Supplementary Information

Qinglin Mei, Guojun Li and Zhengchang Su

Supplementary Figures:

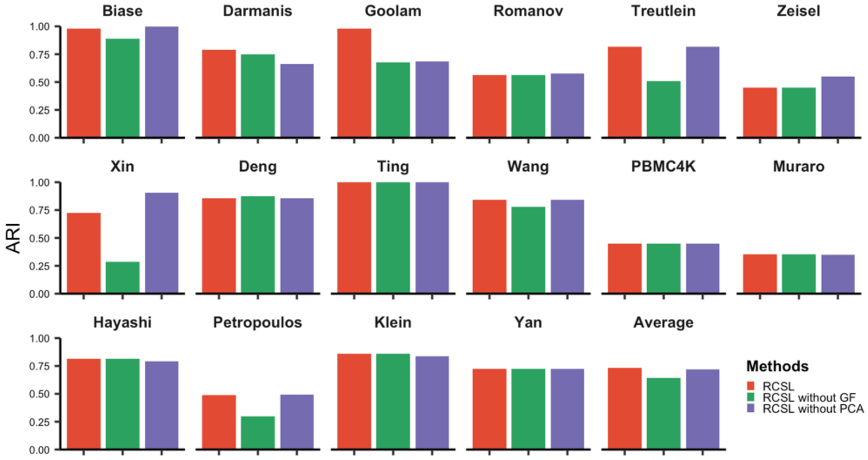

Fig. S1. Effects of the Gene Filter (GF) and PCA step on the accuracy (ARI) of RCSL in the 16 datasets. The ARI values were computed according to the annotated cell types.

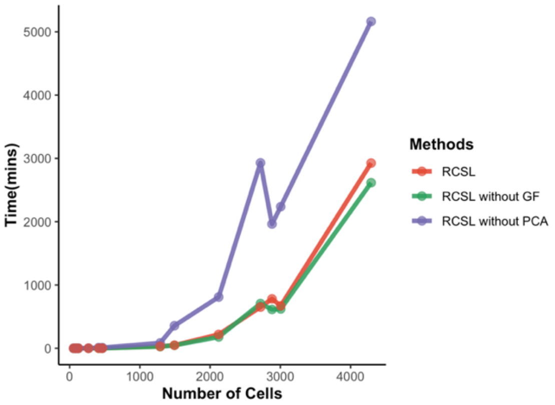

Fig. S2. Comparison of running time of RCSL, RCSL without Gene Filter (GF) and RCSL without PCA.

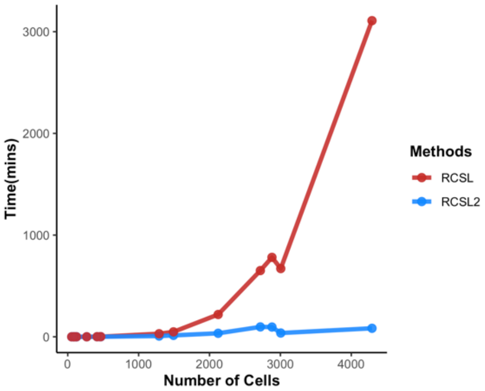

Fig. S3. The running time of RCSL and RCSL2 on the 16 scRNA-seq datasets.

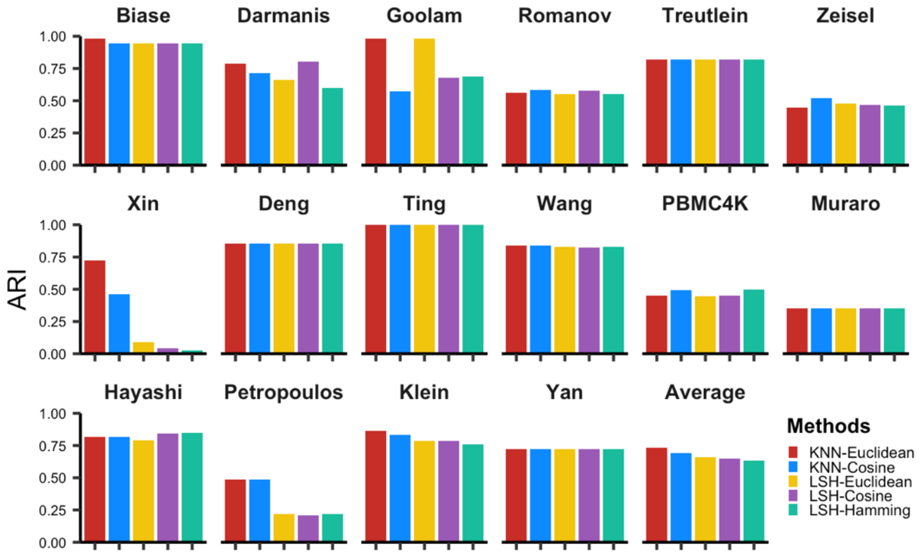

Fig. S4. The comparison of the accuracy of RCSL using KNN and RCSL using LSH for finding *k*-NNs using Euclidean distance, cosine angle distance and Hamming distance.

