nature portfolio

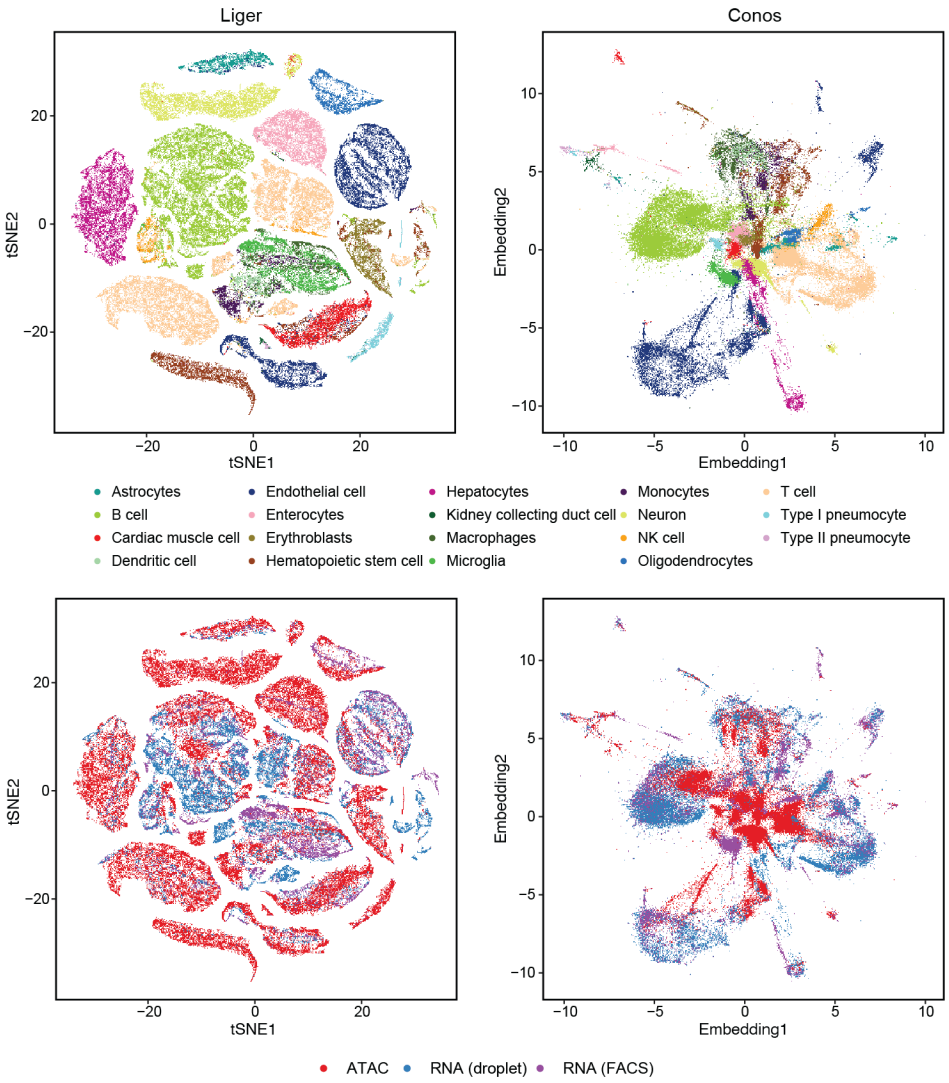
https://doi.org/10.1038/s41587-021-01161-6

Supplementary information

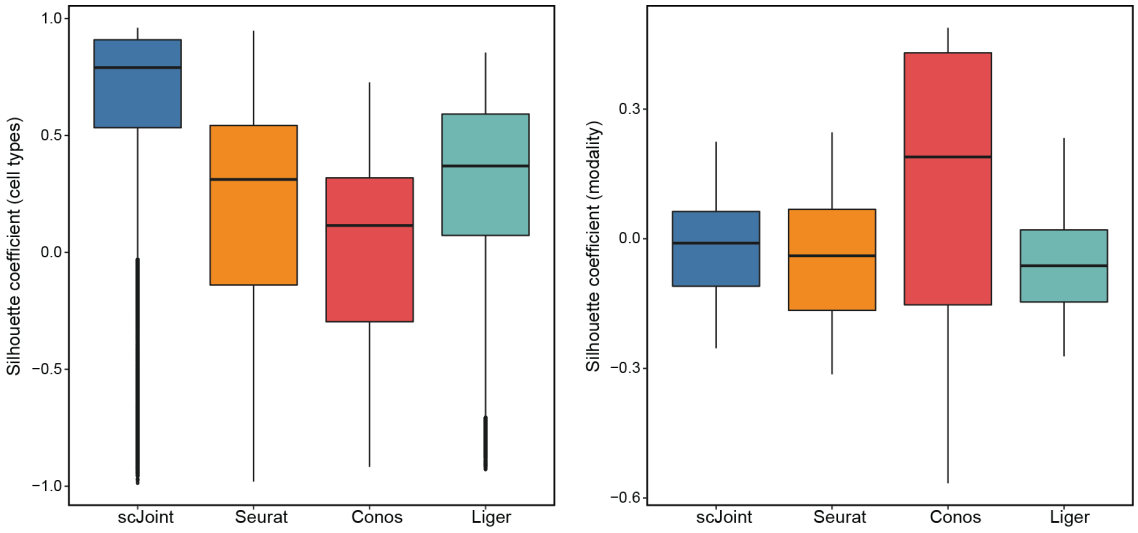
scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning

In the format provided by the authors and unedited

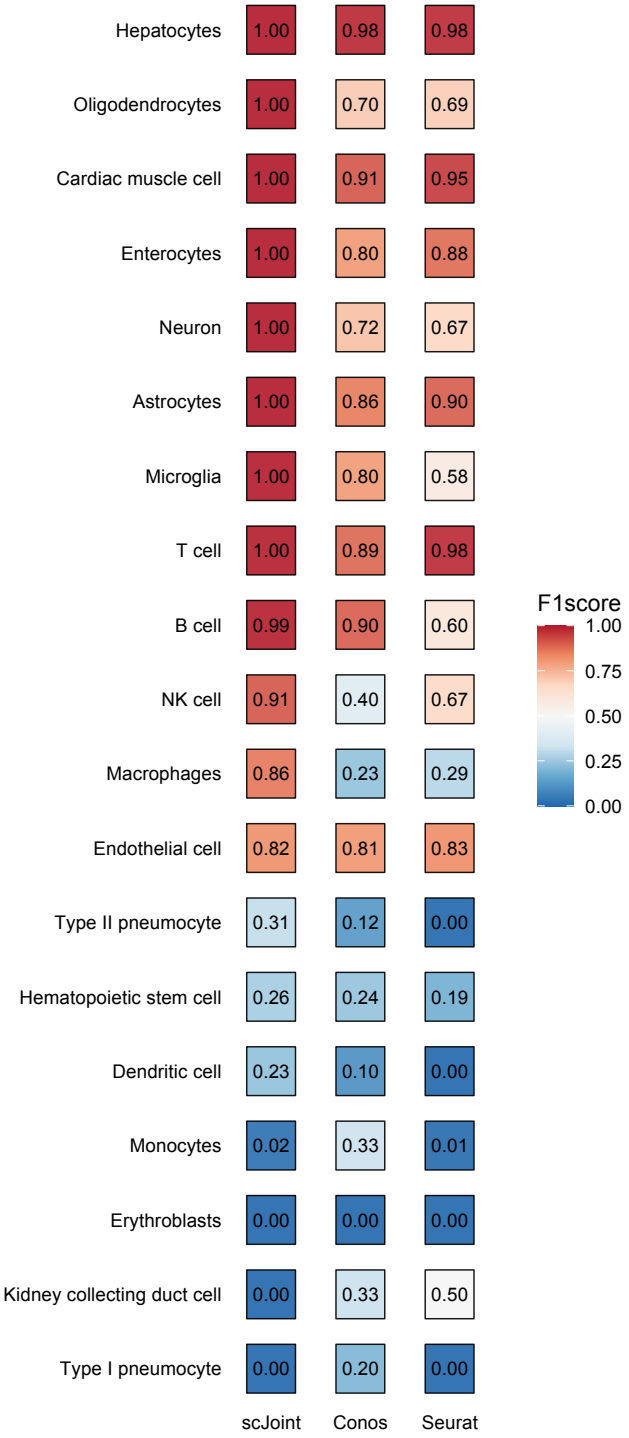
Supplementary Figures



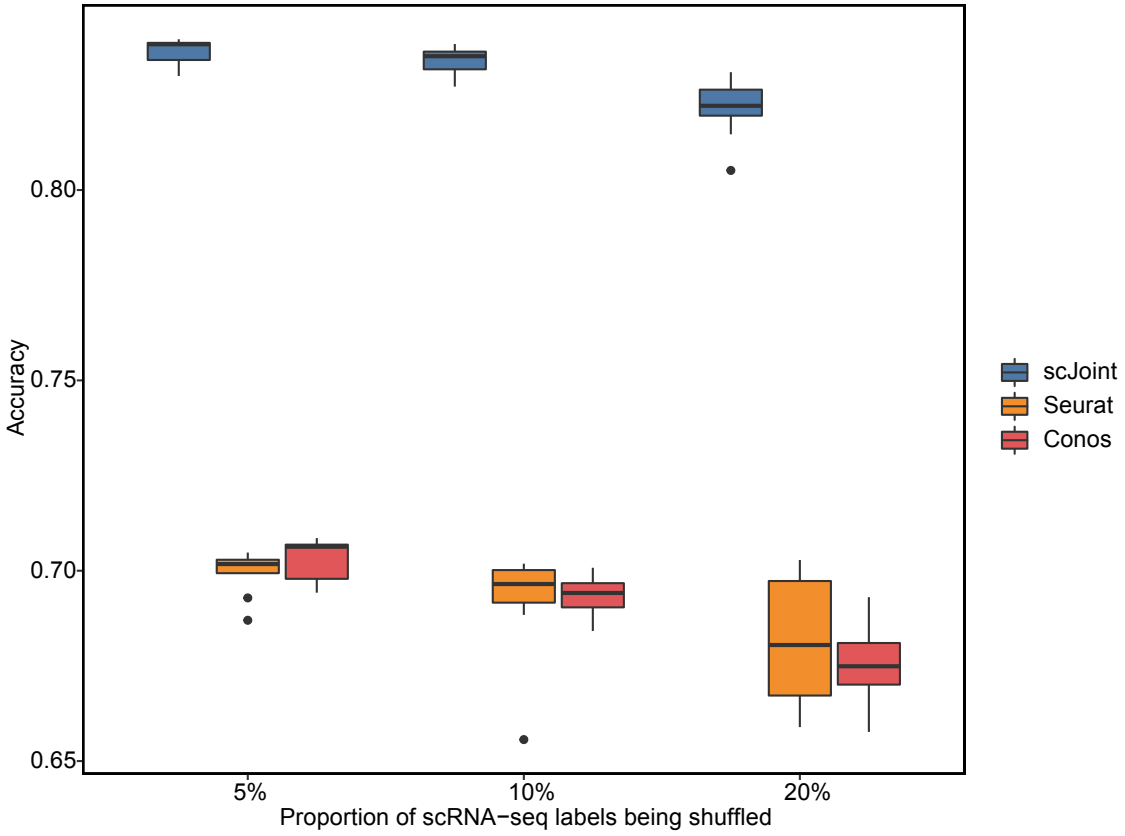
Supplementary Figure S1: tSNE visualization of the overlapping subset data from mouse cell atlases for Liger (first column) and Conos (second column), colored by cell type (first row) and technology (second row).



