A Reproducible Pipeline for Preprocessing and Annotation of scRNA-seq Data Using Seurat and Scanpy

Vladimir Kovačević[2], Andreja Živić[1*], Miloš Ivanović[1], Nevena Milivojević Dimitrijević[3], Marko Živanović[1]

[1] Faculty of Science, University of Kragujevac, Serbia, e-mail: andreja.zivic@pmf.kg.ac.rs
[2] BGI Genomics, Belgrade, Serbia
[3] Institute for Information Technologies, University of Kragujevac, Serbia,
* *Corresponding author*

DOI: 10.46793/ICCBIKG25.372K

Abstract: Single-cell RNA sequencing (scRNA-seq) is now a versatile platform for the dissection of cellular heterogeneity across biological conditions. Standardization of preprocessing and annotation pipelines is still to come. We present here a reproducible and modular workflow that combines the strengths of Seurat (R) and Scanpy (Python) to preprocess, annotate, and prepare scRNA-seq data for downstream analysis.

The workflow begins with raw count matrices from greater than one biological replicates or conditions. Utilizing Seurat, we perform initial quality control, low-quality cell removal, and reference-based cell type annotation from a reference scRNA-seq atlas. The annotated data is re-coded to AnnData format for an easy transition to the Scanpy framework. In Scanpy, additional operations such as normalization, feature selection, dimensionality re- duction (PCA, UMAP), and checking for batch effects are performed. The output data structure is conducive to flexible downstream analysis, including differential expression and pathway enrichment.

This pipeline ensures interoperability, reproducibility, and transparency and is particu- larly suited for group environments and comparative analysis. All of the preprocessing is thoroughly documented and parameterized to be straightforwardly modifiable for a range of datasets and research questions.

Keywords: single-cell RNA sequencing, peripheral blood mononuclear cells, gene expression, pipeline, Seurat, ScanPy.

1. Introduction

Single-cell transcriptomics enables researchers to dissect complex tissues at cellular resolution, providing information on diverse cell states, types, and dynamic processes. Despite its advan- tages, preprocessing of scRNA-seq data is filled with challenges such as dropout events, high dimensionality, and the presence of batch effects due to different replicates or sequencing runs. Several toolkits have been built to process scRNA sequence data. In particular, Seurat (R- based) and Scanpy (Python-based) are two of the most widely used frameworks that offer end- to-end pipelines for data processing, clustering, and visualization. Both tools offer overlapping functionalities but are strong in specific domains. Seurat boasts excellent reference mapping and annotation tools, and Scanpy offers rapid downstream analysis and integration with Python machine learning libraries. Here, we present a reproducible pipeline spanning Seurat and Scanpy for cell annotation and preprocessing. Our pipeline accepts raw count matrices as input, performs rigorous quality control, normalization, and reference-based cell-type annotation, and provides clean and annotated AnnData objects suitable for downstream analysis in Scanpy.

2. Methodlogy

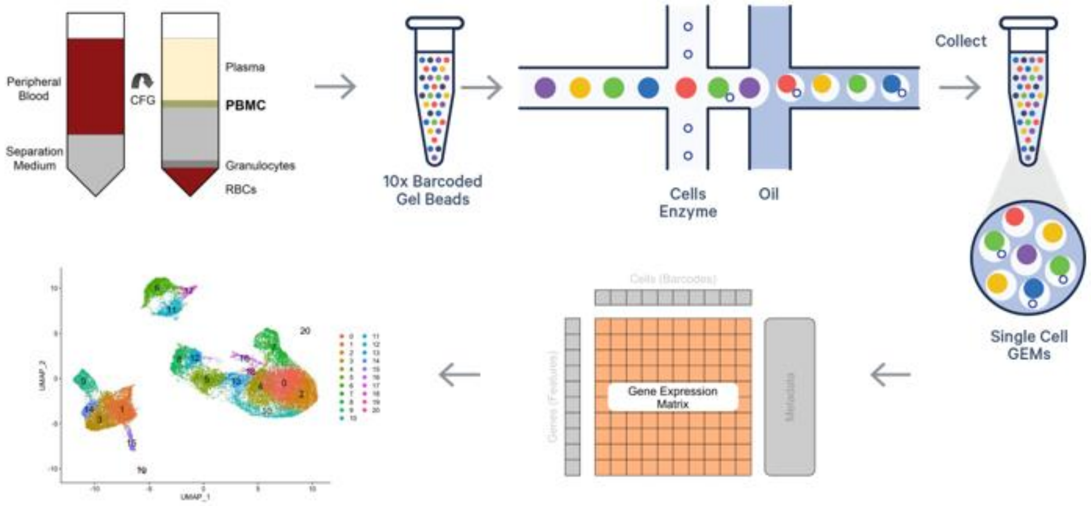

Figure 1: Schematic of the experiment design encompassed perpheral blood extraction, centrifugation and treatment with PSNPs in microfluidic chip-based platform. Single cell RNA sequencing was performed on peripheral blood mononuclear cells sample treated with different size of PSNPs and control sample. On the gene expression profile acquired we conducted bioinformatics analysis to cluster and determine cell types.

