OmicVerse: A single pipeline for exploring the entire transcriptome universe

Zehua Zeng [ab1], Yuqing Ma[c1], Lei Hu[a], Yuanyan Xiong [d*], Hongwu Du [ab*]

*Correspondence: xyyan@mail.sysu.edu.cn (Y. Xiong),

hongwudu@ustb.edu.cn (H. Du)

a School of Chemistry and Biological Engineering, University of Science and Technology Beijing, Beijing 100083, China

[b] Daxing Research Institute, University of Science and Technology Beijing, Beijing 100083, China.

[c] Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, Guangdong Province, 518055, China.

[d] Key Laboratory of Gene Engineering of the Ministry of Education, Institute of Healthy Aging Research, School of Life Sciences ,Sun-Yat-sen University, Guangzhou, Guangdong, 510006, China

[1] These authors contributed equally to this work

Abstract

Single-cell sequencing often results in cellular "interruptions" due to limitations in sequencing throughput, while bulk RNA-seq may contain these potential "interrupted" cells. Here, we propose the BulkTrajBlend algorithm in OmicVerse that combines Beta-Variational AutoEncoder for deconvolution and graph neural networks for overlapping community discovery to effectively interpolate and restore the continuity of "interrupted" cells in the original scRNA-seq data. Additionally, OmicVerse serves as a comprehensive tool, providing standardized and consistent access to a wide range of methods for bulk and single cell RNA-seq analysis, enhancing computational efficiency and enabling beautiful visualization, and helping researchers to explore the transcriptomic universe, unravel novel biological insights, and propel scientific discoveries.

Main

Single-cell RNA sequencing (scRNA-seq) and bulk RNA sequencing (RNA-seq) have become indispensable tools for investigating cellular heterogeneity, differentiation, and disease mechanisms[1-6]. These methods encompass a wide range of applications, including converting bulk-seq data into single-seq[7], performing differential expression analysis[8], pathway enrichment[9], co-expression gene network analysis in bulk-seq[10], cell annotation[11], cell interaction analysis[12], cell-trajectory inference[13], cell-state evaluation in gene sets, and drug response prediction in single cells[14]. Many of

these methods rely on open-source algorithms developed by the scientific community [15,16].

However, the rapidly growing number and diversity of algorithms in the field of omics present challenges in selecting appropriate, accurate, and user-friendly analysis tools. The use of different tools often requires users to learn and master various platform, resulting in computational inefficiencies. Bedides, web servers and the R language are commonly used for low-data quantity analysis[17], while Python is preferred for large-scale data analysis[18].

Furthermore, integrating results from both single-cell and bulk analyses can be complex and produce multi-layered data, making it difficult to derive meaningful biological insights. The limited number of cells available for single-cell sequencing has led to 'interruptions' in the study of cell development and differentiation trajectories. In contrast, bulk RNA-seq sequencing of whole tissues contains, in principle, 'interrupted' cells. To our knowledge, there is currently no algorithm available for extracting 'interrupted' cells from bulk RNA-seq. and there is a lack of tools that effectively bridge the gap between bulk-seq and single-seq analyses.

To address these issues, we introduce OmicVerse (https://omicverse.readthedocs.io/), a Python library designed to explore the entire transcriptome universe. OmicVerse provides standardized access to a range of methods for bulk-seq and single-seq analyses (Supplementary Note 1). All methods have been rewritten for improved computational efficiency and beautiful visualization, and we have also integrated different pre-processing methods based on the results of benchmark tests[19]. Additionally, we developed BulkTrajBlend in OmicVerse, specifically designed to address cell “interruption” in single-cell sequencing. BulkTrajBlend deconvolves single-cell data from Bulk RNA-seq and utlilzes a graph neuron network-based overlapping community discovery algorithm to identify “interrupted” cells in the generated single-cell data. The method of BulkTrajBlend is described in Supplementary Note 2.

To evaluate the efficacy of BulkTrajBlend in generating “interrupted” cells, we utilized single-cell RNA-seq data (GEO accession: GSE95753) obtained from the dentate gyrus of the hippocampus in mice, along with bulk RNA-seq data (GEO accession: GSE74985). In the single-cell data, the neural differentiation trajectory from nIPC (neuronal intermediate progenitor cells) to neuroblast to Granule immature to Granule mature was observed, with other cell types disconnected from nIPC (Fig.1i). A previous study has identified the nIPC to CPN (cortical projection neuron) and gIPC (glial intermediate progenitor cell) to OPC lineages[20]. Our aim was to validate the nIPC differentiation trajectory by linking OPC cells to nIPC cells in the “interruption” single cell data.

