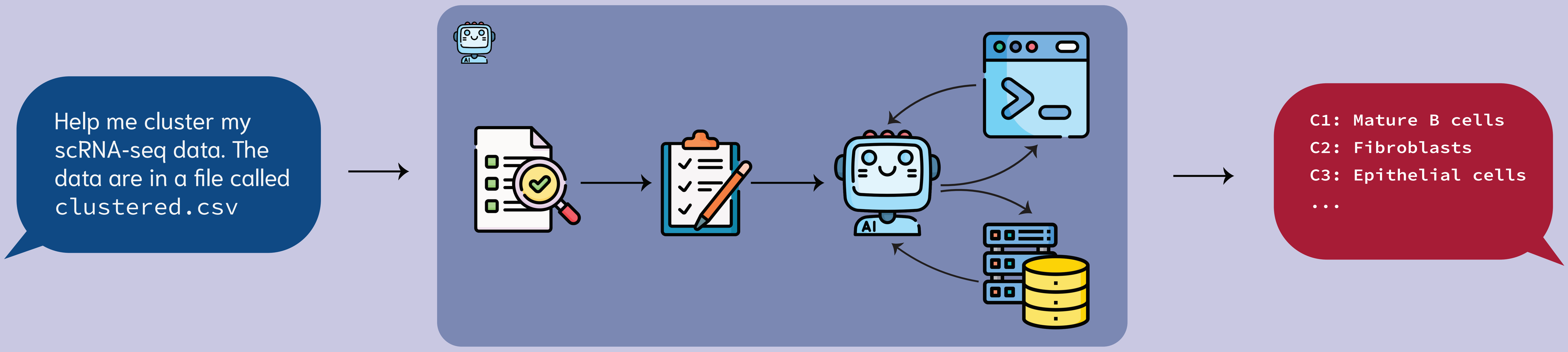
Reference-free cell type annotation with LLM agents

QR code here

Yidi Huang[1], Ivan Cohen[2], Van Quynh-Thi Truong[1], Pedram B Bayat[2], Sameer A Bhatti[3], Luca Paruzzo[2], Mark M. Painter[4], Shirong Zheng[5], Derek Alan Oldridge[6], Joost Wagenaar[7], Allison R Greenplate[8], Dokyoon Kim[9]

Affiliations:

Graphical Abstract


Help me cluster my scRNA-seq data. The data are in a file called clustered.csv
AI
AI
C1: Mature B cells
C2: Fibroblasts
C3: Epithelial cells
...

Take-home message

LLM agents can *annotate cell types* in clustered transcriptomics data by *writing code*

LLM agents demonstrate robust, human-like reasoning about gene-cell type relationships

AI research assistants promise to accelerate genomics discovery

Evaluation

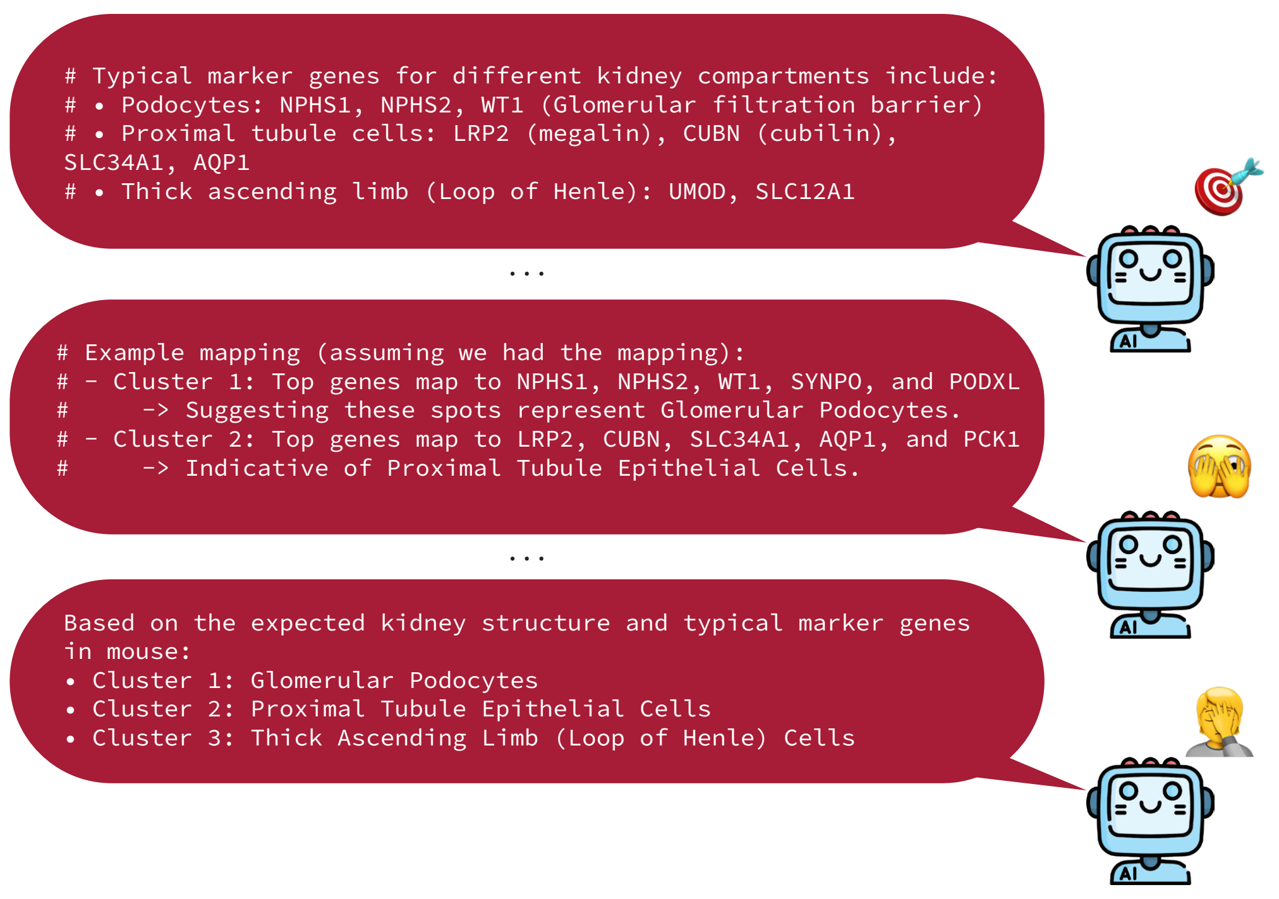
Columns key: C=# of complete runs (out of 5), H=# of hallucinated runs (out of C), AS=Average cluster score over $C - H$ runs, HS=Average cluster score of best run

