Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD): Single Cell Profiling

This document describes the open-access single nucleus -omics data available on Registry of Open Data on AWS (link to landing page; direct link to -omics data). For more information about the SEA-AD cohort of donors, please see the SEA-AD Data and Documentation Page. *Additional spatial transcriptomic and neuropathology data is also available at this location, but is not discussed in this document.*

This bucket contains the processed single-nucleus multiomic datasets generated by SEA-AD. The top-level folders represent either brain regions that data have been generated from or data releases that contain files related to each release. Since the original MTG release, the bucket has been expanded to include the head of the caudate nucleus and a multi-region dataset spanning ten neo- and allocortical brain regions.

Brain region folders

The brain region folders in the bucket are:

- HIP/ – Hippocampus
- MEC/ – Medial entorhinal cortex
- LEC/ – Lateral entorhinal cortex
- ITG/ – Inferior temporal gyrus
- MTG/ – Middle temporal gyrus
- STG/ – Superior temporal gyrus
- FI/ – Frontal insula
- PFC/ – Prefrontal cortex (A9; previously released as DLPFC/)
- AnG/ – Angular gyrus
- V1C/ – Primary visual cortex
- Caudate_Nucleus/ – Head of the caudate

All 84 donors from the previous MTG release had their MTG, DFC, and MEC profiled, while the remaining regions were profiled in a subset of 43 donors without severe comorbidities (the "AD Spectrum" donors). Other folders redirect to the above folders for back-compatibility.

Data releases

There have been five data releases so far. Note that the original MTG release and the AAIC 2025 microglia pre-release have been deprecated and superseded by the Multiregion release; their files have been moved into previous objects subfolders.

Release	Date	Description
MTG (deprecated; superseded by Multiregion)	2022-08-18 (updated 2023-05-05 and 2024-02-13)	First single-nucleus RNA-seq, ATAC-seq, and Multiome dataset release from the middle temporal gyrus (MTG) in 84 donors (Gabitto, Travaglini et al., 2024). Now found in previous_objects within the MTG/ and PFC/ folders.
DREAM	2025-07-15	Streamlined version of the single-nucleus RNA-seq MTG and DFC datasets for the SEA-AD DREAM Challenge. Located in DREAM/.

Microglia and Immune (AAIC 2025 pre-release; deprecated)	2025-07-24	Limited pre-release of microglia from the multiregion dataset for AAIC 2025. Superseded by the Multiregion release; now found in previous_objects within Multiregion_2026/.
Caudate	2025-12-29	Single-nucleus RNA-seq, ATAC-seq, and Multiome dataset release from the head of the caudate nucleus (Kana, Postupna, Close et al., 2026). Located in Caudate_Nucleus/.
Multiregion	2026-06-22	Single-nucleus RNA-seq, ATAC-seq, and Multiome dataset release from ten neo- and allocortical regions spanning the cortical arc of AD neuropathological staging. Supersedes the original MTG release and includes the previously released MTG/ and PFC/ datasets. Regional data are in HIP/, MEC/, LEC/, ITG/, MTG/, STG/, FI/, PFC/, AnG/, and V1C/; subclass-level files are in Multiregion_2026/.

Files available

The table below summarizes the file types available across releases. Specific details about each file, along with definitions for each metadata variable, are detailed in the SEA-AD documentation. .

