BioData Mining

RESEARCH

Open Access

Integrative evidence-knowledge marker selection enhances LLM-based cell type annotation in single-cell RNA-seq analysis

Jiyeong Shin[1], Soyoung Jeong[2,3], Hyun Je Kim[2,4,5] and Minsik Oh[1*]

*Correspondence:
Minsik Oh
msoh@mju.ac.kr
[1]Department of Artificial Intelligence, Myongji University, Seoul, Republic of Korea
[2]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Republic of Korea
[3]Genomic Medicine Institute, Seoul National University Medical Research Center, Seoul, Republic of Korea
[4]Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea
[5]Interdisciplinary Program in Artificial Intelligence (IPAI), Seoul National University, Seoul, Republic of Korea

Abstract

Background Cell type annotation is essential for gaining biological insight from single-cell RNA sequencing data, yet manual labeling remains time-consuming and difficult to reproduce. Various computational approaches have been developed to automate this process, and recent studies suggest that large language models can infer cell types with promising accuracy in single-cell analysis. However, most workflows still rely on cluster-specific markers derived from gene expression alone or manual curation. As a result, marker selection can be sensitive to statistical criteria and dataset-dependent bias, which may lead to the selection of less informative genes or missing important markers, while providing limited biological context.

Results To address this limitation, we introduce CELLIA, an LLM-based workflow for automated and robust cell type annotation. CELLIA employs an integrative evidence-knowledge marker selection strategy that combines statistical differential expression criteria with curated tissue-specific marker resources to identify informative marker genes. In benchmarking analyses of 102 cell types, this approach improved agreement with manual annotations. In addition, CELLIA achieved higher agreement in subtype-level analyses of closely related immune populations and was further evaluated in a non-immune stromal subtype setting, covering 25 cell types in total.

Conclusion By integrating evidence-knowledge from gene expression with curated biological prior knowledge, CELLIA provides a more stable marker selection and improves the reliability of LLM-cell type annotation.

Keywords Single-cell RNA-seq, Large language model, Cell type annotation, Information integration, Marker gene

Background

Cell type annotation is a crucial step in single-cell RNA sequencing (scRNA-seq) analysis, linking cellular identity to disease etiology and enabling targeted treatments [1]. In standard single-cell analysis workflows, cell type annotation is typically performed downstream of preprocessing and clustering [2]. Clustering serves as a key step that partitions heterogeneous tissues into transcriptionally distinct subsets, which are then

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

manually annotated using literature-derived canonical marker genes [3, 4]. However, this process is not only time-consuming but also difficult to reproduce, so automated approaches are increasingly adopted [4].

To improve the substantial time and expertise demands inherent in manual cell type annotation of scRNA-seq data, numerous computational approaches have been developed. These automated methods utilize on diverse strategies, including marker-based, reference-based, supervised classification, and deep learning, to achieve accurate and scalable cell type identification.

Tools that utilize marker-based automatic annotation such as Garnett [5], SCINA [6], and ScType [7] assign cell labels based on the characteristic expression of predefined markers [8]. Although these approaches can abstain from uncertain predictions when marker evidence is insufficient, their applicability is often limited by incomplete marker resources across tissues. Reference-based methods, including scmap [9] and SingleR [10], instead annotate cells by matching expression profiles to labeled reference datasets [8]. Supervised learning frameworks such as CellTypist extend this strategy by training machine-learning models on large reference atlases and providing hierarchical label sets for flexible annotation [11]. More recently, deep learning approaches such as scBERT leverage large-scale pretraining to model gene expression patterns in single-cell data [12]. Despite these advances, many approaches remain sensitive to feature selection and data quality, which can limit their ability to capture subtle biological variation [12, 13].

Recent advances have incorporated large language models (LLMs), opening new possibilities for cell type interpretation. Unlike methods that rely solely on expression matrices, LLMs leverage large-scale pretraining and generative pretrained transformers (GPT) to enable reasoning-based inference and contextual knowledge integration [14, 15]. Benchmarking studies show that GPT models can infer cell types with expert-level accuracy when provided with marker gene inputs, highlighting their practical utility for single-cell analysis [16].

Alongside these developments, several frameworks have further extended LLM-based annotation. LICT introduces a reliability evaluation framework combining multi-model integration and iterative feedback to reconcile LLM predictions with expert annotations [17]. Similarly, mLLMCelltype adopts a multi-LLM consensus strategy to reduce model bias and quantify uncertainty [18]. Agent-based systems such as CASSIA further decompose annotation into specialized reasoning steps, including annotation, validation, and interpretability modules [19]. Together, these approaches highlight a shift from single-pass prompting toward more structured and reliability-aware LLM-based annotation frameworks.

Despite recent advances, most current LLM workflows still inherit a key limitation from traditional pipelines by relying on cluster-specific markers derived from differential expression statistics [20]. Such markers can be unstable or weakly discriminative when clusters are transcriptionally similar or when background heterogeneity is high [15]. To mitigate this issue, several resources curate tissue-specific marker genes. Manually curated databases include singleCellBase, a literature-derived compendium supporting standardized cell type annotation across species [21], and CellMarker 2.0, which integrates markers from scRNA-seq datasets and published studies to provide expanded and updated coverage [22]. Complementarily, MarkerGeneBERT applies an NLP-based pipeline to automatically extract and normalize marker associations from biomedical

literature [23]. While these resources highlight the importance of tissue-specific marker knowledge for cell type annotation, their largely static nature limits the incorporation of newly emerging biological information. To address this limitation, we integrate expert-curated and automatically extracted marker resources within a dynamically updatable framework.

