Lesson 4: Analyzing Transcriptomic Data to Explore Alzhiemer's Disease Pathology

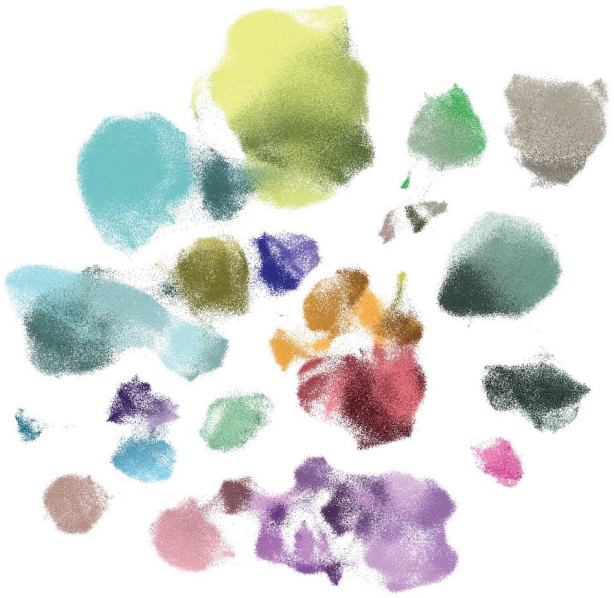


Image is from the Seattle Alzheimer's Disease Brain Cell Atlas consortium

Learning Objectives:

- Students will be able to articulate how transcriptomic data is collected and processed
- Students will be able to interpret transcriptomic data visualized within Uniform Manifold Approximation and Projections (UMAPs)
- Students will be able to navigate the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) database in order to interpret transcriptomic data
- Students will be able to filter data based on specific biomarker and/or demographic characteristics they are interested in exploring
- Students will be able to independently perform an analysis of transcriptomic data using the cellxgene interface
- Students will be able to compare gene expression between cell types using the cellxgene interface
- Students will be able to navigate the NIH gene database to explore the known functions of genes

Introduction

In this lesson, we will explore how scientists use information gathered about a cell's gene expression to explore the pathology of Alzheimer's Disease (AD). AD is a progressive neurological disorder that impacts a person's cognitive and memory abilities.

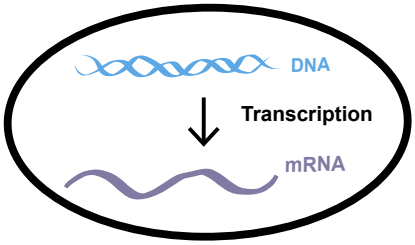
Although AD pathology can be studied in a variety of ways, scientists at the Allen Institute for Brain Science approach this research by paying extremely close attention to the cellular process of transcription analyzing which genes a cell is expressing and in what quantities.

Although transcriptomics can be used across a variety of fields, this type of data is becoming increasingly prominent within neuroscience. Neuroscientists at the Allen Institute who study AD pathology rely heavily on transcriptomic data to study brain tissue donated by individuals both with and without dementia after their deaths.

Review: Transcriptomic Data

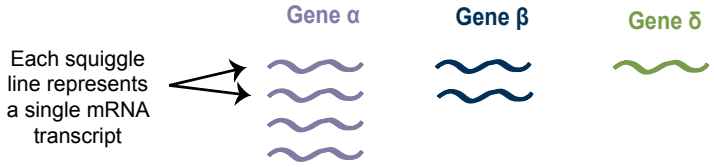
As you learned in lesson 2, transcriptomic data is a type of data that allows scientists to investigate which genes a cell is transcribing/expressing and in what quantities. If a cell, and more specifically, that cell's nucleus, contains a specific RNA transcript, this indicates that the cell is expressing the specific gene associated with that RNA. By (1) isolating nuclei, (2) sequencing the mRNA transcripts found within the nuclei, and (3) counting those transcripts, we can tell **which genes** the cell is expressing and **how much** these cells are expressing these genes. Repeating this process for thousands of cells from a sample of brain tissue allows researchers to find similarities and dissimilarities between cells on the basis of their gene expression. These patterns of similarity and dissimilarity are then used to classify certain cells as specific "types." The graphic below explains in detail how scientists gather and interpret transcriptomic data.

Carefully read through the graphic below to review how transcriptomic data is gathered and interpreted:



Isolate the nuclei from the cells in the sample of brain tissue and extract the RNA found in each nucleus.

Sequence the mRNA transcripts found in each cell's nucleus in order to determine which genes each brain cell was expressing.



Count the number of mRNA transcripts found for each gene. This allows us to quantify how much each cell was expressing each gene.

