bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

BioRxiv, 2021, 1–6

Refer to BioRxiv for DOI
Preprint

PREPRINT

Brain Data Standards Ontology: A data-driven ontology of transcriptomically defined cell types in the primary motor cortex

Shawn Zheng Kai Tan [1,†], Huseyin Kir [1,†], Brian Aevermann [2,†], Tom Gillespie [3], Michael Hawrylycz [4], Ed Lein [4], Nicolas Matentzoglu [5], Jeremy Miller [4], Tyler S. Mollenkopf [4], Christopher J. Mungall [6], Patrick L. Ray [4], Raymond E. A. Sanchez [4], Richard H. Scheuermann [2,3], Brian Staats [4], Yun H. Zhang [4] and David Osumi-Sutherland [1,*]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom and [2]J. Craig Venter Institute (JCVI), La Jolla, CA, USA and [3]University of California San Diego, La Jolla, CA, USA and [4]Allen Institute for Brain Science, Seattle, WA, USA and [5]Semanticly Ltd, London, United Kingdom and [6]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

*davidos@ebi.ac.uk
[†]Authors contributed equally

Abstract

Large scale single cell omics profiling is revolutionising our understanding of cell types, especially in complex organs like the brain. This presents both an opportunity and a challenge for cell ontologies. Annotation of cell types in single cell 'omics data typically uses unstructured free text, making comparison and mapping of annotation between datasets challenging. Annotation with cell ontologies is key to overcoming this challenge, but this will require meeting the challenge of extending cell ontologies representing classically defined cell types by defining and classifying cell types directly from data. Here we present the Brain Data Standards Ontology (BDSO), a data driven ontology that is built as an extension to the Cell Ontology (CL). It supports two major use cases: cell type annotation, and navigation, search, and organisation of a web application integrating single cell omics datasets for the mammalian primary motor cortex. The ontology is built using a semi-automated pipeline that interlinks cell type taxonomies and necessary and sufficient marker genes, and imports relevant ontology modules derived from external ontologies. Overall, the BDS ontology provides an underlying structure that supports these use cases, while remaining sustainable and extensible through automation as our knowledge of brain cell type expands.

Key words: Single Cell Transcriptomics; Ontology; Primary Motor Cortex

Introduction

The large-scale application of omics profiling techniques at the single cell level is producing enormous volumes of data. Cell ontologies are poised to play a critical role in making these data searchable and integratable [1]. At the same time, the application of these techniques is revolutionising our understanding of cell types and cellular heterogeneity [2, 3]. The impact of this revolution is especially dramatic for the brain. Due to the complex cellular architecture of the brain, traditional qualitative, categorical methods of classifying neurons based on location, morphology, marker expression and function have not come close to achieving a coherent, unified view of brain cell types and their classifications. This has begun to change with the application of massively parallel single cell or nucleus RNA sequencing (sc/nRNAseq) methods, measuring the transcript levels of thousands of genes within each of hundreds of thousands of individual cells. The data from these experiments provides the basis for a consensus, data-driven and comprehensive quantitative framework for brain cell-type classification both within and between species. Evidence from sys-