Here, we propose CELLIA (Curated tissue-specific knowledge and evidence Integration for LLM-based cell type Annotation), an LLM-based cell type annotation workflow that uses an integrative evidence-knowledge for marker selection. In this framework, "evidence" refers to combined statistical differential expression criteria, including adjusted P-value (p-value), average $\log_2$ fold-change ($\log_2 \mathrm{FC}$), and expression proportion, whereas "knowledge" denote curated tissue-specific marker database derived from prior biological studies.

CELLIA comprises of two main components: (i) an integrative evidence-knowledge marker selection that prioritize biologically grounded and statistically supported markers for accurate annotation, and (ii) an interactive web interface for reviewing and visualizing annotation results. By integrating evidence-knowledge marker selection, CELLIA moves beyond simple cluster-level heuristics and establishes a more consistent and context-aware basis for LLM-based cell type annotation.

Methods

Overview of CELLIA

The workflow of CELLIA is straightforward. First, CELLIA operates on an scRNA-seq dataset at the cluster level. It identifies differentially expressed gene (DEG) for each cluster (Fig. 1; Step A). To extract tissue-specific markers, CELLIA cross-references each

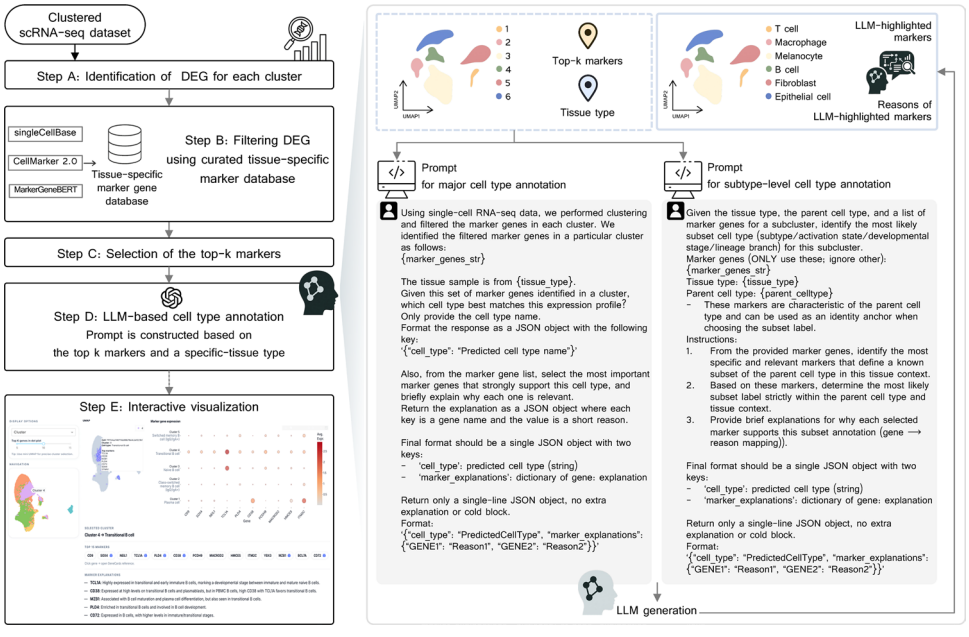


Fig. 1 Overview of CELLIA. CELLIA comprises four main steps (**A**–**D**) plus an interactive web component (**E**). Input: scRNA-seq dataset summarized at the cluster level. Step A: identify DEG for each cluster. Step B and C: filter cluster DEG lists against curated tissue-specific marker resources and retain the top-k markers per cluster. Step D: construct a structured prompt integrating the selected markers with tissue context and submit it to the LLM for annotation. Step E: export annotations in standard formats and provide a web-based interface for interactive visualization and explanations

cluster's DEG list with tissue-specific marker databases and retains the top-k markers (Fig. 1; Step B and C). A structured prompt is then constructed by combining the selected markers with the tissue context and submitted to the LLM for annotation (Fig. 1; Step D). Throughout this entire framework, CELLIA leverages the AnnData. Widely adopted in scRNA-seq analysis, it enables efficient storage of annotated matrices and handling of large-scale datasets [24]. Finally, annotation results can be interactively explored through a web-based interface without requiring computational expertise (Fig. 1; Step E).

Step A. Identification of DEG for each cluster

In scRNA-seq analysis, DEG is defined as those showing statistically significant expression differences between groups [25]. Marker genes represent a more selective subset of DEG that also exhibit cluster specificity and biological interpretability beyond statistical significance [25].

DEG is identified for each cluster using Scanpy's gene ranking function to characterize cluster-level transcriptional programs [26]. Genes are ranked using the Wilcoxon rank-sum test and filtered by p-value and $\log_2 \mathrm{FC}$. Retained genes are subsequently ranked by the $\log_2 \mathrm{FC}$ to prioritize those showing strong and specific upregulation within each cluster. To further refined DEG, we incorporate a percentage metric comprising the proportion of cells within the target cluster ($\mathrm{pct}.1$) and in other clusters ($\mathrm{pct}.2$). Candidate markers are retained using predefined thresholds (p-value < 0.05, $\mathrm{pct}.1 > 25\%$, $\mathrm{pct}.2 < 35\%$, and $\log_2 \mathrm{FC} > 0.3$) and subsequently ranked in descending order of their $\log_2 \mathrm{FC}$.

Step B. Filtering DEG using curated marker database

To prioritize biologically meaningful markers, candidate DEGs are filtered using a curated marker database constructed for the CELLIA framework. The database integrates three complementary tissue-specific marker resources: singleCellBase, CellMarker 2.0, and MarkerGeneBERT. These resources combine manually curated expert knowledge with markers automatically extracted from the biomedical literature, providing both reliable reference information and coverage of newly reported markers.

