Cell Type Inference Using Large Language Models in Single-Cell Data Analysis

Duy Tran[1], Khoi Nguyen[1], Nam Sy Vo[2], Phi Bya[1], and Tin Nguyen[1(✉)]

[1] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA
tinn@auburn.edu

[2] Center for Biomedical Informatics, Vingroup Big Data Institute, Hanoi, Vietnam

Abstract. Single-cell RNA sequencing enables high-resolution analysis of cellular heterogeneity, but a key analysis step, cell type identification, remains a labor-intensive process that requires manual inspection of marker genes. Large Language Models (LLMs) offer a promising solution for automating this critical step. We present a case study using CytoAnalyst, a web-based platform that integrates LLMs for automated cell type annotation. Using a bone marrow organoid dataset, we compared multiple state-of-the-art LLMs in their ability to predict cell types from marker genes identified through differential expression analysis. Our structured prompting approach yielded accurate predictions for common cell types across all models, while performance varied for rare or specialized populations. This work demonstrates that LLMs can significantly reduce manual effort in scRNA-seq analysis, though further improvements are needed for more accurate and robust annotations. Our web-based platform and method are freely available at: https://cytoanalyst.tinnguyen-lab.com/.

Keywords: single-cell RNA sequencing · large language models · cell type prediction · clustering · differential expression analysis

1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables high-throughput gene expression analysis at single-cell resolution, providing unprecedented insights into cellular heterogeneity. Many tools and web-based systems facilitate scRNA-seq data analysis, including Seurat [1], Scanpy [2], ICARUS [3], SingleCAnalyzer [4], and CytoAnalyst [5]. A key analysis step is identifying distinct cell populations through clustering and marker gene identification. Accurate identification requires researchers to manually inspect identified markers and compare them with canonical cell type markers from literature and databases [6]. This process demands extensive biological expertise and is both time-consuming and error-prone.

T. X. Tu et al. (Eds.): ICTA 2025, LNNS 1833, pp. 239–247, 2026.
https://doi.org/10.1007/978-3-032-18162-6_25

Recently, Large Language Models (LLMs) [7] have emerged as promising tools for automating biological data analysis, including cell type annotation. LLMs can leverage their knowledge base to infer cell types from marker genes, potentially reducing manual inspection requirements. CytoAnalyst [5] is among the first platforms integrating LLMs for automated cell type annotation. In this paper, we present a case study using CytoAnalyst for scRNA-seq analysis and cell type annotation with LLMs. We compare CytoAnalyst's built-in LLM with external LLMs, including DeepSeek R1, Claude 3.7 Sonnet, GPT-4o, Llama 4 Scout, and Gemini 2.5 Pro. Our results demonstrate that LLMs are promising tools for cell type annotation, significantly reducing time and effort required for manual inspection.

2 Methods

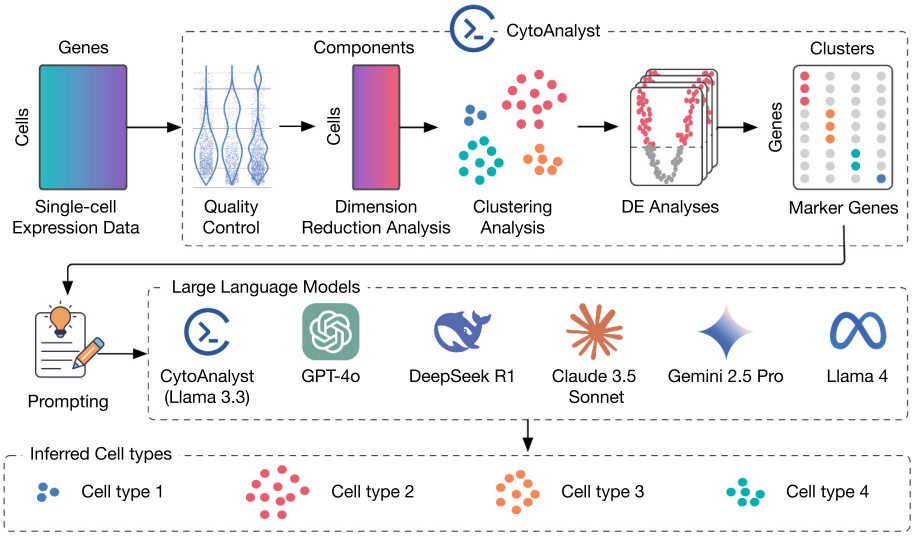

Fig. 1. Overall workflow for cell type inference using CytoAnalyst and large language models (LLMs). First, single-cell RNA sequencing data is imported into CytoAnalyst for quality control, normalization, and dimensionality reduction. Next, distinct cell populations and their marker genes are identified using CytoAnalyst's clustering and differential expression (DE) analysis modules. Finally, these marker genes are then used to annotate cell types with CytoAnalyst's built-in LLM and external LLMs.