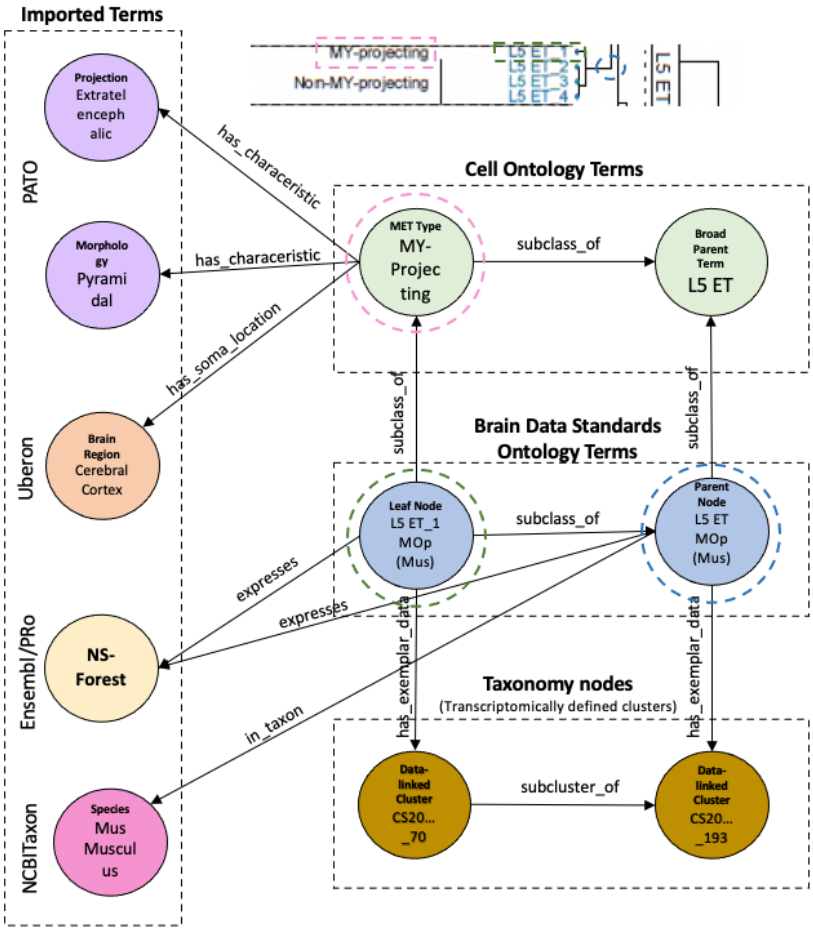

Figure 1. Graph illustrating the BDSO schema. This graph shows the relationship of the BDSO classes (Brain Data Standards Ontology nodes, light blue circles) to OWL Individuals (Taxonomy nodes, brown circles) representing clusters in the data driven taxonomy used as input and to the build process, to classes in the Cell Ontology (green circles) and from external ontologies (blue backgrounds) representing species (NCBITaxon), brain region (UBERON), morphology (PATO), and markers (Ensembl/PRO).

tems in which a more comprehensive classification of cell types has been achieved by classical methods suggests that the classifications resulting from sc/nRNAseq analysis align closely with classically defined types [4].

In parallel with the development and use of sc/nRNAseq, techniques have also been developed that can produce transcriptomic profiles, morphology and/or functional measurements of the same individual single neuron (e.g., Patch-seq), allowing function and morphology to be mapped to cell types defined using sc/nRNAseq data based on similarity in transcriptional profiles. The result is an increasingly consistent, unified and integrated view of mammalian brain cell types.

How can we integrate definitions of cell types from sc/nRNAseq data analysis, which take transcriptomic data from clusters of transcriptomically similar cells as ground truth for cell-typing, into cell ontologies in which cell type/classes are defined using simple, categorical assertions about their morphological and functional properties, location and marker expression? How can we do this in a way that is transparent about the origins and evidence for these classifications? How can we enable users to leverage the data used to define and classify reference cell types in the ontology in order to classify cell types represented in their own data? Here we propose an approach for data-driven cell type classification and semantic representation to address these challenges.

Brain Data Standards Ontology

The Brain Data Standards Ontology (BDSO) is a data-driven extension of the Cell Ontology (CL) [5] that supports the navigation, search, and organisation of information about cell types through an integrated web portal, and also functions as an independent ontology for use in cell-type annotation. The initial focus of work on this ontology utlises data from the BRAIN Initiative Cell Census Network (BICCN) mini-atlas of the mammalian primary motor cortex [6]. It attempts to solve the above-stated problems by using a schema that directly defines cell types via links to reference (exemplar) data and analyses, extending an earlier proposal for defining cell type classes from sc/nRNAseq experiment data and metadata [3].

Cell types in BDSO are defined by reference to clusters of transcriptomically similar cells. Classification in BDSO is derived from the hierarchical relationships between these transcriptomic clusters. The clusters and their hierarchical arrangement derive from unsupervised, hierarchical clusterings of single-cell transcriptomes and epigenetic profiles of the primary motor cortex in mouse, human, and non-human primates [6]. Each individual hierarchical clustering (referred to from here as a taxonomy) is either created from a single data set (e.g., in marmoset) or through a consensus of two (human) or many (mouse) data sets. Leveraging transcriptomic similarity, a subset of clusters in these taxonomies are mapped across species [7]. Finally, using mouse tran-