Candidate DEG is cross-referenced with tissue-specific markers complied from three complementary resources. This step retains genes supported by published, tissue-specific annotations, thereby refining each cluster's DEG list to markers that are biologically plausible and contextually relevant.

Step C. Selection of the top-k markers

From this refined list, the top k markers are selected to represent each cluster. In cases where fewer than k markers are available, the marker set is supplemented with the remaining DEG (i.e., those excluded during database cross-referencing) exhibiting the highest $\log_2 \mathrm{FC}$ until the target size is reached. This approach ensures that each cluster is characterized by a balanced combination of validated tissue-specific markers and data-driven candidates. For our framework, we set $k = 15$, as this value yielded stable and accurate annotation performance.

Step D. LLM-based cell type annotation

CELLIA takes the top 15 markers (from Step C) and the tissue context as input, then constructs a structured prompt based on the template (Fig. 1). In this prompt, 'marker_gene_str' represents a list of markers, and 'tissue_type' and 'parent_celltype' are variables that are replaced with the actual name of the tissue (e.g., human PBMC) and the corresponding major cell group (e.g., B cell). Genes belonging to the same cell population are separated by commas.

The LLM was queried via its API using the constructed prompt. The returned response was programmatically parsed to extract two key components: (i) the predicted cell type annotation and (ii) a subset of highly informative markers with their associated biological rationales. This structured parsing enabled simultaneous retrieval of both annotation results and interpretable reasoning traces underlying the LLM's decision. Specifically, the result included key markers and concise explanations describing their functional roles in supporting the assigned identity. Finally, these results are integrated back into the AnnData as input and exported in JSON and CSV formats to ensure reproducibility and accessibility through the web-based interface (Step E).

Step E. Interactive visualization

We developed an interactive web environment for CELLIA to provide a transparent and user-friendly way to review annotation results without requiring computational expertise (Fig. 2). After completing the CELLIA workflow, the CELLIA's results are supplied to the web interface for visualization and review. The interface utilizes these outputs, including (i) cell-level metadata (cluster identity and predicted label), (ii) cluster-level

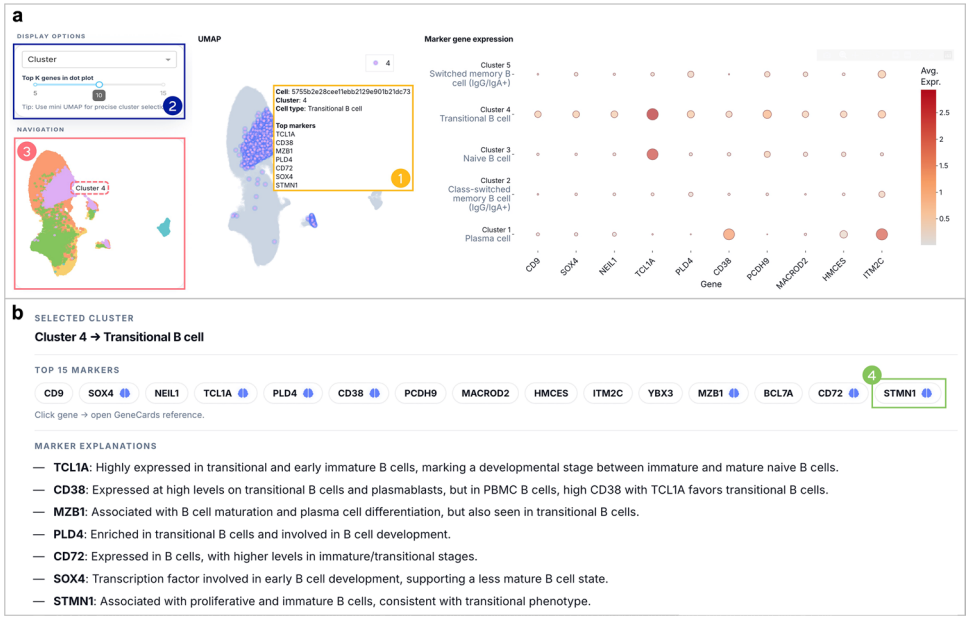


Fig. 2 CELLIA's interactive web-based interface for annotation review. The CELLIA interface provides a web-based, point-and-click environment for inspecting predicted annotations, exploring marker information, and validating results without computational programming requirements. (**a**) Overview of the interface showing the primary UMAP initialized with CELLIA predictions. Interactive controls allow selection of clusters of interest (3), adjustment of the number of displayed markers ($5 \leq k \leq 15$) (2), and visualization of gene expression patterns through cluster-level dot plots. (**b**) Cluster-level annotation panel displaying the predicted label, detailed report for the selected cluster, the top-ranked marker genes, including informative markers highlighted by the LLM during prediction (4), and each gene hyperlinked to the GeneCards database for external reference

informative markers identified by the LLM during the annotation process, with concise rationales, and (iii) a UMAP embedding with associated expression values and metadata.

The environment comprises two coordinated visualization views. The primary UMAP offers an overview of the entire dataset and displays cell-level metadata, including cluster identity and predicted labels (Fig. 2a (1)). The secondary UMAP enables selection of clusters of interest and presents detailed cluster-level profiles (Fig. 2a and b). For each selected cluster, dot plots and corresponding explanations are displayed. The dot plot complements its explanatory features with flexible visualization controls, allowing for the dynamic re-rendering of plots as the parameter k ($5 \leq k \leq 15$) is adjusted (Fig. 2a (2)). The associated explanation view is populated from precomputed CELLIA results, reporting: (i) the cluster identifier, (ii) the predicted cell type label, (iii) the top-k marker list with each gene hyperlinked to the GeneCards database for external reference, and (iv) concise LLM-generated rationales that link informative markers to the prediction and describe the reasoning underlying the assigned identity (Fig. 2b). Informative markers are visually distinguished within the list to provide clear interpretive focus (Fig. 2b (4)). To ensure consistency and reproducibility, all visual elements are deterministically rendered from the stored results.

