# Supplementary Information for "Unbiased integration of single cell multi-omics data"

Jinzhuang Dou[1], jdou1@mdanderson.org

Shaoheng Liang[1], sliang3@mdanderson.org

Vakul Mohanty[1], vmohanty@mdanderson.org

Xuesen Cheng[2], xuesenc@bcm.edu

Sangbae Kim[2], Sangbae.Kim@bcm.edu

Jongsu Choi[2], Jongsu.Choi@bcm.edu

Yumei Li[2], yumeil@bcm.edu

Katayoun Rezvani[4], krezvani@mdanderson.org

Rui Chen[2,3], ruichen@bcm.edu

Ken Chen[1,*]. kchen3@mdanderson.org

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center

[2]HGSC, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA

[3]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, 77030, USA

[4]Department of Stem Cell Transplantation and Cellular Therapy, The University of Texas MD Anderson Cancer Center, Houston, Texas

* Correspondence: kchen3@mdanderson.org

# Supplementary Notes

## Supplementary Note 1 Previous studies on multi-omics integration

Recent study [1] evaluated 14 single-cell batch-effect correction/integration methods, showing that Harmony [2], LIGER [3], and Seurat3.0 [4] are the recommended methods for batch integration in general. Thus, we compare bindSC with these three available state-of-the-art methods.

***Harmony***

Harmony [2] uses an iterative clustering approach to align cells from different batches. The algorithm first combines the batches and projects the data into a dimensionally reduced space using PCA. It then uses an iterative procedure to remove the multi-dataset specific effects. In our analysis, we ran Harmony within the Seurat3.0 based on the guide (http://htmlpreview.github.io/?https://github.com/immunogenomics/harmony/blob/master/docs/SeuratV3.html).

***Seurat3.0***

Seurat [4] uses CCA to first compute the linear combinations of genes with the maximum correlation between batches then adopts mutual nearest neighbor (MNN) to align the cells between datasets based on anchor cells identified. In our analysis, we used the Seurat package version 3.0 in the *R* language environment to perform multi-omics integration. Adhering to the suggested integration workflow (https://satijalab.org/seurat/v3.2/atacseq_integration_vignette.html).

***LIGER***

LIGER [3] uses integrative non-negative matrix factorization (iNMF) to first learn a low-dimensional space where each gene is characterized by two sets of factors. The first set contains dataset-specific factors, and the second contains shared factors. The shared factor space is then used to identify similar cell types across datasets by first constructing a shared factor neighborhood graph to connect cells with similar factor loading patterns. Joint clusters are then identified using the Louvain community detection. Thereafter, the factor loading quantiles are normalized using the largest data batch as a reference to achieve batch-correction. In our work, we followed the LIGER documentation (http://htmlpreview.github.io/?https://github.com/MacoskoLab/liger/blob/master/vignettes/Integrating_scRNA_and_scATAC_data.html). For preprocessing, we used the LIGER preprocessing functions, where we first selected genes with high variance. We then performed iNMF-based factorization using an alternating least squares algorithm, followed by data alignment using joint clustering and quantile alignment.

## Supplementary Note 2 Evaluation of peak-gene correlations based on pseudo-cell profiles

On the DEX-treated lung adenocarcinoma (A549) dataset, we ran bindSC to derived co-embeddings. The shared nearest neighbor (SNN) graph was constructed by calculating the l-nearest neighbors ($l$ = 20 by default) based on the Euclidean distance of L2-normalized co-embedding coordinates. The modularity optimization technique Leiden algorithm was used to group cells into interconnected clusters (termed meta-cluster) based on constructed SNN graph (*resolution* = 0.5). The Leiden algorithm was performed again on each cluster with a higher resolution (= 2) to further generate pseudo-cells. Finally, we got 206 pseudo-cells which included a median of 27 cells from scRNA-seq and 16 cells scATAC-seq dataset (**Supplementary Fig. 6e**). We observed only one cell that was modality specific (scRNA-seq) and removed it for downstream analysis. The RNA-seq and ATAC-seq profiles of each pseudo-cell were aggregated respectively. In this way, each pseudo-cell had paired gene expression and chromatin accessibility profiles. The same strategy was used to construct pseudo-cell profiles for Seurat, LIGER, and Harmony. For

Seurat, LIGER, and Harmony, 41/198, 1/89, and 15/142 modality specific pseudo-cells were removed, respectively. A high proportion of modality-specific pseudo-cells indicates that two modalities were not well aligned in co-embeddings.

