Supplementary materials of "A personalized low-rank subspace clustering method based on locality and similarity constraints for scRNA-seq data analysis"

Tian-Jing Qiao[1], Jin-Xing Liu[1], Member, IEEE, Jun-Liang Shang[1], Shasha Yuan[1], Chun-Hou Zheng[1], Member, IEEE, and Juan Wang[1]*

1. School of Computer Science, Qufu Normal University, Rizhao 276826, China

A. Supplementary Table

TABLE S1

Optimal parameters on twelve scRNA-seq datasets

Dataset	λ	β	γ
Ramskold	10^{1}	1	10^{2}
Li_islet	10^{-1}	1	10^{1}
Treutlein	$10^{1.3}$	1	10^{0}
Ting	10^{0}	1	10^{1}
Goolam	$10^{1.4}$	1	$10^{0.1}$
Deng	10^{-2}	1	10^{0}
Engel4	10^{-2}	1	10^{1}
Pollen	10^{1}	1	10^{1}
Darmanis	$10^{-0.8}$	1	$10^{0.2}$
Kolod	$10^{1.4}$	1	$10^{0.8}$
Tasic	10^{-2}	1	$10^{0.5}$
Zeisel	10^{-2}	1	10^{1}

B. Supplementary Algorithm

Algorithm S1. ADMM for solving the objective function of PLRLS

Input: gene expression matrix $\mathbf{X} \in R^{m \times n}$, parameter λ, β, γ.

Initial:

$\mathbf{J}^0 = \mathbf{C}^0 = \mathbf{Z}^0$, $\mathbf{Y}_1^0 = \mathbf{Y}_2^0 = \mathbf{0}$, $k = 0$, $\mu_0 = 0.01$, $\mu_{\max} = 1e8$, $\rho = 1.2$, $\max_{iter} = 80$, $tol = 1e-5$.

While not converged and $k < \max_{iter}$ do

(1) Update $\mathbf{Z}^{k+1}$ as $\mathbf{Z}^{k+1} = (\gamma \mathbf{X}^T \mathbf{X} + 2\mu_k \mathbf{I} + \beta \mathbf{I})^{-1} \times (\gamma \mathbf{X}^T \mathbf{X} + \beta \mathbf{S} + \mu_k \mathbf{J}_k + \mu_k \mathbf{C}_k - \mathbf{Y}_1^k - \mathbf{Y}_2^k)$

(2) Update $\mathbf{J}^{k+1}$ as $\mathbf{J}^{k+1} = \mathrm{D}_{\eta}(\mathbf{Z}^{k+1} + \mathbf{Y}_1^k / \mu_k)$

(3) Update $\mathbf{C}^{k+1}$ as $\mathbf{C}^{k+1} = \max[P^{W}_{1/\mu}(\mathbf{Z}^{k+1} + \frac{\mathbf{Y}_2^k}{\mu_k}), 0]$

(4) Update $\mathbf{Y}_1^{k+1}$, $\mathbf{Y}_2^{k+1}$ and μ_{k+1} as

$\mathbf{Y}_1^{k+1} = \mathbf{Y}_1^k + \mu_k(\mathbf{Z}^{k+1} - \mathbf{J}^{k+1}),\quad \mathbf{Y}_2^{k+1} = \mathbf{Y}_2^k + \mu_k(\mathbf{Z}^{k+1} - \mathbf{C}^{k+1}),\ \mu_{k+1} = \min(\mu_k\rho, \mu_{\max})$

(5) $k = k+1$

(6) Until the maximum number of iterations is reached or the following convergence conditions are satisfied:

$$\max_{i,j}\left|\mathbf{Z}_{k+1} - \mathbf{Z}_k\right| < tol \ \ and \ \max_{i,j}\left|\mathbf{J}_{k+1} - \mathbf{J}_k\right| < tol$$

$$and \ \max_{i,j}\left|\mathbf{C}_{k+1} - \mathbf{C}_k\right| < tol$$

End while

Output: LRR matrix $\mathbf{Z}$.

Algorithm S2. Normalized Spectral Clustering

Input: Affinity matrix $\mathbf{H}$, number of clusters k.