File	Type	Description
Regional: AnnData object – all nuclei (per region)	h5ad	AnnData file with raw UMI counts (in AnnData.layers) and log-normalized counts (in AnnData.X), plus cell metadata, for all nuclei identified by cellranger in a given brain region (“all-nuclei” objects). Provided for each of the ten profiled regions.
Regional: AnnData object – final nuclei (per region)	h5ad	AnnData file for nuclei passing quality control (“final-nuclei” objects), additionally including the scVI latent space (AnnData.obsm[“X_scVI”]) and UMAP coordinates (AnnData.obsm[“X_umap”]). Used as the basis for the SEA-AD web products.
Multiregion: AnnData object (per subclass)	h5ad	AnnData files containing all nuclei across regions for a given subclass, with the scVI latent space and UMAP computed within that subclass. Located in the Multiregion_2026/ folder.
Multiregion: cell metadata and taxonomy CSVs	csv	CSVs containing the full cell-level metadata and taxonomic annotation, along with the cell type taxonomy order and recommended colors.
Caudate Nucleus: AnnData objects	h5ad	All-nuclei and final-nuclei AnnData files for the head of the caudate nucleus, in the same format as the regional files. Located in the Caudate_Nucleus/ folder.
Reference MTG: Gene expression matrix	csv	Gene expression matrix of counts (UMIs) provided as a csv (rows = genes; columns = cells). These data, and the associated metadata (including cell type assignments) available at brain-map.org, are used as baseline for many of the SEA-AD analyses and web products.

Reference MTG: Seurat object (all data and metadata)	RDS	Seurat (v4.0.4) object with the same UMI counts and cell metadata as in the above file.
Reference MTG: AnnData object (all data and metadata)	h5ad	AnnData hdf5 (version 0.7.8) file with the same UMI counts and cell metadata as in the above files.
Reference MTG: AnnData object (QC'ed data, metadata, and more)	h5ad	AnnData hdf5 (version 0.7.8) file with counts, metadata, UMAP coordinates, and other information for SEA-AD reference cells passing QC.
ATAC-seq: Genome browser tracks	.bw	A series of bigwig files of the format [ADNC]_[cell type].bw to include as tracks for a SEA-AD instance of the UCSC genome browser (e.g., ADNC2_L2-3IT.bigwig).
ATAC-seq: Genome browser support files	.txt	Manifest files indicating how to display tracks; files with peak locations; etc.

AnnData structure

The current data are released as AnnData files containing gene expression information for 36,601 features/genes from the 10x Genomics 2020-A human reference. Alignments for singleome RNA-seq (10x v3.1 3') and Multiome (10x) were performed identically with cellranger (v6.1.2) and cellranger-arc (v2.0.0), respectively. "final-nuclei" objects have had low-quality nuclei removed, while "all-nuclei" objects contain all nuclei identified by cellranger. Within a region, AnnData files follow this structure:

- **AnnData.X** – log-normalized unique molecular identifier (UMI) counts per 10,000.
- **AnnData.layers** – raw counts of unique molecular identifiers (UMIs).
- **AnnData.obs** – per-cell donor-, library-, and cell-level metadata. Notable fields include library_prep, Donor ID, Method ("3' 10x v3.1" or "3' 10x Multiome"), Sex, Age at Death, race and ethnicity fields, Years of education, PMI, APOE Genotype, Thal phase, Braak stage, CERAD score, Overall AD neuropathological Change, LATE stage, Highest Lewy Body Disease, Cognitive Status, and the cell-type annotations Class, Subclass, and Supertype.
- **AnnData.obsm** – the scVI latent representation ("X_scVI") and the UMAP coordinates computed from its nearest-neighbor graph ("X_umap"). For regional objects these are available in the "final-nuclei" objects; for the multiregion subclass-level objects they are computed within each subclass.
- **AnnData.var** – gene names and ensembl IDs for each feature.
- **AnnData.uns** – official taxonomy colors in "Subclass_colors" and "Supertype_colors".

Methods Description

Overview

The Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) is a consortium focused on identifying and characterizing early changes in the brain in Alzheimer's disease and normal aging, that is funded by the National Institute on Aging (NIA U19 AG060909-01A1). SEA-AD initially published data in human middle temporal gyrus (MTG) from an aged cohort of 84 donors who span the full spectrum of disease severity, and from five younger neurotypical

donors. The atlas has since been extended to the head of the caudate nucleus and to a multi-region dataset of ten neo- and allocortical regions. The data available in this bucket are divided by brain region and release as indicated above. Information and methods related to data generation are detailed in the SEA-AD documentation.

