

# scCNMF: an integrated analysis model for paired single-cell RNA sequencing and assay for transposase-accessible chromatin sequencing data leveraging cell similarity and cis-regulatory potential

Yufei Zhang<sup>1,2</sup>, Qiongyu Sheng<sup>1,2</sup>, Huiran Zhan<sup>\*1</sup>, Yiyuan Guo<sup>3</sup>, Xiaoran Shi<sup>1</sup>, and Jing Qi<sup>1,2</sup>

<sup>1</sup>School of Mathematics, Harbin Institute of Technology, Harbin 150000, China

<sup>2</sup>Zhengzhou Research Institute, Harbin Institute of Technology, Zhengzhou 450000, China

<sup>3</sup>Department of Ophthalmology, The First Affiliated Hospital of Harbin Medical University, Harbin 150000, China

Corresponding author:  
Jing Qi<sup>1,2</sup>

Email address: qijing@hit.edu.cn

## ABSTRACT

This file includes the following subsections:  
(1) Details of datasets and preprocessing.  
(2) Preprocessing and initialization.  
(3) Parameter selection.  
(4) Comparative methods.  
(5) Evaluation indicators.  
(6) Supplementary figures.

## DETAILS OF DATASETS AND PREPROCESSING

**PBMC dataset.** Sourced from the 10x Genomics website and sequenced in 10x format, with preprocessing data and cell annotations provided by Mowgli (Huizing et al., 2023). This dataset comprises chromatin accessibility and gene expression data for 9,320 paired PBMC cells. The dataset includes the top 2,500 highly variable genes in the RNA modality and the top 5,000 highly variable peaks in the ATAC modality.

**Mouse kidney dataset.** This dataset is obtained from the GEO repository and sequenced using the sci-CAR method. Cell labels are sourced from the dataset supplement, and gene annotations are provided by Ensembl (Harrison et al., 2024). During preprocessing, duplicate genes are merged by taking the maximum value of the corresponding peaks, and cells labeled as NA are removed, resulting in 8,837 cells for downstream analyses.

**Human brain dataset.** Sourced from the 10x Genomics website and sequenced in 10x format, this dataset comprises data from 3,332 cells. Marker genes and corresponding cell annotations are obtained from CellMarker 2.0 (Hu et al., 2023).

**BMMC dataset.** This dataset is obtained from the GEO repository and sequenced using SHARE-seq. Cell labels are sourced from the dataset supplement, and Ensembl provided gene annotations. During the raw data preprocessing, duplicate genes and their corresponding peaks are merged by selecting their maximum values. Five random cell samples are performed to extract 8,000 cells and retain the top 5,000

---

\*Yufei Zhang, Qiongyu Sheng, and Huiran Zhan contributed equally to this work.

highly variable genes and the top 20,000 highly variable peaks for downstream analyses for each sampled dataset.

## PREPROCESSING AND INITIALIZATION

scCNMF takes as input scRNA-seq and scATAC-seq data read matrices from measurements within the same cell, formatted according to the standard output of 10x Genomics CellRanger. Both data types are log-transformed with a pseudocount of 1, retaining genes expressed in at least one cell and peaks expressed in at least ten cells. Subsequently, normalization and RE-TG pair filtering are performed to retain only those peaks and genes that appear in at least one high-scoring moderator pair. Following this, a TF-IDF transformation is applied to the scATAC-seq data. Additionally, each matrix in the scCNMF model is initialized using values drawn from a uniform distribution over the interval  $[0, 1]$ , providing a non-negative starting point for the iterative optimization process.

## PARAMETER SELECTION

The scCNMF model is relatively insensitive to the parameter  $K$ , however, setting  $K$  to small will lead to excessive enhancement that result in over-smoothing, potentially obscuring true intercellular dissimilarities (van Dijk et al., 2018). Tang et al. (Tang et al., 2024) further emphasized that setting  $K$  too high introduce unnecessary noise. Based on these insights, the default value of  $K$  is set to 20 in this study. The parameters  $\alpha$  and  $\beta$  are non-negative, reflecting the relative contributions of gene expression, chromatin accessibility, and regulatory potential. Since all three aspects are considered equally important, the default settings are  $\alpha = \beta = 1$ . Other default parameter values are  $\gamma = 10^4$ ,  $s = 0.25$ , and the number of RE-TG pairs  $m = 10^5$ .

## COMPARATIVE METHODS

**Seurat.** (Hao et al., 2021) scRNA-seq and scATAC-seq data are analyzed using Seurat v4 (WNN) according to the official tutorial (URL: <https://satijalab.org/seurat/>). The raw 10x data is used as input, with all parameters set to default.

**MOFA+.** (Argelaguet et al., 2020) scRNA-seq and scATAC-seq data processed with MOFA+ following its official tutorial (URL: <https://biofam.gith-ub.io/MOFA2/>). The same features used in scCNMF are input with normalized data. The number of potential factors is set to 20, while all other parameters remain the default. Since MOFA+ provides low-dimensional cell embeddings but lacks a clustering method, we applied KNN + Louvain clustering to ensure consistency with scCNMF, using MOFA+'s low-dimensional representations.

**intNMF.** (Chalise and Fridley, 2017) intNMF is analyzed according to its R package (URL: <https://github.com/cran/IntNMF>). The same features as in scCNMF are used to input normalized data, and all parameter sets are set to the default.

**scAI.** (Jin et al., 2020) Analysis conducted using scAI per its official tutorial (URL: <https://github.com/sqjin/scAI>). The same features as scCNMF are used with normalized input data. The latent factor is set to 20, while other parameters remained at default. Dimensionality reduction and clustering are performed using the reducedDims function with Leiden clustering as provided by scAI.

**scREG** (Duren et al., 2022) The scREG analysis followed the official tutorial (URL: <https://github.com/Durenlab/RegNMF>). The same features as in scCNMF are used with normalized input data. The latent factor is set to 20, with all other parameters set to default. Dimensionality reduction and clustering are performed using the default methods provided by scREG.

## EVALUATION INDICATORS

**Silhouette Score** (Rousseeuw, 1987). For each sample, the silhouette width is defined as  $(b - a) / (\max(a, b))$ , where  $a$  is the average distance from the sample to other samples within the same cluster, and  $b$  is the average distance from the sample to the nearest neighboring cluster. The mean silhouette score (or contour coefficient) is the average silhouette width across all samples. Silhouette scores range from -1 to +1, where higher values indicate better-defined clusters. In this article, we use the Scikit-learn implementation of silhouette\_score.

**Adjusted Rand Index (ARI)** (Hubert and Arabie, 1985). The ARI measures how well different cell types are clustered, compared to the ground truth labels, using low-dimensional embeddings from various integration methods. After clustering cells based on their low-dimensional embeddings, the ARI score is computed by comparing the clustering results with the ground truth cell labels. The ARI is defined as:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$$

where  $RI$  is the Rand index,  $E(RI)$  is the expected Rand index, and  $\max(RI)$  is the maximum possible value of the Rand index. The ARI score ranges from -1 to +1, where 0 indicates random clustering, and higher values indicate better alignment with the ground truth. In this article, we use the Scikit-learn implementation of `adjusted_rand_score`.

**Normalized Mutual Information (NMI)** (Vinh et al., 2010). NMI evaluates how well cells from different types are correctly clustered using the potential embedding method. It compares the clustering labels with the ground truth labels to assess the consistency between the two. The NMI is defined as:

$$NMI(U, V) = 2 \frac{MI(U, V)}{H(U) + H(V)}$$

where  $MI(U, V)$  is the mutual information between the clustering labels  $U$  and the ground truth labels  $V$ ,  $H(U)$  and  $H(V)$  are the entropies of  $U$  and  $V$ , respectively. In this article, we use the Scikit-learn implementation of `normalized_mutual_info_score`.

**Fowlkes-Mallows Index (FMI)** (Fowlkes and Mallows, 1983). The Fowlkes-Mallows Index (FMI) measures the similarity between two clusterings, either between the output of a clustering algorithm and the ground truth or between two hierarchical clusters. It is also related to the confusion matrix, as it reflects the balance between precision and recall. The FMI is defined as:

$$FMI = \sqrt{PPV \cdot TPR} = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

In this article, we use the Scikit-learn implementation of `fowlkes_mallows_score`.

**Calinski-Harabasz Index (CHI)** (Caliński and Harabasz, 1974). Also known as the Variance Ratio Criterion (VRC) is a widely used metric for evaluating the quality of clustering results. It quantifies the separation and cohesion of clusters by calculating the ratio of the between-class variance to the within-class variance. The index is mathematically defined as follows:

$$CHI = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)}$$

Where  $tr(B_k)$  denotes the trace of the interclass dispersion matrix,  $tr(W_k)$  represents the trace of the intraclass dispersion matrix,  $k$  is the number of clusters, and  $n$  is the total number of samples. Higher values of the CHI indicate better clustering performance. In this article, we use the Scikit-learn implementation of `calinski_harabasz_score`.

**Davies-Bouldin Index (DBI)** (Davies and Bouldin, 2009). DBI is a metric used to evaluate clusters' compactness and separation by considering inter-class and intra-class distances. It calculates the similarity between each pair of clusters and selects the minimum similarity score for each cluster. The final DBI value is the average of these minimum similarity scores across all pairs of clusters. The index is defined as:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left( \frac{S_i + S_j}{d(c_i, c_j)} \right)$$

Where  $S_i$  and  $S_j$  are the intra-cluster standard deviations of clusters  $i$  and  $j$ , respectively, and  $d(c_i, c_j)$  represents the distance between the centroids of clusters  $i$  and  $j$ . A smaller DBI value indicates better clustering performance. In this article, we use the Scikit-learn implementation of `davies_bouldin_score`.

**cell-type Local Inverse Simpson's Index (cLISI)** (Korsunsky et al., 2019). The cLISI metric is used to evaluate the local purity of cell-type labels in a low-dimensional embedding by examining the composition of the  $k$  nearest neighbors of each cell. It quantifies how homogeneous or mixed the cell-type

distribution is within the neighborhood of each cell in the embedding space. The index is mathematically defined as follows:

