NLP-based biological cell type annotator scRAGAnnot for single-cell transcriptomics

Nikki Tirrell[a], Rahul Paul[a], Alexander Murray[a], Arya Eskandarian[a], Spyros Karaiskos[a], Luis Santana-Quintero[a], Sreenivas Gannavaram[b]

a – Office of Biostatistics and Pharmacovigilance (OBPV), Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Silver Spring, MD, USA
b – Office of Blood Research and Review (OBRR), Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD, USA

Abstract

Cell type annotation is a crucial step in single-cell transcriptomic analyses.

- To discover safety/efficacy signals, robust tools that best predict cell types from marker genes are needed as the usage of single-cell transcriptomics in clinical trials grows.
- Such tools should consider the diversity of cells types found in different tissues without bias.
- Our proposed solution uses advanced Natural Language Processing (NLP) techniques to annotate cell types.

The proposed approach will enable:

1. Accurate prediction of cell type(s) for cells given as input.
2. Establishment of specific marker genes for known cell types.

Introduction

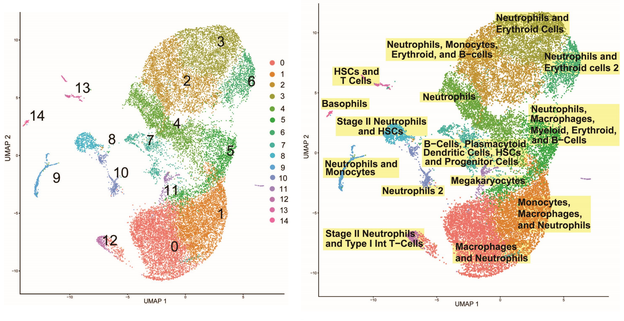
Clustering cells based on gene expression is a standard practice in single-cell expression analysis. Each cluster is distinguished by a certain set of marker genes.

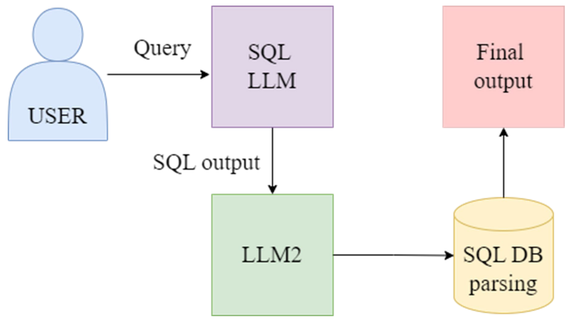
Obstacle: Accurate annotation of said cell clusters from different tissues remains a challenge due to the lack of high-quality reference atlases. Currently available annotation options rely heavily on resources oversaturated with immune cell annotations, leading to "imbalanced datasets".

Objectives: Develop a tool that will assist in annotation of single-cell RNA-seq data.

1. Compile currently published single cell transcriptomics marker gene data into one database for our reference.
2. Utilize Large Language Models (LLMs) to take in marker genes and search the database for associated cell types.
3. Handle searches with multiple genes and multiple databases.

Regulatory Relevance: This effort aims to serve as an internal and independent solution to evaluate submissions that contain single-cell transcriptomics data.





Materials and methods

Data Collection:

Cell marker data was obtained as bulk download files from all publicly available cell marker databases: CellMarker, CellMarker 2.0, Human Protein Atlas, PanglaoDB, and CELLxGENE.
Note: Other single-cell databases for cluster annotation, such as Single Cell Expression Atlas, that do not provide marker gene data, were not considered.

Data Processing:

These bulk data files were parsed for homo sapiens cell type marker genes and saved to a standardized table.

- CellMarker and CellMarker 2.0 data were manually cleaned to remove special characters.
- PanglaoDB data was filtered to only include cell type markers with a specificity ≤ 0.2 and a sensitivity ≥ 0.9[1].
- The top 150 gene markers from the Human Protein Atlas with the highest nTPM were saved for each cell type. Tissue type is not specified, so this was set to 'Undefined'.
- The top 100 gene markers were used from CELLxGENE, then the 'All Tissues' tissue type was removed.

After processing, all relevant data was assembled into a single table. This table was then converted into a SQL database for querying.

scRAGAnnot Tool:

- An SQLite schema from the table was utilized by the LLM SQL-Coder-8b[7] to translate human query to SQL query.
- Then, a second LLM (Mistral-7B[8]), is used to parse the SQL database using the SQL query and return the output.

