Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

scMarkerAgent: An LLM Evidence Agent-based Cell Marker Atlas

Chen Cao

caochen@njmu.edu.cn

Nanjing Medical University

HaoRan Chen

Nanjing Medical University https://orcid.org/0000-0002-1666-3226

Yanqun Sun

Nanjing Medical University

Hangchen Zhang

Nanjing Medical University

Yin Liu

Nanjing Medical University

Hanzi Zhao

Nanjing Medical University

Xu Wang

Nanjing Medical University

Guishen Wang

Changchun University of Technology

Gaoyang Li

Nanjing Medical University

Article

Keywords:

Posted Date: April 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9224570/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

scMarkerAgent: An LLM Evidence Agent-based Cell Marker Atlas

Haoran Chen[1,†], Yanqun Sun[2,†], Hangchen Zhang[1,†],
Yin Liu[1], Hanzi Zhao[1], Xu Wang[2,*],
Guishen Wang[3,*], Gaoyang Li[4,5,*], Chen Cao[1,*]

[1]Key Laboratory for Bio-Electromagnetic Environment and Advanced Medical Theranostics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, Jiangsu, China.
[2]Clinical Medical Research Center, Children's Hospital of Nanjing Medical University, Nanjing 210008, Jiangsu, China.
[3]College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, Jilin, China.
[4]The Second People's Hospital of Changzhou, The Third Affiliated Hospital of Nanjing Medical University, Changzhou 213164, Jiangsu, China.
[5]College of Life Sciences, Nanjing Medical University, Nanjing 211166, Jiangsu, China.

[*]Correspondence: sepnine@njmu.edu.cn; wangguishen@ccut.edu.cn; lgyzngc@njmu.edu.cn; caochen@njmu.edu.cn
[†]These authors contributed equally to this work.

Abstract

Evidence-augmented and reliable cell-type annotation remains a major bottleneck in single-cell RNA-seq analysis, particularly for rare, transitional, and disease-associated populations. To address this, we introduce scMarkerAgent, an evidence-grounded cell marker resource developed using an LLM-assisted literature-curation framework. It integrates 294,692 full-text publications to provide 890,296 high-quality cell type-marker annotations covering 50,233 cell types across human, mouse and rat. scMarkerAgent integrates 82,165 curated negative-marker annotations and 417,812 disease-context annotations, improving disambiguation of homologous cell types and delineation of malignant cells. Every cell type-marker annotation is directly supported by sentence-level literature evidence. In the cell annotation workflow, candidate labels are further refined through an LLM-based reasoning step that jointly evaluates positive and negative markers. Compared with existing resources, scMarkerAgent offers broader coverage of markers, tissues, cell types and diseases. It is released as a FAIR-compliant database together with a code-free web platform that sup-

ports marker retrieval, automated cell annotation and customizable cell scoring (available at `https://www.markeragent.net`).

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the study of cellular heterogeneity in health and disease. A critical step in scRNA-seq analysis is cell type annotation, which fundamentally shapes downstream interpretation of lineage composition, cell states and disease-associated populations. Nevertheless, accurate and reliable annotation remains challenging, particularly in the presence of rare subtypes, transitional states or disease-specific transcriptional programs.

In practice, annotation is supported by multiple computational strategies, including reference mapping, marker-based approaches and, more recently, large language model (LLM)-assisted workflows. Among these, marker-based cell-type annotation remains widely adopted due to its independence from a matched reference and its direct biological interpretability. However, its accuracy depends critically on the completeness, quality and usability of marker knowledge, and the process often still requires expert validation.

To support marker-based annotation, several curated marker resources and tools have been developed, such as Cell Marker Accordion (CMA), CellMarker 2.0 and Annotation of Cell Types (ACT) [1, 2, 3]. Despite improving accessibility, these resources face notable practical limitations. First, manual curation struggles to keep pace with the rapidly expanding literature, resulting in slow updates and incomplete coverage of newly described subtypes, particularly within disease contexts. Second, reproducibility is often hindered because many resources report cell type-marker associations and their source publications, yet frequently omit the exact sentences from the original studies that support these claims. The absence of direct textual evidence makes it challenging to verify key contextual details, such as tissue specificity, disease context or experimental setting, and complicates the resolution of conflicting annotations across studies [4]. Finally, most existing resources focus primarily on positive markers and offer limited support for negative markers, even though markers expected to be absent can improve annotation specificity, especially when closely related cell populations share overlapping positive signatures [5].

LLM-based approaches have been explored to reduce the effort and specialized expertise required for cell type annotation [6]. In particular, AI agent-based methods leverage the strengths of both specialized annotation

algorithms and LLM-enhanced knowledge retrieval, offering an interpretable, verifiable and adaptable workflow [7]. However, relying on an LLM as the primary source of biological knowledge raises concerns regarding transparency and traceability of the underlying training data, susceptibility to hallucinations and biases, and non-deterministic outputs, all of which complicate evaluation and reproducibility [8]. These limitations motivate an alternative paradigm: using LLMs as agents that extract and reason over verifiable evidence from the literature, rather than as opaque sources of knowledge without provenance [9].

Here, we present scMarkerAgent, an evidence-guided cell marker resource that leverages LLM agents to curate peer-reviewed literature, enabling marker retrieval, cell-type annotation and customized cell scoring. Under rigorous quality control, scMarkerAgent links each cell type-marker annotation to sentence-level supporting evidence, allowing users to audit annotations in their original biological context. This framework not only integrates positive and negative markers within a unified structure, but also maps tissue, cell-type and disease terms to established ontologies. Furthermore, scMarkerAgent implements an evidence-grounded LLM workflow, where the model's role is constrained to providing decision support for marker-driven candidate annotations. Across diverse scRNA-seq benchmarks, we demonstrate that scMarkerAgent improves the accuracy, transparency and reproducibility of marker-based cell type annotation compared with widely used marker resources, in both normal and disease contexts. Finally, we provide a user-friendly, code-free web platform that offers rapid access to these resources and supports flexible, personalized analyses.

Results

Large-scale literature retrieval-based cell marker annotation using a multi-agent framework

We collected 294,692 full-text publications and employed an Extract Agent to retrieve 1,352,754 candidate cell type-marker records. Each record includes structured fields covering tissue, cell type, disease, species, marker and marker polarity, and is linked to sentence-level literature evidence to ensure end-to-end traceability. An independent quality-control (QC) Agent was then applied to validate each candidate record against six mandatory rules. Following QC filtering and restricting the dataset to three model species (human, mouse and rat), we retained 890,296 high-confidence records for downstream analyses (Fig. 1A).

To reduce errors introduced by heterogeneous shorthand in the literature, we incorporated an Abbr2full Agent that expands tissue and cell-type abbreviations using publication-specific context (for example, resolving "LSCs" to leukemia stem cells rather than liver stem cells when warranted), thereby improving name consistency prior to downstream normalization (Fig. 1A).

Given the heterogeneity of nomenclature across studies, which can impede cross-dataset integration, we performed candidate mapping followed by the application of a Standardization Agent. This agent normalizes raw names into standardized terminology while retaining non-standard names when no exact match is available (Fig. 1A). Tissues were standardized according to the Uberon Ontology, cell types were mapped to the Cell Ontology, and diseases were aligned with the Disease Ontology [10, 11, 12]. Marker names were standardized via NCBI resources [13] and linked to identifiers such as Entrez and Ensembl where applicable, improving annotation consistency across studies.

In the current release, scMarkerAgent spans three species and provides broad coverage across normal and disease contexts, including 29,277 normal-context cell annotations, 29,552 disease-context cell annotations, 313 normal tissues, 292 diseased tissues, 34,276 cell markers and 15,386 diseases (Fig. 1B).

scMarkerAgent: a FAIR-compliant, literature-curated cell marker repository

Based on the agent-curated data described above, we developed the scMarkerAgent web platform for cell-type annotation and resource search (available at `https://www.markeragent.net`). This platform was designed according to the FAIR (Findable, Accessible, Interoperable and Reusable) principles [14], with all annotations derived from systematically curated, evidence-verified literature; each entry is fully traceable to its original source, ensuring transparency and reliability.

We designed scMarkerAgent as an interactive, code-free platform that supports three tightly integrated workflows: marker search and browsing across species, tissues, cell types, diseases and marker polarity (Supplementary Fig. 1); evidence-based cell type annotation for user-uploaded datasets (Fig. 2A,B); and customizable cell scoring for user-specified cell types with options for positive or negative markers (Fig. 2C,D).

For evidence-based cell-type annotation, the Cell Annotation module allows users to upload Seurat objects (Fig. 2A). Candidate cell-type labels are initially generated with ScType [15], after which an LLM-based annotation

agent performs evidence-grounded selection and produces an interpretable report (Fig. 2B). During this process, positive markers provide supporting signals while negative markers act as exclusion constraints to flag biologically implausible candidates when marker conflicts are present, thereby improving disambiguation of closely related or transitional cell populations.

The Cell Score module enables users to perform custom scoring for cell types of interest within the database (Fig. 2C). It leverages positive-marker signatures to help identify target cell populations and, when necessary, negative-marker signatures to help exclude confounding populations. The scoring results are presented in plots and tables (Fig. 2D).

