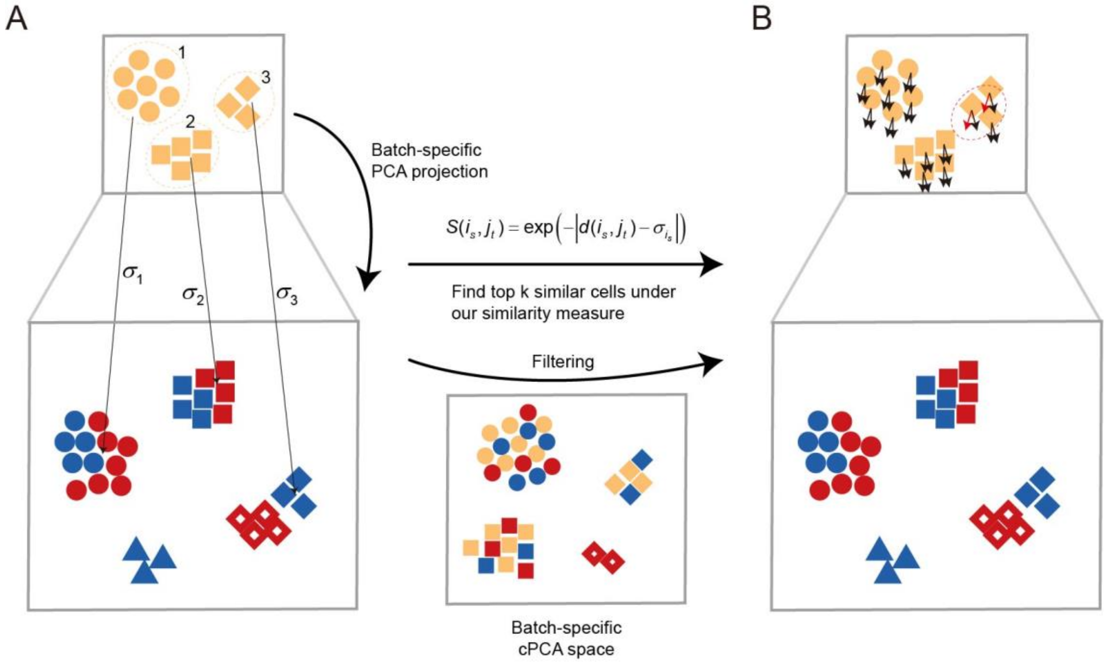
Supplemental Material for “Accurate integration of multiple heterogeneous single-cell RNA-seq data sets by learning contrastive biological variation”

Yang Zhou[1], Qiongyu Sheng[1], Jing Qi[1], Jiao Hua[1], Bo Yang[1], Lei Wan[1], and Shuilin Jin[1*]

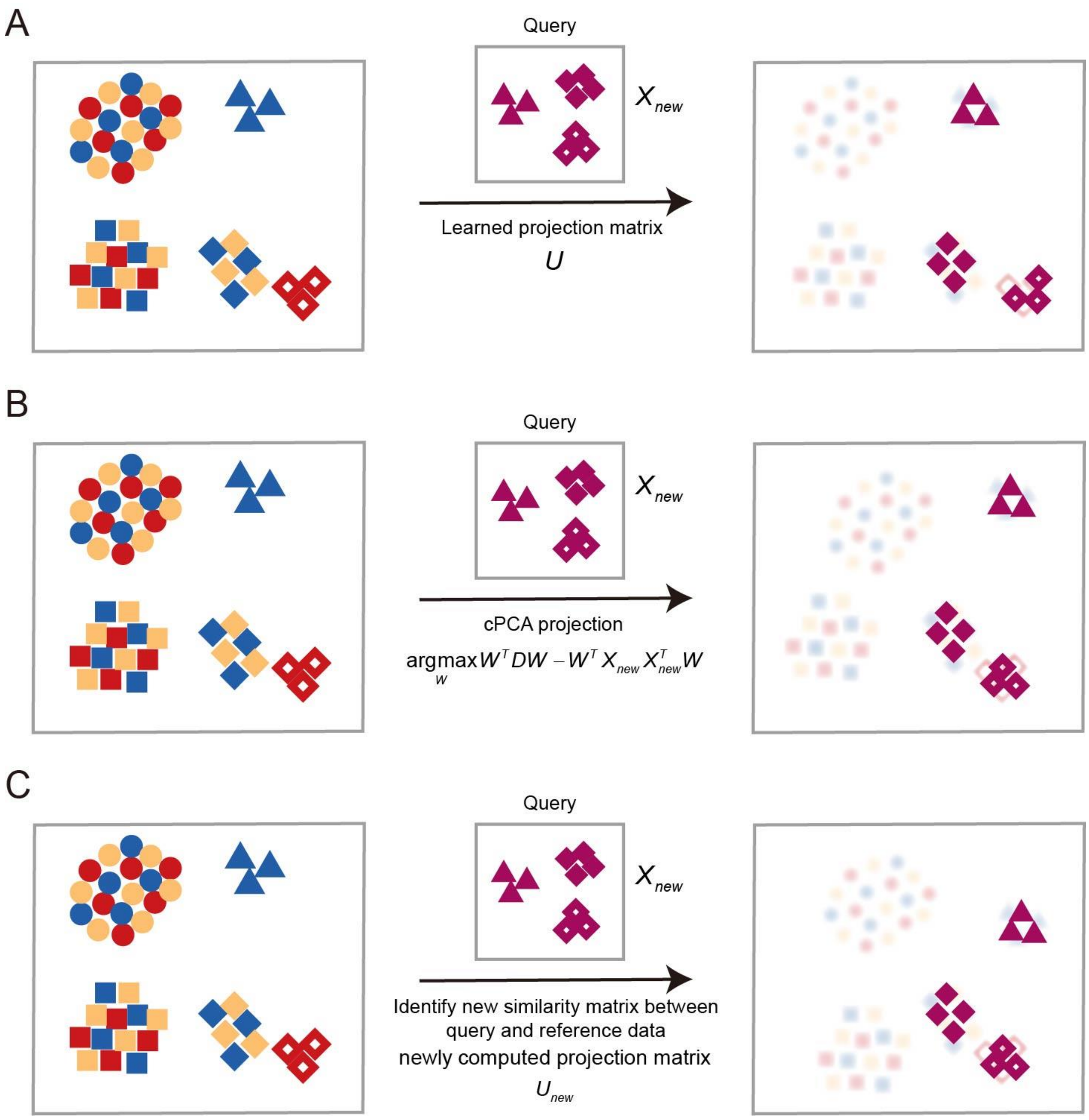
[1] School of Mathematics, Harbin Institute of Technology, Harbin, Heilongjiang Province, China

* Correspondence to:

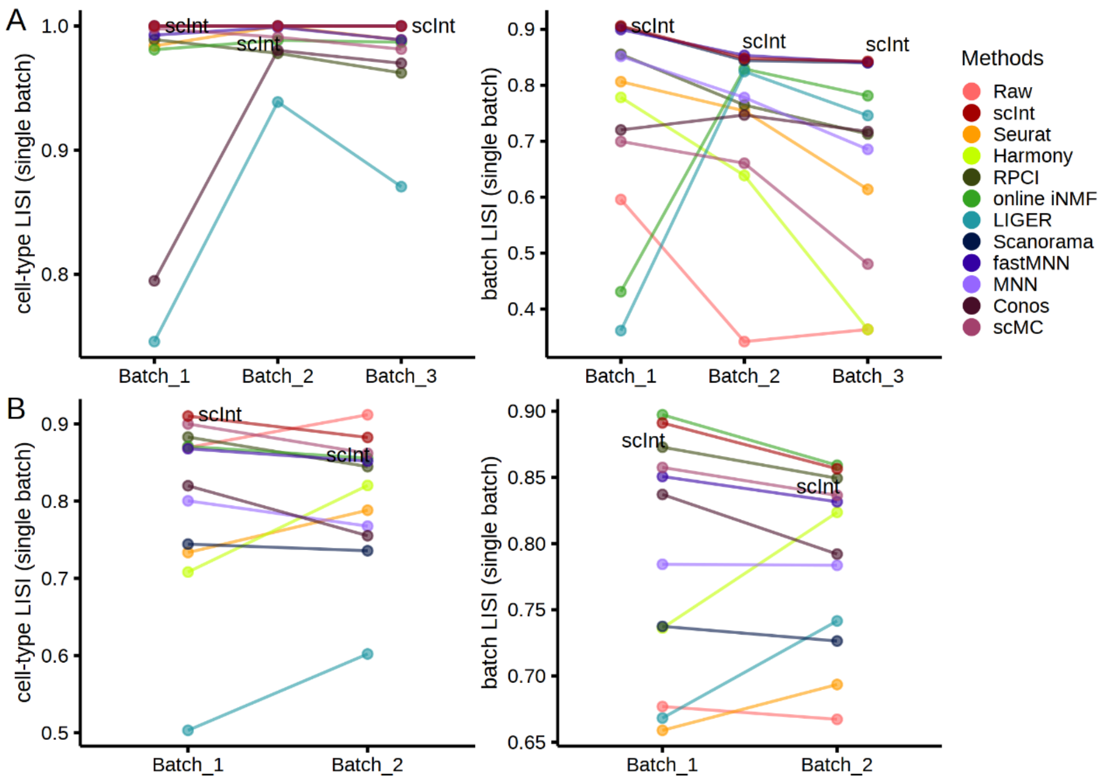
Shuilin Jin: jinsl@hit.edu.cn



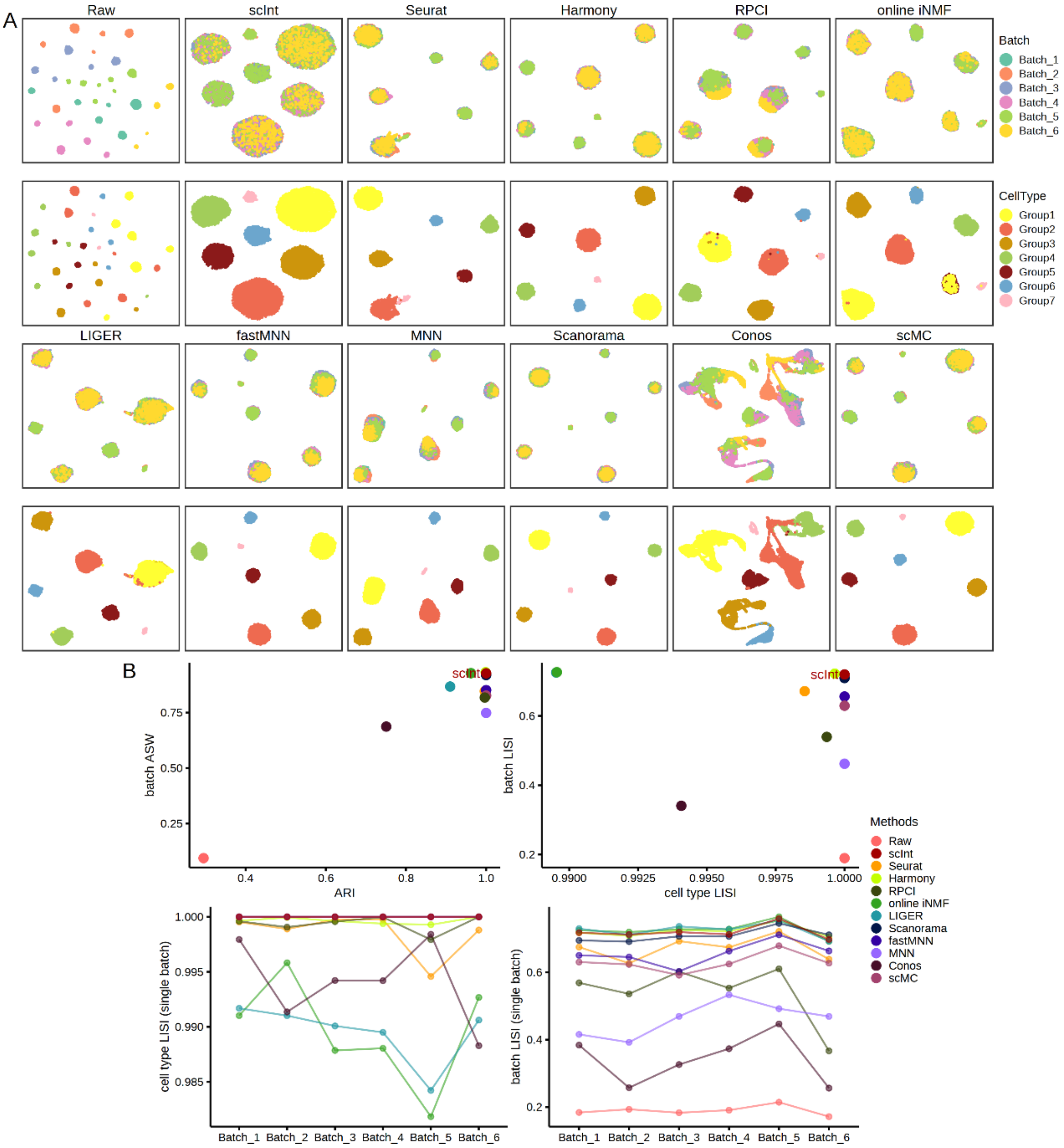
Supplemental Fig. S1. A conceptual schematic of cell-cell similarity relationships identification and filtering in scInt. A. For each batch i , the remaining batches are projected on the same low-dimensional space of i , and the cluster-specific σ_p is identified for each cluster p in batch i . Then the similarity between cell i_s in batch i and j_t in remaining batches is computed using an exponentially measure. The top k similar cells of cell i_s are retained. Further, these cells identified to be similar with the cells in batch i are projected on the cPCA space using batch i . The cosine similarity with the threshold T (default to be 0.6) is used to filter the incorrect-connected cells between batch i and remaining batches. **B**. The final batch-specific identified cell similarity relationships between batch i with its remaining batches.


