

Common cell type nomenclature for the mammalian brain

Jeremy A Miller*, Nathan W Gouwens, Bosiljka Tasic, Forrest Collman, Cindy TJ van Velthoven, Trygve E Bakken, Michael J Hawrylycz, Hongkui Zeng, Ed S Lein, Amy Bernard*

Allen Institute, Seattle, United States

Abstract The advancement of single-cell RNA-sequencing technologies has led to an explosion of cell type definitions across multiple organs and organisms. While standards for data and metadata intake are arising, organization of cell types has largely been left to individual investigators, resulting in widely varying nomenclature and limited alignment between taxonomies. To facilitate cross-dataset comparison, the Allen Institute created the common cell type nomenclature (CCN) for matching and tracking cell types across studies that is qualitatively similar to gene transcript management across different genome builds. The CCN can be readily applied to new or established taxonomies and was applied herein to diverse cell type datasets derived from multiple quantifiable modalities. The CCN facilitates assigning accurate yet flexible cell type names in the mammalian cortex as a step toward community-wide efforts to organize multi-source, data-driven information related to cell type taxonomies from any organism.

Introduction

Cell type classification has been central to understanding biological systems for many tissues (e.g., immune system) (Lees *et al.*, 2015) and organisms (e.g., *Caenorhabditis elegans*) (Packer *et al.*, 2019). Identifying and naming cellular components of the brain has been an integral part of neuroscience since the seminal work of Cajal, 1899. Many neuronal cell types, such as neurogliaform, chandelier, Martinotti, and pyramidal cells, have been identified based on highly distinct shape, location, or electrical properties, providing robust and consistent classifications of neuronal cell types and a common vocabulary (Greig *et al.*, 2013; Markram *et al.*, 2004). However, the recent application of high-throughput, quantitative methods such as single-cell or -nucleus transcriptomics (scRNA-seq) (Hodge *et al.*, 2019; Macosko *et al.*, 2015; Saunders *et al.*, 2018; Tasic *et al.*, 2018; Tasic *et al.*, 2016; Zeisel *et al.*, 2018, Zeisel *et al.*, 2015), electron microscopy (Zheng *et al.*, 2018), and whole brain morphology (Winnubst *et al.*, 2019) to cell type classification is enabling more quantitative measurements of similarities among cells and construction of taxonomies (Zeng and Sanes, 2017). The use of scRNA-seq, in particular, for cell type classification has increased exponentially since its introduction a decade ago (Tang *et al.*, 2009), with nearly 2000 published studies and several hundred tools for data analysis (Zappia *et al.*, 2018). These methodological advances are ushering a new era of data-driven classification, by simultaneously expanding the number of measurable features per cell, the number of cells per study, the number of classification studies, and the computational resources required for storing and analyzing this information.

This data explosion has enriched our collective understanding of biological cell types, while simultaneously introducing challenges in cell type classification within individual studies. In the retina, neurons with shared morphology also have consistent connectivity (Jonas and Kording, 2015), spacing, arbor density, arbor stratification (Seung and Sümbül, 2014), and gene expression signatures (Macosko *et al.*, 2015), often with one-to-one correspondences between phenotype and function (Zeng and Sanes, 2017). However, studies combining scRNA-seq with traditional morphological and

***For correspondence:**

jeremym@alleninstitute.org (JAM); amyb@alleninstitute.org (AB)

Competing interests: The authors declare that no competing interests exist.

Funding: See page 18

Received: 12 June 2020

Accepted: 28 December 2020

Published: 29 December 2020

Reviewing editor: Genevieve Konopka, University of Texas Southwestern Medical Center, United States

© Copyright Miller *et al.* This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

electrophysiological characterizations in the brain have found a more complicated relationship in the brain than in retina, with cell types defined by morphology and electrophysiology sometimes containing cells from several cell types defined using gene expression (Gouwens et al., 2020; Kozareva et al., 2020), and some transcriptomically defined types containing cells with multiple morphologies (Hodge et al., 2020; Hodge et al., 2019). Further complicating classification is the overlay of discrete cell type distinctions with graded/continuous properties such as cortical depth (Berg et al., 2020), anterior/posterior and other trajectories across neocortex (Hawrylycz et al., 2012), activity-dependent cell state (Wu et al., 2017), or all simultaneously (Yao et al., 2020b). Furthermore, functional properties observed in matched cell types may diverge across species (Bakken et al., 2020a; Berg et al., 2020; Boldog et al., 2018; Hodge et al., 2019) and as cells advance along trajectories of development (Nowakowski et al., 2017), aging (Tabula Muris Consortium, 2020), and disease (Mathys et al., 2019).

Given this complex landscape, determining fundamental criteria for cell type definition in a given study, and then establishing correspondence to a cell type defined in another study, is often nontrivial and sometimes impossible. Substantial progress has been made toward solving this challenge of 'alignment', whereby datasets collected with genomics assays such as scRNA-seq or snATAC-seq can be used to anchor diverse cell types in a common analysis space (Barkas et al., 2018; Butler et al., 2018; Johansen and Quon, 2019). Alignment has proven effective for matching cell type sequence data collected on different platforms, across multiple data modalities, and even between species where few homologous marker genes show conserved patterns (Bakken et al., 2020a; Bakken et al., 2020b; Hodge et al., 2020; Hodge et al., 2019; Yao et al., 2020a). When combined with experimental methods such as Patch-seq (Cadwell et al., 2016; Fuzik et al., 2016; Gouwens et al., 2020; Scala et al., 2020), which involves application of electrophysiological recording and morphological analysis of single patch-clamped neurons followed by scRNA-seq of cell contents, autoencoder-based dimensionality reduction (Gala et al., 2019) can extend these alignments to bridge distinct modalities. Such analysis strategies provide a mechanism for classifying cell types using data from disparate data sources, allow for annotation transfer between experiments, and are a critical step toward unifying data-driven cell type definitions. However, as new cell type classifications are continually emerging, it is unrealistic to expect complete alignment of all published datasets, but creation of standardized systems for alignment becomes even more important.

Standardized cell type classification needs to include (1) standard nomenclature and (2) centralized and standardized infrastructure associated with cell type classification. Such standards provide a mechanism for storing key information about cell types and assigning explicit links between common cell types identified in different studies. Currently, no standard convention of naming brain cell types is widely followed. Cell types have historically been named by their shape, location, electrical properties, selective neurochemical markers, or even the scientist who discovered them (Betz, 1874; Szentágothai and Arbib, 1974). Now, quantitative clusters that cannot obviously be matched with these types are named on an ad hoc basis, either by assigning generic names like 'interneuron 1' or 'Ex1' and then linking these names to associated figures, tables, or text (Gouwens et al., 2019; Lake et al., 2016; Zeisel et al., 2015), or by chaining critical cell type features in the name itself, resulting in names like 'Neocortex M1 L6 CT pyramidal, Zfp2 non-adapt GLU' (Shepherd et al., 2019). All of these proposals are reasonable for stand-alone projects but make direct comparisons between studies daunting. While several public databases for data storage have been developed (e.g., dbGaP, NeMO, NeuroElectro, Neuromorpho, HuBMAP, etc.), a community-recognized repository for storing and tracking cell type assignments and associated taxonomies does not currently exist. This challenge has been recognized by many (Armañanzas and Ascoli, 2015; DeFelipe et al., 2013; Shepherd et al., 2019) and has been a focus of recent conferences seeking community participation toward a solution (Yuste et al., 2020). Any solution devised to tackle this question should ideally be effective and user-friendly and should directly address some of the ongoing challenges of ontology, data matching, and cell type naming described above in its implementation, providing some amount of immediate standardization of any cell type classifications included therein. This challenge was also addressed at a *Cell Type Ontology Workshop* (Seattle, June 17–18, 2019; hosted by the Allen Institute, Chan Zuckerberg Initiative [CZI] and the National Institutes of Health [NIH]), where input from representatives from the fields of ontology, taxonomy, and neuroscience made recommendations, highlighted best practices, and proposed conventions for naming cell types.

To begin to address these challenges and driven by a practical need to organize vast amounts of multimodal data generated by the Allen Institute and collaborators, we have developed a nomenclature convention aimed at tracking cell type information across multiple datasets. Here we present a generalizable nomenclature convention, the **common cell type nomenclature** (CCN), for matching and tracking cell types across studies. This convention was motivated by methodologies used for management of gene transcript identity tracked across different versions of GENCODE genome builds, allowing comparison of matched types with a common reference or any other taxonomy (Frankish et al., 2019; Harrow et al., 2012). Motivated by gene nomenclature conventions from HGNC (Bruford et al., 2020), the CCN also facilitates assigning accurate yet flexible cell type names in the mammalian cortex as a step toward community-wide efforts to organize multi-source, data-driven information related to cell type taxonomies from any organism. An initial version of the CCN was introduced at <https://portal.brain-map.org/explore/classes> (October 2019), with the intent to encourage discussion and gather feedback for improving subsequent versions, to facilitate collaboration, and to improve shared understanding of the many cell types in the brain.

Results

Overview of proposed nomenclature convention

The problem of defining and naming cell types has many similarities to those of genes in genomics, where there is a practical need to track individual sequencing and assembly results as distinct and self-contained entities, while simultaneously recognizing the goal for a singular reference that the community can use to map sequencing data into a common context (Frankish et al., 2019; Harrow et al., 2012; Kitts et al., 2016). Here, a similar strategy is proposed for cell type nomenclature: Use of a standardized series of identifiers for tracking cell types referenced to individual studies, in addition to providing a mechanism for defining common identifiers (Figure 1A). At the core of the schema are two key concepts: (1) a **taxonomy**, defined as the output of a computational algorithm applied to a specific **dataset**, which must be generated prior to implementation of this schema, and (2) a **cell set**, which can represent any collection of **cells** within a specific taxonomy (see Table 1 for definitions of key terms). These components are generated through the input of data and information generated from analysis that identifies **provisional cell types** (sometimes called **cell types** for convenience). These are analytically relevant cell sets that represent quantitatively derived data clusters defined by whatever classification algorithm generated the taxonomy. Provisional cell types can be organized as the terminal leaves of a hierarchical taxonomy using a **dendrogram**, as a non-hierarchical **community structure**, or both. Taxonomies and cell sets are assigned unique identifier tags, as described below, and additional metadata can be stored alongside these tags for use with future databasing and **ontology** tools. These properties can be tracked using a relational graph or other database service, in a qualitatively similar manner to how transcripts are tracked across different versions of GENCODE genome builds (Frankish et al., 2019).

