Dimensionality Explorer for Single-Cell Analysis

Haejin Jeong*
Korea University

Hyoung-oh Jeong†
Ulsan National Institute of Science and Technology

Semin Lee†
Ulsan National Institute of Science and Technology

Won-Ki Jeong*‡
Korea University

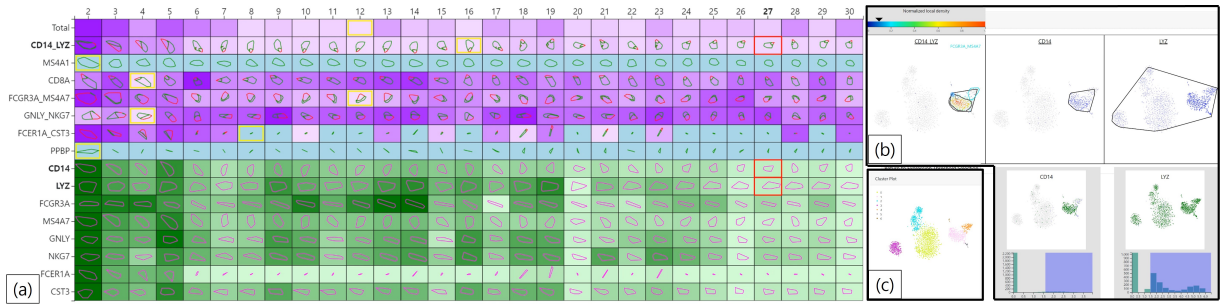

Figure 1: Overview of the proposed system. (a) Hull heatmap shows an overview of overlaps among several cell types for multiple dimensionalities. Each hull represents the target cells where each cell marker or cell type is differentially expressed. (b) Marker expression plots show the expression of cell markers or marker groups. Based on the plots, analysts can check the quality of the hulls and overlaps with other cell types. Analysts identify the distribution of marker expression and modify the hulls by using the cell filtering function. (c) Cluster plot shows a clustering result of cells. Analysts can use this plot when they evaluate the quality of the hulls.

Abstract

Single-cell RNA sequencing (scRNA-seq) is becoming popular in studying the gene expression of cells at the single-cell level. ScRNA-seq enables analysts to characterize cell types, thereby providing a better understanding of dynamic biological processes. In scRNA-seq data analysis, principal component analysis (PCA) is commonly used to reduce at least thousands of dimensions in the raw data to a manageable size so that analysts can visualize and cluster cells to identify different cell types. The conventional process to determine the optimal dimensionality includes a laborious manual review of hundreds of different projection plots. To address this problem, we introduce a dimensionality explorer for single-cell analysis, which is a visualization system that helps analysts to effectively determine the optimal dimensionality of scRNA-seq data. It employs a hull heatmap, which provides a holistic view of overlaps among multiple cell types across various dimensionalities using a convex hull-embedded color map. The hull heatmap effectively reduces the burden of manually reviewing hundreds of projection plots to determine the optimal dimensionality. Our system also provides interactive gene expression level visualization and intuitive lasso selection, thereby allowing analysts to progressively refine the convex hulls of the hull heatmap. We demonstrate the usefulness of the proposed system through a user study and three case studies conducted by domain experts.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Heat maps; Human-centered computing—Visualization—Visualization application domains—Visual analytics

1 Introduction

Single-cell analysis is a detailed analysis of the genome and transcriptome at the single-cell level. Gene expression is the process by which a gene is transcribed into RNA, and cell markers refer to genes specifically expressed in a specific cell (see Figure 2). There exist unique combinations of cell markers for the specific cell type, which are used to identify and classify individual cells. Therefore, to estimate the different types and characteristics of cells, researchers need to discover which cell markers are expressed in the cells based on the expression level (the amount of RNA) of each cell marker. For this, single-cell RNA-seq (scRNA-seq) is a powerful technique that enables the measurement of the gene expression levels of individual cells [39].

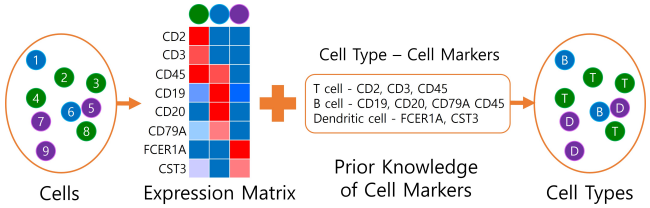

Figure 2: An expression matrix is generated from the collected cells. The cell type is estimated based on the matrix and cell markers.

The scRNA-seq analysis workflow involves several steps, including data pre-processing, feature selection, dimensionality reduction, clustering, and identification of differentially expressed genes. Among them, reducing the dimensionality of data is a crucial step because each cell type is identified in the 2D feature space generated based on the reduced dimensionality. The standard procedure for dimensionality reduction in the single-cell analysis is to select significant principal components (PCs) that are generated using principal component analysis (PCA) to reduce at least thousands of dimensions of input data to a manageable size (usually less than a hundred dimensions) while preserving global and local structures in the 2D projection that characterizes the target cell types most effectively (see Figure 3).

In this work, we specifically focus on how to select the optimal (PCA intermediate) dimensionality, which is one of the most time-consuming and laborious processes in single-cell analysis. Determining a good embedding is not always easy and intuitive, because cell locations in the embedding may change significantly as the dimensionality changes, which results in the gathering or dispersing of the target cells in which a target cell marker is differentially expressed in the 2D embedding space (see Figure 4). Single-cell analysts find multiple target cell types on a dimensionality reduction plot, but sometimes target cell type regions overlap with each other, making it difficult to analyze. Thus, conventional dimensionality selection methods rely on an exhaustive search in the dimensionality

*e-mail: {haejinjeong, wkjeong}@korea.ac.kr
†e-mail: {hyoung-oh, seminlee}@unist.ac.kr
‡Corresponding author.

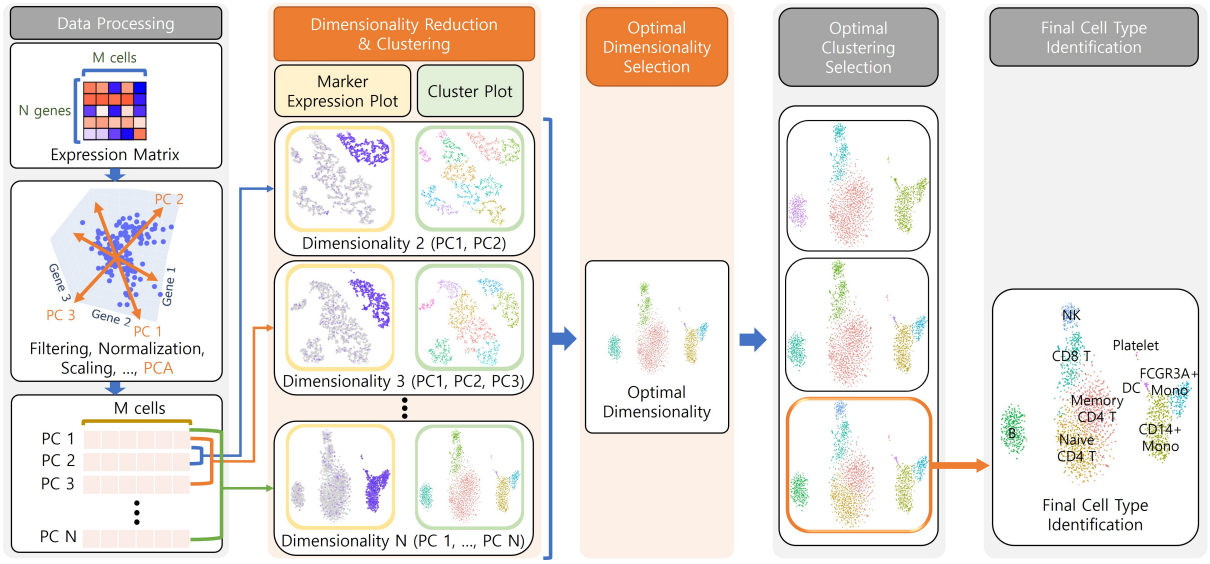

Figure 3: Overview of the single-cell RNA sequencing (scRNA-seq) analysis workflow. ScRNA-seq produces an expression matrix consisting of at least thousands of cells and genes. Data processing including filtering, normalization, scaling and principal component analysis (PCA) is performed on the matrix. Dimensionality reduction and clustering are applied to the top-N principal components (PCs) and the marker expression and cluster plots are generated for each dimensionality (i.e., the number of used PCs) to determine the optimal dimensionality. After selecting the optimal dimensionality, analysts perform cluster analysis and finalize cell type identification. Note that steps marked in orange color (dimensionality reduction & clustering, optimal dimensionality selection) are our work proposed in this paper.

space to find the optimal dimensionality where multiple target cell types are clearly separated. Analysts need to repeatedly review many different embeddings back-and-forth and finalize the best decision sorely based on the analyst's memory.

