EDGE: Ensemble Dimensionality Reduction and Feature Gene Extraction for Single-cell RNA-seq Data

Xiaoxiao Sun*[1], Yiwen Liu*[2], and Lingling An†[1,3]

[1]Department of Epidemiology and Biostatistics, University of Arizona

[2]Department of Mathematics, University of Arizona

[3]Department of Biosystems Engineering, University of Arizona

Abstract

Single-cell RNA sequencing (scRNA-seq) technologies allow researchers to uncover biological states of a single cell in a high resolution. For computational efficiency and easy visualization, dimensionality reduction is needed to capture gene expression patterns in a low dimensional space. In this paper, we propose an ensemble method for simultaneous dimensionality reduction and feature gene extraction (EDGE) of scRNA-seq data. Different from existing dimensionality reduction techniques, the proposed method implements an ensemble learning scheme that utilizes massive weak learners for an accurate similarity search. Based on the similarity matrix constructed by those weak learners, the low-dimensional embeddings of the data are estimated and optimized through the spectral embedding and stochastic gradient descent, respectively. Comprehensive simulation and empirical studies show that EDGE is well suited for searching for the meaningful organization of cells, detecting rare cell types, and identifying essential feature genes associated with certain cell types.

Keywords: single-cell RNA-seq; dimensionality reduction; ensemble method; feature gene selection; rare cell types

*Equal contribution

†To whom correspondence should be addressed

Background

The advent of massive single-cell transcriptomic data provides unprecedented opportunities to study cellular heterogeneity within complex tissues (Tang et al., 2009; Patel et al., 2014; Wagner et al., 2016). A crucial component for scRNA-seq data analysis is dimensionality reduction of the large-scale and feature-rich datasets (Pierson and Yau, 2015; Ding et al., 2018). The dimensionality reduction methods consist of two major types, linear and non-linear techniques. The former one, such as principal component analysis (PCA), has the issue of overcrowding representation for the scRNA-seq data (Tran et al., 2019; Kobak and Berens, 2019). Therefore, nonlinear dimensionality reduction methods such as t-distributed stochastic neighborhood embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are widely used in the scRNA-seq data analysis (Tenenbaum et al., 2000; Coifman et al., 2005; Maaten and Hinton, 2008; Linderman et al., 2019; Becht et al., 2019). The t-SNE method is powerful in preserving the local structure of the data. However, it may suffer from losing the global geometry pattern/property. When the hierarchical structure presents in the data, t-SNE may not be able to capture such a global structure (Kobak and Berens, 2019). The UMAP method has been developed to address those issues (Becht et al., 2019). It preserves both the local and global structures of cell populations and is more efficient in computation, compared with t-SNE. However, UMAP may be inefficient in identifying rare cell types when dominant/common cell types/populations exist. Our simulated and empirical studies demonstrate that UMAP is less efficient in separating rare cell types from dominant ones and preserving locally hierarchical structures. More important, these techniques are not able to detect feature (i.e., differentially expressed) genes that are associated with various cell types while performing dimensionality reduction.

In this paper, we propose an Ensemble Dimensionality reduction and feature Gene Extraction (EDGE) method to perform dimensionality reduction and feature gene identification simultaneously. The ensemble technique allows massive weak learners to vote for cell similarities, while the genes that make great contributions are selected as feature genes. A series of comprehensive simulation and empirical studies demonstrated the effectiveness of EDGE in

identifying rare cell types, preserving local and global structures, and detecting vital feature genes.

Results

Overview of EDGE

To conduct dimensionality reduction, it is pivotal to construct the similarity/distance matrix among cells in an accurate way. Similarity search for massive amounts of feature-rich data emerges as a challenging problem. Several methods have been proposed to perform the similarity search (Guttman, 1984; Krauthgamer and Lee, 2004; Liu et al., 2005). Nevertheless, similarity search in the scRNA-seq data, by its nature, is complicated. Firstly, traditional methods only work well when the dimensionality is relatively low (Gionis et al., 1999; Beygelzimer et al., 2006) thus are not applicable to feature-rich scRNA-seq data. Secondly, due to the dropout effects, it is even more challenging to obtain accurate similarity measures (Kim et al., 2018). Thirdly, the existence of rare cell groups and/or subgroups adds another layer of complications. Due to the data heterogeneity, rare cell groups often get overlooked, while cell subtypes are usually masked or nested within upper-level cell types (Jindal et al., 2018).

To address the above concerns, we propose to use EDGE to learn the similarity among cells. The proposed approach is motivated by a similarity search method, sketches (Lv et al., 2006; Wang et al., 2007). Jindal *et al.* proposed the finder of rare entities (FiRE) algorithm based on the sketches (Jindal et al., 2018). It successfully identified rare cell types in a matter of seconds. Although motivated by sketches, EDGE is entirely different from FiRE in three aspects. Firstly, EDGE starts by generating a number of weak learners to vote for the similarity among cells in an ensemble way. Specifically, a weak learner is formulated by randomly picking up a small group of genes and then constructing a sketch (bit vectors) on those feature vectors. Regardless of the dropout events, those bit vectors are capable of capturing the information contained in nonzero values with higher probability. The weak

learner then votes for the cells if they are assigned to the same hash code (box). With these weak learners, cells of the same type have a higher similarity score. Secondly, EDGE uses the similarity measures to learn and optimize the low-dimensional embedding through minimizing a cross-entropy function by stochastic gradient descent (Bottou, 2010; Becht et al., 2019). Lastly, the proposed method can detect feature genes that contribute to the separation of various cell types. The detection of feature genes relies on trained weak learners in the first step. To this end, EDGE uses the ensemble learning to construct low-dimensional embedding and perform feature gene selection simultaneously. A visual illustration of the EDGE method is provided in Figure 1, and an elaborate explanation of the EDGE method is presented in the Methods section.