Using BulkTrajBlend, we initially deconvoluted the single-cell data from the bulk RNA-seq of the dentate gyrus, which consisted of 13 distinct cell types. We then filtered out noisy subpopulations and identified cell types with similar high expression patterns as those found in the original single-cell data marker genes (Fig. 1c and Fig. 1d, respectively). Subsequently, BulkTrajBlend calculated the overlap of cell types in the generated single-cell data (Fig. 1e) and visualized it as an adjacency matrix in the form

of a heat map (Fig. 1f). Within the generated single-cell data, we annotated the overlapping cell community and specifically selected the single-cell data in which OPCs were associated with nIPC. We integrated this selected data into the original single-cell data of the dentate gyrus. Remarkably, we observed that the OPCs in the integrated scRNA-seq of the dentate gyrus were situated on the differentiated branch of nIPC (Fig. 1i-1j). This finding indicated that BulkTrajBlend effectively generates a continuous cell trajectory, thereby addressing the issue of "interrupted" cells.

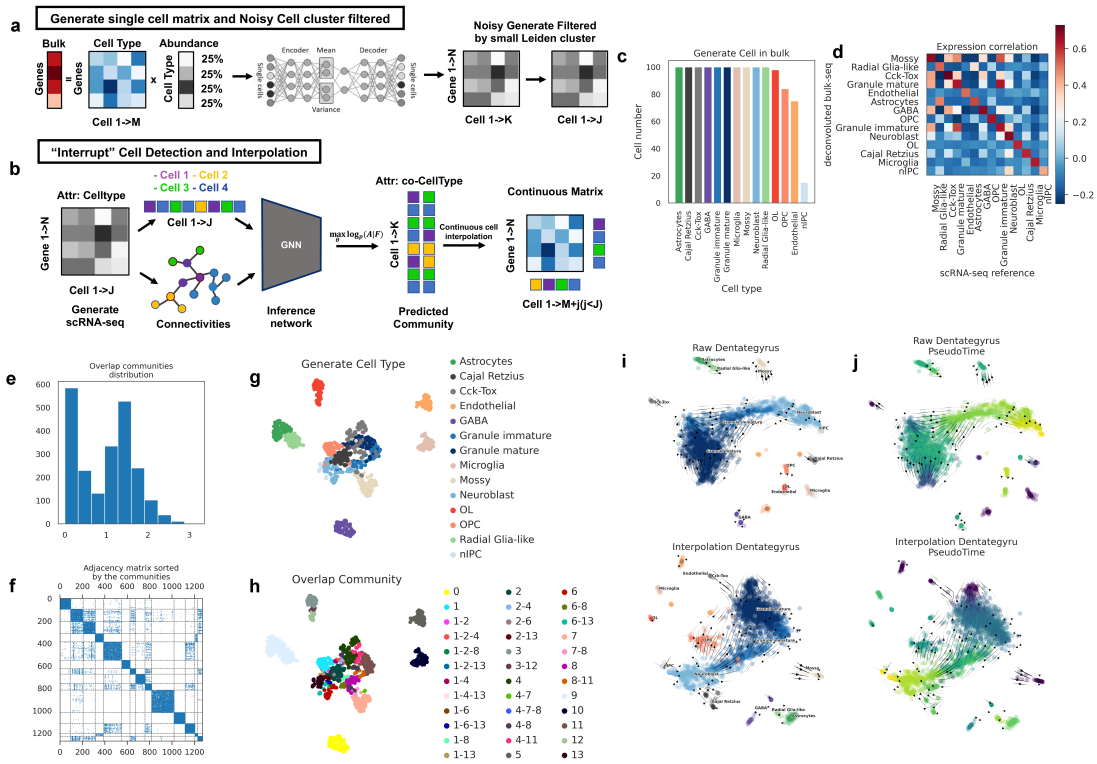

Figure 1 | **Overview of BulkTrajBlend and application to dentate gyrus neurogenesis**. a, Graphical model representation of Generate scRNA-seq in BulkTrajBlend. b, Graphical model representation of "interrupt" cell detection using GNN in BulkTrajBlend. c, The number of celltype in Generate scRNA-seq from Bulk RNA-seq. d, Pairwise expression correlation of cell-type-specific marker genes between single cells generated by BulkTrajBlend and the single-cell reference for dentate gyrus neurogenesis. e, the overlap cell community's distribution in Generate scRNA-seq. f, the heatmap showing each cell in different cell communities of adjacency matrix from GNN. g, UMAP embedding the cell type in generate scRNA-seq. h, UMAP embedding the overlap cell type communities in generate scRNA-seq. i, Force-directed graph of raw scRNA-seq (Upper) and interpolation scRNA-seq (Bottom) of dentate gyrus neurogenesis colored in cell type. j, the pseudo-time of raw scRNA-seq (Upper) and interpolation scRNA-seq (Bottom) of dentate gyrus neurogenesis

OmicVerse establishes a bridge between single-seq and bulk-seq[7]. For each omics layer, OmicVerse constructs an analysis object downstream of the expression matrix (Fig. 2a), enabling independent analysis of each omics layer while maintaining ease of

navigation. Additionally, OmicVerse supports 14 data formats, which can interact with annotated datasets or Pandas formats, ensuring a consistent user interface. (Fig. 2b).

