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.

Cell IO: a database linking single-cell transcriptomes to literature-derived knowledge of cell functions and perturbations

Haruka Ozaki

ai-biology-inquiry@ml.riken.jp

RIKEN Center for Biosystems Dynamics Research https://orcid.org/0000-0002-1606-2762

Research Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11558063/v1

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

Additional Declarations: The authors declare no competing interests.

Cell IO: a database linking single-cell transcriptomes to literature-derived knowledge of cell functions and perturbations

Hirofumi Kariyayama[1], Haruto Ijiri[1,2], Ryosuke Matsuzawa[1,2], Yuya Tahara-Arai[1,2], Haruka Ozaki[1,2,*]

[1]Laboratory for AI Biology, RIKEN Center for Biosystems Dynamics Research (BDR), Japan
[2]University of Tsukuba, Japan
[*]To whom correspondence should be addressed. Email: ai-biology-inquiry@ml.riken.jp

Abstract

Single-cell transcriptome databases tell us which genes a cell type expresses, but what a cell type does, and what happens to it when a gene is perturbed, remains scattered across the literature. Cell IO (`https://cell-io.riken.jp/`) links these two kinds of knowledge through Cell Ontology identifiers. Its transcriptome layer re-processes 10x Chromium single-cell and single-nucleus RNA-seq experiments of human, mouse, and rat from the Sequence Read Archive with a uniform pipeline, annotates clusters with Cell Ontology terms restricted by the tissue of origin, and stores the cell type composition and UMAP embedding of every experiment (currently 14,931 experiments and 82.9 million cells). Its literature layer holds 895,258 claims extracted with a large language model and verified in a second pass from open-access abstracts, review articles, and gene perturbation studies in immunology and neuroscience, each with the verbatim sentence, highlighted entity spans, evidence strength, and ontology-normalized cell, gene, and context. The web interface lets users search cell types by name, synonym, or identifier, find the experiments in which a cell type is most abundant, inspect the composition and embedding of each experiment, read and filter the claims about a cell type by knowledge base, direction, evidence strength, and organism, and identify cell types whose functional characterization is missing from part of the literature. Cell IO is freely available without registration.

1 Introduction

Single-cell RNA sequencing has made the transcriptome of individual cells a routine measurement, and public repositories now hold tens of thousands of experiments and hundreds of millions of cells. Several resources make these data browsable. Portals such as CZ CELLxGENE Discover [1] and the Single Cell Portal [2] collect processed matrices and annotations submitted by authors, and CELLxGENE requires cell type and tissue annotations in the Cell Ontology and Uberon so that collections can be queried together. Other projects re-process raw reads from the Sequence Read Archive with one pipeline so that datasets become comparable: the Single Cell Expression Atlas [3], PanglaoDB [4], DISCO [5], and most recently scBaseCount [6], which uses language-model agents to curate SRA metadata at the scale of hundreds of millions of cells. Uniform re-processing has a longer history in bulk transcriptomics, where recount [7] and ARCHS4 [8] showed that consistent quantification and curated metadata turn a repository into a queryable resource.

What these resources answer is which genes a cell expresses. What a biologist usually wants to know next is what the cell does: which cells it activates or kills, which molecules it secretes, how it behaves in a disease, and how it changes when a gene is deleted or overexpressed. This knowledge is abundant but scattered through the literature. Cell-centered knowledge bases capture parts of it: the Cell Ontology [9, 10] defines cell types and their relations, marker databases

Table 1: **Comparison of Cell IO with related resources.** Re-proc., uniform re-processing from raw sequencing reads; CL, cell type annotation with Cell Ontology identifiers; Tissue ont., tissue annotation with an anatomy ontology; Lit. claims, literature-derived statements about cell functions; Perturb., gene perturbation outcomes; Cross-link, navigation between expression data and literature statements for the same cell type. Parentheses qualify a partial match: Single Cell Portal collects Cell Ontology and Uberon identifiers through an optional metadata convention [20]; scBaseCount predicts cell types with a classifier for human cells only; DISCO labels cell types with its own hierarchical vocabulary built with the Cell Ontology as a guideline; PanglaoDB and CellMarker 2.0 curate marker genes, not functional statements, from the literature; PubTator 3.0 annotates genes, diseases, chemicals, variants, species, and cell lines but not cell types; scPerturb harmonizes author-provided count matrices.

Resource	Re-proc.	CL	Tissue ont.	Lit. claims	Perturb.	Cross-link
CZ CELLxGENE Discover [1]	–	✓	✓	–	–	–
Single Cell Portal [2]	–	(convention)	(convention)	–	–	–
Single Cell Expression Atlas [3]	✓	✓	✓	–	–	–
PanglaoDB [4]	✓	–	–	(markers)	–	–
DISCO [5]	✓	(own vocabulary)	–	–	–	–
scBaseCount [6]	✓	(predicted)	✓	–	–	–
CellMarker 2.0 [11]	–	✓	✓	(markers)	–	–
PubTator 3.0 [14]	–	–	–	(entities, relations; no cell types)	–	–
scPerturb [16]	–	–	–	–	(datasets)	–
Cell IO (this work)	✓	✓	✓	✓	✓	✓

such as CellMarker [11] and PanglaoDB [4] link cell types to genes, and curated interaction databases such as CellPhoneDB [12] and CellTalkDB [13] link cell types through ligand–receptor pairs. Text-mining resources such as PubTator [14] and SemMedDB [15] recognize entities and relations across all of PubMed, but they are not organized around cell types and do not represent what a cell does as a claim. Repositories of perturbation experiments such as scPerturb [16] and PerturBase [17] collect perturbation datasets, but the outcomes of gene perturbations reported in the literature for a given cell type are, to our knowledge, not systematically collected in any resource (Table 1). Large language models have recently made it feasible to extract structured statements from biomedical text at scale with controlled precision [18, 19], which opens the possibility of building a knowledge base of what cells do that is large enough to be matched against single-cell data.

Here we describe Cell IO (`https://cell-io.riken.jp/`), a database that links cell types in uniformly re-processed single-cell transcriptome data to literature-derived statements about their functions and about the outcomes of gene perturbations (Figure 1). Cell IO re-processes 10x Chromium experiments of human, mouse, and rat from the Sequence Read Archive, annotates clusters with Cell Ontology terms restricted by the tissue of origin, and stores the composition and embedding of every experiment. In parallel it extracts, with a large language model and a second verification pass, three knowledge bases of claims from open-access abstracts, review articles, and gene perturbation studies in immunology and neuroscience, and normalizes the cells, genes, and contexts they mention to ontologies. Because both layers use Cell Ontology identifiers, a user can move from a cell type to the datasets in which it is abundant and to the sentences that describe its functions, and back. We describe the construction of the database, its current content, the web interface, and representative use cases, and we discuss the limitations of automated extraction and annotation and the planned extensions.

2 Materials and Methods

An overview of the construction workflow is given in Figure 2. Version numbers of the tools and resources are listed in Supplementary Table S1.

