Database Guide: ABC Atlas - Single-Cell & Spatial Transcriptomics Mouse Datasets, *Background Knowledge*

TABLE OF CONTENTS

Introduction

There are multiple datasets on the Allen Institute website - and while the information you will need for your project is all there, it could be difficult to navigate the website, and some of the scientific concepts might be new to you! We hope this resource guide will be able to provide some additional help in navigating the Allen Institute's various resources, as well as give you some basic information and suggestions of outside resources to further your understanding of the experimental methods used.

Allen Brain Cell (ABC) Atlas Overview

The ABC Atlas is an online, browser-based tool that allows researchers to visualize multiple high-quality single-cell datasets for free. These datasets are from complex and expensive experiments, with unprecedented sample size/depth from high-impact papers. Therefore, to promote open science, the ABC Atlas allows scientists across the globe to visualize this data.

The data in the ABC Atlas shows the transcriptomic profile of brain cells in various brain regions. The ABC Atlas helps scientists find the **cell types** and **gene expression (transcriptomics**) of those cell types for mouse and human brains.

What is a cell type?

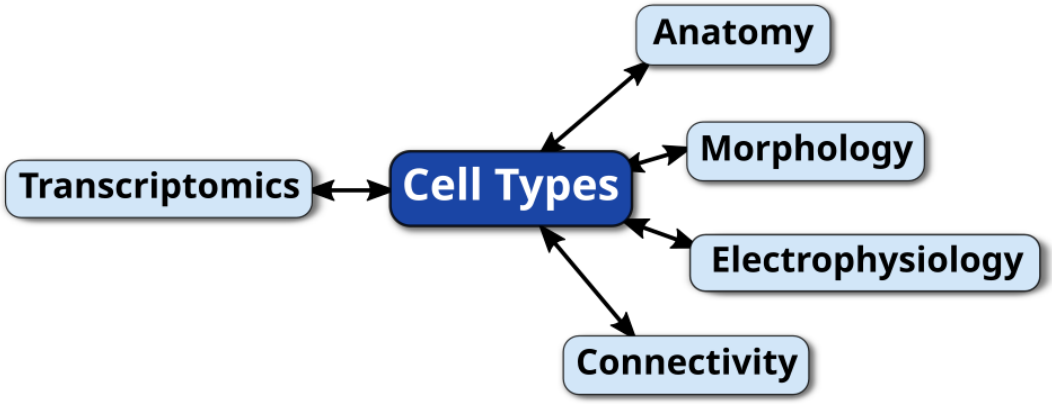
To understand how a multicellular organism works, we must understand each cell's organization and function. Cells can be grouped into types and within a type, the structure and function are distinct from other types (Arendt 2008). *Cells within the same cell type are more similar to each other than they are to cells in a different cell type*.

Evolutionary pressures lead to changes in genetics, development, and the environment in which organisms live. In response to these changes, cells differentiate and have specificity resulting in distinct cell types. Cell types in the brain can be defined by properties of anatomy, morphology, electrophysiology, and connectivity to other cells. These properties determine function.

A cell type would ideally be defined by cellular function, but scientists currently don't fully understand every cell's function. Therefore, we define cell types by the properties we can measure.

Historically, anatomy and morphology were the standard properties used to define cell types; Cajal, Golgi, Nissl, and many other scientists have been drawing neurons for over 200 years. As technology advanced, we added electrophysiology via methods such as patch-clamp that allow us

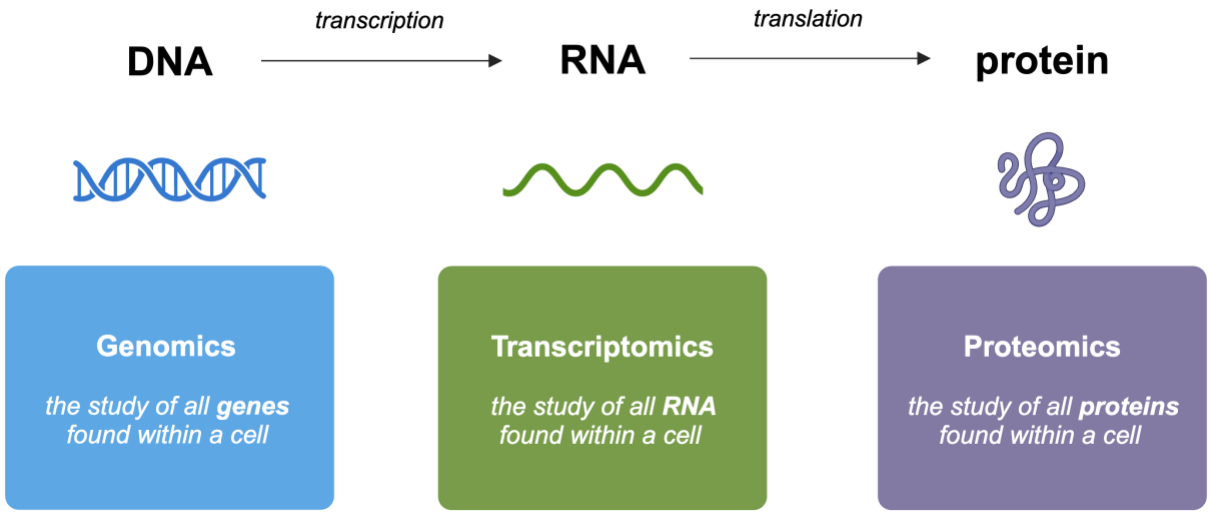
to categorize cells by how they send electrical signals. Technology is still advancing, and in the future, Allen Institute hopes to add connectivity as another cell type defining property, but we don't have fully connectivity yet. The cornerstone of defining cell types, though, is **transcriptomic data**.