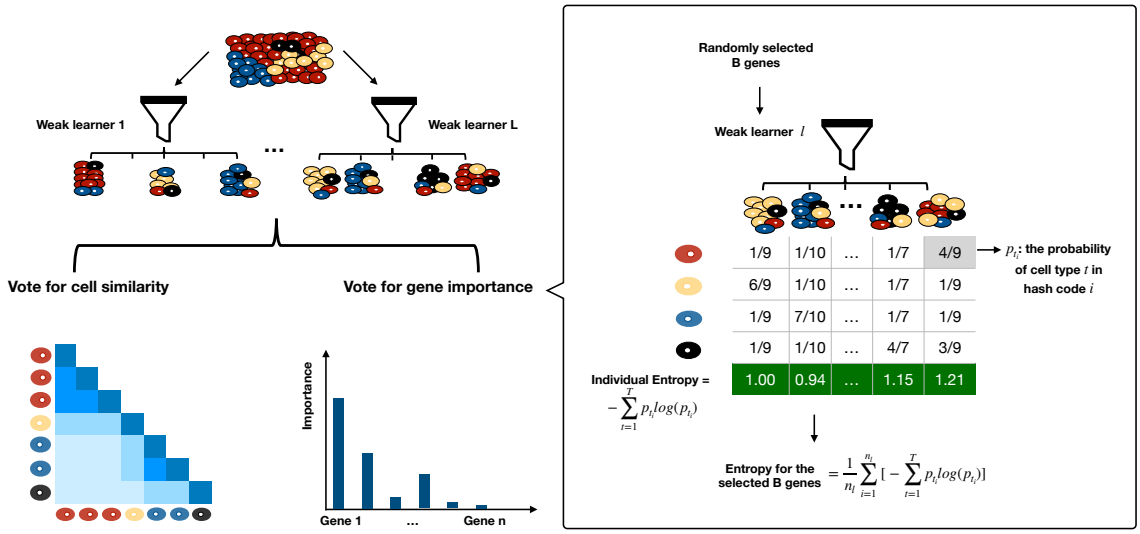

Figure 1: Overview of EDGE. The procedure starts by generating a number of weak learners. Each weak learner constructs a few of hash codes (imaginary boxes, i.e., piles of cells) using randomly selected a certain number of genes, e.g., B genes. The information of these weak learners is utilized to vote for cell similarities, which will then be used in the embedding estimation and optimization. The importance score of these B genes for each weak learner is obtained through the calculation of averaged entropy across the hash codes. When the weak learner well discerns different cell types, the corresponding averaged entropy value will be small. Details of the procedure can be found in Methods.

Benchmarking EDGE in simulated studies

We designed simulated experiments to evaluate the performance of EDGE in embedding the synthetic scRNA-seq count data with the presence of rare cell types and at various rates of dropout events. The proposed method was compared with t-SNE and UMAP, which were widely-used dimensionality reduction methods in scRNA-seq data analysis (Amir et al., 2013; Grün et al., 2015; Vento-Tormo et al., 2018). We generated the simulated data using Splatter under four scenarios (Methods) (Zappia et al., 2017). **Scenario 1.** One thousand cells in four groups, with the ratio of 10:10:10:70 in group size (number of cells) and a low dropout rate in gene expression of 1,500 genes; **Scenario 2.** Four equal-sized groups of cells, i.e., each with a proportion of 25% of the total 1,000 cells, and gene expression of 1,500 genes at a high dropout rate; **Scenario 3.** Four groups of 1,000 cells, with the ratio of 10:10:10:70 in group size and a high dropout rate in gene expression of 1,500 genes; **Scenario 4.** Four equal-sized groups of 1,000 cells with gene expression of 1,500 genes at a low dropout rate. Deferentially expressed (DE) gene proportions for Scenario 1 to 3 were fixed at 35%, while varied from 10% to 25% for Scenario 4. The details of the settings are shown in Table 1 and Methods section.

Table 1: Settings in simulation studies. The standard deviation of the observed zero proportions in simulated data over 100 replications is shown in the parentheses. The observed zero proportions for the high and low dropout rates are around 0.70 and 0.87, respectively.

	Group Proportions (%)	Zero Proportion	DE Gene Proportion (%)
Scenario 1	(10, 10, 10, 70)	0.6993 (0.0229)	35
Scenario 2	(25, 25, 25, 25)	0.8682 (0.0169)	35
Scenario 3	(10, 10, 10, 70)	0.8688 (0.0165)	35
Scenario 4	(25, 25, 25, 25)	0.6961 (0.0233)	10
		0.6965 (0.0234)	15
		0.6968 (0.0231)	20
		0.6974 (0.0233)	25

To investigate the abilities of EDGE to preserve the structure of cell populations, we trained random forests to predict cell clusters' identities using the embeddings generated by the aforementioned methods for comparison (Breiman, 2001). The within-group and overall accuracy of the predictions were measured through out-of-bag (OOB) prediction errors over 100 simulation replicates. When the dropout rates were around 0.7 in the presence of rare cell types (Figure 2, a), EDGE led to higher prediction accuracy, especially for rare cell types. For the case of equal group size, even with high zero proportion (close to 0.9), EDGE appeared more efficient than the other two methods at separating cell types (Figure 2, b). When there existed rare cell types, coupled with the presence of high dropout rates (Figure 2, c), all methods became less efficient, but EDGE still ranked first in terms of prediction accuracy. We also evaluated how reliable the embedding was when the proportions of DE genes varied. In Scenario 4, EDGE and UMAP achieved higher overall prediction accuracy, followed by t-SNE (Figure 2, d). In summary, the results suggest that the embedding generated by EDGE accurately characterizes all cell groups and is effective in separating rare cell types in different scenarios.