1. Single nucleus RNA-seq data collected from younger neurotypical donors

Files from this group are prefaced "Reference MTG". These files include gene expression matrices and associated metadata for 166,868 nuclei derived from five post-mortem young adult human brain specimens and collected using 10x Genomics Chromium Single Cell 3'™ Reagent Kit v3 (link) or v3.1 (link). These data are included in the following formats:

1. A table of comma-separated values, which is widely accessible but quite large.
2. A Seurat (v4.0.4) object, which is a standard format for data processing in R developed by the Satija lab. Several tutorials for use of snRNA-seq data in this format are available on their website.
3. An AnnData hdf5 (version 0.7.8) file, which is a standard format for single nucleus RNA-seq data in python, with associated tutorials. Two key tools used by SEA-AD and others based on this data format are Scanpy and scvi-tools, which include several tutorials. AnnData objects for both the final filtered data set (for most use cases including all SEA-AD web applications) and for the complete data set (intended for reproducibility and to aid in QC of novel data sets) are provided for the reference RNA-seq data and for the SEA-AD cohort (below).

Gene expression matrices provide the number of unique molecular identifiers (UMIs or counts) for each gene in each nucleus, while metadata tables include cell type assignments and donor demographic information. In addition, the AnnData files include coordinates for the scVI latent space and UMAP embedding, nearest neighbor graph, and cluster specific colors, all of which can be used to reproduce the plots and analyses presented on the SEA-AD website. Specific details about each of these files along with definitions for each metadata variable are detailed in the SEA-AD documentation.

As part of the SEA-AD website, we provide tools for visualizing and exploring this reference, transferring labels from this reference to novel datasets, and data download in additional formats. All these tools rely on the same data and metadata provided here.

Creation of cell "supertypes"

We defined "supertypes" as a set of fine-grained cell type annotations for single nucleus expression data that could be reliably predicted on held-out reference data (where "ground truth" labels were assigned as described above) using state-of-the-art machine learning approaches (publication). From 5 neurotypical donors in a related study with roughly 140K nuclei captured with 10x snRNAseq we systematically held out 2 donors and used scANVI to iteratively and probabilistically predict their class (3 labels), subclass (24 labels), and then cluster (151 labels). When predicting each nucleus's class, we selected the top 2,000 highly variable genes along with the top 500 differentially expressed genes unique to each class (calculated from the reference cells which had their labels retained using a Wilcoxon rank sum test) to use as features in training the model and specified the donor name and number of genes detected as categorical and continuous covariates, respectively. Nuclei were then separated by their predicted class and features were re-selected with the same criteria to predict subclasses and again in predicting clusters. A differential expression test was run on clusters with an F1 score below 0.7, and those without 3 positive markers when compared against nuclei from their constituent subclass (corrected p-value <0.05, fraction in group expression >0.7, fraction out of group expression <0.3) were pruned from the taxonomy. Of the

26 clusters flagged, 24 fell below these cutoffs and were pruned from the final supertype taxonomy. The remaining 2 (L2/3 IT_2 and Oligo_3) were retained and recovered after supertype prediction (see below).

2. Single nucleus RNA-seq data collected from 84 aged donors (SEA-AD Cohort)

Files from this group are prefaced "SEA-AD MTG". These files include gene expression matrices and associated metadata for 1,240,908 nuclei derived from 84 aged adult human brain specimens that span the histopathological and cognitive spectrum of Alzheimer's disease, including unaffected and cognitively normal individuals. Nuclei from all donors are collected using the 10x Genomics Chromium Single Cell kits (linked above) that only assess gene expression, and additional nuclei from a subset of 28 donors are collected using Multiome, which assesses both gene expression and chromatin accessibility for the same cell.

Note: these MTG files were part of the original 2022 release and have been deprecated and superseded by the Multiregion release (2026-06-22). The re-released MTG data now reside in the MTG/ folder alongside the other regions, with the original objects retained under previous_objects/. See the changelogs in MTG/ and PFC/ for a list of changes, which include expansion of the taxonomy to additional regional cell types, additional quality-control flagging, and an scVI latent space and UMAP now computed across the entire multi-region dataset.