$$cLISI_i = \left( \sum_l p_l^2 \right)^{-1}$$

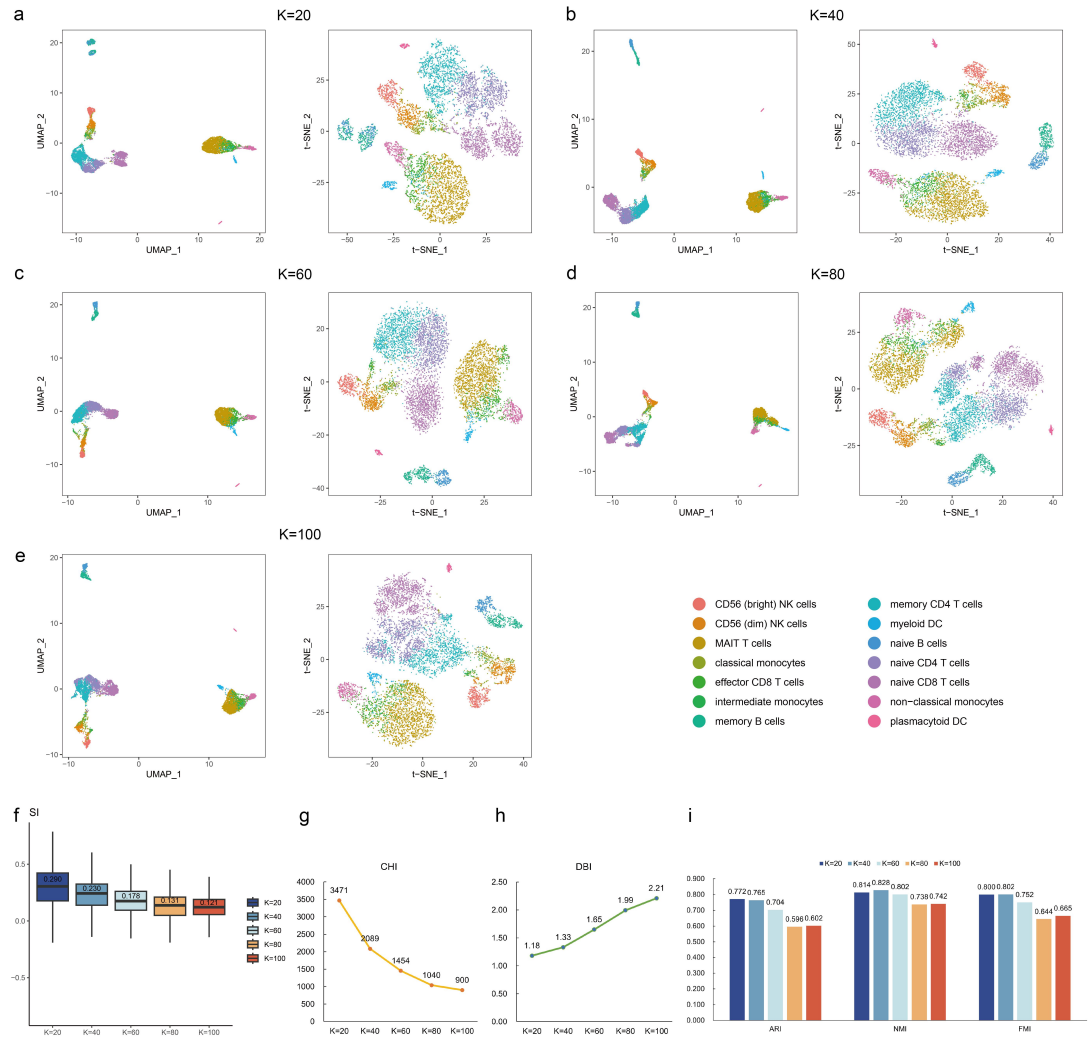
Where  $k$  denotes the number of nearest neighbors considered for each cell,  $l$  indexes over all distinct cell-type labels,  $p_l$  represents the fraction of the  $k$  neighbors assigned to cell type  $l$ .  $cLISI_i$  denotes the  $cLISI$  value computed for cell  $i$  based on the heterogeneity of its neighborhood. For assessing cell-type separation, a  $cLISI$  value closer to 1 indicates better performance, reflecting higher purity and clearer separation between cell types.

## REFERENCES

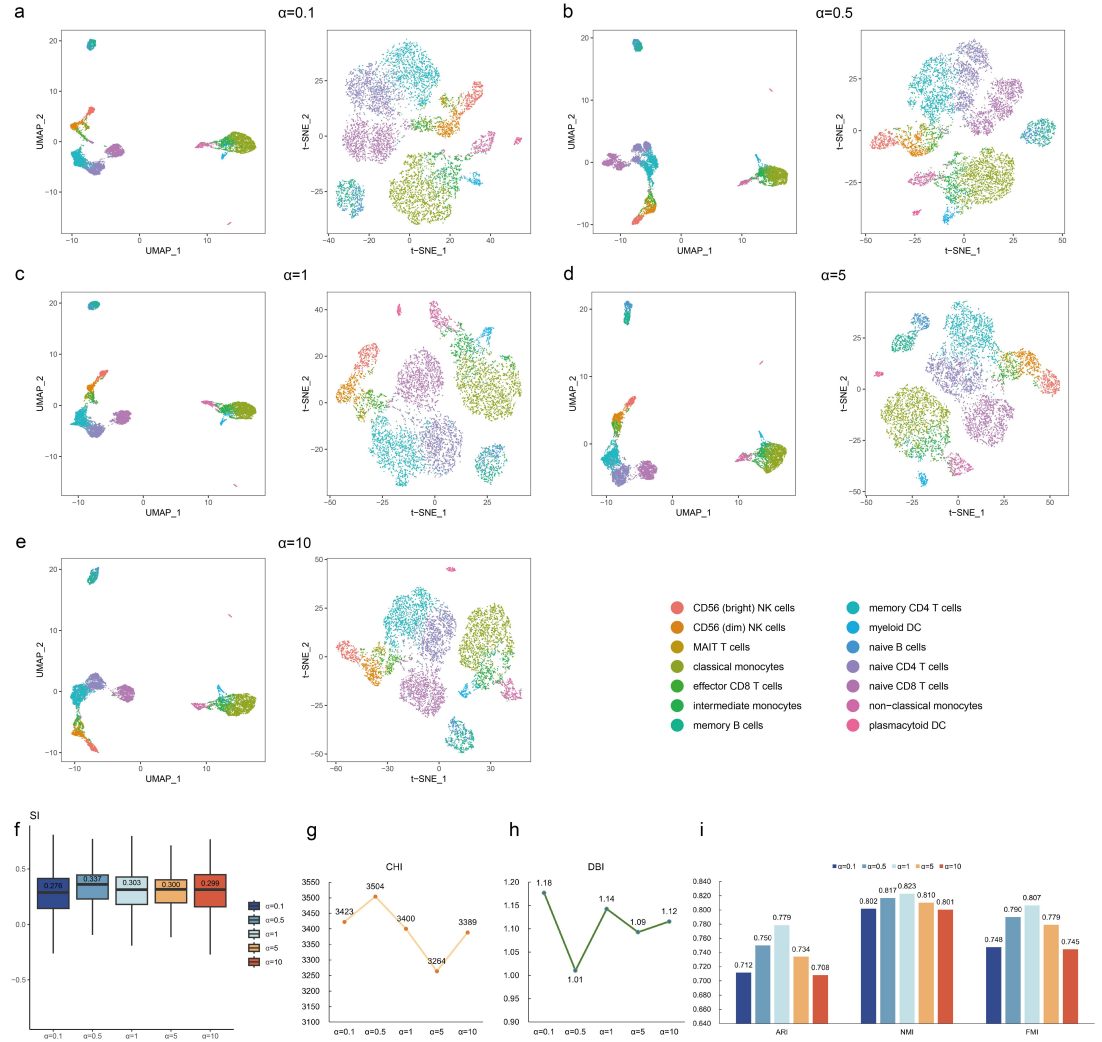
- Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17.
- Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.
- Chalise, P. and Fridley, B. L. (2017). Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. *PloS one*, 12(5):e0176278.
- Davies, D. L. and Bouldin, D. W. (2009). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.
- Duren, Z., Chang, F., Naqing, F., Xin, J., Liu, Q., and Wong, W. H. (2022). Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with screg. *Genome biology*, 23(1):114.
- Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.
- Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Budhanuru Ramaraju, S., Campbell, L. I., Carbajo Martinez, M., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N. H., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Garcia Giron, C., Genez, T., Grigoriadis, D., Ghattaoraya, G. S., Gonzalez Martinez, J., Gurbich, T. A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, G. A., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Ogeh, D. N., Pandian Sakthivel, M., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J. G., Abdul Salam, A. I., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, N. L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G. R., Keatley, J., Loveland, J. E., Moore, B., Mudge, J. M., Naamati, G., Tate, J., Trevanion, S. J., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, R. D., Martin, F. J., and Yates, A. D. (2024). Ensembl 2024. *Nucleic acids research*, 52(D1):D891–D899.
- Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P., and Zhang, Y. (2023). Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic acids research*, 51(D1):D870–D876.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.
- Huizing, G.-J., Deutschmann, I. M., Peyré, G., and Cantini, L. (2023). Paired single-cell multi-omics data integration with mowgli. *Nature Communications*, 14(1):7711.
- Jin, S., Zhang, L., and Nie, Q. (2020). scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome biology*, 21:1–19.
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh,



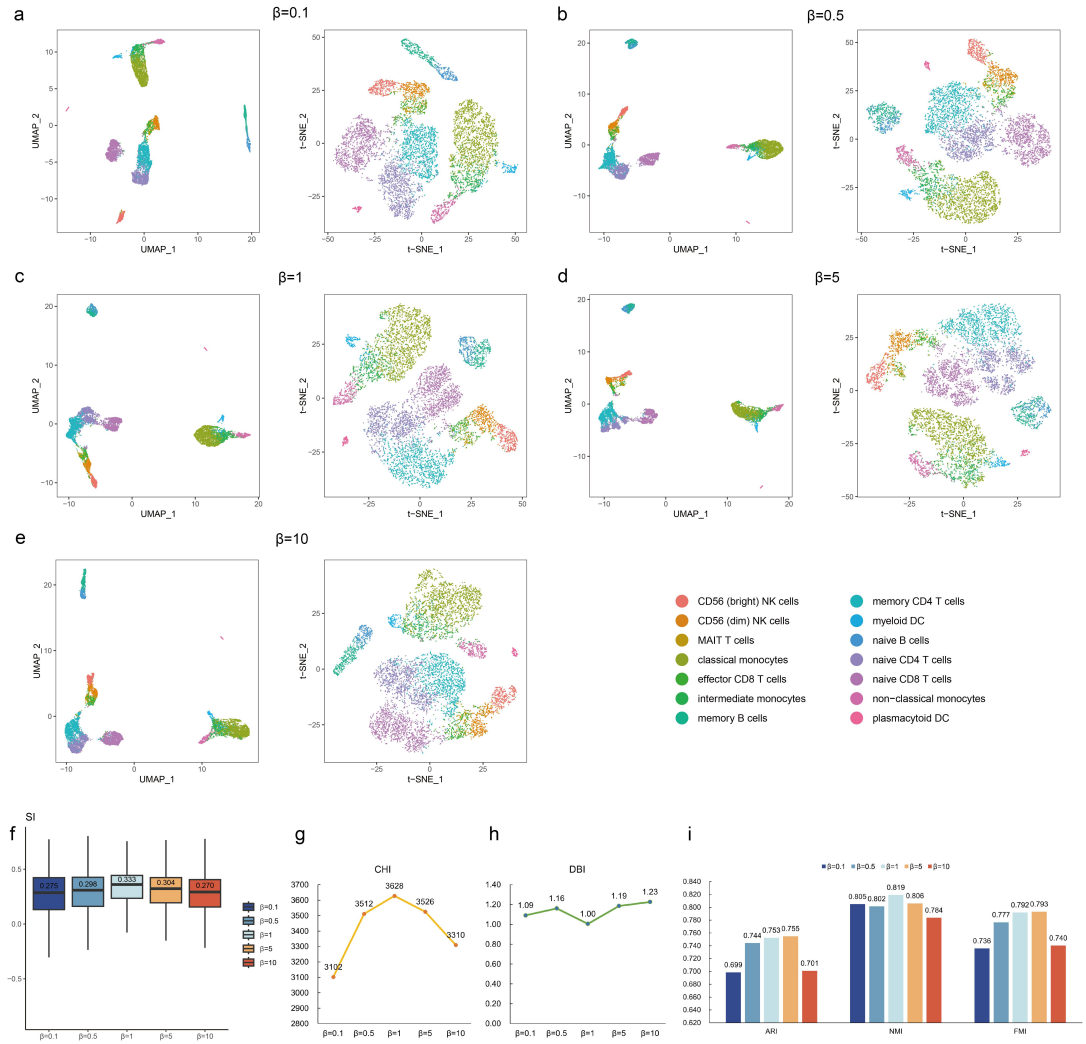
155 P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with  
 156 harmony. *Nature methods*, 16(12):1289–1296.  
 157 Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.  
 158 *Journal of computational and applied mathematics*, 20:53–65.  
 159 Tang, S., Cui, X., Wang, R., Li, S., Li, S., Huang, X., and Chen, S. (2024). sccase: accurate and inter-  
 160 pretable enhancement for single-cell chromatin accessibility sequencing data. *Nature Communications*,  
 161 15(1):1629.  
 162 van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer,  
 163 C. L., Pattabiraman, D., Bieri, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2018).  
 164 Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.  
 165 Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison:  
 166 Variants, properties, normalization and correction for chance. *Journal of machine learning research*,  
 167 11:2837–2854.



