Supplementary information for scMarkerAgent: An LLM Evidence Agent-based Cell Marker Atlas

Table of Contents

Supplementary Note 1

Cell Annotation Agent: The Cell Annotation Agent is a two-stage, cluster-level annotation module. In Stage 1, candidate cell types for each transcriptionally defined cluster are generated by computing ScType marker-gene enrichment scores against a species/tissue/disease-filtered subset of the scMarkerAgent database, followed by differential expression analysis (Wilcoxon test, log-fold-change threshold) and hypergeometric enrichment testing to quantify the overlap between each cluster's significant markers and each candidate's positive and negative gene sets. In Stage 2, each cluster's ranked candidate list—together with the overlap gene information—is submitted to a large language model per cluster, which selects the most biologically appropriate cell type identity and returns a structured JSON response containing the selected cell type, confidence level, biological reasoning, and key validated marker genes.

1. Example Input

For each cluster, the following information is assembled from Stage 1 and passed to Stage 2:

Field	Example (Cluster 3)
Species	Human
Tissue	Brain
Disease	Glioblastoma
Cluster ID	3
Significant cluster markers (n)	87
Rank 1 candidate	Astrocytes; overlap_positive: [GFAP, AQP4, SLC1A2]; overlap_negative: []
Rank 2 candidate	Oligodendrocytes; overlap_positive: [MBP, PLP1]; overlap_negative: [GFAP]
Rank 3 candidate	Neurons; overlap_positive: [SNAP25, RBFOX3]; overlap_negative: [GFAP, AQP4]

2. Constructed Prompt

```
You are a cell biology and single-cell transcriptomics expert.

CRITICAL INSTRUCTIONS:
1. Reply with ONLY a single line of valid JSON - NO other text before or after
2. Do NOT use markdown code blocks (no ```)
3. Do NOT add explanations or comments
4. The JSON must be minified (no line breaks inside the JSON)
5. Start your response directly with { and end with }

Required JSON schema:
{"selected_celltype": "string", "confidence": "high|medium|low", "reasoning": "string",
"key_markers_validated": ["gene1", "gene2"]}

Task input:
{
```

```
  "instruction": "Select the most biologically appropriate cell type annotation for this cluster from the ranked candidates.",
  "rules": [
    "If none of the candidates are biologically reasonable, return 'Unknown' as selected_celltype",
    "Candidates with negative markers overlapping cluster markers must be rejected due to biological conflict",
    "Select the candidate whose overlapping positive markers most coherently represent that cell type's lineage-defining features and functional identity in vivo",
    "Prefer candidates where overlapping markers match known cell-surface phenotypes used in flow cytometry and transcriptional signatures validated in single-cell studies",
    "The selected_celltype must exactly match one candidate celltype name or be 'Unknown'",
    "key_markers_validated must contain only genes from the selected candidate's overlap_positive_markers",
    "Each gene symbol in key_markers_validated must use exactly the same spelling and letter casing as in overlap_positive_markers (do NOT change capitalization, abbreviate, or invent new gene symbols)",
    "Do not introduce any gene in key_markers_validated that is not present in overlap_positive_markers",
    "In the reasoning field, explain both why the selected cell type is most appropriate AND why other highly-ranked candidates were rejected"
  ],
  "inputs": {
    "species": "Human",
    "tissue": "Brain",
    "disease": "Glioblastoma",
    "cluster_id": "3",
    "n_significant_cluster_markers": 87,
    "candidates": [
      {
        "rank": 1,
        "celltype": "Astrocytes",
        "overlap_positive_markers": ["GFAP", "AQP4", "SLC1A2"],
        "overlap_negative_markers": []
      },
      {
        "rank": 2,
        "celltype": "Oligodendrocytes",
        "overlap_positive_markers": ["MBP", "PLP1"],
        "overlap_negative_markers": ["GFAP"]
      },
      {
        "rank": 3,
        "celltype": "Neurons",
```

```
            "overlap_positive_markers": ["SNAP25", "RBFOX3"],
            "overlap_negative_markers": ["GFAP", "AQP4"]
          }
        ]
    },
    "output_format": {
      "format": "minified JSON",
      "schema": {
        "selected_celltype": "string (must be from candidates or 'Unknown')",
        "confidence": "string (must be 'high', 'medium', or 'low')",
        "reasoning": "string (explain why selected cell type is chosen AND why others were rejected)",
        "key_markers_validated": "array of strings (2–5 key gene symbols)"
      }
    }
}
```

Your response (pure JSON only):

3. Example Output

```
{
  "selected_celltype": "Astrocytes",
  "confidence": "high",
  "reasoning": "Cluster 3 expresses canonical astrocyte markers GFAP, AQP4, and SLC1A2, which define astrocyte identity through structural (GFAP intermediate filament), water-transport (AQP4), and glutamate-uptake (SLC1A2) functions well validated in human brain single-cell studies. Oligodendrocytes (Rank 2) are rejected because GFAP appears in their negative marker list, indicating a direct biological conflict with the cluster's expression profile. Neurons (Rank 3) are rejected for the same reason: both GFAP and AQP4 are listed as negative markers, contradicting the observed expression pattern.",
  "key_markers_validated": ["GFAP", "AQP4", "SLC1A2"]
}
```

Supplementary Note 2

Extract Agent: Extract Agent is an automated extraction pipeline that leverages large language models (LLMs) with a rigorous, rule-based prompt engineering framework to identify and structure cell type-specific marker gene information from scientific text sections.

1. Example Input

The pipeline accepts paper metadata and a specific text section as input.

Metadata:

Title: Single-cell RNA sequencing reveals the immune landscape of human lung cancer

PMID: 12345678

Current Section: Results

Section Number: 2 of 10

Text Segment:

> "In the lung cancer microenvironment, we annotated cell types based on canonical markers. T cells were identified by the high expression of CD3D and CD3E. B cells were characterized by MS4A1 (CD20) expression. In contrast, Natural Killer (NK) cells were defined by the presence of NKG7 and the absence of CD3D. "

2. Constructed Prompt

You are analyzing a scientific paper focused on single-cell RNA sequencing. Extract cell annotation information from the provided text section.

