Supplementary

Parameter Selection

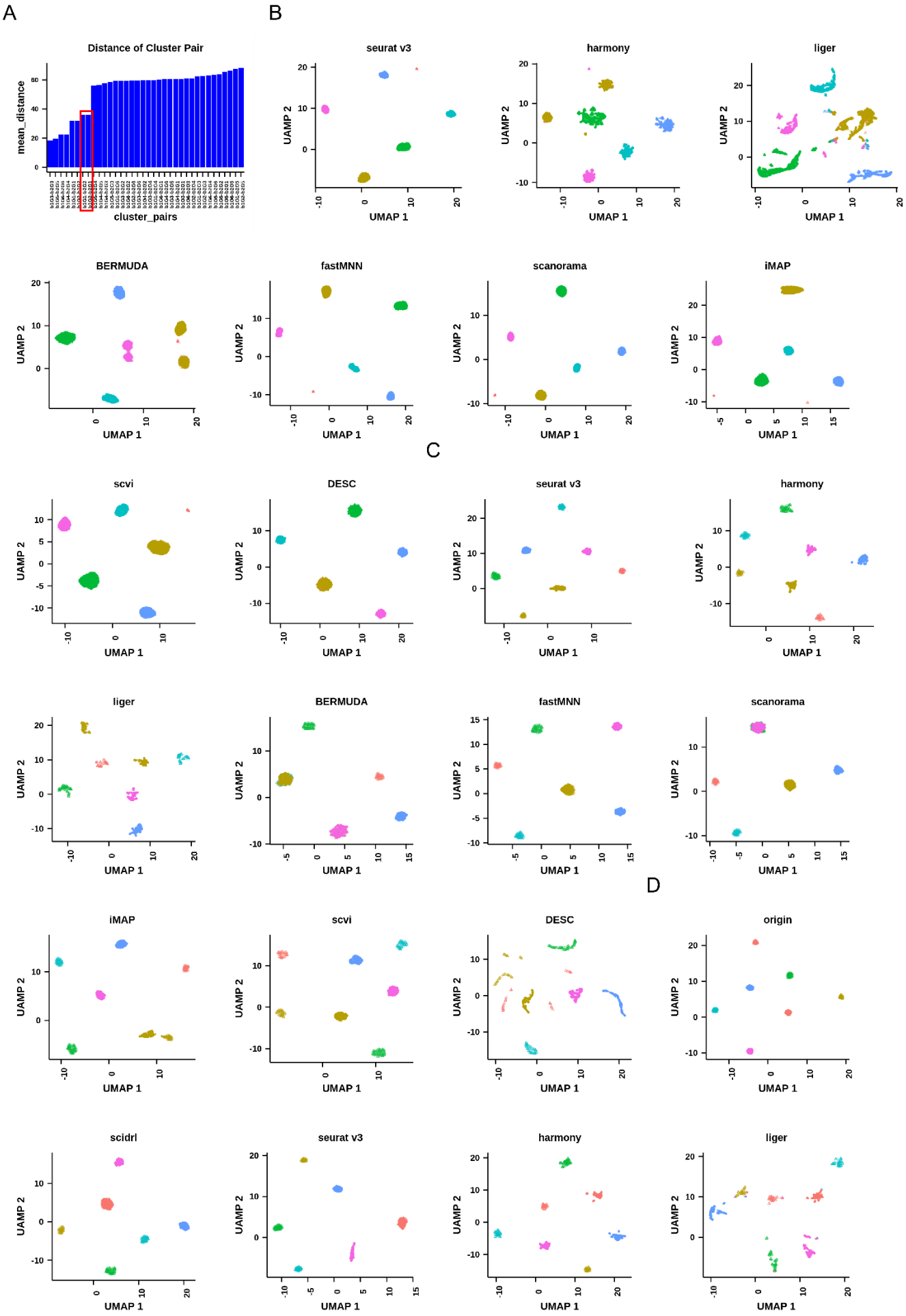
There are several parameters to be determined in SCIDRL, including the hyperparameters of the network and parameters β (the weight for the noise classifier) and λ (the weight for the discriminator). Experiments show that the performances are similar for a wide range of hyperparameters of the network. In our experiments, the default setting of manually optimized parameters is: depth of the autoencoder=3 with 64-10-64 neurons, depth of the noise classifier =1, depth of the discriminator =3 with 32-16-4 neurons, the weight of noise classifier $\beta = 1$.

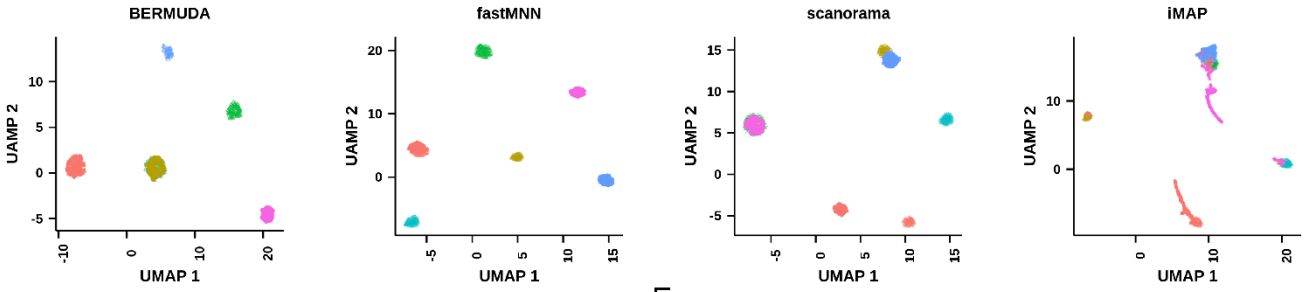
The most important parameter is the weight of the discriminator λ, which controls the trade-off between the generation of biological representations and the discrimination of them in different batches. It varies from 0.1 to 5 experimentally. The default value of $\lambda = 0.1$ is usually sufficient for achieving reasonable results for most analyses. To identify the best parameter for better integration, we proposed a quantitative heuristic strategy. To select the best parameter λ in a data-specific way, 1) we firstly compute the discriminator's loss $\widehat{loss_2}$ over the candidate parameter choices (λ =0.1, 0.5 or 1) and select the minimum value (i.e. λ_{base}), at which $\widehat{loss_2}$ has the tendency to firstly decreases and then increases. The discriminator's loss $\widehat{loss_2}$ measures how well mixtures of cells from different batches. In the training process, the discriminator's loss $\widehat{loss_2}$ usually decreases sharply during the beginning epochs, which indicates that the discriminator is well trained to classify different batches. After several epochs, the discriminator's loss increases to be stable and displays very limited fluctuations. 2) We then just compute the LISI-batch of each cell cluster over a grid of candidate parameter choices ($\lambda = \lambda_{base}:\lambda_{step}:5$, in which λ_{step}=0.1 when λ_{base}=0.1 or 0.5 and λ_{step}=1 when $\lambda_{base} = 1$) and select the elbow value (i.e. $\lambda_{critical}$) that increasing it sharply. The cell clusters are determined by running Seurat with the default resolution value as 0.1 on the cells in the integrated representation space when $\lambda = \lambda_{base}$. Although the cell clusters assigned by Seurat with default parameter may not be in complete agreement with the biological ground-truth, they can still reflect a certain layer of the cell type hierarchy tree. The optimal value λ_{optim} is obtained by $\lambda_{critical}$-λ_{step}. λ_{optim} will be further examined by UMAP visualizations, to be specific, test if some clusters, especially distant clusters, are mixed up when $\lambda = \lambda_{optim}$.