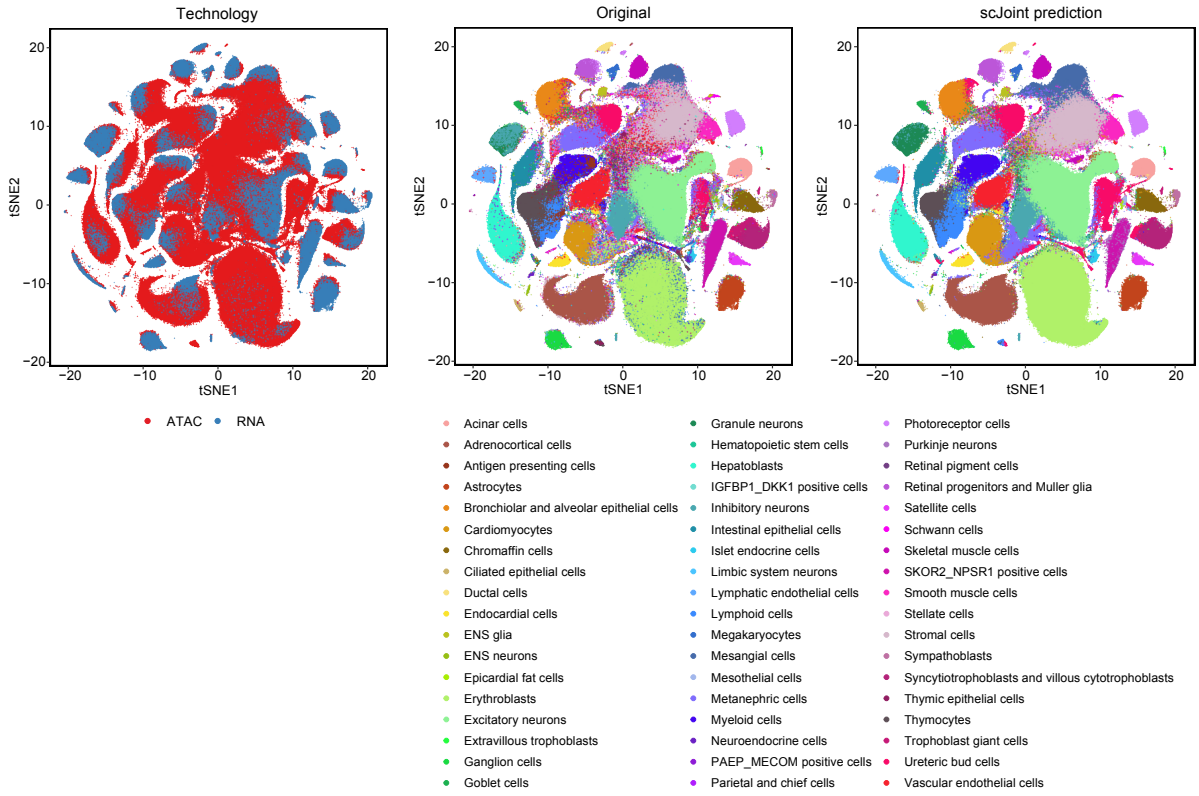
Supplementary Figure S2: Evaluating the joint visualizations of the mouse cell atlas subset data. Boxplots of cell type silhouette coefficients (left) and modality silhouette coefficient (right) for scJoint, Seurat, Conos and Liger (n = 101,692). Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range.



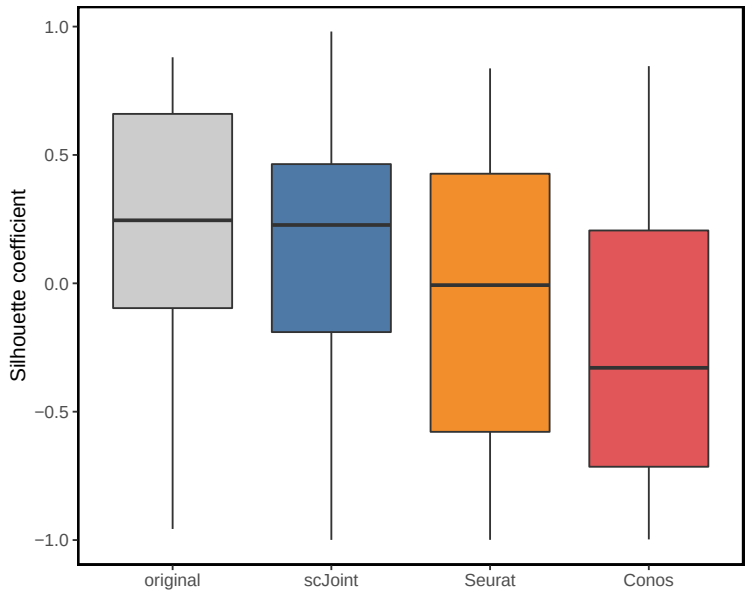
Supplementary Figure S3: Evaluating the accuracy of transferred labels for each cell type in the mouse cell atlas subset data. F1-scores of cell type classification from each method.



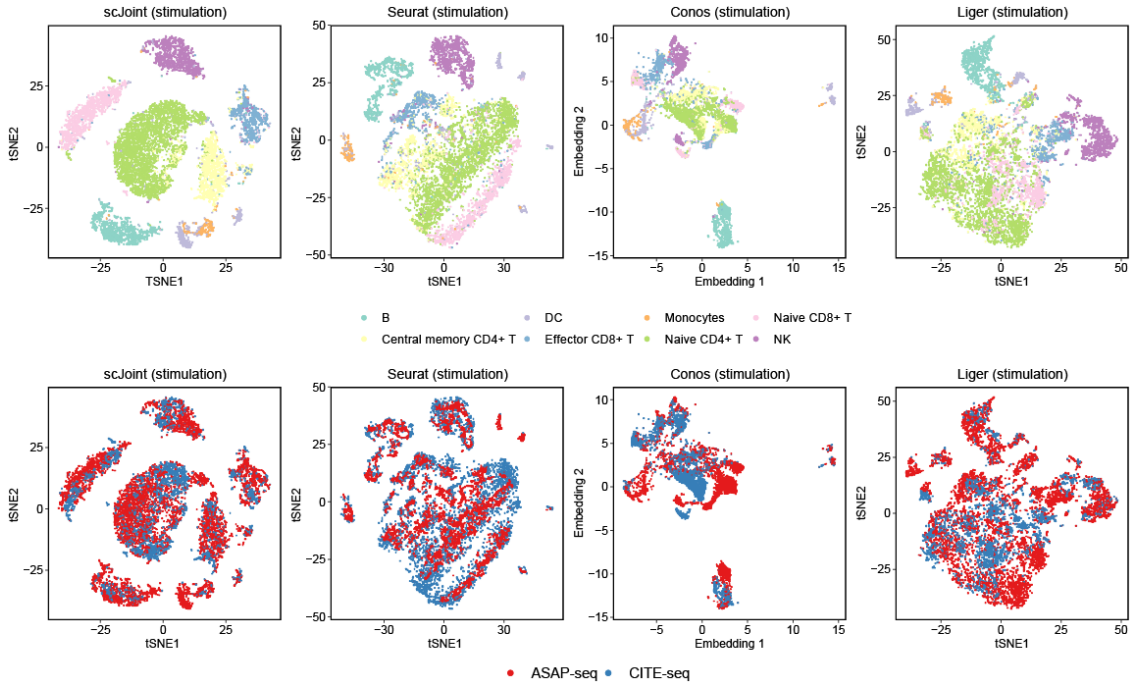
Supplementary Figure S4: Accuracy rates of scJoint, Seurat and Conos using scRNA-seq data with 5%, 10%, and 20% of the cell type labels randomly shuffled in mouse cell atlas subset data. 10 random shuffling were performed for each setting to generate the variance. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range.



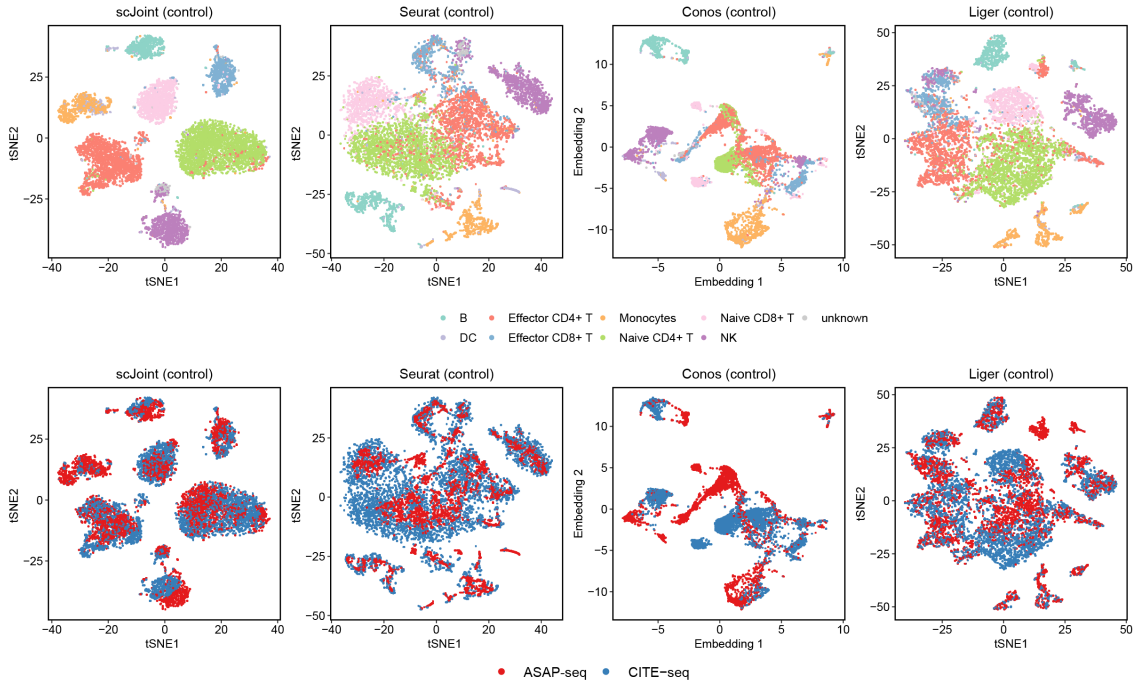
Supplementary Figure S5: tSNE visualization of the overlapping subset data (433,695 cells from scRNA-seq and 656,074 cells from scATAC-seq) from human fetal atlas for scJoint, colored by technology (left), original labels (middle) and scJoint prediction (right).