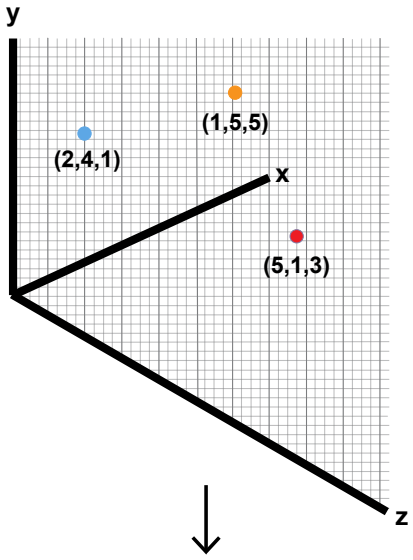
	Gene α	Gene β	Gene δ
# of mRNA transcripts found in Cell 1	4	2	1

Create a table that shows how much each cell was expressing each gene.

	Gene α	Gene β	Gene δ
# of mRNA transcripts found in Cell 1	2	4	1
# of mRNA transcripts found in Cell 2	1	5	5
# of mRNA transcripts found in Cell 3	5	1	3
repeat count for thousands of cells...	...	...	...

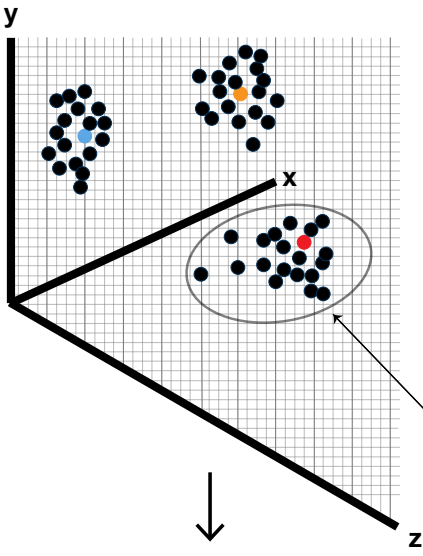
Repeat this process for THOUSANDS of cells. Remember, this means we are counting how much EACH cell was expressing EACH gene. If we wanted to create a table that listed the data in full, this data table would have thousands of rows.

x = gene α value
y = gene β value
z = gene δ value



If we wanted to create a graph that plotted the inital data for cell 1, cell 2, and cell 3 and their relative amount of expression of gene alpha, gene beta, and gene delta, we would need a 3D graph like the one on the left.

x = gene α value
y = gene β value
z = gene δ value



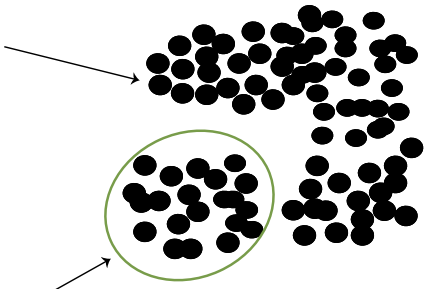
We can repeat this process for the thousands of cells that were collected from the brain tissue sample. Notice that the cells begin to cluster based on how similar their gene expression for gene alpha, gene beta, and gene delta is to one another. These clusters help us identify which cells may be more similar and/or dissimilar to one another!

when we plot the gene expression data for more cells, we notice that cell 3 (red dot) clusters next to these other cells from the sample

	Gene α	Gene β	Gene δ	repeat for thousands of genes...
# of mRNA transcripts found in Cell 1	2	4	1	...
# of mRNA transcripts found in Cell 2	1	5	5	...
# of mRNA transcripts found in Cell 3	5	1	3	...
repeat count for thousands of cells...	...	...	...	...

In addition to collecting data on gene expression for thousands of cells, scientists will add another layer of complexity by measuring the gene expression of these thousands of cells for THOUSANDS of genes. A table displaying this data would have thousands of rows and thousands of columns. Since the graph would now have much more than just 3 dimensions, we will need a special type of tool to graphically represent this data in a way that humans can visualize.

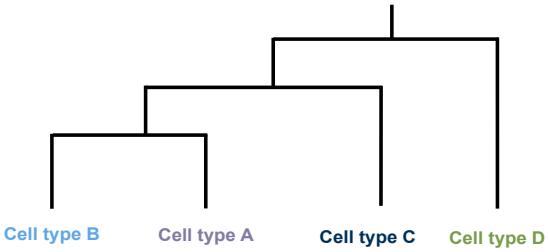
Each dot represents a single nucleus isolated from a single brain cell