Figure 2. Example of an automatically generated class displayed in Protege, an ontology browser. In this example, we show L5 Extratelencephalic (ET), which is a grouping class. The label, definition, and set of synonyms are auto-generated from OWL templates using a Dead simple OWL design patterns (DOSDP) pattern system. Automatic axiomatisation includes brain region, species, NS-forest markers, projection pattern, and has_exemplar_data link to taxonomy node (cluster), using a reification pattern. Other possible automated axiomatisation not shown in this figure include morphology and named markers.

scriptomics clusterings as an anchor, morphological and electrophysiological profiles of single cells are mapped to omics-based types using Patch-seq data [8]. All of this information is available in a standard format developed by the BICCN to represent mammalian brain cell type taxonomies and the relationships between them [9] These taxonomies are inputs to a semi-automated ontology build process that extends the Ontology Development Kit [10]. This process drives templated term addition and import of ontology term modules from the Cell Ontology (CL) [5], Uberon [11], PATO [12], and Ensembl/PRO [13].

Web Ontology Language, OWL2 [14] makes a distinction between individuals, e.g., an individual neuron depicted in a micrograph, and classes, e.g. classes representing canonical types of neurons. Each taxonomy is represented in BDSO as a collection of OWL Individuals, each representing a cluster of single cell transcriptomes. Hierarchical clustering is represented by relating clusters to each other via a transitive subcluster_of relation (OWL objectProperty; see L5 ET_1 and L5 ET nodes in Figure 1). We create classes (data-linked cell types) for all leaf node clusters in each taxonomy and a subset of grouping clusters that map to either known classes or new classes defined by morphology and function based on Patch-seq data. Each data-linked cell type 'C' is linked to a cluster individual 'X' using a value restriction pattern (see L5 ET in Figures 1 and 2):

`{C} EquivalentTo 'native cell' and has_exemplar_data` [1] `value cluster {X}`

With some additional axiomatisation, we can use a standard OWL reasoner to automatically build a classification hierarchy for the BDSO classes, mirroring the cluster hierarchy. Figure 3 illustrates this schema and details how it generates a classification hierarchy. In this illustration, as in the BDSO, not all clusters are used to define cell types, as not all intermediary nodes in a hierarchical clustering are equally biologically informative (in Fig 3, i2 is not used to define cell types). Inferred classification still reflects hierarchical clustering in these cases.

Each class has axiomatisation that records species, brain re-

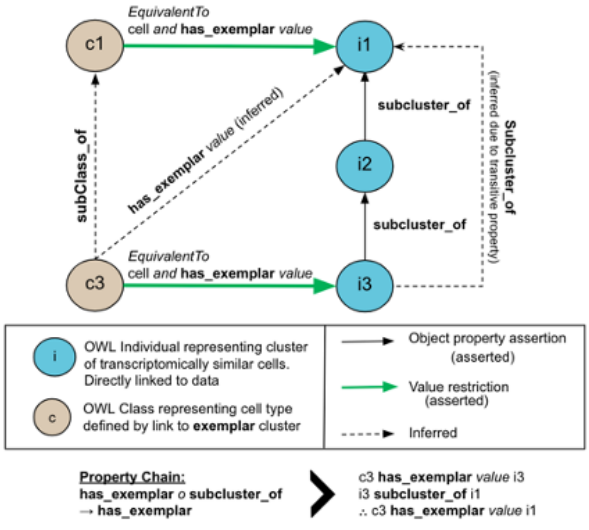

Figure 3. Representative schema for data driven classification. Blue nodes (i1-3) are OWL individuals representing clusters of single cell transcriptomes, while tan nodes (c1, c2) are OWL classes representing cell types. Hierarchical clustering is represented using the transitive subcluster_of relation (objectProperty) to link individuals. Each class is defined by reference to a cluster individual (i), via the relation (objectProperty) as equivalent to (any) cell that has_examplar (value) i. Reasoning via a chain of these two properties (bottom and right sides of the diagram above) is sufficient to infer that c3 has_examplar value i1 and so, combined with the assertion that it is a (type of) cell, fulfils the conditions required to be a subclass of i1.