To address this problem, we introduce a dimensionality explorer for single-cell analysis, which guides analysts in identifying the optimal dimensionality to achieve the best cell type identification result with significantly less effort. One of the primary features of our method is a hull heatmap (Figure 1a), which is a 2D (gene-dimensionality) heatmap augmented by convex hulls of the target types of cells. The hull heatmap shows overlaps of cell type areas across various dimensionalities; consequently, analysts can easily find the dimensionality where the target cell types are clearly separated. Therefore, the labor-intensive manual review of different embeddings can be significantly reduced. Our system also provides intuitive interactive cell filtering functions using the lasso selection (Figure 1b), which allows analysts to progressively update the hull heatmap until they obtain the most satisfactory heatmap and identify the optimal dimensionality. We demonstrate the effectiveness of the proposed system via a user study and three case studies conducted by domain experts.

Our contributions are:

- The problem characterization of scRNA-seq analysis and requirement analysis for system design;
- A novel visualization technique (hull heatmap) that shows the difference in hundreds of marker expression plots according to multiple dimensionalities and cell markers at a glance;
- Interactive cell filtering to update the hull heatmap using user feedback;
- A user study and three case studies showing the effectiveness of the proposed system.

The online demo and source code are publicly available at `https://github.com/hvcl/DESC`.

2 Background

2.1 Overview of Single-Cell Analysis Workflow

Figure 3 shows the overview of the single-cell analysis workflow. ScRNA-seq provides the expression matrix that shows the gene expression levels of each cell. The matrix consists of at least thousands of cells and genes. Once the raw input data are pre-processed (e.g., cell filtering, normalization, identification of highly variable genes, and scaling), dimensionality reduction is performed to reduce the data dimension because the high dimensionality of scRNA-seq data makes clustering and 2D projection results unreliable (i.e., the curse of dimensionality) [19]. Among the several existing dimensionality reduction methods, PCA is commonly used in single-cell analysis to overcome the extensive technical noise in any single feature for scRNA-seq [2].

Because different types of cells can express the same genes, *differentially* expressed cell markers (genes) are used to estimate cell types. A gene is declared differentially expressed in a cell cluster if a significant difference is observed in expression levels among multiple cell clusters. A commonly used method to examine cell marker expression is a 2D projection plot of cells (e.g., t-stochastic neighbor embedding (t-SNE) [42] or uniform manifold approximation and projection (UMAP) [25]) colored by the expression levels of each marker (see Figure 5), i.e., marker expression plot. Because it is difficult to simultaneously visualize the expression level of multiple markers of a cell in a single plot, analysts usually review multiple plots to find cells in which all target markers are differentially expressed.

When visualizing cells in a 2D plot, the appropriate number of PCs (i.e., dimensionality) should be selected because the selection of the dimensionality directly affects the 2D projection result, which eventually affects cell type identification. Single-cell analysts obtain important clues from the spatial distance between cells in the 2D projection plot based on the assumption that similar cells tend to have smaller feature distances. However, certain rare cell types with a small number of cells can be found only in certain specific dimensionalities. Hence, selecting only a few major PCs may not satisfactorily represent such cells [21]. Furthermore, Raimundo *et al.* showed that the selection of the dimensionality strongly affects the performance scores (adjusted mutual information and silhouette) of the embedding of scRNA-seq data [28]. Therefore, selecting the optimal number of PCs for retaining the important information that can efficiently represent the target cell types while removing technical noise or other unwanted sources of variation by reducing

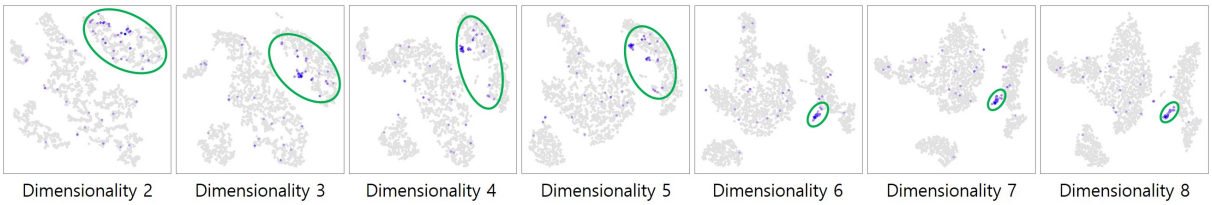

Figure 4: The embedding changes for the cells in which gene FCER1A is differentially expressed (the target cells of FCER1A) as the dimensionality increases. In dimensionalities 2–5, the target cells are spread out; therefore, it is difficult to identify the cell type. Conversely, in dimensionalities 6–8, the target cells start to form a cell cluster; consequently, it is easier to define the cell type. Similar to this example, cell type identification results can differ based on the selection of dimensionality.

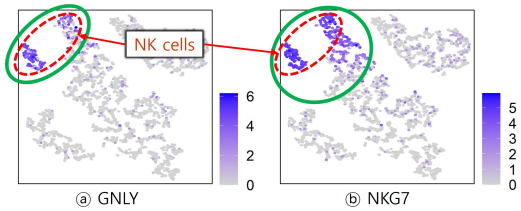

Figure 5: Expression plots of cell markers GNLY (a) and NKG7 (b). The color represents expression level. The green circles represent the cells where each marker is differentially expressed. Because both genes are cell markers of NK cells, the cells where both genes are differentially expressed (i.e., the cells inside the red dashed circles) are estimated as NK cells.

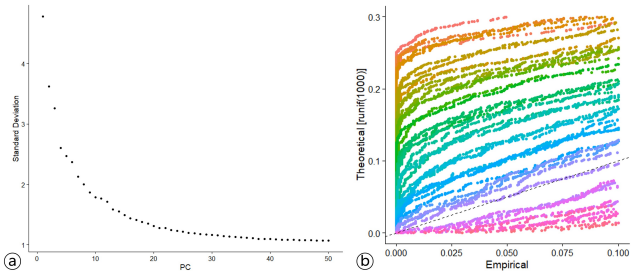

Figure 6: (a) Elbow plot shows the fraction of variance explained by each PC. (b) JackStraw plot shows the distribution of p-values for each PC.

dimensionality is a challenging but crucial process.

2.2 Optimal Dimensionality Selection

The conventional optimal dimensionality selection methods are based on analyzing the changes in the data distribution across different dimensionalities. One method is to use an elbow plot (Figure 6a), which shows the fraction of variance explained by each PC. Researchers are required to visually identify the point where the curve makes a sharp bend, which is often referred to as the "elbow" and retain only PCs before the occurrence of the elbow. Another commonly used method is the JackStraw plot [9] (Figure 6b), which shows the distribution of p-values for each PC. The best PC is found near the sharp drop (change) of p-values in this plot. However, it is not always a straightforward process to determine a proper threshold in these plots because changes are sometimes considerably subtle and consequently accurate differentiation cannot be performed. Therefore, a manual review of 2D expression plots of target cell markers and cluster plots for each PCA intermediate dimensionality is preferred for a more accurate analysis.

