BootCellNet, a resampling-based procedure, promotes unsupervised identification of cell populations via robust inference of gene regulatory networks.

Yutaro Kumagai

Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology, Higashi 1-1-1, Tsukuba, Ibaraki 305-8565, Japan

Correspondence:

Yutaro Kumagai,

Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology, Higashi 1-1-1, Tsukuba, Ibaraki 305-8565, Japan.

E-mail: yutaro.kumagai@aist.go.jp.

Abstract

Recent advances in measurement technologies, particularly single-cell RNA sequencing (scRNA-seq), have revolutionized our ability to acquire large amounts of omics-level data on cellular states. As measurement techniques evolve, there has been an increasing need for data analysis methodologies, especially those focused on cell-type identification and inference of gene regulatory networks (GRNs). We have developed a new method named BootCellNet, which employs smoothing and resampling to infer GRNs. Using the inferred GRNs, BootCellNet further infers the minimum dominating set (MDS), a set of

genes that determines the dynamics of the entire network. We have demonstrated that BootCellNet robustly infers GRNs and their MDSs from scRNA-seq data and facilitates unsupervised cell labeling using scRNA-seq datasets of peripheral blood mononuclear cells and hematopoiesis. It has also identified COVID-19 patient-specific cells and their potential regulatory transcription factors. BootCellNet not only identifies cell types in an unsupervised and explainable way but also provides insights into the characteristics of identified cell types through the inference of GRNs and MDS.

Introduction

Recent advances in measurement technologies have revolutionized our ability to acquire large amounts of omics-level data. Among these technologies, single-cell RNA sequencing (scRNA-seq) has become particularly widespread. The technique allows for an unbiased investigation of the distribution and status of various cell types within a sample, providing insights into the dynamic and heterogeneous nature of cells.

As measurement techniques evolve, there has been an increasing need for data analysis methodologies[1]. One significant challenge is the definition of cell types and states from scRNA-seq data. This requires a balance between distinguishing different cell states and grouping similar ones together. In this regard, the problem essentially boils down to one of clustering. In the analysis of scRNA-seq, cells are clustered in an unsupervised manner by using various algorithms. Among those, *k*-means clustering, hierarchical clustering, spectral clustering, and Louvain clustering are examples of methods utilized frequently. However, these methods do not provide an answer for the appropriate number of clusters and/or an explanation of why some cells are clustered together but others are separated, leaving uncertainty in the analysis. An alternative to

unsupervised clustering is the use of supervised methods. Typically, cells are labeled by using an annotated dataset. However, this approach faces difficulties in identifying novel cell types not included in the given dataset.

The inference of gene regulatory networks (GRNs) from transcriptomic data like scRNA-seq represents another major challenge. Successful unraveling of the complex interactions and regulatory mechanisms that control gene expression within cells will offer profound insights into cellular functions and disease mechanisms. Numerous studies are dedicated to this field, employing a range of computational approaches[2]. Despite the efforts, GRN inference is still challenging, due to the noisy nature of scRNA-seq, characterized by dropouts and strong batch effects.

Once a GRN has been inferred, an essential next step is identifying genes that are crucial for the overall control and regulation of the network. Understanding which genes have a significant influence on the network can offer critical insights into the underlying biological processes and potential targets for therapeutic intervention[3]. There are several methods to explore these pivotal elements within GRNs, and one of the promising approaches relies on the concept of the minimum dominating set (MDS).

The MDS is a graph theoretical concept, defined as the smallest subset of vertices in a graph such that every other vertex is adjacent to at least one vertex in this subset. MDS plays a crucial role in controlling complex networks[4]. In the context of GRNs, this translates to a minimal set of genes whose regulatory influence extends across the entire network. In GRNs, MDS can be used to identify master regulatory genes[5]. In protein interaction networks, MDS proteins are found to be enriched with essential, cancer-related, and virus-targeted genes, showing a higher impact on network resilience than hub proteins[6].

In this study, we introduce BootCellNet, a method that employs smoothing and resampling procedures to reduce the noise in scRNA-seq data. We showed that this approach robustly infers GRN and MDS. Since GRN controls the whole transcriptome by definition, and MDS, in theory, governs the dynamics of the whole GRN, the expression pattern of genes in MDS may reflect the entire transcriptome. Leveraging this premise we utilized MDS for cell clustering. Because clusters are given based on the MDS, the method gives a rational explanation of why some cells are clustered together while others are not. Applying our method to multiple scRNA-seq datasets confirmed its effectiveness. This addresses the aforementioned challenges in GRN identification and clustering methods simultaneously.

Methods

Outline of the BootCellNet procedure.

The proposed procedure is outlined in Figure 1. The procedure, BootCellNet, utilizes smoothing of scRNA-seq expression data and resampling to reduce noise in the dataset and to facilitate a robust inference of gene regulatory network (GRN), in combination with computation of minimum dominating set (MDS). Using the MDS, BootCellNet performs clustering of cells to identify cell types and states.

Estimation of statistical significance by using resampling.

The nature of the BootCellNet procedure enables statistical evaluation of inferred GRNs. BootCellNet repeatedly samples cells from the smoothed scRNA-seq dataset and performs GRN inference. Some nodes (genes) and edges (regulatory interactions) frequently show up in each resampled set, while others do only occasionally. At the same

time, we shuffle gene names and perform the same inferences repeatedly to evaluate the occurrence of false positives.

We first performed the k-NN smoothing on 2-dimensional UMAP space. The k is set to a value small enough to find a cell population having about k cells in a dataset. In this regard, the k determines the resolution of the analysis.

Next, we resample N k-NN smoothed cells for M times to infer GRN or for $\widehat{M}$ times to calculate the support values of appearances of genes/nodes and regulations/edges. In the current study, the resampling number N was set to from one-third to half of the total number of cells in the dataset, but no more than 20000 to reduce the computation load. For each resampled set, we compute variable expression genes and then infer GRN. We also perform shuffling of gene names in each resampled set to calculate the support.

Let the resampled sets be $X_m = \{x_{m1}, x_{m2}, \dots, x_{mN}\}$, where $m = 1, 2, \dots, \widehat{M}$, and x be a vector of k-NN smoothed gene expressions of a cell. Shuffling of gene names in X_m for K times is performed to generate nested resampling sets X_{mk} $(k = 1, 2, \dots K)$[7]. Then we obtain $S(X_{mk}; \theta)$, which takes 1 if a hypothesis θ is supported by the set X_{mk} or 0 if not. In the current case, θ is a hypothesis that we find a gene as a variable expression gene or one that we find a regulatory relationship (i.e. an edge of GRN) in a GRN inferred from X_{mk}. The parameter to be estimated is $\lambda(\theta) = \frac{1}{\widehat{M}K} \sum_{m,k} S(X_{mk}; \theta)$, representing how often we occasionally have θ. Once we estimate the parameter, we can compute a p-value by supposing the event obeys a Poisson distribution with parameter λM. However, the rate of occasional appearances of genes is equal to (the number of variable expression genes selected) / (the number of total genes), and thus there's no need to compute by resampling. Hence, we computed the λ only for regulations/edges in GRNs. The obtained p-values were adjusted by the Benjamini-Yekutieli method to obtain q-

values. Nodes and edges with q<1e-05 were considered significant, and non-significant nodes and edges were removed.

In the implementation of this paper, resampling of cells was performed using a function named BootCellNet, and shuffling of gene names was carried out in a function named AggregateBCN. For inferring GRN, resampling was done with M = 100, but the shuffling of gene names was not performed (K = 0). In the estimation of the support, resampling was performed with $\widehat{M} = 1000$, and gene name shuffling was carried out with K = 100, obtaining 10^5 resampled specimens.

Inference of gene regulatory network.

For each resampled specimen, 2000 genes with variable expression levels were selected by using FindVariableFeature function in Seurat[8] with vst method. From the set of genes, we only focused on transcription factors (TFs). The list of TFs was retrieved from a reported list of human TFs[9] in the case of human dataset analysis, and from AnimalTFDB 4.0[10] in the case of mouse dataset analysis. The TFs with variable expression levels were subjected to GRN inference.

GRN is inferred by using ARACNe method[11] implemented in an R package minet[12] with the estimation of mutual information by Miller-Madow asymptotic bias corrected empirical estimator and an equal width discretization, as recommended in a report[13].

Computation of minimal dominating set (MDS).

The minimum dominating set (MDS) is a graph theoretical concept, which is the smallest subset of vertices in a graph such that every other vertex is adjacent to at least one vertex

in this subset. MDS can be computed exactly by using an integer programming formulation described elsewhere[14]. We implemented the method using the R package lpSolve. Although the problem is known as an NP-hard problem, the number of genes in the GRN is typically less than 200 and thus MDS is computable in a laptop computer.

Clustering of cell clumps.

In the BootCellNet procedure, we start from small clusters and aggregate them. The starting small clusters (here we call them clumps to distinguish them from the final clusters) were computed by using Louvain algorithm with a high resolution implemented in Seurat. The resolution was determined so that the size of the smallest clump was not very different from k, which is the parameter used in the smoothing.

After obtaining the cell clumps, expression levels of MDS genes were averaged over each cluster. Using the expression levels, the clumps are hierarchically clustered by using Ward's algorithm with a value of Pearson correlation subtracted from 1 as a distance. The number of clusters of clumps was determined by visual inspection.

Single-cell RNA-seq data processing.

The peripheral blood mononuclear cell (PBMC) dataset was provided by 10x Genomics (https://cf.10xgenomics.com/samples/cell-exp/3.1.0/manual_5k_pbmc_NGSC3_ch5/manual_5k_pbmc_NGSC3_ch5_filtered_feature_bc_matrix.tar.gz). Expression levels and metadata were aggregated and processed with a standard workflow by using Seurat. The processed data containing 3707 cells was then subjected to BootCellNet with $k = 10$ and $N = 3000$. The starting 63 cell clumps were given with a resolution of 10, which was determined according to the criterion described

above.