For more about cell types visit Cell Types – Transcriptomics, Morphology, & Electrophysiology Lecture Deck or Cell Types 101 Webinar.

WHAT IS TRANSCRIPTOMICS?

Transcriptomics: the study of all the RNA transcripts found within a cell at a given point in time. The genome describes DNA, while the transcriptome describes RNA that was transcribed. Note that "gene expression" often is shorthand referring to RNA levels. Transcriptomics comes from the words "transcription" (the process in which DNA is converted to RNA) and "omics" (study of all of a particular type of molecule found within a living system. To learn more about "omics", here is a guide on "What is Omics".



Why would scientists want to measure RNA expression over DNA?

1. DNA would be the same for each cell within a single organism (minus sex cells)
2. RNA is the code that is translated when ribosomes build proteins
3. Transcriptomics reveals gene expression for the selected tissue or cell

There are two general methods for measuring transcriptomics – microarray and RNA-sequencing (RNAseq). **We will focus on RNAseq, which is the method used in the ABC Atlas.**

There are three methods of RNAseq:

- **Bulk RNAseq:** Measures average expression level of genes across hundreds to millions of cells but there is no single-cell level resolution.
- **Single-cell RNAseq (scRNAseq):** Measures expression of genes within individual cells but is more resource-intensive than bulk RNAseq.
- **Spatial Transcriptomics:** Measures expression of genes for individual cells while keeping each cell's spatial location in the brain. It is currently one of the most expensive methods to measure transcriptomics.

An analogy to explain bulk vs. single-cell vs. spatial is to imagine cells as fruit. There are 3 ways you can consume your fruit.

1. You can drink a fruit smoothie where you blend the fruit, allowing you to drink the smoothie quickly, but you will lose the identity of the fruit (bulk sequencing).
2. You could eat each fruit individually, but this is time-consuming (single-cell RNAseq).
3. You could eat a fancy fruit tart where you know the spatial location of every fruit relative to other fruits, but this is an expensive method (spatial transcriptomics).

We will focus on single-cell RNAseq and spatial transcriptomics as used in the ABC Atlas.

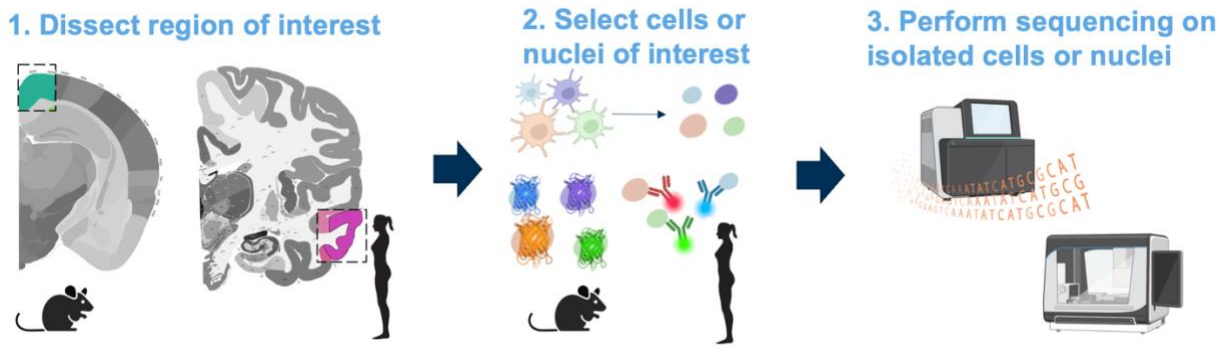
Single-Cell RNA Sequencing

The transcriptomic profile in the ABC Atlas is generated using single-cell/nucleus RNA sequencing (sc/snRNAseq). For more information about the technical differences between single-cell vs single-

nucleus here is a webinar, but a general rule of thumb in the ABC Atlas is that single-cell was used for mouse data collection, and single-nucleus was used for human data collection.

The general workflow to get sc/snRNAseq is:

1. Dissect the tissue for the region of interest
2. Select cells or nuclei of interest via methods like homogenizing and staining or genetically encoded fluorescent reporter
3. Once we have isolated the RNA from our samples, we send our RNA to other biotechnical companies that run established, proprietary protocols, often called "**platforms**". These platforms often involve in-depth protocols that most labs do not have the resources to carry out, so it is a common practice to outsource these steps. The data featured within the ABC Atlas was mostly sequenced using SMART-seq and 10x Chromium platforms. To learn more about these technologies in-depth, here is an informational video.



Because there are thousands of different RNA molecules transcribed within a given cell, the scientist conducting an experiment will select relevant RNA for their project to be sequenced. Since the ABC Atlas was designed for scientists across the globe to help study their cells and genes of interest, a wide range of genes were selected, with over 30,000 genes selected for the scRNAseq mouse dataset.

Spatial Transcriptomics

Cells in different locations are differentially affected by diseases and have different behavioral functions (Topolnik & Tamboli 2022, Gabitto et al, 2023, Dobrzanski et al. 2022). Spatial transcriptomics is a way for us to measure transcriptomics while retaining the *location* within the tissue. Scientists leverage platforms developed by various biotechnical companies that create machines that do both sequencing techniques and microscopy.