Figure 1 illustrates the overall workflow of our cell type inference using large language models (LLMs). The workflow begins with importing single-cell RNA sequencing data into CytoAnalyst, a web-based platform for scRNA-seq data analysis. Using CytoAnalyst, we perform quality control, normalization, dimensionality reduction, clustering, and differential expression (DE) analysis to identify distinct cell populations and their marker genes. We then manually inspect

each analysis result using CytoAnalyst's visualization tools to ensure data quality and accuracy. Finally, we use the identified marker genes with designed prompts to predict cell types using both CytoAnalyst's built-in LLM and external LLMs.

2.1 Data Processing

CytoAnalyst supports data uploads in 10X Genomics Cell Ranger output [8] and AnnData objects [9]. The platform provides quality metrics, including unique gene counts per cell, UMI counts, and mitochondrial/ribosomal gene percentages to help users assess data quality. Using the data processing module, we filter out cells with fewer than 400 detected genes, over 40,000 total counts, and more than 10% mitochondrial gene counts.

2.2 Embedding and Clustering Analysis

Following quality filtering, we proceed with embedding and clustering analysis using CytoAnalyst's modules. The embedding analysis module provides comprehensive processing from normalization to visualization. We normalize each cell

Prompting

What cell type can this marker list be?

Input genes: KIF26B , GULP1 , COL5A2 , COL3A1 , ANTXR1 , COL6A3 , PDZRN3 , ..., SULF1

Target tissue context: bone marrow

Instruction:
- Identify the most likely cell types based ONLY on the provided markers
- Output a maximum 5 cell types with the highest confidence
- For each cell type, provide a complete lineage hierarchy (minimum 4 parent levels)
- You MUST refer to Cell Ontology, CellxGenes.
- All hierarchies must have "Cells" as the highest-level ancestor
- Only include established, well-documented cell types
- Exclude tissue-specific naming conventions
- Exclude cancer or disease-related cell types
- Only use markers that appear in the input list
- List matching markers for each cell type

IMPORTANT: You must strictly follow this exact output format with no deviations:
Cell type --> A --> B --> C --> D --> E --> Cells: [List all Cell type markers]
A is the parent of the cell type
B is the parent of A
C is the parent of B
D is the parent of C
E is the parent of D
Cells is the parent of E

DO NOT add any explanations, notes, or additional text before or after the cell-type listings.
DO NOT include any cell types that are not supported by the marker genes in the input list.

Fig. 2. Prompting structure for guiding large language models (LLMs) in cell type annotation. The prompt uses comma-separated marker genes from CytoAnalyst and specifies tissue context to focus on relevant cell types. It enforces strict output requirements and formatting rules to ensure accurate and structured annotations.

to 10,000 counts, apply log1p transformation, and select the top 4,000 highly variable genes (HVGs). Using these HVGs, we then conduct principal component analysis (PCA) [10] to extract the top 50 components for dimensionality reduction. For visualization, we utilize UMAP [11] to project data into two-dimensional space, though CytoAnalyst also offers t-SNE [12] and PCA options.

CytoAnalyst's clustering module identifies distinct cell populations using various algorithms, including Louvain [13], Leiden [14], and K-means [15]. In our analysis, we apply Louvain clustering at multiple resolutions to define cell clusters with varying granularity. We then manually inspect clustering results using CytoAnalyst's visualization module to ensure clusters correspond to meaningful populations. We also merge similar clusters when necessary, e.g., close and overlapping clusters.

2.3 Marker Genes Identification

After identifying distinct cell populations, we proceed to identify marker genes using CytoAnalyst's differential expression (DE) analysis module. The module supports multiple statistical tests, including Wilcoxon rank-sum test [16] and MAST [17]. Here, we identify marker genes by comparing each population's expression profile against remaining cells using the Wilcoxon rank-sum test with Benjamini-Hochberg adjustment [18] for false discovery rate control. To identify significant markers, we filter DE results based on adjusted p-values, log2 fold change, average expression in target populations, and differences in expression proportions. We manually adjust these filtering criteria and visualize gene expression to ensure biological relevance.

