Supplemental materials

Table of Contents

1. Supplemental Notes

Parameter setup and versions of the clustering methods

Seven popular single-cell clustering algorithms were applied for performance comparison, including SC3 (version 1.24.0), Seurat (Version 4.1.1), Soup (version 0.0.0.9000), CIDR (version 0.1.5), pcaReduce (version 1.0), SIMLR (version 1.22.0), and tSNE+kmeans (Rtsne version 0.16). All the parameters of the methods are utilized with their default values. Moreover, the number of clusters for all methods except Seurat is set to the number of true cell types derived from the raw datasets. For the tSNE+k-means algorithm, we apply the tSNE method 50 times, and then k-means clustering is performed on the tSNE results with the smallest Kullback–Leibler divergence value.

2. Supplemental Methods

Consensus clustering

Combining multiple clustering results can improve the reliability and robustness of the final clustering results (Strehl and Ghosh, 2002). Consensus clustering (Kiselev *et al.*, 2017)aims to integrate multiple clustering results obtained by different algorithms through a probabilistic strategy. For one clustering result, consensus clustering converts it into a binary matrix and the dimension of this binary matrix is $N \times N$ (N is the total number of cells).

Given clustering result $R=\{r_1, r_2,\ldots, r_N\}$, assuming that the i-th cell and j-th cell belong to the same cluster, then the value in the i-th row and j-th column of the corresponding binary matrix is 1; if not, the corresponding value in the binary matrix is 0.

$$w_{ij} = \begin{cases} 1, r_i = r_j \\ 0, r_i \neq r_j \end{cases}$$

where r_i and r_j respectively represent the clustering labels of the i-th cell and j-th cell in the clustering result $R=\{r_1, r_2,\ldots, r_N\}$.

Assuming that we have m clustering results, the final consensus matrix Y is obtained by averaging these binary matrices corresponding to these m clustering results.

$$Y = \frac{\sum_{i=1}^{m} W_i}{m}$$

where w_i is the binary matrix corresponding to the *i*-th clustering result.

The values in the consensus matrix *Y* are in the range of [0,1], representing the probability that the *i*-th cell and the *j*-th cell are clustered into the same cluster. The larger the value of *Y*, the greater the probability that the two cells belong to the same cluster. Therefore, to some extent, this value can be regarded as the similarity coefficient between the *i*-th and *j*-th cells.

3. Supplemental Tables

Table S1. 10 Datasets used for validations.

Dataset	Tissue	Cells	Genes	Cell types
Biase(Blase *et al.*, 2014)	Mouse Embryos	56	25734	4
Darmanis (Darmanis *et al.*, 2015)	Human Brain	466	20214	9
Deng (Deng *et al.*, 2014)	Mouse Embryos	268	22431	6
Muraro (Muraro *et al.*, 2016)	Human Pancreas	2122	19140	10
Usoskin (Usoskin *et al.*, 2015)	Mouse Embryo	622	25334	4
Romanov (Romanov *et al.*, 2017)	Mouse Brain	2881	24341	7
Zeisel (Zeisel *et al.*, 2015)	Mouse Brain	3005	19972	9
Lake (Lake *et al.*, 2016)	Human Brain	3042	25123	16
Buettner (Buettner *et al.*, 2015)	Mouse Embryos	182	8989	3
Baron-mouse (Baron *et al.*, 2016)	Mouse Pancreas	1886	14878	13

4. Supplemental Figures

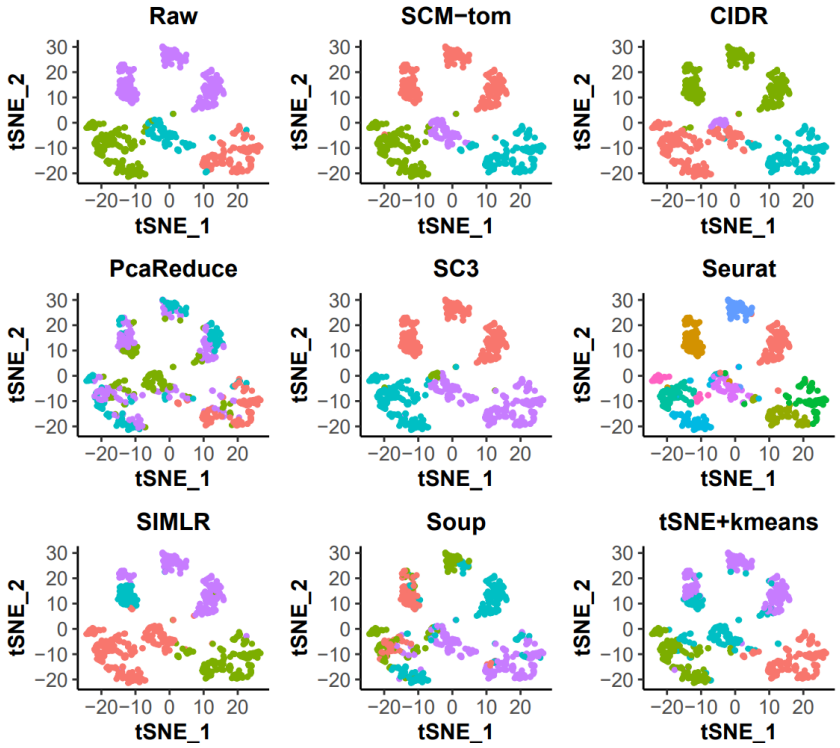

Figure S1. The Visualization of the clustering results on Usoskin dataset. Each panel

represents a algorithm. The first panel represents the true labels annotated by the dataset author.