Results

Benchmark datasets for evaluating CELLIA

All analyses were conducted using publicly available datasets. We benchmarked CELLIA nine different tissue datasets across eight tissue types—colorectal cancer (CRC; GSE132257), lung (GSE131907), breast (GSE167036), peripheral blood mononuclear cells (PBMCs; Human Immune Health Atlas (HIHA) [27]; and 10X), liver (GSM4008623), kidney (GSM4008620), skin (GSM4008710), and pancreas (GSM4008637)—representing 55 unique cell types. For subtype-level evaluation, we focused on immune populations from the HIHA PBMC [27] dataset, assessing four major cell groups—T cell (CD4T, DNT, and Treg; CD8T, gdT, and MAIT), dendritic cell, and B cell—comparing 19 cell types in total (Table 1). To additionally examine a non-immune subtype setting, we also used fibroblasts from the NSCLC subset of the human perturbed-state fibroblast

Table 1 Summary of scRNA-seq datasets used in this study

Dataset	Cell types	Cells	Genes
Major cell type annotation			
CRC	6	18,409	33,694
Lung	5	20,304	29,634
Breast	12	49,141	25,301
PBMC (HIHA)	9	1,821,725	32,357
Liver	16	1,811	12,631
Kidney	20	9,966	26,474
PBMC (10X)	9	5,702	38,606
Skin	13	1,697	20,119
Pancreas	12	9,727	19,707
Subtype-level cell type annotation			
T cell (CD4T, DNT, Treg)	5	743,615	32,357
Dendritic cell	4	23,287	32,357
B cell	5	160,632	32,357
T cell (CD8T, gdT, MAIT)	5	406,730	32,357
Fibroblast	6	2,055	17,410

atlas reported by Buechler et al. [28]. Including this additional fibroblast analysis, the subtype-level evaluation covered 25 unique cell types in total.

Evaluation metrics for cell type annotation

We compared CELLIA against representative automated scRNA-seq cell type annotation tools, including GPTCelltype, CASSIA, LICT, mLLMCelltype, CellTypist, SingleR, and Seurat. All methods were executed using their recommended configurations. For each dataset, we followed the procedures described in the original publications or, when available, used the normalized matrices provided by the authors.

Predicted cell type annotations by CELLIA and the baseline methods were compared against the manual labels provided in the original studies using a three-level agreement scheme adapted from established benchmarking practice [16]. For each evaluated cluster, the predicted label and the corresponding manual label were first matched to an unambiguous Cell Ontology (CL) term and, when applicable, to a broader cell type category. A prediction was classified as fully match if it corresponded to the same annotation term, the same CL term, or a more specific subtype of the manual label. For example, a prediction of $CD4^+$ T cell or $CD8^+$ T cell was considered fully matched when the manual label was T cell. A prediction was classified as partially match if it belonged to the same broad cell type category but differed in annotation and CL term, including cases in which the predicted label was broader than the manual label. For example, a prediction of T cell was considered partially matched when the manual label was $CD4^+$ T cell. A prediction was classified as mismatch if the predicted and manual labels differed in broad cell type category, annotation term, and CL term.

To enable quantitative comparison, each evaluated cluster i was assigned an agreement score:

$$\text{agreement_score}_i = \begin{cases} 1, & \text{if fully match} \\ 0.5, & \text{if partially match} \\ 0, & \text{if mismatch} \end{cases} \tag{1}$$

For a given dataset, the proportion assigned to each agreement category a was calculated at the cluster level as:

$$\text{proportion } (a) = \frac{\text{number of cell types in } a}{\text{total number of cell types}} \tag{2}$$

where $a \in$ {fully match, partially match, mismatch}.

To summarize overall concordance within each dataset, we calculated the average agreement score across all evaluated clusters:

$$\text{average agreement} = \frac{1}{N} \sum_{i=1}^{N} \text{agreement_score}_i \tag{3}$$

where N is the total number of evaluated clusters in the dataset.

Both the agreement proportions and the average agreement score were therefore calculated using clusters as the evaluation unit. When a manually annotated cell type was represented by multiple clusters, each cluster was scored separately and contributed

independently to the dataset-level summary. These metrics summarize annotation outcomes at the cluster level and provide a dataset-level average of agreement.

Comparison with existing annotation methods

We benchmarked CELLIA against three traditional computational methods (CellTypist, SingleR, and Seurat) and four recent LLM-based frameworks (GPTCelltype, CASSIA, LIcT, and mLLMCelltype) using the same pre-clustered, normalized single-cell datasets. This design was intended to focus the comparison on the annotation step while controlling for differences in the input clustering results. For cell-level annotation methods, including CellTypist, SingleR, and Seurat, predictions were first generated at the cell level and then aggregated into cluster-level labels by majority voting for comparison within the cluster-level benchmark. LLM-based methods were evaluated on the same clustered input data following their original annotation workflows. Detailed comparison results for each method are provided in Supplementary Tables 1–9.

Traditional computational methods

CELLIA showed overall favorable performance relative to traditional methods in both average agreement scores and the overall distribution of agreement categories across the evaluated datasets (Fig. 3a). While these tools provide rapid and scalable cell identity assignment, their performance can be sensitive to biological context particularly in complex or heterogeneous tissues. Evaluation of Seurat, SingleR, and CellTypist on the PBMC (HIHA) dataset could not be completed. Seurat and SingleR failed during conversion of the input dataset into their respective R-based object formats, before reference mapping or correlation scoring. CellTypist terminated with a process-level error during execution. All three failures were consistent with memory exhaustion given the limitations of the available shared computing resources (1 TB memory; see Supplementary Table 16 for details). A Seurat-based label transfer baseline was otherwise evaluated only for the lung, kidney, PBMC (10X), and pancreas datasets, for which compatible reference atlases were available. Seurat was not evaluated for CRC, breast, liver, or skin because no suitable reference atlas was available for these tissues.