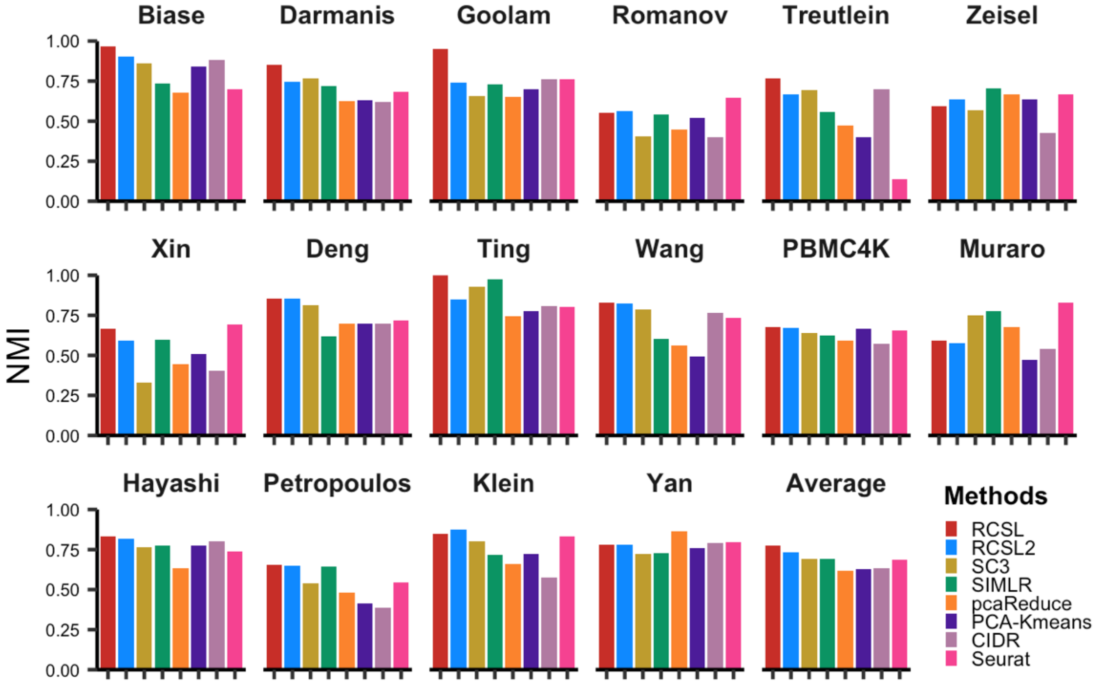

Fig. S5. Performance of the algorithms (RCSL, RCSL2, SC3, SIMLR, pcaReduce, k-means, CIDR, Seurat) on the 16 datasets measured by Normalized Mutual Information (NMI). The last panel shows the average NMI value for each algorithm over the 16 datasets.

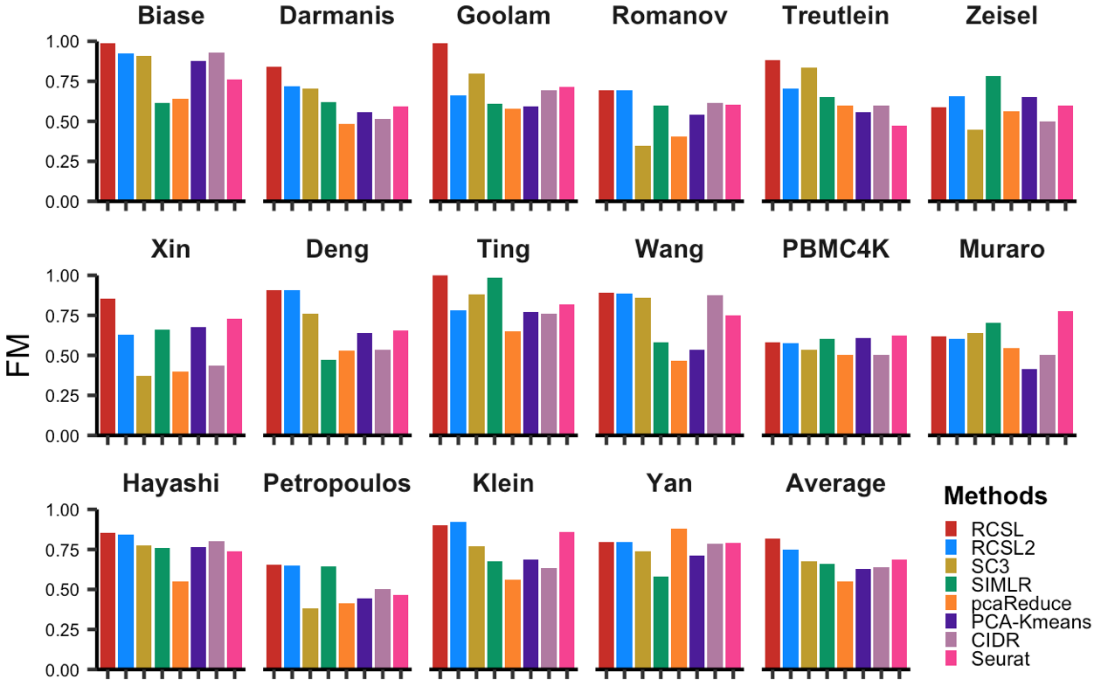

Fig. S6. Performance of the algorithms (RCSL, RCSL2, SC3, SIMLR, pcaReduce, k-means, CIDR, Seurat) on the 16 datasets measured by Fowlkes-Mallows index (FM). The last panel shows the average FM value for each algorithm over the 16 datasets.