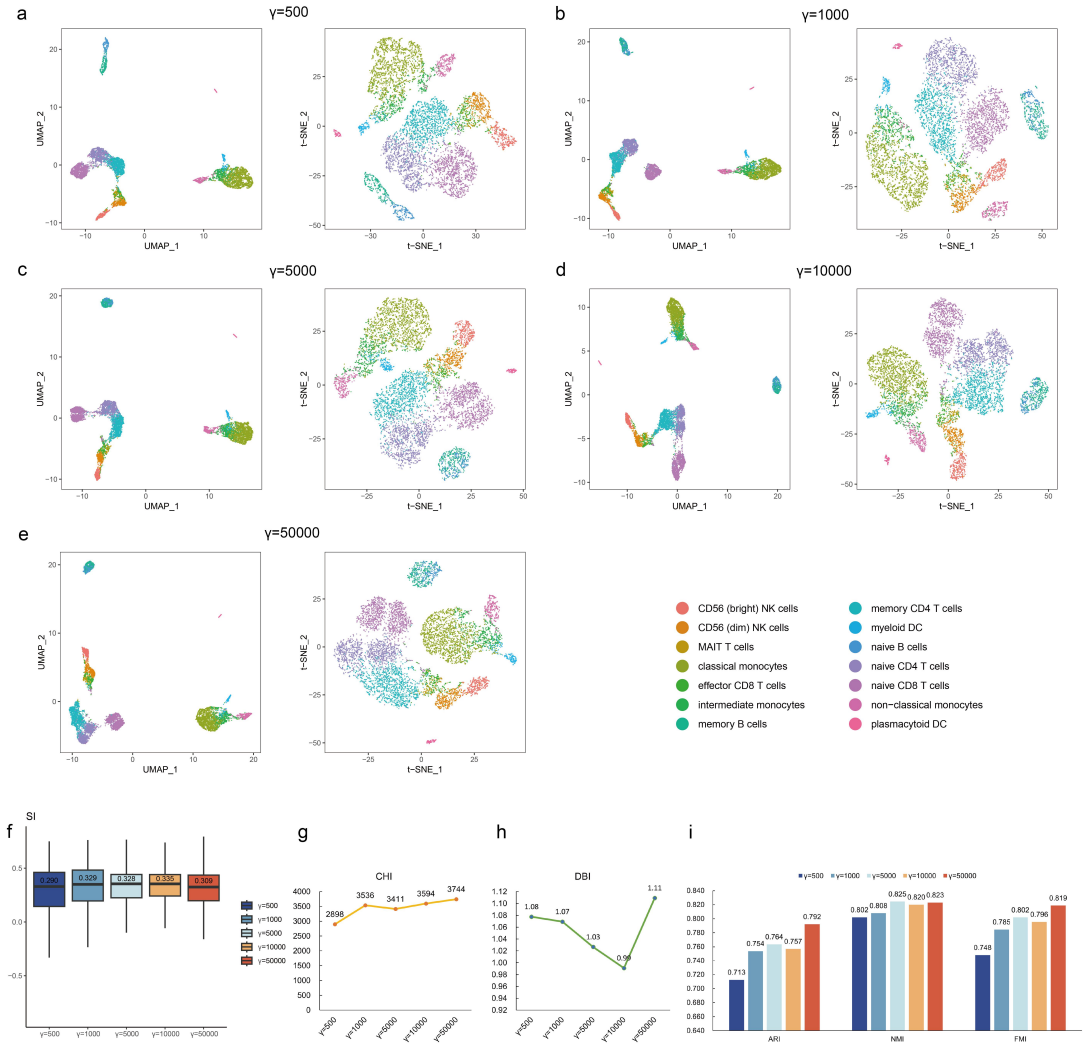
**Supplementary Figure 1.** Effect of the choice of different parameters  $K$  on the integration results. (a) UMAP and t-SNE results for  $K = 20$ . (b) UMAP and t-SNE results for  $K = 40$ . (c) UMAP and t-SNE results for  $K = 60$ . (d) UMAP and t-SNE results for  $K = 80$ . (e) UMAP and t-SNE results for  $K = 100$ . (f) SI for different parameters. (g) CHI for different parameters. (h) DBI for different parameters. (i) ARI, NMI and FMI for different parameters.



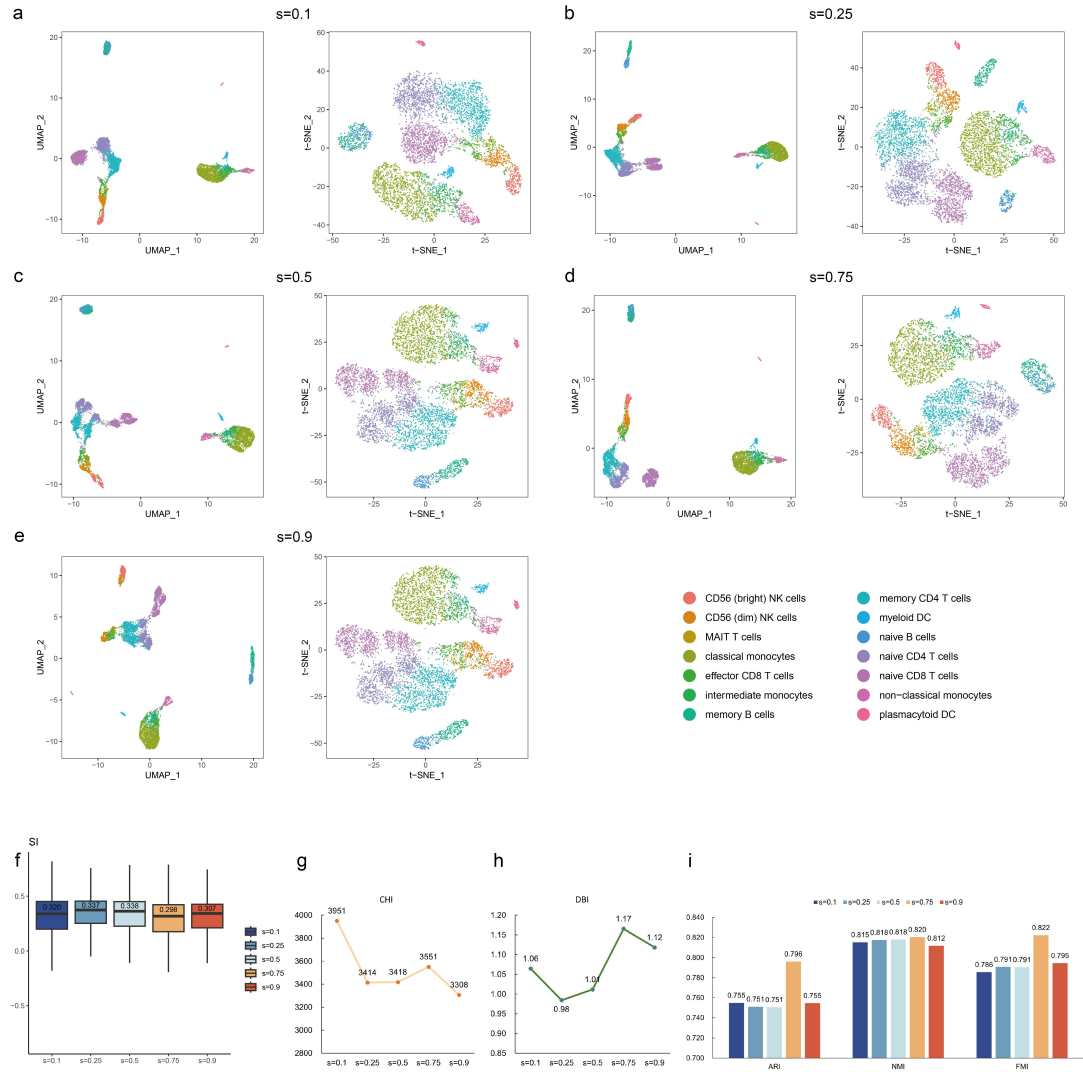
**Supplementary Figure 2.** Effect of the choice of different parameters  $\alpha$  on the integration results. (a) UMAP and t-SNE results for  $\alpha = 0.1$ . (b) UMAP and t-SNE results for  $\alpha = 0.5$ . (c) UMAP and t-SNE results for  $\alpha = 1$ . (d) UMAP and t-SNE results for  $\alpha = 5$ . (e) UMAP and t-SNE results for  $\alpha = 10$ . (f) SI for different parameters. (g) CHI for different parameters. (h) DBI for different parameters. (i) ARI, NMI and FMI for different parameters.



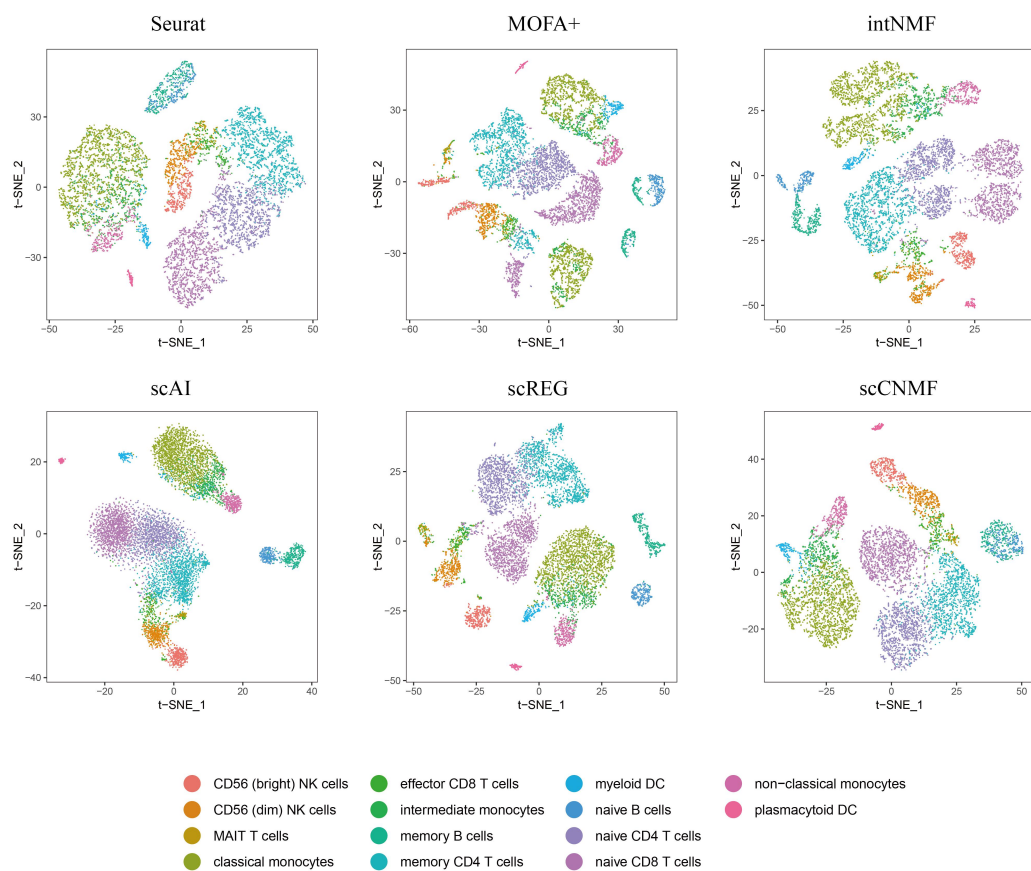
**Supplementary Figure 3.** Effect of the choice of different parameters  $\beta$  on the integration results. (a) UMAP and t-SNE results for  $\beta = 0.1$ . (b) UMAP and t-SNE results for  $\beta = 0.5$ . (c) UMAP and t-SNE results for  $\beta = 1$ . (d) UMAP and t-SNE results for  $\beta = 5$ . (e) UMAP and t-SNE results for  $\beta = 10$ . (f) SI for different parameters. (g) CHI for different parameters. (h) DBI for different parameters. (i) ARI, NMI and FMI for different parameters.



