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 identification of cell clusters 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.

Author Summary

Single-cell omics technologies, such as single-cell RNA-seq (scRNA-seq), are instrumental in identifying novel cell subsets that are involved in various biological processes and diseases. These technologies, however, require further development in data analysis, especially in areas focused on cell-type identification and inference of interactions between genes. The problem of cell-type identification essentially involves clustering, which necessitates a balance between distinguishing different cell states and grouping similar ones together. Current clustering methods still suffer from uncertainty in determining the appropriate number of clusters and in explaining why some cells are clustered together but others are separated. The inference of interactions between genes, gene regulatory network (GRN), remains challenging due to the noisy nature of scRNA-seq. We have developed BootCellNet, a method that infers GRNs and a set of genes dominating the network dynamics and utilizes the set to cluster cells and identify cell types. The method addresses challenges in GRN identification and clustering methods simultaneously and will facilitate the generation of the working hypotheses from a large amount of scRNA-seq data.

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 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, such as the selection of genes based on network centrality measures[4]. Among them, 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[5]. 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[6]. 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[7].

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.

This paper is organized as follows. In the first three sections, we introduced the BootCellNet procedure and applied it to scRNA-seq data of peripheral blood mononuclear cells. The first two sections showed the necessity of steps in BootCellNet for robust GRN inference and MDS computation. The following section showed that robustly computed MDS offers consistent clustering results, and the result was compared to the results of supervised labeling and unsupervised clustering. In the next section, we applied the method to scRNA-seq data of hematopoiesis to show if the method can be applied to continuous spectra of cell states and help to infer developmental course. The following section is devoted to the introduction of BootCellNet2, in which the unique set of control nodes is used instead of MDS, and the clustering is automated. We showed that the use of the new set of nodes and the automated clustering provides more robust clustering results. Validation using CITE-seq data was also performed in the following section to compare to unsupervised clustering. Finally, we applied BootCellNet2 to scRNA-seq data of cells from COVID-19 patients. In each application, the relevance of clustering results was evaluated by its robustness, and the biological relevance of the obtained clusters was evaluated through the expert-based labeling of the obtained clusters.

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 constructed GRN and MDS was computed as shown in Figure 4B. We calculated the average expression of MDS genes in each cell clump and clustered the cell

167 clumps as well as the genes by using hierarchical clustering with Ward's method (Figure
168 4C, left). By visual inspection of the heatmap of the MDS expression, we determine the
169 number of clusters of clumps as 13 (Figure 4C, right). As a comparison, we also carried
170 out the same by using a set of genes with variable expression instead of MDS (Figure 4D).
171 Of 2000 variable genes, we sampled 100 genes (around three times the number of MDS,
172 containing 36.0 genes on average; Figure 3B) and performed the same procedure as for
173 MDS. We then applied the number of clusters of clumps as 13, as in the procedure for
174 MDS. The clustering was compared with each other and adjusted Rand indexes (ARIs)
175 were calculated. As shown in Figure 4E, MDS-based clustering had higher ARIs among
176 the 10 specimens than clustering based on 100 variable genes, indicating the robustness
177 of MDS-based clustering.

178 To evaluate the obtained clusters, we picked one specimen shown in Figure 5A
179 and checked the marker genes, Gene Ontology (GO) terms, and enriched TFBS in each
180 cluster in the specimen. Marker genes, GO terms, and enriched TFBSs are listed in
181 Supplementary Tables 1, 2, and 3, respectively. Cluster 1, 2, 3, and 9 exhibited
182 characteristics of T cells. Cluster 1 is associated with the signature of CD4 T cells,
183 whereas Cluster 2 is associated with that of CD8 T cells. Clusters 3 and 7 exhibit
184 characteristics of activated CD4 and CD8 T cells, respectively. Cluster 8 is also associated
185 with T cells, and based on the MDS expression pattern, it exhibits high RUNX3
186 expression, indicating the cluster is CD8 T cells[8]. Cluster 5 is associated with NK cell-
187 like signatures. Cluster 4 shows B cell characteristics, marked by CD19 and
188 immunoglobulin genes. Clusters 6, 10, and 11 feature markers whose promoters are
189 enriched with PU.1/SPI1 transcription factor binding site (TFBS), suggesting that these
190 clusters represent myeloid cell populations. CLEC10A and CD1C genes emerged as

markers of Cluster 10, pointing to myeloid dendritic cells (DCs)[9]. 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[10]. 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.

The result obtained by BootCellNet was also compared to that of unsupervised

215 Louvain clustering (Figure 5C). We chose a resolution of 1.0 because it gave 14 clusters,
216 a similar number to that of BootCellNet. Although the results were mostly similar, we
217 found two notable differences. First, a cluster of pDC was missing. Those cells were
218 mixed with cDC2 to form a cluster (Cluster 11 in Figure 5C). Second, Louvain clustering
219 could distinguish memory B cells from naïve B cells (Clusters 9 and 3 in Figure 5C,
220 respectively), while BootCellNet could not and mixed into one cluster (Cluster 4 in Figure
221 5A) as described above. These results imply that unsupervised clustering including
222 Louvain clustering is effective enough to identify cell types, though there's no guiding
223 principle to choose resolution and no rational explanation to separate or merge cells at the
224 step of clustering. We will discuss these points later.

225 BootCellNet identifies MDS in GRN. This provides insights into how the
226 differentiation and/or the function of cell types are regulated by checking the expression
227 of genes in MDS as shown in Figure 5D. Comparing the B cell cluster 4 and the pDC
228 cluster 13, TCF4, a TF important for pDC development[10], was expressed more in
229 Cluster 13. Cluster 12, representing platelets, has a high expression of GFI1B which
230 regulates platelet function[11]. We found CEBPB, a TF that controls
231 monocyte/macrophage differentiation, was higher in Clusters 6 and 11 of
232 monocyte/macrophage compared to Cluster 10 of DCs. Id2 is expressed more in the NK
233 cell cluster 5, compared to Clusters 1, 2, 3, 7, 8, and 9, all of which are clusters of T cells.
234 Consistently, the TF is essential for NK cell maturation[12]. Cluster 7 also shows high
235 Id2 expression, thought to a lesser extent. This is consistent with reports that Id2 is
236 important for CD8+ memory T cell development[13] because the cluster represents an
237 activated CD8 T cell population. Further, by assessing RUNX3 expression, we
238 determined Cluster 2, 7, and 8, which show higher RUNX3 expression, to be CD8 T cells,

239 whereas Cluster 1, 3, and 9 are identified as CD4 T cells, given that RUNX3 is important
240 for CD8 T cell development[8]. By examining RORA expression in Cluster 1 and 3, we
241 identified Cluster 3 as activated T cells, since RORA is notably expressed in Th17
242 population[14].

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

247

248 BootCellNet facilitates inference of developmental courses of 249 hematopoiesis.

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

258 The cells were clustered into 9 as shown in Figure 6B. By examining the
259 expressed MDS TFs (Figure 6C), the marker genes (Supplementary Table 4), and
260 associated GO terms (Supplementary Table 5), we identified Cluster 1, 4, 6, and 8 as stem
261 cells, Cluster 2 as erythrocytes, Cluster 5 as megakaryocytes, Cluster 3 as monocytes, and

262 Cluster 7 as neutrophils. Cluster 9 seems to be lymphocytes and dendritic cells, which are
263 removed in the analyses performed in a report using the data[16].

264 Cluster 1 has Flt3 as a marker, as well as Sox4, Runx1, Cebpa, Lmo4, and Gata2,
265 which are TFs known to be expressed in hematopoietic stem cells. Cluster 3, 4, and 7 are
266 high in Sfpi1 expression, indicating that those clusters are myeloid lineages. However,
267 the expression of Sox4 distinguishes Cluster 4 from others as more progenitor-like cells.
268 Cluster 3 is associated with terms related to innate immune responses, and has Csf1r as a
269 marker gene, as well as Irf8. Cluster 7 is associated with myeloid cell-related terms and
270 has Fcnb and Elane genes as markers. The mentioned marker genes are characteristics of
271 monocytic and neutrophilic cells, respectively[17], supporting the identification of
272 Cluster 3 as monocytes and Cluster 7 as neutrophils. Comparing Cluster 3 and 7, the
273 former is high in Irf8 expression, whereas the latter has higher expression of Jdp2 and
274 Cebpe. This expression pattern of TFs is also a characteristic that distinguishes monocytes
275 and neutrophils[17, 18]. The importance of higher expression of Cebpe in the neutrophil
276 lineage has been suggested by using CellOracle, which is a cell lineage analysis method
277 based on GRN and a mathematical dynamical system model[16]. The result demonstrates
278 that BootCellNet efficiently infers MDS and their roles in development, without using
279 mathematical dynamical system models.

280 We inferred the lineage trajectories by using the clusters identified by
281 BootCellNet. Slingshot[19] was applied to the data, setting Cluster 1 as a starting point.
282 This resulted in 4 lineage curves (Figure 6D; one lineage is omitted, see the Figure Legend
283 for details). The result recapitulates known developmental trajectories and aligns with the
284 analysis performed by using a different method[16]. One trajectory represents the
285 development of erythrocytes spanning from Cluster 1 to 2 through 8 and 6. The others

286 represent that of monocytes and neutrophils spanning from Cluster 1, through 4, to Cluster
287 3 and 7, respectively. This result suggests that cell cluster identification by BootCellNet
288 facilitates lineage trajectory inference.
289

290 Automated cell type cluster identification by BootCellNet2.

291 Two ambiguous points remain in the BootCellNet procedure. First, MDS is not
292 necessarily uniquely determined. Second, we determined the number of clusters by visual
293 inspection of the hierarchical clustering heatmap.

294 To circumvent the first problem, we use the idea of critical nodes and intermittent
295 nodes. In the case where multiple MDSs exist, one can define critical nodes and
296 intermittent nodes, which appear in all the MDSs or at least in one MDS[20]. A set of
297 critical and intermittent nodes is uniquely determined. By using those sets, we could
298 circumvent possible inconsistencies derived from the non-uniqueness of MDS.

299 To address the second point, we utilize multiscale bootstrapping to determine
300 statistically significant clusters. The method provides approximately unbiased confidence
301 for each clustering[21]. By providing a cutoff *p*-value in advance, an R package pvclust
302 extracts statistically confident clusters of cell clumps. However, this leaves other clumps
303 unclustered. Thus we compare the clusters given by pvclust with the clusters obtained by
304 hierarchical clustering with varying numbers of clusters, and determine the number of
305 clusters.

306 We implemented the above methods (Figure 7A), and the procedure was applied
307 to the PBMC dataset. Using the same dataset as in Figure 2, we first checked whether the
308 new procedure consistently gives a unique set of nodes. We counted the appearance of
309 nodes, both critical and intermittent, in 10 specimens. As shown in Figure 7B, the number

310 of nodes that consistently appeared 10 times was not significantly different from that of
311 the former procedure. This indicates that the uniquely determined set of nodes is still
312 influenced by noises and resampling in GRN reconstruction.

313 We then performed clustering of cell clumps using the critical and intermittent
314 nodes instead of MDS. Also, we utilized multiscale bootstrap to determine the number of
315 clusters. The R package pvclust gives clusters with relatively low approximately unbiased
316 p-values. Based on the clustering, we determined the number of clusters as described in
317 detail in Methods and shown briefly in Figure 7A. The obtained clusters were compared
318 with each other and ARI was calculated. As shown in Figure 7B, the automated clustering
319 procedure along with the calculation of critical and intermittent nodes (hereafter we call
320 the nodes as control nodes) gave clustering results highly similar to each other. The
321 numbers of clusters were stable, 16 or 17 except for one 14. This indicates that the
322 procedure offered consistent clustering results. If we did not use the unique set of control
323 nodes for clustering but instead used MDS, the number of clusters had more variations.
324 ARIs were significantly higher in the combination of the automated clustering procedure
325 and the use of the unique set of nodes than others (Figure 7C). This suggests the
326 importance of the use of the unique set of nodes, although the set had variation as shown
327 above. Taken together, these results indicate that the use of the automated clustering
328 procedure together with the unique control node set offers more consistent clustering than
329 the former procedure.