RESPONSE CONTRACT (STRICT):

Return ONLY valid JSON using the exact structure below

No markdown, explanations, or extra text

If no data found, return: {"cell_annotations": []}

TASK SCOPE:

Extract how markers are used to IDENTIFY and ANNOTATE cell types in this text (not expression analysis results or functional studies). NUMBERED RULES:

Inclusion (I):

I1: Marker explicitly used to identify/classify/define/characterize/distinguish the specific cell type

I2: Marker described as known/canonical/established/standard/specific/validated for identification/annotation

I3: Marker associated with recognized biological cell type names using standard anatomical/immunological terminology

I4: Marker must be an individual gene symbol or protein name (not descriptive terms or gene families)

Polarity (P):

P1: Positive = presence/expression stated as defining/identifying/characterizing/required for identity; used to include cells

P2: Negative = absence/non-expression stated as defining/identifying/distinguishing; used to exclude or confirm identity

P3: Assign polarity solely based on the identification context for the specific cell type being annotated

Exclusion (E):

E1: Expression analysis only (levels/abundance/enrichment/DE/fold-changes) without identification context
E2: Functional role context (functions/pathways/processes/physiology) rather than identification
E3: Computational classifications (numbered clusters/experimental groups/technical samples/analysis results)
E4: Uncertain relationships (speculative/conditional/uncertain language)
E5: Experimental findings lacking established identification validation
E6: Non-specific references (vague terms/no specific genes/generic categories)
E7: Methodological context (methods/technical procedures/research tools)
E8: Pure expression description: absence mentioned without explicit use for identification/distinction
E9: Low-expression only: markers described as low/lowly expressed or reduced levels; such "low" does not satisfy Positive/Negative and must be excluded
Source Evidence (S):
S1: Copy text verbatim; preserve characters/casing/punctuation/spacing
S2: Use complete sentence(s) as the minimal evidence unit; one or more sentences (optionally joined using "...") must evidence: tissue_class, tissue_type, disease_type, cell_name, marker, marker_polarity
S3: Be minimal; include only clauses necessary to evidence fields (no unrelated sentences)
S4: Ellipsis "..." may connect multiple sentences from the same section while preserving original order; irrelevant sentences may be omitted; do not modify any words within sentences
S5: No paraphrasing/normalization/inference/brackets/commentary; only text from the paper
S6: If the section lacks sufficient text to support all required fields, output no annotation for it
Field Definitions (F):
F1: species = Organism studied (Human/Mouse/Rat preferred; else "Other")
F2: tissue_class = Broad tissue category
F3: tissue_type = Specific tissue subtype/location
F4: disease_type = Disease context of the sample
If source mentions patients with specific diseases, extract the full disease name (use full name over abbreviation when possible)
For ambiguous abbreviations without clear evidence, you may use the abbreviation
If source describes healthy donors, controls, or normal samples without disease context, return "Normal"
If disease context is unclear or not mentioned, return "Normal"
F6: cell_name = Biological cell type name (not cluster/sample IDs) exactly as in text (no generalization; no cross-section normalization)
F7: marker = Specific gene symbol or protein name used for identification (never descriptive terms; only individual genes explicitly used to identify/classify/annotate)
F8: marker_polarity = Determined by identification usage (Positive/Negative per P1/P2)
F9: source = Minimal verbatim sentence(s) evidencing all required fields; may join multiple sentences using "..." in original order
Naming Rules (N):
N1: tissue_class/tissue_type use canonical anatomical nouns (structures/organs/regions); avoid verbs/standalone adjectives

N2: Choose the most accurate cell type name supported by the current section
N3: Use the original description for cell_name; do not generalize subtypes to parent types
N4: Do not normalize across sections; align only when entity and specificity are identical
Context Usage (C):
C1: Base all field values on the current section text
C2: Use previous sections only as spelling/terminology reference at the same specificity level
C3: Do not propagate/infer from previous sections; do not collapse subtypes into broader parents
INTERNAL SELF-CHECK (do not output this checklist):
Before producing JSON, internally verify compliance with all I, *P*, E, *S*, F, *N*, C* rules.
PAPER INFORMATION:
Title: Single-cell RNA sequencing reveals the immune landscape of human lung cancer
PMID: 12345678
Current Section: Results
Section Number: 2 of 10
TEXT TO ANALYZE:
In the lung cancer microenvironment, we annotated cell types based on canonical markers. T cells were identified by the high expression of CD3D and CD3E. B cells were characterized by MS4A1 (CD20) expression. In contrast, Natural Killer (NK) cells were defined by the presence of NKG7 and the absence of CD3D.
REQUIRED JSON STRUCTURE:
{
"cell_annotations": [
{
"species": "Organism species name",
"tissue_class": "Broad tissue system category",
"tissue_type": "Specific tissue subtype or location",
"disease_type": "Full disease name if diseased; Normal if healthy",
"cell_name": "Biological cell type identifier (if subtype, preserve subtype specificity exactly as supported by the row source)",
"marker": "Gene symbol or protein name",
"marker_polarity": "Positive or Negative",
"source": "Original text supporting the extraction, MUST be original text from the paper"
}
]
}

3. Example Output (JSON format)
{
"cell_annotations": [
{
"species": "Human",
"tissue_class": "Lung",
"tissue_type": "Lung",
"disease_type": "Lung cancer",
"cell_name": "T cells",

