluster, the maximum of the two weights is determined:

$$\underline{w}_{ij} = max(w_{ij}^{DG}, w_{ij}^{DF})$$

with $\underline{w}_{ij}$ being the output probability for each cell in clusterC_f.

E. Final cluster generation

Upon the creation of a singular consensus matrix, cluster assignment can be determined by the associated higher probability between the weights defined by the DESC-Gini and DESC-Fano pipelines. The final consensus matrix CM_{ij} is clustered using K-means and projected onto a two-dimensional Uniform Manifold Approximation and Projection (UMAP) display. Clusters that represent < 1% of the overall cell population are tagged as rare.

C. IMPLEMENTATION

A. *Generation of "rare cell clusters" using Splatter*

In order to evaluate algorithm performance, we generated several artificial single cell RNA-sequencing datasets using Splatter[11] (R package). The simulated datasets contain 500 cells and 2000 genes with highly unbalanced cell group proportion. The simulated datasets contain three cell types, the proportion of the smallest cluster ranges from 1% to 10%, while the proportion of the two major ones are around 50%. In the preprocessing step, genes for which expression counts exceeded 200 in at least 3 cells were reserved for downstream analysis.

B. Rare cluster generation using our own method

To further develop rare cluster generation, we implemented our own clustering technique using An annotated scRNA-seq dataset found on NCBI's Gene

Expression Omnibus (GSE130710). This data was taken from sinus node and atrial muscle biopsies from 4 mice with 4 replicates each for a total of 16 tissue samples. The data comprises 5357 cells and covers a total of 27,998 genes, with a per-cell average of 2715 expressed genes. Based on expression profiles, the authors of the dataset identified 12 distinct cell clusters containing the following cell types: epicardial cells, vascular endothelial cells, sinus node myocytes, epithelial cells, neurons, endocardial cells, three subpopulations of adipocytes, and two subpopulations of fibroblasts, as shown in Table 1.

Table

Cluster Number	Cell Type
0	Epicardial Cells
1	Fibroblasts I
2	Adipocytes III
3	Vascular Endothelial Cells
4	Sinus Node Myocytes
5	Adipocytes II
6	Adipocytes I
7	Epithelial Cells
8	Neurons
9	Endocardial Cells
10	Fibroblasts II
11	Macrophages
12	Sinus Node Myocytes

Table 1: Table of annotated cell types present in NCBI Gene Expression Omnibus dataset (GSE130710) and respective cluster it belongs to.

The data was available as a normalized expression matrix and metadata csv file with cell source information as well as annotated cell type. A Uniform Manifold Approximation and Projection (UMAP) visualization of the normalized data colored by cell type is shown in Figure 1.

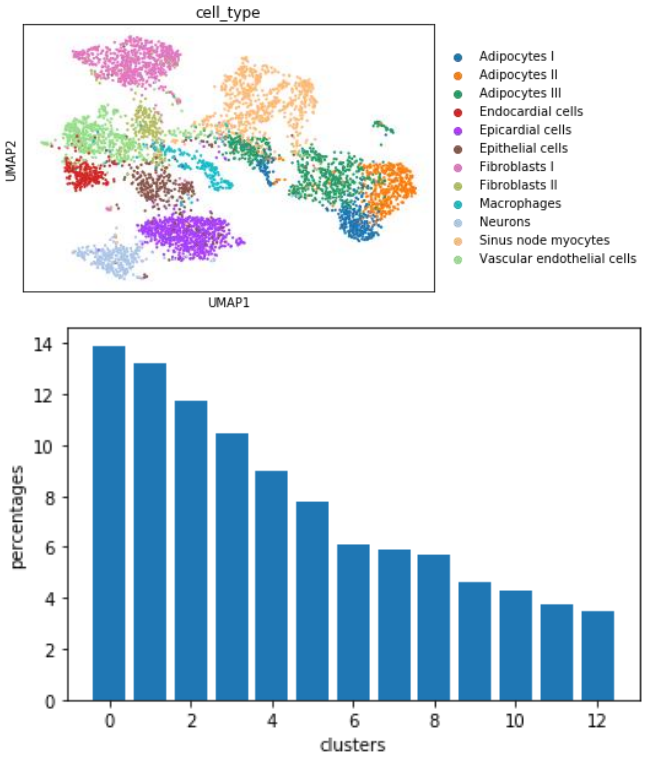


Figure 1: Representation of scRNA-seq dataset. (Top) UMAP projection of clusters and (bottom) cluster cell percentage of total cell population.