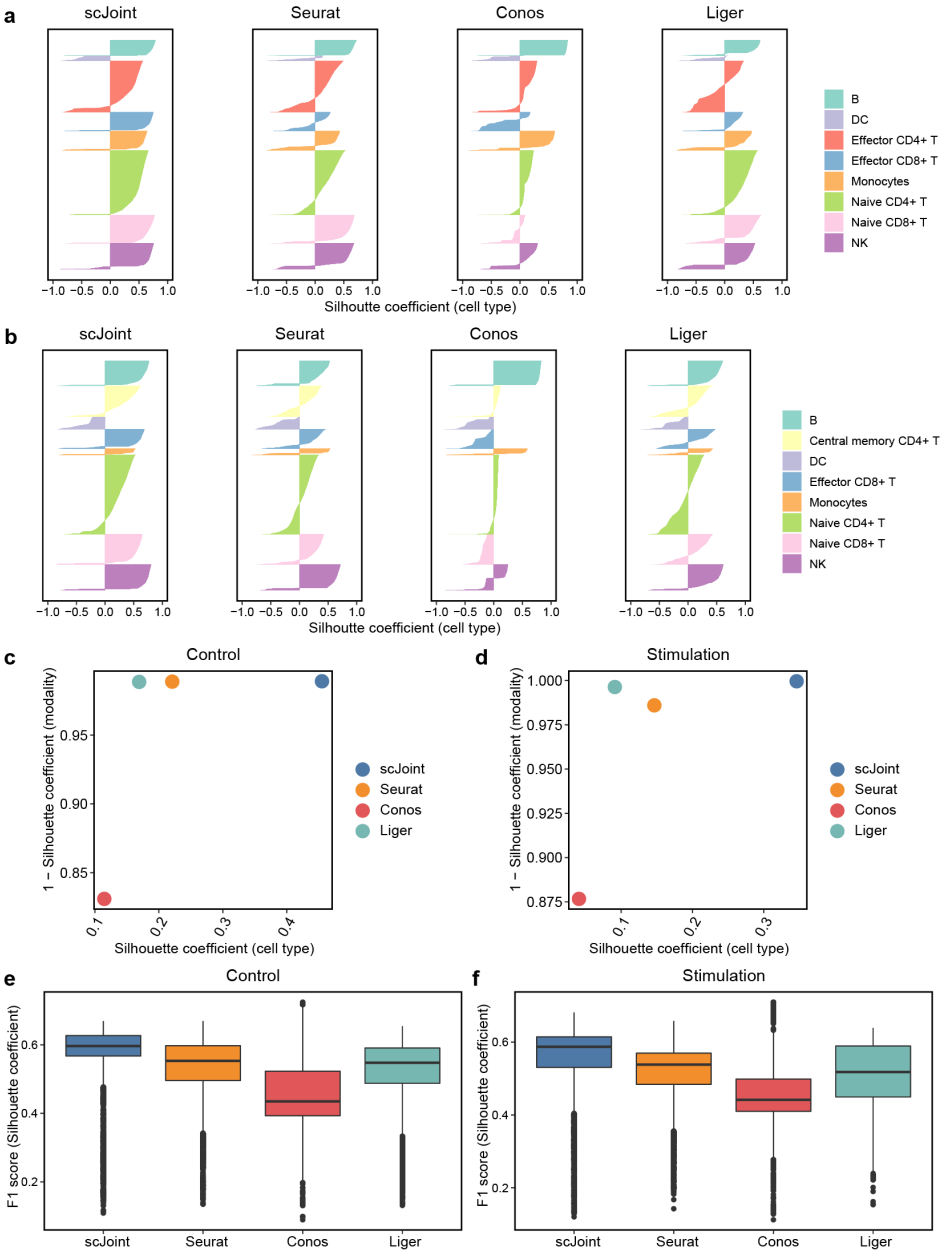
Supplementary Figure S6: Evaluating the joint visualizations of the mouse cell atlas data: boxplots of cell type silhouette coefficients. The distances between cells are calculated based on the tSNE plot of scATAC-seq data (n = 81,173), and the predicted labels from scJoint, Seurat and Conos are used as grouping information, with the cell type labels from the original scATAC-seq study used as the golden standard for comparison. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range.



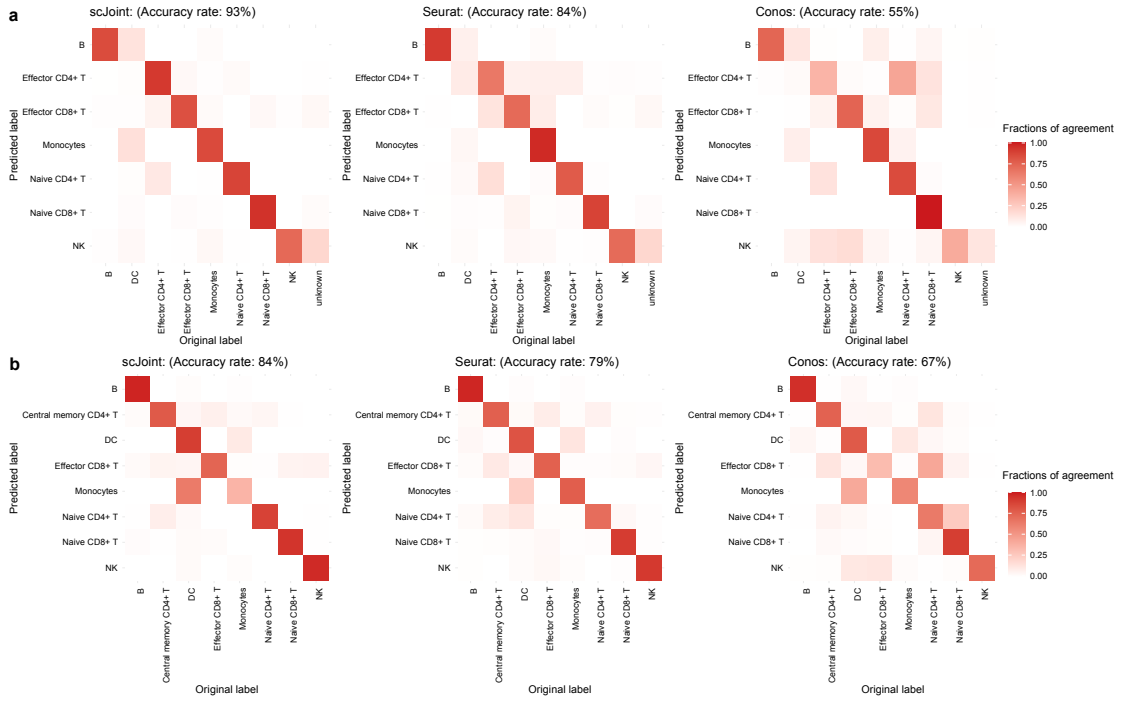
Supplementary Figure S7: tSNE visualization of CITE-seq and ASAP-seq PBMC data under stimulation, generated by scJoint (first column), Seurat (second column), Conos (third column) and Liger (fourth column), colored by cell types (first row) from CiteFuse and manual annotations, and technology (second row).



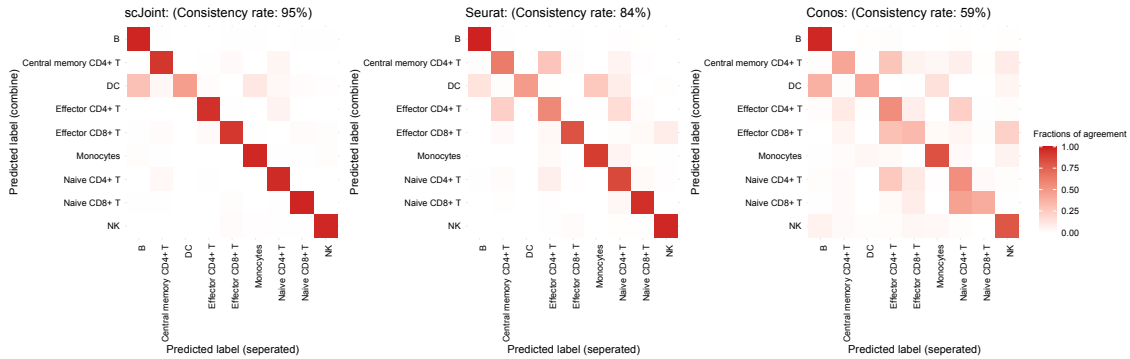
Supplementary Figure S8: tSNE visualization of CITE-seq and ASAP-seq PBMC data under the control condition, generated by scJoint (first column), Seurat (second column), Conos (third column) and Liger (fourth column), colored by cell types (first row) from CiteFuse and manual annotations, and technology (second row).



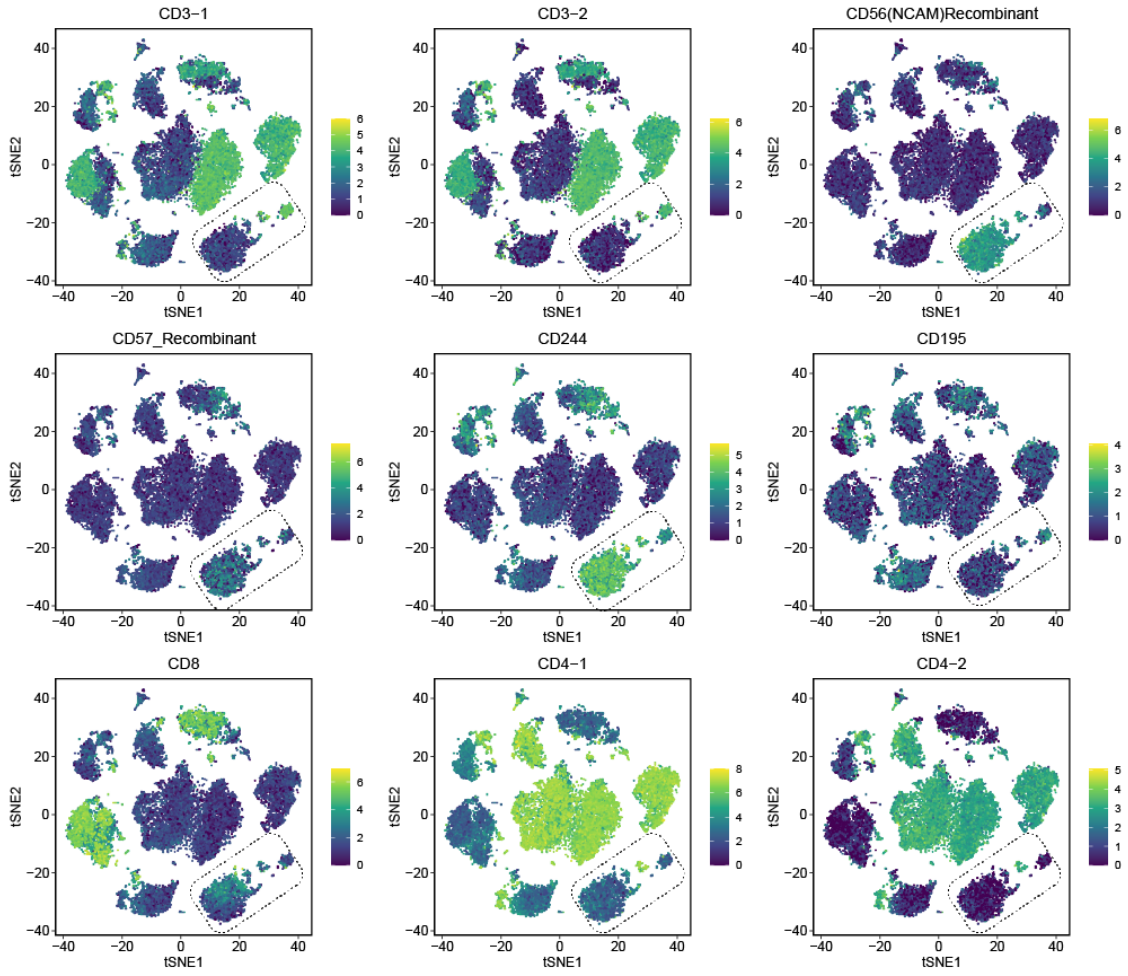
Supplementary Figure S9: Evaluating the joint visualizations of CITE-seq and ASAP-seq PBMC data. (a-b) Barplots of cell type silhouette coefficients for scJoint, Seurat, Conos and Liger for all cells, colored by cell types under (a) control; (b) stimulation. (c-d) Scatter plot of mean silhouette coefficients for scJoint, Seurat, Conos and Liger (left), where the x-axis denotes the mean silhouette coefficients of cell types and the y-axis denotes 1 - mean modality silhouette coefficients under two conditions: (c) control; (d) stimulation; (e-f) Boxplots of F1 scores of silhouette coefficients for scJoint, Liger, Seurat, and Conos, under (e) control (n = 9,146) and (f) stimulation (n = 8,942). Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range.