Importantly, the entire workflow can be completed online without requiring specialized computational skills, making scMarkerAgent accessible to users from diverse backgrounds. Interactive reports are automatically generated, facilitating rapid validation and iterative refinement of annotation results within a straightforward, user-friendly environment.

Benchmarking evaluation of scMarkerAgent

To evaluate the performance of scMarkerAgent, we systematically compared it with three published high-quality reference atlases, Cell Marker Accordion (CMA), CellMarker 2.0 and Annotation of Cell Types (ACT), across 15 different dimensions (see Methods for details) [1, 2, 3]. First, while CMA was constructed by integrating 23 existing cell marker databases and cell-sorting resources, scMarkerAgent contains approximately 3.66 times more total annotations than CMA and substantially more than the manually curated CellMarker 2.0 and ACT (Fig. 3A). Second, scMarkerAgent covers the largest number of distinct markers among the four databases (Fig. 3B). Although CellMarker 2.0 and ACT also curated cell markers from the literature and performed final integration, each incorporated fewer than 10,000 publications. By contrast, scMarkerAgent includes approximately 17.74 times more publications than CellMarker 2.0 and 34.89 times more than ACT. Because CMA does not provide literature-source annotations, it is shown as zero (Fig. 3C).

Moreover, scMarkerAgent contains substantially higher numbers of tissues, cell types and diseases than the three other databases (Fig. 3D–F). Given potential naming inconsistencies for the same biological entities across studies, we standardized tissue, cell and disease names using the Uberon Ontology, Cell Ontology and Disease Ontology, respectively, while retaining and merging non-standardized names (see Methods for details). Notably, both the total and standardized counts of tissues, cell types and diseases are higher

in scMarkerAgent, thereby broadening the reference coverage available for cell annotation and downstream analyses (Fig. 3G–I).

In terms of species coverage, scMarkerAgent provides more annotations for human, mouse and rat than the other databases, and uniquely includes rat annotations, a model species not covered by the other resources (Fig. 3J–L). In addition, scMarkerAgent contains the largest number of annotations derived from diseased samples, which is 7.94-fold greater than the second-ranked CMA, providing a larger disease-related reference resource for cell-type annotation in pathological contexts (Fig. 3M). scMarkerAgent also contains the largest number of positive marker annotations (Fig. 3N) and the largest number of negative marker annotations (Fig. 3O), which may be useful for distinguishing highly similar cell types.

scMarkerAgent provides more accurate cell type-marker annotations

To assess the accuracy and completeness of scMarkerAgent-derived annotations, we benchmarked marker-database-based workflows alongside a database-free, LLM-only annotation workflow across a diverse collection of scRNA-seq datasets spanning different tissues, species and disease contexts. To enable a controlled comparison across marker resources, we used ScType as the unified annotation algorithm for the marker-database-based workflows. ScType supports user-defined marker inputs and incorporates both positive and negative markers via specificity-based weighting across clusters and candidate cell types [15]. Specifically, we compared ScType using markers from ACT, CellMarker 2.0, CMA and scMarkerAgent. We additionally evaluated a Cell Annotation Agent workflow (ScType+scMarkerAgent+LLM), a two-stage, cluster-level framework in which ScType first ranks candidate cell types for each cluster using species-, tissue- and disease-matched markers from scMarkerAgent, after which an LLM then selects the final identity by interpreting the candidate list along with the overlapping positive and negative markers and their supporting evidence within the relevant biological context (see Supplementary Note 1 for details). In parallel, where evaluated, we benchmarked CASSIA as a database-free, LLM-driven annotation workflow for comparison [7].

We assembled benchmark datasets comprising normal human peripheral blood mononuclear cells (PBMC), normal human lung, normal human and mouse pancreas, mouse type 2 diabetes (T2D) pancreas, rat lung and human glioblastoma [16, 17, 18, 19, 20, 21]. All datasets were processed using the Seurat v5.3.0 pipeline [22]. To ensure comparability across studies, we

evaluated annotations at the cluster level and defined the ground-truth annotation for each cluster as the most abundant author-validated or expert-consensus cell type within that cluster. Following a previously described scoring scheme, each predicted annotation was assigned a score of 1, 0.5 or 0, corresponding to a full match, partial match or mismatch to the ground truth, respectively [6, 7].

As an example, we evaluated these workflows on a published human lung single-cell atlas (584,944 cells), which provides expert consensus annotations for 24 cell clusters that served as ground truth [23, 17]. All methods achieved high-level concordance with the expected broad lineages (epithelial, immune, stromal and others), but we observed clear differences in fine subtype resolution and error rates (Fig. 4A–F, Supplementary Fig. 2 and Supplementary Tables S1 and S2).

Benchmark results demonstrated that ScType powered by scMarkerAgent achieved consistently strong annotation accuracy relative to the public marker databases across the benchmark panel, and this advantage was further enhanced by an LLM-guided refinement grounded in scMarkerAgent-derived marker evidence. In the human lung dataset (24 clusters), ScType+scMarkerAgent correctly annotated the vast majority of cell clusters (21 full matches, 2 partial matches and 1 mismatch), outperforming all other reference databases. This included the correct identification of specialized cell types that were missed or misannotated by other methods. For instance, a proliferating alveolar macrophage population was recognized by scMarkerAgent, whereas some baseline methods erroneously annotated this cluster as proliferating T cells or undefined cycling cells. Similarly, a mast cell cluster was accurately annotated by scMarkerAgent, while two public databases (CellMarker 2.0 and CMA) misclassified these cells as basophils.

The subsequent LLM-based refinement resolved several of the remaining errors from the initial ScType+scMarkerAgent predictions. Notably, one cluster initially identified as alveolar type I (AT1) cells by ScType+scMarkerAgent was reclassified by the LLM as alveolar intermediate (ATi) cells, consistent with the co-expression of both AT1 and AT2 marker genes in that cluster. More broadly, the LLM's inferential refinements were grounded in explicit sentence-level evidence retrieved by scMarkerAgent, and representative supporting references and statements are provided in Table 1. In another case, the LLM corrected a ScType+scMarkerAgent misassignment of cycling basal epithelial cells (originally misannotated as macrophages) to their proper identity as respiratory basal cells.

Across the other benchmark datasets (human PBMC, human and mouse pancreas, mouse T2D pancreas, rat lung and human glioblastoma; Sup-

plementary Figs. 3–8), ScType+scMarkerAgent again achieved the highest overall concordance with the reference annotations (Fig. 4G): in the human PBMC and pancreas datasets, at least 75% of clusters were full matches to the ground truth across scMarkerAgent, with the remaining clusters receiving partial-match or mismatch scores, and performance was broadly consistent across species. For example, in the mouse pancreas dataset, ScType+scMarkerAgent yielded full matches for 8 of 10 clusters (80%) without any mismatches. By contrast, in the glioblastoma dataset all methods showed reduced accuracy relative to normal tissues, with ScType+scMarkerAgent still yielding the largest number of expert-consistent annotations but only approximately 46% of clusters (16/35) full matches and approximately 26% mismatches, which may be related to increased heterogeneity and state plasticity in malignant tissue. Notably, and contrary to the common expectation that aggregating more sources necessarily improves performance, CMA, despite integrating 23 marker resources including CellMarker 2.0, achieved lower annotation scores than CellMarker 2.0 in both the mouse pancreas and human glioblastoma benchmarks, suggesting that this aggregation may have introduced lower-quality or inconsistent marker annotations.

In addition, to benchmark against a database-free, LLM-driven annotation paradigm, we included CASSIA (accessed at `https://www.cassia.bio/` on 17 December 2025), a recently proposed multi-agent LLM framework for cell-type annotation [7]. Overall, across the datasets where both methods were evaluated, CASSIA exhibited cluster-level annotation performance broadly comparable to that of ScType+scMarkerAgent, but it did not surpass ScType+scMarkerAgent+LLM (Fig. 4G and Supplementary Table S2). Relative to ScType workflows that rely on public marker databases, CASSIA's clearest advantage was observed in glioblastoma. Unlike the ScType-based workflows, CASSIA does not use an explicit marker reference database; instead, it relies on the LLM's parametric knowledge. In practice, its errors more often reflected lineage-level misassignments or state-first outputs, and its free-text annotations were less standardized across clusters. For example, in the human lung dataset, CASSIA alternated between multiple descriptions for macrophage populations rather than using consistent labels such as "alveolar macrophage", and for a proliferative macrophage cluster it returned the state-only label "Cycling / mitotic cells (G2/M-phase enriched)" instead of the lineage-aware annotation "cycling alveolar macrophage". In contrast, ScType+scMarkerAgent+LLM produced more standardized, lineage-aware labels such as "alveolar macrophage" and "cycling alveolar macrophage" (Supplementary Table S2).

Negative markers improve cell-type disambiguation and malignancy delineation

To further understand why the cluster originally annotated as AT1 was corrected to ATi cells by the ScType+scMarkerAgent+LLM framework, we examined the LLM's reasoning process (see Supplementary Table S1):

> The cluster co-expresses AT1 markers (AGER) with transitional or AT2 features (KRT8, SFTPB), which is characteristic of ATi cells. Top-ranked mature AT1 was rejected because its negative marker KRT8 overlaps the cluster, indicating biological conflict; iAT1 YAP5SA was rejected due to overlapping negatives SFTPB and NAPSA. Endothelial stalk cell is unlikely given epithelial lineage markers (for example, NKX2-1 and EPCAM) and absence of endothelial hallmarks. Alveolar epithelial progenitor lacks the key transitional markers (KRT8 and SFTPB) seen here. Induced AT1 and broad pulmonary alveolar epithelial labels are less specific than ATi for this marker pattern. Mesothelial and goblet cell candidates do not match the alveolar epithelial signature.