330 To evaluate the newly obtained clusters, we picked the same specimen as shown
331 in Figure 5A. The resulting GRN with critical nodes and intermittent nodes is shown in
332 Figure 8A. The GRN contains 186 TFs with 22 critical and 33 intermittent TFs. Using the
333 nodes, cell clumps were clustered, obtaining 17 clusters. The clusters are mostly the same

as those in Figure 5A. We checked the marker genes and associated GO terms (Supplementary Tables 6 and 7, respectively). Based on the marker genes and GO terms, the cell clusters were labeled as shown in Figure 8B. Clusters of T cells were more separated than in Figure 5A, and the novel clustering result enabled the identification of naïve, activated, central memory, and effector memory subsets, along with Cluster 1, 4, and 11, which represent naïve CD4 T cell subsets. As in the above analysis, we made use of markers as well as the set of nodes to explain why those clusters are separated (Figure 8C). Cluster 1 is high in JUNB which is critical for the survival of T cells[22], while Cluster 11 is high in NFE2L2/NRF2 which interfers activation of T cells[23] and Cluster 4 is not high for either, suggesting differential activation states of the three naïve T cell clusters. These results together suggest that the novel procedure BootCellNet2 offers automated and robust clustering and identification of biologically meaningful cell types.

BootCellNet2 identified cell type clusters complying with CITE-seq only from transcriptome data.

We further validated BootCellNet2 using CITE-seq data, which allows us to evaluate clustering results with antibody-derived tags. The BootCellNet2 procedure was applied to transcriptome data from the CITE-seq study reported[24]. The analysis identified a set of 29 control nodes, consisting of 19 critical nodes and 10 intermittent nodes (Supplementary Figure 1), and generated 10 clusters as shown in Figure 9A. We labeled the clusters based on their marker genes and associated GO terms for each cluster (Supplementary Tables 8 and 9, respectively), as well as CITE-seq results (Figure 9C). In

comparison, we also performed clustering using the Louvain algorithm with a resolution of 0.7, which resulted in 10 clusters, the same number of clusters as BootCellNet2. The clusters were labeled as shown in Figure 9D, according to the marker genes, associated GO terms, and CITE-seq results (Supplementary Tables 10 and 11, and Figure 9E, respectively). The results suggest that the two methods offered largely comparable results.

A few key differences, however, were found. BootCellNet2 identified distinct states of T cells similar to those observed in another PBMC dataset analysis shown above (see Figure 8B), which was not the case with Louvain clustering. Moreover, CITE-seq results suggest clear separation in CD4 and CD8 expressions within the T cell clusters identified by BootCellNet2 (Cluster 1, 2, 3, and 5, as shown in Figure 9B); Clusters 3 and 5 exhibited low CD4 and high CD8, whereas Clusters 1 and 2 showed high CD4 and low CD8 expression. The difference between the states of CD4 and CD8 T cells was also evident from CD45RA expression known to be high in naïve T cells (high in Cluster 1 and low in Cluster 2) and CD57 expression known to be high in activated T cells (high in Cluster 3 and low in Cluster 5), respectively. Expression of HMGB2, which is highly expressed in effector and memory T cells compared to naïve T cells[25], further supports these findings (Figure 9C). Additionally, platelets were identified by BootCellNet2 but not by Louvain. These results collectively imply that BootCellNet2 clustering achieves more accurate cell type delineation, aligning more closely with antibody-derived tag results than unsupervised clustering.

Identification of characteristic cell types in COVID-19 patients by BootCellNet2.

380 We applied BootCellNet2 to scRNA-seq data of bronchoalveolar lavage fluid cells from
381 COVID-19 patients[26] to test whether BootCellNet2 can identify cell types in scRNA-
382 seq data where discrete and continuous cell types/states are mixed. The cells were
383 clustered into 32 in the original article. BootCellNet2 gave the GRN of 161 TFs with 21
384 critical nodes and 13 intermittent nodes (Figure 10A). The automated clustering
385 procedure offered 17 clusters of the cell clumps (Figure 10B). Marker genes were
386 identified for each cluster, and the GO terms associated with the markers of each cluster
387 were checked (Supplementary Tables 12 and 13, respectively).

388 Clusters 1, 2, 3, 5, 8, 13, 15, and 16 were associated with myeloid cell-related
389 signatures. Compared to the clusters in the original report (reproduced in Supplementary
390 Figure 2, encircled as “macrophage/monocyte”, “neutrophil”, and “DC”), the clustering
391 by BootCellNet2 has a smaller number of clusters while identifying myeloid cells with
392 different characteristics.

393 Clusters 1, 3, and 8 were associated with inflammatory immune responses and/or
394 characteristics of macrophage/monocyte/DC thus identified as myeloid cells. Clusters 1,
395 3, and 8 have high expression of FOS, IRF1, and NFKB2, which are TFs involved in
396 inflammatory immune response (Figure 10C). The clusters are significantly enriched with
397 cells from severe patients (Figure 10D). The associated terms are also related to
398 inflammatory responses. In contrast, Clusters 2, 5, 13, 15, and 16 are enriched with cells
399 from mild and/or healthy individuals. Clusters 5, 6, 13, 14, and 16 have HLA genes as
400 markers and are associated with DC/monocyte signatures, implying the clusters contain
401 antigen-presenting myeloid cells. Clusters 5, 13, and 16 have either MARCO, CEBPA,
402 and PPARG as markers, indicating that the clusters are alveolar macrophages since
403 PPARG is an essential TF for alveolar macrophage development[27]. Clusters 5, 6, 13,

404 14, and 16 did not show high expression of inflammatory TFs such as FOS, IRF1, and
405 NFKB2, in contrast to Clusters 1, 4, and 8. This implies that the myeloid cells in healthy
406 individuals or patients with mild symptoms are less inflammatory than those in
407 individuals with severe disease, consistent with former reports showing that inflammation
408 and cytokine storms are the cause of severe COVID-19[28].

409 Cluster 2 is associated with myeloid and dendritic cell signatures. It has CD19
410 as a marker. Indeed, the cluster may contain pDC, myeloid DC, and B cells, which were
411 clustered separately in the original report (Supplementary Figure 2), suggesting a
412 probable failure of the BootCellNet2 procedure in the identification of cell types.
413 Clusters 4 and 15 are associated with T cell signature and have genes for components of
414 T cell receptors such as CD3D/E/G, LCK, LAT, and ZAP70 as markers. They also have
415 CD8A/B and RUNX3 as markers, indicating that those clusters are CD8 T cell
416 populations. Further, all of them have genes for granzymes and IFN-g as markers,
417 suggesting that the clusters represent activated CD8 T cells.

418 To examine the differences among those activated T cells clustered separately,
419 we compared Cluster 4 to Cluster 15 and found that the latter, which is enriched with cells
420 from patients with mild symptoms, but not the former, which is enriched with cells from
421 patients with severe symptoms, has ID2, TCF7, and CCR5 as markers, which are highly
422 expressed in effector memory CD8 T cells[13, 29, 30]. We also examined the expression
423 of IL7R (encoding IL-7Ralpha) and PDCD1 (encoding PD-1) and found that Cluster 15
424 has both as markers while Cluster 4 has only PDCD1, indicating that Cluster 4 is
425 exhausted T cells [31]. Notably, examination of the expression of the control nodes in the
426 GRN also showed that HMGB2, which is indispensable for the development of exhausted
427 T cells[25], is higher in Cluster 4 than in Cluster 15 (Figure 10C). These results

428 collectively suggest that induction of effector memory CD8 T cell upon SARS-CoV-2
429 infection is critical for controlling disease to mild symptoms, and an increase of exhausted
430 T cells is detrimental for disease control, consistent with former reports[26, 32, 33].
431 Taken together, the results indicate that BootCellNet2 successfully identifies
432 clusters of cells specific to patients with different symptoms, and offers explanations of
433 the difference in symptoms that are consistent with preceding reports.
434

435 Discussion.

436 In the current study, a method named BootCellNet has been developed, which combines
437 smoothing, resampling, GRN inference, computation of a set of genes controlling the
438 entire network, and clustering. We demonstrated the robustness of our method,
439 particularly given the probabilistic steps involved and the presence of noise within the
440 data. We also performed an expert-based evaluation of the clusters by manually labeling
441 the clusters. This ensured that despite the probabilistic nature of our method, the clusters
442 identified are not only statistically robust but also biologically meaningful.
443 Since the method consists of multiple steps, the choice of algorithms will affect
444 the results of the procedure. We used *k*-NN smoothing to reduce the noise in scRNA-seq.
445 We performed the smoothing on the 2-dimensional UMAP plane, but one can do the same
446 in a higher dimension[34]. One can also use different smoothing methods, such as *k*-NN
447 graph-based MetaCell algorithm[35], or data imputation methods such as MAGIC[36]
448 and DeepImpute[37].
449 In BootCellNet, GRN reconstruction is performed with the ARACNe method,
450 because of its low computational load. As in the current study, NestBoot utilizes a nested
451 bootstrap to control FDR in GRN inference[38], and they showed that the bootstrapping

452 procedure improved the accuracy of the GRN inference by various inference methods
453 such as GENIE3[39]. Thus, the incorporation of methods other than ARACNe in
454 BootCellNet may improve the GRN inference accuracy.

455 The use of MDS as a set of genes that controls the dynamics of the whole
456 network needs some caution. The controllability of a network by its MDS requires a
457 premise that each MDS node can control its adjacent nodes individually[5]. In the case
458 where nodes are TFs, this may be justified by assuming that a TF controls each target
459 gene through individually controlled accessibility, although the assumption should be
460 verified independently.

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

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

473 Despite the room for improvement and caution, and although the method is
474 relatively simpler than other sophisticated methods, we demonstrated that BootCellNet
475 effectively worked for a variety of datasets with different characteristics. Notably, the

proposed method excels in explainability, not only clustering cells but also explaining why they are clustered based on the GRN, and providing characterization of what type of population they represent. Along with automating the determination of the number of clusters, the method resolves the common problem in unsupervised clustering which does 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. For example, the clusters can be explained by factors such as GFI1B, CEBPB, and RUNX3 in clustering PBMC data, by Cebpe in clustering hematopoiesis data, and by HMGB2 in clustering COVID-19 data, allowing the identification of what type of cells the clusters represent and what characteristics they possess, even within similar cell types, without looking at GO terms, using information from GRN/MDS. This demonstrates that BootCellNet can be a useful complementary for other methods of clustering, supervised labeling, and lineage inference.

Also, BootCellNet, especially BootCellNet2, is suitable for automation. The method will be utilized for automated and routine annotation of large-scale scRNA-seq cell data. Moreover, BootCellNet addresses an increasing need for explainable automated data processing[40], 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.

Materials and 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.

BootCellNet2 is outlined in Figure 7A. The procedure utilizes a unique set of control nodes consisting of critical nodes and intermittent nodes, instead of MDS, and the clustering is automated as described below.

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. For given k and 2-dimensional UMAP space, for each cell, k-nearest neighbors in the UMAP space with Euclidean distance are selected, and the average of the expression levels is computed to obtain k-NN smoothed cells. The decimal part of the averaged expression levels was truncated to obtain expression levels in integers.

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} $(k = 1, 2, \dots K)$[41]. 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[42] 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[43] in the case of human dataset analysis, and from AnimalTFDB 4.0[44] 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[45] implemented in an R package minet[46] with the estimation of mutual information by Miller-Madow asymptotic bias corrected empirical estimator and an equal width discretization, as recommended in a report[47].

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[48]. 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. We also implemented the computation of critical and intermittent nodes. To obtain the nodes,

the computation of MDS is repeated for the defined times (15 times as default, by setting num.bin.solns = 15 in the function lp of the lpSolve package), and nodes identified as MDS in all the computations are defined as critical nodes, and those in at least one computation are defined as intermittent nodes.

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 much lower than k (we use 80 % of k as a default procedure), 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.

In the BootCellNet2 procedure, an R package pvclust is used to compute approximately unbiased p values for clusters of clumps with the multiscale bootstrap algorithm. A function pvpick in the pvclust package extracts clusters having p values less than a given cutoff to offer a high-confident clustering. ARIs are computed between the high-confident clustering result given by pvclust/pvpick and those given by a function rect.hclust with varying numbers of clusters. Then the number of clusters of clumps is determined as the lowest number with a maximal ARI (see Figure 7A).