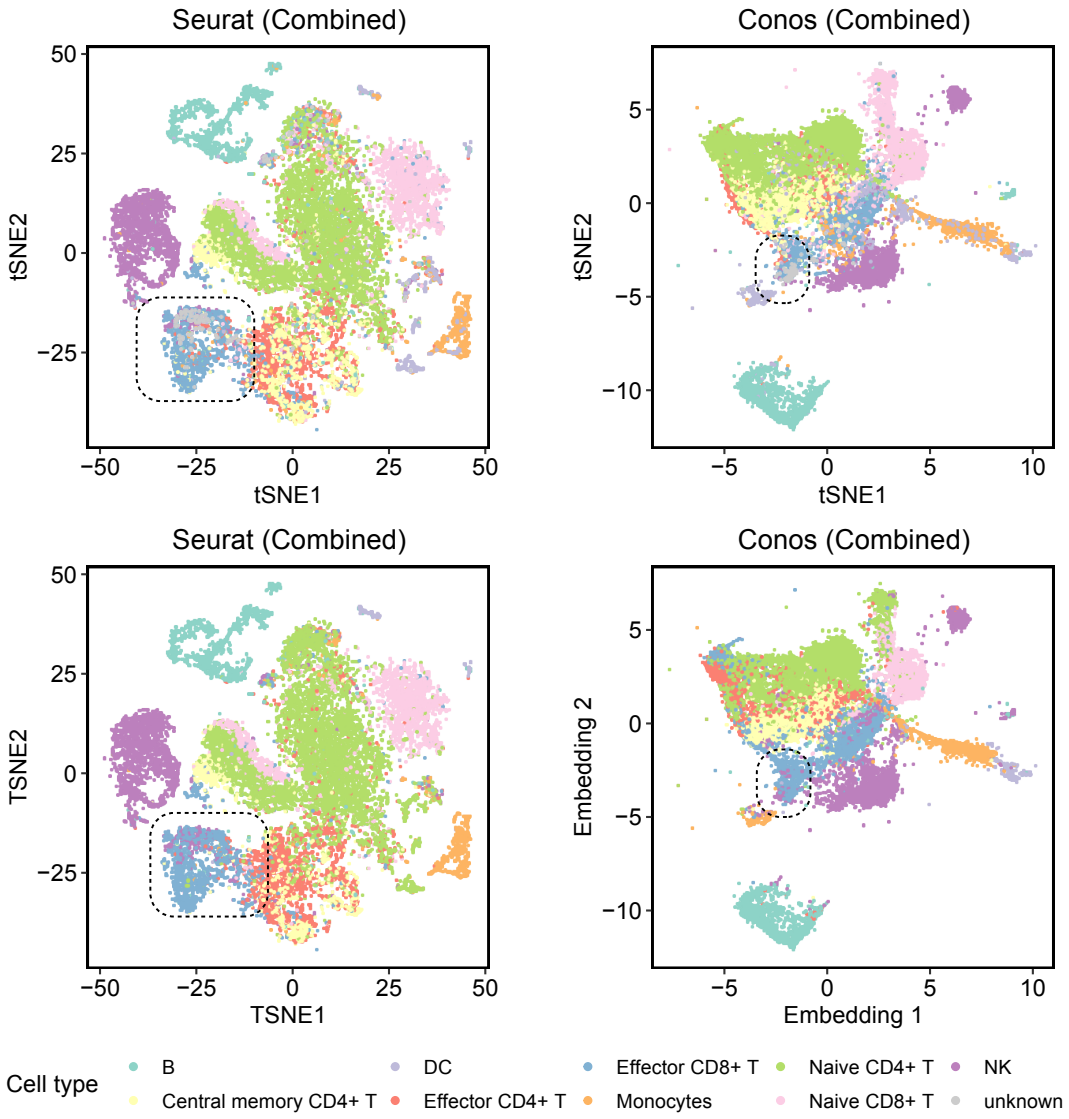
Supplementary Figure S10: Label transfer accuracy in CITE-seq and ASAP-seq PBMC data. Heatmaps show fractions of agreement between the original labels and the transferred labels of scJoint, Seurat and Conos: (a) control; (b) stimulation.



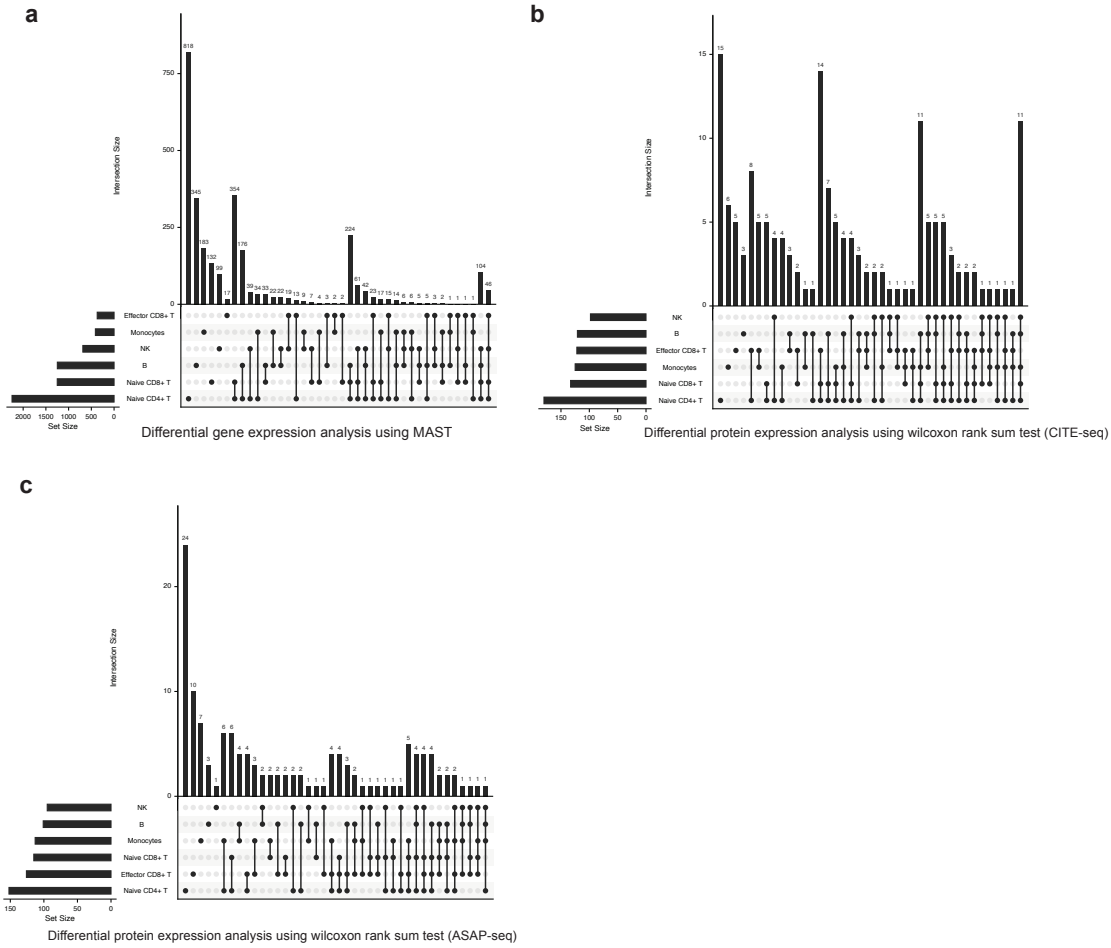
Supplementary Figure S11: Consistency of label transfer in CITE-seq and ASAP-seq PBMC data. Heatmaps show fractions of agreement among the transferred labels from running each method on control / stimulation separately and two conditions jointly: scJoint (left), Seurat (middle) and Conos (right).



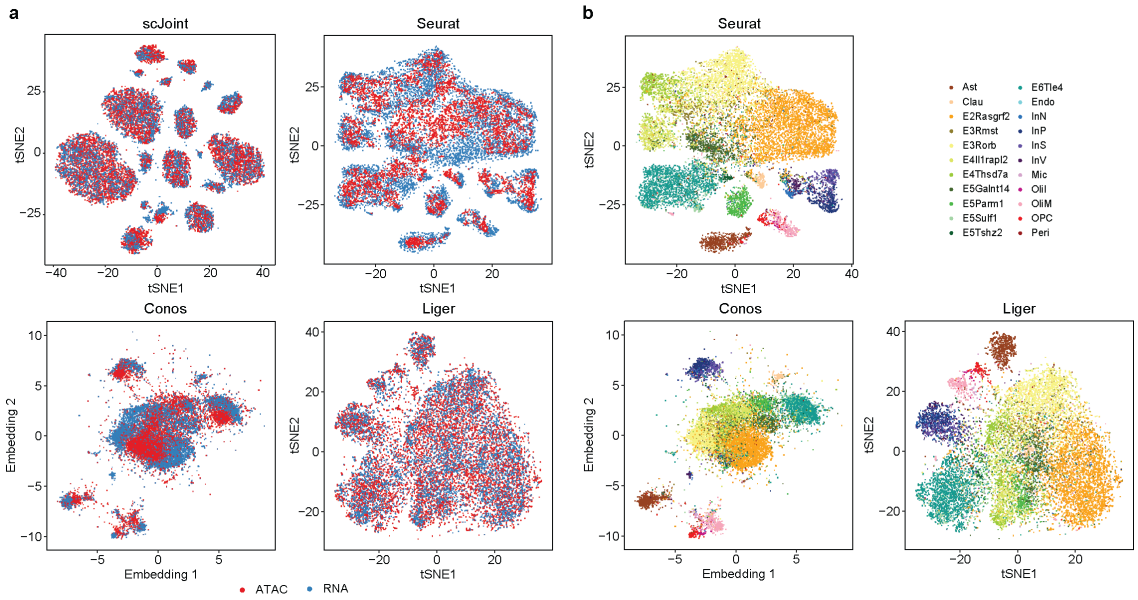
Supplementary Figure S12: ADT expression of NK T cells in CITE-seq and ATAC-seq data.



