ScRNA-seq Analysis Copilot: Facilitating scRNA-seq Analysis Learning Through Interactive Chat-bot

Chien-Yueh Liu
Pui Ching Middle School, Macau
chienyuehliu@gmail.com

October 4, 2024

Abstract

Single-cell RNA sequencing (scRNA-seq) is a powerful technology that enhances the field of medicine, including analysis of tissue, cancer development and therapeutic evaluation. However, users must have programming skills and go through a long learning process. We propose an AI agent aimed at empowering the Large Language Model (LLM) with domain knowledge and scRNA-seq analytical ability through an Application Programming Interface (API). Through the Human-Computer Interaction (HCI), new learners could practice and interact with the gene expression matrix using natural language. Clinical healthcare crews and medical laboratory practitioners will be equipped with a much more efficient and user-friendly interface, facilitating the development of drugs, treatment, and groundbreaking research.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a novel approach to investigating tissue micro-environments and gene expressions. The RNA of single cells from tissue can be screened, mapped, and counted, followed by a series of analyses of the expression matrix [1]. Despite its wide application and extensive use, scRNA-seq analysis is neither accessible nor easy to perform. Users are required to possess an intermediate level of programming skills in Python or R [2] and an understanding of the data structure, which can be bewildering and challenging for those with limited background knowledge of bioinformatics and data science. Additionally, the analysis consists of multiple steps of data processing and often requires using various methods to validate the results in parallel. Since most analyses rely on manual annotation for cell cluster identification (comparing the expression level of marker genes), the analysis could be unrepeatable and time-consuming [3]. Human-Computer Interaction (HCI), a subfield of computer

science concerned with the interaction between the computer and the users, demonstrates the design, evaluation, and implementation of the user interface of computer systems that respond to users' needs and habits [4]. In this project, we utilize HCI, integrating R functions to analyze single-cell RNA data with an AI agent powered by LLM. Specifically, users can interact with datasets through natural language and practice analyzing scRNA-seq data without programming skills.

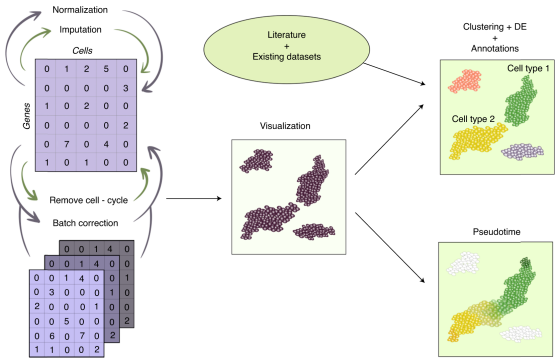

Figure 1: The entire sequence analyzing process includes the following stages: 1. Quality control (QC), 2. Normalization, 3. Visualization (PCA, Principal Component Analysis, for dimensional reduction), 4. Clustering, and 5. Annotation. These procedures are usually performed using R packages like Seurat or the Python library Scanpy, thus requiring programming skills and experience to annotate cells. Source: Andrews, T.S., Kiselev, V.Y., McCarthy, D. et al.

2 The scRNA-seq Analysis Copilot

The main target of this agent is to assist new learners of scRNA-seq analysis and facilitate the operation of such analysis by practitioners. To achieve this, the agent has two main goals: processing the given scRNA-seq files and answering related questions. This agent is integrated on the Dify platform [1] and powered by Azure OpenAI Service gpt-4o model, the latest OpenAI multimodal model [6]. The general structure of the copilot is composed of a backbone LLM agent with engineered prompt, a knowledge database and APIs to execute R functions.

2.1 Knowledge Database

Answering the user with precise and accurate information is crucial. Nine academic papers [1][2][3][8][9][10][11][13][14] addressing scRNA-seq (total: 658,018 words) are processed and separated into chunks with the Dify Economical Index

[1]Dify is an open-source large language model (LLM) application development platform which combines Backend-as-a-Service and LLM Operation and Practices.

mode and retrieved with Vector Search, generating query embeddings to match its vector representation [15].

2.2 APIs

Another essential component of the agent for it to process the data is the API. The agent can send HTTP requests to these APIs on the server, and without computing them on the client end, the user can alter the uploaded scRNA dataset. R package Plumber [16] is applied to transform R code into APIs. On the Dify interface, the agent sends requests to API and receives responses (in JSON format), thus interacting with the given Seurat Object [2] [17].

2.3 Prompt

Azure OpenAI Service gpt-4o model is trained to follow the workflow, assist the user with a structured prompt, and send API requests to input the variables for analysis. Detailed prompts can be assessed in the Data and Code Availability section. The basic framework of the prompts is composed of five parts 1.
Choosing file: First, the user is asked to provide a direct link to the .RDS file of the data or use the default dataset (pbmc3k, pancref, kidneyref, or lungref). These datasets are derived from the Azimuth reference dataset [18] and 10x Genomics. The agent downloads the file from the URL to the server, opens the Seurat object, and plots an overview QC metrics visualization.
Quality control: Then, the agent asks the user to conduct QC by providing the minimum and maximum numbers of genes (features) and the maximum percentage of mitochondrial genes.
Normalization and PCA: Once the user enters the numbers, the agent selects the cells in the object by the quality control parameters. Proceeding to the normalization and PCA, the user should provide the normalization factor, normalization method, selection method and the number of highly variable genes. The agent applies these parameters to normalize the data and run PCA before showing the elbow plot of standard deviation versus the number of principal components.
Clustering: The agent asks the user to provide the number of dimensions with the elbow point in the previous graph and the desired resolution. Next, by using FindNeighbors, FindClusters, and runUMAP R functions, unsupervised clustering is accomplished. [10] [19]
Annotation: Finally, scType cell annotation [10] is adopted to classify the cells. The agent calculates the scType score of each cell cluster by comparing the expression of positive and negative marker gene lists. The highest score for each cluster determines the cell type, which is be labeled and colored on the clustered graph. This final graph, a preliminary analysis, is completed. Downstream analysis is not yet available, but the agent can still answer questions related to it.

[2] A representation of single-cell expression data in R.

3 Results

3.1 Demonstration

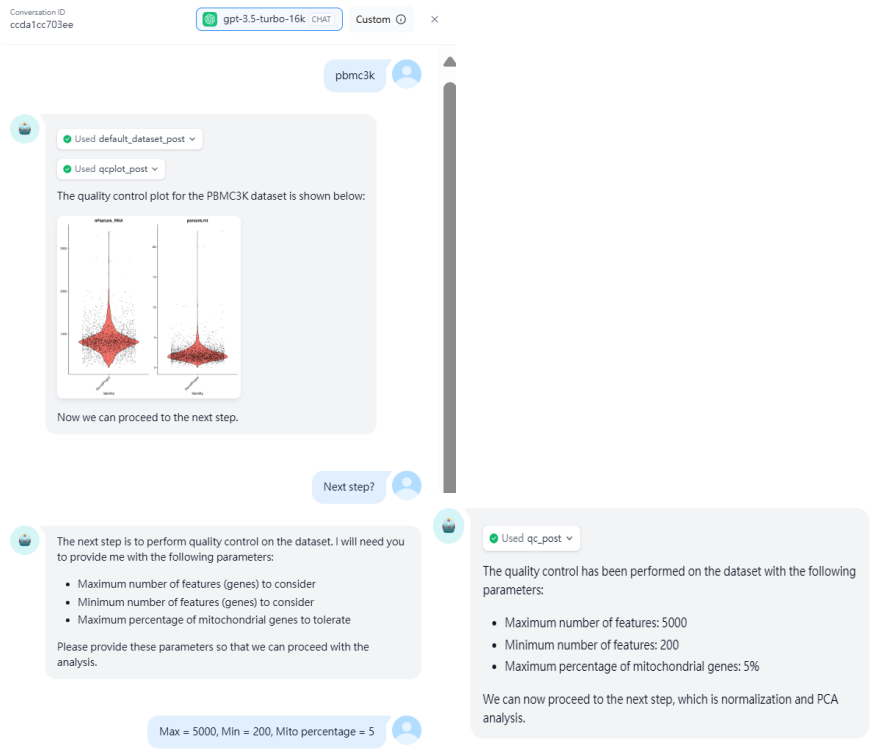

Figure 2: The conversation between the user and scRNA-seq Analysis Copilot about running quality control. Captured on Dify "Logs & Ann." section of the agent. The agent can send requests corresponding to the user's message, show the graph generated, and give instructions to the user for inputting parameters.

The copilot first asks the user to provide the necessary parameters, then generates results and explains them to the user. As is demonstrated in Fig. 2, the agent is capable of understanding the workflow, reading instructions, giving instructive responses and showing respective QC metrics to the user.