The process of determining the optimal dimensionality by reviewing the marker expression and cluster plots is performed as follows. The marker expression and cluster plots share the same 2D embedding from the projection of the selected PCA intermediate dimensionality. First, target cell types and markers need to be listed. Analysts usually find known cell markers of target cell types through literature review. *Target cell* refers to the cell in which a target cell marker is differentially expressed, and is the candidate for the target

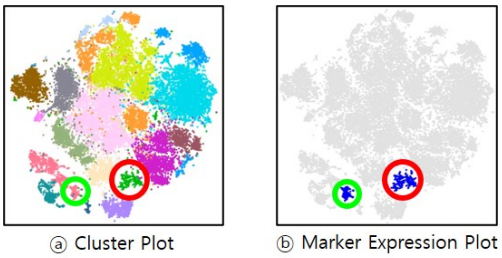

Figure 7: Marker expression (b) shows the two target cell clusters (green and red circles) of ILC1-like cells. However, on cluster plot (a), the cells inside the green circle form a cluster with other cells that are not the target cells. Thus, the cells in the red circle are more likely to be ILC1-like cells.

cell type. Second, analysts identify the target cells of each target cell type in each dimensionality based on the marker expression plots. Figure 5 shows sample expression plots of GNLY, and NKG7 genes. Each dot represents a single cell, and its color represents the expression level (i.e., the degree to which a particular gene is expressed) of each cell. Analysts identify the target cells where each marker is differentially expressed (i.e., the cells in the green circles in Figure 5). Note that GNLY and NKG7 are cell markers of NK cells. Once analysts identify the target cells of GNLY and NKG7, then they collect cells commonly included in both target cells and assign the final cell type to them (in this example, NK cells, Figure 5 red dashed circle). Certain cell types are often not found in some dimensionalities. As shown in Figure 4, a few target cells of cell marker FCER1A are scattered in dimensionality 2–5; therefore, it is difficult to define a tight cluster where FCER1A is differentially expressed (i.e., the green circle contains cells that are not differentially expressed). In this case, analysts should search for other dimensionalities for a tighter cluster, such as dimensionality 6–8 in Figure 4.

After finding all target cells, analysts then select the optimal dimensionality that makes the best target cell clusters (i.e., cells belonging to the same cell type are gathered closely together while different cell types are clearly separated, and all the target cell types are found). Also, they evaluate the dimensionalities using cluster plots (Figure 1c). The cluster plot is colored by the clustering results of the selected PCA intermediate dimensionality. If the target cells estimated as the same cell type are differently clustered or clustered with the target cells of other cell types, some of the target cells may not be the correct target type of the cells (see Figure 7). In this case, analysts can redefine the target cells based on the cluster plot or review other cluster plots in the other dimensionalities.

The position of the target cells usually changes depending on the dimensionality (see Figure 4). Thus, some target cell types may not be found in some dimensionalities, or the target cells from different cell types can overlap. In addition, the cell type identification result can be different from the clustering results. Therefore, at least hundreds of plots must be examined to determine the optimal dimensionality for at least tens of dimensionalities and target cell markers.

3 Related Work

3.1 Single-Cell Analysis Methods

Clustering analysis is commonly used to handle multidimensional scRNA-seq data from various cell types. Several publicly available software packages for single-cell analysis such as Seurat [5], Scanpy [48], and SINCERA [13] provide different clustering methods. These software tools focus on clustering rather than the dimensionality of data. Some of these tools provide information on each dimensionality, but analysts are still required to employ heuristic methods to determine the optimal dimensionality. These tools are useful for determining significant cell clusters to obtain more information about cell types after selecting the optimal dimensionality.

In single-cell analysis, most scRNA-seq studies rely on t-SNE for 2D projection [22]; however, more recent methods such as UMAP [25] are also emerging. Deep learning-based dimensionality reduction methods are also used for single-cell analysis (e.g., parametric t-SNE [41], parametric UMAP [31], scvis [11], DR-A [24], and VASC [46]). These methods need more scRNA-seq datasets for good performance. There are various dimensionality reduction methods for single-cell analysis [3, 29, 37, 40, 43, 44, 49]. However, to use these methods, analysts should manually fine-tune the parameters of the algorithms and the dimensionality of the data, which severely affects the projection results. Choosing the appropriate parameters remains an unsolved problem [19].

Most existing R or Python packages for single-cell analysis require analysts to write codes. In addition, they should select suitable analysis algorithms and visualization methods on their own. To address these problems, various visual analytical tools were developed to guide analysts in examining single-cell data. Cerebro [14] enables analysts to investigate single-cell data through an intuitive graphical interface. CyteGuide [16] guides analysts in exploring the hierarchy of single-cell data in a single view. Cytosplore [15] provides multiple linked views that enable analysts to identify both known and unknown cell types easily. VDJView [32] visualizes scRNA-seq data with metadata profiles to ease the process of hypothesis testing, data interpretation, and the discovery of cellular heterogeneity. scQuery [1] provides an automated pipeline for downloading and analyzing publicly available scRNA-seq datasets. scSVA [38] supports interactive three-dimensional visualization and exploration of massive single-cell data. Even though there are several visual analytical tools for single-cell analysis, none of them focus on dimensionality selection, which makes it inappropriate to directly compare our system.

3.2 Visual Analysis of Dimensionality Reduction and Clustering

There exist various methods to analyze dimensionality reduction or clustering using visualization. PRIM-9 [12] allows analysts to examine multidimensional data using continuously updated projections. Asimov *et al.* proposes a sequence of orthogonal projections of multidimensional data and searches for desirable sequences to understand the shape of data [4]. The Hierarchical Clustering Explorer is a visualization tool for hierarchical clustering, which provides an overview of large data sets, dynamic query controls, coordinated displays, and cluster comparisons [34]. Cluster Sculptor [27] is a cluster analysis framework to enable analysts to interactively adjust clustering parameters based on the visualization of characteristics of high-dimensional data. Clustrophile 2 [7] recommends various clustering parameters and continues to improve the clustering results based on user feedback. INCREMENT [26] refines clustering results based on user feedback by training a feature embedder to map the input features into a new feature space. TINDER [36] is a Bayesian prior elicitation framework based on user feedback. Analysts can reject clustering results and obtain new results.

Similar to these methods, ours enables users to easily explore multiple projection plots and guides users to search for optimal solutions based on user feedback. However, the major difference is that our method provides an intuitive visualization-guided approach (i.e., the hull heatmap and interactive visual interface) to compare more than a hundred projection plots in a single view, which saves both time and labor used in comparing many marker expression plots for multiple dimensionalities and cell markers.

4 Requirement Analysis for System Design

To design our visualization system, we interviewed five domain experts who routinely use Seurat [5] and SC3 [20] for single-cell analysis, and derived the requirements as listed below.

R1: Visualization of multiple cell marker expressions over various dimensionalities in a single view: To determine an optimal dimensionality, analysts should generate and review expression plots for tens of markers and dimensionalities, which require a significant amount of time and effort. In the conventional workflow, analysts compare multiple marker expression plots displayed on a screen. Owing to the limited screen size, analysts should frequently change their views to compare thousands of plots. In addition, analysts are required to memorize previous views and compare them using their mental images, which is not easy considering the large number of plots that must be analyzed. Therefore, identification of the expression of multiple cell markers at various dimensionalities in a single view using a compact visual representation will be helpful for analysts to compare multiple expression plots.

R2: Tracking target cells of target cell types: To identify a specific cell type, analysts must identify the expression of relevant cell markers. In each dimensionality, they should compare expression plots of multiple cell markers and identify the target cells where all the markers for a target cell type are differentially expressed. Then, they keep track of those cells over at least tens of dimensionalities to determine the optimal dimensionality, which consumes a large amount of time and memory. Highlighting the target cells relieves them of having to compare multiple plots.

