[BIOINFO 2025]

scTAG-DB: Tissue-specific marker gene database approach for LLM-based cell type annotation in single-cell analysis

Jiyeong Shin[1], Minsik Oh[1,*]

[1] *Department of Data Technology, Myongji University*
**Corresponding author: msoh@mju.ac.kr*

Cell type annotation is a crucial step in single-cell RNA sequencing (scRNA-seq) analysis. Traditionally, this process begins with dimensionality reduction followed by manual annotation of cell clusters using canonical marker genes identified through literature search. While widely used, manual annotation is time-consuming, requires domain expertise, and lacks scalability. To overcome these challenges, computational methods have been developed for automatic annotation. These methods can be broadly divided into marker gene–based, reference-based, and supervised learning approaches. Recently, large language model (LLM)–based methods have emerged, leveraging advances in generative pre-trained transformers (GPT) to enhance annotation accuracy and interpretability in biomedical applications.

Here, we present scTAG-DB, an LLM-based workflow that integrates tissue-specific marker databases with GPT-4.1 to perform automated cell type annotation. To enable accurate annotation, scTAG-DB combines differential gene expression analysis with curated marker databases. At the cluster level, scTAG-DB identifies differentially expressed (DE) genes and matches them with tissue-specific marker genes curated from three resources: singleCellBase, CellMarker, and MarkerGeneBERT. These markers, together with tissue context, are incorporated into a structured prompt provided to the LLM. The model outputs predicted cell types, highlights supporting marker genes, and provides explanations for its decisions. Annotation results are exported in standard formats (CSV, JSON, and AnnData), and a web-based interactive environment enables easy exploration of results.

We evaluated scTAG-DB across eight tissue datasets and benchmarked it against GPTCelltype, CellTypist, and SingleR. Predicted labels were compared with manual annotations from the original studies and classified as full matches, partial matches, or mismatches. Across datasets, scTAG-DB showed strong concordance with manual annotations and improved accuracy relative to existing

methods, demonstrating how database-filtered prompting strategies can enhance LLM-based cell type annotation.