LLM-based methods

To evaluate performance across different LLMs, we benchmarked CELLIA using two models: GPT-4.1 (June 2024) and Gemini 2.5 Flash (June 2025).

Across the benchmark datasets, CELLIA showed generally strong agreement with manual labels (Fig. 3b). The relative differences between methods were modest overall and varied across datasets. Most methods showed broadly similar agreement patterns, although some dataset-specific differences in agreement composition were observed in kidney and immune-related populations. Overall, CELLIA remained comparable to the other LLM-based methods. Supplementary Figure 1 is presented to show dataset-specific agreement patterns. In some datasets, CELLIA showed a higher proportion of fully matched annotations or a lower proportion of mismatches (Supplementary Tables 1–9).

To further examine the effect of cluster definition on annotation results, we additionally evaluated CELLIA and the baseline LLM-based methods under alternative Leiden clustering resolutions (0.05, 0.1, and 0.3). Because changing clustering resolution alters cluster number and composition, each newly generated Leiden cluster was assigned a

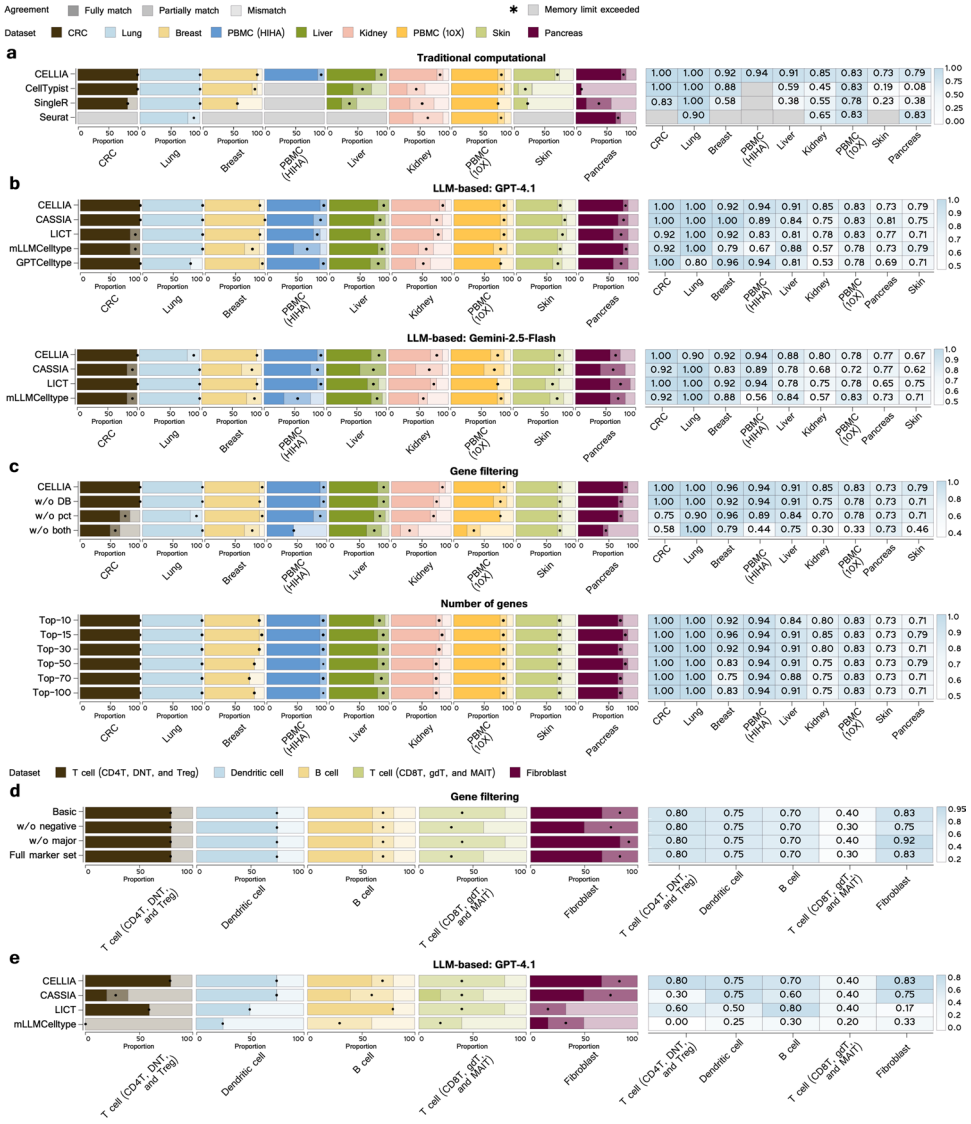


Fig. 3 Performance evaluation of CELLIA. Proportion of cell types with varying agreement levels (left) and average agreement scores (right). Agreement scores for major cell type annotation across 9 datasets, (**a**) comparing CELLIA with traditional computational methods (CellTypist, SingleR, and Seurat) and (**b**) LLM-based annotation frameworks (GPTCelltype, mLlmcelltype, LICT, and CASSIA). (**c**) Effect of gene filtering strategies on annotation performance, comparing CELLIA with ablated variants lacking curated database filtering (w/o DB), proportion filtering (w/o pct), or both. Sensitivity analysis of the number of top-ranked marker genes provided to the LLM, evaluated across tissues using different input sizes. Subtype-level annotation performance across immune and non-immune subtype datasets, (**d**) impact of subtype-level marker configurations, comparing the Basic filtering strategy with variants incorporating major cell type markers, negative markers, or both, and (**e**) comparison of CELLIA with other LLM-based frameworks. Subtype-level analyses included four immune cell group datasets from PBMC (HIHA) and an additional fibroblast subtype setting from the NSCLC subset of the human perturbed-state fibroblast atlas. Agreement categories are shown as fully matched, partially matched, or mismatched