We also evaluated the performance of EDGE for detecting the feature (DE) genes in simulated studies. Since the t-SNE and UMAP methods could not identify feature genes, we only presented the performance of EDGE in identifying true feature genes in the simulation studies. Due to the multi-layer data generation structure, it is difficult to obtain the true feature genes using the original data generation procedure in Splatter (Zappia et al., 2017). We modified the procedure (Methods) and simulated scRNA-seq datasets in four scenarios. In the first two scenarios, two types of $1,000$ cells with proportions of about 80% and 20% were generated. Among the total of 500 genes, 30 of them were set as true feature genes in each two-group scenario. We repeated the simulation 100 times and reported results in Table 2. The means of zero proportions over 100 replications for the high and low dropout scenarios were 78.83% and 50.55%. The proposed algorithm has a high accuracy rate in detecting true feature genes. In the low dropout scenario, when the top 15 genes were chosen based on the importance scores, all of them were the true feature genes; and when

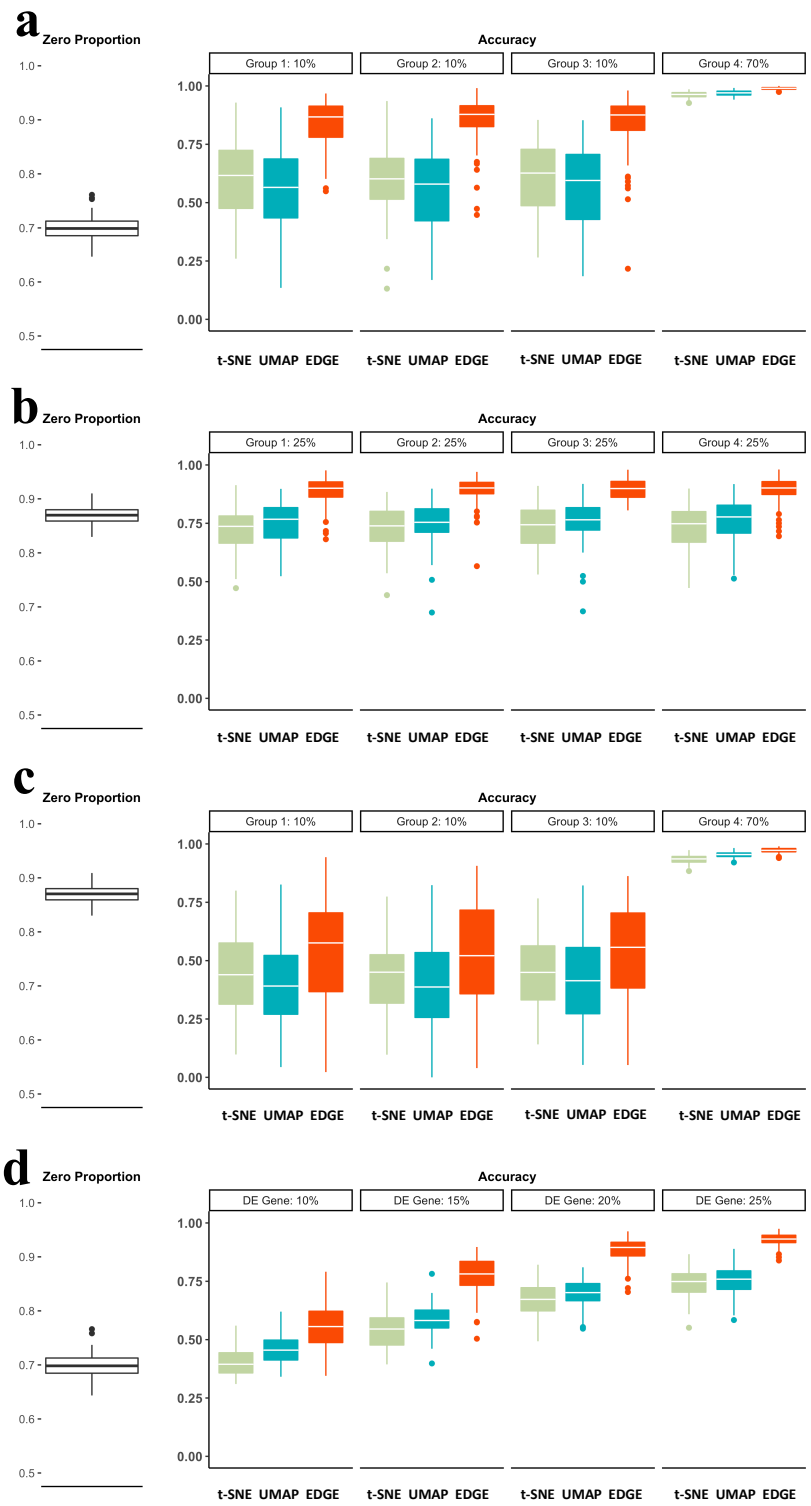

Figure 2: The accuracy of random forests classifiers in predicting cell cluster labels using the learnt embeddings as input in Scenario 1 (a), 2 (b), 3 (c), and 4(d).The left plot in each panel shows observed zero proportions in the corresponding datasets.

we selected top 30 genes, 27.06 on average were. In the high dropout scenario, although the signal-to-noise ratio was decreased, we still detected 14.79 and 25.74 true feature genes on average. The simulation results for three cell types are shown in Table S1 in Supplementary Information.

Table 2: The performance of EDGE in detecting the feature genes in simulated studies with two cell types. The standard deviation of the number of identified true feature genes is shown in the parentheses.

