Enhancing Rare Cell Type Identification in Single-Cell Data: An Innovative Gene Filtering Approach using Bipartite Cell-Gene Relation Graph

Maziyar Baranpouyan[1] Hossein Mohammadi[2]
Hojjat Torabi Goudarzi[3] Krishnaprasad Thirunarayan[2] and Lingwei Chen[2]

***Abstract*— A useful tool for examining cellular diversity is single cell RNA sequencing (scRNA-seq). However, the high dimensionality and technical noise of scRNA-seq data make analysis difficult. To address this issue, gene filtering has been widely adopted to minimize single cell data noise and enhance the quality of subsequent analyses. Nonetheless, existing gene filtering techniques may inadvertently omit critical but rare genes which are necessary for identifying rare cell types that play a pivotal role in comprehending many biological processes. A novel graph-based gene selection technique is suggested in this study with the aim of preserving the informative genes to better identify rare cell types. Our findings demonstrate that this technique enhances the identification of rare cell populations, providing a refined approach for scRNA-seq data analysis and potentially enabling earlier and more reliable disease detection.**

I. INTRODUCTION

Single cell sequencing has become an effective method for examining cellular heterogeneity and identifying cell types. The volume of data created has multiplied exponentially as single cell RNA sequencing (scRNA-seq) has become popular. However, analysing scRNA-seq data is complex and challenging due to its high-dimensional nature and the presence of technical noise [1]. Gene filtering is one of the most crucial steps in the processing of scRNA-seq data because single cell data has a high dropout rate and there are problems related to technical noise in this type of data as well [2]. Technical noises related to batch effects, sequencing depth, and amplification bias are also present. There is a chance that these variables will interfere with later analysis, which could result in inaccurate results and misinterpretations of biological occurrences. Therefore, gene filtering algorithms attempt to find and remove low quality genes in order to mitigate the impact of dropout events and technical noise [3].

We usually refer to rare cell types as those with low abundance. These cell types may have gene expression profiles that differ from other cell types in this sample. Despite their rarity in the sample, they can be important in some biological activities such as disease progression and immunological response. So discovering these sparse cell types contributes to a significantly better understanding of these biological processes [4]. Furthermore, these rare cell types may be present in low quantities in blood samples or other physiological fluids, and being able to detect them aids in the early detection of a related disease and allows for early intervention and therapy to manage or treat the disease.

In this respect, if we remove low expression genes during the gene filtering process confounding them with noise, we are likely to delete key genes that belong to these rare cell types, which can lead to misidentification or complete loss of them in downstream analysis. To summarize, we need to ensure that during gene filtering we eliminate ”noise” while preserving ”signal” even if the latter is not abundant, to preserve informative rare cell types. It is thus imperative to carefully design the gene filtering process to lower the risk of missing uncommon cell type identification and, to increase the precision of future analysis [2].

One of the most prevalent ways is to use thresholding to filter out noisy and uninformative genes. Currently, some well-known approaches are based on expression level [5], variance [6], and fold changes [7]. Also, researchers [2] have introduced a new method for automatically generating suitable thresholding curves for the input dataset.

Our study introduces a novel graph-based gene selection strategy tailored for single cell RNA sequencing (scRNA-seq) data analysis. Through rigorous comparative evaluations, we demonstrate its superior performance in dimension reduction and the identification of rare cell types, enhancing our understanding of cellular heterogeneity. Our approach's emphasis on preserving essential ”signal” while eliminating ”noise” during gene filtering ensures accurate identification and retention of informative rare cell types. This contribution advances the field by providing a more effective framework for scRNA-seq data analysis and biomedical research.