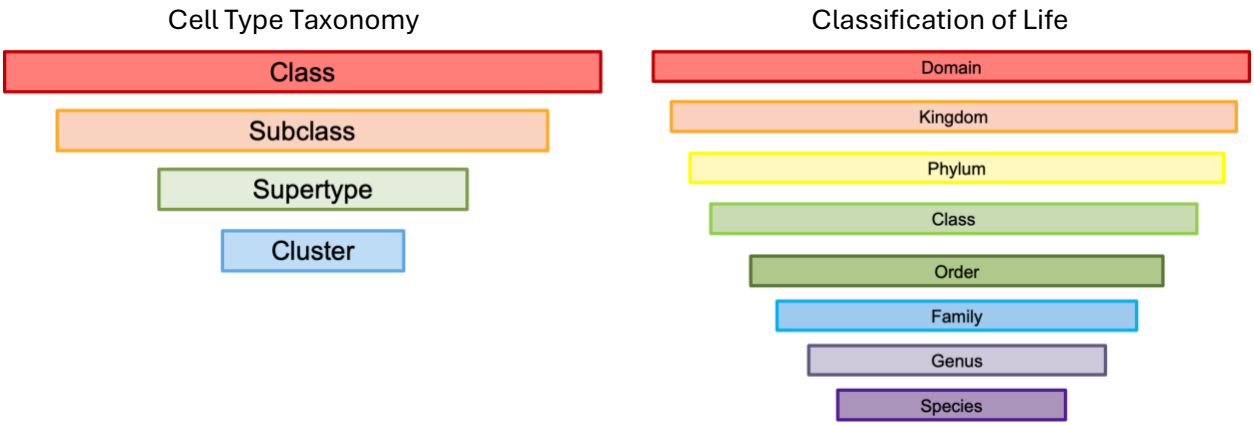
Here are some additional resources to learn more about spatial transcriptomics:

- "The" Spatial Transcriptomics Paper: Kok Hao Chen et al. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348,aaa6090(2015). DOI:10.1126/science.aaa6090
- Vizgen: https://vizgen.com/resources/how-merfish-technology-works/

- Nanostring: https://nanostring.com/research-focus/spatial-transcriptomics/
- 10x Genomics: https://www.10xgenomics.com/spatial-transcriptomics

WHAT IS A TAXONOMY?

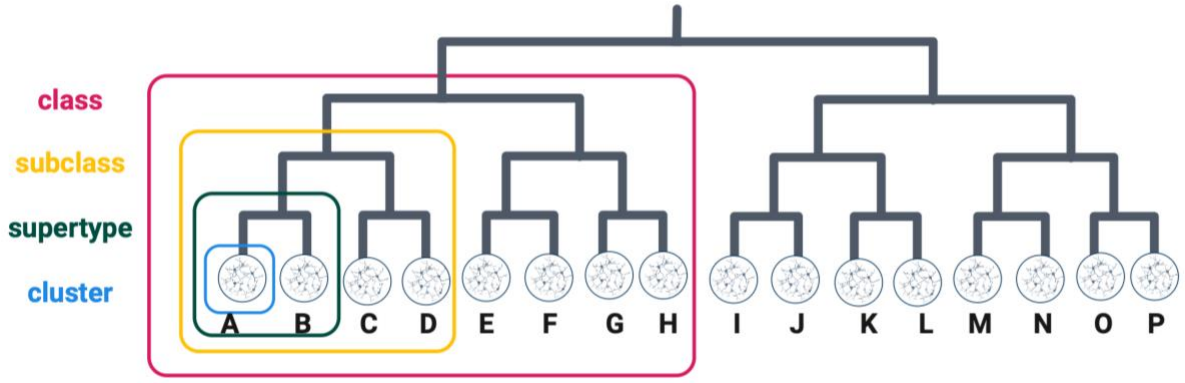
Taxonomy is broadly defined as the classification, organization, and naming of living things. A popular taxonomy taught in general biology classrooms is the "classification of life" (bottom image on right). The 'Classification of Life' is a taxonomy because there are different levels of classifications, where each level is a subset of the level above it. For cell biology, we can use the transcriptomic profiles of individual cells to create a taxonomy. A *cell type taxonomy* (bottom image on left) is a specific analysis of organizing cells into groups (or types) applied to a specific set of data and saved in a standard format.



The data showed in ABC Atlas was used to build the mouse brain taxonomy ("Whole Mouse Brain Atlas") and the first version of the human brain taxonomy ("Human Whole Brain Atlas"). To create a taxonomy that is robust and captures variations, it is important to have a high sample size. The "Whole Mouse Brain" Taxonomy is made with 317 mice from the Allen Institute, but done as a part of a larger collaboration with other leading neuroscience institutions. The "Human Whole Brain Atlas" was with 3 full male brains, hence why it is still considered the first version.

Creating cell-type transcriptomic taxonomy is a complicated process that is explained in-depth in this webinar called *"What is a taxonomy?"*.

CLASS, SUBCLASS, SUPERTYPE, CLUSTER



Cluster A and Cluster B are both in the same supertype, as shown by the green box. Cluster A and B are in the same subclass as Clusters C & D, as shown by the orange box. Clusters A, B, C, D, E, F, G, H are all in the same class, as shown by the red box.

Class: The top categorical level is *class* and is the broadest classification. Definitions like "GABAergic", "Dopaminergic", "Glutamatergic" would be equivalent to this level of classification.

Subclass: The next categorical level is *subclass*. These are *supertypes* grouped by mathematical formulas that are driven by known marker genes and a cell's anatomical location. Definitions like "somatostatin", "parvalbumin" would be equivalent to this level of classification.

Supertype: The next categorical level is *supertype*. These are *clusters* grouped by mathematical formulas that are driven by the transcriptomic data. This is usually the finest resolution scientist would refer too, as cluster is too specific.

Cluster: The lowest categorical level with the highest resolution is *cluster*. These are *cells* grouped by mathematical formulas that are driven by the transcriptomic data. Clusters are groups of cells, while supertypes are groups of groups of cells.

DATASETS IN ABC ATLAS

The ABC Atlas hosts multiple datasets. As of April 2025, there are 8 types of datasets in the ABC Atlas. Each dataset is a different project that has different data. Here is a summary of the datasets in ABC Atlas. Each dataset is a different experiment, with some of the data done by the Allen Institute's collaborators.

Dataset Name in ABC Atlas	**Species**	**Technique**	**Number of Genes**	**Number of Samples or Donors**	**Plane of brain section**	**Primary Publication**
10x scRNAseq whole brain	Mouse	Single-cell RNAseq: 10x Chromium	31,113	317 mice	n/a	Yao et al., 2023
MERFISH-C57BL6J-638850 Reconstructed Coordinates OR MERFISH-C57BL6J-638850	Mouse	Spatial Transcrptomics: MERSCOPE	500	1 mouse	Coronal	Yao et al., 2023
MERFISH-C57BL6J-638850 with Imputed Genes + Reconstructed Coordinates OR MERFISH-C57BL6J-638850 with Imputed Genes	Mouse	Spatial Transcrptomics: MERSCOPE	8,640* **imputed genes*	1 mouse	Coronal	Yao et al., 2023
Zhuang-ABCA-1/2/3/4	Mouse	Spatial Transcriptomics: MERFISH	1,122	4 mice	Coronal & Sagittal	Zhang et al., 2023
10x-scRNAseq-aged-adult	Mouse (young and old)	Single-cell RNAseq: 10x Chromium	30,824	44 mice	n/a	Jin et al., 2025
Neurons AND Non-Neuronal Cells	Human	Single-nucleus RNAseq: 10x Chromium	51,820	3 whole brains, 1 primary motor cortex section	N/A	Siletti et al., 2023

SEA-AD snRNAseq - MTG and DLPFC	Human (Alzheimer's Disease)	Single-nucleus RNAseq: 10x Chromium	36,266	84, only medial temporal gyrus and dorsal lateral prefrontal cortex	N/A	Gabitto & Travaglini et al., 2024
SEA-AD MERFISH - MTG	Human (Alzheimer's Disease)	Spatial transcriptomics: MERSCOPE	140	24, only medial temporal gyrus	Coronal	Gabitto & Travaglini et al., 2024

This guide is focused on the "Whole Mouse Brain (Yao: et al. 2023)" projects of single-cell RNAseq and Spatial Transcriptomics.

Mouse, Single-Cell Transcriptomics Dataset

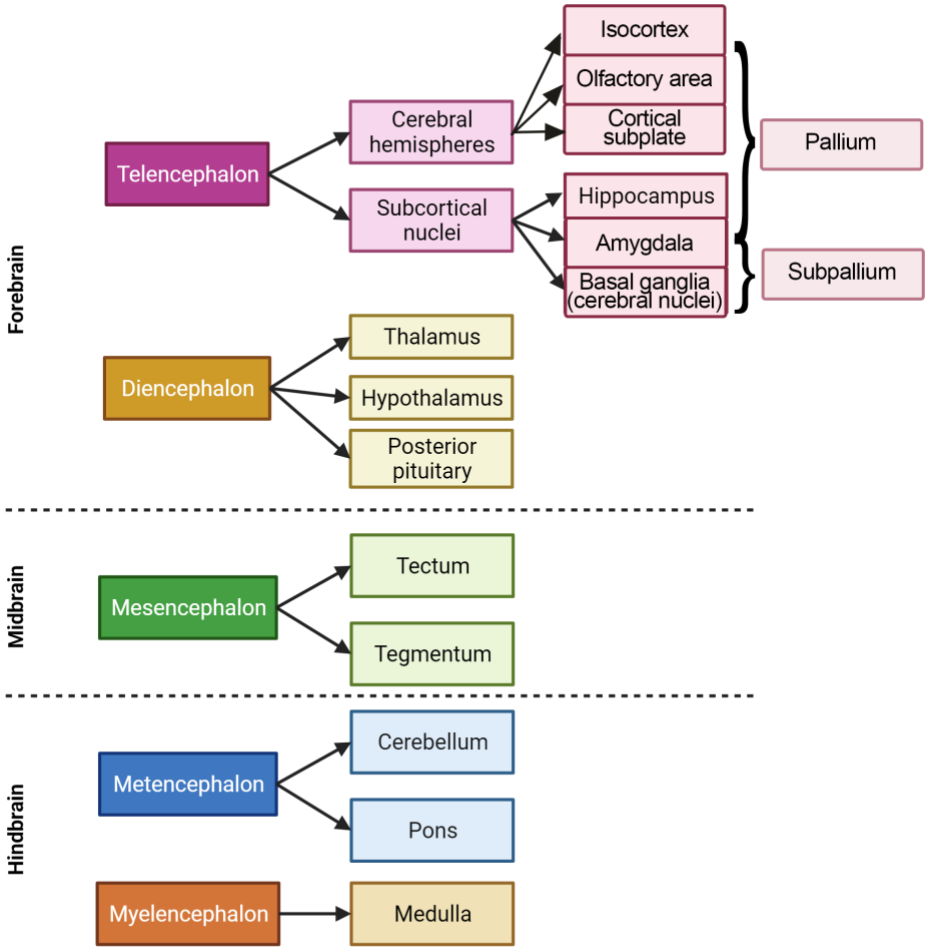
The mouse single-cell transcriptomics dataset is from the paper, "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain" (Yao et al., 2023). This paper contains information about both the single-cell RNAseq and the spatial RNAseq datasets. Although these were separate experiments, as different mice and sequencing techniques were used to generate the data, the spatial transcriptomics taxonomy is built from the single-cell RNAseq data. This paper was done in collaboration with the BICAN consortium, meaning the data was generated as part of a larger global effort by scientists to map the parts list for the human and mouse brains.