In order to generate a dataset that contains a "rare cell type," defined as a cell type representing fewer than 1% of the overall population within the dataset, a cluster index was randomly selected from the 12 within the dataset. Within that cluster, a threshold was set for the number of cells to keep. From that value, random index values were generated using the numpy random.int function, which generates random integers using a discrete uniform distribution. These index values indicate the specific indices of cells to keep within the selected cluster. Using these indices, the normalized expression matrix was then updated to keep only those cells, yielding an expression matrix containing the new filtered cluster. A UMAP visualization of an example "rare cell type"-containing expression matrix generated using the aforementioned approach is shown in Figure 2. Additionally, the number of rare cell type clusters that are generated can be altered to generate multiple rare cell types.

C. Real dataset

In order to estimate the performance of our model when dealing with real problems, we downloaded the dataset consisting in 20K Neurons, downsampled from 1.3 Million Brain Cells from E18 Mice from 10X genomics website[12]. This dataset contains 20k cells obtained from cortex, hippocampus and ventricular zones of E19 mice with 27998 genes expression. The preprocessing procedure is the same as the simulated dataset.

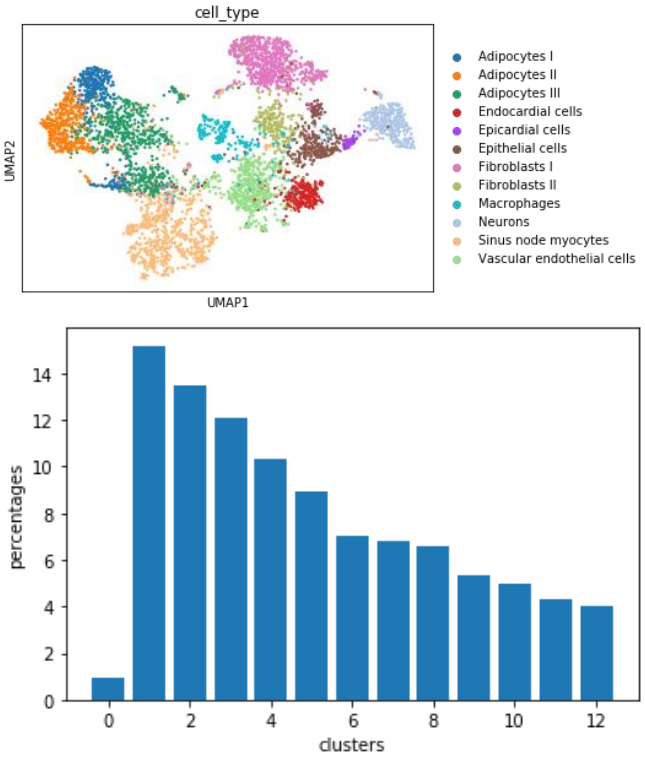


Figure 2: Representation of scRNA-seq dataset with Epicardial Cells selected as rare cluster. (Top) UMAP projection of clusters and (bottom) cluster cell percentage of total cell population. Cluster 0 has been reduced to 1% of total population, making it a rare cluster.

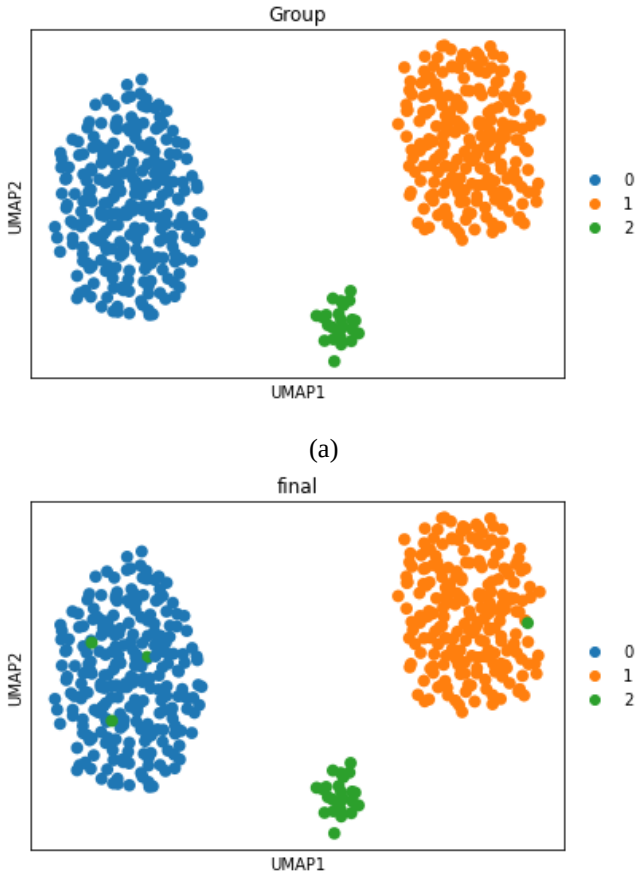
D. Comparison of DL-integrated algorithm and GiniClust3 performance on generated rare cell cluster dataset