"marker": "CD3D",
"marker_polarity": "Positive",
"source": "In the lung cancer microenvironment... T cells were identified by the high expression of CD3D and CD3E."
},
{
"species": "Human",
"tissue_class": "Lung",
"tissue_type": "Lung",
"disease_type": "Lung cancer",
"cell_name": "T cells",
"marker": "CD3E",
"marker_polarity": "Positive",
"source": "In the lung cancer microenvironment... T cells were identified by the high expression of CD3D and CD3E."
},
{
"species": "Human",
"tissue_class": "Lung",
"tissue_type": "Lung",
"disease_type": "Lung cancer",
"cell_name": "B cells",
"marker": "MS4A1",
"marker_polarity": "Positive",
"source": "In the lung cancer microenvironment... B cells were characterized by MS4A1 (CD20) expression."
},
{
"species": "Human",
"tissue_class": "Lung",
"tissue_type": "Lung",
"disease_type": "Lung cancer",
"cell_name": "Natural Killer (NK) cells",
"marker": "NKG7",
"marker_polarity": "Positive",
"source": "In the lung cancer microenvironment... Natural Killer (NK) cells were defined by the presence of NKG7 and the absence of CD3D."
},
{
"species": "Human",
"tissue_class": "Lung",
"tissue_type": "Lung",
"disease_type": "Lung cancer",
"cell_name": "Natural Killer (NK) cells",

"marker": "CD3D",
"marker_polarity": "Negative",
"source": "In the lung cancer microenvironment... Natural Killer (NK) cells were defined by the presence of NKG7 and the absence of CD3D."
}
]
}

Supplementary Note 3

QC Agent: Quality Control Agent is a strict, post-extraction verification module that audits each annotation entry against six mandatory validity rules to ensure high-precision data integrity before database insertion.

1. Example Input

The controller receives a batch of candidate annotations extracted from the **Extract Agent**.

Row ID	Cell Name	Marker	Polarity	Source Text
1	T cells	CD3D	Positive	"T cells were identified by the high expression of CD3D."
2	Macrophages	low levels of CD14	Positive	"Macrophages expressed low levels of CD14 compared to monocytes."
3	B cells	B cell receptor genes	Positive	"B cells were defined by the expression of B cell receptor genes."
4	NK cells	CD3D	Negative	"NK cells were distinguishable by the absence of CD3D."

2. Constructed Prompt

You are performing strict quality control on cell annotation data. Your task is to validate and correct annotation rows based on mandatory rules, ensuring only high-quality annotations are retained.

RESPONSE FORMAT: You MUST respond with ONLY valid JSON. No other text, explanations, or markdown formatting.

MANDATORY VALIDATION RULES:

Marker Specificity: marker must be a single specific gene symbol or protein name; never gene families, descriptive terms, or compound designations

Identification Purpose Clarity: A marker in the source text must be explicitly used for cell type identification, classification, definition, or distinction. It must not be used solely for expression analysis or functional studies. Exclude markers that are only described as "low" or "reduced expression," since such usage does not meet the criteria for Positive/Negative definition.

Source Evidence Sufficiency: source must be minimal verbatim complete sentence(s) explicitly supporting the marker–cell_name identification relationship; multiple sentences may be joined with "..." in original order; do not modify any words

Marker–Cell Association Clarity: source must explicitly demonstrate association of the marker with the specific cell_name for identification purposes

Marker-Polarity Relationship Accuracy: marker_polarity must be determined by identification usage: positive = presence/expression defines identity; negative = absence/non-expression distinguishes identity. Note: "low" or "reduced" expression is NOT negative.

Marker Certainty: source cannot be speculative, conditional, or uncertain descriptions of marker–cell relationships

ROW INDEPENDENCE AND SPECIFICITY:

Evaluate each row independently using only its source; do not infer or propagate across rows

Preserve subtype specificity present in the source; do not generalize a subtype to a broader parent term

PROCESSING INSTRUCTIONS:

For each row:
If ALL rules are satisfied: return action "pass" with no field modifications
If some rules fail but can be corrected while maintaining rule compliance and specificity: return action "pass" with only the corrected field values
If rules cannot be satisfied: return action "fail" with failure_reason explaining why it fails validation
EFFICIENCY OPTIMIZATION:
Only return field values that need to be changed, not complete rows
For "pass" actions, only include fields that were corrected from original values
Do not return fields that remain unchanged
INPUT DATA:
[JSON representation of the input table above]
REQUIRED OUTPUT FORMAT:
{
"validation_result": {
"processed_rows": [
{
"row_id": <integer>,
"action": "pass|fail",
"failure_reason": "explanation if action is fail",
"corrections": {
"cell_name": "KEEP_ORIGINAL or corrected value",
"marker": "KEEP_ORIGINAL or corrected value",
"marker_polarity": "KEEP_ORIGINAL or corrected value"
}
}
]
}
}

3. Example Output


```
{
  "validation_result": {
    "processed_rows": [
      {
        "row_id": 1,
        "action": "pass",
        "failure_reason": null,
        "corrections": {
          "cell_name": "KEEP_ORIGINAL",
          "marker": "KEEP_ORIGINAL",
          "marker_polarity": "KEEP_ORIGINAL"
        }
      },
      {
```

```
      "row_id": 2,
      "action": "fail",
      "failure_reason": "Marker described as 'low levels' does not satisfy Positive/Negative identification criteria (Rule 2 & 5).",
      "corrections": {}
    },
    {
      "row_id": 3,
      "action": "fail",
      "failure_reason": "Marker is a descriptive term/gene family, not a specific gene symbol (Rule 1).",
      "corrections": {}
    },
    {
      "row_id": 4,
      "action": "pass",
      "failure_reason": null,
      "corrections": {
        "cell_name": "KEEP_ORIGINAL",
        "marker": "KEEP_ORIGINAL",
        "marker_polarity": "KEEP_ORIGINAL"
      }
    }
  ]
 }
}
```

Supplementary Note 4

Abbr2full Agent: The Abbr2full Agent is a two-stage, publication-specific abbreviation expansion module that resolves tissue and cell-type abbreviations into their unambiguous full names to prevent erroneous marker–cell associations (e.g., "LSCs" could denote either *leukemic stem cells* or *liver stem cells* depending on the publication context). In Stage 1 (Classification), each unique term from the three annotation fields—tissue class, tissue type, and cell type—is submitted in batch to the LLM, which independently judges whether each term is an abbreviation/acronym (ABBREV) or already a full name (FULL). In Stage 2 (Resolution), for each term judged as ABBREV, up to three contextual sentences containing the abbreviation are retrieved from the full text of the corresponding paper (via a precomputed full-text SQLite database indexed by PMC ID), and both the abbreviation and its retrieved contexts are submitted to the LLM for expansion. The LLM returns the expanded full name, a resolution status, and the minimal supporting source sentence. Terms whose contexts contain multiple irresolvable plausible expansions are marked as "failed" and retained in their original form.