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 CITE-seq data of PBMC[24] was retrieved from NCBI GEO under accession number GSE100866. The data is comprised of scRNA-seq data and antibody-derived tag sequencing data. These data were aggregated and processed with a standard workflow by using Seurat. The processed data containing 7690 cells after filtering was then subjected to BootCellNet with $k = 10$ and $N = 5000$. The starting 66 cell clumps were obtained by using Louvain clustering with a resolution of 9.

The single-cell RNA-seq data of COVID-19 patients[26] 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

618 and 6 severe patients. The data were analyzed by using Seurat and were subjected to
619 BootCellNet with $k = 20$ and $N = 20000$. The starting 113 cell clumps were given by using
620 Louvain clustering with a resolution of 10.

621 As a comparison, supervised cell type annotation was performed by using
622 SingleR[49] with Human Primary Cell Atlas fine subtype labels in an R package
623 celldex[49].

624

625 Transcription factor binding site (TFBS), Gene Ontology (GO) 626 enrichment and lineage analyses.

627 Marker genes were obtained by using FindAllMarkers function in Seurat. We used
628 Homer[50] for TFBS and GO analysis of the marker genes of each cluster. Promoters are
629 defined as from -500 to +200 of transcription start sites. For lineage analysis,
630 Slingshot[19] was utilized with default parameters.

631

632 Statistical comparison of clustering results.

633 Adjusted Rand index[51] was calculated to compare two clustering results.

634

635

636 Author Contributions

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

639

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 and by Human Genome Center, the Institute of Medical Science, the University of Tokyo.

References

1. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24(8):550-72. doi: 10.1038/s41576-023-00586-w. PubMed PMID: 37002403; PubMed Central PMCID: PMCPMC10066026 Janssen, and received consulting fees from Chan-Zuckerberg Initiative. F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Omniscope Ltd, and has ownership interest in Dermagnostix GmbH and Cellarity. M.G.J. consults for and has ownership interests in Vevo Therapeutics. L. Heumos has received

speaker's honorarium from Vesalius Therapeutics. Single-Cell Best Practices Consortium: M.G.J. consults for and has ownership interests in Vevo Therapeutics. R.P. is co-founder of Ocean Genomics, Inc. The other authors declare no competing interests.

2. Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief Bioinform. 2021;22(3). doi: 10.1093/bib/bbaa190. PubMed PMID: 34020546; PubMed Central PMCID: PMCPMC8138892.

3. Guo WF, Zhang SW, Zeng T, Akutsu T, Chen L. Network control principles for identifying personalized driver genes in cancer. Brief Bioinform. 2020;21(5):1641-62. doi: 10.1093/bib/bbz089. PubMed PMID: 31711128.

4. Badia IMP, Wessels L, Muller-Dott S, Trimbour R, Ramirez Flores RO, Argelaguet R, et al. Gene regulatory network inference in the era of single-cell multi-omics. Nat Rev Genet. 2023;24(11):739-54. doi: 10.1038/s41576-023-00618-5. PubMed PMID: 37365273.

5. Nacher JC, Akutsu T. Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. New Journal of Physics. 2012;14(7):073005.

6. Nazarieh M, Wiese A, Will T, Hamed M, Helms V. Identification of key player genes in gene regulatory networks. BMC Syst Biol. 2016;10(1):88. doi: 10.1186/s12918-016-0329-5. PubMed PMID: 27599550; PubMed Central PMCID: PMCPMC5011974.

7. Wuchty S. Controllability in protein interaction networks. Proc Natl Acad Sci U S A. 2014;111(19):7156-60. doi: 10.1073/pnas.1311231111. PubMed PMID: 24778220; PubMed Central PMCID: PMCPMC4024882.

8. Egawa T, Tillman RE, Naoe Y, Taniuchi I, Littman DR. The role of the Runx transcription factors in thymocyte differentiation and in homeostasis of naive T cells. J Exp Med. 2007;204(8):1945-57. doi: 10.1084/jem.20070133. PubMed PMID: 17646406; PubMed Central PMCID: PMCPMC2118679.

9. Collin M, Bigley V. Human dendritic cell subsets: an update. Immunology. 2018;154(1):3-20. doi: 10.1111/imm.12888. PubMed PMID: 29313948; PubMed Central PMCID: PMCPMC5904714.

10. Reizis B. Plasmacytoid Dendritic Cells: Development, Regulation, and Function. Immunity. 2019;50(1):37-50. doi: 10.1016/j.immuni.2018.12.027. PubMed PMID: 30650380; PubMed Central PMCID: PMCPMC6342491.

11. Stevenson WS, Morel-Kopp MC, Chen Q, Liang HP, Bromhead CJ, Wright S, et al. GFI1B mutation causes a bleeding disorder with abnormal platelet function. J Thromb Haemost. 2013;11(11):2039-47. doi: 10.1111/jth.12368. PubMed PMID: 23927492.

12. Li ZY, Morman RE, Hegermiller E, Sun M, Bartom ET, Maienschein-Cline M, et al. The transcriptional repressor ID2 supports natural killer cell maturation by controlling TCF1 amplitude. J Exp Med. 2021;218(6). doi: 10.1084/jem.20202032. PubMed PMID: 33857289; PubMed Central PMCID: PMCPMC8056751 the submitted work. No other disclosures were reported.

13. Cannarile MA, Lind NA, Rivera R, Sheridan AD, Camfield KA, Wu BB, et al. Transcriptional regulator Id2 mediates CD8+ T cell immunity. Nat Immunol. 2006;7(12):1317-25. doi: 10.1038/ni1403. PubMed PMID: 17086188.

14. Yang XO, Pappu BP, Nurieva R, Akimzhanov A, Kang HS, Chung Y, et al. T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. Immunity. 2008;28(1):29-39. doi: 10.1016/j.immuni.2007.11.016. PubMed PMID: 18164222; PubMed Central PMCID: PMCPMC2587175.

15. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell. 2015;163(7):1663-77. doi: 10.1016/j.cell.2015.11.013. PubMed PMID: 26627738.

16. Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. Dissecting cell identity via network inference and in silico gene perturbation. Nature. 2023;614(7949):742-51. doi: 10.1038/s41586-022-05688-9. PubMed PMID: 36755098; PubMed Central PMCID: PMCPMC9946838.

17. Kwok I, Becht E, Xia Y, Ng M, Teh YC, Tan L, et al. Combinatorial Single-Cell Analyses of Granulocyte-Monocyte Progenitor Heterogeneity Reveals an Early Uni-potent Neutrophil Progenitor. Immunity. 2020;53(2):303-18 e5. doi: 10.1016/j.immuni.2020.06.005. PubMed PMID: 32579887.

18. Maruyama K, Fukasaka M, Vandenbon A, Saitoh T, Kawasaki T, Kondo T, et al. The transcription factor Jdp2 controls bone homeostasis and antibacterial immunity by regulating osteoclast and neutrophil differentiation. Immunity. 2012;37(6):1024-36. doi: 10.1016/j.immuni.2012.08.022. PubMed PMID: 23200825.

19. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477. doi: 10.1186/s12864-018-4772-0. PubMed PMID: 29914354; PubMed Central PMCID: PMCPMC6007078.

20. Nacher JC, Akutsu T. Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. Journal of Complex Networks. 2014;2(4):394-412. doi: 10.1093/comnet/cnu029.

21. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and evolution.

1999;16(8):1114.

22. Hsieh T, Sasaki D, Taira N, Chien H, Sarkar S, Seto Y, et al. JunB Is Critical for Survival of T Helper Cells. Front Immunol. 2022;13:901030. doi: 10.3389/fimmu.2022.901030. PubMed PMID: 35837408; PubMed Central PMCID: PMCPMC9273772.

23. Muri J, Kopf M. Redox regulation of immunometabolism. Nat Rev Immunol. 2021;21(6):363-81. doi: 10.1038/s41577-020-00478-8. PubMed PMID: 33340021.

24. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865-8. doi: 10.1038/nmeth.4380. PubMed PMID: 28759029; PubMed Central PMCID: PMCPMC5669064.

25. Neubert EN, DeRogatis JM, Lewis SA, Viramontes KM, Ortega P, Henriquez ML, et al. HMGB2 regulates the differentiation and stemness of exhausted CD8(+) T cells during chronic viral infection and cancer. Nat Commun. 2023;14(1):5631. doi: 10.1038/s41467-023-41352-0. PubMed PMID: 37704621; PubMed Central PMCID: PMCPMC10499904.

26. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26(6):842-4. doi: 10.1038/s41591-020-0901-9. PubMed PMID: 32398875.

27. Schneider C, Nobs SP, Kurrer M, Rehrauer H, Thiele C, Kopf M. 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. 2014;15(11):1026-37. doi: 10.1038/ni.3005. PubMed PMID: 25263125.

28. Vanderbeke L, Van Mol P, Van Herck Y, De Smet F, Humblet-Baron S, Martinod K, et al. Monocyte-driven atypical cytokine storm and aberrant neutrophil activation as key mediators of COVID-19 disease severity. Nat Commun. 2021;12(1):4117. doi: 10.1038/s41467-021-24360-w. PubMed PMID: 34226537; PubMed Central PMCID: PMCPMC8257697.

29. Zhou X, Yu S, Zhao DM, Harty JT, Badovinac VP, Xue HH. Differentiation and persistence of memory CD8(+) T cells depend on T cell factor 1. Immunity. 2010;33(2):229-40. doi: 10.1016/j.immuni.2010.08.002. PubMed PMID: 20727791; PubMed Central PMCID: PMCPMC2928475.

30. Kohlmeier JE, Miller SC, Smith J, Lu B, Gerard C, Cookenham T, et al. The chemokine receptor CCR5 plays a key role in the early memory CD8+ T cell response to respiratory virus infections. Immunity. 2008;29(1):101-13. doi: 10.1016/j.immuni.2008.05.011. PubMed PMID: 18617426; PubMed Central PMCID:

773 PMCPMC2519120.
774 31. Seo W, Jerin C, Nishikawa H. Transcriptional regulatory network for the
775 establishment of CD8(+) T cell exhaustion. Exp Mol Med. 2021;53(2):202-9. doi:
776 10.1038/s12276-021-00568-0. PubMed PMID: 33627794; PubMed Central PMCID:
777 PMCPMC8080584 Pharmaceutical, MSD and Bristol-Myers Squibb and research
778 funding from Taiho Pharmaceutical, Daiichi-Sankyo, Kyowa Kirin, Zenyaku Kogyo,
779 Oncolys BioPharma, Debiopharma, Asahi-Kasei, Sysmex, Fujifilm, SRL, Astellas
780 Pharmaceutical, Sumitomo Dainippon Pharma and BD Japan outside of this study. The
781 other authors declare no competing interests.
782 32. Odak I, Barros-Martins J, Bosnjak B, Stahl K, David S, Wiesner O, et al.
783 Reappearance of effector T cells is associated with recovery from COVID-19.
784 EBioMedicine. 2020;57:102885. doi: 10.1016/j.ebiom.2020.102885. PubMed PMID:
785 32650275; PubMed Central PMCID: PMCPMC7338277.
786 33. Wauters E, Van Mol P, Garg AD, Jansen S, Van Herck Y, Vanderbeke L, et al.
787 Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell
788 profiling of bronchoalveolar lavages. Cell Res. 2021;31(3):272-90. doi: 10.1038/s41422-
789 020-00455-9. PubMed PMID: 33473155; PubMed Central PMCID: PMCPMC8027624.
790 34. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput
791 single-cell RNA-Seq data. BioRxiv. 2017:217737.
792 35. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al.
793 MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol.
794 2019;20(1):206. doi: 10.1186/s13059-019-1812-2. PubMed PMID: 31604482; PubMed
795 Central PMCID: PMCPMC6790056.
796 36. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering
797 Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174(3):716-
798 29 e27. doi: 10.1016/j.cell.2018.05.061. PubMed PMID: 29961576; PubMed Central
799 PMCID: PMCPMC6771278.
800 37. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an
801 accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq
802 data. Genome Biol. 2019;20(1):211. doi: 10.1186/s13059-019-1837-6. PubMed PMID:
803 31627739; PubMed Central PMCID: PMCPMC6798445.
804 38. Morgan D, Tjarnberg A, Nordling TEM, Sonnhammer ELL. A generalized
805 framework for controlling FDR in gene regulatory network inference. Bioinformatics.
806 2019;35(6):1026-32. doi: 10.1093/bioinformatics/bty764. PubMed PMID: 30169550.
807 39. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks
808 from expression data using tree-based methods. PLoS One. 2010;5(9). doi:

10.1371/journal.pone.0012776. PubMed PMID: 20927193; PubMed Central PMCID: PMCPMC2946910.

