

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

Single Cell Profiling

For more information about the SEA-AD cohort of donors, please see the [SEA-AD Documentation Page](#)

Files available:

File	Type	Description
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 the 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 above files.
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 above files
Reference MTG: AnnData Object (QC'ed data, metadata, and more)	h5ad	AnnData hdf5 (version 0.7.8) file counts, metadata, UMAP coordinates, and other information for SEA-AD cells passing QC.
SEA-AD MTG: AnnData Object (all data and metadata)	h5ad	Compressed AnnData hdf5 (version 0.7.8) file with UMI counts, cell metadata, UMAP coordinates, and other information for all cells from 84 aged donors in the SEA-AD. Metadata includes extended calls for low quality and doublet clusters for reproducibility and to aid in QC of novel data sets.
SEA-AD MTG: AnnData Object (QC'ed data, metadata, and more)	h5ad	Compressed AnnData hdf5 (version 0.7.8) file counts, metadata, UMAP coordinates, and other information for SEA-AD cells passing QC, matching results presented in the Comparative Viewer and CZI cellxgene. Final cell type assignments match cell types presented in the Reference MTG taxonomy with expanded non-neuronal types.
ATAC-seq MTG: Genome browser tracks for ATAC-seq data	.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 MTG: Genome browser support files	.txt	Manifest files indicating how to display tracks; files with peak locations; etc.

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). So far SEA-AD has 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 data available in this bucket are divided by brain region and as indicated below.

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 is available [on their website](#).
- (3) An AnnData hdf5 ([version 0.7.8](#)) file, which is a standard format for data 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 includes [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.

Due to its size, this data set is only provided as an AnnData hdf5 ([version 0.7.8](#)) file, with versions for the complete and filtered 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 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 supported 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.

Tutorial on how to use the 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 on the pre-computed UMAP coordinates

```
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 (APOE in this case) grouping by AD Neuropathologic Change (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'
)
```

3. Chromatin accessibility in the MTG of aged human donors

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.

Each individual genome browser track file is 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 (e.g., microglia, astrocytes, and oligodendrocytes) separately for

donors from each ADNC score for exploration of how chromatic accessibility changes with increasing AD pathology in oligo cells. These manifest files will be directly linked from [the SEA-AD website](#) by late summer 2022, 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.