1 Currently being requested from the Relations Ontology (RO).

gion, and gross classification, based on division of the taxonomy into gross cell types including GABAergic neuron, glutamatergic neuron, and oligodendrocyte. Additionally, we add marker expression axioms corresponding to the minimal set of markers required to distinguish the exemplar from all other clusters in the taxonomy, generated using NS-Forest [15]. Data-driven classes defined for intermediate nodes in the hierarchy are further classified using classes added to the Cell Ontology as part of this work (e.g., see 'L5 extratelencephalic' class in Figure 1). These include classes that are defined by expression of classical markers (e.g., VIP expressing GABAergic neurons), morphology (pyramidal) or projection pattern (extratelencephalic projecting), mapped based on co-collected transcriptomic profiles [6]. Each BDS class also has an auto-generated label, definition, and set of synonyms driven by an OWL Template through a Dead simple OWL design patterns (DOSDP) system [16]. An example of a semi-automatically generated class can be found in Figure 2 shown through an ontology browser, Protégé [17].

The BDSO's code base is available at GitHub (http://purl.obolibrary.org/obo/cl/bds/) including documentation of the full technology stack and details of the approach. A provisional release of the ontology is available for download from http://purl.obolibrary.org/obo/cl/bds/bds.owl and is hosted on a dedicated instance of the EMBL-EBI ontology lookup service (OLS) [18] at http://purl.obolibrary.org/obo/cl/bds/browser/. OLS provides ontology search, browsing, visualisation capabilities and enables web services driven programmatic access to the BDS Ontology.

Integration of BDSO and brain cell type data in a web application.

A key function of the BDSO is to support organisation, navigation and searching of data in a community-accessible view of the cell types defined in the BICCN mini-atlas of the mammalian primary motor cortex [6] through a web-based application (web-app) that integrates cell type descriptions and related data, provisionally known as "Cell Type Cards". Each page in this web-app corresponds to a cell type defined with reference to a cluster in one of the BICCN taxonomies, represented in the BDSO, and features a wide range of data and analysis from multiple cross integrated datasets. The aim of the ontology driven search and navigation tools is to support access to these pages in the web-app.

While expressiveness of ontologies is an advantage for semantic data processing, using ontologies as the data layer of a web application brings several challenges (such as blank nodes, existential restrictions, annotations, entailments and