A major goal of the CCN is to track taxonomies and their associated cell sets by providing an easy-to-understand schema that is widely applicable to new and published taxonomies and that can be implemented through a user-friendly code base. The CCN is compatible with taxonomies generated from either single or multiple modalities, taxonomies applied to cells from overlapping datasets, and **reference taxonomies** (discussed in detail below). Each taxonomy is assigned a unique **taxonomy id** of the format CCN[YYYYMMDD][#], where 'CCN' denotes this nomenclature convention; Y, M, and D represent year, month, and day, respectively; and # is an index for compiling multiple taxonomies on a single day. Each taxonomy can also be assigned metadata, such as species, but such details are outside the scope of the CCN. Within each taxonomy, cell sets (and therefore also provisional cell types) are assigned multiple identifier tags, which are used for different purposes. **Cell set accession IDs** track unique cell sets across the entire universe of taxonomies and are defined as CS[YYYYMMDD][#]_[unique # within taxonomy], where CS stands for 'cell set' and the date and number match the taxonomy id. **Cell set labels** are useful for constructing cell sets from groups of provisional cell types, but can otherwise be ignored. **Cell set aliases** represent descriptors intended for public use and communication, including data-driven terms, historical names, or more generic cell type nomenclature. For convenience these are split into at most one **preferred alias**, which represents the primary tag for public consumption (e.g., the cell type names used in a manuscript), and

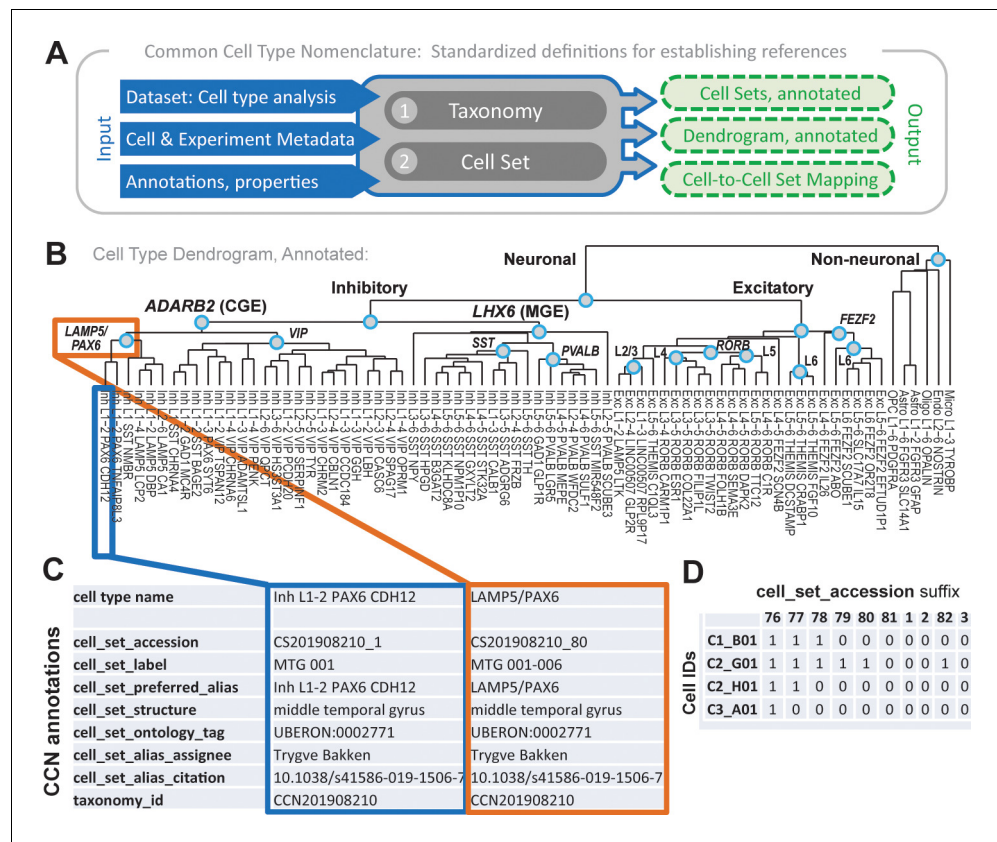


Figure 1. Overview of common cell type nomenclature (CCN) and application to human middle temporal gyrus (MTG). (A) Schematic of CCN components and process. (B–D) Example outputs from the CCN. (B) Annotated dendrogram of cell types in human MTG, along with associated cell type names, reproduced from *Hodge et al., 2019*. Internal nodes with a term (teal circles) represent cell sets with preferred alias tags. (C) CCN annotations for a putative cell type (outlined in blue) and an internal node (outlined in orange) of this dendrogram. (D) Snippet of an output file from the CCN showing cell to cell set mappings as applied to human MTG.

any other **additional aliases**. Additionally, each cell set can have at most one **aligned alias**, which is a biologically driven term that is selected from a controlled vocabulary. Aligned aliases generally are assigned to only a subset of cell sets by alignment to a reference taxonomy, but in principle can be assigned in any taxonomy or taxonomies (e.g., if a rare type is identified that is missing from the reference). The CCN includes a specific system for assigning such aliases in the mammalian cortex using properties that are predicted to be largely preserved across development, anatomical area, and species, which will be discussed in detail. Furthermore, the CCN includes a series of metadata tags tracking the provenance and anatomy of cell sets. The **cell set alias assignee** and **cell set alias citation** indicate the person and permanent data identifier associated with each cell set alias. The **cell set structure** indicates the location in the brain (or body) from where associated cells were primarily collected. Ideally, this will be paired to an established ontology using the **cell set ontology tag**; in this case, we use UBERON since it is designed as an integrated cross-species anatomy ontology (*Haendel et al., 2014*). Finally, the CCN is compatible with incorporation of additional taxonomy-specific or future global cell set metadata or descriptors. This could include donor metadata (e.g., age or sex), summarized cell metadata (e.g., cortical layer or average reads), or additional cell set tags. In particular, the concept of a cell set level is often useful for distinguishing highly specific but statistically less confident provisional cell types from the more general and more statistically robust cell sets.

The CCN is currently in use by the Allen Cell Types Database for transcriptomic taxonomies (<http://celltypes.brain-map.org/rnaseq/>) and is being applied to taxonomies generated by the BRAIN Initiative Cell Census Network (BICCN; <https://biccn.org/>) (*Bakken et al., 2020a*;

Table 1. Glossary of terms.

Terminology used with the common cell type nomenclature (CCN), definitions for use, and examples of how terms are applied. Terms are presented in bold upon first use in the text. This glossary is intended to clarify use for the purposes of the CCN since some terms are open to multiple interpretations, and effective classification requires disambiguation. Asterisks denote terms that represent specific components of the CCN.

Term	Definition	Example
Taxonomy	Set of quantitatively derived data clusters defined by a specific computational algorithm on a specific dataset(s). Taxonomies are given a unique label and can be annotated with metadata about the taxonomy, including details of the algorithms and relevant cell and cell set IDs.	Any clustering result in a cell type classification manuscript
Dataset	Feature information (e.g., gene expression) and associated metadata from a set of cells collected as part of a single project.	Gene expression from 6000 human MOp nuclei
Ontology	A structured controlled vocabulary for cell types.	Cell Ontology
Marker gene(s)	A gene (gene set) which, when expressed in a cell, can be used to accurately assign that cell to a specific cell set.	GAD2; PVALB; CHODL
Taxonomy ID*	An identifier uniquely tagging a taxonomy of the format CCN[YYYYMMDD][#].	CCN201910120
Cell	A single entry in a taxonomy representing data from a single cell (or cell compartment, such as the nucleus). Cells have metadata including a unique ID.	N/A
Cell set	Any tagged group of cells in a taxonomy. This includes cell types, groups of cell types, and potentially other informative groupings (e.g., all cells from one donor, organ, cortical layer, or transgenic line). Cell sets have several IDs and descriptors (as discussed below) and can also have other metadata.	A cell type; a group of cell types; all cells from layer two in MTG; all cells from donor X
Provisional cell type	Quantitatively derived data cluster defined within a taxonomy. This is a specific example of a cell set that is of high importance, as most other cell sets are groupings of one or more provisional cell types. Here, the term 'cell type' is synonymous with 'provisional cell type.'	A cell type defined in a specific study
Dendrogram	A hierarchical organization of provisional cell types defined for a specific taxonomy. Dendrograms have a specific semantic and visualizable structure and include nodes (representing multiple provisional cell types) and leaves (representing exactly one). Not all taxonomies include a dendrogram (e.g., if the structure of cell sets is non-hierarchical).	N/A
Community structure	Non-hierarchical relationships between cell types defined as groups of cell types in a graph.	N/A
Cell set accession ID*	A unique ID across all tracked datasets and taxonomies. This tag labels the taxonomy and numbers each cell type. CS[taxonomy id]_[unique # within taxonomy]	CS201910120_1
Cell set label*	An ID unique within a single taxonomy that is used for assigning cells to cell sets defined as a combination of multiple 'provisional cell types'.	MTG 12 MTG 01–08
Cell set alias*	Any cell set descriptor. It can be defined computationally from the data, or manually based on new experiments, prior knowledge, or a combination of both. Cell aliases beyond the 'preferred' or 'aligned' are defined as 'cell set additional aliases'.	(Any 'cell set aligned alias'); Interneuron 1; Rosehip
Cell set preferred alias*	The primary cell set alias (e.g., what cell types might be called in a publication). This can sometimes match the aligned alias, but not always, and can be left unassigned.	Inh L1-2 PAX6 CDH12; ADARB2 (CGE); Chandelier; [blank]
Cell set aligned alias*	Analogous to 'gene symbol'. At most one biologically driven name for linking matching cell sets across taxonomies and with a reference taxonomy.	L2/3 IT 4; Pvalb 3; Microglia 2
Cell set structure*	The location in the brain (or body) from where cells in the associated set were primarily collected.	Neocortex
Cell set ontology tag*	A tag from a standard ontology (e.g., UBERON) corresponding to the listed cell set structure.	UBERON:0001950
Cell set alias assignee*	Person responsible for assigning a specific cell set alias in a specific taxonomy (e.g., the person who built the taxonomy or uploaded the data, or a field expert).	(First author of manuscript)
Cell set alias citation*	The citation or permanent data identifier corresponding to the taxonomy where the cell set was originally reported.	(Manuscript DOI); [blank]
Reference taxonomy	A taxonomy based on one or a combination of high-confidence datasets, to be used as a baseline of comparison for datasets collected from the same organ system.	Cross-species cortical cell type classification
Morpho-electric(ME) type	A provisional cell type defined using a combination of morphological and electrophysiological features.	ME_Exc_7
Governing body	A forum of subject-matter experts to guide policy and manage change of the CCN and associated ontologies and databasing efforts.	N/A

Adkins et al., 2020; Yao et al., 2020a), a consortia of centers and laboratories working collaboratively to generate, analyze, and share data about brain cell types in human, mouse, macaque, and other non-human primates.

Application of the CCN to cell types in human middle temporal gyrus

A detailed walk-through of how to apply the CCN to a published study on cell types in human middle temporal gyrus (MTG) (*Hodge et al., 2019*) is presented in Materials and methods. In short, **Figure 1B** recapitulates the cell types and associated hierarchy previously published for MTG (*Hodge et al., 2019*). After applying the CCN, each leaf (provisional cell type) and internal node of the dendrogram is assigned the series of cell set tags described above (**Figure 1C**), and every cell is mapped to every cell set (**Figure 1D**). This was all done using a user-friendly set of scripts (<https://github.com/AllenInstitute/nomenclature>). These output files are intended to be directly included as supplemental materials in manuscripts performing cell type classification in any species, and such output for human MTG (and for 17 additional taxonomies) is presented in **Supplementary file 1**.

Naming cell types in mammalian cortex

Mammalian brain cell types inhabit a complex landscape with fuzzy boundaries and complicated correspondences between species and modalities, leading to a variety of disparate solutions for naming cell types. Thus, a challenging and potentially contentious question in cell type classification is how these newly identified cell types should be named, or in the context of the CCN, what should be put in the 'cell set aligned alias' identifier. The CCN utilizes a strategy for naming cell types in the mammalian cortex that includes properties that are cell intrinsic and potentially well conserved between species (**Table 2**). This convention is used as the cell set aligned alias tag in the CCN and ideally should directly map to cell types defined in a relevant ontology (i.e., Cell Ontology [*Diehl et al., 2016*] or Neuron Phenotype Ontology [*Gillespie et al., 2020*]). While admittedly underdeveloped, this convention has been applied to multiple studies of the primary motor cortex (M1; as discussed below) and represents only a starting point for discussion.

For glutamatergic neurons, cell types are named based on predominant layer(s) of localization of cell body (soma) and their predicted projection patterns. The relatively robust laminarity of glutamatergic cell types has been described based on cytoarchitecture in multiple mammalian species for many years (e.g., *Rakic, 1984*), and has been confirmed using RNA in situ hybridization (*Hodge et al., 2019; Tasic et al., 2018; Zeng et al., 2012*), and a combination of layer dissections and scRNA-seq (*Hodge et al., 2019; Tasic et al., 2018*). While in humans many cell types do not follow the layer boundaries defined by cytoarchitecture entirely, laminar patterning is still generally well conserved between human donors and mice (*Hodge et al., 2019*). In adult mouse visual cortex, projection targets for cell types have been explicitly measured using a combination of retrograde labeling and scRNA-seq (*Tasic et al., 2018; Tasic et al., 2016*). By aligning cell types across species, the projection targets in mice can be hypothetically extrapolated to putative projection targets in human or other mammalian species. For example, von Economo neurons are likely to project sub-cortically (*Hodge et al., 2020*). For GABAergic interneurons, developmental origin may define cell types by their canonical marker gene profile established early in development, with *Pvalb* and *Sst* labeling cell types derived from the medial ganglionic eminence and *Vip*, *Sncg*, and *Lamp5* labeling cell types derived from the caudal ganglionic eminence (*DeFelipe et al., 2013*). Non-neuronal cell types have not been a focus of the studies cited and hence they are labeled at a broad cell type level only. However, knowledge from other single-cell transcriptomics studies on microglia (*Hammond et al., 2019; Li et al., 2019*), astrocytes (*Batiuk et al., 2020*), and oligodendrocytes

Table 2. Proposed strategy for naming cortical cell types.