1. Example Input

Stage 1 — Classification receives a batch of (label, pmcid) pairs for the cell_name field (each field is processed independently):

Row	label	pmcid
1	LSCs	PMC10867020
2	blast cells	PMC10867020
3	HSCs	PMC10867020

Stage 2 — Resolution receives, for each term classified as ABBREV, the abbreviation together with up to 3 contextual sentences retrieved from the paper's full text (CTX_LIMIT = 3):

pmcid	label	contexts
PMC10867020	LSCs	["patients harbour a population of quiescent leukemic stem cells (LSCs) which can emerge from quiescence to trigger relapse after therapy.", "This blast population is replenished by rare leukemia initiating cells, called leukemic stem cells (LSCs).", "Similar to healthy hematopoietic stem cells (HSCs), LSCs are generally quiescent and are therefore thought to be responsible for relapse following chemotherapy which targets rapidly proliferating cells."]

2. Constructed Prompt

Stage 1 — Classification Prompt (cell_name field):

Task: For each row, decide if 'label' is an abbreviation/acronym/initialism or a full name.

Decision basis: Use biomedical and anatomical naming conventions in scholarly literature; avoid relying solely on typography.

Output format: Return ONLY a JSON array of strings.

Cardinality: The array length must be exactly 3. Do not add or drop elements.

Ordering: Preserve input order exactly (position i in output corresponds to input row i).

De-duplication policy: Do not merge or deduplicate rows; output one decision per input row.

Each element must be exactly 'ABBREV' or 'FULL' (uppercase).

Row independence: Judge each row independently without cross-reference.

Conventional term policy: Treat standardized names that are commonly written in their short canonical form in the literature as FULL; do not fabricate expansions when none is universally used.

Ambiguity policy: If the string denotes a descriptive biological/anatomical name, choose 'FULL'; if it is a short-form label that is normally expanded to a specific established full name, choose 'ABBREV'.

Policy: Ignore gene/protein names when making the judgment.

Constraints: No markdown, no explanations, no extra keys or text; return only valid JSON.

Fallback: If uncertain, still output either 'ABBREV' or 'FULL' for that row.

Input:

[{"label":"LSCs","pmcid":"PMC10867020"},{"label":"blast cells","pmcid":"PMC10867020"},{"label":"HSCs","pmcid":"PMC10867020"}]

Stage 2 — Resolution Prompt (submitted individually per ABBREV term):

Task: For each item, expand 'label' (abbreviation) to its full name using the provided 'contexts' from the specified paper (pmcid). All contexts are provided at once.

Order: Scan 'contexts' in sequence. If none contains an explicit full-name expansion, infer using widely accepted biomedical knowledge.

Evidence policy: Provide 'source' as minimal sentence(s) from the selected context that justify the expansion; you may use '...' to omit unrelated parts and connect sentences. If no explicit expansion is found and you infer, leave 'source' as an empty string.

Ambiguity policy: Do not annotate cases where multiple plausible expansions exist and the contexts cannot disambiguate; mark as failed.

Output format: Return ONLY a JSON array of objects.

Cardinality: The array length must be exactly 1. Do not add or drop elements.

Ordering: Preserve input order exactly (position i in output corresponds to input item i).

De-duplication policy: Do not merge or deduplicate items; output one object per input item.

Each object must have exactly these keys: 'full_name', 'source', 'status'.

Allowed status values: 'paper sourced', 'knowledge based (high confidence)', 'knowledge based (low confidence)', 'failed'.

Row independence: Judge each item independently.

Policy: Ignore gene/protein names during expansion; do not expand gene/protein-only labels.

Constraints: No markdown, no explanations, no extra keys or text; return only valid JSON.

Input:

[{"pmcid":"PMC10867020","label":"LSCs","contexts":["patients harbour a population of quiescent leukemic stem cells (LSCs) which can emerge from quiescence to trigger relapse after therapy.","This blast population is replenished by rare leukemia initiating cells, called leukemic stem cells (LSCs).","Similar to healthy hematopoietic stem cells (HSCs), LSCs are generally quiescent and are therefore thought to be responsible for relapse following chemotherapy which targets rapidly proliferating cells."]}]

3. Example Output

Stage 1 output (JSON array, length = 3, preserving input order):

["ABBREV", "FULL", "ABBREV"]

Stage 2 output for LSCs (JSON array, length = 1):

[
  {
    "full_name": "leukemic stem cells",
    "source": "patients harbour a population of quiescent leukemic stem cells (LSCs) which can emerge from quiescence to trigger relapse after therapy.",
    "status": "paper sourced"
  }
]

Supplementary Note 5

Standardization Agent: The Standardization Agent maps each raw cell type name, tissue term, and disease term into a corresponding controlled vocabulary entry from a target ontology. For each term, the agent first submits the raw label directly as a search query to the corresponding ontology API via the EBI OLS4 interface (Cell Ontology for cell types, Uberon Ontology for tissues, Disease Ontology for diseases), retrieving up to 30 candidate terms. A large language model then selects the single best-matching candidate by index, guided by rules that preserve the granularity and essential qualifiers of the original label and permit tissue context only when strictly necessary to disambiguate. If the initial query returns no candidates, the agent falls back to first prompting the LLM to generate a concise natural-language biological description of the term (≤ 35 words), submitting that description as a new OLS4 query, and repeating the candidate-selection step. Each term is resolved to a standardized ontology label, an OBO ID, and a mapping status (pass or failed).

1. Example Input

Each term submitted to the Standardization Agent carries the following fields:

Field	Example Value
cell_name_full	CD8+ cytotoxic T cell
tissue_class_uberon	blood
tissue_type_uberon	peripheral blood
Ontology target	Cell Ontology (CL)

The OLS4 API is queried with "CD8+ cytotoxic T cell" as the search string and returns the following candidate list (up to 30; truncated here for brevity):

Index	OLS4 candidate label
1	cytotoxic T cell
2	CD8-positive, alpha-beta cytotoxic T cell
3	CD8-positive T cell
4	effector CD8-positive, alpha-beta T cell

2. Constructed Prompt

You are a precise ontology assistant. Reply with ONLY minified JSON.