40. Occhipinti A, Verma S, Doan LMT, Angione C. Mechanism-aware and multimodal AI: beyond model-agnostic interpretation. Trends Cell Biol. 2024;34(2):85-9. doi: 10.1016/j.tcb.2023.11.002. PubMed PMID: 38087709.

41. Hinkley D, Shi S. Importance sampling and the nested bootstrap. Biometrika. 1989;76(3):435-46.

42. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29. doi: 10.1016/j.cell.2021.04.048. PubMed PMID: 34062119; PubMed Central PMCID: PMCPMC8238499.

43. Joung J, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, et al. A transcription factor atlas of directed differentiation. Cell. 2023;186(1):209-29 e26. doi: 10.1016/j.cell.2022.11.026. PubMed PMID: 36608654; PubMed Central PMCID: PMCPMC10344468.

44. Shen WK, Chen SY, Gan ZQ, Zhang YZ, Yue T, Chen MM, et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Res. 2023;51(D1):D39-D45. doi: 10.1093/nar/gkac907. PubMed PMID: 36268869; PubMed Central PMCID: PMCPMC9825474.

45. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7 Suppl 1(Suppl 1):S7. doi: 10.1186/1471-2105-7-S1-S7. PubMed PMID: 16723010; PubMed Central PMCID: PMCPMC1810318.

46. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008;9:461. doi: 10.1186/1471-2105-9-461. PubMed PMID: 18959772; PubMed Central PMCID: PMCPMC2630331.

47. de Matos Simoes R, Emmert-Streib F. Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS One. 2011;6(12):e29279. doi: 10.1371/journal.pone.0029279. PubMed PMID: 22242113; PubMed Central PMCID: PMCPMC3248437.

48. Nacher JC, Akutsu T, editors. Analysis on critical nodes in controlling complex networks using dominating sets. 2013 International Conference on Signal-Image Technology & Internet-Based Systems; 2013: IEEE.

49. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-72. doi: 10.1038/s41590-018-0276-y. PubMed PMID: 30643263; PubMed Central PMCID: PMCPMC6340744.

50. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576-89. doi: 10.1016/j.molcel.2010.05.004. PubMed PMID: 20513432; PubMed Central PMCID: PMCPMC2898526.

51. Hubert L, Arabie P. Comparing partitions. Journal of classification. 1985;2:193-218.

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 and unsupervised clustering. (A) The resulting clusters obtained by BootCellNet and their annotations described in the main text were shown. (B) Annotation of cells obtained by SingleR with the fine HumanCellAtlas dataset. (C) Clustering result obtained by Louvain clustering implemented in Seurat, with a resolution of 1.0. (D) Heatmap of average expression of MDS genes in each cluster obtained by BootCellNet. Color codes for columns are the

896 same as those in panel A.

897

898 **Figure 6. BootCellNet facilitates the inference of the course of hematopoiesis.** (A)
899 GRN and its MDS (red squares) identified by BootCellNet. (B) The resulting cell clusters.
900 (C) Heatmap of average expression of MDS genes in each cluster obtained by
901 BootCellNet. Color codes for columns are the same as those in panel B. (D) Lineage
902 curves were obtained by Slingshot using the clusters by BootCellNet. One lineage curve,
903 which spans from Cluster 1 to Cluster 9 is omitted, since Cluster 9 encompasses lymphoid
904 cells.

905

906 **Figure 7. BootCellNet2 and its validation.** (A) Outline of the procedure. Upto the GRN
907 construction, the procedures are the same. From the constructed GRN, the unique set of
908 control nodes consisting of critical nodes and intermittent nodes was computed. Cell
909 clumps are obtained in the same manner as BootCellNet. Averaged expression levels of
910 genes in the unique set are calculated and used for the clustering of cell clumps. Here the
911 clustering was done with multiscale bootstrap to obtain highly confident clusters. The
912 resulting high-confident clustering results are compared to clustering with varying
913 numbers of clusters. The optimal number of clusters is determined as the lowest number
914 of clusters with maximal similarities. (B) Number of appearances of genes in the unique
915 set of control nodes or MDS in 10 specimens. No statistically significant difference was
916 observed by Welch's two-sided *t*-test. (C) Adjusted Rand indexes calculated among 10
917 automated clusterings with (left) or without (right) using the unique set of control nodes
918 are shown. The numbers above the heatmaps represent the number of clusters determined
919 by the BootCellNet2 procedure. (D) Adjusted Rand indexes calculated among 10

920 automated clusterings with (left) or without (middle) using the unique set of control nodes,
921 along with those without using either automated clustering and the unique set (right) are
922 shown. *p < 5x10^-5 by Welch's two-sided *t*-test.
923
924 **Figure 8. Identification of cell type clusters by BootCellNet2.** (A) GRN and its unique
925 set of control nodes (critical nodes as red squares, intermittent nodes as red circles)
926 identified by BootCellNet2 are shown. (B) The resulting cell clusters. Annotations for
927 each cluster described in the main text were shown. (C) Heatmap of average expression
928 of genes in the unique set in each cluster obtained by BootCellNet2. Color codes for
929 columns are the same as those in panel B.
930
931 **Figure 9. BootCellNet2 gives clustering that comply more with CITE-seq results.** (A)
932 The cell clusters obtained by BootCellNet2. Annotations for each cluster according to GO
933 terms, marker genes, and antibody-derived tags were shown. (B) Heatmap of average
934 expression of antibody-derived tags in each cluster obtained by BootCellNet2. Color
935 codes for columns are the same as those in panel A. (C) Heatmap of average expression
936 of the genes in the unique control set in each cluster obtained by BootCellNet2. Color
937 codes for columns are the same as those in panel A. (D) The cell clusters obtained by the
938 Louvain algorithm with a resolution of 0.7. Annotations for each cluster according to GO
939 terms, marker genes, and antibody-derived tags were shown. (E) Heatmap of average
940 expression of antibody-derived tags in each cluster obtained by the Louvain algorithm.
941 Color codes for columns are the same as those in panel D.
942
943 **Figure 10. Identification of cell types in bronchoalveolar lavage fluid scRNA-seq**

data by BootCellNet2. (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. Statistical significances of enrichment of cells from each origin were calculated by Fisher's exact test with Bonferroni correction. Asterisks were shown when $p<5x10^{-4}$.

Supplementary Figure 1. GRN reconstructed from CITE-seq data and its unique set of control nodes (critical nodes as red squares, intermittent nodes as red circles) identified by BootCellNet2 are shown.

Supplementary Figure 2. The original clustering result of BAL scRNA-seq in Liao et al. The result was reproduced from the deposited data and is shown (left). Cell type labeling based on the original paper is shown. As a comparison, the MDS-based clustering (same as Figure 9B) 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.

968

969 **Supplementary Table 4. Marker genes of each MDS-based cluster of cells**
970 **undergoing hematopoiesis.** The output of FindAllMarkers function in Seurat was
971 printed as a table.

972

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

976 **Supplementary Table 6. Marker genes of each BootCellNet2 clustering results of**
977 **PBMCs.** The output of FindAllMarkers function in Seurat was printed as a table.

978

979 **Supplementary Table 7. Gene Ontology terms for marker genes of PBMCs.** Each
980 sheet corresponds to each cluster.

981

982 **Supplementary Table 8. Marker genes of each BootCellNet2 clustering results of**
983 **PBMCs based on CITE-seq.** The output of FindAllMarkers function in Seurat was
984 printed as a table.

985

986 **Supplementary Table 9. Gene Ontology terms for marker genes as in Supplementary**
987 **Table 8.** Each sheet corresponds to each cluster.

988

989 **Supplementary Table 10. Marker genes of each Louvain clustering results of PBMCs**
990 **based on CITE-seq.** The output of FindAllMarkers function in Seurat was printed as a
991 table.

992

993 **Supplementary Table 11. Gene Ontology terms for marker genes as in**
994 **Supplementary Table 10.** Each sheet corresponds to each cluster.

995

996 **Supplementary Table 12. Marker genes of each BootCellNet2 cluster of BAL cells**
997 **from COVID-19 patients.** The output of FindAllMarkers function in Seurat was printed
998 as a table.

999

1000 **Supplementary Table 13. Gene Ontology terms for marker genes of BAL cells from**
1001 **COVID-19 patients.** Each sheet corresponds to each cluster.

