Cell Systems
Previews

CellPress

Data-driven batch detection enhances single-cell omics data analysis

Ziqi Zhang[1] and Xiuwei Zhang[1,*]
[1]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
*Correspondence: xiuwei.zhang@gatech.edu
https://doi.org/10.1016/j.cels.2024.09.011

In single-cell omics studies, data are typically collected across multiple batches, resulting in batch effects: technical confounders that introduce noise and distort data distribution. Correcting these effects is challenging due to their unknown sources, nonlinear distortions, and the difficulty of accurately assigning data to batches that are optimal for integration methods.

In recent years, an increasing number of single-cell sequencing datasets have been made available by labs and organizations worldwide, enabling large-scale biological studies using data atlases.[1] Furthermore, some datasets are multi-omic, where more than one modality of cells is measured, e.g., the transcriptome and chromatin accessibility, while single-omic datasets measure one of the two modalities. However, integrating datasets from different sequencing batches is not straightforward. This is because different batches are often collected under different experimental environments (e.g., sequencing platforms, harvest time points, handling personnel, etc.), which introduce technical confounders, also known as batch effects.[2] Batch effects cause the distortion of data distributions between sequencing batches, where even cells of the same cell type can have different gene expression behaviors across batches (Figure 1A). Therefore, the presence of batch effects renders data from different batches difficult to compare.

Computational methods have been proposed for batch effect removal, which enables the integration of datasets into a common space to obtain a large cell atlas and downstream exploratory analysis on the integrated dataset, including joint cell-type identification, marker gene detection, and gene module detection (Figure 1B). These methods either model multi-batch datasets through traditional machine learning algorithms (e.g., canonical correlation analysis and matrix factorization) for better stability[3–5] or cling to deep neural networks to model nonlinearity.[6,7] Regardless of different modeling strategies, most existing methods assume that batch effects always and only exist between experimental batches. However, this assumption is problematic in multiple ways.

First, when the datasets include samples under different conditions, e.g., when the donors are at different stages of a disease, the differences between data of different batches can be caused by both technical confounders and biological factors associated with disease progression. For such scenarios, methods have been recently proposed to consider the existence of biological factors across batches.[8,9] These methods aim to remove batch effects while preserving the differences between batches caused by biological factors.

Second, batch effects observed in data do not always reflect the recorded experimental batches: batch effects can exist between samples within the same experimental batch, and different experimental batches may not show clear batch effects in data. These issues can cause "over-correction" or "under-correction" of batch effects when applying batch removal or integration algorithms. To tackle this problem, a recent work published in *Cell Systems* explores a data-driven approach to infer batches, such that batch effects can be observed between inferred batches from data.[10]

In their work, Wang et al. defined two batch concepts: experimentally recorded batches and data-defined batches. While experimentally recorded batches are the batches recorded during experiments, data-defined batches are the batches between which the batch effect exists (Figure 1C). Different from experimentally recorded batches, the information from data-defined batches is not directly provided with the experiment and does not always align with the experimental batches. Wang et al. proposed a batch identification algorithm to infer the data-defined batches from data. The algorithm first separates all cells from all batches into low-resolution clusters, then locally assigns batches for cells within each cluster according to an information-theory-based metric and combines local batch assignment into globally data-defined batches for the dataset. In principle, using the inferred data-defined batches should improve the performance of all existing batch effect removal algorithms since it is consistent with observed variation across samples. Wang et al. conducted tests by running several batch effect removal methods using the inferred batches and observed consistent performance improvement compared to using experimentally recorded batches.

Another major contribution of the work by Wang et al. is that they built an end-to-end pipeline, from data pre-processing to downstream analysis, to facilitate single-cell omics data analysis. The proposed batch inference method is included in the pipeline before applying data integration methods. The pipeline is named SPEEDI (single-cell pipeline for end-to-end data integration) (Figure 1C), and the goal of SPEEDI is to streamline steps in data analysis, which are otherwise performed separately. Particularly, in SPEEDI, the authors introduced an automated parameter selection strategy, which significantly reduced the burden of parameter tuning for users. The pipeline is broadly applicable to single-cell RNA sequencing, single-cell assay for transposase-accessible chromatin sequencing, and single-cell multi-omic datasets. It is available as a web portal for user convenience.