R3: Visual cues to help identify target cells: Analysts identify the target cells using their colors (expression levels) and locations in a marker expression plot. As there are no widely accepted criteria (e.g., distance between cells, intensity level, etc) for the target cells, analysts subjectively define the target cells by measuring the relative distance and difference in expression levels between cells. In such a scenario, visual cues provide better judgment for extra information about the distribution of expression levels and local densities.

R4: Visual cues to help identify overlaps between target cell types: Analysts find the dimensionality in which all target cell types are clearly separated. However, there are two challenges. First, the positions of cells change depending on the dimensionality. Thus, analysts do not know which dimensionalities have overlaps until they review all dimensionalities. Second, cell markers from different cell types can be expressed in a cell. In this case, analysts should find the type of the cell by reviewing several marker expression levels and clustering results. To address these challenges, visual cues for avoiding unwanted overlaps are required.

5 Dimensionality Explorer for Single-Cell Analysis

Our method guides analysts in determining the optimal dimensionality through a novel visualization technique that are specifically designed for the exploration of the dimensionality space. Hull heatmap provides an overview of overlaps among several cell types over multiple dimensionalities. By using cell filtering, analysts can update the hull heatmap by interactively modifying the parameters defining the hulls based on their expertise. All the dimensionality reduction plots (t-SNE plots) used in our system were generated by the RunTSNE function in Seurat [5] with default parameters. In the following sections, we explain the design rationale of the proposed system

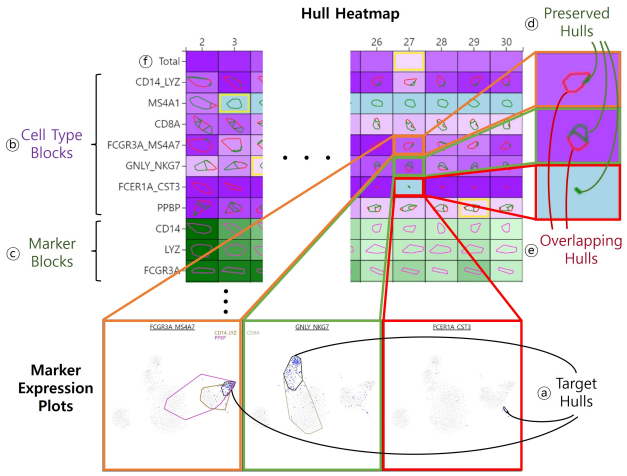

Figure 8: Components of hull heatmap. Hull heatmap consists of (rows of) cell type blocks and marker blocks. A cell type block shows the target hulls of the corresponding marker group. A marker block shows the target hulls of a marker. Each target hull can be divided into the preserved hulls and overlapping hulls. The smaller the overlap between cell types is, the brighter the color of the cell type block becomes. When there is no overlap, the cell type block has a light blue color. The smaller the target hull, the brighter the color of the marker block.

and the method used to search the optimal dimensionality using the proposed system in detail.

5.1 Hull Heatmap

The hull heatmap was designed to show the change of marker expression for multiple dimensionalities and cell markers in a single view (**R1**). The row of the heatmap represents a cell marker or cell type and the column represents the dimensionality. The target cells of each cell marker are visualized in the form of a convex hull in each dimensionality and the hull is called the target hull (Figure 8a). In order to show hundreds of the target cell regions on a limited visualization space, we chose convex hulls, which can visualize the regions relatively simpler than other methods such as concave hulls and alpha-shapes. We call a *cell* in the hull heatmap a *block* to avoid confusion with the word "cell" also used in the single cell we analyze. Then, each target hull is drawn in each block. Thus, analysts can easily identify the area of the target cells over multiple dimensionalities without the requirement of reviewing many plots (**R2**). Also, the heatmap enables analysts to track the target cells of a target cell type by making a group of the target cell markers (for example, IL7R_CCR7 is a marker group consisting of two markers, IL7R and CCR7), so the blocks are divided into cell type blocks (marker group blocks) (Figure 8b) and marker blocks (Figure 8c). Each block of the heatmap also includes overlap information between target cell types (**R4**). A target hull is divided into preserved hulls and overlapping hulls. Preserved hulls (Figure 8d) represent the area of the target hulls not overlapped by the target hulls of other target cell types. Overlapping hulls (Figure 8e) represent the area of the target hulls overlapped by the target hulls of other target cell types. The outline color of preserved hulls is green, and the outline color of overlapping hulls is red. Thus, analysts can easily identify which area is overlapped by other target cell types. The color of the block also provides information on the overlap. When a cell type block has only preserved hulls, the color of the block is light blue. When a cell type block has both preserved and overlapping hulls, its color is decided by the preservation ratio. The ratio is defined as the area of the preserved hulls over the area of the target hulls. For coloring, the ratio is normalized in each row of the heatmap. The cell type block with the lowest ratio in the row has a dark purple color, and the cell with the highest ratio has a light purple color. The intermediate cells get colors by linear interpolation. The color of the marker block is decided by the area of the target hulls. The total row (Figure 8f) shows the average preservation ratio of all target cell types. Based on this color visualization, analysts can easily identify how much different target hulls overlap each other in each dimensionality. The yellow edge highlights a cell type block that has the largest preservation ratio in the row of the heatmap, so analysts can search the optimal dimensionality based on the edge.

To generate a convex hull, we define the target cells as follows. First, we identify the cells where the expression level of each cell marker is higher than the expression level threshold (each cell marker is differentially expressed). Then, we compute the local density of each cell using Gaussian kernel density estimation [35] and normalize the density values. Finally, we collect the target cells by filtering cells whose local density is lower than the user-given density threshold to avoid the generation of a large hull by a few outlier cells. When creating target hulls of a cell type block, we identify the cells whose expression levels of all group member markers are higher than the given expression level threshold, filter the cells based on the local density, and create convex hulls of the cells. The default expression level threshold and density thresholds are 0 and 0.1, respectively. Analysts can change both the expression level and density thresholds using the cell filtering function (Section 5.4).

5.2 Marker Expression Plot

The marker expression plot (see Figure 8) is a dimensionality reduction plot that is colored based on the expression level of a cell marker or the target cells of a marker group. On a marker expression plot of a cell type, the target hulls of the cell type and other overlapping cell types are drawn. With the marker expression plots, analysts can identify the target cells and overlapping cell types.

5.3 Cluster Plot

The cluster plot (Figure 10a) is a t-SNE plot that is colored according to the clustering result of each dimensionality (number of PCs). The embedding is the same as the marker expression plot. The clustering result is generated using the clustering function of Seurat [5] with the default parameters. The function clusters cells using a shared nearest neighbor modularity optimization-based clustering algorithm [45]. Analysts can change the dimensionality in the hull heatmap; then, the cluster plot is updated into the dimensionality reduction plot for the selected dimensionality.

The cluster plot is used to evaluate each 2D embedding. Analysts can check how cells that are estimated as a particular cell type form clusters in the cluster plot and obtain information for further analysis. If the cells of a specific cell type are divided into multiple clusters in the cluster plot, it may imply that the cell type can be divided into multiple sub-types. If the cells of a cell type are included in a single cluster with other cell types, it may imply that the cell types in the cluster have a close relationship. Analysts explore various dimensionalities to select the optimal one in which only cells of a target cell type form a single cluster.

After the optimal dimensionality is selected, analysts commonly explore clustering results performed on the selected dimensionality with various clustering algorithms and parameters and compare cell type identification results. Because this step is beyond the scope of this work, we used the default clustering algorithm and parameters provided in Seurat for generating cluster plots.

5.4 Cell Filtering

Before cell filtering, the target hulls depend on the default expression level threshold and density threshold. Analysts usually subjectively judge which cell marker is differentially expressed in which cell cluster. When the hulls and their judgments do not match, the hulls of the hull heatmap and marker expression plots may not be

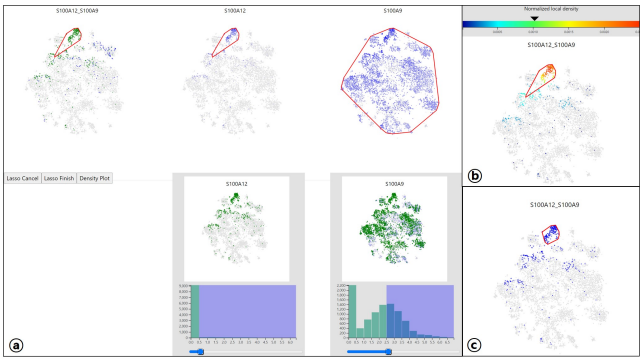

Figure 9: Example of the cell filtering process for the S100A12/S100A9 group. (a) In each cell filtering window of the member markers, analysts can change each expression threshold. The target cells of the group are updated based on the new thresholds and the color of the cells changes to green. (b) In the next step, the color of the target cells changes based on their local densities. Analysts can select a new density threshold using the density color bar. (c) Final target cells and hull of the group.

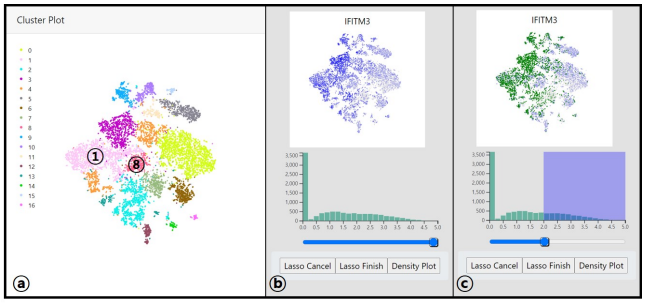

Figure 10: By selecting a proper expression level threshold, analysts can figure out expression pattern. It can be easily observed that the clusters 1 and 8 in (a) have different expression levels in (c).

useful. Analysts want to focus on certain significant (important) cells and follow the positional changes of only those cells when the dimensionality changes (**R2**).

To address this issue, we propose cell filtering (Figure 9). A cell filtering window consists of a marker expression plot, an expression level histogram, and an expression threshold slider. When we perform cell filtering on a marker group, one cell filtering window appears for each member marker. The expression level histogram is a histogram of the marker expression levels of all the cells in the dataset (**R4**). Analysts can move the slider to change the marker expression threshold after checking the expression level histogram. The target cells whose expression level is higher than the updated threshold are colored in green in the marker expression plot of the window. Whenever the expression level threshold of the member marker is updated, the target cells of the group are also updated and colored in green in the marker expression plot (Figure 9a). When the user clicks a button, the target cells of the group are colored according to their local densities (**R4**) (Figure 9b), and the density color bar appears. The local densities are calculated using Gaussian kernel density estimation [33]. Analysts can easily identify regions where the target cells are gathered closely. Then, analysts can select a new density threshold by selecting a density color on the density color bar. By selecting a proper density threshold, analysts can filter out the few target cells located far from the remaining cells, thereby preventing hulls from becoming too large. Another option for selecting the cells of interest is to use a lasso tool to directly select cells in the marker expression plot. This allows analysts to effectively filter out less relevant cells. For all dimensionalities, our system updates the target cells and hulls of the group based on the new density threshold and cell selection (Figure 9c).

Analysts can use cell filtering to examine the expression level pattern of a marker. Figure 10b shows the marker expression plot of IFITM3. The cells in clusters 1 and 8 in Figure 10b appear to have similar expression levels with no expression level threshold in Figure 10b. After increasing the threshold, we found that IFITM3 was differentially expressed in the cells of cluster 1 but not in the cells of cluster 8. Therefore, the cell types of the two clusters are likely different.

6 Evaluation

To demonstrate the usefulness of our system, we conducted a user study with three domain experts (P1-3) and three case studies with four domain experts (P1-4). The information of the participants is listed in Supplemental Table 1. The user study focuses on a comparison between the state-of-the-art clustering methods for scRNA-seq data and our method, and the case studies focus on demonstrating how our method is useful for determining the optimal dimensionality.

6.1 User Study Design

The main goal of this user study is to quantitatively assess the performance of our method. For this, we compared our method with the state-of-the-art scRNA-seq data clustering methods to show that our method enables analysts to get similar or better cell type identification results in a shorter time. Thus, we compared the clustering accuracy and analysis time. The methods compared in this study are scziDesk [8], scGNN [47], and graph-sc [10], which are the state-of-the-art deep learning-based scRNA-seq data clustering methods. The clustering performance of the methods depends on both 2D embedding (projection) and clustering (assigning labels) results. When using the clustering methods, participants obtained various clustering plots by changing parameters; among them, the best one is selected. In our method, participants also compared different plots with our visualization tool and selected a plot with the best clustering result. In this study, three participants analyzed two publicly available datasets: epithelial cells and endothelial cells datasets [18]. The epithelial cells dataset contains 3643 cells and 29634 genes, and the endothelial cells dataset contains 2107 cells and 29634 genes. In the epithelial cells dataset, the participants identified the following four target cell types: AT1, AT2, Club, and Ciliated. The cell markers used to identify the above cell types were AGER, SFTPC/LAMP3, SCGB1A1, and FOXJ1/RFX2, respectively. In the endothelial cells dataset, the participants identified the following five target cell types: Tip-like ECs, Stalk-like ECs, Lymphatic ECs, and EPCs. The cell markers used to identify the above cell types were RAMP3/RGCC/ADM, SELP/ACKR1, CCL21/ LYVE1, and TYROBP/C1QB, respectively. The participants searched the best clustering plot for four methods (ours and the three state-of-the-art methods) and two datasets, 8 times in total. A dry run was first conducted on the test dataset with our tool to familiarize the participants with the system for performing the main tasks.

6.2 User Study Results

The clustering performance is measured with two external score metrics, adjusted rand index (ARI) [17] and normalized mutual information (NMI), and two internal score metrics, Silhouette score [30] and Calinski-Harabasz (CH) index [6]. The external scores are used to evaluate the agreement with the ground truth and the internal scores are used to evaluate the cluster compactness. For all scores, a higher value means better performance.

Table 1 shows the average clustering performances of each method with the epithelial and endothelial cells dataset. In the result of the epithelial cells dataset, scziDesk showed the highest ARI and NMI scores. Ours and graph-sc produced similar ARI and NMI scores to scziDesk, while scGNN showed the lowest scores. Also, graph-sc produced the highest Silhouette and CH scores. In the result of the endothelial cells dataset, ours has the highest ARI

Table 1: The average clustering performances of each method with the epithelial and endothelial cells dataset.