A
Query
X_{new}
Learned projection matrix
U
B
Query
X_{new}
cPCA projection
$\underset{W}{\operatorname{argmax}} W^T D W - W^T X_{new} X_{new}^T W$
C
Query
X_{new}
Identify new similarity matrix between
query and reference data
newly computed projection matrix
U_{new}

Supplemental Fig. S2. Three kind of reference-based mapping models are available in scInt framework. A. “Projection” model. **B**. “cPCA” model. **C**. “Global” model.

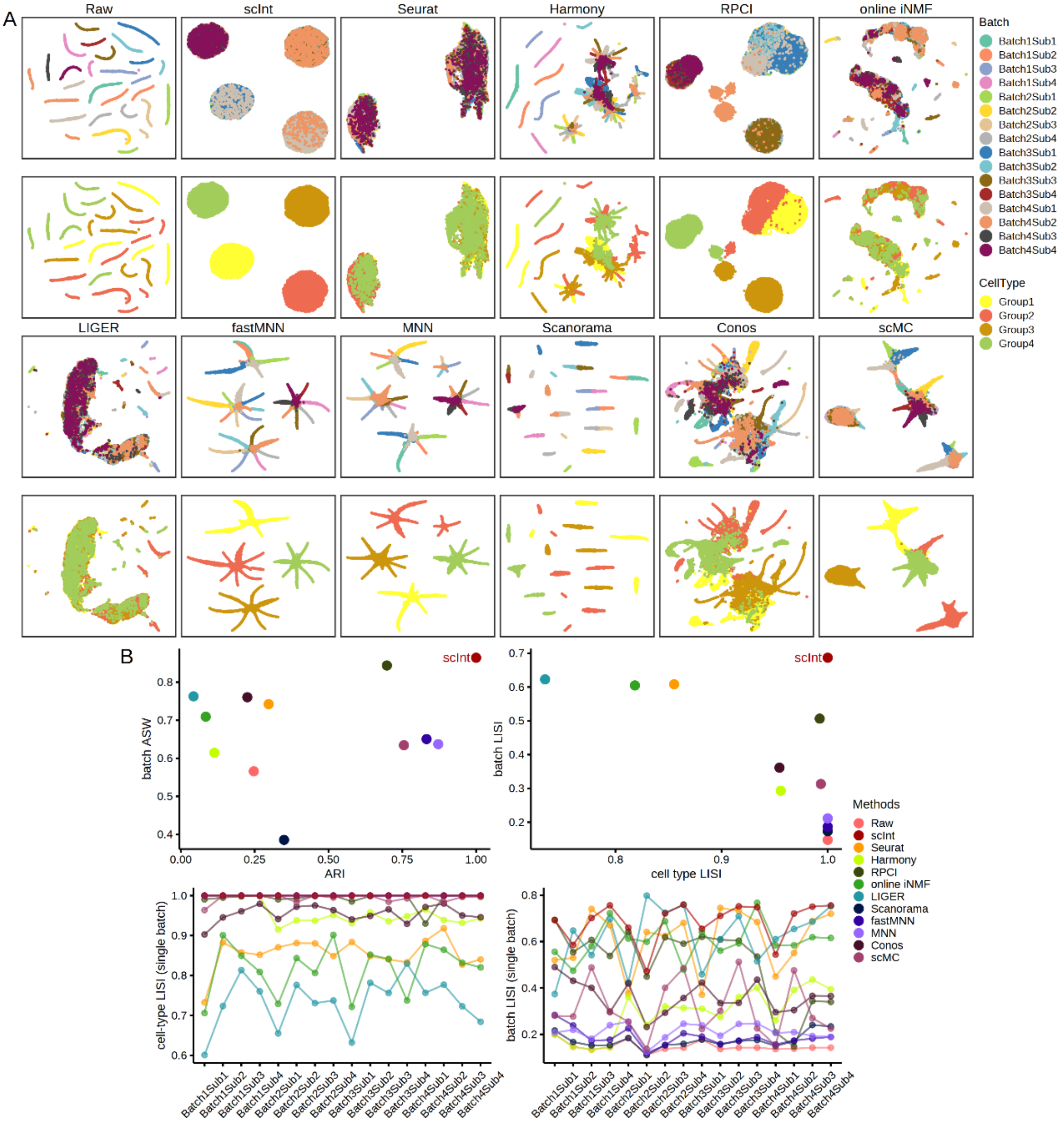


Supplemental Fig. S3. Comparisons of cell type LISI and batch LISI specific to each batch on simulation 1 (A) and human dendritic data (B).



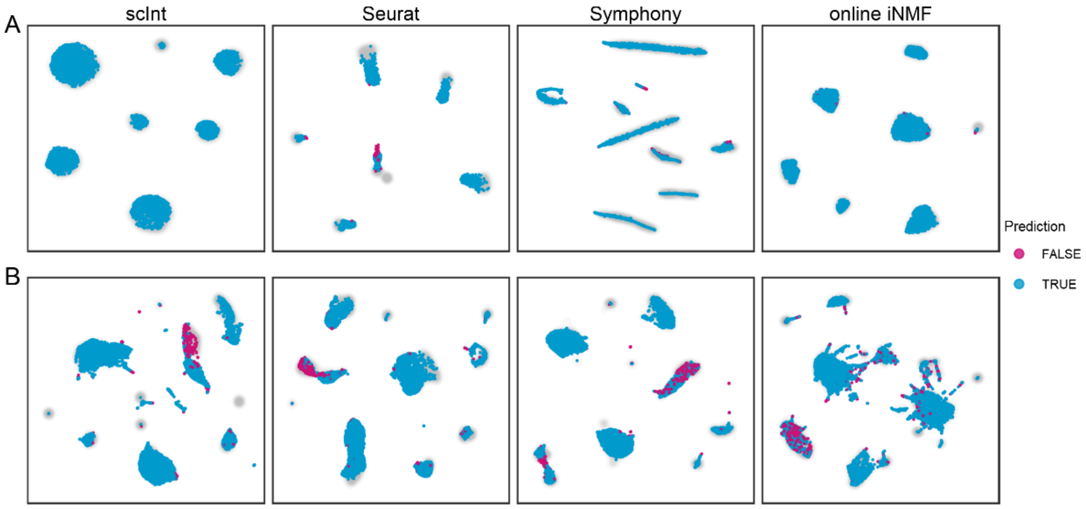
Supplemental Fig. S4. Comparisons of scInt against other methods on simulation 2.

A. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on simulation 2. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows). **B**. The comparisons of the integration results using overall metrics, including ARI, batch ASW, cell type LISI, and batch LISI, and metrics specific to each batch, including cell type LISI (single batch) and batch LISI (single batch).

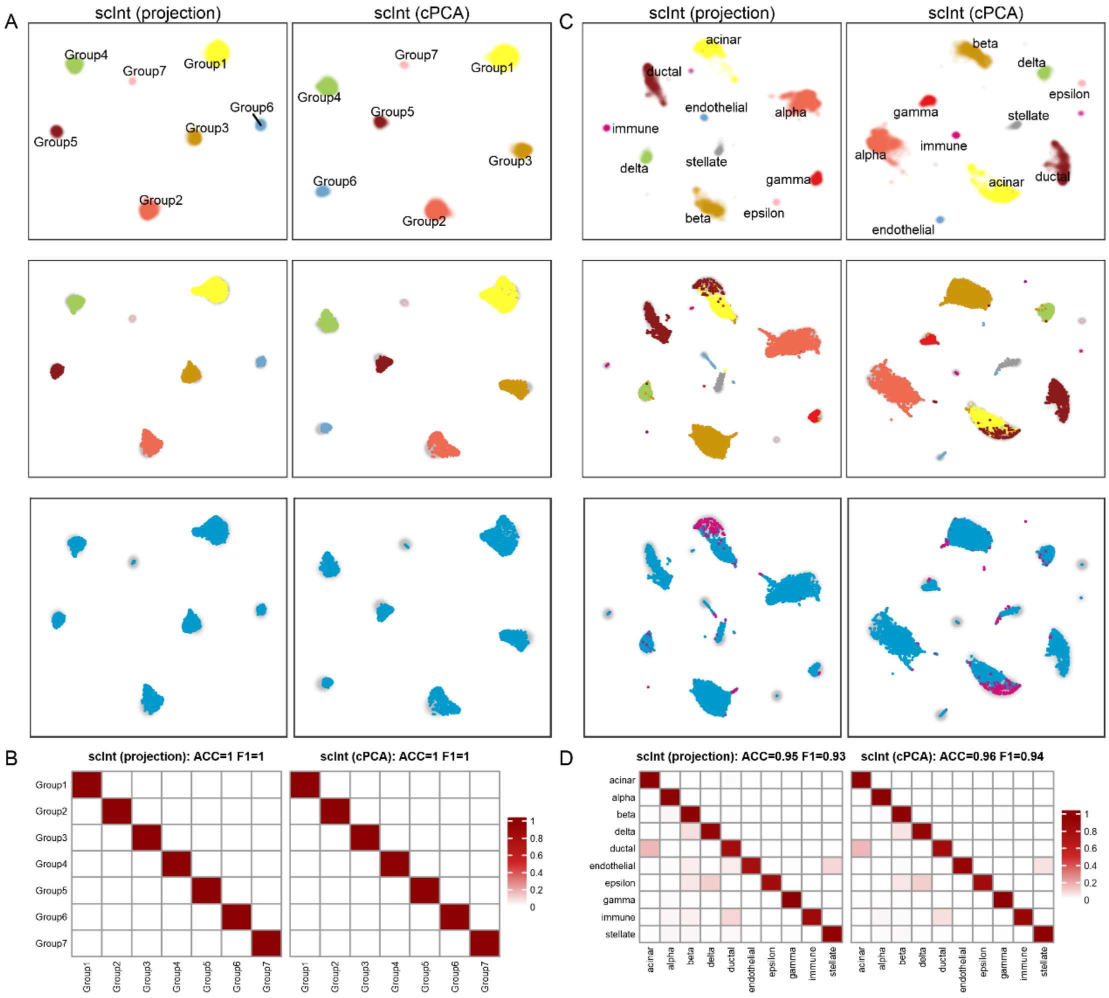


Supplemental Fig. S5. Comparisons of scInt against other methods on simulation 3.