II. METHODS

Let's say we have gene expression values for G genes across C cells organized in a $G \times C$ expression matrix X. The next step involves creating a bipartite graph. This graph includes two types of entities, cells and genes, represented as vertices. We refer to this graph as the Cell-Gene Relation

[1]Maziyar Baranpouyan is with the Accenture Technology Labs, San Francisco, CA 94105 USA. maziyar.baran.pouyan@accenture.com

[2]Hossein Mohammadi, Krishnaprasad Thirunarayan and Lingwei Chen are with Department of Computer Science and Engineering, Wright State University, Fairborn, OH 45435 USA. mohammadi.5@wright.edu, t.k.prasad@wright.edu, lingwei.chen@wright.edu

[3]Hojjat Torabi Goudarzi is with Department of Electrical and Computer Engineering, The University of Oklahoma, 660 Parrington Oval, Norman, OK 73019 USA. tg.hojjat@ou.edu

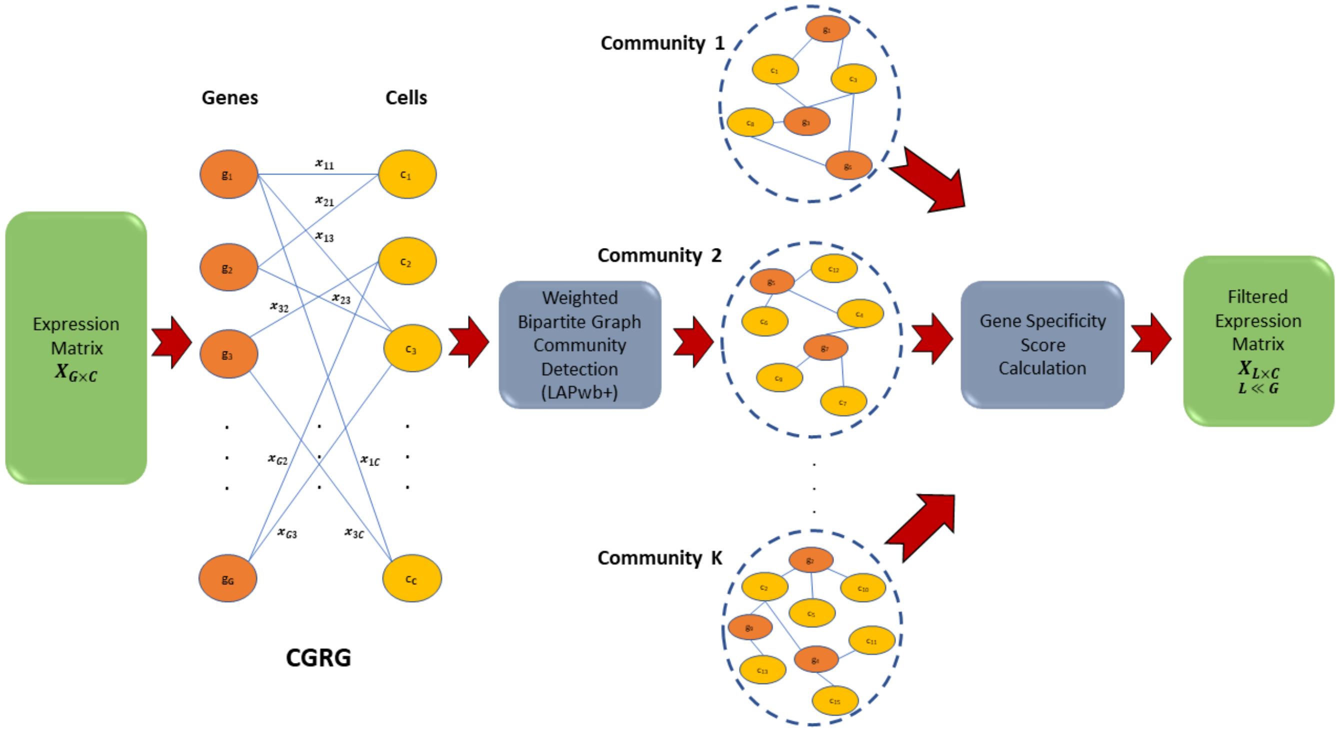

Genes
Cells
x_{11}
x_{21}
x_{13}
x_{32}
x_{23}
x_{G2}
x_{1C}
x_{G3}
x_{3C}
Expression Matrix $X_{G\times C}$
CGRG
Weighted Bipartite Graph Community Detection (LAPwb+)
Community 1
Community 2
Community K
Gene Specificity Score Calculation
Filtered Expression Matrix $X_{L\times C}$ $L \ll G$

Fig. 1: Illustrative diagram of the procedural steps of the model.