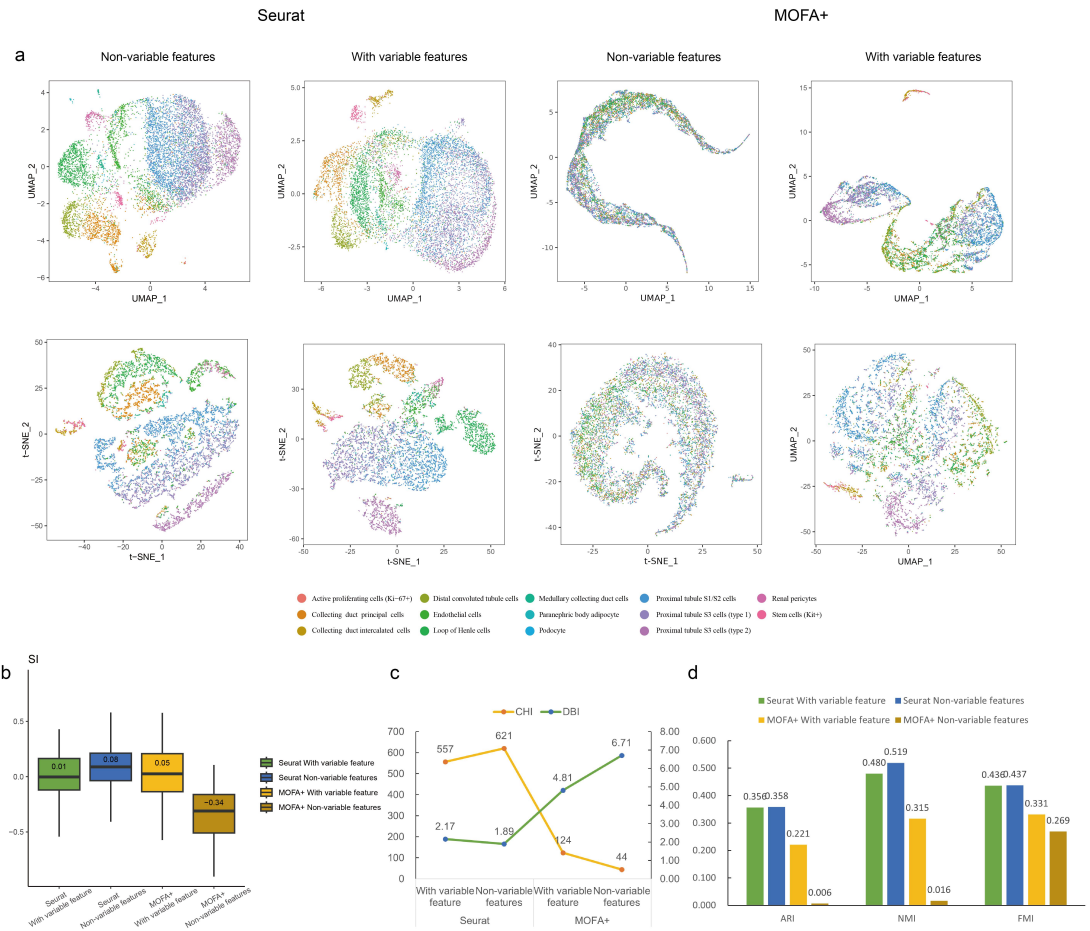
**Supplementary Figure 4.** Effect of the choice of different parameters  $\gamma$  on the integration results. (a) UMAP and t-SNE results for  $\gamma = 500$ . (b) UMAP and t-SNE results for  $\gamma = 1000$ . (c) UMAP and t-SNE results for  $\gamma = 5000$ . (d) UMAP and t-SNE results for  $\gamma = 10000$ . (e) UMAP and t-SNE results for  $\gamma = 50000$ . (f) SI for different parameters. (g) CHI for different parameters. (h) DBI for different parameters. (i) ARI, NMI and FMI for different parameters.



**Supplementary Figure 5.** Effect of the choice of different parameters  $s$  on the integration results. (a) UMAP and t-SNE results for  $s = 0.1$ . (b) UMAP and t-SNE results for  $s = 0.25$ . (c) UMAP and t-SNE results for  $s = 0.5$ . (d) UMAP and t-SNE results for  $s = 0.75$ . (e) UMAP and t-SNE results for  $s = 0.9$ . (f) SI for different parameters. (g) CHI for different parameters. (h) DBI for different parameters. (i) ARI, NMI and FMI for different parameters.

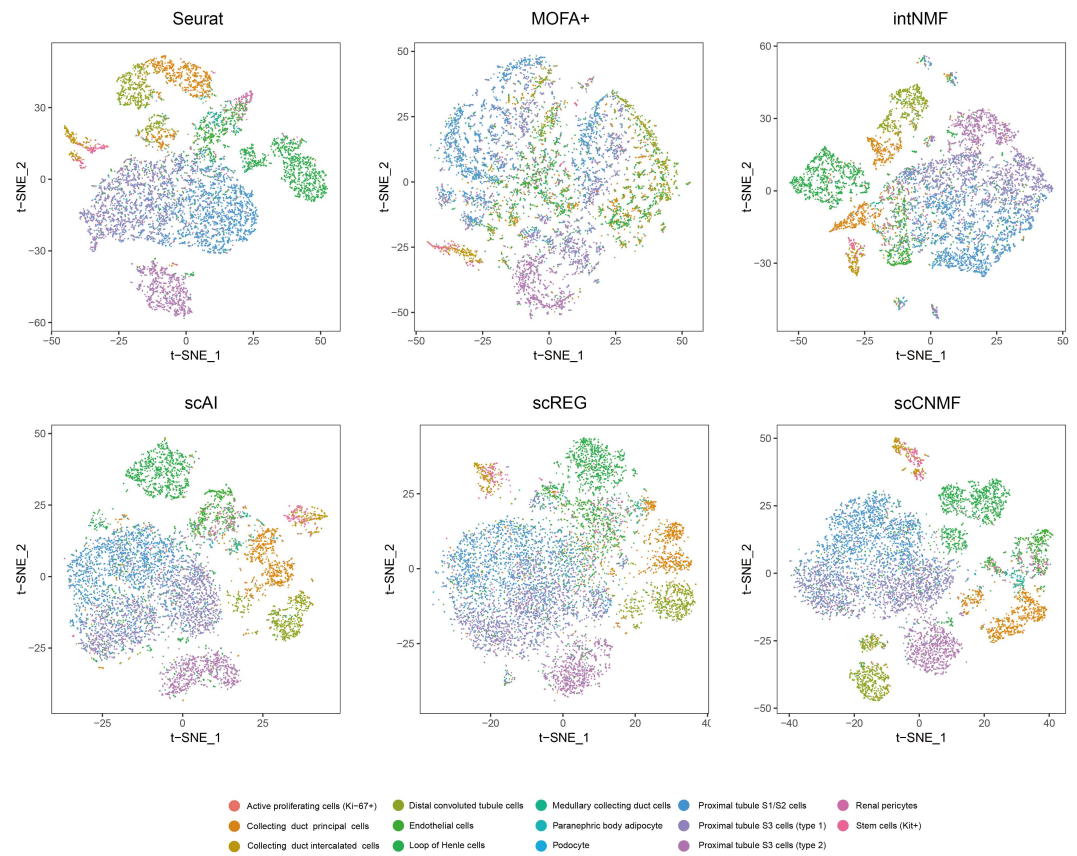


**Supplementary Figure 6.** t-SNE visualization of the integration results on the PBMC dataset using different methods.



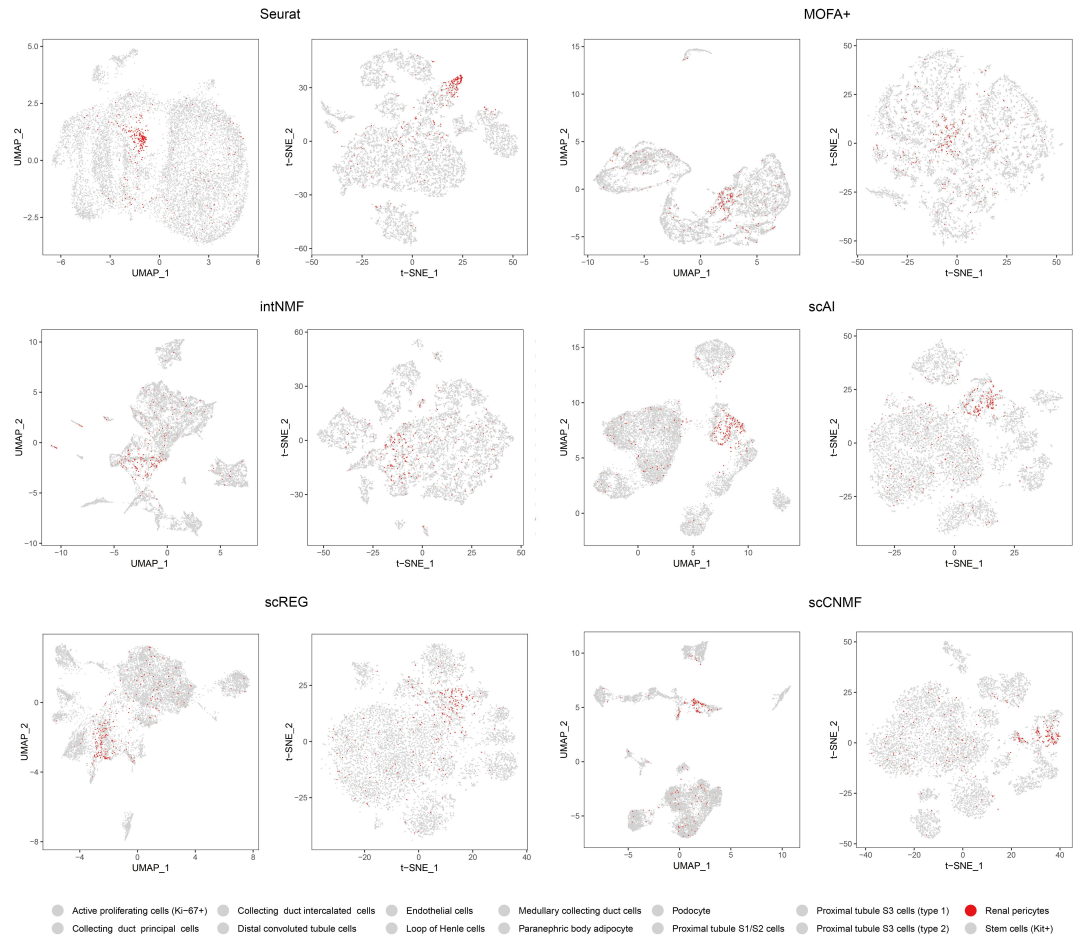
**Supplementary Figure 7.** Integration analysis by Seurat and MOFA+ with and without the selection of highly variable features. (a) UMAP and t-SNE visualizations before and after the selection of highly variable features for both methods. (b) Comparison of SI before and after feature selection for both methods. (c) Comparison of CHI and DBI before and after feature selection for both methods. (d) ARI, NMI and FMI before and after feature selection for both methods.



