

Lawrence Berkeley National Laboratory

LBL Publications

Title

The Cell Ontology in the age of single-cell omics

Permalink

<https://escholarship.org/uc/item/3qm903wf>

Journal

Scientific Data, 13(1)

ISSN

2052-4463

Authors

Tan, Shawn Zheng Kai

Puig-Barbe, Aleix

Goutte-Gattat, Damien

et al.

Publication Date

2026-04-01

DOI

10.1038/s41597-026-07173-8

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

OPEN
ARTICLE

The Cell Ontology in the age of single-cell omics

Shawn Zheng Kai Tan^{1,2,25}, Aleix Puig-Barbe^{2,25}, Damien Goutte-Gattat^{3,4}, Caroline Eastwood⁵, Brian Aeevermann⁶, Alida Avola⁵, James P. Balhoff⁷, Ismail Ugur Bayindir⁵, Jasmine Belfiore⁵, Anita Reane Caron², David S. Fischer⁸, Nancy George⁹, Benjamin M. Gyori¹⁰, Melissa A. Haendel¹¹, Charles Tapley Hoyt¹², Huseyin Kir⁵, Tiago Lubiana¹³, Nicolas Matentzoglou¹⁴, James A. Overton¹⁵, Beverly Peng¹⁶, Bjoern Peters¹⁷, Ellen M. Quardokus¹⁸, Patrick L. Ray¹⁹, Paola Roncaglia², Andrea D. Rivera⁵, Ray Stefancsik², Wei Kheng Teh², Sabrina Toro¹¹, Nicole Vasilevsky²⁰, Chuan Xu²¹, Yun Zhang²², Richard H. Scheuermann²²✉, Christopher J. Mungall²³✉, Alexander D. Diehl²⁴✉ & David Osumi-Sutherland⁵✉

Single-cell omics technologies have transformed our understanding of cellular diversity by enabling high-resolution profiling of individual cells. However, the unprecedented scale and heterogeneity of these datasets demand robust frameworks for data integration and annotation. The Cell Ontology (CL) has emerged as a pivotal resource for achieving FAIR (Findable, Accessible, Interoperable, and Reusable) data principles by providing standardized, species-agnostic terms for canonical cell types, forming a core component of a wide range of platforms and tools. In this paper, we describe the wide variety of uses of CL in these platforms and tools and detail ongoing work to improve and extend CL content including the addition of transcriptomic types, working closely with major atlasing efforts including the Human Cell Atlas and the Brain Initiative Cell Atlas Network to support their needs. We cover the challenges and future plans for harmonising classical and transcriptomic cell type definitions, integrating markers and using Large Language Models (LLMs) to improve content and efficiency of CL workflows.

¹SigmaMind, Singapore, Singapore. ²European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ³Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3DY, UK. ⁴German Biolmaging GMB e.V., c/o University of Konstanz, Box 604, 78454, Konstanz, Germany. ⁵Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1RQ, UK. ⁶Chan Zuckerberg Initiative, Redwood City, USA. ⁷Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC, USA. ⁸Medical University of Vienna, Institute of Artificial Intelligence, Center for Medical Data Science, Vienna, Austria. ⁹Syngenta, Jealott's Hill, Warfield, Bracknell, UK. ¹⁰Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA. ¹¹University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ¹²RWTH Aachen University, Institute of Inorganic Chemistry, Landoltweg 1a, 52074, Aachen, Germany. ¹³University of São Paulo, São Paulo, Brazil. ¹⁴Semanticly, Athens, Greece. ¹⁵Knocean Inc., Toronto, Ontario, Canada. ¹⁶Department of Informatics, J. Craig Venter Institute, La Jolla, CA, USA. ¹⁷La Jolla Institute for Immunology, 9420 Athena Circle, La Jolla, CA, 92037, USA. ¹⁸Indiana University: Bloomington, Indiana, US. ¹⁹Allen Institute for Brain Science, Seattle, WA, USA. ²⁰Critical Path Institute, Tucson, AZ, USA. ²¹Cambridge Stem Cell Institute and Department of Medicine, University of Cambridge, Trinity Ln, Cambridge, UK. ²²Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ²³Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA, 94720, USA. ²⁴Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, 14203, USA. ²⁵These authors contributed equally: Shawn Zheng Kai Tan, Aleix Puig-Barbe. ✉e-mail: richard.scheuermann@nih.gov; cjmungall@lbl.gov; addiehl@buffalo.edu; do12@sanger.ac.uk

Introduction

An ever growing array of single cell omics techniques are being used to generate massive and diverse datasets profiling single cells. These include techniques for single cell transcriptomics^{1,2}, proteomics³, multiplexed imaging of protein location⁴, metabolomics⁵, spatially resolved transcriptomics⁶, connectomics^{7,8}, epigenetic profiling⁹, and multimodal techniques such as patch-sequencing (patch-seq)¹⁰.

The substantial volumes of data generated by these techniques pose a significant challenge for biologists, clinicians, and bioinformaticians aiming to identify and integrate relevant data for analysis, generate hypotheses, and apply machine learning techniques across datasets. To best utilise these new datasets for hypothesis generation in biology and medicine, we need ways to connect omics data to traditional knowledge of cell types, including their anatomy and physiology. This prior knowledge is primarily found in the research literature as free text in non-standardised forms. Integrating this knowledge with quantifiable 'omics datasets is therefore a challenge.

Findable, Accessible, Interoperable, and Reusable (FAIR) data annotation principles¹¹ can help to solve these problems by ensuring that data is structured and annotated consistently. Annotation with ontology terms is a key component of these principles. In the context of single cell data, ontologies provide unambiguous identifiers¹² for consistent annotation of cell type, tissue, developmental stage, and disease. Referenced textual definitions attached to these identifiers ensure that the meaning of ontology terms is clear, while attached names and synonyms support searching with a wide range of commonly used names.

Ontologies also relate terms to each other in biologically meaningful ways. As a result, ontology annotation not only integrates data annotated with the same term, it can also integrate data annotated with related terms. This underlies the widespread use of the Gene Ontology for gene enrichment analysis^{13,14} - summarising the functions of genes in lists by mapping up from specific annotations of individual genes to more general terms in the ontology, e.g. mapping 'CMG complex assembly' to 'lagging strand initiation' to 'DNA replication'. In the context of a cell ontology, a query for data annotated with an ontology term for 'T cell' might use the structure of the ontology to return data annotated with terms for the many known subtypes of 'T cell'.

Relationships between ontology terms can extend beyond a single type of term, for example linking a cell type to terms recording its anatomical location, characteristic components, functions, and cell surface markers (see Fig. 1). Collectively, these relationships form a graph that can be used to support applications ranging from database search to gene set enrichment to machine learning and which is used as the backbone of knowledge graphs^{15,16}. In the context of cell type ontologies, these relationships, combined with referenced textual definitions, provide links to classical and contextual knowledge about cell types. This context is useful not only to biologists but also as input to machine learning^{17,18}.

The Cell Ontology. The Cell Ontology (CL), the first version of which was released in 2004¹⁹, is a foundational resource for FAIR annotation of cell type data, and widely used to drive data integration and search in biomedical research by academia and their partners, government supported initiatives (Table S1), and industry (eg pharmaceutical companies, tech companies like Benevolent AI, and many others)²⁰. CL represents canonical (non-pathological/wild-type) cell types. It covers granular cell types in mammals, while also supporting general classes for other metazoa (e.g., muscle cell, neuron). It facilitates interoperability through mappings to species-specific ontologies for model organisms such as those for *Xenopus*, Zebrafish, and *Drosophila*.

As part of the OBO (Open Biological and Biomedical Ontology) Foundry^{21,22}, CL adheres to standardized methods for defining relationships and classifications^{23,24} and integrates tightly with other OBO Foundry ontologies. CL is also used by phenotype²⁵, disease²⁶, and trait ontologies²⁷ to record the cellular location of disease and phenotype and by the Gene Ontology (GO) to record the location of processes and cellular components²⁸ - see Fig. 1 for examples. CL in turn uses GO to capture cell type functions and the characteristic components of cell types, and the anatomy ontology Uberon²⁹ to record anatomical location of cell types. As the GO is a rich source of gene product annotation, links between CL and GO provide a potential route to map genes relevant to the transcriptomic and proteomic signatures of cell types.

Terms in CL are predominantly defined and classified based on classical structural, functional, and molecular criteria (classical cell types). For instance, pulmonary alveolar type 2 cells (AT2 cell) (Fig. 1) are defined by their location in the alveolar epithelium, their columnar morphology and their function in surfactant production, secretion and homeostasis³⁰. Surfactant production requires the presence of an alveolar lamellar body, an organelle that is specific to these cells. These properties are all reflected in the CL textual and formal definition - which classifies the cell type as a columnar/cuboidal epithelial cell, records its alveolar location, and links it to GO terms for alveolar lamellar body and surfactant homeostasis (http://purl.obolibrary.org/obo/CL_0002063). As shown in Fig. 1 and Table 1, we can find genes that reflect the transcriptomic signatures underpinning the structure and function of this cell type by intersecting the set of genes annotated to these GO terms with markers derived from analysis of transcriptomics data annotated with the AT2 CL term.

CL supports diverse use cases by allowing terms to have more than one parent classification; for example, AT2 cells are classified both as alveolar epithelial cells and cuboidal epithelial cells (other alveolar epithelial cells are squamous). Similarly tracheal goblet cells are both goblet cells (which are found in many mucosal epithelia) and tracheal epithelial cells, reflecting both function and anatomical attributes. Such classifications leverage OWL logic to automate error checking and ensure consistency²¹⁻²⁴.

Markers are practical tools for identifying cell types. Their practical application depends on both anatomical and experimental context. A gene (or set of genes) is a useful marker if its expression can be used to unambiguously identify (mark) a cell type in some context. Markers are often species-specific. Proteins are commonly used as cell-type markers, but protein and transcript levels often do not correlate well and proteins may be present when transcript has degraded³¹. Markers that are specific in a narrow anatomical context (e.g. a single brain region) may not be in the context of a whole brain (see Fig. 2). Many classical markers have their origin in the context of imaging prior to the advent of spatial transcriptomics. In this context, markers typically need

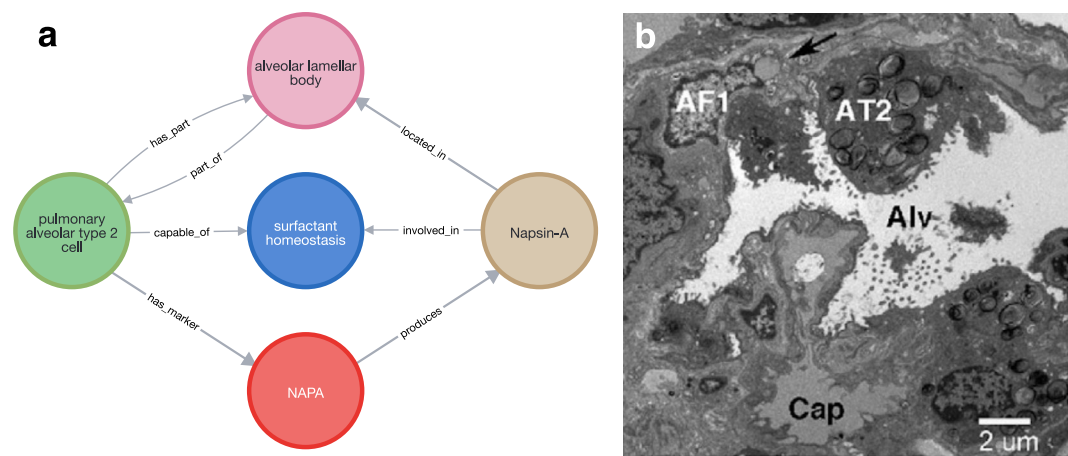


Fig. 1 ‘Classical’, property-based definition of pulmonary alveolar type 2 cell (AT2 cell) in the Cell Ontology. In the Cell Ontology this term (CL:0002063) is defined as “A pulmonary alveolar epithelial cell that modulates the fluid surrounding the alveolar epithelium by secreting and recycling surfactants via specialized organelles called alveolar lamellar bodies... This cuboidal cell is thicker than squamous alveolar cells (AT1), has a rounded apical surface that projects above the level of the surrounding epithelium, and its free surface is covered by short microvilli”. **Panel a** shows this property-based definition as a graph linking the Cell Ontology term ‘pulmonary alveolar type 2 cell’ to the GO terms for ‘surfactant homeostasis’ and ‘alveolar lamellar body’. GO annotation records the function and of the Napsin-A protein in surfactant homeostasis and its location in the alveolar lamellar body. NAPA, the gene encoding Napsin-A, was identified as an AT2 cell marker via analysis of single cell transcriptomics data annotated with this cell type³⁵. **Panel b**, shows an electron micrograph of an AT2 cell with many alveolar lamellar bodies (ALM) (adapted from Sun *et al.*¹³⁸ licensed under CC-BY4.0).

Cell Type	GO Term	Mouse Markers	Human Markers
pulmonary alveolar type 2 cell	surfactant homeostasis	Lamp3, Abca3, Lpcat1, Ctsh, Sftpd, Napsa	SFTPD, NAPSA, ABCA3, CTSH
pulmonary alveolar type 2 cell	alveolar lamellar body	Abca3, Napsa, Sftpc, Sftpb, Ctsh	ABCA3, NAPSA, SFTPB, SFTPC, CTSH

Table 1. Mouse and human markers of AT2 cells that are annotated to a GO terms for a defining structure or function of this cell type.

to distinguish cell types within the narrow anatomical context of a single imaged tissue slice or block, but large differences in expression compared to background are needed for thresholding to be able to mark cells. In contrast, markers derived from single cell transcriptomics often need to distinguish cells in the context of large and diverse tissue samples, and the typical approach to deriving markers, differential expression analysis, can take advantage of relative expression levels that would be too subtle to clearly mark cells in a traditional imaging context. Variation in documented markers can also come from different marker differential expression algorithms producing different top marker gene lists⁶.

CL uses gene expression as one criterion for defining cell types in cases where expression is widely considered definitional for the cell type, for instance for immune cells, where cell type definitions have long been dominated by cell surface markers used for isolating cell populations. As detailed in this paper, we also use markers as a component of cell types defined with reference to transcriptomic data. However, CL does not aim to be a comprehensive source of cell type markers. The highly contextual and diverse nature of markers means that storing all markers would bloat the ontology and make it hard to use. A number of resources use CL terms to record markers in their species and anatomical context, including the HubMap Human Reference Atlas^{32,33}, CellMarker³⁴ and CELLxGENE Expression³⁵. The SOULCAP initiative (Standardised Ontology Unique Labelling for Cytometry Annotation of Populations, <https://soulcap.org/>) is an international consortium of academic and industry researchers in flow cytometry who are developing an expert curated database of marker-cell type associations, primarily for immune cells, in collaboration with CL developers, as well as bioinformatics methods to standardize cell type identification across flow cytometry results.