Graph (CGRG). Formally, CGRG can be represented as $CGRG = \{V, E\}$ where $V = \{v_1, v_2, \ldots, v_G, \ldots, v_{G+C}\}$ is the set of $G + C$ vertices and $E \subseteq V \times V$ denotes the set of edges connecting the vertices. Each vertex v_i indicates a cell or a gene in expression matrix X. Let v_i^g and v_j^c represent gene and cell vertices, respectively.

To establish connections in CGRG, we connect each gene vertex (v_i^g) to its top θ expressing cells, and each cell vertex (v_j^c) to its top θ highly expressed genes. The weight values of the connections are determined by the corresponding expression counts in X. This direct interaction between cells and genes in the CGRG is facilitated by these connections. This type of bipartite modeling not only minimizes the impact of noisy expressions but also highlights the role of crucial genes in identifying rare cell populations. The optimal θ value is pivotal; too large, and rare cell information is obscured by dominant cell relations, making rare cells hard to identify. Conversely, too low can lead to an excessive number of communities. In this study, θ is empirically chosen as 200 through experimentation.

Next, we try to identify the mutually exclusive communities of vertices within CGRG that exhibit as tendency to interact more frequently with one another. Since CGRC is a bipartite graph, its adjacency matrix A can be represented in a block diagonal format as follow:

$$A = \begin{bmatrix} 0_{G\times G} & X_{G\times C} \\ X_{C\times G}^T & 0_{C\times C} \end{bmatrix} \tag{1}$$

Using this representation, we can formulate weighted bipartite modularity as follow [8]:

$$Q = (\sum_{i=1}^{G}\sum_{j=1}^{C}\left(x_{ij} - \frac{k_i k_j}{2M}\right)\delta(h_i, h_j'))\frac{1}{2M} \tag{2}$$

TABLE I: Summary of datasets.

Dataset	# Cells	# Genes	# Cell type	# Rare cell type	Reference
CL1	4999	11499	8	1	[12]
CL2	3989	11499	8	4	[12]
Zeisel	3005	19972	9	3	[13]

In this equation, Q is the modularity score, M is the total number of edges in the CGRC, x_{ij} is the expression of gene i in cell j (i.e., the weight of the edge between gene i and cell j in CGRC), k_i and k_j are the sums of the weights of the edges connected to vertices i and j, respectively. h_i and h_j' are the communities to which gene i and cell j belong, respectively. Finally, $\delta(h_i, h_j')$ is the delta function returning 1 if gene i and cell j belong to the same community and 0 otherwise.

Weighted bipartite modularity (Q) measures how well a weighted bipartite graph can be partitioned into communities. These communities consist of vertices with strong connections within their community but relatively weak edges with vertices outside their community [8].

To find the partition that maximizes the modularity score (Eq. 2) for CGRG, we apply LAPwb+ algorithm [9] which has been successfully used in ecology domain [9]. Briefly, LAPwb+ is a process that involves two iterative steps. Firstly, it uses label propagation to update the labels of

TABLE II: Nearest neighbor error values for dimension reduction on all cell types (in percent, lower is better)

Method	FRQ (tSNE)	HiE (tSNE)	HVG (tSNE)	OGFSC (tSNE)	M3Drop (tSNE)	BCGGS (tSNE)	No filtering	FRQ (PCA)	HiE (PCA)	HVG (PCA)	OGFSC (PCA)	M3Drop (PCA)	BCGGS (PCA)
CL1	3.3	2.8	2.5	2.5	0.3	0	12.3	17.2	13.4	12.6	10.7	9.2	9
CL2	4.4	4.1	4.5	5.1	1.2	1	12.4	15.2	22.8	17.3	14.4	7.4	8.8
Zeisel	4.1	3.7	5.4	4.2	3.1	2.9	4.5	16.5	15.7	19.2	17.1	18.2	14.4