	Top 15 Genes	Top 30 Genes
Low Dropout	15.00 (0.00)	27.06 (1.38)
High Dropout	14.79 (1.46)	25.74 (3.03)

EDGE is accurate in embedding rare cell types

To investigate the performance of EDGE in handling rare cell populations, we applied EDGE alongside t-SNE and UMAP to two scRNA-seq datasets (the Jurkat dataset and the PBMC dataset). The Jurkat dataset contains two cell types, 293T and Jurkat cells (Methods), and the proportion of Jurkat cells is around 2.5% (Jindal et al., 2018; Zheng et al., 2017). The PBMC dataset contains nine cell types, among which four are rare, with the concentration varying from 0.89% to 2.2% (Methods). For each dataset, we projected cells into a two-dimensional space based on the embeddings generated by the three methods. Figure 3 demonstrates the performances of those methods in separating rare cell types from the other cell types. EDGE captured much information in the Jurkat dataset and outperformed t-SNE and UMAP (Figure 3, a, b, and c). It successfully separated all Jurkat cells (the rare cell type, in red) from 293T cells (the dominant type, in blue) (Figure 3, a). However, t-SNE and UMAP mapped some 293T cells to the cluster of rare cells (Figure 3, b and c). In analyzing the PBMC dataset, the four rare cell types were separated into more compact clusters with clearer patterns by EDGE than by t-SNE and UMAP. For instance, dendritic and megakaryocyte cells formed distinct clusters based on the low-dimensional representation of EDGE (in yellow and brown, respectively, Figure 3, d). Nonetheless, these cells were close to other cell types in the low-dimensional subspace generated by t-SNE and UMAP

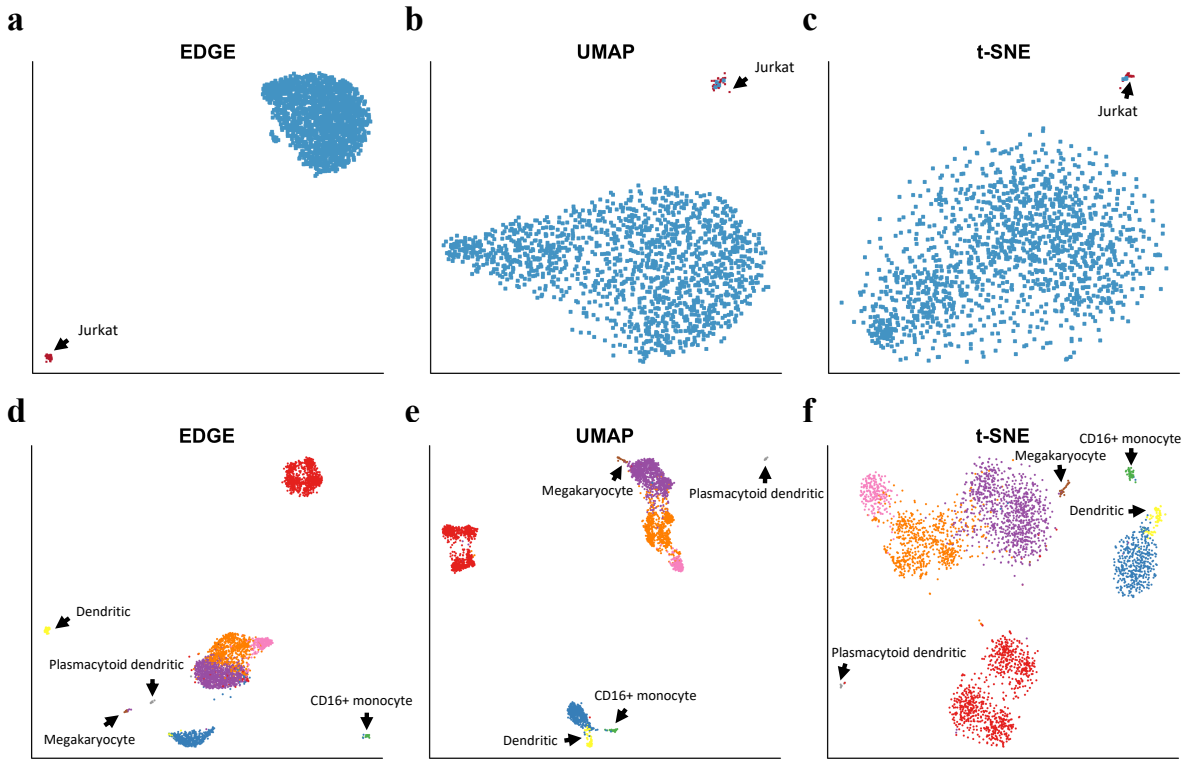

Figure 3: Detection of rare cell types. **a-c**: The two-dimensional embeddings learnt by EDGE, UMAP, and t-SNE respectively on the Jurkat dataset. **d-f**: The two-dimensional embeddings learnt by EDGE, UMAP, and t-SNE respectively on the PBMC dataset. Rare cell types are annotated by black arrows.

(Figure 3, e and f). In addition, plasmacytoid dendritic cells (in gray) formed a distinct cluster in all low-dimensional subspaces. However, some B cells (in red) were mapped to the corresponding regions in the two-dimensional scatter plots of t-SNE (Figure 3, f). We also took the embeddings estimated by the methods for comparison as the input of a shared nearest neighbor modularity based clustering algorithm to predict cell types (Waltman and Van Eck, 2013; Stuart et al., 2019). For a fair comparison, the same parameters were used in the clustering algorithm. Based on the predicted cell labels, the prediction accuracy of EDGE for the four rare cell types was 92.02%, whereas t-SNE and UMAP achieved the prediction accuracy of 88.83% and 88.30%, respectively. Taken together, these analyses demonstrate that EDGE is an accurate method in separating rare cell types compared to t-SNE and UMAP.