Transcriptomic cell types. Recent advances in single-cell omics technologies have transformed how scientists study cell types by providing unbiased ways of characterising cell identity without relying on prior knowledge. Progress has been made most rapidly in single cell transcriptomics, which has provided new means for typing cells based on shared transcriptomic profiles. This typically involves selection of informative (highly variable) genes, dimensionality reduction, and clustering in latent space³⁶. Types defined in this way can be mapped back to classically defined types based on expression of known markers, sometimes combined with gene set enrichment analysis providing evidence for functions/pathways diagnostic of particular cell types³⁷.

Clusters of sufficient granularity often correspond well to classical cell types or subdivide them^{38,39}. Where they do not, some researchers merge or split clusters to more closely match classical cell types or align with other

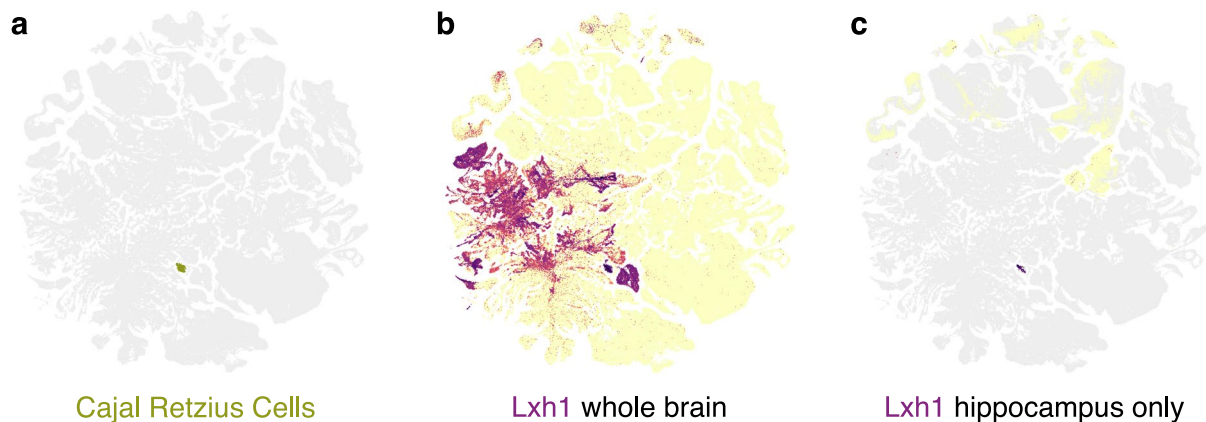


Fig. 2 The contextual nature of markers. **Panel a** shows a UMAP embedding of >4m cells covering the whole mouse brain⁴¹, with cells annotated as Cajal Retzius cells (CR) highlighted. Lhx1 is a classically known marker used to identify CR cells in microscopy¹³⁹. It is not specific in the context of the whole brain (**Panel b**). However, the vast majority of CR cells surviving in the adult mouse brain are present in the hippocampus¹³⁹. When we filter out cells from other brain regions, Lhx1 is a much more specific marker (**Panel c**). This figure was generated using the Allen Brain Cell Atlas (<https://portal.brain-map.org/atlas-and-data/bkp/abc-atlas>).

modalities, whereas others treat cell types that are transcriptomically defined (T-types) as primary⁴⁰. The latter approach is especially important in the central nervous system, where classical cell typing has only ever been able to provide a very limited and incomplete picture of the cell types present⁴⁰. As well as defining types, transcriptomic analysis is also used to define hierarchies of cell types based on transcriptomic similarity, including transcriptomic hierarchies covering the entire brain⁴¹.

In this paper we detail how CL is currently being used to help make data about cell types Findable, Accessible, Interoperable, and Reusable (FAIR). We explore how CL helps to form a bridge between different data modalities, and between 'omics views of cell types and classical cell type definitions. We also detail how CL is evolving to face new challenges, including extending to include T-types, defined via links to reference data and marker sets derived from reference data. Finally we outline how improved technical and community coordination and the use of large language models in the curation process^{42,43} is driving content improvements and new opportunities for CL use.

Results

Application. *CL drives FAIR across multiple platforms, species and modalities.* With the huge influx of single cell 'omics data, annotation with CL has become an important step in making single cell/nucleus RNA-seq (scRNAseq) data FAIR, and has become a key use of CL.

Consortia and platforms like HuBMAP (the Human BioMolecular Atlas Program)³², the Human Cell Atlas (HCA)⁴⁴, and CELLxGENE^{35,45} mandate the use of CL IDs to annotate cell type terms in datasets—this has allowed better integration and findability of datasets. (For a more comprehensive list covering 19 different platforms and tools, see Table S1).

This is perhaps best illustrated by its use in the CZ CELLxGENE (CxG) platform³⁵, the largest and most widely used, consistently annotated database of single cell transcriptomics data. It uses CL to annotate single cell transcriptomics data and to drive many aspects of its functionality.

CxG mandates that CL be used for annotation of cell types in all submitted datasets via a well-defined standard for submission of annotated cell-by-gene matrices³⁵. By providing terms for annotation at various levels of granularity, it allows for accurate but less specific annotations where annotators are not confident of the precise identity of the cells being annotated. On the CxG Discover platform, users can find datasets by the cell types they contain via a faceted browsing interface, which supports searching for cell types by name and browsing via a simplified version of CL hierarchy (Fig. 3c). An alternative view is provided by CxG CellGuide, which displays extended information and analysis about cell types and links to datasets featuring data about these cell types. CellGuide supports search for cell type and dataset by cell type name or synonym (from CL) and browsing of cell types and datasets via a more complete, interactive graph view of CL hierarchy (Fig. 3a&b). CellGuide also takes advantage of CL annotation and hierarchy to calculate marker genes by aggregating data about cell types across multiple datasets and multiple levels of granularity (Fig. 3d). CxG also provides integrated embeddings, combining data about specific cell types from across the CxG corpus. For example, the integrated embedding page for human T cells (https://cellxgene.cziscience.com/e/cellguide-cxgs/homo_sapiens/CL_0000084.cxg/) aggregates data annotated with over 60 terms for types of T cells from CL. CL annotation also plays important roles in CxG Census API functionality, supporting the generation of synthetic cell-by-gene matrices for specific cell types from across the census. In summary, CL is an integral component of CELLxGENE that is critical to much of its functionality.

CL is also used, along with Uberon, by the HuBMAP Human Reference Atlas^{32,46} and CellMarker³⁴ to curate sets of cell-type specific markers in various tissue contexts from the literature. This allows the integration of

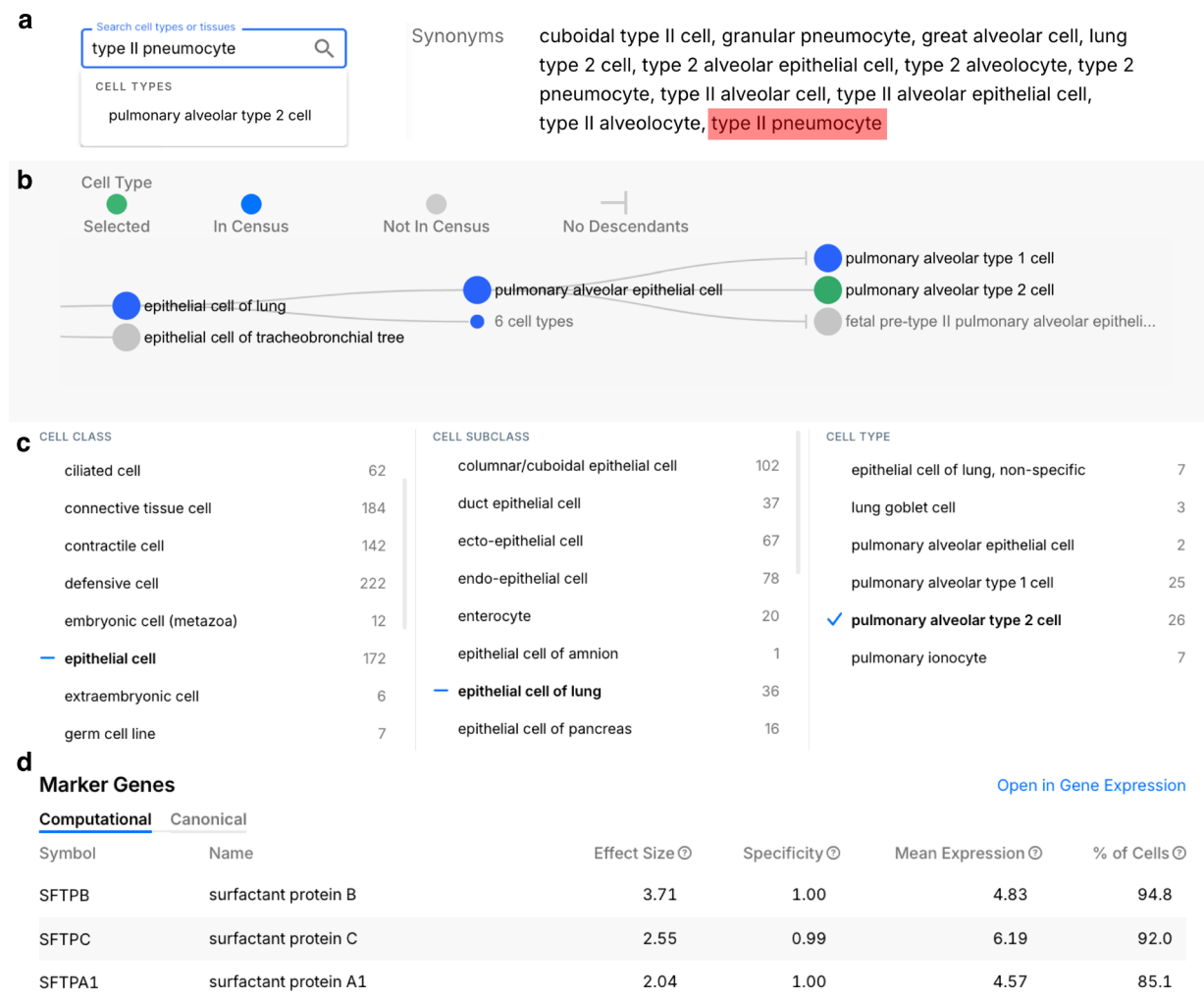


Fig. 3 The Cell Ontology supports findability of scRNAseq data on the CZ CELLxGENE platform³⁵. CELLxGENE CellGuide (panels a and b) provides data about markers and other summary information about cell types as well as links to datasets. **Panel a:** CellGuide Search supports search by synonym, in this case the term ‘pulmonary alveolar type 2 cell’ is found via its synonym ‘type II pneumocyte’. **Panel b:** CellGuide supports browsing of related terms via a compact, interactive tree that shows the classification hierarchy of CL, annotated to show which terms have associated data. Selecting a term returns datasets annotated with that term and its subclasses. **Panel c** shows a search for datasets by cell type using the “CELLxGENE Discover” search interface. An initial lexical search step narrows down search terms from all those used. For ease of browsing the results are arranged in a 3-level hierarchy that combines curation of terms into 3 lists of terms at different levels of granularity (“class”, “subclass”, “type”) with links between the levels driven by the structure of CL. In the above example the ‘Class’ level term selected is ‘epithelial cell’, which restricts the ‘Subclass’ column terms to subclasses of ‘epithelial cell’ in CL including the selected term ‘lung epithelial cell’ of which the selected “Cell Type” term, ‘pulmonary alveolar type 2 cell’ has been used in 15 of >1000 datasets on CELLxGENE. **Panel d:** Annotation with CL terms (in this case pulmonary alveolar type 2 cell) is used to aggregate single cell transcriptomic data to calculate marker genes.

community curated markers from the HuBMAP Human Reference Atlas into multiple resources including the CxG CellGuide and the gene enrichment knowledge graph Enrichr-KG⁴⁷.

While the primary focus of CL is on vertebrate and especially mammalian cell types, precise description of specialized non-mammalian cell types is delegated to species-specific ontologies. CL provides “bridges” between its general cell type terms to their species-specific counterparts. These bridges are routinely curated to ensure their continued accuracy, and the workflow used to produce them has recently been modernized (as further described in the Methods section). This supports integration of cell types specific to a much broader range of species including Zebrafish⁴⁸, *Xenopus*⁴⁹, *Drosophila*^{16,50}, and *C.elegans*⁵¹. While some of these bridges are likely to represent convergent phenotypes, especially between vertebrates and invertebrates, recent transcriptomic evidence supports conservation of broad cell types (e.g. neurons and muscle cells) across the bilateria⁵².

These cross-species bridges do not explicitly encode whether related cell types are homologous or analogous. However, single-cell transcriptomics is enabling more systematic evaluation of cross-species relationships.