The myeloid/erythroid development scRNA-seq dataset[15] was retrieved from a repository (Paul_Cell_MARSseq_GSE72857.RData, https://github.com/theislab/scAnalysisTutorial). The data, comprised of 2730 cells, were processed with a standard workflow by using Seurat. The processed data were subjected to BootCellNet procedure with k = 10 and N = 2000. The starting 62 cell clumps were obtained with a resolution of 10.

The single-cell RNA-seq data of COVID-19 patients[16] were obtained from UCSC Cell Browser (http://cells.ucsc.edu/covid19-balf/nCoV.rds). The data is comprised of 66452 cells in total of bronchoalveolar lavage fluid from 4 healthy individuals, 3 mild and 6 severe patients. The data were analyzed by using Seurat and were subjected to BootCellNet with k = 20 and N = 20000. The starting 113 cell clumps were given by using Louvain clustering with a resolution of 10.

As a comparison, supervised cell type annotation was performed by using SingleR[17] with Human Primary Cell Atlas fine subtype labels in an R package celldex[17].

Transcription factor binding site (TFBS), Gene Ontology (GO) enrichment and lineage analyses.

Marker genes were obtained by using FindAllMarkers function in Seurat. We used Homer[18] for TFBS and GO analysis of the marker genes of each cluster. Promoters are defined as from -500 to +200 of transcription start site. For lineage analysis, Slingshot[19] was utilized with default parameters.

Statistical comparison of GRNs, MDSs, and clustering results.

Similarity between two GRNs or MDSs were compared by using Jaccard index. Adjusted Rand index[20] was calculated to compare two clustering results.

Results.

BootCellNet enables robust GRN inference.

One of the problems in GRN inference from scRNA-seq data is its noisy and unstable nature, making the inferred GRN unreliable. To overcome the noise, BootCellNet employs two procedures, smoothing of scRNA-seq data and resampling-shuffling of data to estimate *p*-values of genes and their regulations, as detailed in Methods (Figure 1).

To confirm the importance of each step, we first compared GRNs constructed with or without smoothing. GRNs were inferred 10 times each with or without smoothing, and appearances of genes/nodes and regulations/edges were counted (Figure 2A). As shown in Figure 2B, the number of appeared genes was almost the same in GRNs with smoothing as in those without smoothing before selection using *q*-values. After the selection of genes using *q*-values, however, the number of genes was higher in GRNs with smoothing. This result indicates that the smoothing procedure promotes the selection of consistent genes and stable inference of GRN.

The above result shown in Figure 2B also implies that the smoothing alone is sufficiently effective in extracting reliable genes/nodes because the resampling-based selection of significant genes/nodes did not alter the number of genes/nodes drastically. We examined the regulations/edges in the networks with or without resampling-based selection. As shown in Figure 2C, GRNs with a resampling-based selection have a higher burden of regulations/edges that appeared 10 times, and a lower burden of those that appeared only once, indicating that the selection efficiently enriches regulations/edges

with high consistency.

BootCellNet enables robust MDS inference.

Once we have robustly inferred GRN, we further infer how the GRN regulates gene expression by identifying MDS in the GRN. We implemented exact MDS computation based on an integer programming formulation. GRNs inferred above were subjected to the computation of MDS, and then the appearance of genes in each MDS was counted. As shown in Figure 3A, the resampling-based selection increased the number of genes in MDS that appeared in all 10 specimens. Moreover, the sizes of MDS in GRNs with selection were larger than those without selection (Figure 3B), despite the selection having reduced the number of genes/nodes and regulations/edges. Furthermore, in the GRNs with the selection, MDS genes with a low degree rank appeared more consistently than in the GRNs without the selection (Figure 3C). These results indicate that the resampling-based selection of nodes and edges leads to consistent GRN structures and robust inference of MDS.

BootCellNet determines cell type clusters in combination with the computation of MDS.

Since MDS, in theory, controls the dynamics of the whole network, the expression pattern of genes in MDS may represent the whole transcriptome. Thus, it is assumed that the expression pattern of genes in MDS can classify cells. To answer if GRN and MDS that are robustly inferred by the BootCellNet can facilitate rational cell type classification, we tested the idea on the PBMC scRNA-seq dataset. We first obtained small clusters of cells, by using Louvain algorithm with high resolution. The resolution

of 10 gave 63 clusters of cells (Figure 4A). Hereafter, we call these original clusters "clumps" to distinguish them from clusters obtained later.

Then we calculated the average expression of MDS genes in each cell clump and clustered the cell clumps as well as the genes by using hierarchical clustering with Ward's method (Figure 4B, left). By visual inspection of the heatmap of the MDS expression, we determine the number of clusters of clumps as 13 (Figure 4B, right). As a comparison, we also carried out the same by using a set of genes with variable expression instead of MDS (Figure 4C). Of 2000 variable genes, we sampled 100 genes (around three times the number of MDS, containing 36.0 genes on average; Figure 3B) and performed the same procedure as for MDS. We then applied the number of clusters of clumps as 13, as in the procedure for MDS. The clustering was compared with each other and adjusted Rand indexes (ARIs) were calculated. As shown in Figure 4D, MDS-based clustering had higher ARIs among the 10 specimens than did clustering based on 100 variable genes, indicating the robustness of MDS-based clustering.

To evaluate the obtained clusters, we picked one specimen shown in Figure 5A and checked the marker genes, Gene Ontology (GO) terms, and enriched TFBS in each cluster in the specimen. Marker genes, GO terms, and enriched TFBSs are listed in Supplementary Tables 1, 2, and 3, respectively. Cluster 1, 2, 3, and 9 exhibited characteristics of T cells. Cluster 1 is associated with the signature of CD4 T cells, whereas Cluster 2 is associated with that of CD8 T cells. Clusters 3 and 7 exhibit characteristics of activated CD4 and CD8 T cells, respectively. Cluster 8 is also associated with T cells, and based on the MDS expression pattern, it exhibits high RUNX3 expression, indicating the cluster is CD8 T cells[21]. Cluster 5 is associated with NK cell-like signatures. Cluster 4 shows B cell characteristics, marked by CD19 and

immunoglobulin genes. Clusters 6, 10, and 11 feature markers whose promoters are enriched with PU.1/SPI1 transcription factor binding site (TFBS), suggesting that these clusters represent myeloid cell populations. CLEC10A and CD1C genes emerged as markers of Cluster 10, pointing to myeloid dendritic cells (DCs)[22]. This aligns with the cluster's association with myeloid DC signatures. Cluster 12, marked with PDGFRA and GATA1, is linked to a platelet signature. Cluster 13, identified by SPIB, TCF4, and IRF7 as markers, all of which are well-known TFs involved in plasmacytoid DC (pDC) development and function[23]. Consistently, the cluster is associated with pDC gene signature, indicating that the cluster represents pDC population.

To compare this result with reference-based supervised cell labeling, the cells were labeled by using SingleR with Human Cell Atlas dataset in celldex R package. The main dataset, encompassing 37 cell types, labeled the PBMC scRNA-seq data with 7 cell types: T cells, B cells, monocytes, platelets, NK cells, CMP, GMP, and pre-B cells (data not shown). The fine dataset contains 157 subtypes, resulting in 19 cell types as shown in Figure 5B. Most labels are consistent with the result from BootCellNet. However, the reference-based labeling misclassified certain populations, such as clusters of myeloid DCs (Cluster 10) and pDCs (Cluster 13) as monocytes and CD34- GMP/Pre-B cells, respectively. Moreover, it tended to subdivide cells into smaller populations. NK cells, represented by Cluster 5 in MDS-based clustering, were subdivided into several clusters with similar markers, such as killer-cell receptors (KLRs) and granzymes (GZMs). The activated CD8 T cell Cluster 7 in MDS-based clustering was mislabeled as a CD4 T cell. B cells, represented by Cluster 4 in MDS-based clustering, were also divided into several subpopulations, most with nearly identical markers. A population, B cell with memory phenotype, has unique markers IGHA1/2, implying class switching which is a

characteristic of activated B cells. This implies that some populations might not be identified by MDS-based clustering.

BootCellNet identifies MDS in GRN. This provides insights into how the differentiation and/or the function of cell types are regulated. Comparing the B cell cluster 4 and the pDC cluster 13, TCF4, a TF important for pDC development[23], was expressed more in Cluster 13. Cluster 12, representing platelets, has a high expression of GFI1B which regulates platelet function[24]. We found CEBPB, a TF that controls monocyte/macrophage differentiation, was higher in Clusters 6 and 11 of monocyte/macrophage compared to Cluster 10 of DCs. Id2 is expressed more in the NK cell cluster 5, compared to Clusters 1, 2, 3, 7, 8, and 9, all of which are clusters of T cells. Consistently, the TF is essential for NK cell maturation[25]. Cluster 7 also shows high Id2 expression, thought to a lesser extent. This is consistent with reports that Id2 is important for CD8+ memory T cell development[26] because the cluster represents an activated CD8 T cell population. Further, by assessing RUNX3 expression, we determined Cluster 2, 7, and 8, which show higher RUNX3 expression, to be CD8 T cells, whereas Cluster 1, 3, and 9 are identified as CD4 T cells, given that RUNX3 is important for CD8 T cell development[21]. By examining RORA expression in Cluster 1 and 3, we identified Cluster 3 as activated T cells. RORA is notably expressed in Th17 population[27].

These findings collectively suggest that MDS-based clustering offers robust and accurate cell type identification, largely aligning with reference-based supervised labeling, while providing insights into the regulatory mechanisms of the cells through the identification of GRN and MDS.

BootCellNet facilitates inference of developmental courses of hematopoiesis.

The above result indicates that BootCellNet can identify cell types from scRNA-seq data containing multiple discrete cell types. To examine if BootCellNet can help to infer continuous spectra of cell states, we applied BootCellNet to scRNA-seq hematopoiesis data. The scRNA-seq data of myeloid cell development[15] was subjected to BootCellNet. The procedure resulted in a GRN of 110 TFs and MDS of 22 TFs (Figure 6A). The MDS includes TFs that are known for their importance in myeloid/erythroid development, such as Cebpe, Gata2, Gfi1b, Irf8, Jdp2, Klf1, Meis1, Myb, Sfpi1, and Sox4, further suggesting the robust inference GRN and its dominating TFs by BootCellNet.