Class	Format	Example
Glutamatergic	[Layer] [Projection] #	L2/3 IT 4
GABAergic	[Canonical gene(s)] #	Pvalb 3
Non-neuronal	[Cell class] #	Microglia 2
Any class	[Historical name] #	Chandelier 1

(*Marques et al., 2016*) could be included in subsequent versions of this convention. In all cases, multiple cell types are present within a given class. While it may not be possible to directly translate every feature to other brain structure or other organs, most of the concepts proposed here could still be followed.

Alignment of established cell sets using reference taxonomies

The CCN presents a flexible data structure for storing key information about taxonomies and cell sets, implemented through reproducible code with standard output files, along with a specific convention for naming mammalian neocortical cell types. It is applicable to taxonomies defined on any data type using any classification algorithm, including hierarchical cell type classification using scRNA-seq. While useful for these reasons alone, a primary utility of the CCN is to facilitate cross-study integration of cell type classifications, in particular when applied in the framework of a reference taxonomy. A reference taxonomy (or reference cell type classification) is any taxonomy based on one or a combination of high-confidence datasets, which can be used as a baseline of comparison for other datasets collected from the same organ system. For example, many researchers favor building a gene expression-based reference taxonomy based on high-throughput, high-resolution single-cell transcriptomics assays and then layering on additional phenotypic data as they become available (*Yuste et al., 2020*). Molecular, physiological, and morphological characteristics of cortical neurons are highly correlated based on simultaneous measurement in individual cells using Patch-seq (*Berg et al., 2020; Gouwens et al., 2020; Scala et al., 2020*), making such a strategy feasible. Many groups are currently performing scRNA-seq analysis in different areas of the brain, from all organs in the human body (*Rozenblatt-Rosen et al., 2017*), from multiple mammalian species (*Geirsdottir et al., 2019*), and across trajectories of development (*Nowakowski et al., 2017*), aging (*Tabula Muris Consortium, 2020*), and disease (*Mathys et al., 2019*). Application of the CCN to these datasets will allow future reference taxonomies to evolve to accommodate these additional complexities by overlaying a common data structure and associated nomenclature.

Reference taxonomies and the CCN are two components of a multi-staged analysis workflow for aligning cell type classifications using datasets collected across multiple labs, from multiple experimental platforms, and from multiple data modalities (*Figure 2*). This workflow accommodates methodological differences in cell type definitions across studies and accommodates changes in reference taxonomies over time. The proposed workflow can be broken down into four broad stages:

1. First, many research teams will independently define cell types, identify their discriminating features, and name them using one of many available experimental and computational strategies. This represents the current state of the field. The CCN may be applied to each dataset independently at this stage.
2. Second, an initial reference cell type classification will be defined by taking the results from one or more (ideally validated) datasets and integrating these data together in a single analysis, if needed. Being high dimensional, high throughput, and relatively low cost, transcriptomics strategies are immediately applicable to many organs and species, and the goal is for reference cell types to be defined using this modality (*Yuste et al., 2020*). The CCN will then be applied to the reference taxonomy as described above – the CCN treats reference taxonomies identically to any other taxonomy. Importantly, aligned aliases should be defined in the reference taxonomy at this stage using a standard naming convention such as the one proposed above.
3. This reference cell type classification can now be used as a comparator for any related datasets, providing a mechanism for transferring prior knowledge about cell types across datasets. Cell sets from existing taxonomies can be renamed using one of the many validated alignment algorithms (e.g., *Barkas et al., 2018; Butler et al., 2018; Gala et al., 2019; Johansen and Quon, 2019*) by integrating data from this taxonomy with the reference, and then updating the cell set aligned alias to match terms defined in the reference. For new datasets, taxonomies can be generated using any clustering or alignment strategy followed by the same mapping and annotation transfer steps.
4. Finally, new versions of the reference cell type classification should be periodically generated using additional data and/or computational methods, and this new classification will now be used as comparator for related datasets. Steps 3 and 4 can iterate at some to-be-defined cadence.

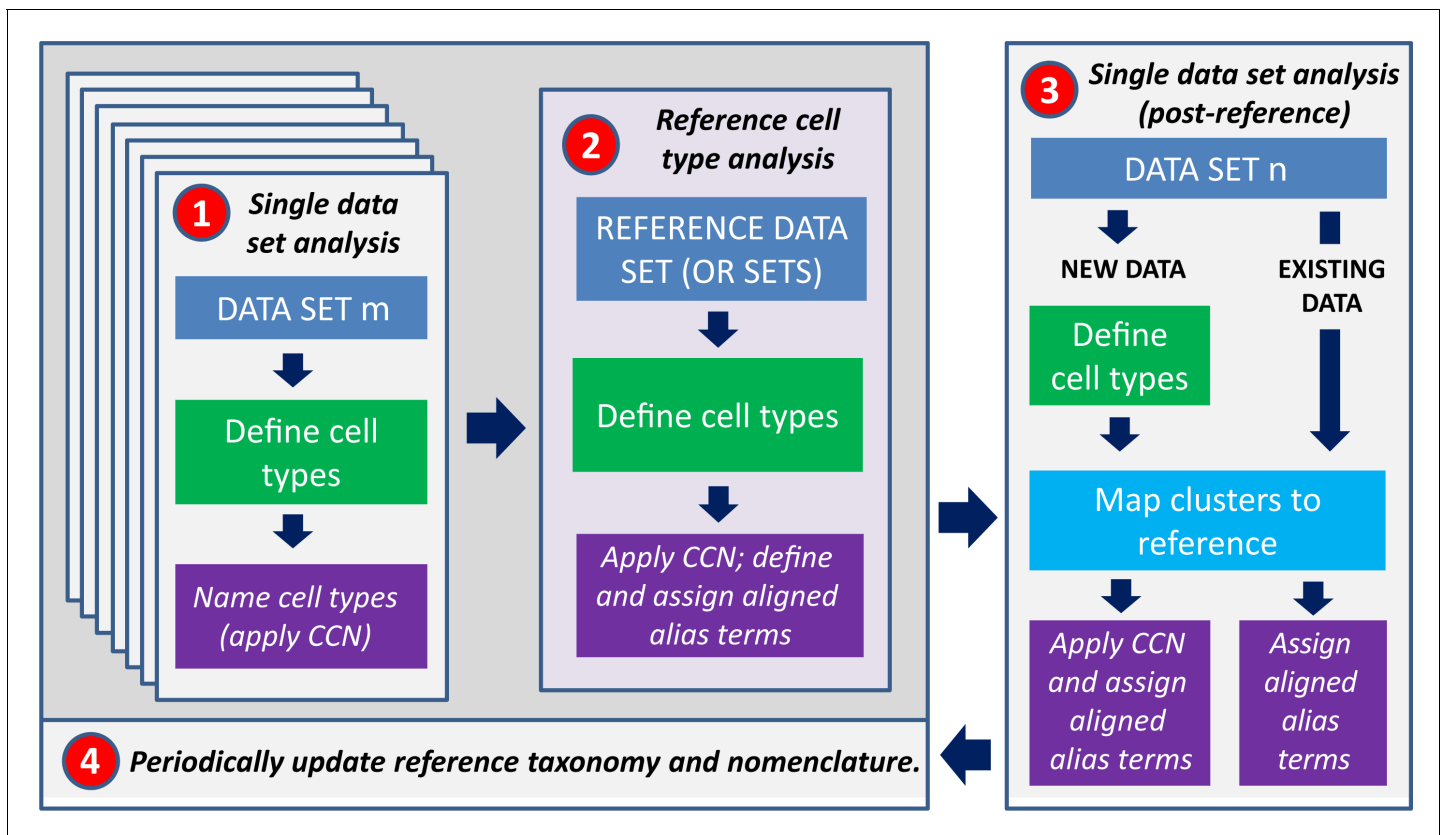


Figure 2. Workflow for assigning types to a given dataset with taxonomy. (1) Cell type classification will initially be performed separately on all taxonomies. (2) One, some, or all of these datasets will be combined into a high-confidence reference taxonomy which can be used as a comparator for any related datasets, by (3) mapping existing and new datasets to the reference taxonomy. (4) The reference will periodically be updated as new datasets and taxonomies are generated.

This workflow provides two complementary strategies to compare between taxonomies without needing to look at gene expression or other quantitative features. First, each taxonomy draws upon a common set of aligned alias terms, which allows for immediate linking of common cell sets between taxonomies (in cases where such information can be reliably assigned). A second strategy is through inclusion of common datasets across multiple taxonomies (reference or otherwise); if cells are assigned to the same cell sets in more than one taxonomy, then the cell sets can be directly linked. As a whole, this workflow provides a general outline for versioned cell type classification that could be specialized as needed for communities studying different organ systems and that provides a starting point for design of future cell taxonomy and nomenclature databases.

Defining a cross-species reference taxonomy in M1

A recent study profiling nearly half a million nuclei in primate and mouse primary motor cortex (M1) presents a taxonomy suitable for defining as a reference taxonomy (Bakken et al., 2020a). This study included single cell data from three separate 'omics' modalities (transcriptomics, epigenetics, and methylation) for mouse, marmoset, and human. Datasets were integrated in two ways. First, epigenetics and methylation datasets were integrated with snRNA-seq data within mouse, marmoset, and human independently (as shown in Figure 3A for human), which demonstrates a consistent genomic profile of cell types within species. Second, snRNA-seq from each species were aligned into a single integrated reference, which identifies cell type homologies across species that were presumably present in the mammalian ancestor to rodents and primates. This evidence-based assumption of cross-species homologies provides a strategy for transferring cell type characteristics from rodent studies (e.g., projection targets) into human, where experiments for making such