We displayed some examples in Figure S4. Firstly, determine λ_{base} in all datasets we used. We find that λ_{base} of simulated, pancreas, DC, cell line, mouse hematopoietic, human cerebral organoids, mouse retina, mouse brain, PBMC and eight organ datasets can be set as 0.1, whereas, it can be set as 0.5 of mouse atlas and mouse cortex datasets, which can be intuitively visualized by the variation curves of $\widehat{loss_2}$ in Figure S4A&S4B. Secondly, determine $\lambda_{critical}$ from UMAP visualizations and LISI-batch. In simulated 2 dataset (Figure S4C), when $\lambda = \lambda_{base}$, the LISI-batch of cluster 0 is close to 2 and are close to 1 for the remaining clusters (left first), which indicates cluster 0 has two batches and the other clusters have one batch. This phenomenon is further proved by UMAP visualization that only cluster 0 has two batches mixed (left second). Observing the curve, we find that when $\lambda = 0.5$, the LISI-batch of cluster 1 and cluster 5 have remarkable rise from 1 to 2, and it has obvious reduction from 2 to 1 for cluster 0, which warns us $\lambda_{critical}$ may be 0.5 (left first). As expected, the UMAP visualization of $\lambda = 0.5$ shows that cluster 0,1 and 5 are blended (left third), and the cell type labels indicate that Group 1, Group 2 and Group 5 are grouped together (left fourth). So, we recommend $\lambda_{optim} = 0.4$ for this dataset. In DC dataset (Figure S4D), when $\lambda = 0.1$, the LISI-batch of cluster 0,1 and 2 are 1.13, 1.78 and 1.85, which is consistent with UMAP that cluster 1 and 2 have two batches mixed, and cluster 0 has two batches separated (left first and second). When $\lambda = 0.5$, the LISI-batch of cluster 0 has an obvious increase from 1.13 to 1.79, which indicates fault mixture happens in cluster 0. This inference is testified by the batch mixtures of cluster 0 from UMAP visualization (left third). This occurs due to the fault mixtures of CD141 and CD1C (left fourth). For this data, we set $\lambda_{optim} = 0.4$. For mouse hematopoietic dataset (Figure S4E), when $\lambda = 0.1$, the UMAP visualization shows that cluster 2 is the only cluster that belongs to one batch (left second), which is consistent with 1.06 of LISI-batch (left first). When $\lambda = 0.3$, the LISI-batch of cluster 2 has a quick growth from 1.06 to 1.38 (left first). Inspecting the UMAP visualization, we find that the cells of cluster 2 are blended with cells from another batch (left third). The cell type labels show fault mixture of CMP, MPP and LTHSC (left fourth). So, we set λ_{optim} as 0.2. It worth noting that, at base

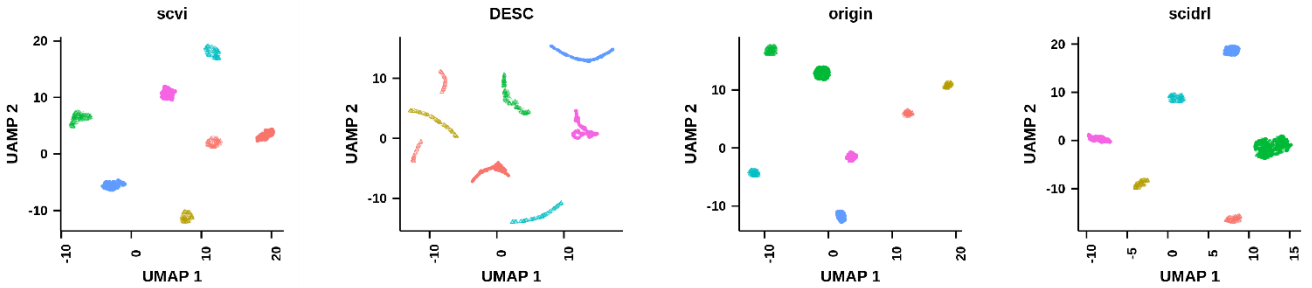
value, the LISI-batch of cluster 0 and 1 are 1.14 and 1.15, which indicates cluster 0 and 1 include one batch empirically. However, LISI-batch is not an absolute standard because when two batches are close but not mixed, the LISI-batch is still small. So further testified by UMAP visualization is necessary. For down-sampled pancreas dataset with alpha as shared cell type (Figure S4F), when $\lambda = \lambda_{base}$, we find that cluster 0 and 1 contain cells from two batches and cluster 6 belongs to 'baron' batch from UMAP (left second), the corresponding LISI-batch are 1.02, 1.17 and 1.007 (left first). Observing the variation of LISI-batch, we find that cluster 6 has steep rise from 1 to 1.73 when $\lambda = 0.6$ (left first). Inspecting the UMAP, we find that cluster 6 mixes up with cluster 1 and 0 (left third), which is caused by the fault mixtures of alpha, gamma and epsilon cells (left fourth). It is worth noting that, the LISI-batch of cluster 8 has a relatively severe oscillation though, the UMAP visualization shows that there is no fault mixture of clusters. So, $\lambda_{optim} = 0.5$ is an appropriate selection. As for multiple batches, we discussed cell line dataset and human cerebral organoid dataset. In cell line dataset (Figure S4G), when $\lambda = \lambda_{base}$, the UMAP visualization and LISI-batch indicate cluster 0, 1 and 2 contain cells from two batches (left second and first). When $\lambda = 0.3$, the LISI-batch of cluster 0 and 1 rise dramatically from 1.7 to 2.5 (left first), whose UMAP visualization also indicates wrong mixtures of 293t cells and jurkat cells (left fourth). So, we set $\lambda_{optim} = 0.2$ for our comparison. In human cerebral organoid dataset (Figure S4H), when $\lambda = 0.6$, we find that the LISI-batch of cluster 2, 3, 6 and 8 have obvious rise, especially for cluster 8, which is from 1.9 to 2.3 (left first). The UMAP visualization of $\lambda = 0.6$ shows that cluster 2, 3 and 8 are mixed (left third), which is caused by the mixing of GE NPCs, cortical NPCs and Non-telencephalon NPCs (left fourth). For this dataset, we select $\lambda_{optim} = 0.5$. For Pancreas dataset (Figure S4I), when $\lambda = \lambda_{base}$, we find all clusters have two batches mixed (left second), although some mixtures are insufficient from UMAP visualization. When $\lambda = 1$, no cluster is mixed wrongly, and the batches are mixed more thoroughly (left third). In this case, the rise of LISI-batch indicates more thorough integration. In a word, these examples testified the effectiveness of our strategy.