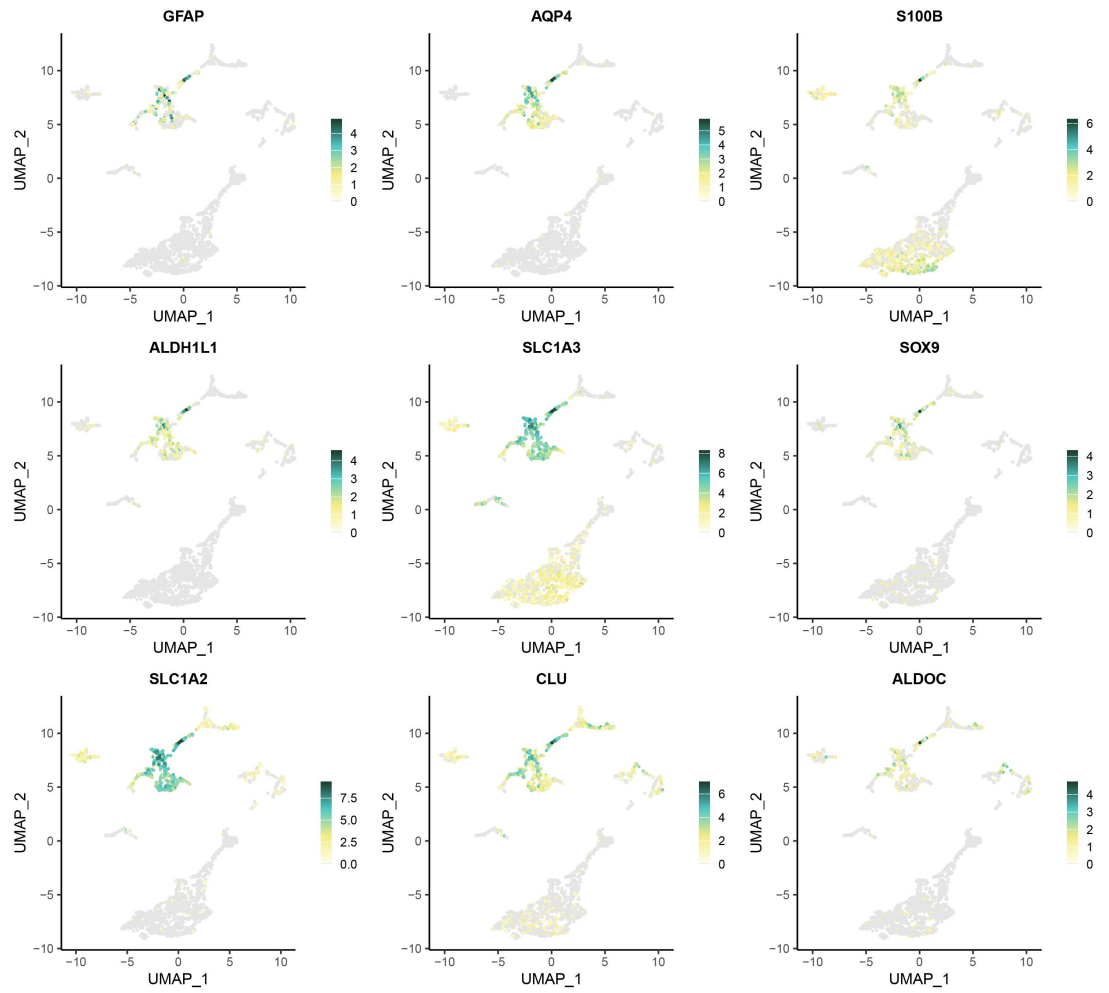


**Supplementary Figure 8.** t-SNE visualization of the integration results on the mouse kidney dataset using different methods.

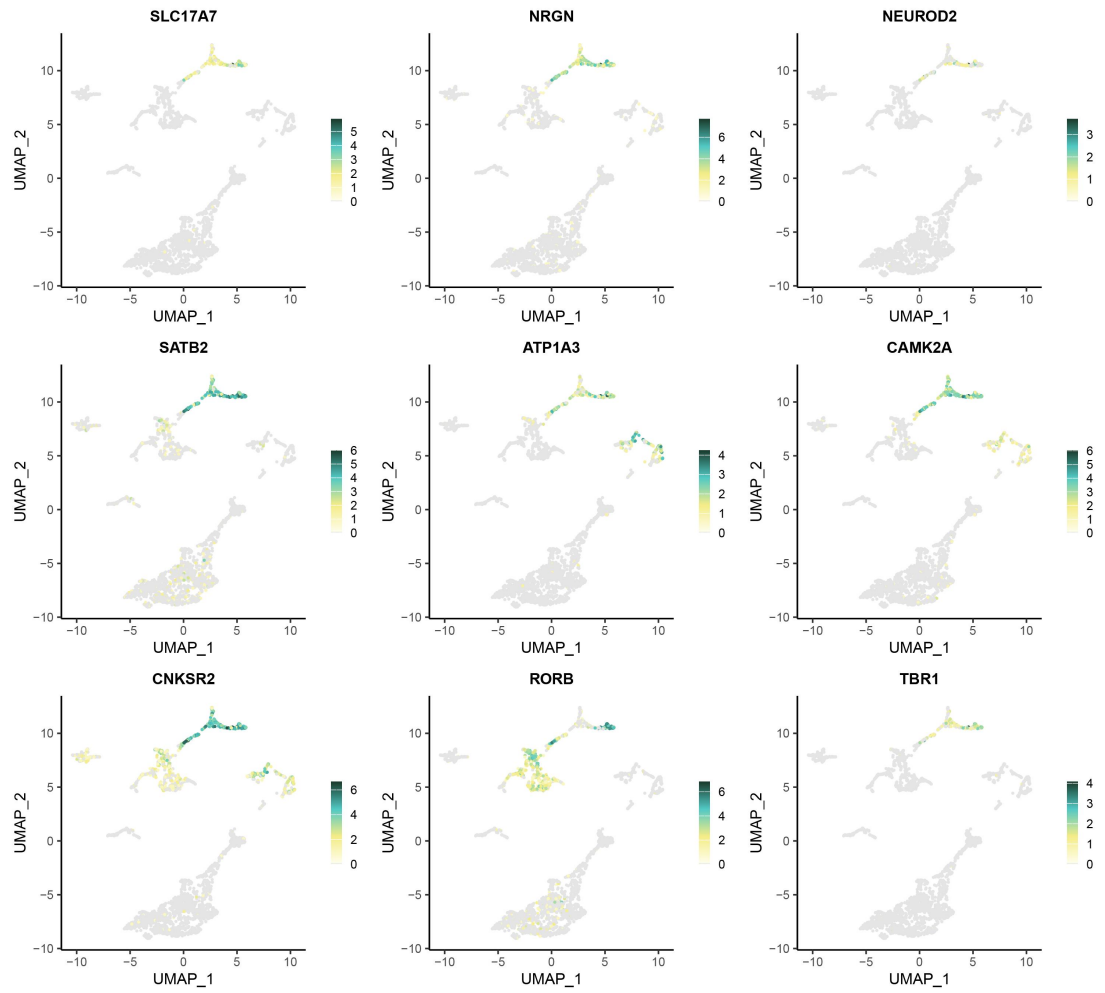




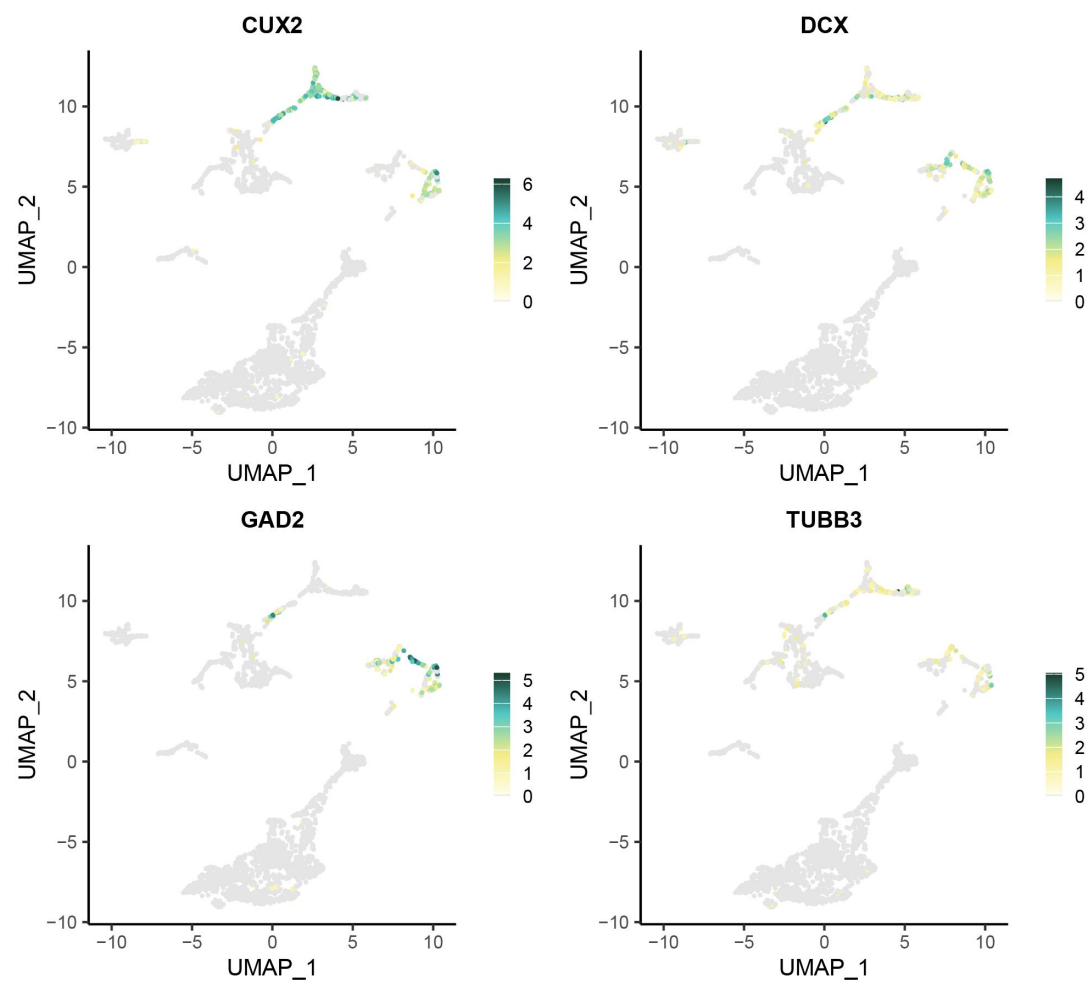
**Supplementary Figure 9.** Clustering of Renal pericytes cells on the mouse kidney dataset using different methods.



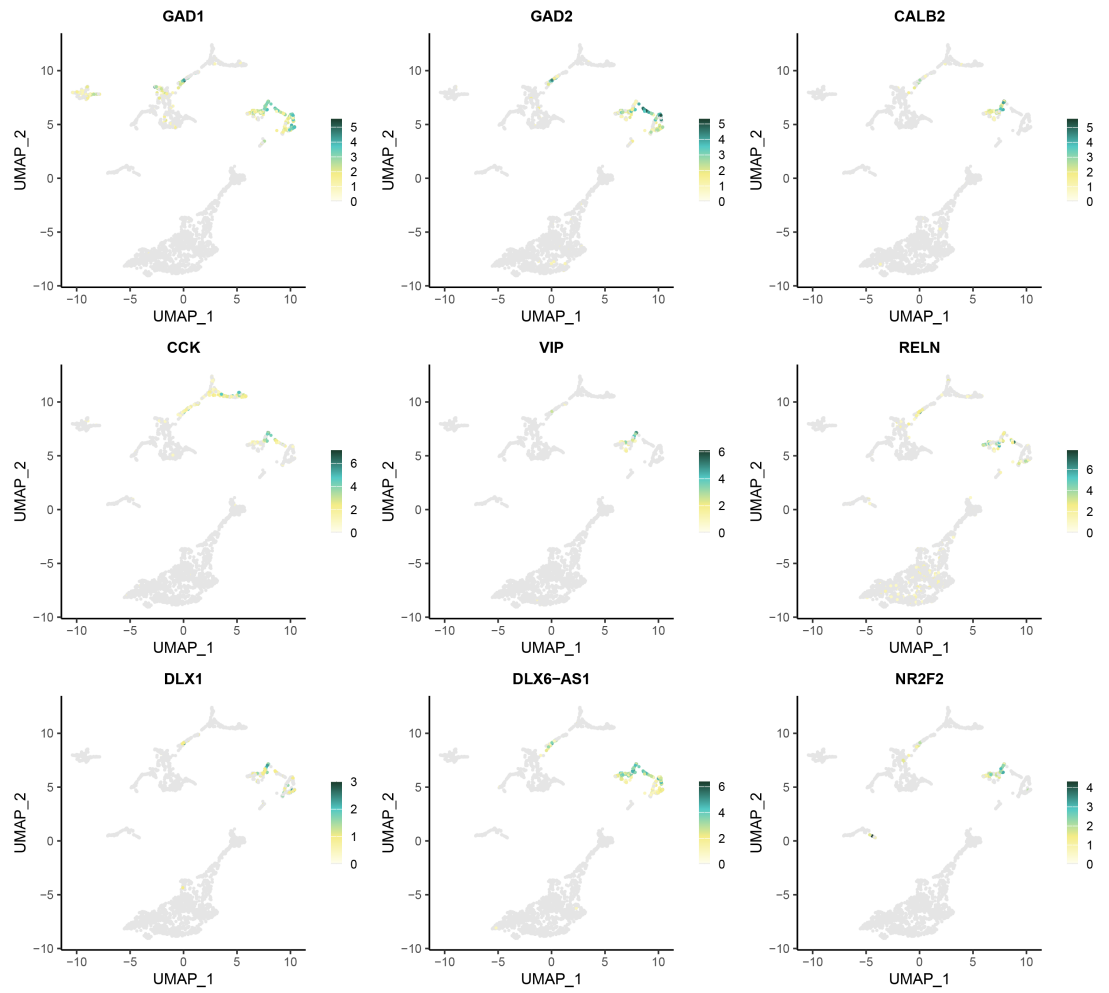
**Supplementary Figure 10.** Distribution of marker genes for Astrocyte.