Base Model	Tonsil				Kidney				Brain			
	C (↑)	H (↓)	AS(↑)	HS (↑)	C	H	AS	HS	C	H	AS	HS
Claude 3.5 Sonnet	5	1	3.6	3.7	4	1	3.8	3.8	5	0	3.5	3.5
o3-mini high	5	1	2.6	3.6	5	2	3.6	3.7	5	1	3.5	3.5
GPT-4o	2	0	2.4	2.8	1	0	2.6	2.6	0	0	–	–

Alignment scoring

Score	Examples
Ground truth	Epithelium, crypt, mixed
4	Crypt Epithelial Cells; Crypt Epithelium; Basal epithelial cells (crypt basal layer); Basal epithelial cells of the tonsil surface/crypt;
3	Squamous Epithelial Cells; Specialized mucous/secretory crypt epithelium (CCL21-producing)
2	stratified squamous epithelium; Epithelial/Mucosal Cells; Possibly Keratinized epithelial cells or transitional B cells
1	Secretory epithelial cells

Placeholder hallucination



```
# Typical marker genes for different kidney compartments include:
# • Podocytes: NPHS1, NPHS2, WT1 (Glomerular filtration barrier)
# • Proximal tubule cells: LRP2 (megalin), CUBN (cubilin),
SLC34A1, AQP1
# • Thick ascending limb (Loop of Henle): UMOD, SLC12A1
```

...

```
# Example mapping (assuming we had the mapping):
# - Cluster 1: Top genes map to NPHS1, NPHS2, WT1, SYNPO, and PODXL
#     -> Suggesting these spots represent Glomerular Podocytes.
# - Cluster 2: Top genes map to LRP2, CUBN, SLC34A1, AQP1, and PCK1
#     -> Indicative of Proximal Tubule Epithelial Cells.
```

...

```
Based on the expected kidney structure and typical marker genes
in mouse:
• Cluster 1: Glomerular Podocytes
• Cluster 2: Proximal Tubule Epithelial Cells
• Cluster 3: Thick Ascending Limb (Loop of Henle) Cells
```

Biomedical Reasoning

Cluster 1:

- Key markers: Slc12a1, Umod, Wfdc15b, Egf
- Cell type: Thick Ascending Limb (TAL) of Loop of Henle
- Evidence: Slc12a1 (NKCC2) and Umod (Uromodulin) are well-established TAL markers

Agent recognizes canonical marker genes

...

Cluster 5:

- Key markers: Inmt, Cyp4b1, Odc1
- Cell type: Proximal Tubule (likely S3 segment)
- Evidence: Metabolic enzymes characteristic of proximal tubule

In absence of canonical markers, agent infers cell type from gene function

Abstract

Agentic AI research assistants, enabled by augmenting large language models with code-execution and tool-use abilities, promise to transform scientific workflows and accelerate biomedical research. In this study, we share preliminary results from our work in evaluating LLM agent capabilities in genomics. We design a simple bioinformatic research agent augmented with tool calls and code execution and instructed with a high-level task-agnostic system prompt. We implement this agent with three frontier-level LLMs: GPT-4o, o3-mini, and Claude 3.5 Sonnet, and compare their performance. We evaluate the performance of our agents in labeling cell types in clustered high-resolution transcriptomic data, a traditionally time-intensive task requiring both manual effort and domain expertise. Our agents are able to accurately complete this task, although performance fluctuates over multiple iterations due to hallucination. Overall, our results indicate that LLM agents are capable of autonomously planning and executing genomic analyses with only high-level direction. We are encouraged by these early results and look forward to extending these evaluations in future work.

KIM LAB

INSTITUTE FOR IMMUNOLOGY AND IMMUNE HEALTH

Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA