SCassist: An AI Based Workflow **Assist**ant for **S**ingle-**C**ell Analysis

Vijayaraj Nagarajan[1*], Guangpu Shi[1], Samyuktha Arunkumar[1], Chunhong Liu[3], Jaanam Gopalakrishnan[2], Pulak R Nath[4], Junseok Jang[1], Rachel R Caspi[1]

[1] Laboratory of Immunology, National Eye Institute, NIH, Bethesda 20892, USA
[2] Neuro-Immune Regulome Unit, National Eye Institute, NIH, Bethesda 20892, USA
[3] Neuro-Immune Regulome Unit (Alumni), National Eye Institute, NIH, Bethesda 20892, USA
[4] Laboratory of Immunology (Alumni), National Eye Institute, NIH, Bethesda 20892, USA

* Correspondence to: Vijayaraj Nagarajan, National Eye Institute, NIH, Bethesda, 20892.
Email: vijayaraj.nagarajan@nih.gov

Abstract

Summary

Single-cell RNA sequencing (scRNA-seq) data analysis often involves complex iterative workflow, requiring significant expertise and time. To navigate this complexity, we have developed SCassist, an R package that leverages the power of the large language models (LLM's) to guide and enhance scRNA-seq analysis. SCassist integrates LLM's into key workflow steps, to analyze user data and provide relevant recommendations for filtering, normalization and clustering parameters. It also provides LLM guided insightful interpretations of variable features and principal components, along with cell type annotations and enrichment analysis. SCassist provides intelligent assistance using popular LLM's like Google's Gemini and Meta's Llama3, making scRNA-seq analysis accessible to researchers at all levels.

Availability and implementation

The SCassist package, along with the detailed tutorials, is available at GitHub.
https://github.com/NIH-NEI/SCassist

Contact

* Correspondence to: Vijayaraj Nagarajan, National Eye Institute, NIH, Bethesda, 20892.
Email: vijayaraj.nagarajan@nih.gov

Supplementary information

SCassist_Supplementary_Data.pdf

Introduction

The standard Single-cell analysis workflow includes several steps that require researchers to identify and use appropriate parameters. These steps include quality filtering, normalization, dimension reduction and clustering (Nath et al. 2024) . Other aspects, such as understanding the context of variable genes, principal components, cell type markers and enrichment analysis, require extensive knowledge of the background biology to extract meaningful insights from the data.

Large Language Models (LLM's) have significantly advanced text generation, text summarization, translation, question answering, image generation, audio/video generation, code generation etc., (Tom B. Brown 2020). In biomedical research and clinical settings, LLM based applications are being explored for their potential in drug design, drug discovery, sequence analysis, target discovery, clinical diagnosis, treatment recommendations, and outcome predictions (Ji et al. 2021; Lee et al. 2020; Thirunavukarasu et al. 2023).

In the field of single-cell analysis, various tools and methods have been developed using specialized, foundation or fine-tuned models to address challenges such as cell type annotation (Fu et al. 2024), integration, clustering, dimensionality reduction, and trajectory analysis (Qiu et al. 2024; Shen et al. 2023; Szalata et al. 2024). Researchers have also integrated LLM-generated results into radiology reporting workflows (Jorg et al. 2024). Although 'hallucination' is recognized as an inherent limitation of LLM's (Kankanhalli 2024), new methods and tools to mitigate this and take advantage of the LLM's abilities, like retrieval augmentation, are also being established (Jin et al. 2024) .

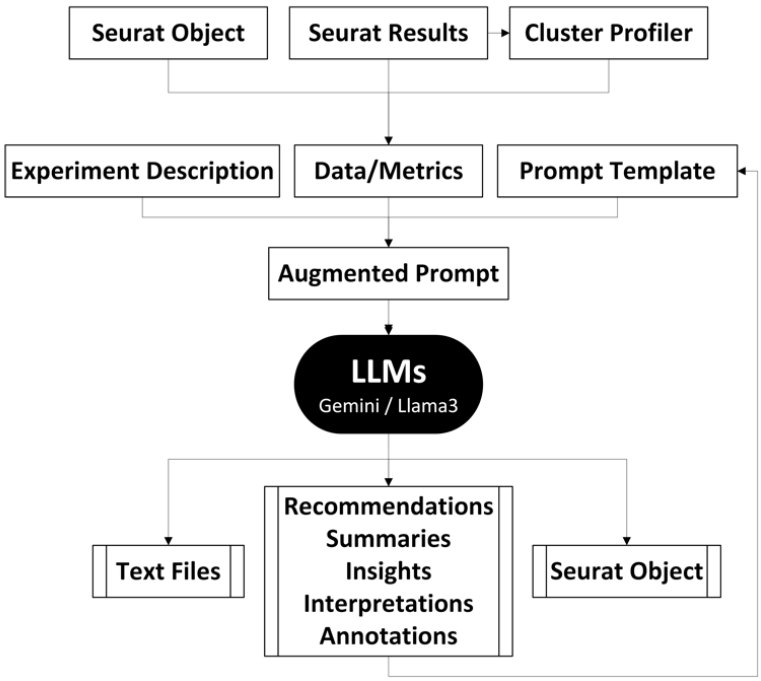
In this work, we have developed a novel workflow assistant, SCassist, as an opensource R package. It helps the researchers easily navigate through the complex and time-consuming components of the single-cell transcriptomics data analysis. SCassist leverages the retrieval and tool-based augmentation techniques to seamlessly integrate the power of the LLM's into the standard single-cell analysis workflow. This enables data exploration and understanding through insightful descriptions and enhanced interpretability of complex concepts, accelerating discovery.