{

"instruction": "Select the single best Cell Ontology (CL) candidate that is strictly equivalent to the original label under the rules.",

```
  "rules": [
    "Base the decision on the original label; use tissue context only when strictly necessary to disambiguate.",
    "Preserve essential qualifiers, markers, and granularity in the original label; do not broaden or narrow.",
    "Choose only from the provided candidates by their index.",
    "If no candidate satisfies all constraints, return null.",
    "Output only minified JSON: {\"best_index\": integer|null}"
  ],
  "original_label": "CD8+ cytotoxic T cell",
  "context": {"tissue_class": "blood", "tissue_type": "peripheral blood"},
  "candidates": [
    {"index": 1, "label": "cytotoxic T cell"},
    {"index": 2, "label": "CD8-positive, alpha-beta cytotoxic T cell"},
    {"index": 3, "label": "CD8-positive T cell"},
    {"index": 4, "label": "effector CD8-positive, alpha-beta T cell"}
  ]
}
```

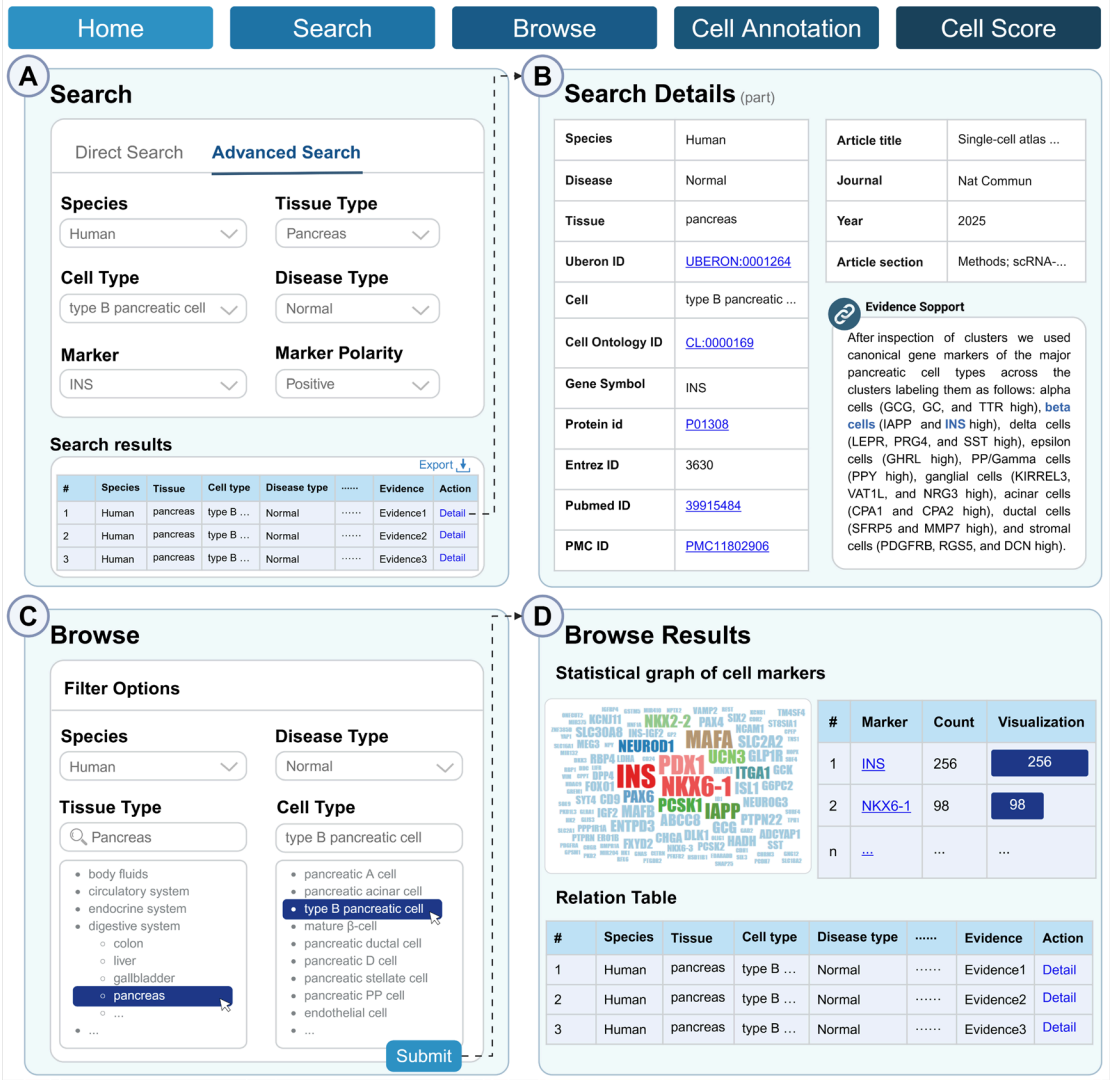

3. Example Output


```
{"best_index": 2}
```

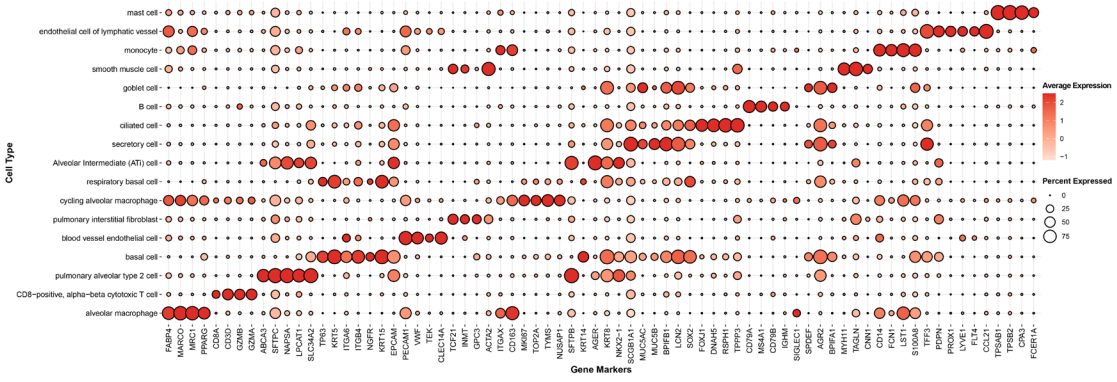

Resolved result: label = CD8-positive, alpha-beta cytotoxic T cell, OBO ID = CL:0001049, status = pass.