Cell type results of enterocyte

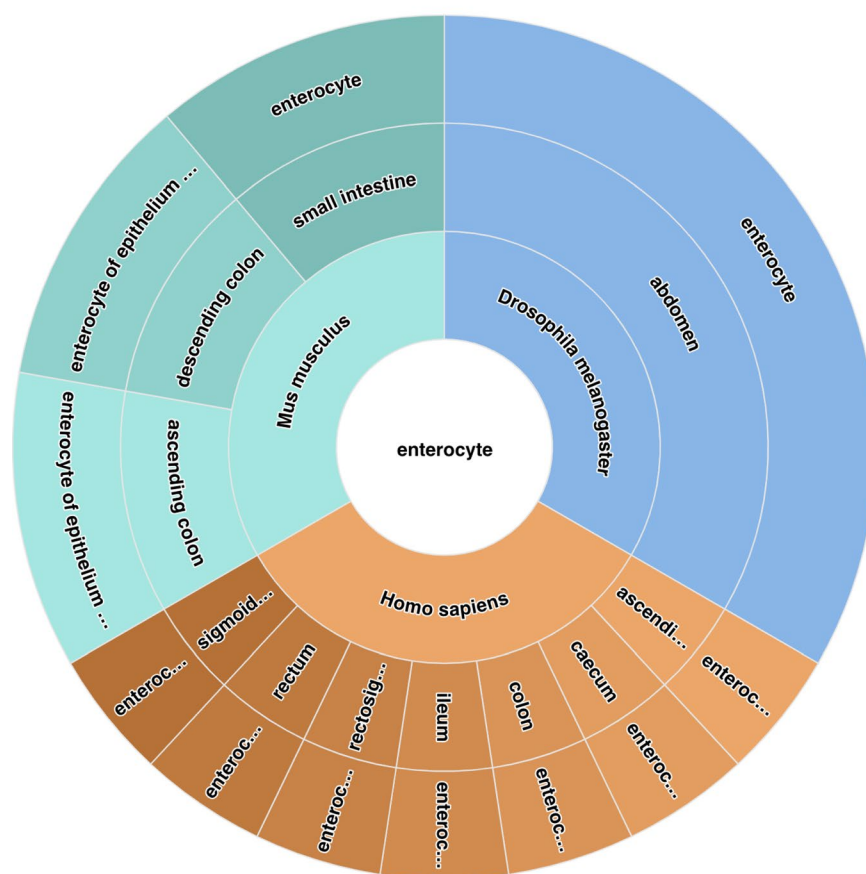


Fig. 4 Use of extended CL for cross-species data integration on the EBI single cell expression atlas. The central circle shows a general cell type (in this case enterocyte) found by lexical search on the atlas. The inner and middle wheel show the species and tissue of origin of the sample used as input to scRNAseq. The outer wheel displays subtypes of enterocyte used to annotate single cell transcriptomics data. Clicking on the outer wheel takes users to a list of highly expressed genes in the cell type and to individual studies. This search is made possible by an extended version of CL that folds in cell type terms from other, single-species ontologies (in the above example, the *Drosophila* Anatomy Ontology) via curated bridges between the two ontologies using SSSOM⁸² (see Methods for details).

Where transcriptomic analyses provide strong evidence for orthology, we are beginning to add grouping terms that unify cell types across species, as discussed further in the Discussion section.

CL support for integration across diverse species is well illustrated by the EBI Single Cell Expression Atlas⁵³. For example, a search for enterocyte, for which evidence exists for conservation between vertebrates and invertebrates⁵⁴, finds data annotated with specific types of enterocyte in *Drosophila*, humans and mice (Fig. 4).

While much recent usage of CL has focussed on annotating transcriptomics data, one great strength of CL and the ontologies that extend it is their utility for annotating data across modalities. This can perhaps be best seen in the use of the *Drosophila* Anatomy Ontology to annotate 3D images (>100,000 images from confocal and EM data), connectomics (>4m pairwise connections) and transcriptomics data (240 cell types, 1560 annotations, 2.5m gene expression links) using cell type terms that extend CL¹⁶. CL is used directly to annotate microscopy images and support their Findability on the Cell Image Library⁵⁵ and to annotate schematic images of cell types in anatomical context as part of the Single Cell Expression Atlas⁵³.

CL use in Automated Cell Type Annotation. Manual annotation of scRNAseq data is slow and labour intensive. In tissues for which high-quality annotated reference data is available, automated annotation is critical to efficiently interpreting scRNAseq data in the context of disease and experimental manipulation⁵⁶. CL is increasingly integrated into tools for automated cell type annotation including the Azimuth annotation transfer tool⁵⁷, CellTypist^{58,59}, and scTAB⁶⁰ which both feature mappings to CL. It is also used directly by multiple automated annotation tools as a source of knowledge to help drive annotation.

OnClass is an automated annotation system, built on annotated training data¹⁸, that uses CL as an input. Typically such systems are limited to annotating cell types that are present and annotated in their training data. OnClass takes advantage of the observation that relatedness in the ontology graph correlates with transcriptomic

similarity, along with similar correlations with text similarity of labels and definitions, to support annotation of cell types in transcriptomic data that are present in CL but not in the training data. The same team has further extended their work in a successor software that uses 8 different annotation methods, including OnClass, and CL, to create an ensemble of prediction models with a voting scheme⁶¹.

A similar approach is taken by Cello⁶², a discriminative, supervised machine learning approach for classifying clusters of cells in human scRNA-seq data. Cello uses the structure of CL combined with extensive training data to support automated hierarchical annotation with CL terms.

Sfaira^{17,60} is Python software designed to enable scRNA-seq data curation, ontology-based database queries and automated analysis with pre-trained neural networks using data from a variety of curated sources. It takes advantage of the structure of CL beyond the definition of sets of mutually exclusive cell types. Sfaira requires only that annotated cell types be related via a directed acyclic graph and so can take advantage of the multi-inheritance classifications hierarchies in CL. Sfaira also allows for ontology-aware subsetting queries, for example allowing for the selection of any T cell or subtype thereof, to build cross-study meta analyses.

Given that many standards like CELLxGENE⁴⁵ require the use of CL for cell type annotation, by providing mappings to CL, these tools greatly aid in preparing data for submission.

Applications for viewing and querying the Cell Ontology. The content of CL can be viewed on a variety of ontology portals (Bioportal^{63,64}), Ontology Lookup Service^{65–67}) and queried via their APIs. More powerful query options are supported by UberGraph.

As already described, CL is tightly integrated with other OBO Foundry ontologies, which makes powerful new types of queries possible, leveraging the knowledge encoded in multiple linked ontologies. For example, we can take advantage of the thousands of asserted and inferred links from Uberon to CL to query for all cell types found in a specific tissue. These queries are exposed through Ubergraph⁶⁸, which integrates a large and growing set of OBO Foundry ontologies and precomputes inferred relationships. The result is queryable via a SPARQL endpoint and a REST-API that wraps precomposed SPARQL queries. Wrapped queries include a query for cell type by location.

Community. In keeping with the ethos of the OBO Foundry²², CL is built as an open, community-driven resource where the democratization of, contribution to, and editing of the ontology is a key consideration. This approach supports distributed maintenance and extension by a team of editors, funded by and engaging with multiple projects and organizations with an interest in its development, including the BRAIN Initiative Cell Atlas Network (BICAN)⁶⁹, HuBMAP^{32,70}, the HCA⁴⁴, Cell Annotation Platform (CAP; <https://celltype.info/>), and the Chan Zuckerberg Initiative (CZI). As a result of this we have seen a large increase in engagement as measured by increased GitHub requests and substantial improvements in responsiveness since 2018 (see supplementary fig. 1 for details).

The HuBMAP Human Reference Atlas³² is a major source of expert input to CL content. Groups of subject matter experts work with HuBMAP curators and developers to curate hierarchies of anatomical terms, cell types and markers covering human anatomy, mapping these terms to CL and Uberon. These efforts are a major source of expert feedback on missing terms, missing relationships, term naming, and term definitions. Additionally, validating these curated hierarchies against both asserted and inferred relationships in CL and Uberon provides feedback and correction for both CL and the curated hierarchies⁷¹.

CL editors also work closely with the Human Cell Atlas data-integration efforts to develop integrated atlases for specific organs and tissues, combining multiple independently generated datasets. One early success of this effort is an integrated atlas of the human respiratory system⁷². Ongoing efforts are centered around annotation of integrated reference datasets on the Cell Annotation Platform (CAP) including gut, eye, liver, lung, kidney, and breast. This collaboration has created a productive feedback cycle: CL editors manually review annotations from CAP-hosted atlases, using publicly available spreadsheets to suggest more granular existing CL terms or flag missing ones that are then added to CL to improve coverage across tissues. This hands-on process directly improves CAP annotations and expands the ontology. For example, while reviewing the trabecular meshwork dataset on CAP, we identified gaps and subsequently added missing cell types such as “Schlemm’s canal endothelial cell”, “juxtacanalicular tissue cell”, and “beam cells”.

In a similarly productive cycle with Brain Initiative Atlases, CL editors have utilised LLM Deep Search methods^{73,74} with information from transcriptomic type annotations (markers, neurotransmitters, location, neuro-peptide) to search the literature to identify potential mappings between transcriptomic and classical types along with data and criteria that can be used to test those mappings. In an increasing number of cases, this supports both improved atlas annotation and the definition of terms in CL that unify transcriptomic and classical definitions. For example, we used Deep Search to identify papers and datasets that bridge classical interneuron types and transcriptomic types in the mouse Cerebellum. We then used annotation transfer to identify the same types in the BICAN mouse whole brain atlas, validating results against location from spatial transcriptomics. As a result we were able to improve annotation of the mouse whole brain atlas and types for mouse Lugaro and Candelabrum interneurons that bridge classical and transcriptomic definitions^{75–77}.

The Cell Ontology increasingly provides a bridge between these atlas efforts by defining and linking terms for common cell types found across these tissues. For example, CL has a general term for ‘goblet cell’, which classifies terms for more specialised goblet cells present in the gut, lung and eye. In parallel, an ongoing collaboration aims to mine CZI datasets (covering a much wider range of datasets than those covered by HCA and BICAN) for new terms and to improve definitions of existing terms, with input from CZI data curators and experts as part of CZI organised expert jamborees.

The BICAN-funded Human and Mammalian Brain Atlas (HMBA) is developing atlases of transcriptomic cell types across mice, humans and non-human primate brains, integrating these atlases across species. Extensive spatial transcriptomics and patch-seq data are being used to map the cell types defined in these atlases to brain regions, functions, and morphology. CL is working with the community developing these atlases to extend CL with classically defined cell types and, as consensus atlases become sufficiently stable, with data-linked transcriptomic cell types.

Supporting community input to CL. Building high-quality biomedical ontologies often requires close collaboration between expert biologists, who may not have the technical expertise needed for editing ontologies, and ontology editors/engineers, who may not be as well-versed in specific biological knowledge. CL has taken a number of approaches to help bridge this gap.

A range of GitHub issue templates guide users to provide the required information for various types of edit, including addition of new terms and synonyms and edits to existing definitions and relationships. For example, when requesting a new term, we now ask for a preferred term label, synonyms with references, a definition with reference, the anatomical structure the cell is part of with instructions to use UBERON terms, the submitter's ORCID, and any notes or concerns. Submitters providing ORCIDs get attribution on edited terms, ensuring due credit and fostering better community relations.

Alternatively, users can request batches of related terms using a spreadsheet based standard: Minimal Information Reporting About a Cell (MIRACL)⁷⁸. This has been successfully used to facilitate the addition of sets of terms specified by experts and curators from the HuBMAP Human Reference Atlas^{32,46,70}. A Google Sheets plugin is available to support this⁷⁹.

CL also supports the use of Knowledge Graph Change Language (KGCL)⁸⁰ to facilitate direct editing of CL by non-experts. KGCL describes ontology editing operations (e.g. renaming a class, changing its parent, adding a synonym, etc.) using a syntax that is very close to natural English language. We allow the use of KGCL instructions on our issue tracker, so that users can directly specify edits in tickets, for example including a KGCL instruction such as “rename CL:0002075 to ‘brush cell of tracheobronchial tree’”. The specified edits are automatically implemented (via GitHub Actions) and are instantly available for review by submitters and editors (via GitHub Pull Requests).

The Biomappings project⁸¹ enables community-driven curation of high quality, precise ontology mappings using the Simple Standard for Sharing Ontological Mappings (SSSOM)⁸². CL imports these using the ODK plugin for SSSOM, resulting in the addition of 81 mappings to the Medical Subject Headings (MeSH) and 74 mappings to Brenda Tissue Ontology (BTO) at the time of writing. Relatedly, CL has used the Bioregistry⁸³ to standardize references to external ontologies and vocabularies in its curated database cross-references.

Finally, we facilitate community review of proposed changes to CL (on GitHub Pull Requests) by displaying human readable summaries of changes to the ontology and their implications for automated classification of cell types. This makes review by editors and contributing experts easier, aided by GitHub social media functionality.

Improvements to quality control and release frequency. Recent technological improvements have made the process of editing and maintaining CL more efficient and consistent. By adopting tools like the Ontology Development Kit (ODK)⁸⁴, the editorial team can release updates more frequently—around once a month compared to one or two releases per year before. The ODK also supports automated quality checks that catch common technical issues before they are incorporated and ensure that release artifacts are compliant with the OBO Foundry principles²². Additionally, CL ensures its compatibility with other related ontologies through advanced tools like Ubergraph, which conducts daily checks for cross-ontology consistency⁸⁵.

Content. The Cell Ontology has significantly expanded over the past few years, with a total of 3228 terms and 3205 synonyms (release v2025-12-17) based on 1448 unique references⁸⁶. Since 2019, we have added over 1000 new cell type terms, and 1051 synonyms. Over 1100 definitions have been added or revised.

While many of these improvements to content have been driven by the needs of single cell transcriptomics atlasing projects, the majority⁸⁶ of this work to date has focused on adding classically defined cell types—specifying a general classification and a set of properties which distinguish that cell type and are practically useful in its identification. These definitions, as well as changes to names and the collection of synonyms, reflect evolving domain knowledge and terminology from various disciplines. This is sufficient to support annotation of single cell transcriptomics data wherever annotated cell sets map cleanly to classically-defined cell types (subject to consensus agreement, backed by multi-modal marker evidence) or when further investigation of cell types defined by single cell transcriptomics determines properties that can be used for new classical cell-type definitions in CL. An alternative approach, trialled in the Provisional Cell Ontology⁶⁹, extends CL with cell types and classifications defined by reference single cell transcriptomics atlases. This approach is especially important for tissues and where classical cell typing has been overwhelmed by complexity and scale of cell types and tissue structure. This is a major issue for neuron types in the brain and CNS cell types^{40,87}. It may also be key where transcriptomics is uncovering a wealth of previously uncharacterised diversity, for example, among fibroblasts⁸⁸ and amacrine cells of the retina⁸⁹.

CL also remains committed to representing vetted and standards-compliant cell types defined by various communities of biologists. This work runs in parallel with work to associate findings of scRNAseq and other omics approaches with existing cell types and identify novel cell types. In this section we discuss some of the key improvements in the content of CL.

Immune refactoring. Significant prior work has gone into representing immune cells of various lineages in CL^{24,90,91}, and it remains important for CL to stay up-to-date with cell type names that align with standard nomenclature used by immunologists. For example, innate lymphoid cells (ILC) are a group of related lymphocytes separate from B cells and T cells that have become widely studied in the past 15 years^{92,93}. ILCs lack the adaptive immune receptors of T cells and B cells and rely on combinations of other receptors for activation and inhibition. ILCs are divided into three main groups, the ILC1, ILC2, and ILC3, based on functional and marker-based criteria^{92,94}. Cell types within one ILC lineage, the natural killer cells, have been extensively studied over a much longer period and have been retrospectively grouped with the larger set of ILC1 cells. In collaboration with the Human Immunology Project Consortium and subject matter experts (Dr. Lewis L. Lanier), CL curators enriched CL with a full representation of these cell types, including both generic and human-specific forms axiomatized with their distinguishing surface markers and formal statements of species specificity. This resulted in the addition of new cell types, the revision of existing cell types, and the revision of the lymphocyte hierarchy to properly accommodate the ILC subgroup.