We also highlight several new features of OmicVerse that are applicable to bulk-seq and single-seq analyses. The first feature enables the deconvolution of bulk-seq data, providing an integrated "query" scRNA-seq map. In Supplementary Note 3, we apply this feature to an Alzheimer's disease (AD) development dataset and pancreatic ductal adenocarcinoma (PDAC) dataset to verify cell types identified in RNA-seq analysis. We validate these results using scRNA-seq data from the same tissue. The second feature extends a robust pipeline for scRNA-seq analysis, incorporating pySCSA, pyAUCell, pyVIA, and CellPhoneDB. This feature aids researchers in evaluating the quality and efficiency of scRNA-seq data. Supplementary Notes 4 present a case study demonstrating the application of this pipeline by projecting data from colorectal cancer patients into a tumor microenvironment cell atlas. OmicVerse also includes a feature for performing comprehensive bulk RNA-seq analysis using pyDEG, pyWGCNA, pyPPI, and pyGSEA. Supplementary Note 5 presents a case study using Alzheimer' disease patients to illustrate the application of this pipeline. Furthermore, OmicVerse integrates MOFA[21] for multi-omics analysis, including scRNA-seq, scATAC-seq, and spatial-seq, allowing the identification of variant factors across different omics layers. Supplementary Note 6 showcases a case study on Alzheimer's disease, capturing variant factors paired with GLUE[22].

The OmicVerse documentation website provides an application programming interface (API) reference for each algorithm, along with tutorials describing the functionality and limitations of each algorithm and its interaction with other bulk/single-seq tools. These tutorials are also accessible through Google Colab, providing a free computing environment to support pipeline studies.

Our aim was to create an ecosystem for bulk/single-seq analysis and beautiful visualization within Python environments. Users can perform comprehensive transcriptome analysis using a single programming language, leveraging the machine-learning models and expertise of the Python community. As OmicVerse continues to evolve, we anticipate ongoing updates that incorporate new algorithms, features, and models. OmicVerse is expected to benefit the bulk/single-seq community by facilitating the prototyping of new models, establishing standards for the analysis of trans omics, and enhancing the scientific discovery pipeline.

