Supplementary Materials for: An Effective Biclustering-based Framework for Identifying Cell Subpopulations from scRNA-seq Data[*]

Qiong Fang[†] Dewei Su, Wilfred Ng, Jianlin Feng

January 19, 2020

[*] **Availability and Implementation:** DivBiclust, including source codes, preprocessed data, experiment results, and a supplementary document, is freely available at https://www.github.com/Qiong-Fang/DivBiclust.

[†] Correspondence concerning this article should be addressed to Qiong Fang. E-mail: fang.qiong@gmail.com

Contents

1 Data Preparation

Ten real scRNA-seq datasets are used in our experiments. Three datasets with smaller cell sets were generated by low-throughput plate-based scRNA-Seq methods. They respectively contain the transcriptomes of *human cancer/somatic cells* [9], the transcriptomes of *human embryonic cells* [14], and the transcriptomes of *human pancreatic islet cells* [7]. Thus we call them the *HCC*, *HEC*, and *HPIC* datasets, respectively. One dataset was generated by Illumina high-throughput sequencing method and contains the transcriptomes of *mouse brain cells* [16]. We call this dataset the *MBC* dataset. The remaining six datasets were generated by the high-throughput InDrop [6] method, which respectively contain the transcriptomes of pancreatic cells from four human donors and two mouse strains [1]. We denote these six datasets as the *HUM1*, *HUM2*, *HUM3*, *HUM4*, *MOU1*, and *MOU2* datasets, respectively. In this section, we introduce the preprocessing steps of these scRNA-Seq datasets.

1.1 HCC Dataset

The HCC dataset contains the transcriptomes of circulating tumor cells. It was downloaded from the Gene Expression Ominbus (GEO) database with the accession number GSE38495. We follow the preprocessing steps in [13] and only keep the genes with RPKM ≥ 20 to avoid the influence of low-expressed genes. Then, a log transformation is performed to reduce the effect of extreme values. The resulting gene expression matrix contains 86 cells and 4501 genes with 8.1% missing values. The 86 cells respectively belong to 11 classes with the transcriptomes shown in Table 1.

Table 1: Transcriptomes of Human Cancer/Somatic Cells

Cell type	#Samples
Brain	16
Bladder cancer cell line (T24)	4
Embryonic stem cells	8
Melanoma cancer (SKMEL5)	4
Melanoma cancer (UACC)	3
Melanoma-derived circulating tumor cells (CTC)	6
Melanocytes	2
Prostate cancer cell line (PC3)	4
Prostate cancer cell (picked from Petri dish)	8
Prostate cancer cell (isolated by EPCAM marker)	11
Universal human reference RNA	20

The abbreviations for the names of 11 cell types are respectively "*Brain*", "*Bladder(T24)*", "*ESC*", "*Melanoma(SKMEL5)*", "*Melanoma(UACC)*", "*Melanoma (CTC)*", "*Melanocytes*", "*Prostate(PC3)*", "*Prostate*", "*Prostate(EPCAM)*", and "*refRNA*". They are used in the experiment section for simplicity in presentation.

1.2 HEC Dataset

The HEC dataset includes the transcriptomes of human oocytes and early embryos at seven developmental stages plus the primary outgrowth during hESC derivation. It was downloaded from the GEO database with the accession number GSE36552. We also follow the preprocessing steps in [13] and filter out those genes whose log-transformed RPKM under all the cell samples are smaller than 0.1. A gene expression matrix with 124 cells, 20 018 genes and 36.9% missing values is produced. The cells are divided into 9 classes and the transcriptome details are presented in Table 2.

Table 2: Transcriptomes of Human Embryonic Cells

Cell type	#Samples
Oocyte	3
Zygote	3
2-cell	6
4-cell	12
8-cell	20
Morulae	16
Late-blastocyst	30
hESC passage 0	8
hESC passage 10	26

1.3 HPIC Dataset

The HPIC dataset includes the expression profiles of endocrine and exocrine cells in human pancreatic islets. We obtained a preprocessd HPIC expression matrix from Lin et al. [8], in which undefined cells and bulk RNA-seq samples were excluded. It consists of 60 cells and 180 253 transcripts, and contains 77.8% missing values. The 60 cells are divided into 6 classes with the transcriptome details shown in Table 3.

Table 3: Transcriptomes of Human Pancreatic Islet Cells

Cell type	Biological hormone	#Samples
Acinar	NA	11
Alpha	Glucagon	18
Beta	Insulin	12
Delta	Somatostatin	2
Duct	NA	8
PP	Pancreatic Polypeptide	9

1.4 MBC Dataset

The MBC dataset contains the transcriptomes of the cells in the mouse somatosensory cortex and hippocampal CA1 region [16]. It was downloaded from the GEO database with the accession number GSE60361. We follow the preprocessing method in [16] and select 5000 genes which are regarded the most informative for further analysis. Consequently, an expression matrix with $3,005$ cells, $5,000$ genes, and 84.2% missing values is generated. The cells belong to 9 classes and the transcriptome details are shown in Table 4.

Table 4: Transcriptomes of Mouse Brain Cells

Cell type	#Samples
Interneurons	290
S1 Pyramidal	390
CA1 Pyramidal	948
Oligodendrocytes	820
Microglia	98
Endothelial	175
Astrocytes	198
Ependymal	26
Mural	60

1.5 HUM and MOU Datasets

The four HUM datasets and two MOU datasets respectively contain the transcriptomes of the pancreas islets cells from four human donors and two mouse strains [1]. They were downloaded from the GEO database with the accession number GSE84133. The cells in each HUM dataset belong to 14 classes and the cells in each MOU dataset belong to 13 classes. The transcriptome details of the six datasets are listed in Tables 5 and 6, respectively.

