Integrating multiple single-cell multi-omics samples with Smmit

Changxin Wan[1] and Zhicheng Ji[1,2,*]

[1]Program of Computational Biology and Bioinformatics, Duke University School of Medicine, Durham, 27705, NC, USA and [2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, 27705, NC, USA
[*]Corresponding author. zhicheng.ji@duke.edu

Abstract

Multi-sample single-cell multi-omics datasets, which simultaneously measure multiple data modalities in the same cells and in multiple samples, facilitate the study of gene expression and gene regulatory activities on a population scale. Existing integration methods can integrate either multiple samples or multiple modalities, but not both simultaneously. To address this limitation, we developed Smmit, a computational pipeline that leverages existing integration methods to simultaneously integrate both samples and modalities and produces a unified representation of reduced dimensions. We demonstrate Smmit's capability to integrate information across samples and modalities while preserving cell type differences in two real datasets. Smmit is an R software package that is freely available at Github: https://github.com/zji90/Smmit

Key words: single-cell sequencing, multi-omics, integration

Introduction

Single-cell multi-omics sequencing measures multiple modalities of molecular profiles in the same cells. Examples of such technologies include joint profiling of gene expression and protein abundances (e.g., CITE-seq[1]) and joint profiling of gene expression and chromatin accessibility (e.g., SHARE-seq[2] and SNARE-seq[3]). Multi-sample single-cell multi-omics datasets have been generated to study and compare gene expression and gene regulatory activities across samples in neural development[4], leukemia[5], skin fibroblast[6], and other biological systems. These datasets enable population-level studies of cells' comprehensive functions and characteristics under different conditions, providing more in-depth insights than data derived from a single modality or a single sample.

To effectively analyze single-cell sequencing data from multiple samples and modalities, one needs to first integrate information across samples and modalities to generate a single representation of reduced dimensions. The integration process harmonizes discrepancies across samples and modalities, facilitating downstream analysis such as cell clustering, cell type identification, and characterization of the regulatory behavior within cell subpopulations. Several integration methods have been proposed that either integrate multiple samples within a single modalities (e.g., Seurat V3[7] and Harmony[8]) or integrate multiple modalities within a single sample (e.g., Seurat weighted nearest neighbour [9]). However, no method can perform integration across both samples and modalities simultaneously. Without integration across modalities, data from multiple modalities have to be analyzed separately. Without integration across samples, technical variation across samples and sequencing platforms may bias the results of downstream analysis.

To address this challenge, we introduce Smmit, a computational pipeline designed for single-cell multi-sample and multi-omics integration. Smmit builds upon existing integration methods of Harmony[8] and Seurat[9]. Smmit first employs Harmony to integrate multiple samples within each data modality. Smmit then applies Seurat's weighted nearest neighbor (WNN) function to integrate multiple data modalities and produces a single UMAP space. We demonstrate that Smmit effectively integrates multi-sample single-cell multi-omics data in real applications.

Methods

Smmit pipeline

The input for Smmit is a list of Seurat objects prepared by the user. Each Seurat object contains single-cell multi-omics data from a single sample, processed using the standard Seurat[9] or Signac[10] pipeline. Smmit then uses the `Seurat::merge()` function to merge the input list of Seurat objects into a single merged Seurat object. For antibody-derived tags (ADT) in CITE-seq[1], only proteins that are shared across all Seurat objects are used in the merging process.

For merged RNA data, the following Seurat functions are applied sequentially with default parameters to produce

a joint principal component (PC) space: `NormalizeData()`, `FindVariableFeatures()`, `ScaleData()`, `RunPCA()`. For merged ADT data, the following Seurat functions are applied sequentially to produce a joint PC space: `NormalizeData()`, `ScaleData()`, `RunPCA()`. The `NormalizeData()` function is executed using a centered log ratio transformation that normalizes across cells, while the other two functions employ default parameters. For merged ATAC data, the following Signac functions are applied sequentially with default parameters to produce a joint latent semantic indexing (LSI) space: `FindTopFeatures()`, `RunTFIDF()`, `RunSVD()`. Within each modality, the `harmony::RunHarmony()` function is then applied to the joint space of reduced dimensions to integrate samples.