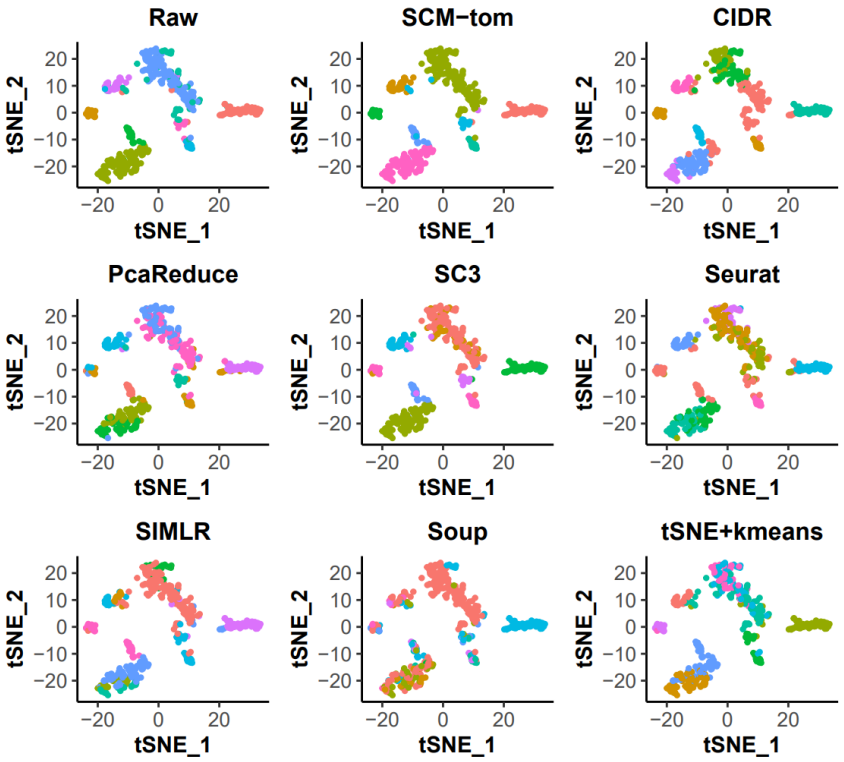

Figure S2. The Visualization of the clustering results on Darmanis dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

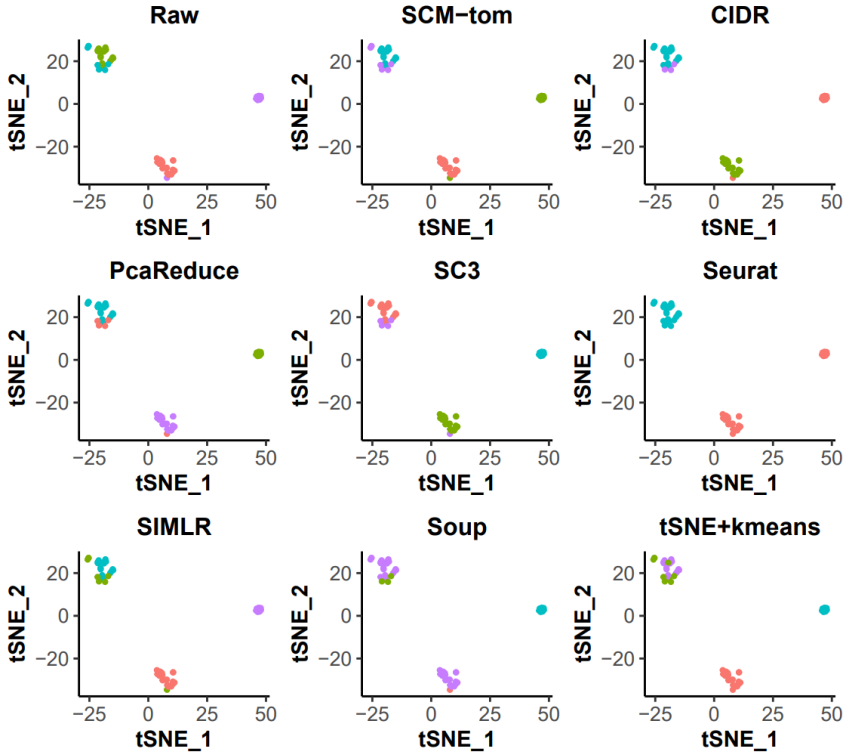

Figure S3. The Visualization of the clustering results on Biase dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