Table 5: Transcriptomes of four HUM datasets

Cell type	# Samples			
	HUM1	HUM2	HUM3	HUM4
Alpha	236	676	1130	284
Beta	872	371	787	495
Gamma	70	86	36	63
Delta	214	125	161	101
Acinar	110	3	843	2
Ductal	120	301	376	280
Activated_stellate	51	81	100	52
Quiescent_stellate	92	22	54	5
Schwann	5	6	1	1
Endothelial	130	23	92	7
Macrophage	14	17	14	10
Mast	8	9	7	1
T_cell	2	2	2	1
Epsilon	13	2	2	1

Table 6: Transcriptomes of two MOU datasets

Cell type	# Samples	
	MOU1	MOU2
Alpha	9	182
Beta	343	551
Gamma	14	27
Delta	85	133
Ductal	236	39
Activated_stellate	4	10
Quiescent_stellate	29	18
Schwann	3	3
Endothelial	72	67
Macrophage	17	19
B_cell	2	8
T_cell	4	3
Immune_other	4	4

We first take a simple preprocessing strategy by filtering out the genes with zero expression values under all the cell samples, and generate four datasets with large gene sets and higher dropout rates. On the other hand, gene selection has become an indispensable preprocessing step in many cell subpopulation identification methods, which filters out those low-variance or low-abundance genes before the cell clustering analysis is performed. Therefore, we adopt the gene selection strategy in [10] and keep top 10% of the genes which have the maximum count across all the cells. Accordingly, four datasets with smaller gene sets and lower dropout rates are generated. We call the datasets generated by the first preprocessing strategy as the HUM1_NZ, HUM2_NZ, HUM3_NZ, and HUM4_NZ datasets, respectively, and use HUM1, HUM2, HUM3, and HUM4 to refer to the datasets generated by the second preprocessing strategy. The statistics of the datasets generated by two different preprocessing strategies are listed in Table 7. Because some of the related methods we compare with cannot deal with the datasets with larger gene sets, the experiment results presented in the paper are conducted on the datasets generated by the second preprocessing strategy. We show the experiment results on the NZ datasets in Section 4 in this document.

Table 7: Statistics of the HUM and MOU Datasets

Dataset	# Cells	#Classes	nonzero data (NZ)		Top 10% data	
			# Genes	Dropout rate	# Genes	Dropout Rate
HUM1	1,937	14	16,016	88.0%	2,012	63.7%
HUM2	1,724	14	16,310	88.4%	2,012	64.1%
HUM3	3,605	14	16,678	89.5%	2,012	63.6%
HUM4	1,303	14	15,720	86.0%	2,012	59.0%
MOU1	822	13	14,359	90.1%	1,487	67.9%
MOU2	1,064	13	14,646	87.6%	1,487	67.9%

1.6 Analysis of Data Noise

We conduct an analysis over the datasets to study their noise levels and the results are used to help determine a proper setting for the similarity threshold α used in the pattern mining phase. The analysis consists of three steps as follows.

- For each cell c in the dataset, a set T of 100 genes are selected whose expression levels under c are maximal. The vector of the expression values of genes in T under the cell c is called the expression profile of c over T.
- Then, the distance between the cell c and every other cell is computed, and the distance function is defined as follows.

$$D(c, c_i) = \frac{1}{|T|} \sum_{t_j \in T} \frac{|M(c, t_j) - M(c_i, t_j)|}{M(c, t_j)}. \tag{1}$$

- In the dataset with high dropout rate, it is possible that the expression values of genes in T under a certain cell c_i are all zeros. Such cells never have the chance to be covered by the bicluster with cell c as the seed. Thus, we only take into accounts the cells whose expression profiles over T contain at least 50% valid entries. The noise level of a cell c is the average of the distance between c and every other cell, and the noise level of a dataset is the average of the noise levels of all the cells in the data.

The distance function is formulated based on the definition of the sc-bicluster and is similar with the computation of α. Thus, the noise level of a dataset can be a reference for setting the similar threshold α.

Figure 1 shows the noise levels of all the sixteen datasets, including the nonzero versions of the four HUM and two MOU datasets.

2 Experiment Results on the MBC Dataset

2.1 Impact of Thresholds

According to the noise analysis in Section 1.6, the noise level of the MBC dataset is 0.463 which is the highest among all the sixteen datasets. The high noise level of the MBC dataset is not only

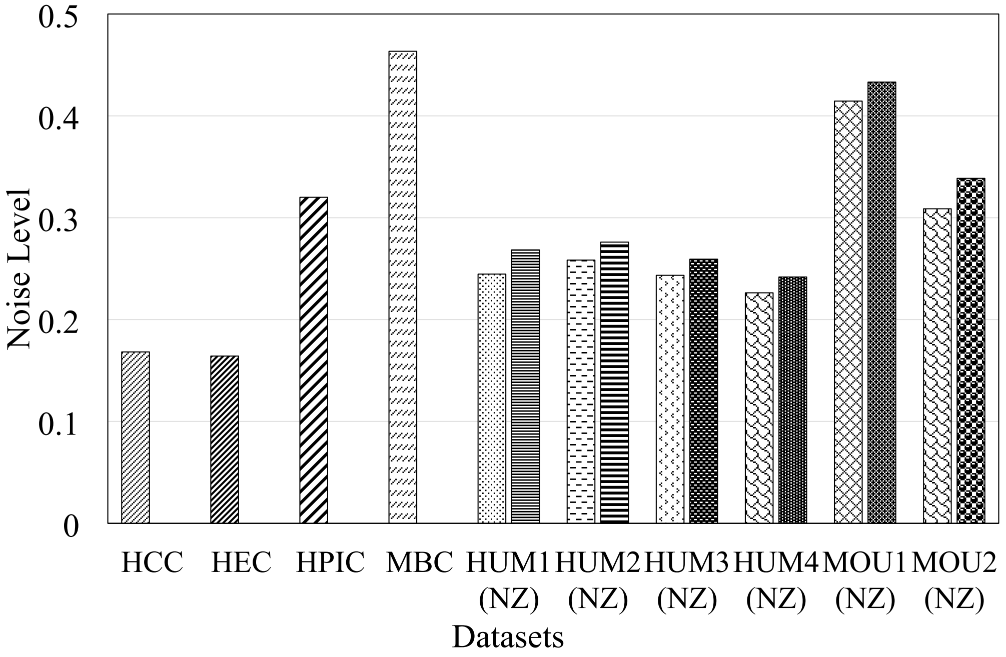


Figure 1: Noise Levels of Sixteen scRNA-seq Datasets

due to the large fraction of dropouts but also caused by the high variability of the valid expression values.