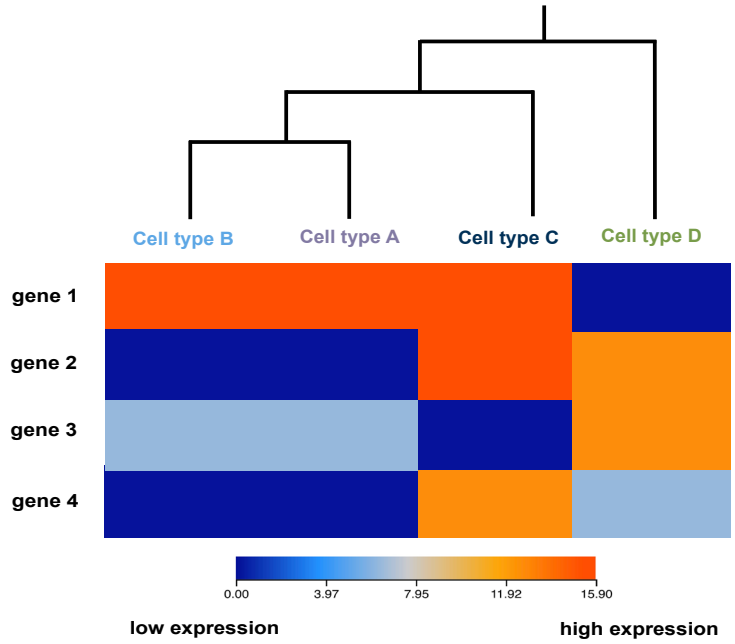
Identify clusters in the data--these clusters represent cells that are more like each other than they are like any other cells

UMAP

In order to plot this many-dimensional graph in a way humans can visualize, we use a dimensionality reduction tool, such as a UMAP, to plot it in a 2D space. Dimensionality reduction is a technique that helps represent many-dimensional data in just two or three dimensions.



Organize the clusters identified in the UMAP to construct a dendrogram that displays hierarchical relationships between the clusters based on each cell type's similarity and dissimilarity of gene expression.



Use a heatmap below the dendrogram to compare the level of gene expression between each cell type for specific genes of interest.

Exploring the different cell types of the brain helps us to understand how healthy brains function. This information is crucial to gather to understand what changes in the brain when certain neurological diseases arise. A common neurological disease that impacts memory capabilities and cognitive functioning is AD. Understanding the different cell types of the brain is crucial to AD research, as researchers can ask questions such as:

- *Do aged brains from people with dementia contain the same types of cells as brains from young adults?*
- *Do specific types of cells show changes in gene expression between a healthy brain and a brain with AD?*

Mapping the Brain:

Researchers at the Allen Institute are exploring these questions and more as a part of the Allen Institute for Brain Science's Seattle Alzheimer's Disease Brain Cell Atlas, which is abbreviated as SEA-AD. As a part of this project, scientists gather large amounts of transcriptomic data from post-mortem donated brain tissue from individuals within the Seattle area. This data is then uploaded into an online database called the SEA-AD Cell Atlas which is freely available to the public. The SEA-AD Cell Atlas consists of data from 84 older adults with and without AD. As of the writing of this lesson in 2022, the database features data from the middle temporal gyrus (MTG), and more data of the same types from other brain regions will be added in later updates. The Allen Institute for Brain Science has previously studied the MTG in depth in healthy donors, and it is affected relatively early in the progression of AD.

The SEA-AD project strives to gain a deep molecular and cellular understanding of the early pathogenesis of AD. The data collected within this study are derived from a full spectrum of aged donors, from healthy controls to those with high AD pathology and dementia. Data and specimens were obtained from the Adult Changes in Thought (ACT) Study from Kaiser Permanente Washington Health Research Institute (KPWHRI), and the University of Washington Alzheimer's Disease Research Center (ADRC). The ACT study is a study that follows initially healthy donors at age 65 for the rest of their lives. This study records a large range of medical information about each donor, each donor's demographic characteristics, and collects vital information about each donor while they are living and after they have passed.

In order to strengthen your understanding of AD, we will complete a guided tutorial of the SEA-AD database. This database is available online at: SEA-AD.org.

The SEA-AD database is vast and full of many different resources. Given the time constraints of today's lesson, we will focus on only one of the databases that contains the SEA-AD data: **CZ cellxgene**