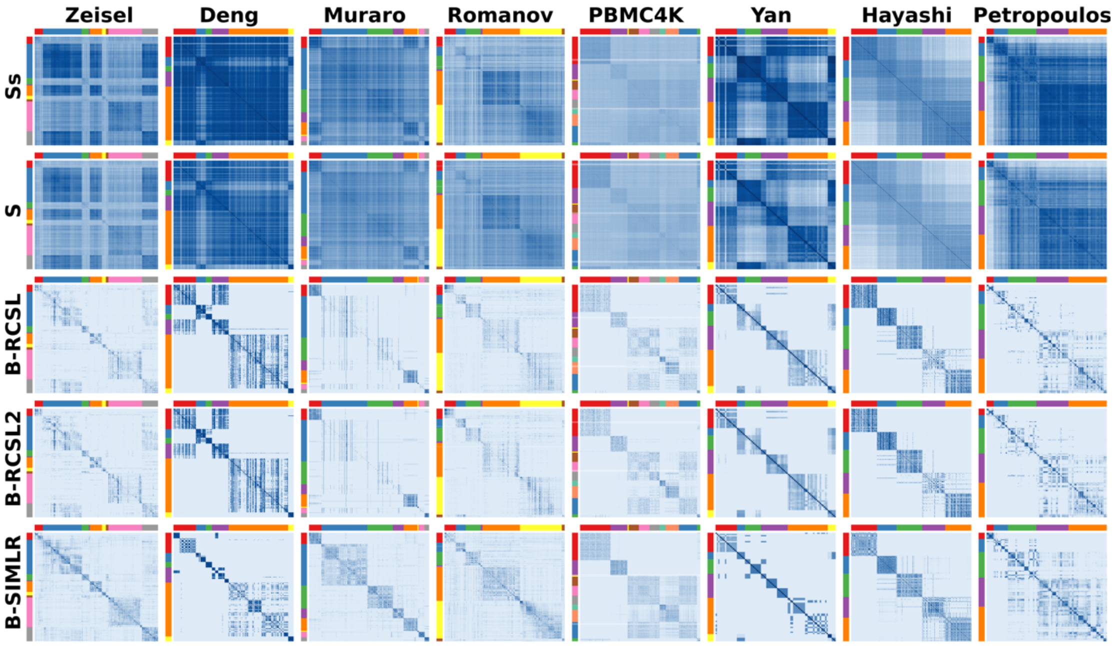

Fig. S7. Heatmap of Spearman's rank correlation matrix $\boldsymbol{S}_s$ and similarity matrix $\boldsymbol{S}$ in RCSL as well as the block-diagonal similarity matrices $\boldsymbol{B}$ learned by RCSL, RCSL2, SIMLR in the indicated eight datasets. Cells are arranged according to their annotated types indicated by differently colored bars at the top and left of the matrices.

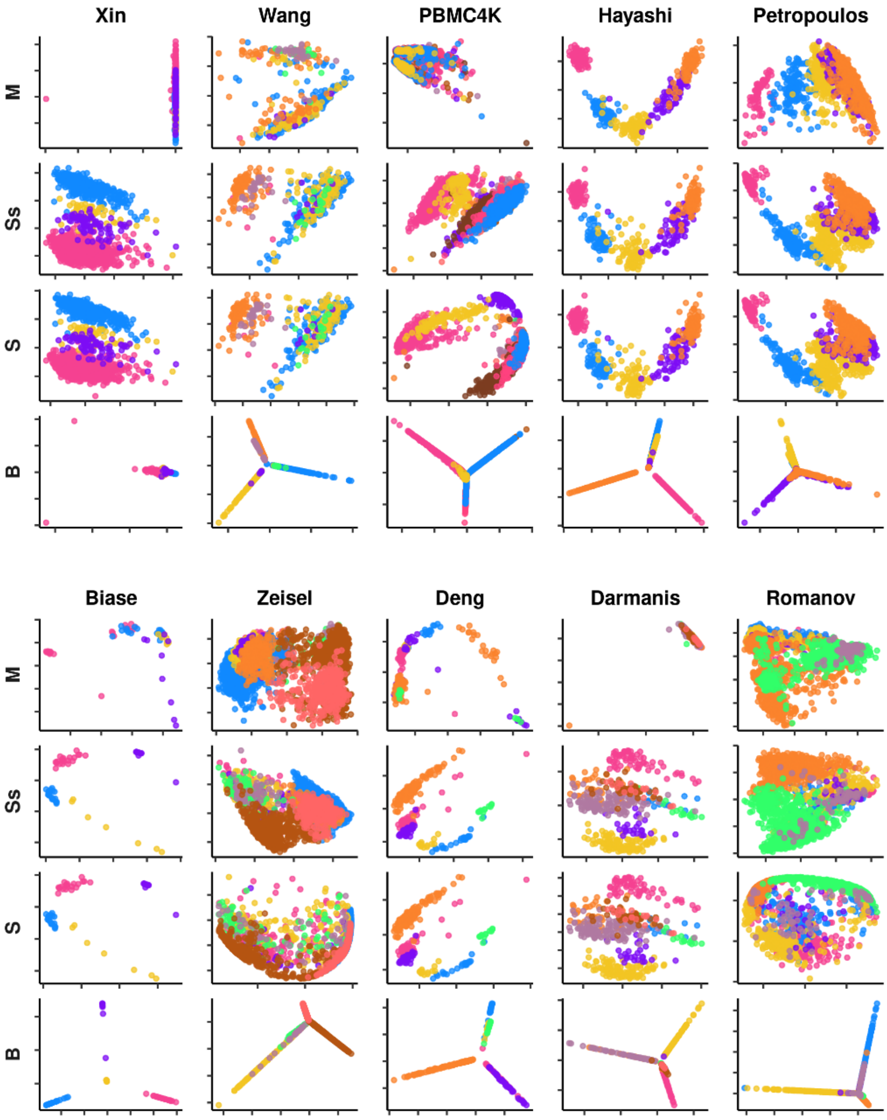

Fig. S8. 2-D PCA displays of the expression data matrices and matrices produced by RCSL in the 10 datasets.