The analysis runs on a Ubuntu server (Ubuntu 22.04.1 LTS (GNU/Linux 5.15.0-91-generic x86_64)). A total of 30,640 tokens (average: 3404 tokens per message) and 126.74 seconds (average: 14.8s per message) is required for one annotation analysis. One question accounts for approximately 1,500 tokens.

3.2 Methodology of Evaluation

We hypothesized that by equipping learners with the copilot, they would have a deeper understanding of scRNA-seq. To evaluate the effectiveness of the agent, we designed a questionnaire consisting of 20 Multiple Choice Questions (MCQ). The scores were calculated and compared (one point for each MCQ, totaling 20 points). 31 volunteers with no previous understanding of scRNA-seq were randomly assigned to three groups: Experiment, Control and Baseline. Their consent and parent consent were confirmed to allow data collection during the research. They were also informed that this experiment did not pose any risks to them. Both the Control and Experiment group (11 participants each) can use the Internet and other AI products, but only the Experiment group has access to the copilot. An additional group with nine participants, Baseline, was asked to complete the final test in 10 minutes without using a search engine or any type of AI. After a brief introduction of single cell transcriptome, research motivation, and experiment instruction, participants from Control and Experiment groups were instructed to learn, explore, and search "scRNA-seq, its application, and analysis procedure" in 30 minutes. After 30 minutes, they finished the questionnaire in 10 minutes. The results of the questionnaire were anonymous, and the scores were calculated automatically.

3.3 Evaluation Results

Before analyzing the significance of the data, we preprocessed the data to ensure that the data is distributed normally and not skewed by extreme data points. We excluded a biased response whose respondent clicked all options in the first two questions. Similarly, we removed an extreme case which scored 15 out of 20, as the median number is 6 in that group). As can be seen from Fig. 3, the average performance of the Experiment group is the best, with a mean value of 6.36. Specifically, the mean value of the Experiment group is 2.36 points higher than that of the Baseline group. While the mean value of the Control group is 1.5 points higher than that of the Baseline group, it is 0.86 points lower than the average score of the Experiment group. To evaluate whether our results are significant, we first used Mann-Whitney test [20] because the number of data points is less than 30. The U-value of Experiment and Control group is 43.5. The critical value of U at $p < 0.05$ is 31. Therefore, the result is not significant at $p < 0.05$. As for Experiment and Baseline, their U-value is 23.5. The critical value of U at $p < 0.05$ is 27, proving that the result is significant at $p < 0.05$. Lastly, when we compared Control and Baseline, the U-value is 26.5. The critical value of U at $p < .05$ is 24, so the result is not significant at $p < 0.05$. In short, the results of the survey illustrated that the performance of Experiment and Control groups have no significant difference. On the contrary, the scores of group Experiment and Baseline show a statistical significance when that of Control and Baseline do not. Therefore, we can conclude that the agent can facilitate the users' understanding of scRNA-seq analysis when compared to those who have no knowledge of it, but not people who use the Internet or

artificial general intelligence. Although the copilot may not outcompete other AI or search tools, it has several advantages, including direct analysis practices and personalized response with domain knowledge. This will benefit the users in the longer term and enhance their ability to analyze scRNA-seq data.

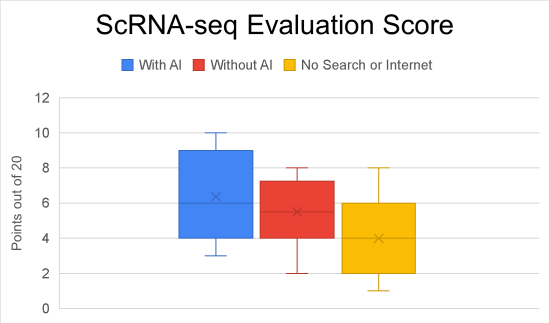

Figure 3: The box plot of scores of groups: Experiment (with access to AI), Control (without AI) and Baseline (without search or internet). The number presented on the box is the mean value of each group.

4 Conclusion

By incorporating a LLM into scRNA-seq analysis, new learners and scientists without relevant experiences can explore scRNA-seq and familiarize themselves with this field efficiently. The evaluation results demonstrate that using scRNA-seq Analysis Copilot, users' understanding of scRNA-seq is significantly enhanced. This application will facilitate the exploration of scRNA-seq data, enhance the productivity of clinical investigations, and provide medical practitioners with insights into cancer or disease development and gene expression at the cell level.