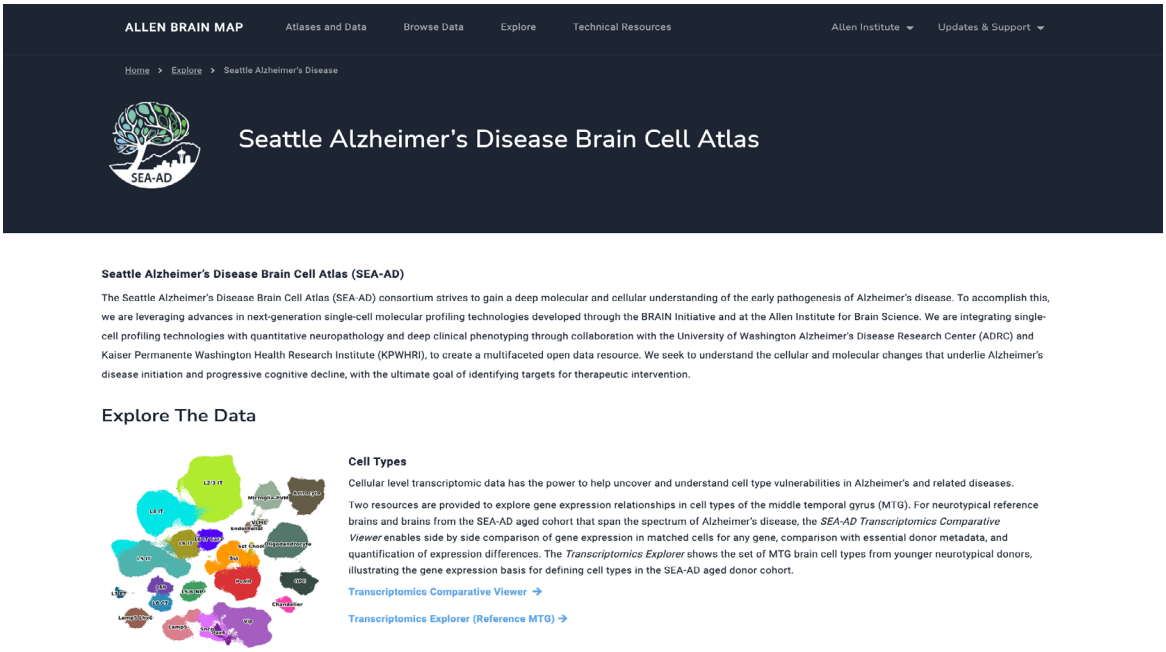
In order to orient you to the CZ cellxgene interface and how it can be used to explore questions related to Alzheimer's disease pathology, the following section provides you with a detailed tour of the database.

Database Tutorial:

In this section, we will explore how to use the ***cellxgene*** *interface. We will learn how to interpret UMAPs and how to filter the data to explore gene expression across cell types, donors, and genes of interest.*

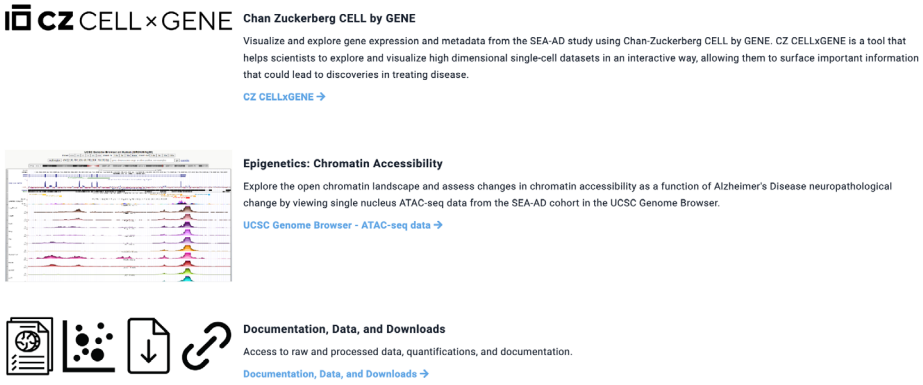
1. In order to access the Allen Institute for Brain Science's Seattle Alzheimer's Disease Brain Cell Atlas, go to SEA-AD.org.

Your web page should look like this:

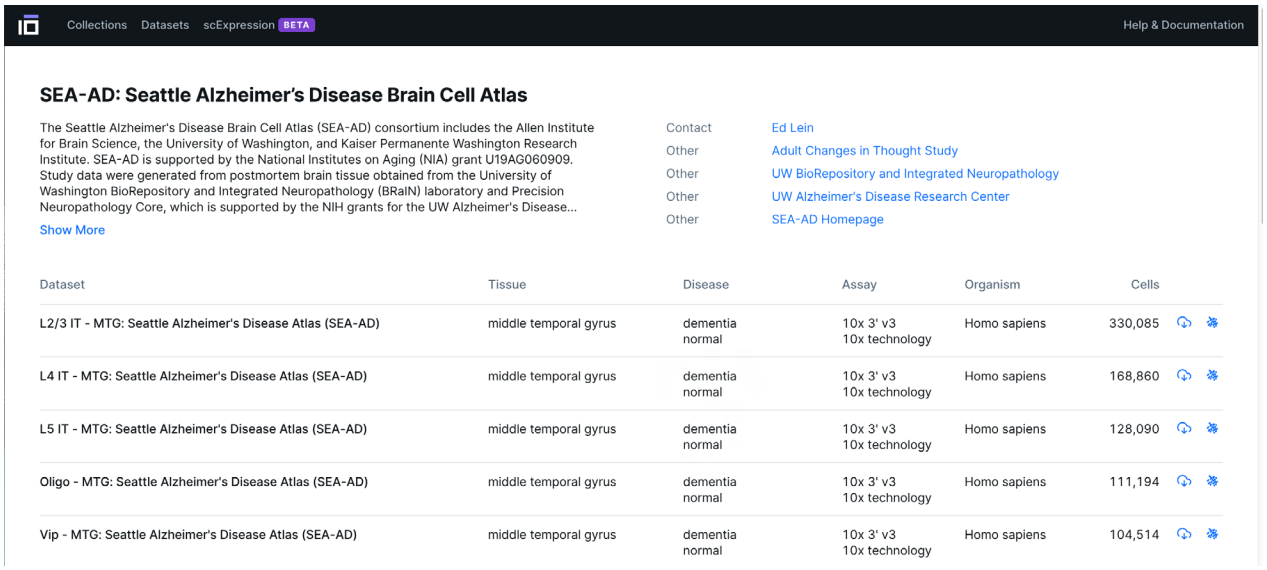


2. Scroll down until you see the **CZ cellxgene** link under the "Explore The Data" header.

3. Click on the "**CZ cellxgene**" link that is in blue.



After clicking the CZ cellxgene link, your page should look like this:



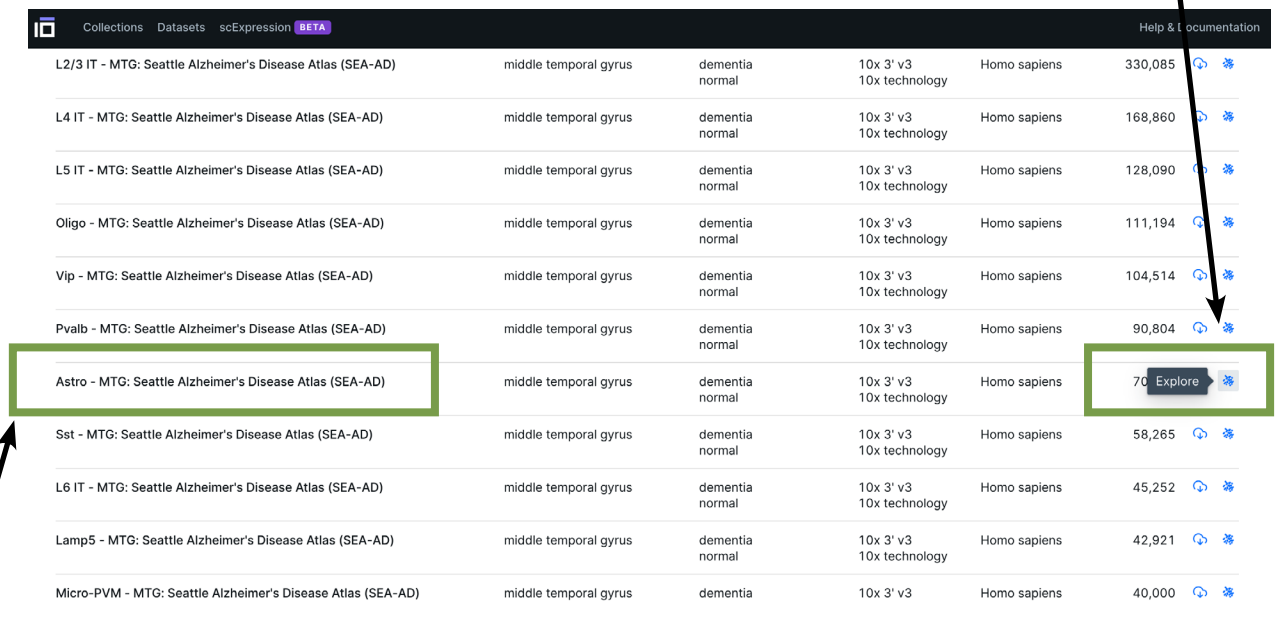
Each dataset linked on this page is for a different **cell type** within the brain. For example, the first dataset is titled "L2/3 IT – MTG: Seattle Alzheimer's Disease Atlas (SEA-AD)." This is the transcriptomic data from cells within the L2/3 layer of the middle temporal gyrus (MTG) within the human brain.

Cell types of interest: Astrocytes and Microglia

Although there are several different cell types listed within the cellxgene database, there are two particular cell types that are of interest to AD researchers. Scientists hypothesize that **astrocytes** (abbreviated as "Astro") and **microglia** (abbreviated as "micro") are two specific types of cells that may play a role in AD pathology. The SEA-AD database contains data for multiple types of brain cells, but for the purposes of this tutorial, we will look at the data for astrocytes and microglia.