(1) Construct the normalized Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ of the matrix $\mathbf{H}$.

(2) Compute the top k eigenvalues $\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_k$ of the matrix $\mathbf{L}$, and such that the matrix $\mathbf{U} \in \mathbb{R}^{n\times k} = \{\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_k\}$.

(3) Construct the matrix $\mathbf{T} \in \mathbb{R}^{n\times k}$ from $\mathbf{U}$ by normalizing the rows to norm 1, that is $t_{ij} = u_{ij} / (\sum_k u_{ik}^2)^{1/2}$.

(4) For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the $i-th$ row of $\mathbf{T}$.

(5) Cluster the points $(y_i)_{i=1,...,n}$ with the K-means algorithm into clusters $B_1, ..., B_k$.

Output: Clusters $O_1, ..., O_k$ with $O_i = \{j \mid y_i \in B_i\}$.

C. Details of Comparison Methods

In the comparative analysis, we considered twelve comparison methods: t-SNE+K-means, SC, SC3, SIMLR, Corr, MPSSC, Seurat, SinNLRR, SCCLRR, Spectrum, SHARP and CDC. For the input data of t-SNE and SC (constructing the input affinity matrix using the Pearson correlation coefficient) methods, we perform gene filtering and normalization. To ensure fairness of the comparison, for the processing of the input data for the comparison methods SC3, SIMLR, Corr, MPSSC, Seurat, SinNLRR, SCCLRR, Spectrum, SHARP and CDC, we followed the instructions for each method. Table S2 shows the download links and data processing instructions for each method. For each baseline parameter setting, we use either the recommended parameters or the derived formulas for the given parameters to assign values. Table S3 shows the detailed parameter descriptions that need to be set by the user for each method.

TABLE S2

Download links and preprocessing methods for the comparison methods

Methods	Download Links	Pretreatment
t-SNE	https://github.com/qiao1002/plrls222/tree/main/tSNE	Gene filtering Normalization
SC	https://github.com/qiao1002/plrls222/tree/main	Gene filtering Normalization
SC3	https://github.com/hemberg-lab/SC3	Log transformation
SIMLR	https://github.com/BatzoglouLabSU/SIMLR	Log transformation
Corr	http://sysbio.sibcb.ac.cn/	-
MPSSC	https://github.com/ishspsy/project/tree/master/MPSSC	-
Seurat	https://github.com/satijalab/seurat	Normalization Find Variable Features Scaling PCA
SinNLRR	https://github.com/zrq0123/SinNLRR	Gene filtering Normalization
SCCLRR	https://github.com/wzhangwhu/SCCLRR	Scaling
Spectrum	https://cran.rproject.org/web/packages/Spectrum/index.html	log2-normalised counts and selected the top 100 most variable genes for analysis
SHARP	https:/ /github.com/shibiaowan/SHARP	-
CDC	https://github.com/ZPGuiGroupWhu/ClusteringDirectionCentrality	Normalization Find Variable Features Scaling PCA

TABLE S3

Detailed description of the parameters of the comparison methods

Methods	Parameter Settings
t-SNE	The number of clusters: k= real cluster number.
SC	The number of clusters: k= real cluster number.
SC3	The number of clusters: k= real cluster number.
SIMLR	Nonnegative tuning parameter: β=0.8 and γ=-1; Kernel function parameters: k = 10,12,14，…; 30, σ = 1.0,1.25,1.5,1.75,2; The number of neighbors k: 10~30; The number of clusters C= real cluster number; The dimension B embedded in the similar points learned by SIMLR is equal to C by default; The smoothing parameters in similarities diffusion τ, default τ= 0.8; The moment parameters in similarity updatingα: default α=0.8.