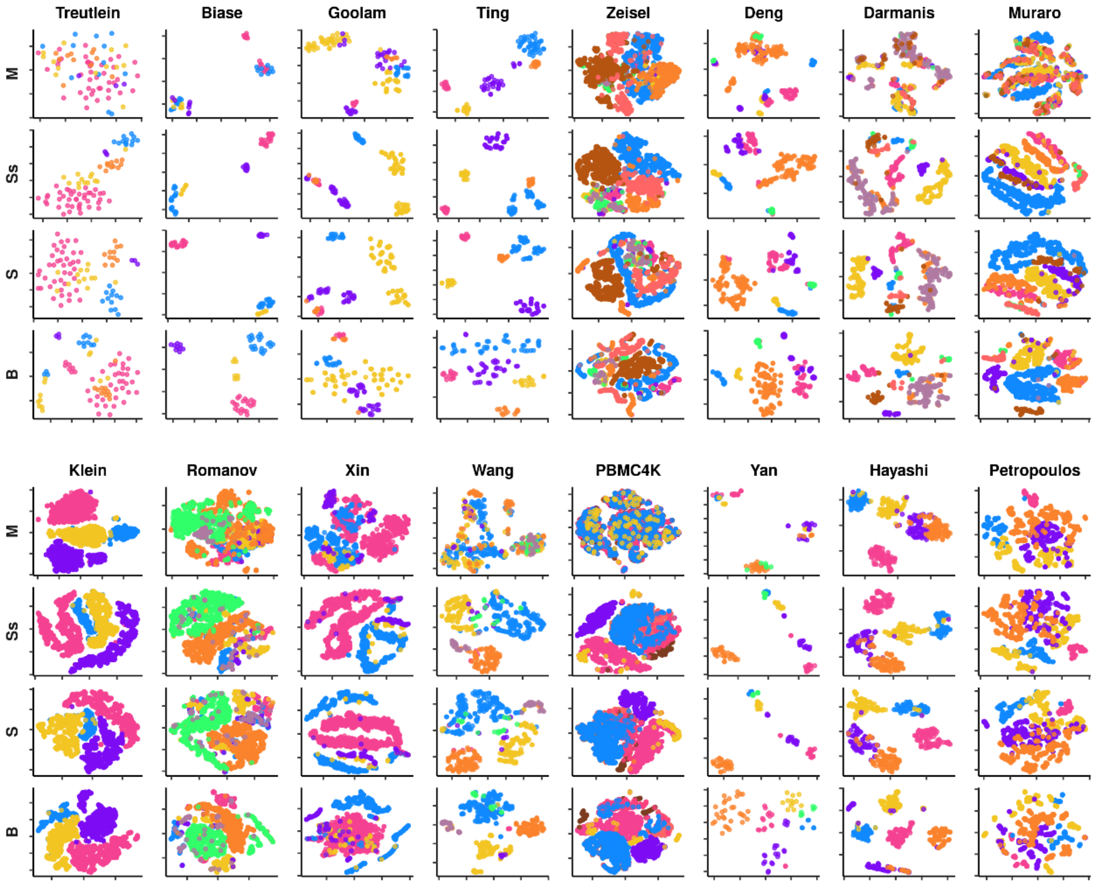

Fig. S9. 2-D t-SNE displays of the expression data matrices and matrices produced by RCSL in the 16 datasets.

