Supporting Materials

Word Counts

Main Text: 1010
Supporting Materials: Reflection (191),

Reflection

In this mini-project, we benchmarked five integration tools, Scanorama, scVI, Harmony, CCA, and RPCA to evaluate their effectiveness in correcting batch effects in scRNA-seq data, particularly in the context of immune cell heterogeneity. By comparing performance across metrics such as batch mixing, biological conservation, clustering stability, and cross-species integration, we identified distinct strengths and trade-offs among the tools.

However, this study has several limitations. We did not evaluate computational efficiency due to differences between Seurat and Scanpy data structures may introduce biases. Incorporating rpy2 may ensure fairer comparisons across platforms. Second, the dataset focused on immune cells, limiting generalizability. Future benchmarks should include more tissues and larger datasets. Additionally, we applied default parameters across all methods, which may not reflect optimal performance. Lastly, current clustering stability evaluation was basic and could be improved with more comprehensive statistical analysis.

This project not only strengthened my skills in single-cell data processing and computational benchmarking but also deepened my understanding of batch correction methods and the importance of preserving biological interpretation. It has inspired me to further explore integrative modeling strategies to enhance cross-species integration performance and better balance technical correction with biological fidelity.

Supplementary Methods

Data Acquisition and Preprocessing

We benchmarked data integration methods on 7 real datasets. We downloaded published scRNA seq data of 17 samples (see Supplementary Table 1 for an overview of datasets). All scRNA-seq datasets were quality controlled and normalized in the same way according to published best practices (Luecken *et al.*, 2022). Specifically, we used the general preprocessing pipeline implemented in OmicVerse, which applies scran pooling normalization followed by a $\log_{10}(x + K)$ transformation (K = 1 by default) on count data. For data solely available in TPM or RPKM units, we applied the same $\log_{10}(x + K)$ transformation without further normalization. Cell identity annotations were primarily harmonized according to published annotations and established best practices (Luecken *et al.*, 2022).

Integration methods

We ran the 5 embedding-based integration methods according to default parameterizations obtained from available tutorials or paper methods (Butler *et al.*, 2018; Lopez *et al.*, 2018; Korsunsky *et al.*, 2019; Stuart *et al.*, 2019; Hie *et al.*, 2024). To ensure fair comparison, highly variable genes were first identified independently within each batch, and genes that were detected as highly variable in more than two batches were retained. From this intersection, the top 2,000 genes were selected and used as input for all integration methods. Principal component analysis (PCA) was performed on

these genes, and the top 30 principal components were used for UMAP visualization. The output embedding dimensionality was also fixed at 30 for all methods. Further implementation details are provided in the accompanying Jupyter Notebook.

Evaluation Metrics

We evaluated integration performance using a set of metrics grouped into two broad categories: batch effect removal and biological variance conservation, following the recommendations from the scIB benchmarking framework (Luecken *et al.*, 2022). For biological conservation, we applied K-means NMI, K-means ARI, Silhouette (label), and cLISI. To assess batch effect removal, we separately evaluated correction for donor and protocol effects using Silhouette (batch), iLISI, graph connectivity, and PCR comparison. This set of metrics provides a balanced evaluation of how well integration methods preserve biological structure while mitigating batch-specific variation.

Cluster stability was assessed using a bootstrap-based approach with Jaccard similarity. For each method, K-means clustering (k = 20) was performed on the low-dimensional embedding. In each of 100 bootstrap iterations, cells were resampled with replacement and reclustered. The original and bootstrapped clusters were compared using Jaccard index, and the average score across iterations was used to quantify stability

Visualizationand

All visualizations were generated in R using base plotting functions and the ggplot2 package. UMAP plots, performance metrics, and cluster stability results were visualized consistently across methods to facilitate comparison.

Data and Code Availability

All code and data generated during this project have been uploaded to the following GitHub repository: https://github.com/SYD0831/CMML3.git. In addition, the Jupyter notebook is provided in the supplementary materials for direct access and reproduction.

Reference

Butler, A. *et al.* (2018) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*, 36(5), pp. 411–420. Available at: https://doi.org/10.1038/nbt.4096.

Hie, B.L. *et al.* (2024) 'Scanorama: integrating large and diverse single-cell transcriptomic datasets', *Nature Protocols*, 19(8), pp. 2283–2297. Available at: https://doi.org/10.1038/s41596-024-00991-3.

Korsunsky, I. *et al.* (2019) 'Fast, sensitive and accurate integration of single-cell data with Harmony', *Nature Methods*, 16(12), pp. 1289–1296. Available at: https://doi.org/10.1038/s41592-019-0619-0.

Lopez, R. *et al.* (2018) 'Deep generative modeling for single-cell transcriptomics', *Nature Methods*, 15(12), pp. 1053–1058. Available at: https://doi.org/10.1038/s41592-018-0229-2.

Luecken, M.D. *et al.* (2022) 'Benchmarking atlas-level data integration in single-cell genomics', *Nature Methods*, 19(1), pp. 41–50. Available at: https://doi.org/10.1038/s41592-021-01336-8.

Stuart, T. *et al.* (2019) 'Comprehensive Integration of Single-Cell Data', *Cell*, 177(7), pp. 1888-1902.e21. Available at: https://doi.org/10.1016/j.cell.2019.05.031.