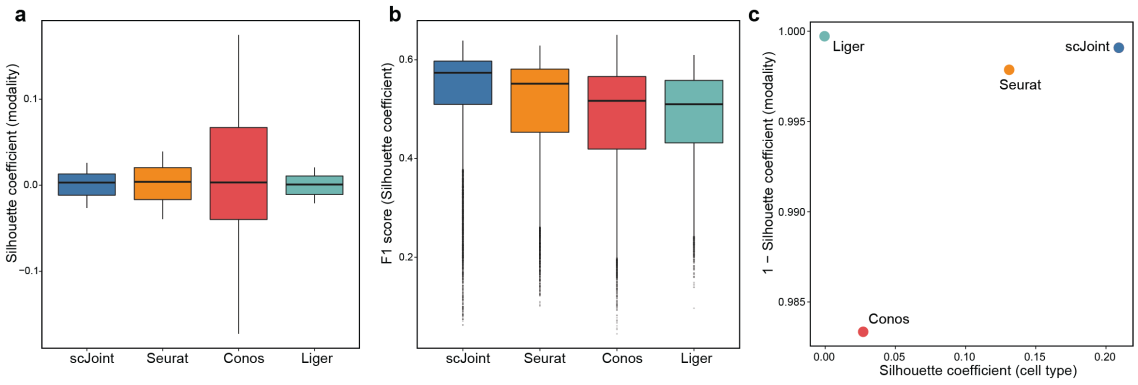
Supplementary Figure S13: tSNE visualization of CITE-seq and ASAP-seq PBMC data under combined conditions, generated by Seurat (first column) and Conos (second column), colored by original cell types (first row) from CiteFuse and manual annotations, and predicted cell types. Cells identified as NK T cells in scJoint visualization are mixed with Effector CD8+ T cells by other methods.



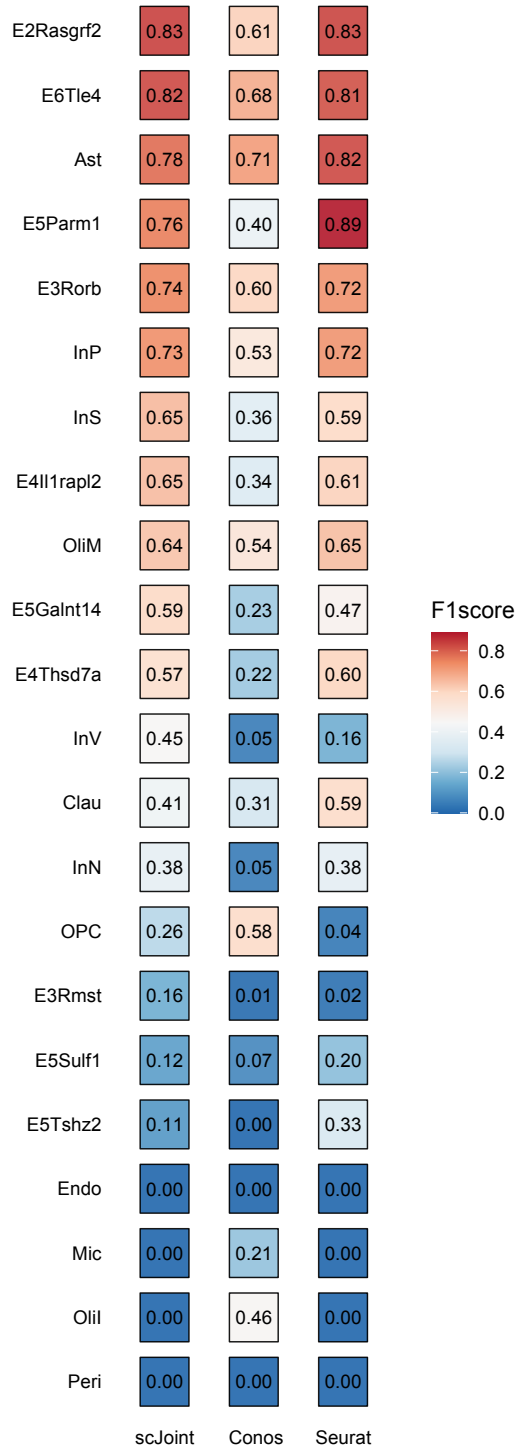
Supplementary Figure S14: Differential expression (DE) analysis across two conditions of CITE-seq and ASAP-seq.



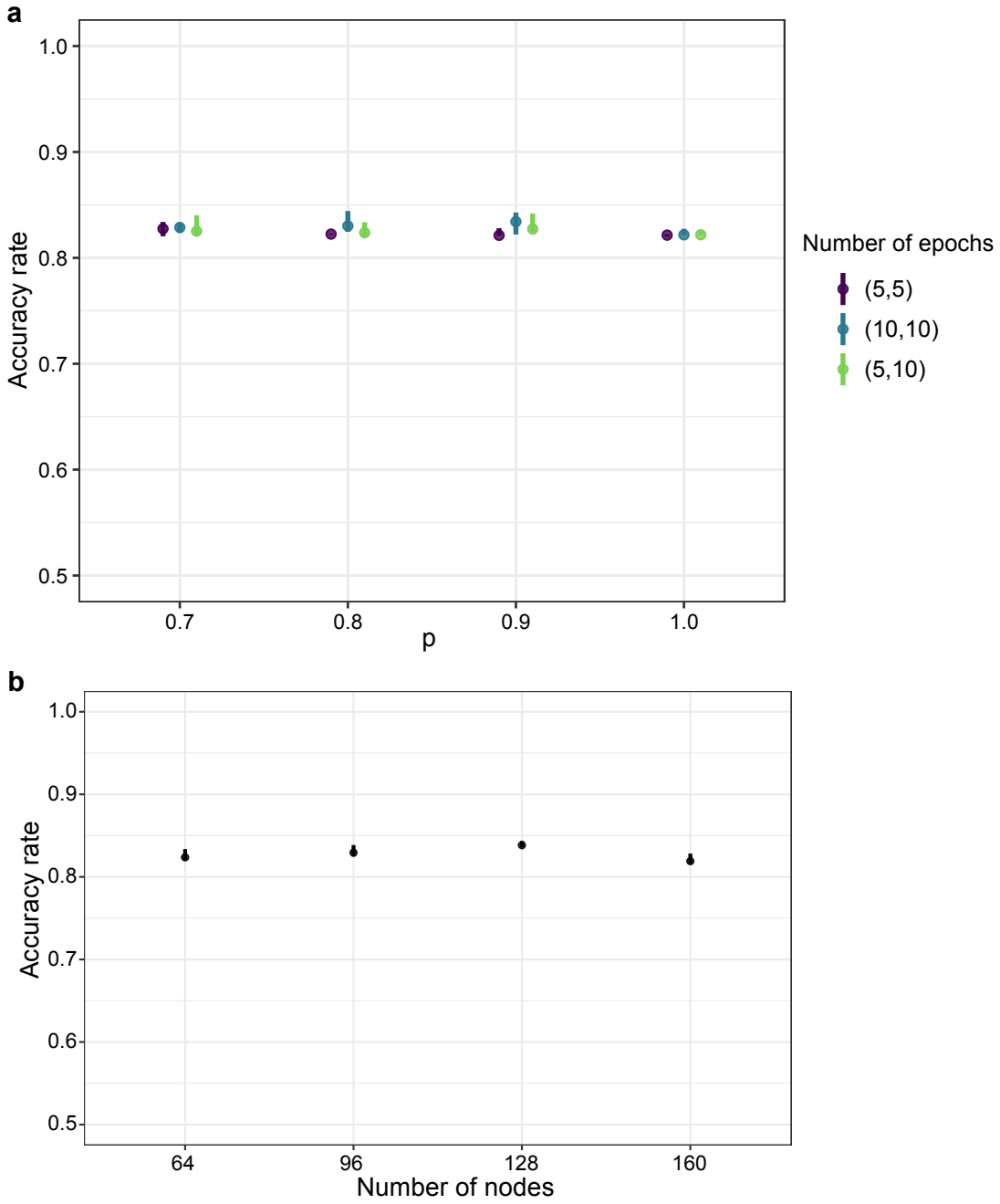
Supplementary Figure S15: (a) tSNE visualization of SNARE-seq data with the RNA and ATAC parts colored separately for unpaired methods: scJoint (top left), Seurat (top right), Conos (bottom left) and Liger (bottom right). (b) tSNE visualization of SNARE-seq data colored by original cell types, generated by Seurat (top), Conos (bottom left), Liger (bottom right).



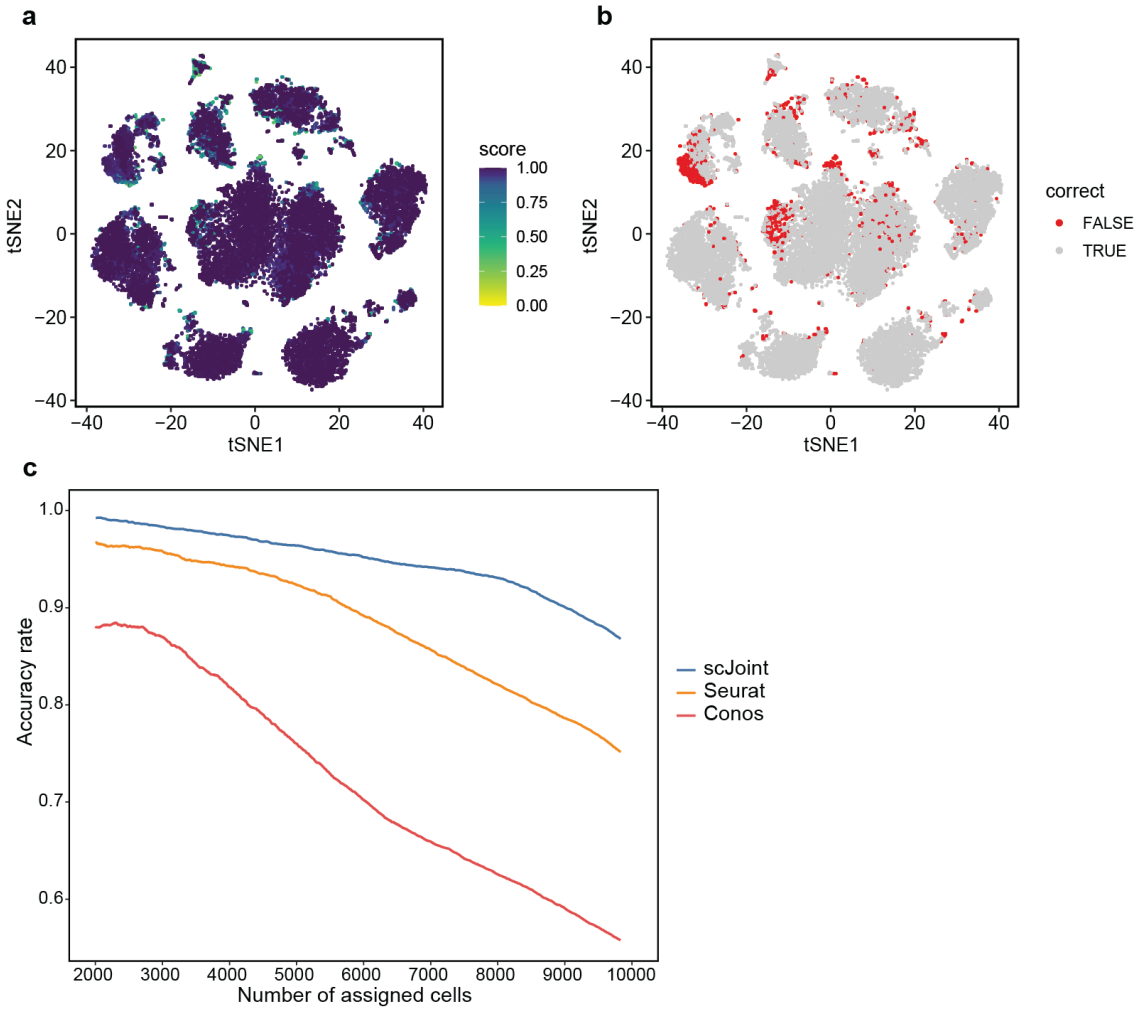
Supplementary Figure S16: Evaluating the joint tSNE visualizations for SNARE-seq data. (a) Boxplots of modality silhouette coefficients for scJoint, Seurat, Conos, and Liger (n = 9,190); smaller values indicate better mixing. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range. (b) Boxplots of F1 scores of silhouette coefficients for scJoint, Seurat, Conos, and Liger (n = 9,190); larger values indicate better balance. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend 1.5 times the interquartile range. (c) Scatter plot of mean silhouette coefficients for scJoint, Liger, Seurat, and Conos, where the x-axis shows the mean cell type silhouette coefficients and the y-axis shows 1 - mean modality silhouette coefficients; ideal outcomes would lie in the top right corner.