2.4 Cell Type Annotation

To identify cell types for each population, we leverage CytoAnalyst's built-in inference tool integrating Meta's Llama 3.3 [19]. The tool uses structured prompting to guide LLMs in generating accurate annotations, as illustrated in Fig. 2. It generates hierarchical cell type lists where the first level represents the most specific type. For each population, we provide marker genes from DE analysis and specify tissue context to predict possible cell types. We then select the cell type appearing most frequently across LLM results as the final annotation.

3 Results

In this section, we present the results of our analysis using CytoAnalyst on a single-cell RNA sequencing dataset. We also compare the performance of CytoAnalyst's built-in LLM with several external LLMs, including DeepSeek R1, Claude 3.7 Sonnet, GPT-4o, Llama 4 Scout, and Gemini 2.5 Pro, in annotating cell types based on identified marker genes.

We analyze a bone marrow organoid dataset from Frenz-Wiessner et al. [20] containing 31,040 cells from human induced pluripotent stem cells. Following the

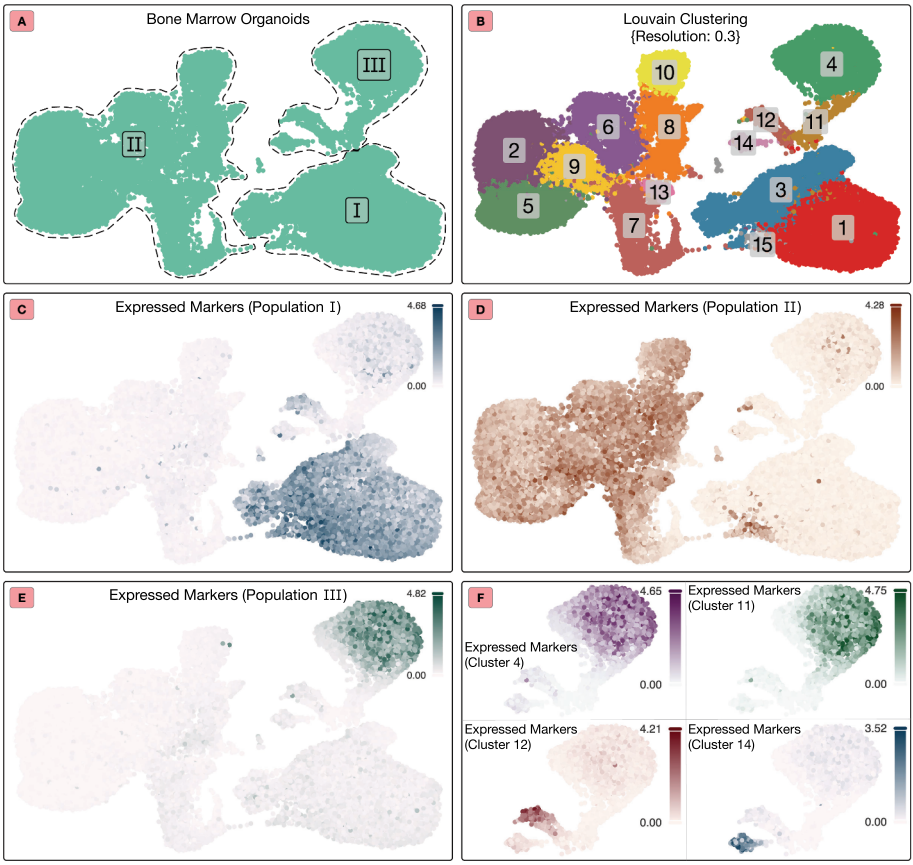

Fig. 3. Visualizations of single-cell transcriptomic data analysis using CytoAnalyst. A) UMAP projection showing three distinct cellular populations (I, II, III) with dashed boundaries. B) Louvain clustering at resolution 0.3 identifying 15 clusters. CE) Marker gene expression for populations I, II, and III. F) Marker gene expression for population III subpopulations (clusters 4, 11, 12, 14).

methods described in Sects. 2.1 and 2.2, we perform quality control, normalization, dimensionality reduction, and visualization using CytoAnalyst. The UMAP projection reveals three major cell populations (I, II, and III) as shown in Fig. 3A. We then apply Louvain clustering [13] at multiple resolutions (0.1–0.5) and manually inspect results using CytoAnalyst's visualization module. Resolution 0.3 provides optimal clustering, grouping cells within the three major populations without excessive fragmentation (Fig. 3B). This resolution identified 15 distinct clusters corresponding to: population I (clusters 1, 3, 15), population II (clusters 2, 5–10, 13), and population III (clusters 4, 11, 12, 14).