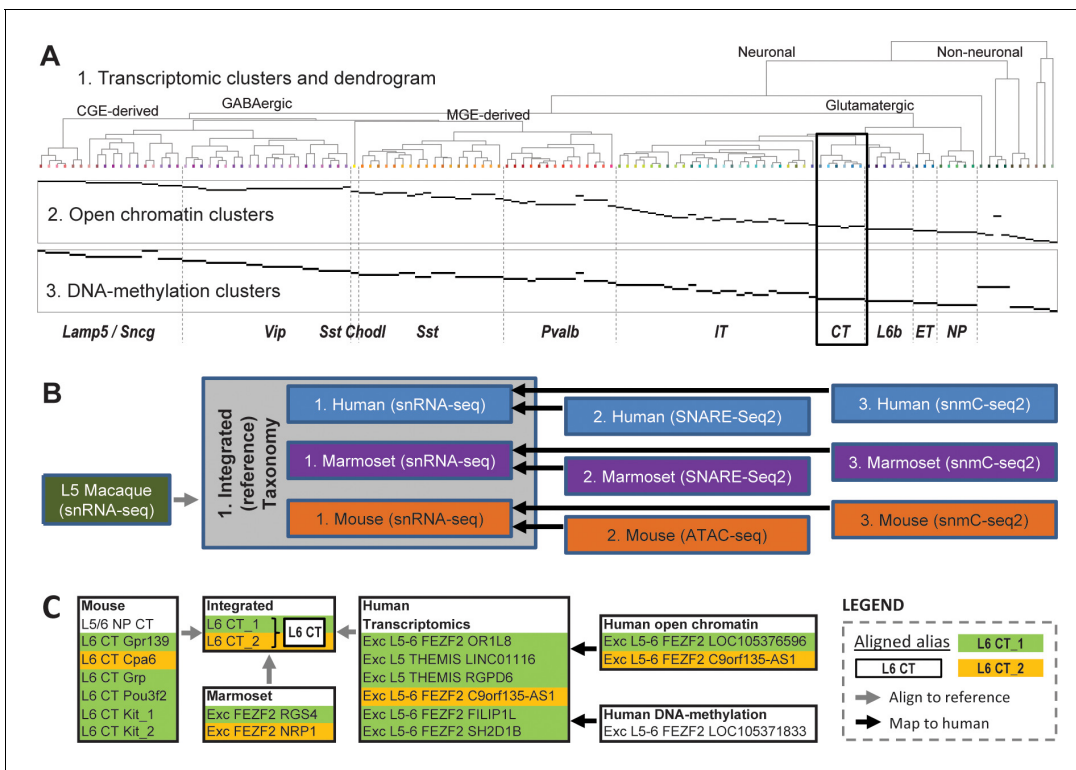


Figure 3. Series of multimodal, cross-species taxonomies in primary motor cortex (M1) demonstrates utility of nomenclature schema. (A) Taxonomies based on transcriptomic ('1'; top), open chromatin ('2'; middle), and DNA methylation ('3'; bottom) in human M1. Epigenomic clusters ('2', '3'; in rows) aligned to RNA-seq clusters ('1') as indicated by horizontal black bars and are also assigned matching cell sets in the relevant taxonomies. Adapted from *Bakken et al., 2020a*. (B) Flow chart showing all 11 taxonomies generated for this project and their connections. The integrated (reference) taxonomy included nuclei collected using snRNA-seq from three species (gray box), with nuclei collected from layer five in macaque mapped to this space post hoc (gray line). Separately, epigenetics taxonomies from human, marmoset, and mouse were aligned to their respective transcriptomics taxonomies (black lines). This entire taxonomic structure is captured by the CCN (see *Supplementary file 1*). (C) An example mapping of corticothalamic (L6 CT) provisional cell types across the human and transcriptomics taxonomies using the CCN (black box in A). Preferred aliases for each taxonomy are used for clarity.

measurements are not yet possible. A total of 11 taxonomies were generated (*Figure 3B*), and all were included in the same nomenclature schema, and the CCN was applied to this set of taxonomies as described above (see *Supplementary Table 3* in *Bakken et al., 2020a* and *Supplementary file 1*). *Figure 3C* shows an example of these cross-species and cross-modality alignments for L6 CT cells, which are divided into two cell sets in the integrated taxonomy (and assigned the aligned alias tags L6 CT_1 and L6 CT_2) and include between one and seven cell sets in the single-modality taxonomies.

This integrated taxonomy (*Figure 3B*, gray box) represents a suitable reference taxonomy for several reasons: first, the data generation, data analysis, and write-up spanned multiple BICCN-funded labs as part of a coordinated consortium project, indicating that this taxonomy was approved by a large subset of the neocortex cell typing community; second, while a number of differences were found between species, 45 core provisional cell types could be aligned across all species with transcriptomics; third, the taxonomies generated using epigenetics and methylation are largely consistent with results of this integrated taxonomy (*Figure 3A*, bottom panels and *Yao et al., 2020a*); and finally, this taxonomy can be linked with other quantitative features (such as morphology, electrophysiology, and expected projection targets) through comparison with mouse studies using complementary modalities such as Patch-seq (*Gouwens et al., 2020; Scala et al., 2020*) and Retro-seq (*Tasic et al., 2018; Tasic et al., 2016*). Using these linkages, aligned aliases of the format proposed in *Table 2* were assigned to cell sets in the integrated taxonomy along with the 10 other species-specific taxonomies using a combination of (1) robust gene markers from the literature, (2) highly

discriminating gene markers in these data, (3) projection targets in mouse, (4) historical names based on cell shape, and (5) broad or low-resolution cell type names (that directly map to ontologies), providing a starting point for how brain cell types could be named. A complete list of aligned aliases used is shown in *Supplementary file 2*.

Applying the CCN to existing and new datasets

For a specific convention to be adopted, both in general or in place of other competing conventions, it needs to be easy to use and immediately useful to the community. For example, many cell type classification studies now use Seurat (*Butler et al., 2018*) for clustering and alignment because it produces believable biological results, and it is implemented in intuitive R code with extensive user guides for non-specialists. As such, Seurat visualizations appear frequently in manuscripts and its file format is used as input for several analysis pipelines. While the usability of the CCN has been established above, the utility of applying it to a single taxonomy in the absence of a centralized database of taxonomies may be less clear. Here five use cases are presented to illustrate how the CCN can be applied to published datasets using scRNA-seq and electrophysiology and morphology in multiple species. These use cases provide immediate utility and also lay a foundation for future databasing and ontology efforts.

Use case 1: Alignment of human MTG taxonomy to M1 reference

The M1 reference taxonomy includes a validated set of aligned aliases that follows the proposed nomenclature for mammalian cortex (*Table 2*) and that can be applied to any other taxonomy. As part of the original analysis (*Bakken et al., 2020a*), nuclei from human MTG (*Hodge et al., 2019*) were aligned to the human M1 dataset. This analysis provides a perfect use case for transferring cell set aligned alias tags from the reference taxonomy to the MTG data (as was done; see Materials and methods). *Figure 4* shows a visualization of glutamatergic types in M1 and MTG, with the color of each square representing the fraction of cells from provisional cell types in each brain region that are assigned to the same alignment cluster, and boxes indicating the aligned alias calls in M1 and their corresponding calls in MTG. While alignment is not perfect for provisional cell types, it is sufficient for matching aligned aliases between cortical areas. These mappings enable biological insights such as the presence of L4-like neurons in M1, where an anatomically defined L4 is not apparent. Likewise, such alignment enables prediction of cell properties such as long-range connectivity (e.g., thalamic inputs), as well as electrophysiology measurements in primary sensorimotor cortices or other brain regions inaccessible to techniques such as Patch-seq. Similar alignments were performed for GABAergic interneurons and non-neuronal cell types (*Supplementary file 1*). Such tagging allows cell sets in human MTG to be directly compared to cell sets from any other taxonomy with the same aligned alias, for example to infer morphological or electrophysiological properties (see Use case 2) or cell class persistence across multimodal phenotypes and developmental stages (see Use case 3) in mouse. Cell sets can even be matched with more distant species using the CCN (see Use case 4), to the extent that such alignment is possible based on the data.

Use case 2: Building a morpho-electric taxonomy

While much effort for cell typing is currently focused on taxonomies based on scRNA-seq datasets, the CCN can equally apply to non-transcriptomic and non-hierarchical taxonomies. For example, a study of mouse visual cortex examined ~1800 cells characterized electrophysiologically by whole-cell patch clamp recordings, and for a subset of these (450 cells), morphological reconstructions were also performed (*Gouwens et al., 2019*). Using a multimodal unsupervised clustering method, the authors identified 20 excitatory and 26 inhibitory **morpho-electric types** (or **me-types**), which are cell types defined using a combination of morphological and electrophysiological features. *Figure 5* shows the application of the CCN to a subset of excitatory (glutamatergic) me-types of that study (see *Supplementary file 1* for application to remaining me-types). The preferred alias and inferred subclass columns show the organization scheme; me-types were organized by broader cell types inferred from transgenic labels, but not placed into a binary hierarchical taxonomic tree (*Gouwens et al., 2019*). Through application of the aligned alias tag, these cell types can be directly linked to cell types defined based on transcriptomics.

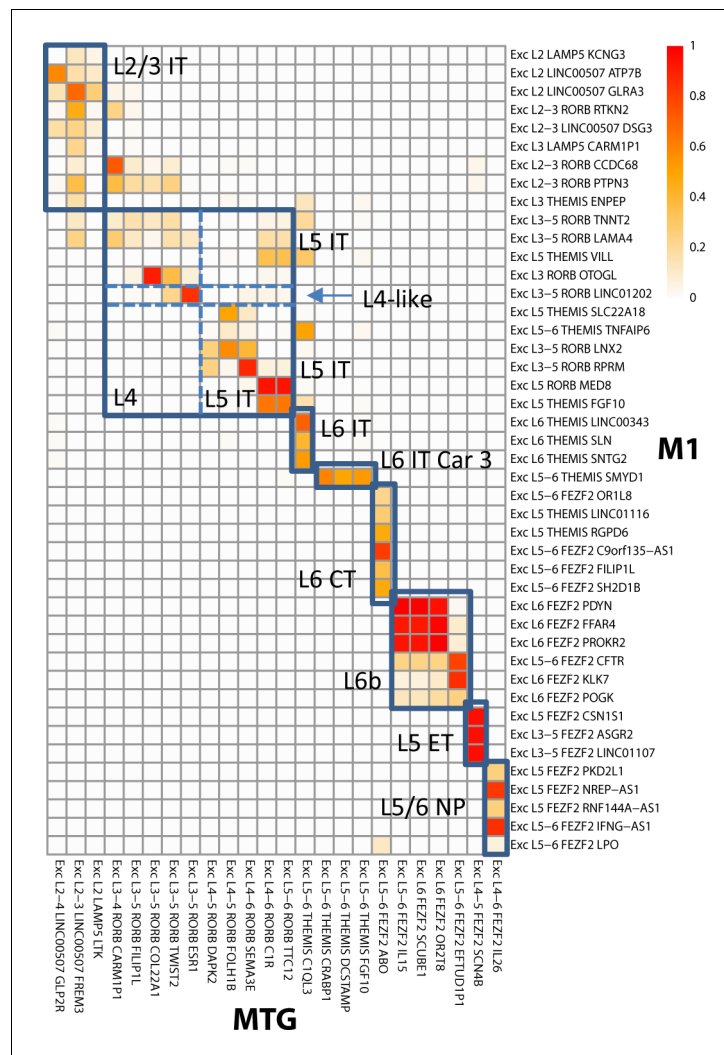


Figure 4. Alignment of glutamatergic cell sets in human middle temporal gyrus (MTG) to a reference primary motor cortex (M1) taxonomy. Cluster overlap heatmap showing the proportion of nuclei from MTG clusters and the reference (M1) clusters that coalesce with a given aligned cluster. Cell sets corresponding to aligned aliases in the MTG and M1 taxonomies are labeled and indicated by blue boxes. Adapted from *Bakken et al., 2020a*.

Use case 3: Exploring an interneuron subclass using multimodal attributes: The ‘Sst Chodl’ class persists across cross-taxonomy matching

Somatostatin-expressing interneurons can be divided into multiple cell types (the specific number differs by taxonomy), some of which include cells that express Chodl in the mouse cerebral cortex (*Tasic et al., 2018; Tasic et al., 2016*). These ‘Sst Chodl’ neurons are rare and, based on expression of specific marker genes, correspond to the only known cortical interneurons with long-range projections (*Tomioka et al., 2005*). Recent studies using the multimodal cell phenotyping method Patch-seq (*Gouwens et al., 2020*) confirmed that ‘Sst Chodl’ cell sets characterized based on morphology and electrophysiology (*Gouwens et al., 2019*) match those defined by transcriptomic profiles (*Tasic et al., 2018; Tasic et al., 2016*). The CCN can be applied to readily represent these ‘Sst Chodl’ cells (and other cell types) matched between all relevant taxonomies, regardless of species or modality through the use of aligned alias tags. For example, *Table 3* shows all cell sets from *Bakken et al., 2020a* (*Figure 3B*) associated with Sst Chodl cells, which all have ‘Sst Chodl’ in the aligned alias (with one exception noted below). In mouse, all three modalities have a single ‘Sst Chodl’ cell type, which can be linked to a matched type in V1Sp due to its highly distinct gene expression patterning that is conserved across brain regions (*Tasic et al., 2018; Yao et al., 2020b*).