Supplementary Figure S17: Evaluating the accuracy of transferred labels for each cell type in the SNARE-seq data. F1-scores of cell type classification from each method.



Supplementary Figure S18: Robustness to tuning parameters. Label transfer accuracy of scJoint on the overlapping subset data from the mouse cell atlases when varying (a) the fraction p of data pairs included in the cosine similarity loss ($p = 0.7, 0.8, 0.9, 1.0$) and number of training epochs in Step 1 and 3 ($(5, 5)$, $(10, 10)$, and $(5, 10)$); (b) the number of nodes in the embedding (hidden) layer (number of nodes $= 64, 96, 128, 160$). The dots indicate the medians and the bars indicate the interquartile range from 10 independent runs.



Supplementary Figure S19: Evaluating the probability scores for label transfer in CITE-seq and ASAP-seq PBMC data. (a-b) tSNE visualization using scJoint colored by: (a) probability score of cell type prediction; (b) the correctness of transferred labels. (c) Accuracy rate changes as we change the threshold for probability scores in each method. The x-axis shows the number of cells whose probability scores exceed a given threshold and were assigned a prediction; the y-axis shows the corresponding accuracy rate.

Dataset	S	T	# epochs (Step 1)	lr (Step 1)	# epochs (Step 3)	lr (Step 3)	λ
Mouse atlas full	2	1	10	0.01	10	0.01	10
Mouse atlas subset	2	1	10	0.01	10	0.01	10
SNARE-seq	1	1	10	0.01	10	0.01	1
Multi-modal control	1	1	20	0.01	20	0.01	1
Multi-modal stimulation	1	1	20	0.01	20	0.001	1
Multi-modal combined	1	1	20	0.01	20	0.01	1

Table S1: Training details for each data listing the number of scRNA-seq datasets (S), number of scATAC-seq datasets (T), learning rate (lr) and number of training epochs used in Step 1 and Step 3.

Number of cells	batch size	# epochs (Step 1)	lr (Step 1)	# epochs (Step 3)	lr (Step 3)
$\leq$ 50k	256	10	0.01	10	0.01
50k - 500k	512	10	0.01	10	0.01
$\geq$ 500k	1024	10	0.01	10	0.01

Table S2: Training details for the human fetal atlas data, including the batch size, learning rate (lr) and number of training epochs used in Step 1 and Step 3.

Supplementary Note

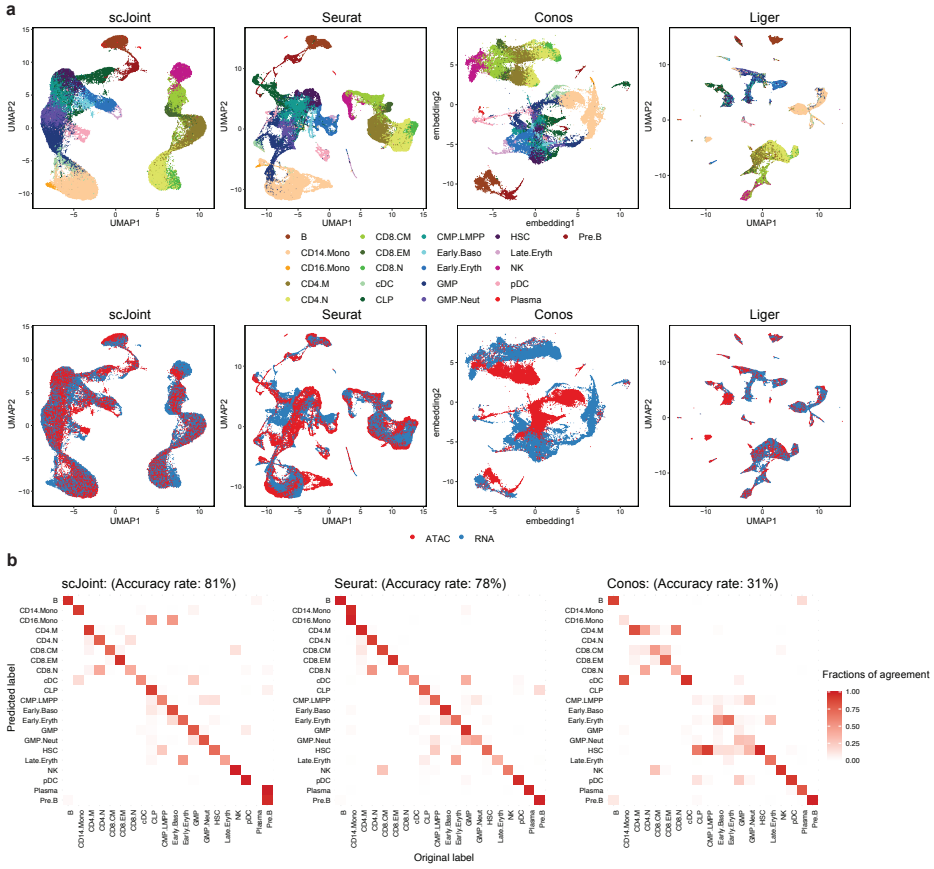
A: Integrative analysis of human hematopoiesis data

As a case study to demonstrate the trajectory mode of scJoint, we performed integrative analysis of human hematopoiesis data from healthy donors generated by scRNA-seq and scATAC-seq [1]. The data contains 35038 cells for scRNA-seq data and 35582 cells for scATAC-seq data from multiple hematopoietic lineages, including the B cell lineage (HSC - CMP/LMPP - CLP - Pre B - B) and the monocytic lineage (HSC - CMP/LMPP - GMP - CD14 Mono).

We first integrated the full data to evaluate the label transfer accuracy rate and the joint embedding of the two modalities. Supplementary Figure S20 shows that the trajectory mode of scJoint effectively mixes the two modalities while multiple lineage trajectories are also well represented in the joint visualization. Moreover, scJoint achieves a higher accuracy rate (81%) than Seurat (78%) and Conos (31%).

Next, we performed trajectory analysis on the B cell lineage and monocytic lineage respectively to evaluate the conservation of trajectory after integration based on the original labels. Similar to Luecken et al.[2], the trajectory for each lineage was generated by the diffusion map of the joint embedding space from each method; diffusion pseudotime was then obtained using the R package *destiny* [3]. The cells labeled hematopoietic stem cells (HSC) were considered as root cells. The diffusion pseudotime was then ranked and normalized. For the unintegrated data, the diffusion map was performed on the first 50 principal components for the scRNA-seq and scATAC-seq data respectively. We then used the trajectory conservation score to evaluate the preservation of the trajectory structure. The trajectory conservation score is defined similarly as in Luecken et al. [2], quantified by the Spearman correlation between the trajectories from the integrated data and the unintegrated data. Note that the trajectory conservation score proposed assumes the trajectory from unintegrated data as the golden standard, but our analyses (panel c of Supplementary Figures S21, S22) suggest unintegrated data do not necessarily contain the clearest trajectory information.

The diffusion maps of scJoint reveal biological trajectories following the correct order of lineage development for both the B cell lineage and monocytic lineages (panel b of Supplementary Figures S21, S22, first column) and also broadly consistent with the unintegrated data (panel



Supplementary Figure S20: (a) UMAP plots of human hematopoiesis developmental data for scJoint, Seurat, Conos, and Liger, colored by cell types defined in the original study [1] (first row), and cell types (second row). (b) Predicted cell types and their fractions of agreement with the original cell types given in the original study [1] for scJoint (left panel), Seurat (middle panel) and Conos (right panel). Clearer diagonal structure indicates better agreement.

a of Supplementary Figures S21, S22). In terms of the trajectory conservation scores, scJoint performs consistently with scores greater than 0.7 in all cases; scJoint and Seurat are the top two methods and perform most consistently in both lineages (panel d of Supplementary Figures S21, S22). Although Conos has high conservation scores for the B cell lineage, it fails to integrate the two modalities (panel b of Supplementary Figures S21, S22, last column). Looking closely at the distribution of the pseudotime, scJoint shows clearer shifts in distribution as the cell types evolve along the correct developmental path than the unintegrated data (panel c of Supplementary Figures S21, S22, RNA and ATAC). This suggests integration of the two modalities brings about better reconstruction of developmental processes and the use of unintegrated data as the golden standard for comparison in the conservation scores may not be optimal. In this sense, the plots in panel c of Supplementary Figures S21, S22 can be considered as an alternative way of assessing

the pseudotime from each method, where we can see scJoint provides mostly uni-modal distributions with clear shifts, while the other methods often have multiple modes in each cell type.

Together, the analysis of the human hematopoiesis data illustrates that the trajectory version of scJoint is capable of integrating data from continuous biological processes.

B: scJoint loss functions

The purpose of each loss function component in $\mathcal{L}_{\text{scJoint}}$:

1. The **NNDR loss** performs dimension reduction on data adapting intuition from PCA and is a novel loss that integrates dimension reduction into the whole training framework. Compared to existing approaches in computer vision, which apply PCA separately followed by CNN (known as whitening), our loss directly uses the neural network itself for feature extraction, thus allowing the low dimensional features to be jointly updated throughout training.