Methods

Architecture

SCassist is written in the R programming language. SCassist takes the Seurat single-cell object and related Seurat results to generate relevant data/metrics. Seurat results are also used to perform enrichment analysis with ClusterProfiler, generating additional tools-based data/metrics. An augmented prompt is then constructed, by appropriately combining the data/metrics and experiment description using the prompt template. The contextualized augmented prompt is then submitted to the LLM. The LLM's response is parsed, and the results are displayed on the screen and/or saved them to a file, or appended to the Seurat object, as metadata. In some functions, the parsed response from the LLM is also used as the input for SCassist's second round of LLM queries. The general architecture of SCassist is illustrated in Figure 1. The SCassist package was built using R 4.4.1 running on the macOS Ventura platform.

Figure 1: The general architecture of the SCassist algorithm



Large Language Models

SCassist allows users to choose between "Google" or "Ollama" as the LLM server. The default models are set as "gemini-1.5-flash-latest" for Google and "llama3" for Ollama. If using Google as the server, the user is expected to provide an API key from Google. For Ollama, the user is expected to have it installed and running locally. We chose Google as a commercial online LLM server, due to their large context window availability. Ollama was chosen as an open-source, offline LLM server, providing free access to open-source models like Llama3 locally, without sending the data to a remote server. Detailed installation and setup instructions are available on our project's GitHub page.

Implementation

The SCassist package implements several key functions as described below:

- **SCassist_analyze_quality():** Analyzes the quality of the single-cell data using metrics like nCount_RNA, nFeature_RNA, and any user-specified quality metrics generated through Seurat's PercentageFeatureSet, to recommend filtering cutoffs.
- **SCassist_recommend_normalization():** Analyzes the characteristics of the dataset (number of cells, gene expression distribution, library size variation) to recommend the most suitable normalization method from Seurat's options.
- **SCassist_analyze_variable_features():** Analyzes the top variable features identified by Seurat to identify enriched gene ontologies or pathways, based on known feature characteristics and explain their relevance to the experimental design.
- **SCassist_recommend_pcs():** Analyzes the variance explained by each principal component (PC) to recommend the optimal number of PCs for downstream analysis.
- **SCassist_analyze_pcs():** Analyzes the top PCs to understand the driving biological processes based on top contributing genes and their associated pathways.
- **SCassist_recommend_k():** Analyzes the number of cells, PCs used, and clustering goals to recommend a range of potential k.param values for FindNeighbors

- **SCassist_recommend_res():** Analyzes the number of cells, mean expression variability, and median neighbor distance to recommend a suitable resolution range for FindClusters.
- **SCassist_analyze_and_annotate():** Analyzes top markers for each cluster to predict potential cell types based on the markers and provide reasoning. Optionally annotates the Seurat object with the predicted cell types.
- **SCassist_analyze_enrichment():** Runs ClusterProfiler for KEGG pathway and Gene Ontology enrichment. Analyzes and processes the ClusterProfiler results to integrate and generate insights on significant pathways, potential regulators, key genes or targets, and a summary network.
- **SCassist_summary_network():** Uses the LLM to extract network data from the output of SCassist_analyze_enrichment and creates an interactive network visualization that summarizes the system.

Detailed documentation for each of the functions and methods, along with usage and examples, are provided with corresponding R help files, as part of the SCassist package. Each of the above steps can be run in sequence, along with the appropriate standard Seurat single-cell analysis workflow steps. Each of the above steps can also be run independently at any appropriate step of the standard workflow. SCassist also provides options for advanced users to choose specific models and set custom model parameters. We have engineered appropriate template prompts for each of the SCassist functions, to obtain an optimal response from the LLM. Detailed template prompts for all the functions are provided in the project's GitHub page.