For the single-cell transcriptomics dataset, 317 mice were used to measure 31,113 genes. The mice were all adult, with a total of 164 male mice and 153 female mice used. The sequencing technological platform was 10x Genomics Chromium, with around 7.0 million single-cell transcriptomes measured, representing about 5% of the cells in the mouse brain. There is one "parent" single-cell transcriptomics dataset called "10x scRNAseq whole brain" that contains the transcriptomic data for the entire brain region and all the cells sequenced. From this dataset, the whole mouse brain taxonomy was built. As of 2025, the whole, healthy adult mouse brain taxonomy is considered a near-completed resource due to its extensive sample size and rigorous quality control.

Subsets of the Single-Cell Transcriptomics Dataset

7 sub-datasets focus on particular brain regions and neurotransmitters. It is important to understand that the 7 subsets are the same cells from the "parent" dataset, with the cells being filtered to a particular region & transmitter, which causes the visualization to change, but it is still the same data. For reference, these subsets are called "neighborhoods" in the Yao, et al. 2023 paper.

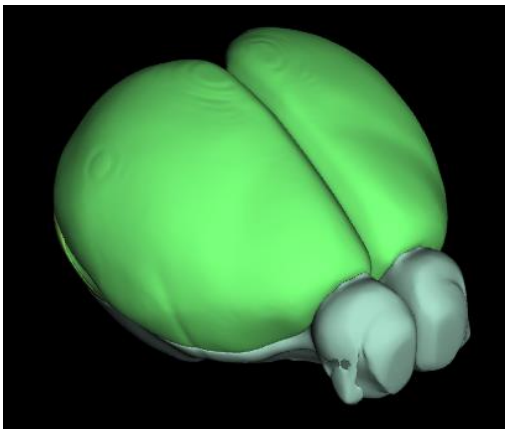
These 7 subsets' naming of brain regions comes from the anatomical architecture that is defined by cross-species evolution and the brain's development. Therefore, it is easier to make future cross-species comparisons.



Adapted from "Introduction to Behavioral Neuroscience", Chapter 1.4 Figure 1.28

Below is a summary of the 7 subsets, along with the full name, what the dataset contains, and an image of the subset's mouse brain region as generated using the Allen Brain Atlas 3D Viewer. There are also links to the Allen Brain Atlas 3D Viewer to see the images.

1. Pallium-Glut: Pallium-Glutamatergic. Contains all pallium structures.



Allen Brain Atlas 3D Viewer link

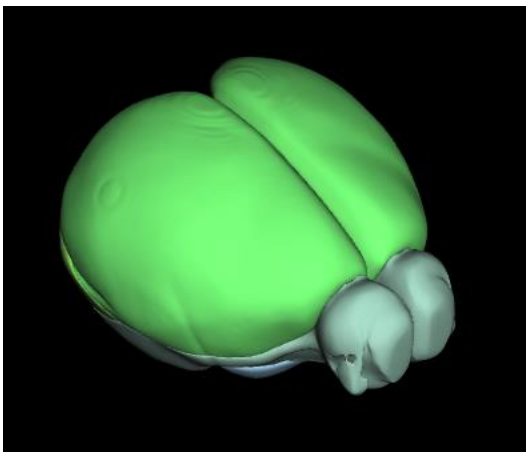
Isocortex:

Olfactory Area:

Hippocampal formation:

Cortical subplate:

2. Subpallium-GABA: Subpallium-GABAergic. Contains all GABAergic neurons found in pallium structures and those in the subpallial cerebral nuclei.



Allen Brain Atlas 3D Viewer link

Isocortex:

Olfactory Area:

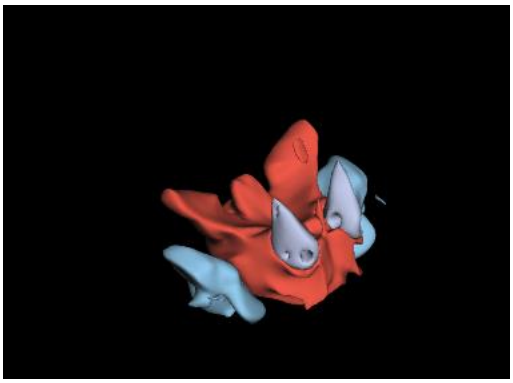
Hippocampal formation:

Cortical subplate:

Amygdala:

Striatum:

3. HY-EA-Glut-GABA: Hypothalamus-Extended Amygdala-Glutamatergic-GABAergic. Contains neurons found in hypothalamus (HY) and extended amygdala (EA).