Figure 1. UMAP plots of bone marrow scRNA-seq data before (left) and after (right) cell type cluster annotation

Results and discussion

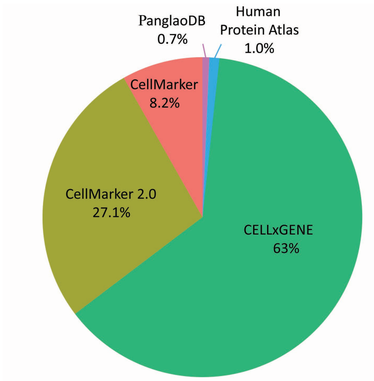


Figure 3. Proportion of total entries for each individual resource in the aggregate database

Database Statistics:

- 5 public resources were utilized to populate this database.
- There are a total of 7,934 cell types with marker genes in this combined database.
- The marker gene lists have a minimum of 3 genes, a median of 32 marker genes, and a mode of 100 marker genes.
- There are 30,375 marker genes in this combined database.
- The most common gene marker is CD3D, found in 656 cell type lists across the databases.

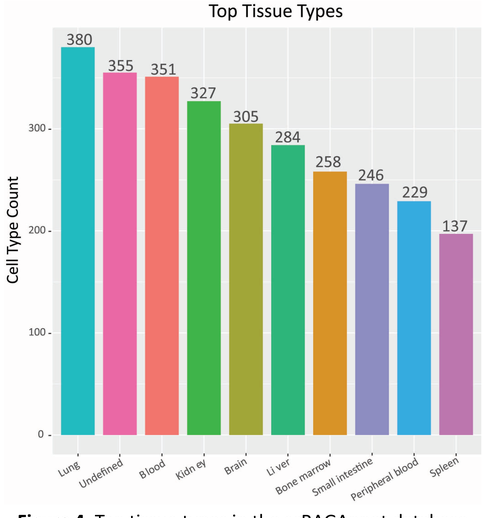


Figure 4. Top tissue types in the scRAGAnnot database

scRAGAnnot Results:

- We have prepared and assembled all sources to minimize bias. Pilot studies demonstrated that an NLP-based QA pipeline could effectively parse this database.
- **scRAGAnnot** is accessible via HIVE's Jupyter platform.
- The user enters a query on genes or cell types and can specify to limit results to certain tissue types and/or databases.
- **scRAGAnnot** parses the database.
- **scRAGAnnot** returns the relevant information to the user.

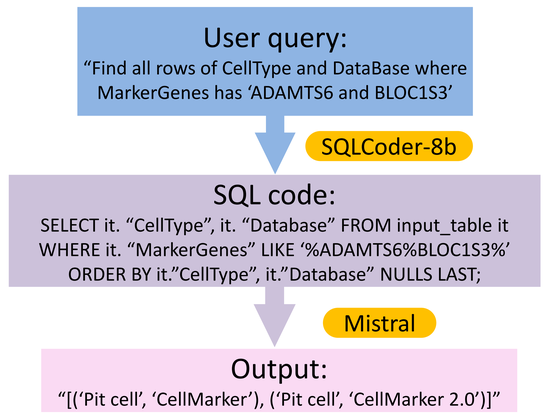


Figure 5. Sample use of **scRAGAnnot**

Conclusion

- Developed an NLP-based tool to provide cell-type annotations for single-cell RNA-seq data.
- Our tool will provide a simple interface for users to determine cluster cell types based on provided marker genes.
- The initial implementation (the prototype) of the cell annotator tool is going to be hosted at the HIVE platform, and currently is available to registered users for testing via a Jupyter notebook.
- Next steps include validating **scRAGAnnot**, creating a user interface for this tool, and expanding cell type annotations to additional species starting with M*us musculus*.

References

1. https://www.nature.com/articles/s41587-021-01188-9

THE HUMAN PROTEIN ATLAS · Hugging Face · CellMarker · CellMarker 2.0 · PanglaoDB · CZ CELL×GENE DISCOVER

Disclaimer: *The information in this presentation represents the opinions of the speaker and does not necessarily represent FDA's position or policy.*