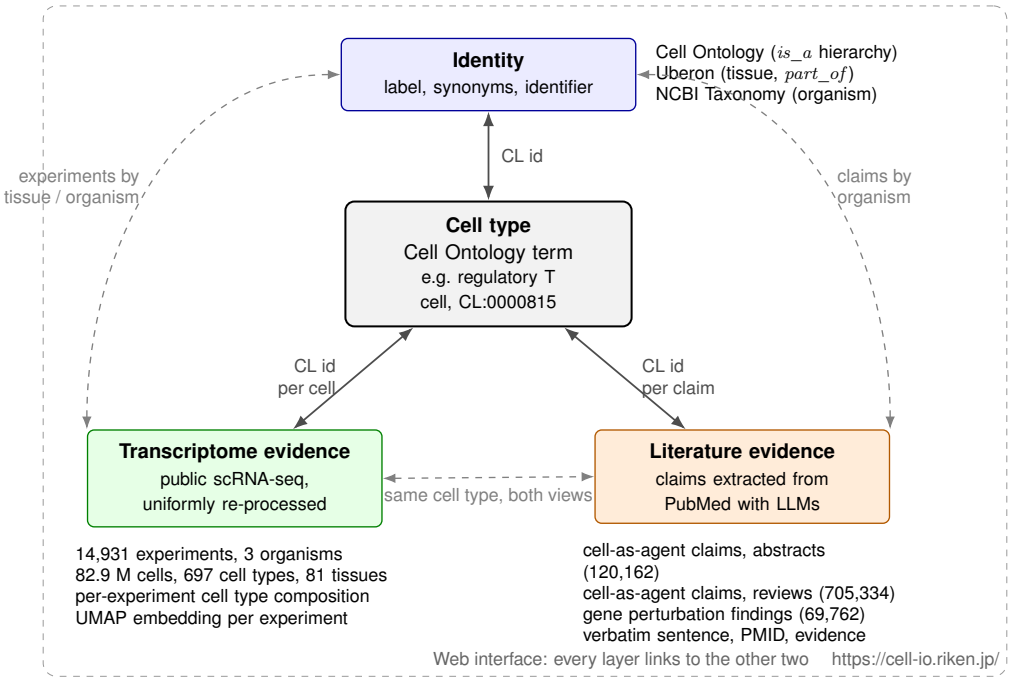


Figure 1: **Concept of Cell IO.** Cell IO organizes evidence around cell types defined in the Cell Ontology. Each cell type is connected to three layers: its position in the Cell Ontology, Uberon, and NCBI Taxonomy hierarchies (identity); the public single-cell RNA-seq experiments in which it was identified, with per-experiment composition and embedding (transcriptome evidence); and the claims about it extracted from the literature by three knowledge bases (literature evidence). The layers are joined by Cell Ontology identifiers, and in the web interface every layer links to the other two. Counts refer to the transcriptome release of 2026-10-29 and the literature release of 2026-10-06.

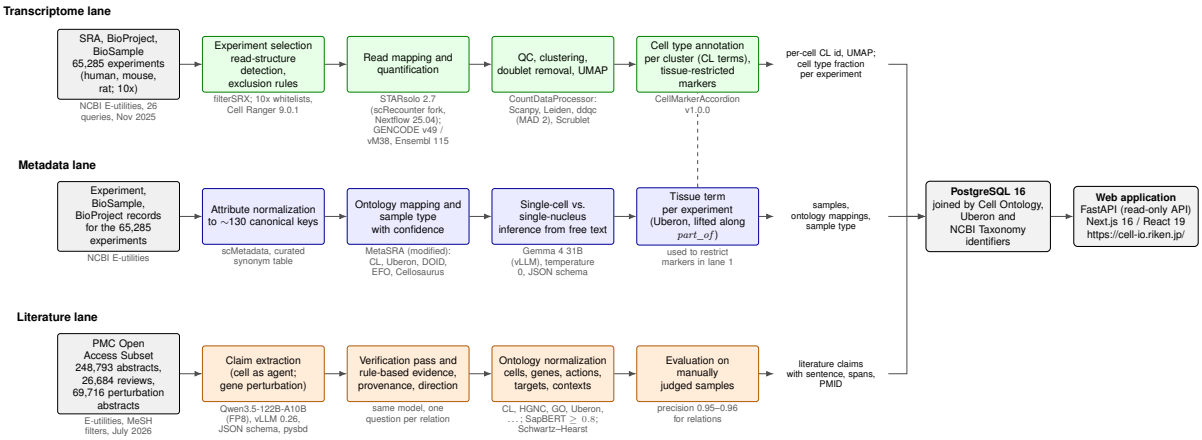


Figure 2: **Construction workflow of Cell IO.** Transcriptome lane: 65,285 candidate experiments were collected from the Sequence Read Archive with the NCBI E-utilities, selected by read-structure detection and exclusion rules, mapped from raw reads with STARsolo (STAR 2.7, in a fork of the scRecounter workflow run with Nextflow 25.04; GENCODE v49 and vM38, Ensembl release 115), processed per experiment with Scanpy (Leiden clustering, ddqc quality control, Scrublet doublet removal, UMAP), and annotated per cluster with Cell Ontology terms from the CellMarkerAccordion database (v1.0.0), restricted to the markers of the sample's tissue. Metadata lane: Experiment, BioSample, and BioProject records were normalized to about 130 attribute keys, mapped to ontologies with a modified MetaSRA pipeline (Cell Ontology, Uberon, Disease Ontology, Experimental Factor Ontology, Cellosaurus) that also predicts a sample type with a confidence score, and the single-cell versus single-nucleus label was inferred with Gemma 4 31B served by vLLM; the Uberon term of each experiment restricts the markers used in the transcriptome lane (dashed arrow). Literature lane: 248,793 abstracts, 26,684 review articles, and 69,716 perturbation-related abstracts from the PubMed Central Open Access Subset were processed with Qwen3.5-122B-A10B served by vLLM 0.26 under a JSON schema, verified in a second pass, annotated by rules for evidence strength, provenance, and direction, normalized to ontologies (Cell Ontology, HGNC, Gene Ontology, Uberon, and others, with SapBERT nearest neighbors at cosine similarity of at least 0.8 and Schwartz–Hearst abbreviation expansion), and evaluated on manually judged samples. The three lanes are joined by shared Cell Ontology, Uberon, and NCBI Taxonomy identifiers in one PostgreSQL 16 database that serves a read-only FastAPI interface and a Next.js 16 / React 19 frontend.

2.1 Transcriptome data collection and processing

2.1.1 Selection of experiments

Candidate experiments were collected from the Sequence Read Archive (SRA) through the NCBI E-utilities [21] with 26 queries, each restricted to *Homo sapiens*, *Mus musculus*, or *Rattus norvegicus*, to the terms "10x Genomics" or "10x Chromium", and to "RNA-seq", combined with one of 25 nervous-system keywords or the keyword "immune". Records submitted to ENA or DDBJ were included through the INSDC mirror at NCBI. The queries, run in November 2025, returned 65,285 experiments (211,793 runs, 57,148 BioSamples, 2,824 BioProjects). Runs were split by organism and library layout, runs larger than 300 GB and reads shorter than 26 bp were dropped, and an experiment entered the mapping stage only if a 10x barcode whitelist and read structure could be identified from a sample of one million reads (at least 30% of reads with a valid barcode). After processing, experiments whose SRA or BioSample text indicated non-expression libraries were excluded by rule: V(D)J, antibody- or hashtag-derived tag libraries, Perturb-seq and CROP-seq, AAV-based screens, MULTI-seq and CellPlex, pooled libraries, and suppressed records. 10x Multiome gene expression libraries were kept.

2.1.2 Read mapping and quantification

Reads were downloaded with the SRA Toolkit and mapped with STARsolo [22, 23] inside a fork of the scRecounter workflow of the scBaseCamp project [6], orchestrated with Nextflow [24]. STARsolo was run with Cell Ranger-compatible settings (CB_UMI_Simple, CellRanger4 adapter clipping, 1MM_CR UMI deduplication, EmptyDrops_CR cell calling [25]) and the GeneFull_Ex50pAS feature set, which counts intronic reads and is closest to Cell Ranger 7 and later. References were GENCODE v49 on GRCh38 for human, GENCODE vM38 on GRCm39 for mouse [26], and GRCr8 with Ensembl release 115 annotation for rat [27]. Barcode whitelists were taken from Cell Ranger 9.0.1. Mapping ran on eight virtual machines of the RIKEN HOKUSAI Sailing Ship system sharing a Lustre file system. A first pass that inadvertently capped the number of reads per run was discarded, and the whole corpus was re-mapped from the full reads between July and October 2026.

2.1.3 Quality control, clustering, and cell type annotation

Each experiment was processed independently with Scanpy [28] in a workflow we call CountDataProcessor. Counts were normalized to 10,000 per cell and log-transformed, 2,000 highly variable genes were selected, and 50 principal components were computed. Cells were clustered with the Leiden algorithm [29] at the default resolution on a 10-nearest-neighbor graph over 30 components. Low-quality cells were removed cluster-wise with a data-driven procedure (ddqc [30]) using a median absolute deviation threshold of 2 on counts, detected genes, and the fractions of mitochondrial (upper bound 10%) and ribosomal reads. Doublets were removed with Scrublet [31]. The retained cells were re-clustered and embedded with UMAP [32]. Clusters were annotated with the CellMarkerAccordion marker database (v1.0.0) [33], scoring each candidate Cell Ontology term by the weighted mean expression of its positive markers and assigning the best-scoring term to the cluster. Candidate terms were restricted to the markers recorded for the tissue of the sample, identified by the Uberon term derived from its metadata (see below); samples without a tissue term were scored against all tissues of the species. Clusters without applicable markers were labelled "Unknown". Rat markers were derived from mouse markers through Ensembl orthology. For every experiment the pipeline stores the Cell Ontology identifier of each cell, the UMAP coordinates, and the number and fraction of cells per Cell Ontology term. Pipeline runs are versioned per experiment, and the web application always reads the most recent run.