We then explored peak-gene correlation based on pseudo-cell profiles from each method. For each peak-gene pair, Spearman rank correlation coefficients (SRCC) between a normalized ATAC peak level and a gene expression levels of all the pseudo-cells were calculated. There are 4,759 genes and 24,953 peaks in the peak-gene correlation matrix. The SRCC of each peak-gene pair calculated based on 1,429 co-assayed cell profiles was used as the gold standard including 1,836,974 cis peak-gene pairs and 118.7 million trans peak-gene pairs. The overall concordance between each method and the gold-standard was further quantified using a single SRCC across all peak-gene pairs (**Fig. 3c**). In most cases, the correlation of peak-gene may include many false positive and indirect targets. We therefore focused on peak-gene pairs that were supported by Hi-C data from an independent study [5]. There were 585 trans peak-gene pairs associated with the top 200 *NR3C1* target binding genes identified. Among these trans peak-gene pairs, bindSC has the best agreement with that from co-assayed cell profiles among all methods (**Supplementary Fig. 5**).

To explore TF-gene correlation at the pseudo-cell level, we obtained motif-based TF activity matrix calculated based on peak profiles using ChromVAR [6]. The final TF activity matrix included profiles for 386 TFs. Pseudo-cell level TF activity was obtained by aggregating cell profiles in each pseudo-cell. The SRCC was calculated for each TF-gene pair on pseudo-cell level. Overall, SRCC was 0.67 for bindSC and less than 0.59 for other methods (**Fig. 3c**). The SRCC of TF-gene pairs was higher than that from peak-gene pairs partly due to the fact that motif-based TF activity was derived from genome-wide motif regions and it was less noisy than single peak region.

**Supplementary Note 3 Joint profiling of chromatin accessibility and transcription in DEX-treated A549 cells**

Besides using the DEX-treated A549 cell dataset as the gold standard for method performance evaluation, we performed downstream analysis to show how bindSC improved previous studies by integrating transcriptomic and epigenomic datasets. Joint clustering module in bindSC defined 5 clusters (**Supplementary Fig. 6a**). Cells from the two technologies were well mixed together within each cluster. This classification result was in good concordance with the treatment time: cluster 1 consists of cells from mostly 0-hour (92%), and clusters 3-5 include cells from 1 and 3 hours (> 99%). Clusters 2 included cells from multiple time points and may represent transitional states (**Supplementary Fig. 7b**). The list of transcription factors (TFs) that are associated with the joint chromatin accessibility and gene expression changes and their activity levels across states can be derived at pseudo-cell resolution, and so can the genes differentially expressed in each cluster (**Supplementary Fig. 7d)**. Such co-embedding yielded higher granularity in delineating cell states and associated TFs than did embeddings derived from only one modality or based on the treatment times.

**Supplementary Note 4 Integrating single cell epigenomic data with single cell transcriptomic data on the mouse skin cell dataset**

We examined the performance of bindSC in integrating the scRNA-seq and scATAC-seq data derived from mouse skin tissue. This dataset was generated using SHARE-seq [7] which included

34,774 cells that have joint profiles of RNA and ATAC profiles. The final ATAC-seq matrix (i.e., **Y**) includes 25,594 cells on 74,161 peaks after quality control (including removing cells with less than 350 genes expressed; peaks that exist in less than 500 cells). In addition, 4,894 genes were identified that were highly variable in both gene expression and gene activity profiles (i.e., both **X** and **Z** includes 25,594 cells on 4,894 genes). For this evaluation, we only focused on the third metric (i.e., anchoring accuracy) that represents the chance for the two instances of a co-assayed cell to appear from the co-embeddings. The dimensionality $E$ was set to 15 for bindSC. BindSC achieved substantially shorter anchoring distance than the other methods (**Supplementary Fig. 7**).