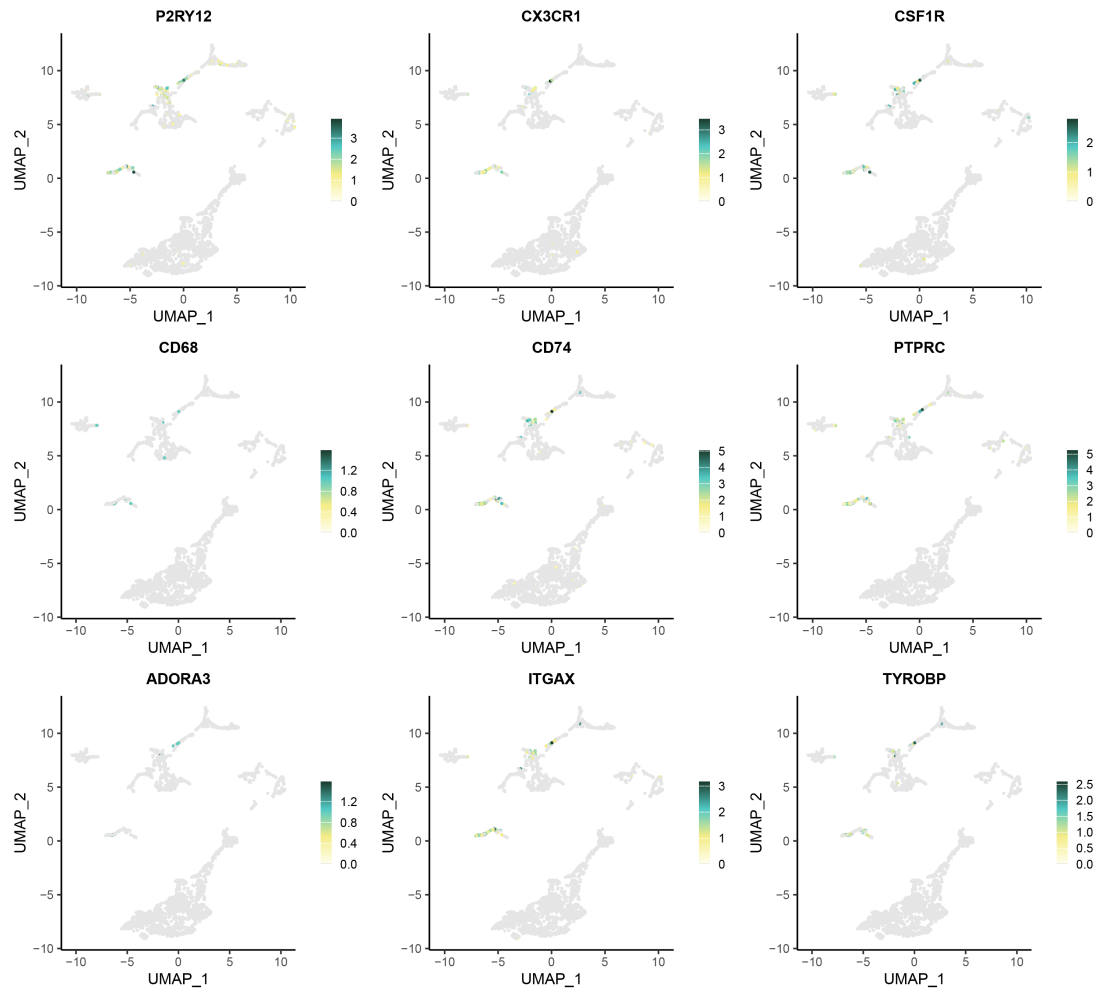
**Supplementary Figure 11.** Distribution of marker genes for Excitatory neuron.



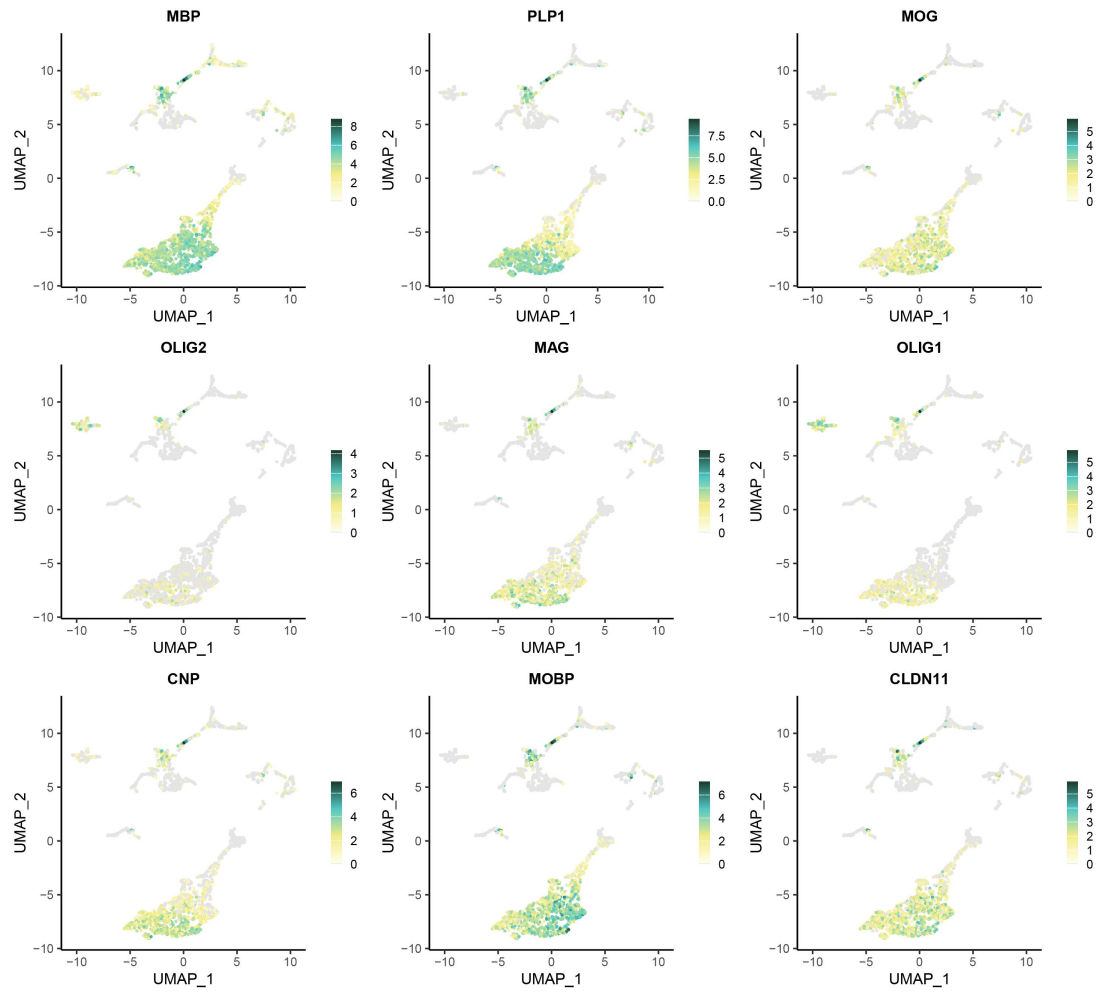
**Supplementary Figure 12.** Distribution of marker genes for Immature neuron.



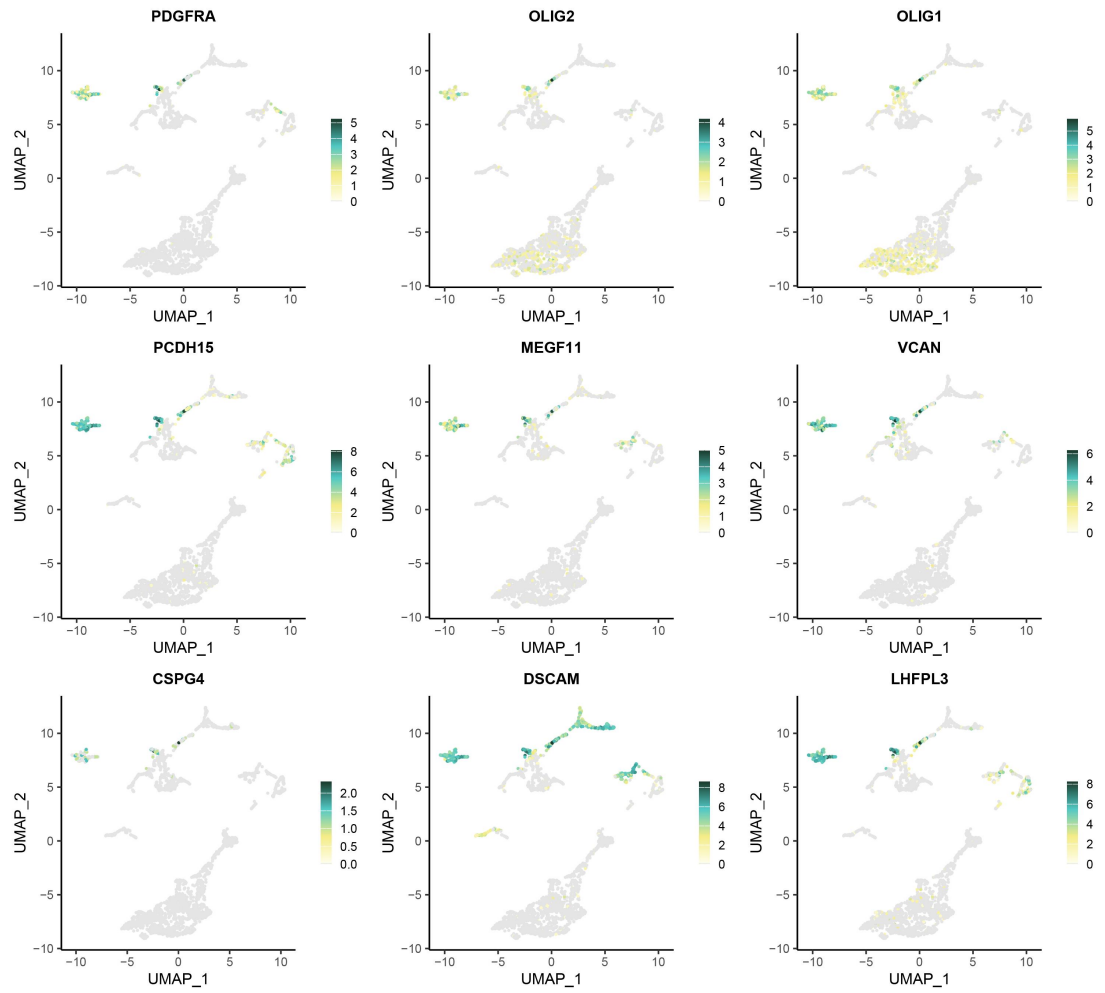
**Supplementary Figure 13.** Distribution of marker genes for Inhibitory neuron.