2.2 Transcriptome metadata curation

Run information, Experiment packages, BioSample records, and BioProject summaries were retrieved for all 65,285 experiments through the E-utilities [21]. BioSample attribute names, which vary freely between submitters, were normalized to about 130 canonical keys with a curated synonym table (for example "tissue source" and "source" to `source_name`) before any further processing. The normalized attributes were mapped to ontology terms with the MetaSRA pipeline [34], using the Cell Ontology [9, 10], Uberon [35], the Disease Ontology [36], the Experimental Factor Ontology [37], and Cellosaurus [38], and MetaSRA's sample type classifier assigned one of six categories (tissue, primary cells, cell line, in vitro differentiated cells, stem cells, induced pluripotent stem cells) with a confidence score to each sample. We modified MetaSRA to make tie-breaking between source attributes deterministic and to tolerate failures in batch mode. The tissue used to restrict marker-based annotation was chosen per experiment as the Uberon term, or failing that the Cell Ontology term, mapped by MetaSRA, lifted along Uberon partonomy relations to the nearest tissue present in the CellMarkerAccordion database. Whether an experiment profiled whole cells or nuclei was inferred from the Experiment title, sample title, design description, library construction protocol, and sample attributes with the instruction-tuned Gemma 4 31B model [39] served locally with vLLM [40], at temperature 0 and with the output constrained to a fixed JSON schema. On 200 manually labelled experiments (100 human, 100 mouse) the inferred label agreed with the manual label in 198 cases. The label is shown on each experiment page together with the 10x chemistry version identified during experiment selection.

2.3 Literature data collection and processing

2.3.1 Corpora

Three corpora were built from the PubMed Central Open Access Subset through the E-utilities in July 2026. The abstract corpus comprises 248,793 original research articles indexed with the MeSH term "Cells", with "Humans", "Mice", or "Rats", and with MeSH terms of the immune system (157,376 articles) or the nervous system (105,692 articles), after excluding reviews, case reports, clinical trials, meta-analyses, letters, and editorials. The review corpus comprises the 26,684 review articles excluded by that step; full-text JATS XML was available for 26,267 of them and yielded 1,005,106 body paragraphs after removing non-body sections, citation markers, and duplicates. The perturbation corpus is the subset of 69,716 abstracts (28.0%) of the abstract corpus that contain a regular-expression cue for gene deletion, knockdown, overexpression, mutation, or editing.

2.3.2 Extraction of claims

All language model stages used Qwen3.5-122B-A10B (FP8; the Qwen3 series [41]) served with vLLM [40] on two NVIDIA GB200 GPUs of the RIKEN RIKYU system, at temperature 0, with JSON-schema constrained decoding and without fine-tuning. Each abstract or paragraph was first classified for the presence of a statement in which a cell type acts on something ("cell as agent"), under a written definition shared by all prompts. Positive units were split into sentences with pysbd [42] and shown to the model with sentence identifiers. The model returned, per relation, the sentence identifier, the subject cell, the action, the target, and optional context, condition, and organism, together with an expanded sentence in which anaphora and abbreviations are resolved by insertion only. Spans were verified mechanically to occur verbatim in the sentence, and relations failing this check were discarded. For the perturbation corpus the model returned findings with the perturbed gene, the kind of perturbation, the readout cell type, the outcome, and its direction (increase, decrease, no change, change, or context-dependent); findings without a named cell type were discarded. A second, independent pass asked one yes-or-no

question per relation to reject statements about the cell's own properties, expression, or location, statements in which the subject is not a cell or not the agent, and mere associations, and recorded an enumerated reason for every rejection. Rule-based annotation then assigned to each claim an evidence strength (asserted, suggested, speculative), flags for negated, correlational, and attributed statements, and a provenance (own result, prior knowledge, unclear). Actions were further grouped into a taxonomy of 13 major and 49 minor concept classes, and targets and outcomes into a shared taxonomy of 11 major and 51 minor classes, by the same model.

2.3.3 Ontology normalization

Cell mentions were mapped to the Cell Ontology and, where no term exists, to Cellosaurus, BTO, the Gene Ontology cellular component branch, MeSH, or NCIt; actions to Gene Ontology biological processes; contexts to Uberon, MONDO, and related vocabularies; genes to HGNC [43]; and targets and outcomes to a wider set of vocabularies. Resolution proceeded by lexical matching with variant generation, abbreviation expansion with the Schwartz–Hearst algorithm [44] supplemented by model-derived and corpus-derived long forms, molecule detection, nearest-neighbor matching with SapBERT embeddings [45] at cosine similarity of at least 0.8, and phrase decomposition. Non-exact matches were confirmed or rejected by the language model in a one-question-per-pair format. Rows were never dropped by normalization; unmapped fields carry a null identifier.

2.3.4 Evaluation

Precision was estimated on manually judged samples. On 200 abstracts, classification reached precision 0.91 and recall 0.77, sentence-level extraction precision 0.90 and recall 0.52, and relation precision 0.96 after verification (0.72 before). On 100 review paragraphs relation precision was 0.95 after verification. Ontology mapping precision was 0.87 on 118 field–string pairs, concept classification accuracy 0.92 (lenient) or 0.74 (strict) on 90 items, and evidence strength 0.82 on 28 items. These gold sets were judged by a single annotator, one of the authors (H.O.); the current knowledge bases have not been reviewed by a panel of domain experts (see Discussion).

2.4 Data integration through ontology

The transcriptome and literature layers are joined by shared ontology identifiers rather than by foreign keys. Cell Ontology identifiers connect per-experiment cell type fractions to literature claims by exact match. Uberon identifiers connect experiments to tissues, and NCBI Taxonomy identifiers connect experiments to organisms, which are matched to the literature's organism terms "Humans", "Mice", and "Rats". Labels, synonyms, and *is_a* relations of the Cell Ontology (release 2026-06-08; 3,335 terms) and labels, synonyms, *is_a* and *part_of* relations of Uberon (release 2026-06-19; 14,975 terms) were loaded from the OBO releases to support search by synonym and navigation of the hierarchies. Hierarchical relations are used for navigation and are not used to aggregate evidence to ancestor terms (see Discussion).

2.5 Frontend

Data are stored in PostgreSQL 16. A read-only API implemented with FastAPI (Python 3.12) serves a Next.js 16 and React 19 frontend behind an nginx reverse proxy; all services run in Docker Compose on an Amazon EC2 instance and are exposed through a Cloudflare tunnel. Aggregates for the literature browser are precomputed as materialized views. The UMAP viewer uses Plotly.js.

Table 2: **Summary of the content of Cell IO.** Transcriptome counts refer to the release of 2026-10-29 (live since Cell IO v0.9.1) and literature counts to the knowledge-base release of 2026-10-06; values from figures/data/content_statistics.json.

Layer	Item	Count
Transcriptome	Experiments (SRX)	14,931
	Cells	82,852,125
	Organisms	3
	Tissues (Uberon terms)	81
	Cell types (Cell Ontology terms)	697
Literature	Claims, cell as agent (abstracts)	120,162
	Claims, cell as agent (reviews)	705,334
	Claims, gene perturbation	69,762
	Publications with ≥1 claim (abstracts / reviews / perturbation)	49,724 / 22,638 / 27,382
	Distinct Cell Ontology terms (any knowledge base)	1,783
	Distinct identifiers, all vocabularies	3,615
Both layers	Cell Ontology terms with transcriptome and literature evidence	585

3 Database content

3.1 Overview

Cell IO (`https://cell-io.riken.jp/`) organizes three kinds of evidence about cell types: their position in reference ontologies, their occurrence and abundance in uniformly re-processed public single-cell RNA-seq experiments, and the statements made about them in the literature (Figure 1). All three are keyed by Cell Ontology identifiers, so that a cell type page, an experiment page, and a literature claim refer to the same term. Table 2 summarizes the content.