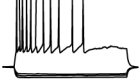

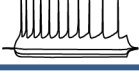









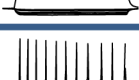



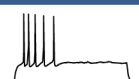

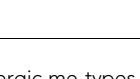
Preferred alias	Inferred subclass	Description	Example Morphology	Example Electrophys.
ME_Exc_7	L2/3 IT	Wide, short L2/3; RS adapting		
ME_Exc_14	L4	Tufted (sparse) L4; RS adapting		
ME_Exc_18	L4	Non-tufted L4; RS adapting		
ME_Exc_13	L5 IT	Tufted L5; RS adapting		
ME_Exc_1	L5 CF	Thick-tufted L5; RS low R _p , sharp sag		
ME_Exc_8	L5 NP	Tufted (sparse basal) L5; RS adapting, large sag		
ME_Exc_6	L6 IT	Wide, short L6a & tufted (large basal) L5; RS adapting		
ME_Exc_11	L6 IT	Inverted L6a, b; RS adapting		
ME_Exc_2	L6 CT	Narrow L6a; RS adapting		
ME_Exc_4	L6 CT	Narrow L6a; RS transient		
ME_Exc_9	L6b	Subplate L6b		

Figure 5. Application of common cell type nomenclature (CCN) to glutamatergic me-types in the mouse visual cortex. Excitatory (glutamatergic) me-types from *Gouwens et al., 2019* that have been incorporated into the nomenclature schema. Eleven of the original 20 excitatory me-types are shown as examples. Representative morphologies and electrophysiological responses are shown to illustrate the differences between types. The ‘inferred subclass’ calls perfectly map to cell set aligned aliases from the reference M1 taxonomy in *Figure 3*, except that L5 CF (corticofugal) is an additional alias for L5 ET, and cells sets corresponding to L4, L6 IT, and L6 CT (blue boxes) have been added to the taxonomy.

This transcriptomic cell type is similarly linked to the ‘Sst Chodl’ cell type in the integrated transcriptomic (reference) taxonomy, which lists ‘long-range projecting Sst’ as an additional alias to formalize the cross-modal correspondence. In human, the RNA-seq and ATAC-seq have one-to-one correspondences, but for DNA methylation (DNAm), Inh L1-5 SST AHR aligns with several Sst cell types including Sst Chodl (likely due to the rarity of this cell type). Cell sets from the methylation- and epigenetics-based taxonomies include an additional alias that list the cell set labels in transcriptomics taxonomy, directly linking these cell types. Therefore, while Inh L1-5 SST AHR does not have ‘Sst Chodl’ as its aligned alias, the cell set label ‘RNA-seq 040, 046–047, 050–052, 068 in CCN201912131’ indicates the inclusion of ‘Sst Chodl’ cells (RNA-seq 040). In marmoset, where fewer

Table 3. Nomenclature for ‘Sst Chodl’ cell sets cited in *Bakken et al., 2020a*.

Relevant common cell type nomenclature (CCN) entities and taxonomy metadata, including the cell set additional alias that links to cell set labels from relevant transcriptomics taxonomies. All listed cell sets have a cell set structure of ‘primary motor cortex’ and a cell set ontology tag of ‘UBERON:0001384’.

#	Cell set preferred alias	Cell set label	Cell set accession	Cell set aligned alias	Cell set additional alias
1	Inh L1-6 SST NPY	RNA-seq 040	CS201912131_40	Sst Chodl	
2	Inh L1-5 SST AHR	DNA _m 12	CS202002272_12		RNA-seq 040, 046–047, 050–052, 068 in CCN201912131
3	Inh L1-6 SST NPY	ATAC-seq 08	CS202002273_8	Sst Chodl	RNA-seq 040 in CCN201912131
4	Inh SST NPY	RNA-seq 01	CS201912132_1	Sst Chodl	
5	Sst Chodl	RNA-seq 028	CS202002013_28	Sst Chodl	
6	Sst Chodl	DNA _m 09	CS202002276_9	Sst Chodl	RNA-seq 028 in CCN202002013
7	Sst Chodl	ATAC-seq 10	CS202002277_10	Sst Chodl	RNA-seq 028 in CCN202002013
8	Sst Chodl	Integrated 14	CS202002270_14	Sst Chodl	Long-range projecting Sst

#	Cell set alias assignee	Cell set alias citation	Taxonomy id	Species	Modality
1	Nikolas Jorstad	10.1101/2020.03.31.016972	CCN201912131	Human	RNA-seq
2	Wei Tian	10.1101/2020.03.31.016972	CCN202002272	Human	DNA _m
3	Blue Lake	10.1101/2020.03.31.016972	CCN202002273	Human	ATAC-seq
4	Fenna Krienen	10.1101/2020.03.31.016972	CCN201912132	Marmoset	RNA-seq
5	Zizhen Yao	10.1101/2020.02.29.970558	CCN202002013	Mouse	RNA-seq
6	Hanqing Liu	10.1101/2020.02.29.970558	CCN202002276	Mouse	DNA _m
7	Yang Li	10.1101/2020.02.29.970558	CCN202002277	Mouse	ATAC-seq
8	Nikolas Jorstad	10.1101/2020.03.31.016972	CCN202002270	All	RNA-seq

cells were collected, an ‘Sst Chodl’ cell set is only found with transcriptomics. Explicitly linking cell sets in this way provides multiple potential points of comparison with other studies, including studies of disease or development. For example, a study of interneuron development in E14 mice found that the ‘Sst Chodl’ cells were severely affected by *Sox6* removal during interneuron migration (*Munguba et al., 2019*), and cell class definitions observed in the mature brain may have foundational roles in cortical patterning.

Use case 4: Alignment of cell types from reptilian and mammalian cortex using the CCN

While the focus of this study is the mammalian cortex, the CCN framework is applicable to other organs and more distant species. As an example use case, a single-cell transcriptomics study of turtle and lizard pallium found GABAergic interneuron and non-neuronal cell types to be homologous with those seen in mouse cortex (*Tosches et al., 2018*). In many cases, these cell types expressed shared gene markers, suggesting a shared evolutionary origin across 320 million years of evolution in amniote vertebrates. These types include astrocytes (GFAP), oligodendrocytes (MBP), oligodendrocyte precursor cells (OLIG1 and PDGFRA), microglia (C1QC), GABAergic interneurons as a whole (GAD1 and GAD2), and Sst+ interneurons (SST). Reptilian analogs for other CGE- and MGE-derived GABAergic types were also identified, although interestingly neither VIP nor PVALB marker genes are expressed in reptiles. Application of the CCN to the taxonomies presented for the turtle demonstrates the utility of this approach (*Supplementary file 1*).

Assignment of aligned aliases for non-neuronal cells and GABAergic interneurons is straightforward, with ‘PV-like’ interneurons (cell types i11–i13 from *Tosches et al., 2018*) assigned ‘PVALB’, and similar alignments for other types. In contrast, the correspondence between reptilian and mammalian glutamatergic cells is more complicated. Reptiles have a three layer pallium and only the anterior dorsal cortex (representing a small fraction of pallium) is comparable with the six-layer mammalian neocortex (*Jarvis, 2009; Tosches et al., 2018*). RNA-seq in combination with in situ hybridization identified two distinct sublayers of turtle layer 2: a superficial L2a (cell types e07–e08) and a

deeper L2b (e13–e16), which seem to correspond with mammalian deep layer and upper layer neurons, respectively, suggesting there was likely an inversion of layers in one clade. However, all of these cell types coexpress genes found in mutually exclusive L2/3, L4, and L5a intra-telencephalic neurons (e.g., SATB2, RORB, and RFX3) along with extra-telencephalic projection neurons (e.g., BCL11B, TBR1, and SOX5), suggesting either a lack of homologous cell types between clades or at least a change in the core transcription factor regulatory programs. Thus, with the level of resolution presented in this study, no aligned aliases (beyond the broadest) are assigned for glutamatergic types. This points to the importance of having measurements in other modalities, for example, local and long-range connectivity, that may help establish homologies or bolster claims of clade-specific cellular innovations. Overall, the CCN provides a mechanism for assigning a standard nomenclature for cell types found in the reptilian cortex and linking these types with a mammalian neocortical reference at the level of resolution resolved in the taxonomy.

Use case 5: Comparison of novel to existing taxonomies

The first four use cases represent specific examples of how taxonomies from different brain regions, modalities, and species can be presented in the framework of the CCN to make published inferences more easily accessible to a naive reader. These represent specific examples of a more general use case for scientists, who may want to compare their newly generated taxonomy to what is currently known about cell types. The ideal application for this scenario is a centralized database for taxonomy integration with an associated ontology and annotation capabilities; such a framework is well beyond the scope of this manuscript, but solutions are underway. As a starting point for this goal, **Supplementary file 1** presents output files from the CCN for 18 taxonomies (including all taxonomies discussed herein; **Table 4**) that have been annotated with the aligned aliases from the M1 reference taxonomy presented in **Figure 3**. Transcriptomics-based taxonomies were collected from human, non-human primate, mouse, and reptile, and span multiple neocortical areas. In addition, several of these are matched to taxonomies collected using other modalities such as morphology,

Table 4. Taxonomies with applied CCN.

Table showing the set of taxonomies included in **Supplementary file 1**. All taxonomies include the annotated nomenclature table. Asterisk (*) and carrot (⌘) indicate that the updated dendrogram and cell to cell set mapping files are also included for that taxonomy, respectively. CCN202002270 is the reference taxonomy presented in **Figure 3B**.

Taxonomy id	Description	Reference
CCN201810310*	Mouse VISp + ALM (from the <i>Tasic et al., 2018</i>)	<i>Tasic et al., 2018</i>
CCN201908210*	Human MTG (from the <i>Tasic et al., 2018</i>)	<i>Hodge et al., 2019</i>
CCN201908211*	Joint mouse/human analysis (slight modification from <i>Hodge et al., 2019</i>)	<i>Hodge et al., 2019</i>
CCN201912130	Human M1 taxonomy using 10× data	<i>Bakken et al., 2020a</i>
CCN201912131	Human M1 taxonomy using Smart-seq and 10x data	<i>Bakken et al., 2020a</i>
CCN201912132	Marmoset M1 taxonomy using 10× data	<i>Bakken et al., 2020a</i>
CCN202002013*	Mouse MOp BICCN taxonomy using multiple RNAseq datasets	<i>Yao et al., 2020a</i>
CCN202002270	Cross species (integrated) transcriptomics taxonomy	<i>Bakken et al., 2020a</i>
CCN202002271	Macaque transcriptomics taxonomy, layer 5/6 only	<i>Bakken et al., 2020a</i>
CCN202002272	Human DNA methylation taxonomy	<i>Bakken et al., 2020a</i>
CCN202002273	Human ATAC-seq taxonomy	<i>Bakken et al., 2020a</i>
CCN202002274	Marmoset DNA methylation taxonomy	<i>Bakken et al., 2020a</i>
CCN202002275	Marmoset ATAC-seq taxonomy	<i>Bakken et al., 2020a</i>
CCN202002276	Mouse DNA methylation taxonomy	<i>Yao et al., 2020a</i>
CCN202002277	Mouse ATAC-seq taxonomy	<i>Yao et al., 2020a</i>
CCN202005150	Mouse inhibitory neurons in VISp defined using electrophysiology, morphology, and transcriptomics	<i>Gouwens et al., 2020</i>
CCN201906170	Mouse neurons in VISp defined using electrophysiology and morphology	<i>Gouwens et al., 2019</i>
CCN201805250	Turtle pallium transcriptomics taxonomy	<i>Tosches et al., 2018</i>

electrophysiology, epigenetics, and methylation. Such breadth provides multiple avenues of entry into this framework for annotation of novel datasets and allows for a more flexible implementation of the specific analysis workflow described in **Figure 2**. In particular, instead of requiring alignment of new datasets to the reference taxonomy, new datasets can be aligned with any taxonomy from **Table 4**, and information about cell type can then be inferred from any cell sets in any included taxonomy with a common aligned alias as the matched cell set. If this process is applied to novel taxonomies and the output files are included as supplemental materials in any resulting manuscript, then these taxonomies can be included in any future centralized database with minimal effort, providing a richer reference for further study.

Discussion

The complexity of cell types taxonomies and their generation now requires conventions and methodology to capture and communicate essential knowledge derived from experiments. The CCN provides a schema and workflow that allows scientists to organize their cell types within a single dataset and to link taxonomies using the aligned alias and other alias terms. However, the CCN is currently a stand-alone nomenclature schema that lacks the centralization and governance of gene-based standards proposed by the HUGO Gene Nomenclature Committee (HGNC) (**Bruford et al., 2020**) and does not yet have a mechanism for integrating with underlying data and metadata.