Corr	The algorithm can automatically determine the optimal number of clusters without setting additional parameters.
MPSSC	Regularization parameter λ ∈ {0.01,0.001, ⋯, 0.00001}, default λ= 0.0001; ρ ∈ [0.01,4], default ρ= 0.2; μ ∈ {0.01,0.001, ⋯, 0.00001}, default μ= 0.0001; The parameters in kernel functions σ ∈ {1,1.25, ⋯, 2} and k ∈ {10,12, ⋯, 30}; The number of clusters C= real cluster number.
Seurat	The dimension parameter n_components ∈ [2,5] of the UMAP space to be embedded, default n_components=2; The adjacent point parameter n_neighbors ∈ [5,50] of UMAP, default n_ neighbors=30; The parameter min_dist ∈ [0.1,1], which controls the degree to which UMAP embedding allows points to be compressed together, is the default value of min_ dist=0.3; Set the number of clusters k according to the community based detection technology.
SinNLRR	Regularization parameter λ ∈ [2,5] and γ ∈[0.5,5]; automatically select parameter λ with NoMN, default γ= 1; The number of clusters k= real cluster number.
SCCLRR	λ1, λ2 ∈ [5e-4, 1e-3, 3e-3, 5e-3, 8e-3, 1e-2, 3e-2, 5e-2, 7e-2, 1e-1, 5e-1], α ∈（0,，1）， λ1=1e-2; λ2=5e-3; α=0.4; The number of clusters c= real cluster number.
Spectrum	Set the number of clusters k according to the multimodality gap heuristic algorithm.
SHARP	Set the number of clusters k according to the Silhouette, Calinski-Harabasz index and hierarchical heights of the clustering.

CDC	seurat_dim: dimension of reduction to use as input in Seurat (Default: seq(20, 50, 5));
	seurat_resolution: resolution of Louvain algorithm in Seurat (Default: seq(0.1, 1, 0.1));
	snnwalk_k: k of SNN graph in SNN-Walktrap (Default: seq(5, 30, 5));
	snnlouv_k: k of SNN graph in SNN-Louvain (Default: seq(5, 30, 5));
	kmeans_k: k of K-means (Default: seq(2, 50));
	CDC_k: k of KNN in CDC (Default: seq(30, 50, 10));
	CDC_ratio: ratio of CDC (Default: seq(0.85, 0.99, 0.02)).

D. Evaluation Measurements

The clustering performance of the proposed PLRLS method is assessed using two widely used assessment criteria: Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). The bigger the value, the more accurate the forecast, and NMI belongs to [0,1] and ARI belongs to [-1,1] respectively.

For the true label $\mathbf{T} = \{T_1, T_2, ..., T_k\}$ and the predicted label $\mathbf{P} = \{P_1, P_2, ..., P_k\}$, the formulas of NMI and ARI are as follows:

$$\mathrm{NMI}(\mathbf{T},\mathbf{P})=\frac{\mathrm{M}(\mathbf{T},\mathbf{P})}{[\mathrm{E}(\mathbf{T})+\mathrm{E}(\mathbf{P})]/2}, \tag{1}$$

$$\mathrm{ARI}(\mathbf{T},\mathbf{P})= \frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_{ij}\binom{n_{ij}}{2}\sum_{ij}\binom{n_{ij}}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_i}{2}\right]-\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_i}{2}\right]\Big/\binom{n}{2}}. \tag{2}$$

In (1), $\mathrm{E}(\cdots)$ is the correlation entropy function and $\mathrm{M}(\mathbf{T},\mathbf{P})$ is the mutual information of $\mathbf{T}$ and $\mathbf{P}$. In (2), n_{ij} is the number of cells in T_i and P_i, a_i means the number of cells in T_i, b_i represents the number of cells in P_i, $\binom{n}{2}=n(n-1)/2$.