EDGE better preserves global and local structures

We applied EDGE, t-SNE, and UMAP to the mouse brain dataset (Zeisel et al., 2015) and illustrated that EDGE better preserved global and local structures. The dataset consists of

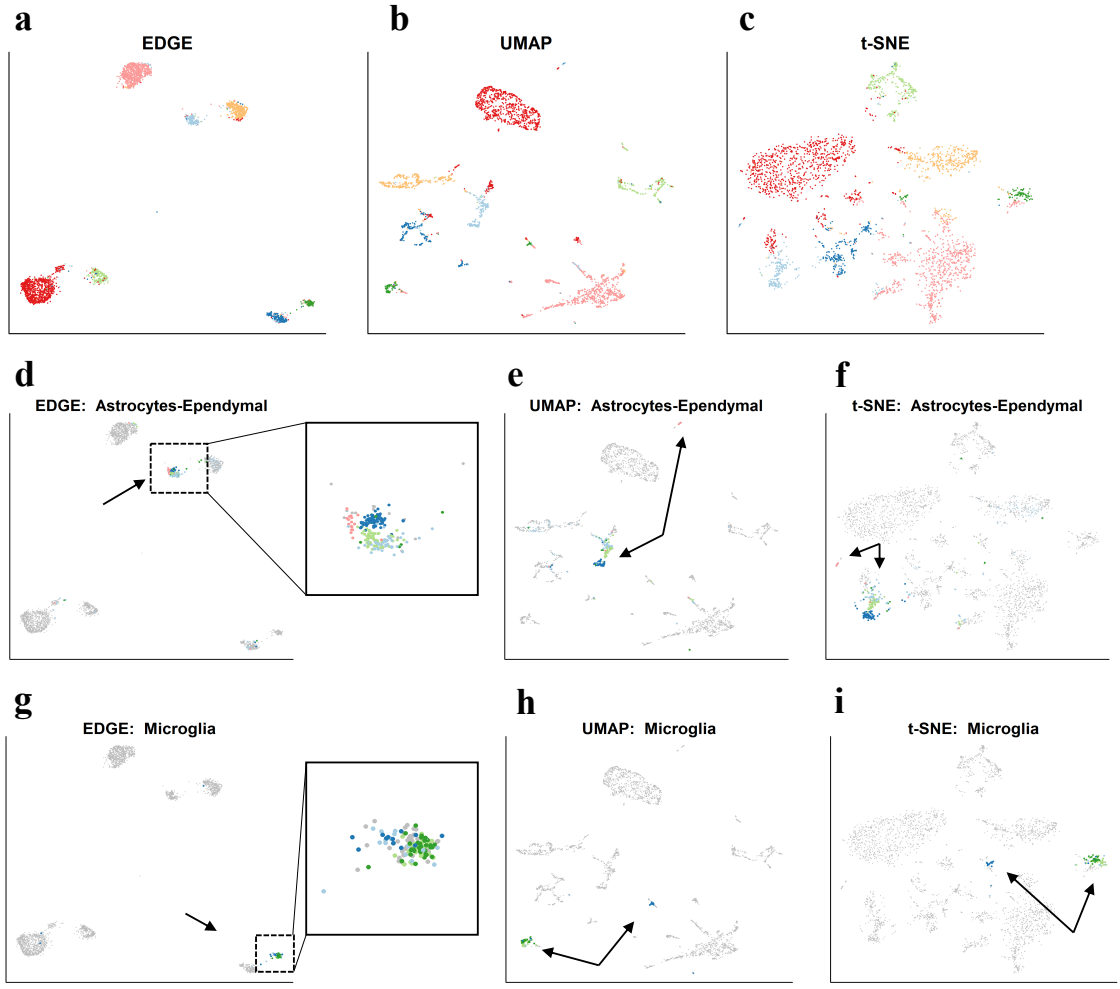

Figure 4: Local and global structure can be preserved by EDGE in the mouse brain dataset. **a-c**: The two dimensional representations learnt by EDGE, UMAP, and t-SNE respectively. **d-f**: same as a-c, but only the astrocytes-ependymal cells are highlighted with their subtype status . **g-i**: same as a-c, but only the microglia cells are highlighted with their subtype status.

3,005 cells in the mouse cortex and hippocampus (Methods) (Zeisel et al., 2015). We first fitted a 30-dimensional ellipsoid to the dataset using PCA. Then cells in the 30-dimensional subspace were projected onto a two-dimensional space generated by the embeddings from

the three methods. As demonstrated in Figure 4, seven types of cells identified in the original study were mapped to seven compact regions in the EDGE-based two-dimensional subspace (Figure 4, a). Compared to EDGE, some clusters were scattered in the two-dimensional subspace estimated by t-SNE and UMAP (Figure 4, b and c). We also applied the shared nearest neighbor modularity based clustering algorithm to the estimated embeddings by three methods. The overall prediction accuracy of all cell types was 93.31% for EDGE, while 79.83% for UMAP, and 82.43% for t-SNE. In addition to preserving the global structure, EDGE shows promising in maintaining hierarchical structures. It maps cell subtypes within the same cell type together in the low-dimensional space. For instance, five subtypes of astrocytes-ependymal cells were present within a compact region annotated by the black arrow (Figure 4, d). However, in the resulting plots of UMAP and t-SNE, one subtype (annotated by the black arrow) was mapped to a region far away from the region of other subtypes (Figure 4, e and f). Similarly, four subtypes of microglia cells were also projected to a clearly distinguishable region in the EDGE-based two-dimensional subspace (Figure 4, g), whereas those four cell subtypes were mapped to different regions in the UMAP-based and t-SNE-based subspaces (Figure 4, h and i).

EDGE is capable of detecting feature genes

EDGE is able to identify feature genes that are responsible for the separation of different cell populations. It ranks all candidate genes according to their contribution to predict cell identities, which is referred to as the importance score. Then the number of feature genes is determined by the distribution of those importance scores (Methods). We identified 17, 35, and 43 feature genes for the Jurkat, PBMC, and mouse brain datasets (Table S2). In the 35 detected genes in PBMC dataset, *PPBP*, *LYZ*, *CST3*, and *NKG7* were the marker genes for platelet, CD14+ monocyte, dendritic cells, and natural killer cells, respectively (Figure 5) (Stuart et al., 2019). Feature genes detected by EDGE were classified into two types. For the first type, genes such as *ACRBP* and *IGLC3* were solely expressed in a specific cell type. Such genes could be detected using standard methods, e.g., fold change (Wang et al., 2019).