These shortcomings would be addressed through linking cell type ontology curation with corresponding databases. Ontology curation would allow users to associate data-derived cell sets to common usage terms from prior knowledge, and connect directly with the well-annotated ontology tools that are available for broader classifications (e.g., the Cell Ontology, <http://www.obofoundry.org/ontology/>). In addition, aligned aliases defined in reference taxonomies could represent a starting point for expansion of existing ontologies to presumptive cell types defined using other data-driven approaches (such as the terms in **Table 2** for cortical neurons). Centralizing a location for taxonomies, their associated cell sets, and underlying datasets could provide a more robust ecosystem for comparing relevant nomenclature information, other metadata, and the primary data itself. Such databases can be implemented using knowledge graph-based models (**Alshahrani et al., 2017**; **Waagmeester et al., 2020**), permitting traversal across a *data, information, knowledge, and wisdom* hierarchy (**Rowley, 2007**). A potential presentation could be a 'Cell Type Card', instantiated as a web-accessible reference that compiles information about a specific cell set in a standardized summary. Not unlike a periodic table in structure, this concept has been implemented for genes (<http://www.genecards.org>), and as a prototype using transcriptomically defined cell types in mouse hippocampus and cortex (**Yao et al., 2020b**).

Incorporating community input on the definition and management of cell type standards will be necessary as new experiments are performed and additional evidence emerges. A cell type standards **governing body** would ideally be responsible for vetting ontologies for organizing data, controlled vocabulary for assigning cell type nomenclature, and will need to define a process for submission to ensure that critical data and metadata can be stored in a robust database. Deciding which taxonomies to include as reference taxonomies, along with frequency of updates, and how to address the breadth of brain regions, data modalities, cross-species reconciliation, and stochasticity of developmental and disease trajectories is essential. Organizing such a governance framework represents an important step and efforts are under way through BRAIN Initiative-funded initiatives, but is beyond the scope of work presented here.

This work presents a framework that is a modest step in a long and iterative process. With cross-disciplinary partnership and ever-increasing data, refinement of this proposed convention is expected. Together with collaborators, the Allen Institute has begun to combine ontology development, data integration, and nomenclature formalization efforts with the aim of facilitating cell type standards for the neuroscience community. Together with the goals articulated as part of the NIH BICCN and Brain Cell Data Center (BCDC) (<https://biccn.org/>), we seek to provide access to the diverse cell types in the human, mouse, and marmoset brain. The Allen Brain Map Community Forum (<https://community.brain-map.org/c/cell-taxonomies/>) has a dedicated space for discussion related to cell taxonomy refinement, to promote open and accessible opportunity for exchanging ideas and suggesting improvements. Beyond brain, whole-body projects seeking to categorize cell types, such as the NIH Common Fund-supported Human BioMolecular Atlas Program (HuBMAP, <https://>

hubmapconsortium.org/) and the Human Cell Atlas consortium (<https://www.humancellatlas.org/>), will also need to leverage organizational conventions such as this, for comparable purposes that are practical and promote scientific rigor. The authors look forward to engagement with emerging communities and forums as evolution of cell classification methods continues.

Materials and methods

User-friendly executable code for applying the CCN is available on GitHub (<https://github.com/AllenInstitute/nomenclature>). This repository aims at providing a set of standardized terms and files that are immediately useful and also formatted to seed any future or in-process platform for cell type characterization and annotation. It is written as a user-friendly script in the R programming language (<https://www.R-project.org>) that includes specific details for how to apply the CCN, along with a set of example input files from a published study on cell types in human MTG (*Hodge et al., 2019*).

Step-by-step application of the CCN to human MTG

This section addresses how to apply the CCN to an example taxonomy, from human MTG. Three inputs are required to run the CCN: (1) a cell type taxonomy (not necessarily hierarchical), (2) a cell metadata file with cluster assignments (and optionally additional information), and (3) optional manual annotations of cell sets (e.g., aliases), which typically would be completed during taxonomy generation. Example files for human MTG are saved in the repository's data folder. Once all files are downloaded and the workspace is set up, several global variables are set, which propagate to each cell set as a starting point, and which can be updated for specific cell sets later in the process. A unique taxonomy_id of the format CCN[YYYYMMDD][#] is chosen, which will match the prefix for cell set accession IDs. To ensure uniqueness across all taxonomies, taxonomy_ids are tracked in a public-facing database, with future plans to transfer these to a more permanent solution that will also provide storage for accompanying CCN output files and relevant metadata. In addition, values for the cell set assignee, citation, structure, and ontology tag are defined, along with the prefix(es) for the cell set label. For human MTG, 'CCN201908210', 'Trygve Bakken', '10.1038/s41586-019-1506-7', 'middle temporal gyrus', 'UBERON:0002771', and 'MTG' are used, respectively. Next, the dendrogram is read in as the starting point for defining cell sets by including both provisional cell types (terminal leaves) and groups of cell types with similar expression patterns (internal nodes). **Figure 1B** shows the annotated dendrogram in human MTG provided in the GitHub repository, under which are displayed the names of cell types presented in *Hodge et al., 2019*. These provisional cell types were named using an entirely data-driven strategy: (cell class) (L)(cortical layers of localization) (canonical marker gene) ([optional] specific marker gene), as discussed in *Hodge et al., 2019*.

The main script takes the preset values and dendrogram as input, assigns accession ids and labels for each cell set, and then outputs an intermediate table and a dendrogram with all CCN labels defined in the previous section (**Figure 1C**). By default, the provisional cell types are assigned their original name from the dendrogram as preferred alias (e.g., 'Inh L1-2 PAX6 CDH12'), while this field is left blank for internal nodes. For all cell sets, fields for additional and aligned alias are also initially left blank. Cell set labels are formatted as the label prefix (e.g., 'MTG') followed by a list of the cell set labels of all included provisional cell types. For example, the 'LAMP5/PAX6' node in human MTG includes the first six cell types in the tree and therefore has the cell set label of 'MTG 001–006'. The table with these CCN tags for each cell set is then written to a csv file for manual annotation, which includes two critical aspects: (1) creation of new cell sets and (2) updating CCN tags for any cell sets. Cell sets corresponding to groups of relevant cell types either based on biological relevance (e.g., LAMP5-associated cell types in MTG) or as defined using a non-hierarchical computational strategy can be added at this step. In addition, cell sets corresponding to metadata rather than cell types can also be added. For example, in human MTG, 'CS201908210_154' corresponds to the set of nuclei collected from neurosurgical tissue and is given a cell set label of 'Metadata 1' and a preferred alias of 'Neurosurgical'.

After finalizing these cell sets, they can then be annotated to include additional aliases based on known literature (e.g., assigning 'basket' or 'fast-spiking' to relevant PVALB+ cell sets), along with the assignees and citations from which such aliases were derived (e.g., 'Nathan Gouwens' and '10.1101/2020.02.03.932244'). As another example, Inh L1-4 LAMP5 LCP2 corresponds to Rosehip

cells (see [Boldog et al., 2018](#)) and therefore an additional alias for this cell type is 'Rosehip'. The structures and associated ontology tags could also be updated at this stage. For example, previous studies in mice suggest that most non-neuronal and GABAergic cell types are conserved across cortical areas ([Tasic et al., 2018](#); [Yao et al., 2020b](#)). Although not done here, relevant cell sets could be generalized to an anatomic structure such as 'Neocortex' (UBERON:0001950). A final component of manual annotation is to update relevant cell sets with an aligned alias (e.g., a common usage term), which is critical for comparison of taxonomies in the CCN. In this case, aligned aliases for all cell sets were assigned by comparison with the human M1 reference ([Bakken et al., 2020a](#)), as shown in [Figure 4](#) for glutamatergic neurons and as described in Use case 1. It is important to note that this step requires a previous computational alignment (or some other strategy to match cell sets) to use as evidence prior to assignment of the aligned alias; cell set alignment itself is not performed as part of the CCN.

After completing the manual annotations, the updated table is read back into R for additional dendrogram annotation and for mapping of cells to cell sets. Dendrograms are revised to include the new cell sets and annotations, and then output in a few standard formats (.RData, .json, and .pdf) for ontology construction and other downstream uses. Individual cells are then mapped to cell sets using the cell metadata table, which includes a unique cell identifier, provisional cell type classification, and other optional metadata. Cells are then mapped to cell sets representing one or more provisional cell type using the annotated dendrogram and/or the updated nomenclature table using the cell set label tag. Finally, cells are mapped to remaining cell sets (if any) using custom scripts. This results in a table of binary calls (0 = no, 1 = yes), indicating exclusion or inclusion of each cell in each cell set ([Figure 1D](#)), which is written to another csv file as part of the process. This format is designed to allow for probabilistic mapping of cells to cell sets, which is beyond the scope of this manuscript. *These output files are intended to be directly included as supplemental materials in manuscripts performing cell type classification in any species.* In addition, the GitHub repository will be updated to include conversion functions to allow input into future community-accepted cell type databases, as such resources become available. [Supplementary file 1](#) includes a table of applied nomenclature for all taxonomies discussed in this manuscript, along with cell to cell set mappings for a few example taxonomies.

Acknowledgements

The authors would like to acknowledge general input and considerations on aspects of cell nomenclature from attendees and affiliates of the workshop, 'Defining an Ontological Framework for a Brain Cell Type Taxonomy: Single-Cell -omics and Data-Driven Nomenclature', held in Seattle, WA, June 2019, including Alex Pollen, Alex Wiltschko, Alexander Diehl, Andrea Beckel-Mitchener, Angela Pisco, Anna Maria Masci, Anna-Kristin Kaufmann, Anton Arkhipov, Aviv Regev, Becky Steck, Bishen Singh, Brad Spiers, Chris Mungall, Christophe Benoist, Cole Trapnell, Dan Geschwind, David Holmes, David Osumi-Sutherland, Davide Risso, Deep Ganguli, Detlev Arendt, Ed Callaway, Eran Mukamel, Evan Macosko, Fenna Krienen, Gerald Quon, Giorgio Ascoli, Gordon Shepherd, Guoping Feng, Hanqing Liu, Jay Shendure, Jens Hjerling-Leffler, Jessica Peterson, Joe Ecker, John Feo, John Marioni, John Ngai, Jonah Cool, Josh Huang, Junhyong Kim, Kelly Street, Kelsey Montgomery, Kara Woo, Lindsay Cowell, Lucy Wang, Luis De La Torre Ubieta, Mark Musen, Maryann Martone, Michele Solis, Ming Zhan, Nicole Vasilevsky, Olga Botvinnik, Olivier Bodenreider, Owen White, Peter Hunter, Peter Kharchenko, Rafael Yuste, Rahul Satija, Richard Scheuermann, Samuel Kerrien, Sean Hill, Sean Mooney, Sten Linnarsson, Tim Jacobs, Tim Tickle, Tom Nowakowski, Uygur Sümbül, Vilas Menon, and Yong Yao. We thank the NIH and CZI for generous co-sponsorship of this workshop. We would also like to acknowledge the many members of the Allen Institute, past and present, who contribute to or support the development of data and analysis of brain cell types - and the organization of this information, particularly Christof Koch, Kimberly Smith, Zizhen Yao, Carol Thompson, Rebecca Hodge, Jonathan Ting, Lucas Graybuck, Thuc Nguyen, Jim Berg, Staci Sorensen, Nik Jorstad, Susan Sunkin, Stefan Mihalas, Rob Young, Tim Fliss, Lydia Ng, Shoaib Mufti, and Stephanie Mok. Research and methods reported in this publication were supported by the Allen Institute, and the National Institute of Mental Health of the National Institutes of Health under award numbers U19MH114830 (to HZ) and U01MH114812 (to ESL). The content is solely the responsibility of the authors and does

not necessarily represent the official views of the National Institutes of Health. The authors would like to thank the Allen Institute founder, Paul G Allen, for his vision, encouragement and support.