E. Prediction of Clustering Number

In clustering, the number of clusters needs to be determined in advance. To solve this problem, we use the eigengap method [1] to confirm the number of clusters in the scRNA-seq dataset. This method has been widely used to determine the number of clusters in single-cell clustering analysis [2-7]. First, we use a large enough number k' as a target number. Then, we determine the number of clusters k by maximizing the eigenvalues gap $|\lambda_k-\lambda_{k-1}|$, where $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n\geq 0$ is the eigenvalues of the affinity matrix $\mathbf{H}$. Furthermore, we choose SIMLR, MPSSC and SinNLRR methods as competing methods, which also focus on similarity learning. For these comparative methods, we execute the corresponding estimation algorithms provided by ourselves. Table S4 summarizes the prediction results of twelve real datasets, in which the prediction results consistent with the known number of clusters are highlighted in bold. It can be seen from Table S4 that the proposed method PLRLS obtains the same numbers as the pre-annotated numbers in five datasets (Ramskold, Li_islet, Ting, Engel4, and Kolod) and is closest to the pre-annotated numbers in most datasets (Treutlein, Goolam, Deng, Pollen, and Darmnis). SIMLR correctly estimated the number of clusters in Ting and Deng datasets, MPSSC correctly estimated the number of clusters in Li_islet datasets, and SinNLRR correctly estimated the number of clusters in Pollen and Engel4 datasets. It can be seen that no method can accurately estimate the number of clusters of all datasets. On the whole, the proposed framework has obvious advantages.

TABLE S4

Estimated number of clusters on twelve scRNA-seq datasets

Dataset	Known	SIMLR	MPSSC	SinNLRR	PLRLS
Ramskold	7	3	3	2	**7**
Li_islet	6	2	**6**	5	**6**
Treutlein	5	15	10	3	3
Ting	5	**5**	7	4	**5**
Goolam	5	11	15	4	3
Deng	7	**7**	8	3	6

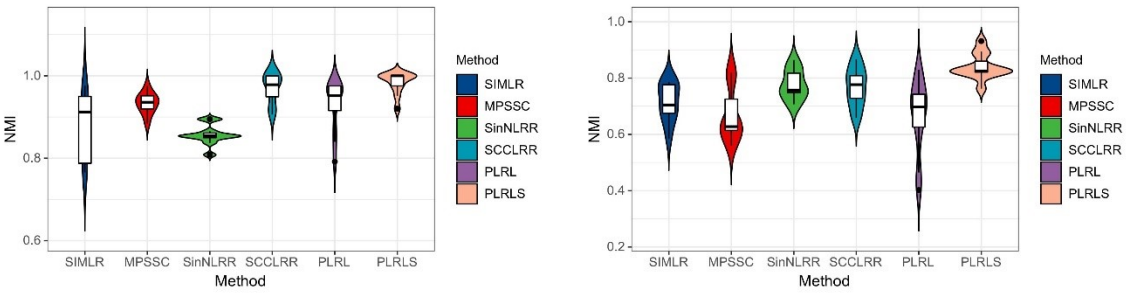
Engel4	4	3	2	**4**	**4**
Pollen	11	19	21	**11**	9
Darmanis	8	18	13	9	6
Kolod	3	4	4	5	**3**
Tasic	48	7	7	22	8
Zeisel	48	10	3	11	9

F. Robustness Tests

We will talk about how robust the proposed PLRLS approach is to dropout and noise in this part. It is because dropout and noise are intrinsic properties of scRNA-seq data. Dropout event refers to the phenomenon that the gene expression value in a cell is false or close to zero, usually due to the low total RNA content in a single cell and the technical limitations, which presents computational challenge. Noise in scRNA-seq data may come from experimental design, human errors, or biological factors.

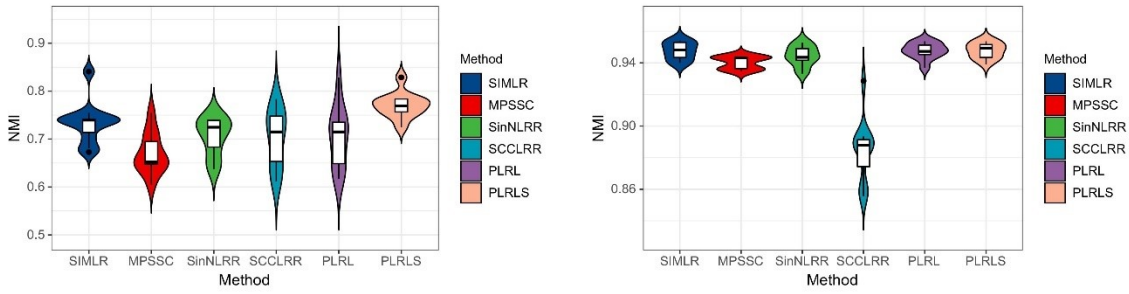
In the robustness testing experiments, on the Treutlein, Ting, Deng and Pollen datasets, the comparison methods used are SIMLR, MPSSC, SinNLRR, SCCLRR and PLRL. The PLRL is included here to study whether the introduction of the similarity constraint contributes to the robustness of the model.