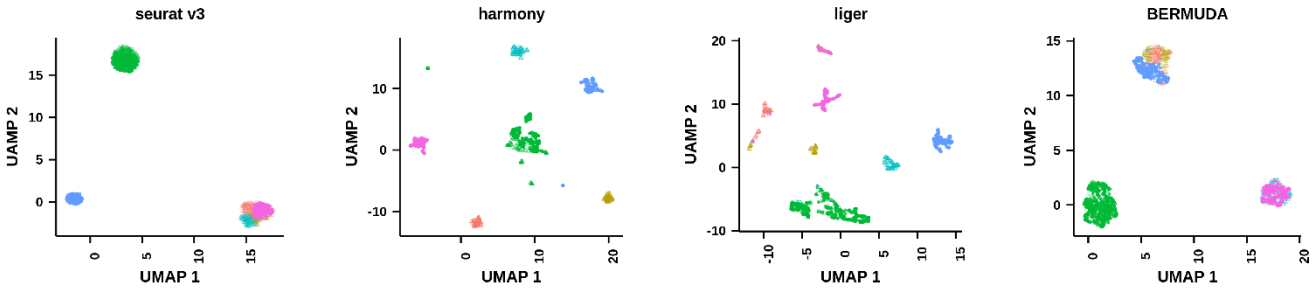
FIGURES AND TABLES

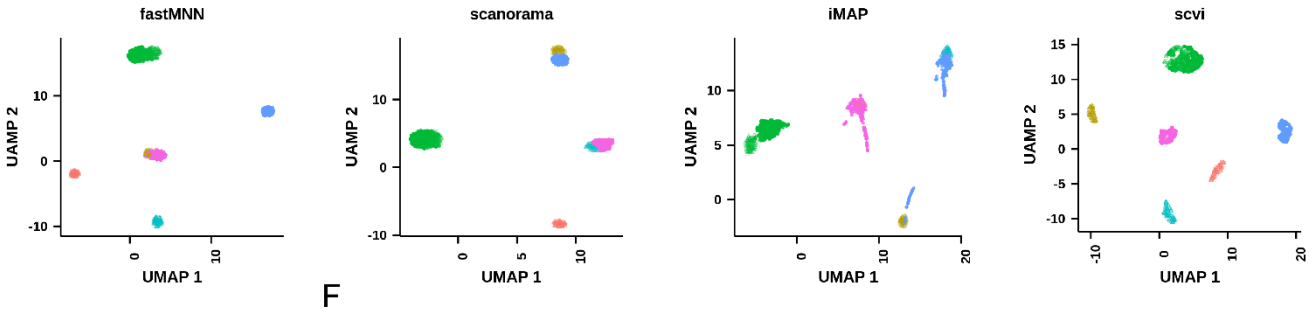



BERMUDA
fastMNN
scanorama
iMAP
UMAP 2
UMAP 1

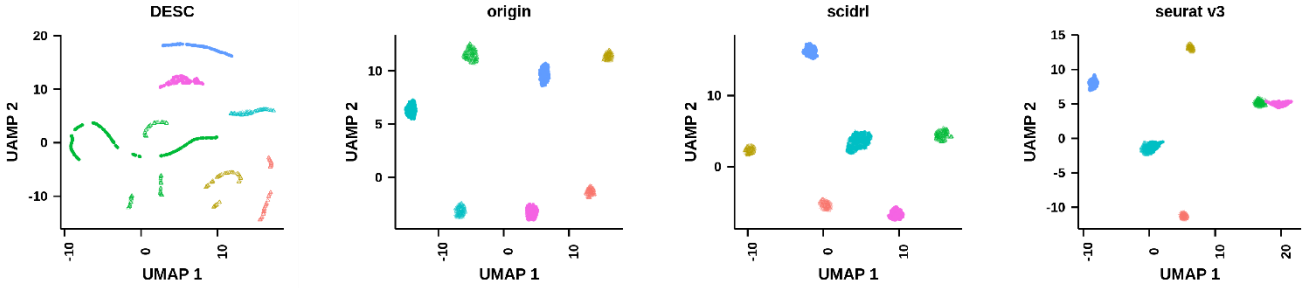
E

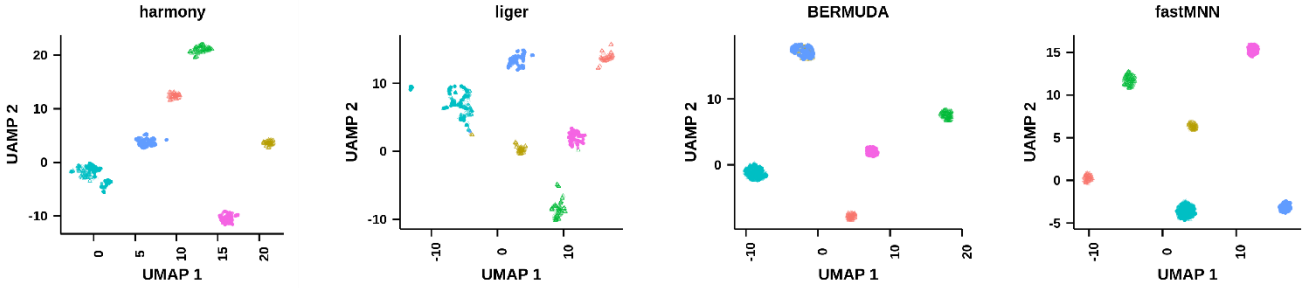

scvi
DESC
origin
scidrl
UMAP 2
UMAP 1


seurat v3
harmony
liger
BERMUDA
UMAP 2
UMAP 1


fastMNN
scanorama
iMAP
scvi
UMAP 2
UMAP 1

F


DESC
origin
scidrl
seurat v3
UMAP 2
UMAP 1


harmony
liger
BERMUDA
fastMNN
UMAP 2
UMAP 1

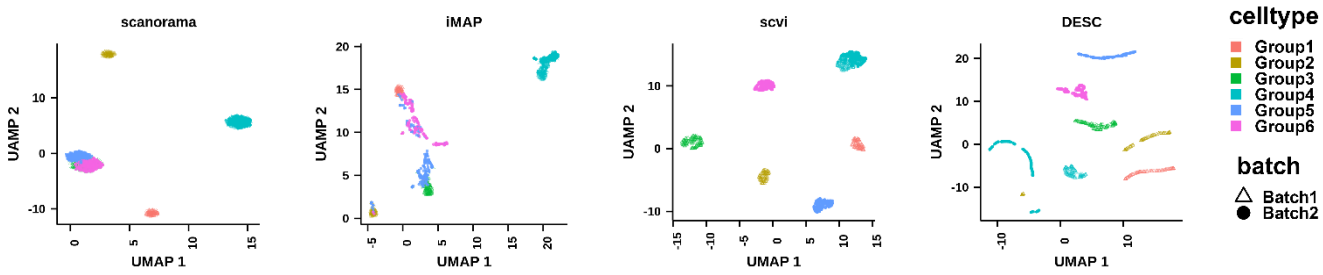
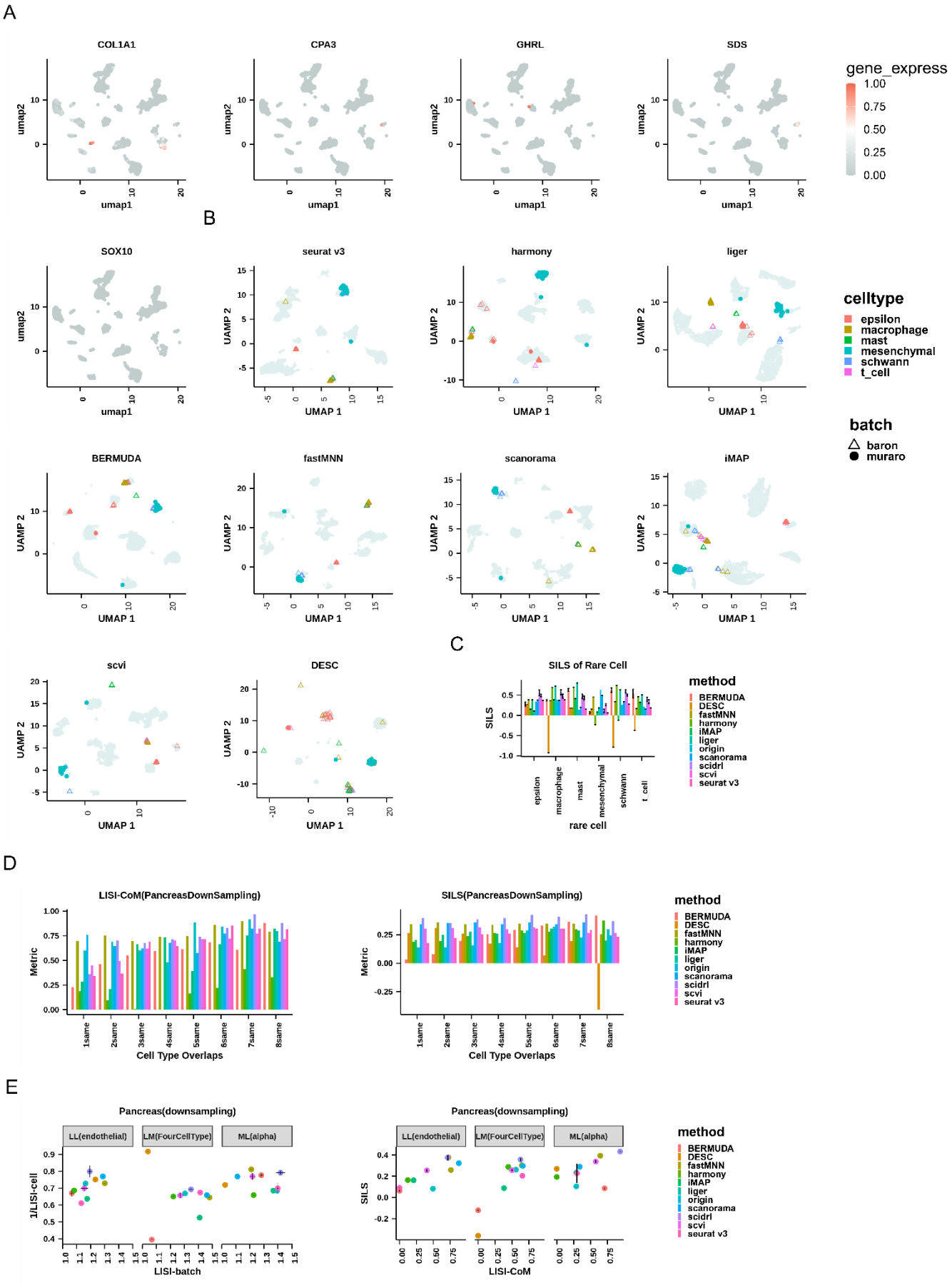
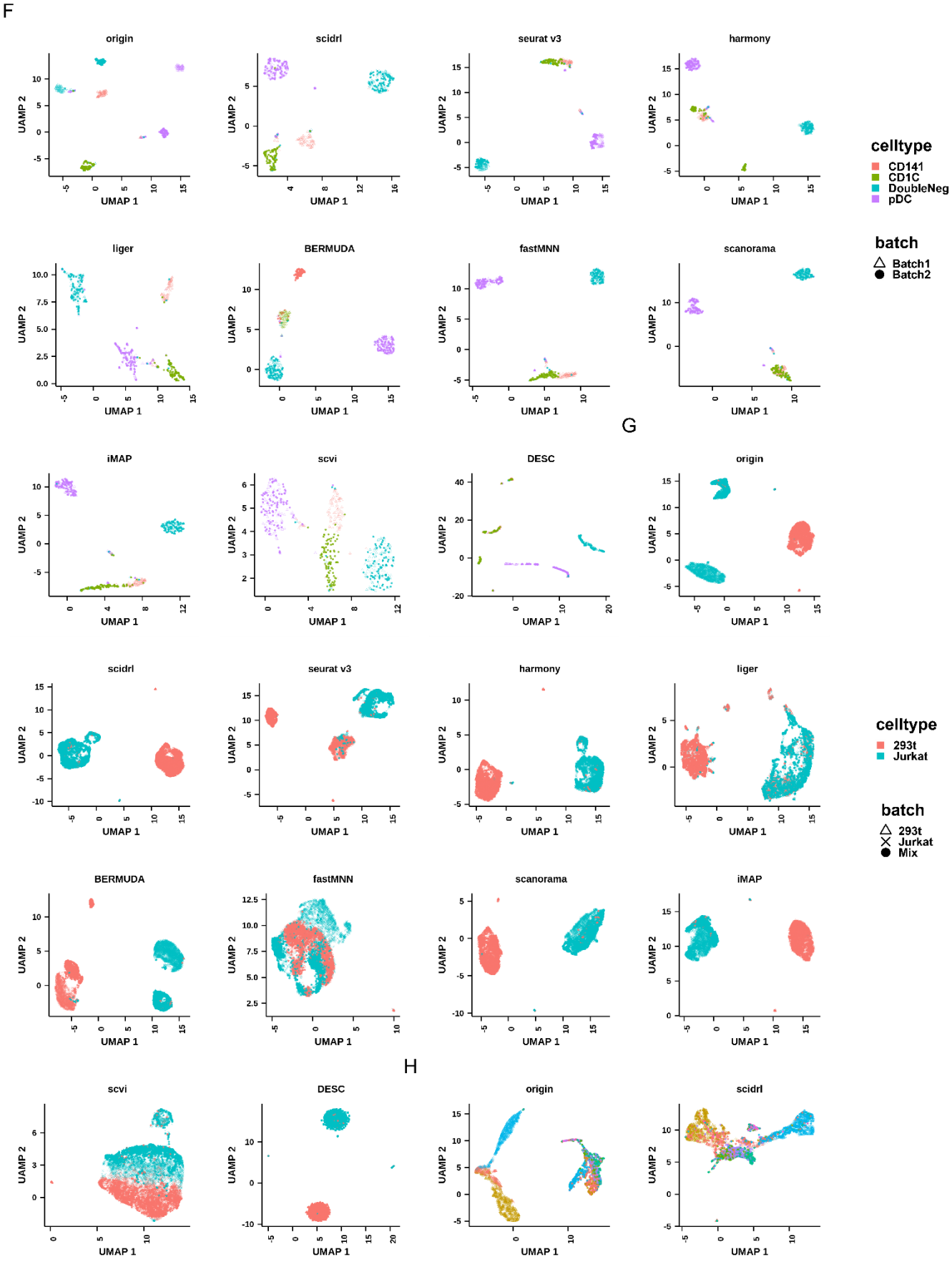