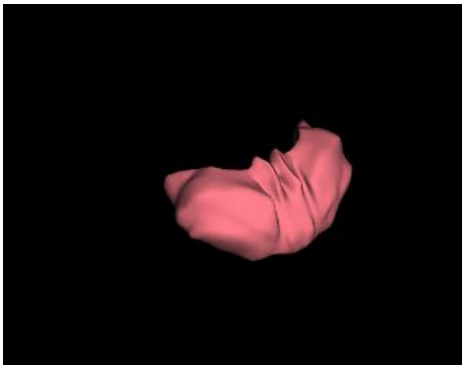


Allen Brain Atlas 3D Viewer link

Hypothalamus:

Extended amygdala:

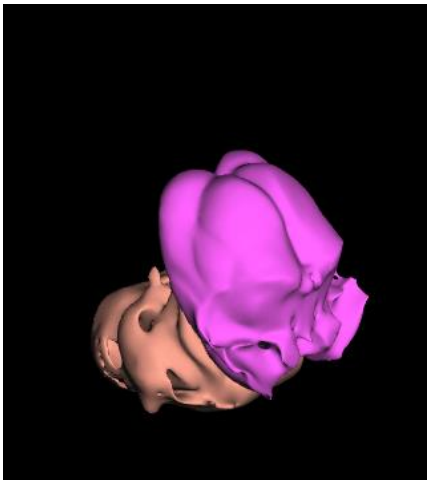
4. TH-EPI-Glut: Thalamus-Epithalamus-Glutamatergic. Contains all glutamatergic neuronal subclasses located in the thalamus (TH).



Allen Brain Atlas 3D Viewer link

Thalamus:

5. MB-HB-Glut-Sero-Dopa: Midbrain-Hindbrain-Glutmate-Serotonergic-Dopaminergic. Contains all glutamatergic, serotonergic and dopaminergic neuronal types in midbrain (MB) and hindbrain (HB).

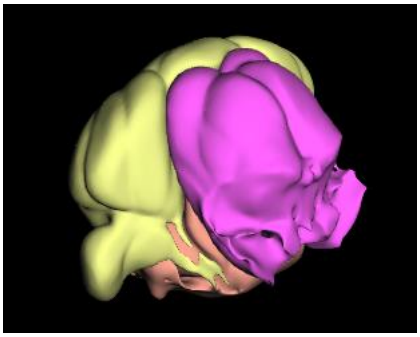


Allen Brain Atlas 3D Viewer link

Midbrain:

Hindbrain:

6. MB–HB–CB-GABA: Midbrain-Hindbrain-Cerebellum-GABAergic. Contains all GABAergic subclasses located in midbrain, hindbrain and cerebellum.



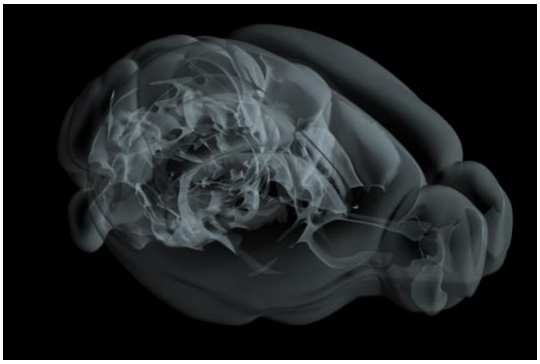
Allen Brain Atlas 3D Viewer

Midbrain:

Hindbrain:

Cerebellum:

7. NN–IMN–GC: Non Neuronal-Immature Neurons-Granule Cell. Contains a mixed collection of highly distinct non-neuronal cell types, immature neuronal types, and granule cell types throughout the entire brain.



Allen Brain Atlas 3D Viewer

Mouse, Spatial Transcriptomic Dataset

The mouse spatial transcriptomic dataset is from the paper "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain" (Yao et al., 2023). This paper contains information about both the single-cell RNAseq and the spatial RNAseq datasets. Although these were separate experiments, as different mice and sequencing techniques were used to generate the data, the spatial transcriptomics taxonomy is built from the single-cell RNAseq data. This paper was done in collaboration with BICAN consortium, meaning the data was generated as part of a larger global effort by scientists to map the part list for the human and mouse brain.

For the spatial transcriptomics dataset, 1 male mouse was used to measure 500 genes spatially to get 59 coronal brain sections with 4.3 million high-quality single cells using Vizgen MERSCOPE platform. Compared to the single-cell RNA transcriptomics dataset sample size (317 mice, 31113 genes), the spatial transcriptomics sample size is small. However, we must consider how much more time-consuming and expensive spatial transcriptomics is. For context, in 2025, to image a single adult coronal brain slice with a 500-gene panel limit, it would cost around $6,000. To run a sample of single-cell RNA seq that can measure around 5,000 cells with 5,000 gene panel, it would cost $3,000. This is a rough estimate, with costs varying by sample size, gene count, and technology platform. Additional expenses, such as labor and pre-imaging/sequencing steps (e.g., animal care, tissue processing), are excluded. In short, spatial transcriptomics is a laborious and expensive process but one that yields extremely rich data.

Mouse, Spatial Transcriptomic Dataset with Anatomical Overlay

Scientists asked the Allen Institute to have anatomical boundaries on top of the spatial transcriptomic dataset within the ABC Atlas. This request is understandable but was technologically more complicated than it first seemed to ensure the anatomical boundaries were accurate. In summary, anatomical boundaries were added, and the spatial transcriptomic data itself did not change (i.e., no new cells were

measured); however, about 300,000 cells had to be removed due to uncertainty in precise anatomical locations.

To achieve accurate anatomical overlay, the 2D slices were transformed to 3D space by referring to multiple reference data (e.g., histological stains, immunohistochemistry, transgene expression, connectivity patterns, endogenous gene expression) and labeling every voxel (pixel for a 3D image). Then the 3D brain was registered to our **Common Coordinate Framework (CCF)** to account for brain size, sectioning angle, and location of each major anatomical division.