In the dropout robustness test, each original scRNA-seq dataset is considered as a population, and 5%-50% of the data are randomly removed. Fig. S1 shows the test results under different dropout rates. In the noise robustness test, we add an independent zero-mean Gaussian noise with various variances $x_{ij} = x_{ij} + N(0, \sigma_{ij}), 0 \le \sigma_{ij} \le 2$ to the scRNA-seq data's raw expression matrix. Fig. S2 shows the results of the robustness tests for all methods in terms of noise. By analyzing Fig. S1 and Fig. S2, the following conclusions can be drawn: (i) the introduction of similarity constraint based on fractional function can enhance the robustness of the model to dropout events and noise; (ii) the PLRLS method is more robust compared to other clustering methods based on similarity learning.



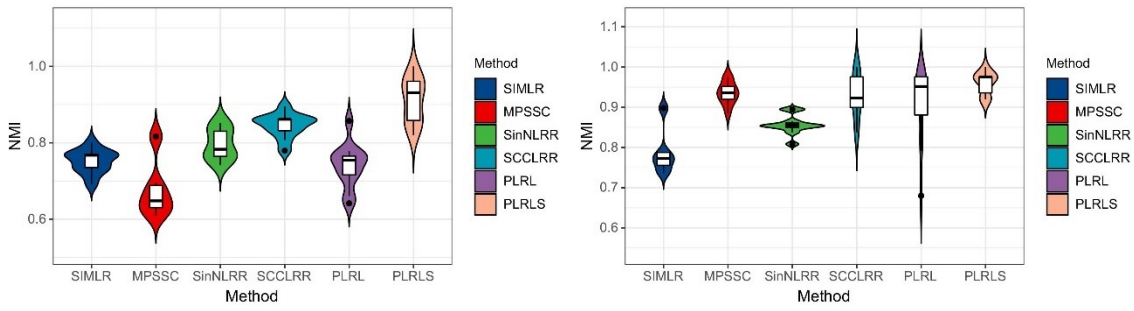
(a) Treutlein

(b) Ting

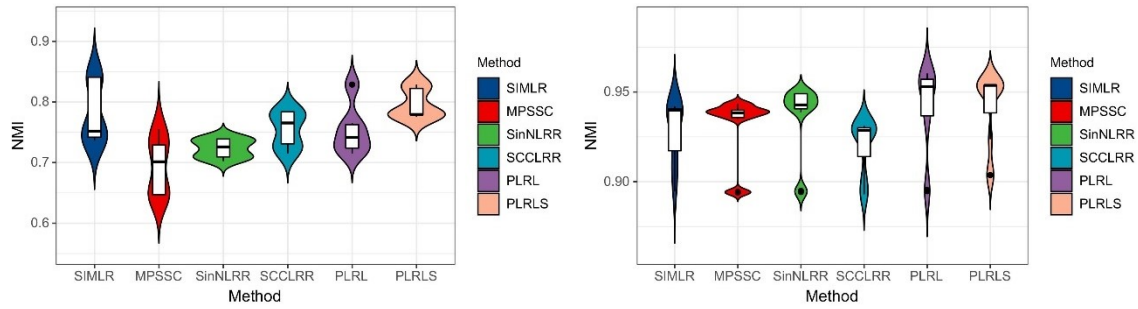


(c) Deng (d) Pollen

Fig. S1. Comparison of the robustness test results under different dropout rates.



(a) Treutlein (b) Ting



(c) Deng (d) Pollen

Fig. S2. Comparison of robustness test results in the presence of noise involvement.