Candidate 2 is selected because it exactly preserves both the CD8 marker specification and the cytotoxic functional qualifier present in the original label. Candidate 1 (cytotoxic T cell) is rejected for being too broad (lacks CD8 specificity). Candidate 3 (CD8-positive T cell) is rejected for lacking the cytotoxic qualifier. Candidate 4 (effector CD8-positive, alpha-beta T cell) is rejected because "effector" is an additional qualifier not present in the original label and would over-specify the mapping.

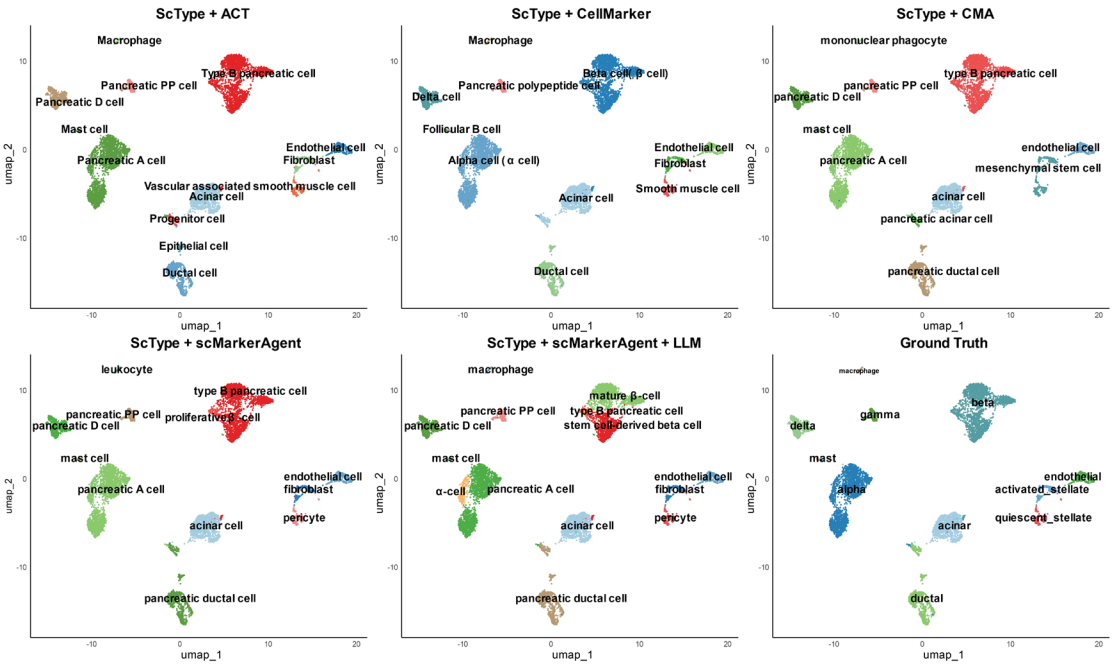
Supplementary Figures



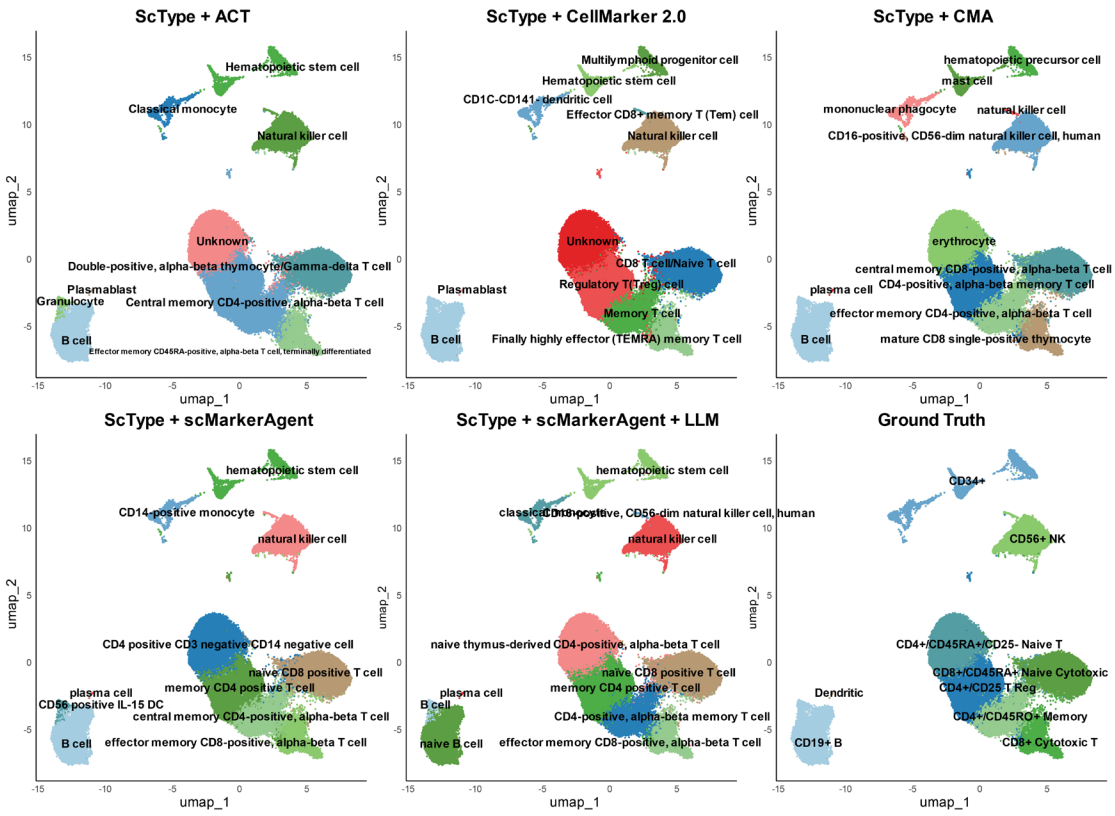
Supplementary Fig. 1. The Search and Browse pages of scMarkerAgent. **(A)** On the Search page, users can perform a Direct Search or an Advanced Search with six conditions; the search results are displayed in a table. **(B)** For each result, detailed information for the corresponding cell type–marker annotation can be viewed. **(C)** On the Browse page, users can browse cell lists by different tissue types. **(D)** The browse results page shows the occurrence frequency of markers and provides the corresponding annotation details. Created in BioRender. Chen, H. (2026) https://BioRender.com/nt1rexl.