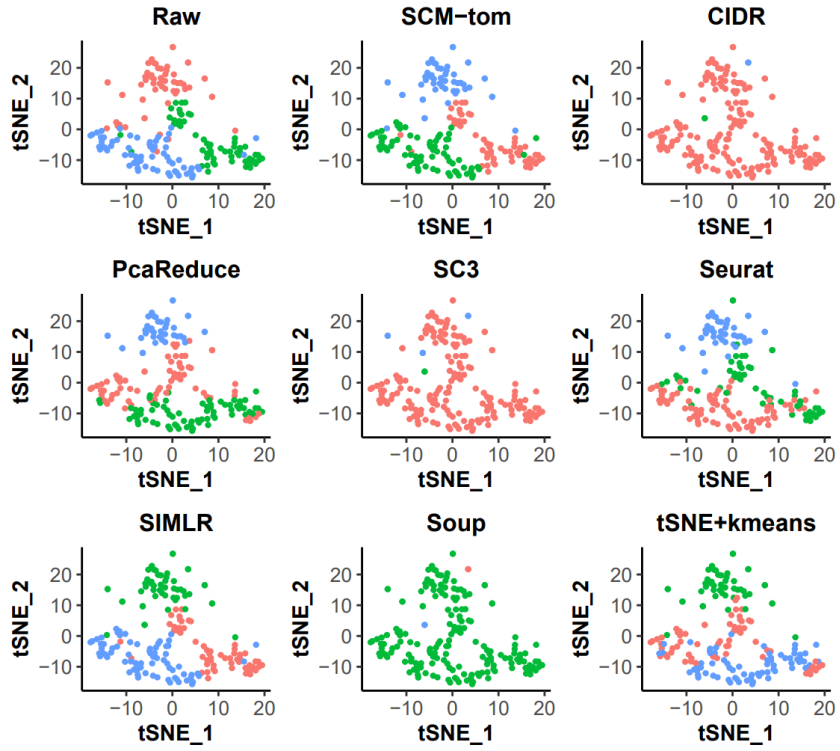

Figure S4. The Visualization of the clustering results on Buettner dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

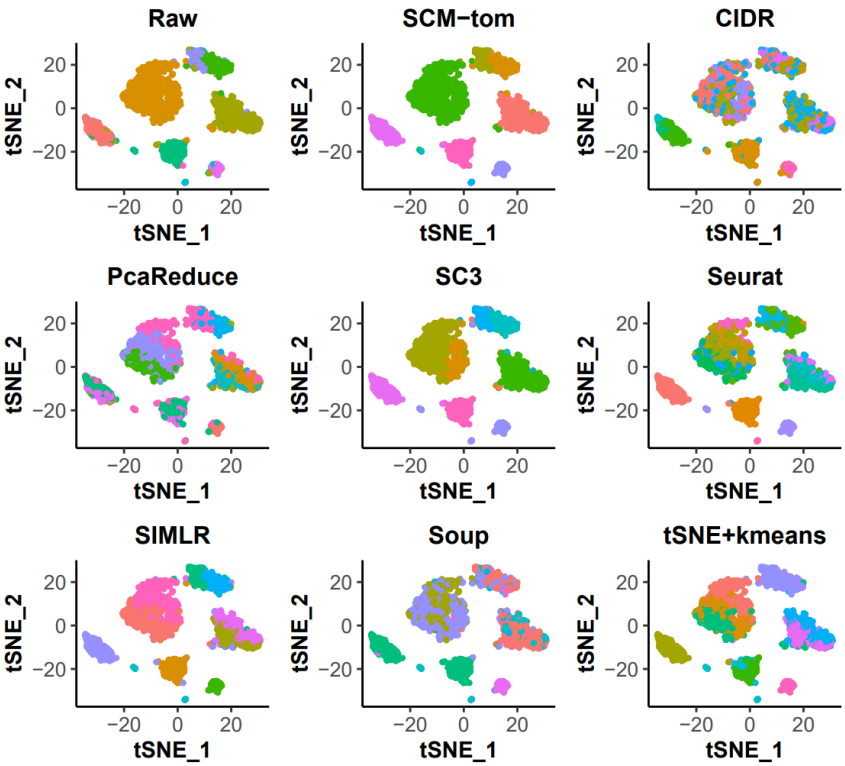

Figure S5. The Visualization of the clustering results on Muraro dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

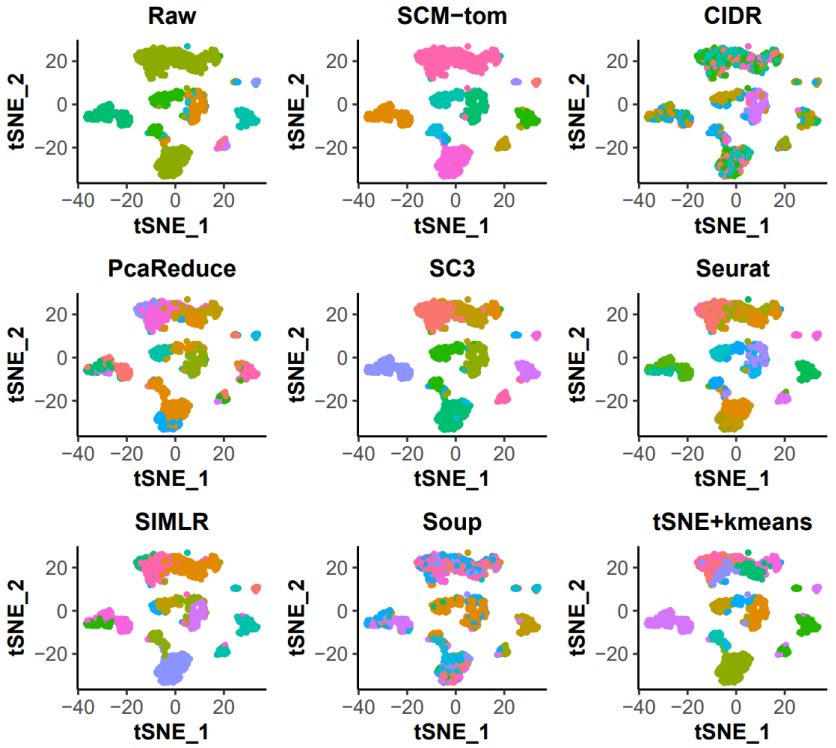

Figure S6. The Visualization of the clustering results on Baron-mouse dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