2. The **cosine similarity loss** aligns cells across the two modalities (RNA and ATAC) with similar low dimensional representations. It is commonly used in the computer vision literature for face recognition (e.g. Wang et al. [4]), but underexplored in the single-cell deep learning literature and novel in its application.

3. The **cross entropy loss** is a commonly used loss for classification. It enforces clear separation between different cell types in the case with well-differentiated tissues. This loss can be removed in Step 3 for developmental data since the underlying cell states are more continuous.

4. The **center loss** removes batch effects by encouraging cells of the same cell type to be close to their cluster center in the low dimensional space representation, regardless of their batch or modality labels. This loss only comes in Step 3 as it requires a reasonably accurate estimation of ATAC cell labels, which are obtained in Step 2, for calculating the cluster centers.

C: Additional assessment of scJoint

C1: Evaluation of activation functions

We examined two types of nonlinear activation functions popular in the single cell deep learning literature:

- (1) LeakyReLU with slope equal to 0.05, 0.1 and 0.5, and
- (2) ELU with default setting.

The results of the mouse atlas overlapping subset data are summarized in the table below. We observe that for LeakyReLU, the accuracy increases as the slope increases and as the function becomes closer to a linear function; for ELU, it achieves the same accuracy rate as using the linear activation function. Intuitively, the nonlinear activation functions do not outperform the linear activation functions because these nonlinear activation functions can be well approximated by a few linear function pieces and we only have one hidden layer. In this sense, we do not expect nonlinearity to add extra representation power since we already have an overparametrized neural network (80 million parameters). Thus, the linear activation function is one of the optimal settings for scJoint.

	LeakyReLU (slope = 0.05)	LeakyReLU (slope = 0.1)	LeakyReLU (slope = 0.5)	ELU (default in PyTorch)
Accuracy	69%	82%	82%	0.84%

Table S3: Label transfer accuracy table of nonlinear activation functions or mouse atlas overlapping subset data.

C2: Choice of the weight for the center loss

We examined the choice of weight for the center loss. By default, we set the weight of the center loss λ as 1. For all the datasets that used $\lambda = 1$ in the manuscript (Supplementary Table S1), we found larger λ did not change their label transfer accuracy or embedding visualizations. However, a larger λ should be considered when different technologies were used in generating the omics leading to batch effects (e.g. droplet-based and plate-based in scRNA-seq). In the mouse atlas data analysis, which contains scRNA-seq data from 10x Genomics and SMART-seq2, we set λ as 10. Supplementary Figure S23 shows that batch effects are still apparent with λ equal to

1. Good mixing starts with $\lambda = 5$, and the visualizations are very stable for lambda larger than 10. ($\lambda = 10$ shown in Figure 2a.)

In the example of human hematopoiesis developmental data, we set $\lambda = 1$ during training since the scRNA-seq data was generated by one technology. Supplementary Figure S24 shows UMAP plots with larger λ ($\lambda = 5, 10$) show similar cell type development structure and batch mixing as the UMAP plot with λ = 1 (shown in Supplementary Figure S20). Thus similar to the case for well differentiated tissues, the results are very stable to larger λ values.

C3: Evaluation of binarization

To investigate the impact of different ways to binarize data as input for scJoint, we performed additional experiments using the mouse cell atlas subset data from the 19 overlapping cell types to compare the performance of:

1. Binarized vs. non-binarized scRNA-seq data
2. Different higher expression thresholds for non-binarized scRNA-seq data
3. Different lower expression thresholds for binarized scRNA-seq data

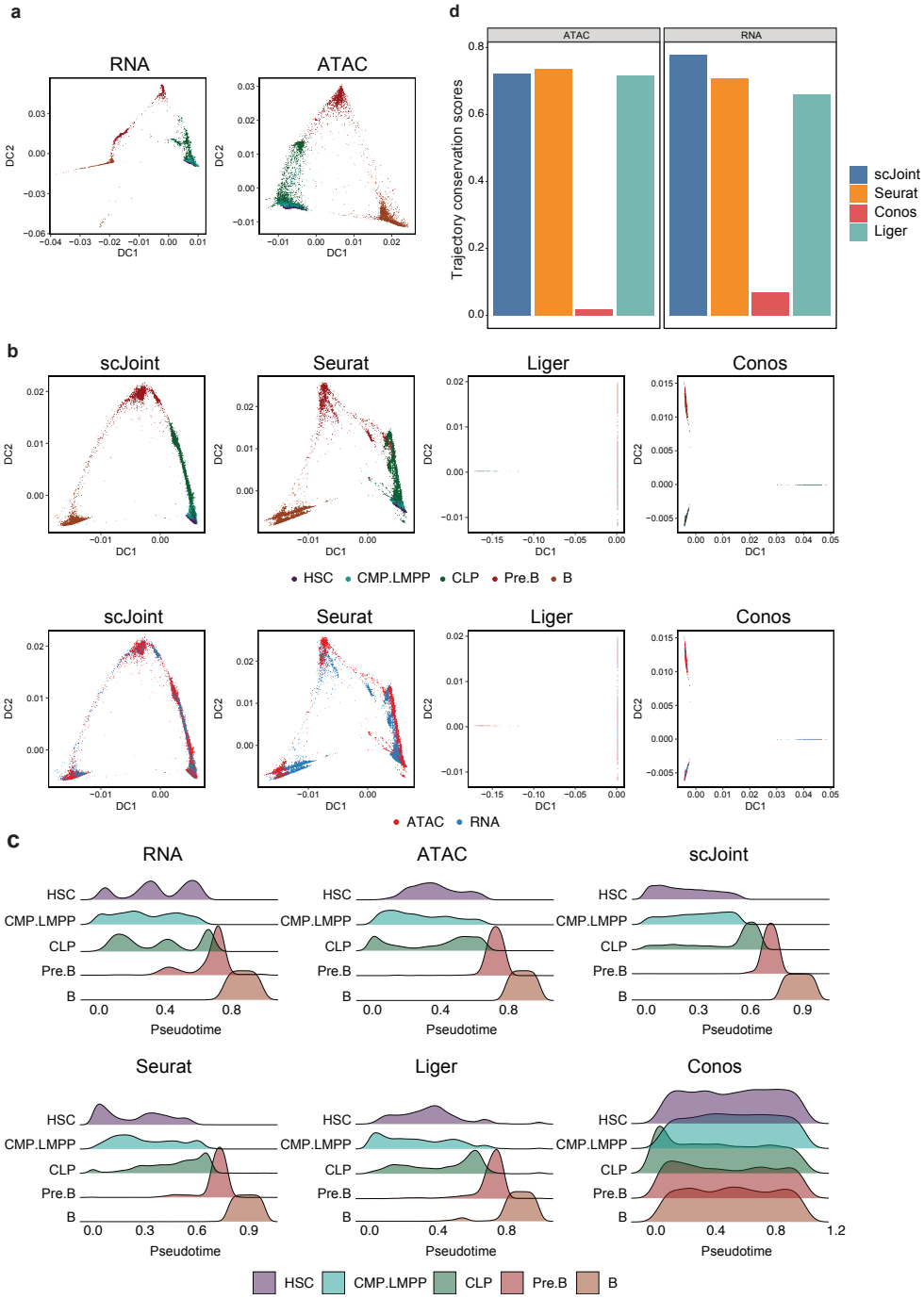
The performance of these different forms of scRNA-seq as input data is evaluated by the overall label transfer accuracy and the F1-score per cell type in the scATAC-seq data. Overall, these results suggest that the binarized scRNA-seq is optimal for scJoint, and the label transfer performance of scJoint is robust to how the binary matrix is constructed. The details of the experiments are:

Case 1: Binarized vs. non-binarized scRNA-seq data. For the non-binarized data, we scaled the log-transformed gene expression data into the range of 0 to 1 using min-max scaling so that it has the same scale as the binary scATAC-seq data. The label transfer accuracy of the non-binarized scRNA-seq is 71.8%, which is significantly lower than using the binarized matrix (84%) and using the binarized matrix as input outperforms the non-binarized for all cell types (Supplementary Figure S25a).

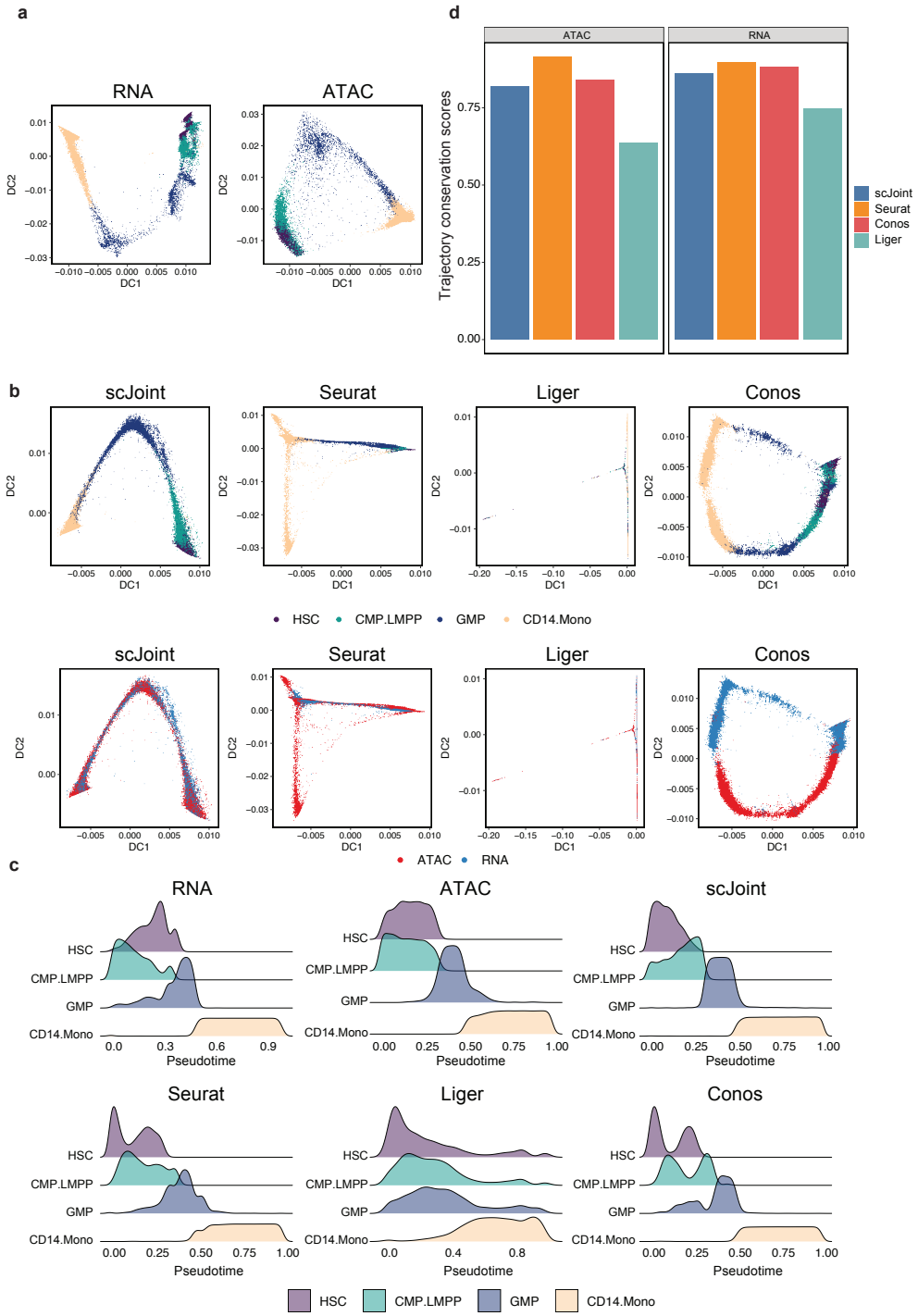
Case 2: Different higher expression thresholds for non-binarized. scRNA-seq data. First we truncated the expression values with maximum value set to a threshold t. We then performed min-max transformation to scale the values to between 0 and 1. As this threshold decreases, the input matrix becomes more similar to the binarized matrix. We varied this threshold from 3 to 8.