G. Runtime Comparisons

The proposed method PLRLS is run on a PC with Core i7-10700K CPU @ 3.80GHz and 96.0G RAM. We test the running time of the algorithm developed by MATLAB on ten scRNA-seq datasets with different cell counts. To ensure the accuracy of the results, we take the average of 100 runs as the final run time. Table S5 shows the actual computation times. We found that the running time of most algorithms increases with the number of samples. SinNLRR, SCCLRR and the proposed method PLRLS are single-cell clustering methods based on the LRR model. Due to the introduction of the local structure constraint and the similarity constraint, PLRLS takes a little more time compared to SinNLRR and SCCLRR. However, it is found from the clustering experiments that our method can obtain more accurate clustering results. In

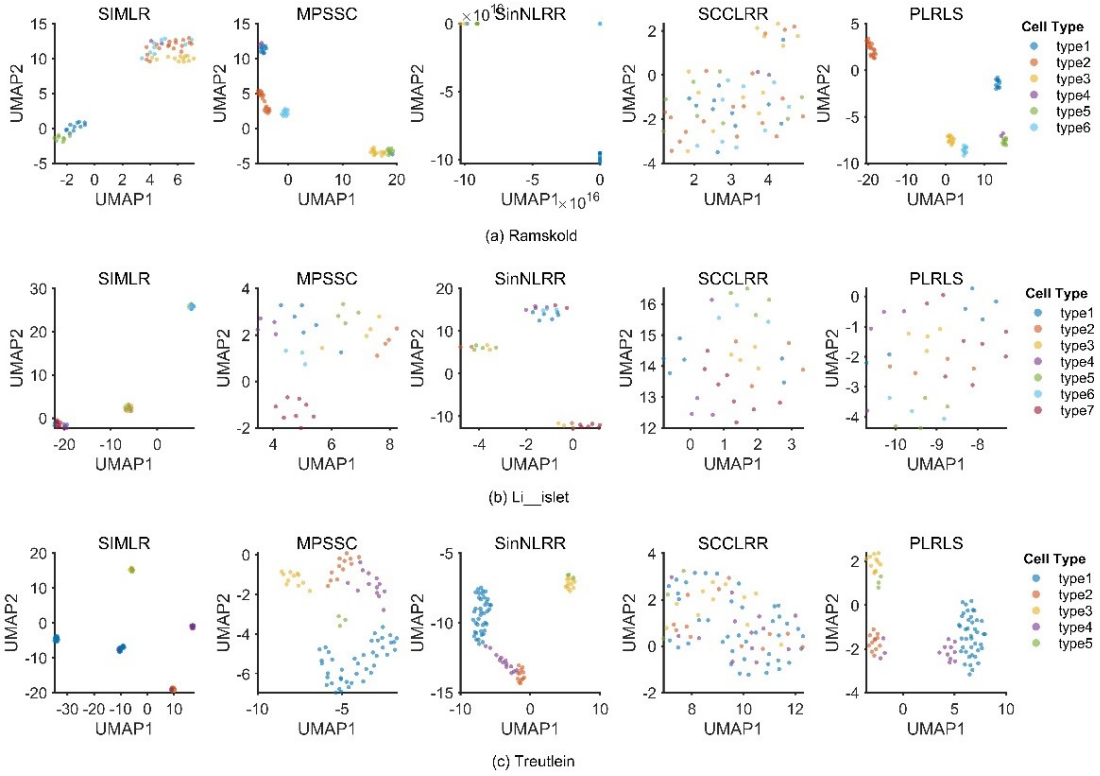
particular, the running time of the proposed method PLRLS is similar to that of the SCCLRR method, but the clustering accuracy of our method has a significant advantage.

TABLE S5

The running times (in seconds) of Corr, MPSSC, SinNLRR, SCCLRR and PLRLS on ten scRNA-seq datasets.