Unifying the representation of classical cell types across tissues. CL editors collaborate with experts in the HCA integration networks, HuBMAP and community meetings organised by CELLxGENE to ensure that the representation, naming and classification of classical cell types matches the needs of these atlases as closely as possible and is integrated into broader, cross tissue classifications.

This work includes ensuring term labels and synonyms correspond to current community usage (e.g. replacing the outdated label ‘pneumocyte’ with the more widely used ‘pulmonary alveolar epithelial cell’), or to new standards emerging from integration efforts. Editors improve cell type classifications and definitions, aligning them across tissues, enhancing links to the Gene Ontology, and improving the accuracy of anatomical location links. Efficient improvement of textual definitions is supported by the use of Large Language Models (LLM) and standardised prompts, which provide first-pass textual descriptions with links to the literature. Editors then check the assertions in these descriptions for accuracy and relevance, refining the definition. We also make use of OntoGPT⁴³ to mine the text for candidate links to Gene Ontology terms.

Textual definitions in CL are increasingly supplemented by longer, LLM generated, expert reviewed descriptions from CellGuide. These are further reviewed for inclusion in CL by editors, and only rejected where there are clear errors or where very general term definitions are simply enumeration of subclasses. To date, 434 out of the 607 reviewed Cellguide definitions (71%) have been accepted and integrated into CL, with the remaining 183 definitions awaiting future review and integration. While definition content needs to be relatively minimal, species-general, and apply to all subclasses, these descriptions have no such restrictions. They can therefore include broad, medically relevant context that would not be appropriate in a core definition. This is made clear by additional text attached to descriptions.

LLMs also help editors search for openly licensed figures to incorporate into Uberon and the Cell Ontology. Schematic figures that illustrate the location and anatomical context of cell types are particularly valuable for clarifying the meaning of terms to both editors and users. These figures are linked to terms via a standard schema supported by ontology portals including the Ontology Lookup Service, which makes linked figures visible on term pages.

A major focus of recent work has been on mucosal epithelial cells across many tissues including brush border cells, found in the choroid plexus, kidney tubules, and gut (enterocytes); goblet cells found in multiple mucosa, multiciliated epithelial cells found in many organs⁹⁵ and tuft cells. Work on tuft cells (also known as “brush cells”) across various epithelial tissues and organs—including those in the gut, respiratory tract, gallbladder, thymus, pancreas, salivary gland and urethra—illustrates this approach well. Tuft cells are characterised by a distinctive “tuft” on their apical surface, allowing them to serve as chemosensors and modulate immune responses across tissues⁹⁶. Through recent edits, we have transformed an incomplete set of terms that previously lacked key cell types, had missing links to GO, and only connected to general anatomical locations into a unified hierarchy of 23 tuft cell terms across 10 different organs. We added 12 new CL terms and revised existing terms within the hierarchy—ensuring each term has a detailed definition, appropriate links to GO terms (tuft, cytoplasm), and a part of the relationship to a specific epithelium. Accompanying work in Uberon has ensured that the epithelia themselves are represented and classified correctly.

Our improvements highlight the unique characteristics and diverse roles of tuft cells across different tissues. In the thymus, tuft cells are distinguished by their lateral microvilli structure and MHC class II expression⁹⁷. In recent edits, we have added this term and linked it to a richly-annotated GO process term ‘antigen processing and presentation of endogenous peptide antigen via MHC class II’, highlighting its role in central tolerance. In the intestine, we have expanded developmental relationships, linking intestinal tuft cells to intestinal crypt stem cells, reflecting their potential as reserve stem cells after damage⁹⁸. In the nasal cavity, ‘tuft cell of olfactory epithelium’⁹⁹, has a definition leveraging both single cell transcriptomics data and microscopy data using markers uncovered by transcriptomics, and knockout phenotypes. The resulting definition includes molecular, functional properties, and location, as well as recording the terms used in scRNAseq annotation as synonyms. These changes collectively set a precedent for our future work on unifying the representation of cell types across tissues in CL.

To streamline the process of adding terms to CL, generating textual definitions for tuft cells across the hierarchy was supported by the use of LLMs with standardised prompts. A manual editing step was included to validate the provided literature and refine the definitions as needed. An example of thymic tuft cells is illustrated in Fig. 5.

This process of refactoring and extending CL has been driven by the implementation of unified design patterns for terms. One critical aspect has been unifying the general cell type (genus) terms used in logical definitions, which is critical to ensure that classification can be automated. For example, in earlier version of CL, many

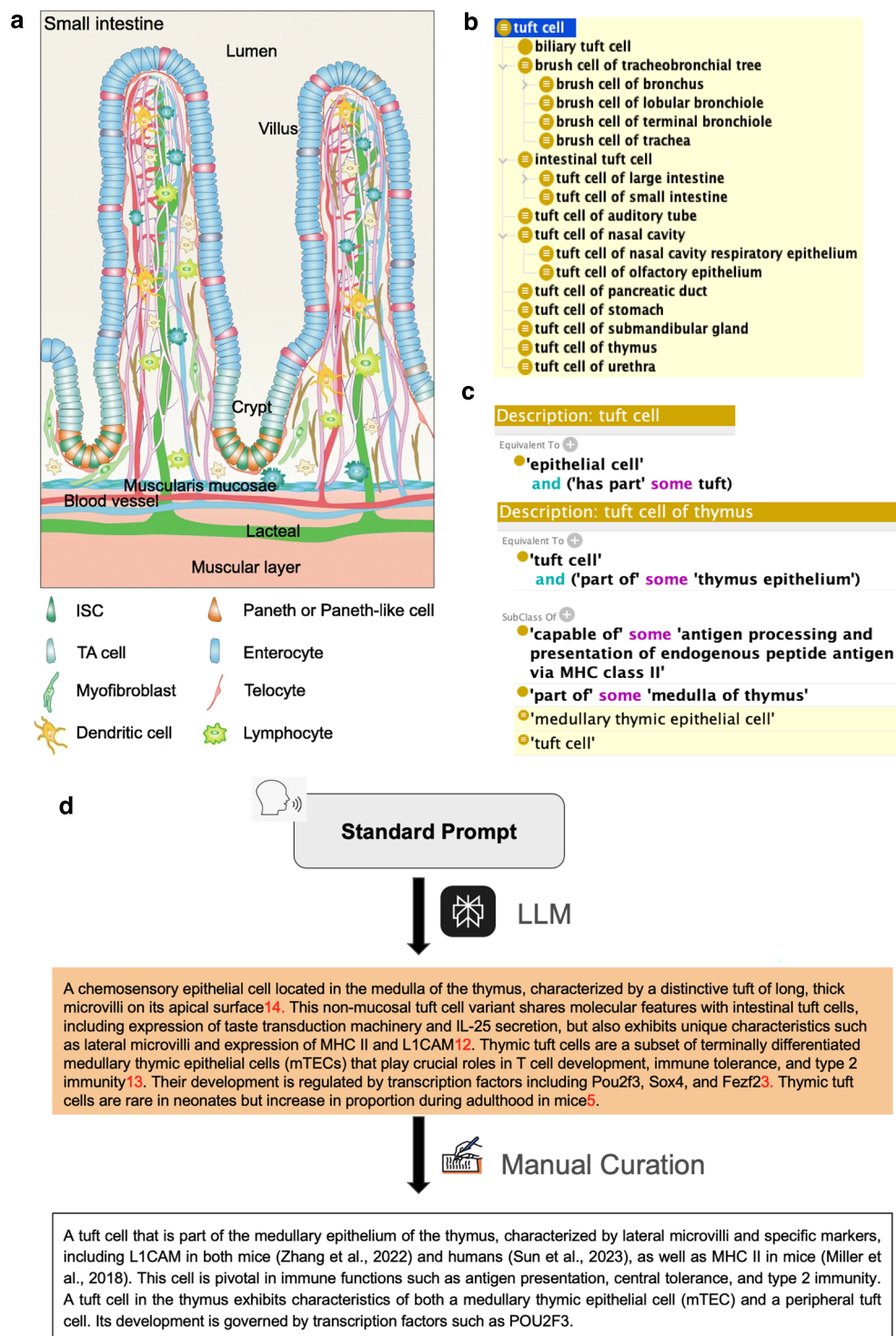


Fig. 5 **Panel a:** A schematic of cell types in the intestinal villi of the small intestine (adapted from Zhu *et al.*¹⁴⁰, licensed under CC-BY-4.0). This schematic is linked to gut cell type terms within CL. Many of these cell types are specific examples of more general types found across tissues, e.g. tuft cell and goblet cell. Edits to CL have focused on standardising the naming, representation and classification of these cell types across tissues—linking the terms to specific epithelial locations and GO terms for parts and functions. **Panel b:** A unified hierarchy of tuft (bush) cells. **Panel c:** Logical definition of the term 'tuft cell' and of one of its subclasses, 'tuft cell of thymus', following a standard design pattern. **Panel d:** An example for generating and refining cell type definitions in CL using large language models, illustrated by the tuft cell of thymus.

epithelial cell types, including multiciliated cells, were manually classified using a mix of subtypes of epithelial cell differentiated by origin (e.g. endo-epithelium), morphology (e.g. columnar), anatomical location or the

presence of specialized cell components. As we work on improving epithelial cell type definitions for specific tissues or cross tissue type, we switch to classifying with a general epithelial cell term and classifying dynamically by the properties asserted for individual cell types using an OWL reasoner. Figure 5 shows some example patterns for multiciliated cells.

To enable correct automated classification with a reasoner, sometimes an overhaul of more generalised classification is needed. For example, CL previously classified endothelial cells as a type of epithelial cell. While this is a defensible position, the majority of the biological literature and working biologists treat epithelial and endothelial cells as distinct and non-overlapping categories. By switching CL to this view we are now able to dynamically classify epithelial cell types without automatically classifying endothelial cells under them. This work resulted in automated reclassification of over 40 cell types.

Extending the Cell Ontology with T-types. Single cell transcriptomics data provides an opportunity to root cell type definition and classification in unbiased analysis of data covering near-complete profiles of gene expression across large populations of individual cells. CL is therefore committed to incorporating T-types—cell types largely defined by their shared transcriptional profiles. CL incorporates T-types based on an interconnected set of evidence criteria. First, we prioritize multimodal evidence beyond transcriptomics alone, incorporating anatomical location, morphology, or functional properties wherever possible. Second, we require that T-types be defined in an integrated sc/snRNAseq atlas from multiple sources or be supported by multiple independent atlases. Third, we require community agreement on reference datasets, typically mediated by official organizations such as the Human Cell Atlas, LungMap, or the Brain Initiative Cell Atlas Network (BICAN). Importantly, we treat all mappings as hypotheses and provide explicit documentation of supporting evidence, allowing users to assess reference quality for themselves.

CL terms require clear definitions that distinguish the cell types they denote from other cell types and can guide accurate annotation of data. Before we can generate definitions for T-types, we need sufficient community consensus on which reference data and analysis to use. Defining a T-type with reference to multiple, independently generated datasets is challenging. While independently generated datasets may be mapped to each other via annotation transfer following statistical matching⁵⁶, alignment is often imperfect due to experimental and computational differences. These include differences in anatomical regions used as input, e.g. whole cortex vs cortical region¹⁰⁰, single cell vs single nucleus¹⁰¹, library preparation and sequencing differences. And some annotation transfer tools, like Azimuth and CellTypist, force annotations of novel cell types onto related cell types present in their reference⁵⁶. Clustering can differ due to study-specific choices of parameters for dimensionality reduction, nearest neighbor search, community detection (Louvain or Leiden), nonlinear embedding (tSNE or UMAP), and iterative/hierarchical clustering scheme (different resolution in each iteration)¹⁰².

Given this variation, we rely on community agreement on reference datasets, typically a single annotated dataset or integration of multiple datasets sanctioned via an official organization such as the Human Cell Atlas, LungMap, or the Brain Initiative Cell Atlas Network (BICAN) and released with cleanly referenceable versions.

Once we have a reference dataset, we define T-types via links to reference data (in versioned reference datasets) and via marker sets derived from this reference data, folding in additional classical properties where these map clearly, as illustrated in Fig. 6. Links to data follow a formal schema established in prior work on the Provisional Cell Ontology^{69,103}. Marker sets are recorded along with key contextual information including species and anatomical context and confidence scores (F-beta, precision, and recall) calculated using NS-Forest¹⁰⁴, providing an objective measure of how useful marker sets are in identifying a cell type within the reference dataset. By selecting sets of markers, NS-Forest can provide the highest possible confidence marker sets for a given tissue context. Given that NS-Forest marker sets vary considerably in their F-beta scores, depending on data quality, cluster separability, and biological context, we provide the F-beta scores alongside the marker sets so that users can make informed decisions in their own analytical context. Wherever possible, we also fold transcriptomic classification into CL using the schema established for PCL⁶⁹. Figure 6 shows an example and also shows how single cell transcriptomics can inform classification. Two transcriptomically distinct populations of GABAergic cortical interneurons express LAMP5. Classifying based on properties, we could define a Lamp5 neuron as any CGE derived cortical GABAergic interneuron that expresses LAMP5. However, doing so would define a class that subsumes Lamp5 Lhx6 neurons. In this case and for transcriptomic types in general, the transcriptomic hierarchy is primary.

Naming of T-types is often challenging. Where reference datasets are the result of dataset integration efforts, as is the case for the Human Cell Atlas integration efforts including the Human Lung Cell Atlas¹⁰⁵, or naming is based on extensive annotation transfer, as in the Brain Initiative, the variety of names used in different source datasets need to be reconciled. In some cases, a large number of cell subtypes are identified, as is the case for amacrine cell subtypes in the Human Retina Atlas⁸⁹, where CL editors are using cell type-specific marker genes from NS-Forest¹⁰⁴ combined with the parent cell type in the ontology as an objective means for cell type naming. CL editors are actively involved in many of these naming efforts as part of annotation jamborees. A successful example of this approach is the Breast Atlas Jamboree organized by CZI, which facilitated data integration across studies^{89,106}. This effort resulted in clusters that were previously named differently being unified under a single label approved by the mammary gland community¹⁰⁷. Annotation of single reference datasets rarely provides names that are clear and unambiguous in the broader context of cell types defined in CL. In these cases, CL keeps names of cell types from reference datasets as synonyms and official symbols and generates longer, more informative official names that frequently include transcriptional biomarkers for these T-types.