Comparison of our improved DESC-integrated clustering algorithm with GiniClust3 must first be considered on a synthetically developed rare cluster dataset to determine whether improved performance has been developed.

A simulated dataset using Splatter [11] was utilized to introduce a synthetic three class scRNA-seq data with 500 cells and 2000 total genes. A rare cell cluster was introduced that represents a proportion of the total cell population ranging from 10-1%. The data was pre-processed to filter cells with no genes present and genes that have a total UMI count of less than 3. The resulting expression matrix was then normalized to 10000 UMIs. Apart from the preprocessing, we used the default parameters when implementing GiniClust3.

The normalized expression matrix was used to calculate the fano factor and Gini indices. For the GiniClust3 pipeline, a consensus matrix was produced using the resultant indexed expression matrices, creating a GiniClust3 consensus matrix. For our method, we introduced the respective indexed expression matrices into the DESC model and created our consensus matrix from the output.

For each tested cell proportion, the consensus matrix produced from GiniClust3 and our improved method were compared with the metrics: area under the curve (AUC), adjusted random index (ARI), normalized mutual information (NMI), and F1 Score. The F1 Score for both GiniClust3 and alternative model can be seen in Figure 5 at the tested cell proportion percentages. Furthermore, the calculated metrics at each rare cell proportion for both methods are located in Table 2. The UMAP for a rare cluster proportion of 5% for the improved Gini scheme with ground truth embedding shown in figure 3. The respective confusion matrix is also described in Figure 4.



(a)

(b)

Figure 3: UMAP embedding of cells from simulated dataset produced by Splatter. (a) UMAP with the true label, (b) UMAP with the label predicted by DESC-integrated clustering algorithm, both corresponding to a rare cell proportion of 5%.

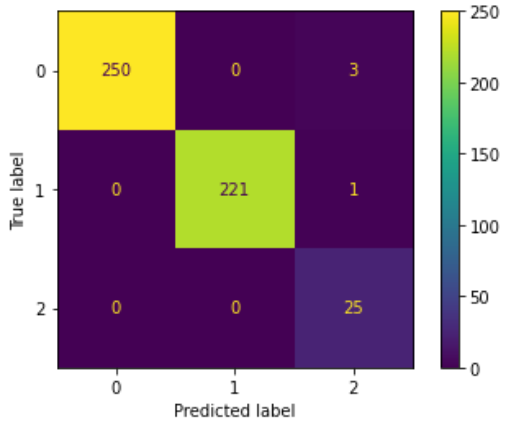


Figure 4: Confusion matrix for DESC-integrated clustering algorithm, corresponding to a rare cell proportion of 5%

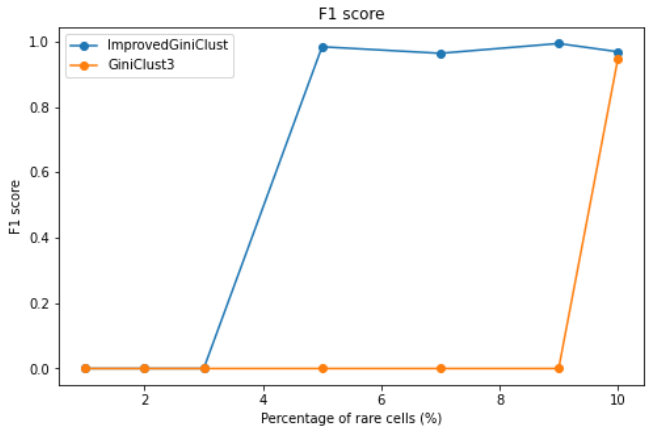


Figure 5: F1 scores were calculated with regard to the rare cell cluster, although both GiniClust3 and DESC-integrated clustering algorithm perform not well when the proportion of rare cells is very low, DESC-integrated clustering algorithm has lower threshold then GiniClust3 indicating higher sensitivity to rare cells than GiniClust3