1002

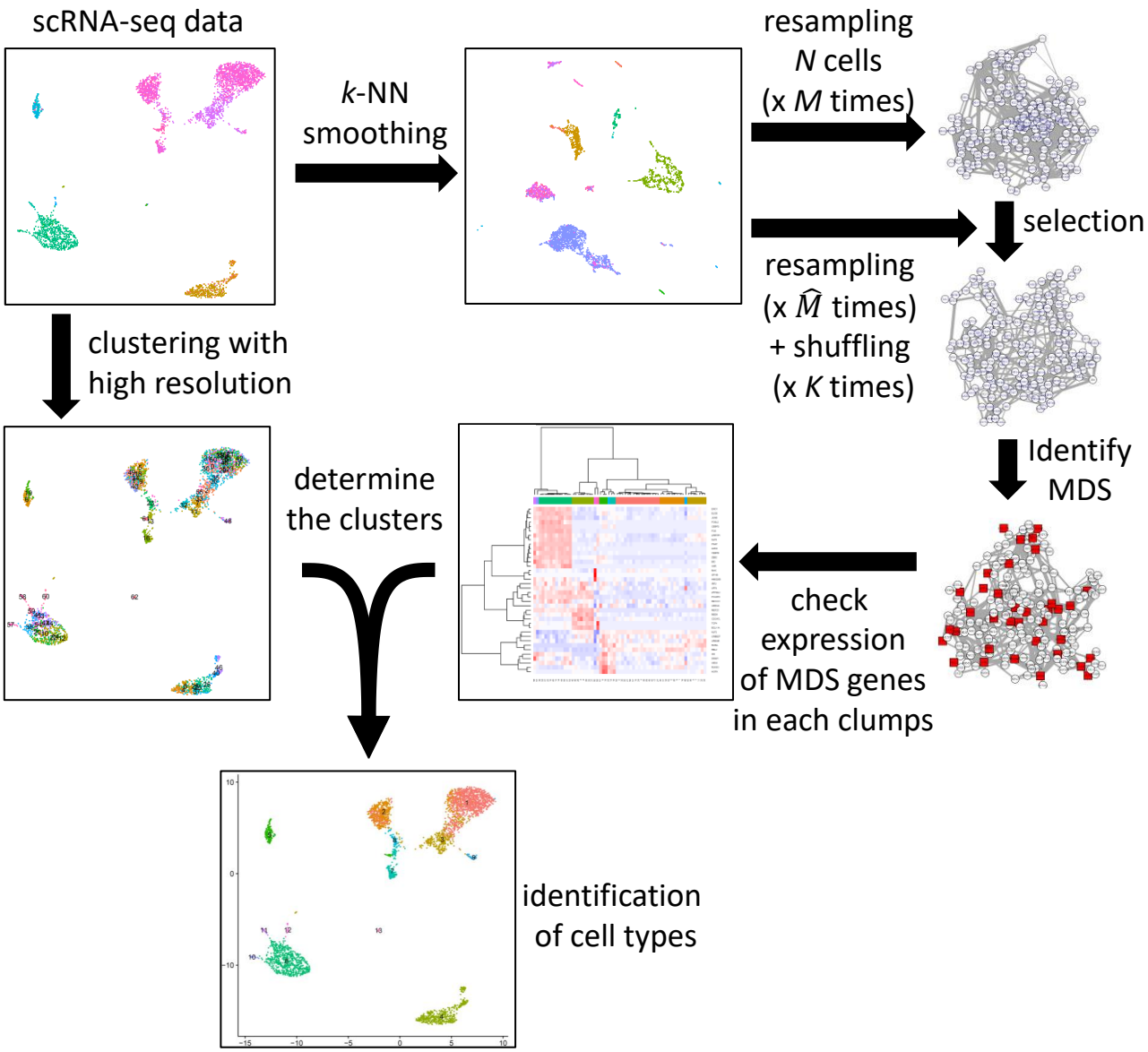
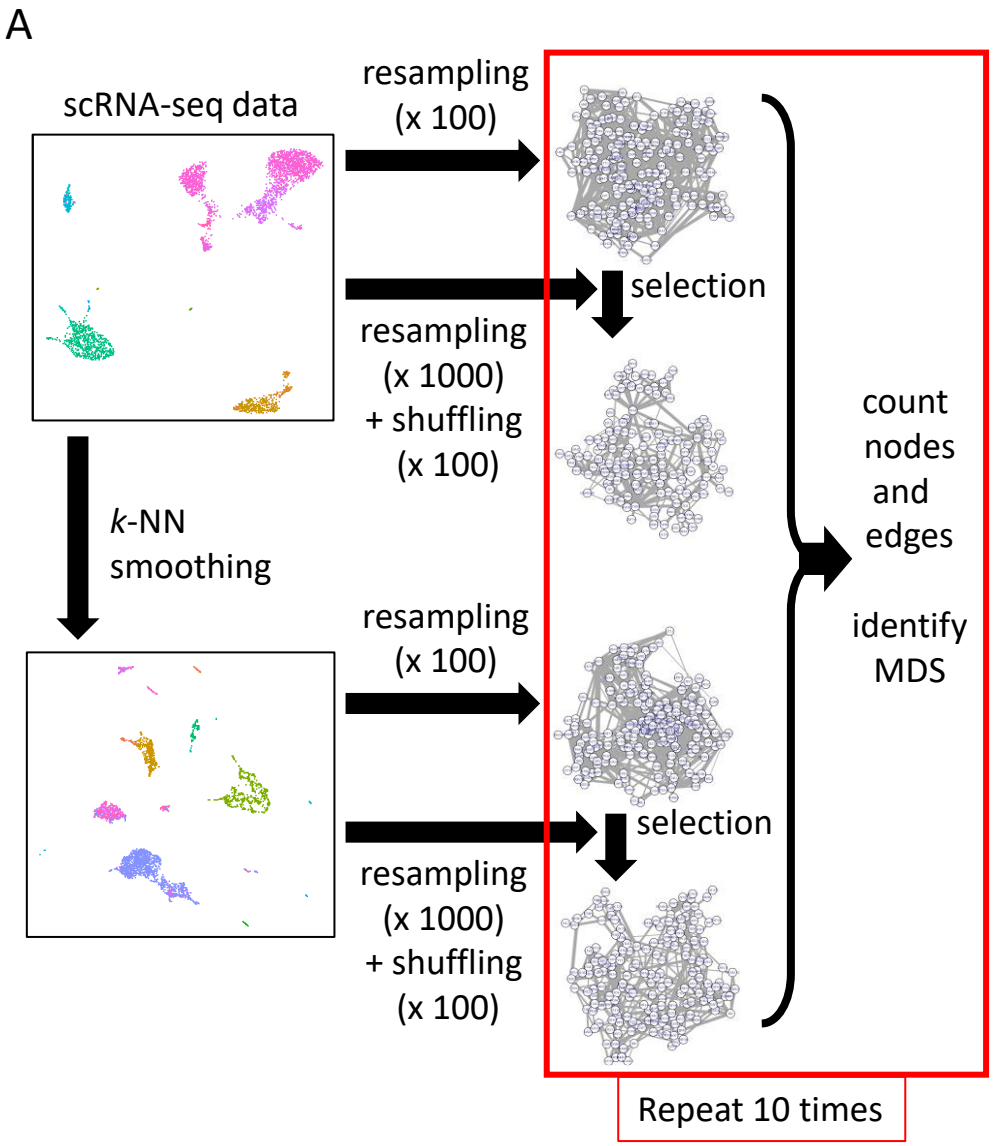
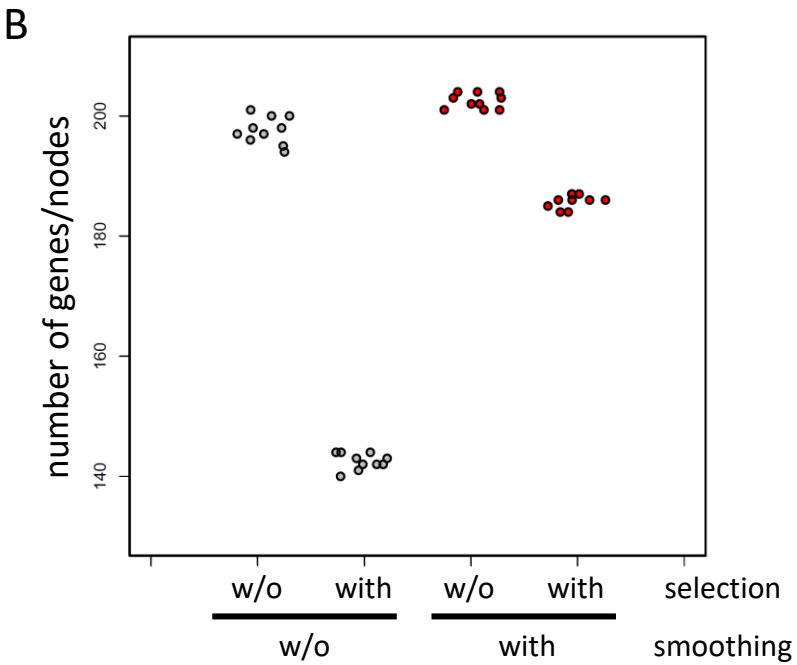

scRNA-seq data
k-NN
smoothing
resampling
N cells
(x M times)
selection
resampling
(x $\widehat{M}$ times)
+ shuffling
(x K times)
clustering with
high resolution
Identify
MDS
determine
the clusters
check
expression
of MDS genes
in each clumps
identification
of cell types

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

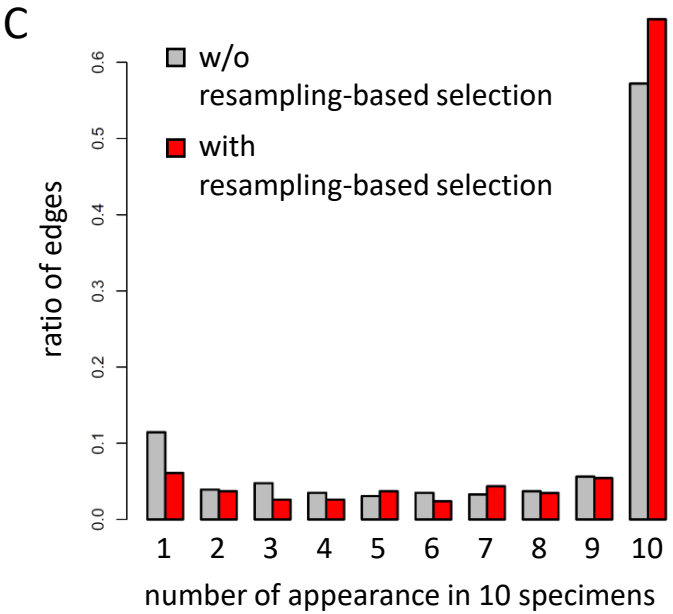

C
w/o
resampling-based selection
with
resampling-based selection
ratio of edges
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
number of appearance in 10 specimens

Figure 2

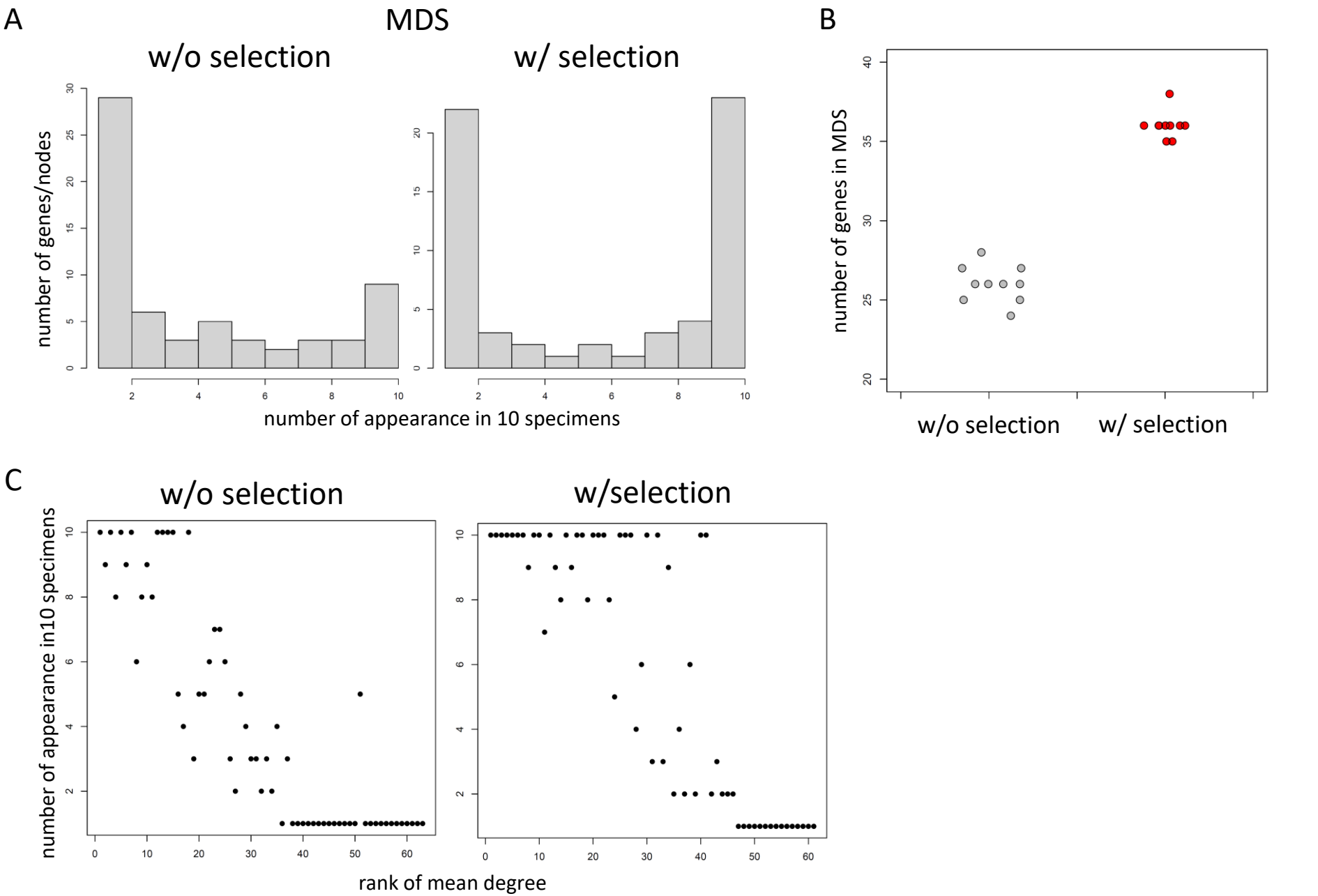

A
MDS
w/o selection
w/ selection
number of genes/nodes
number of appearance in 10 specimens
B
number of genes in MDS
w/o selection
w/ selection
C
w/o selection
w/selection
number of appearance in10 specimens
rank of mean degree