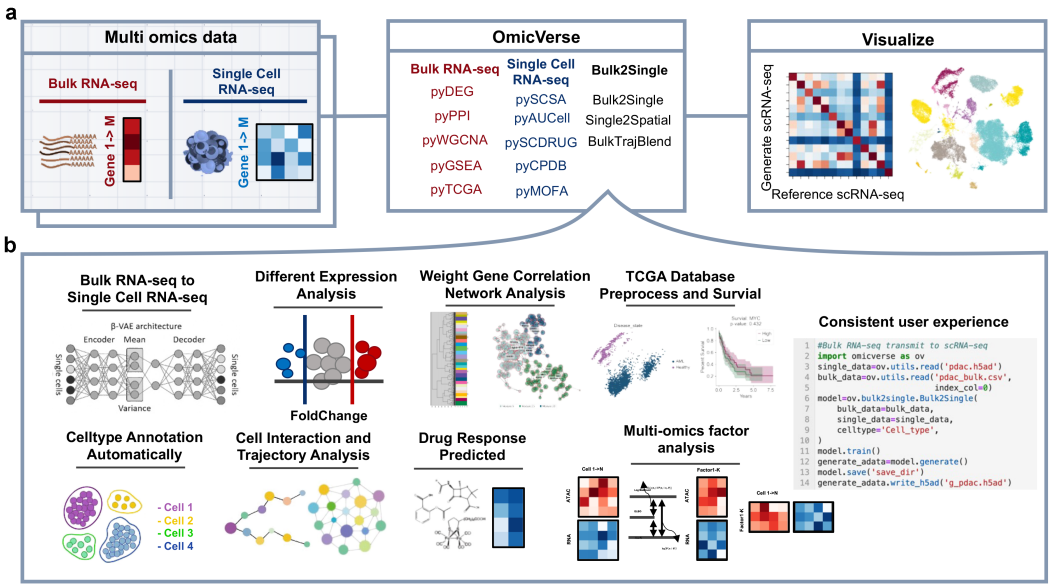

Figure 2 | **OmicVerse overview. a**, OmicVerse establishes a bridge between the bulk and single-cell RNA-seq using a comprehensive set of algorithms offered by OmicVerse. The results can be visually represented with OmicVerse's aesthetically pleasing color palette. **b**, OmicVerse offers a comprehensive set of analysis, commonly applied in bulk and single-cell RNA-seq analysis. On the right of the a sample code snippet is provide, demonstrating the application of the "bulk2single" algorithm to a dataset read from both h5ad and csv formats.

Acknowledgments

We thank all the Github users who contributed code to omicverse over the years. This work was financially supported by the Fundamental Research Funds for the Central Universities (FRF-TP-22-007A1), the Hebei Provincial Department of Science and Technology (No.19942410G), and the Student Research Training Program of University of Science and Technology Beijing.

Author contributions

Z.Z., and Y.M., contributed equally. Z.Z. designed the omicverse application programma interface, integrated the bulk2single and single2spatial model from bulk2space by optimization deep-learning model. Y.M. designed and implemented the overlap cell community of BulkTrajBlend, Z.Z. implemented and test the single-cell RNA-seq pipeline. L.H. and Z.Z. implemented and test the multi omics module of pyMOFA. H.D. and Z.Z. conceived, implemented and test bulk RNA-seq pipeline. Y.X., H.D. and Z.Z. supervised and conceived the concept of omicverse. Z.Z., Y.M., L.H., H.D., and Y.X. wrote the manuscript.

Competing intersets

The authors declare no competing interests.

Reference

1 Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods* **18**, 723-732 (2021).

2 Peng, L. *et al.* Single-cell RNA-seq clustering: datasets, models, and algorithms. *RNA biology* **17**, 765-783 (2020).

3 Xu, X., Hua, X., Mo, H., Hu, S. & Song, J. Single-cell RNA sequencing to identify cellular heterogeneity and targ ets in cardiovascular diseases: from bench to bedside. *Basic Research in Cardiology* **118**, 7.

4 Derakhshan, T., Boyce, J. A. & Dwyer, D. F. Defining mast cell differentiation and heterogeneity through single-ce ll transcriptomics analysis. *Journal of Allergy and Clinical Immunology* **150**, 739-747, doi:10.1016/j.jaci.2022.08.011.

5 Zeng, L. *et al.* Research progress of single-cell transcriptome sequencing in autoimmun e diseases and autoinflammatory disease: A review. *J Autoimmun* **133**, 102919, doi:10.1016/j.jaut.2022.102919.

6 Thind, A. S. *et al.* Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. *Briefings in bioinformatics* **22**, bbab259 (2021).

7 Liao, J. *et al.* De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution. *Nature Communications* **13**, 6498, doi:10.1038/s41467-022-34271-z (2022).

8 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

9 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpret ing genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102.

10 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

11 Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research* **51**, D870-D876, doi:10.1093/nar/gkac947 (2023).

12 Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols* **15**, 1484-1506, doi:10.1038/s41596-020-0292-x (2020).

13 Stassen, S. V., Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K. & Tsia, K. K. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nature Communications* **12**, 5528, doi:10.1038/s41467-021-25773-3 (2021).

14 Hsieh, C.-Y. *et al.* scDrug: From single-cell RNA-seq to drug response prediction. *Computational and Structural Biotechnology Journal* **21**, 150-157, doi:10.1016/j.csbj.2022.11.055.

15 Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nature Methods* **17**, 137-145, doi:10.1038/s41592-019-0654-x (2020).

16 Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology* **41**, 604-606, doi:10.1038/s41587-023-01733-8 (2023).

17 Giorgi, F. M., Ceraolo, C. & Mercatelli, D. The R Language: An Engine for Bioinformatics and Data Science. *Life (Basel)* **12**, 648, doi:10.3390/life12050648.

18 Brittain, J., Cendon, M., Nizzi, J. & Pleis, J. Data scientist's analysis toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Science Review* **1**, 7 (2018).

19 Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods* **20**, 665-672, doi:10.1038/s41592-023-01814-1 (2023).

20 Ramos, S. I. *et al.* An atlas of late prenatal human neurodevelopment resolved by single-nu cleus transcriptomics. *Nature communications* **13**, 7671, doi:10.1038/s41467-022-34975-2.

21 Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology* **21**, 1-17 (2020).

22 Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 1-9 (2022).