Aligned with the LLM's reasoning, we visualized the expression of AGER, KRT8 and SFTPB in the human lung dataset (Fig. 5A–C). The AT1 marker AGER showed strong and specific upregulation in the AT1 cluster ($\log_2$FC = 8.67, PCT = 0.955; Fig. 5A), which likely explains why other methods annotated this cluster as pulmonary alveolar type I (AT1) cells. However, those methods did not incorporate a key biological rule: AT1 cells should not express AT2-associated markers such as KRT8 and SFTPB. As shown in Fig. 5B,C, both KRT8 and SFTPB were also highly expressed in this cluster. Notably, KRT8 is listed as a negative marker for AT1 cells in scMarkerAgent; therefore, incorporating negative-marker constraints enabled the framework to resolve the marker conflict and yield a more accurate annotation as ATi cells.

Beyond their utility as exclusion constraints for cell-type disambiguation, negative markers are also valuable for malignant-cell identification in tumor scRNA-seq, helping to reduce potential misclassification when positive markers are used alone. Accurate identification of malignant cells in tumor scRNA-seq is a critical prerequisite for disentangling tumor-intrinsic transcriptional programs from signals originating from non-malignant cells in the tumor microenvironment. This enables reliable characterization of intratumoral heterogeneity and tumor cell states associated with progression and

therapy resistance, thereby informing mechanism-based therapeutic strategies [24].

scMarkerAgent contains 417,812 curated cell type-marker annotations in disease-associated contexts. Using glioblastoma as an example, we retrieved glioblastoma-specific positive and negative marker sets from scMarkerAgent and used UCell to compute per-cell signature scores [25]. Positive-marker scores (Fig. 5D) were generally higher in malignant cells and broadly aligned with the malignancy labels originally assigned by the authors using SCEVAN iCNV analysis (Fig. 5F), whereas negative-marker scores showed the opposite trend (Fig. 5E). Together, these complementary signals indicate that negative markers can serve as exclusion constraints and improve the robustness of malignant-cell identification beyond the use of positive markers alone (Fig. 5G).

Taken together, these results support the breadth and reliability of scMarkerAgent annotations and underscore that jointly incorporating negative and positive markers can improve cell-type disambiguation and malignant-cell delineation in scRNA-seq analyses.

scMarkerAgent+LLM improves annotation performance and standardization in brain spatial transcriptomics

Using the spatial transcriptomics dataset from Sun et al. [26], we evaluated scMarkerAgent+LLM on a Seurat object containing 2,327,742 spatial profiles across 32 sections, and all results together with supporting evidence can be explored interactively at `https://spatial.markeragent.net` (Fig. 6A). On this challenging brain spatial dataset, scMarkerAgent achieved a mean annotation score of 0.63, outperforming conventional marker-database workflows by more than 10%, and LLM-based refinement further increased performance to 0.87 (Fig. 6B). Together with the observation that the purely LLM-driven CASSIA workflow also ranked ahead of conventional marker-database methods, this pattern suggests that reasoning-based annotation is particularly helpful in heterogeneous, state-rich settings, where marker-only workflows may be less able to resolve fine-grained cell identities. A purely LLM-driven approach (CASSIA) ranked second, underscoring the importance of reasoning during annotation; however, its outputs were comparatively non-standardized (for example, cluster 4 is returned as a long, region- and marker-qualified narrative label; Supplementary Table S3), whereas scMarkerAgent provides a controlled vocabulary that supports consistent labeling and cross-cluster comparison (Fig. 6B). Overall, we observed that scMarkerAgent+LLM increased subtype and state resolution relative to the au-

thors' largely marker-based cluster labels while remaining anchored to differentially expressed markers and explicit marker-consistency checks (Fig. 6C).

In the oligodendrocyte lineage, the authors' broad oligodendrocyte labels were refined where mature myelin programs were unambiguous: for example, clusters 0 and 22 were upgraded from an oligodendrocyte category to mature myelinating oligodendrocytes, supported by canonical mature myelin and compaction markers (Mog, Plp1, Mag, Cldn11 and Mal in cluster 22), whereas cluster 21 was intentionally retained as a more general oligodendrocyte because the marker evidence did not justify a fully terminal-state label.

In neuronal compartments, where the authors frequently used marker-centered names (for example, Neuron_) that do not explicitly encode lineage or functional class, our framework converted marker combinations into standardized, interpretable identities consistent with established neurobiology. These included an L6 corticothalamic-projecting glutamatergic neuron for cluster 6 (driven by Rprm), PV interneuron specializations such as a chandelier cell for cluster 7 (Vipr2, Pvalb) and a Pvalb GABAergic cortical interneuron for cluster 14 (Pvalb, Lhx6, Sox6, Tac1), as well as somatostatin-expressing interneurons for clusters 12 and 28 (for example, an Sst program with Sox6/Npy/Crhbp in cluster 12, and Chodl/Crhbp in cluster 28) and a VIP interneuron identity for cluster 16 (Htr3a, Vip). Importantly, these refinements reflect lineage-consistent combinations rather than single-marker naming and therefore improve cross-cluster comparability (Fig. 6D,E).

scMarkerAgent+LLM corrected lineage assignments in non-neuronal clusters when the authors' labels conflicted with canonical marker modules. This was most evident in cluster 24, where a choroid plexus epithelial identity (Ttr, Folr1, Krt18, Cldn2, Igfbp2) provided a more coherent epithelial and secretory program than an ependymal-type label, and in cluster 8, where a mural or pericyte program (Acta2, Myl9, Cspg4, Rgs5) supported vascular mural identity rather than a broader leptomeningeal-associated designation.

Together, these results show that scMarkerAgent+LLM improves both accuracy and subtype or state resolution in brain spatial transcriptomics. Moreover, unlike purely LLM-driven annotation, it maintains standardized, comparable cell-type names by grounding decisions in differential markers and explicit consistency checks.

Discussion

In summary, the major bottleneck in marker-based annotation is not merely a scarcity of markers, but rather the absence of auditable, context-aware evidence and principled conflict handling [27]. By linking each record to sentence-level evidence and applying QC filtering, scMarkerAgent converts literature-derived cell type-marker statements into traceable, structured annotations. Building on this framework, ScType-based pre-filtering, negative-marker exclusion and LLM reasoning grounded in scMarkerAgent-derived marker evidence collectively improve the accuracy and interpretability of cell-type annotation in scRNA-seq benchmarks and spatial transcriptomics.

The value of sentence-level evidence extends beyond finer-grained citations. It enables direct verification of marker knowledge and supports interpretation of why annotation disagreements arise. When conflicts arise over a cell type-marker annotation, users can trace back to the original evidence sentences to examine factors such as tissue, disease background and experimental context, thereby discerning whether disagreements stem from contextual variation, naming inconsistencies or insufficient support. Importantly, this evidence-centric marker paradigm differs from reference-based, training-driven approaches. Most atlas mapping or supervised classifiers rely on labeled training data, are sensitive to the representativeness of the training reference, and typically require retraining or adaptation when novel cell types or disease states emerge [28]. In contrast, scMarkerAgent can be expanded by adding new cell type-marker annotations along with their supporting sentences, without the need for model pretraining, enabling faster incorporation of newly characterized populations while maintaining interpretability and traceable provenance.

Our analyses suggest that fine-grained misannotations often arise not from missing strong positive markers, but from the absence of negative-marker evidence capable of explicitly excluding biologically implausible candidate annotations [15, 29]. In the AT1-to-ATi disambiguation example, strong positives such as AGER can dominate decisions in fixed-weight scoring schemes and bias the label toward AT1. However, the concurrent expression of KRT8 and SFTPB argues against a canonical AT1 program: SFTPB provides negative evidence against AT1 by reflecting retained AT2-associated features, and KRT8 provides positive support for an intermediate, ATi-like identity consistent with a mixed or transitional transcriptional program, in line with prior studies [30, 31, 32]. By systematically curating negative markers, scMarkerAgent renders such incompatibilities explicit and shifts the decision logic from “which candidate scores higher” to “which

candidates are inconsistent with the observed marker pattern". In the scMarkerAgent+LLM workflow, the LLM does not merely follow ScType's fixed weights; instead, it performs evidence-guided reasoning based on the retrieved sentences, evaluates lineage specificity, and accounts for both positive support and negative-marker conflicts. More broadly, this conflict-aware strategy reduces over-reliance on a few dominant positives and improves robustness and interpretability in fine-grained settings where closely related populations share markers or transitional states mix transcriptional programs.

In disease settings such as tumors and diabetic pancreas, we observed that annotation accuracy can decline across methods, consistent with increased transcriptional plasticity, the intermixing of lineage and state programs, and entangled disease-microenvironment signals [33]. scMarkerAgent addresses these challenges through complementary data and algorithmic innovations. On the data side, it incorporates substantially more disease-context marker annotations, aligning candidate identities and evidence with the non-canonical expression combinations commonly observed in diseased samples. On the algorithmic side, we jointly leverage positive and negative markers while implementing evidence-constrained candidate filtering and conflict resolution, such that malignant-cell delineation is guided by both capturing malignant programs and excluding microenvironmental cells. This approach reduces ambiguity and misclassification that are more likely to occur when relying solely on positive markers [34].