To study the effect of the thresholds to the clustering accuracy, we first fix the dropout threshold γ to 0 and set the similarity threshold α to be close to the noise level of the dataset. Then, we set a nonzero dropout threshold, i.e., 0.1, and take more strict settings for α. The similarity threshold β for pattern merging is fixed to 0.5. The ARI scores of *DivBiclust* on the MBC dataset with respect to the size of initial gene set $|T_s|$ are presented in Figure 2. The ARI scores under four settings all remain stable and exhibit a small decrease when $|T_s|$ gets larger than 70. The best ARI score is 0.654 which is achieved when $\alpha = 0.45$, $\gamma = 0$, and $|T_s| = 20$. When $\alpha = 0.35$, $\gamma = 0.1$, and $|T_s| = 60$, the ARI score is 0.622. Therefore, when the settings of α and γ match the noise level of the dataset, the clustering results with high enough ARI scores can be generated under different combinations of the threshold values. When γ is set to 0, the cells covered by the same bicluster should have similar enough expression profiles over a subset of genes and no missing values are allowed. Therefore, $|T_s|$ can only be set to a small value. When γ is set to 0.1, missing values can exist in the expression profiles of the cells that are covered by the same bicluster, and thus the similarity between the expression profiles of the cells can span a larger gene set and hence a larger $|T_s|$ value.

Next, we study the impact of the similarity threshold β to the clustering accuracy. We vary β from 0.4 to 0.6 and fix α and γ to 0.45 and 0, respectively. The ARI scores with respect to the size of initial gene set are presented in Figures 3. The ARI scores under three β settings exhibit very small difference for each $|T_s|$ value. The ARI scores are all larger than 0.6 when $|T_s|$ is smaller than 50, and slightly decrease when $|T_s|$ further increases. Therefore, the influence of β is limited and thus we fix it to 0.5 as declared in the paper.

2.2 Comparison with Related Methods

The optimal ARI scores achieved by *DivBiclust* and nine state-of-the-art methods on the MBC dataset are listed in Table 8. The ARI score of *DivBiclust* is 0.654 which is the second highest among ten cell clustering methods. SC3 achieves the ARI score of 0.810 when the number of cell

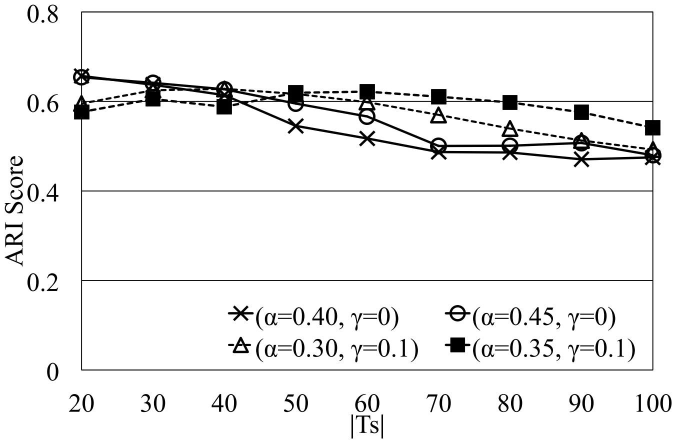


Figure 2: ARI w.r.t. similarity threshold α, dropout threshold γ, and the size of initial gene set $|T_s|$ ($\beta = 0.5$).

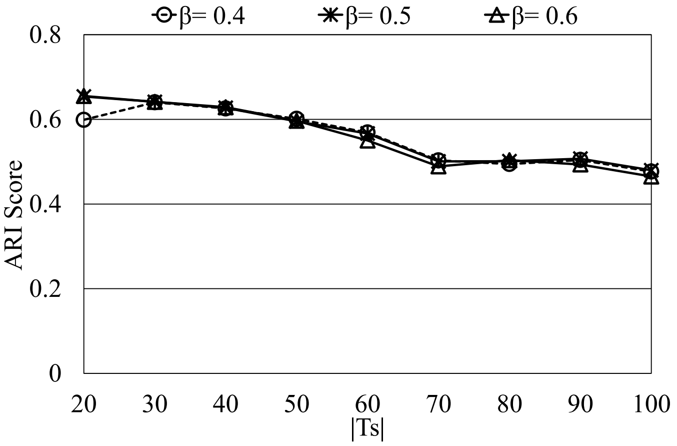


Figure 3: ARI w.r.t. similarity threshold β and the size of initial gene set $|T_s|$ ($\alpha = 0.45$, γ=0).

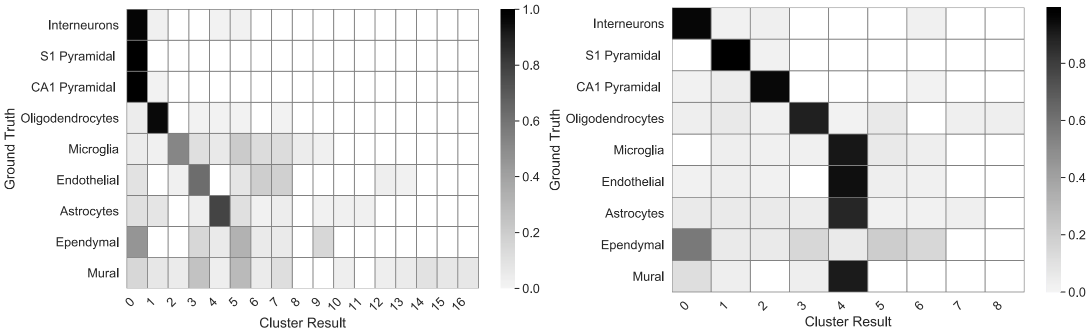
clusters is set to be the same as the number of true cell classes in the MBC data. Its ARI score is 0.507 if the final number of cell clusters is estimated automatically.