TABLE III: Nearest neighbor error values for dimension reduction on rare cell types (in percent, lower is better)

Method	FRQ (tSNE)	HiE (tSNE)	HVG (tSNE)	OGFSC (tSNE)	M3Drop (tSNE)	BCGGS (tSNE)	No filtering	FRQ (PCA)	HiE (PCA)	HVG (PCA)	OGFSC (PCA)	M3Drop (PCA)	BCGGS (PCA)	No filtering
CL1	1.1	1.3	1.3	1.5	0	0	3.8	5.2	5	6.1	4.6	2.9	2.5	9.2
CL2	1.8	2.1	1.5	1.5	0	0	6.2	7.7	9.2	5.8	6.6	2.2	2.6	13.5
Zeisel	14.2	15.7	9.1	9.6	6.1	4.9	28.5	27.3	20.6	23.7	21.5	11.3	10.2	34.7

TABLE IV: Clustering results using different methods on the CL1 dataset.

Method	**SC3**	**Seurat**	**pcaReduce**	**RAFSIL**	**DBSCAN**	**RaceID**	**CELLSIUS**
FRQ	76.2	92.4	64.8	95.1	95.7	73.5	85.3
HiE	81.1	**100**	64.8	97.4	**100**	78.5	**94.6**
HVG	83.9	96.4	72.2	**100**	**100**	80.2	90.5
OGFSC	74.6	93.7	70.2	**100**	94.6	76.9	80.3
M3Drop	85.1	**100**	78.2	**100**	**100**	83.7	88.2
BCGGS	**90.2**	**100**	**81.9**	**100**	**100**	**87.4**	92.6
No filtering	75.1	86.2	58.5	96.5	83.5	62.3	77.1

TABLE V: Clustering results using different methods on the CL2 dataset.

Method	**SC3**	**Seurat**	**pcaReduce**	**RAFSIL**	**DBSCAN**	**RaceID**	**CELLSIUS**
FRQ	50.2	77.4	63.2	86.4	95.3	61.8	65.4
HiE	54.7	79.1	68.8	90.2	**100**	59.8	70.5
HVG	**60.3**	82.5	64.5	92.3	90.2	**65.2**	70.5
OGFSC	44.7	80.6	**70.3**	84.4	88.6	62.4	71.6
M3Drop	56.5	79.2	47.3	90.2	98.4	62.5	71.4
BCGGS	**60.3**	**83.4**	67.5	**94.5**	99.2	**65.2**	**75.3**
No filtering	40.5	63.2	57.5	76.5	84.4	56.5	60.1

genes and cells in order to maximize Q locally. Secondly, it uses agglomeration to merge some communities together and prevent the algorithm from getting stuck in local maxima. These two steps are repeated until it is not possible to increase Q any further by merging communities. More details can be found in [9] and LAPwb+ R code is available in https://github.com/sjbeckett/weighted-modularity-LPAwbPLUS.

Once we extract the communities within CGRG, then we calculate the *weighted degree centrality* for each gene vertex in each community. Specifically, weighted degree centrality is the sum of edge weights for edges incident to the vertex. This would give greater importance to genes that are not just connected to many cells, but also have high expression levels as well. However, genes that are highly expressed in many cells across different communities might not be as informative as genes that are highly expressed only in a few communities. Hence, we create a *measure of specificity* by dividing the weighted degree centrality within a community by the weighted degree centrality across all communities [10]. After calculating these specificity gene scores, we rank all genes based on these scores and report the top 10% as our final selection. Note that CGRG is distinct from the KNN graph, which is typically constructed after gene filtering to identify cell clusters (e.g., phonograph method [11]), since CGRG incorporates both genes and cells, with interactions restricted solely to genes and cells and no direct connections between cells. This modelling approach can enhance the identification of rare cell types associated with specific genes, as they will have more interconnected zones centred around them. The name we have given to our proposed method is BCGGS, which stands for Bipartite Cell-Gene Graph Selection. An illustrative diagram of the procedural steps of the proposed model is shown in Figure 1.