reference label by majority voting over the original manual clusters contained within it. These additional analyses suggested that the influence of clustering resolution varied across datasets, with some settings showing relatively similar agreement across resolutions and others showing more variation, particularly at higher resolution. A similar tendency was observed in subtype-level analyses, where finer-grained annotation produced more variable agreement patterns across resolutions. The corresponding major cell-type and subtype-level results are shown in Supplementary Figures 2–5.

Ablation analysis of marker selection

To systematically evaluate how input marker list affect LLM performance, we benchmarked several key parameters for cell type assignment. Each statistical criterion is detailed in the methods section. Ablation analyses are summarized in Supplementary Tables 10 and 11.

Gene filtering methods

We assessed four distinct methods for filtering identified marker genes. All methods ranked genes by descending $\log_2$ FC as the final step.

(i) p-value+pct+$\log_2$ FC+Database (CELLIA): Genes were filtered by p-value, pct.1, pct.2 and $\log_2$ FC, then cross-referenced using a curated database corresponding to the input tissue type.
(ii) p-value+pct+log2FC (w/o DB): Genes were first filtered by p-value, then by pct.1, pct.2, and $\log_2$ FC.
(iii) p-value+$\log_2$ FC+Database (w/o pct): Genes were first filtered by p-value and $\log_2$ FC, and cross-referenced using a curated database corresponding to the input tissue type.
(iv) p-value+$\log_2$ FC (w/o both): Genes were filtered solely using differential expression statistics (p-value and $\log_2$ FC).

These comparisons show the specific impact of evidence-knowledge marker selection, evaluating their individual and combined roles in CELLIA's overall performance. Across tissues, CELLIA achieved higher agreement with manual annotations than most ablated variants (Fig. 3c). The advantage was most apparent in kidney tissue for Loop of Henle–related cell types, where CELLIA produced full matches across all evaluated cases, while other methods frequently yielded partial or incorrect annotations (Fig. 3c and Supplementary Table 10). The removal of the database filter (w/o DB) or the expression proportion metric (w/o pct) resulted in a reduction in agreement across most tissues (Fig. 3c). The simultaneous removal of both components (w/o both) resulted in the largest reduction in agreement (Fig. 3c). These findings suggest that differential expression statistics alone are not enough for reliable annotation, and that both biological knowledge and evidence filtering are required for stable marker selection and accurate cell type annotation.

Number of genes

We evaluated the impact of the number of top genes provided to the LLM annotation performance, testing input sizes of 10, 15, 30, 50, 70, and 100 genes under the CELLIA gene filtering method (Fig. 3c). Performance was broadly stable across input sizes, although slight decreases were observed in breast and kidney datasets as the number of top-ranked genes increased.

Subtype-level cell type annotation

Reliable annotation at the subtype level remains challenging, particularly for rare or less well-characterized populations, which often lack well-established markers or reference datasets [29]. We therefore extended our evaluation beyond major cell type annotation to assess LLM performance at the subtype level. For subtype-level benchmarking, we

primarily used the PBMC subtype dataset from the Human Immune Health Atlas, in which major immune cell classes were selectively subset and iteratively reprocessed by reclustering. Specific subpopulations were first identified within each major class using marker-based criteria, and each selected subset was subsequently reanalyzed with subtype-specific Leiden resolutions. To additionally examine a non-immune subtype setting, we also included fibroblasts from the NSCLC subset of the human perturbed-state fibroblast atlas reported by Buechler et al. [28].

Subtype-level marker genes were than selected from the resulting subtype clusters using the following criteria: adjusted p-value < 0.1, pct.1> 0.25, pct.2< 0.7, and average $\log_2 \mathrm{FC} > 0.3$. These subtype-level annotation analyses, including gene filtering ablation experiment and comparisons with existing LLM-based annotation methods, are presented in Supplementary Tables 12 and 13.

Gene filtering methods

We assessed four distinct methods for filtering identified marker genes.

(i) p-value+pct+$\log_2 \mathrm{FC}$+Database (Basic): Genes were first filtered by p-value, $\mathrm{pct}.1$, $\mathrm{pct}.2$, and $\log_2 \mathrm{FC}$, then cross-referenced using a curated database corresponding to the input tissue type, and finally ranked by descending $\log_2 \mathrm{FC}$.

(ii) Basic+major cell type marker (w/o negative): Genes from the basic filtering method were used as the primary input, with 15 markers for each major cell group included as reference.

(iii) Basic+negative marker (w/o major): Genes from the basic filtering method were used as the primary input, with 15 cluster-specific negative markers included in the prompt to represent low or absent expression.

(iii) Basic+major/negative marker (Full marker set): Genes from the basic filtering method were used as the primary input, with both major cell group's markers and cluster-specific negative markers included in the prompt.

Subtype-level annotation performance was evaluated across both immune and non-immune settings (Fig. 3d). In the immune subtype datasets, performance was broadly similar across the four filtering strategies. In the additional fibroblast subtype setting, the relative performance of the configurations differed, indicating that the contribution of supplemental marker information may vary depending on the subtype context. Notably, in the fibroblast setting, the w/o major configuration yielded the highest average agreement score, further suggesting that the relative contribution of each marker component may differ across subtype settings.