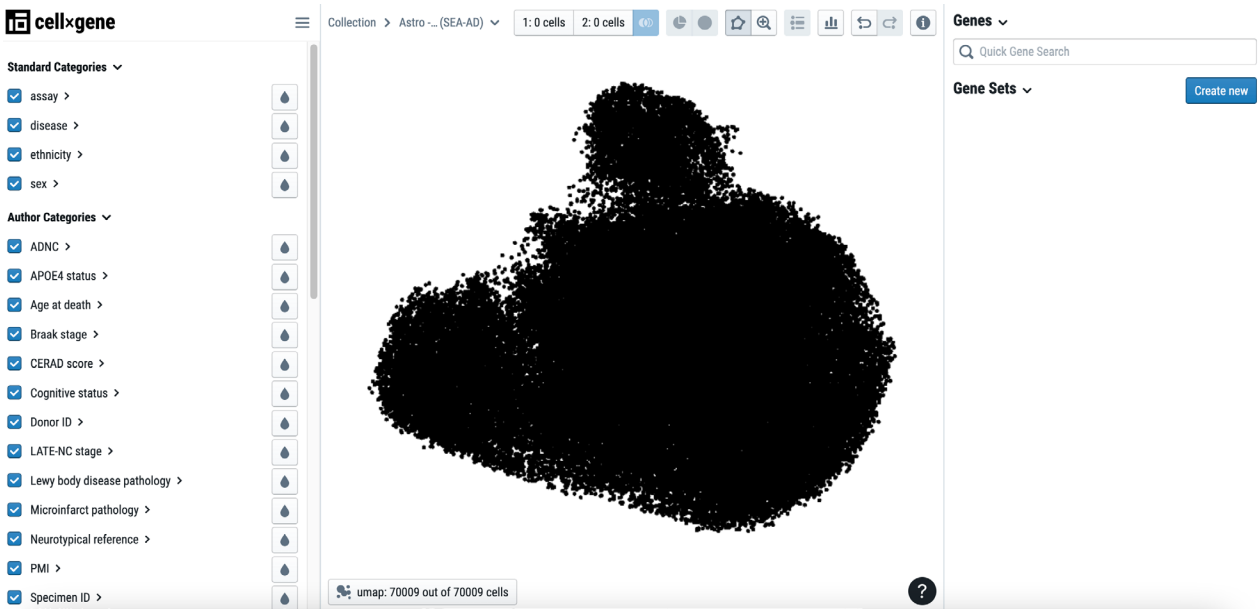
4. For the purposes of this tutorial, scroll down on the page until you find the dataset labeled “Astro-MTG: Seattle Alzheimer’s Disease Atlas (SEA-AD).” On the right-hand side, click on the “explore” icon.

the explore button can be found here next to the download icon



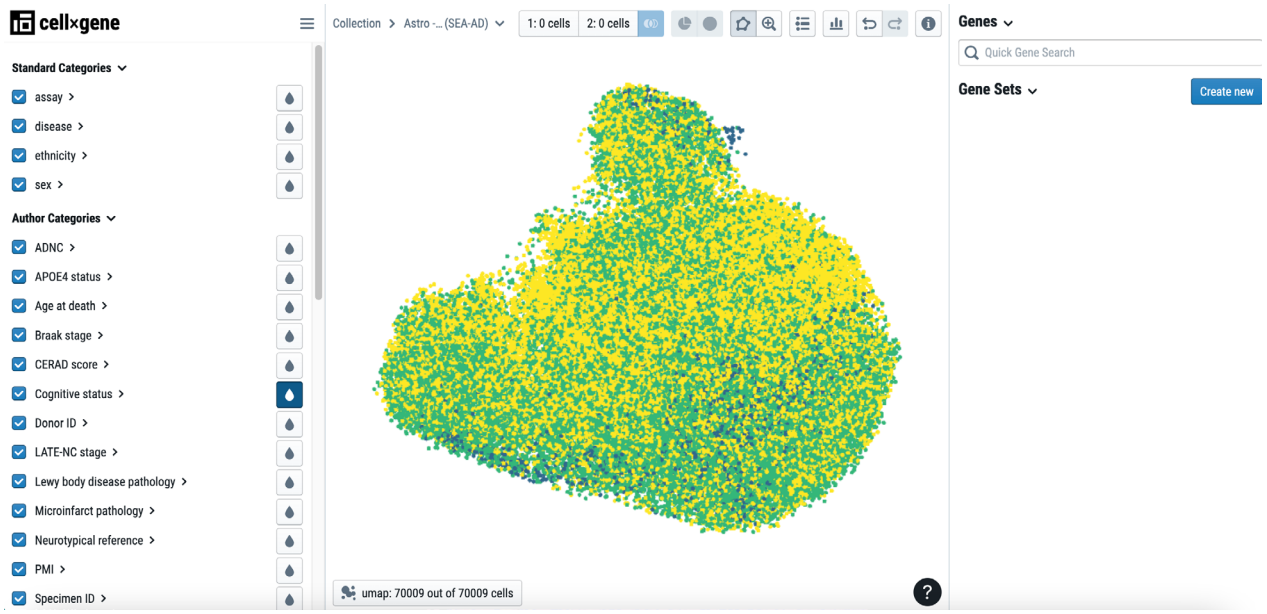
Scroll until you find the dataset labeled “Astro-MTG”

After clicking on “explore,” your screen should look like this:



In order to analyze this figure, we can apply filters. Select the paint drop icon next to each of the categories. For this tutorial, click on the paint drop icon next to “cognitive status.”