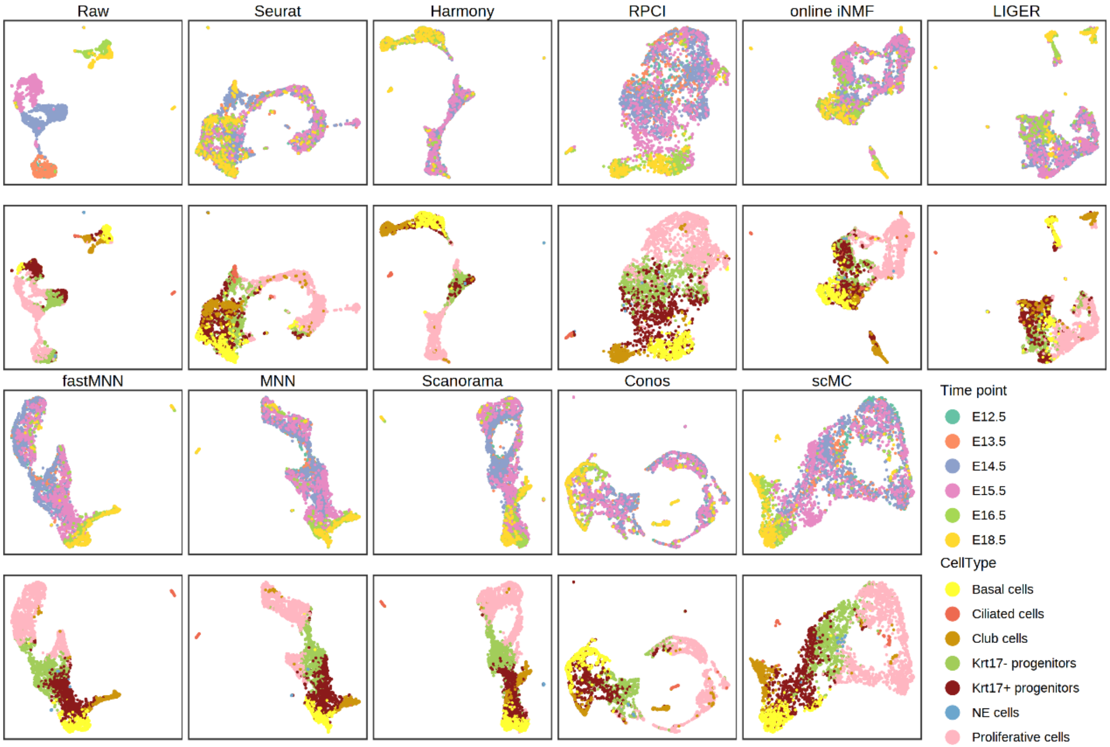
A. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on simulation 3. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows). **B**. The comparisons of the integration results using overall metrics, including ARI, batch ASW, cell type LISI, and batch LISI, and metrics specific to each batch, including cell type LISI (single batch) and batch LISI (single batch).



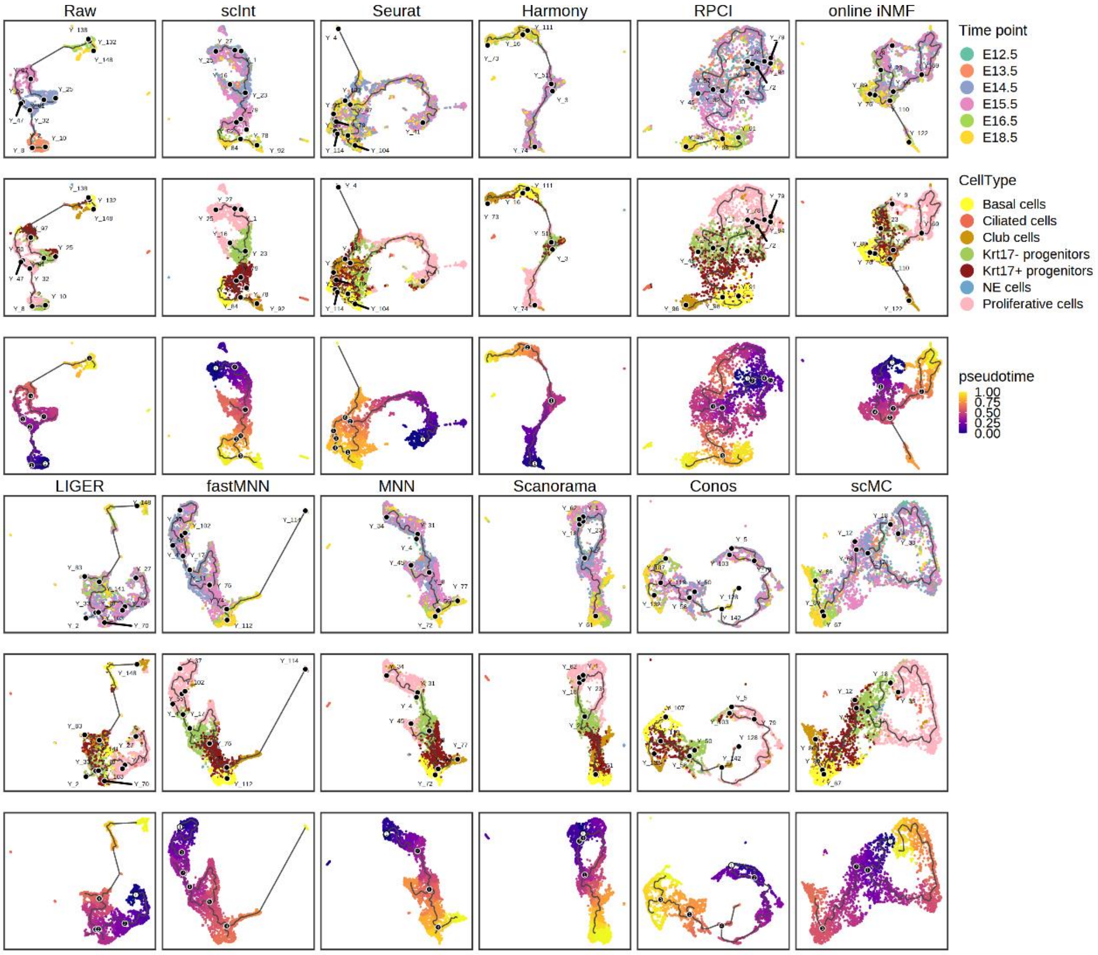
Supplemental Fig. S6. Comparisons of the label prediction accuracy (by 5-NN classifier) of query cells in the reference-based mapping tasks. A, B. UMAP visualizations of the joint low-dimensional embeddings of reference and query for (**A**) simulation 2 and (**B**) human pancreas data. Reference cells are gray, and query cells are colored by incorrect (deep pink) and correct (deep sky blue), which are predicted by the 5-NN classifier in the joint low-dimensional embeddings of reference and query.



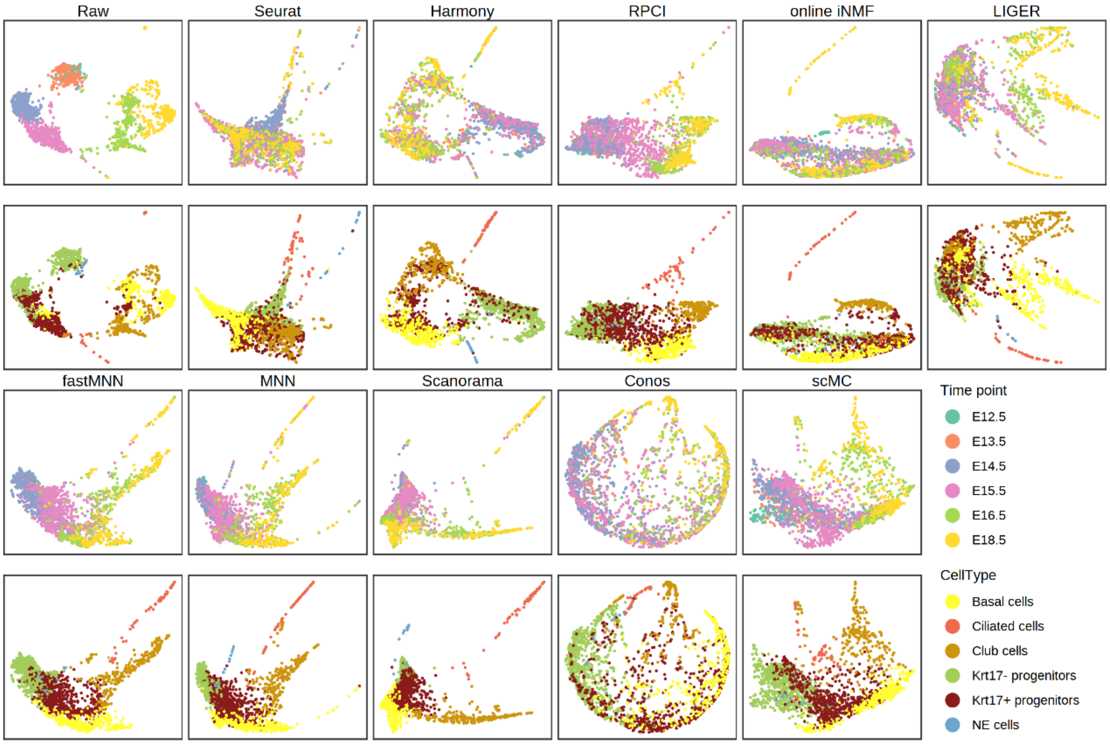
Supplemental Fig. S7. "Projection" and "cPCA" mapping models of scInt in the reference-based mapping tasks. A, C. UMAP visualizations of the density of integrated reference cells (the first row), the scatters of mapped query cells (the second row), and the label prediction accuracy (the third row) of the (**A**) simulation 2 and (**C**) human pancreas data, using "projection" model and "cPCA" model. Cells are colored by ground-truth cell type labels, with gray shadows representing the reference. **B, D.** Heatmap comparing 5-NN predicted labels (columns) and the original labels (rows) of the query. The color bar indicates the proportion of query cells per original cell type label that was predicted to be of each reference label (rows sum to 1).



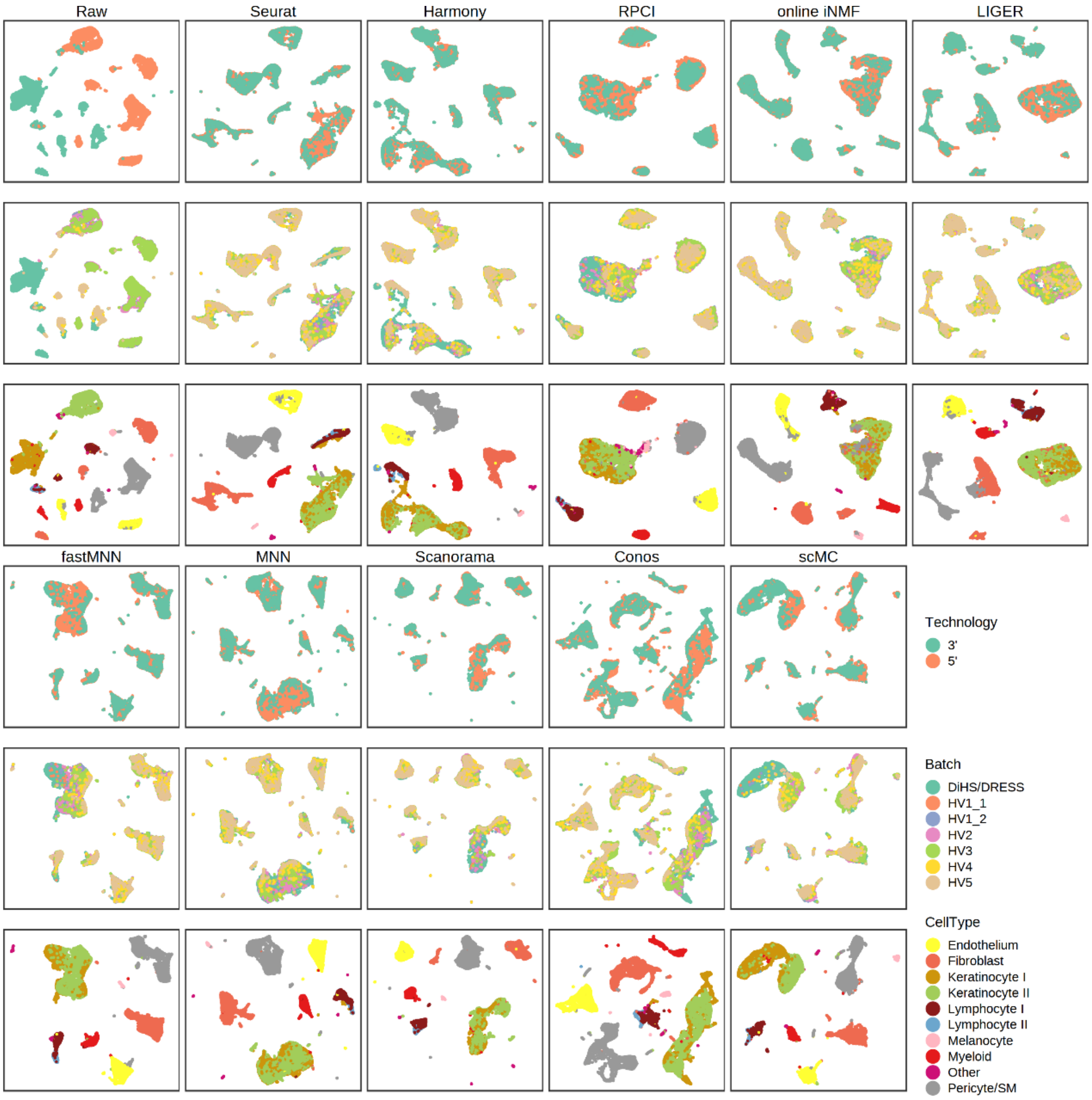
Supplemental Fig. S8. Comparisons of integrated results of other methods on mouse developing tracheal epithelial data. UMAP visualizations of the raw data and corrected data by 10 integration methods on mouse developing tracheal epithelial data. Cells are colored by time points (the first and third rows) and cell type labels (the second and fourth rows).



