A simple way to understand and annotate scRNAseq data with the help of open-source language models

Wei Kevin Zhang

Guangzhou National Laboratory, No. 9 XingDaoHuanBei Road, Guangzhou International Bio Island, Guangzhou, 510005, Guangdong Province, China

*Correspondence to: Wei Kevin Zhang

E-mail: zhang_wei@gzlab.ac.cn

Key words: open-source LLM, scRNA, annotation, GPT-OSS, cell function

Summary

Since its introduction in 2009, single-cell RNA sequencing (scRNA-seq) has transformed biological research by enabling large-scale cell mapping, uncovering novel cell subtypes, and integrating multiomics data. A critical step in this process is cell annotation, which links gene expression profiles to biological identities, aiding in understanding cellular heterogeneity and disease mechanisms. This step also serves as a valuable learning opportunity for researchers. To improve efficiency and security, we propose a semi-automated cell annotation workflow using OpenAI open-source GPT-OSS-20B model, which can be run locally via Ollama and integrated into R using the rollama package. This method is cost-effective, user-friendly, and provides precise annotations with biological insights, as demonstrated in our results. This approach simplifies scRNA-seq data analysis, promoting broader adoption and further advancements in the field.

Main text

Ever since the first true scRNAseq protocol published in 2009, this remarkable technique has revolutionized biological research. The primary applications of scRNA-seq fall into at least three key areas:

- Cell Atlas Projects: enabling large-scale mapping of cell types in organs and model organisms.
- Biological Insights: revealing novel cell subtypes, developmental trajectories, and gene regulatory networks.
- Multimodal Integration: expanding into "multiomics" studies by combining scRNA-seq with epigenomics and proteomics.

However, these advancements largely depend on the speed and accuracy of cell annotation—the process of assigning biological identities (e.g., cell types, states, or functional subsets) to individual cells or clusters based on their gene expression profiles. Cell annotation is a critical step that bridges raw sequencing data with biological interpretation, enabling researchers to understand cellular heterogeneity, developmental trajectories, and disease mechanisms.

Moreover, cell annotation should not be treated merely as an experimental step but also as an opportunity to extract biological insights, serving as a valuable learning process for biologists. From a practical standpoint, using locally deployed LLMs is more economical and secure than relying on commercial APIs, as all data can be processed and stored locally, minimizing risks associated with data transfer.

Hence, we present here a simple, semi-automated cell annotation workflow using an open-source large language model recently released by OpenAI, which offers excellent precision and a user-friendly interface.

Installation procedure

The preparation on environment and installation is quite simple and straight forward. Ollama is used to run the foundation of open-source LLMs. A large variety of AIs could be retrieved from the website of ollama without any payment. Here we strongly recommend to use gpt-oss:20b as our foundation model due to its high accuracy and low cost (meaning either time and hardware). rollama is an implementable package bridging ollama and the R environment, which could be retrieved directly from CRAN. All the information was listed in figure 1. Please be noted that to run gpt-oss:20b locally, an GPU with compute capability larger than 8 is highly recommended.

Annotation procedure

The annotation method mentioned in this manuscript is quite simple and easy to conduct. The whole procedure is composed of several lines of R commands as listed in supplemental file 1. The following prompt options could be modified case by case due to the specific demand:

You are a helpful assistant on gene functions and cellular markers.

Every answer must follow this format:

- H2 bold title.

- Exactly three bullet points (1–2 sentences each).

- Total length ≤ 300 words.

Results of annotation could be seen in figure 2. Gpt-oss:20b model gives a brief title on every cluster of cells, and also provided three informative bullet points for biological implications. One who have some biological background, can easily figure out and verify the cell types with marker genes and cell type briefly provided by the model. Moreover, one can also learn more information about the cellular functions and potential molecular function with the description provided by the model.

In summary, we present an easy-to-use method for understanding and annotating scRNA-seq data using the open-source language model GPT-OSS-20B. We hope this approach will facilitate further advancements in scRNA-seq technology and its applications.