What is CCF?

Despite research animals having the same genetic strain, there can be a slight variance in brain anatomy. To address this, the Allen Institute released the Common Coordinate Framework (CCF). The CCF is a 3D brain template with an average of 1,675 young adult male mouse brains at 10um voxel resolution. By registering brain images to the CCF, a voxel will have anatomical descriptions, allow quantitative analysis, and visualization of the data. The Allen Institute uses the CCF to register all our mouse brain data.

Mouse, Spatial Transcriptomic Dataset with Imputed Genes

The original mouse spatial transcriptomic dataset was made by running a 500-gene panel. Since the mouse spatial transcriptomic dataset was built from the same taxonomy created by the larger single-cell RNAseq dataset, scientists could computationally impute 8,460 genes into spatial space. The math to do this is beyond the scope of this guide, but we can think about it as 8,640 genes were modeled into the spatial space without the need to run additional experiments. Below is image from Heumos (2023) that overviews have imputed genes are mapped from spatial data to reference scRNA-seq data.

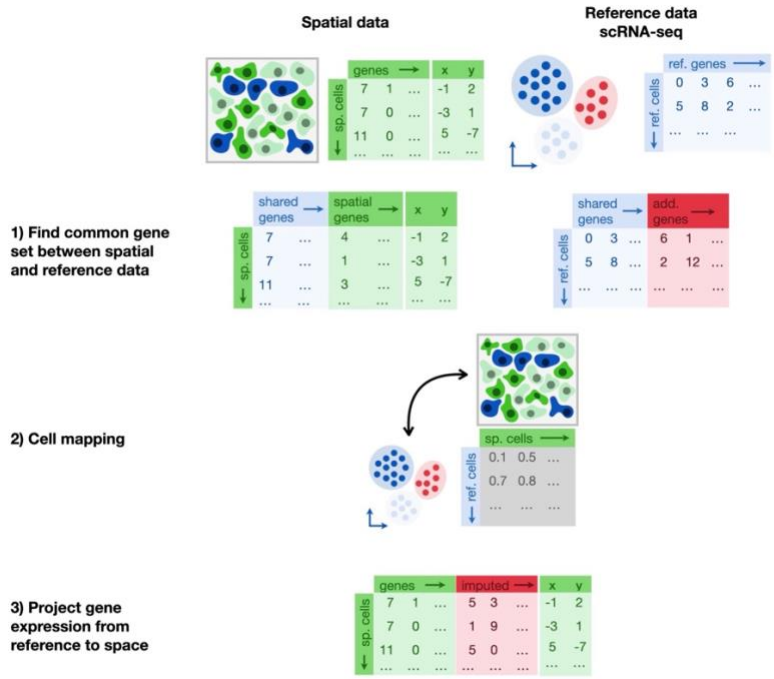


Image from Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet (2023).

EXAMPLES OF SCIENTISTS USING THE ABC ATLAS IN THEIR RESEARCH

- Micoli (2023) used ABC Atlas spatial transcriptomic data to look at their cells of interest, long-range projection type of somatostatin-expressing inhibitory neurons, to see the spatial distribution (Figure 1B)

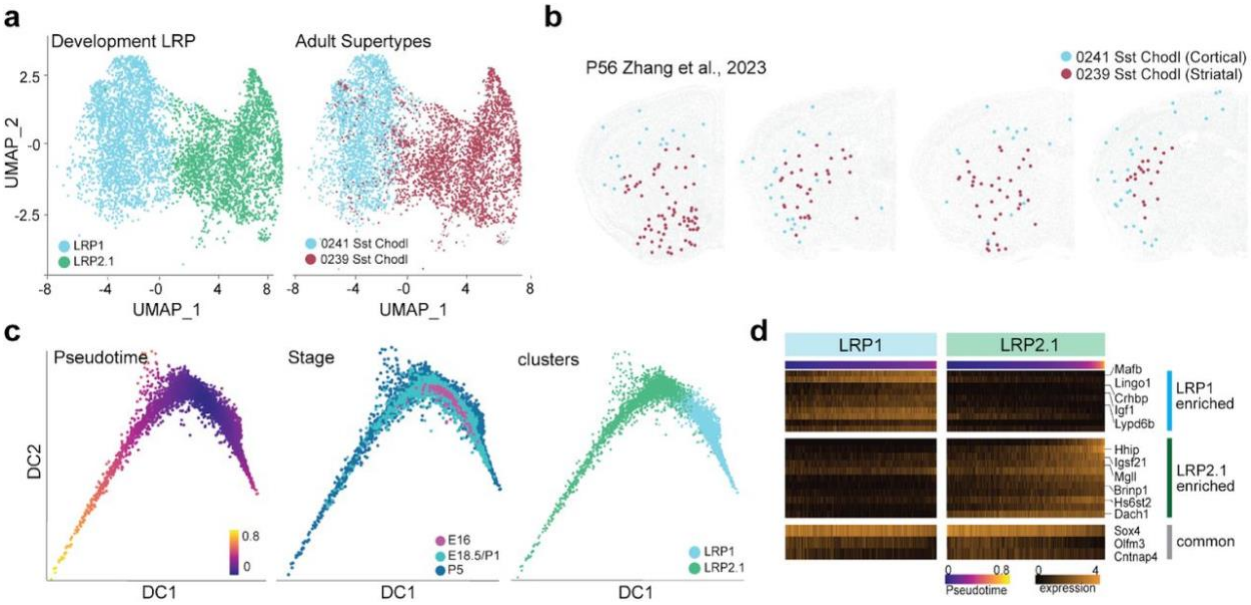


Figure 1 from https://doi.org/10.1101/2025.02.19.636192

- Forzisi-Kathera-Ibarra (2024) used ABC Atlas used the single-cell RNAseq data to visualize their cells of interest, hypothalamic neurons, to see the difference classes and subclasses (Figure S6A-B) and their genes of interest, KCNB1, POMC, LepR, and co-expression KCNB+LepR+POMC, expression levels in their cells of interest (Figure S6C)