Taxon constraints. As CL aims to cover cell types across broad ranges of species, care must be taken to ensure that the ontology does not assert anything about a cell type that is not valid in all the species where that cell type is known to exist.

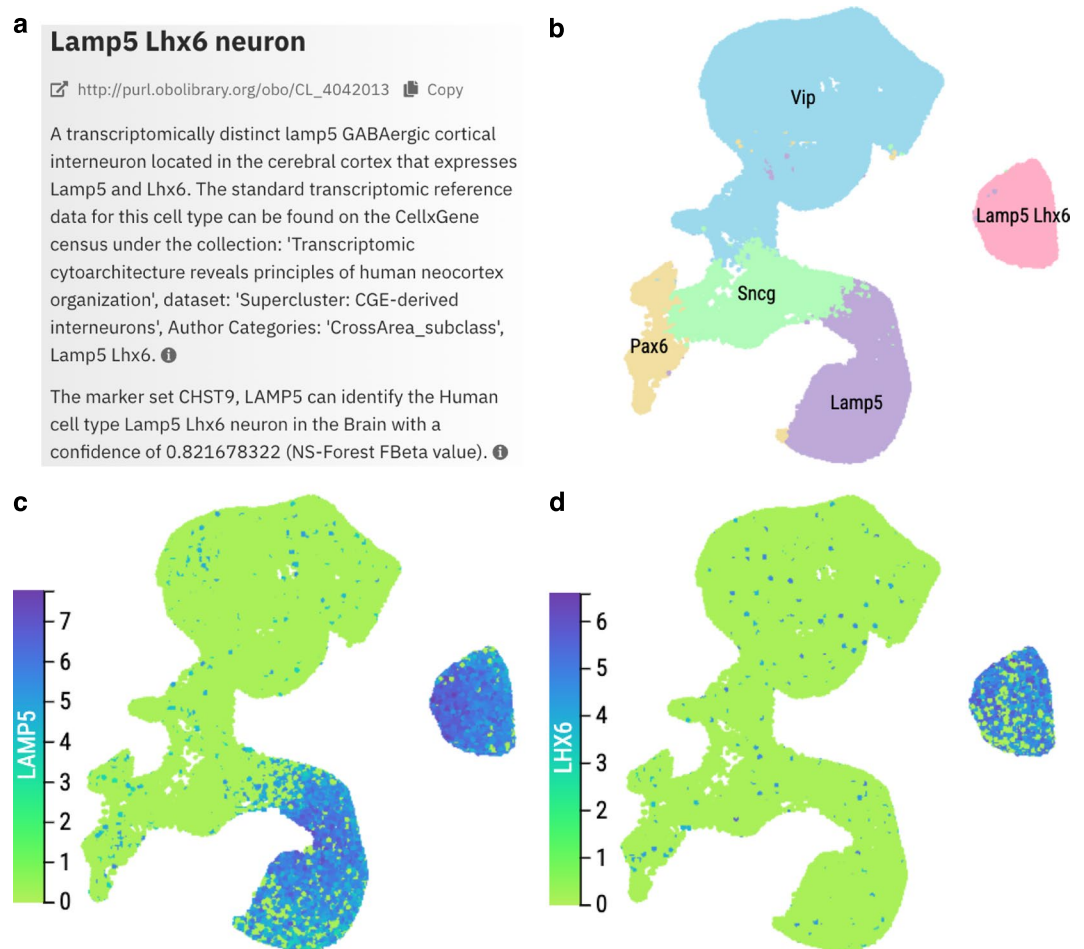


Fig. 6 **Panel a** shows the definition of a T-type (cell type defined in part by a transcriptomic reference) Lamp5 Lhx6 neuron. Note the actionable link to transcriptomic data (linkouts include a dataset version specific link to data on CELLxGENE, not shown) and the NS-Forest markers derived from data¹⁴¹, confidence score and relevant tissue context. **Panel b**: UMAP of transcriptomic clusters in the reference dataset: CGE derived cortical interneurons from integrated samples from across the human cortex¹⁰⁰. **Panel c**: LAMP5 expression. **Panel d**: LHX6 expression. Note that Lamp5 Lhx6 forms a distinct transcriptomic cluster that is not a subset of LAMP5 expressing cells.

For example, ‘taste receptor cell’ (CL:0000209), which is defined as a chemoreceptor cell involved in the sensory perception of taste, is a very generic cell type that has a broad applicability across the animal kingdom (many animal species have such cells). It would therefore be wrong to assert (as has been the case in earlier versions of CL) that taste receptor cells are part of the ‘tongue’ (UBERON:0001723), because not all species that have taste receptor cells have them in the tongue, or even have a tongue to begin with. Notably, arthropods do have taste receptor cells, yet lack a tongue, which is an anatomical feature specific to chordates.

To facilitate the detection of such incorrect assertions, CL formally describes whenever possible the known taxonomic constraints¹⁰⁸ that apply to a cell type, such as “this cell type is known to exist in that taxon”, “this cell type is known *not* to exist in that taxon”, or “this cell type is known to be *specific* to that taxon” (see Methods for details on how those constraints are formalized). This makes it possible for a reasoner to flag any assertion that would violate said taxonomic constraints, and thereby allows CL editors to fix such an incorrect assertion before it is even committed to the ontology.

CL bridges classical definitions and transcriptomic definitions. Links between GO and CL are an important part of defining terms in both ontologies. In CL they constitute a formalised part of the classical definition, defining key functions and cell components. By combining annotation of genes to these GO terms with transcriptomic data, we can extract the transcriptomic signatures of these functions and components. We ran this analysis across markers from the majority of which come from single cell data. We assessed markers from CELLxGENE, based on automatically generated transcriptomic data annotated with CL terms³⁵ and CellMarker2.0³⁴, supported by high confidence annotation of genes to GO terms (those supported by experimental or evolutionary evidence, see methods for details) with direct links to CL terms. This approach finds transcriptomic signatures (600 markers in total; 315 from CELLxGENE) for 252 distinct structures and functions of 136 cell types

supported by 428 references. An example of this is shown in Fig. 1. Complete results are shown in supplementary tables S2 and S3.

In future we will extend this analysis to broader gene expression from transcriptomics data, concentrating on well-validated reference datasets to provide more general transcriptomic signatures of defining structures and functions of cell types. We will also extend it to take advantage of the broader semantics of GO, for example leveraging relationships between Cellular Components and Biological Processes for their assembly and maintenance and functions. We will also make it a priority to extend GO-CL links, leveraging agentic AI approaches to make this efficient. This will make an increasingly rich set of transcriptomic profiles for the functions of cell type defined in CL available through the CL knowledge graph or via independent analysis.

Discussion

Data-linked definitions. As cell typing becomes more data driven, CL faces new challenges in defining and naming cell types in a way that is relevant to data-driven annotation. CL and its PCL extension have risen to this challenge by developing and using formal data-linked cell-type definitions: computationally tractable links to reference data that can be used to retrieve and analyze this data and use it as a reference in automated annotation. At a minimum this requires resolvable links between ontology terms and files containing the relevant reference data and mechanisms to refer to specific defining data for a cell type. Data archives such as Nemo¹⁰⁹ currently lack sufficient granularity for referencing files, and standardization of files is not always sufficient to unambiguously reference data within them. The development of resources, such as CELLxGENE Census, that harmonize representation of single cell transcriptomics data across large numbers of datasets and provide APIs for accessing data provides easy solutions to this problem, as long as the reference data is deposited on such a resource.

Choosing reference data is also a challenge. Reference standards need to be sufficiently stable—a reference that changes rapidly in radical ways without clean mappings and migration paths between versions is not suitable for supporting reliable searching and browsing in integrative resources such as CELLxGENE Discover. For this reason, we have mostly so far restricted these types of definitions to the PCL. However, standard and relatively stable reference datasets are beginning to emerge from the work of the Human Cell Atlas integration efforts and the BRAIN Initiative. For example, CL uses standard references for gross cell types in the human cortex¹⁰⁰ to define cortical cell type via a combination of textual definitions and formal links to data hosted on CELLxGENE Discover (see Fig. 6).

T-type hierarchy and naming. Decisions on nomenclature are challenging for any ontology, but the rise of single-cell transcriptomics, the use of clustering in latent space to define T-types, and the increasing use of annotation transfer have brought new challenges. Classically defined cell types are typically named for their intrinsic and extrinsic properties. For example, L4 IT neurons in the cortex were originally named for their intra-telencephalic projection pattern (IT) and soma location in cortical layer 4 (L4). Similarly, VIP interneurons in the cortex were originally named for their expression of Vasoactive Peptide. However, recent papers have published T-types with L4 IT in the name that refer to neurons in regions with no layer 4^{100,110}, and T-types with VIP in the name where VIP expression is no higher than in many interneuron types not named VIP¹¹¹. These names make perfect sense based on annotation transfer and clustering, but potentially violate community expectations around the meaning of names.

This poses a challenge for standard ontology definitions, where properties in the name are typically reflected in properties asserted for members of the class. One potential solution, trialed in PCL, is to append ‘-like’ to terms defined by links to reference transcriptomics that have subclasses that do not share the properties in the name. For example we could use L4 IT-like for the L4-IT T-type named above. However, this results in terms that do not reflect usage. Alternatively, CL can accept that names do not necessarily reflect actual properties and expression, acknowledging this in the formal structure of the ontology (only linking VIP expression to T-types where expression is above background) and making exceptions clear in comments attached to terms.

Transcriptomic types are also frequently organised into hierarchies based on transcriptomic similarity. This might be overall transcriptomic similarity based on highly variable genes or a more restricted subset of gene expression, such as transcription factor expression profiles^{41,110}. These hierarchies are arguably more scientifically grounded than traditional property-based classification and so it is important to represent them in cell ontologies. PCL already represents many such hierarchies representing assertions of homology between cell types using simple pairwise relationships⁶⁹. With the emergence of increasing consensus on types in the mammalian brain, CL plans to integrate standard reference T-types and hierarchies from BICAN.

However, there are dangers in assuming that these hierarchies reflect simple subClassOf hierarchies down which properties are inherited. Markers that can be used to identify a very general T-type with high confidence may not be expressed in rare types nested under them in the hierarchy. Similarly, assertions about location or function based on mapping to spatial transcriptomics and patch-seq to one level in a taxonomy may not hold for all subtypes. This can be solved by moving many of these assertions to non-logical relationships (annotation properties), which are not inherited down the subclass hierarchy, or by replacing subClassOf relationships between hierarchical T-types with a non-logical relation that specifically reflects transcriptomic hierarchy.

Bridging modalities with data. Our understanding of the degree of correspondence between transcriptomic and classically defined cell types is likely to increase with the use of multimodal techniques that provide a link to anatomy, morphology, and function. For example, location plays a critical role in defining and mapping many cell types in a combined single cell and spatial transcriptomics atlas of the mouse brain—classical markers that are insufficient to identify cell types in the context of the whole brain effectively identify them in a specific

anatomical location (Fig. 2)^{41,112}. Similarly, morphology plays a critical role in defining and mapping cell types in a combined patch-seq and single cell atlas of the mouse motor cortex^{10,112}.

Challenges of harmonizing cross-species cell types in CL. General cell types are typically very well conserved across vertebrates even when their subtypes show species-specific specialization. This is most apparent in organs whose function is under strong selective pressure. For example, the retina is specialised in different mammals to support the acuity, color vision, motion tracking, contrast detection and energy efficiency required for their survival^{113–115}. The three-layer structure of the retina and its major cell types (rod and cone photoreceptors, Müller glia, amacrine cells, bipolar cells, retinal ganglion cells and horizontal cells) is conserved across vertebrates¹¹⁵. However, there is significant variation in the structure, function, distribution, and connectivity of subtypes of these general classes particularly among the amacrine and retinal ganglion cell (RGC) subtypes.

Until recently, we had limited ability to map these more specialised cell types across species, relying on morphology, function and a limited set of markers to do so^{116,117}. As a result, there is a proliferation of RGC terms in CL derived from studies of single species, with only very limited cross species grouping. There are 68 RGC terms in total, the majority from studies in mice, but with non-overlapping sets of terms for rabbits and primates (including humans).

Single cell transcriptomics now provides data-rich evidence for cell type homology across species. For example, Hahn *et al.*¹¹⁵ define RGC “orthotypes” based on single cell RNAseq analysis of 12 different mammalian species. Using this approach they find strong evidence that the midget ON & OFF RGCs of primates are homologous to much rarer cell types found in the mouse retina (c42 and c43) that have decreased in size and increased in number as primates evolved. Similarly, the Brain Initiative Cell Atlas Network is generating cross-species consensus atlases integrating human and non-human primate brains, mapping types to mice where possible. This has resulted in the recent publication of a cross-species consensus atlas of neuron types in the Basal Ganglion¹¹⁸.

CL has recently added cross-species types, defined in part via links to reference atlas for both RGCs and basal ganglion neurons using standard formal schemas for linking⁶⁹. However, we exercise caution in interpretation: while transcriptomic similarity provides strong evidence for homology, it could reflect convergent evolution, particularly between distantly related taxa. These grouping terms link to the data and analyses supporting homology assertions, representing an initial step toward distinguishing homologous and analogous cell types without yet encoding this distinction explicitly in the ontology structure.

Markers. While CL has always supported recording cell type markers, the proliferation of markers from omics data and their dependence on tissue context assay and analysis algorithms means that it is increasingly challenging to record markers consistently and comprehensively enough within CL to meet community needs without bloating CL content.

LLM pipelines show some promise in direct annotation of cell types with reasonable free text terms¹¹⁹. However, these lack the advantage of controlled annotation, are currently only accurate for very general cell types and are potentially prone to hallucination in the absence of curation.

To support community need for queries of markers linked to CL, we are building a Knowledge Graph extension to the Cell Ontology that folds in multiple sources of markers, including our own LLM aided curation, as well as anatomical and experimental context, evidence and provenance, including links to reference data and data driven validation via links to analysis of data on CELLxGENE. Where available, we will also fold in NS-Forest markers^{104,120}, which provide machine learning-based optimal combinations for identification of markers in the context of a dataset, along with a confidence score for marker combinations relative to the dataset that they are derived from. This, we believe, provides a practical and objective standard for markers, while allowing access to a broad range of markers used by the community.

The knowledge graph API will support searching for cell type markers by anatomical context, taking advantage of Uberon structure to support a wide range of granularities. Results will be ranked based on the weight of supporting evidence and can be filtered for official sources (e.g. HCA official integration markers). The provenance model employed will support full attribution and links to all original sources.