Due to its size, this data set is only provided as an AnnData hdf5 file, with versions for the complete ("all-nuclei") and filtered ("final-nuclei") data sets, as described above. AnnData files include a gene expression matrix, cell metadata, coordinates for the scVI latent space and UMAP embedding, nearest neighbor graph, and cluster specific colors, all of which can be used to reproduce the plots and analyses presented on the SEA-AD website. In addition to the metadata provided for nuclei in the reference data set, cell metadata from this aged cohort include (1) additional donor metadata related to disease pathology and associated demographic and cognitive metrics and (2) RNA metrics potentially associated with disease (e.g., fraction of mitochondrial reads). Specific details about each of these files along with definitions for each metadata variable are detailed in the SEA-AD documentation.

As part of the SEA-AD website, we provide two tools for visualizing and exploring the SEA-AD MTG data provided here. The Comparative Viewer is a tool aimed at allowing users to explore gene expression in aged donors in the context of donor metadata and cell types. Cellxgene is an interactive browser that enables scientists to annotate, publish, find, download, explore and analyze single cell datasets (reference). Both tools directly apply the information provided in the above hdf5 files, and present visualizations based on the same UMAP coordinates.

Quality control and cell type assignment of nuclei from aged donors using snRNA-seq

This section applies to nuclei collected using singleome 10xV3 snRNA-seq (all donors) and 10x Multiome (28 donors), which are then mapped to the above reference transcriptome using only gene expression. The results of these mappings are extensively used in the SEA-AD web product as described below. A separate mapping of these data to cell types using a combination of gene expression and chromatin accessibility is described below, which is used for assigning cell types to ATAC-seq data and for assessment of cell type assignment confidence.

Initial removal of low-quality nuclei

SEA-AD nuclei with fewer than 500 genes detected were removed upstream of supertype mapping (see below).

Mapping SEA-AD nuclei to reference supertypes

After defining supertypes in neurotypical donors, we iteratively and probabilistically predicted each SEA-AD nucleus's class, subclass, and supertype using scANVI, as above. Each SEA-AD nucleus's class was predicted after projecting them into a shared latent space with reference nuclei using models trained with 2000 highly variable genes and 500 differentially expressed genes per class (from reference data, where donor name and number of genes were passed as categorical and continuous covariates, respectively). Nuclei were then split by predicted class, projected into a new class-specific latent space where subclass was predicted, and again for supertype. The subclass-specific latent spaces were then used to compute two-dimensional uniform manifold approximation and projections (UMAPs) and the scANVI predictions were evaluated by known marker gene expression (using signature scores defined by differentially expressed genes in reference nuclei). In regions reference nuclei occupied there was strong agreement in signature gene expression with SEA-AD nuclei, indicating accurate prediction by scANVI, but there was more variable expression in regions with poor reference support (which also had higher uncertainty in their predictions). These areas represented either droplets with ambient RNA, multiple nuclei, dying cells, or transcriptional states missing from the reference, unique to a donor or found only in aging or disease. To triage these possibilities, we fractured the graph into tens to hundreds of clusters (called "metacells") using high resolution Leiden clustering (resolution=5, k=15) and then merged them based on differential gene expression using the defaults in the transcriptomics_clustering package. Clusters and metacells were then flagged and removed if they had poor group doublet scores, fraction of mitochondrial reads, number of genes detected, or donor entropy, eliminating common technical sources of transcriptional heterogeneity.

Expanding the reference taxonomy for non-neuronal cells

With common technical axes of variation removed, we then sought to identify nuclei that were transcriptionally distinct from the reference and add them to our supertype taxonomy. We constructed a new latent space for each subclass using scVI, where the model was aware of the supertype prediction for each nucleus, gene dispersion was allowed to vary per supertype, donor name, sex, race and 10x technology (multiome versus singlome) were passed as categorical covariates, and the number of genes detected in each nucleus and the donor age at death were passed as continuous covariates. Using the neighborhood graph from this latent space, we clustered the nuclei into tens to hundreds of groups and merged them based on differential gene expression, as above. We defined merged clusters with fewer than 10% of all reference cells or of any single supertype as having poor reference support and added them to the taxonomy (systematically named Subclass_Number-SEAAD). In cases where more than 90% of SEA-AD nuclei within these poorly support groups were predicted to be one supertype, their new label reflected that assignment (e.g. Subclass_SupertypeNumber_Number-SEAAD). These cell type assignments are used as baseline for the analyses, plots, and tools developed for the web product and in-processed scientific manuscripts.