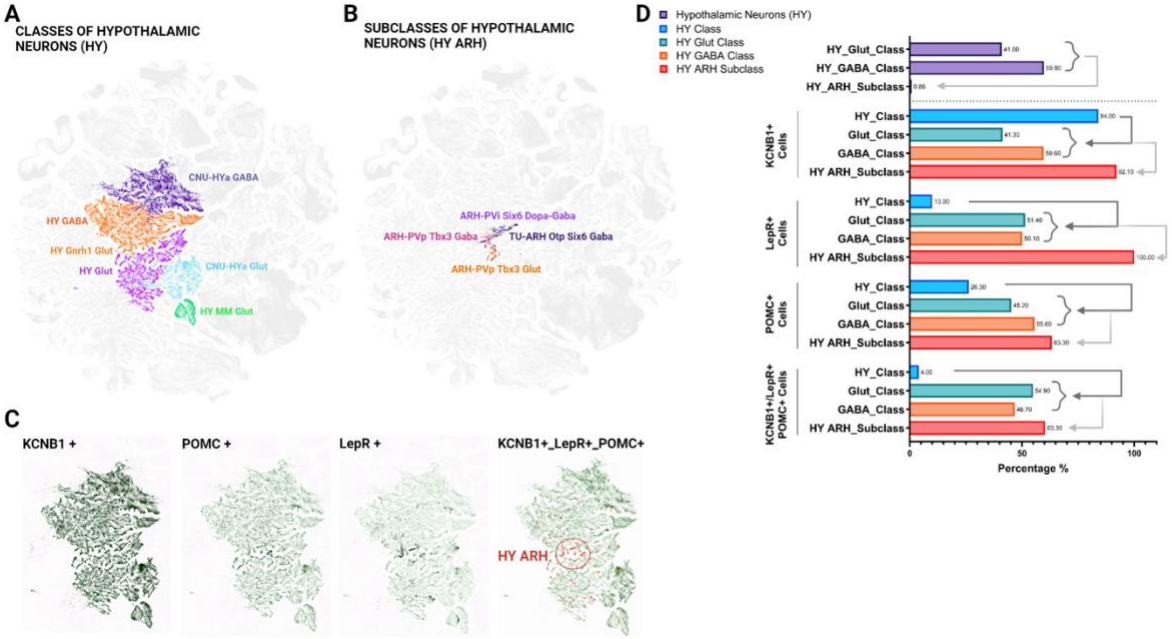


Figure S6 from https://doi.org/10.1096/fj.202401931R

- Ottenheimer (2024) used ABC Atlas to examine their cells of interest, ventral pallidum neuronal subclass, spatial location to show lack of regional specificity.

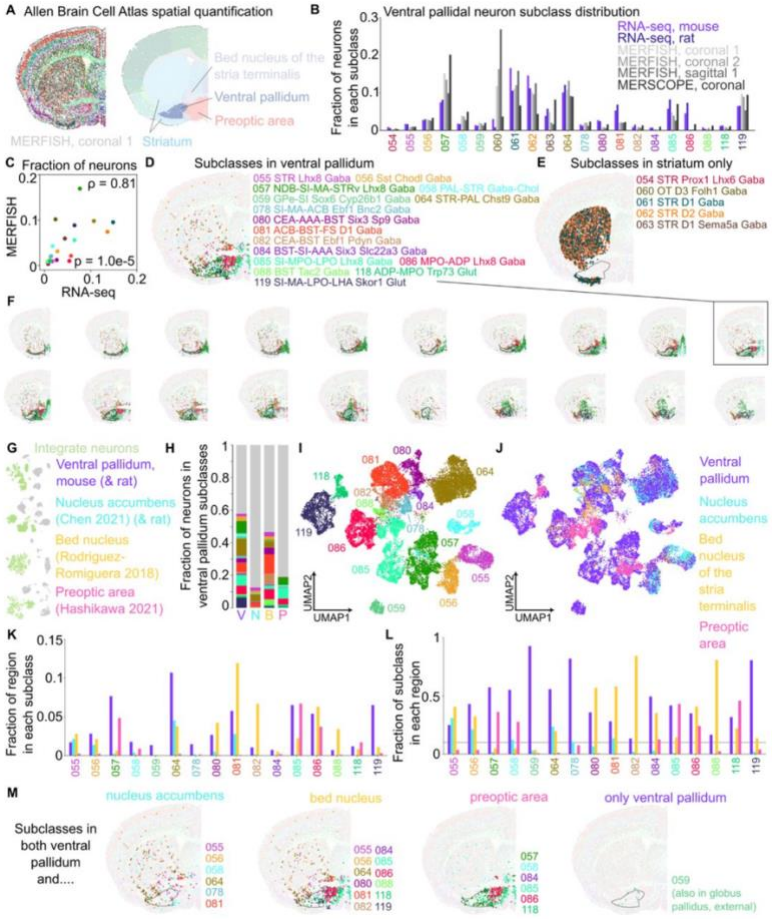


Figure 2 from https://doi.org/10.1101/2024.03.18.585611