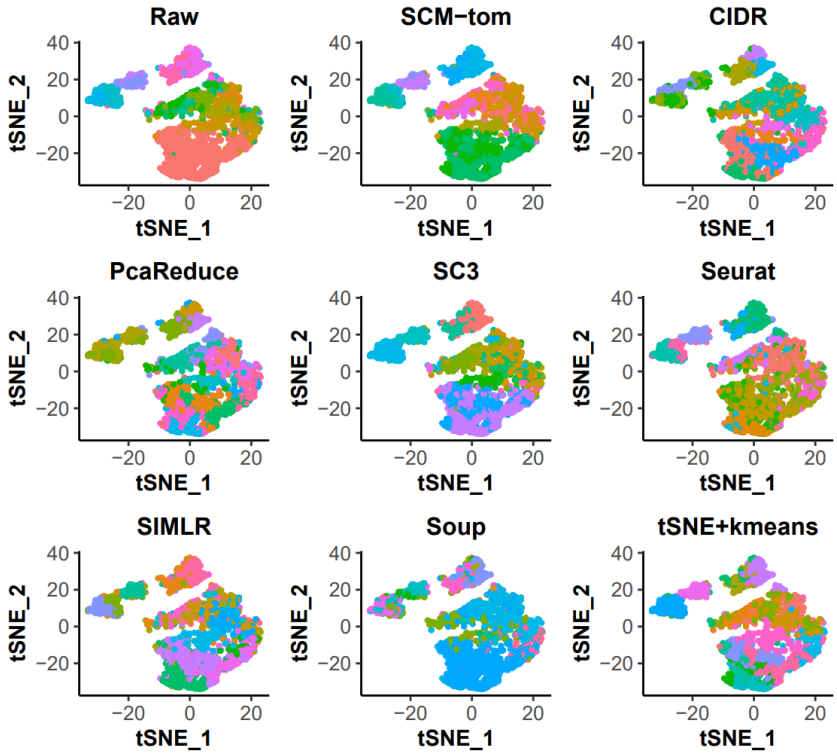

Figure S7. The Visualization of the clustering results on Lake dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

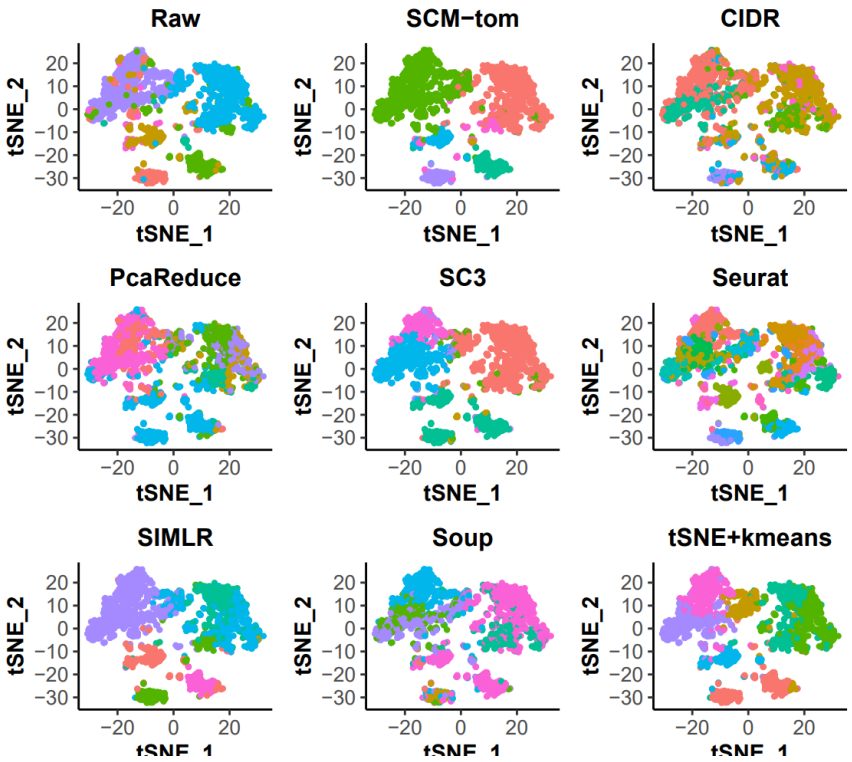

Figure S8. The Visualization of the clustering results on Romanov dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