Figure S1. Removing batch effect in simulated data.

A. Distances of cell types in different batches. The red box highlights the nearest cell type pairs in different batches.

B-F. Performance comparison of ten integrated methods for UMAP visualizations on datasets with rare cell types (B) and datasets with one shared cell types: Group2, Group1, Group3 or Group4 (C-F). Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.


A
COL1A1
CPA3
GHRL
SDS
umap1
umap2
gene_express
1.00
0.75
0.50
0.25
0.00
B
SOX10
seurat v3
harmony
liger
UMAP 1
UMAP 2
celltype
epsilon
macrophage
mast
mesenchymal
schwann
t_cell
batch
baron
muraro
BERMUDA
fastMNN
scanorama
iMAP
C
scvi
DESC
SILS of Rare Cell
SILS
rare cell
method
BERMUDA
DESC
fastMNN
harmony
iMAP
liger
origin
scanorama
scidrl
scvi
seurat v3
D
LISI-CoM(PancreasDownSampling)
SILS(PancreasDownSampling)
Metric
Cell Type Overlaps
1same
2same
3same
4same
5same
6same
7same
8same
E
Pancreas(downsampling)
LL(endothelial)
LM(FourCellType)
ML(alpha)
1/LISI-cell
LISI-batch
SILS
LISI-CoM


F
origin
scidrl
seurat v3
harmony
liger
BERMUDA
fastMNN
scanorama
UMAP 1
UMAP 2
celltype
CD14!
CD1C
DoubleNeg
pDC
batch
Batch1
Batch2
G
iMAP
scvi
DESC
origin
scidrl
seurat v3
harmony
liger
celltype
293t
Jurkat
batch
293t
Jurkat
Mix
BERMUDA
fastMNN
scanorama
iMAP
H
scvi
DESC
origin
scidrl

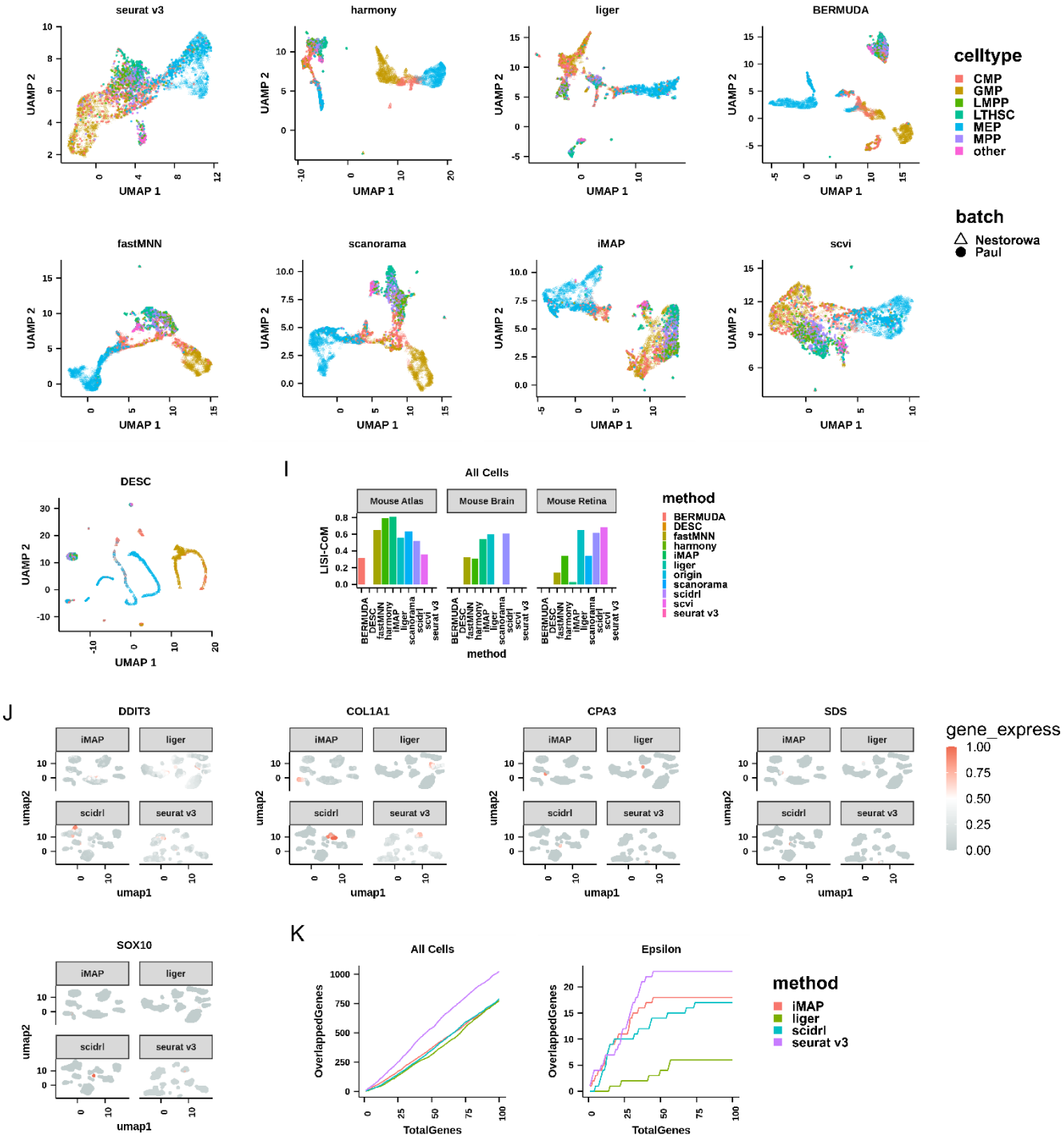


Figure S2. Removing batch effect in two- or three-batch datasets.

A. Expression patterns of marker genes of mesenchymal, macrophage, mast, epsilon and schwann cells. Each point represents a cell and the cell is colored according to the expressions their marker genes: COL1A1, SDS, CPA3, GHRL and SOX10.

B. Performance comparison of nine methods for UMAP visualizations on pancreas dataset. Each point represents a cell, the rare cells are colored according to their known cell type labels and all cells are shaped according to their batch labels.

C. Performance comparison of ten methods for SILS on six rare cell types of pancreas dataset. The x-axis represents rare cell types and the y-axis represents SILS (larger value means better performance). Different colors represent different methods.

D. Performance comparison of the ten integrated methods for two metrics (SILS and LISI-CoM) on 57 down-sampling pancreas datasets. The number of share cell types in each dataset ranges from one to eight. The x-axis represents LISI-batch and the y-axis represents 1/LISI-cell. Different colors represent different methods.