We use heatmap to demonstrate the correlations between the clustering result generated by *DivBiclust* and the ground-truth cell classes in Figure 4. The heatmap that visualizes the clustering result generated by SC3 is also presented for comparison. Although *DivBiclust* successfully identifies over 99% of interneuron cells, S1 pyramidal and CA1 pyramidal cells, it fails to separate them into three different clusters. The reason is that the expression profiles of the cells from different classes possess different levels of variability. When we set a more strict similarity threshold α for bicluster mining and pattern merging, the intermediate cell clusters generated before the post-hoc assignment phase cover around 1000 cells, all of which belong to the classes of interneurons, S1 and CA1 pyramidal neurons, and oligodendrocytes. The intermediate clustering result accurately separates the cells of these types and achieves the ARI score over 0.86. However, the cells of other types are failed to be clustered under the strict setting of the similarity threshold due to higher varability of their expression profiles. At the final post-hoc assignment phase, these cells are assigned to the clusters in which the cells of other types are contained. As a result, the accuracy of the final clustering result degrade. When we set a larger α value, i.e., 0.45, 96.1% oligodendrocytes cells and 71.2% astrocytes cells are accurately identified, but the interneuron cells, S1 pyramidal, and CA1 pyramidal cells are clustered into the same group, as shown in Fig-

Table 8: Comparisons of Clustering Accuracy

Methods	MBC Data
DivBiclust	<u>0.654</u>
SNN-Cliq	0.037
BackSPIN	0.610
CIDR	0.104
SC3	**0.810**
pcaReduce	0.469
TSCAN	0.265
Seurat v3	0.375
GiniClust2	0.650
SOUP	0.553

* The largest ARI score over each dataset is highlighted using bolded font, and the second largest one is underlined.



(a) MBC - DivBiclust

(b) MBC - SC3

Figure 4: Comparisons with the ground-truth clustering of the MBC dataset

ure 4(a). It would be of interest and one of our future work to study the automatic adaptation of the similarity threshold to different levels of variability in the expression profiles of the cells of different types in the single-cell dataset during the bicluster mining process.

3 Experiment Results on the MOU Datasets

3.1 Impact of α and $|T_s|$

According to the data noise analysis shown in Figure 1, the noise level of the MOU1 dataset exceeds 0.4. The noise level of the MOU2 dataset is 0.31 and is similar with that of the HPIC dataset. Therefore, we refer to the setting for the HPIC dataset and set the dropout threshold γ to 0.1 for the two MOU datasets as well. The similarity threshold β for pattern merging is fixed to 0.5.

The ARI scores of *DivBiclust* on the two MOU datasets with respect to the similarity threshold α and the size of initial gene set $|T_s|$ are presented in Figure 5. Due to the higher noise level of the MOU1 dataset, the ARI scores on this dataset is generally lower than the ARI scores on the MOU2 dataset under most of the settings. In addition, as listed in Table 9, the best ARI score on the MOU1 dataset is 0.670 which is achieved when $\alpha = 0.25$ and $|T_s| = 30$, and the best ARI score on the MOU2 dataset is 0.841 and it is achieved when $\alpha = 0.15$ and $|T_s| = 30$. For the more noisy MOU1 dataset, a more relaxed similarity threshold is needed.

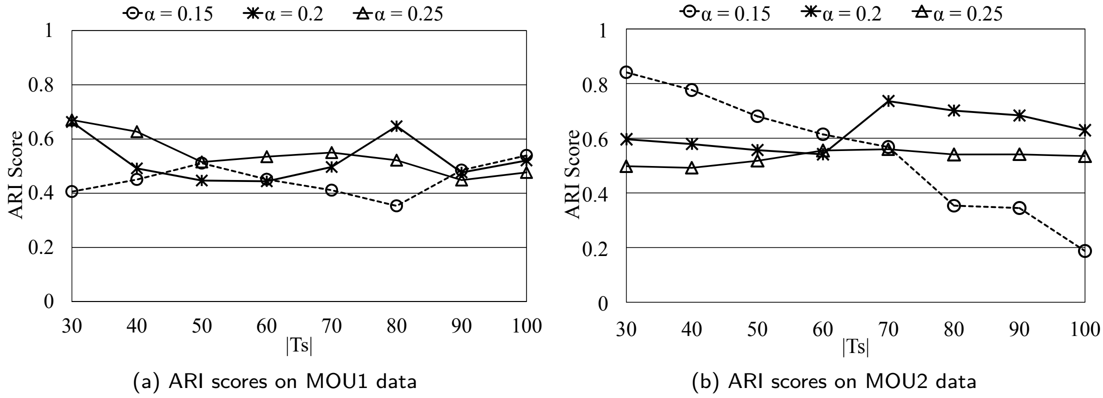


Figure 5: ARI w.r.t. similarity threshold α and the size of initial gene set $|T_s|$ ($\beta = 0.5$).

Table 9: The Highest ARI Scores on the MOU1 and MOU2 Data

Dataset	ARI Score	Parameter Settings		
		α	$\|T_s\|$	γ
MOU1	0.670	0.25	30	0.1
MOU2	0.841	0.15	30	0.1

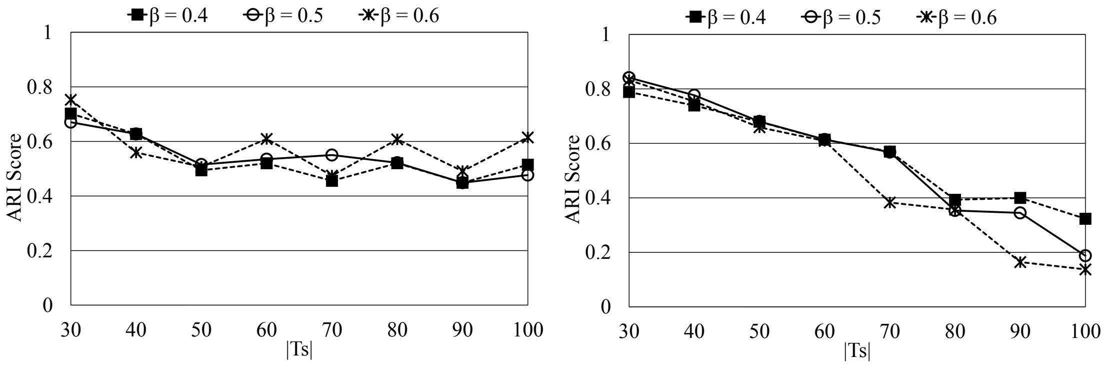
3.2 Impact of β

Next, we study how the setting of the similarity threshold β affects the clustering accuracy using the MOU datasets. The threshold α and γ are respectively fixed to the optimal settings for individual dataset as listed in Table 9. Then, the threshold β varies in $\{0.4, 0.5, 0.6\}$ and $|T_s|$ increases from 30 to 100. The experiment results are shown in Figure 6. Similar to the results on the four HUM datasets, the impact of β on the clustering accuracy is limited over the two MOU datasets as well. On both datasets, as β increases from 0.4 to 0.6, the ARI scores under all the $|T_s|$ values simply show a small change. In addition, the highest ARI scores under different β settings all occur with $|T_s|$ equal to 30. Thus, the setting of β does not affect the the optimal settings of other parameters. Therefore, β is fixed to 0.5 as we declared in the paper.