Based on our comparisons, LLMs appear to be useful for reasoning over complex marker combinations and for producing standardized outputs, whereas their main risks stem from unverifiable knowledge sources and output instability. Foundation models typically rely on large-scale pretraining and downstream fine-tuning, involve substantial computational cost, and their performance is tightly coupled to the coverage of training data [35]. In single-cell applications, transformer-based approaches can treat genes as tokens to learn expression patterns and achieve a degree of transfer across tissues and conditions, motivating a growing ecosystem of LLM and foundation-model tools [7, 36]. However, purely LLM-driven or purely generative workflows can still face practical limitations. First, predictions often rely on parametric knowledge without an auditable evidence chain, making it difficult to justify why a label was chosen [37, 38]. Second, outputs may drift into non-standard free-text naming, reducing consistency across clusters and datasets [39, 40]. Third, when new cell types or disease subtypes are not represented in training corpora, models may produce plausible-sounding but weakly supported inferences [27, 41]. We therefore invoke the LLM only

after evidence retrieval and candidate constraint. The model then resolves conflicts and assigns names only using scMarkerAgent-supported marker evidence and a predefined candidate set, preserving the benefits of LLM reasoning while reducing non-traceable or inconsistent outputs in both scRNA-seq and spatial transcriptomics.

Despite these improvements, scMarkerAgent remains subject to literature and coverage bias. Common tissues and cell types are associated with more annotations, whereas uncommon tissues or newly reported findings may lack sufficient support. Moreover, as scRNA-seq advances, outdated Cell Ontology terms may not keep pace with the emergence of new cell types. Improving the update speed of cell ontologies would benefit cross-study analyses and cell atlas construction [42]. In addition, this study evaluated only the application of ScType within the scMarkerAgent database, and other marker-based annotation methods were not tested. Further algorithmic optimization may improve annotation accuracy.

Future work will extend scMarkerAgent in two areas. First, beyond the current coverage of human, mouse and rat, we will incorporate marker annotations for additional species to broaden its applicability to studies in diverse organisms. Second, we plan to establish an automated, versioned release mechanism with quarterly updates to continuously integrate newly published evidence, enabling timely inclusion of newly reported cell types and markers.

In conclusion, scMarkerAgent provides a sentence-level, evidence-based marker resource that improves the transparency and interpretability of marker-driven cell type annotation. By integrating positive and negative markers with candidate-constrained LLM reasoning, it enables more reliable fine-grained annotation across multi-species single-cell datasets, including in disease contexts, and in spatial transcriptomics.

Methods

Data collection

We used the following search terms: ("single cell" OR "single-cell" OR "single cell RNA-seq" OR "scRNA-seq" OR "single-cell RNA" OR "scRNA sequencing" OR "single cell sequencing" OR "single cell RNA sequencing" OR "single cell RNAseq" OR "scRNAseq" OR "scRNA seq" OR "cell marker" OR "cell markers" OR "cellular marker" OR "cellular markers" OR "surface marker" OR "surface markers" OR "specific marker" OR "cell signature" OR "cell signatures" OR "cellular signature" OR "cellular signatures" OR "surface sig-

nature" OR "surface signatures" OR "cell specific signature" OR "cell specific signatures" OR "identify cell marker" OR "identify cell markers" OR "identify surface marker" OR "identify surface markers" OR "identify cell specific marker" OR "identify cell specific markers" OR "identify cell signature" OR "identify cell signatures" OR "identify cellular signature" OR "identify cellular signatures" OR "identify surface signature" OR "identify surface signatures" OR "identify cell specific signature" OR "identify cell specific signatures" OR "classify cell type"). A total of 404,362 annotations related to cell markers were retrieved from the PubMed Central database. After excluding publications unrelated to cell markers, 294,692 relevant articles were included (`https://ftp.ncbi.nlm.nih.gov/pub/pmc/`, accessed 22 September 2025).

ChatGPT-5-based framework for cell marker annotation

To convert publications containing cell marker information into a standardized database format, we developed a multi-agent framework based on ChatGPT-5 (version gpt-5-2025-08-07). We also built a robust, high-throughput parallel pipeline to extract key information related to cell type annotation from large-scale full-text databases. The entire annotation workflow follows an evidence-based principle, and all annotations are derived directly from traceable textual evidence in the original publications (Fig. 1A).

The core component of this framework is the Extract Agent, which runs within a high-throughput parallel worker pipeline to extract structured annotations from large-scale full-text corpora in an evidence-based manner. For each paper, the agent first aggregates the full text by section and performs section filtering to prioritize marker-relevant content while excluding sections that are less likely to contain marker-based annotation evidence. It is instructed to extract species, tissue class, tissue type, disease type, cell type, marker and marker polarity, while restricting markers to single, specific gene symbols or protein names that are explicitly used for cell-type identification, definition, classification or distinction rather than expression-only descriptions or functional studies, and it explicitly captures negative markers when absence or non-expression is used to distinguish identities. To ensure traceability, every extracted record is accompanied by minimal verbatim sentence-level evidence from the original text, together with the originating section and publication metadata. Results from all processed sections are then consolidated and deduplicated while preserving their associated section information and evidence text for downstream processing (see Supplementary Note 2 for details).

The QC Agent performs evidence-based, independent and stringent qual-

ity control for each annotation record by validating the cell type, marker and marker polarity fields. Each record must satisfy six mandatory rules that collectively enforce sentence-level evidence, direct association, non-speculative wording, single-symbol markers, polarity consistency and explicit use for cell-type identification. (1) The marker must be a single, specific gene symbol or protein name, not a gene family, descriptive term or compound designation. (2) The marker must be explicitly used in the source text for cell-type identification, classification, definition or distinction, rather than only for expression analysis or functional studies. (3) The source evidence must consist of complete sentence(s) that directly support the cell type-marker annotation claim. (4) The evidence must clearly demonstrate the association between the marker and the cell type. (5) Marker polarity must be determined based on identification usage: a positive marker means presence or expression defines identity, whereas a negative marker means absence or non-expression is used to distinguish identity. (6) The source text must not be speculative, conditional or uncertain about the cell type-marker annotation. During validation, the QC Agent evaluates each record independently. Records that can satisfy all rules with limited field correction are automatically corrected and marked as passed; records that cannot satisfy the rules are marked as failed with a specific failure reason for human review. Only annotations that pass quality control are retained for database construction (see Supplementary Note 3 for details).

During the annotation process, we observed that abbreviations for tissue and cell types are frequently used in papers, which may lead to misinterpretation. For instance, LSCs could refer to either leukemia stem cells or liver stem cells, potentially resulting in erroneous cell type-marker annotations. To address this issue, we developed the Abbr2full Agent, which expands tissue and cell-type abbreviations using publication-specific context. This module first classifies each term in the three fields of tissue class, tissue type and cell type as either an abbreviation or a full name, then retrieves contextual sentences containing the abbreviation from the full text of the corresponding paper for terms identified as abbreviations, and subsequently submits both the abbreviation and its contexts to the LLM for expansion, enabling conversion to the full name based on the paper context (see Supplementary Note 4 for details).

We used a Standardization Agent to standardize raw names into controlled-vocabulary terms. Specifically, we selected the Cell Ontology to standardize cell types (accessed on 18 October 2025), applied the Uberon Ontology for tissue standardization (accessed on 1 October 2025), and used the Disease Ontology for disease standardization (accessed on 13 October 2025) [11, 10,

12]. Further details are provided in Supplementary Note 5.

Classification and standardization of marker names

In this study, we designed a standardization workflow for gene symbols and protein names of cell markers to address potential inconsistencies and ambiguities in nomenclature across data sources and species. At the gene level, we performed species-specific searches using the NCBI E-utilities API to obtain standardized gene symbols, descriptions and aliases [13]. When direct matching failed, we supplemented the results using three R packages, `org.Hs.eg.db`, `org.Mm.eg.db` and `org.Rn.eg.db`. At the protein level, we annotated proteins using the UniProt database under species-specific conditions, prioritizing reviewed entries [43]. To improve reliability, we conducted stringent quality control for both gene and protein annotations to ensure high accuracy and biological plausibility. For each marker, we ultimately provided standardized gene symbols, Entrez IDs, Ensembl IDs, full names and gene types, and explicitly recorded the corresponding data sources. We also ensured that the entire standardization procedure was traceable and reproducible to facilitate downstream analyses and verification.

Statistical comparisons

To systematically evaluate the overall performance of scMarkerAgent relative to comparable resources, we selected three publicly available databases as benchmarks: Cell Marker Accordion (CMA), CellMarker 2.0 and ACT [1, 2, 3]. ACT and CellMarker 2.0 are manually curated databases of annotated cell type-marker associations. CMA is an integrated database that aggregates 23 published human and mouse cell-type gene marker databases, together with cell-sorting marker sources.