Methods	Corr	MPSSC	SinNLRR	SCCLRR	PLRLS
Ramskold	12.72	0.23	0.23	0.17	0.11
Li_islet	16.63	0.49	0.33	0.52	0.59
Treutlein	4.06	0.48	0.31	0.27	0.28
Ting	61.16	0.52	0.38	2.71	2.86
Goolam	194.03	1.20	0.35	7.67	8.15
Deng	56.49	0.95	0.38	3.13	3.90
Engel4	198.02	1.69	0.84	9.50	9.87
Pollen	250.60	6.29	0.83	9.05	9.45
Darmnis	286.25	6.45	18.98	33.55	38.05
Kolod	290.74	28.16	13.44	48.09	57.09

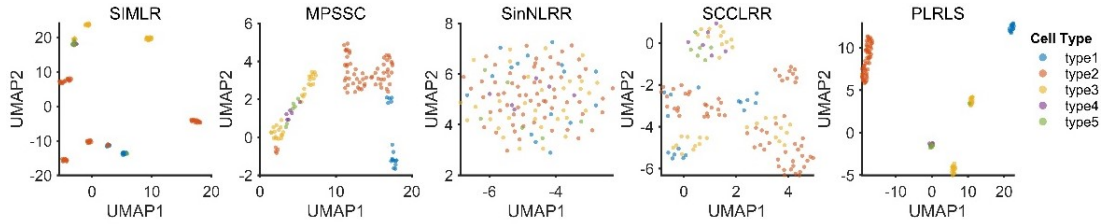
Visualization



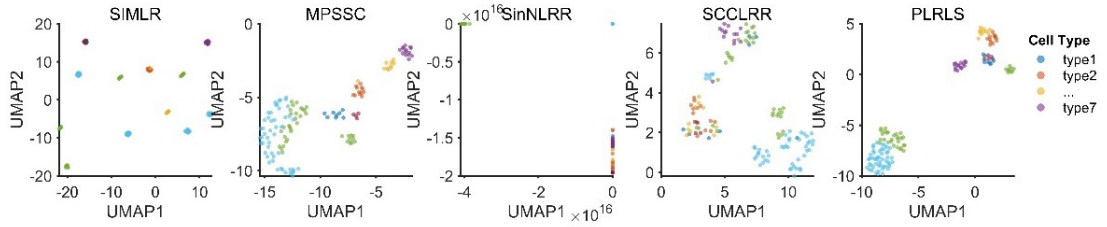
(a) Ramskold

(b) Li__islet

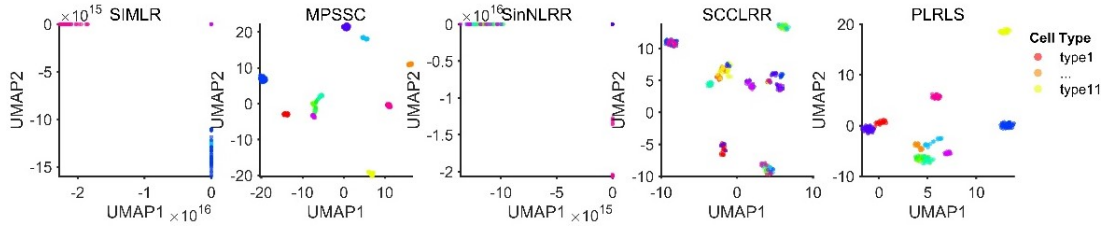
(c) Treutlein



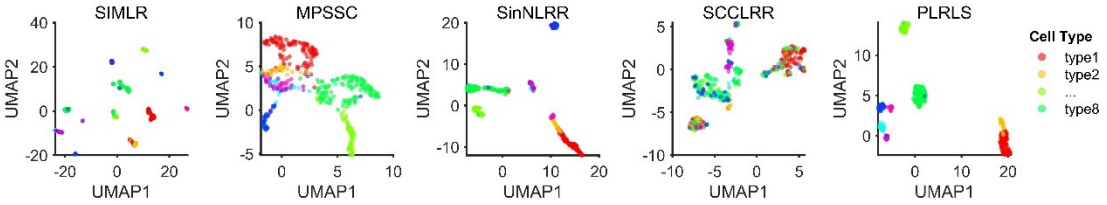
(d) Goolam



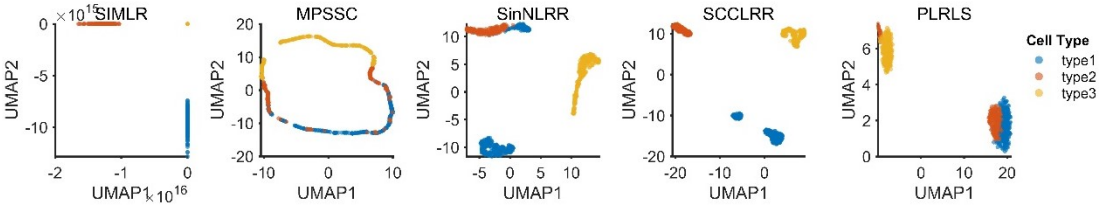
(e) Deng



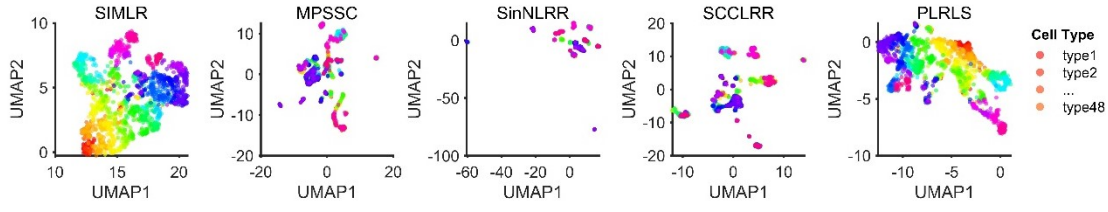
(f) Pollen



(g) Darmanis



(h) Kolod



(i) Tasic

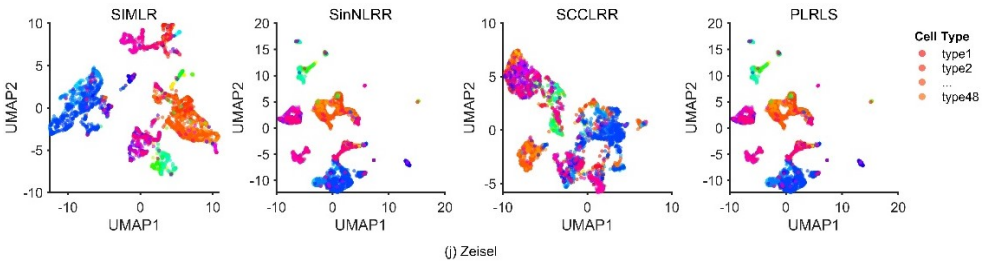


Fig. S3. UMAP visualization comparison of affinity matrices obtained from ten different scRNA-seq datasets with the chosen comparison methods SIMLR, MPSSC, SinNLRR and SCCLRR. And different colors indicate different cell subpopulations.

I. References

[1] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing,* vol. 17, no. 4, pp. 395-416, 2007.