Comparison with existing LLM-based methods

We evaluated subtype-level annotation across closely related immune populations and an additional non-immune fibroblast subtype setting, and compared CELLIA with recent LLM-based annotation frameworks. Across T cell subtypes (CD4T, DNT, and Treg), CELLIA consistently achieved high agreement with manual annotations, while alternative approaches frequently produced partial matches or incorrect assignments (Fig. 3e and Supplementary Table 13). Similarly, for dendritic cell subtypes, CELLIA showed higher agreement and maintained stable performance (Fig. 3e and Supplementary Table 13). Collectively, these results indicate that CELLIA can support fine-grained

subtype annotation not only in closely related immune populations but also in an additional non-immune stromal setting, although performance remained dependent on the biological context and subtype granularity.

Reproducibility of CELLIA annotation

We assessed consistency and reproducibility of LLM-based annotations by repeating CELLIA predictions 10 times across 9 datasets (Supplementary Table 14 and 15). Across tissues, predictions were identical for the vast majority of clusters, with residual variation largely confined to nomenclature (e.g., "NK cell" and "Natural Killer (NK) cell") rather than substantive changes in cell identity. Notably, these subtle discrepancies mirror the inconsistencies inherent in manual annotations, reflecting the broader absence of standardized naming conventions in single-cell studies, where cell type nomenclature is often individual investigator-defined and taxonomic alignment remains limited [30]. Rare variations were instead observed primarily in biologically similar contexts. For example, in PBMC (10X) dataset, the first and second runs yielded "Monocyte" in place of "Classical Monocyte", consistent with the subtype-level (Supplementary Table 14). We also evaluated CELLIA using multiple API models, including GPT-5.2 (Dec 2025), GPT-5.4 (Mar 2026), GPT-5.4 mini (Mar 2026), and GPT-5.4 nano (Mar 2026), and observed broadly similar annotation patterns across models (Supplementary Table 15). This suggests that the overall annotation behavior of CELLIA is preserved not only across repeated inference runs under a fixed setup, but also across different API model settings, with the remaining variation largely limited to semantic synonyms or differences in nomenclature rather than substantive changes in cell identity.

Discussion

Cell type annotation is essential for interpreting single-cell analysis, yet manual annotation remains time-consuming and dependent on expert knowledge. To overcome these limitations, automated and LLM-based approaches have expanded rapidly, interpretation of cluster-specific marker gene often still involves manual choices, including which markers to prioritize and how tissue context should be incorporated. CELLIA was designed in part to make this preparatory step more structured by filtering and prioritizing marker genes using a curated tissue-specific marker database together with tissue context before the evidence is presented to the LLM.

This evidence-knowledge integration strategy forms the basis of CELLIA. The database combines manually curated and automatically extracted tissue-specific marker resources, allowing biologically relevant evidence to be organized in a more structured form prior to annotation. In this sense, the database functions not only as a source of marker information, but also as part of a framework for making the preparation of annotation evidence more consistent.

Benchmarking showed that CELLIA achieved strong overall agreement with manual annotations across datasets, although differences from other recent LLM-based methods were generally modest. Gene filtering analyses provide context for this result. Removing either the knowledge-based or evidence-based component reduced agreement across datasets, with the largest reduction observed when both were removed, suggesting that the two components provide complementary information. Taken together, these findings suggest that the contribution of CELLIA lies in part in how marker evidence is

filtered and organized before LLM-based interpretation. This does not imply that differential expression statistics alone are insufficient, but it does suggest that their interpretation can benefit from additional biological context.

CELLIA is a knowledge-guided annotation framework that prioritizes database-supported markers during cluster interpretation. This design may be particularly useful in well-characterized tissues, where marker knowledge is relatively well established, but may be more limited in out-of-distribution settings, such as rare tissue states, transitional populations, unusual tumor microenvironments, or non-model species, where marker coverage is sparse. In these contexts, biologically meaningful but less-characterized markers may be deprioritized, potentially biasing annotation toward better-represented canonical labels.

Accordingly, CELLIA is intended for knowledge-guided annotation rather than de novo identification of novel cell types. When available evidence is insufficient for confident assignment to a known cell identity, retaining a cluster as unknown is more appropriate than assigning it to the nearest canonical label. Applications in database-poor contexts therefore require more cautious interpretation.

The current workflow assumes pre-clustered input data and does not include the initial clustering step. Because CELLIA derives the DEG list, marker filtering, top-k marker selection, and final prompt from cluster-level definitions, clustering directly shapes the input evidence used for annotation. An additional sensitivity analysis using alternative clustering algorithms and resolutions showed that some datasets remained highly stable across low resolution to moderate resolution settings, whereas others showed greater sensitivity, particularly under stronger cluster splitting. These findings indicate that CELLIA is not invariant to clustering strategy or resolution, even though it can remain stable under some clustering conditions. Annotation results should therefore be interpreted together with the clustering procedure used to define the input clusters. Future work may focus on closer integration with clustering pipelines and continued expansion of curated marker resources.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-026-00593-7.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors thank Seri Oh for her contribution during the early phase of this work, in generating a preliminary marker-gene list using a marker-gene selection tool.

Author contribution

JS (Writing—original draft, Investigation, Data curation, Methodology, Software), SJ (Writing—review, Investigation, Data curation, Validation), HK (Conceptualization, Investigation), and MO (Writing—review & editing, Conceptualization, Methodology, Supervision).

Funding

This work was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number RS-2024-00403375], and by the National Research Foundation of Korea (NRF) grant, funded by the Korea government (MSIT) [grant number RS-2025-24534272].