Supplementary Figure S25b shows that the label transfer accuracy rate increases as this threshold decreases, that is, as more high expression values are capped at 1. For performance per cell type, the binarized data has higher F1-scores than the thresholded non-binarized data in all cell types except monocytes (Supplementary Figure S25a, columns 1, 7-12). While monocytes is one of the cell types with a small number of cells in the training data, the binarized data performs well in all the other cell types with fewer cells than monocytes.

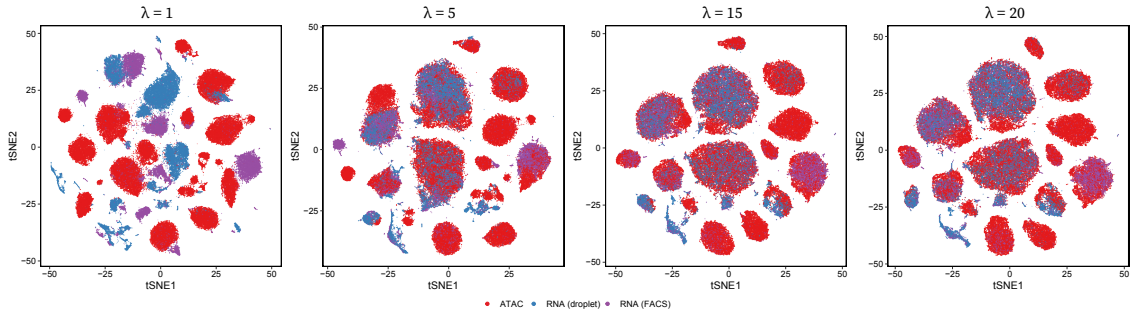
Case 3: Different lower expression thresholds for binarized scRNA-seq data. We explored different lower thresholds to binarize data. If the log-transformed gene expression value is greater than this threshold, we set the value as 1; otherwise we set the value as 0. As this threshold increases, the input matrix becomes sparser. (Our binarization is equivalent to setting this threshold as 0.) We varied this threshold from 1 to 4. Note the 5%, 10%, 20% and 30% quantiles of the non-zero expression value in the scRNA-seq SMART-seq2 data are $1.1, 1.73.1$ and 4.2 respectively. Therefore, using a threshold of 1 is equivalent to roughly setting 5% of non-zero gene expression values as 0. As shown in Supplementary Figure S25a (columns 3-6) and Supplementary Figure S25c, we found that both the overall label transfer accuracy and F1-score per cell type are robust to thresholds from 1 to 3. As expected, the accuracy rate drops at the threshold 4 as the input data becomes too sparse and loses too much information due to thresholding.



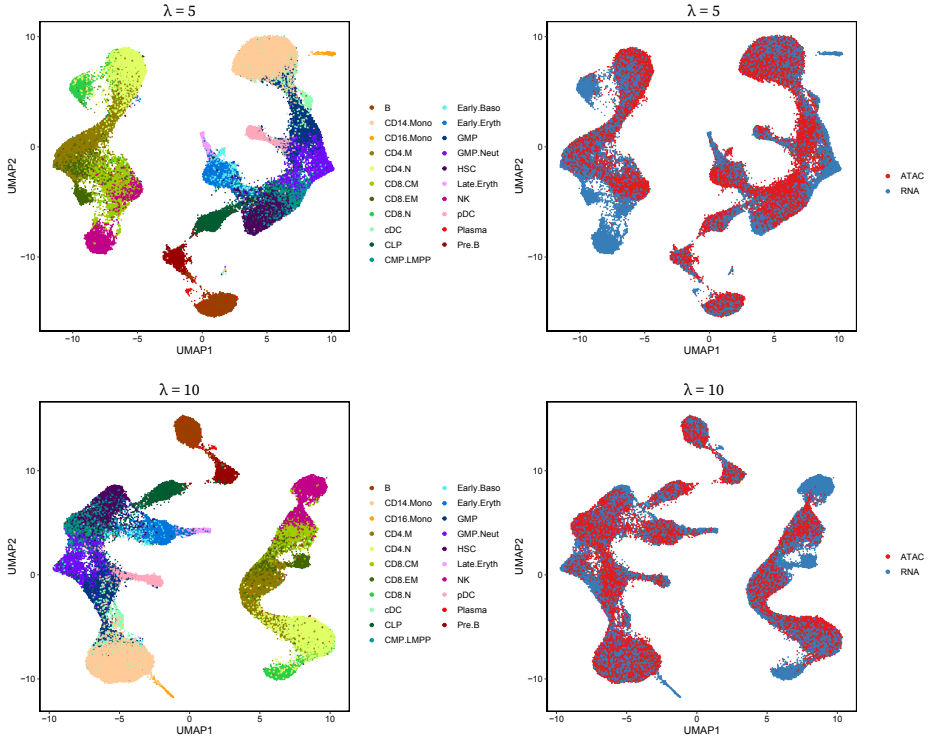
Supplementary Figure S21: Analaysis of B lineages. (a) Diffusion maps of unintegrated scRNA-seq (left) and scATAC-seq data (right). (b) Diffusion maps of integrated data, colored by cell types (top) and modality (bottom), generated by scJoint, Seurat, Liger and Conos. (c) Distribution of pseudotime for scRNA-seq, scATAC-seq, scJoint, Seurat, Liger and Conos. (d) Trajectory conservation score for scJoint, Seurat, Liger and Conos, using the trajectories built by scATAC-seq (left) and scRNA-seq (right) as reference.



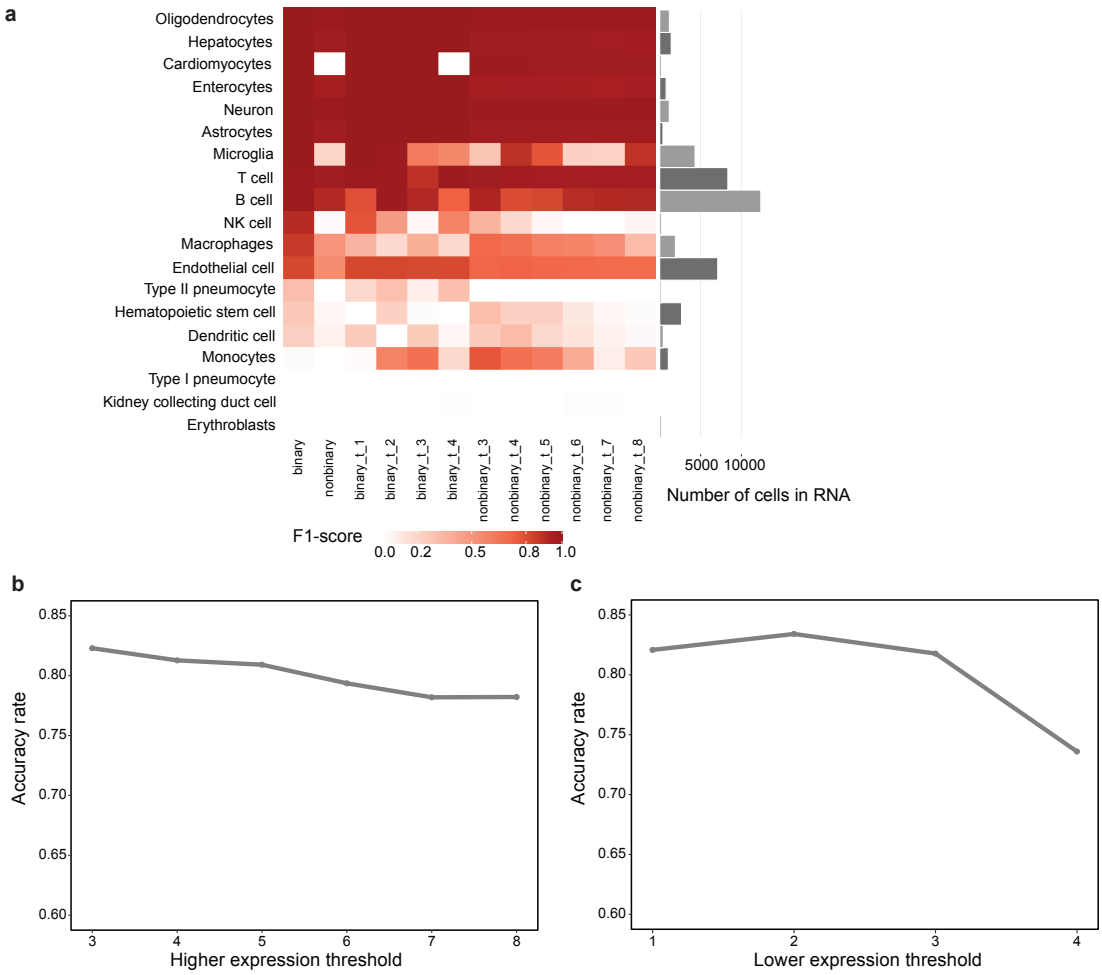
Supplementary Figure S22: Analaysis of monocyte lineages. (a) Diffusion maps of unintegrated scRNA-seq (left) and scATAC-seq data (right). (b) Diffusion maps of integrated data, colored by cell types (top) and modality (bottom), generated by scJoint, Seurat, Liger and Conos. (c) Distribution of pseudotime for scRNA-seq, scATAC-seq, scJoint, Seurat, Liger and Conos. (d) Trajectory conservation score for scJoint, Seurat, Liger and Conos, using the trajectories built by scATAC-seq (left) and scRNA-seq (right) as reference.



Supplementary Figure S23: tSNE plots of mouse cell atlas subset data with $\lambda = 1, 5, 15, 20$



Supplementary Figure S24: UMAP plots of human hematopoiesis developmental data with $\lambda = 5, 10$



Supplementary Figure S25: (a) Heatmap of F1-score for each cell type using different binarization construction methods. (b-c) Accuracy rate changes as the (b) higher expression threshold and (c) lower expression threshold varies to construct binarized matrix.

References

1. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature biotechnology* **37,** 1458–1465 (2019).

2. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *BioRxiv* (2020).

3. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13,** 845–848 (2016).

4. Wang, H. *et al. Cosface: Large margin cosine loss for deep face recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 5265–5274.