Integration of large language models. Large Language Models (LLMs) are now a core component of our ontology editing pipelines. Editors primarily use Deep Search⁷³ to generate draft cell type definitions with literature references, then review the output — checking the references, removing unsupported or irrelevant assertions and ensuring definitions align with CL standards. Rather than relying on one reference at a time, LLMs can quickly summarise current knowledge across multiple sources around a cell type, speeding up the drafting process. With well-engineered prompts, editors can produce definitions that closely follow the standards of CL. However, research on LLM-generated definitions shows that users with lower domain expertise are more likely to accept a generated definition on face value, even if incorrect⁴², so expert review remains essential. While Deep Search minimises LLM hallucinations, this can be further reduced by integrating it with Retrieval Augmented Generation (RAG), which automatically checks generated referenced text for consistency with references⁷⁴.

To directly address this, we have developed **CLARA (CL Accuracy Reference Agent)** (<https://github.com/Cellular-Semantics/clara-agent>), an agentic pipeline built using PaperQA⁷⁴ that validates both formal and textual definitions against their supporting references. CLARA automatically retrieves references associated with a candidate term, breaks the text into individual assertions, and checks whether each assertion is supported or contradicted by the literature. In an initial pilot evaluation of 689 assertions, CLARA showed high precision and recall (over 0.9 for both). Future work will focus on more extensive validation of CLARA and integrating CLARA into GitHub Actions so that every proposed new term or edit in CL — including those generated by LLMs — is automatically validated, with any errors flagged before an editor reviews. This provides an additional layer of quality control and significantly reduces the risk of hallucinated or unsupported content entering the

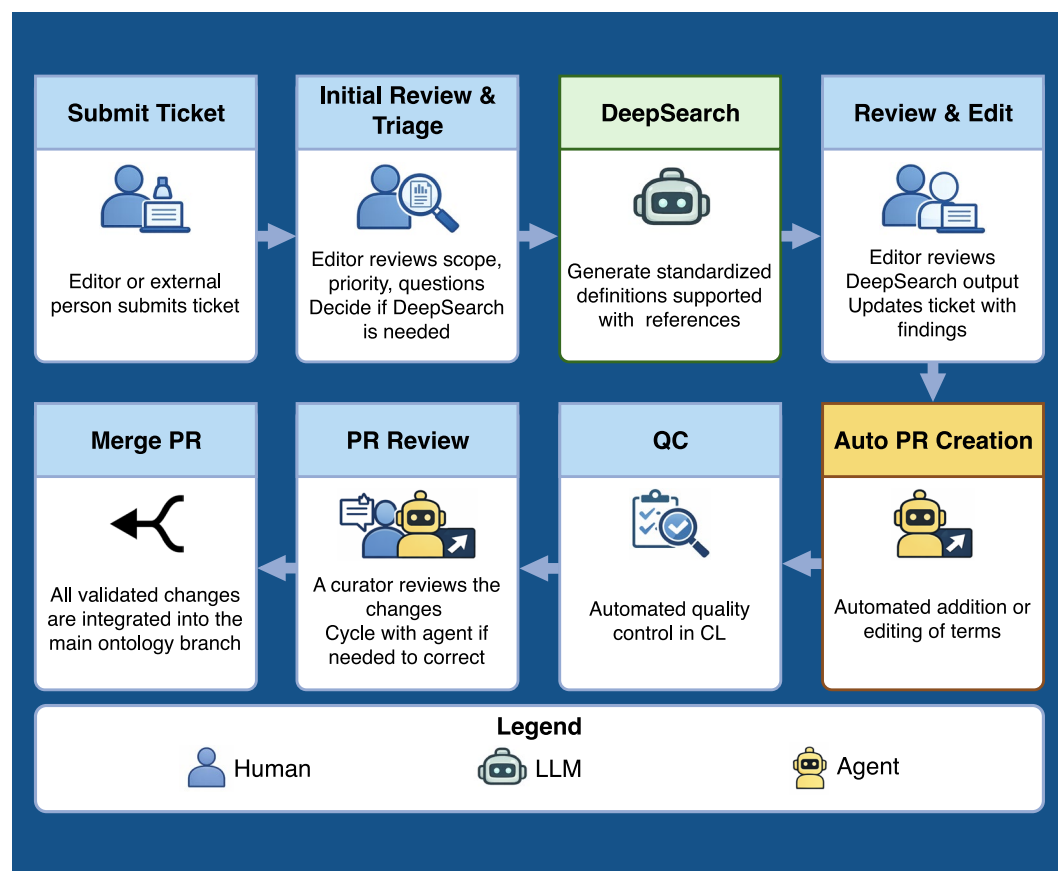


Fig. 7 Workflow for the creation of new ontology terms. A ticket is submitted by an editor or external contributor describing the proposed term(s). An editor then performs an initial review and triage to assess scope, priority, and outstanding questions, and to determine whether automated literature support is required. When appropriate, Deep Search is used to generate standardized draft definitions supported by literature references, which are subsequently reviewed and edited by the editor. Once all required details are present in the ticket, an AI agent implements the changes in the ontology and creates a Pull Request (PR). The ontology edits then undergo automated quality control to ensure that no structural or semantic errors are introduced that could affect reasoning or downstream data integration. Finally, a human curator reviews the PR and iterates with the agent until the changes are approved and merged into the main ontology branch.

ontology. Importantly, all content generated using LLM and agentic tools undergoes the same rigorous expert review process as the manually curated content prior to inclusion in CL.

Agentic AI is now fully integrated into our ontology editing workflows using GitHub Copilot (Fig. 7). These agentic tools operate within GitHub to interpret issues created by editors and automatically generate OWL edits via Pull Requests (PRs), allowing editors to focus on review rather than drafting changes from scratch. To date, 73 PRs created by GitHub Copilot have been merged into CL, showing how agentic tools can speed up routine ontology editing. Although they are currently used for well-defined tasks, these tools have the potential to handle broader, higher-level instructions and turn them into accurate, context-appropriate ontology updates as the technology continues to mature. In parallel, we are exploring multi-agent frameworks such as Aurelian (<https://github.com/monarch-initiative/aurelian>), which coordinate complex tasks such as literature review, semantic annotation, and ontology navigation. These emerging approaches may help automate literature extraction for new cell types, propose candidate definitions and mappings, and support consistency checking across ontologies such as GO and Uberon. Ultimately, adopting scalable, agent-assisted curation workflows will be key for future data integration, allowing CL to evolve quickly and keep pace while providing accurate, standardised grounding for the rapidly growing number of cell types emerging from single-cell atlases.

Future prospects. CL is a critical resource for making data about cell types FAIR. It provides consistent, well-defined cell type terms for large scale atlasing efforts, particularly those built on transcriptomics. By providing stable identifiers with machine readable axiomatization, CL provides the underpinnings for semantic data products that can be reliably reused, combined, and queried. Such semantically aligned data products not only enable symbolic reasoning, but are particularly well suited for connectionist AI systems which benefit from input data being structured, aligned, and grounded in ontologies^{121,122}. Bridging classical and transcriptomic definitions it distills key prior knowledge about cell types using both well-referenced text and formal assertions (e.g. via links

to the Gene Ontology. Increasing adoption of agentic workflows to edit CL combined with extensive community engagement are rapidly improving the scalability of CL development and its coverage of biology.

Rapid advances in Machine Learning (ML) and agentic AI are transforming the landscape of single cell research, bringing a constantly evolving set of new tools supporting data integration, analysis and interpretation. This generates new challenges and opportunities for CL use and further development.

Major training datasets for Machine Learning are ready-annotated with CL terms, the most prominent example being the CELLxGENE corpus. AlphaGenome¹²³ is a model for predicting cell type and tissue specific prediction of regulatory effects on mutations, and uses the CL both in the preparation of training data, and at inference time. New ML models built on multi-modal embeddings identify cell types, the tissue ‘niches’ in which they are located and both known and uncharacterized gene programs that underpin their structure and function^{124–127}. At least some of these models use morphology and transcriptomics to define cell types¹²⁴ arguably reflecting a more classical approach to cell type definition than recent work in single cell transcriptomics.

New agentic frameworks are capable of running analysis and annotation pipelines and aiding atlas interpretation using information mined from the literature and public databases and using gene enrichment analysis to map to pathways and biological processes^{128–130}. At least some also use the cell ontology directly for annotation^{128,129} or indirectly via tools that do¹³⁰.

With all these advances, the need for high-quality standard frameworks as a locus of integration for data and prior knowledge remains. CL is uniquely well positioned to play this role while the workflows developed to extend and adapt it will ensure it has sufficient coverage and quality to remain relevant.

Methods

Implementing human readable diffs on GitHub PRs. In order to simplify the review process, a GitHub actions automation automatically posts a comment on pull requests into the main branch, when given a specific command in the comment, that shows a human-readable report of differences between both unreasoned and reasoned files in the pulling branch and the main branch, prepared using a WHELK reasoner¹³¹). This is achieved by using the ROBOT “diff” function through the ODK on GitHub actions^{84,132}.

Implementing taxon constraints. CL, in conjunction with the Uberon anatomy ontology, makes use of a handful of properties to formally represent known constraints between a cell type and a taxon. They are (1) “in_taxon” (RO:0002162), used to indicate that a cell type is specific to a given taxon; (2) “never_in_taxon” (RO:0002161), used to indicate, conversely, that a cell type does not exist in a given taxon; and (3) “present_in_taxon” (RO:0002175), used to indicate that a cell type is known to exist in a given taxon, but without implying, contrary to RO:0002162, that it is specific to that taxon.

Those properties are intended to be used conjointly with a specific variant of the NCBI Taxonomy ontology that contains additional disjointness axioms between sibling taxa, and that can be used by an EL reasoner to detect whenever taxon constraints are violated.

RO:0002161 and RO:0002175 are annotation properties that are “expanded” at build-time into logical axioms by ROBOT’s “expand” command, according to a principle similar to OWLTools’ macros¹³³. An annotation assertion like “CL:0002174 never_in_taxon NCBITaxon:50577” (“follicular cells of ovary don’t exist in insects”) is expanded into the following two axioms:

CL:0002174 DisjointWith: in_taxon some NCBITaxon:50577

CL:0002174 SubClassOf: in_taxon some (not NCBITaxon:50577)

Conversely, an annotation assertion like “CL:0000711 present_in_taxon NCBITaxon:9606” (“cumulus cells are known to exist in humans”) is expanded into a “witness class” *W* such as

W SubClassOf: CL:0000711

W SubClassOf: in_taxon some NCBITaxon:9606

The purpose of that witness class is solely to deliberately cause an incoherency in the ontology, should the taxon constraint ever be violated (for example, if CL:0000711 ended up being mistakenly classified as a subclass of a cell type that is specific to a taxon disjoint with humans).

Integration with taxon-specific ontologies. CL maintains mappings between its own terms and equivalent terms in taxon-specific ontologies. Historically and in many OBO ontologies, inter-ontology mappings have been represented by *oboInOwl:hasDbXref* annotations (aka “cross-references”). Recently, the Simple Standard for Sharing Ontological Mappings (SSSOM) has been devised to represent cross-ontology mappings without the shortcomings of cross-references⁸². The *Drosophila* Anatomy Ontology (DAO) has already switched to using SSSOM for its mappings with Uberon and CL, and CL has been updated to automatically fetch and use the DAO-provided, SSSOM-formatted mapping set. SSSOM brings several benefits over the old cross-reference format, such as the possibility to use meaningful mapping predicates from the SEMAPV vocabulary¹³⁴, specifically intended to represent cross-species mappings, and the possibility to annotate each mapping with rich provenance metadata. We are currently using the DAO-CL mappings as a testbed for using SSSOM-maintained mappings, and we hope to progressively generalize the use of SSSOM to the mappings between CL and all the other taxon-specific ontologies.

Regardless of their formats (cross-references or SSSOM), all mappings are ultimately used to generate “bridging axioms” between CL and the taxon-specific ontologies. For example, a mapping between CL term for ‘neuron’ (CL:0000540) and the corresponding term in DAO (FBbt:00005106) is derived into an equivalence axiom of the form *FBbt:00005106 EquivalentTo: CL:0000540 and (part_of some NCBITaxon:7227)*, in effect stating that a fly neuron (FBbt:00005106) is a generic neuron (CL:0000540) that is part of a fruit fly (NCBITaxon:7227). This ensures that when CL and DAO are merged together, the DAO term is correctly positioned within the resulting merged hierarchy (Fig. 8). We expose these mappings in a number of ways in order to provide infrastructure for

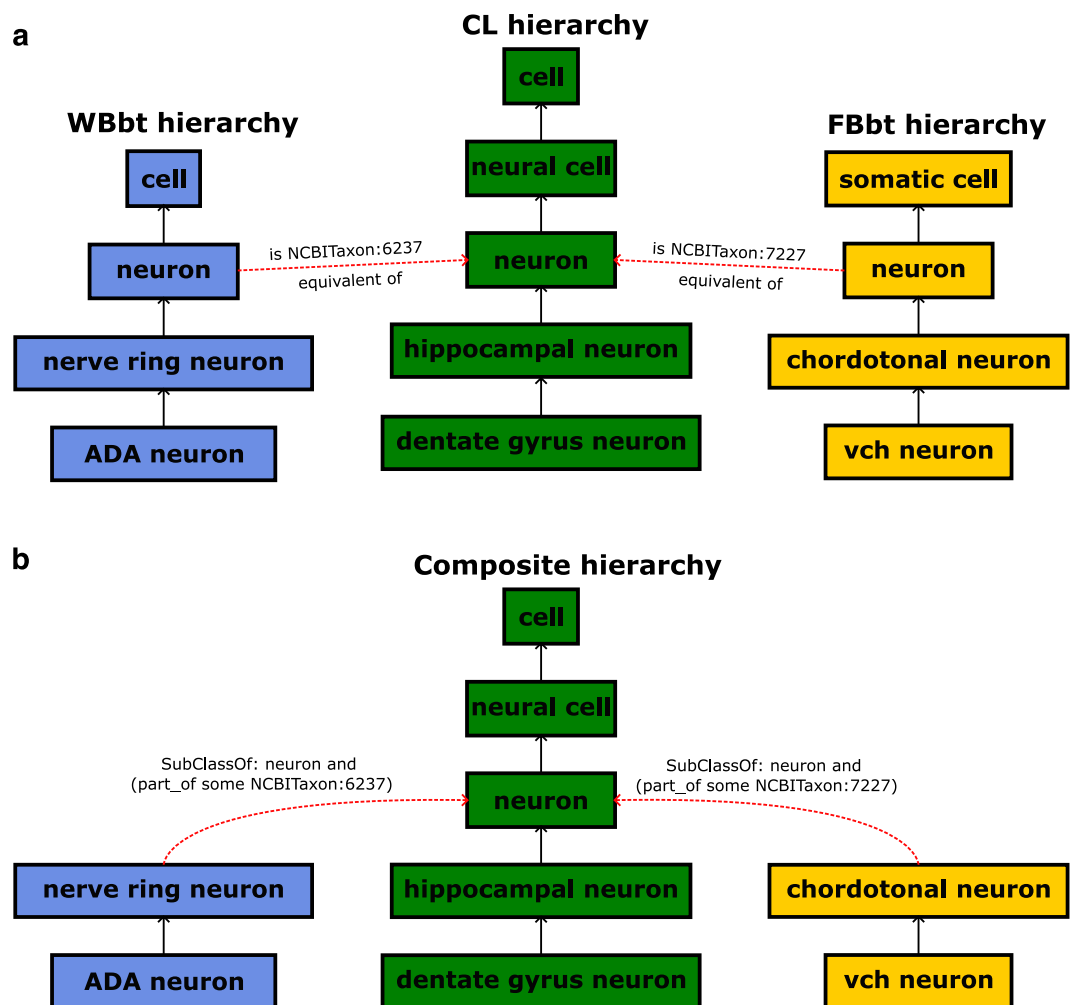


Fig. 8 CL and the species-specific anatomy ontologies each maintain their own hierarchy of terms separately. (Top) The hierarchies are bridged by means of cross-species equivalence axioms that connect species-specific terms to their CL species-neutral counterparts. (Bottom) A single, unified hierarchy is obtained by replacing the species-specific terms by class expressions that use the appropriate species-neutral term from CL.