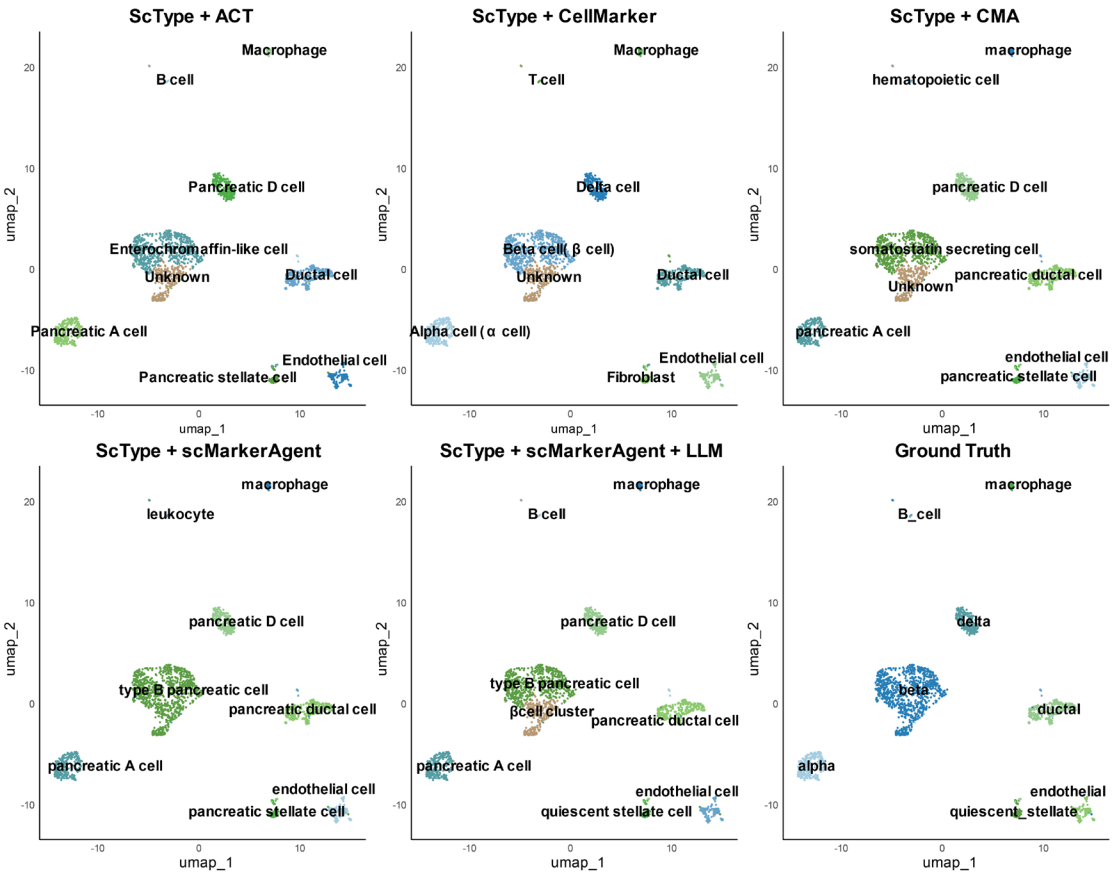
Supplementary Fig. 2. Dot plot of marker gene expression for cell types annotated by ScType + scMarkerAgent + LLM.



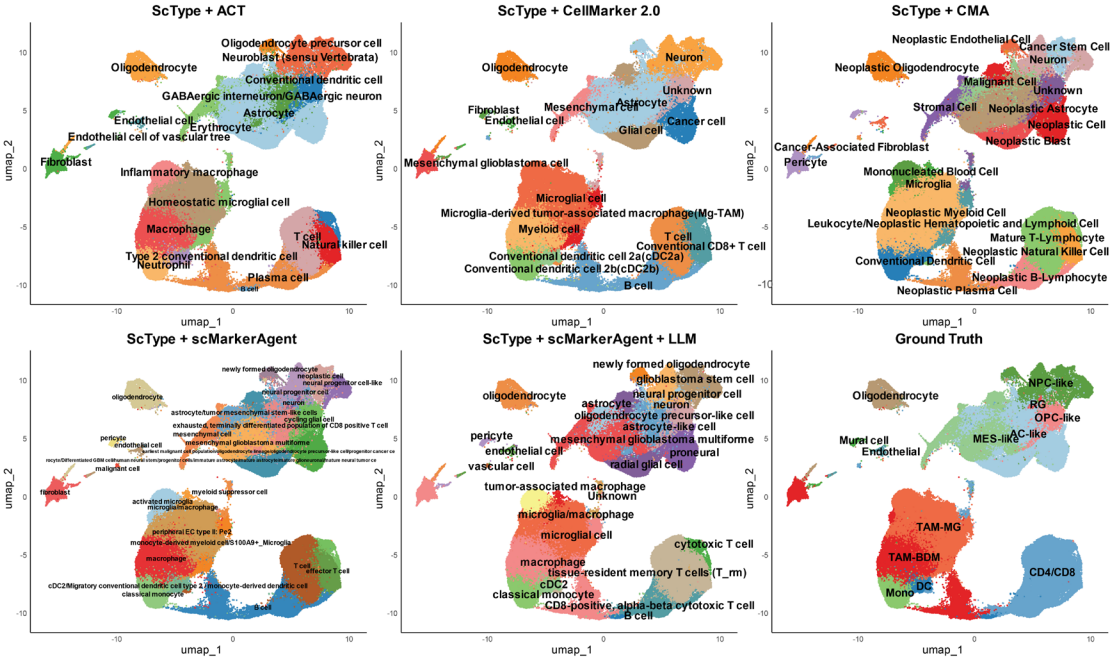
Supplementary Fig. 3: Cell-type annotations of the human pancreas dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.