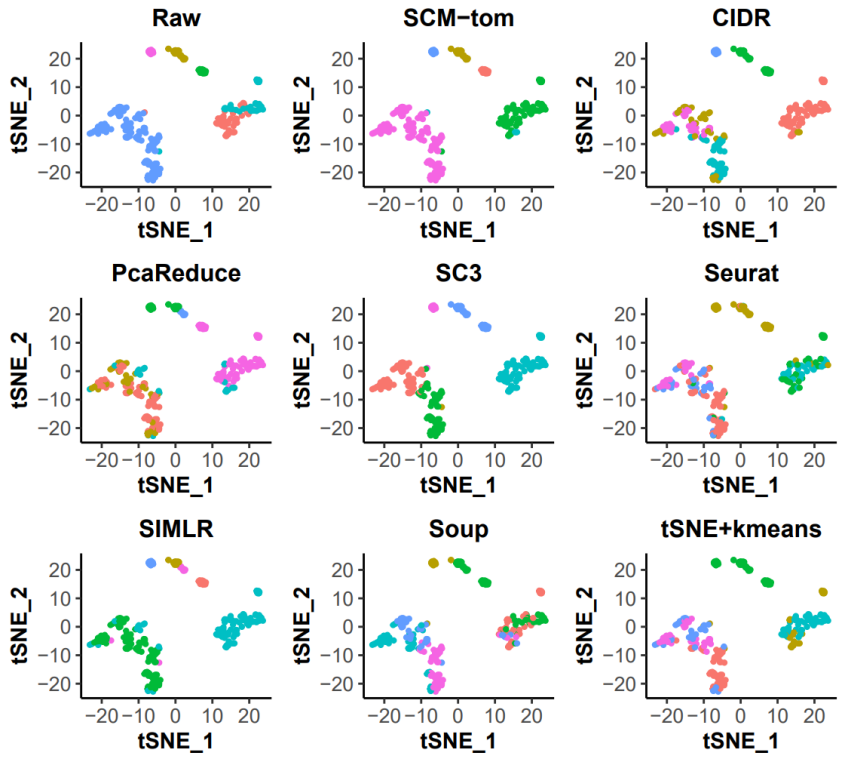

Figure S9. The Visualization of the clustering results on Deng dataset. Each panel represents a algorithm. The first panel represents the true labels annotated by the dataset author.

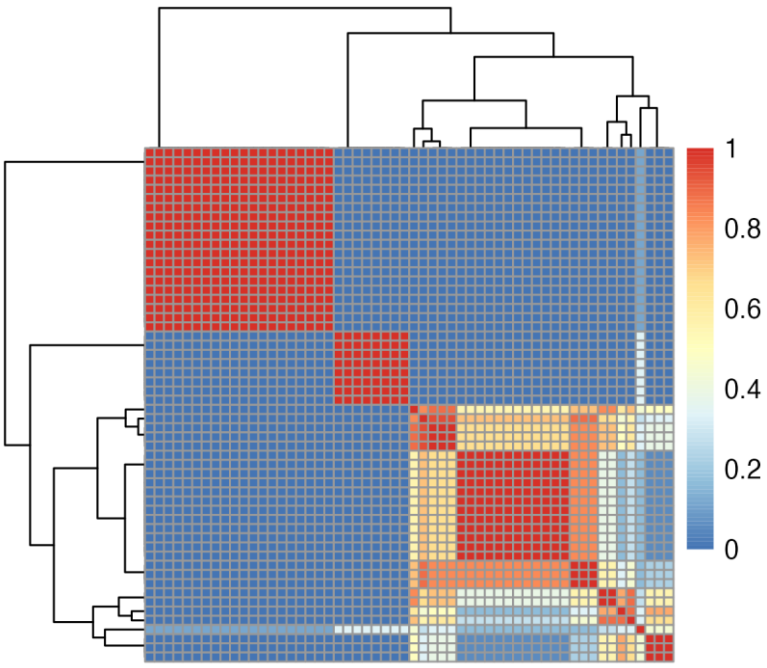

Figure S10. Heatmap of the optimal consensus matrix on Biase dataset.

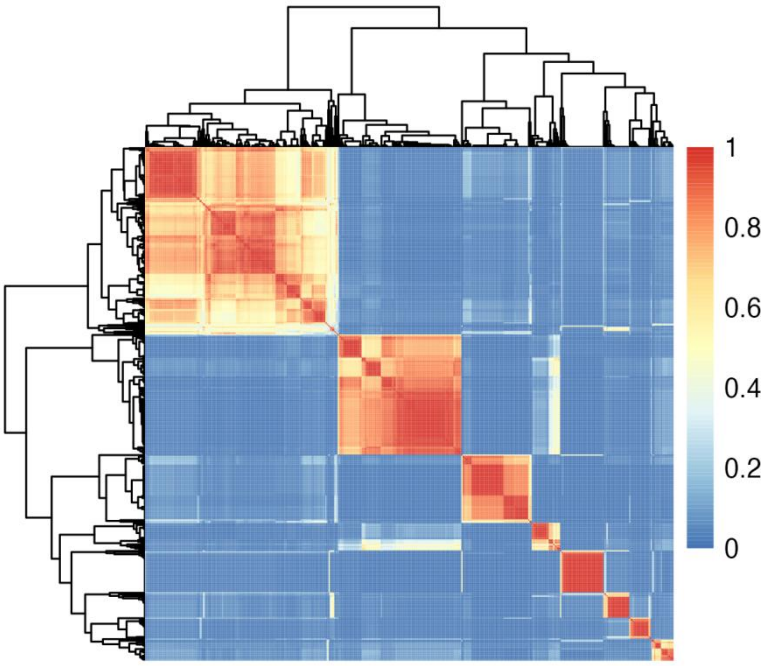

Figure S11. Heatmap of the optimal consensus matrix on Darmanis dataset.