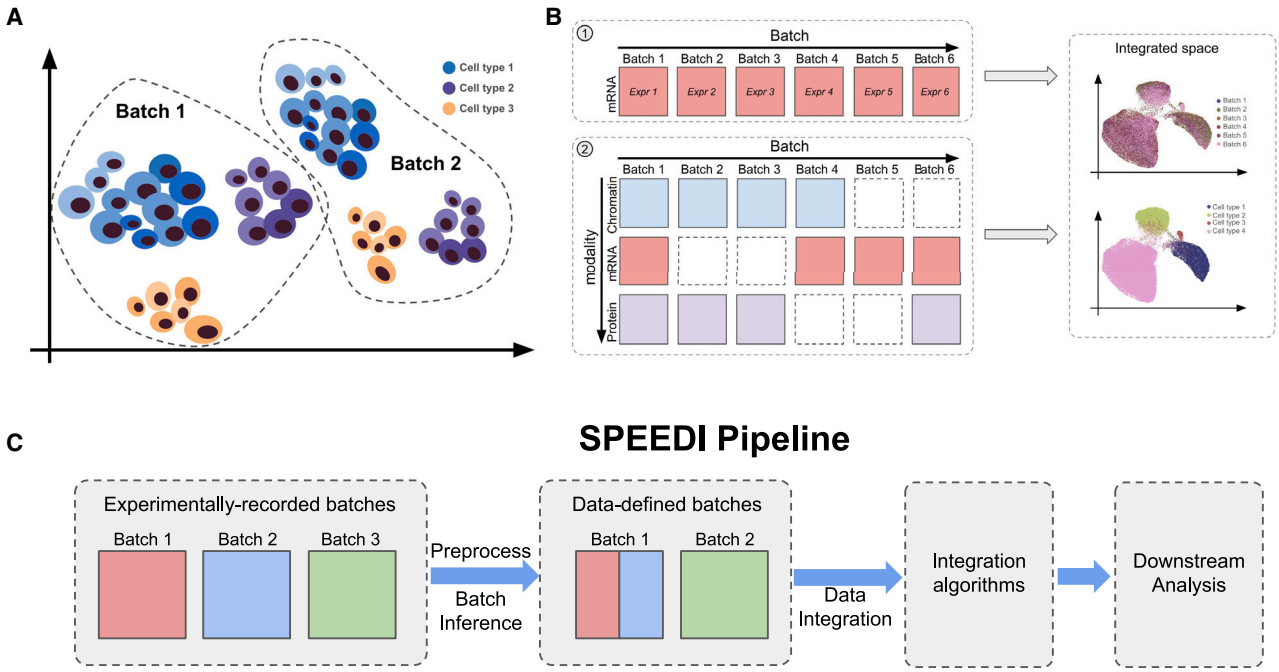

Figure 1. Data integration on single-cell sequencing datasets

(A) Batch effects cause distortion of data distributions across cell batches.

(B) Data integration algorithms integrate multiple batches of single-cell sequencing datasets from (1) the same modality or (2) multiple modalities to learn the unified representation of cells for batch effect removal and cell-type identification.

(C) SPEEDI pipeline includes data pre-processing, batch inference, data integration, and downstream analysis.

Unlike other batch removal methods, which focus on improving the algorithms for higher integration accuracy using provided experimental batch assignment, SPEEDI considers a fundamental question, which is how to define a batch for batch removal methods. Wang et al. showed improved performance with data-defined batches. In the meantime, the challenge of data integration is still far from being solved, as the sources of variations between samples can come from a complex mix of batch effects and biological effects. While Wang et al. used the variation between samples to define data-driven batches, future work should continue to study the decomposition of different sources, which can enhance understanding in disease research and clinical studies. Biological effects can be associated with different characteristics of donors in disease studies, including age, gender, disease severity, drug treatment, and others. Therefore, such studies can also benefit from appropriate wet-lab experimental design when selecting donors and organizing samples into experimental batches. While the batch inference method proposed in SPEEDI can be used in future studies, researchers can also consider developing data-driven approaches to label biological conditions for samples in cases where different biological conditions do not lead to even variation in data.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation (DBI-2145736).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Elmentaite, R., Domínguez Conde, C., Yang, L., and Teichmann, S.A. (2022). Single-cell atlases: shared and tissue-specific cell types across human organs. Nat. Rev. Genet. *23*, 395–410.

2. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

3. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.

4. Zhang, Z., Sun, H., Mariappan, R., Chen, X., Chen, X., Jain, M.S., Efremova, M., Teichmann, S.A., Rajan, V., and Zhang, X. (2023). scMoMaT jointly performs single cell mosaic integration and multi-modal biomarker detection. Nat. Commun. *14*, 384.

5. Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E.Z., and Welch, J.D. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. Nat. Protoc. *15*, 3632–3662.

6. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

7. Zhang, Z., Yang, C., and Zhang, X. (2022). scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. Genome Biol. *23*, 139.

8. Zhang, Z., Zhao, X., Bindra, M., Qiu, P., and Zhang, X. (2024). scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data. Nat. Commun. *15*, 912.

9. Liu, R., Qian, K., He, X., and Li, H. (2024). Integration of scrna-seq data by disentangled representation learning with condition domain adaptation. BMC Bioinformatics *25*, 116.

10. Wang, Y., Thistlethwaite, W., Tadych, A., Ruf-Zamojski, F., Bernard, D.J., Cappuccio, A., Zaslavsky, E., Chen, X., Sealfon, S.C., and Troyanskaya, O.G. (2024). Automated single-cell omics end-to-end framework with data-driven batch inference. Cell Syst. *15*, 982–990.e5.