3. Single nucleus RNA-seq data collected from the head of the caudate nucleus

Files from this group are located in the Caudate_Nucleus/ folder and described in Kana, Postupna, Close et al. (2026). These files include gene expression matrices and associated metadata for nuclei derived from the head of the caudate nucleus in a subset of the SEA-AD cohort, collected using the same 10x Genomics singleome and Multiome chemistries described above and provided as all-nuclei and final-nuclei AnnData files in the standard structure.

Because the caudate is a basal ganglia structure whose neuronal composition differs from cortex, its cell types were assigned using a combined reference that pairs the SEA-AD MTG taxonomy with the BRAIN Initiative Cell Atlas Network's (BICAN) Human and Mammalian Brain Cell Atlas (HMBA) cross-species consensus taxonomy of the basal ganglia (Johansen et al. 2026). Basal-ganglia-specific neuronal populations, principally lateral ganglionic eminence (LGE)-derived medium spiny neurons, are labeled using the HMBA taxonomy, while non-neuronal populations, which are largely shared with cortex, are labeled using the SEA-AD MTG taxonomy; a subset of inhibitory interneuron types (SST+ CHODL+ and a subset of VIP+) are shared across both. Cell type assignment otherwise follows the same iterative scANVI mapping and quality-control approach used for the MTG cohort.

4. Single nucleus RNA-seq data collected across ten neo- and allocortical regions (Multiregion)

The multiregion release encompasses ten neo- and allocortical brain regions that span the cortical arc of AD neuropathological staging. It includes the previously released middle temporal gyrus (MTG) and prefrontal cortex (PFC/A9) data along with eight additional regions: hippocampus (HIP), medial entorhinal cortex (MEC), lateral entorhinal cortex (LEC), inferior temporal gyrus (ITG), superior temporal gyrus (STG), frontal insula (FI), angular gyrus (AnG), and primary visual cortex (V1C). As noted above, all 84 donors had MTG, DFC, and MEC profiled, while the remaining regions were profiled in the subset of 43 "AD Spectrum" donors without severe comorbidities. This release supersedes the original MTG release.

Nuclei from every region were mapped to the SEA-AD MTG/PFC taxonomy to hierarchically assign a major cell "Class", a "Subclass" beneath it, and a mappable, fine-grained "Supertype" below that. The previous taxonomy was expanded to include new types that were not present in the previous MTG/PFC taxonomy (e.g. region-specific excitatory types in HIP, MEC, and V1C), while cell labels that are common across regions are retained (e.g. Sst_25 still refers to the same kind of cell type in the updated taxonomy). In total the taxonomy comprises 207 cell types organized into 28 subclasses and 3 higher-level classes.

The dataset is provided as a series of AnnData files, either as all nuclei from a given brain region (in that region's folder) or all nuclei across regions from a given subclass (in the Multiregion_2026/ folder); for the subclass-level objects, the scVI latent space and UMAP are computed within each subclass. SEA-AD additionally provides CSVs containing the full cell-level metadata and taxonomic annotation, as well as the cell type taxonomy order and recommended colors. The AnnData structure is consistent with the previous release and is described in the AnnData structure section above.

Tutorials on how to use the single nucleus transcriptomics data

Accessing SEA-AD data through the ABC Atlas (multi-region and caudate studies)

The ABC Atlas aims to empower researchers worldwide to explore and analyze multiple whole-brain datasets simultaneously. The ABC Atlas can be accessed here and is the primary method of interacting with single cell and spatial transcriptomics from the Allen Institute. For SEA-AD, snRNA-seq data from the multi-region and caudate studies are available through the Allen Brain Cell (ABC) Atlas Access for data access, visualization, and basic analysis; these data sets will shortly be added to the ABC Atlas web application as well. This entry point includes a subset of the full data compendium described above for the multi-region and caudate studies,

so for identical data points the source you start from should not matter. Any questions or issues with this tool are best directed to the Allen Institute Community Forum.