		Ours	graph-sc	scziDesk	scGNN
Epithelial	ARI	0.911	0.906	**0.930**	0.344
	NMI	0.847	0.875	**0.886**	0.486
	Silhouette	0.480	**0.568**	0.335	0.223
	CH	3284.369	**4710.016**	1480.184	1637.838
Endothelial	ARI	**0.553**	0.479	0.435	0.234
	NMI	0.581	**0.589**	0.553	0.310
	Silhouette	**0.275**	0.208	0.126	0.240
	CH	486.898	419.986	203.555	**942.409**

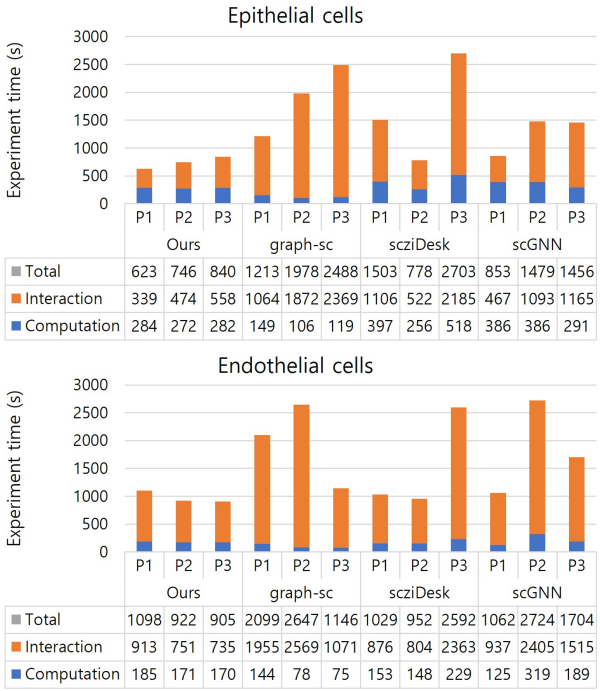

Figure 11: The experiment time of each method on the two datasets. Participants were able to finish the task up to three times faster using our method compared to SOTA clustering methods for single-cell analysis.

and Silhouette scores. Also, NMI of ours is similar to the highest NMI. scGNN has the highest CH score because it tends to form much more separate clusters than other methods.

In the experiments of other methods except scGNN, the number of clusters was set to the number of target cell types. However, scGNN cannot specify the number of clusters, so some participants did not get accurate results in some experiments. Therefore, the average ARI and NMI of scGNN were always the lowest.

Fig. 11 shows the experiment time of each method with the two datasets. The time includes both computation time by a machine and interaction time by participants. With both datasets, the participants took the shortest time on average when using ours. It shows that ours enables analysts to obtain better or similar clustering results to the state-of-the-art clustering methods for scRNA-seq data in a shorter amount of time.

6.3 Case Studies

We conducted three case studies with four domain experts. The goal of the case studies is to conduct an in-depth analysis of how our method can help analysts identify the optimal dimensionality. In each case study, three experts (P1–P3) individually identified the target cell types and optimal dimensionality using our method, and the most experienced expert (P4) validated the optimal dimensionalities found by the three experts. The experts analyzed three publicly available scRNA-seq datasets [23,50], i.e., peripheral blood mononuclear cells (PBMC), T cells and myeloid-like cells in the lung tumor microenvironment. The analysis dimensionalities ranged from 2 to 30 for PBMC dataset, and from 2 to 50 for T cells and myeloid-like cells datasets. Each case study consists of the following three steps: 1) loading the given dataset to our system, 2) modifying hulls of target cell types, and 3) searching the optimal dimensionality.

6.3.1 PBMC Dataset

The target cell types used in this study are as follows: CD14+ Mono, B, CD8+ T, FCGR3A+ Mono, NK, DC, and Platelet cells. The cell markers used to identify the above cell types were CD14/LYZ, MS4A1, CD8A, FCGR3A/MS4A7, GNLY/NKG7, FCER1A/CST3, and PPBP, respectively.

After creating a heatmap (Figure 12a), P1 found that the dimensionality 27 in the total row was highlighted by the yellow edge. He started to analyze plots in this dimensionality and found that the groups except for MS4A1 and FCER1A/CST3 had overlaps with other groups. P2 explored several dimensionalities at the beginning, but he was not able to not find any significant difference and started cell filtering in dimensionality 6. P3 started to analyze the dataset in dimensionality 11, which was highlighted by the yellow edge in the CD8A row. In dimensionality 6 and 11, all groups except MS4A1 had overlaps.

All participants identified the initial hulls of all target marker groups (Figure 12b). They started to filter cells to reduce the overlaps. They adjusted the expression threshold and selected only cells that are highly differentially expressed. Based on the density plot, they created hulls consisting of cells with high local density.

Initially, P1 found that PPBP overlapped with CD14/LYZ, FCGR3A/MS4A7, and CD8A (see Figure 12c). After cell filtering, three overlaps were removed and the preservation ratio of PPBP and MS4A1 became 1 for all dimensionalities. Next, he modified the hull of CD14/LYZ to remove overlaps between CD14/LYZ and FCGR3A/MS4A7 (see Figure 12d). Also, the preservation ratio of FCGR3A/MS4A7 became 1 for dimensionalities higher than 5. Last, he modified the hull of GNLY/NKG7 to remove overlaps between GNLY/NKG7 and CD8A (see Figure 12e). The preservation ratio of GNLY/NKG7 and CD8A became 1 for dimensionalities higher than 2. P2 and P3 also performed similar cell filtering processes to modify hulls and remove all overlaps.

When hull modification is done, all participants checked the dimensionality highlighted by the yellow edge (i.e., yellow dimensionality) in the total row of the hull heatmap (see Figure 12f). Also, they compared the final hulls and cluster plots in this dimensionality to check whether the hulls accord with the clustering result (i.e., cells belonging to the same label in the cluster plot are grouped together with a hull). Finally, they selected their yellow dimensionality as the optimal one. P1 and P3 selected dimensionality 8 and P2 selected dimensionality 6.

P4 confirmed that both 6 and 8 are the optimal dimensionality because all target cell types were well-separated and all hulls accorded with the clustering result. The main reason for the mismatch between participants was due to different cell filtering. In dimensionality 6 (Supplemental Figure 7a), a few cells of the hull of FCER1A/CST3 were inside the hull of CD14/LYZ. Because the hulls of FCER1A/CST3 and CD14/LYZ did not overlap in dimensionality 8 (see Supplemental Figure 7b) and the overlapping region in dimensionality 6 was too small, P1 and P3 did not modify the hull of FCER1A/CST3. On the other hand, P2 removed this overlap, which resulted in the yellow dimensionality 6.

6.3.2 Myeloid-like Cells Dataset

After creating a heatmap, P1 started to analyze plots in dimensionality 31 (the yellow dimensionality). He identified that