We compared scMarkerAgent with CMA, CellMarker 2.0 and ACT across 15 quantitative dimensions: (1) the total number of cell type-marker annotations; (2) the number of unique markers after de-duplication; (3) the number of unique publications; (4) the number of unique tissue types after de-duplication; (5) the number of unique cell types after de-duplication; (6) the number of unique disease types after de-duplication; (7) the number of standardized tissue types based on the Uberon Ontology; (8) the number of standardized cell types based on the Cell Ontology; (9) the number of standardized disease types based on the Disease Ontology; (10) the number of human annotations; (11) the number of mouse annotations; (12) the number of rat annotations; (13) the total number of disease annotations; (14)

the number of positive marker annotations; and (15) the number of negative marker annotations.

Comparison of scRNA-seq annotation performance in scMarkerAgent

To evaluate the applicability and generalization performance of scMarkerAgent for scRNA-seq cell-type annotation, we analyzed public datasets spanning multiple species, tissues and disease states. To enable a controlled comparison across marker resources, we used ScType as the unified annotation algorithm because it supports user-defined marker inputs and incorporates both positive and negative markers via specificity-based weighting across clusters and candidate cell types [15]. In addition, we implemented a ScType+scMarkerAgent+LLM framework in which ScType first proposes the top candidate cell types based on scMarkerAgent markers, and an LLM then selects the final identity by interpreting marker patterns in context. The ScType+scMarkerAgent+LLM framework was designed to increase the likelihood that the correct identity appeared among the proposed candidates and enables the LLM to interpret marker polarity and specificity when resolving candidate identities, whereas ScType's raw scoring assigns equal weight to individual markers. ScType-based candidate pre-filtering keeps the LLM input concise, and the evidence-derived marker sets in scMarkerAgent ground the LLM's predictions to reduce spurious annotations [44]. In parallel, where evaluated, we benchmarked CASSIA (accessed at `https://www.cassia.bio/` on 17 December 2025) as a database-free, LLM-driven annotation workflow for comparison [7].

We selected seven benchmark scRNA-seq datasets with author-validated or expert-consensus annotations to evaluate annotation performance. To ensure comparability across studies, we used the cluster as the unit of comparison and defined the ground truth for each cluster as the most frequent known cell type within that cluster. For scoring, we followed the approach described by Hou et al. and evaluated agreement between predicted annotations and ground truth, assigning 1 point for an exact match, 0.5 points for a partial match and 0 points for a mismatch [6]. The benchmark datasets included: Sikkema et al., a core dataset from a healthy human lung cell atlas that includes lung tissue samples from 107 healthy individuals and provides manual cell-type annotations based on consensus from six independent experts [17]; Baron et al., a human pancreas single-cell transcriptome dataset comprising three healthy donors [18]; Baron et al., a mouse pancreas single-cell transcriptome dataset comprising two mice [18]; Hrovatin et al., a

mouse T2D pancreas single-cell dataset [20]; Hong et al., a rat lung single-cell dataset [21]; Zheng et al., a reference expression dataset of normal human PBMC subpopulations generated using magnetic bead enrichment [16]; and Ruiz-Moreno et al., GBmap, a multimodal single-cell and spatial atlas of IDH-wild-type glioblastoma integrating data from 240 patients and more than 1.1 million cells [19].

Comparison of spatial transcriptomics annotation performance in scMarkerAgent

To evaluate the applicability of scMarkerAgent in spatial transcriptomics, we analyzed the spatial transcriptomics dataset from Sun et al. [26] using a Seurat object containing 2,327,742 spatial profiles across 32 sections. We compared the ScType+scMarkerAgent workflow, the ScType+scMarkerAgent+LLM workflow, representative database- or LLM-based annotation workflows, and CASSIA at the cluster level using the same scoring scheme described above.

scMarkerAgent for custom cell scoring

In disease research, particularly in cancer studies, distinguishing malignant cells from normal cells is essential for accurate diagnosis, personalized treatment strategies, understanding the disease microenvironment, and improving therapeutic efficacy [45]. Malignant cells in tumors can be identified by examining cell-of-origin marker genes, tumor heterogeneity and copy number alterations [34]. However, traditional manual annotation methods are limited by incomplete marker knowledge and subjectivity, and there is currently no standardized large-scale disease-context cell markers database available to overcome these challenges [46].

Therefore, we generated 417,812 curated cell type-marker annotations in disease contexts, covering 68,456 publications. We recorded 15,386 diseases, including 3,667 standardized disease terms, compared with only 155 diseases in CMA (Fig. 3I).

To assess the accuracy of our annotations, we collected and organized the glioblastoma scRNA-seq dataset published in GBmap. We then filtered scMarkerAgent to obtain the marker set with tissue set to brain, cell type set to malignant cell, and disease name set to glioblastoma. We used the UCell algorithm to score each cell [25]. UCell is an R package designed for scRNA-seq data to evaluate gene-signature enrichment. Higher scores indicate a higher probability that a cell is malignant. We then compared these scores against the malignant cells annotated by experts in GBmap to

evaluate malignant scoring patterns.

Analysis of scRNA-seq data

We performed a unified scRNA-seq analysis workflow using the Seurat package (version 5.3.0) in R (version 4.5.1) [22]. Data were first log-normalized using the `NormalizeData` function, and the 2,000 most highly variable features were identified using the variance-stabilizing transformation selection method and scaled. Linear dimensionality reduction through principal component analysis was performed on scaled data, and scRNA-seq cells were clustered based on a shared nearest-neighbor graph (obtained with the `FindNeighbors` function using the first 30 principal components and `k`=20) employing the `FindClusters` function with the default Louvain algorithm for community detection (random seed set to 1234). For cluster visualization, we used uniform manifold approximation and projection (UMAP) with the same 30 principal components.

Implementation of scMarkerAgent

The frontend of scMarkerAgent is built with Vue.js and the Element UI component library, with ECharts used for interactive data visualization. The backend analytical pipeline is implemented in R, with a SQLite database storing curated cell-type-marker annotations. The platform supports task-based execution for both cell annotation and cell scoring modules. Each analysis generates a unique task ID and a persistent result URL, enabling users to monitor job status through the web interface. Users receive email notifications at task initiation and completion. The system is deployed on a high-performance server with 96 GB RAM and 2 TB storage.

Funding

This work was supported by the Natural Science Foundation of China (Grant Nos. 62471240 and 62231013 to Chen Cao) and the New Technology Project of Children's Hospital of Nanjing Medical University (XJS2024-108 to Xu Wang). Funding for open access charge: Natural Science Foundation of China (Grant Nos. 62471240 and 62231013).

Author contributions

H.C., Y.S. and H.Zhang conceived the study and designed the overall framework. H.C., Y.S., H.Zhang, Y.Liu and H.Zhao carried out data curation, method development, analysis and validation. X.W., G.W., G.L. and C.C. supervised the study. H.C., Y.S. and H.Zhang drafted the manuscript. Y.Liu, H.Zhao, X.W., G.W., G.L. and C.C. critically revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The data generated in this study have been deposited in Zenodo and are available at `https://doi.org/10.5281/zenodo.18825921`.

Code availability

The code used in this study is publicly available at `https://github.com/HaoRan-Code/scMarkerAgent` under the MIT license.

Table 1: Sentence-level evidence retrieved by scMarkerAgent for LLM-guided annotation refinement in ScType+scMarkerAgent+LLM.

PMC ID	Marker	Sentence-based evidence
PMC10116630	AGER	Early progenitors AT1/2 cells co-expressed the AT1 marker AGER as well as surfactant proteins SFTPD, SFTPC and SFTPB.
PMC11355357	KRT8	The epithelial cells could be divided into alveolar type 1 cells (AT1), alveolar type 2 cells (AT2) and KRT8+ PATS (KRT8+ pre-alveolar type 1 transitional cell state) cells based on specific marker expressions.
PMC10287567	SFTPB	We were further able to detect six cell identities that were not previously found in the human lung or were only recently described in individual studies, including the recently described alveolar type 0 cells (n = 1,440, expressing SFTPB+, SCGB3A2+, SFTPChigh and SCGB3A1low).