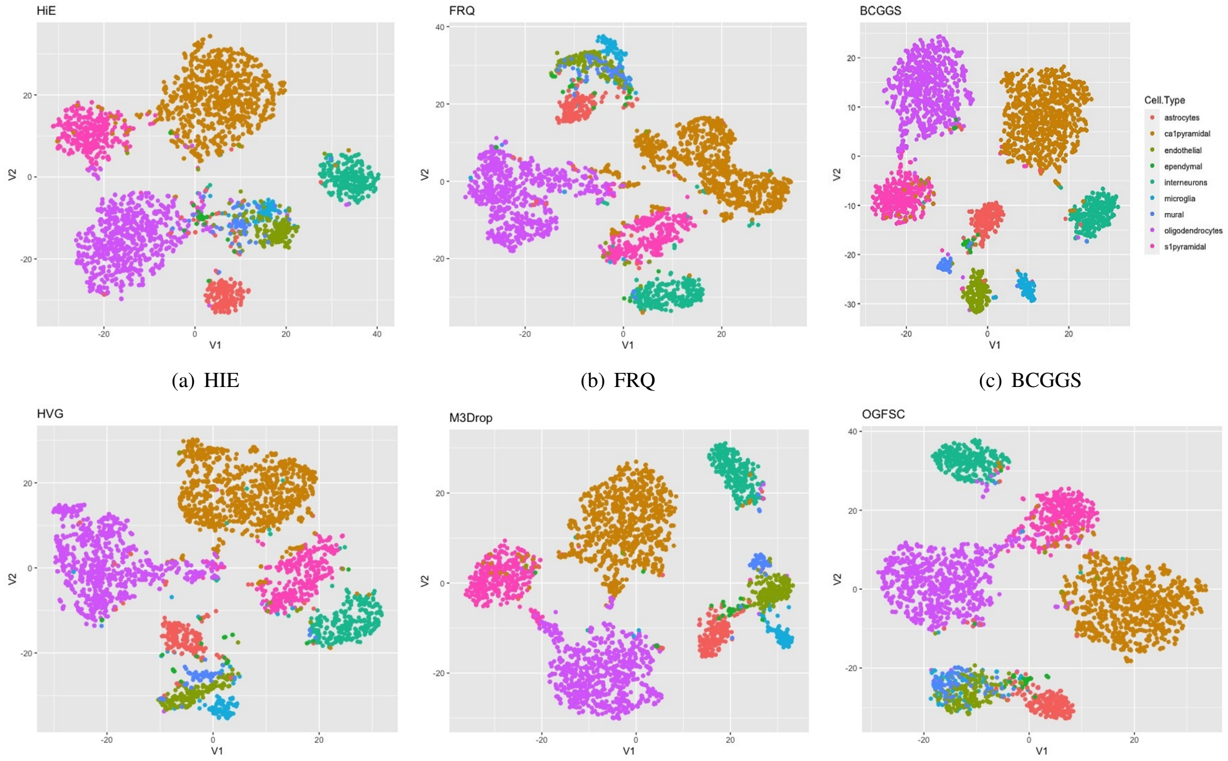
III. Experimental Results

For our experiments, we examine the performance of each gene filtering method using two alternative scenarios: dimension reduction for visualization, and unsupervised clustering to detect unusual cell groups. We have selected three benchmarks of rare cell types as illustrated in Table I.

In the first two benchmarks (CL1, CL2), there are eight human cell lines (A549, H1437, HCT116, HEK293, IMR90, Jurkat, K562, and Ramos). Among them, Jurkat accounts for 2% of CL1, while A549 (2.01%), H1437 (0.06%), Jurkat (0.15%), and K562 (1.76%) represent our rare cell populations in CL2 [12]. The third data (Zeisel) was obtained from the mouse cortex and hippocampus with 9 main cell types

TABLE VI: Clustering results using different methods on the Zeisel dataset.