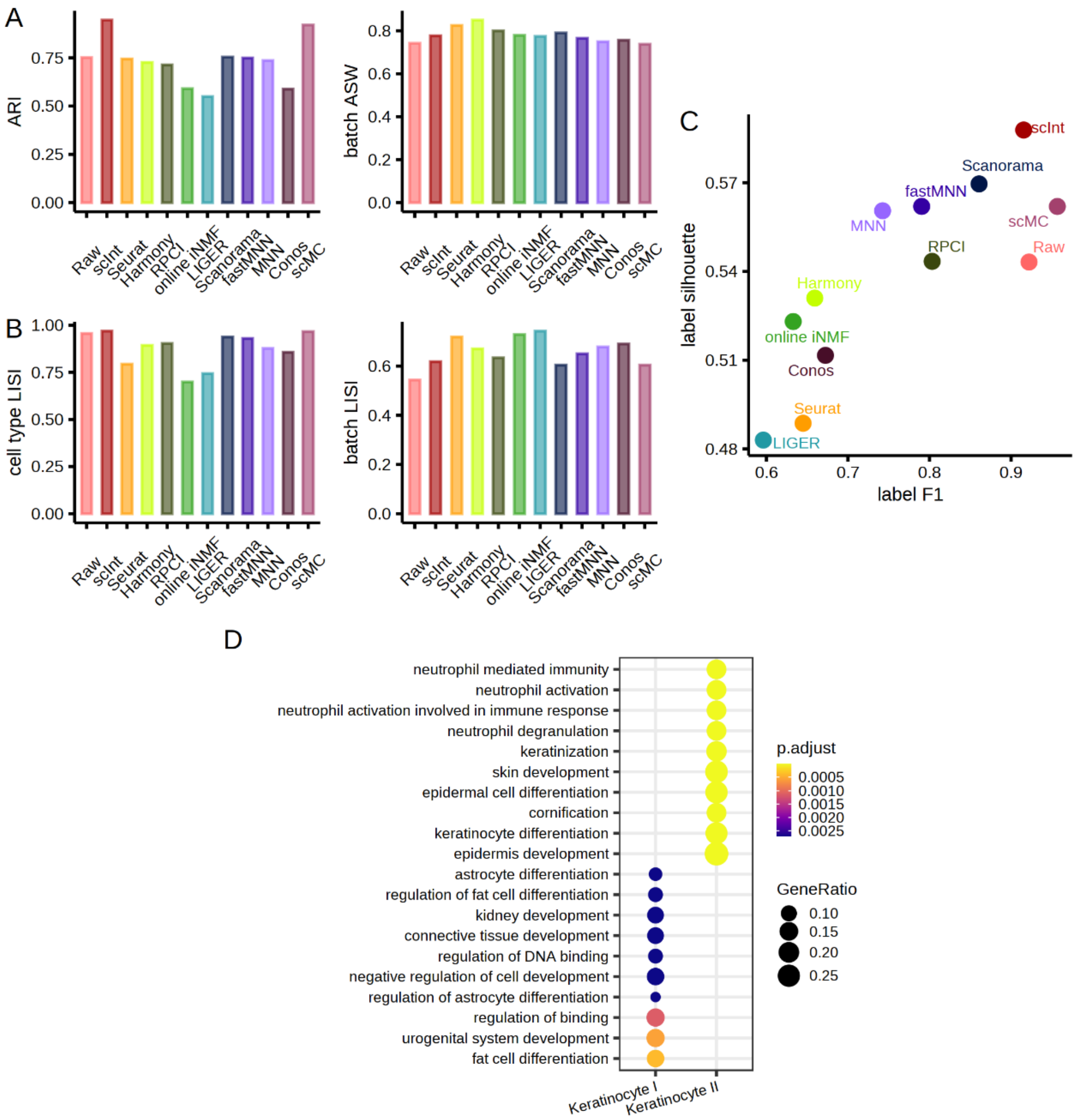
Supplemental Fig. S9. Comparisons of inferred trajectories of integrated results on mouse developing tracheal epithelial data. UMAP visualizations of inferred trajectories of the raw data and integrated data using scInt and other 10 integration methods on mouse developing tracheal epithelial data. Cells are colored by time points (the first and fourth rows), cell type labels (the second and fifth rows), and pseudotime (the third and sixth rows).



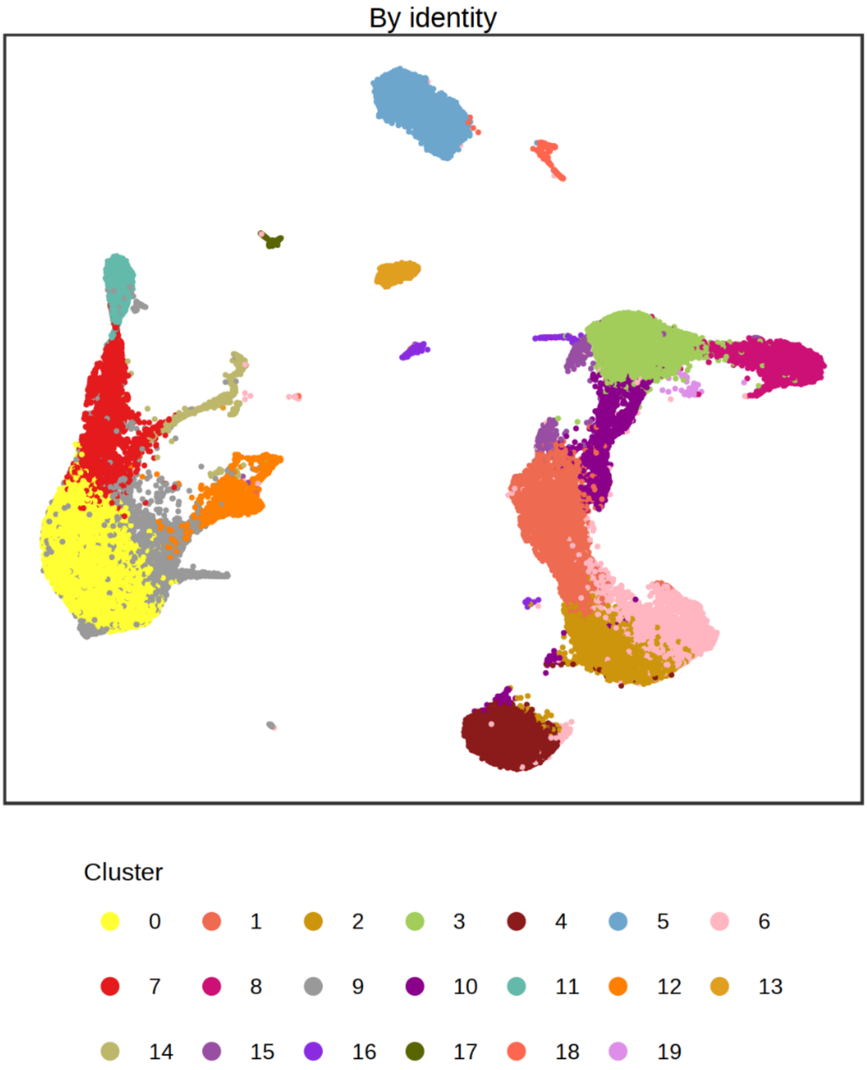
Supplemental Fig. S10. PHATE visualizations of integrated results of other methods on mouse developing tracheal epithelial data. PHATE visualizations of the raw data and corrected data by 10 integration methods on mouse developing tracheal epithelial data except for proliferative cells. Cells are colored by time points (the first and third rows) and cell type labels (the second and fourth rows).



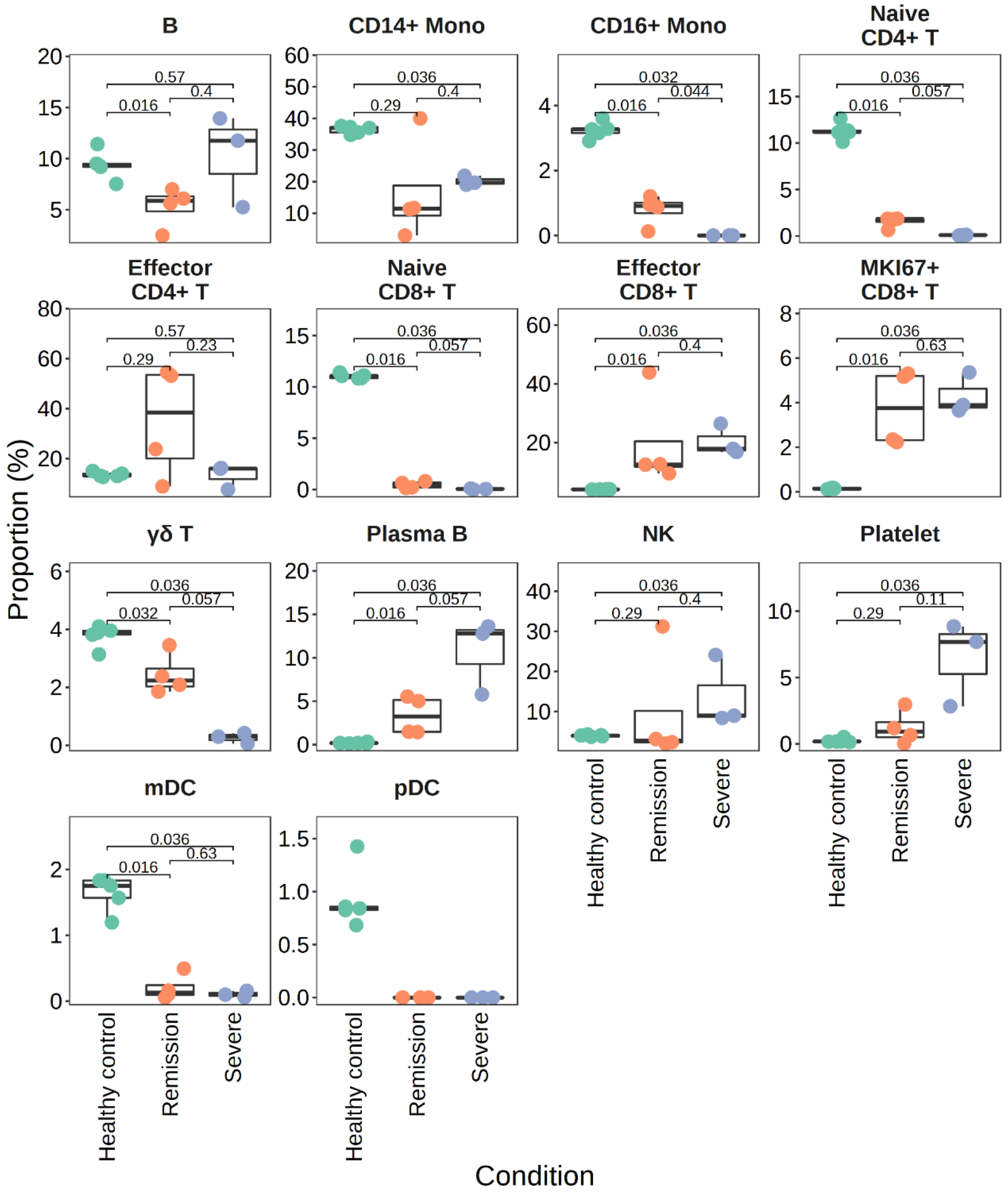
Supplemental Fig. S11. Comparisons of integrated results of other methods on human DiHS/DRESS skin data. UMAP visualizations of the raw data and corrected data by 10 integration methods on human DiHS/DRESS skin data. Cells are colored by technologies (the first and fourth rows), batch labels (the second and fifth rows), and cell type labels (the third and sixth rows).



Supplemental Fig. S12. scInt reveals condition-specific subpopulations in DiHS/DRESS and health human skin. A, B. The comparisons of (**A**) ARI and batch ASW, (**B**) cell type LISI and batch LISI for integrated results. **C.** The comparisons of the label F1 and silhouette for integrated four condition-specific subpopulations. **D.** The top 10 enriched GO biological processes of the marker genes associated with the keratinocyte I and keratinocyte II subpopulations.