Additional information

Funding

Funder	Grant reference number	Author
Allen Institute		Jeremy A Miller Nathan W Gouwens Bosiljka Tasic Forrest Collman Cindy TJ van Velthoven Trygve E Bakken Michael J Hawrylycz Hongkui Zeng Ed S Lein Amy Bernard
National Institute of Mental Health	U19MH114830	Hongkui Zeng
National Institute of Mental Health	U01MH114812	Ed S Lein

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Jeremy A Miller, Conceptualization, Resources, Data curation, Software, Formal analysis, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing; Nathan W Gouwens, Conceptualization, Resources, Data curation, Software, Formal analysis, Visualization, Methodology, Writing - original draft, Writing - review and editing; Bosiljka Tasic, Conceptualization, Resources, Funding acquisition, Methodology, Writing - review and editing; Forrest Collman, Cindy TJ van Velthoven, Trygve E Bakken, Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Visualization, Methodology, Writing - review and editing; Michael J Hawrylycz, Writing - review and editing; Hongkui Zeng, Ed S Lein, Resources, Funding acquisition, Writing - review and editing; Amy Bernard, Conceptualization, Resources, Methodology, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Jeremy A Miller  <https://orcid.org/0000-0003-4549-588X>
 Nathan W Gouwens  <https://orcid.org/0000-0001-8429-4090>
 Bosiljka Tasic  <http://orcid.org/0000-0002-6861-4506>
 Forrest Collman  <http://orcid.org/0000-0002-0280-7022>
 Cindy TJ van Velthoven  <http://orcid.org/0000-0001-5120-4546>
 Trygve E Bakken  <http://orcid.org/0000-0003-3373-7386>
 Michael J Hawrylycz  <http://orcid.org/0000-0002-5741-8024>
 Hongkui Zeng  <http://orcid.org/0000-0002-0326-5878>
 Ed S Lein  <http://orcid.org/0000-0001-9012-6552>
 Amy Bernard  <https://orcid.org/0000-0003-2540-1153>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.59928.sa1>

Author response <https://doi.org/10.7554/eLife.59928.sa2>

Additional files

Supplementary files

- Supplementary file 1. Output files from applying the CCN on 17 taxonomies. This file contains annotated cell sets from all 17 taxonomies shown in **Table 4** along with annotated dendrograms and cell to cell set assignments for a subset of these taxonomies. This file is available on GitHub (<https://github.com/AllenInstitute/nomenclature>).
- Supplementary file 2. A set of aligned aliases in mammalian M1, reproduced from **Bakken et al., 2020a**. These terms are also applicable to other cortical areas, representing a starting point for future cell type classification efforts and for ontology curation. InterLex identifiers are provided in parentheses when available (**Adkins et al., 2020**).
- Transparent reporting form

Data availability

This work describes the creation of a cell type nomenclature convention that will, with adoption by the community, become a standard. The data cited is open data available through the Allen Institute web portal, <https://brain-map.org>. An open Forum is available to engage the community in further development, at <https://portal.brain-map.org/explore/classes/nomenclature>. Data referenced in this study is also made available according to the terms of NIH's Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN), through the Brain Cell Data Center portal, <https://biccn.org/> and <https://biccn.org/data>.

References

- Adkins RS**, Aldridge AI, Allen S, Ament SA, An X, Armand E, Ascoli GA, Bakken TE, Bandrowski A, Banerjee S, Barkas N, Bartlett A, Bateup HS, Margarita Behrens M, Berens P, Berg J, Bernabucci M, Bornaert Y, Bertagnolli D, Biancalani T, et al. 2020. A multimodal cell census and atlas of the mammalian primary motor cortex. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.10.19.343129>
- Alshahrani M**, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R. 2017. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* **33**:2723–2730. DOI: <https://doi.org/10.1093/bioinformatics/btx275>, PMID: 28449114
- Armañanzas R**, Ascoli GA. 2015. Towards the automatic classification of neurons. *Trends in Neurosciences* **38**: 307–318. DOI: <https://doi.org/10.1016/j.tins.2015.02.004>, PMID: 25765323
- Bakken TE**, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, Crow M, Hodge RD, Krienen FM, Sorensen SA, Eggermont J, Yao Z, Aevermann BD, Aldridge AI, Bartlett A, Bertagnolli D, Casper T, Castanon RG, Crichton K, Daigle TL, et al. 2020a. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.03.31.016972>
- Bakken TE**, van Velthoven CTJ, Menon V, Hodge RD, Yao Z, Nguyen TN, Grayback LT, Horwitz GD, Bertagnolli D, Goldy J, Garren E, Parry S, Casper T, Shehata SI, Barkan ER, Szafer A, Levi BP, Dee N, Smith KA, Sunkin SM, et al. 2020b. Single-cell RNA-seq uncovers shared and distinct axes of variation in dorsal LGN neurons in mice, non-human primates and humans. Cold Spring Harbor Laboratory.
- Barkas N**, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, Kharchenko PV. 2018. Wiring together large single-cell RNA-seq sample collections. *bioRxiv*. DOI: <https://doi.org/10.1101/460246>
- Batiuk MY**, Martirosyan A, Wahis J, de Vin F, Marneffe C, Kusserow C, Koeppen J, Viana JF, Oliveira JF, Voet T, Ponting CP, Belgard TG, Holt MG. 2020. Identification of region-specific astrocyte subtypes at single cell resolution. *Nature Communications* **11**:1220. DOI: <https://doi.org/10.1038/s41467-019-14198-8>
- Berg J**, Sorensen SA, Ting JT, Miller JA, Chartrand T, Buchin A, Bakken TE, Budzillo A, Dee N, Ding S-L, Gouwens NW, Hodge RD, Kalmbach B, Lee C, Lee BR, Alfiler L, Baker K, Barkan E, Beller A, Berry K. 2020. Human cortical expansion involves diversification and specialization of supragranular intratelencephalic-projecting neurons. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.03.31.018820>
- Betz W**. 1874. Anatomischer nachweis zweier gehirncentra. *Centralblatt Für Die Medizinischen Wissenschaften* **12**:578–599.
- Boldog E**, Bakken TE, Hodge RD, Novotny M, Aevermann BD, Baka J, Bordé S, Close JL, Diez-Fuertes F, Ding SL, Faragó N, Kocsis ÁK, Kovács B, Maltzer Z, McCarrison JM, Miller JA, Molnár G, Oláh G, Ozsvár A, Rózsa M, et al. 2018. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. *Nature Neuroscience* **21**:1185–1195. DOI: <https://doi.org/10.1038/s41593-018-0205-2>, PMID: 30150662
- Bruford EA**, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. 2020. Guidelines for human gene nomenclature. *Nature Genetics* **52**:754–758. DOI: <https://doi.org/10.1038/s41588-020-0669-3>, PMID: 32747822

- Butler A**, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**:411–420. DOI: <https://doi.org/10.1038/nbt.4096>, PMID: 29608179
- Cadwell CR**, Palasantza A, Jiang X, Berens P, Deng Q, Yilmaz M, Reimer J, Shen S, Bethge M, Tolias KF, Sandberg R, Tolias AS. 2016. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature Biotechnology* **34**:199–203. DOI: <https://doi.org/10.1038/nbt.3445>, PMID: 26689543
- Cajal R**. 1899. *La Textura Del Sistema Nerviosa Del Hombre Y Los Vertebrados*. Nicolas Moya.
- DeFelipe J**, López-Cruz PL, Benavides-Piccione R, Bielza C, Larrañaga P, Anderson S, Burkhalter A, Cauli B, Fairén A, Feldmeyer D, Fishell G, Fitzpatrick D, Freund TF, González-Burgos G, Hestrin S, Hill S, Hof PR, Huang J, Jones EG, Kawaguchi Y, et al. 2013. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Reviews Neuroscience* **14**:202–216. DOI: <https://doi.org/10.1038/nrn3444>, PMID: 23385869
- Diehl AD**, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, Van Slyke CE, Vasilevsky NA, Haendel MA, Blake JA, Mungall CJ. 2016. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics* **7**:44. DOI: <https://doi.org/10.1186/s13326-016-0088-7>, PMID: 27377652
- Frankish A**, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**:D766–D773. DOI: <https://doi.org/10.1093/nar/gky955>, PMID: 30357393
- Fuzik J**, Zeisel A, Máté Z, Calvigioni D, Yanagawa Y, Szabó G, Linnarsson S, Harkany T. 2016. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nature Biotechnology* **34**:175–183. DOI: <https://doi.org/10.1038/nbt.3443>, PMID: 26689544
- Gala R**, Gouwens N, Yao Z, Budzillo A, Penn O, Tasic B, Murphy G, Zeng H, Sümbül U. 2019. A coupled autoencoder approach for multi-modal analysis of cell types. In: Wallach H, Larochelle H, Beygelzimer A, Fox E, Garnett R (Eds). *Advances in Neural Information Processing Systems*. Curran Associates, Inc. p. 9267–9276.
- Geirsdottir L**, David E, Keren-Shaul H, Weiner A, Bohlen SC, Neuber J, Balic A, Giladi A, Sheban F, Dutertre CA, Pfeifle C, Peri F, Raffo-Romero A, Vizioli J, Matiassek K, Scheiwe C, Meckel S, Mätz-Rensing K, van der Meer F, Thormodsson FR, et al. 2019. Cross-Species Single-Cell analysis reveals divergence of the primate microglia program. *Cell* **179**:1609–1622. DOI: <https://doi.org/10.1016/j.cell.2019.11.010>, PMID: 31835035
- Gillespie TH**, Tripathy S, Mf S, Martone ME, Hill SL. 2020. The neuron phenotype ontology: a FAIR approach to proposing and classifying neuronal types. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.09.01.278879>
- Gouwens NW**, Sorensen SA, Berg J, Lee C, Jarsky T, Ting J, Sunkin SM, Feng D, Anastassiou CA, Barkan E, Bickley K, Blesie N, Braun T, Brouner K, Budzillo A, Caldejon S, Casper T, Castelli D, Chong P, Crichton K, et al. 2019. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature Neuroscience* **22**:1182–1195. DOI: <https://doi.org/10.1038/s41593-019-0417-0>, PMID: 31209381
- Gouwens NW**, Sorensen SA, Baftizadeh F, Budzillo A, Lee BR, Jarsky T, Alfiler L, Baker K, Barkan E, Berry K, Bertagnolli D, Bickley K, Bomben J, Braun T, Brouner K, Casper T, Crichton K, Daigle TL, Dalley R, de Frates RA, et al. 2020. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. *Cell* **183**:935–953. DOI: <https://doi.org/10.1016/j.cell.2020.09.057>, PMID: 33186530
- Greig LC**, Woodworth MB, Galazo MJ, Padmanabhan H, Macklis JD. 2013. Molecular logic of neocortical projection neuron specification, development and diversity. *Nature Reviews Neuroscience* **14**:755–769. DOI: <https://doi.org/10.1038/nrn3586>, PMID: 24105342
- Haendel MA**, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics* **5**:21. DOI: <https://doi.org/10.1186/2041-1480-5-21>, PMID: 25009735
- Hammond TR**, Dufort C, Dissing-Olesen L, Giera S, Young A, Wysoker A, Walker AJ, Gergits F, Segel M, Nemesh J, Marsh SE, Saunders A, Macosko E, Ginhoux F, Chen J, Franklin RJM, Piao X, McCarroll SA, Stevens B. 2019. Single-Cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex Cell-State changes. *Immunity* **50**:253–271. DOI: <https://doi.org/10.1016/j.immuni.2018.11.004>, PMID: 30471926
- Harrow J**, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, et al. 2012. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research* **22**:1760–1774. DOI: <https://doi.org/10.1101/gr.135350.111>, PMID: 22955987
- Hawrylycz MJ**, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, et al. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**:391–399. DOI: <https://doi.org/10.1038/nature11405>, PMID: 22996553
- Hodge RD**, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, Yao Z, Eggermont J, Höllt T, Levi BP, Shehata SI, Aevermann B, Beller A, Bertagnolli D, Brouner K, Casper T, et al. 2019. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**:61–68. DOI: <https://doi.org/10.1038/s41586-019-1506-7>, PMID: 31435019
- Hodge RD**, Miller JA, Novotny M, Kalmbach BE, Ting JT, Bakken TE, Aevermann BD, Barkan ER, Berkowitz-Cerasano ML, Cobbs C, Diez-Fuertes F, Ding SL, McCarrison J, Schork NJ, Shehata SI, Smith KA, Sunkin SM, Tran DN, Venepally P, Yanny AM, et al. 2020. Transcriptomic evidence that von economo neurons are