Table 3: Ten most enriched GO biological processes for the PBMC scRNA-seq dataset.

GO Biological Process	Fold Enrichment	Raw P-value	FDR [1]
defense response GO:0006952	9.22	4.02E-15	3.20E-11
leukocyte mediated immunity GO:0002443	13.81	3.52E-15	5.60E-11
immune effector process GO:0002252	10.33	3.74E-14	1.99E-10
immune response GO:0006955	6.67	1.30E-12	3.45E-09
response to external biotic stimulus GO:0043207	8.37	1.09E-12	3.47E-09
response to biotic stimulus GO:0009607	8.17	1.59E-12	3.62E-09
response to other organism GO:0051707	8.38	1.06E-12	4.24E-09
defense response to other organism GO:0098542	10.22	2.50E-12	4.97E-09
humoral immune response GO:0006959	19.4	5.81E-12	1.03E-08
regulated exocytosis GO:0045055	12.27	1.23E-11	1.96E-08

[1]False Discovery Rate.

Genes of the second type separated different cell types based on their unique distribution patterns of gene expression values in some cell types. For instance, the most important gene *S100A9* (leftmost gene in Figure 5) was highly expressed in CD14+ monocyte, CD16+ monocyte, and dendritic cells. While this gene distinguished these three cell types from the remaining, the unique distribution patterns of expression levels in these three cell types (violin shapes in Figure 5) were beneficial to further differentiate three of them. These two types of genes were also found in Figures S5 in Supplementary Information, for example, *MPB* and *ARAP3*.

Furthermore, we performed gene ontology (GO) enrichment analysis for the 35 detected genes in PBMC dataset (Ashburner et al., 2000; Mi et al., 2018) and showed ten most enriched GO biological processes in Table 3. All ten enriched biological processes were

related to immune response and response to stimulus. Since PBMC cells such as B cells and T cells initiated or got involved in immune responses, the enriched biological processes were highly correlated with the biological functions of PBMC cells (Cooper, 2015).

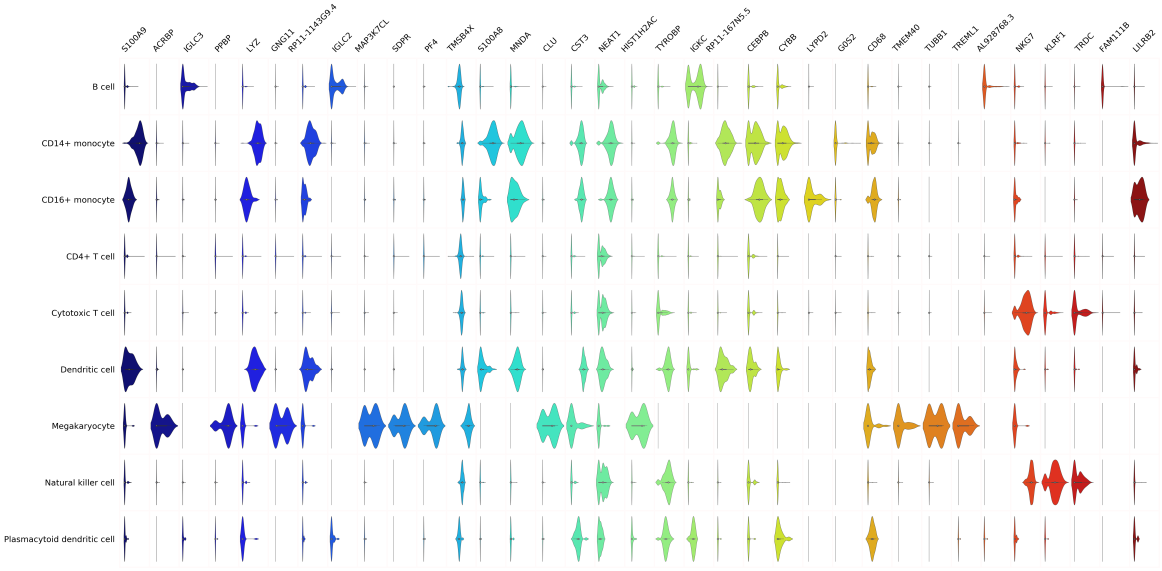

Figure 5: Normalized expression levels of 35 top-ranked feature genes detected by EDGE for PBMC dataset. Genes are ordered by their importance scores with *S100A9* (the most left on top) having the highest importance score.

Discussion

We have developed an ensemble method, EDGE, for simultaneous dimensionality reduction and feature gene extraction in scRNA-seq data. It exploits a large number of weak learners to study the similarities among cells. Those massive weak learners not only vote for cell similarities but also detect feature genes that contribute the most in the voting. Furthermore, those learners are constructed through sketches, and EDGE thus is sensitive to rare cell types. There are three major contributions of this study.

- EDGE provides an accurate way for similarity search. In all three datasets, EDGE showed more compactly projected structures compared with t-SNE and UMAP (Figure 3 and 4, a-c).

- EDGE preserves both local and global structures in scRNA-seq data. When there existed rare cell types, EDGE provided a neat separation between rare and common cell types (Figure 3, a and d). For cells that belong to different subtypes, EDGE maintains the hierarchical structures of cell types successfully.

- Compared to popular embedding methods such as t-SNE and UMAP, EDGE can effectively detect feature genes that are crucial for discerning different cell types.

The way we vote for cell similarity and gene importance has a natural connection with the popular ensemble method, random forests. The success of random forests relies on two prerequisites: 1) the decision trees are weak learners, and 2) the predictions made by the decision trees have low correlations (Wyner et al., 2017). In the proposed algorithm, we build each learner by randomly selecting B genes. Since each learner uses different sets of genes, cell type predictions for the learners are not highly correlated. Besides, the number of selected genes, B, is much smaller than the total number of genes. Thus, each learner is weak in terms of prediction performance. Based on these two prerequisites, EDGE are accurate in estimating cell similarities and gene importance scores.