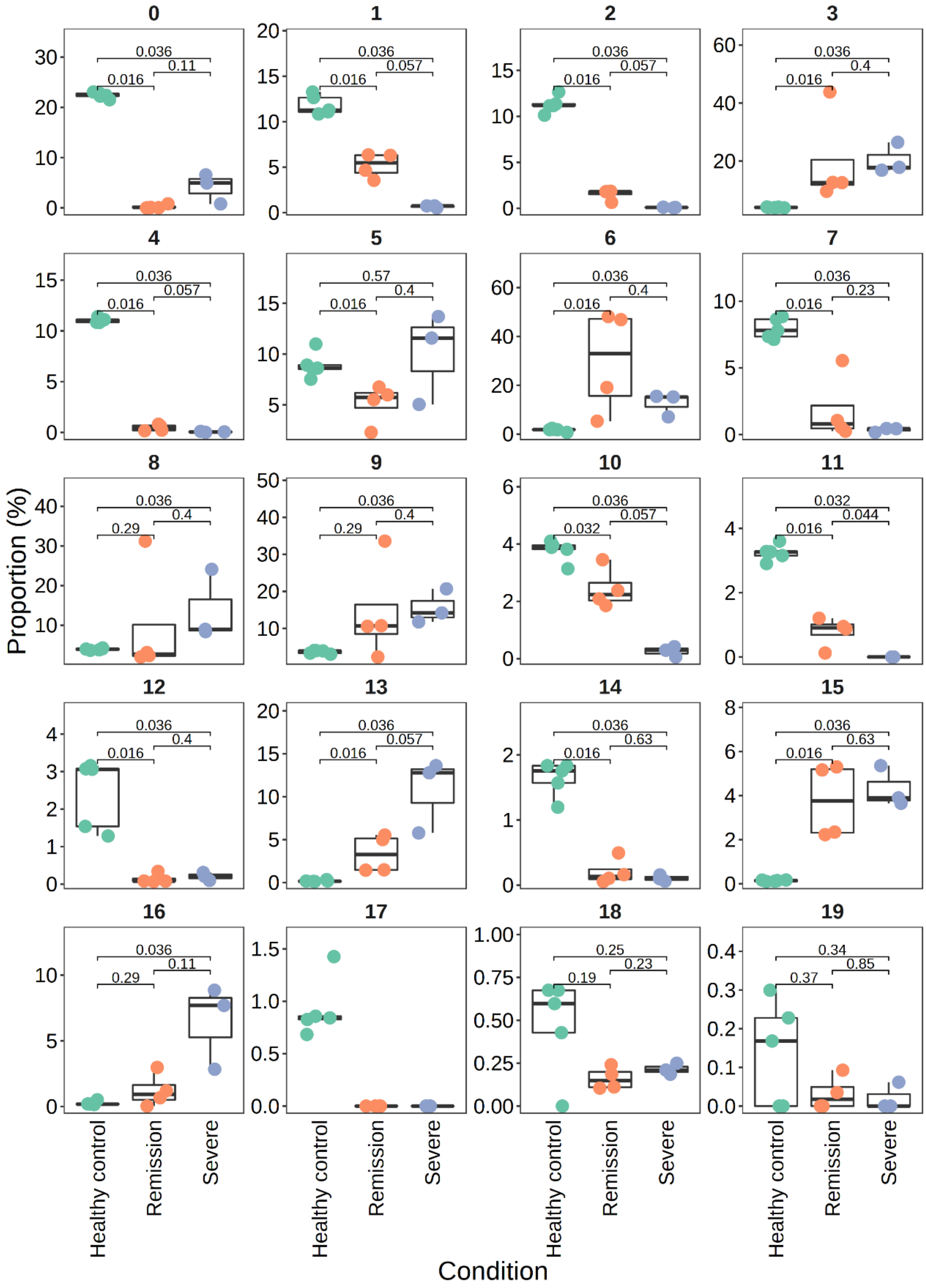


Supplemental Fig. S13. The clustering result of scInt integrated COVID-19 PBMC data.



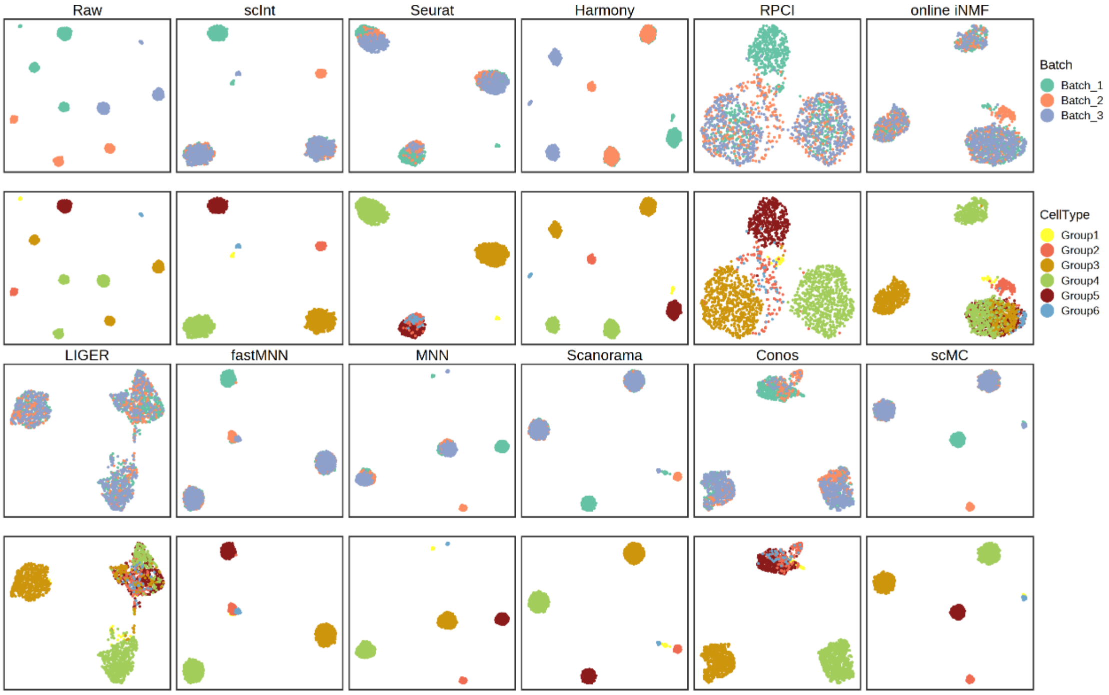
Supplemental Fig. S14. Percentage of cell types identified by scInt in the 12 batches.

Shown are exact two-sided P values by the Wilcoxon rank-sum test.

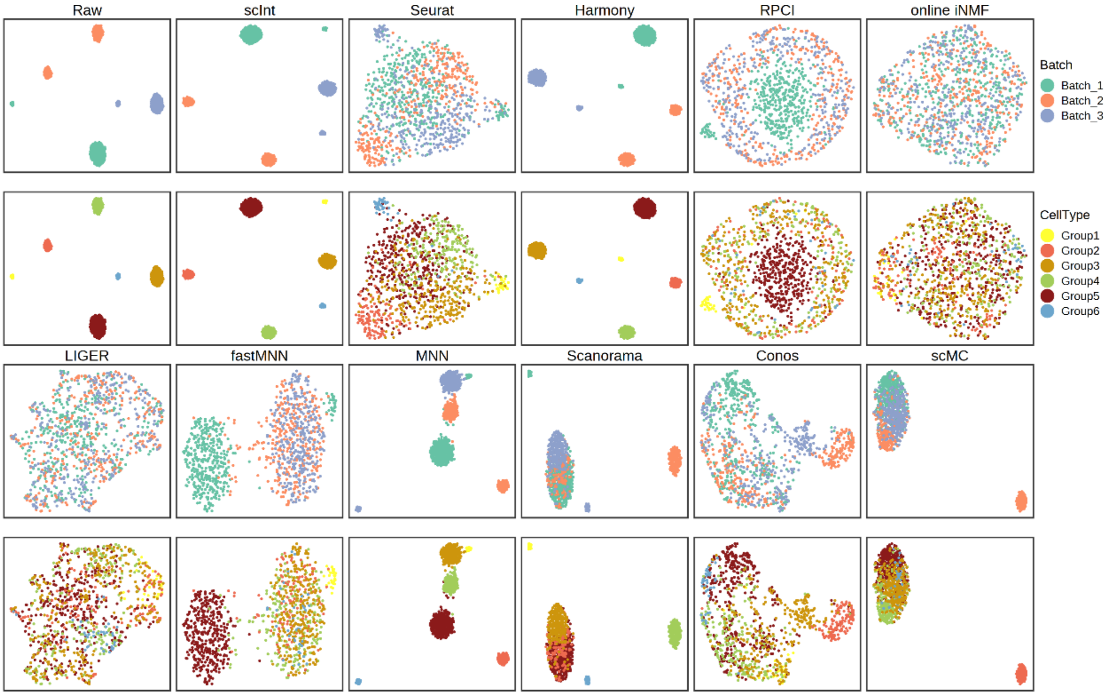


Supplemental Fig. S15. Percentage of clusters identified by scInt in the 12 batches.

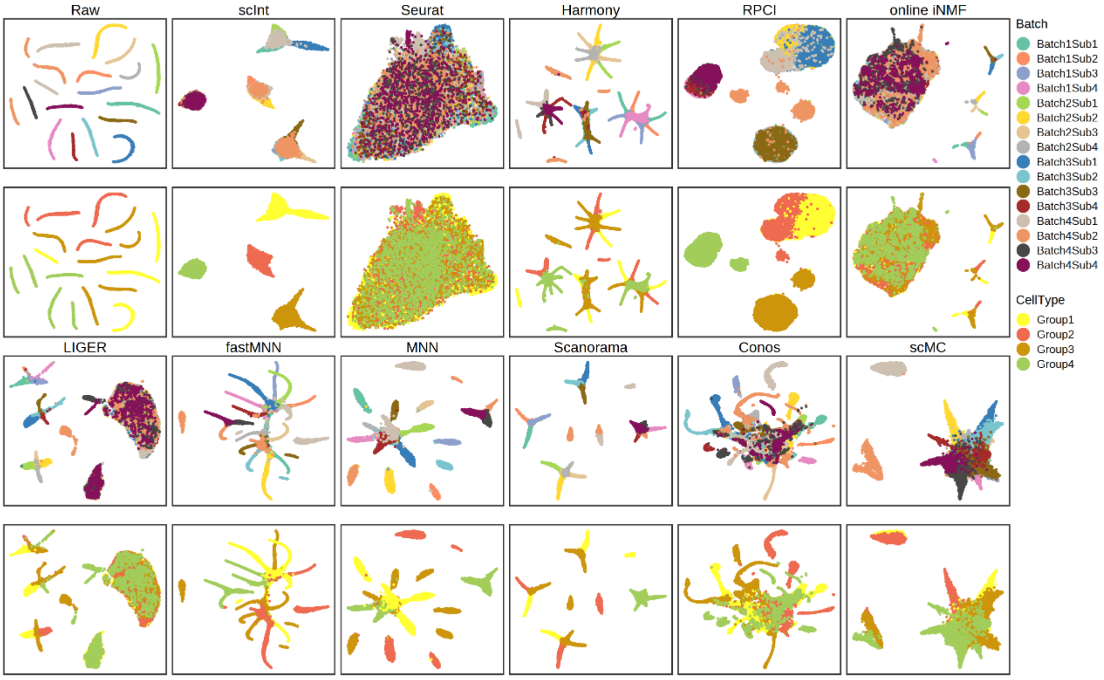
Shown are exact two-sided P values by the Wilcoxon rank-sum test.