Next, we perform differential expression analyses to identify marker genes for each population using the criteria described in Sect. 2.3: log2 fold change $\geq$3,

CytoAnalyst Built-in LLM (Llama 3.3)	DeepSeek R1
Pericyte → **Mesenchymal Cell** → Connective Tissue Cell → Somatic Cell → Eukaryotic Cell → Cells: PDGFRB, COL1A1, COL1A2, ANTXR1, PDZRN3	Fibroblast --> Stromal cell --> **Mesenchymal cell** --> Connective tissue cell --> Animal cell --> Native cell --> Cells: [COL5A2, COL3A1, COL6A3, PDGFRB, COL1A2]
Smooth Muscle Cell → **Mesenchymal Cell** → Connective Tissue Cell → Somatic Cell → Eukaryotic Cell → Cells: CALD1, COL1A1, COL1A2, COL3A1, COL5A1	Pericyte --> Contractile cell --> **Mesenchymal cell** --> Connective tissue cell --> Animal cell --> Native cell --> Cells: [PDGFRB, ANTXR1, CALD1, VCAN, EDNRA]
Fibroblast → **Mesenchymal Cell** → Connective Tissue Cell → Somatic Cell → Eukaryotic Cell → Cells: COL1A1, COL1A2, COL3A1, COL5A1, COL5A2	Osteoblast --> Secretory cell --> Connective tissue cell --> Animal cell --> Native cell --> Cells: [COL1A1, COL1A2, SPARC, POSTN, CDH11]
Mesenchymal Stem Cell → **Mesenchymal Cell** → Connective Tissue Cell → Somatic Cell → Eukaryotic Cell → Cells: COL1A1, COL1A2, PDGFRB, VCAN	Mesenchymal Stromal Cell --> Stromal cell --> **Mesenchymal cell** --> Connective tissue cell --> Animal cell --> Native cell --> Cells: [PDGFRB, VCAN, COL1A1]
Osteoblast → **Mesenchymal Cell** →Connective Tissue Cell → Somatic Cell → Eukaryotic Cell → Cells: COL1A1, COL1A2, SPARC, POSTN, DCN	Smooth Muscle Progenitor Cell --> Contractile cell --> **Mesenchymal cell** --> Connective tissue cell --> Animal cell --> Native cell --> Cells: [CALD1, EDNRA, ITGA1]
Claude 3.7 Sonnet	**GPT-4o**
mesenchymal stromal cell --> stromal cell --> **mesenchymal cell** --> mesoderm-derived cell --> multi-potent progenitor cell --> Cells: [COL1A1, COL1A2]	Pericyte --> Mural cell --> **Mesenchymal cell** --> Mesodermal cell --> Embryonic cell --> Cells: [PDGFRB, RGS5, COL1A1, COL1A2, COL3A1, COL5A1, COL5A2]
fibroblast --> **connective tissue cell --> mesenchymal cell** --> mesoderm-derived cell --> stem cell --> Cells: [COL1A1, COL1A2, COL3A1, COL5A1, COL5A2]	Mesenchymal stem cell --> **Mesenchymal cell** --> Mesodermal cell --> Embryonic cell --> Cells: [COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, DCN, LUM]
pericyte --> mural cell --> vascular associated smooth muscle cell --> smooth muscle cell --> mesenchymal cell --> stem cell --> Cells: [PDGFRB, CALD1, EDNRA]	Fibroblast --> **Connective tissue cell --> Mesenchymal cell** --> Mesodermal cell --> Embryonic cell --> Cells: [COL1A1, COL1A2, COL3A1, COL5A1, COL6A2, DCN]
reticular cell --> stromal cell --> **mesenchymal cell** --> mesoderm-derived cell --> multi-potent progenitor cell --> stem cell --> Cells: [COL1A1, COL3A1, SPARC, LUM]	Smooth muscle cell --> Muscle cell --> Myocyte --> Mesodermal cell --> Embryonic cell --> Cells: [CALD1, COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, SPARC]
adventitial cell --> stromal cell --> **mesenchymal cell** --> mesoderm-derived cell --> multi-potent progenitor cell --> stem cell --> Cells: [COL1A1, COL1A2, COL3A1]	Stromal cell --> **Connective tissue cell --> Mesenchymal cell** --> Mesodermal cell --> Embryonic cell --> Cells: [COL1A1, COL6A1, COL6A2, DCN]
Llama 4 Scout	**Gemini 2.5 Pro**
Fibroblast --> **Connective Tissue Cell --> Mesenchymal Cell** --> Stromal Cell --> Cells: COL1A1, COL1A2, COL3A1, COL5A1, DCN, LUM, POSTN, SPARC	Fibroblast --> Stromal cell --> **Mesenchymal cell** --> Eukaryotic cell --> Somatic cell --> Cells: [COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL6A1]
Tenocyte --> **Connective Tissue Cell --> Mesenchymal Cell** --> Stromal Cell --> Cells: COL1A1, COL3A1, COL5A1, COL6A1, DCN, LUM, POSTN, SPARC	Mesenchymal stem cell --> Stromal cell --> **Mesenchymal cell** --> Eukaryotic cell --> Somatic cell --> Cells: [PDGFRB, VCAN, SPARC, COL1A1, DCN]
Endothelial Cell --> Lymphatic Endothelial Cell --> **Endothelial Cell --> Mesenchymal Cell** --> Stromal Cell --> Cells: ANTXR1, CDH11, COL1A1, COL1A2, COL3A1	Pericyte --> **Connective tissue cell --> Mesenchymal cell** --> Eukaryotic cell --> Somatic cell --> Cells: [PDGFRB, VCAN, SPARC, ANTXR1, CALD1]
Pericyte --> **Mesenchymal Cell** --> Stromal Cell --> Cells: COL1A1, COL3A1, COL5A1, COL5A2, COL6A1, DCN, LUM, PDGFRB, POSTN, SPARC	Endothelial cell --> Vascular cell --> **Mesenchymal cell** --> Eukaryotic cell --> Somatic cell --> Cells: [EDNRA, VCAN, SPARC, ANTXR1, ITGA1, SDC2, NID2, CDH11]
Adipocyte --> **Mesenchymal Cell** --> Stromal Cell --> Cells: COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL6A1, COL6A2, COL6A3, DCN, LUM, POSTN, SPARC	Smooth muscle cell --> Muscle cell --> **Mesenchymal cell** --> Eukaryotic cell --> Somatic cell --> Cells: [CALD1, PDGFRB, EDNRA, VCAN, SPARC]