The cells were clustered into 9 as shown in Figure 6B. By examining the expressed MDS TFs (Figure 6C), the marker genes (Supplementary Table 4), and associated GO terms (Supplementary Table 5), we identified Cluster 1, 4, 6, and 8 as stem cells, Cluster 2 as erythrocytes, Cluster 5 as megakaryocytes, Cluster 3 as monocytes, and Cluster 7 as neutrophils. Cluster 9 seems to be lymphocytes and dendritic cells, which are removed in the analyses performed in a report using the data[28].

Cluster 1 has Flt3 as a marker, as well as Sox4, Runx1, Cebpa, Lmo4, and Gata2, which are TFs known to be expressed in hematopoietic stem cells. Cluster 3, 4, and 7 are high in Sfpi1 expression, indicating that those clusters are myeloid lineages. However, the expression of Sox4 distinguishes Cluster 4 from others as more progenitor-like cells. Cluster 3 is associated with terms related to innate immune responses, and has Csf1r as a marker gene, as well as Irf8. Cluster 7 is associated with myeloid cell-related terms and has Fcnb and Elane genes as markers. The mentioned marker genes are characteristics of monocytic and neutrophilic cells, respectively[29], supporting the identification of Cluster 3 as monocytes and Cluster 7 as neutrophils. Comparing Cluster 3 and 7, the former is high in Irf8 expression, whereas the latter has higher expression of Jdp2 and Cebpe. This

expression pattern of TFs is also a characteristic that distinguishes monocytes and neutrophils[29,30]. The importance of higher expression of Cebpe in the neutrophil lineage has been suggested by using CellOracle, which is a cell lineage analysis method based on GRN and a mathematical dynamical system model[28]. The result demonstrates that BootCellNet efficiently infers MDS and their roles in development, without using mathematical dynamical system models.

We inferred the lineage trajectories by using the clusters identified by BootCellNet. Slingshot[19] was applied to the data, setting Cluster 1 as a starting point. This resulted in 4 lineage curves (Figure 6D; one is omitted, see the Figure Legend for details). The result recapitulates known developmental trajectories and aligns with the analysis performed by using a different method[28]. One trajectory represents the development of erythrocytes spanning from Cluster 1 to 2 through 8 and 6. The others represent that of monocytes and neutrophils spanning from Cluster 1, through 4, to Cluster 3 and 7, respectively. This result suggests that cell cluster identification by BootCellNet facilitates lineage trajectory inference.

Identification of characteristic cell types in COVID-19 patients by BootCellNet.

We applied BootCellNet to scRNA-seq data of bronchoalveolar lavage fluid cells from COVID-19 patients[16] to test whether BootCellNet can identify cell types in scRNA-seq data where discrete and continuous cell types/states are mixed. The cells were originally clustered into 32. BootCellNet gave the GRN of 161 TFs with the MDS of 27 TFs (Figure 7A) and we clustered the cell clumps into 19 (Figure 7B). Marker genes were identified for each cluster, and the GO terms as well as TFBS enrichment associated with the markers of each cluster were checked (Supplementary Tables 6, 7, and 8, respectively).

Clusters 1, 3, 4, 5, and 8 were associated with inflammatory immune responses and/or characteristics of macrophage/monocyte/DC thus identified as myeloid cells. Compared to the clusters in the original report (reproduced in Supplementary Figure 1), the clustering by BootCellNet has a smaller number of clusters while identifying myeloid cells with different characteristics. Clusters 1, 4, and 8 have high expression of FOS, IRF1, and NFKB2, which are TFs involved in inflammatory immune response (Figure 7C). The clusters mostly consist of cells from severe patients (Figure 7D). The associated terms are also related to inflammatory responses. Promoters of the marker genes were enriched with NF-kB and ISRE motifs, further supporting that the clusters are inflammatory myeloid populations. In contrast, Clusters 3 and 5 mostly consist of cells from mild and healthy individuals. Cluster 3 has HLA-DQ as a marker and is associated with DC/monocyte signatures, implying the cluster contains antigen-presenting myeloid cells. Cluster 5 has MARCO, CEBPA, and PPARG as markers and is associated with inflammatory macrophage signatures. This indicates that the cluster is alveolar macrophages since PPARG is an essential TF for alveolar macrophage development[31]. Cluster 2, which is in the vicinity of those myeloid cell clusters, is also associated with DC maturation and MHC class II signatures and has MARCO, HLA-DRA, and PPARG as marker genes. Although the cluster is distant from the mentioned clusters in the tree of the hierarchical clustering based on MDS gene expression (Figure 7C), these results suggest that the cluster is also an antigen-presenting myeloid population. Clusters 2, 3, and 5 did not show high expression of inflammatory TFs such as FOS, IRF1, and NFKB2, in contrast to Clusters 1, 4, and 8. This implies that the myeloid cells in healthy individuals or patients with mild symptoms are less inflammatory than those in individuals with severe disease, consistent with former reports showing that inflammation and cytokine storms are the

cause of severe COVID-19[32].

Clusters 9, 13, and 16 are other clusters with cells from severe patients. Cluster 6 is in the vicinity of those clusters both on the UMAP plane and in the hierarchical clustering tree but has fewer cells from severe patients. Cluster 13 has genes for gamma immunoglobulin, suggesting the cluster is a plasma cell population. Clusters 6, 9, and 16 are associated with T cell signature and have genes for components of T cell receptors such as CD3D/E/G, LCK, LAT, and ZAP70 as markers. They also have CD8A/B and RUNX3 as markers, indicating that those clusters are CD8 T cell populations. Further, all of them have genes for granzymes and IFN-g as markers, suggesting that the clusters represent activated CD8 T cells.

To examine the differences among those activated T cells clustered separately, we compared Cluster 6 to Clusters 9 and 16 and found that the former but not the latter has ID2, TCF7, and CCR5 as markers, which are highly expressed in effector memory CD8 T cells[26,33,34]. Clusters 9 and 16 have markers for activation and exhaustion. Specifically, when we examined IL7R (encoding IL-7Ralpha) and PDCD1 (encoding PD-1), Cluster 6 has both, while Cluster 9 has only PDCD1 and Cluster 16 has only IL7R. This indicates that Cluster 9 is exhausted T cells and Cluster 16 is activated naïve T cells[35]. These results collectively suggest that induction of effector memory CD8 T cell upon SARS-CoV-2 infection is critical for controlling disease to mild symptoms, and an increase of exhausted T cells is detrimental for disease control, consistent with former reports[16,36,37].

Clusters 12 and 17 are rich in cells from patients with mild symptoms. Cluster 12 has CD1C, CLEC9A, CLEC10A, and HLA-related genes as markers and promoters of the marker genes are enriched with STAT and NF-kB binding motifs, suggesting that

the cluster is an activated dendritic cell population. However, CD1C and CLEC10A are markers for cDC2, while CLEC9A is a marker for cDC1[22], implying that those DCs are mixed in the cluster. Cluster 17 has immunoglobulin genes as well as some known pDC markers, implying that the cluster is a mixture of B cells and pDCs. These results suggest the importance of DC populations and B cells in the control of COVID-19.

Taken together, the results indicate that BootCellNet successfully identifies clusters of cells specific to patients with different symptoms, and offers explanations on the difference in symptoms that are consistent with preceding reports.

Discussion.

In the current study, a method named BootCellNet has been developed, which combines smoothing, resampling, GRN inference, computation of a set of genes controlling the entire network, and clustering.

Since the method consists of multiple steps, the choice of algorithms will affect the results of the procedure. We used *k*-NN smoothing to reduce the noise in scRNA-seq. We performed the smoothing on the 2-dimensional UMAP plane, but one can do the same in a higher dimension[38]. One can also use different smoothing methods, such as *k*-NN graph-based MetaCell algorithm[39], or data imputation methods such as MAGIC[40] and DeepImpute[41].

In BootCellNet, GRN reconstruction is performed with the ARACNe method, because of its low computational load. As in the current study, NestBoot utilizes a nested bootstrap to control FDR in GRN inference[42], and they showed that the bootstrapping procedure improved the accuracy of the GRN inference by various inference methods such as GENIE3[43]. Thus, it is possible that the incorporation of methods other than

ARACNe in BootCellNet will improve the GRN inference accuracy.

The use of MDS as a set of genes that controls the dynamics of the whole network needs some caution. First, MDS is not necessarily uniquely determined. In the case where multiple MDSs exist, one can determine critical nodes and intermittent nodes, which appear in all the MDSs or at least in one MDS[44]. A set of critical and intermittent nodes is uniquely determined. It may be possible to improve the clustering results and their explainability by using these uniquely determinable sets. Second, the controllability of a network by its MDS requires a premise that each MDS node can control its adjacent nodes individually[4]. In the case where nodes are TFs, this may be justified by assuming that a TF controls each target gene through individually controlled accessibility, although the assumption should be verified independently.

Nevertheless, we demonstrated that the utilization of expression levels of only several tens of MDS genes was sufficient for robust clustering of cells while using one hundred genes with variable expression was not. This implies that MDS can provide a dimensionality reduction of scRNA-seq data.

In the current study, we inferred the network of TF genes, because we used transcriptome data, which is principally controlled by TFs. But this could be extended by incorporating not only TFs, but also co-factors, post-transcriptional regulators, and soluble mediators that control TF expression and/or activity. The incorporation of these factors also conforms to the picture where transcriptome activity is controlled principally on the mRNA level. Indeed, some post-transcriptional regulators, such as ZFP36L1 and ZFP36L2 were included and found to be a part of MDS. The incorporation of more factors will help to uncover dynamic picture of the regulatory network.

The last clustering step also has room for improvement. We utilized hierarchical

clustering with the number of clusters determined by visual inspection in this study. The number can be determined without visual inspection, for example, by using multiscale bootstrapping, implemented in an R package pvclust, which provides approximately unbiased confidence for each clustering[45]. This may in addition offer an automated determination of the number of clusters if we provide a cutoff *p*-value in advance.