5 Data and Code Availability

The AI agent is developed on the Dify platform. R code (API descriptions), detailed prompt, and evaluation data are available here: `https://github.com/Gina727/scRNA-seq-analyst` and the basic code is derived from Harvard Single-cell RNA-seq data analysis workshop (`https://github.com/hbctraining/scRNA-seq_online`) and scType (`https://github.com/IanevskiAleksandr/sc-type`).

I would like to thank Mr. Ng Ka Long, Mr. Chan Iat Chong and DocAI Technology Pte. Ltd. for providing model training resources, answering my questions, and teaching me how to build the APIs.

References

[1] Andrews T, Kiselev V, McCarthy D, et al. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc 16, 1–9 (2021). https://doi.org/10.1038/s41596-020-00409-w

[2] Hao Yu, Yuqing Wang, Xi Zhang, Zheng Wang, GRACE: a comprehensive web-based platform for integrative single-cell transcriptome analysis, NAR Genomics and Bioinformatics, Volume 5, Issue 2, June 2023, lqad050, https://doi.org/10.1093/nargab/lqad050

[3] Abdelaal, T., Michielsen, L., Cats, D. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194 (2019). https://doi.org/10.1186/s13059-019-1795-z

[4] Ritvo, S. and Allison, R. Designing for the exceptional user: Nonhuman animal-computer interaction (ACI), Computers in Human Behavior, Volume 70, 2017, Pages 222-233, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2016.12.062.

[5] Brey P, Søraker JH. Philosophy of Computing and Information Technology. In: Meijers A (ed.), Philosophy of Technology and Engineering Sciences. Amsterdam: North-Holland; 2009. Pages 1341-1407. DOI: 10.1016/B978-0-444-51667-1.50051-3

[6] Boyd E. Introducing GPT-4O: OpenAI's new flagship multimodal model now in preview on Azure [Internet]. 2024. Available from: https://azure.microsoft.com/en-us/blog/introducing-gpt-4o-openais-new-flagship-multimodal-model-now-in-preview-on-azure/?msockid=0fac5c86b977624c38124f53b8ae637f

[7] Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics, Volume 40, Issue 2, February 2024, btae075.

[8] Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med. 2022 Mar;12(3):e694. doi: 10.1002/ctm2.694. PMID: 35352511; PMCID: PMC8964935.

[9] Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Apr 13;2015(11):951-69. doi: 10.1101/pdb.top084970. PMID: 25870306; PMCID: PMC4863231.

[10] Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022 Mar 10;13(1):1246. doi: 10.1038/s41467-022-28803-w. PMID: 35273156; PMCID: PMC8913782.

[11] Yu H, Wang Y, Zhang X, Wang Z. GRACE: a comprehensive web-based platform for integrative single-cell transcriptome analysis. NAR Genom Bioinform. 2023 Jun 9;5(2):lqad050. doi: 10.1093/nargab/lqad050. PMID: 37305171; PMCID: PMC10251641.

[12] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746. doi: 10.15252/msb.20188746. PMID: 31217225; PMCID: PMC6582955.

[13] Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, Winter DR. A Beginner's Guide to Analysis of RNA Sequencing Data. Am J Respir Cell Mol Biol. 2018 Aug;59(2):145-157. doi: 10.1165/rcmb.2017-0430TR. PMID: 29624415; PMCID: PMC6096346.

[14] Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics. 2021 May 17;37(7):963-967. doi: 10.1093/bioinformatics/btaa751. PMID: 32840568; PMCID: PMC8599307.

[15] Schwaber-Cohen R. Vector Similarity Explained. Pinecone. 2023.Retrieved from: https://www.pinecone.io/learn/vector-similarity/

[16] Schloerke B, Allen J. plumber: An API Generator for R. R package version 1.2.1.9000. GitHub repository. (2024).

[17] Satija, R., Farrell, J., Gennert, D. et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33, 495–502 (2015). https://doi.org/10.1038/nbt.3192

[18] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.

[19] Piper M, Mistry M, Liu J, Gammerdinger W, Khetani R. scRNA-seq Lessons from HCBC (first release). Zenodo. 2022. DOI: 10.5281/zenodo.5826256

[20] Mann-Whitney U Test Calculator. Stangroom J. 2024. From: Mann-Whitney U Test Calculator (socscistatistics.com)