3.3 Comparison with Related Methods

The optimal ARI scores achieved by *DivBiclust* and nine state-of-the-art methods on the two MOU datasets are listed in Table 10. The results again effectively demonstrate the superiority of our *DivBiclust* method in identifying the cell subpopulations. The ARI scores achieved by *DivBiclust* on the MOU1 and MOU2 datasets are 0.670 and 0.841, which are respectively 0.198 and 0.363 higher than the best results achieved by the counterpart methods.

The heatmaps that demonstrate the correlations between the cell clusters generated by *DivBiclust* and the ground-truth cell classes of the datasets are shown in Figure 7. The clustering result of a related method that achieves the second highest ARI score is also presented for comparison. Because both Seurat and SOUP involve a parameter that can be used to control the final number of cell clusters, these two methods perform better at identifying the final number of cell subpopulations. However, the clustering result generated by *DivBiclust* still achieves much higher ARI score. The reason is that, *DivBiclust* successfully identifies several largest cell classes with high accuracy from both MOU1 and MOU2 data. For example, the three largest cell classes in the MOU1 data are respectively "*beta cells*", "*ductal cells*", and "*Endothelial*". *DivBiclust* manages to cluster together 72.0% beta cells, 94.9% ductal cells, and 93.1% endothelial cells, individually. In contrast, the clustering result generated by Seurat v3 only successfully clusters 44.0% beta cells, 53.8% ductal cells, and 97.2% endothelial cells. Similar advantage is observed from the results on



(a) ARI scores on MOU1 data

(b) ARI scores on MOU2 data

Figure 6: ARI w.r.t. the threshold β for pattern merging and the size of initial gene set $|T_s|$

Table 10: Comparisons of Clustering Accuracy

Methods	Datasets	
	MOU1	MOU2
DivBiclust	**0.670**	**0.841**
SNN-Cliq	0.063	0.065
BackSPIN	0.006	0.003
CIDR	0.322	0.272
SC3	0.439	0.262
pcaReduce	0.317	0.221
TSCAN	0.404	0.412
Seurat v3	0.472	0.298
GiniClust2	0.383	0.282
SOUP	0.332	0.478

* The largest ARI score over each dataset is highlighted using bolded font, and the second largest one is underlined.

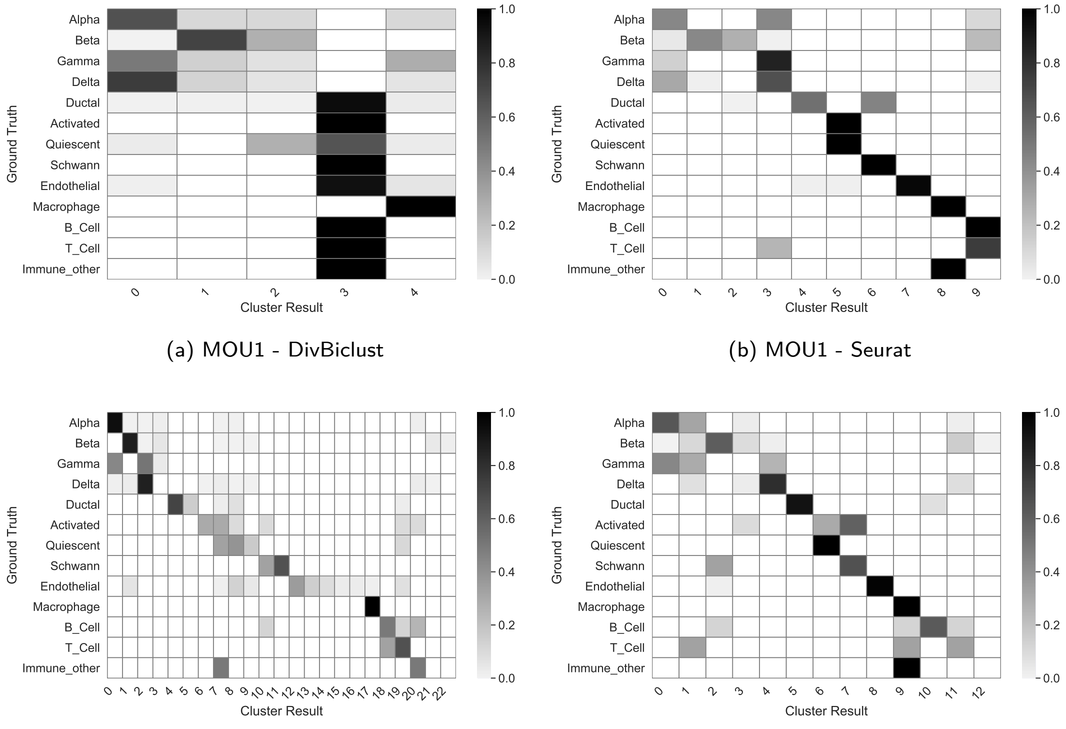
the MOU2 dataset. *DivBiclust* successfully clusters 481 beta cells, 171 alpha cells, and 115 delta cells, which respectively account for 87.3% of the beta cells, 94.0% of the alpha cells, and 86.5% of the delta cells in the dataset.

4 Experiments on the HUM_NZ and MOU_NZ Datasets

As discussed in Section 1.5, two preprocessing methods are adopted to handle the data of human and mouse pancreatic cells. The first preprocessing method simply removes those genes which have zero expression values under all cell samples, and the resultant nonzero datasets are denoted as the HUM1_NZ, HUM2_NZ, HUM3_NZ, HUM4_NZ, MOU1_NZ, and MOU2_NZ datasets. Compared to the datasets generated by the second preprocessing method, these nonzero datasets have much larger gene sets and much higher dropout rates. In this set of experiments, we study whether or not the adoption of different preprocessing steps for the datasets would affect the performance of *DivBiclust* in identifying cell subpopulations.