- regionally specialized extratelencephalic-projecting excitatory neurons. *Nature Communications* **11**:1172. DOI: <https://doi.org/10.1038/s41467-020-14952-3>, PMID: 32127543
- Jarvis ED. 2009. Evolution of the Pallium in Birds and Reptiles. In: Binder M. D, Hirokawa N, Windhorst U (Eds). *Encyclopedia of Neuroscience*. Springer. p. 1390–1400. DOI: https://doi.org/10.1007/978-3-540-29678-2_3165
- Johansen N, Quon G. 2019. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biology* **20**:166. DOI: <https://doi.org/10.1186/s13059-019-1766-4>, PMID: 31412909
- Jonas E, Kording K. 2015. Automatic discovery of cell types and microcircuitry from neural connectomics. *eLife* **4**:e04250. DOI: <https://doi.org/10.7554/eLife.04250>, PMID: 25928186
- Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research* **44**:D73–D80. DOI: <https://doi.org/10.1093/nar/gkv1226>, PMID: 26578580
- Kozareva V, Martin C, Osorno T, Rudolph S, Guo C, Vanderburg C, Nadaf N, Regev A, Regehr W, Macosko E. 2020. A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. Cold Spring Harbor Laboratory.
- Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung HL, Chen S, Vijayaraghavan R, Wong J, Chen A, Sheng X, Kaper F, Shen R, Ronaghi M, Fan JB, Wang W, Chun J, et al. 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**:1586–1590. DOI: <https://doi.org/10.1126/science.aaf1204>, PMID: 27339989
- Lees JR, Azimzadeh AM, Ding Y, Webb TJ, Bromberg JS. 2015. Cells of the immune system. In: Li CX, Jevnikar A. M (Eds). *Transplant Immunology*. John Wiley & Sons, Ltd. p. 25–47.
- Li Q, Cheng Z, Zhou L, Darmanis S, Neff NF, Okamoto J, Gulati G, Bennett ML, Sun LO, Clarke LE, Marschallinger J, Yu G, Quake SR, Wyss-Coray T, Barres BA. 2019. Developmental heterogeneity of microglia and brain myeloid cells revealed by deep Single-Cell RNA sequencing. *Neuron* **101**:207–223. DOI: <https://doi.org/10.1016/j.neuron.2018.12.006>, PMID: 30606613
- Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly parallel Genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**:1202–1214. DOI: <https://doi.org/10.1016/j.cell.2015.05.002>, PMID: 26000488
- Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C. 2004. Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience* **5**:793–807. DOI: <https://doi.org/10.1038/nrn1519>, PMID: 15378039
- Marques S, Zeisel A, Codeluppi S, van Bruggen D, Mendanha Falcão A, Xiao L, Li H, Häring M, Hochgerner H, Romanov RA, Gyllborg D, Muñoz Manchado A, La Manno G, Lönnerberg P, Floriddia EM, Rezayee F, Ernfors P, Arenas E, Hjerling-Leffler J, Harkany T, et al. 2016. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**:1326–1329. DOI: <https://doi.org/10.1126/science.aaf6463>, PMID: 27284195
- Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, Martorell AJ, Ransohoff RM, Hafler BP, Bennett DA, Kellis M, Tsai LH. 2019. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**:332–337. DOI: <https://doi.org/10.1038/s41586-019-1195-2>, PMID: 31042697
- Munguba H, Nikouei K, Hochgerner H, Oberst P, Kouznetsova A, Ryge J, Batista-Brito R, Munoz-Manchado AB, Close J, Linnarsson S, Hjerling-Leffler J. 2019. Transcriptional maintenance of cortical somatostatin interneuron subtype identity during migration. *bioRxiv*. DOI: <https://doi.org/10.1101/593285>
- Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, Haeussler M, Sandoval-Espinosa C, Liu SJ, Velmeshev D, Ounadjela JR, Shuga J, Wang X, Lim DA, West JA, Leyrat AA, Kent WJ, Kriegstein AR. 2017. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**:1318–1323. DOI: <https://doi.org/10.1126/science.aap8809>, PMID: 29217575
- Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J, Waterston RH, Murray JI. 2019. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**:eaax1971. DOI: <https://doi.org/10.1126/science.aax1971>, PMID: 31488706
- Rakic P. 1984. Organizing Principles for Development of Primate Cerebral Cortex. In: Sharma S. C (Ed). *Organizing Principles of Neural Development*. Springer. p. 21–48. DOI: https://doi.org/10.1007/978-1-4684-4802-3_2
- Rowley J. 2007. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science* **33**:163–180. DOI: <https://doi.org/10.1177/0165551506070706>
- Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. 2017. The human cell atlas: from vision to reality. *Nature* **550**:451–453. DOI: <https://doi.org/10.1038/550451a>, PMID: 29072289
- Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, Goeva A, Nemes J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll SA. 2018. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**:1015–1030. DOI: <https://doi.org/10.1016/j.cell.2018.07.028>, PMID: 30096299
- Scala F, Kobak D, Bernabucci M, Bernaerts Y, Cadwell CR, Castro JR, Hartmanis L, Jiang X, Laturus S, Miranda E, Mulherkar S, Tan ZH, Yao Z, Zeng H, Sandberg R, Berens P, Tolias AS. 2020. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*. DOI: <https://doi.org/10.1038/s41586-020-2907-3>, PMID: 33184512
- Seung HS, Sümbül U. 2014. Neuronal cell types and connectivity: lessons from the retina. *Neuron* **83**:1262–1272. DOI: <https://doi.org/10.1016/j.neuron.2014.08.054>, PMID: 25233310

- Shepherd GM**, Marenco L, Hines ML, Migliore M, McDougal RA, Carnevale NT, Newton AJH, Surles-Zeigler M, Ascoli GA. 2019. Neuron names: a gene- and Property-Based name format, with special reference to cortical neurons. *Frontiers in Neuroanatomy* **13**:25. DOI: <https://doi.org/10.3389/fnana.2019.00025>, PMID: 30949034
- Szentágothai J**, Arbib MA. 1974. Conceptual models of neural organization. *Neurosciences Research Program Bulletin* **12**:305–510. PMID: 4437759
- Tabula Muris Consortium**. 2020. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**:590–595. DOI: <https://doi.org/10.1038/s41586-020-2496-1>, PMID: 32669714
- Tang F**, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**:377–382. DOI: <https://doi.org/10.1038/nmeth.1315>, PMID: 19349980
- Tasic B**, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, et al. 2016. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* **19**:335–346. DOI: <https://doi.org/10.1038/nn.4216>, PMID: 26727548
- Tasic B**, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, Goldy J, Garren E, Economo MN, Viswanathan S, Penn O, Bakken T, Menon V, Miller J, Fong O, Hirokawa KE, Lathia K, Rimorin C, Tieu M, Larsen R, et al. 2018. Shared and distinct transcriptomic cell types across neocortical Areas. *Nature* **563**:72–78. DOI: <https://doi.org/10.1038/s41586-018-0654-5>, PMID: 30382198
- Tomioaka R**, Okamoto K, Furuta T, Fujiyama F, Iwasato T, Yanagawa Y, Obata K, Kaneko T, Tamamaki N. 2005. Demonstration of long-range GABAergic connections distributed throughout the mouse neocortex. *European Journal of Neuroscience* **21**:1587–1600. DOI: <https://doi.org/10.1111/j.1460-9568.2005.03989.x>
- Tosches MA**, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G. 2018. Evolution of pallium, Hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**:881–888. DOI: <https://doi.org/10.1126/science.aar4237>, PMID: 29724907
- Waagmeester A**, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mittra E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, et al. 2020. Wikidata as a knowledge graph for the life sciences. *eLife* **9**:e52614. DOI: <https://doi.org/10.7554/eLife.52614>, PMID: 32180547
- Winnubst J**, Bas E, Ferreira TA, Wu Z, Economo MN, Edson P, Arthur BJ, Bruns C, Rokicki K, Schauder D, Olbris DJ, Murphy SD, Ackerman DG, Arshadi C, Baldwin P, Blake R, Elsayed A, Hasan M, Ramirez D, Dos Santos B, et al. 2019. Reconstruction of 1,000 projection neurons reveals new cell types and organization of Long-Range connectivity in the mouse brain. *Cell* **179**:268–281. DOI: <https://doi.org/10.1016/j.cell.2019.07.042>, PMID: 31495573
- Wu YE**, Pan L, Zuo Y, Li X, Hong W. 2017. Detecting activated cell populations using Single-Cell RNA-Seq. *Neuron* **96**:313–329. DOI: <https://doi.org/10.1016/j.neuron.2017.09.026>, PMID: 29024657
- Yao Z**, Liu H, Xie F, Fischer S, Sina Boeshaghi A, Adkins RS, Aldridge AI, Ament SA, Pinto-Duarte A, Bartlett A, Margarita Behrens M, Van den Berge K, Bertagnolli D, Biancalani T, Bravo HC, Casper T, Colantuoni C, Creasy H, Crichton K, Crow M, et al. 2020a. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.02.29.970558>
- Yao Z**, Nguyen TN, van Velthoven CTJ, Goldy J, Sedeno-Cortes AE, Baftizadeh F, Bertagnolli D, Casper T, Crichton K, Ding S-L, Fong O, Garren E, Glandon A, Gray J, Graybuck LT, Hawrylycz MJ, Hirschstein D, Kroll M, Lathia K, Levi B, et al. 2020b. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *bioRxiv*. DOI: <https://doi.org/10.1101/2020.03.30.015214>
- Yuste R**, Hawrylycz M, Aalling N, Aguilar-Valles A, Arendt D, Arnedillo RA, Ascoli GA, Bielza C, Bokharaie V, Bergmann TB, Bystron I, Capogna M, Chang Y, Clemens A, de Kock CPJ, DeFelipe J, Dos Santos SE, Dunville K, Feldmeyer D, Fiáth R, et al. 2020. A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nature Neuroscience* **23**:1456–1468. DOI: <https://doi.org/10.1038/s41593-020-0685-8>, PMID: 32839617
- Zappia L**, Phipson B, Oshlack A. 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology* **14**:e1006245. DOI: <https://doi.org/10.1371/journal.pcbi.1006245>, PMID: 29939984
- Zeisel A**, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S. 2015. Brain structure. cell types in the mouse cortex and Hippocampus revealed by single-cell RNA-seq. *Science* **347**:1138–1142. DOI: <https://doi.org/10.1126/science.aaa1934>, PMID: 25700174
- Zeisel A**, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. 2018. Molecular architecture of the mouse nervous system. *Cell* **174**:999–1014. DOI: <https://doi.org/10.1016/j.cell.2018.06.021>, PMID: 30096314
- Zeng H**, Shen EH, Hohmann JG, Oh SW, Bernard A, Royall JJ, Glattfelder KJ, Sunkin SM, Morris JA, Guillozet-Bongaarts AL, Smith KA, Ebbert AJ, Swanson B, Kuan L, Page DT, Overly CC, Lein ES, Hawrylycz MJ, Hof PR, Hyde TM, et al. 2012. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**:483–496. DOI: <https://doi.org/10.1016/j.cell.2012.02.052>, PMID: 22500809
- Zeng H**, Sanes JR. 2017. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience* **18**:530–546. DOI: <https://doi.org/10.1038/nrn.2017.85>, PMID: 28775344
- Zheng Z**, Lauritzen JS, Perlman E, Robinson CG, Nichols M, Milkie D, Torrens O, Price J, Fisher CB, Sharifi N, Calle-Schuler SA, Kmecova L, Ali IJ, Karsh B, Trautman ET, Bogovic JA, Hanslovsky P, Jefferis G, Kazhdan M,

Khaury K, et al. 2018. A complete electron microscopy volume of the brain of adult *Drosophila melanogaster*. *Cell* **174**:730–743. DOI: <https://doi.org/10.1016/j.cell.2018.06.019>, PMID: 30033368