Finally, the Seurat function `FindMultiModalNeighbors()` is used to integrate the Harmony integrated reduced dimensions of RNA and ADT or RNA and ATAC. For RNA and ADT integration, if ADT has less than 30 PCs, all PCs in ADT are used. The number of PCs used for RNA and ADT is set to be the same. For RNA and ATAC integration, the top 30 PCs are used for RNA, and the top 2 to 30 LSI components are used for ATAC. The Seurat function `RunUMAP()` is then used to produce a UMAP space based on the WNN results.

Data processing

To demonstrate the utility of Smmit, a processed CITE-seq dataset with two samples of human peripheral blood mononuclear cells (PBMC) and human cord blood mononuclear cells (CBMC) was downloaded from [1]. Only human genes were retained, and cells with positive RNA expression in at least 500 human genes were retained. A processed single-cell multi-omics dataset with two mouse cerebral cortex samples was downloaded from the ENCODE portal[11]. Cells with ATAC read counts between 1000 and 100000, RNA read counts between 1000 and 25000, nucleosome signals less than 2, and TSS enrichment scores greater than 1 were retained.

Results

We first applied Smmit to a CITE-seq dataset with two samples of PBMC and CBMC. The UMAP of RNA and ADT modalities clearly show separation of cells from two samples before sample integration (Figure 1a). Smmit first applies Harmony for sample integration within each data modality (Figure 1b) and subsequently applies WNN to integrate the two data modalities into the same UMAP space (Figure 1c). Cells from the two samples are mixed (Figure 1c), while different cell types of monocytes, B cells, T cells, and NK cells are separated in the UMAP space (Figure 1d). These results demonstrate that Smmit successfully integrate information across multiple samples and modalities while preserving differences across cell types.

We then applied Smmit to a dataset with two mouse cerebral cortex samples that profiles both gene expression and chromatin accessibility. Cells from the two samples are separated in the ATAC UMAP space (Figure 1e). Smmit was applied to integrate multiple samples (Figure 1f) and subsequently multiple modalities (Figure 1g). Again, cells from the two samples are mixed (figure 1g), while different brain cell types are separated in the UMAP space (Figure 1h).