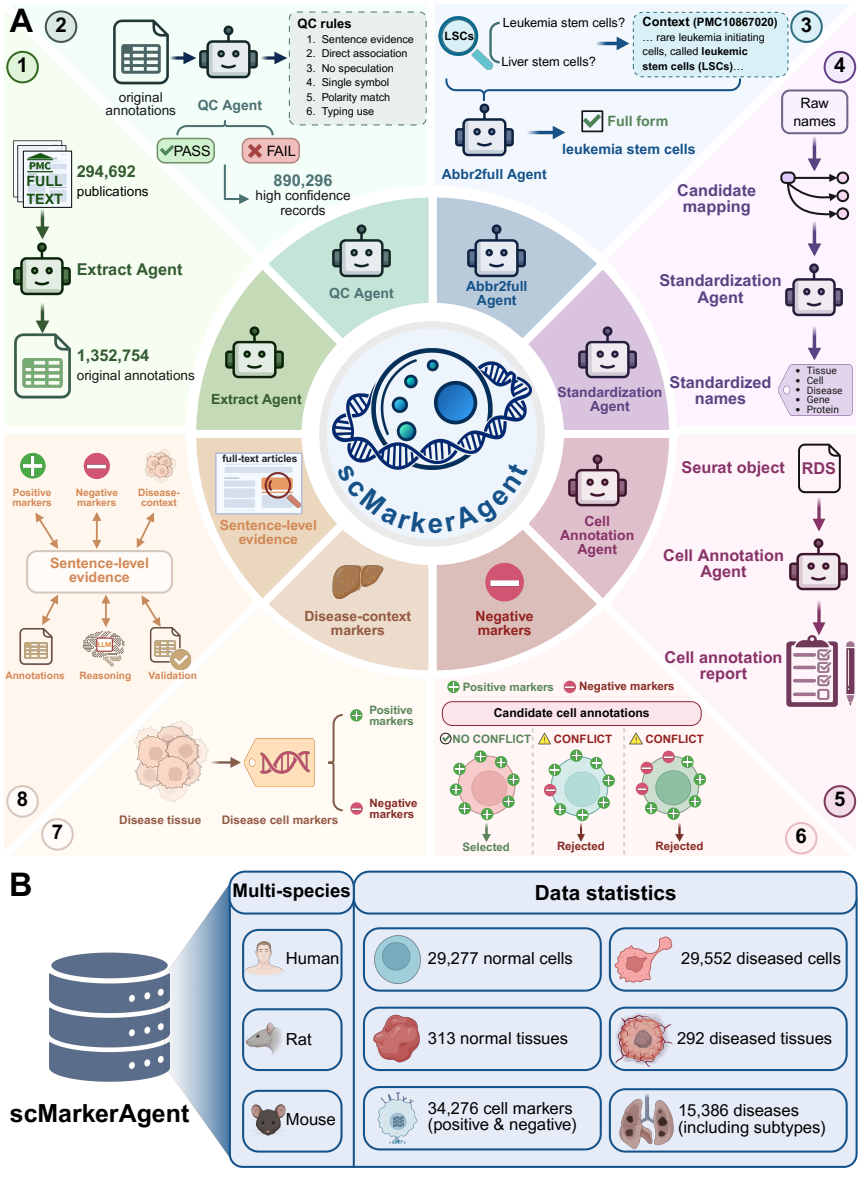


Figure 1: Overview of scMarkerAgent. (A) scMarkerAgent multi-agent framework and core features. Four LLM-based agents, Extract Agent, QC Agent, Abbr2full Agent and Standardization Agent, are employed to support evidence-guided knowledge construction, while the Cell Annotation Agent is integrated into the web platform for automated cell annotation and report generation. Three key features are highlighted: sentence-level evidence, disease-context markers and negative markers. (B) Multi-species summary statistics of scMarkerAgent. Created in BioRender. Chen, H. (2026) `https://BioRender.com/6lka6zy`.

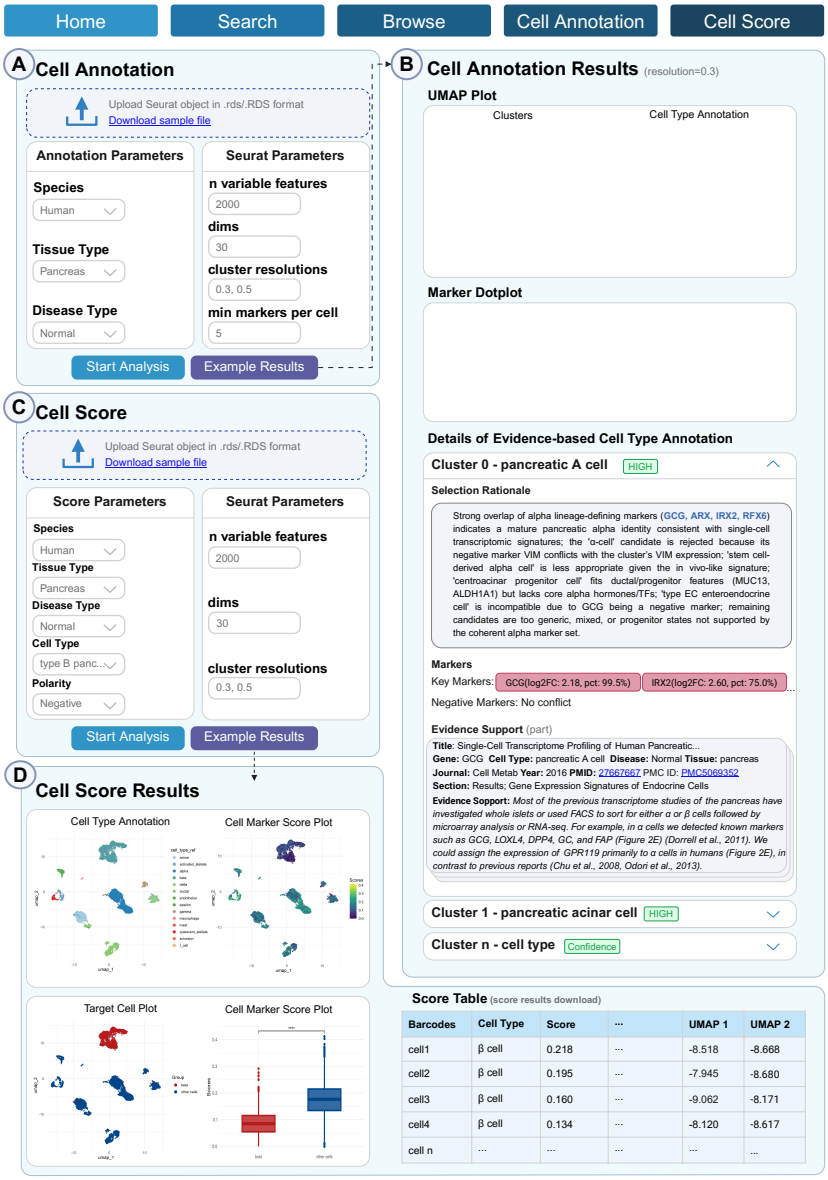


Figure 2: The Cell Annotation and Cell Score pages of scMarkerAgent. (A) The Cell Annotation page allows users to upload a Seurat object and set annotation-related and Seurat parameters for cell annotation using the Cell Annotation Agent. (B) The Cell Annotation Results page displays UMAP plots based on clusters and cell type annotation, together with a marker dot plot showing cell-marker relationships. For each cluster, the results provide details of the evidence-based cell type annotation. (C) The Cell Score page allows users to upload a Seurat object and set score-related and Seurat parameters for scoring a specified cell type. (D) The Cell Score Results page presents the scoring results in plots and tables. Created in BioRender. Chen, H. (2026) `https://BioRender.com/k7954ak`.

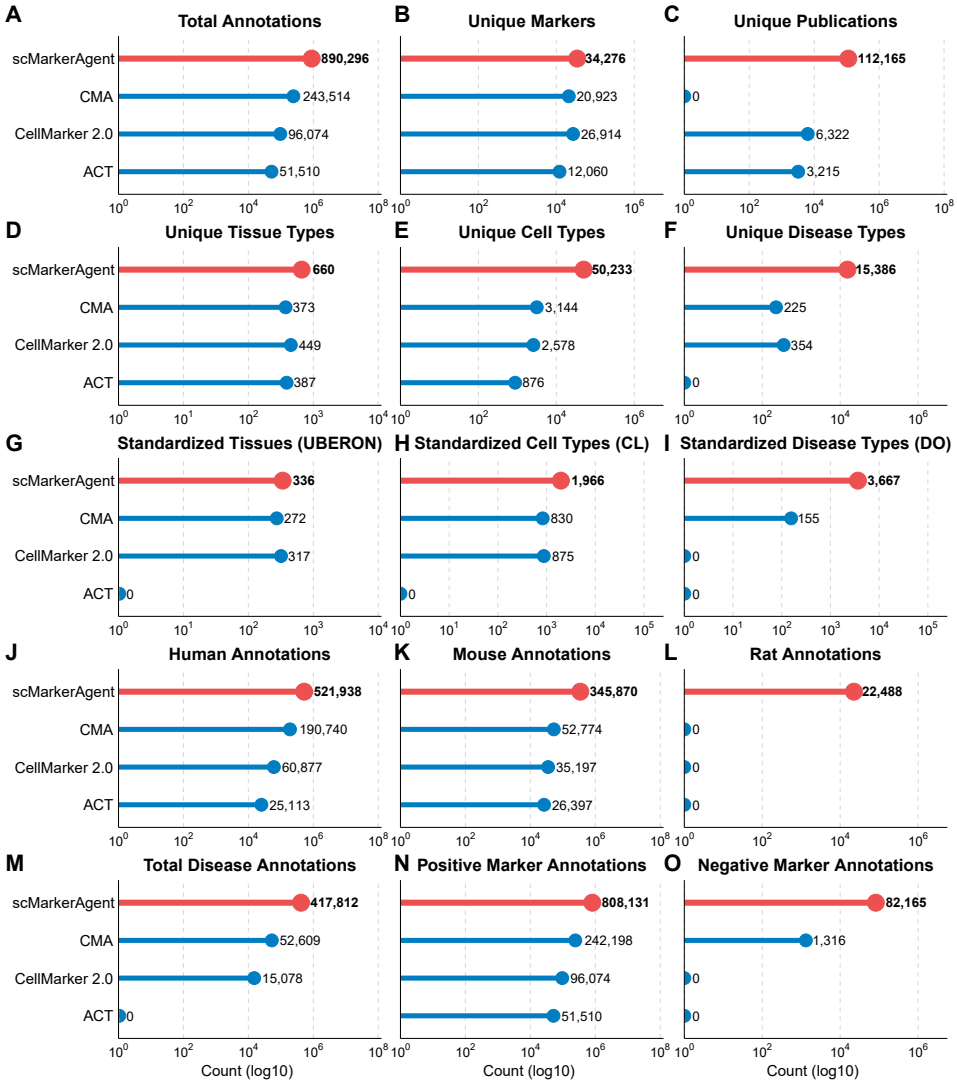


Figure 3: Comparison of scMarkerAgent with CMA, CellMarker 2.0 and ACT across 15 dimensions.