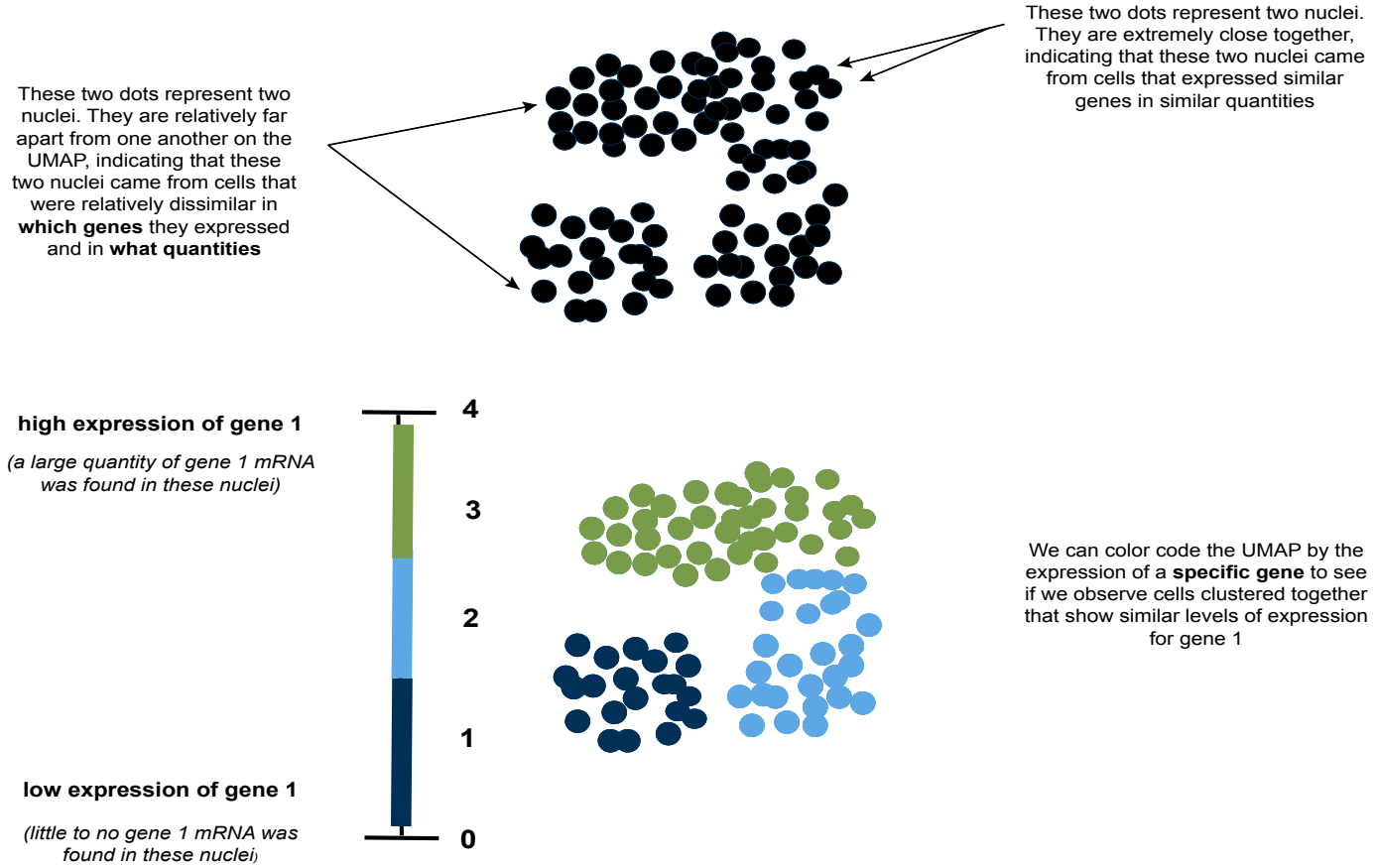
The color of the UMAP should change, and your screen should now look like this:



This is a special type of graph, called a **Uniform Manifold Approximation and Projection (UMAP)**. UMAPs are helpful ways of displaying many types of data, including transcriptomic data. In other words, these graphs help us to compare gene expression between cells. To review how UMAPs are generated and the type of data they display, refer to the transcriptomic data infographic included at the start of the lesson.