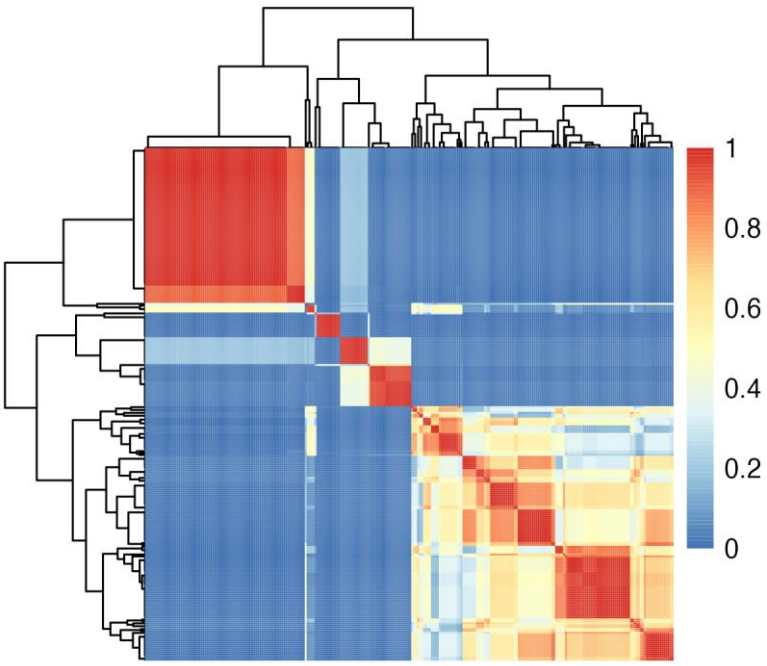

Figure S12. Heatmap of the optimal consensus matrix on Deng dataset.

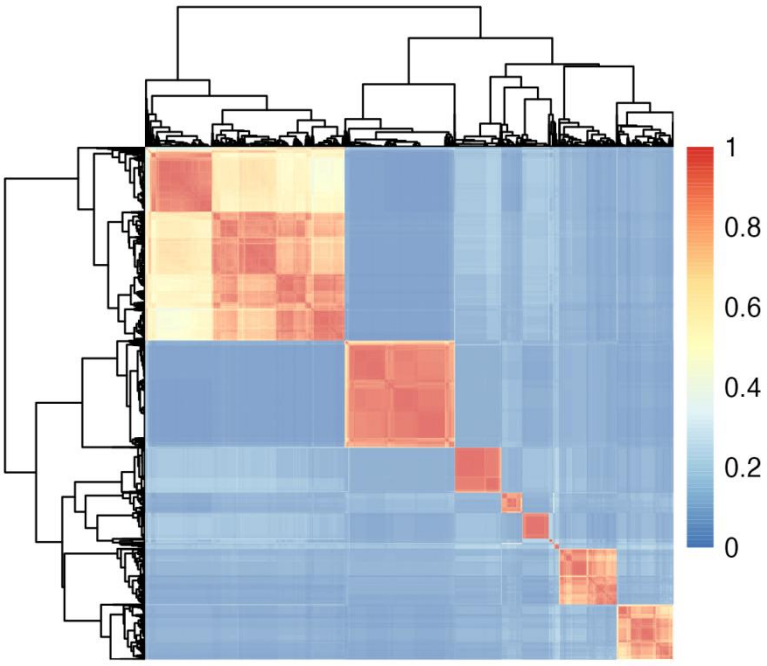

Figure S13. Heatmap of the optimal consensus matrix on Muraro dataset.

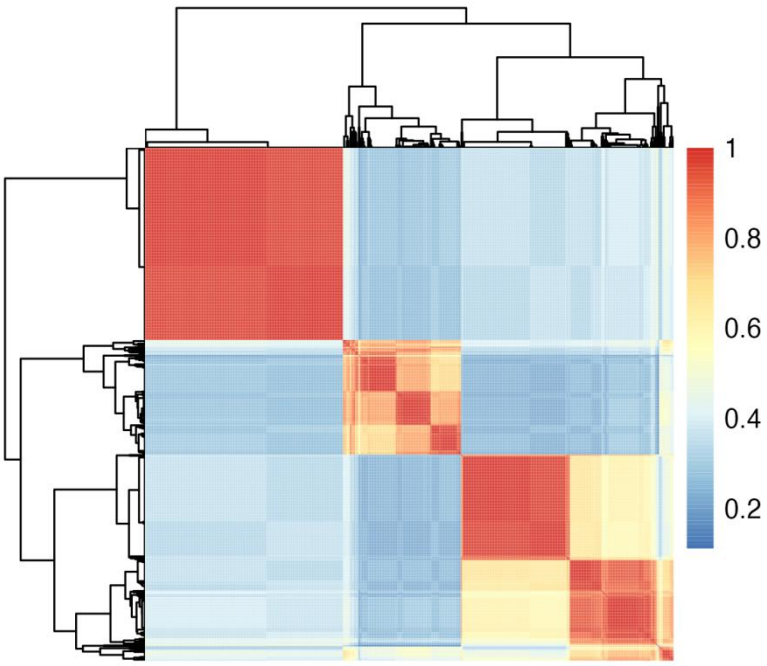

Figure S14. Heatmap of the optimal consensus matrix on Usoskin dataset.