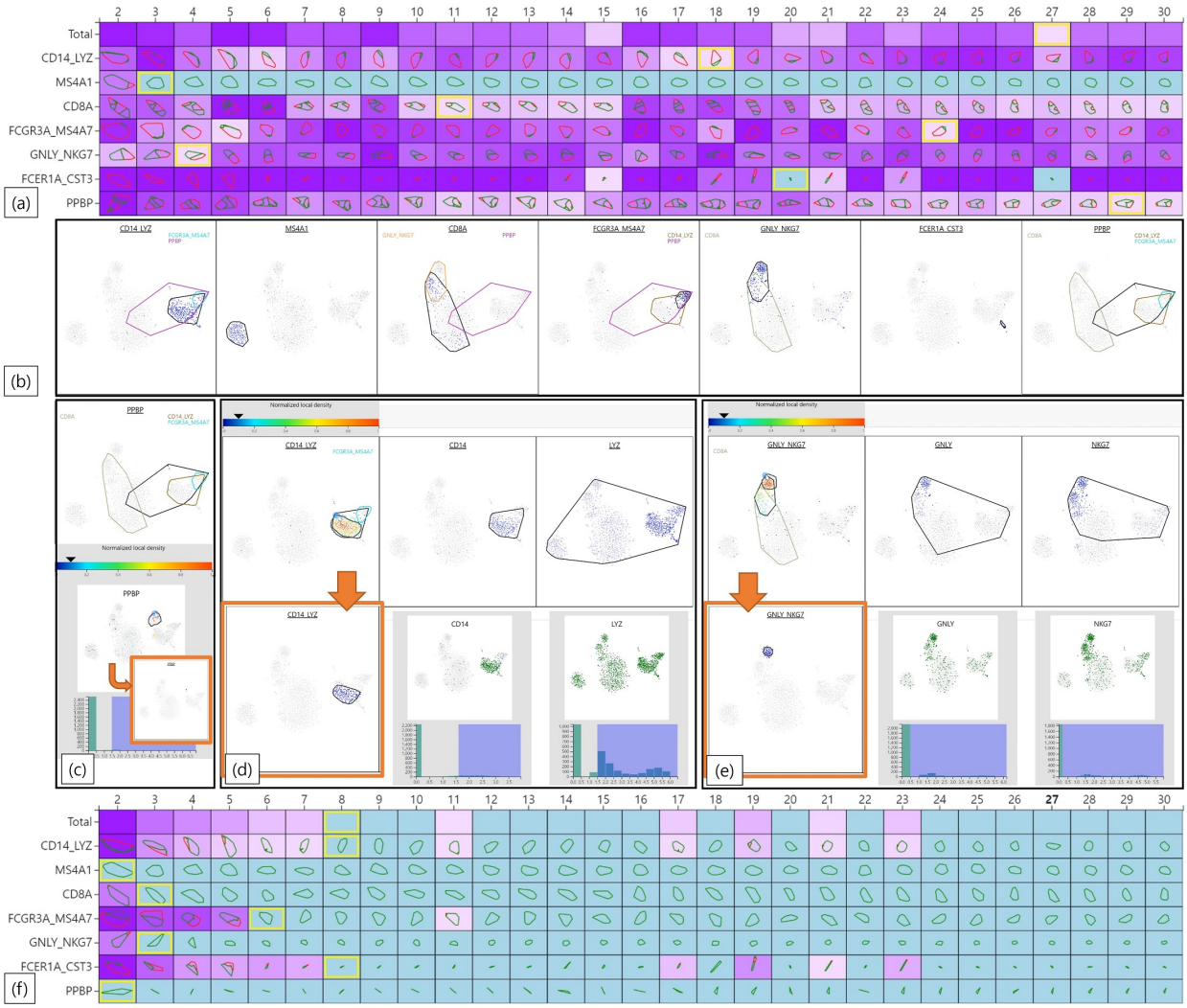

Figure 12: The overview of the first case study using the PBMC dataset. (a) The initial hull heatmap and (b) hulls of target groups. P1 modified the hull of (c) PPBP, (d) CD14/LYZ, and (e) GNLY/NKG7. (f) The final hull heatmap. See Supplemental Figure 1-6 for enlarged figures of it.

FCER1A/CD1C/CD1A/CD207 and CLEC9A/XCR1 had no overlaps while the other four groups had overlaps. P2 found that the preservation ratio of CLE9A/XCR1 is 1 for the dimensionalities larger than 16, so he started cell filtering in dimensionality 17 in which all groups except CLEC9A/XCR1 had overlaps. P3 started to analyze plots in dimensionality 7, which was highlighted by the yellow edge in the CD163/IFITM3 row. All groups overlapped each other in this dimensionality.

All participants continued to modify hulls and removed overlaps similarly as in the first study (with the PBMC dataset). Participants expressed difficulty in defining hulls in this study compared to the first study because the overlapping regions among target groups were larger. Therefore, when defining hulls, they referred to the cluster plot more frequently.

When P1 modified the hull of FCGR3A/CYTIP, they selected only one of two clusters with similar expression levels and density because the selected cluster (Supplemental Figure 8a) did not overlap with other hulls unlike the other one (Supplemental Figure 8b). He said he usually selected cell clusters with a smaller overlap with other types of cells even though the cell clusters have similar expression levels or local density. Owing to the overlapped hulls drawn on the marker expression plot, he was able to filter out another cell-type cluster without comparing other marker expression plots.

When P2 modified the hull of FCGR3A/CYTIP, FCGR3A and CYTIP were expressed in the middle of their marker expression plots. Thus, P2 thought that the target cells of FCGR3A/CYTIP would be in the middle when he reviewed the two expression plots individually. However, our system showed that there were few target cells in which both genes are differentially expressed in the middle of the plot (see Supplemental Figure 9), which helped P2 to find the cell type area more accurately.

After removing all overlapping regions in the hull, the yellow dimensionality of P1 was 8 and that of P2 and P3 was 7. By comparing the hulls and cluster plots in the selected yellow dimensionality, the participants discovered that FCER1A/CD1C/CD1A/CD207 and CLE9A/XCR1 were clustered together, and FCGR3A/CYTIP and S100A12/S100A9 were clustered together (see Supplemental Figure 10a). Thus, they tried to find the dimensionality in which all target groups are matched with different clusters in the cluster plot. Based on the light blue colors in the total row, the participants observed that no overlap happened in the dimensionalities larger than their yellow dimensionality. They compared the cluster plots by increasing the dimensionality by one at a time. They found that FCER1A/CD1C/CD1A/CD207 and CLE9A/XCR1 were included in different clusters in the dimensionality 9 or larger (see Supplemental Figure 10b), and FCGR3A/CYTIP and S100A12/S100A9 were included in different clusters starting in the dimensionality 15 or larger (see Supplemental Figure 10c). Therefore, all participants selected

dimensionality 15 as the optimal one, which is later confirmed by P4 as well.

6.3.3 T Cells Dataset

In the beginning, all groups had overlaps for all dimensionalities. P1 started to analyze plots in dimensionality 48 (the yellow dimensionality). P2 found that the heatmap color of TRDC/TRGC2 began to brighten from dimensionality 29, so he modified the hulls in that dimensionality. P3 started to analyze plots in dimensionality 5 because it was highlighted by the yellow edge in the CD3D/CD4 row. All participants continued to modify hulls and removed overlaps similarly as in the previous case studies.

After removing all overlaps, the yellow dimensionality of P1 and P3 was 33 and the yellow dimensionality of P2 was 29. By comparing the hulls and cluster plots, they observed that CD3D/CD8A and TRDC/TRGC2 were all clustered together. To find the dimensionality that separates these two groups in different clusters, they compared cluster plots, starting from dimensionality 3 and gradually increasing the dimensionality. Finally, P1 selected dimensionality 15, and P2 and P3 selected dimensionality 14, making CD3D/CD8A and TRDC/TRGC2 included in different clusters. Because the hulls were created by thresholding and lasso selection, small differences among the hulls of participants made them select different optimal dimensionalities. P4 confirmed that both dimensionality 14 and 15 can be optimal because there was little difference between them.

6.4 Overall Feedback

In the conventional analyses, the participants expressed that it was difficult to compare all dimensionalities because they were unable to review all plots simultaneously; they had to compare plots using their mental images. In addition, they faced difficulty in finding cells where cell markers were differentially expressed when cell colors looked similar in the marker expression plots. In general, they thought that it was difficult to present a final decision using this method. Furthermore, frequent user interactions requiring a switch between plots resulted in the participants becoming tired.

In the analyses using our method, the participants felt that the analysis using our method was more straightforward than the conventional analyses because the hull heatmap enabled them to quickly identify overlaps of the target cells of multiple dimensionalities in a single view, which reduced the range of exploration. The participants were confident in the analysis results because they can clearly watch the overlaps and set the expression level and density thresholds after checking the effects of the thresholds shown in the plots. Furthermore, the grouping of multiple markers of the same cell type enabled the participants to track the target cells without comparing multiple marker expression plots in each dimensionality. In the conventional analyses, the participants had to return to the initial step to filter certain cells and repeat data processing, including PCA, 2D projection, and clustering. In contrast, using our method, they were able to filter out certain cells during the exploration of the projection plots and examine the results in a few seconds.

7 Discussion & Limitation

In this work, we focused on selecting the optimal number of PCs for the t-SNE projection. Even though it is possible to use the raw input data directly for 2D projection using t-SNE, such a naive application of t-SNE does not work well because the high dimensionality (i.e., the large number of genes) of scRNA-seq data results in the distances between cells appearing similar; thus, global structures are not well preserved. Therefore, using PCs for the input of t-SNE is considered a standard process in single-cell analysis [19, 21].