Despite the room for improvement and caution, and although the method is relatively simpler than other sophisticated methods, we demonstrated that BootCellNet effectively worked for a variety of datasets with different characteristics, namely, the PBMC dataset containing multiple discrete cell types, the hematopoiesis dataset containing continuous cell states, and the dataset of lung BAL cells from COVID-19 patients containing multiple cell types from multiple individuals. BootCellNet can be a useful complementary for other methods of clustering, supervised labeling, and lineage inference.

Also, BootCellNet is suitable for automation. Once automated as discussed above, the method will be utilized for automatic and routine annotation of large-scale scRNA-seq cell data. Moreover, BootCellNet addresses an increasing need for explainable automatic data processing[46], since it has a firm basis in statistics and network dynamics. Although there is a strict need for manual inspection and experimental verification[1], the proposed method will accelerate the generation of the working hypotheses from a large amount of scRNA-seq data.

Author Contributions

Conceptualization, algorithm development and coding, data analysis, funding acquisition, manuscript preparation; Y. K.

Competing Interests

None declared.

Data availability

The code for BootCellNet is deposited in the author's GitHub (https://github.com/oursvallee/BootCellNet). All the data used are publicly available as described in Methods.

Acknowledgment.

The authors thank Drs. Yasuyuki Kida, Nobuhito Mori, and Kentaro Kawata for discussions, Mr. Akihiko Fujii for technical assistance, and Dr. Yasuko Ozaki for secretarial assistance. This research has partly been supported by Kakenhi 22H04925 (PAGS) and 23H04486, and by a Grant from International Joint Usage/Research Center, the Institute of Medical Science, the University of Tokyo. Computational time was in part provided by the NIG supercomputer at ROIS National Institute of Genetics.

References

1. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lucken, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. Nat Rev Genet *24*, 550-572. 10.1038/s41576-023-00586-w.
2. Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief Bioinform *22*. 10.1093/bib/bbaa190.
3. Guo, W.F., Zhang, S.W., Zeng, T., Akutsu, T., and Chen, L. (2020). Network control principles for identifying personalized driver genes in cancer. Brief

Bioinform *21*, 1641-1662. 10.1093/bib/bbz089.

4. Nacher, J.C., and Akutsu, T. (2012). Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. New Journal of Physics *14*, 073005.
5. Nazarieh, M., Wiese, A., Will, T., Hamed, M., and Helms, V. (2016). Identification of key player genes in gene regulatory networks. BMC Syst Biol *10*, 88. 10.1186/s12918-016-0329-5.
6. Wuchty, S. (2014). Controllability in protein interaction networks. Proc Natl Acad Sci U S A *111*, 7156-7160. 10.1073/pnas.1311231111.
7. Hinkley, D., and Shi, S. (1989). Importance sampling and the nested bootstrap. Biometrika *76*, 435-446.
8. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.
9. Joung, J., Ma, S., Tay, T., Geiger-Schuller, K.R., Kirchgatterer, P.C., Verdine, V.K., Guo, B., Arias-Garcia, M.A., Allen, W.E., Singh, A., et al. (2023). A transcription factor atlas of directed differentiation. Cell *186*, 209-229 e226. 10.1016/j.cell.2022.11.026.
10. Shen, W.K., Chen, S.Y., Gan, Z.Q., Zhang, Y.Z., Yue, T., Chen, M.M., Xue, Y., Hu, H., and Guo, A.Y. (2023). AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Res *51*, D39-D45. 10.1093/nar/gkac907.
11. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7 Suppl 1*, S7. 10.1186/1471-2105-7-S1-S7.
12. Meyer, P.E., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics *9*, 461. 10.1186/1471-2105-9-461.
13. de Matos Simoes, R., and Emmert-Streib, F. (2011). Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS One *6*, e29279. 10.1371/journal.pone.0029279.
14. Nacher, J.C., and Akutsu, T. (2013). Analysis on critical nodes in controlling complex networks using dominating sets. (IEEE), pp. 649-654.
15. Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H.,

Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell *163*, 1663-1677. 10.1016/j.cell.2015.11.013.

16. Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med *26*, 842-844. 10.1038/s41591-020-0901-9.
17. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol *20*, 163-172. 10.1038/s41590-018-0276-y.
18. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589. 10.1016/j.molcel.2010.05.004.
19. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477. 10.1186/s12864-018-4772-0.
20. Hubert, L., and Arabie, P. (1985). Comparing partitions. Journal of classification *2*, 193-218.
21. Egawa, T., Tillman, R.E., Naoe, Y., Taniuchi, I., and Littman, D.R. (2007). The role of the Runx transcription factors in thymocyte differentiation and in homeostasis of naive T cells. J Exp Med *204*, 1945-1957. 10.1084/jem.20070133.
22. Collin, M., and Bigley, V. (2018). Human dendritic cell subsets: an update. Immunology *154*, 3-20. 10.1111/imm.12888.
23. Reizis, B. (2019). Plasmacytoid Dendritic Cells: Development, Regulation, and Function. Immunity *50*, 37-50. 10.1016/j.immuni.2018.12.027.
24. Stevenson, W.S., Morel-Kopp, M.C., Chen, Q., Liang, H.P., Bromhead, C.J., Wright, S., Turakulov, R., Ng, A.P., Roberts, A.W., Bahlo, M., and Ward, C.M. (2013). GFI1B mutation causes a bleeding disorder with abnormal platelet function. J Thromb Haemost *11*, 2039-2047. 10.1111/jth.12368.
25. Li, Z.Y., Morman, R.E., Hegermiller, E., Sun, M., Bartom, E.T., Maienschein-Cline, M., Sigvardsson, M., and Kee, B.L. (2021). The transcriptional repressor ID2 supports natural killer cell maturation by controlling TCF1 amplitude. J Exp Med *218*. 10.1084/jem.20202032.
26. Cannarile, M.A., Lind, N.A., Rivera, R., Sheridan, A.D., Camfield, K.A., Wu,

B.B., Cheung, K.P., Ding, Z., and Goldrath, A.W. (2006). Transcriptional regulator Id2 mediates CD8+ T cell immunity. Nat Immunol *7*, 1317-1325. 10.1038/ni1403.

27. Yang, X.O., Pappu, B.P., Nurieva, R., Akimzhanov, A., Kang, H.S., Chung, Y., Ma, L., Shah, B., Panopoulos, A.D., Schluns, K.S., et al. (2008). T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. Immunity *28*, 29-39. 10.1016/j.immuni.2007.11.016.
28. Kamimoto, K., Stringa, B., Hoffmann, C.M., Jindal, K., Solnica-Krezel, L., and Morris, S.A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. Nature *614*, 742-751. 10.1038/s41586-022-05688-9.
29. Kwok, I., Becht, E., Xia, Y., Ng, M., Teh, Y.C., Tan, L., Evrard, M., Li, J.L.Y., Tran, H.T.N., Tan, Y., et al. (2020). Combinatorial Single-Cell Analyses of Granulocyte-Monocyte Progenitor Heterogeneity Reveals an Early Uni-potent Neutrophil Progenitor. Immunity *53*, 303-318 e305. 10.1016/j.immuni.2020.06.005.
30. Maruyama, K., Fukasaka, M., Vandenbon, A., Saitoh, T., Kawasaki, T., Kondo, T., Yokoyama, K.K., Kidoya, H., Takakura, N., Standley, D., et al. (2012). The transcription factor Jdp2 controls bone homeostasis and antibacterial immunity by regulating osteoclast and neutrophil differentiation. Immunity *37*, 1024-1036. 10.1016/j.immuni.2012.08.022.
31. Schneider, C., Nobs, S.P., Kurrer, M., Rehrauer, H., Thiele, C., and Kopf, M. (2014). Induction of the nuclear receptor PPAR-gamma by the cytokine GM-CSF is critical for the differentiation of fetal monocytes into alveolar macrophages. Nat Immunol *15*, 1026-1037. 10.1038/ni.3005.
32. Vanderbeke, L., Van Mol, P., Van Herck, Y., De Smet, F., Humblet-Baron, S., Martinod, K., Antoranz, A., Arijs, I., Boeckx, B., Bosisio, F.M., et al. (2021). Monocyte-driven atypical cytokine storm and aberrant neutrophil activation as key mediators of COVID-19 disease severity. Nat Commun *12*, 4117. 10.1038/s41467-021-24360-w.
33. Zhou, X., Yu, S., Zhao, D.M., Harty, J.T., Badovinac, V.P., and Xue, H.H. (2010). Differentiation and persistence of memory CD8(+) T cells depend on T cell factor 1. Immunity *33*, 229-240. 10.1016/j.immuni.2010.08.002.
34. Kohlmeier, J.E., Miller, S.C., Smith, J., Lu, B., Gerard, C., Cookenham, T., Roberts, A.D., and Woodland, D.L. (2008). The chemokine receptor CCR5 plays a key role in the early memory CD8+ T cell response to respiratory virus infections. Immunity *29*, 101-113. 10.1016/j.immuni.2008.05.011.

35. Seo, W., Jerin, C., and Nishikawa, H. (2021). Transcriptional regulatory network for the establishment of CD8(+) T cell exhaustion. Exp Mol Med *53*, 202-209. 10.1038/s12276-021-00568-0.
36. Odak, I., Barros-Martins, J., Bosnjak, B., Stahl, K., David, S., Wiesner, O., Busch, M., Hoeper, M.M., Pink, I., Welte, T., et al. (2020). Reappearance of effector T cells is associated with recovery from COVID-19. EBioMedicine *57*, 102885. 10.1016/j.ebiom.2020.102885.
37. Wauters, E., Van Mol, P., Garg, A.D., Jansen, S., Van Herck, Y., Vanderbeke, L., Bassez, A., Boeckx, B., Malengier-Devlies, B., Timmerman, A., et al. (2021). Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. Cell Res *31*, 272-290. 10.1038/s41422-020-00455-9.
38. Wagner, F., Yan, Y., and Yanai, I. (2017). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. BioRxiv, 217737.
39. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol *20*, 206. 10.1186/s13059-019-1812-2.
40. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716-729 e727. 10.1016/j.cell.2018.05.061.
41. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L.X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol *20*, 211. 10.1186/s13059-019-1837-6.
42. Morgan, D., Tjarnberg, A., Nordling, T.E.M., and Sonnhammer, E.L.L. (2019). A generalized framework for controlling FDR in gene regulatory network inference. Bioinformatics *35*, 1026-1032. 10.1093/bioinformatics/bty764.
43. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*. 10.1371/journal.pone.0012776.
44. Nacher, J.C., and Akutsu, T. (2014). Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. Journal of Complex Networks *2*, 394-412. 10.1093/comnet/cnu029.
45. Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and

evolution *16*, 1114.

46. Occhipinti, A., Verma, S., Doan, L.M.T., and Angione, C. (2024). Mechanism-aware and multimodal AI: beyond model-agnostic interpretation. Trends Cell Biol *34*, 85-89. 10.1016/j.tcb.2023.11.002.

Figure Legends.

Figure 1. The BootCellNet procedure. Starting from scRNA-seq data (top left), *k*-NN smoothing is carried out to reduce noise. From the smoothed data, *N* cells are sampled and GRNs are computed *M* times. Parallelly, sampling and GRN construction are performed $\widehat{M}$ times with shuffling gene names *K* times, resulting in $\widehat{M}K$ shuffled specimens. Using the shuffled data, statistically significant genes and their regulatory relationships are selected. The resulting GRN is subjected to the computation of MDS. At the same time, cells in the original scRNA-seq data are clustered into small clumps by using Louvain clustering with a high resolution. Averaged expression levels of genes in MDS are calculated and used for the clustering of cell clumps. The resulting clusters are regarded as cell types.

Figure 2. Smoothing and resampling in BootCellNet are important for the robust selection of genes and regulations. (A) Schematics of the computer experiment. (B) Numbers of genes/nodes in GRNs generated 10 times with or without smoothing and with or without selection. (C) Distribution of number of appearances of edges in 10 specimens, with (red) or without (gray) resampling-based selection.

Figure 3. BootCellNet robustly identifies MDS. (A) Distribution of the number of the appearance of genes/nodes in 10 MDS specimens. (B) Numbers of genes in MDS with

(red) or without (gray) selection. (C) Relationship between rank of mean degree and number of appearances in 10 MDS specimens of each gene.

Figure 4. BootCellNet robustly identifies cell types. (A) Clumps of cells by using Louvain clustering with a resolution of 10. (B) GRN and its MDS (red squares) identified by BootCellNet. (C) An example of the heatmap representation of hierarchical clustering of cell clumps shown in panel A (left) and the resulting clusters (right) based on MDS. (D) An example of the heatmap of hierarchical clustering of cell clumps (left) and the resulting clusters (right) based on 100 genes with variable expression. (E) Adjusted Rand indexes calculated among 10 MDS-based clusterings (left) and 10 variable gene-based ones (right) were shown.

Figure 5. Cell types in peripheral blood mononuclear cells identified by BootCellNet and its comparison to reference-based labeling. (A) The resulting clusters obtained by BootCellNet. Annotations for each cluster described in the main text were shown. (B) Annotation of cells obtained by SingleR with the fine HumanCellAtlas dataset. (C) Heatmap of average expression of MDS genes in each cluster obtained by BootCellNet. Color codes for columns are the same as those in panel A.

Figure 6. BootCellNet facilitates the inference of the course of hematopoiesis. (A) GRN and its MDS (red squares) identified by BootCellNet. (B) The resulting cell clusters. (C) Heatmap of average expression of MDS genes in each cluster obtained by BootCellNet. Color codes for columns are the same as those in panel B. (D) Lineage curves were obtained by Slingshot using the clusters by BootCellNet. One lineage curve,

which spans from Cluster 1 to Cluster 9 is omitted, since Cluster 9 encompasses lymphoid cells.

Figure 7. Identification of cell types in bronchoalveolar lavage fluid scRNA-seq data by BootCellNet. (A) GRN and its MDS (red squares) identified by BootCellNet. (B) The resulting cell clusters. (C) Heatmap of average expression of MDS genes in each cluster obtained by BootCellNet. Color codes for columns are the same as those in panel B. (D) Distribution of cells from each class of patients in each cluster.

Supplementary Figure 1. The original clustering result of BAL scRNA-seq in Liao et al. The result was reproduced from the deposited data and is shown (left). As a comparison, the MDS-based clustering (same as Figure 7B) is also shown.

Supplementary Table 1. Marker genes of each MDS-based cluster of PBMCs. The output of FindAllMarkers function in Seurat was printed as a table.

Supplementary Table 2. Gene Ontology terms for marker genes of each MDS-based cluster of PBMCs. Each sheet corresponds to each cluster.

Supplementary Table 3. TFBSs enriched in promters of marker genes of each MDS-based cluster of PBMCs. Each sheet corresponds to each cluster.

Supplementary Table 4. Marker genes of each MDS-based cluster of cells undergoing hematopoiesis. The output of FindAllMarkers function in Seurat was

printed as a table.

Supplementary Table 5. Gene Ontology terms for marker genes of each MDS-based cluster cluster of cells undergoing hematopoiesis. Each sheet corresponds to each cluster.

Supplementary Table 6. Marker genes of each MDS-based cluster of BAL cells from COVID-19 patients. The output of FindAllMarkers function in Seurat was printed as a table.

Supplementary Table 7. Gene Ontology terms for marker genes of BAL cells from COVID-19 patients. Each sheet corresponds to each cluster.

Supplementary Table 8. TFBSs enriched in promters of marker genes of BAL cells from COVID-19 patients. Each sheet corresponds to each cluster.

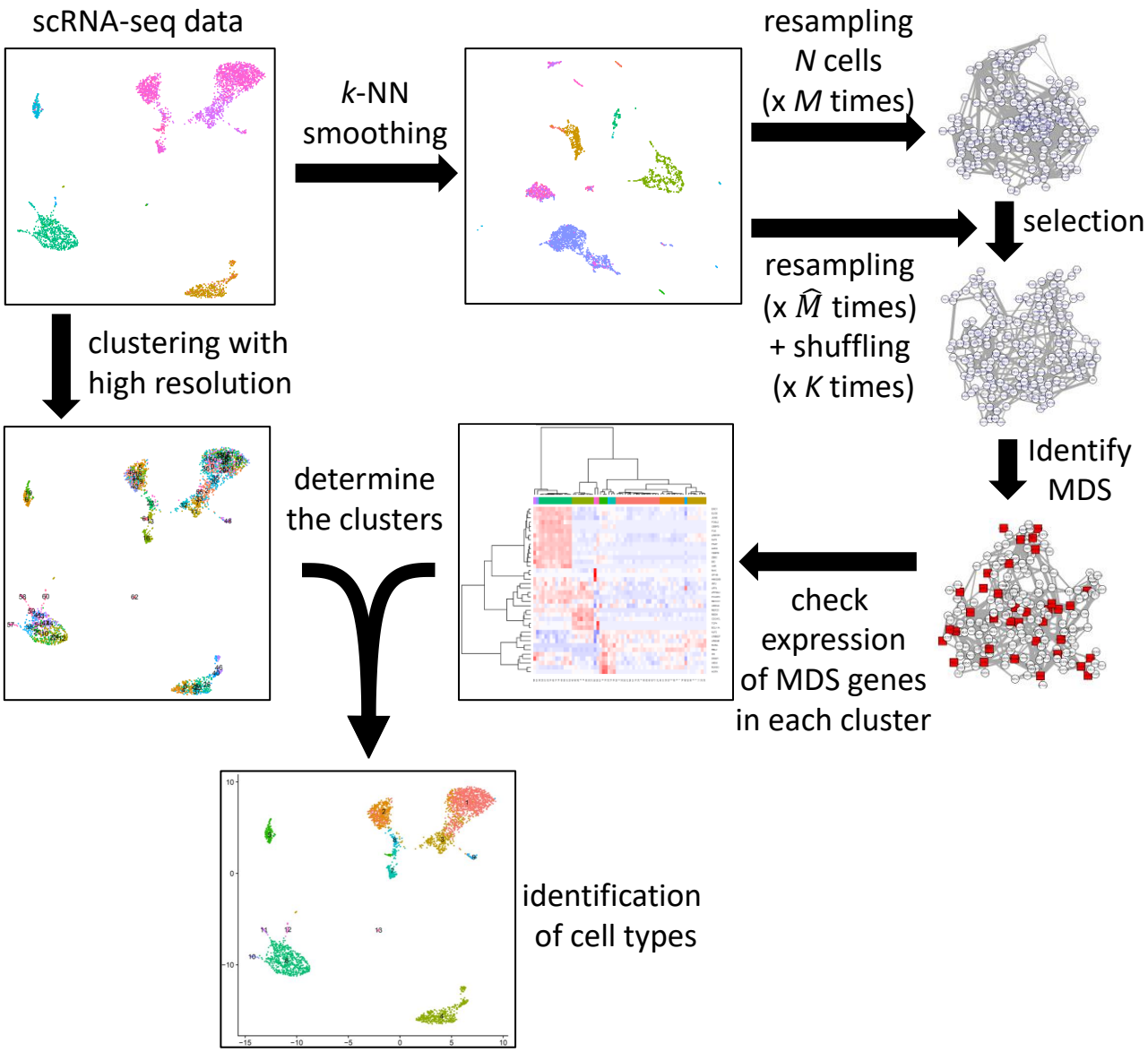
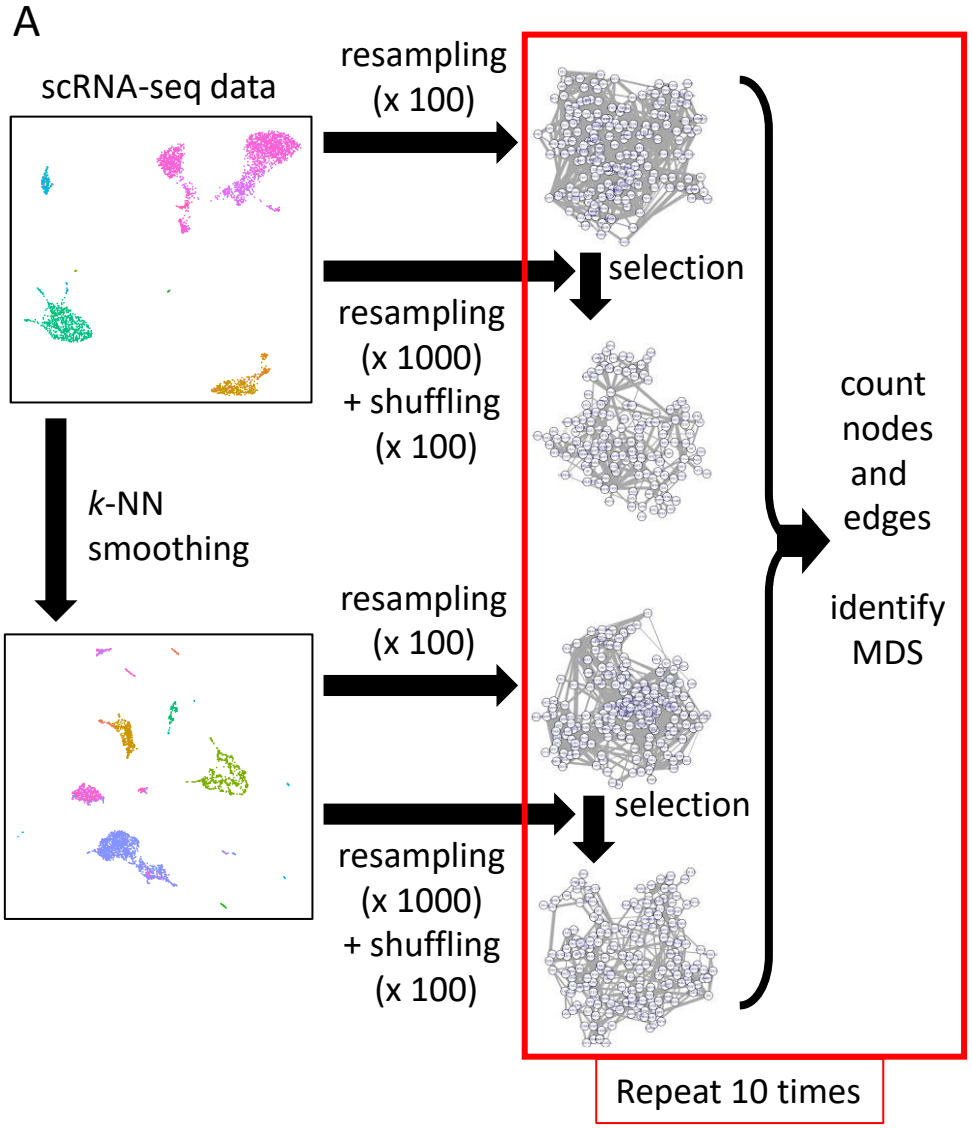
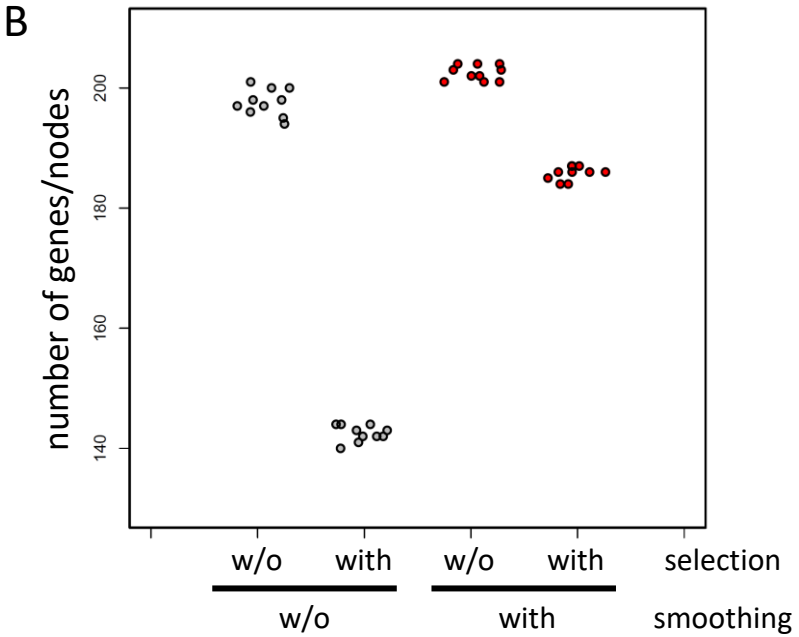

scRNA-seq data
k-NN
smoothing
resampling
N cells
(x M times)
selection
resampling
(x $\widehat{M}$ times)
+ shuffling
(x K times)
clustering with
high resolution
Identify
MDS
determine
the clusters
check
expression
of MDS genes
in each cluster
identification
of cell types

Figure 1


A
scRNA-seq data
resampling
(x 100)
selection
resampling
(x 1000)
+ shuffling
(x 100)
k-NN
smoothing
resampling
(x 100)
selection
resampling
(x 1000)
+ shuffling
(x 100)
count
nodes
and
edges
identify
MDS
Repeat 10 times


B
number of genes/nodes
140
160
180
200
w/o
with
w/o
with
selection
w/o
with
smoothing

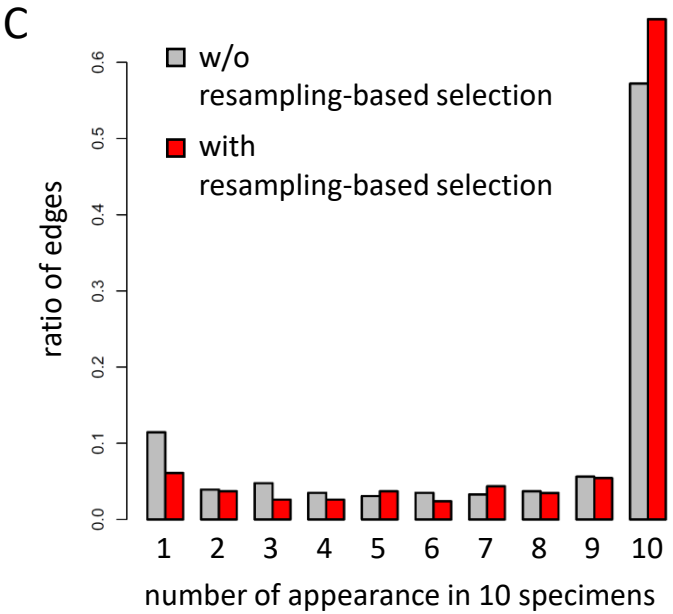

C
w/o
resampling-based selection
with
resampling-based selection
ratio of edges
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
number of appearance in 10 specimens

Figure 2

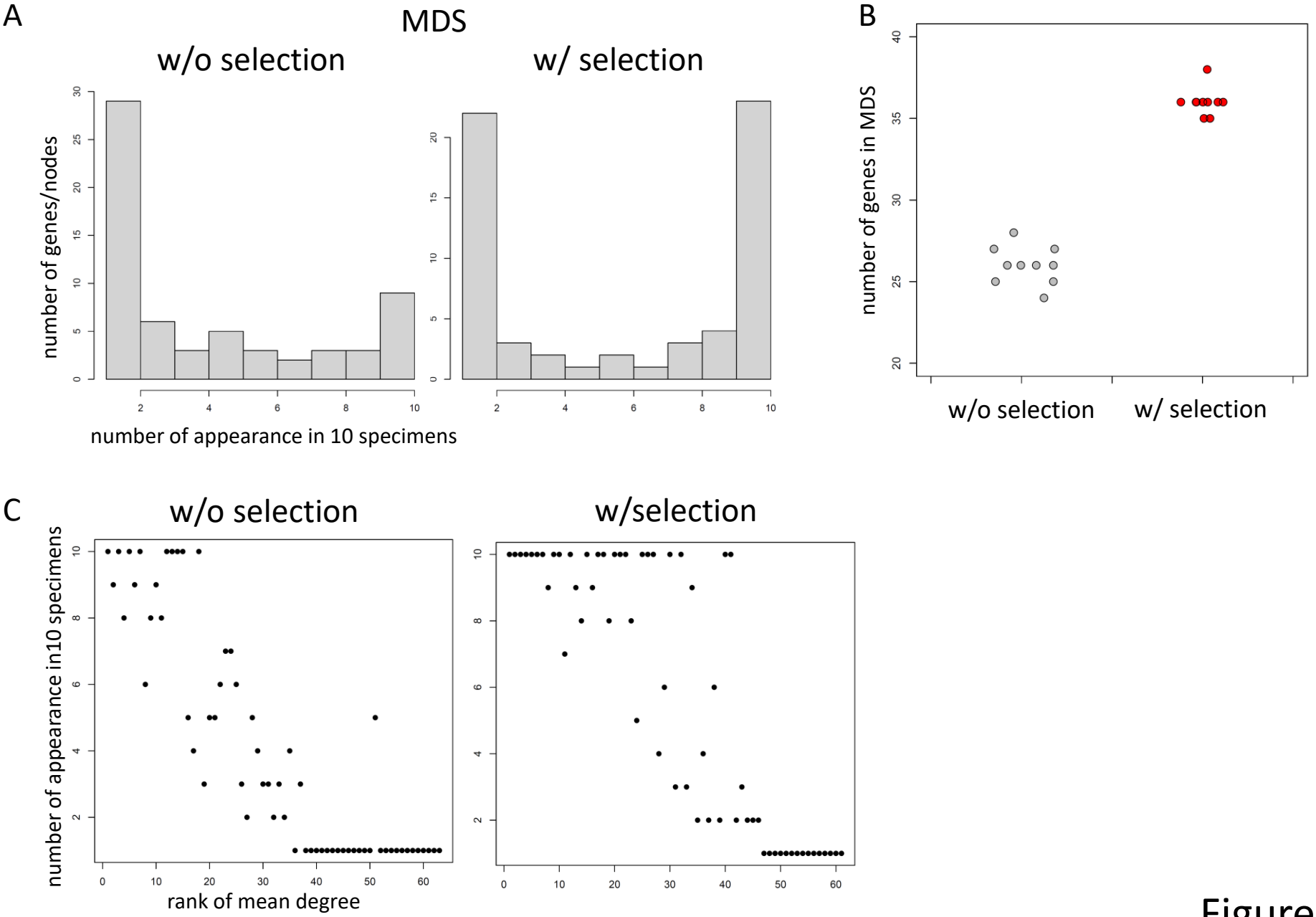

A
MDS
w/o selection
w/ selection
number of genes/nodes
number of appearance in 10 specimens
B
number of genes in MDS
w/o selection
w/ selection
C
w/o selection
w/selection
number of appearance in10 specimens
rank of mean degree