3.2 Transcriptome layer

Of the 65,285 candidate experiments, 16,850 were mapped, quality-controlled, and annotated, and 1,919 of these were then excluded by the metadata rules, leaving 14,931 experiments. The transcriptome layer therefore holds 14,931 experiments from human, mouse, and rat, each a 10x Chromium single-cell or single-nucleus RNA-seq library deposited in SRA and re-processed from raw reads, comprising 82,852,125 quality-controlled cells (release of 2026-10-29; Figure 3A–C). Of these, 6,757 experiments are human (33,810,355 cells), 7,919 mouse (47,790,054 cells), and 255 rat (1,251,716 cells), and an experiment contains a median of 4,587 cells. Every experiment carries one of 81 distinct Uberon tissue terms, the tissue whose marker set was used for annotation; it is the MetaSRA tissue when the sample metadata yielded one, and otherwise the tissue whose markers scored best across all tissues (Figure 3B). The most frequent tissues are brain (4,432 experiments), blood (1,833), lung (1,276), and pancreas (1,034). Across experiments, 697 distinct Cell Ontology terms were assigned to clusters (573 in human, 455 in mouse, and 242 in rat), with a median of 13 terms per experiment (Figure 3D). For each experiment the database stores the number and fraction of cells per term, which drives the composition views and the cell-type-centric sorting of the experiment browser, and the per-cell UMAP coordinates and labels, which drive the embedding viewer. A MetaSRA sample type is available for 14,613 of the 14,931 experiments: 7,567 are classified as tissue, 4,170 as primary cells, 2,202 as cell lines, and 674 as stem cells, induced pluripotent stem cells, or cells differentiated in vitro (Figure 3E).

3.3 Literature layer

The literature layer holds 895,258 claims from three knowledge bases: 120,162 claims from 49,724 original research abstracts (cell as agent, abstracts), 705,334 claims from 22,638 review articles (cell as agent, reviews), and 69,762 gene perturbation findings from 27,382 abstracts (Figure 3F).

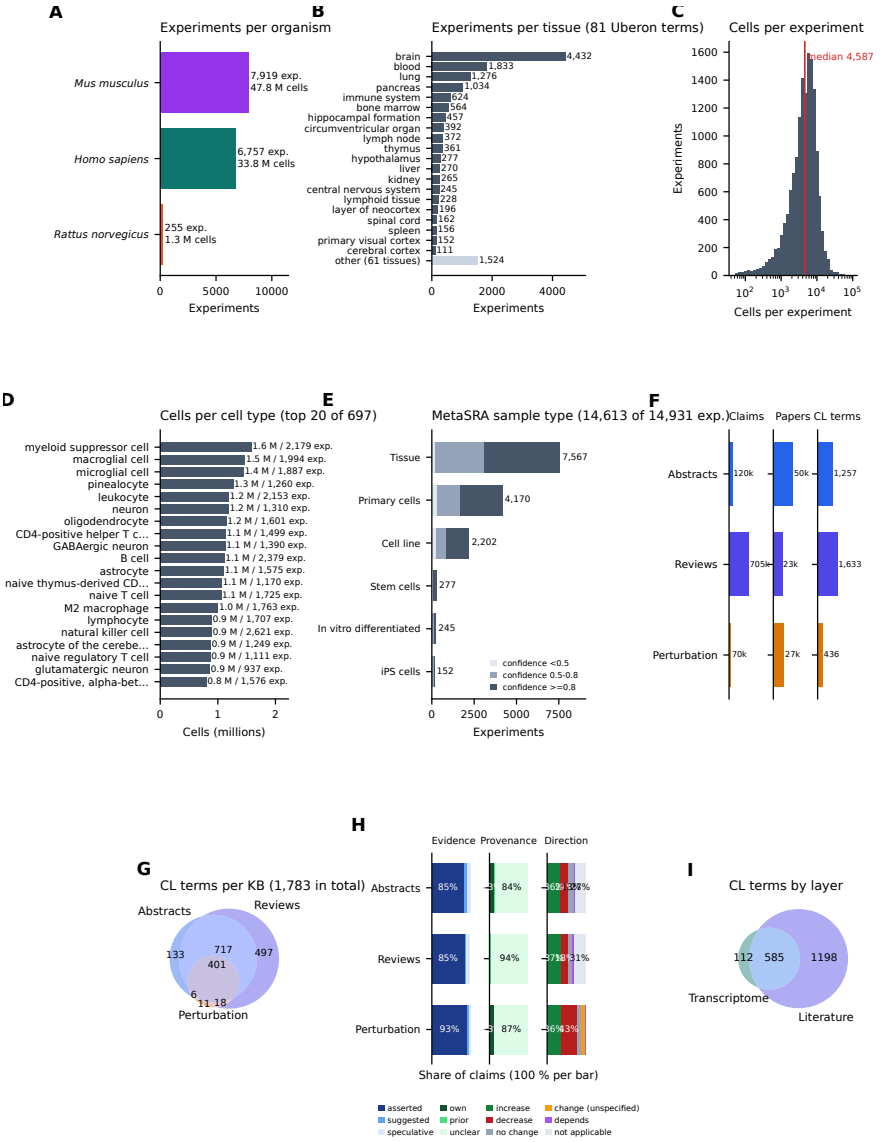


Figure 3: **Content of Cell IO (transcriptome release of 2026-10-29; literature release of 2026-10-06).** (**A**) Number of experiments per organism. (**B**) Number of experiments per tissue for the 20 most frequent Uberon terms; the remaining 61 terms are pooled. (**C**) Distribution of the number of cells per experiment (logarithmic axis); the red line marks the median of 4,587 cells. (**D**) Number of cells per cell type for the 20 most abundant Cell Ontology terms, out of 697 terms in total. (**E**) Number of experiments per MetaSRA sample type, stratified by the confidence of the prediction, for the 14,613 experiments with a sample type. (**F**) Size of the three literature knowledge bases in claims, distinct publications, and distinct Cell Ontology identifiers. (**G**) Overlap of Cell Ontology identifiers among the three knowledge bases (1,783 identifiers in total). (**H**) Distribution of evidence strength, provenance, and direction of claims within each knowledge base; for cell-as-agent claims the direction is the polarity of the extracted action, for perturbation findings the direction of the outcome. (**I**) Overlap between Cell Ontology terms that have transcriptome evidence and those that have literature evidence; the intersection (585 of 697 transcriptome terms) is the set of cell types for which both views can be cross-linked.

Every claim retains the verbatim sentence, character spans for the cell, action, and target or outcome, the source PMID, and the full output of the extraction pipeline. The subject cell is mapped to an ontology term in 76% of abstract claims and 85% of review claims; in 67% and 73% the term is a Cell Ontology term, and the remainder are cell lines, tissues, and disease-specific populations mapped to Cellosaurus, the BRENDA Tissue Ontology, MeSH, or NCIt. In perturbation findings the perturbed gene is mapped to an HGNC identifier in 78% of cases and the readout cell to a Cell Ontology term in 39%. In total 1,783 Cell Ontology terms carry at least one claim (3,615 distinct identifiers when every vocabulary is counted); 401 of them are covered by all three knowledge bases and 1,382 are missing from at least one (Figure 3G). The two cell-as-agent knowledge bases differ in character: an abstract yields on average 0.7 claims that mostly report the paper's own result, whereas a review yields about 27 claims that restate consensus knowledge, so their counts are reported separately throughout the interface and should not be summed as independent observations. Among perturbation findings, 50% describe gene deletion, 23% knockdown, 20% overexpression, and 5% mutation; the outcome decreased in 43% and increased in 36% of findings, and 9% report no change (Figure 3H). Claims are labelled as asserted in 85% of cell-as-agent claims and 93% of perturbation findings, suggested in 4–6%, and speculative in 3–10%.

3.4 Overlap between the layers

Of the 697 Cell Ontology terms with transcriptome evidence, 585 (84%) also carry at least one literature claim, 112 have transcriptome evidence only, and 1,198 terms have literature evidence only (Figure 3I). For these terms the cell type page shows both evidence summaries side by side, and the experiment browser and literature browser can be entered from the same page.

3.5 Web interface

Cell IO is delivered as a public, read-only web application (`https://cell-io.riken.jp/`) in which a cell type, tissue, organism, or experiment can be viewed from three complementary angles: its identity in a reference ontology, the single-cell transcriptome experiments in which it appears, and the claims about it extracted from the literature (Figure 4). Every page links each of these angles to the other two. A user who starts from an expression dataset can therefore reach the published knowledge about the cells it contains, and a user who starts from a literature claim can reach the experiments that measured those cells. The interface requires no login or installation. All state that defines a view, such as filters, sort order, and page, is encoded in the URL, so any result set can be bookmarked or shared as a link.