user facing applications. Firstly, we release a combined product “Composite Metazoan”, a single unified anatomy and cell type ontology that integrates a dozen of taxon-specific ontologies into the taxon-neutral framework of CL and Uberon. As such, Composite Metazoan is a convenient, readily available option for applications that require broad coverage of the animal kingdom. Second, we release taxon-specific bridge files, each containing the bridging axioms for a given taxon — these may be used by a third-party application to build an integrated cross-species ontology covering precisely the taxa that the application required. Lastly, we also publish the mapping themselves, as a single SSSOM⁸² mapping set, for downstream applications that require more flexibility than what the pre-made bridge files can provide.

Generating CL textual definitions using LLMs. The improvements in the tuft cell hierarchy illustrated above was supported by LLMs. A standardised prompt was created, tested and refined to align with CL standards. This prompt is input into the default Perplexity Pro model to generate textual definitions for a given {cell type} (<https://www.perplexity.ai/search/system-role-you-are-an-expert-V1D0txTvQxuE0ApqaG3ZTw>). Editors may run the prompt multiple times for a better output. Editors then assess the LLM-generated definitions and supporting references, refining the definition accordingly. During manual editing, editors focus on removing redundant general information, validating assertions against provided references, and including species-specific information (with references) if unique markers are specified. An example of thymic tuft cell definition is shown above (Fig. 5). The text below represents the standardised prompt currently in use by CL to create LLM-generated textual definitions.

System Role: You are an expert cell biologist with extensive experience in creating precise and informative descriptions of cell types for ontologies.

User Role: I need you to create definitions for specific cell types to be included in the Cell Ontology.

Each definition should:

1. Avoid naming the cell type being defined directly. It should start with a statement of a general classification for the cell type being defined, followed by the characteristics that distinguish it from other cell types within the same general classification.
2. Describe distinguishing characteristics, including structural features, functional roles, and anatomical context.
3. Include species-specific information when relevant, noting presence or absence in different organisms.
4. Mention key molecular markers, transcription factors, or genes only if they are crucial for identification or development of the cell type. When including molecular markers, specify the species in which they have been identified (e.g., “marker X in mice”, “marker Y in humans”).
5. For general cell types, focus on common features across different tissues or organs.
6. Include supporting references to key statements of the definition rather than listing them at the end.
7. Be concise yet comprehensive, aiming for 80–120 words in a single paragraph.
8. Use clear, scientific language accessible to biologists across various specialties.

Example output for a specific cell type: “A tuft cell that is part of the epithelium of pancreatic duct. Present in humans and rats, this cell is absent in the murine pancreas under normal conditions but emerges during acinar-to-ductal metaplasia triggered by injury, inflammation, or oncogenic mutations (1,2). It modulates the immune response and protects against pancreatic ductal adenocarcinoma progression by producing suppressive eicosanoids, such as prostaglandin D2 (3). A tuft cell in the pancreatic duct highly expresses the transcription factor POU2F3, which is essential for its development and presence (4).”

“Please generate a definition for the following cell type: ‘tuft cell of thymus’”

CL-Knowledge graph and GO-Marker cross-analysis. The Ontology Based Application Starter Kit, OBASK¹³⁵, is a well-tested, containerised pipeline that supports conversion of OWL ontologies and RDF into annotated nodes and edges in a Neo4J graph using a standard set of mappings from OWL axioms that relate only 2 entities (classes, individuals). We used this pipeline to generate a graph representing a merge between the Cell Ontology (including markers recorded in CL) and the Gene Ontology, including relationships between terms in these ontologies in both directions.

We then folded additional edges into this graph representing links between cell types and their markers and between GO terms and genes annotated to them, focusing on GO terms that are directly related to CL terms in the graph.

More specifically, we added gene:cell type edges representing markers from CELLxGENE (generated from the CxG corpus of transcriptomic data annotated with CL³⁵), CellMarker2.0³⁴ and the Scheurmann lab⁵⁶. We queried the QuickGO API¹³⁶ to identify all mouse and human genes annotated to the selected GO terms, where GO annotation used the evidence codes ‘biological aspect of ancestor evidence used in manual assertion’ (ECO:0000318) and ‘experimental evidence used in manual assertion’ (ECO:0000269) or their descendant codes. These were loaded into the graph as GO-Gene edges.

Finally, we used Cypher to query the resulting knowledge graph for genes that are both markers of a specific cell type and implicated in a GO process or Cell Component directly linked to that cell type.

Keeping this updated with source ontologies and data sources simply requires re-running the pipeline.

Data availability

CL⁸⁶ can be accessed via a variety of routes. It can be browsed on the Ontology Lookup Service^{65,66} and Bioportal^{64,137}. A stripped down version of CL with links to markers can also be browsed on CZ CELLxGENE CellGuide (see Fig. 3).

CL is available for download from a set of standard persistent URLs in a variety of forms (see below) and formats (OWL, OBO and OBOgraphs JSON). The main CL release is available from:

<http://purl.obolibrary.org/obo/cl.obo/owl/json>

All other releases are available via

<http://purl.obolibrary.org/obo/cl/cl-{form}.obo/owl/json>

Each download also includes a version URL—a permanently resolvable link to the version that can be used to reference versions used in analysis and annotation.

e.g.

<http://purl.obolibrary.org/obo/cl/releases/2025-12-17/cl.owl>

Downloads available by content (form):

- cl - The Cell Ontology plus merged imports, including from Uberon and the Gene Ontology
- cl-plus - combines the full Cell Ontology with the provisional cell ontology
- cl-base - The Cell Ontology with links to external ontologies recorded as bare IRIs. This is designed to be combined with other base ontologies to make integrated products and for use in generating import modules from CL.
- cl-simple - only CL terms and their relationships. No imports
- cl-basic - only CL terms and their relationships. Additionally the graph is guaranteed to lack cycles (these are perfectly legal in OWL but removing them makes it easier for some software to operate on it).

Each release also comes with a detailed, automated report of changes since the last release.

CL is also available as part of ‘composite-metazoan’—an ontology that combines, Uberon, CL and species-specific ontologies: *Drosophila* Anatomy Ontology (FBbt); *C. elegans* Gross Anatomy Ontology (WBbt); Zebrafish Anatomy and Development Ontology (ZFA); *Xenopus* Anatomy Ontology (XAO); Mouse Adult Gross Anatomy (MA) and Mouse Developmental Anatomy Ontology (EMAPA); Human Developmental Anatomy (EHDAA2);

and the ontologies derived from the Allen Institute's brain atlases. The composite ontology is available at <http://purl.obolibrary.org/obo/uberon/composite-metazoan.owl>.

Programmatic access is available via the Ontology Lookup Service API (<https://www.ebi.ac.uk/ols4/api/ontologies/cl/>), the Biportal API (see <http://data.bioontology.org/documentation>), via the UberGraph⁶⁸ SPARQL endpoint (<https://ubergraph.apps.renci.org/sparql>), and via the Ontology Access Kit (OAK). UberGraph supports queries of all asserted and inferred relationships between terms in CL and terms in a broad range of OBO-Foundry ontologies that reference CL, including the Gene Ontology, Uberon, Human Phenotype Ontology and MONDO. A REST API is also available, featuring pre-rolled template queries of Ubergraph, for example allowing users to query for all cell types in a specific anatomical structure (<https://grlc.io/api-git/INCATools/ubergraph/subdir/sparql>).

Code availability

The Cell Ontology is developed, documented and released via a GitHub repository (<https://github.com/obophenotype/cell-ontology>) implemented using the Ontology Development Kit⁸⁴. This includes all code and ontology files used in the development of the ontology, available under an open license, as well as a tracker for community requests. Code and data used for stats in this paper can be found at <https://github.com/shawntanzk/cl-manuscript>.

All code used to build the knowledge-base can be found at https://github.com/Cellular-Semantics/CL_KG.

Received: 26 June 2025; Accepted: 30 March 2026;

Published online: 24 April 2026

References

1. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Front Cell Dev Biol* **6**, 108 (2018).
2. Shekhar, K. & Menon, V. Identification of cell types from single-cell transcriptomic data. *Methods Mol. Biol.* **1935**, 45–77 (2019).
3. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* **20**, 363–374 (2023).
4. Hickey, J. W. *et al.* Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nat. Methods* **19**, 284–295 (2022).
5. Shrestha, B. Single-Cell Metabolomics by Mass Spectrometry. *Methods Mol. Biol.* **2064**, 1–8 (2020).
6. Wang, Y. *et al.* Spatial transcriptomics: Technologies, applications and experimental considerations. *Genomics* **115**, 110671 (2023).
7. Winding, M. *et al.* The connectome of an insect brain. *Science* **379**, eadd9330 (2023).
8. Scheffer, L. K. *et al.* A connectome and analysis of the adult drosophila central brain. *Elife* **9**, 1–74 (2020).
9. Baek, S. & Lee, I. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Comput. Struct. Biotechnol. J.* **18**, 1429–1439 (2020).
10. Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* **34**, 199–203 (2016).
11. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 1–9 (2016).
12. McMurtry, J. A. *et al.* Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol.* **15**, e2001414 (2017).
13. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*. **102**, 15545–15550 (2005).
14. Hu, Y. *et al.* PANGEA: a new gene set enrichment tool for Drosophila and common research organisms. *Nucleic Acids Res.* **51**, W419–W426 (2023).
15. Putman, T. E. *et al.* The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Res.* **52**, D938–D949 (2024).
16. Court, R. *et al.* Virtual Fly Brain—An interactive atlas of the Drosophila nervous system. *Front. Physiol.* **14** (2023).
17. Fischer, D. S. *et al.* Sfaira accelerates data and model reuse in single cell genomics. *Genome Biol.* **22**, 248 (2021).
18. Wang, S. *et al.* Leveraging the Cell Ontology to classify unseen cell types. *Nat. Commun.* **12**, 5556 (2021).
19. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
20. Tan, S. Z. K. *et al.* Digital evolution: Novo Nordisk's shift to ontology-based data management. *J. Biomed. Semantics* **16**, 6 (2025).
21. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
22. Jackson, R. *et al.* OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database* **2021** (2021).
23. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
24. Meehan, T. F. *et al.* Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).
25. Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
26. Vasilevsky, N. A. *et al.* Mondo: Integrating disease terminology across communities. *Genetics* <https://doi.org/10.1093/genetics/iyaf215> (2025).
27. Stefancsik, R. *et al.* The Ontology of Biological Attributes (OBA)—computational traits for the life sciences. *Mamm. Genome* <https://doi.org/10.1007/s00335-023-09992-1> (2023).
28. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224** (2023).
29. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
30. Pranty, A. I., Cuevas Ocaña, S. & Dickens, J. A. Mechanisms of surfactant maturation in development and diseases. *Physiology (Bethesda)*, <https://doi.org/10.1152/physiol.00033.2025> (2025).
31. Vallejo, J. *et al.* Combined protein and transcript single-cell RNA sequencing in human peripheral blood mononuclear cells. *BMC Biol.* **20**, 193 (2022).
32. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
33. Bueckle, A. *et al.* Construction, Deployment, and Usage of the Human Reference Atlas Knowledge Graph for Linked Open Data. *Bioinformatics* (2024).

34. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).
35. CZI Single-Cell Biology Program *et al.* CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv* 2023.10.30.563174 <https://doi.org/10.1101/2023.10.30.563174> (2023).
36. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
37. Franchini, M., Pellicchia, S., Viscido, G. & Gambardella, G. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. *NAR Genom Bioinform* **5**, lqad024 (2023).
38. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-. *Cell Transcriptomics. Cell* **166**, 1308–1323. e30 (2016).
39. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).
40. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).
41. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).
42. Toro, S. *et al.* Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *J Biomed Semantics* **15**, 19 (2024).
43. Caufield, J. H. *et al.* Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *arXiv [cs.AI]* (2023).
44. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6** (2017).
45. McGill, C. *et al.* cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv* 2021.04.05.438318, <https://doi.org/10.1101/2021.04.05.438318> (2021).
46. Börner, K. *et al.* Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas construction and usage. *Nat Methods* **22**, 845–860 (2025).
47. Evangelista, J. E. *et al.* Enrichr-KG: bridging enrichment analysis across multiple libraries. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkad393> (2023).
48. Van Slyke, C. E., Bradford, Y. M., Westerfield, M. & Haendel, M. A. The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *J. Biomed. Semantics* **5**, 12 (2014).
49. Segerdell, E. *et al.* Enhanced XAO: the ontology of Xenopus anatomy and development underpins more accurate annotation of gene expression and queries on Xenbase. *J. Biomed. Semantics* **4**, 31 (2013).
50. Costa, M., Reeve, S., Grumblin, G. & Osumi-Sutherland, D. The Drosophila anatomy ontology. *J. Biomed. Semantics* **4**, 32 (2013).
51. Davis, P. *et al.* WormBase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics* **220** (2022).
52. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10** (2021).
53. George, N. *et al.* Expression Atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Res.* **52**, D107–D114 (2024).
54. Ohlstein, B. & Spradling, A. The adult Drosophila posterior midgut is maintained by pluripotent stem cells. *Nature* **439**, 470–474 (2006).
55. Orloff, D. N., Iwasa, J. H., Martone, M. E., Ellisman, M. H. & Kane, C. M. The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Res.* **41**, D1241–50 (2013).
56. Hu, J. *et al.* Benchmarking single cell transcriptome matching methods for incremental growth of reference atlases. *bioRxiv* <https://doi.org/10.1101/2025.04.10.648034> (2025).
57. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
58. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
59. Xu, C. *et al.* Automatic cell-type harmonization and integration across Human. *Cell Atlas datasets. Cell* **186**, 5876–5891.e20 (2023).
60. Fischer, F. *et al.* scTab: Scaling cross-tissue single-cell annotation models. *Nat. Commun.* **15**, 6611 (2024).
61. Ergen, C. *et al.* Consensus prediction of cell type labels in single-cell data with popV. *Nat. Genet.* <https://doi.org/10.1038/s41588-024-01993-3> (2024).
62. Bernstein, M. N., Ma, Z., Gleicher, M. & Dewey, C. N. CellO: comprehensive and hierarchical cell type classification of human cells with the Cell Ontology. *iScience* **24**, 101913 (2021).
63. Noy, N. F. *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37**, W170–3 (2009).
64. Cell Ontology in Bioportal, <https://bioportal.bioontology.org/ontologies/CL>.
65. Jupp, S., Burdett, T., Leroy, C. & Parkinson, H. E. A new Ontology Lookup Service at EMBL-EBI. *SWAT4LS* **2**, 118–119 (2015).
66. McLaughlin, J., Lagrimas, J., Iqbal, H., Parkinson, H. & Harmse, H. OLS4: A new Ontology Lookup Service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics* **41**, btaf279 (2025).
67. Cell Ontology in Ontology Lookup Service (OLS). <https://www.ebi.ac.uk/ols4/ontologies/cl>.
68. Balhoff, J. P. *et al.* Ubergraph: integrating OBO ontologies into a unified semantic graph. *CEUR Workshop Proceedings* (2022).
69. Tan, S. Z. K. *et al.* Brain Data Standards - A method for building data-driven cell-type ontologies. *Sci. Data* **10** (2023).
70. Börner, K. *et al.* Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nat. Cell Biol.* **23**, 1117–1128 (2021).
71. Caron, A. R. *et al.* A general strategy for generating expert-guided, simplified views of ontologies. *bioRxiv*, <https://doi.org/10.1101/2024.12.13.628309> (2024).
72. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *bioRxiv* 2022.03.10.483747 (2022).
73. Xu, R. & Peng, J. A comprehensive survey of Deep Research: Systems, methodologies, and applications. *arXiv [cs.AI]* (2025).
74. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge. *arXiv [cs.CL]*, <https://doi.org/10.48550/ARXIV.2409.13740> (2024).
75. Osorno, T. *et al.* Candelabrum cells are ubiquitous cerebellar cortex interneurons with specialized circuit properties. *Nat. Neurosci.* **25**, 702–713 (2022).
76. Kozareva, V. *et al.* A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219 (2021).
77. Rivera, A. D. & Osumi-Sutherland, D. Bioinformatic analysis of Purkinje and Molecular Layer interneurons for WMB annotation. Preprint at, <https://doi.org/10.5281/ZENODO.17543831> (2025).
78. Lubiana, T. *et al.* Guidelines for reporting cell types: the MIRACL standard. *arXiv [q-bio.OT]* (2022).
79. Bayindir, İ. U. Ontology-Autosuggest Metadata Annotation Spreadsheet. Preprint at, <https://doi.org/10.5281/ZENODO.15387766> (2025).
80. Hegde, H. *et al.* A Change Language for Ontologies and Knowledge Graphs (2024).
81. Hoyt, C. T., Hoyt, A. L. & Gyori, B. M. Prediction and curation of missing biomedical identifier mappings with Biomappings. *Bioinformatics* **39** (2023).
82. Matentzoglou, N. *et al.* A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database* **2022** (2022).
83. Hoyt, C. T. *et al.* Unifying the identification of biomedical entities with the Bioregistry. *Sci. Data* **9**, 714 (2022).
84. Matentzoglou, N. *et al.* Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database* **2022**, baac087 (2022).
85. Balhoff, J. & Curtis, C. K. INCATools/ubergraph: Release 2021-03-26, <https://doi.org/10.5281/zenodo.4641309> (2021).
86. Tan, S. *et al.* Obophenotype/cell-Ontology: 2025-12-17 Release, <https://doi.org/10.5281/ZENODO.17966975> (Zenodo, 2025).

87. Keefe, M. G. & Nowakowski, T. J. A recipe book for cell types in the human brain. *Nature* **573**, 36–37 (2019).
88. Brügger, M. D. & Basler, K. The diverse nature of intestinal fibroblasts in development, homeostasis, and disease. *Trends Cell Biol.* **33**, 834–849 (2023).
89. Li, J. *et al.* Integrated multi-omics single cell atlas of the human retina. *Molecular Biology* (2023).
90. Masci, A. M. *et al.* An improved ontological representation of dendritic cells as a paradigm for all cell types. *BMC Bioinformatics* **10**, 70 (2009).
91. Diehl, A. D. *et al.* Hematopoietic cell types: prototype for a revised cell ontology. *J. Biomed. Inform.* **44**, 75–79 (2011).
92. Artis, D. & Spits, H. The biology of innate lymphoid cells. *Nature* **517**, 293–301 (2015).
93. Ebbo, M., Crinier, A., Vély, F. & Vivier, E. Innate lymphoid cells: major players in inflammatory diseases. *Nat. Rev. Immunol.* **17**, 665–678 (2017).
94. Spits, H. *et al.* Innate lymphoid cells—a proposal for uniform nomenclature. *Nat. Rev. Immunol.* **13**, 145–149 (2013).
95. Chae, S., Park, T. J. & Kwon, T. Convergent differentiation of multiciliated cells. *Sci Rep* **13**, 23028 (2023).
96. Schneider, C., O’Leary, C. E. & Locksley, R. M. Regulation of immune responses by tuft cells. *Nat Rev Immunol* **19**, 584–593 (2019).
97. Miller, C. N. *et al.* Thymic tuft cells promote an IL-4-enriched medulla and shape thymocyte development. *Nature* **559**, 627–631 (2018).
98. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cellular and Molecular Life Sciences: CMLS* **69**, 2907 (2012).
99. Ualiyeva, S. *et al.* A nasal cell atlas reveals heterogeneity of tuft cells and their role in directing olfactory stem cell proliferation. *Science immunology* **9**, eabq4341 (2024).
100. Jorstad, N. L. *et al.* Transcriptomic cytoarchitecture reveals principles of human neocortex organization. *Science* **382**, eadf6812 (2023).
101. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
102. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
103. Bakken, T. *et al.* Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* **18**, 559 (2017).
104. Liu, A. *et al.* Discovery of optimal cell type classification marker genes from single cell RNA sequencing data. *BMC Methods* **1** (2024).
105. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
106. Reed, A. D. *et al.* A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nature Genetics* **56**, 652–662 (2024).
107. Gray, G. K. *et al.* Defining breast epithelial cell types in the single-cell era. *Dev. Cell* **60**, 2218–2236 (2025).
108. Deegan née Clark, J. I., Dimmer, E. C. & Mungall, C. J. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics* **11**, 530 (2010).
109. Ament, S. A. *et al.* The Neuroscience Multi-Omic Archive: a BRAIN Initiative resource for single-cell transcriptomic and epigenomic data from the mammalian brain. *Nucleic Acids Res.* **51**, D1075–D1085 (2023).
110. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
111. Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv* 2020.02.29.970558 (2020).
112. Scala, F. *et al.* Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, <https://doi.org/10.1038/s41586-020-2907-3> (2020).
113. Collin, S. P. & Trezise, A. E. O. The origins of colour vision in vertebrates. *Clin Exp Optom* **87**, 217–223 (2004).
114. Shekhar, K. & Sanes, J. R. Generating and Using Transcriptomically Based Retinal Cell Atlases. *Annu Rev Vis Sci* **7**, 43–72 (2021).
115. Hahn, J. *et al.* Evolution of neuronal cell classes and types in the vertebrate retina. *Nature* **624**, 415–424 (2023).
116. Masland, R. H. The neuronal organization of the retina. *Neuron* **76**, 266–280 (2012).
117. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
118. Johansen, N. J. *et al.* Cross-species consensus atlas of the primate basal ganglia. *bioRxiv*, <https://doi.org/10.64898/2025.12.15.694496> (2025).
119. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods* **21**, 1462–1465 (2024).
120. Aevermann, B. *et al.* A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Res.* **31**, 1767–1780 (2021).
121. Sahoo, S. S. *et al.* Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records. *Sci. Rep.* **12**, 19430 (2022).
122. Kulmanov, M., Smaili, F. Z., Gao, X. & Hoehndorf, R. Semantic similarity and machine learning with ontologies. *Brief. Bioinform.* **22** (2021).
123. Avsec, Ž. *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
124. Acosta, P. H. *et al.* CellSymphony: Deciphering the molecular and phenotypic orchestration of cells with single-cell pathomics. *arXiv [cs.CV]* (2025).
125. Zhao, S. *et al.* SToFM: a Multi-scale Foundation Model for Spatial Transcriptomics. *arXiv [q-bio.GN]* (2025).
126. Schaar, A. C. *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv* 2024.04.15.589472 <https://doi.org/10.1101/2024.04.15.589472> (2024).
127. Birk, S. *et al.* Quantitative characterization of cell niches in spatially resolved omics data. *Nat. Genet.* **57**, 897–909 (2025).
128. Ahuja, G. *et al.* Multi-agent AI enables evidence-based cell annotation in single-cell transcriptomics. *bioRxiv*, <https://doi.org/10.1101/2025.11.06.686964> (2025).
129. Wang, H. *et al.* SpatialAgent: An Autonomous AI Agent for Spatial Biology. *bioRxiv*, <https://doi.org/10.1101/2025.04.03.646459> (2025).
130. Nouri, N., Artzi, R. & Savova, V. An agentic AI framework for ingestion and standardization of single-cell RNA-seq data analysis. *bioRxiv*, <https://doi.org/10.1101/2025.07.31.667880> (2025).
131. Balhoff, J. P. & Mungall, C. J. Whelk: An OWL EL+RL reasoner enabling new use cases. Preprint at, <https://doi.org/10.4230/TGDK.2.2.7> (2024).
132. Jackson, R. C. *et al.* ROBOT: A Tool for Automating Ontology Workflows. *BMC Bioinformatics* **20**, 407 (2019).
133. Mungall, C., Rutenberg, A. & Osumi-Sutherland, D. Taking shortcuts with OWL using safe macros. *Derm. Helv.*, <https://doi.org/10.1038/npre.2011.5292.2> (2011).
134. Matentzoglou, N. *et al.* A Simple Standard for Ontological Mappings 2022: Updates of data model and outlook. in *17th International Workshop on Ontology Matching, OM 2022* vol. 3324 61–66 (CEUR-WS, 2022).
135. Kir, H., Bayindir, I. U. & Osumi-Sutherland, D. *Ontology Based Application Starter Kit (OBASK)*, <https://doi.org/10.5281/zenodo.10518990> (2024).
136. Munoz-Torres, M. & Carbon, S. Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools. *Methods Mol. Biol.* **1446**, 149–160 (2017).

137. Vendetti, J. *et al.* BioPortal: an open community resource for sharing, searching, and utilizing biomedical ontologies. *Nucleic Acids Res.*, <https://doi.org/10.1093/nar/gkaf402> (2025).
138. Sun, X. *et al.* A census of the lung: CellCards from LungMAP. *Dev. Cell* **57**, 112–145.e2 (2022).
139. Causeret, F., Moreau, M. X., Pierani, A. & Blanquie, O. The multiple facets of Cajal-Retzius neurons. *Development* **148** (2021).
140. Zhu, G., Hu, J. & Xi, R. The cellular niche for intestinal stem cells: a team effort. *Cell Regen. (Lond.)* **10**, 1 (2021).
141. Peng, B. NS-forest marker genes for human neocortex cross-areal atlas. Preprint at, <https://doi.org/10.5281/ZENODO.15319198> (2025).

Acknowledgements

We thank Aidan Maartens (Wellcome Sanger Institute) for scientific writing support. A.P.B. and A.R.C. were funded by NIH via the HuBMAP project (OT2OD033756 and OT2OD026671) and EMBL-EBI core funds. CE and DOS contribution was supported by grant number 2023-221471 from the Chan Zuckerberg Initiative D.A.F. J.P.B. was partly supported by the GO consortium that is funded by the National Human Genome Research Institute (NHGRI), US National Institutes of Health, grant number HG012212. IUB and JJB's work was funded by a gift from Schmidt Futures (<https://www.schmidtfutures.org>) to support the 'Cell Annotation Platform'. MAH, ST, and RS were supported by NIH Office of the Director Grant #5R24OD011883 for the Monarch Initiative. SZKT, HK, ADR & DOS work was supported in part by NIH 1UM1MH130981-0. TL was supported by a grant from the São Paulo Research Foundation (#19/26284-1). ADD, JAO, and BP were supported in part by the National Institute of Allergy and Infectious Disease [4U19AI118610]. CJM was supported by the National Institutes of Health National Human Genome Research Institute [HG010860, HG012212, HG010859]; National Institutes of Health Office of the Director [R24 OD011883]; and the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy. PR was supported by a grant from Chan Zuckerberg Initiative DAF to support the Human Cell Atlas data coordination platform. Y.Z. was supported, in part, by the Division of Intramural Research of the National Library of Medicine (NLM), National Institutes of Health. RS was supported by the NIH National Human Genome Research Institute Phenomics First Resource, NIH-NHGRI # 5RM1 HG010860, a Center of Excellence in Genomic Science. The work of D.O.S., H.K., U.B., J.B. & C.E. was funded, in part, by 220540/Z.20/A, "Wellcome Sanger Institute Quinquennial Review 2021-2026. DGG was supported by grant BB/T014008 from the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the US National Science Foundation Directorate of Biological Sciences (NSF/BIO).

Author contributions

S.Z.K.T., A.P.B., D.G.G., C.E., A.D.D. and D.O.S. wrote the manuscript, and all authors reviewed the manuscript. All authors contributed to the development of C.L. and/or major tools that use C.L. and are described in the manuscript.

Competing interests

MAH is the founder of Alamy Health.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-07173-8>.

Correspondence and requests for materials should be addressed to R.H.S., C.J.M., A.D.D. or D.O.-S.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026