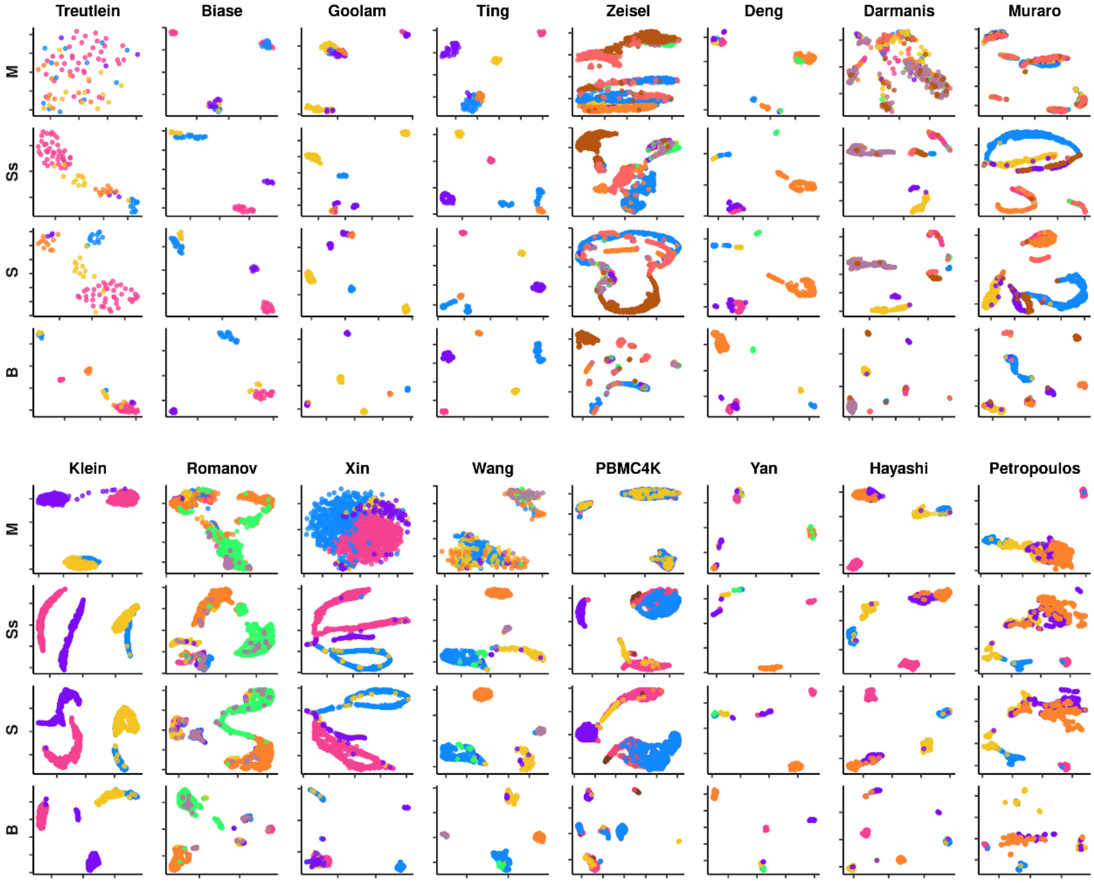

Fig. S10. 2-D UMAP displays of the expression data matrices and matrices produced by RCSL in the 16 datasets.

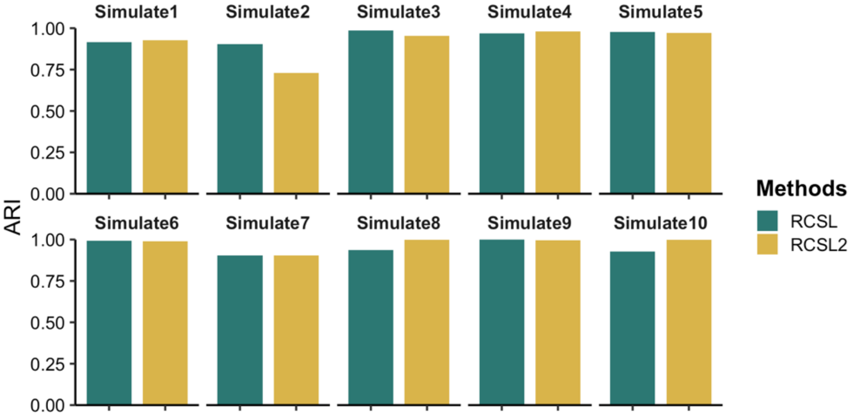

Fig. S11. Performance of RCSL and RCSL2 on 10 simulated datasets measured by adjusted rand index (ARI).