Data availability

The datasets used during the current study are available from public repositories: PBMC data via 10x Genomics (https://www.10xgenomics.com/datasets/5k_Human_Donor1_PBMC_3p_gem-x), Human Immune Health Atlas (HIHA) website (https://apps.allenimmunology.org/aifi/resources/imm-health-atlas/), a cross-organ fibroblast dataset, accessible at (https://www.fibroxplorer.com), and the Gene Expression Omnibus database under accession numbers

GSE132257(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132257), GSE131907(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907), GSE167036(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167036), GSM4008623(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134355), GSM4008620(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134355), GSM4008710(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134355), and GSM4008637(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134355). All data supporting the findings of this study are available within the paper.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 March 2026 / Accepted: 1 August 2026
Published online: 11 August 2026

References

1. Elsborg J, Salvatore M. Using LLMs and explainable ML to analyze biomarkers at single-cell level for improved understanding of diseases. Biomolecules. 2023;13(10):1516. https://doi.org/10.3390/biom13101516.
2. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746. https://doi.org/10.15252/msb.20188746.
3. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–45. https://doi.org/10.1038/nrg3833.
4. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20(1):194. https://doi.org/10.1186/s13059-019-1795-z.
5. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nat Methods. 2019;16(10):983–86. https://doi.org/10.1038/s41592-019-0535-3.
6. Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. Genes. 2019;10(7):531. https://doi.org/10.3390/genes10070531.
7. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246. https://doi.org/10.1038/s41467-022-28803-w.
8. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc. 2021;16(6):2749–64. https://doi.org/10.1038/s41596-021-00534-0.
9. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62. https://doi.org/10.1038/nmeth.4644.
10. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72. https://doi.org/10.1038/s41590-018-0276-y.
11. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197. https://doi.org/10.1126/science.abl5197.
12. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4(10):852–66. https://doi.org/10.1038/s42256-022-00534-z.
13. Xia Y, Liu Y, Li T, He S, Chang H, Wang Y, et al. Assessing parameter efficient methods for pre-trained language model in annotating scRNA-seq data. Methods. 2024;228:12–21. https://doi.org/10.1016/j.ymeth.2024.05.007.
14. Li R, Zhang J, Li Z. EasyCellType: marker-based cell-type annotation by automatically querying multiple databases. Bioinf Adv. 2023;3(1):vbad029. https://doi.org/10.1093/bioadv/vbad029.
15. Lin A, Ye J, Qi C, Zhu L, Mou W, Gan W, et al. Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. Briefings Bioinf. 2025;26(4):bbaf357. https://doi.org/10.1093/bib/bbaf357.
16. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat Methods. 2024;21(8):1462–65. https://doi.org/10.1038/s41592-024-02235-4.
17. Ye W, Ma Y, Xiang J, Liang H, Luo J, Li Y, et al. Evaluation of cell type annotation reliability using a large language model-based identifier. Commun Biol. 2025;8(1):1360. https://doi.org/10.1038/s42003-025-08745-x.
18. Yang C, Zhang X, Chen J. Large language Model consensus substantially improves the cell type annotation accuracy for scRNA-seq data. bioRxiv. 2025;2025–04. https://doi.org/10.1101/2025.04.10.647852.
19. Xie E, Cheng L, Shireman J, Cai Y, Liu J, Mohanty C, et al. CASSIA: a multi-agent large language model for automated and interpretable cell annotation. Nat Commun. 2025;17(1). https://doi.org/10.1038/s41467-025-67084-x.
20. Crowley G, Quake SR, Krasnow M, Pisco AO, Salzman J, Yosef N, et al. Benchmarking cell type and gene set annotation by large language models with AnnDictionary. Nat Commun. 2025;16(1):9511. https://doi.org/10.1038/s41467-025-64511-x.
21. Meng FL, Huang XL, Qin WY, Liu KB, Wang Y, Li M, et al. singleCellbase: a high-quality manually curated database of cell markers for single cell annotation across multiple species. Biomark Res. 2023;11(1):83. https://doi.org/10.1186/s40364-023-00523-3.
22. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51(D1):D870–76. https://doi.org/10.1093/nar/gkac947.

23. Cheng P, Peng Y, Zhang XL, Chen S, Fang BB, Li YZ, et al. A natural language processing system for the efficient extraction of cell markers. Sci Rep. 2024;14(1):21183. https://doi.org/10.1038/s41598-024-72204-6.
24. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. Anndata: access and store annotated datamatrices. J Open Source Softw. 2024;9(101):4371. https://doi.org/10.21105/joss.04371.
25. Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. Genome Biol. 2024;25(1):56. https://doi.org/10.1186/s13059-024-03183-0.
26. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0.
27. Gong Q, Sharma M, Glass MC, Kuan EL, Chander A, Singh M, et al. Multi-omic profiling reveals age-related immune dynamics in healthy adults. Nature. 2025;648(8094):696–706. https://doi.org/10.1038/s41586-025-09686-5.
28. Buechler MB, Pradhan RN, Krishnamurty AT, Cox C, Calviello AK, Wang AW, et al. Cross-tissue organization of the fibroblast lineage. Nature. 2021;593(7860):575–79. https://doi.org/10.1038/s41586-021-03549-5.
29. Cheng C, Chen W, Jin H, Chen X. A review of single-cell rna-seq annotation, integration, and cell–cell communication. Cells. 2023;12(15):1970. https://doi.org/10.3390/cells12151970.
30. Miller JA, Gouwens NW, Tasic B, Collman F, van Velthoven CT, Bakken TE, et al. Common cell type nomenclature for the mammalian brain. Elife. 2020;9:e59928. https://doi.org/10.7554/eLife.59928.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.