Figure 3

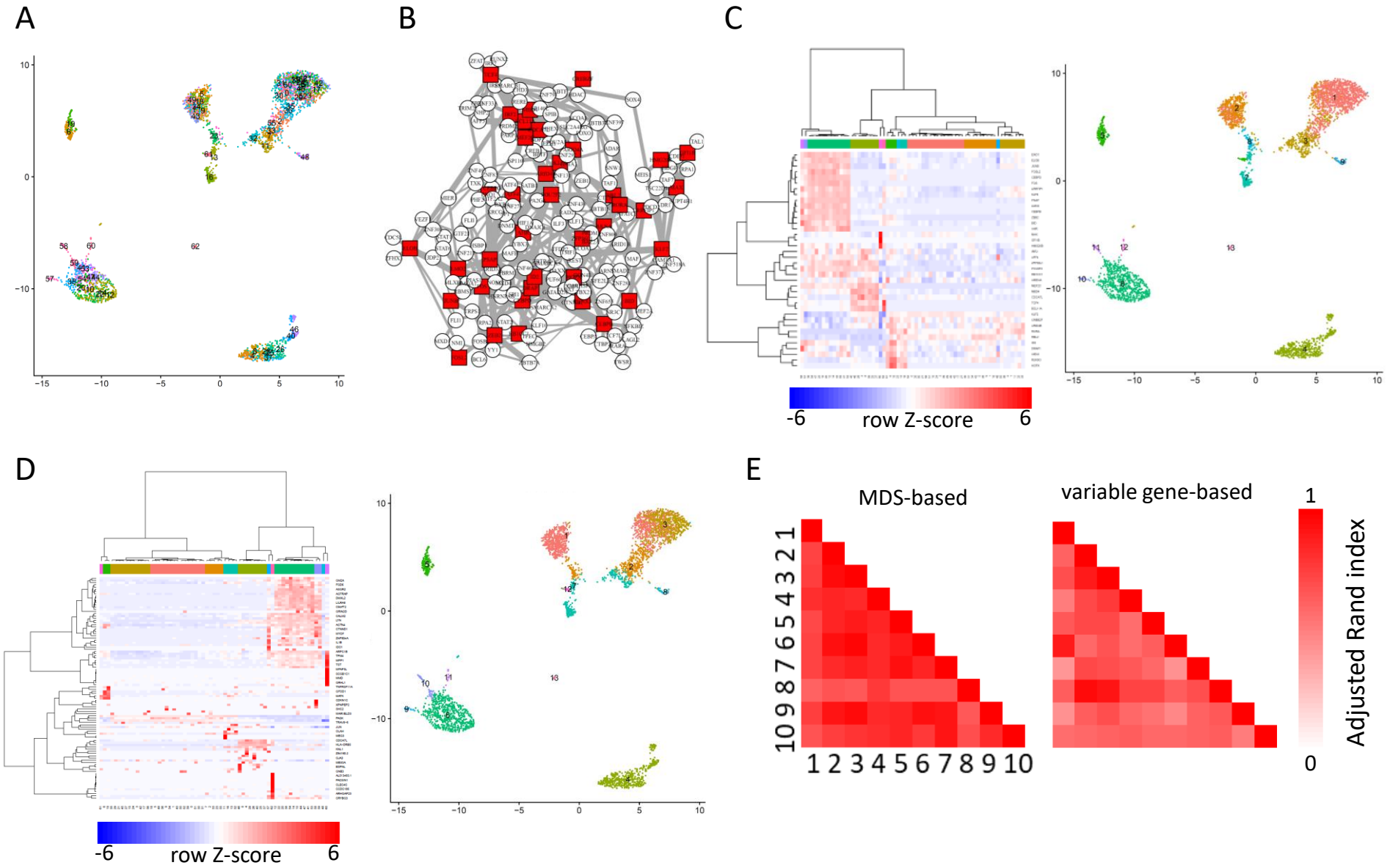

A
B
C
D
E
-6
row Z-score
6
MDS-based
variable gene-based
1
0
Adjusted Rand index
1 2 3 4 5 6 7 8 9 10

Figure 4

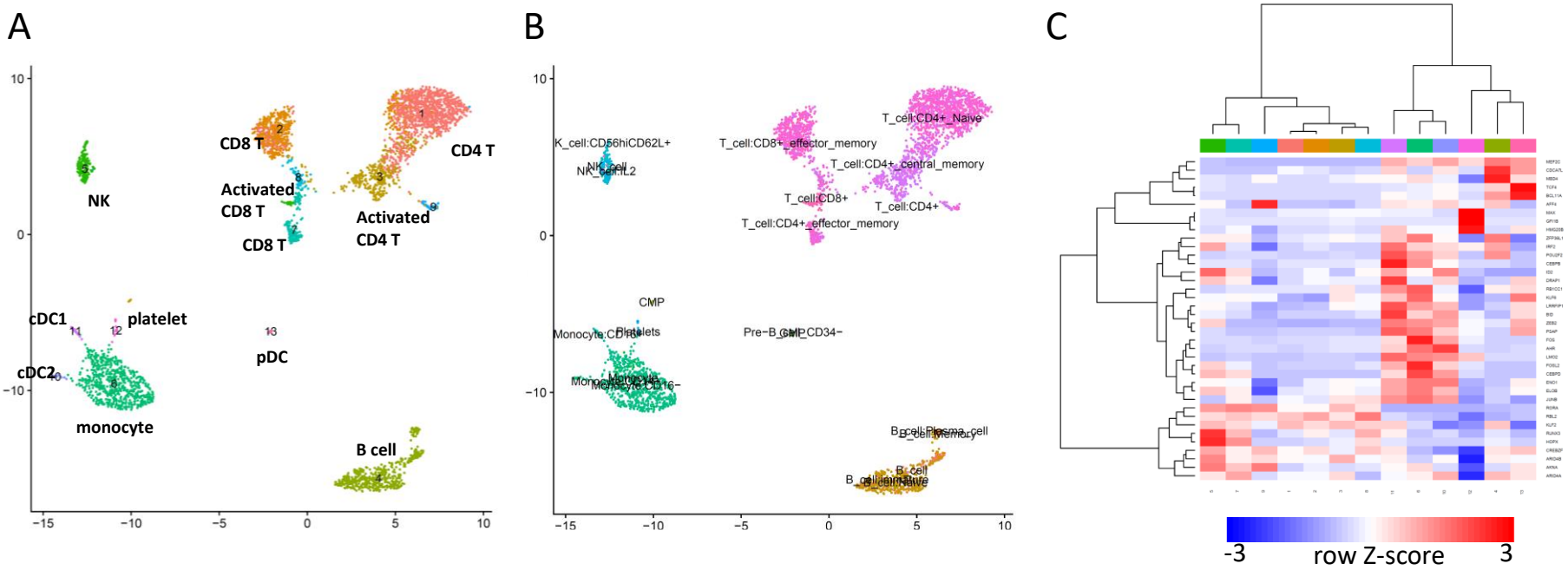

A
CD8 T
CD4 T
NK
Activated CD8 T
Activated CD4 T
CD8 T
cDC1
platelet
pDC
cDC2
monocyte
B cell
B
T_cell:CD4+_Naive
T_cell:CD8+_effector_memory
T_cell:CD4+_central_memory
T_cell:CD8+
T_cell:CD4+
T_cell:CD4+_effector_memory
CMP
Pre-B_cell_CD34-
C
row Z-score
-3
3