**Supplementary Figure 14.** Distribution of marker genes for Microglial cell.

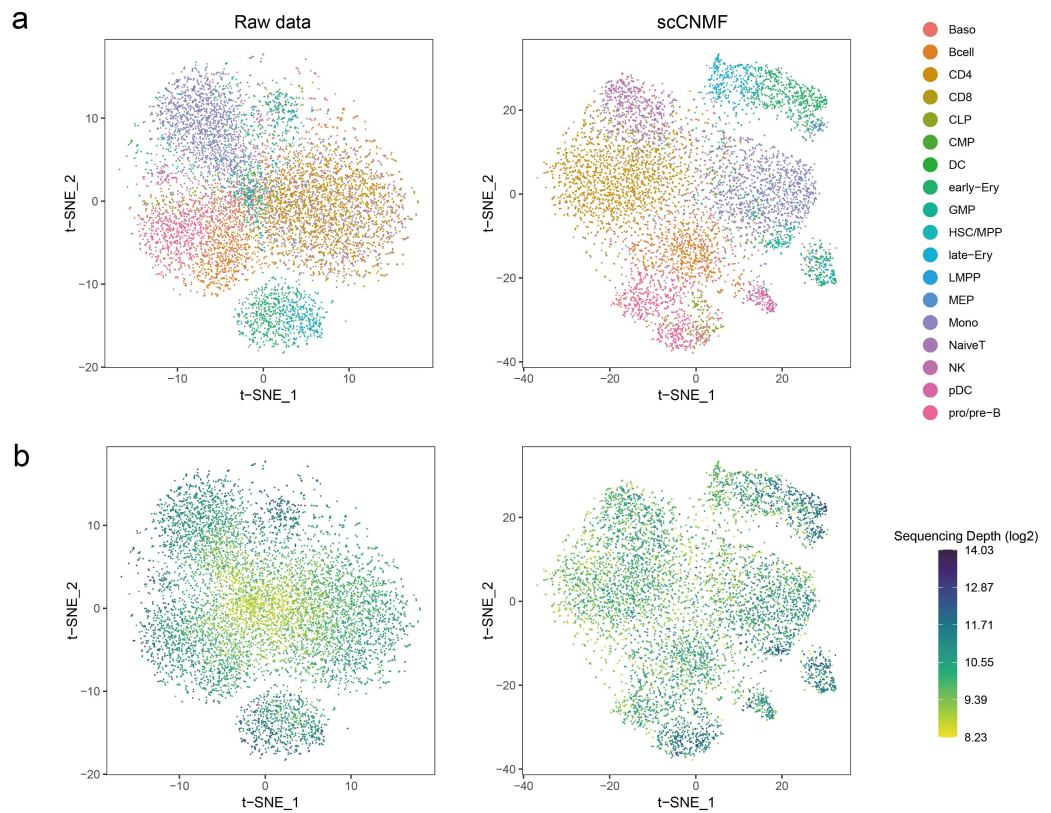


**Supplementary Figure 15.** Distribution of marker genes for Oligodendrocyte.

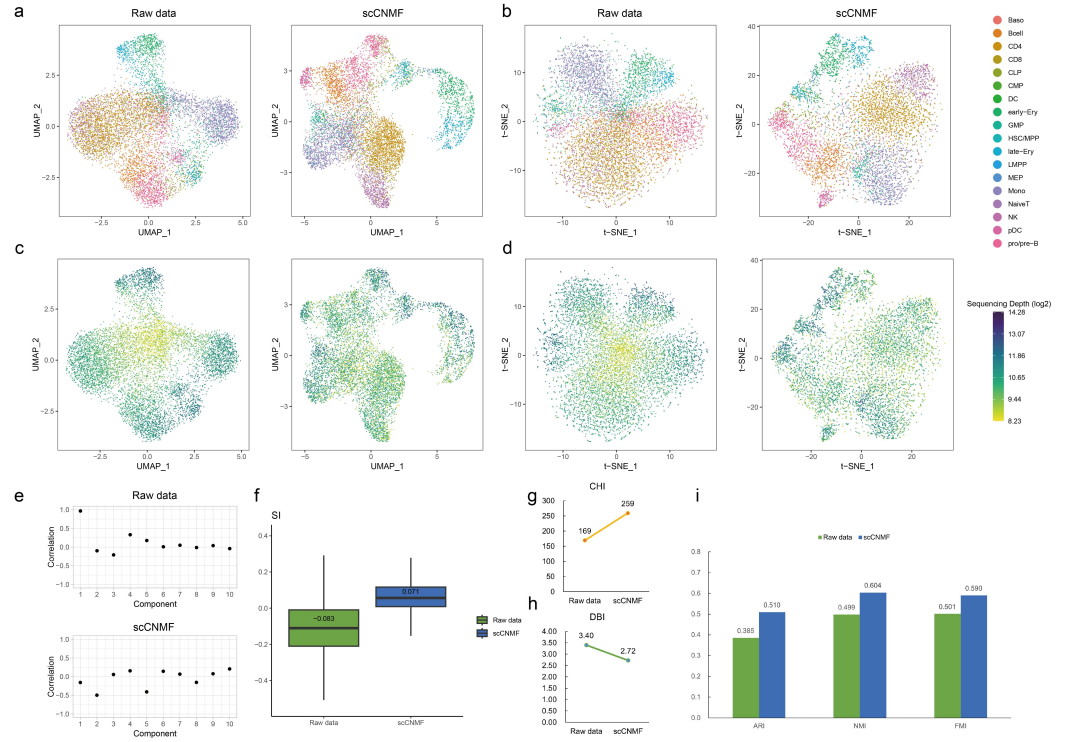


**Supplementary Figure 16.** Distribution of marker genes for Oligodendrocyte precursor.

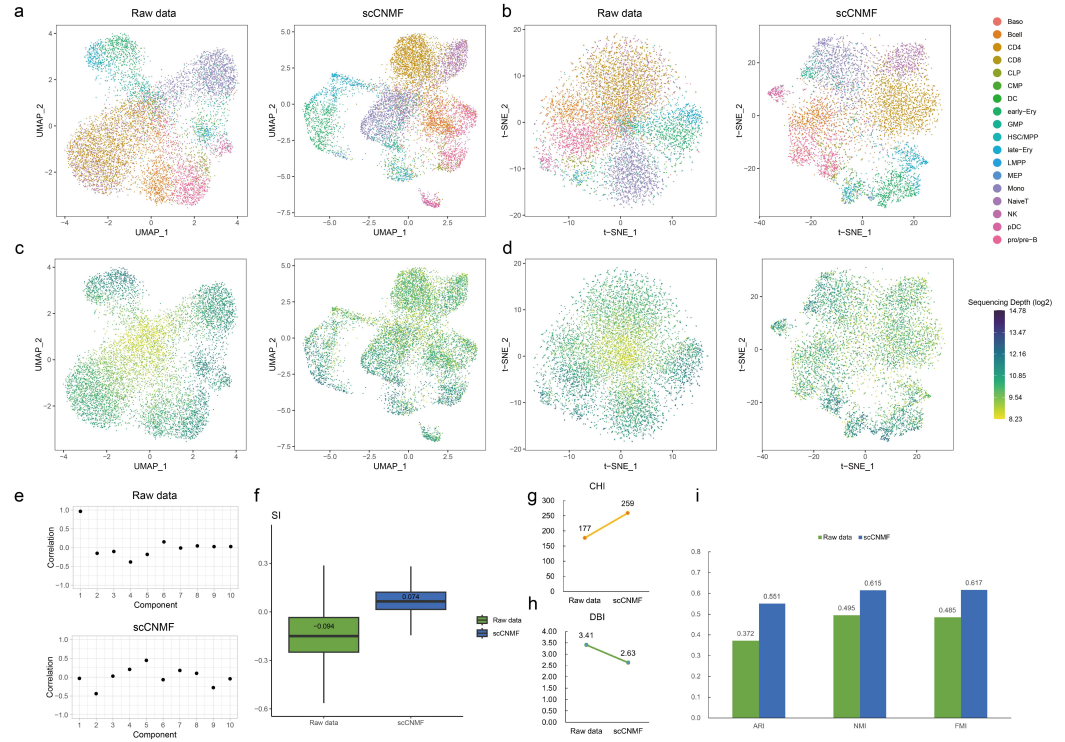




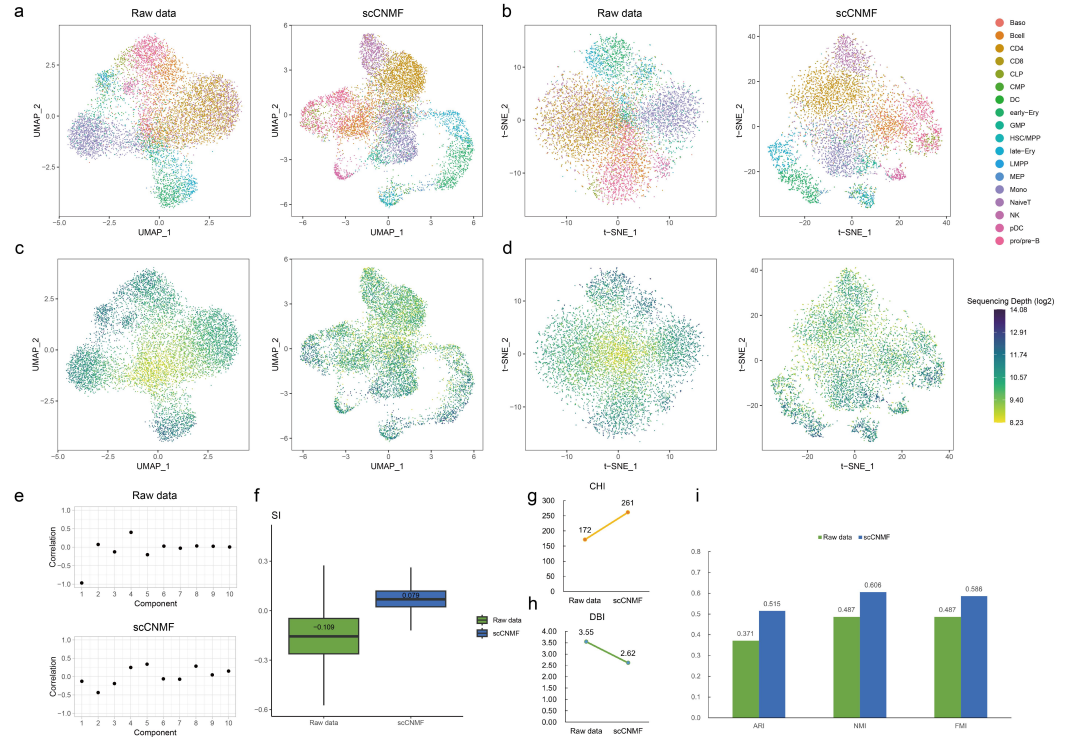
**Supplementary Figure 17.** t-SNE visualization results of scATAC-seq data before and after enhancement by scCNMF. (a) Colored according to real cell labeling. (b) Colored according to sequencing depth.