Table (a)

Proportion of Rare Cell (%)	AUC	F1_Score	ARI	NMI
10	0.984	0.981	0.986	0.963
9	0.997	0.994	0.987	0.967
7	0.981	0.964	0.922	0.894
5	0.992	0.984	0.966	0.927
3	0.485	0	0.949	0.907
2	0.603	0	0.965	0.931
1	0.491	0	0.976	0.949

Table (b)

Proportion of Rare Cell (%)	AUC	F1_Score	ARI	NMI
10	0.267	0.946	0.898	0.878
9	0.772	=0	0.847	0.852
7	0.286	0	0.875	0.864
5	0.267	0	0.898	0.878
3	0.745	0	0.948	0.931
2	0.770	0	0.967	0.948
1	0.408	0	0.976	0.950

Table 2: Scores of several metrics for evaluating the performance of (a) DESC-integrated clustering algorithm, (b) GiniClust3

D. RESULTS

A. *Comparison of DL-integrated algorithm and GiniClust3 performance on 20k Brain cells from E18 Mice*

Performance of our proposed DESC method against GiniClust3 was performed on an un-annotated scRNA-seq dataset containing 20K Neurons, downsampled from 1.3 Million Brain Cells from E18 Mice, found on 10X Genomics' scRNA database [12]. The cells were taken from the cortex, hippocampus, and subventricular zone of two E18 mice.

The initial gene expression matrix consisted of 20,000 cells and 27,998 genes. Pre-processing was performed to filter cells with fewer than 200 genes present and filtering genes affecting less than 3 cells. The resulting expression matrix consisted of 19,793 cells and 17,968 genes which was then normalized to a UMI count of 10,000. All initial parameters considering GiniClust3 were left at default.

The Fano factor and Gini indices were then calculated using the normalized expression matrix and inserted into the DESC model as input. The model was parametrized to have a batch size of 300, k-neighbors of 10, pre-train epoch of 200, drop rate of 0.2, 1000 maximum iterations, alpha of 1, tolerance of 0.005, and a louvain resolution of 0.8 for Fano training and 0.4 for Gini. The model was implemented on random seed 4480 for reproducibility. After training, a consensus matrix of the DESC-Fano and DESC-Gini matrices was produced.

A UMAP of the resulting matrix was produced alongside a bar plot representing the percentage of cells present in each cluster, found in Figure 6. Clusters

representing < 2% of the overall cell population were identified as rare clusters, ranging from clusters 14 - 25. The top 3 genes in each cluster were determined using Wilcoxon Rank-Sum and a dot plot describing the proportion of cells expressing each clusters' top genes plotted based on a two-fold log change is shown in Figure 7. Furthemore, a dot plot of the top 4 genes in just the rare cell clusters was also plotted in Figure 8 with the same two-fold log change.