Methods

Sketches

In a gene expression matrix $X \in \mathbb{R}^{C \times G}$, we randomly select B genes for each cell c. The sketch size of B is much smaller than the number of genes G. For any gene g in the selected B genes of cell c, we randomly pick a threshold within the minimum and maximum gene expression levels of X and check if the gene expression level of gene g is great (1) or less (0) than the threshold. Then, for each cell, we calculate a sketch vector $V \in \mathbb{R}^B$ containing $0s$ and $1s$. Let $W \in \mathbb{R}^B$ be the weight vector generated randomly. We use modulo hashing technique to map $V \cdot W$ to one of the predefined hash codes, where $\cdot$ represents dot product.

Similarity search

A hash code can be viewed as a box, in which similar cells are stored. We add value one to the i, jth entry of similarity matrix $S_l \in \mathbb{R}^{C \times C}, \quad l = 1, \cdots, L$, if cell i and cell j are mapped to the same hash code. We construct L weak learners and similarity matrices in the same way. The final similarity matrix $S = 1/L \sum_{l=1}^{L} S_l$. The detailed process is described in Algorithm S1 in Supplementary Information.

Spectral embedding

The next stage of the proposed method is the construction of a k-nearest neighbor (k-NNG) graph, which has the weighted adjacency matrix S. Once the graph is constructed, the spectral embedding is performed on the normalized Laplacian $D^{1/2}(D - S)D^{1/2}$, where D is the degree matrix for S. The output of this stage is top d eigenvectors of normalized Laplacian, E_d. The detailed process is described in Algorithm S2 in Supplementary Information.

Embedding optimization

Briefly, the optimization stage keeps similar cells close to each other and dissimilar cells far apart in the low-dimensional space. The optimization algorithm includes two stages in which a stochastic gradient descent algorithm with the decreasing step size is implemented. In the first stage, one updates according to the value of $\log(\Phi(e_m, e_n))$, where $\Phi(e_m, e_n) = (1 + a(\|e_m - e_n\|_2^2)^b)^{-1}$, $e_\cdot$ is the eigenvector of the cell $\cdot$ in E_d, n is the nearest neighbor of m based on the k-NNG graph. The hyperparameters a and b are estimated by non-linear least squares in McInnes et al. (2018). In the second stage, one updates according to the value of $\log(1 - \Phi(e_m, e_o))$, where o is one of negative samples, i.e., membership strength is 0. For sufficiently large samples, the negative samples are randomly selected using a uniform distribution.

Feature gene selection

A two-stage algorithm is implemented to identify feature genes. In the first stage, we apply a shared nearest neighbor modularity optimization based clustering algorithm to the optimized embedding matrix (Waltman and Van Eck, 2013; Stuart et al., 2019). The predicted class labels of cells are the input of the second stage, in which we select feature genes based on the measure of information entropy. Particularly, the entropy for weak learner l based on randomly selected B genes is measured by

$$\frac{1}{n_l}\sum_{i=1}^{n_l}(-\sum_{t=1}^{T} p_{t_i}\log p_{t_i}),$$

where n_l is the number of hash codes, T is the number of cell types predicted in the first stage, and p_{t_i} is the proportion of cell type t at the ith hash code. We then assign the values of entropy to the selected B genes. The genes that are randomly picked up in different weak learners are varied. With enough L, every gene should be selected at least one time. The averaged entropy values over L weak learners for G genes are used to identify feature genes. Let μ and σ be the mean and standard deviation of the averaged entropy values over L weak learners. We choose $\mu - 1.5 \times \sigma$ as the cutoff value to select top feature genes.

Computational complexities

One advantage of sketches methods is that they are computationally efficient and can build the learners in linear, i.e., $O(C)$, time (Jindal et al., 2018). For the optimization of embedding, the computational complexity is $O(kC)$ (Mikolov et al., 2013; Tang et al., 2016). The most time-consuming stage in our algorithm is spectral embedding. To significantly reduce the computational burden in this step, we use the RSpectra package to perform the large-scale eigenvalue decomposition (Qiu and Mei, 2019). The user time of EDGE for the varied number of weak learners was shown in Figure S3. It took less than 5 seconds when 500 weak learners were constructed for datasets with 1,000 cells and 500 genes. Note that all weak learners are constructed independently. It is thus convenient to implement our method

under the parallel computing framework, which could make EDGE much faster.

Simulation studies

We utilized the R package Splatter to simulate scRNA-seq datasets (Zappia et al., 2017). In all scenarios for dimensionality reduction, we generated 1000 cells and 1500 genes per cell and varied the proportion of differentially expressed genes, the group structure of cell populations, and the dropout rate.

To simulate scRNA-seq data for feature gene identification by EDGE, we first simulated single population scRNA-seq data with $1,000$ cells and 500 genes. Secondly, cell type labels were randomly assigned to the cells. For the scenarios with two cell types, the ratio of two cell types was $80 : 20$, whereas the ratio, $30 : 30 : 40$, was used in the scenarios with three cell types. Lastly, we converted nonzero values to zeros for the feature genes in the control group.

Real datasets

Below we describe all of the real scRNA-seq datasets used in the current study. All datasets are publicly available. The original Jurkat dataset contains about $3,200$ cells and the expression of $32,738$ genes. These two types of cells (Jurkat and 293T) are mixed at the ratio of $50 : 50$ (Zheng et al., 2017). To have the rare cell phenomenon, we used the dataset with the Jurkat cell proportion of $\sim 2.5\%$ (Jindal et al., 2018). The PBMC dataset consists of $3,362$ cells and $33,694$ genes sequenced by the 10x Chromium method (Ding et al., 2019). The Cell Ranger pipeline (v2.0.0) was used to process the PBMC dataset. Nine cell types were detected by the authors. For the mouse brain dataset, there are $3,005$ cells and $19,972$ genes (Zeisel et al., 2015). Seven major cell types and 47 molecularly subtypes were identified for the dataset.