The SCassist R package can be installed and run on any operating system, using R versions 4.4.1 and above. The R packages Seurat, rollama, httr, jsonlite, visNetwork, clusterProfiler and BiocManager are required for the full functioning of SCassist.

Evaluation

Dataset

To evaluate SCassist's performance, we used datasets from two of our recently published single-cell transcriptome studies (Supplementary Table 1). One study identified an NK cell subset associated with disease activity in human Uveitis patients (Nath et al. 2024) and the other identified a CTCF binding motif site critical for mouse Th1 cell fate specification (Liu et al. 2024). We reanalyzed these datasets using SCassist and evaluated the SCassist-generated workflow reports against the previously published standard workflow reports for these two datasets.

Groundedness score

SCassist's ability to use the provided content was measured using the groundedness score. This score serves as an indirect measure of hallucination by estimating how well the LLM's response is grounded in the provided information. As the ground truth, we used all the data-derived metrics, lists of variable genes, lists of genes from PCs, lists of genes from cluster markers, lists of genes from ClusterProfiler enrichment results, and lists of genes derived from SCassist's enrichment analysis summary reports. We used GeneVenn (Pirooznia et al. 2007) to compare the terms in the SCassist's responses with the corresponding terms in our ground truth, to score for the overlaps. The groundedness score (G) was calculated as the cardinality of the intersection between the ground truth (GT) and the LLM's response (LLM), divided by the cardinality of the LLM's response: $G = |GT \cap LLM / |LLM|$.

Contextual relevance

The contextual relevance of the LLM's response in SCassist was measured using semantic similarity. We compared the SCassist' enrichment analysis summary response to that of the input document containing enriched pathways and GO terms, using the transformer-based language model BERT (Bidirectional encoder representations from transformers) (Devlin et al. 2018). Briefly, the contents were tokenized, added with relevant special tokens to mark the start and end of the contents and chunked to fit BERT's input length. We

then converted the chunks into input ids and extracted relevant embeddings. The chunked embeddings were combined and used to compute the cosine similarity. The semantic similarity analysis was performed in a python virtual environment using the transformers, torch and sklearn.metrics.pairwise packages. The fully documented python script used for this analysis is provided in our project's GitHub page.

Human Evaluation

We assessed the quality of the SCassist generated workflow reports against the standard single-cell analysis workflow reports using the Likert scale (1 to 5), to account for Accuracy, Relevance, Clarity, Trustworthiness and Overall satisfaction (Likert 1932). We designed our scoring sheet with clear anchors, describing how to differentiate the scores, including specific descriptions to reduce any ambiguity. We chose four senior level scientists (two Senior Scientist's from BioTech industry and two Staff Scientist's from the NIH) and four junior level scientists (Post Doctoral Fellows and Post Baccalaureate Fellows from the NIH) to evaluate our reports. Two of the senior scientists were the original authors of the published workflows that are analyzed with SCassist in this study. The template scoring sheet is provided in Supplementary Table 2.

Statistics

For the expert human evaluation, we performed the Wilcoxon Signed-Rank test to assess whether the overall Likert scale scores were significantly skewed towards higher values, indicating positive user perception of the LLM's performance. We also performed the Wilcoxon Rank-Sum test to evaluate if the Likert scale scores are affected by the level of expertise and by the dataset evaluated. Summary statistics, along with frequency distribution plots were used to understand the expert human evaluation scores.

Results

SCassist is available as an opensource R package at our project GitHub page: https://github.com/NIH-NEI/SCassist . Detailed tutorials, walking through the step-by-step workflows, using example data, are also provided at the GitHub project site. SCassist package could be installed using the R command:
`devtools::install_github("NIH-NEI/SCassist").`
Detailed tutorials describing the usage of commands and corresponding responses are shared through the project GitHub page.

Evaluation

Groundedness score

Using a total of 4136 tokens as ground truth, for different categories, SCassist scored an average of 98.7% groundedness for the LCMV. For the BCRUV data, using a total of 4110 tokens, SCassist scored an impressive average of 99.9% groundedness. Detailed scores for different categories are presented in the Supplemental Table 3.

Semantic similarity

Using BERT's uncased model, SCassist scored 76% on the semantic similarity between the terms used and the reports generated, for the LCMV data. SCassist scored a similar 74% for the BCRUV data. Detailed scores for different categories are presented in the Supplemental Table 4.