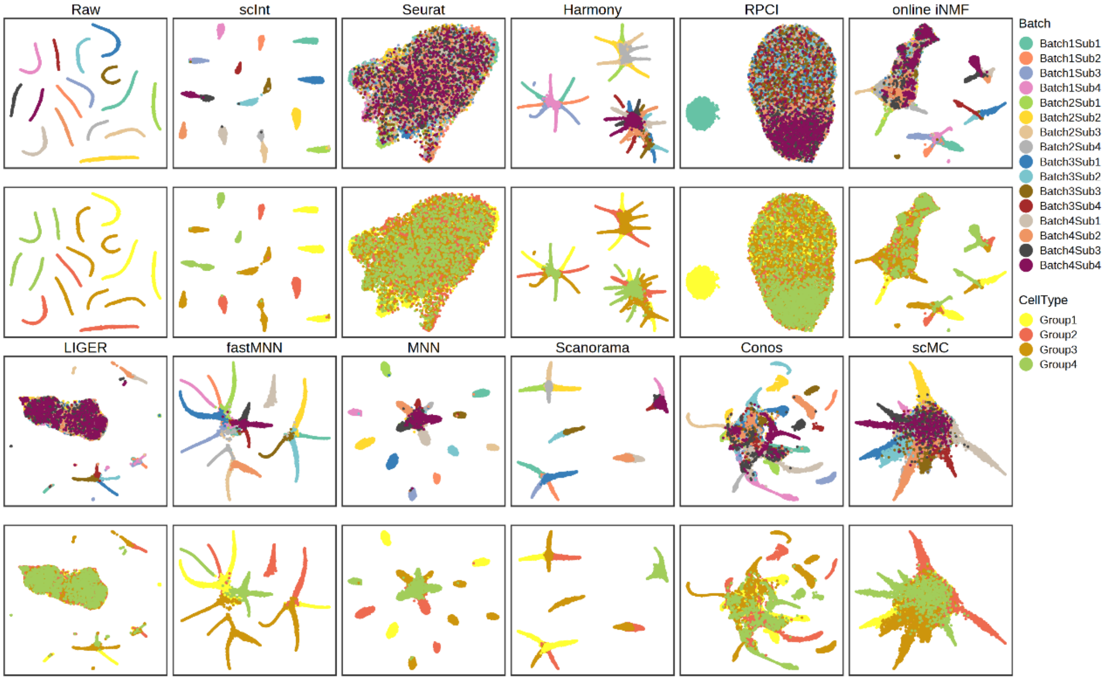
Supplemental Fig. S16. Comparisons of scInt against other methods on variant 1 of simulation 1. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 1 of simulation 1. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



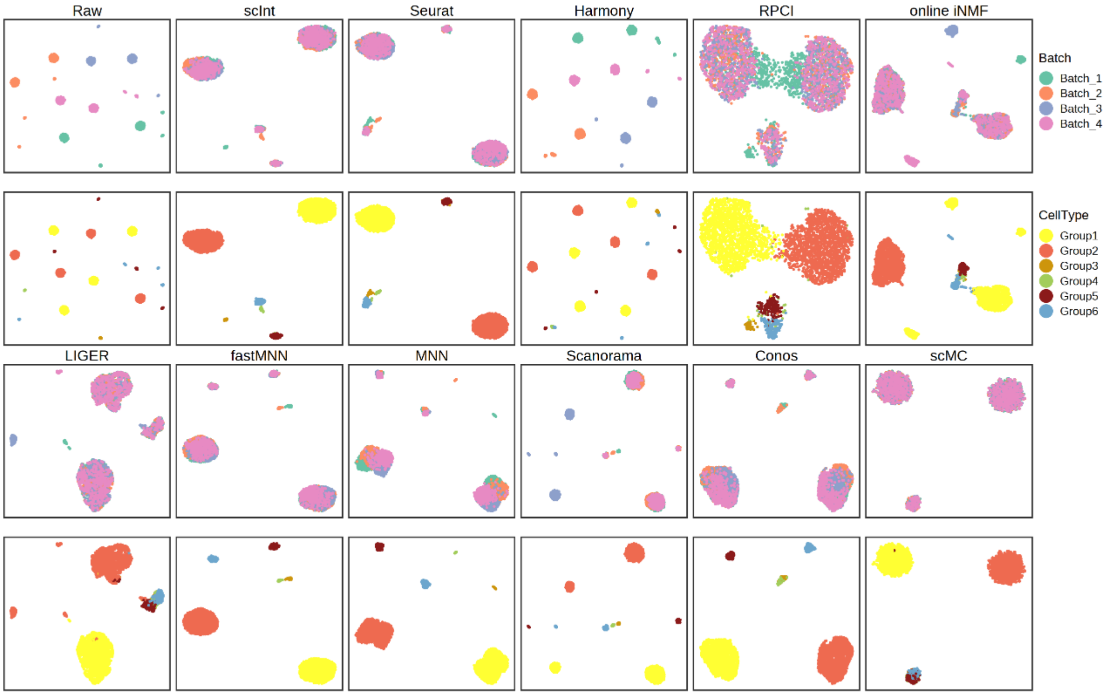
Supplemental Fig. S17. Comparisons of scInt against other methods on variant 2 of simulation 1. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 2 of simulation 1. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



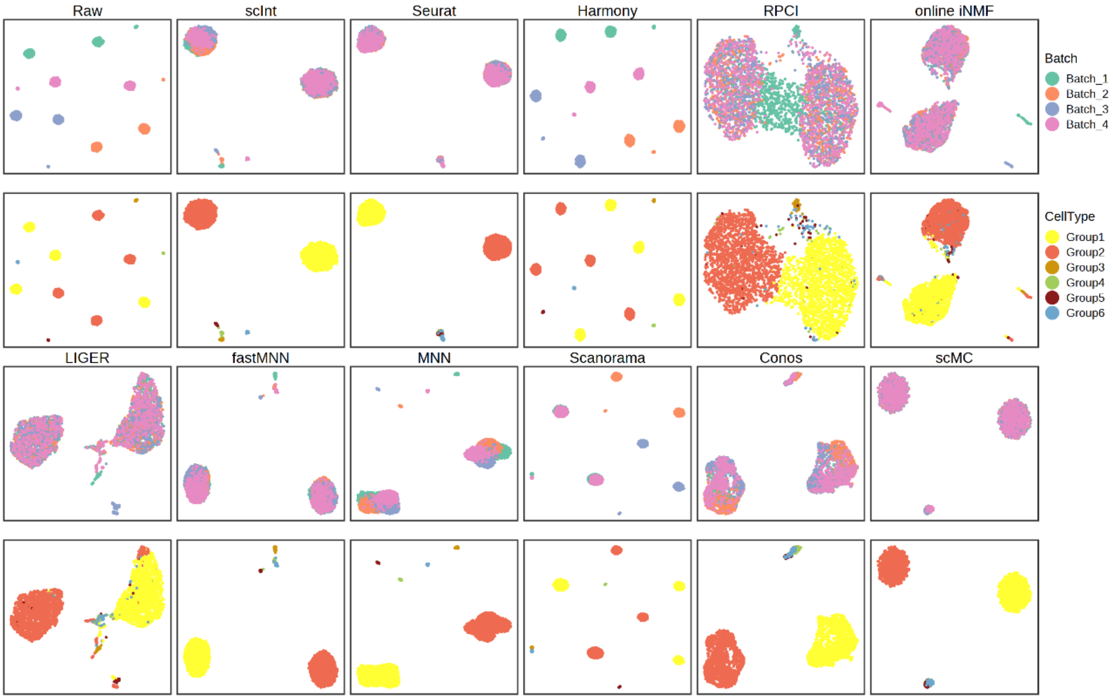
Supplemental Fig. S18. Comparisons of scInt against other methods on variant 1 of simulation 3. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 1 of simulation 3. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



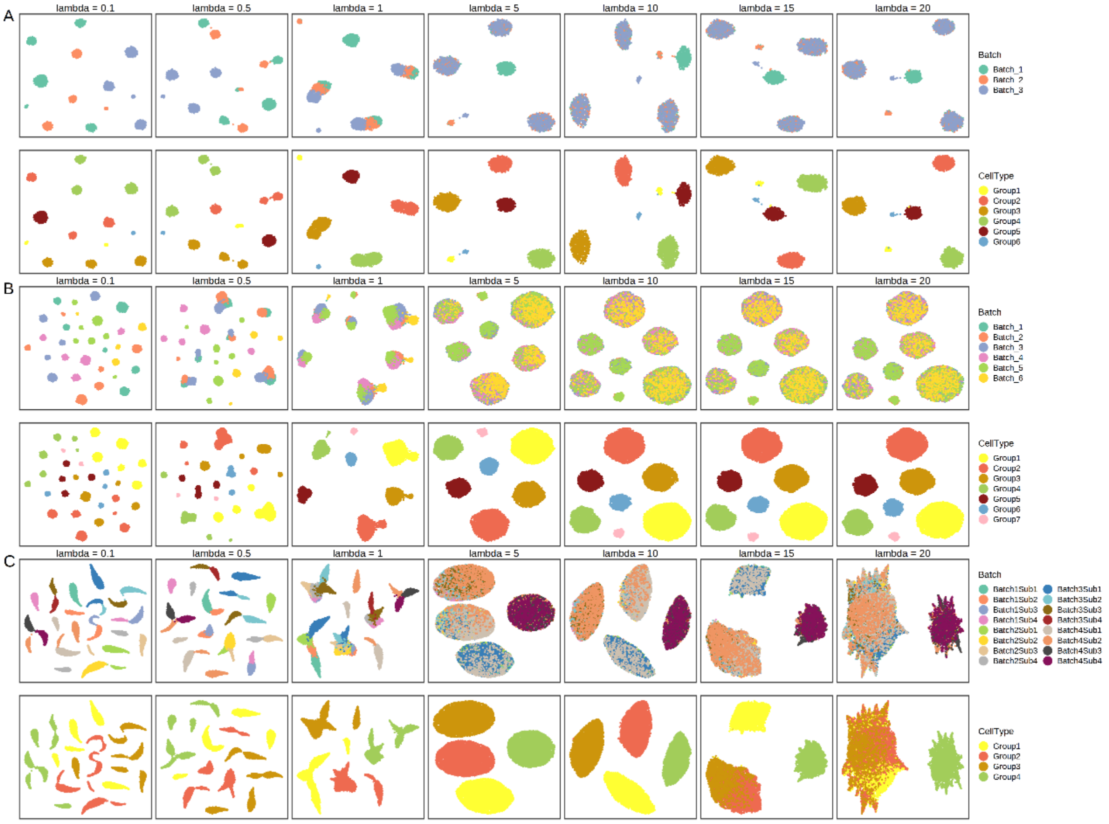
Supplemental Fig. S19. Comparisons of scInt against other methods on variant 2 of simulation 3. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 2 of simulation 3. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



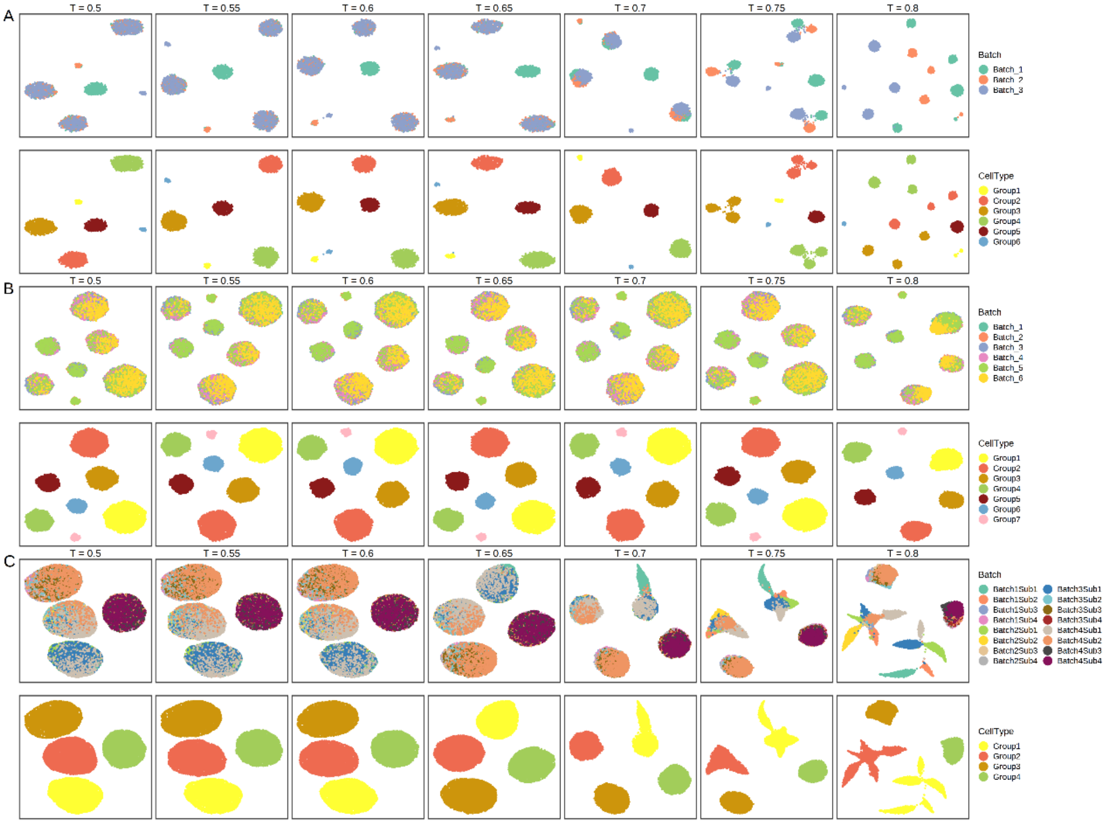
Supplemental Fig. S20. Comparisons of scInt against other methods on variant 1 of simulation 4. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 1 of simulation 4. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