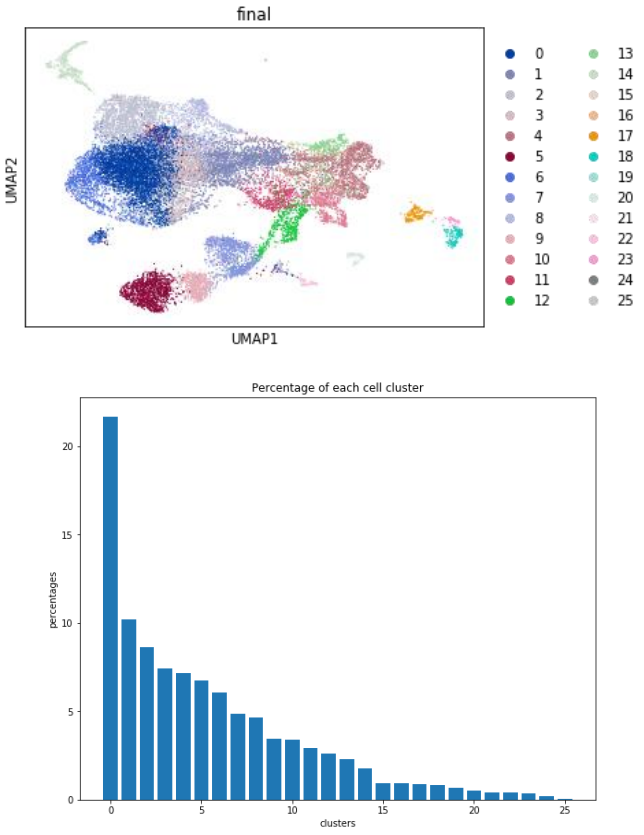


Figure 6: Representation of DESC-Integrated clustering algorithm on 20K Neurons downsampled from E18 Mice 10x Genomics dataset. (Top) UMAP projection of clusters and (bottom) cluster cell percentage of total cell population.

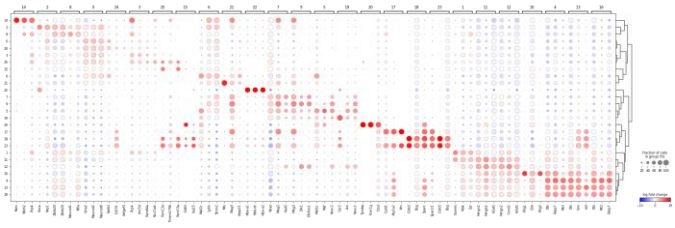


Figure 7: Dot Plot describing gene expression of top 3 genes represented in each cluster along all clusters. Gene expression was manipulated to two-fold log change.

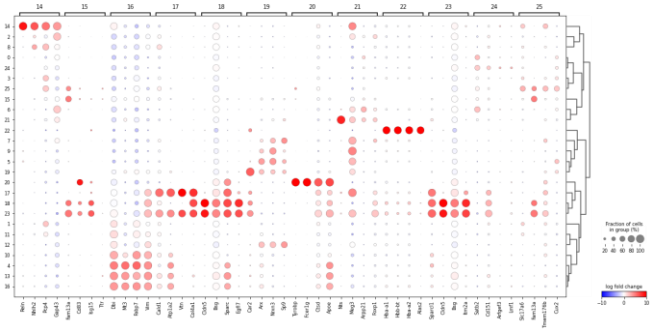


Figure 8: Dot Plot describing gene expression of top 4 genes represented in rare clusters along all other clusters. Gene expression was manipulated to two-fold log change.

E. Discussion

With the development of technology, it is gradually discovered that rare cells play an important role in determining the pathogenesis of cancer, mediating immune responses and other diseases, thus detecting rare cells could help us to have a new and deeper understanding of their role in biological functions and disease progression. Yet how to efficiently detect rare genes remains a pressing problem. Numerous methods have been created to characterize common cell populations, but lack the sensitivity to identify rare populations. Alternative methods designed specifically for rare cell cluster identification, such as RaceID and later GiniClust3, were proposed to alleviate the issues with traditional common clustering schemes, but fail to identify potential dependencies between gene-gene interactions due to their assumption on genes being independent of each other. Our DESC-integrated method aims to improve upon the existing GiniClust3 framework by utilizing a more sophisticated deep learning approach. Our intention is to improve the clustering step to allow for more sensitive identification of rare cell populations within scRNA-seq datasets.

In proving the validity of our method, we applied our DESC-integrated algorithm to an annotated scRNA-seq dataset and compared the area under the curve (AUC), adjusted random index (ARI), normalized mutual information (NMI), and F1 Score to GiniClust3, as shown in Table 2. At higher rare cell proportion populations, our algorithm (Table a) outperformed GiniClust3 (Table b) in all metrics.

The algorithm's performance was also tested on a 10X Genomics dataset containing 20K Neurons, downsampled from 1.3 Million Brain Cells from E18 Mice [12]. 25 clusters were generated with 12 identified rare cell clusters, as shown in Figure 6. The two-fold log change dot plot was shown (Figure 7) to illustrate unique major gene expression among each identified cluster, and further elaborated to rare cells clusters in Figure 8. The results indicate that there could potentially be rare clusters present, such as clusters 14, 22, 17, 18, and 23, while other rare clusters could either be overfitted and part of a major common

cluster, or simply just technical noise. More research is needed to further explore the presence of these identified rare clusters.