ABC Atlas Access provides separate Jupyter notebooks describing how to download the data and how to perform some basic visualization and analysis, and therefore such examples are not presented here. If you would rather use the AWS buckets described above, or if your primary interest is the MTG study, the two Use Cases below describe two entry points into common questions with the SEA-AD snRNA-seq data.

Use Case 1: What kinds of cells express a gene of interest?

1. Install scanpy (installation instructions).
2. Load scanpy and read the AnnData object:
   ```
   import scanpy as sc
   adata = sc.read_h5ad('anndata.h5ad')
   ```
3. Display the gene (APOE in this case) on the pre-computed UMAP coordinates:
   ```
   sc.pl.umap(adata, color=['APOE'])
   ```
4. Display the gene across cell types as a violin plot:
   ```
   sc.pl.violin(adata, color=['APOE'], groupby='Supertype')
   ```

Use Case 2: How does a gene change across Alzheimer's disease (AD) pathology?

1. Install scanpy (installation instructions).
2. Load scanpy and read the AnnData object:
   ```
   import scanpy as sc
   adata = sc.read_h5ad('anndata.h5ad')
   ```
3. Display the gene (APOE in this case) grouping by AD Neuropathologic Change (ADNC):
   ```
   sc.pl.violin(adata, color=['APOE'], groupby='ADNC')
   ```
4. Display the gene grouping by ADNC in a specific Supertype (Microglia-PVM_1 in this case):
   ```
   sc.pl.violin(
       adata[adata.obs['Supertype'] == 'Microglia-PVM_1'],
       color=['APOE'],
       groupby='ADNC'
   )
   ```

Chromatin accessibility in aged human donors

Donor-level data from middle temporal gyrus

Files from this group are prefaced "ATAC-seq MTG". Collectively, these files present chromatin accessibility information from the SEA-AD cohort in a format that will be displayed in the UCSC Genome Browser and that is de-identified to retain donor privacy. The data that feeds these files is single nucleus ATAC-seq data derived from the same 84 aged adult human brain specimens described above. Nuclei from all donors are collected using the 10x Genomics Chromium Single Cell ATAC-seq kits (here) that only assess chromatin accessibility, and additional nuclei from a subset of 28 donors are collected using Multiome, which assesses both gene expression and chromatin accessibility for the same cell. *In late 2026 or early 2027, ATAC-seq data will be released for all other available brain regions using this (or a similar) format.*

Displaying summary data via a genome browser

Separately for MTG and DFC, individual genome browser track files are created by combining data from cell nuclei assigned to a specific subclass for a specific set of donors specified by a donor metadata field. For example, the bigwig file "ADNC3_L2-3IT.bw" would contain chromatin accessibility information for all cells of any Layer 2/3 IT cell type from all donors with an ADNC score of 3 (which represents high pathology). Additional manifest files are provided which, when added as "custom tracks" to the UCSC Genome Browser (as described here), allow viewing of sensible subsets of these genome browser track files at the same time. For example, the file "ADNC2_oligo.manifest" would provide an access point to the UCSC genome browser to view tracks for each oligo cell type separately for donors from each ADNC score for exploration of how chromatin accessibility changes with increasing AD pathology in oligo cells. These manifest files are linked from the SEA-AD website, and we recommend accessing the UCSC genome browser from there. Track files can also be downloaded separately for programmatic analysis of ATAC-seq data, and for visualization in other genome browsers. Finally, we provide prebuilt link-outs to UCSC instances focused on MTG and DFC in the "Epigenetics: Chromatin Accessibility" section of the "Our Resources" page for interactive viewing of ATAC-seq information per subclass in the context of ADNC, as described.