Figure 3

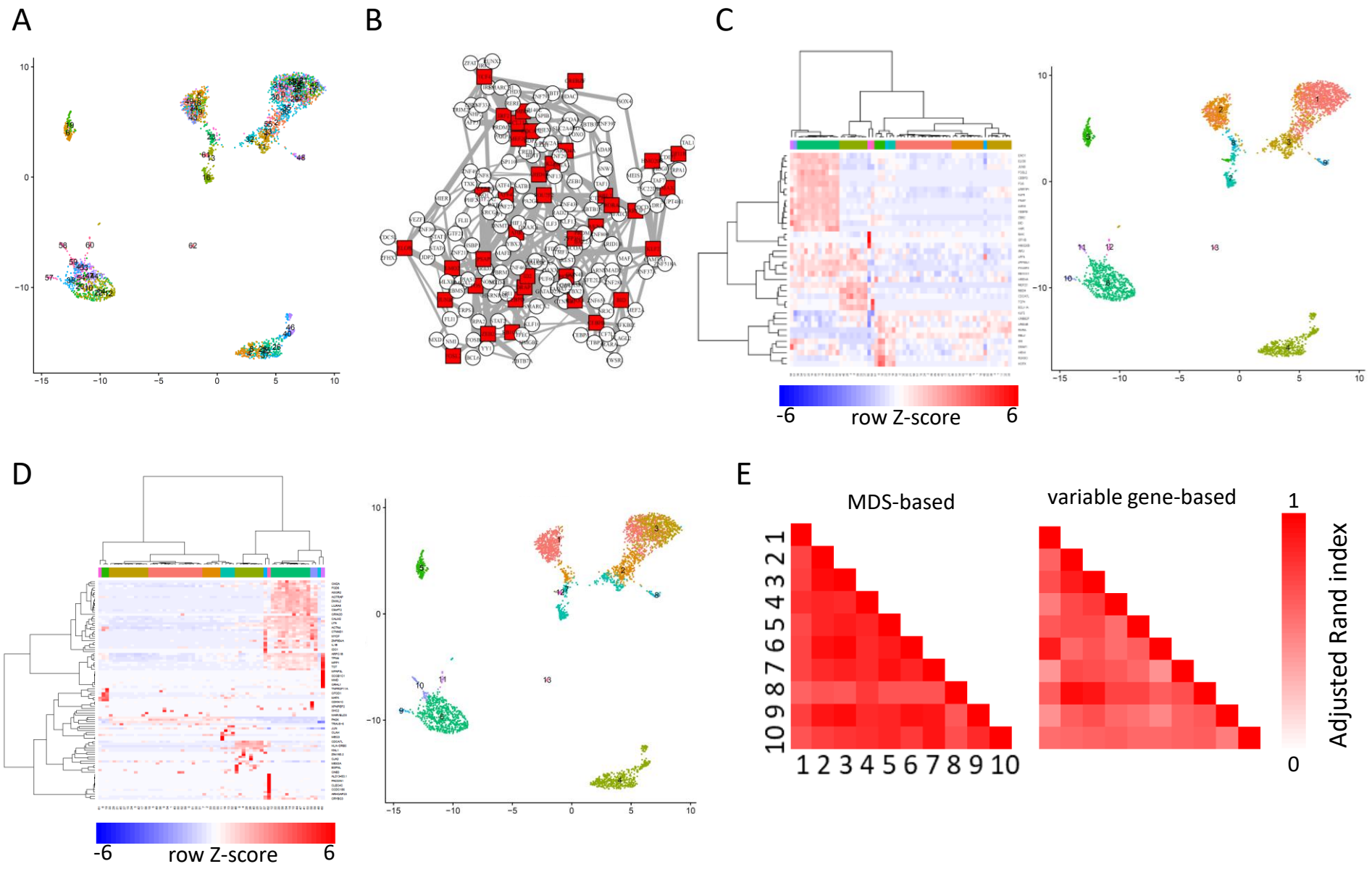

A
B
C
D
E
-6
row Z-score
6
-6
row Z-score
6
MDS-based
variable gene-based
1
0
Adjusted Rand index
1 2 3 4 5 6 7 8 9 10
10 9 8 7 6 5 4 3 2 1

Figure 4

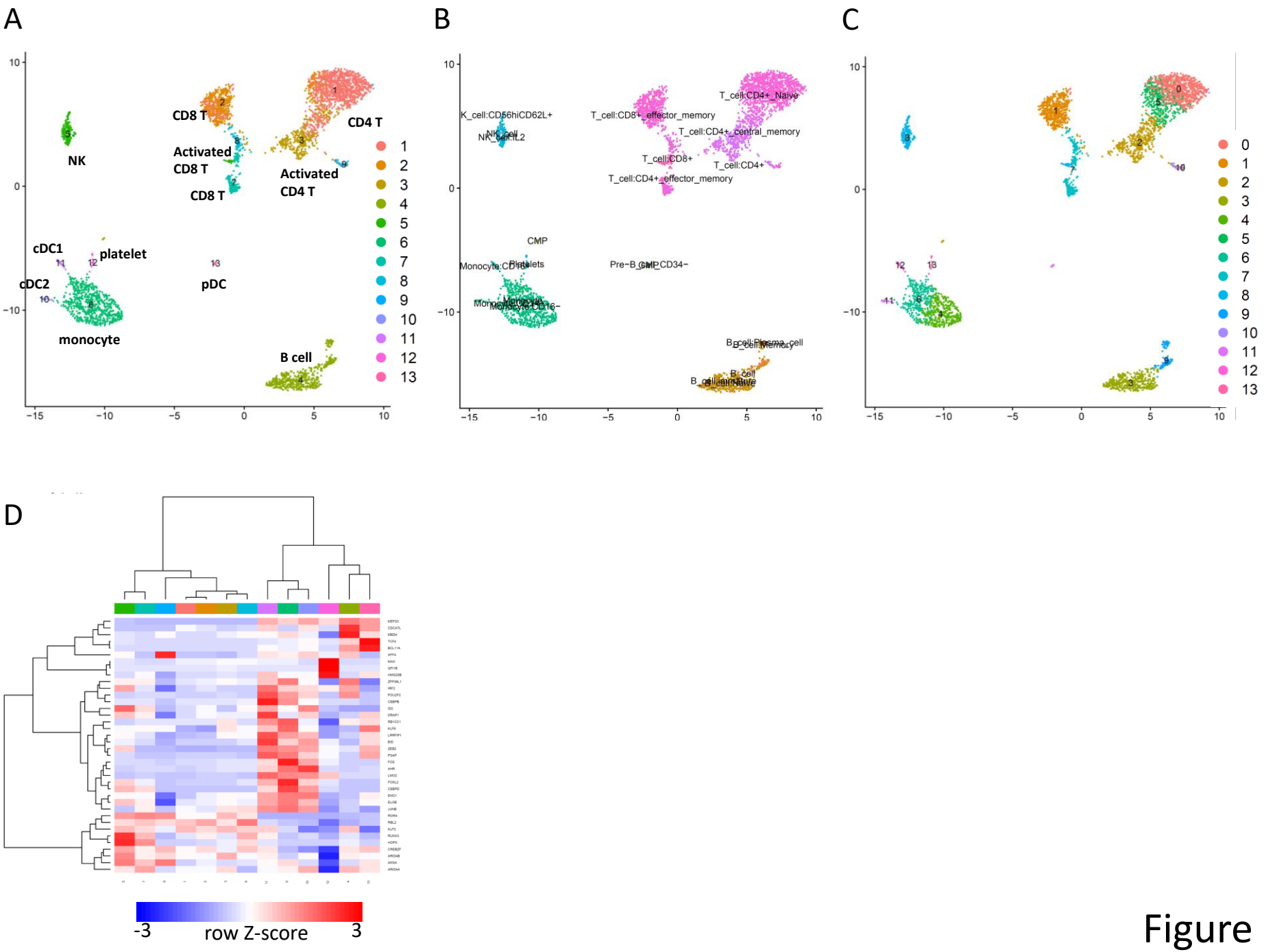

A
B
C
D
NK
CD8 T
Activated CD8 T
CD8 T
CD4 T
Activated CD4 T
cDC1
platelet
cDC2
pDC
monocyte
B cell
K_cell:CD56hiCD62L+
T_cell:CD8+_effector_memory
T_cell:CD4+_Naive
T_cell:CD4+_central_memory
T_cell:CD8+
T_cell:CD4+
T_cell:CD4+_effector_memory
CMP
Pre-B_cell_CD34-
B_cell
-3
row Z-score
3