Fig. 4. Cell type inference results for population I (clusters 1, 3, 15) using CytoAnalyst and external LLMs (DeepSeek R1, Claude 3.7 Sonnet, GPT-4o, Llama 4 Scout, Gemini 2.5 Pro). Each model generates five predicted cell types. Bold text indicates correctly predicted cell type and ontology hierarchy per ground truth. Red bold text indicates correct identification with incorrect hierarchical order.

adjusted p-value ≤ 0.05, average expression ≥ 1, and $\geq 50\%$ difference in expression proportion. Figure 3C–E illustrates the expression levels of identified marker genes for populations I, II, and III, respectively. It is clear that populations I

and II exhibit homogeneous marker expression, indicating a single cell type for each. However, population III shows heterogeneous expression, with cluster 4 displaying distinct patterns from clusters 11, 12, and 14, suggesting multiple cell types within this population. Thus, we performed additional DE analyses for each cluster in population III. The analyses reveal that clusters 4 and 11 share similar marker expression patterns, i.e., the same cell type, while clusters 12 and 14 represent unique cell types (Fig. 3F). This analysis yields five final populations: I (clusters 1, 3, 15), II (clusters 2, 5–10, 13), III (clusters 4, 11), IV (cluster 12), and V (cluster 14).

Finally, using marker genes and bone marrow tissue context, we predict cell types with CytoAnalyst's built-in LLM and external LLMs following the prompting strategy in Sect. 2.4. Figure 4 shows an example of the predictions for population I, where all models correctly identify mesenchymal stem cells (MSCs). However, most external models struggle with maintaining correct hierarchical order, as indicated by red bold text.