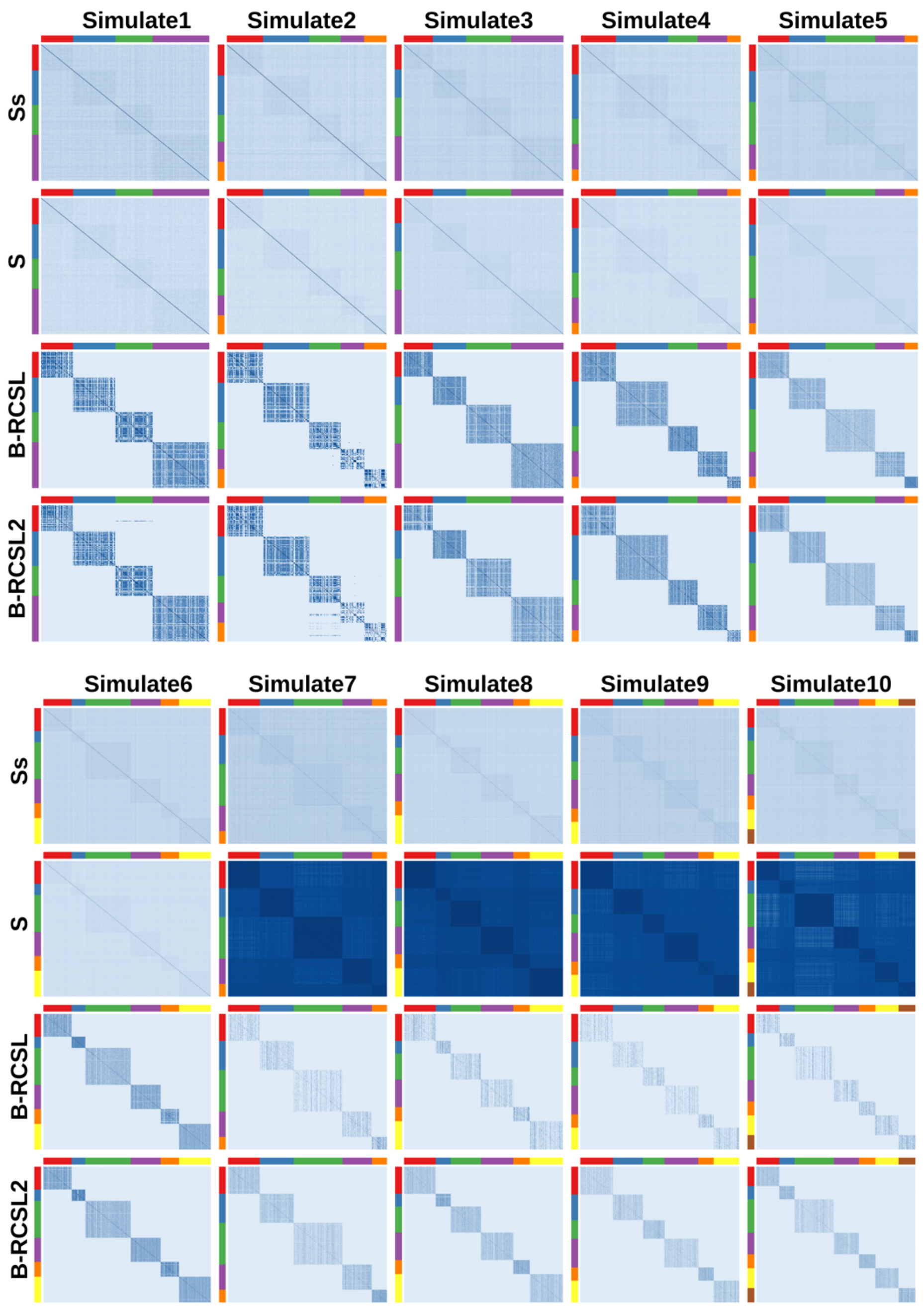

Fig. S12. Heatmap of Spearman's rank correlation matrix $\boldsymbol{S}_s$ and similarity matrix $\boldsymbol{S}$ in RCSL as well as the block-diagonal similarity matrices $\boldsymbol{B}$ learned by RCSL and RCSL2 in the 10 simulated datasets. Cells are arranged according to their types indicated by differently colored bars at the top and left of the matrices.

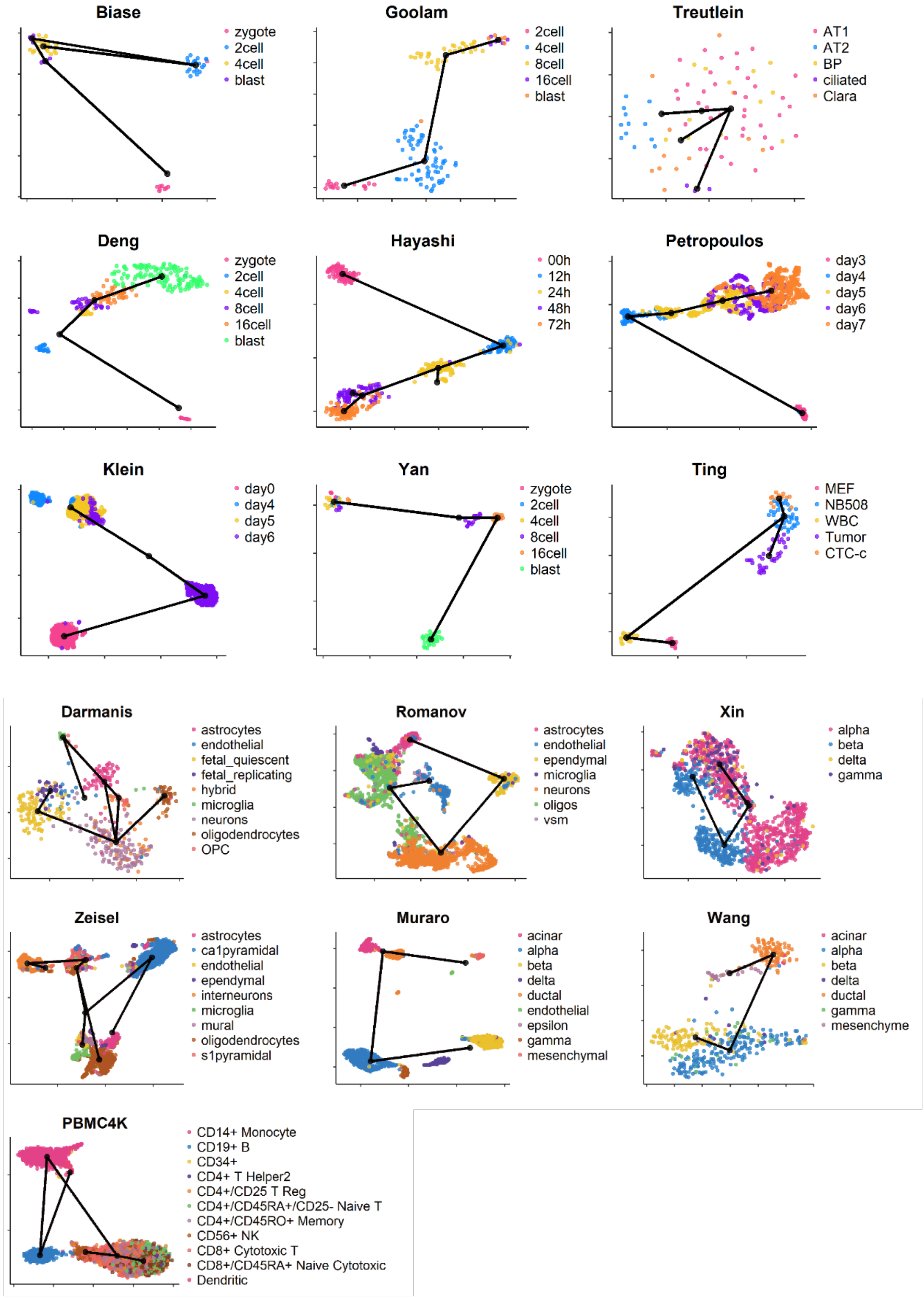

Fig. S13. UMAP visualization of constructed MSTs (Minimum Spanning Trees) based on the clustering results of RCSL.

Supplementary Tables:

Table S1: Extended summary of the 16 scRNA-seq datasets used in this study.