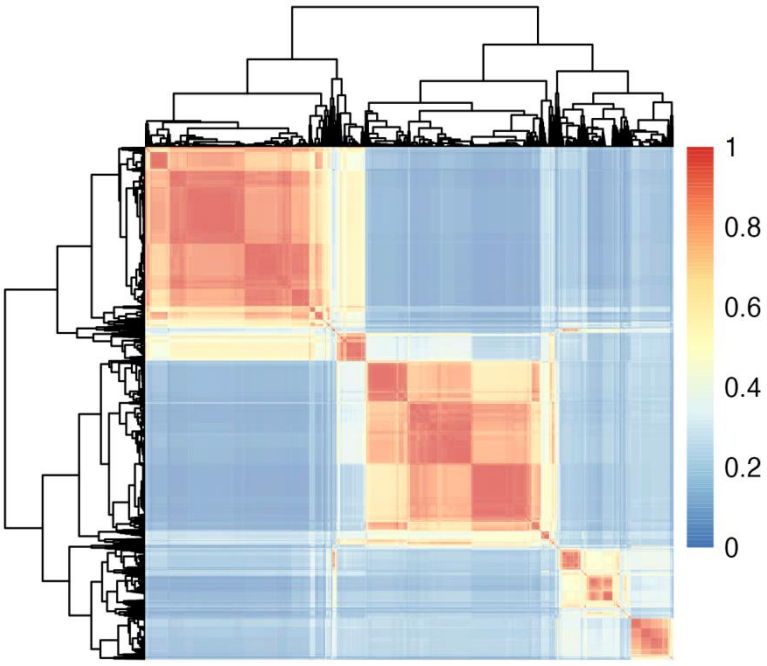

Figure S15. Heatmap of the optimal consensus matrix on Romanov dataset.

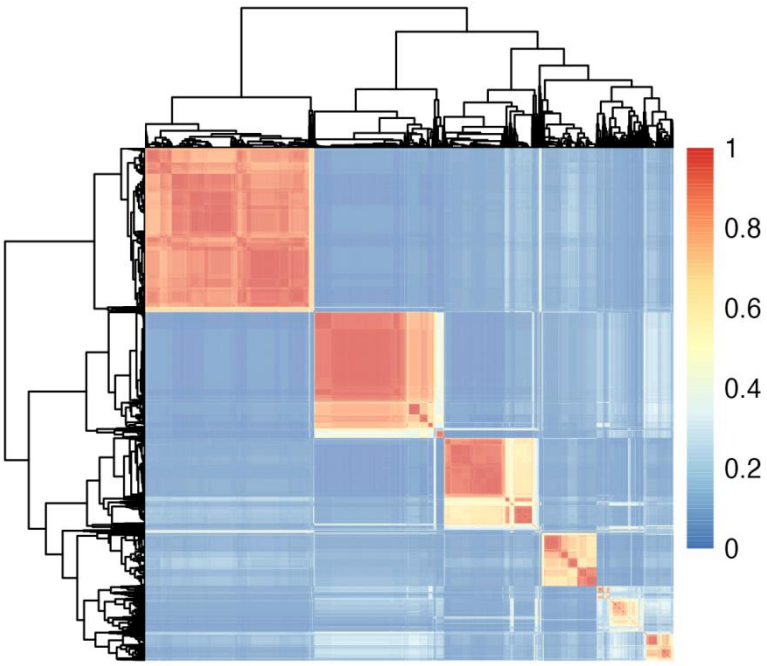

Figure S16. Heatmap of the optimal consensus matrix on Zeisel dataset.

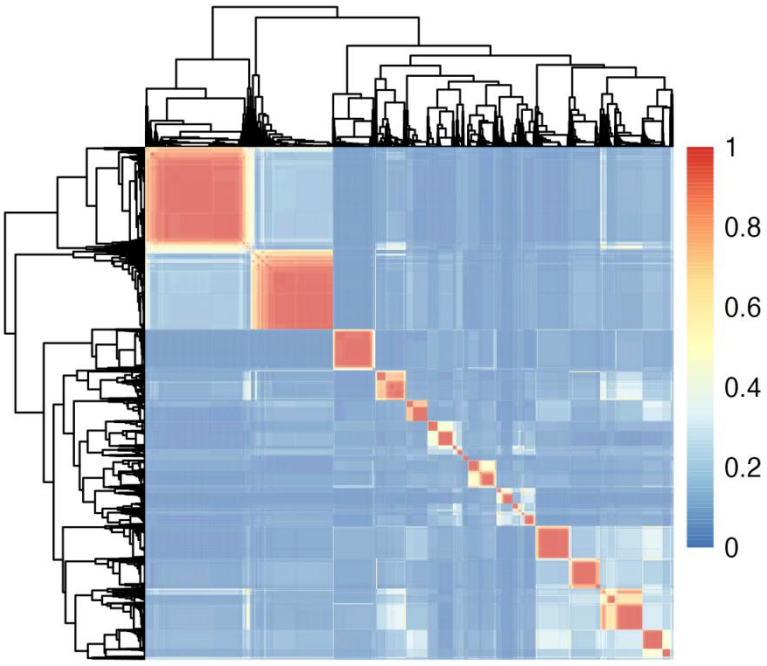

Figure S17. Heatmap of the optimal consensus matrix on Lake dataset.