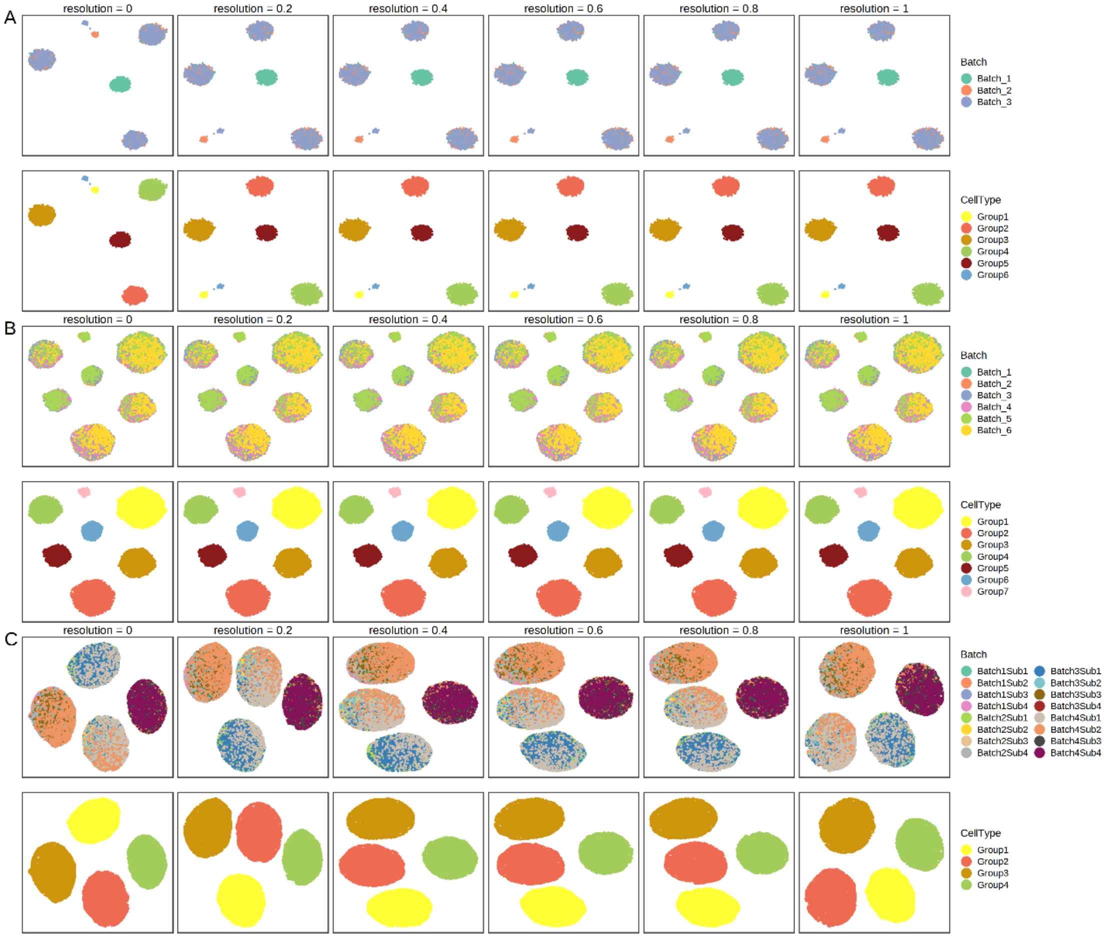
Supplemental Fig. S21. Comparisons of scInt against other methods on variant 2 of simulation 4. UMAP visualizations of the raw data and integrated data using scInt and other 10 integration methods on variant 2 of simulation 4. Cells are colored by batch labels (the first and third rows) and cell type labels (the second and fourth rows).



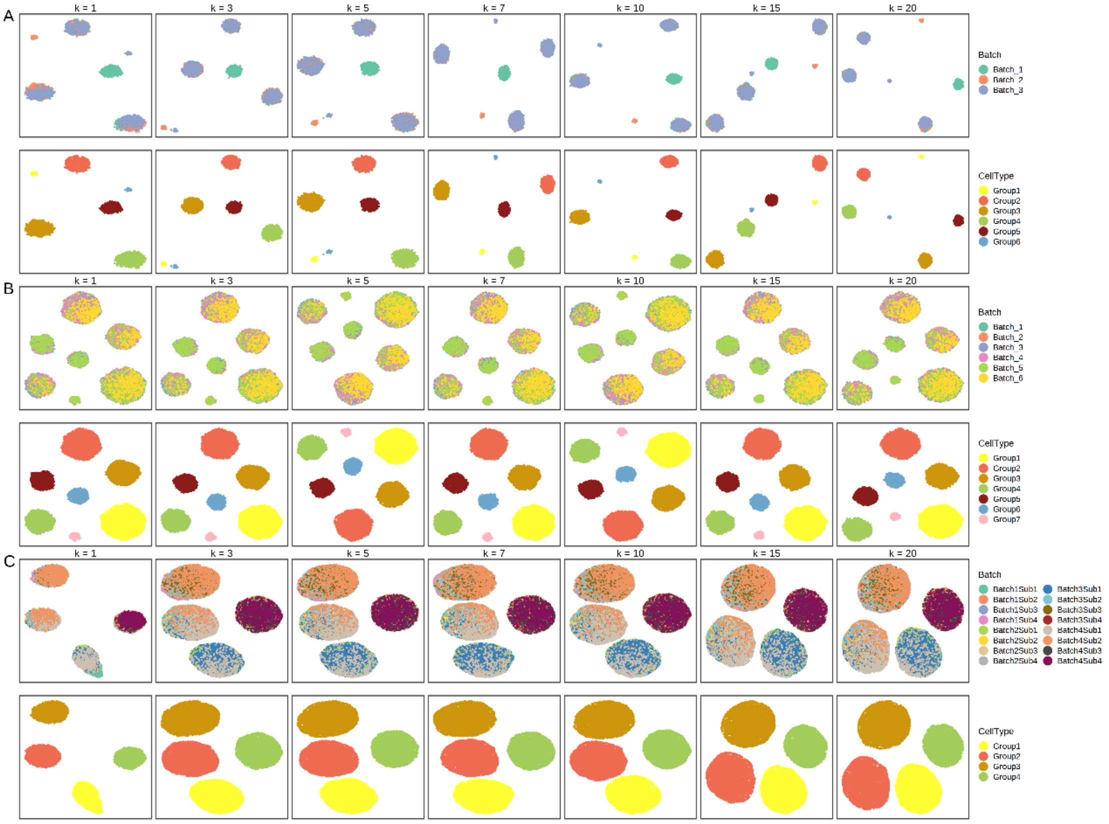
Supplemental Fig. S22. The performance of scInt with the varied λ on simulations 1-3. A. UMAP visualizations of the corrected data of simulation 1 using scInt with λ varied 0.1 to 20. Cells are colored by batch label (top row) and cell type labels (bottom row). **B.** UMAP visualizations of the corrected data of simulation 2 using scInt with λ varied 0.1 to 20. **C.** UMAP visualizations of the corrected data of simulation 3 using scInt with λ varied 0.1 to 20.



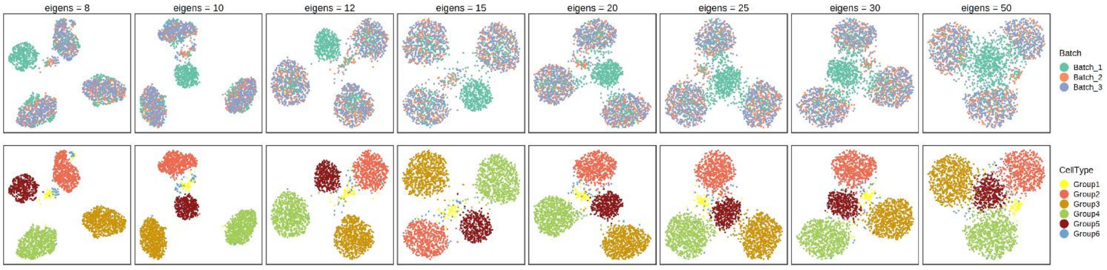
Supplemental Fig. S23. The performance of scInt with the varied *T* on simulations 1-3. A. UMAP visualizations of the corrected data of simulation 1 using scInt with *T* varied 0.5 to 0.8. Cells are colored by batch label (top row) and cell type labels (bottom row). **B.** UMAP visualizations of the corrected data of simulation 2 using scInt with *T* varied 0.5 to 0.8. **C.** UMAP visualizations of the corrected data of simulation 3 using scInt with *T* varied 0.5 to 0.8.



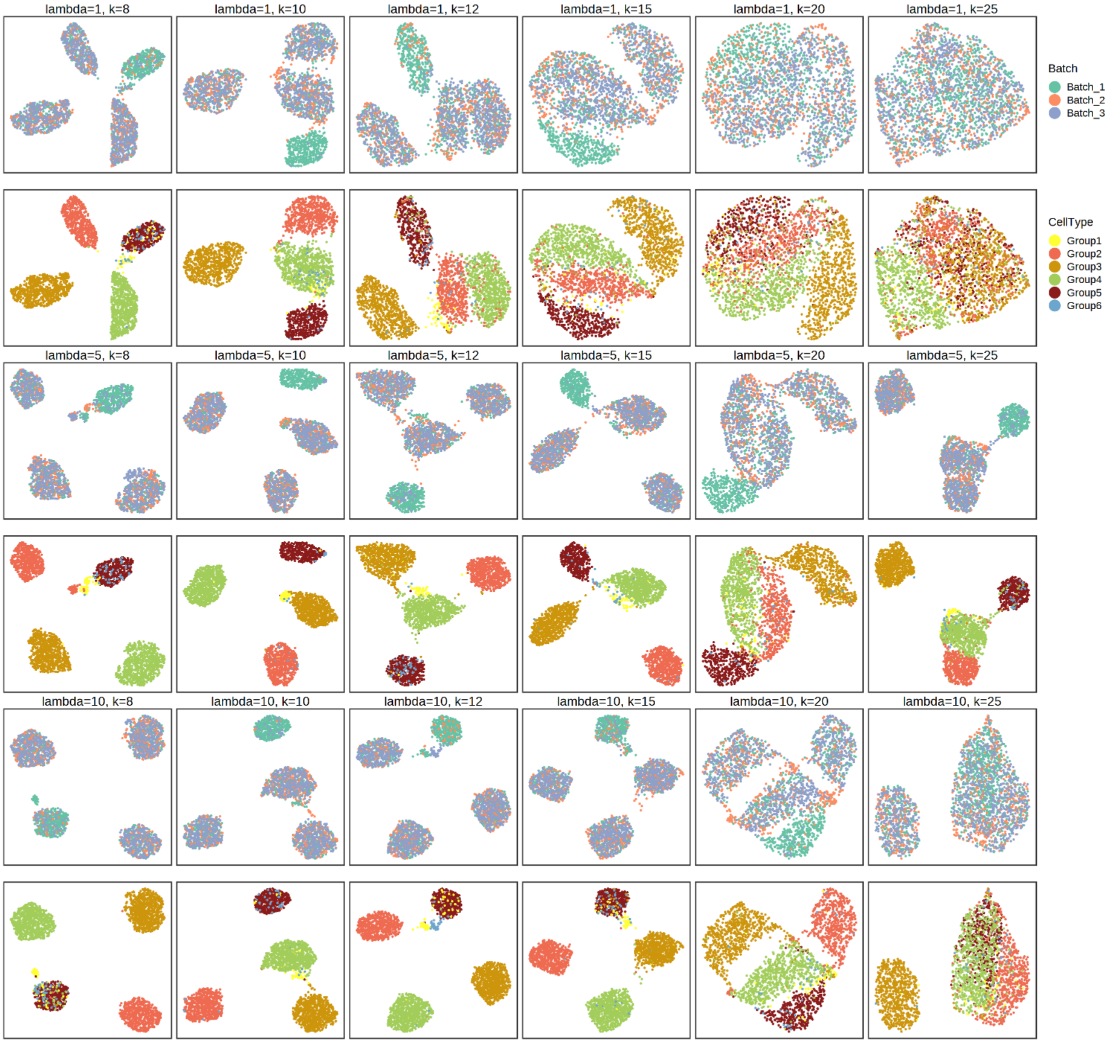
Supplemental Fig. S24. The performance of scInt with the varied resolution on simulations 1-3. A. UMAP visualizations of the corrected data of simulation 1 using scInt with resolution varied 0 to 1. Cells are colored by batch label (top row) and cell type labels (bottom row). **B.** UMAP visualizations of the corrected data of simulation 2 using scInt with resolution varied 0 to 1. **C.** UMAP visualizations of the corrected data of simulation 3 using scInt with resolution varied 0 to 1.