Datasets	Accession ID	Species	# Cells	# Classes	Protocol	UMI	Cell types										
Treutlein	GSE52583	Mouse	80	5	SMARTer	No	AT1 (41)	AT2 (12)	BP (13)	ciliated (3)	Clara (11)						
Biase	GSE57249	Mouse	56	4	SMARTer	No	Zygote (7)	2-cell (20)	4-cell (20)	Blast (9)							
				5			Zygote (7)	2-cell (20)	4-cell (20)	ICM (4)	TE (3)						
Goolam	E-MTAB-3321	Mouse	124	5	Smart-Seq2	No	2-cell (16)	4-cell (64)	8-cell (32)	16-cell (6)	Blast (6)						
Ting	GSE51372	Mouse	114	5	Tang	No	16-cell (6)	2-cell (16)	4-celll (64)	8-cell (32)	Blast (6)						
Zeisel	GSE60361	Mouse	3005	9	Smart-Seq STRT-Seq UMI	Yes	astrocytes (198)	ca1pyramidal (948)	endothelial (175)	ependymal (26)	interneurons (290)	microglia (98)	mural (60)	oligodendrocytes (820)	slpyramidal (390)		
Deng	GSE45719	Mouse	268	6	Smart-Seq2	No	Zygote (12)		2-cell (22)		4-cell (14)	8-cell (37)	16-cell (50)	Blast (133)			
				10	Drop-seq	No	Zygote (4)	early 2-cell (8)	Mid 2-cell (12)	Late 2-cell (10)	4-cell (14)	8-cell (37)	16-cell (50)	Early blast (43)	Mid blast (60)	Late blast (30)	
Darmanis	GSE67835	Human	466	9	SMARTer	No	astrocytes (62)	endothelial (20)	fetal_quiescent (110)	fetal_replicating (25)	hybrid (46)	microglia (16)	neurons (131)	oligodendrocytes (38)	OPC (18)		
Muraro	GSE85241	Human	2122	9	CEL-Seq2	No	acinar (219)	alpha (812)	beta (448)	delta (193)	ductal (245)	endothelial (21)	epsilon (3)	gmma (101)	mesenchymal (80)		
Klein	GSE65525	Mouse	2717	4	inDrop	No	d0 (933)	d2 (303)	d4 (683)	d7 (798)							
Romanov	GSE74672	Mouse	2881	7			astrocytes (267)	endothelial (240)	ependymal (356)	microglia (48)	neurons (898)	oligos (1001)	vsm (71)				
Xin	GSE81608	Human	1492	4	SMARTer	No	alpha (886)	beta (472)	delta (49)	gamma (85)							
Wang	GSE83139	Human	457	7	SMARTer	No	acinar (6)	alpha (190)	beta (111)	delta (9)	ductal (96)	gamma (18)	mesenchyme (27)				
PBMC4K	SRP073767	Human	4292	11	10xGenomics Chromium	Yes	CD14+ Monocyte (1083)	CD19+ B (606)	CD34+ (11)	CD4+ T Helper2 (36)	CD4+/CD25 T Reg (363)	CD4+/CD45RA+/CD25- Naive T (386)	Dendritic (120)	CD4+/CD45RO+ Memory (353)	CD56+ NK (220)	CD8+ Cytotoxic T (473)	CD8+/CD45RA+ Naive Cytotoxic (641)
Yan	GSE36552	Human	90	6	Tang	No	Zygote (6)		2-cell (6)	4-celll (12)	8-cell (20)	16-cell (16)	Blast (30)				
							Oocyte (3)	Zygote (3)	2-cell (6)	4-celll (12)	8-cell (20)	morula (16)	Late blast (30)				
Hayashi	GSE98664	Mouse	414	5	RamDA-seq	No	00h (89)	12h (67)	24h (89)	48h (79)	72h (90)						
Petropoulos	E-MTAB-3929	Human	1289	5	Smart-Seq2	No	Embryonic Day3 (75)	Embryonic Day4 (154)	Embryonic Day5 (304)	Embryonic Day6 (345)	Embryonic Day7 (411)						

Table S2: Cells types and subtypes annotated in the Biase, Deng and Yan datasets.

Biase	**Cell type**	Zygote (7)	2-cell (20)	4-cell (20)	Blast (9)	
	Sub-type	Zygote (7)	2-cell (20)	4-cell (20)	ICM (4)	TE (3)

Deng	**Cell type**	Zygote (12)		2-cell (22)		4-cell (14)	8-cell (37)	16-cell (50)	Blast (133)		
	Sub-type	Zygote (4)	Early 2-cell (8)	Mid 2-cell (12)	Late 2-cell (10)	4-cell (14)	8-cell (37)	16-cell (50)	Early blast (43)	Mid blast (60)	Late blast (30)

Yan	**Cell type**	Zygote (6)		2-cell (6)	4-cell (12)	8-cell (20)	16-cell (16)	Blast (30)
	Sub-type	Oocyte (3)	Zygote (3)	2-cell (6)	4-cell (12)	8-cell (20)	Morula (16)	Late blast (30)

Table S3. Summary of the 10 simulated datasets and the ARI value of RCSL and RCSL2.