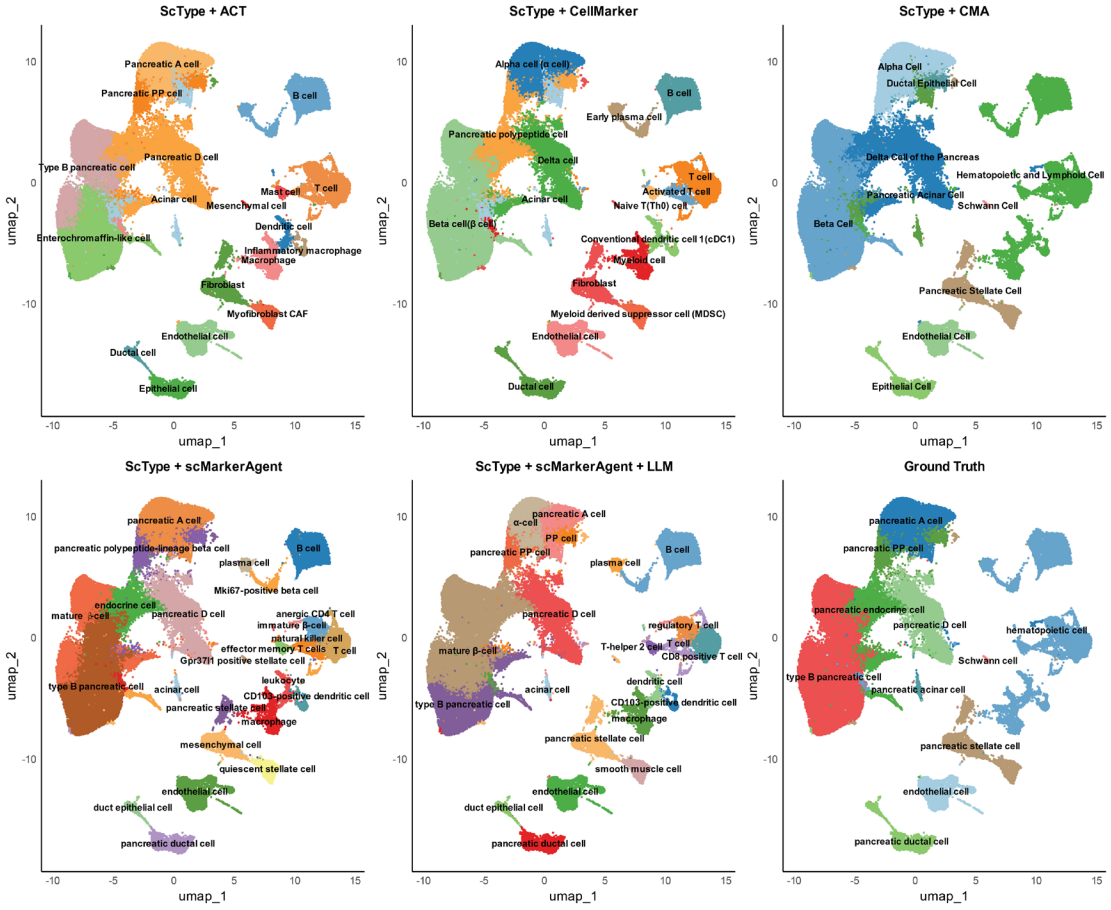
Supplementary Fig. 4: Cell-type annotations of the human PBMC dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.



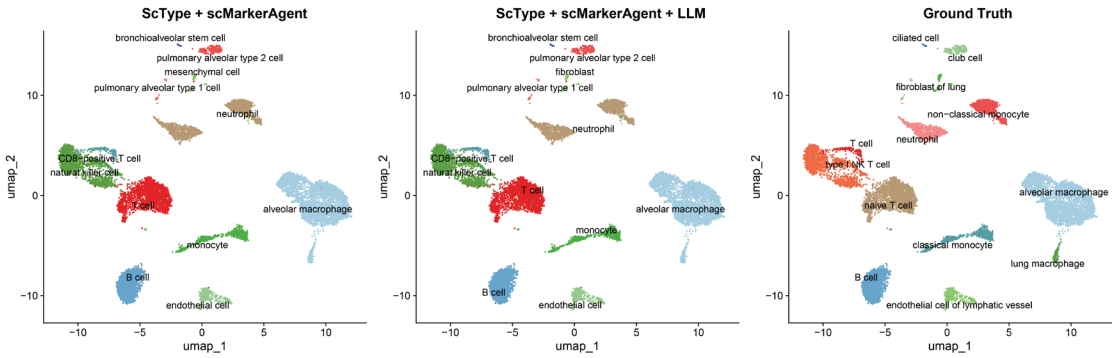
Supplementary Fig. 5: Cell-type annotations of the mouse pancreas dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.



Supplementary Fig. 6: Cell-type annotations of the human glioblastoma dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.



Supplementary Fig. 7: Cell-type annotations of the mouse type 2 diabetes dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.



Supplementary Fig. 8: Cell-type annotations of the rat lung dataset generated by ScType using markers from ACT, CellMarker 2.0, CMA, or scMarkerAgent, and by ScType + scMarkerAgent + LLM. Ground truth indicates the reference annotations provided by the original authors or by expert consensus.