Supplemental Fig. S25. The performance of scInt with the varied *k* on simulations 1-3. A. UMAP visualizations of the corrected data of simulation 1 using scInt with *k* varied 1 to 20. Cells are colored by batch label (top row) and cell type labels (bottom row). **B.** UMAP visualizations of the corrected data of simulation 2 using scInt with *k* varied 1 to 20. **C.** UMAP visualizations of the corrected data of simulation 3 using scInt with *k* varied 1 to 20.



Supplemental Fig. S26. UMAP visualizations of tuning results of the "eigens" parameter of RPCI on simulation 1.



Supplemental Fig. S27. UMAP visualizations of tuning results of k and lambda of online iNMF on simulation 1.

Supplemental Table 1: Details of scRNA-seq data sets used in scInt manuscript.

Data set	Batch number	Total cell number	Scenario	Source	Batch name	Cell number
Simulation 1	3	2,401	Unbalanced cell subpopulation compositions & rare cell subpopulation across batches	Generated by Splatter package (see Supplemental Methods)	Batch_1	978
					Batch_2	599
					Batch_3	824
Simulation 2	6	12,097	Unbalanced cell subpopulation compositions & reference-based mapping	https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_datasets_Immune_and_pancreas_/12420968	Batch_1	2908
					Batch_2	2422
					Batch_3	2120
					Batch_4	1929
					Batch_5	1761
					Batch_6	957
Simulation 3	16	19,318	Unbalanced cell subpopulation compositions & nested batch effects	https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_datasets_Immune_and_pancreas_/12420968	Batch1Sub1	1200
					Batch1Sub2	1204
					Batch1Sub3	1210
					Batch1Sub4	1192
					Batch2Sub1	1466
					Batch2Sub2	1471
					Batch2Sub3	984
					Batch2Sub4	984
					Batch3Sub1	1930
					Batch3Sub2	1205
					Batch3Sub3	1189
					Batch3Sub4	485
					Batch4Sub1	1687
					Batch4Sub2	1676
					Batch4Sub3	716
					Batch4Sub4	719
Simulation 4	4	4000	Rare cell subpopulations	Generated by Splatter package (see Supplemental Methods)	Batch_1	1000
					Batch_2	1000
					Batch_3	1000
					Batch_4	1000
Human dendritic cells	2	576	Unbalanced cell subpopulation compositions & biologically similar cell types	Gene Expression Omnibus database	Batch_1	288
					Batch_2	288

				accession code: GSE94820		
Human pancreas cells	8	14,456	Different sequencing platforms & reference-based mapping	SeruatData R package (“panc8” data)	c1	625
					celseq	946
					celseq2	2238
					smartseq	2078
					indrop 1	1937
					indrop 2	1724
					indrop 3	3605
					indrop 4	1303
mouse developing tracheal epithelial data	6	3,508	Developmental data	Gene Expression Omnibus database accession code: GSE152692	E12.5	239
					E13.5	455
					E14.5	1232
					E15.5	931
					E16.5	365
					E18.5	286
Human DiHS/DRESS skin data	7	20,415	Disease and healthy conditions & different sequencing platforms	Gene Expression Omnibus database accession code: GSE132802	DiHS/DRESS	5653
					HV1_1	1152
					HV1_2	1917
					HV2	4016
					HV3	4923
					HV4	1589
					HV5	1165
Human COVID-19 and healthy PBMCs data	12	67,362	Disease and healthy conditions & complex batch correction	Gene Expression Omnibus database accession code: GSE150861; https://support.10xgenomics.com/single-cell-gene-expression/datasets	10k_1	11485
					10k_2	11996
					20k	23837
					500_1	705
					500_2	587
					P1-day1-rep1	2591
					P1-day1-rep2	2081
					P1-day5-rep1	3129
					P1-day5-rep2	3851
					P2-day1	2280
					P2-day5	1488
					P2-day7	3332

Supplemental Methods

Details of Simulated data sets

Simulation 1. Simulation 1 contains three batches generated by Splatter package. Three batches with six cell groups were generated with parameters batchCells = (1000, 1000, 1000), and group.prob = (0.03, 0.15, 0.2, 0.2, 0.4, 0.02). Then the Group6 in batch 1, Group5 and Group6 in batch 2, Group1 and Group5 in batch 3 were removed.

Simulation 2 and Simulation 3. These two simulation data sets were downloaded from https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_datasets_Immune_and_pancreas_/12420968. Detailed information can be found in a recent benchmark study (Luecken et al. 2022).

Simulation 4. Simulation 4 contains four batches generated by Splatter package. Four batches with six cell groups were generated with parameters batchCells = (1000, 1000, 1000, 1000), and group.prob = (0.42, 0.42, 0.04, 0.04, 0.04, 0.04).

Pathological cases

Variants of simulation 1. Simulation 1 contains three batches and six cell groups. Group5 and Group6 of these six cell groups are specific to batches 1 and 3, respectively. We removed Group1 in batch 2 and Group2 in batches 1 and 3 from simulation 1 to get variant 1. Then, we removed Group3 in batches 1 and 2 and Group4 in batches 1 and 3 from variant 1 to get variant 2. Thus, variants 1 and 2 contain four and six cell groups that are only present in a single batch, respectively.

Variants of simulation 3. Simulation 3 contains 16 batches, 11 of which have more than one cell group. For these 11 batches, we removed all cell groups except the dominant cell group from a batch at a time. We repeated this step 11 times and referred to the ninth and last generated data sets as variants 1 and 2 of simulation 3, respectively. Thus, variants 1 and 2 contain 14 and 16 batches, respectively, each of which has only one cell group.

Variants of simulation 4. Simulation 4 contains four batches and four cell groups at low proportions (Group3, Group4, Group5, and Group6). We removed Group4, Group5, and Group6 from batch 1, and Group3, Group5, and Group6 from batch 2 of simulation 1 to get variant 1. Then, we removed Group3, Group4, and Group6 from batch 3, and Group3, Group4, and Group5 from batch 4 of variant 1 to get variant 2. Thus, variants 1 and 2 contain two and four rare cell groups, respectively, with each rare cell group only present in a single batch.

cPCA can remove the technical effects between data sets

Here, we show that the technical effects between data sets can be removed by cPCA, giving the theoretical reasonability of the "cell similarities filtering on the cPCA space" step in the scInt model.

Given two scRNA-seq data sets $X \in \mathbb{R}^{p\times n_1}$ and $Y \in \mathbb{R}^{p\times n_2}$, with X as target data and Y as background data. Then, they can be formulated as:

$$\begin{aligned} X &= WZ_X + W_X U_X + \varepsilon_X \\ Y &= WZ_Y + W_Y U_Y + \varepsilon_Y \end{aligned}$$

where $W \in \mathbb{R}^{p\times d}$ represents the shared effects between the target and background data, $W_X \in \mathbb{R}^{p\times d_X}$ represents the specific effects of the target data relative to the background data, and $W_Y \in \mathbb{R}^{p\times d_Y}$ represents the specific effects of the background data relative to the target data. It assumes that the target and background data follow the Gaussian distribution, i.e.,

$Z, Z', U, V \sim_{i.i.d.} \mathrm{N}(0, I)$, $\varepsilon_X, \varepsilon_Y \sim_{i.i.d.} \mathrm{N}(0, \sigma^2 I)$. Without loss of generality, let $span(W \cup W_X) \cap span(W_Y) = \varnothing$. Because if $span(W \cup W_X) \cap span(W_Y) \neq \varnothing$, then the above equation can be rewritten as:

$$\begin{aligned} X &= WZ + W_{W_X \cap W_Y} U + W_{W_X \setminus W_Y} U + \varepsilon_X \\ Y &= WZ' + W_{W_Y \cap (W \cup W_X)} V + W_{W_Y \setminus (W \cup W_X)} V + \varepsilon_Y \end{aligned}$$

Then the covariance matrices of the target and background data can be written as:

$$\begin{aligned} C_X &= XX^T = WW^T + W_X W_X^T + \sigma^2 I \\ C_Y &= YY^T = WW^T + W_Y W_Y^T + \sigma^2 I \end{aligned}$$

$$C_X - C_Y = W_X W_X^T - W_Y W_Y^T$$

Consider the optimization problem:

$$\begin{aligned} &\mathrm{argmax}_{v \in \mathbb{R}^p_{unit}} \; v^T (C_X - C_Y) v = v^T W_X W_X^T v - v^T W_Y W_Y^T v \\ &s.t. \; \|v\|_2^2 \leq 1 \end{aligned}$$

The obtained contrastive principal components v_i 's are the directions of maximization of target-specific variations $W_X W_X^T$ meanwhile minimization of background-specific variations $W_Y W_Y^T$. By our assumption, $W_Y W_Y^T$ can be the background-specific variation and background-to-target technical variation. Thus, the common dimensional reduction of the target and background data using the matrix $V \in \mathbb{R}^{p \times l}$ of first l contrastive principal component, $L = V^T [X, Y]$, remove the technical variations between X and Y .

Selection of scInt parameters

To select the optimal key parameters from sets of candidates, we propose the following methods.

For T , we expect 1) it big enough to preserve high reliability of retained similarities; 2) retained similarities are sufficient to capture the technical effects. For each batch i , as T increases, both the number of retained similarities and the number of cell identities of identified similar cells are non-increasing. Without loss of generality, given a set of incremental candidate values $SetT = \{T_s : s = 1, \cdots, S\}$, we aim to select optimal T satisfied 1) and 2). First, we delete the T_s's which filter out excessive similarities. For each batch i, we filter the identified similarities using each candidate T_s and obtain the number $Num_{T_s}^{i}$ of retained similar cells in remaining batches. We calculate the decrease ratio of $Num_{T_s}^{i}$ for each increase of $T_s : (Num_{T_s}^{i} - Num_{T_{s+1}}^{i}) / Num_{T_s}^{i}$. We delete all T_s's after T_s whose decrease ratio is greater than 50%, as the bigger T_s will filter out excessive similarities. We take the intersection of the update sets of candidate T_s for all batches denoted by $SetT*$. Then, we select optimal T which captures most similarities. As we pre-cluster each batch in the previous section, the pre-clustered labels are also available in our pipeline. We find the smallest T_s^{i} for each batch i, which preserve the identified cell label identities using the maximum T_s in $SetT*$. The median of T_s^{i}'s is identified as the optimal T.

For λ, given a set of incremental candidate values $Set\lambda = \{\lambda_v : v = 1, \cdots, V\}$, we aim to select optimal λ satisfied 1) big enough to remove technical effects; and 2) enable sufficient number of cPCs. We simply select the biggest λ in $Set\lambda$, while $XX^T - \lambda ZZ^T$ has more than l (default by 40) positive eigenvalues.

Selecting parameters for RPCI, LIGER, and online iNMF

We selected the optimal key parameters of RPCI, LIGER, and online iNMF. For RPCI, we set the reference batch as the batch that contained the most cell types and selected the optimal "eigens" parameter for each data set from 8, 10, 12, 15, 20, 25, 30, and 50. The final "eigens" parameter was tuned to be 12, 15, 8, 8, 12, and 15 for simulation 1, simulation 2, simulation 3, dendritic, developing tracheal epithelial, and DiHS/DRESS skin data sets, respectively. For LIGER and online iNMF, since they employed the same integrative nonnegative matrix factorization model, the same key parameters k and lambda were applied for these two methods. We selected the optimal lambda from 1, 5, and 10, and k from 8, 10, 12, 15, 20, and 25. Finally, lambda was tuned to be 5 for all data sets, and k was tuned to be 10, 20, 20, 15, 20, and 20 for simulation 1, simulation 2, simulation 3, dendritic, developing tracheal epithelial, and DiHS/DRESS skin data sets, respectively. In particular, the UMAP visualizations of the parameter tuning results of RPCI and online iNMF on simulation 1 were shown in Supplemental Figs. S26 and S27.

Supplementary References

Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**: 41-50.