Supplementary Figure and Table

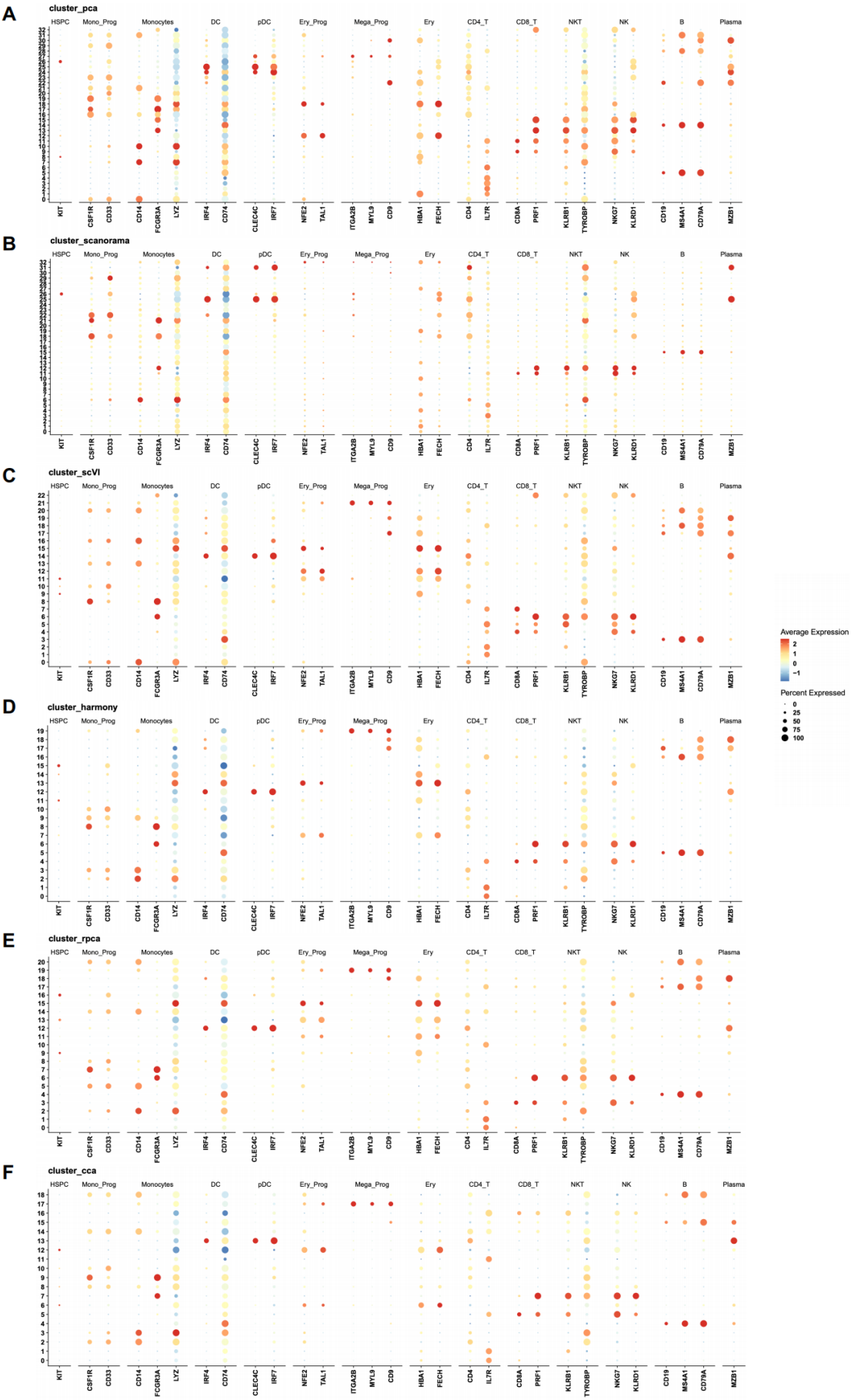

Figure S1: Dot plot of canonical marker gene expression across clusters identified by different integration methods.

Dot plots show the expression of canonical immune cell marker genes across clusters identified by (A) RPCA, (B) Scanorama, (C) scVI, (D) Harmony, (E) RPCA (Seurat v3), and (F) CCA (Seurat v3). Dot size indicates the percentage of cells expressing each gene; color represents average expression level.

Table S1: Overview of the datasets used to benchmark data integration methods

Study ID	Sample Name	Cell number	Species	Chemistry	Tissue
PMID: 30518681	Oetjen_A	2586	Human	v2_10X	Bone Marrow
PMID: 30518681	Oetjen_P	3265	Human	v2_10X	Bone Marrow
PMID: 30518681	Oetjen_U	3730	Human	v2_10X	Bone Marrow
PMID: 30967541	Sun_sample4_TC	2420	Human	10X	PBMC
PMID: 30967541	Sun_sample3_TB	2403	Human	10X	PBMC
PMID: 30967541	Sun_sample2_KC	2281	Human	10X	PBMC
PMID: 30967541	Sun_sample1_CS	1725	Human	10X	PBMC
10X	10X	10727	Human	v3_10X	PBMC
PMID: 30228881	Freytag	3347	Human	v2_10X	PBMC
PMID: 28428369	Villani	1022	Human	smart-seq2	PBMC
PMID: 29588278	Dahlin_1	7836	Mice	v2_10X	Bone Marrow
PMID: 29588278	Dahlin_2	7575	Mice	v2_10X	Bone Marrow
PMID: 29588278	Dahlin_3	7542	Mice	v2_10X	Bone Marrow
PMID: 29588278	Dahlin_4	7444	Mice	v2_10X	Bone Marrow
PMID: 29775597	MCA_BM_1	1546	Mice	microwell-seq	Bone Marrow
PMID: 29775597	MCA_BM_2	2813	Mice	microwell-seq	Bone Marrow
PMID: 29775597	MCA_BM_3	380	Mice	microwell-seq	Bone Marrow