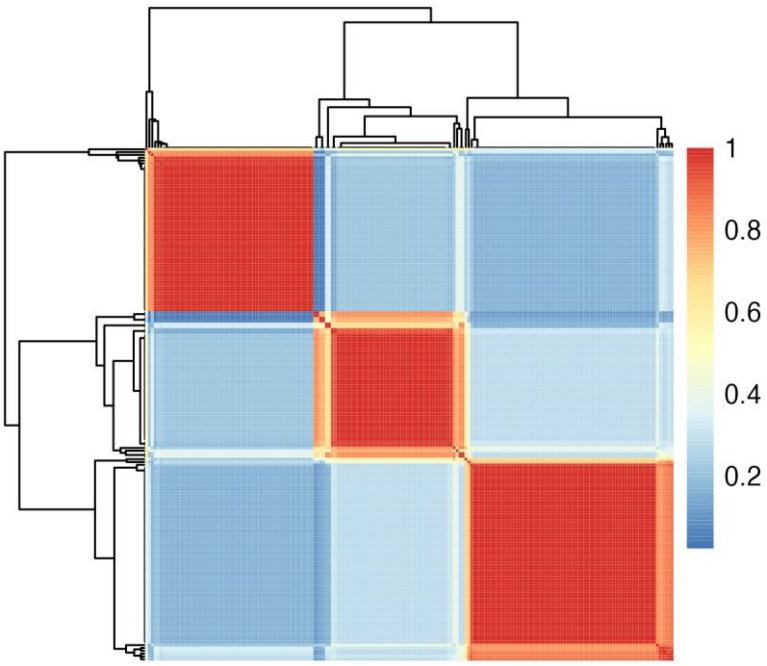

Figure S18. Heatmap of the optimal consensus matrix on Buettner dataset.

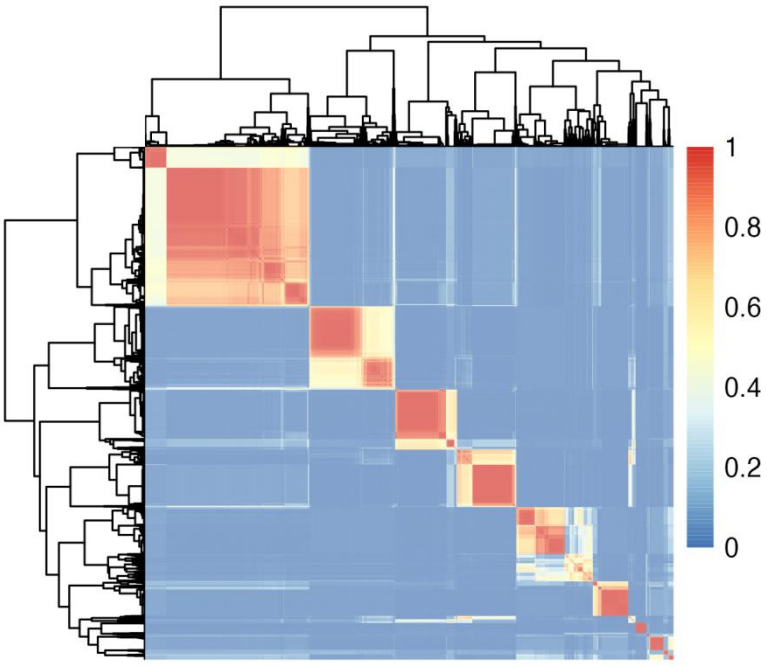

Figure S19. Heatmap of the optimal consensus matrix on Baron-mouse dataset.

5. References

Baron,M. *et al.* (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst*, **3**, 346-360.e4.

Blase,F.H. *et al.* (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.*, **24**, 1787–1796.

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell

RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

Darmanis,S. *et al.* (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 7285–7290.

Deng,Q. *et al.* (2014) Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*, **343**, 193–196.

Kiselev,V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483-+.

Lake,B.B. *et al.* (2016) Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, **352**, 1586–1590.

Muraro,M.J. *et al.* (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.*, **3**, 385-+.

Romanov,R.A. *et al.* (2017) Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.*, **20**, 176–188.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles - A knowledge reuse framework for combining partitionings. In, *Eighteenth National Conference on Artificial Intelligence (aaai-02)/Fourteenth Innovative Applications of Artificial Intelligence Conference (iaai-02), Proceedings*. M I T Press, Cambridge, pp. 93–98.

Usoskin,D. *et al.* (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.*, **18**, 145-+.

Zeisel,A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.