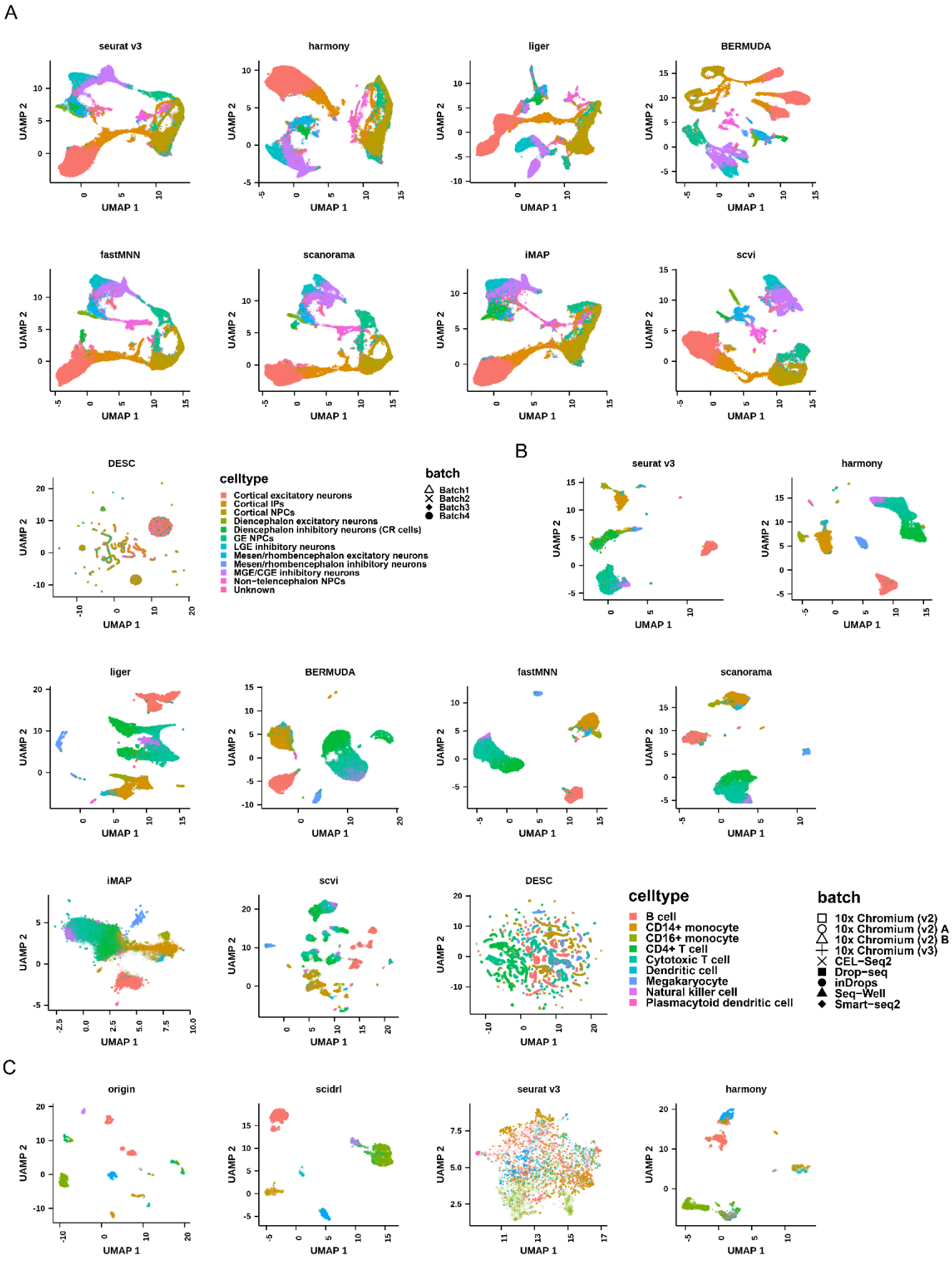
E. Performance comparison of ten integrated methods for four metrics LISI-batch & LISI-cell (left) , SILS & LISI-CoM (right) on different down-sampling pancreas datasets: dataset with alpha as shared cell type (right), dataset with endothelial as shared cell type (middle), dataset with four cell types as shared cell types (left). The x-axis represents LISI-batch or LISI-CoM and the y-axis represents 1/LISI-cell or SILS. Different colors represent different methods.

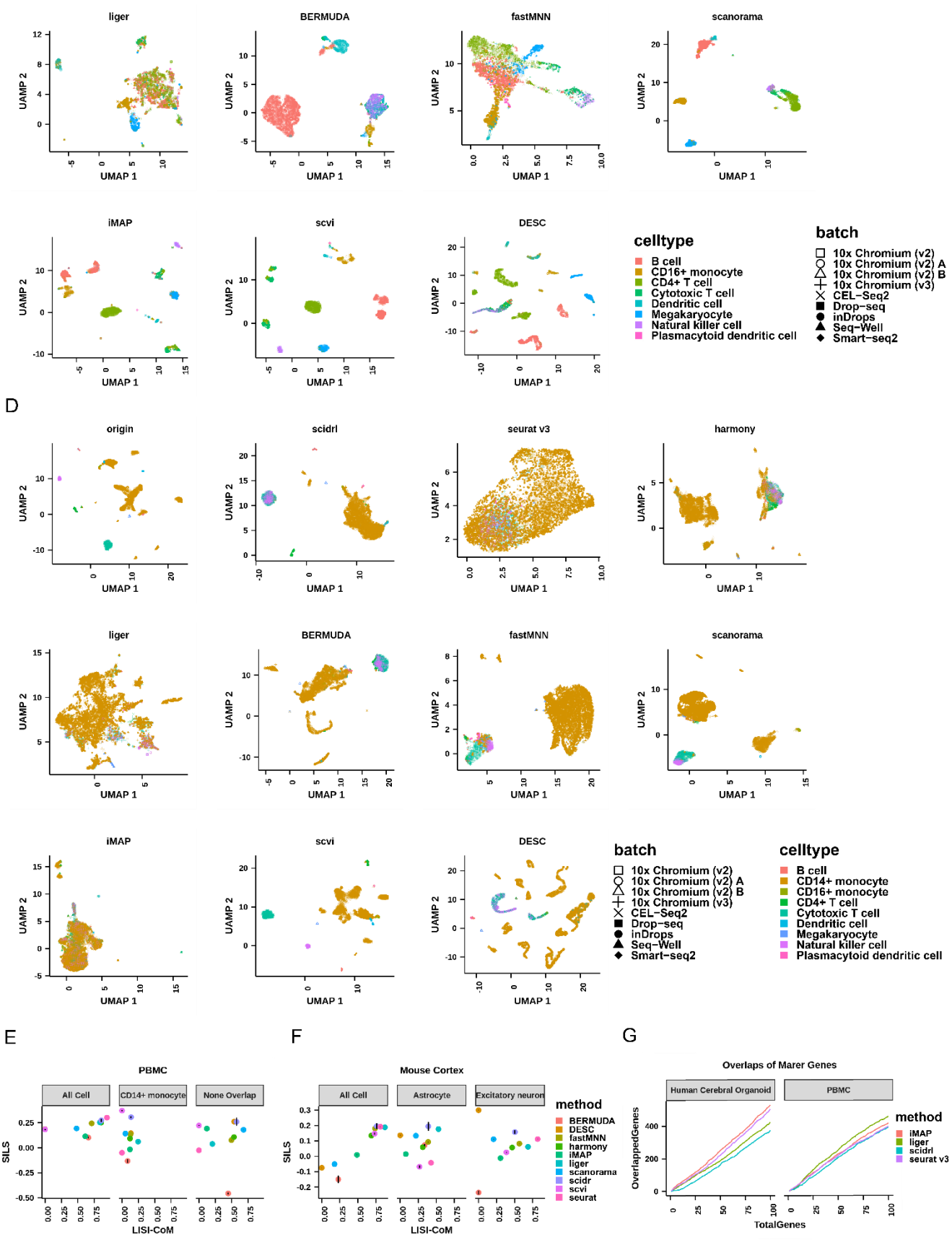
F-H. Performance comparison of ten methods for UMAP visualizations on DC (F), cell line (G) and mouse hematopoietic (H) datasets. Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.