We used single-cell RNA-seq data collected from four experimental setups with PBMCs treated with nanoparticles of differing sizes. Each experimental setup is a biological replicate. The raw data were preprocessed to generate count matrices in .h5ad [5] and .csv formats.

Initial filtering measures were taken with Seurat on:

- Removal of cells with low UMI counts or with too few genes expressed (minimum of 200 expressed genes).
- Removal of genes expressed in very few cells (minimum of 3 cells).
- Depleting potential doublets and high mitochondrial content cells.

Total count normalization and log transform were done by the pipeline. Highly variable genes (HVGs) in Seurat and Scanpy were found using the Seurat method (flavor='seurat') and used for dimensionality reduction.

Principal Component Analysis (PCA) was applied to the HVGs, neighbor graph construction and UMAP for visualization later. Leiden algorithm was used for clustering in Scanpy.

Annotated PBMC reference datasets from Broad Institute (SCP424) [1] were used for cell annotation by supervised mapping using Seurat's label transfer [2, 3]. Two other annotation processes (CoDi and CoDi_dist) were also run and compared with reference-based prediction. The annotations were merged back to Scanpy-compatible .h5ad format for further downstream analysis.

UMAP plots, violin plots for quality scores, bar plots for cell proportions, and volcano plots for differential gene expression were produced using Matplotlib, Seaborn, and Scanpy's built-in functions. All intermediate results and final annotated datasets are saved in the pipeline for reproducibility and reuse.

3. Results and Discussion

The four samples were merged successfully, UMAP plots with overlap of cell types between replicates, and no batch effects suggested (Figure 2). Seuratbased cell type annotation demonstrated robust concordance with established PBMC subtypes [1]. Violin plots showed homogenous gene and UMI counts across samples. Median gene counts per cell were within the expected range for PBMCs (around 500–2,500 genes/cell).

Comparison of Seurat annotation output to CoDi [4] showed that whereas both identified the same large cell types, CoDi offered finer grain in T cell subsets, and reference based mapping by Seu- rat yielded more accurate overall identification in comparison to known markers(Table 1).

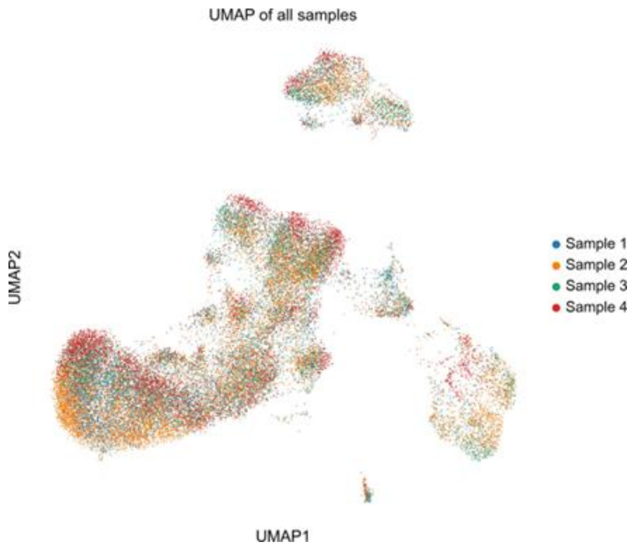

Figure 2: UMAP plot with all sample

Our approach exploits the complementary features of Seurat and Scanpy—R-based annotation and integration, coupled with scalable visualization and differential analysis within Python. Such a hybrid workflow is particularly valuable for research groups that require flexibility as well as performance across both environments.

Table 1: Scores for capturing known marker genes for detected cell types.

Annotation type	**B cell**	**CD14+ monocyte**	**CD4+ T cell**	**Cytotoxic T cell**	**Natural killer cell**	**Average**
CoDi	0.62	0.1225	0.45	0.4025	0.36	0.391
Seurat	0.62	0.1225	0.45	0.42	*Not detected*	0.4031

Even though we focused on nanoparticle-exposed PBMCs, the pipeline can be extended and applied to other tissue types or disease states. Subsequent versions may include batch correction automatically and integration with spatial transcriptomics.

4. Conclusion

We introduce an open, reproducible scRNA-seq preprocessing pipeline that integrates Seurat for cell-type annotation and Scanpy for scalable analysis. It ensures data integrity, robust annotation, and seamless entry into downstream analyses such as clustering, visualization, and differential expression.

Acknowledgment

This paper is supported through the EIT's Higher Education Initiative SMART-2M, DEEPTECH- 2M and A-SIDE projects, coordinated by EIT RawMaterials, funded by the European Union and the i-GREENPHARM project, HORIZON-MSCA-2023-SE-01-01, Grant No. 101182850 and supported by the Ministry of Education and Ministry of Science, Technological Development and Innovation, Republic of Serbia, Grants: No. 451-03-136/2025-03/200122.

References

[1] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Ne- manja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparative analysis of single cell rna-sequencing methods. *bioRxiv*, 2019.

[2] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, An- drew 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.

[3] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2):293–304, 2024.

[4] Vladimir Kovacevic, Marija Bezulj, Nikola Milicevic, Bojana Josic, Shuangsang Fang, Yong Zhang, and Junhua Li. Codi: Contrastive distance cell type annotation for spatially resolved transcriptomics. *Preprint*, 2024.

[5] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *BioRxiv*, pages 2021–12, 2021.