3.5.1 Entry points

The home page offers one search box and six entry panels. The search box has two targets. In *Cell type* mode it matches the names, synonyms, and Cell Ontology (CL) identifiers of the cell types represented in the database. After two characters it suggests up to six candidates and shows which synonym was matched, for example the broad synonym "Treg" for *regulatory T cell* (Figure 4A). A user who enters an exact CL identifier or selects a single candidate is taken directly to that cell type's page. In *Literature* mode the same box runs a keyword search over the literature claims. Below the search box, six panels report the current size of the database, namely the numbers of experiments, cell types, organisms, tissues, literature claims, and cells (14,931 experiments, 697 cell types, 3 organisms, 81 tissues, 895,258 claims, and 82,852,125 cells for the data generation described here). The first five panels double as links into the corresponding browsers. A persistent header gives one-click access to the experiment browser, the cell type list, the literature browser, and the About page.

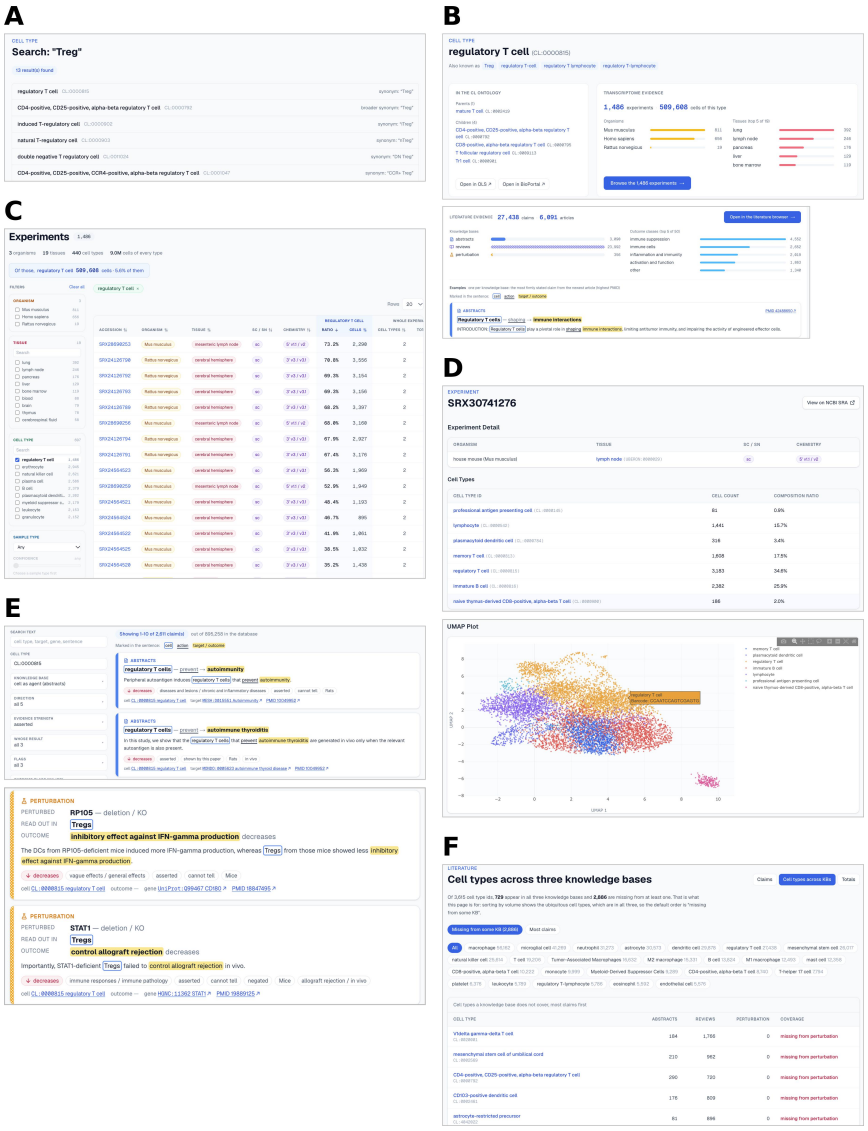


Figure 4: **The Cell IO web interface (v0.9.1, screenshots of 2026-10-30).** (**A**) Cell type search results for the query "Treg": 13 Cell Ontology terms matched through their labels or synonyms, each shown with its identifier and the synonym that matched. (**B**) Cell type page for regulatory T cell (CL:0000815): synonyms, parents and children in the Cell Ontology, Transcriptome Evidence with the numbers of experiments and cells and their breakdown by organism and tissue, and Literature Evidence with the number of claims per knowledge base, the most frequent outcome classes, and one example claim per knowledge base. (**C**) Experiment browser filtered to regulatory T cell and sorted by composition ratio: facet filters for organism, tissue, cell type, and sample type with its confidence threshold on the left; on the right the two cell-type-specific columns giving the fraction and number of cells of that type in each experiment. (**D**) Experiment page of SRX30741276 (mouse lymph node): organism, tissue, single-cell versus single-nucleus label and 10x chemistry; the cell type composition table; and the UMAP embedding colored by cell type with the hover tooltip showing the cell type and barcode of one cell. (**E**) Literature browser (Claims view) for regulatory T cell with the facet rail; two cell-as-agent claims from abstracts restricted to asserted evidence, with the cell, action, and target marked in the sentence, and two gene perturbation findings in the three-line layout of perturbed gene, readout cell, and outcome. (**F**) Cross view listing, for each cell type, the number of claims in the three knowledge bases and the knowledge base from which it is missing; the default order shows the cell types with the most claims that are absent from some knowledge base.

3.5.2 Browsing by cell type, tissue, and organism

Cell IO provides a list page and a detail page for each of three organizing dimensions: cell type, tissue, and organism. The list pages show, for every entity represented in the database, the number of experiments and the number of cells, and can be filtered by name or identifier and sorted by experiments, cells, or name.

The detail pages share a common three-part layout (Figure 4B). The first part states the entity's identity: label, identifier, and exact synonyms, drawn from the Cell Ontology for cell types, from Uberon for tissues, and from the NCBI Taxonomy for organisms, with outbound links to OLS4, BioPortal, or NCBI Taxonomy. The second part places the entity in its ontology. Cell type pages list the immediate parents and children in the CL *is_a* hierarchy. Tissue pages separate *Anatomy* (Uberon *part_of*: "part of" and "contains") from *Classification* (*is_a*: "type of" and "subtypes"), so that a lymph node is shown both as a type of secondary lymphoid organ and as a part of the lymphatic system. Each neighbor is itself a link, so the hierarchy can be walked without leaving the browser. The third part, *Transcriptome Evidence*, summarizes what the expression data say about the entity. It reports the number of experiments and cells together with horizontal-bar breakdowns of the other dimensions: for a cell type, the organisms and tissues in which it was observed; for a tissue, its top cell types by cell count and organisms by experiment count; for an organism, its tissues and cell types. A button labelled "Browse the N experiments" opens the experiment browser pre-filtered to that entity.

Cell type and organism pages add a fourth part, *Literature Evidence*, which summarizes the literature claims attached to the entity in the same "count, breakdown, link" form. It reports the number of claims and of distinct publications (PMIDs), the number of claims from each of the three knowledge bases including those with zero so that gaps in coverage are visible, the five most frequent outcome classes, and one example claim per knowledge base rendered as a compact claim card. A button opens the literature browser pre-filtered to the entity. Because literature claims carry the organism as a free-text term such as "Humans" or "Mice" rather than a taxonomy identifier, the organism page states explicitly which term was counted.

3.5.3 Browsing experiments

The experiment browser is the single table of experiments in Cell IO (Figure 4C). The cell type, tissue, and organism pages deliberately do not duplicate it and instead link into it. Each row is one SRA experiment (SRX accession) with its organism, tissue, number of cell types, and number of cells. Experiments are filtered through three facets, organism, tissue, and cell type, with the usual faceted-search semantics: values within one facet are combined with OR, facets are combined with AND, and each facet's counts are computed with only its own selection removed so that alternative values remain selectable. A fourth filter, *Sample type*, exposes the MetaSRA sample type prediction together with a confidence threshold, so that users can, for example, exclude or isolate cells differentiated from induced pluripotent or embryonic stem cells. The six values are tissue, primary cells, cell line, stem cells, in vitro differentiated cells, and induced pluripotent stem cells. The table can be sorted by cell count, number of cell types, or accession, and the page size can be set to 20, 50, or 100 rows.