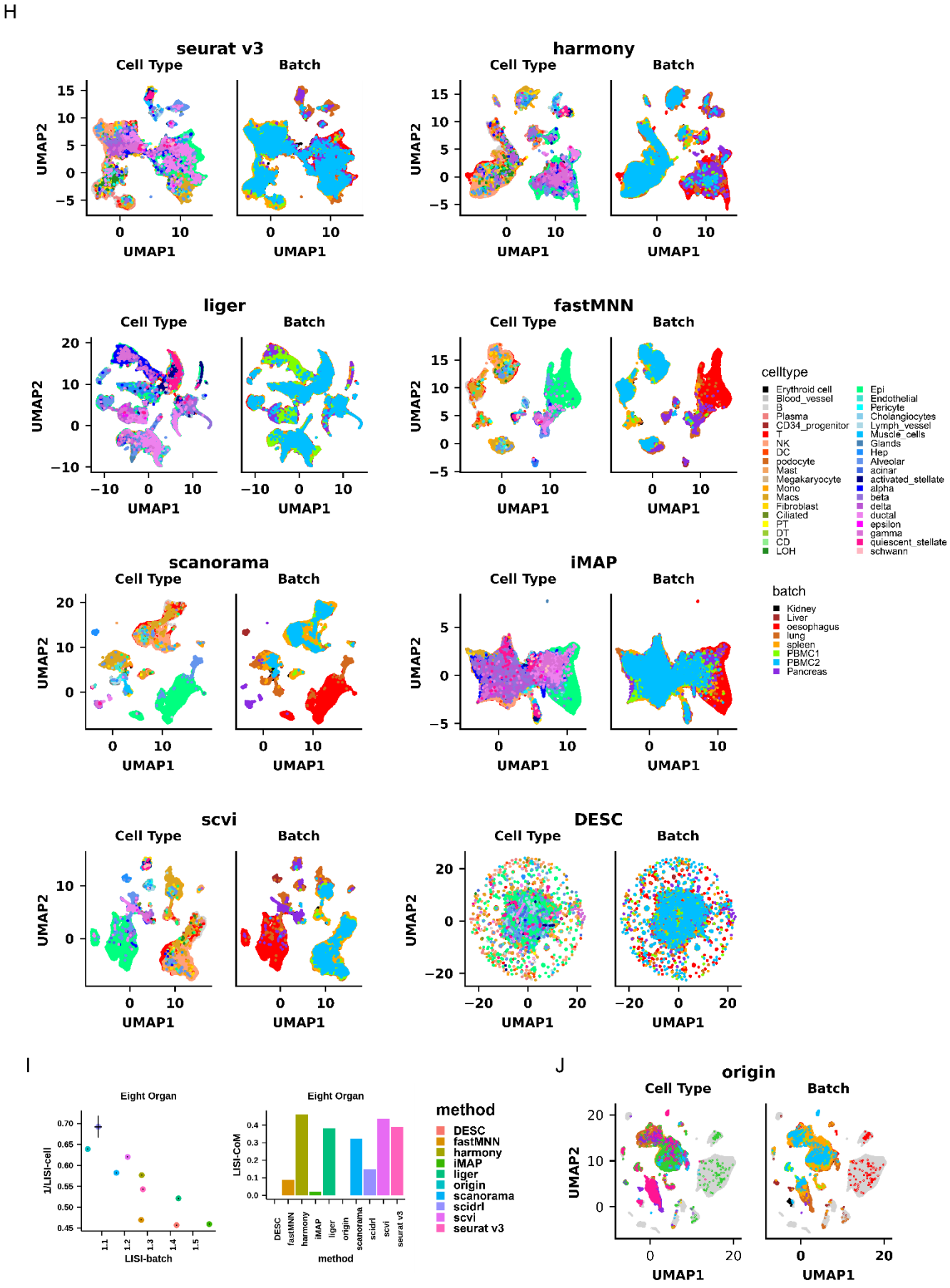
I. Performance comparison of ten methods for LISI-CoM on mouse atlas (left), mouse brain (middle) and mouse retina dataset (right). The x-axis represents methods and the y-axis represents SILS. Different colors represent different methods.

J. Performance comparison of four methods (SCIDRL, Seurat v3, Liger and iMAP) on integrated gene expression patterns of marker genes of ER-stress beta, mesenchymal, mast, macrophage and schwann cells. Each point represents a cell and the cell is colored according to the expression value of its marker gene.

K. Performance comparison of the four integrated methods (SCIDRL, Seurat v3, Liger and iMAP) on overlaps between marker genes found by these methods and found by original data for each cell type. The x-axis represents top N genes and the y-axis represents the total number of overlaps in top N genes (N ranges from 1 to 100). The left panel shows the results of all cells and the right panel shows the results of epsilon cells.


A
seurat v3
harmony
liger
BERMUDA
fastMNN
scanorama
iMAP
scvi
DESC
UMAP 1
UMAP 2
celltype
Cortical excitatory neurons
Cortical IPs
Cortical NPCs
Diencephalon excitatory neurons
Diencephalon inhibitory neurons (CR cells)
GE NPCs
LGE inhibitory neurons
Mesen/rhombencephalon excitatory neurons
Mesen/rhombencephalon inhibitory neurons
MGE/CGE inhibitory neurons
Non-telencephalon NPCs
Unknown
batch
Batch1
Batch2
Batch3
Batch4
B
seurat v3
harmony
liger
BERMUDA
fastMNN
scanorama
iMAP
scvi
DESC
celltype
B cell
CD14+ monocyte
CD16+ monocyte
CD4+ T cell
Cytotoxic T cell
Dendritic cell
Megakaryocyte
Natural killer cell
Plasmacytoid dendritic cell
batch
10x Chromium (v2)
10x Chromium (v2) A
10x Chromium (v2) B
10x Chromium (v3)
CEL-Seq2
Drop-seq
inDrops
Seq-Well
Smart-seq2
C
origin
scidrl
seurat v3
harmony


liger
BERMUDA
fastMNN
scanorama
iMAP
scvi
DESC
UMAP 1
UMAP 2
celltype
B cell
CD16+ monocyte
CD4+ T cell
Cytotoxic T cell
Dendritic cell
Megakaryocyte
Natural killer cell
Plasmacytoid dendritic cell
batch
10x Chromium (v2)
10x Chromium (v2) A
10x Chromium (v2) B
10x Chromium (v3)
CEL−Seq2
Drop−seq
inDrops
Seq−Well
Smart−seq2
D
origin
scidrl
seurat v3
harmony
liger
BERMUDA
fastMNN
scanorama
iMAP
scvi
DESC
batch
celltype
E
PBMC
All Cell
CD14+ monocyte
None Overlap
SILS
LISI-CoM
F
Mouse Cortex
All Cell
Astrocyte
Excitatory neuron
SILS
LISI-CoM
method
BERMUDA
DESC
fastMNN
harmony
iMAP
liger
scanorama
scidr
scvi
seurat
G
Overlaps of Marer Genes
Human Cerebral Organoid
PBMC
OverlappedGenes
TotalGenes
method
iMAP
liger
scidrl
seurat v3


H
seurat v3
harmony
liger
fastMNN
scanorama
iMAP
scvi
DESC
Cell Type
Batch
UMAP1
UMAP2
celltype
Erythroid cell
Blood_vessel
B
Plasma
CD34_progenitor
T
NK
DC
podocyte
Mast
Megakaryocyte
Mono
Macs
Fibroblast
Ciliated
PT
DT
CD
LOH
Epi
Endothelial
Pericyte
Cholangiocytes
Lymph_vessel
Muscle_cells
Glands
Hep
Alveolar
acinar
activated_stellate
alpha
beta
delta
ductal
epsilon
gamma
quiescent_stellate
schwann
batch
Kidney
Liver
oesophagus
lung
spleen
PBMC1
PBMC2
Pancreas
I
Eight Organ
1/LISI-cell
LISI-batch
LISI-CoM
method
DESC
fastMNN
harmony
iMAP
liger
origin
scanorama
scidrl
scvi
seurat v3
J
origin