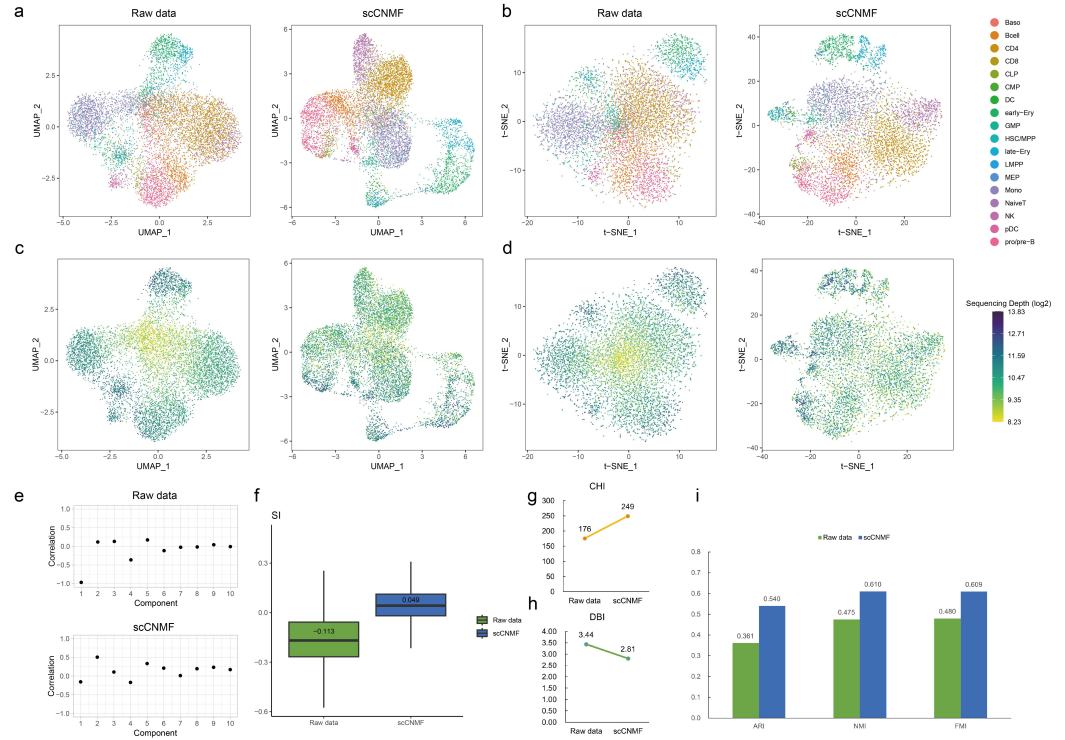
**Supplementary Figure 18.** Evaluation of enhancement performance of scCNMF data on sampled dataset 2. (a) UMAP results of pre-and post-enhancement data, colored by cell-true labeling. (b) t-SNE results of pre-and post-enhancement data, colored by cell-true labeling. (c) UMAP results of pre-and post-enhancement data are colored by sequencing depth. (d) t-SNE results of pre-and post-enhancement data, colored by sequencing depth. (e) Correlation coefficients between SVD components and sequencing depth before and after data enhancement. (f) SI before and after data enhancement. (g) CHI before and after data enhancement. (h) DBI before and after data enhancement. (i) ARI, NMI, and FMI before and after data enhancement.



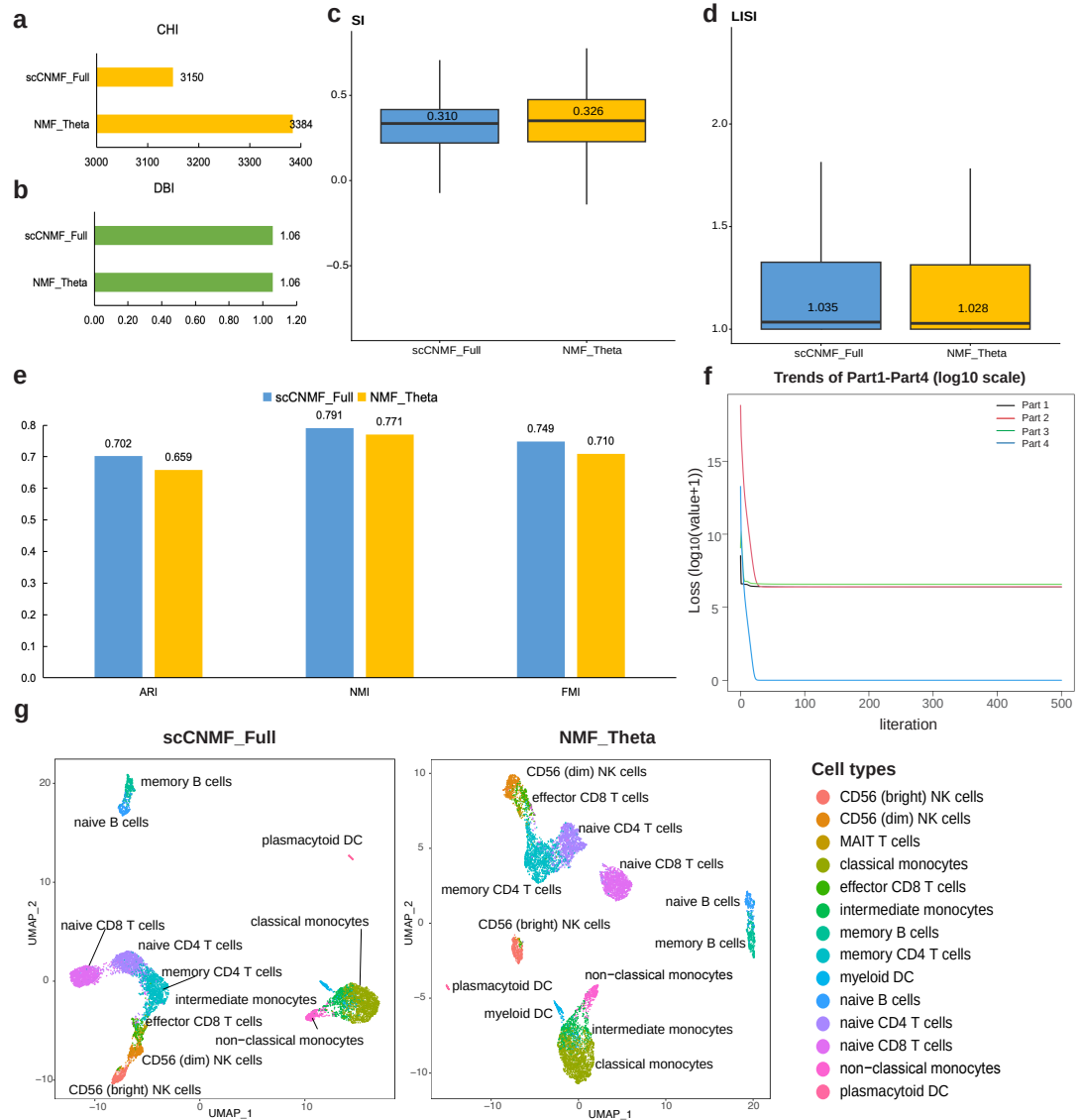
**Supplementary Figure 19.** Evaluation of enhancement performance of scCNMF data on sampled dataset 3. (a) UMAP results of pre-and post-enhancement data, colored by cell-true labeling. (b) t-SNE results of pre-and post-enhancement data, colored by cell-true labeling. (c) UMAP results of pre-and post-enhancement data are colored by sequencing depth. (d) t-SNE results of pre-and post-enhancement data, colored by sequencing depth. (e) Correlation coefficients between SVD components and sequencing depth before and after data enhancement. (f) SI before and after data enhancement. (g) CHI before and after data enhancement. (h) DBI before and after data enhancement. (i) ARI, NMI, and FMI before and after data enhancement.



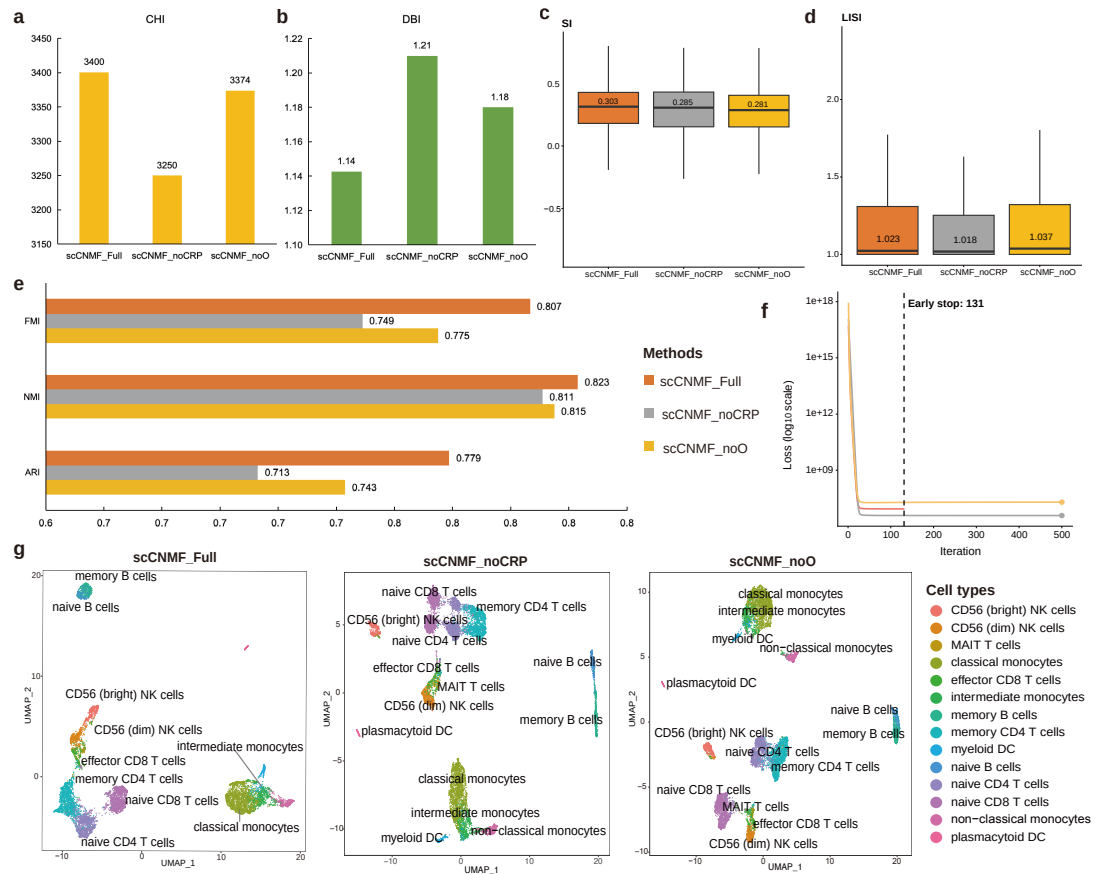
**Supplementary Figure 20.** Evaluation of enhancement performance of scCNMF data on sampled dataset 4. (a) UMAP results of pre-and post-enhancement data, colored by cell-true labeling. (b) t-SNE results of pre-and post-enhancement data, colored by cell-true labeling. (c) UMAP results of pre-and post-enhancement data are colored by sequencing depth. (d) t-SNE results of pre-and post-enhancement data, colored by sequencing depth. (e) Correlation coefficients between SVD components and sequencing depth before and after data enhancement. (f) SI before and after data enhancement. (g) CHI before and after data enhancement. (h) DBI before and after data enhancement. (i) ARI, NMI, and FMI before and after data enhancement.



**Supplementary Figure 21.** Evaluation of enhancement performance of scCNMF data on sampled dataset 5. (a) UMAP results of pre-and post-enhancement data, colored by cell-true labeling. (b) t-SNE results of pre-and post-enhancement data, colored by cell-true labeling. (c) UMAP results of pre-and post-enhancement data are colored by sequencing depth. (d) t-SNE results of pre-and post-enhancement data, colored by sequencing depth. (e) Correlation coefficients between SVD components and sequencing depth before and after data enhancement. (f) SI before and after data enhancement. (g) CHI before and after data enhancement. (h) DBI before and after data enhancement. (i) ARI, NMI, and FMI before and after data enhancement.



**Supplementary Figure 22.** Comparison of full scCNMF (scCNMF\_Full) and NMF applied to the final similarity matrix (NMF\_Theta). (a) CHI scores of the two methods. (b) DBI scores of the two methods. (c) Silhouette Index (SI) of the two methods. (d) Cell-type Lisi values measuring local purity. (e) Label-consistency metrics (ARI, NMI, FMI). (f) Optimization trajectories of the four objective-function components (Part 1–Part 4) across 500 iterations, shown on a log10 scale. All four terms decrease substantially, indicating that the -weighted term does not dominate optimization. (g) UMAP visualizations generated from the full scCNMF (scCNMF\_Full) model and the NMF\_Theta baseline.



**Supplementary Figure 23.** Ablation study evaluating the contributions of the CRP component and the sparse similarity matrix  $O$ . (a) CHI scores for the full model (scCNMF\_Full), the model without CRP (scCNMF\_noCRP), and the model without  $O$  (scCNMF\_noO). (b) DBI scores for the three models. (c) SI values for the three models. (d) LISI scores for the three models. (e) External label-consistency metrics (ARI, NMI, FMI). (f) Optimization trajectories (log<sub>10</sub> scale) showing much faster convergence of the full model. (g) UMAP visualizations generated from the three models.