When exactly one cell type is selected, the browser switches to a cell-type-centric view. Two additional columns show the number and the fraction of cells of that type in each experiment, two additional sort orders ("highest composition ratio" and "most cells of this type") become available, and a separate summary box totals the cells of that type across the matching experiments. This is the view reached from a cell type page, and it answers the question of which datasets are rich in a given cell type.

Clicking an accession opens the experiment page (Figure 4D). It shows the organism and tissue, a table of the cell types identified in the experiment with their cell counts and fractions, each linked to its cell type page, and nine curated sample metadata fields (organism, source

name, tissue, cell type, age, sex, treatment, cell line, and developmental stage). An interactive two-dimensional UMAP embedding of all cells, colored by cell type, lets the user hover over a point to see its cell type and barcode. A link leads to the original record in the NCBI Sequence Read Archive.

3.5.4 Browsing literature claims

The literature browser gives access to the three literature-derived knowledge bases, "cell as agent (abstracts)", "cell as agent (reviews)", and "gene perturbation", through three views that answer three different questions.

The *Claims* view answers what the literature says (Figure 4E). Each claim is rendered as a card showing the source sentence with the cell, the action, and the target or outcome highlighted in place using character offsets from the extraction pipeline. Roles are distinguished by shape (outline, underline, fill) rather than by color alone. Chips on the card give the direction of the effect (up, down, no change, context-dependent, or other), the evidence strength (asserted, suggested, or speculative), whether the sentence reports the paper's own result or recites established knowledge, and flags for negated or correlational statements. Claims from the perturbation knowledge base use a three-line layout ("perturbed", "read out in", "outcome") instead of an arrow, so that the card does not assert a causal relation the sentence did not make. Normalized terms are linked: the cell to its Cell IO cell type page, genes to HGNC, other ontology terms to OLS4, and the PMID to PubMed. A left rail supports free-text search over the cell, target, and gene text and over the full sentence, a cell type filter that accepts either a name or a CL identifier, and ten multi-select facets (knowledge base, direction, evidence strength, provenance, flags, major and minor outcome class, organism, perturbation kind, and verification status). Facet counts are shown against the whole collection so that they remain stable while filters are toggled.

The *Cross* view answers which cell types are covered by which knowledge base (Figure 4F). It lists, for each cell type, the number of claims in each of the three knowledge bases and names the knowledge bases in which it is absent. By default it shows only cell types missing from at least one knowledge base, since the most-cited cell types are present in all three and would otherwise dominate the list. Selecting a cell type adds a breakdown of its claims by outcome class and links to the filtered Claims view.

The *Totals* view answers how large and how balanced the knowledge bases are, with six summary tables: the size of each knowledge base in claims, cell types, and publications; the distributions of direction, evidence strength, provenance, and outcome class per knowledge base; and the overlap of cell type identifiers across the three, including what would be lost by dropping any one of them.

3.5.5 Accessibility and design conventions

The home page, header, and footer were redesigned to conform to WCAG 2.2 Level AA. The search box is a keyboard-operable combobox whose suggestions are announced to screen readers, the entry panels remain usable while counts are loading or unavailable, and the layout, including Windows forced-colors (high-contrast) mode, degrades without losing the outlines of controls. Conformance of the remaining pages has not yet been assessed. Two conventions are, however, applied throughout the site. Information is never encoded by color alone: the three knowledge bases are also distinguished by line pattern and icon, claim roles by shape, and effect direction by arrow glyphs. All links are underlined, with an explicit marker for external destinations. The layout is responsive from phone widths, where the header collapses into a menu, to wide screens. The About page gives the contact address, funding, and the licences of the ontologies used. The site has no separate help or tutorial page yet (see Discussion).

Cell IO is intended for interactive use. The frontend is backed by a read-only HTTP API,

but the public site sits behind a bot-management challenge that admits browsers and verified search-engine crawlers only, so the API is not currently offered for programmatic access. The content of the database is released under the Creative Commons Attribution 4.0 International licence (see Data availability).

4 Use cases

4.1 From a cell type to datasets that contain it

A researcher planning to study regulatory T cells can type "Treg" into the search box; the synonym is resolved to *regulatory T cell* (CL:0000815) and the cell type page reports in how many experiments and cells the term was assigned (Figure 5A). "Browse the N experiments" opens the experiment browser sorted by the fraction of regulatory T cells, which identifies the datasets in which the population is most abundant and shows their organism, tissue, and sample type (Figure 5B). Opening one experiment shows the UMAP embedding with the population highlighted and the link to the SRA record from which the raw data can be downloaded. In the current release the term is assigned in 1,486 experiments and to 509,608 cells, and the experiment shown in Figure 4D, SRX30741276, is a mouse lymph node dataset in which regulatory T cells make up 34.6% of the cells (3,183 cells).

4.2 From a cell type to what the literature says it does

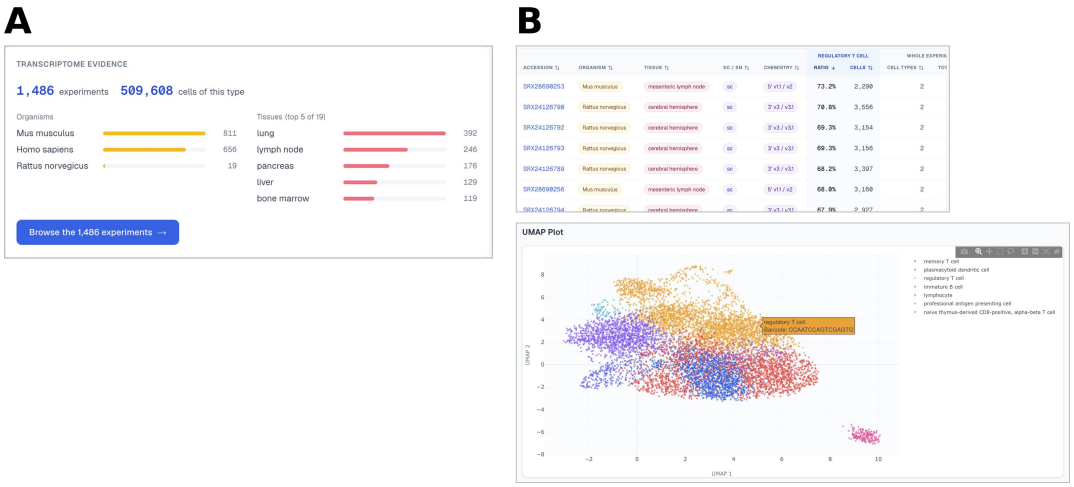
The Literature Evidence panel of the same page summarizes 27,438 claims from 6,091 publications (3,090 from abstracts, 23,992 from reviews, and 356 perturbation findings) and lists the most frequent outcome classes, led by immune suppression (4,552 claims). Following the link to the literature browser, the researcher can restrict the claims to asserted statements reporting the authors' own results, to a given organism, or to a direction of effect, and read each sentence with the cell, action, and target highlighted (Figure 5C). Switching to the perturbation knowledge base lists the genes whose deletion, knockdown, or overexpression changed a property of regulatory T cells, with the direction of the change (Figure 5D). In the current release 356 findings name regulatory T cells as the readout; for example, deletion of Nr4a2 in mouse regulatory T cells reduced Foxp3 expression and suppressive ability (PMID 21468021), and knockdown of miR-126 reduced Foxp3 expression and impaired suppressive function (PMID 23301798, recorded as speculative because the sentence is hedged). Negative results are kept as well: depletion of Foxp3^{+} regulatory T cells did not exacerbate allergic airway inflammation in a mouse model (PMID 23071726), stored with the direction "no change" and the flag "negated". The genes with the most findings for this cell type are NAC1 (NACC1), Nox2 (CYBB), SRC-3 (NCOA3), and miR-126, each with 8 to 11 findings, which illustrates that the knowledge base reflects the sampled open-access corpus rather than a curated gene list.

4.3 From a tissue to its cellular composition across studies

Starting from a tissue, for example *hippocampal formation*, the tissue page shows its position in Uberon, the cell types identified across all experiments from that tissue ranked by cell count, and the organisms studied. The experiment browser then lists the individual datasets, and the sample type filter separates tissue samples from cell lines and from cells differentiated in vitro, a distinction that users requested at a recent symposium.