Table 1 summarizes cell type annotations for all populations. All models accurately identify populations I, II, and III as mesenchymal stem cells, hematopoietic cells, and endothelial cells, respectively. CytoAnalyst misidentifies population IV, while all external LLMs incorrectly annotate population V. This highlights the challenges of annotating rare cell types.

Table 1. Cell type annotations for all populations using CytoAnalyst and external LLMs. Light gray shading indicates incorrect annotations versus ground truth. Abbreviations: EC, endothelial cell; EpC, epithelial cell; HC, hematopoietic cell; MDC, mesodermal cell; MSC, mesenchymal cell; SSC, somatic stem cell.

Population	**I**	**II**	**III**	**IV**	**V**
Ground Truth	MSC	HC	EC	EpC	MDC
CytoAnalyst	MSC	HC	EC	MSC	MDC
DeepSeek R1	MSC	HC	EC	EpC	HC
GPT-4o	MSC	HC	EC	EpC	HC
Gemini 2.5 Pro	MSC	HC	EC	EpC	SSC
Claude 3.7 Sonnet	MSC	HC	EC	EpC	SSC
Llama 4 Scout	MSC	HC	EC	EpC	N/A

4 Conclusion

We presented a case study using CytoAnalyst for single-cell RNA sequencing analysis and cell type annotation with large language models. We demonstrated the effectiveness of CytoAnalyst's built-in LLM and external LLMs (DeepSeek R1, Claude 3.7 Sonnet, GPT-4o, Llama 4 Scout, and Gemini 2.5 Pro) in predicting cell types from marker genes. Most LLMs accurately identified common cell

types (mesenchymal, hematopoietic, and endothelial cells) but struggled with rare or specific types like mesodermal cells. This highlights LLMs' potential for automating cell type annotation while indicating the need for improved performance, particularly for uncommon cell types.

Acknowledgment. This work was partially supported by National Science Foundation (2343019 and 2203236), National Cancer Institute (U01CA274573), National Institute of General Medical Sciences (R44GM152152), and National Institute of Food and Agriculture (2023-67022-40041). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Satija, R., Farrell, J.A., et al.: Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. **33**, 495–502 (2015)
2. Wolf, F.A., Angerer, P., et al.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**, 1–5 (2018)
3. Jiang, A., Lehnert, K., et al.: ICARUS, an interactive web server for single cell RNA-seq analysis. Nucleic Acids Res. **50**(W1), W427–W433 (2022)
4. Prieto, C., Barrios, D., et al.: SingleCAnalyzer: interactive analysis of single cell RNA-Seq data on the cloud. Front. Bioinform. **2**, 793309 (2022)
5. Bya, P., Tran, D., et al.: CytoAnalyst: a web-based platform for comprehensive single-cell RNA sequencing analysis. bioRxiv 2025–04 (2025)
6. Hu, C., Li, T., et al.: Cell Marker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. **51**(D1), D870–D876 (2023)
7. Naveed, H., Khan, A.U., et al.: A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435 (2023)
8. Zheng, G.X.Y., Terry, J.M., et al.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**(1), 14049 (2017)
9. Virshup, I., Rybakov, S., et al.: anndata: Access and store annotated data matrices. J. Open Source Softw. **9**(101), 4371 (2024)
10. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisc. Rev.: Comput. Stat. **2**(4), 433–459 (2010)
11. McInnes, L., Healy, J., et al.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
13. Blondel, V.D., Guillaume, J.L., et al.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
14. Traag, V.A., Waltman, L., et al.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. **9**(1), 1–12 (2019)
15. Kodinariya, T., Makwana, P.: Review on determining number of Cluster in K-Means Clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. **1**, 90–95 (2013)
16. Wilcoxon, F., Katti, S., et al.: Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. Sel. Tables Math. Stat. **1**, 171–259 (1970)

17. Finak, G., McDavid, A., et al.: MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. **16**, 278 (2015)
18. Thissen, D., Steinberg, L., et al.: Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. J. Educ. Behav. Stat. **27**(1), 77–83 (2002)
19. Grattafiori, A., Dubey, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
20. Frenz-Wiessner, S., Fairley, S.D., et al.: Generation of complex bone marrow organoids from human induced pluripotent stem cells. Nat. Methods **21**, 868–881 (2024)