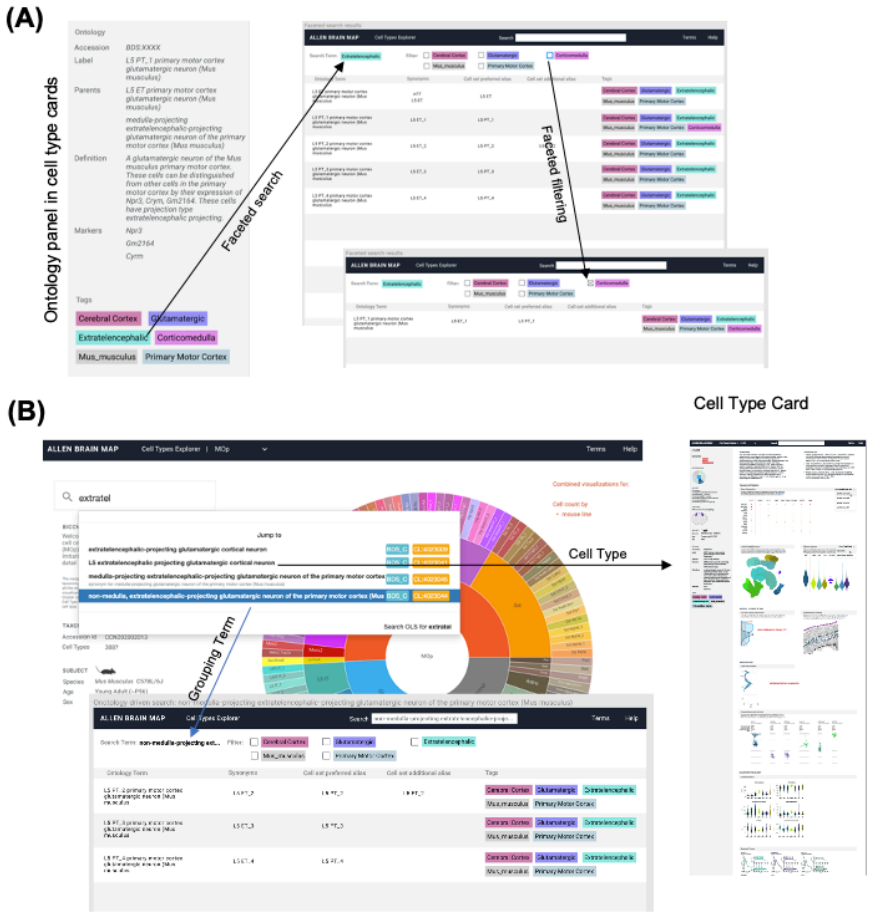

Figure 4. A mockup of the cell type cards web app, incorporating planned search and navigation functionality driven by the BDSO. (A) The panel on the left shows information about an ontology term associated with a cell type card (the corresponding card is shown in panel B on the right). This includes the ontology term ID, name, synonyms, definition, parent cell types, location, species and markers. It also includes a set of semantic tags corresponding to species, brain region, and cell properties such as morphology (pyramidal) and projection pattern (extratelencephalic). Clicking on one of these panels drives faceted search, prompting display of a results page listing all cards with that tag. This can be further refined by selecting additional tags as facets. Panel (B) shows an autocomplete search. Selecting a BDSO class corresponding to a cell-type card (arrow with "Cell Type" label) displays that card. Cell type cards will display the ontology panel seen in (A), but also summary data related to that cell type (e.g. transcriptomics profiles, example morphology, electrophysiology, etc.). Selecting a BDSO class that does not correspond to a single card (arrow with "Grouping Term" label), but which subsumes classes that do, prompts display of a results page listing subsumed cards. This can be further refined via faceted search as shown in panel A.

scalability) compared to conventional data persistence solutions. For this purpose, we extended a library, neo4j2owl (https://github.com/VirtualFlyBrain/neo4j2owl), developed for the Virtual Fly Brain project [19, 20], that ensures logical projection of OWL ontologies into labelled property graphs. Neo4j2owl imports the BDSO and associated ontologies into Neo4j in a way that preserves entailments and annotations, but not the syntactic complexities. Neo4j2OWL also allows the addition of semantic tags, driven by OWL DL or SPARQL queries, that can be used to drive faceted search. For example we can tag all classes corresponding to subclasses of GABAergic neuron, or all classes fulfilling an OWL DL query for classes of neuron with pyramidal morphology (see Figure 4b).

A sample of the resulting property graph is shown in Figure 1. Cloud hosted property graphs RESTfully serve knowledge to cell-type cards in web-friendly formats such as JSON.

Ontology based navigation and search function through two mechanisms - autocomplete (which takes advantage of curation of synonyms in the ontology) and faceted search (Fig 4). Autocomplete allows users to search for cell-type ontology terms, displaying a list of lexical matches for users to choose from. Choosing a term corresponding to a single cell type card takes the user to that card, whereas choosing a term that subsumes multiple cards prompts display of a list of terms corresponding to subsumed cards. A mockup of the interface, illustrating this behavior, is shown in Figure 4a. Faceted search of cell type cards works via a set of tags corresponding to gross classifications (e.g. GABAergic), intrinsic properties (e.g. pyramidal morphology) and extrinsic properties (brain region location, species) of cell types, added to cell type neo4j nodes via OWL DL queries of the underlying ontologies. This allows users to take advantage of the semantics of OWL for faceted search at a practical level of granularity/partitioning. A mock-up of faceted search implementation in cell-type cards is shown in Figure 4b.

Conclusion

The BDSO is a faithful representation of the data driven, consensus cell type classification that constitutes the BICCN mini-atlas of the mammalian motor cortex [6]. By using a schema that defines classes logically via links to an OWL representation of data and analyses, the BDSO is able to directly leverage data-driven classification using OWL reasoning. As a result, classes retain direct links to the data and analyses that define them and the origins of this classification are transparent and insulated from the manual editing process which might alter or obfuscate them. Using templated specification of ontology classes, the BDSO build process is scalable and extensible and allows a flexible mix of automation and manual curation. It also makes it possible to update as new, improved versions of data driven classifications of the same cell types are released.

The linked data can potentially be used to replicate analyses and to map cell types represented in other datasets (e.g. Azimuth [21], NS-forest [15], FR-match [22]). The addition of NS-Forest Markers [15], representing minimal markers for distinguishing, with high confidence, cell types from other cell types defined in the analysis, provides a simple mechanism for mapping cell types from third party transcriptomics data to the BDSO.