4.4 Finding gaps in the literature

The Cross view lists cell types that are described in one knowledge base but absent from another, for example cell types with many perturbation findings but no review-level statements, or vice versa (Figure 4F). Combined with the transcriptome counts, this identifies cell types that are



C

Cell (as written)	Action	Target	Direction	Organism	PMID
Tregs	attenuated	oxLDL-induced pro-inflammatory responses	down	Humans	41181110
Treg	inhibiting	tumor associated antigen-specific immune responses	down	Humans	17205133
T reg cells	restraint of	an MNP-IFN-I-driven $CD8^+$ T cell response	down	Mice	29980582
T-bet^+ regulatory T (Treg) cells	inhibit	Th1 immune responses	down	Mice	25658840
Tregs	prevent	excessive immune responses	down	Mice	25807185

D

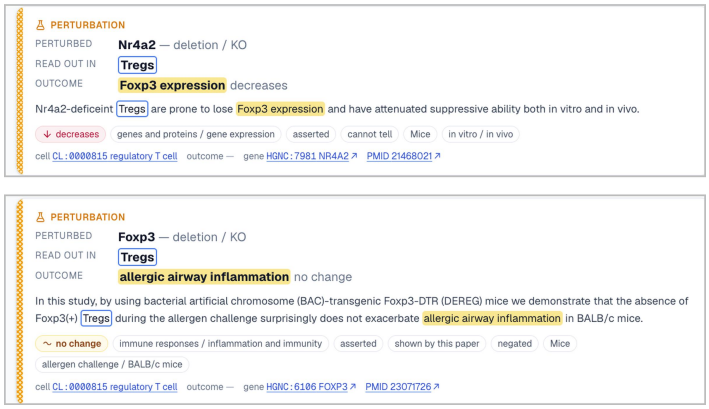


Figure 5: **Use case: from a cell type to datasets and knowledge.** (**A**) The Transcriptome Evidence panel of the regulatory T cell page (CL:0000815) reports 1,486 experiments and 509,608 cells of this type and links to the experiment browser. (**B**) The experiment browser sorted by composition ratio lists the datasets most enriched in regulatory T cells with their organism, tissue, and chemistry; the UMAP of one experiment (SRX30741276) shows the population. (**C**) Five cell-as-agent claims about regulatory T cells from the abstract knowledge base, restricted to asserted statements that report the authors' own result and whose outcome class is immune suppression; the cell, action, and target are given as written in the sentence, and the direction is the polarity of the action. (**D**) Two gene perturbation findings returned by the text search "FOXP3" in the perturbation knowledge base: deletion of Nr4a2 reduced Foxp3 expression in mouse regulatory T cells (PMID 21468021), and depletion of $Foxp3^+$ regulatory T cells did not change allergic airway inflammation (PMID 23071726), a negative result retained with the direction "no change" and the flag "negated".

frequently observed in single-cell data but rarely characterized functionally in the literature, which are natural targets for further study. For example, the finer term CD4-positive, CD25-positive, alpha-beta regulatory T cell (CL:0000792) carries 720 review claims and 290 abstract claims but no perturbation finding, whereas its perturbation findings are attached to the coarser term regulatory T cell, and embryonic fibroblast (CL:2000042) has 101 perturbation findings but only 8 review claims. On the transcriptome side, pinealocyte (CL:0000652) is assigned to 1,269,211 cells in 1,260 experiments but has four literature claims, and 112 of the 697 cell types with transcriptome evidence have no claim in any knowledge base.

5 Discussion

Cell IO differs from existing single-cell portals in what it places next to the expression data. Where CZ CELLxGENE Discover [1] or the Single Cell Expression Atlas [3] answer which genes a cell type expresses, Cell IO adds, for the same Cell Ontology term, what the literature says the cell type does and how it responds to gene perturbations (Table 1). The two layers were built independently and meet only through shared ontology identifiers, which keeps each layer replaceable: a new annotation run or a new extraction run replaces one side without touching the other. Several limitations follow from this design and from the automated construction, and they define the planned extensions.

5.1 Coverage of the transcriptome layer

The transcriptome layer is restricted to 10x Chromium libraries of three species found by keyword searches centered on the nervous and immune systems, so it is not a census of public single-cell data. Some libraries that are not gene expression assays, in particular V(D)J libraries, were found in the corpus, because SRA metadata do not flag them reliably; a rule-based exclusion removed 1,919 experiments, and an expression-based classifier is being developed to catch the remainder. Cell lines are mapped by MetaSRA but not yet searchable as such, and extension to further species, to non-Chromium protocols, to CITE-seq, and to perturbation screens is planned. Users have also asked for download of the processed count matrices, which amount to about 30 TB in h5ad format; a distribution route is under evaluation.

5.2 Quality of cell type annotation

Clusters are annotated by marker scoring against the CellMarkerAccordion database at a single clustering resolution, without reference mapping or manual review. Annotations therefore disagree with the original publications in some experiments, for example by assigning a developmental cell type to adult tissue, and populations that are not defined in the Cell Ontology cannot be named. We evaluated reference mapping with scArches [46] and factor-based discovery of unannotated populations, and neither was adopted; a recursive clustering in a shared latent space is the current direction for defining subtypes below the Cell Ontology terms. Hierarchical relations are loaded but evidence is not aggregated to ancestor terms, so a search for a coarse term such as “T cell” does not return claims attached to its subtypes; hierarchy-aware aggregation is planned.

5.3 Reliability of literature-derived claims

The literature knowledge bases were produced by a single large language model with a verification pass, and their precision was estimated on samples judged by one annotator: 0.96 for abstract claims, 0.95 for review claims, and lower for ontology mapping of targets (0.73) and for the provenance label. Recall at the sentence level is about 0.5, so the knowledge bases should be read as a precise sample of the literature rather than a complete record. Provenance is “unclear” for

most claims, because rule-based detection of whether a sentence reports the authors' own result is weak. Review-derived claims restate consensus and are counted separately from abstract-derived claims for this reason. The knowledge bases capture statements in which a cell acts on something and findings in which a gene perturbation changes a cell; they do not capture expression facts, changes in cell abundance under a condition, or cell–cell interactions as such, and the corpus is limited to open-access articles in immunology and neuroscience. The knowledge bases have not been curated by domain experts: the gold sets behind the precision estimates were judged by one of the authors, and the immunologists and neuroscientists acknowledged below reviewed output of an earlier version of the pipeline rather than the released claims. We therefore present the claims as machine-extracted evidence that points to its source, not as expert-validated facts. Every claim keeps the verbatim sentence, the character spans, and the PMID, so that users can check a statement against its source in one click, and rejected relations are kept with their reasons so that the verification step itself can be audited.

5.4 Planned extensions

Beyond the items above, we plan to release the database as RDF with a SPARQL endpoint in collaboration with DBCLS, to offer the single-cell versus single-nucleus and 3′ versus 5′ chemistry labels, which are currently displayed per experiment, as experiment filters, to report the overlap between transcriptome and literature coverage per cell type inside the interface, and to add help material for first-time users. Users at two recent meetings asked for four entry points that the current interface only partly supports: identifying a user's own cluster from a marker set, listing characteristic genes and functions for a cell type, finding the cell types associated with a gene or signature, and interpreting a user's own data against the database. These define the roadmap for the next release.

5.5 Conclusion

Cell IO provides, for human, mouse, and rat cell types defined in the Cell Ontology, a uniformly processed view of their occurrence in public single-cell data and a verifiable, sentence-level view of what the literature says they do. It is freely available at `https://cell-io.riken.jp/`.

Data availability

Cell IO is freely accessible at `https://cell-io.riken.jp/` without registration. The content of the database is released under the Creative Commons Attribution 4.0 International licence (CC BY 4.0). The underlying sequencing data are available from the NCBI Sequence Read Archive under the accessions listed on each experiment page, and every literature claim links to its source article through its PubMed identifier.

Funding

This work was supported by JST, NBDC Japan Grant Number JPMJND2402 to H.O. and the RIKEN TRIP Advanced General Intelligence for Science Program (AGIS) to H.O.