In the last half-decade, scRNA-seq data exploration has grown immensely. Despite this, the rate in which this data can be adequately analyzed is heavily lagging behind its accessibility, ease of collection, and utilization. Our DESC-integrated algorithm scheme may serve as a valuable tool for researchers to better understand heterogeneity in cellular systems and will hopefully help propel the research of scRNA-seq data into newer frontiers.

CODE AND DATA AVAILABILITY

The code of our project is available at: https://github.com/ChenghaoZhang97/BMENE4480_Statistical-machine-learning-for-genomic_Final_Project.git.

The dataset used in our project is publicly available. The dataset consisting 20K Neurons is available from: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons. The annotated scRNA-seq dataset used for creating synthetic dataset can be accessed on NCBI's Gene Expression Omnibus (GSE130710).

REFERENCES

[1] Kiselev, V., Kirschner, K., Schaub, M. et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 14, 483–486 (2017). https://doi.org/10.1038/nmeth.4236

[2] Giecold, G., Marco, E., Garcia, S. P., Trippa, L., & Yuan, G. C. (2016). Robust lineage reconstruction from high-dimensional single-cell data. Nucleic acids research, 44(14), e122. https://doi.org/10.1093/nar/gkw452

[3] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347:1138–42. https://doi.org/10.1126/science.aaa1934

[4] Kuo, Y. H., Lin, C. H., Shau, W. Y., Chen, T. J., Yang, S. H., Huang, S. M., Hsu, C., Lu, Y. S., & Cheng, A. L. (2012). Dynamics of circulating endothelial cells and endothelial progenitor cells in breast cancer patients receiving cytotoxic chemotherapy. BMC cancer, 12, 620. https://doi.org/10.1186/1471-2407-12-620

[5] Grün, D., Lyubimova, A., Kester, L. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255 (2015). https://doi.org/10.1038/nature14966

[6] Jindal, A., Gupta, P., Jayadeva et al. Discovery of rare cells from voluminous single cell expression data. Nat Commun 9, 4719 (2018). https://doi.org/10.1038/s41467-018-07234-6

[7] Grün, Dominic, et al. "De Novo Prediction of Stem Cell Identity Using Single-Cell Transcriptome Data." Cell Stem Cell, vol. 19, no. 2, 2016, pp. 266–277., https://doi.org/10.1016/j.stem.2016.05.010.

[8] Wegmann, R., Neri, M., Schuierer, S. et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol 20, 142 (2019). https://doi.org/10.1186/s13059-019-1739-7

[9] D. Tsoucas and G.-C. Yuan, "GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection," *Genome Biol.*, vol. 19, no. 1, p. 58, May 2018, doi: 10.1186/s13059-018-1431-3.

[10] L. Ceriani and P. Verme, "The origins of the Gini index: extracts from VariabilitA e MutabilitA (1912) by Corrado Gini," Journal of Economic Inequality - J ECON INEQUAL, vol. 10, pp. 1–23, Sep. 2012, doi: 10.1007/s10888-011-9188-x.

[11] Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 18, 174 (2017). https://doi.org/10.1186/s13059-017-1305-0

[12] Brain Cells from E18/C57BL/6 mice, Single Cell Gene Expression Dataset by Cell Ranger 1.3.0, 10x Genomics, (2017, February 9).

[13] Li, X., Wang, K., Lyu, Y. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun 11, 2338 (2020). https://doi.org/10.1038/s41467-020-15851-3

[14] Kaikun Xie, Yu Huang, Feng Zeng, Zehua Liu, Ting Chen, scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types, NAR Genomics and Bioinformatics, Volume 2, Issue 4, December 2020, lqaa082, https://doi.org/10.1093/nargab/lqaa082

[15] Bao S, Li K, Yan C, Zhang Z, Qu J, Zhou M. Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. Brief Bioinform. 2022 Jan 17;23(1):bbab473. doi: 10.1093/bib/bbab473. PMID: 34849562.

[16] Fa, B., Wei, T., Zhou, Y. et al. GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. Nat Commun 12, 4197 (2021). https://doi.org/10.1038/s41467-021-24489-8