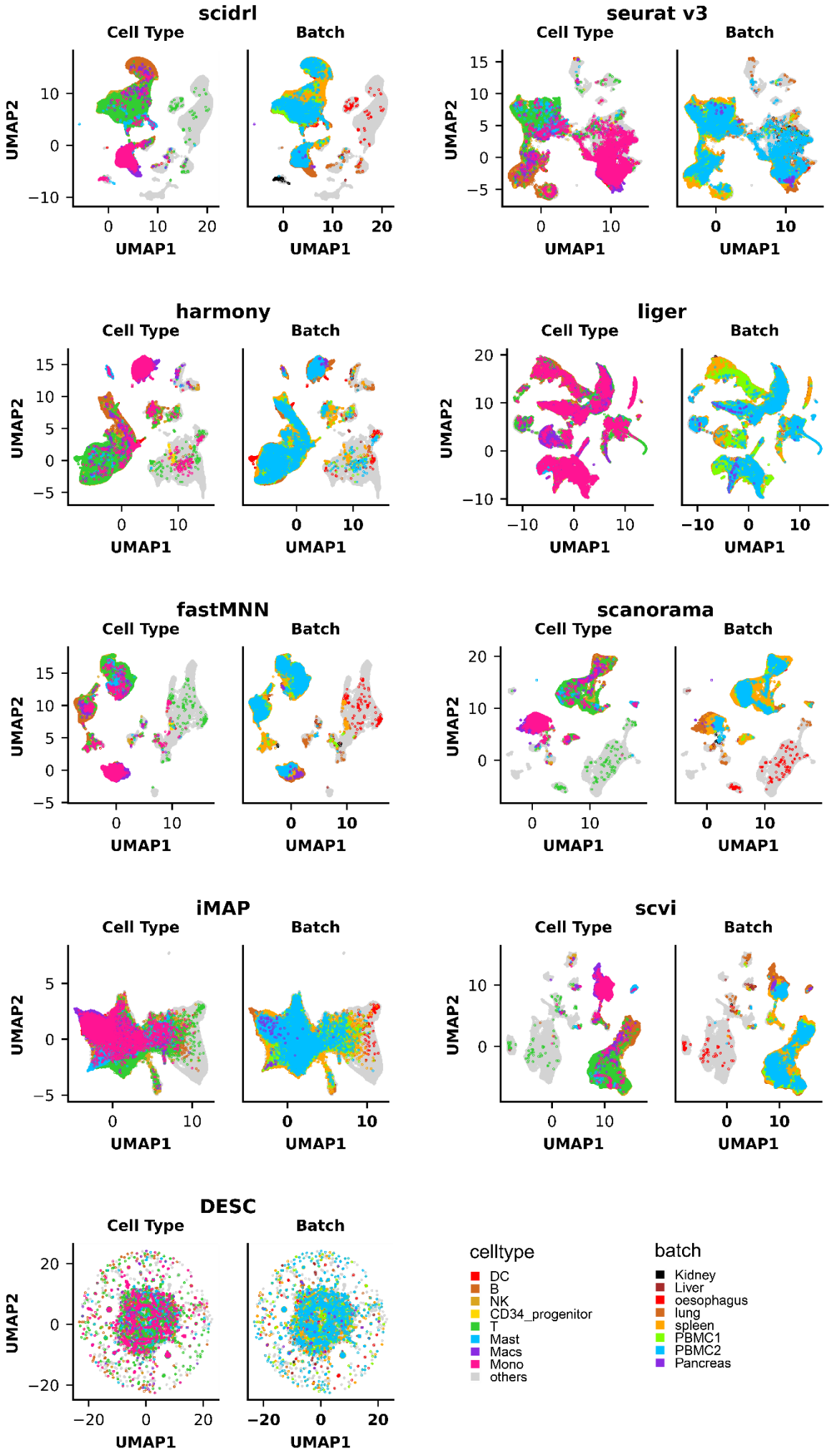

scidrl
Cell Type
Batch
seurat v3
harmony
liger
fastMNN
scanorama
iMAP
scvi
DESC
UMAP1
UMAP2
celltype
DC
B
NK
CD34_progenitor
T
Mast
Macs
Mono
others
batch
Kidney
Liver
oesophagus
lung
spleen
PBMC1
PBMC2
Pancreas

Figure S3. Removing batch effect in multiple datasets.

A. Performance comparison of nine methods for UMAP visualizations on human cerebral organoids dataset. Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.

B. Performance comparison of nine methods for UMAP visualizations on PBMC dataset. Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.

C. Performance comparison of ten methods for UMAP visualizations on down-sampling PBMC dataset with none shared cell type in nine batches. Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.

D. Performance comparison of ten methods for UMAP visualizations on down-sampling PBMC dataset with CD14+ monocyte as shared cell type in nine batches. Each point represents a cell and the cell is colored according to its known cell type label and shaped according to its batch label.

E. Performance comparison of the ten integrated methods for two metrics (SILS and LISI-CoM) on PBMC (dataset with all cells (left), dataset with CD14+ monocyte as shared cell type (middle) and dataset with zero shared cell type (right)). The x-axis represents LISI-CoM and the y-axis represents SILS. Different colors represent different methods.

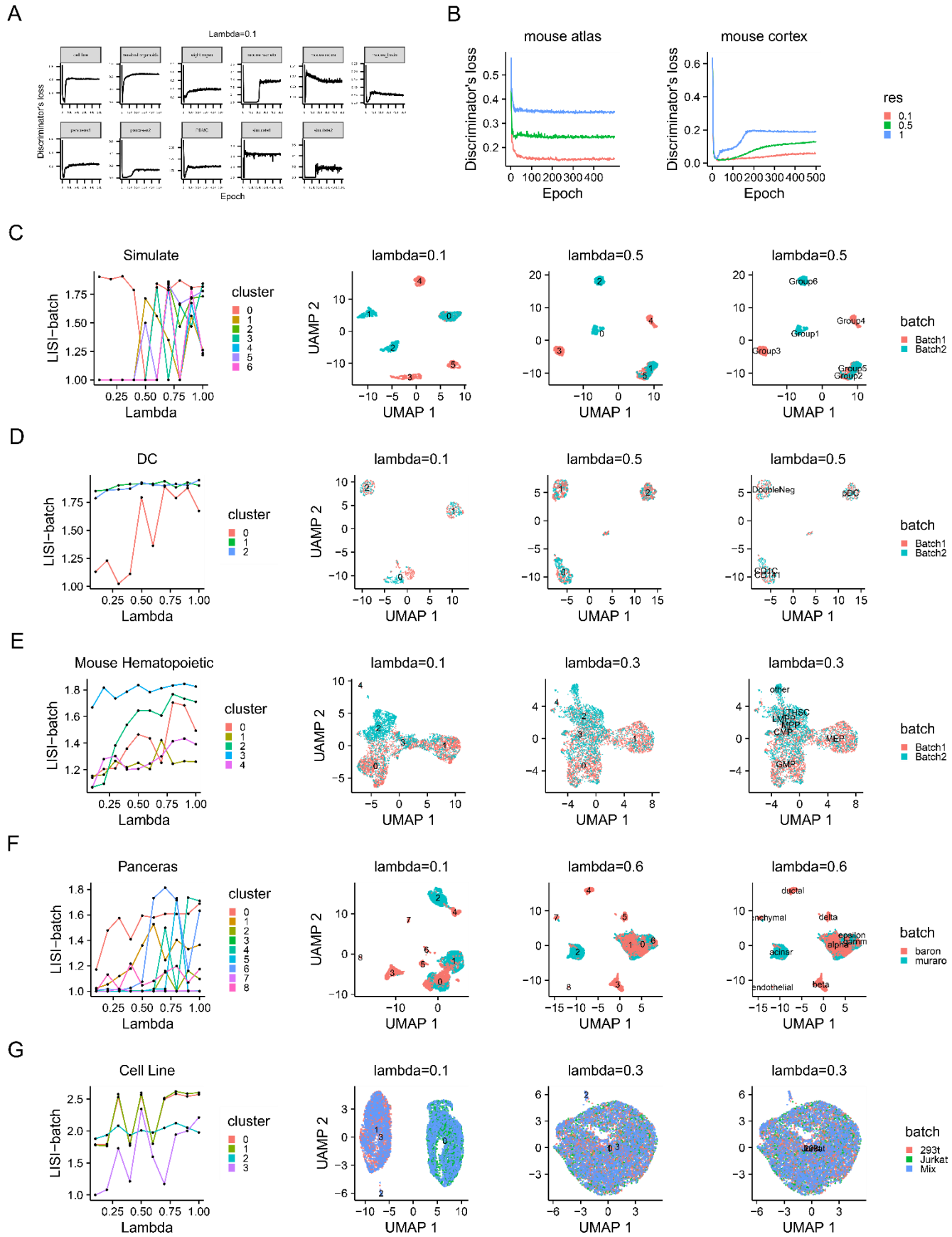
F. Performance comparison of the ten integrated methods for two metrics (SILS and LISI-CoM) on mouse coretex (dataset with all cells (left), dataset with astrocyte as shared cell type (middle) and dataset with excitatory as shared cell type (right)). The x-axis represents LISI-CoM and the y-axis represents SILS. Different color represents different methods.

G. Performance comparison of the four integrated methods (SCIDRL, Seurat v3, Liger and iMAP) on overlaps between marker genes found by these methods and found by original data for each cell type on human cerebral organoids (left) and PBMC dataset (right). The x-axis represents top N genes and the y-axis represents the total number of overlaps in top N genes (N ranges from 1 to 100).

H. Performance comparison of nine methods (except for Bermuda) for UMAP visualizations on eight-organ dataset. Each point represents a cell and the cell is colored according to its known cell type label (left) and its batch label (right).

I. Performance comparison of the ten integrated methods for three metrics on eight-organ dataset (Left: LISI-batch and 1/LISI-cell, Right: LISI-CoM). The x-axis represents LISI-batch or methods and the y-axis represents 1/LISI-cell or LISI-CoM. Different colors represent different methods.

J. Performance comparison of the nine integrated methods (except for Bermuda) for UMAP visualization on eight-organ dataset. Each point represents a cell, the immune cells are colored according to their known cell type labels and the remaining cells are colored using gray (left), all cells are colored according to its batch label (right).