Acknowledgements

We thank Ryo Ishibashi for prototyping, Ryota Yamada (Science Aid, Inc.) for the design of text mining, and Manabu Ishii (Tech V Inc.) for database security. We thank the domain experts who reviewed extracted text during the development of the literature layer: Shinya Nakata, Kazuya Miyanishi, Ami Kaneko, Yoshihiko Sakaguchi, and Haruto Ijiri (neuroscience), and Naganari

Ohkura, Ryoji Kawakami, Masaki Hagihara, Reo Morimoto, and Daiki Takeuchi (immunology). This research used computational resources of the RIKEN HOKUSAI SS supercomputer system and, in part, of the supercomputer RIKYU provided by the RIKEN Center for Computational Science.

Conflict of interest statement

None declared.

References

[1] CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

[2] Leyla Tarhan, Jon Bistline, Jean Chang, Bryan Galloway, Emily Hanna, and Eric Weitz. Single cell portal: an interactive home for single-cell genomics data. *bioRxiv*, 2023. doi: 10.1101/2023.07.13.548886.

[3] Nancy George, Silvie Fexova, Alfonso Munoz Fuentes, Pedro Madrigal, Yalan Bi, Haider Iqbal, Upendra Kumbham, Nadja Francesca Nolte, Lingyun Zhao, Anil S Thanki, et al. Expression atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Research*, 52(D1):D107–D114, 2024. doi: 10.1093/nar/gkad1021.

[4] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019, 2019. doi: 10.1093/database/baz046.

[5] Mengwei Li, Kok Siong Ang, Brian Teo, Uddamvathanak Rom, Minh N Nguyen, Sebastian Maurer-Stroh, and Jinmiao Chen. Rediscovering publicly available single-cell data with the disco platform. *Nucleic Acids Research*, 53(D1):D932–D938, 2025. doi: 10.1093/nar/gkae1108.

[6] Nicholas D. Youngblut, Christopher Carpenter, Arshia Nayebnazar, Abhinav Adduri, Rohan Shah, Chiara Ricci-Tam, Jaanak Prashar, Rajesh Ilango, Noam Teyssier, Silvana Konermann, et al. scbasecount: an ai agent-curated, uniformly processed, and autonomously updated single cell data repository. *bioRxiv*, 2025. doi: 10.1101/2025.02.27.640494.

[7] Christopher Wilks, Shijie C. Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P. Ling, Eddie Luidy Imada, David Zhang, Lance Joseph, Jeffrey T. Leek, et al. recount3: summaries and queries for large-scale rna-seq expression and splicing. *Genome Biology*, 22(1):323, 2021. doi: 10.1186/s13059-021-02533-6.

[8] Alexander Lachmann, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hoyjin J. Lee, Lily Wang, Moshe C. Silverstein, and Avi Ma'ayan. Massive mining of publicly available rna-seq data from human and mouse. *Nature Communications*, 9(1):1366, 2018. doi: 10.1038/s41467-018-03751-6.

[9] Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7.

[10] Shawn Zheng Kai Tan, Aleix Puig-Barbe, Damien Goutte-Gattat, Caroline Eastwood, Brian Aevermann, Alida Avola, James P. Balhoff, Ismail Ugur Bayindir, Jasmine Belfiore, Anita Reane Caron, et al. The cell ontology in the age of single-cell omics. *Scientific Data*, 13(1):946, 2026. doi: 10.1038/s41597-026-07173-8.

[11] Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, 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(D1):D870–D876, 2023. doi: 10.1093/nar/gkac947.

[12] Kevin Troulé, Robert Petryszak, Batuhan Cakir, James Cranley, Alicia Harasty, Martin Prete, Zewen Kelvin Tuong, Sarah A. Teichmann, Luz Garcia-Alonso, and Roser Vento-Tormo. Cellphonedb v5: inferring cell–cell communication from single-cell multiomics data. *Nature Protocols*, 20(12):3412–3440, 2025. doi: 10.1038/s41596-024-01137-1.

[13] Xin Shao, Jie Liao, Chengyu Li, Xiaoyan Lu, Junyun Cheng, and Xiaohui Fan. Celltalkdb: a manually curated database of ligand–receptor interactions in humans and mice. *Briefings in Bioinformatics*, 22(4):bbaa269, 2021. doi: 10.1093/bib/bbaa269.

[14] Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546, 2024. doi: 10.1093/nar/gkae235.

[15] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012. doi: 10.1093/bioinformatics/bts591.

[16] Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

[17] Zhiting Wei, Duanmiao Si, Bin Duan, Yicheng Gao, Qian Yu, Zhenbo Zhang, Ling Guo, and Qi Liu. Perturbase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Research*, 53(D1):D1099–D1111, 2025. doi: 10.1093/nar/gkae858.

[18] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024. doi: 10.1038/s41467-024-45563-x.

[19] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. Structured prompt interrogation and recursive extraction of semantics (spires): a method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3): btae104, 2024. doi: 10.1093/bioinformatics/btae104.

[20] Single Cell Portal. SCP convention – required metadata. `https://singlecell.zendesk.com/hc/en-us/articles/360060609852`. Accessed 30 October 2026.

[21] Eric W Sayers, Jeffrey Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Ryan Connor, Michael Feldgarden, Anna M Fine, Kathryn Funk, Jinna Hoffman, et al. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Research*, 53 (D1):D20–D29, 2025. doi: 10.1093/nar/gkae979.

[22] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. doi: 10.1093/bioinformatics/bts635.

[23] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data. *bioRxiv*, 2021. doi: 10.1101/2021.05.05.442755.

[24] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. doi: 10.1038/nbt.3820.

[25] Aaron T. L. Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C. Marioni. Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell rna sequencing data. *Genome Biology*, 20(1):63, 2019. doi: 10.1186/s13059-019-1662-y.

[26] Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E Loveland, Carme Arnan, If Barnes, Ruth Bennett, et al. Gencode 2025: reference gene annotation for human and mouse. *Nucleic Acids Research*, 53(D1):D966–D975, 2025. doi: 10.1093/nar/gkae1078.

[27] Andrew D Yates, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Vianey Paola Barrera-Enriquez, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2026. *Nucleic Acids Research*, 54(D1):D1053–D1060, 2026. doi: 10.1093/nar/gkaf1239.

[28] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

[29] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

[30] Ayshwarya Subramanian, Mikhail Alperovich, Yiming Yang, and Bo Li. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, 23(1):267, 2022. doi: 10.1186/s13059-022-02820-w.

[31] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291.e9, 2019. doi: 10.1016/j.cels.2018.11.005.

[32] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

[33] Emma Busarello, Giulia Biancon, Ilaria Cimignolo, Fabio Lauria, Zuhairia Ibnat, Christian Ramirez, Gabriele Tomè, Marianna Ciuffreda, Giorgia Bucciarelli, Alessandro Pilli, et al. Cell marker accordion: interpretable single-cell and spatial omics annotation in health and disease. *Nature Communications*, 16(1):5399, 2025. doi: 10.1038/s41467-025-60900-4.

[34] Matthew N Bernstein, AnHai Doan, and Colin N Dewey. Metasra: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics*, 33(18):2914–2923, 2017. doi: 10.1093/bioinformatics/btx334.

[35] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1): R5, 2012. doi: 10.1186/gb-2012-13-1-r5.

[36] Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50(D1):D1255–D1261, 2022. doi: 10.1093/nar/gkab1063.

[37] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010. doi: 10.1093/bioinformatics/btq099.

[38] Amos Bairoch. The cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques : JBT*, 29(2):25–38, 2018. doi: 10.7171/jbt.18-2902-002.

[39] Gemma Team and Google DeepMind. Gemma 4 technical report, 2026. URL `https://arxiv.org/abs/2607.02770`.

[40] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*, pages 611–626, 2023. doi: 10.1145/3600006.3613165.

[41] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[42] Nipun Sadvilkar and Mark Neumann. Pysbd: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, 2020. doi: 10.18653/v1/2020.nlposs-1.15.

[43] Ruth L Seal, Bryony Braschi, Kristian Gray, James McClay, Susan Tweedie, and Elspeth A Bruford. Genenames.org: the hgnc and pgnc resources in 2026. *Nucleic Acids Research*, 54 (D1):D1098–D1107, 2026. doi: 10.1093/nar/gkaf1229.

[44] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003: Proceedings of the Pacific Symposium*, pages 451–462, 2003. doi: 10.1142/9789812776303_0042.

[45] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, pages 4228–4238, 2021. doi: 10.18653/v1/2021.naacl-main.334.

[46] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40 (1):121–130, 2022. doi: 10.1038/s41587-021-01001-7.