[2] D.-J. Zhang, Y.-L. Gao, J.-X. Zhao, C.-H. Zheng, and J.-X. Liu, "A New Graph Autoencoder-Based Consensus-Guided Model for scRNA-seq Cell Type Detection," *IEEE Transactions on Neural Networks and Learning Systems,* 2022.

[3] H. Li, C. R. Brouwer, and W. Luo, "A universal deep neural network for in-depth cleaning of single-cell RNA-Seq data," *Nature Communications,* vol. 13, no. 1, pp. 1-11, 2022.

[4] W. Zhang, X. Xue, X. Zheng, and Z. Fan, "NMFLRR: Clustering scRNA-seq data by integrating nonnegative matrix factorization with low rank representation," *IEEE Journal of Biomedical and Health Informatics,* vol. 26, no. 3, pp. 1394-1405, 2021.

[5] Y. Cui, S. Zhang, Y. Liang, X. Wang, T. N. Ferraro, and Y. Chen, "Consensus clustering of single-cell RNA-seq data by enhancing network affinity," *Briefings in bioinformatics,* vol. 22, no. 6, p. bbab236, 2021.

[6] W. Zhang, Y. Li, and X. Zou, "SCCLRR: a robust computational method for accurate clustering single cell RNA-seq data," *IEEE journal of biomedical and health informatics,* vol. 25, no. 1, pp. 247-256, 2020.

[7] R. Zheng, M. Li, Z. Liang, F.-X. Wu, Y. Pan, and J. Wang, "SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation," *Bioinformatics,* vol. 35, no. 19, pp. 3642-3650, 2019.