## Reference

1 Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology* **21**, 1-32 (2020).

2 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 1-8 (2019).

3 Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).

4 Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).

5 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

6 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).

7 Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* (2020).

## Supplementary Figure/Table Legends

**Fig. S1 Implementation of bindSC.** Bi-CCA iteration procedure (**a**). Implementation of divide-and-conquer SVD in bi-CCA for large matrix SVD decomposition **(b)**.

**Fig. S2 Simulation settings.** Simulation of gene score matrix **Z (a)**. Each row in **X** denotes a gene (feature) and each column a cell. MR: misalignment rate; SNR: Signal-Noise-Ratio. Previous methods including CCA, Seurat, LIGER, and Harmony take **X** and **Z** as input assuming that cell alignment is unknown **(b)**. bindSC takes two parts as input: 1) **X** and **Z** with cell alignment unknown; 2) **X** and **Z** with feature alignment unknown **(c)**.

**Fig. S3 Benchmarking bindSC performance on simulation datasets.** Comparison of bindSC to CCA, Seurat, LIGER, and Harmony based on Silhouette score and alignment mixing score **(a)**. The dataset contains 1,000 genes and 1,000 cells equally distributed in 3 cell types. Signal-to-noise ratio (SNR) was set at 0. X-axes denote the misalignment rates (MR) between features in the two datasets, which ranges from 0 to 1. The features between two datasets have perfect match if MR = 0 and are unrelated if MR = 1. UMAP views of the co-embeddings generated by bindSC, CCA, Seurat, LIGER, and Harmony **(b)**. From top to bottom are results with MR = 0.1, 0.5, and 0.9, respectively. Each point denotes one cell that is colored based on its true cell type label (red, green, or cyan).

**Fig. S4 Benchmarking bindSC performance on simulation datasets.** Comparison of bindSC to CCA, Seurat, LIGER, and Harmony based on Silhouette score and alignment mixing score **(a)**. The dataset contains 1,000 genes and 1,000 cells equally distributed in 3 cell types. Signal-to-noise ratio (SNR) was set at 0.5. X-axes denote the misalignment rates (MR) between features in the two datasets, which ranges from 0 to 1. The features between two datasets have perfect match if MR = 0 and are unrelated if MR = 1. UMAP

views of the co-embeddings generated by bindSC, CCA, Seurat, LIGER, and Harmony **(b)**. From top to bottom are results with MR = 0.1, 0.5, and 0.9, respectively. Each point denotes one cell that is colored based on its true cell type label (red, green, or cyan).

**Fig. S5 Estimation of trans peak-gene regulatory elements supported by the Hi-C data.** We selected the top 200 *NR3C1* target genes based on co-assayed cell profiles and identified 585 trans peak-gene regulatory elements that were supported by the published Hi-C data [5]. X-axes are the SRCCs of peak-gene pairs estimated from the co-assayed cells, which serve as the gold standard, while Y-axes are the SRCCs estimated from pseudo-cells generated by each method. The overall concordance between X and Y are further quantified using a single SRCC shown on the up-left corner of each subfigure. Also, the peak-gene pair CFLAR@chr2:217,704,437-201,770,992 is highlighted in each subfigure. Pearson's correlation was performed to produce the coefficients (R) and the P values.