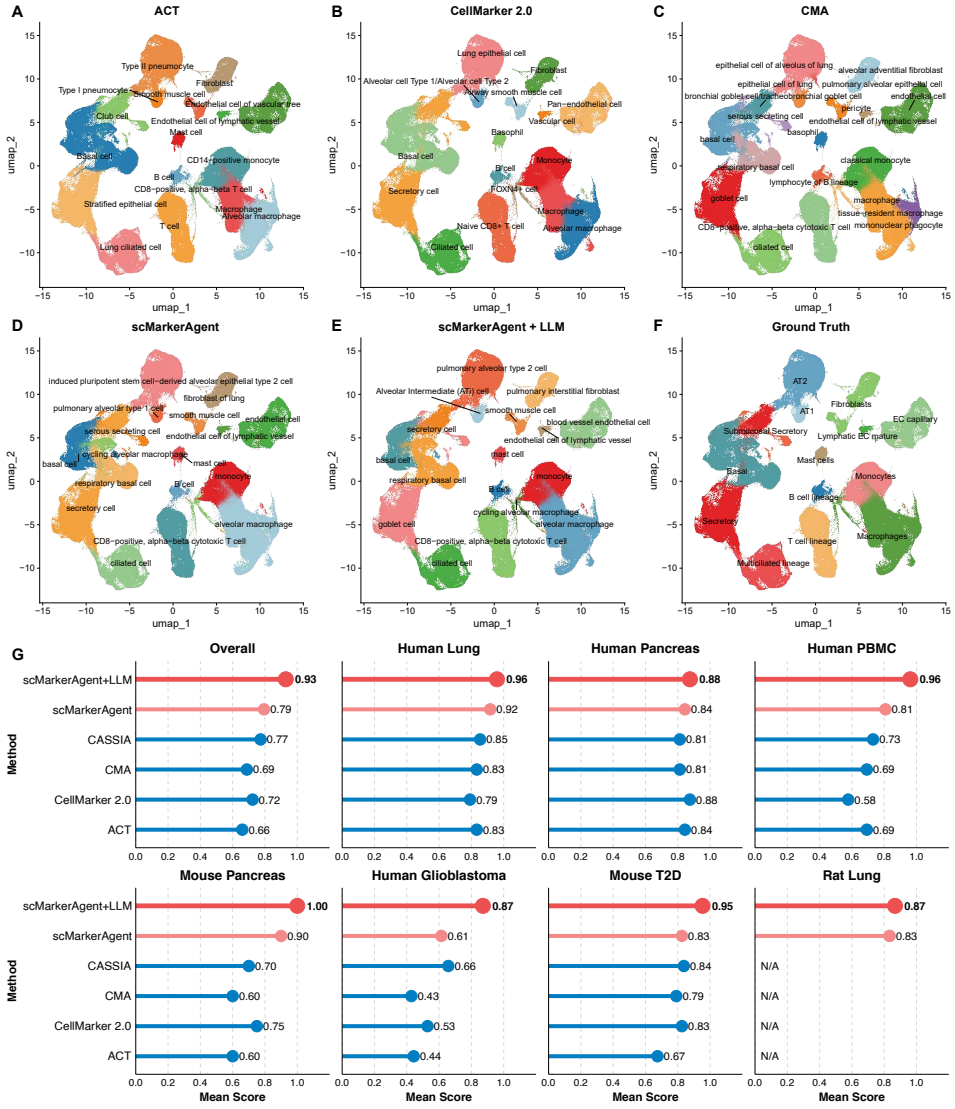


Figure 4: Performance of the Cell Annotation Agent (ScType+scMarkerAgent+LLM) on real-world cell-type annotation and comparison with peer marker or annotation resources. (A) ACT. (B) CellMarker 2.0. (C) CMA. (D) scMarkerAgent. (E) scMarkerAgent+LLM. (F) Authors' annotation (ground truth). (G) Comparison of the mean annotation scores of the scMarkerAgent+LLM workflow versus other databases or annotation workflows. Except for CASSIA, all methods shown here are ScType-based workflows; the "ScType+" prefix is omitted from panel labels for clarity.

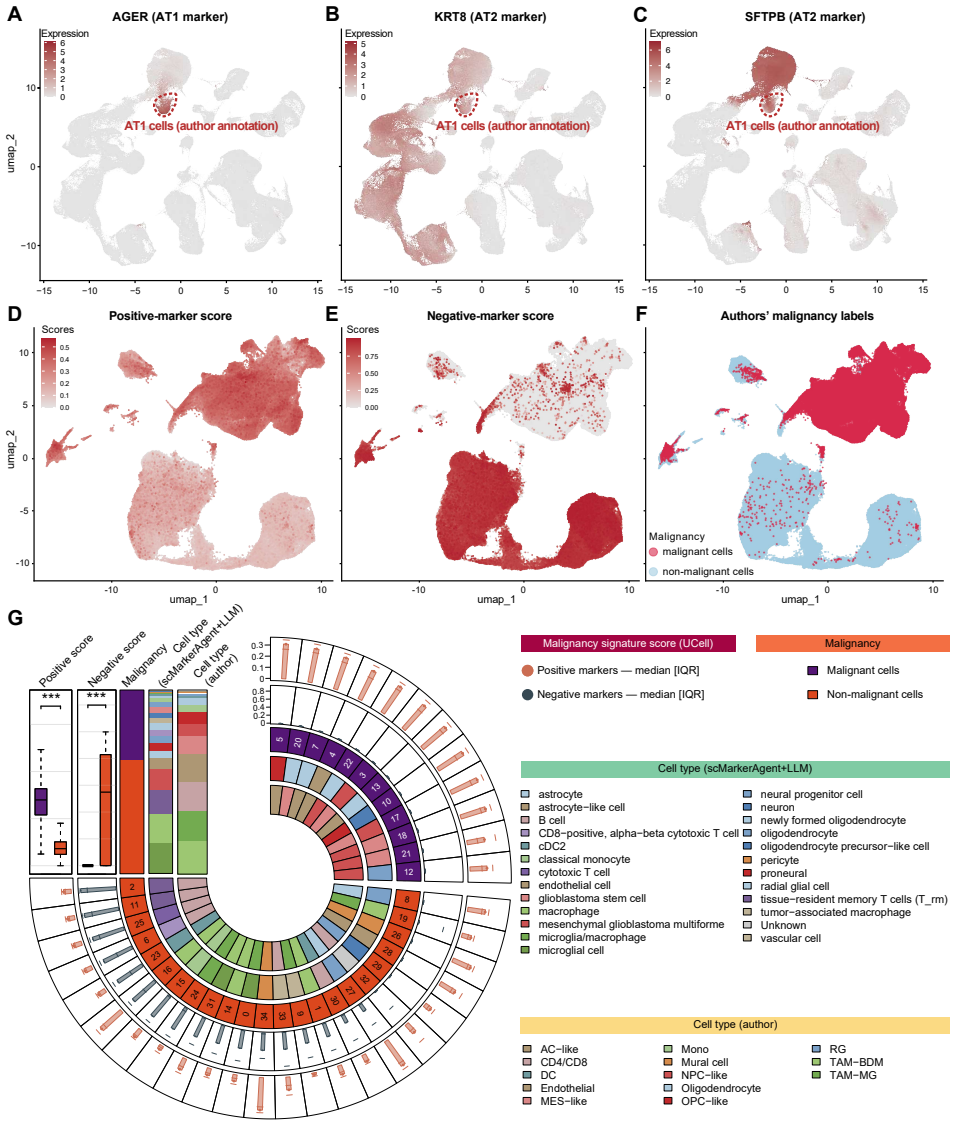


Figure 5: Expression of negative markers in the dataset. (A–C) Expression of (A) AGER, (B) KRT8 and (C) SFTPB in the human lung dataset. The red dashed line outlines the area annotated as AT1 (alveolar type I cells) by the authors. (D–F) Gene-set scoring results on the human glioblastoma dataset using glioblastoma (D) positive and (E) negative markers from scMarkerAgent. (F) Malignancy labels annotated by the authors. (G) Summary of the human glioblastoma dataset. From the outer to inner rings, the panel shows positive-marker score (median [IQR]), negative-marker score (median [IQR]), malignancy (malignant versus non-malignant), cell-type predictions (scMarkerAgent+LLM) and authors' cell-type annotations.