Challenges remain. The current representation lacks links to representations of transcriptomic data from Patch-seq data used to map morphologically defined types. Furthermore, accurately mapping cell types that were historically derived through categorical assertions, before the age of single cell transcriptomics, to cell types that are defined with reference to algorithmic clustering of transcriptomic profiles presents a challenge and requires consensus by the community. Using transcriptomics clustering as ground truth for an ontology also comes with its inherent challenges. Penetrance of marker expression and location to a specific cortical layer varies across clusters, so quantified assertions of marker expression in OWL will always be an approximation if ground truth is defined by clustering on similarity and will always require some assessment of thresholds - either automated or qualitative. Finally, nomenclature issues frequently arise when data driven classifications are mapped onto classically defined classes. For example, the literature is full of references to the VIP-expressing GABAergic neurons, identified using VIP as a marker, but clustering defines a broader group of related GABAergic neurons including some subtypes that do not express VIP. This leaves difficult questions around how such grouping classes should be named to reflect their close relationships to the classically defined class.

Acknowledgements

This work was funded by NIMH:1RF1MH123220-01 - "A Community Framework for Data-driven Brain Transcriptomic Cell Type Definition, Ontology, and Nomenclature." We thank Maryann Martone, Carol Thompson, Nomi Harris for their invaluable contributions to discussions of the work described here.

References

1. Osumi-Sutherland D, Xu C, Keays M, Kharchenko PV, Regev A, Lein E, et al. Cell types and ontologies of the Human Cell Atlas. arXiv 2021 Jun;.
2. Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. Front Cell Dev Biol 2018 Sep;6:108.
3. Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, Miller JA, et al. Cell type discovery and representation in the era of high-content single cell phenotyping. BMC Bioinformatics 2017 Dec;18(Suppl 17):559.
4. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 2016 Aug;166(5):1308–1323.e30.
5. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol 2005 Jan;6(2):R21.
6. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 2021 Oct;598(7879):86–102.
7. Bakken T, Jorstad N, Hu Q, Lake B, Tian W, Kalmbach B, et al. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse; 2020.
8. Scala F, Kobak D, Bernabucci M, Bernaerts Y, Cadwell CR, Castro JR, et al. Phenotypic variation of transcriptomic cell types in mouse motor cortex. Nature 2020 Nov;.
9. Miller JA, Gouwens NW, Tasic B, Collman F, van Velthoven CT, Bakken TE, et al. Common cell type nomenclature for the mammalian brain. Elife 2020 Dec;9.
10. Matentzoglu N, Mungall C, Goutte-Gattat D. Ontology Development Kit. zenodo, DOI:105281/zenodo4973944 2021 Jul;.
11. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012 Jan;13(1):R5.
12. Gkoutos GV, Green ECJ, Mallon AM, Hancock JM, Davidson D. Using ontologies to describe mouse phenotypes. Genome Biol 2005;6(1):R8.
13. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. Nucleic Acids Res 2011 Jan;39(Database issue):D539–45.
14. Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph

S, Others. OWL 2 web ontology language primer. W3C recommendation 2009;27(1):123.
15. Aevermann BD, Zhang Y, Novotny M, Keshk M, Bakken TE, Miller JA, 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 2021 Jun;.
16. Osumi-Sutherland D, Courtot M, Balhoff JP, Mungall C. Dead simple OWL design patterns. J Biomed Semantics 2017 Jun;8(1):18.
17. Musen MA, Protégé Team. The Protégé Project: A Look Back and a Look Forward. AI Matters 2015 Jun;1(4):4–12.
18. Jupp S, Burdett T, Leroy C, Parkinson HE. A new Ontology Lookup Service at EMBL-EBI. SWAT4LS 2015;2:118–119.
19. Milyaev N, Osumi-Sutherland D, Reeve S, Burton N, Baldock RA, Armstrong JD. The Virtual Fly Brain browser and query interface. Bioinformatics 2012 Feb;28(3):411–415.
20. Osumi-Sutherland D, Costa M, Court R, O'Kane C. Virtual Fly Brain-Using OWL to support the mapping and genetic dissection of the Drosophila brain. In: C Maria Keet, editor. Proceedings of OWLED 2014 CEUR workshop proceedings; 2014. p. 85–96.
21. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021 Jun;184(13):3573–3587.e29.
22. Zhang Y, Aevermann BD, Bakken TE, Miller JA, Hodge RD, Lein ES, et al. FR-Match: robust matching of cell type clusters from single cell RNA sequencing data using the Friedman-Rafsky non-parametric test. Brief Bioinform 2021 Jul;22(4).