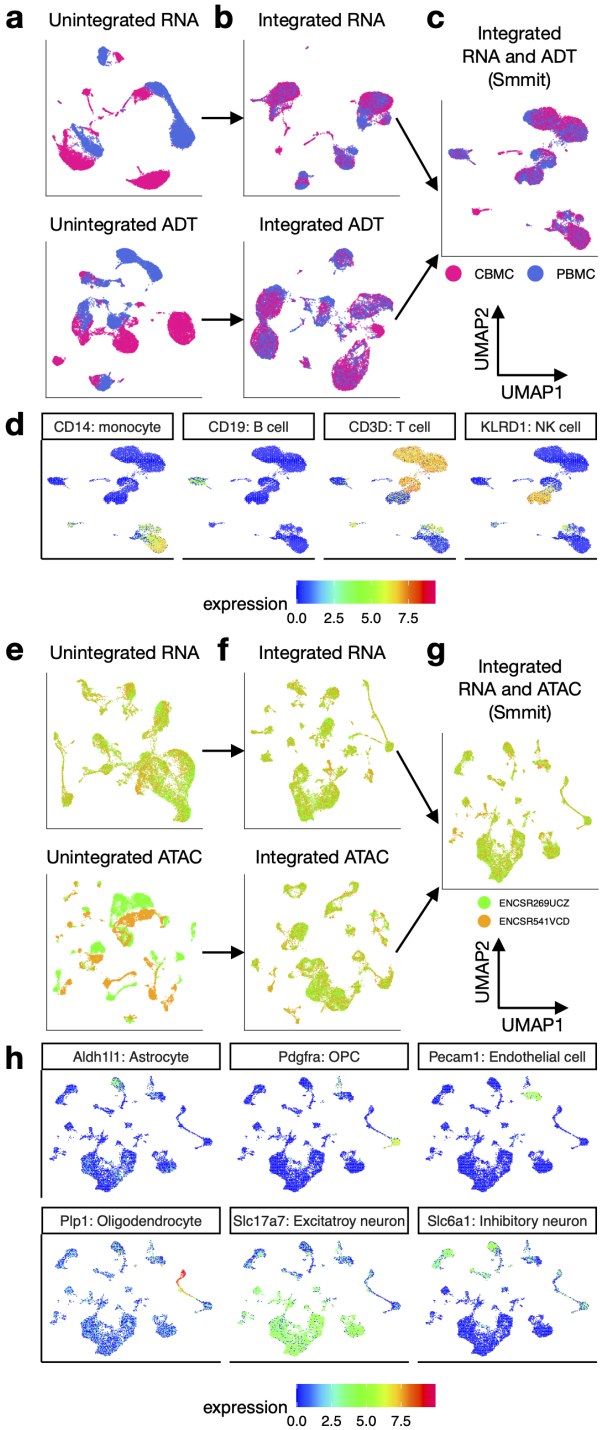

Fig. 1. a-d, Human PBMC and CBMC dataset. **a,** UMAP plot of RNA and ADT modalities before sample integration. **b,** UMAP plot of RNA and ADT modalities after sample integration. **c,** UMAP plot integrating both samples and modalities after running the Smmit pipeline. **d,** Marker gene expression of blood cell types shown on the Smmit UMAP plot using SCUBI visualization[12]. **e-h,** Mouse cerebral cortex dataset. **e,** UMAP plot of RNA and ATAC modalities before sample integration. **f,** UMAP plot of RNA and ATAC modalities after sample integration. **g,** UMAP plot integrating both samples and modalities after running the Smmit pipeline. **h,** Marker gene expression of brain cell types shown on the Smmit UMAP plot using SCUBI visualization[12].

Funding

This work was supported by the National Institutes of Health under Award Number 1U54AG075936-01.

Data availability statement

The processed CITE-seq dataset was downloaded from Gene Expression Omnibus under accession GSE100866. The processed mouse cerebral cortex single-cell multi-omics dataset was downloaded from the ENCODE portal under accession ENCSR269UCZ and ENCSR541VCD.

Competing interests

No competing interest is declared.

Author contributions statement

All authors conceived the study, conducted the analysis, and wrote the manuscript.

References

1. Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.
2. Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.
3. Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.
4. Kathryn C Allaway, Mariano I Gabitto, Orly Wapinski, Giuseppe Saldi, Chen-Yu Wang, Rachel C Bandler, Sherry Jingjing Wu, Richard Bonneau, and Gord Fishell. Genetic and epigenetic coordination of cortical interneuron development. *Nature*, 597(7878):693–697, 2021.
5. Lindsey E Montefiori, Sonja Bendig, Zhaohui Gu, Xiaolong Chen, Petri Pölönen, Xiaotu Ma, Alex Murison, Andy Zeng, Laura Garcia-Prat, Kirsten Dickerson, et al. Enhancer hijacking drives oncogenic bcl11b expression in lineage-ambiguous stem cell leukemiabcl11b hijacking in lineage-ambiguous leukemia. *Cancer discovery*, 11(11):2846–2867, 2021.
6. Chamutal Gur, Shuang-Yin Wang, Fadi Sheban, Mor Zada, Baoguo Li, Fadi Kharouf, Hagit Peleg, Suhail Aamar, Adam Yalin, Daniel Kirschenbaum, et al. Lgr5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell*, 185(8):1373–1388, 2022.
7. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
8. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.
9. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.
10. Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature methods*, 18(11):1333–1341, 2021.
11. Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, et al. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic acids research*, 48(D1):D882–D889, 2020.
12. Wenpin Hou and Zhicheng Ji. Unbiased visualization of single-cell genomic data with scubi. *Cell reports methods*, 2(1):100135, 2022.