4.1 Clustering Accuracy

Although the dropout rates of the datasets generated by two different preprocessing steps are quite different, according to the data noise analysis performed in Section 1.6, we find that the



(a) MOU1 - DivBiclust

(b) MOU1 - Seurat

(c) MOU2 - DivBiclust

(d) MOU2 - SOUP

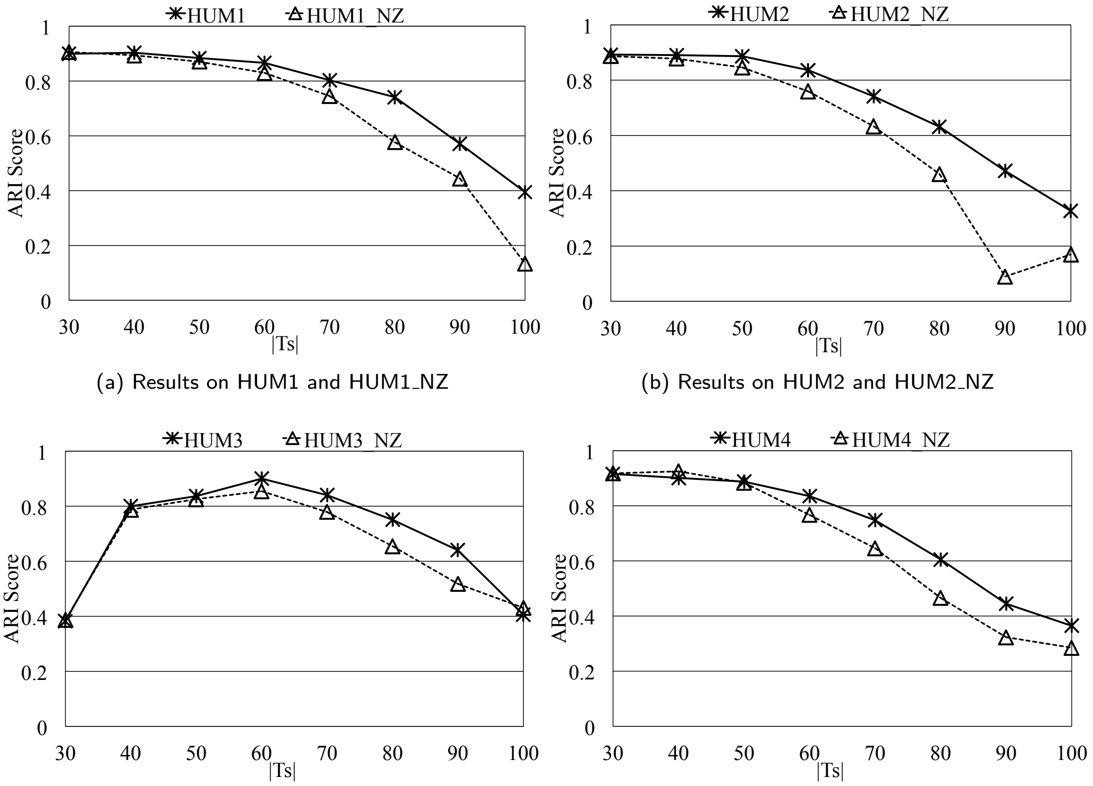
Figure 7: Comparisons with the ground-truth clustering of the MOU1 and MOU2 datasets

noise level of each nonzero dataset is only slightly higher than the noise level of the corresponding dataset generated by the second preprocessing strategy. Thus, we take the parameter settings for the HUM and MOU datasets when running *DivBiclust* over the corresponding HUM_NZ and MOU_NZ datasets.

Specifically, we fix the similarity threshold α to 0.2 and fix the dropout threshold γ to 0 for the four HUM_NZ datasets. Then, we respectively set α to 0.25 and 0.15 and fix γ to 0.1 for MOU1_NZ and MOU2_NZ datasets. The ARI scores with respect to the size of initial gene set $|T_s|$ are shown in Figures 8 and 9. The ARI scores on the nonzero datasets are generally smaller than the ARI scores on the preprocessed datasets with gene selection. The reason is that, the gene selection strategy filters out those low-abundance and low-variance genes and thus reduces the noise level in the datasets. However, the varying trend of the ARI scores on two types of datasets are similar, and the difference between the scores under the same $|T_S|$ value is small . For example, on the HUM1_NZ, HUM2_NZ, and HUM4_NZ datasets, the ARI scores with $|T_s|$ equal to 30 or 40 remain high and are at least 0.88. The highest ARI score over the HUM3_NZ dataset is achieved when $|T_s|$ is equal to 60, and the highest ARI scores on both MOU1_NZ and MOU2_NZ datasets are achieved when $|T_s|$ equals to 30. These results are accordant with what we observed from the experiments over the HUM1, HUM2, HUM3, and HUM4 datasets. Therefore, we are still able to effectively identify the cell subpopulations using the nonzero datasets.

4.2 Running time

Figures 10 and 11 show the running time of *DivBiclust* on two different types of datasets. Because the HUM and MOU datasets have much smaller gene sets compared to the corresponding HUM_NZ and MOU_NZ datasets, the total running time is generally shorter. The only exception is that the running time on the HUM3 and HUM3_NZ datasets are roughly the same. The reason is that,