Expert human evaluation

Evaluation of SCassist's workflow reports, against the standard workflow reports, by eight human expert evaluators with different levels of expertise show that SCassist has good accuracy and generally performs well on relevance and trustworthiness. SCassist performs exceptionally well on clarity, with most evaluators

(7 out of 8) finding it useful (Supplementary Figure 1). Detailed scores are provided in Supplementary Table 5. The Wilcoxon Signed-Rank test with a p-value of 0.0001122 revealed a statistically significant skew towards higher scores, indicating that SCassist performs well overall. The Wilcoxon Rank-Sum test revealed no statistically significant differences in the LLM's performance across expert levels and across the two different datasets (Supplementary Table 6).

Cost

We made a total of 1,978 API calls to Google Gemini, in the two-month testing period during this project. The total cost for the 1,978 API calls was $2.07 excluding taxes. Running SCassist using the local Ollama server option did not incur any direct costs.

Limitations

Due to the fast-changing landscape of the LLM's, SCassist's responses in the future cannot be predicted, due to possible changes in the underlying models. That said, we believe the responses would only get better. We commit to monitoring the test workflow for any substantial changes in the responses at a minimum once every three months and to provide appropriate and diligent updates to the SCassist package. We will strive to fix any issues reported by the users, at the earliest possible time.

Conclusion

SCassist consistently achieved high scores across groundedness, semantic similarity and expert human evaluation, indicating its ability to generate accurate and meaningful insights. Our analysis also revealed that SCassist's performance is robust across different datasets and the evaluation scores are not affected by differences in the expertise level of the evaluator. The user-friendly interface and integration with popular LLM's make SCassist accessible to researchers at all levels. It empowers them to leverage AI-driven insights to accelerate their single-cell analysis workflow, while at the same time uncovering deeper biological insights and advancing scientific discovery. Our future development efforts will extend SCassist's capabilities to support multi-modal and spatial workflows, including interaction and trajectory analyses. We also plan to integrate the LLMs' function-calling abilities to offer automation.

Data availability

All data and code discussed in this work are available in our GitHub page at :
https://github.com/NIH-NEI/SCassist/

Acknowledgments

This work utilized the computational resources of the NIH HPC Biowulf cluster. The authors sincerely thank Drs. Charles Egwuagu and Han-Yu Shih for their support and guidance throughout the related historical work. The authors also thank Zixuan Peng for help with evaluation and review of the manuscript. We also thank the NEI IRP program for funding this research.

Funding

This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Eye Institute [EY000184, R01 EY032482].

References:

Devlin, Jacob, et al. (2018), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arxiv.org*.

Fu, Q., et al. (2024), 'A comparison of scRNA-seq annotation methods based on experimentally labeled immune cell subtype dataset', *Brief Bioinform,* 25 (5).

Ji, Y., et al. (2021), 'DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome', *Bioinformatics,* 37 (15), 2112-20.

Jin, Q., et al. (2024), 'GeneGPT: augmenting large language models with domain tools for improved access to biomedical information', *Bioinformatics,* 40 (2).

Jorg, T., et al. (2024), 'A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports', *Insights Imaging,* 15 (1), 80.

Kankanhalli, Ziwei Xu and Sanjay Jain and Mohan (2024), 'Hallucination is Inevitable: An Innate Limitation of Large Language Models', *arXiv*.

Lee, J., et al. (2020), 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics,* 36 (4), 1234-40.

Likert, Rensis (1932), 'A Technique for the Measurement of Attitudes', *Archives of Psychology,* 140, 1–55.

Liu, C., et al. (2024), 'A CTCF-binding site in the Mdm1-Il22-Ifng locus shapes cytokine expression profiles and plays a critical role in early Th1 cell fate specification', *Immunity,* 57 (5), 1005-18 e7.

Nath, P. R., et al. (2024), 'Single-cell profiling identifies a CD8(bright) CD244(bright) Natural Killer cell subset that reflects disease activity in HLA-A29-positive birdshot chorioretinopathy', *Nat Commun,* 15 (1), 6443.

Pirooznia, M., Nagarajan, V., and Deng, Y. (2007), 'GeneVenn - A web application for comparing gene lists using Venn diagrams', *Bioinformation,* 1 (10), 420-2.

Qiu, Y., et al. (2024), 'scTPC: a novel semisupervised deep clustering model for scRNA-seq data', *Bioinformatics,* 40 (5).

Shen, H., et al. (2023), 'Generative pretraining from large-scale transcriptomes for single-cell deciphering', *iScience,* 26 (5), 106536.

Szalata, A., et al. (2024), 'Transformers in single-cell omics: a review and new perspectives', *Nat Methods,* 21 (8), 1430-43.

Thirunavukarasu, A. J., et al. (2023), 'Large language models in medicine', *Nat Med,* 29 (8), 1930-40.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020), 'Language models are few-shot learners', *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (159; Vancouver, BC, Canada), 1877 - 901.