Data preprocessing

We used the median normalization method to normalize simulated and real datasets (Jindal et al., 2018). The normalized datasets were then $\log_2$ transformed after the addition of one. We selected $1,000$ top variable genes for real datasets and 500 for simulated datasets using the variance of standardized values, which were calculated by the *FindVariableFeatures* function in the Seurat (Stuart et al., 2019).

Hyperparameters

The EDGE algorithm takes five important hyperparameters:

L: the number of weak learners;

H: the hash table size;

B: the number of genes to construct weak learners;

k: the number of nearest neighbors; and

d: the number of eigenvectors.

The number of weak learners, L, represents some degree of trade-off between low variance estimation and high computational cost. The default value for L in our algorithm is 500 (Figure S1, S3). The hash table size, i.e. H, is set 1,017,881 for all datasets (Jindal et al., 2018). A large B makes the algorithm sensitive to noises. Thus, values for B are typically less than $0.05 * G$. The EDGE method is robust to the number of nearest neighbors, k. The typical values are from 10 to 50 (Figure S2, S3). The dimensionality d is usually set to be two for visualization or be the number of target clusters plus one for prediction.

In real datasets, the default parameters were used for the t-SNE and UMAP methods except for the number of nearest neighbors. For a fair comparison, we utilized the same number of nearest neighbors or perplexity for all three methods.

Code availability

The EDGE R package is freely available from the link EDGE.

Competing interests

The authors declare no competing interests.

Author contributions

All authors contributed equally in conceiving the project. X.S. and Y.L. designed the algorithm and software and performed data analysis. All authors wrote and revised the manuscript.

Funding

This work has been partially supported by the United States Department of Agriculture [ARZT-1360830-H22-138 and ARZT-1361620-H22-149] to L.A.

Additional information

Supplementary Information is available for this paper.

References

Amir, E.-a. D., K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology 31*(6), 545.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*(1), 25.

Becht, E., L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology 37*(1), 38.

Beygelzimer, A., S. Kakade, and J. Langford (2006). Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 97–104.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Coifman, R. R., S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences 102*(21), 7426–7431.

Cooper, M. D. (2015). The early history of B cells. *Nature Reviews Immunology 15*(3), 191.

Ding, J., X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, et al. (2019). Systematic comparative analysis of single cell RNA-sequencing methods. *BioRxiv*, 632216.

Ding, J., A. Condon, and S. P. Shah (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications 9*(1), 2002.

Gionis, A., P. Indyk, R. Motwani, et al. (1999). Similarity search in high dimensions via hashing. In *Proceedings of the Twenty Fifth VLDB Conference*, Volume 99, pp. 518–529.

Grün, D., A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature 525*(7568), 251.

Guttman, A. (1984). R-Trees: A dynamic index structure for spatial searching. *SIGMOD Record 14*(2), 47–57.

Jindal, A., P. Gupta, D. Sengupta, et al. (2018). Discovery of rare cells from voluminous single cell expression data. *Nature Communications 9*(1), 4719.

Kim, T., I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang (2018). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics*.

Kobak, D. and P. Berens (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications 10*(1), 1–14.

Krauthgamer, R. and J. R. Lee (2004). Navigating nets: simple algorithms for proximity search. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 798–807.

Linderman, G. C., M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods 16*(3), 243.

Liu, T., A. W. Moore, K. Yang, and A. G. Gray (2005). An investigation of practical approximate nearest neighbor algorithms. In *Advances in Neural Information Processing Systems*, pp. 825–832.

Lv, Q., W. Josephson, Z. Wang, M. Charikar, and K. Li (2006). Ferret: a toolkit for content-based similarity search of feature-rich data. *ACM SIGOPS Operating Systems Review 40*(4), 317–330.

Maaten, L. v. d. and G. Hinton (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 2579–2605.

McInnes, L., J. Healy, and J. Melville (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mi, H., A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research 47*(D1), D419–D426.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science 344*(6190), 1396–1401.

Pierson, E. and C. Yau (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology 16*(1), 241.

Qiu, Y. and J. Mei (2019). *RSpectra: solvers for large-scale eigenvalue and SVD problems*. R package version 0.15-0.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija (2019). Comprehensive integration of single-cell data. *Cell*.

Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods 6*(5), 377.

Tang, J., J. Liu, M. Zhang, and Q. Mei (2016). Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297.

Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319–2323.

Tran, D., H. Nguyen, B. Tran, and T. Nguyen (2019). Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*.

Vento-Tormo, R., M. Efremova, R. A. Botting, M. Y. Turco, M. Vento-Tormo, K. B. Meyer, J.-E. Park, E. Stephenson, K. Polański, A. Goncalves, et al. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature 563*(7731), 347.

Wagner, A., A. Regev, and N. Yosef (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology 34*(11), 1145.

Waltman, L. and N. J. Van Eck (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B 86*(11), 471.

Wang, T., B. Li, C. E. Nelson, and S. Nabavi (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics 20*(1), 40.

Wang, Z., W. Dong, W. Josephson, Q. Lv, M. Charikar, and K. Li (2007). Sizing sketches: a rank-based analysis for similarity search. In *ACM SIGMETRICS Performance Evaluation Review*, Volume 35, pp. 157–168.

Wyner, A. J., M. Olson, J. Bleich, and D. Mease (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research 18*(1), 1558–1590.

Zappia, L., B. Phipson, and A. Oshlack (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biology 18*(1), 174.

Zeisel, A., A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science 347*(6226), 1138–1142.

Zheng, G. X., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 8*, 14049.