(a) Results on HUM1 and HUM1_NZ

(b) Results on HUM2 and HUM2_NZ

(c) Results on HUM3 and HUM3_NZ

(d) Results on HUM4 and HUM4_NZ

Figure 8: Comparison of ARI scores on HUM and HUM_NZ Datasets

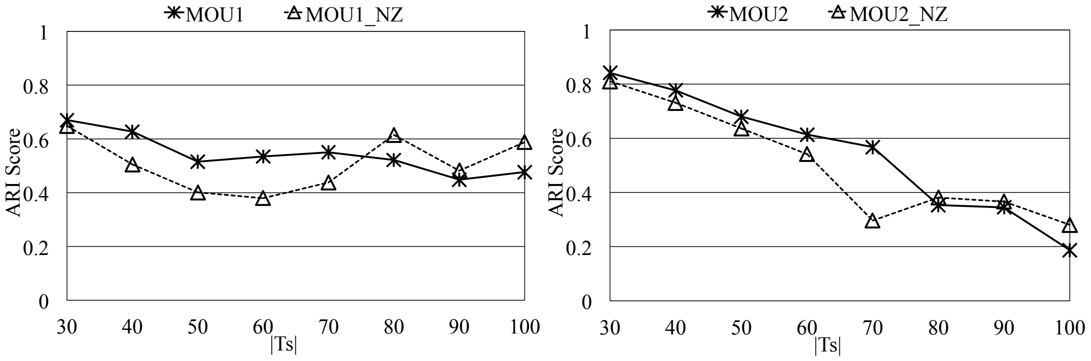
when dealing with these two datasets with large cell sets, the main time consumption is for pattern merging. Because the number of cells in these two datasets are the same, the number of mined sc-biclusters for merging are very close and thus the time cost is similar with each other. On the other hand, even when $|T_s|$ is set to 30, the running time of *DivBiclust* on the HUM3_NZ dataset with more than 3000 cells is around 130 seconds. The running time on all the other datasets is no longer than 30 seconds. Thus, *DivBiclust* is efficient in dealing with both types of datasets.

4.3 Comparison with Related Methods

We list the ARI scores achieved by *DivBiclust* and nine state-of-the-art methods on the six nonzero datasets in Table 11. For each dataset, *DivBiclust* takes exactly the same parameter settings for the corresponding HUM or MOU datasets. The ARI scores achieved by *DivBiclust* are much higher than the scores achieved by the counterpart methods over all the datasets. The experiment results again well demonstrate the effectiveness of *DivBiclust* in identifying cell subpopulations from the scRNA-seq datasets.

5 Parameter Settings of Related Methods

In the paper, we compare *DivBiclust* with nine state-of-the-art method in terms of the accuracy for identifying cell subpopulations. These nine related methods are SNN-Cliq [13], BackSPIN [16], CIDR [8], SC3 [5], pcaReduce [18], TSCAN [4], Seurat v3 [2,11], GiniClust2 [12], and SOUP [17]. The source codes of all these methods are publicly downloaded, and their parameter settings are as follows.



(a) Results on MOU1 and MOU1_NZ

(b) Results on MOU2 and MOU2_NZ

Figure 9: Comparison of ARI scores on MOU and MOU_NZ Datasets

Table 11: Comparisons of Clustering Accuracy*

Methods	Datasets					
	HUM1_NZ	HUM2_NZ	HUM3_NZ	HUM4_NZ	MOU1_NZ	MOU2_NZ
DivBiclust	**0.894**	**0.887**	**0.854**	**0.925**	**0.649**	**0.810**
SNN-Cliq	0.052	0.085	$-^{\dagger}$	0.334	−0.005	0.056
BackSPIN	0.016	0.001	0.038	0.004	0.007	0.007
CIDR	0.188	0.238	0.295	0.449	0.112	0.025
SC3	0.529	0.573	0.559	0.510	0.394	0.276
pcaReduce	0.307	0.351	0.392	0.339	0.327	0.214
TSCAN	0.407	0.570	0.614	0.532	0.456	0.325
Seurat v3	0.383	0.403	0.513	0.493	0.481	0.315
GiniClust2	0.396	0.470	0.423	0.299	0.551	0.305
SOUP	0.621	0.764	0.486	0.556	0.527	0.467

* The largest ARI score over each dataset is highlighted using bolded font, and the second largest one is underlined.

† SNN-Cliq cannot handle the HUM3_NZ dataset and generate the cell clustering.

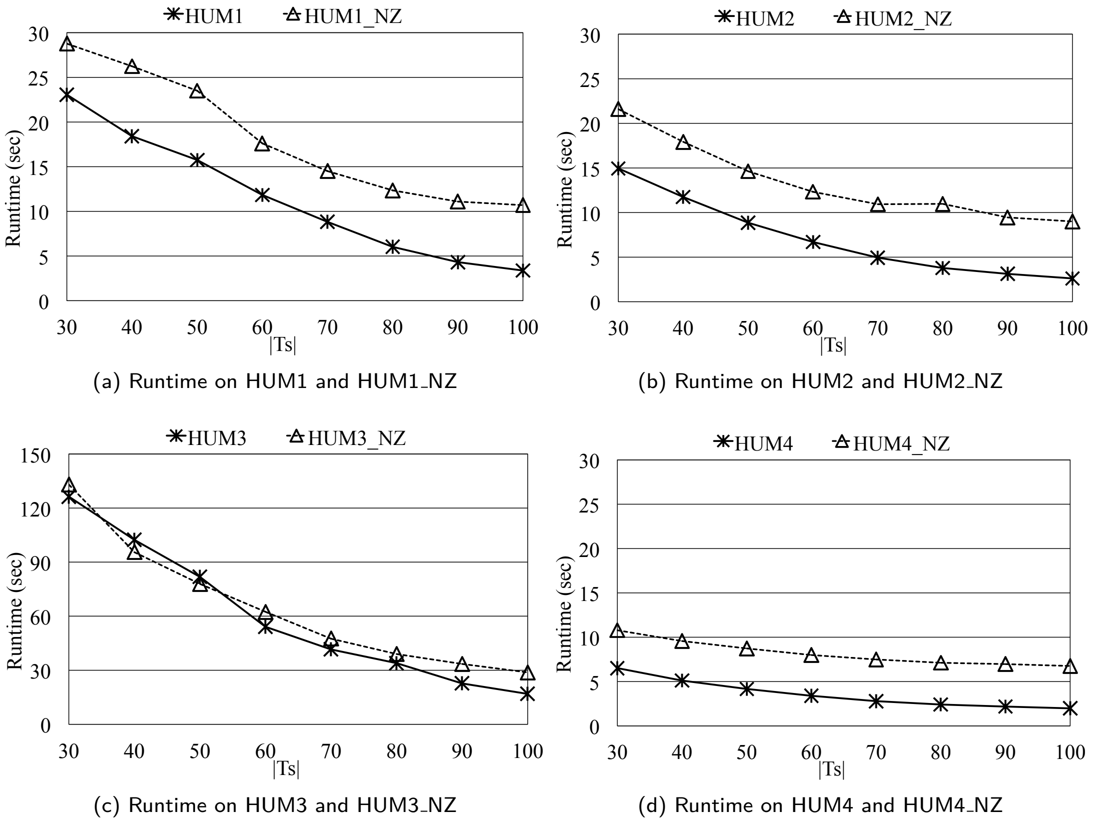
- The SNN-Cliq method utilized an SNN-based graph and the SNN threshold is set to 3 as suggested in [13]. The similarity threshold for merging the quasi-cliques is fixed to 0.5, and this threshold plays the same role as the threshold β for pattern merging in our *DivBiclust* method.