Method	SC3	Seurat	pcaReduce	RAFSIL	DBSCAN	RaceID	CELLSIUS
FRQ	74.1	60.1	51	64.2	53.5	36.2	54.3
HiE	75.4	62.5	45.3	66.3	**60.4**	40.1	52.9
HVG	73.3	57.4	47.9	60.2	58.9	34.9	57.5
OGFSC	74.2	59.5	52.8	70.8	55.4	40.2	50.2
M3Drop	**79.1**	66.3	47.3	68.8	**60.4**	41.3	55.4
BCGGS	77.2	**72.2**	**57.2**	**74.2**	58.1	**50.2**	**62.3**
No filtering	74.8	60.2	40.2	60.2	43.4	38.4	47.6



(a) HIE

(b) FRQ

(c) BCGGS

(d) HVG

(e) M3Drop

(f) OGFSC

Fig. 2: Visualization results of Cell Embeddings in the Zeisel Dataset using tSNE with Gene Filtering Approaches, including our proposed method.

(interneurons, s1pyramidal, ca1pyramidal, oligodendrocytes, microglia, endothelial, astrocytes, ependymal, and mural) [13]. Among these cell types, microglia (0.03%), ependymal (0.008%), and mural (0.02%) are our rare cell populations. We compare BCGGS with five other gene filtering approaches as follows:

Frequency filtering (FRQ): In this strategy, we consider only genes that are expressed in a certain fraction of cells. We choose 6% as threshold [5].

Highly expressed genes (HiE): This method selects the top 10% of genes based on expression frequency [14].

Highly Variable Genes (HVG): It employs variance divided by mean as a selection metric [15].

Optimal Gene Filtering for Single-Cell Data (OGFSC): This method employs a regression-based gene filtering methodology to exclude genes with minimal biological relevance [2].

M3Drop: This method uses a negative binomial distribution model to identify genes with a high dropout rate, as these genes are more likely to be linked with technical noise and may affect downstream analysis [16].

A. Dimension Reduction

We apply each gene filtering method on benchmark datasets, and then utilize tSNE [17] and PCA [18] to reduce the dimensionality of the data. Next, we assesse the performance of the dimensionality reduction using the Nearest Neighbour Error metric (NNE) [14], which calculates the misclassification rate of the nearest neighbor classifier using Euclidean distances in the reduced-dimensional space (in 2D). As a comparison to our proposed approach, we also include a baseline approach that involves dimensionality reduction directly on the expression data. All results are reported in Tables II and III. Our findings show that BCGGS achieves the best overall performance in terms of decreased NNE for all datasets, and furthermore, it achieves better outcomes for rare cell types than using M3Drop method. In addition, we use tSNE to show the cell embeddings in the Zeisel dataset while adding several gene filtering approaches. Figure 2 depicts the findings through visualization. The results suggest that BCGGS and M3Drop are superior to other methods in terms of providing better visualization for all cell types. However, BCGGS is the only method that achieves better separation for rare cell types in this dataset.

B. Cell Clustering

We investigate the performance of cell population identification for seven unsupervised clustering algorithms in our benchmarks: SC3 [5], Seurat [19], pcaReduce [20], RAFSIL [14], DBSCAN [21], RaceID [22], and CELLSIUS [12]. Note that RaceID, and CELLSIUS are two methods specifically designed to identify rare cell types. Briefly, both methods combine a global clustering with a second assessment technique that is designed to detect rare cell population. Once we have applied gene filtering to each dataset, we utilize PCA to convert the original data into vectors in a lower-dimensional space (50 PCs), and then proceed to perform clustering.