Figure 5

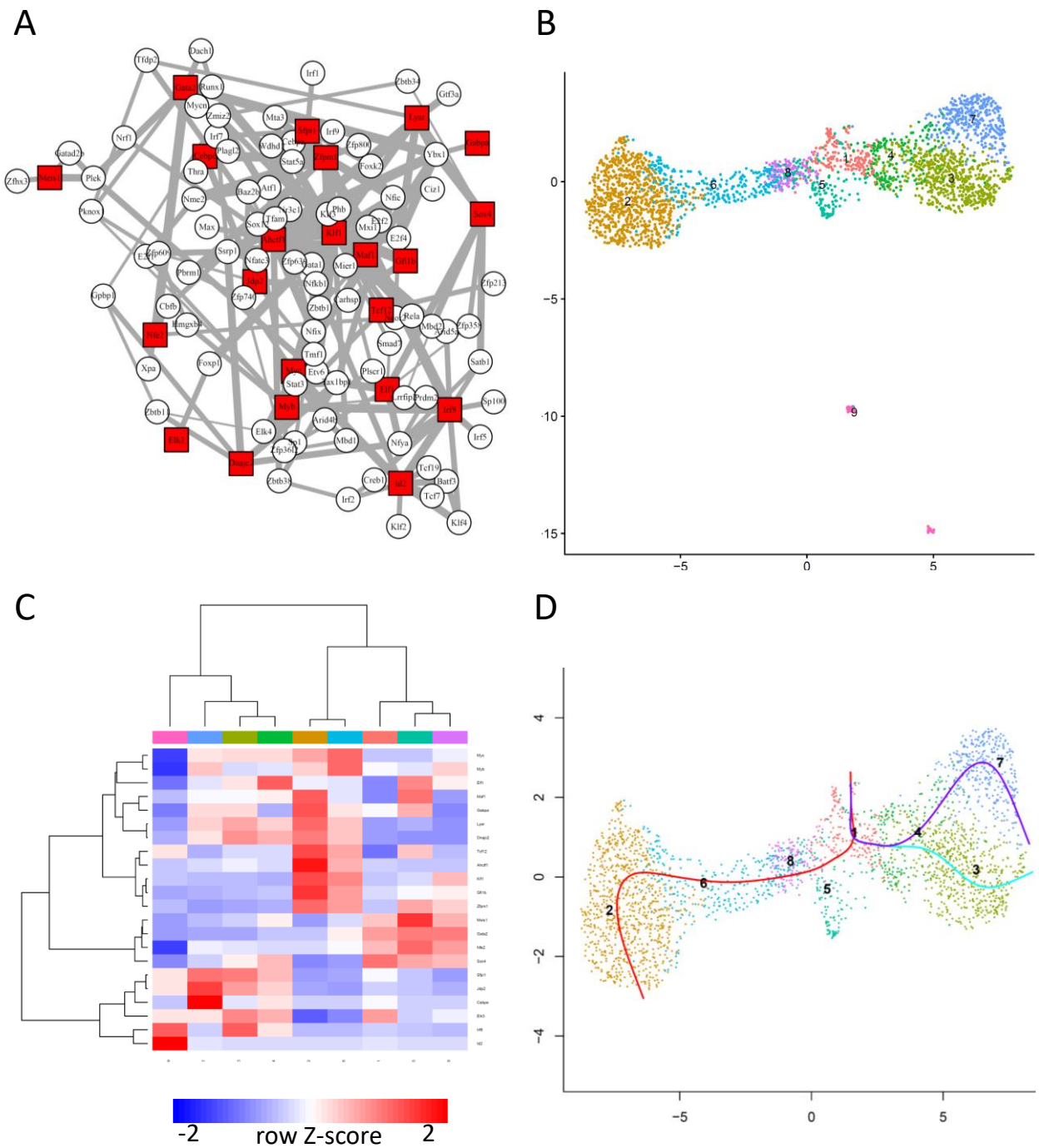


Figure 6

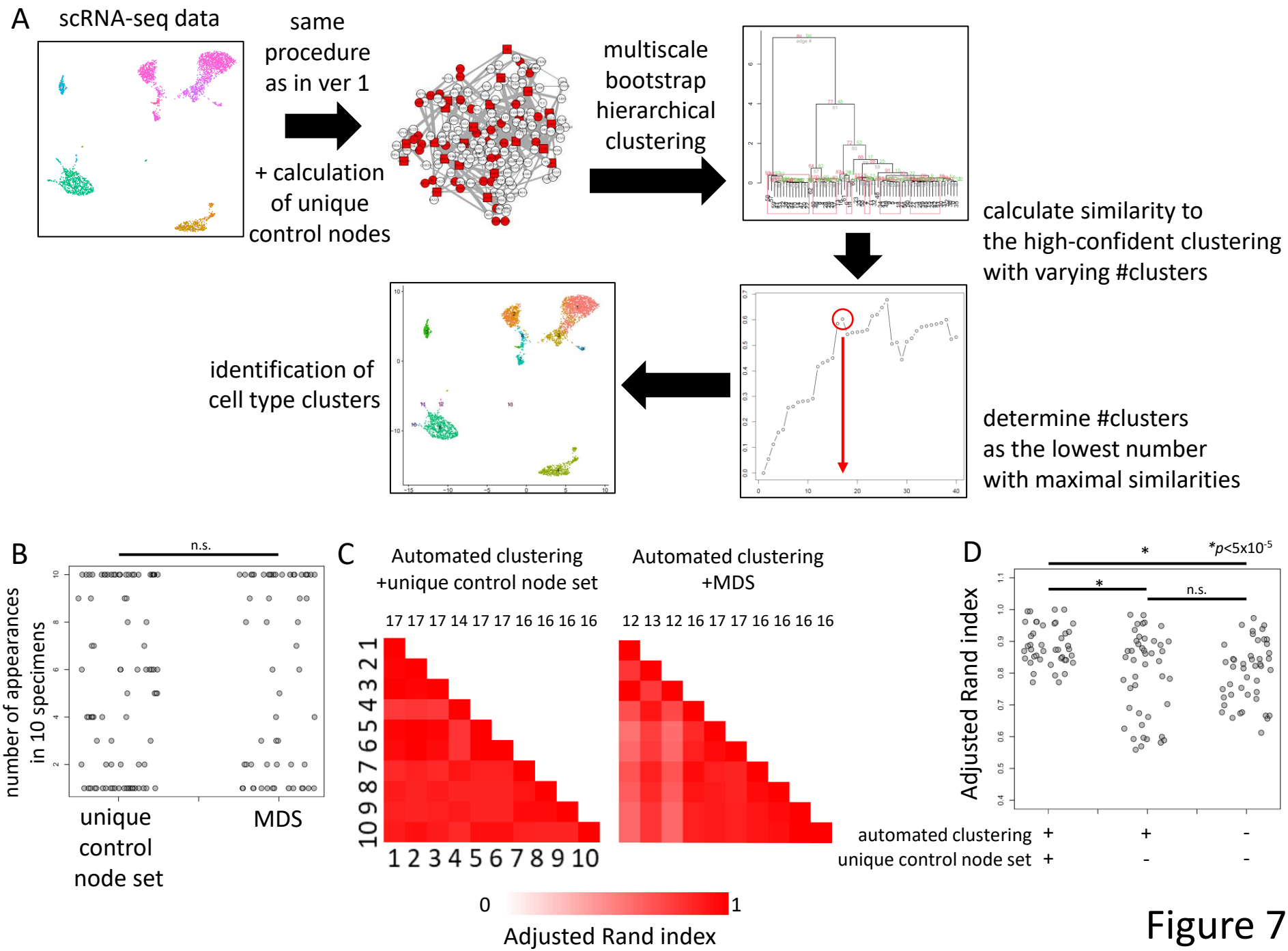
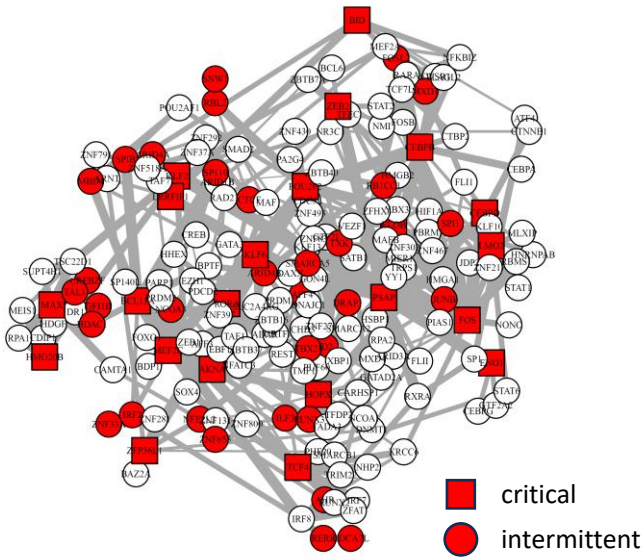
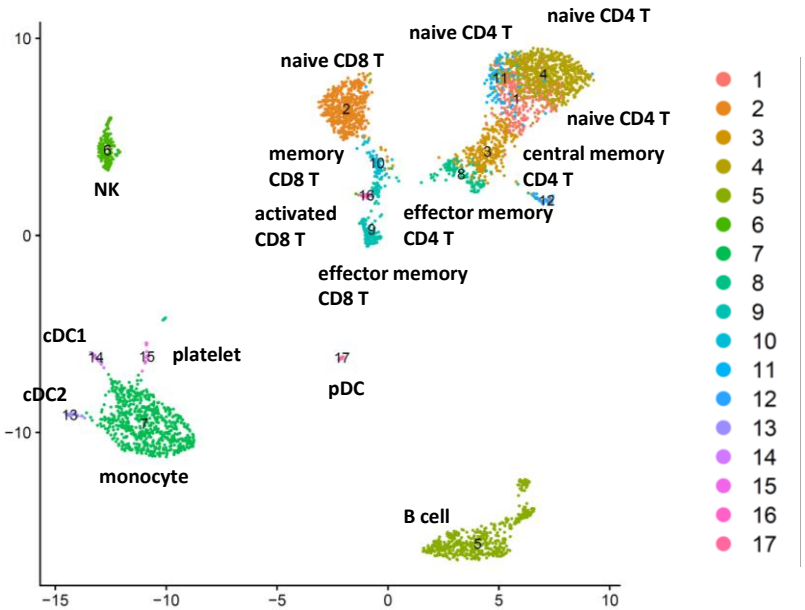

A
scRNA-seq data
same procedure as in ver 1
+ calculation of unique control nodes
multiscale bootstrap hierarchical clustering
calculate similarity to the high-confident clustering with varying #clusters
determine #clusters as the lowest number with maximal similarities
identification of cell type clusters
B
n.s.
number of appearances in 10 specimens
unique control node set
MDS
C
Automated clustering +unique control node set
Automated clustering +MDS
17 17 17 14 17 17 16 16 16 16
12 13 12 16 17 17 16 16 16 16
1 2 3 4 5 6 7 8 9 10
0
1
Adjusted Rand index
D
*
*p<5x10-5
*
n.s.
Adjusted Rand index
automated clustering + + -
unique control node set + - -

Figure 7

A


critical
intermittent

B


naive CD4 T
naive CD4 T
naive CD8 T
naive CD4 T
memory CD8 T
central memory CD4 T
NK
activated CD8 T
effector memory CD4 T
effector memory CD8 T
cDC1
platelet
pDC
cDC2
monocyte
B cell
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

C

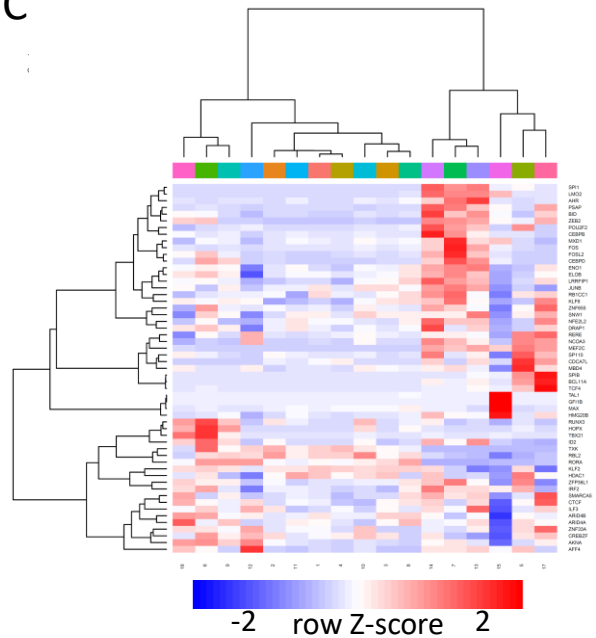

-2
row Z-score
2

Figure 8

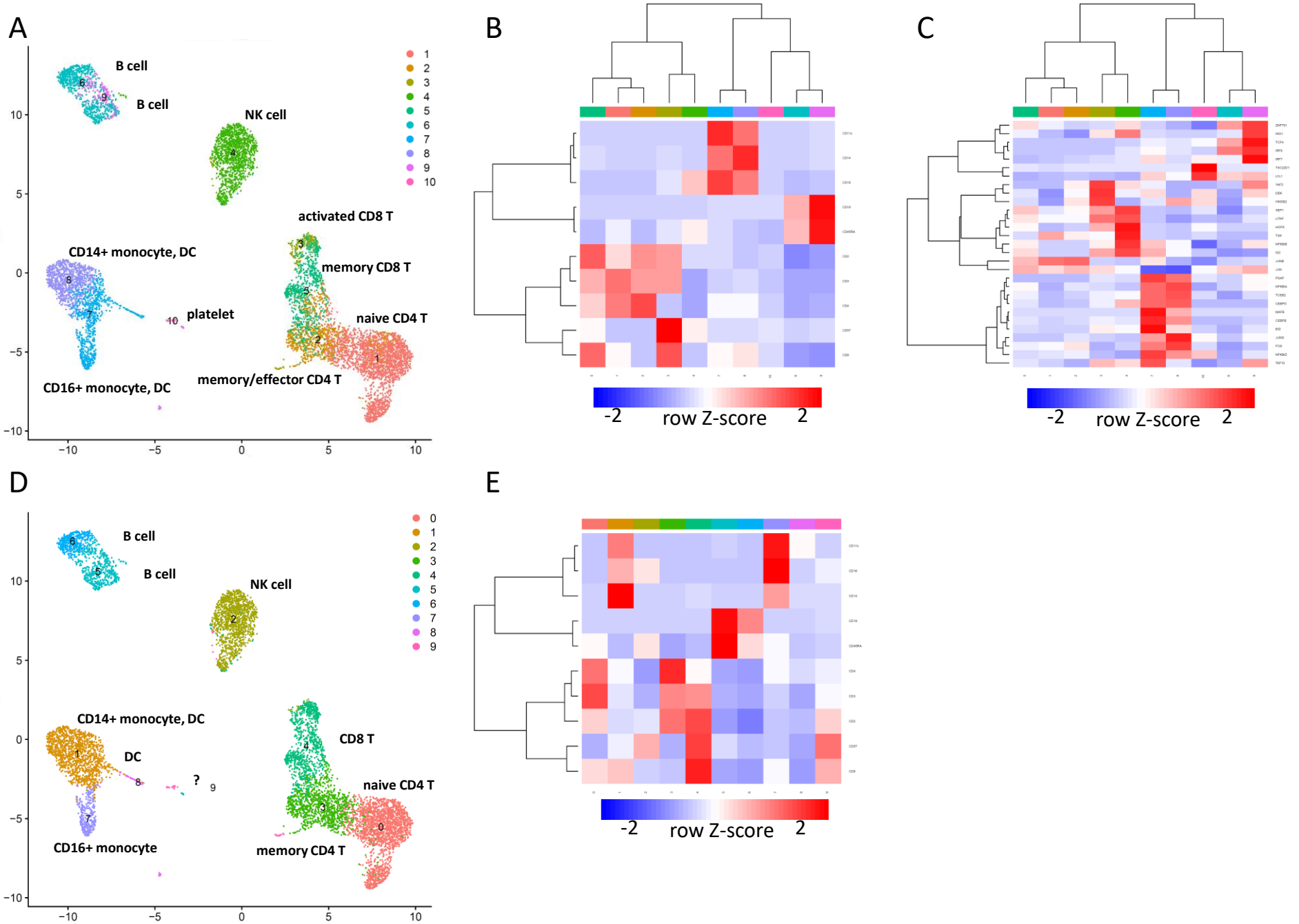

A
B cell
B cell
NK cell
activated CD8 T
CD14+ monocyte, DC
memory CD8 T
platelet
naive CD4 T
CD16+ monocyte, DC
memory/effector CD4 T
1
2
3
4
5
6
7
8
9
10
B
-2
row Z-score
2
C
-2
row Z-score
2
D
B cell
B cell
NK cell
CD14+ monocyte, DC
CD8 T
DC
?
naive CD4 T
CD16+ monocyte
memory CD4 T
0
1
2
3
4
5
6
7
8
9
E
-2
row Z-score
2