A
Lambda=0.1
Discriminator's loss
Epoch
B
mouse atlas
mouse cortex
Discriminator's loss
Epoch
res
0.1
0.5
1
C
Simulate
lambda=0.1
lambda=0.5
lambda=0.5
LISI−batch
Lambda
cluster
UMAP 2
UMAP 1
batch
Batch1
Batch2
Group6
Group4
Group1
Group3
Group5
Group2
D
DC
lambda=0.1
lambda=0.5
lambda=0.5
DoubleNeg
pDC
cDC1
cDC2
E
Mouse Hematopoietic
lambda=0.1
lambda=0.3
lambda=0.3
other
THSC
LMPP
MEP
CMP
GMP
F
Panceras
lambda=0.1
lambda=0.6
lambda=0.6
ductal
nchymal
delta
epsilon
gamm
alpha
acinar
endothelial
beta
baron
muraro
G
Cell Line
lambda=0.1
lambda=0.3
lambda=0.3
293t
Jurkat
Mix

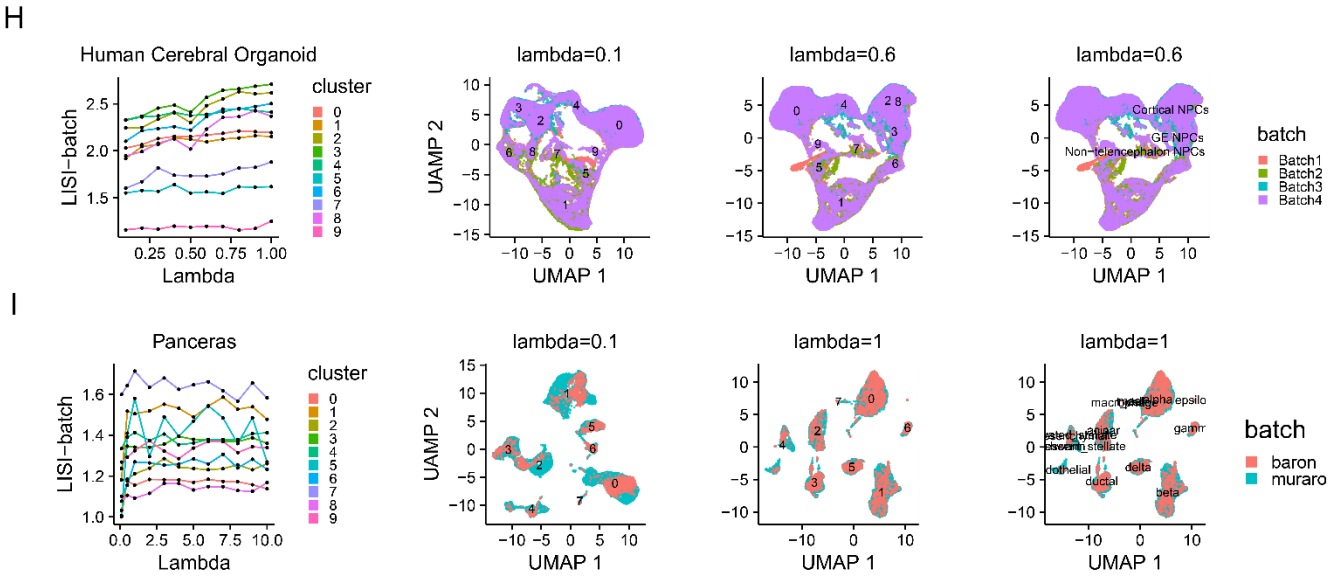


Figure S4. Parameter selections on datasets.

A. Variation curves of Discriminator's loss $\widehat{loss_2}$ when $\lambda = 0.1$ for datasets with 0.1 as base value. The x-axis represents the number of epochs and the y-axis represents the value of $\widehat{loss_2}$.

B. Variation curves of Discriminator's loss $\widehat{loss_2}$ for mouse atlas (left) and mouse cortex data (right) for different λ. The x-axis represents the number of epochs and the y-axis represents the value of $\widehat{loss_2}$. The colors represent different values of λ.

C. Parameter selections on Simulated 2 data: The x-axis corresponds to the values (from λ_{base} to 1 or 10) of λ. The y-axis represents the LISI-batch value for the dataset (first on the left). UMAP visualizations of different values of λ, which are λ_{base} (second on the left), $\lambda_{critical}$ for cluster labels (third on the left) and for known cell type labels (fourth on the left). Different colors represent different batches. The clusters are identified by Louvain clustering.

D-I. the same as (C) on different datasets: DC (D), mouse hematopoietic (E), down-sampled pancreas (F), cell line (G), human cerebral organoid (H) and pancreas (I) datasets.

Dataset	Batch	Cell	Gene	Technology	Percentage of shared cells (types) (%)	Number of rare cell types (proportion)	λ
simulated 1	2	4,000 4,500	500	simulated	100 (100)	1 (0.01)	0.1
simulated 2	2	695 665	500	simulated	27.1 (16.7)	0	0.1
Pancreas	2	8,654 2,122	1,398	CEL-seq2 Drop-seq	94 (53.3)	6 (0.0006,0.001,0.001,0.002, 0.005,0.007)	1
DC	2	283, 286	1,596	Smart-seq Smart-seq	66.8 (50)	0	0.4
Mouse hemato	2	2,729 1,920	4,649	Smart-seq2 MARS-seq	76.2 (42.8)	0	0.2
Mouse Atlas	2	21,855 13,320	20,00	Micro-seq Smart-seq	100 (100)	0	1
Mouse Brain	2	302,175 156,049	1,970	Drop-seq SPLiT-seq	99.4 (64.2)	8 (e-3, e-4, e-5)	1
Mouse Retina	2	44,808 27,499	1,460	Drop-seq Drop-seq	97.4 (38.5)	9 (e-3, e-4, e-5)	0.1
Cell Line	3	3,053 2,676 3,162	2,000	10x 10x 10x	Median:70.8 (Median:50)	0	0.2
Human Cerebral Organoids	4	17,019 8,581 9,433 14,120	2,000	10x 10x 10x 10x	Median:99.2 (Median:91.3)	1 (9e-3)	0.4
PBMC	18	526 526 3,222 3,222 3,222 6,584 3,727 6,584 3,362	2,146	10x (v2)_A 10x (v2)_B 10x (v2) 10x (v3) inDrops Seq-Well Drop-seq CEL-Seq2 Smart-seq2	Median:99 (Median:82.6)	1 (0.005)	5
Mouse Cortex	8	644 5,571 3,130 5,599	2,000	Smart-seq2 DroNC-seq 10x (v2) sci-RNA-seq	Median:87.1 (Median:57.1)	0	5
Eight Organ	8	4,487 8,367 87,947 57,020 94,257 6,584 8,569	1,319	Drop-seq inDrops 10x (v3)	Median:21.4 (Median:20.6)	0	0.1

Table S1 – Statistics of Datasets.

Each column corresponds to one statistic, each row corresponds to one dataset.

Dataset	Percentage of shared cells (%)	Percentage of shared cell types (%)	LISI-CoM (ranking)	SILS (ranking)	Combination (ranking)	Median/Mean (ranking)
PBMC (None Overlap)	0	0	2	2	2	3/3.5
Pancreas-LL (endothelial)	7.7	11.1	4	2	3	
Eight Organ	Median:21.4	Median:20.6	6		6	
Simulated 2	27.1	16.7	1	5	3	
Pancreas-LM (delta, endothelial, epsilon, gamma)	24.5	26.67	4	1	2.5	2.5/2.5
Pancreas-ML (alpha)	60.5	11.1	1	1	1	2/2
Mouse Cortex (Astrocyte)	Median: 63.3	Median: 25	3	1	2	
PBMC (CD14+ monocyte)	79.5	33.3	2	2	2	
Mouse Retina	97.4	38.5	3		3	
Mouse Hemato	76.2	42.8	1	1	1	
Mouse Cortex (Excitatory neuron)	Median: 89.5	Median: 45	4	2	3	
DC	66.8	50	1	6	3.5	2/2.44
Cell Line	Median:70.8	Median:50	1	6	3.5	
Pancreas	94	53.3	1	3	2	
Mouse Cortex	Median:87.1	Median:57.1	3	1	2	
Mouse Brain	99.4	64.2	1		1	
PBMC	Median:99	Median:82.6	2	2	2	
Human Cerebral Organoids	Median:99.2	Median:91.3	1	1	1	
Simulated1	100	100	1	6	3.5	
Mouse Atalas	100	100	6	6	3.5	

Table S2 – Performance of SCIDRL on different categories of datasets

Each column corresponds to one statistic, each row corresponds to one dataset. The colors represent different categories of datasets, which is little shared cells and little shared cell types (LL), little shared cells and many shared cell types (LM), many shared cells and little shared cell types (ML) and many shared cells and many shared cell types (MM).