Finally, we calculate the adjusted Rand index (ARI) [23], which assesses the agreement between the assigned labels and the genuine labels, to formalize clustering quality. It is important to consider that certain methods, such as SC3 and RAFSIL, are subject to randomness. In order to account for this, we run each algorithm 20 times and calculate the median value of the Adjusted Rand Index (ARI), which is reported in Tables IV, V, and VI for datasets CL1, CL2, and Zeisel respectively. Our results show that, once again, the GHVG and M3Drop techniques lead to improvements in the performance of most clustering methods. In addition, we observe that, in average, HiE method degrades the clustering performance in identification of rare cell types, as this method tends to remove low expressed genes, which may remove genes that are unique to rare cell population. Additionally, we discovere that while RaceID and CELLSIUS exhibit superior performance in identifying rare cell types compared to other methods, these approaches (especially RaceID) tend to partition large cell populations and are not very effective in extracting big clusters, as cautioned in a previous study [24].

IV. Conclusion

In this research, we propose a novel computational method that utilizes a cell-gene association graph to identify the most informative genes. Our findings demonstrate that the suggested filtering strategy can improve single cell downstream analysis, such as cell type identification and visualization. However, we observe that the majority of examined clustering techniques perform well in finding populations characterized by more than 2% of all cells, but struggle to identify less prevalent cell populations, emphasizing the need for creating specialized tools aimed at improving the detection of rare cell populations.

References

[1] Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, Pinello L. Eleven grand challenges in single-cell data science. Genome biology. 2020 Dec;21(1):1-35.

[2] Hao J, Cao W, Huang J, Zou X, Han ZG. Optimal Gene Filtering for Single-Cell data (OGFSC)—a gene filtering algorithm for single-cell RNA-seq data. Bioinformatics. 2019 Aug 1;35(15):2602-9.

[3] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology. 2019 Jun;15(6):e8746.

[4] Jindal A, Gupta P, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nature communications. 2018 Nov 9;9(1):4719.

[5] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. SC3: consensus clustering of single-cell RNA-seq data. Nature methods. 2017 May;14(5):483-6.

[6] Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV, Linnarsson S. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nature neuroscience. 2015 Jan;18(1):145-53.

[7] Martinez-Jimenez CP, Eling N, Chen HC, Vallejos CA, Kolodziejczyk AA, Connor F, Stojic L, Rayner TF, Stubbington MJ, Teichmann SA, de La Roche M. Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science. 2017 Mar 31;355(6332):1433-6.

[8] Dormann CF, Strauss R. A method for detecting modules in quantitative bipartite networks. Methods in Ecology and Evolution. 2014 Jan;5(1):90-8.

[9] Beckett SJ. Improved community detection in weighted bipartite networks. Royal Society open science. 2016 Jan 20;3(1):140536.

[10] Gross JL, Yellen J. Graph theory and its applications. CRC press; 2005 Sep 22.

[11] Levine JH, Simonds EF, Bendall SC, Davis KL, El-ad DA, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015 Jul 2;162(1):184-97.

[12] Wegmann R, Neri M, Schuierer S, Bilican B, Hartkopf H, Nigsch F, Mapa F, Waldt A, Cuttat R, Salick MR, Raymond J. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome biology. 2019 Dec;20(1):1-21.

[13] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138-42.

[14] Pouyan MB, Kostka D. Random forest based similarity learning for single cell RNA sequencing data. Bioinformatics. 2018 Jul 1;34(13):i79-88.

[15] Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Briefings in bioinformatics. 2019 Jul;20(4):1583-9.

[16] Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics. 2019 Aug 15;35(16):2865-7.

[17] Van der Maaten L, Hinton G. Visualizing data using tSNE. Journal of machine learning research. 2008 Nov 1;9(11).

[18] Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010 Jul;2(4):433-59.

[19] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015 May 21;161(5):1202-14.

[20] Žurauskienė J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC bioinformatics. 2016 Dec;17:1-1.

[21] Hahsler M, Piekenbrock M, Doran D. dbscan: Fast density-based clustering with R. Journal of Statistical Software. 2019 Oct 31;91:1-30.

[22] Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, Van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015 Sep 10;525(7568):251-5.

[23] Hubert L, Arabie P. Comparing partitions. Journal of classification. 1985 Dec;2:193-218.

[24] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics. 2019 May;20(5):273-82.