T-SNE has multiple hyper-parameters, including perplexity, learning rate, and number of iterations. We observed that the changes in hyper-parameters did not result in noticeable differences in our experiments, and the default parameters worked well. We believe this is because the data size is not too big due to dimensionality reduction by PCA. Some literature reported that parameter adjustments might be required for extremely large data (e.g., over n$\gg$100,000) [21]; however, this is not the case in this study.

We observed that, for certain rare cell types, selecting the optimal number of PCs can be difficult and tricky in conventional methods because large changes occur in most significant PCs. However, less significant PCs may represent important information for such rare cell types. Our method allows analysts to easily identify position changes of target cells for a wide range of dimensionalities and even helps analysts identify less significant PCs where all target cell types are found.

After defining target cells, analysts reviewed the cluster plot to check whether the target cells formed a single cluster in the clustering result. When target cells are separated into multiple clusters, this implies that the cells can be divided into multiple sub-cell types. Because the clustering result changes depending on the dimensionality, analysts repeatedly check the cluster plot for several dimensionalities. Similar to the hull heatmap, highlighting the changes in clustering results can be helpful in determining the optimal dimensionality.

8 Conclusion & Future Work

In this paper, we introduced a novel visual analytics system, for interactively determining the optimal dimensionality for dimensionality reduction of the data for the task of cell type identification. By providing a novel visualization scheme, such as the hull heatmap and several interactive cell filtering methods, our method significantly reduces the effort required to review a large number of dimensionality reduction plots. We demonstrated the efficacy and usefulness of the proposed system through one user study and three case studies. In the future, we plan to extend our system using various 2D projection methods, including UMAP. The current system only supports the default t-SNE parameter; however, an in-depth analysis of the effect of both t-SNE parameters and PCA dimensionality would be an interesting future research direction.

Acknowledgments

This work was partially supported by the Bio & Medical Technology Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2019M3E5D2A01063819), the Basic Science Research Program through the NRF funded by the Ministry of Education (NRF-2021R1A6A1A13044830), a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (HI18C0316), the ICT Creative Consilience program (IITP-2023-2020-0-01819) of the Institute for Information & communications Technology Planning & Evaluation (IITP) funded by MSIT, and a Korea University Grant.

References

[1] A. Alavi, M. Ruffalo, A. Parvangada, Z. Huang, and Z. Bar-Joseph. scquery: a web server for comparative analysis of single-cell rna-seq data. *bioRxiv*, p. 323238, 2018.

[2] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. Tutorial: guidelines for the computational analysis of single-cell rna sequencing data. *Nature Protocols*, pp. 1–9, 2020.

[3] P. Angerer, L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, 12 2015. doi: 10.1093/bioinformatics/btv715

[4] D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing*, 6(1):128–143, 1985.

[5] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[6] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[7] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided visual clustering analysis. *IEEE transactions on visualization and computer graphics*, 25(1):267–276, 2018.

[8] L. Chen, W. Wang, Y. Zhai, and M. Deng. Deep soft k-means clustering with self-training for single-cell rna sequence data. *NAR genomics and bioinformatics*, 2(2):lqaa039, 2020.

[9] N. C. Chung and J. D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, 2015.

[10] M. Ciortan and M. Defrance. Gnn-based embedding for clustering scrna-seq data. *Bioinformatics*, 38(4):1037–1044, 2022.

[11] J. Ding, A. Condon, and S. P. Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, 9(1):2002, May 2018. doi: 10.1038/s41467-018-04368-5

[12] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. An interactive multidimensional data display and analysis system. Technical report, SLAC National Accelerator Lab., Menlo Park, CA (United States), 1974.

[13] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11), 2015.

[14] R. Hillje, P. G. Pelicci, and L. Luzi. Cerebro: interactive visualization of scrna-seq data. *Bioinformatics*, 36(7):2311–2313, 2020.

[15] T. Höllt, N. Pezzotti, V. van Unen, F. Koning, E. Eisemann, B. Lelieveldt, and A. Vilanova. Cytosplore: interactive immune cell phenotyping for large single-cell datasets. In *Computer Graphics Forum*, vol. 35, pp. 171–180. Wiley Online Library, 2016.

[16] T. Höllt, N. Pezzotti, V. van Unen, F. Koning, B. P. Lelieveldt, and A. Vilanova. Cyteguide: Visual guidance for hierarchical single-cell analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):739–748, 2017.

[17] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[18] N. Kim, H. K. Kim, K. Lee, Y. Hong, J. H. Cho, J. W. Choi, J.-I. Lee, Y.-L. Suh, B. M. Ku, H. H. Eum, et al. Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications*, 11(1):1–15, 2020.

[19] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[20] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[21] D. Kobak and P. Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14, 2019.

[22] A. Kulkarni, A. G. Anderson, D. P. Merullo, and G. Konopka. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current opinion in biotechnology*, 58:129–136, 2019.

[23] D. Lambrechts, E. Wauters, B. Boeckx, S. Aibar, D. Nittner, O. Burton, A. Bassez, H. Decaluwé, A. Pircher, K. Van den Eynde, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature medicine*, 24(8):1277–1289, 2018.

[24] E. Lin, S. Mukherjee, and S. Kannan. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell rna sequencing analysis. *BMC bioinformatics*, 21(1):1–11, 2020.

[25] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[26] L. A. Mitchell. Increment-interactive cluster refinement. 2016.

[27] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. Clustersculptor: A visual analytics tool for high-dimensional data. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pp. 75–82. IEEE, 2007.

[28] F. Raimundo, C. Vallot, and J.-P. Vert. Tuning parameters of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology*, 21(1):1–17, 2020.

[29] R. Rostom, V. Svensson, S. A. Teichmann, and G. Kar. Computational approaches for interpreting scrna-seq data. *FEBS letters*, 591(15):2213–2225, 2017.

[30] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[31] T. Sainburg, L. McInnes, and T. Q. Gentner. Parametric umap: learning embeddings with deep neural networks for representation and semi-supervised learning. *arXiv preprint arXiv:2009.12981*, 2020.

[32] J. Samir, S. Rizzetto, M. Gupta, and F. Luciani. Exploring and analysing single cell multi-omics data with vdjview. *BMC medical genomics*, 13(1):1–9, 2020.

[33] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[34] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, 2002.

[35] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[36] A. Srivastava, J. Zou, and C. Sutton. Clustering with a reject option: Interactive clustering as bayesian prior elicitation. *arXiv preprint arXiv:1602.06886*, 2016.

[37] B. Szubert, J. E. Cole, C. Monaco, and I. Drozdov. Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific reports*, 9(1):8914, June 2019. doi: 10.1038/s41598-019-45301-0

[38] M. Tabaka, J. Gould, and A. Regev. scsva: an interactive tool for big data visualization and exploration in single-cell omics. *bioRxiv*, p. 512582, 2019.

[39] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[40] S. K. Tasoulis, A. G. Vrahatis, S. V. Georgakopoulos, and V. P. Plagianakos. Visualizing high-dimensional single-cell rna-sequencing data through multiple random projections. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5448–5450, 2018.

[41] L. Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pp. 384–391. PMLR, 2009.

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] A. G. Vrahatis, S. K. Tasoulis, G. N. Dimitrakopoulos, and V. P. Plagianakos. Visualizing high-dimensional single-cell rna-seq data via random projections and geodesic distances. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–6. IEEE, 2019.

[44] A. G. Vrahatis, S. K. Tasoulis, I. Maglogiannis, and V. P. Plagianakos. Recent machine learning approaches for single-cell rna-seq data analysis. In *Advanced Computational Intelligence in Healthcare-7*, pp. 65–79. Springer, 2020.

[45] L. Waltman and N. J. Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B*, 86(11):1–14, 2013.

[46] D. Wang and J. Gu. Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, 16(5):320–331, 2018.

[47] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, and D. Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1–11, 2021.

[48] F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[49] Y. Wu, P. Tamayo, and K. Zhang. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. *Cell systems*, 7(6):656–666, 2018.

[50] G. X. Zheng, 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. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.