# Cells		**300**		**500**		**1000**		**2000**		**3000**		
# Cell types		**4**	**5**	**4**	**5**	**5**	**6**	**5**	**6**	**6**	**7**	
Datasets		**Simulate1**	**Simulate2**	**Simulate3**	**Simulate4**	**Simulate5**	**Simulate6**	**Simulate7**	**Simulate8**	**Simulate9**	**Simulate10**	**Average**
ARI	**RCSL**	0.9142	**0.9043**	**0.9858**	0. 9680	**0. 9785**	**0.9938**	**0.9048**	**0.9375**	1	**0.9268**	**0.9514**
	RCSL2	**0.9268**	0.7285	0.9524	**0. 9794**	0. 9725	0.9903	**0.9028**	**0.9987**	0.9951	**0.9976**	0.9444

Supplementary Note:

Procedure for constructing the block-diagonal matrix B

Note that Eq. (4) is difficult to solve since $\boldsymbol{L}_B = \boldsymbol{D}_B - \frac{\boldsymbol{B}^T + \boldsymbol{B}}{2}$, and $\boldsymbol{D}_B$ also depends on B, which leads to the constraint $rank(\boldsymbol{L}_B) = N - C$ becomes a complex nonlinear constraint. Let $\sigma_i(\boldsymbol{L}_B)$ be the i-th smallest eigenvalue of $\boldsymbol{L}_B$. Since $\boldsymbol{L}_B$ is positive semidefinite, $\sigma_i(\boldsymbol{L}_B) \geq 0$. Thus the constraint $rank(\boldsymbol{L}_B) = N - C$ in problem (4) can be satisfied if $\sum_{i=1}^{C} \sigma_i(\boldsymbol{L}_B) = 0$. Imposing a large enough value β, Eq. (4) can be transformed into an optimization problem,

$$\min_{\boldsymbol{B}} \| \boldsymbol{B} - \boldsymbol{S} \|_1 + 2\beta \sum_{i=1}^{C} \sigma_i(\boldsymbol{L}_B) \;\; s.t. \sum_{j=1}^{N} b_{ij} = 1, b_{ij} \geq 0. \tag{5}$$

Note that when β is large enough, the sum of the C smallest eigenvalues in $\boldsymbol{L}_B$ is forced to zero. Let $\boldsymbol{Y}_{N\times C}$ be the class indicator matrix, where $y_{il} = 1$ indicates that cell i is assigned to the cluster l. CLR finds $\boldsymbol{B}$ by solving the following constrained minimization problem,

$$\min_{\boldsymbol{B},\boldsymbol{Y}} \| \boldsymbol{B} - \boldsymbol{S} \|_1 + 2\beta Tr(\boldsymbol{Y}^T \boldsymbol{L}_B \boldsymbol{Y}) \;\; s.t. \sum_{j=1}^{N} b_{ij} = 1, b_{ij} \geq 0, \boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{I}. \tag{6}$$

To do so, we fix $\boldsymbol{B}$ and update $\boldsymbol{Y}$: when $\boldsymbol{B}$ is fixed, problem (6) becomes,

$$\min_{\boldsymbol{Y}} Tr(\boldsymbol{Y}^T \boldsymbol{L}_B \boldsymbol{Y}) \;\; s.t. \boldsymbol{Y} \in \boldsymbol{R}^{N\times C}, \boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{I}. \tag{7}$$

The optimal solution of $\boldsymbol{Y}$ is the C eigenvectors of $\boldsymbol{L}_B$ corresponding to the C smallest eigenvalues. CLR then fix $\boldsymbol{Y}$ and update $\boldsymbol{B}$: when $\boldsymbol{Y}$ is fixed, problem (7) can be transformed into,

$$\min_{b_{ij}} \sum_{i,j=1}^{N} | b_{ij} - s_{ij} | + \beta \sum_{i,j=1}^{N} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|_2^2 \, b_{ij} \;\; s.t. \sum_{j=1}^{N} b_{ij} = 1, b_{ij} \geq 0, \tag{8}$$

where $\boldsymbol{y}_i$ is the class indicator vector of cell i. Note that problem (8) is independent for different i, so we can parallel the calculations by solving each i independently, and rewrite (8) in the vector form for each i.

$$\min_{\boldsymbol{b}_i} \| \boldsymbol{b}_i - \boldsymbol{s}_i \|_1 + \beta \boldsymbol{b}_i^T \boldsymbol{f}_i \;\; s.t. \boldsymbol{b}_i^T \boldsymbol{1} = 1, \boldsymbol{b}_i \geq 0, \tag{9}$$

where $f_{ij} = \| y_i - y_j \|^2$, and $\boldsymbol{f}_i$ is a vector with the j-th element equal to f_{ij} (similarly to $\boldsymbol{b}_i$ and $\boldsymbol{s}_i$). Problem (9) can be solved by using an iterative reweighted method described in (Nie, et al., 2016), which grantees converge to the optimal solution. Let $\boldsymbol{U}$ be the $N\times N$ diagonal matrix whose diagonal entry $u_{ii} = \frac{1}{2 | \tilde{b}_{ij} - s_{ij} |}$, and the $\tilde{b}_{ij}$ is the current value. Then, problem (9) can be solved by iteratively solving the following problem:

$$\min_{\boldsymbol{b}_i} Tr(\boldsymbol{b}_i - \boldsymbol{s}_i)^T \boldsymbol{U} (\boldsymbol{b}_i - \boldsymbol{s}_i) + \beta \boldsymbol{b}_i^T \boldsymbol{f}_i \;\; s.t. \boldsymbol{b}_i^T \boldsymbol{1} = 1, \boldsymbol{b}_i \geq 0. \tag{10}$$

It has been proved that this iterative method decreases the objective of (9) in each iteration and it will converge to the optimal solution (Nie, et al., 2010). Then the problem with a convex objective and linear constraints can be solved efficiently by the standard convex optimization method using existing algorithm CLR. In our implementation, we set the largest number of iterations to 30.

References:

Nie, F.*, et al.* Efficient and robust feature selection via joint ℓ2, 1-norms minimization. In, *Advances in neural information processing systems*. 2010. p. 1813-1821.

Nie, F.*, et al.* The Constrained Laplacian Rank algorithm for graph-based clustering. In, *Thirtieth AAAI Conference on Artificial Intelligence*. 2016. p. 1969-1976.