**Fig. S6 Joint profiling of gene expression and chromatin accessibility data at the pseudo-cell resolution on the A549 lung cancer cell-line.** UMAPs of cells coloring by cluster IDs obtained from unsupervised clustering (meta-cluster) in the bindSC co-embedding **(a)**. Proportion of cells from the 3 treatment times in each of the meta-cluster **(b)**. Histogram showing the number of cells in each pseudo-cell **(c)**. Heatmap showing known genes and TFs associated with glucocorticoid receptor (GR) activation process **(d)**. Each row is one gene/TF and each column is one pseudo cell, grouped/colored by cluster ID. Scatter plot showing the number of cells derived from the scRNA-seq and the scATAC-seq data for each pseudo-cell **(e)**. Each dot denotes one pseudo-cell and the dot size denotes number of cells included in it.

**Fig. S7 Integrating single-cell RNA-seq and ATAC-seq data of a mouse skin cell atlas.** UMAP of the mouse skin cells before performing integration, colored by clusters deriving from unsupervised clustering

of the RNA data and the ATAC data, respectively **(a)**. Anchoring distances achieved by bindSC, Seurat, LIGER and Harmony **(b)**. UMAP of cells in the multiomics co-embeddings generated by bindSC **(c)**, Seurat **(d)**, LIGER **(e)**, and Harmony **(f)**, respectively. For each method, the left panel shows cells from the RNA-seq data and the right panel shows cells from the ATAC-seq data.

**Fig. S8 Cell type annotation for cells in the mouse retina cell atlas**. In the heatmap, X-axes denote cluster IDs in the RNA clusters, while Y-axes denotes known retinal cell-type-specific marker genes for the AC, BC, cone, HC, RGC, rod, and RPC cells, respectively. The color gradient in each dot denotes the expression level and the dot size denotes percentage of cells that express the gene.

**Fig. S9 UMAP visualization of mouse retina cells in the *in silico* co-embeddings generated by Seurat, LIGER, and Harmony.** From top to bottom are the results for Seurat (**a**), LIGER (**b**), and Harmony (**c**) respectively. The left panel shows cells from the RNA-seq data. The right panel shows cells from the ATAC-seq data. Cells were colored based on cell types identified in **Supplementary Fig. 8**. The oval regions were zoomed in **Fig. 4 g-j**.

**Fig. S10 Integrating 10x Visium spatial transcriptomics data with SMART-Seq2 scRNA-seq data from mouse frontal cortex cells.** Schematic representation of data used for integration **(a)**. Histology image of mouse frontal cortex overlaying with cells from the 10x Visium technology **(b)**. UMAP view of the RNA expression of the 1,072 spots in the 10x Visium data **(c)**. UMAP view of the transcriptomes of 14,249 frontal cortex cells produced by SMART-Seq2 technology **(d)**. Cell-type labels in **(d)** are derived from the published SMART-Seq2 dataset.

**Fig. S11 Performance of four methods on integrating spatially resolved transcriptomic (ST) data with dissociated scRNA-seq data from mouse frontal cortex cells.** Related to **Fig. 5a**. UMAP of cells from

mouse frontal cortex datasets, separated by sequencing technology with ST on the top panel and scRNA-seq data on the bottom panel **(a)**. Cell-type labels are consistent with those from **Supplementary Fig. 10c-d**. Comparison of cell-type classification based on Silhouette scores **(b)**. Comparison of dataset alignment based on alignment mixing scores **(c)**. Gene expression profiles of three Lamp5-related marker genes *Lsp1*, *Npy2r*, and *Dock5* from the scRNA-seq data (**d**) and the ST data (**e**).

**Fig. S12 Cell types mapped by Seurat onto mouse brain histology images.**

**Fig. S13 Cell types mapped by LIGER onto mouse brain histology images.**

**Fig. S14 Cell types mapped by Harmony onto mouse brain histology images.**

**Fig. S15 Performance of three methods on integration of transcriptomic and proteomic data.** The cluster colors for each modality are consistent with those in **Fig. 6**.

**Table S1 Summary of datasets evaluated in this study.** Also listed are the key parameters for running bindSC, Seurat, LIGER, and Harmony on each dataset.

**Table S2 Simulation results with 5,000 cells**.

**Table S3 Simulation results with 10,000 cells**.