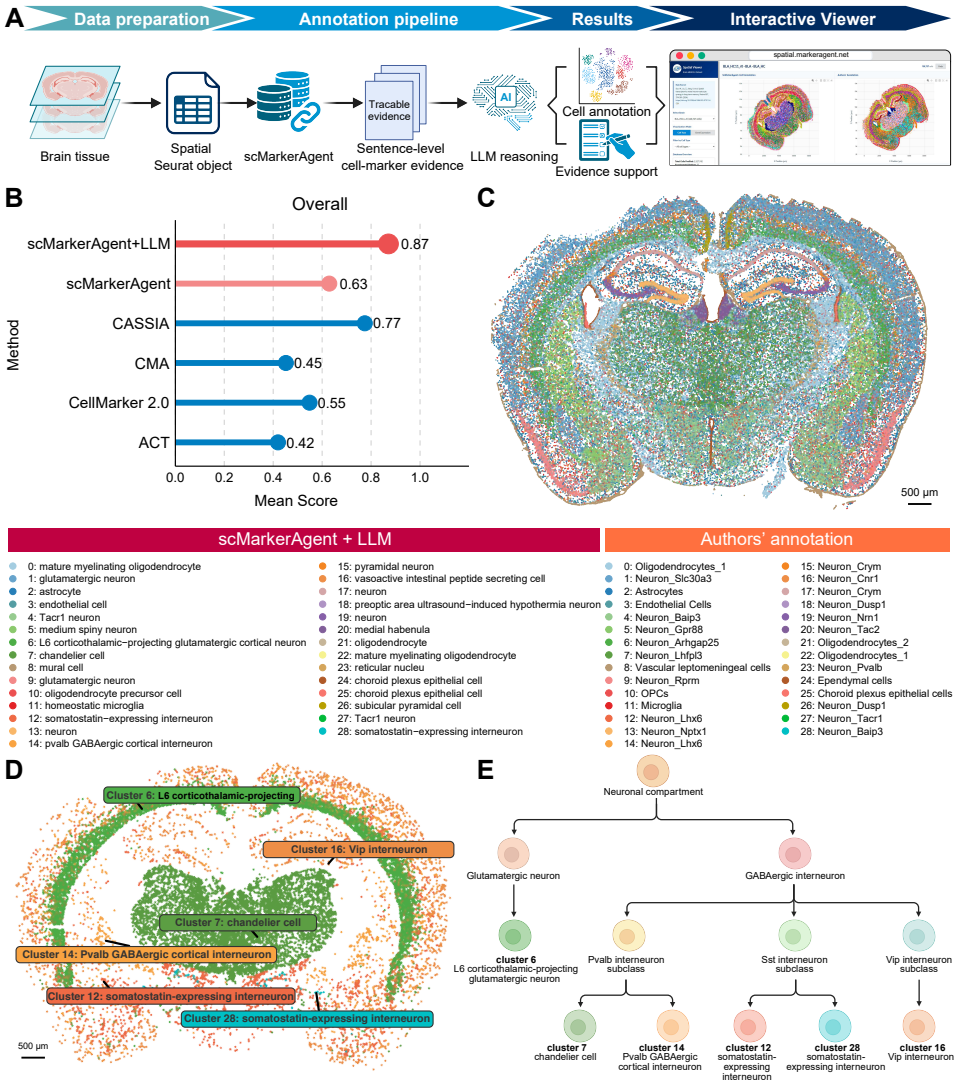


Figure 6: Performance of the scMarkerAgent+LLM workflow on spatial transcriptomics. (A) Data processing pipeline. Created in BioRender. Chen, H. (2026) `https://BioRender.com/ul4hfe7`. (B) Comparison of mean cluster-level annotation scores for scMarkerAgent+LLM versus scMarkerAgent alone and representative database- or LLM-based annotation workflows. (C) Spatial coordinates colored by cluster identity, with cell-type legends shown for the scMarkerAgent+LLM annotations (left) and the original Sun et al. annotations (right); cluster IDs (0–28) are provided for direct correspondence with the displayed regions. (D) DimPlot of selected refined neuronal annotations. (E) Classification schematic of the selected refined neuronal annotations. Created in BioRender. Chen, H. (2026) `https://BioRender.com/5fd7mlm`.

References

[1] Busarello, E. et al. Cell Marker Accordion: interpretable single-cell and spatial omics annotation in health and disease. *Nature Communications* **16**, 5399 (2025).

[2] Hu, C. 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 Research* **51**, D870–D876 (2023).

[3] Quan, F. et al. Annotation of cell types (ACT): a convenient web server for cell type annotation. *Genome Medicine* **15**, 91 (2023).

[4] Clarke, Z.A. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols* **16**, 2749–2764 (2021).

[5] Nouri, N., Gaglia, G., Kurlovs, A.H., de Rinaldis, E. & Savova, V. A marker gene-based method for identifying the cell-type of origin from single-cell RNA sequencing data. *MethodsX* **10**, 102196 (2023).

[6] Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods* **21**, 1462–1465 (2024).

[7] Xie, E. et al. CASSIA: a multi-agent large language model for automated and interpretable cell annotation. *Nature Communications* (2025).

[8] Kazmi, A., Singh, D., Jatav, S. & Luthra, S. Beyond the hype: The complexity of automated cell type annotations with GPT-4. *bioRxiv* 2025.02.11.637659 (2025).

[9] Truhn, D., Reis-Filho, J.S. & Kather, J.N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine* **29**, 2983–2984 (2023).

[10] Haendel, M.A. et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics* **5**, 21 (2014).

[11] Diehl, A.D. et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics* **7**, 44 (2016).

[12] Schriml, L.M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research* **47**, D955–D962 (2019).

[13] Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research* **53**, D20–D29 (2025).

[14] Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship: Comment. *Scientific Data* **3**, 160018 (2016).

[15] Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications* **13**, 1246 (2022).

[16] Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

[17] Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine* **29**, 1563–1577 (2023).

[18] Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems* **3**, 346–360.e4 (2016).

[19] Ruiz-Moreno, C. et al. Charting the single-cell and spatial landscape of IDH-wild-type glioblastoma with GBmap. *Neuro-Oncology* **27**, 2281–2295 (2025).

[20] Hrovatin, K. et al. Delineating mouse β-cell identity during lifetime and in diabetes with a single cell atlas. *Nature Metabolism* **5**, 1615–1637 (2023).

[21] Hong, J. et al. Single-cell study of two rat models of pulmonary arterial hypertension reveals connections to human pathobiology and drug repositioning. *American Journal of Respiratory and Critical Care Medicine* **203**, 1006–1022 (2021).

[22] Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**, 293–304 (2024).

[23] Program, C.C.S. et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research* **53**, D886–D900 (2025).

[24] Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).

[25] Andreatta, M. & Carmona, S.J. UCell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal* **19**, 3796–3798 (2021).

[26] Sun, W. et al. Spatial transcriptomics reveal neuron-astrocyte synergy in long-term memory. *Nature* **627**, 374–381 (2024).

[27] Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology* **20**, 194 (2019).

[28] Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24**, 550–572 (2023).

[29] Lee, J., Kim, M., Kang, K., Yang, C.-S. & Yoon, S. Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing. *Briefings in Bioinformatics* **24**, bbad006 (2023).

[30] Liebler, J.M. et al. Combinations of differentiation markers distinguish subpopulations of alveolar epithelial cells in adult lung. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **310**, L114–L120 (2016).

[31] Strunz, M. et al. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nature Communications* **11**, 3559 (2020).

[32] Han, S., Budinger, G.S. & Gottardi, C.J. Alveolar epithelial regeneration in the aging lung. *The Journal of Clinical Investigation* **133** (2023).

[33] Zhong, Z. et al. Domain generalization enables general cancer cell annotation in single-cell and spatial transcriptomics. *Nature Communications* **15**, 1929 (2024).

[34] Andreatta, M., Garnica, J. & Carmona, S.J. Identification of malignant cells in single-cell transcriptomics data. *Communications Biology* **8**, 1264 (2025).

[35] Cui, H. et al. scGPT: toward building a foundation model for single-cell multiomics using generative AI. *Nature Methods* **21**, 1470–1480 (2024).

[36] Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

[37] Walters, W.H. & Wilder, E.I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* **13**, 14045 (2023).

[38] Wu, K. et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications* **16**, 3615 (2025).

[39] Ye, W. et al. Evaluation of cell type annotation reliability using a large language model-based identifier. *Communications Biology* **8**, 1360 (2025).

[40] Yang, C., Zhang, X. & Chen, J. Large language model consensus substantially improves the cell type annotation accuracy for scRNA-seq data. *bioRxiv* 2025.04.10.647852 (2025).

[41] Crowley, G. & Quake, S.R. Benchmarking cell type and gene set annotation by large language models with AnnDictionary. *Nature Communications* **16**, 9511 (2025).

[42] Wang, S. et al. Leveraging the Cell Ontology to classify unseen cell types. *Nature Communications* **12**, 5556 (2021).

[43] Ahmad, S. et al. The UniProt website API: facilitating programmatic access to protein knowledge. *Nucleic Acids Research* **53**, W547–W553 (2025).

[44] Levy, M., Jacoby, A. & Goldberg, Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15339–15353 (2024).

[45] Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52**, 1452–1465 (2020).

[46] Cheng, C., Chen, W., Jin, H. & Chen, X. A review of single-cell RNA-seq annotation, integration, and cell-cell communication. *Cells* **12**, 1970 (2023).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryInformation.pdf