Figure 5

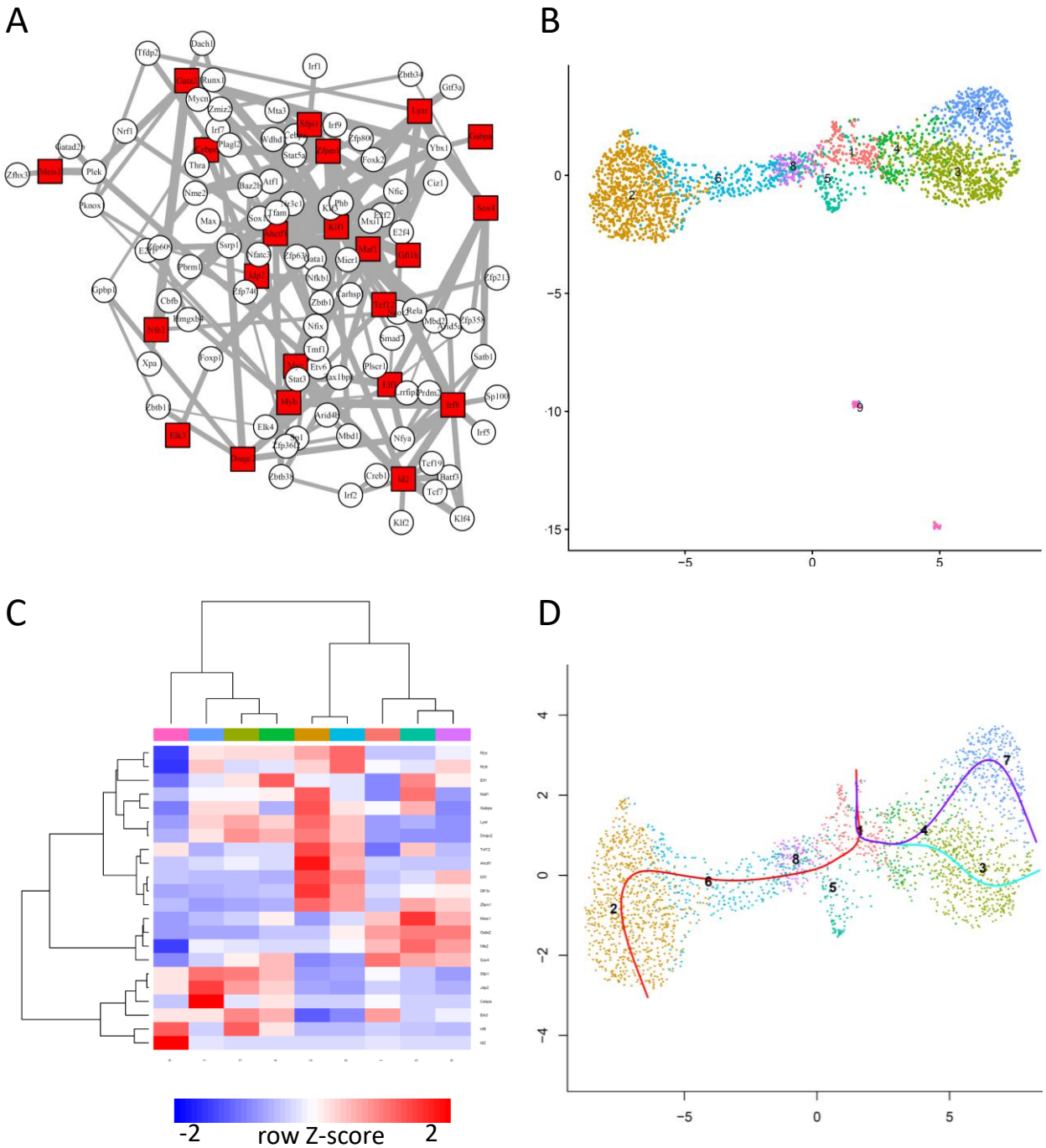


Figure 6

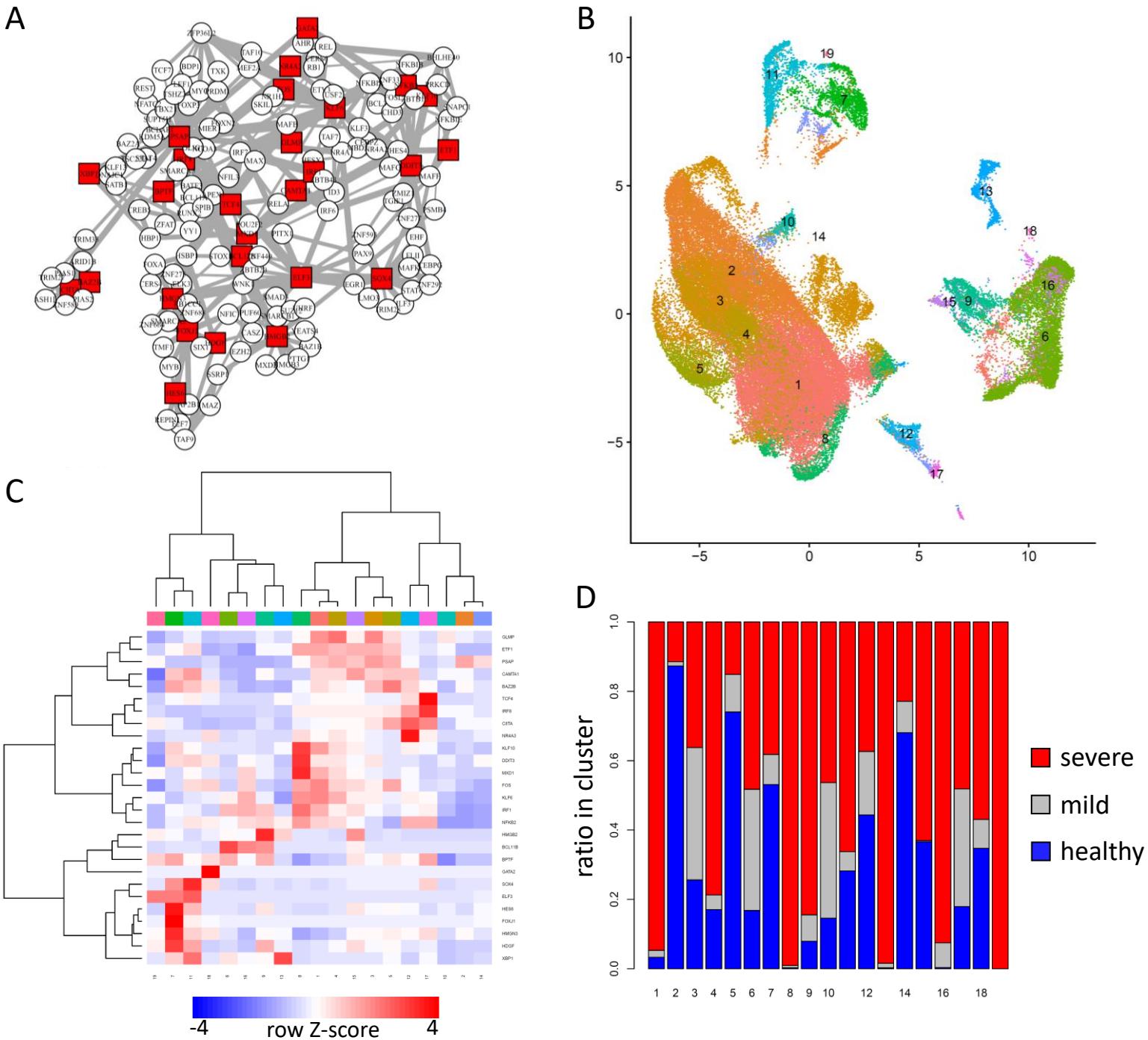
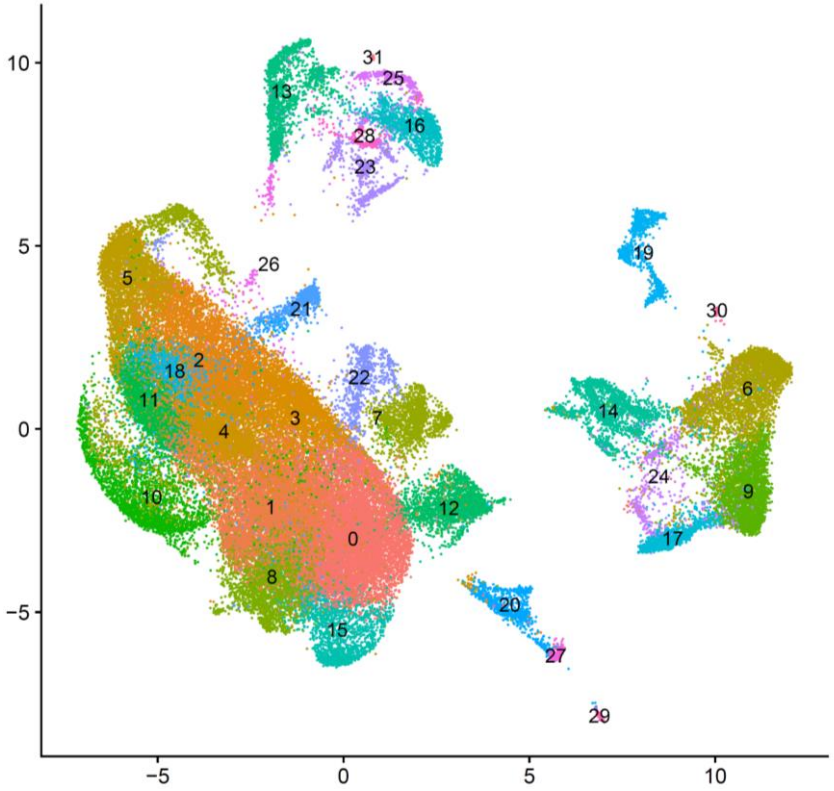
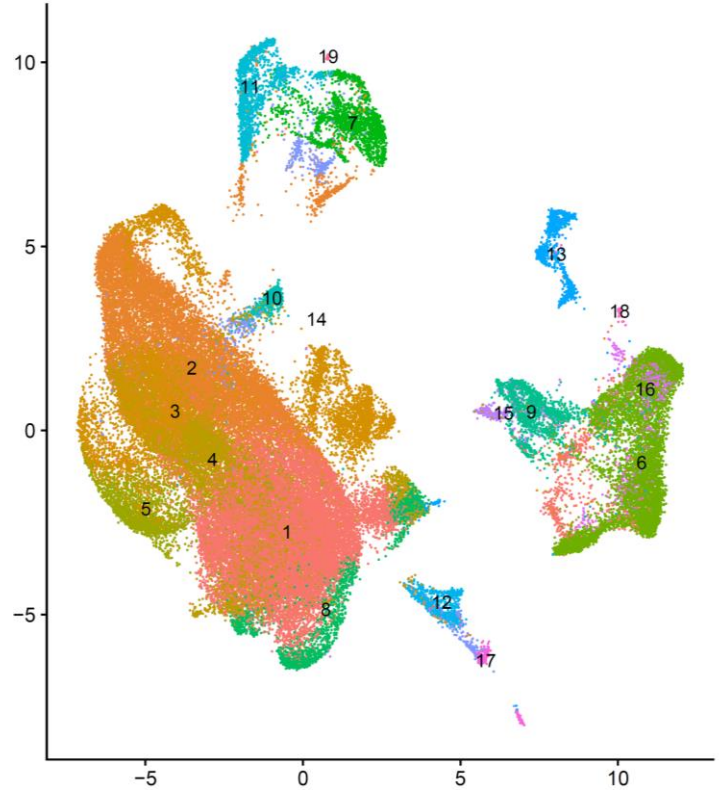


Figure 9

Original clustering in Liao et al.



MDS-based clustering



Supplementary Figure 1