Figure 9

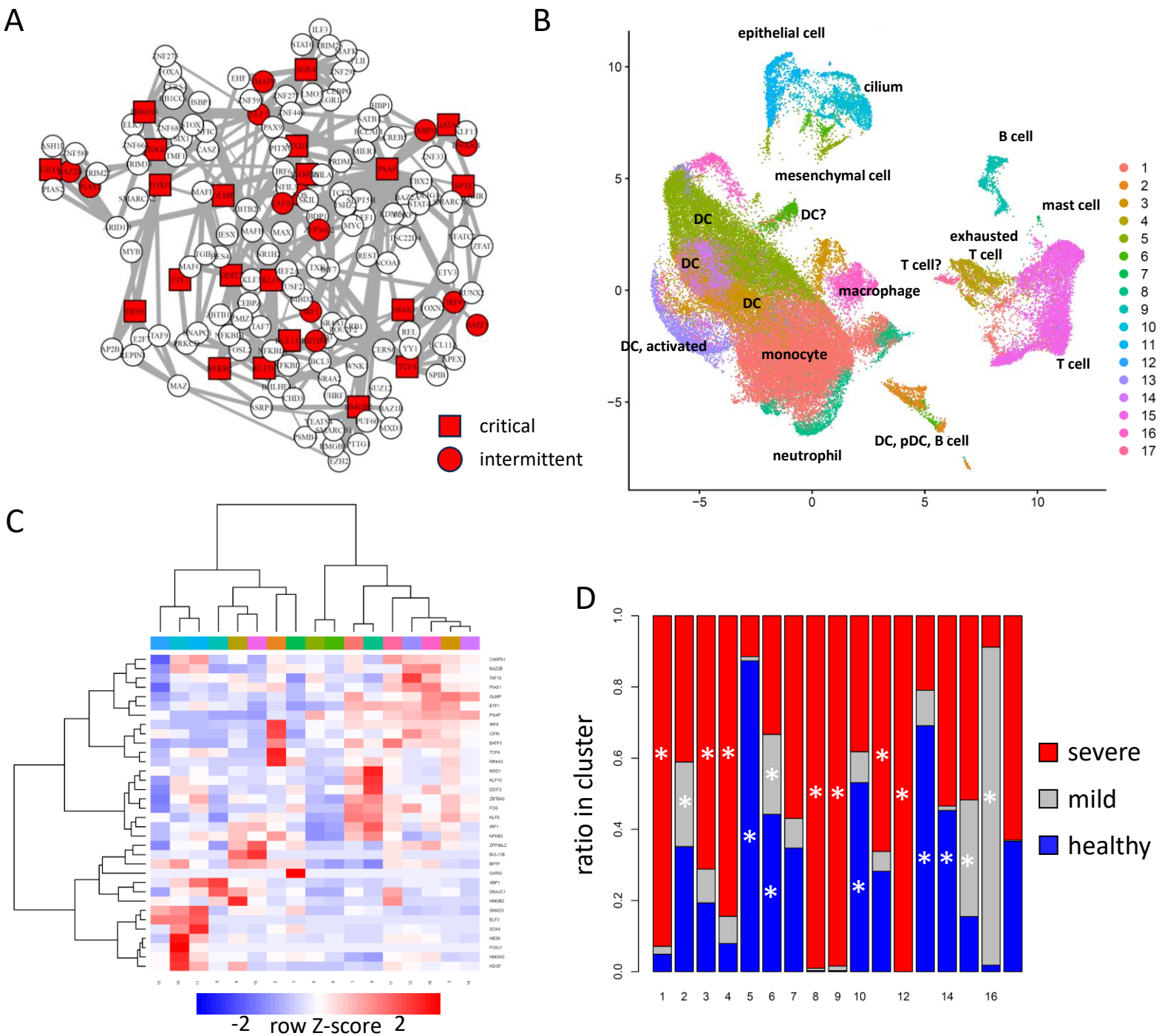
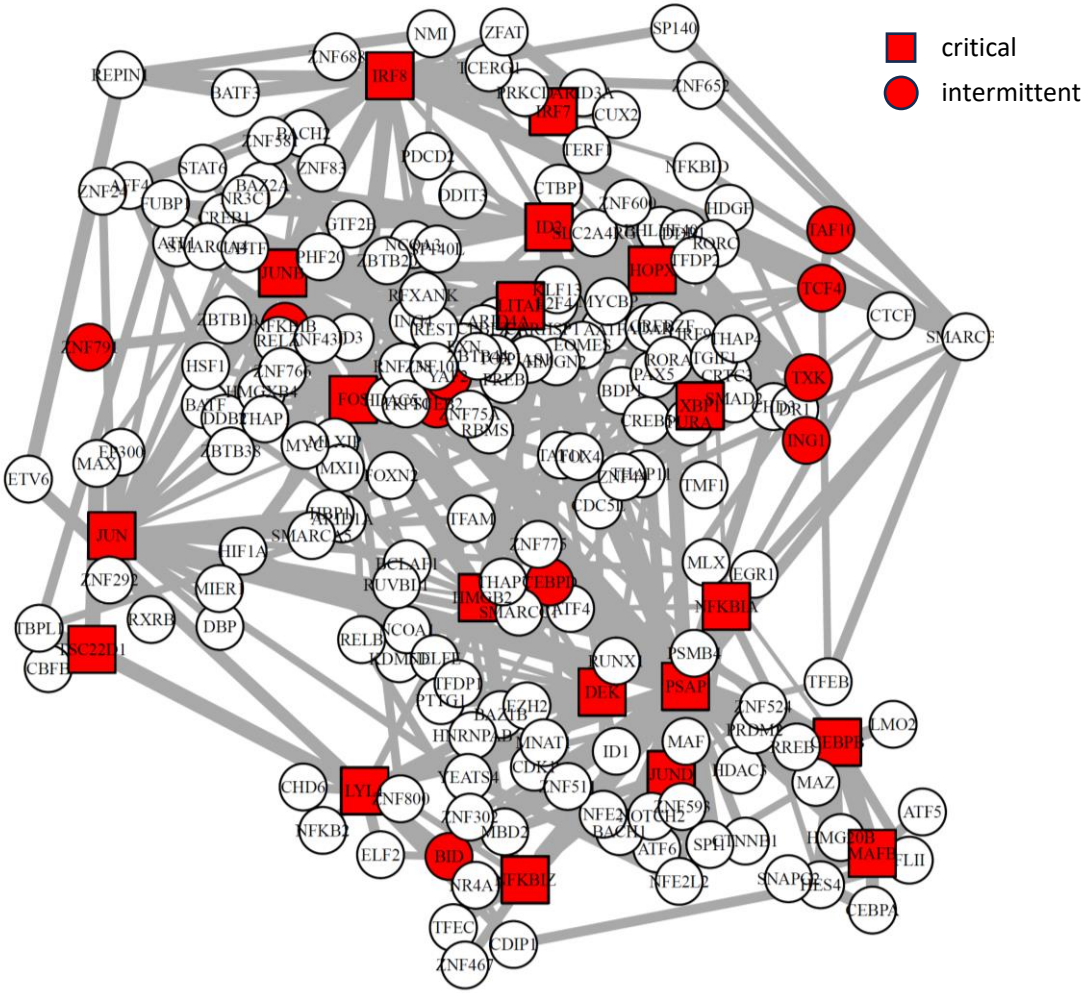
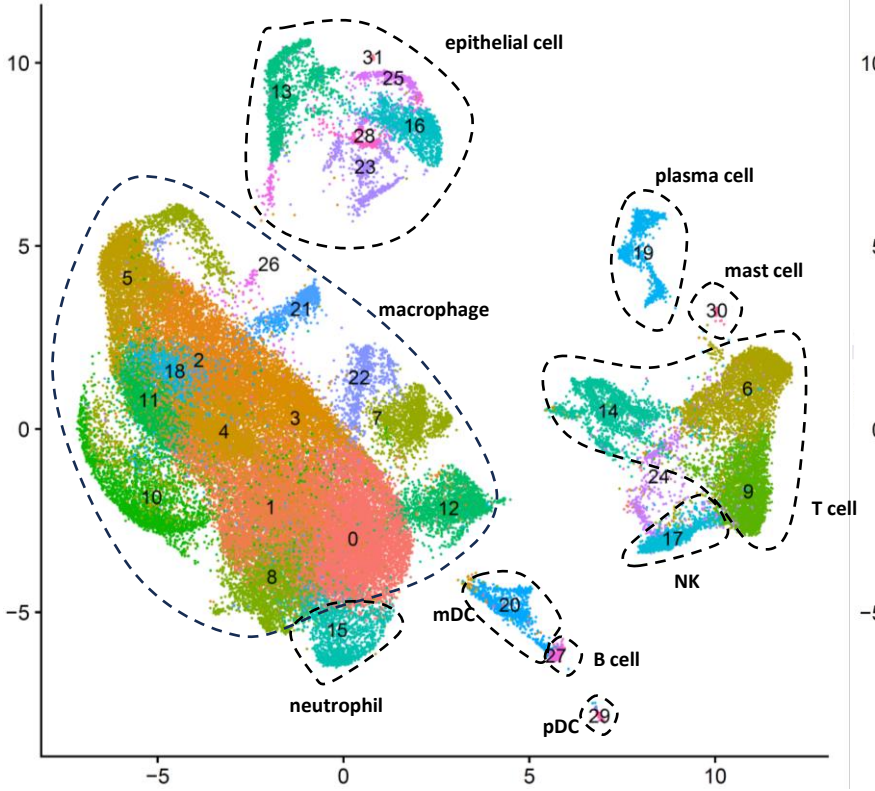

A
critical
intermittent
B
epithelial cell
cilium
mesenchymal cell
B cell
mast cell
DC
DC?
DC
DC
exhausted
T cell
T cell?
macrophage
DC, activated
monocyte
T cell
DC, pDC, B cell
neutrophil
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
C
-2
row Z-score
2
D
ratio in cluster
severe
mild
healthy

Figure 10

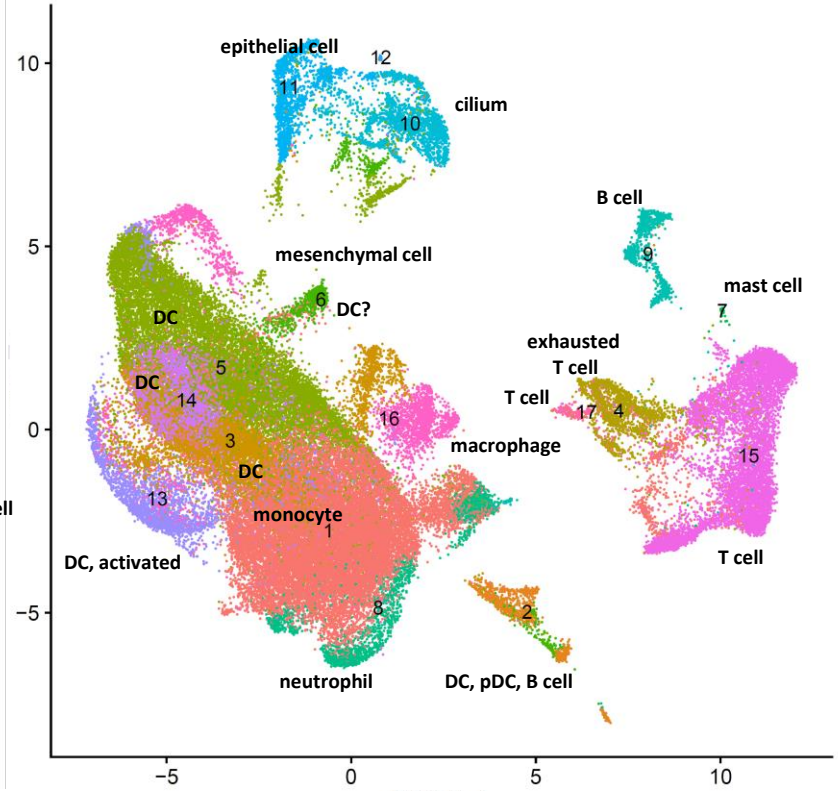


Supplementary Figure 1

Original clustering in Liao et al.



BootCellNet2 (same as Figure 10B)



Supplementary Figure 2