- In the BackSPIN method, two parameters need to be specified. The number of selected genes determines the size of the matrix taken for partition-based biclustering. We either take 5000 genes as suggested in [16] or use the complete gene set. The depth of clustering controls the number of matrix splits, which is also tuned with the best ARI score being reported for every dataset.

- The CIDR method combines PCA with hierarchical clustering. We set the cluster number to be the same as the number of ground-truth cell clusters. The principal coordinates that are used in clustering are automatically estimated by a variation of the scree [3] method.

- The SC3 method involves a parameter K for K-means clustering. We studied the performance by either estimating K automatically or setting it to the number of cell clusters in the ground truth, and take the better result for each dataset.

- The pcaReduce method cannot determine the number of final cell clusters, and thus we set it to be the same as the number of clusters in the ground truth. Moreover, pcaReduce is a



(a) Runtime on HUM1 and HUM1_NZ

(b) Runtime on HUM2 and HUM2_NZ

(c) Runtime on HUM3 and HUM3_NZ

(d) Runtime on HUM4 and HUM4_NZ

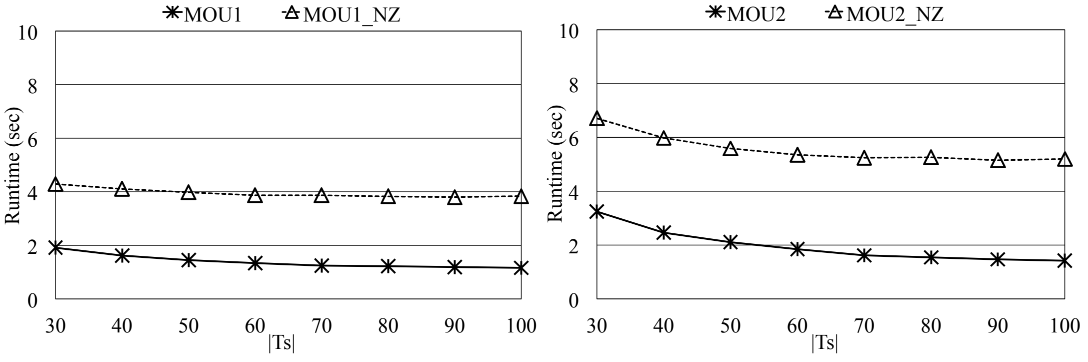
Figure 10: Comparison of Running Time on HUM and HUM_NZ Datasets

randomized algorithm, and so we run it for 10 times and report the highest ARI score for each dataset.

- The TSCAN method adopts a two-way clustering, where the number of gene clusters is fixed to 5% of the total number of genes. The optimal cell cluster number is selected by Bayesian Information Criterion (BIC) from a range of possible cluster numbers defined by user. We set the range to be $[1, 15]$, which covers the number of ground-truth cell clusters in all the datasets.
- The Seurat v3 method involves a parameter to tune the granularity of the clustering result. The larger the value is, the more number of clusters is generated by the graph-based clustering algorithm Louvain. As suggested in [15], we set the parameter to 0.9.
- GiniClust2 is the integration of two clustering methods, GiniClust and K-means, where K is set to the number of cell clusters in the ground-truth.
- The SOUP method first performs K-means over a set of "pure cells", where K is set to the number of cell clusters in the ground-truth. The proportion of pure cells is set to 50% of the total number of cells in the dataset as suggested by the author.

References

[1] M. Baron, A. Veres, S. L. Wolock, and et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346 – 360.e4, 2016.



(a) Runtime on MOU1 and MOU1_NZ

(b) Runtime on MOU2 and MOU2_NZ

Figure 11: Comparison of Running Time on MOU and MOU_NZ Datasets

[2] A. Butler, P. Hoffman, P. Smibert, and et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

[3] R. B. Cattell. The scree test for the number of factors. *Multivar Behav Res.*, 1:245–276, 1966.

[4] Z. Ji and H. Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Research*, 44:gkw430, 05 2016.

[5] V. Yu. Kiselev, K. Kirschner, M. T. Schaub, and et al. Sc3 - consensus clustering of single-cell rna-seq data. *Nature Methods*, 14(5):483–486, 2017.

[6] A.M. Klein, L. Mazutis, I. Akartuna, and et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[7] J. Li, J. Klughammer, M. Farlik, and et al. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Reports*, 17(2):178–187, 2016.

[8] P. Lin, M. Troup, and J. WK Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):59, 2017.

[9] D. Ramskolds, S. Luo, and Y. C. Wang. Full-length mRNA-Seq from single-cell levels of rna and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, 2012.

[10] K. Sato, K. Tsuyuzaki, K. Shimizu, and I. Nikaido. Cellfishing.jl: an ultrafast and scalable cell search method for single-cell rna sequencing. *Genome Biology*, 20(31), 2019.

[11] T. Stuart, A. Butler, P. Hoffman, and et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888 – 1902.e21, 2019.

[12] D. Tsoucas and G. Yuan. Giniclust2: A cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biology*, 19(58), 12 2018.

[13] C. Xu and Z. Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015.

[14] L. Yan, M. Yang, and H. Guo. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural and Molecular Biology*, 20:1131–1139, 2013.

[15] Yuchen Yang, Ruth Huh, Houston W Culpepper, Yuan Lin, Michael I Love, and Yun Li. SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*, 35(8):1269–1277, 09 2018.

[16] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, and et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[17] L. Zhu, J. Lei, L. Klei, B. Devlin, and K. Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.

[18] J. Žurauskienė and C. Yau. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, 17(1):140, Mar 2016.