In order to learn how to analyze a UMAP, read through the figure listed below and discuss it with a neighbor/classmate/friend.

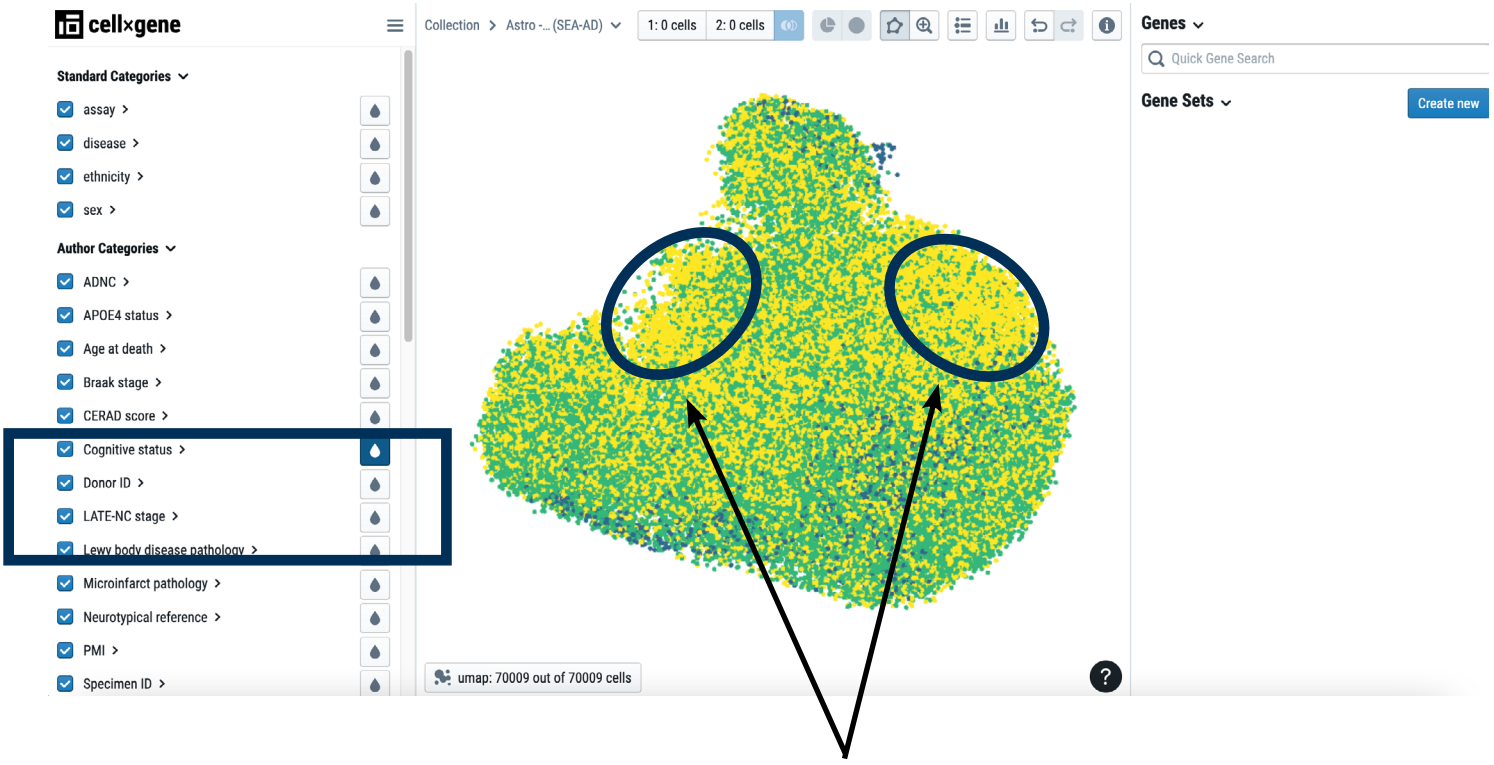
How do we analyze a UMAP?



Filtering by donor characteristics:

Dementia Status

Now that we know how UMAPs are generated and interpreted, let's return to the cellxgene data. By looking at the colors and clusters in the UMAP below, we can compare gene expression between cells taken from patients with dementia (yellow) and without (green).



Yellow dots represent nuclei from patients with dementia. Notice here we see these nuclei clustered together away from the green nuclei. The green nuclei represent nuclei isolated from brain tissue from donors who did not have dementia.

Knowledge Check

1. Does there appear to be a difference in gene expression between cells taken from patients with dementia compared to those without dementia? How can you tell?

2. Is this UMAP comparing gene expression generally, or is it looking at a specific gene?

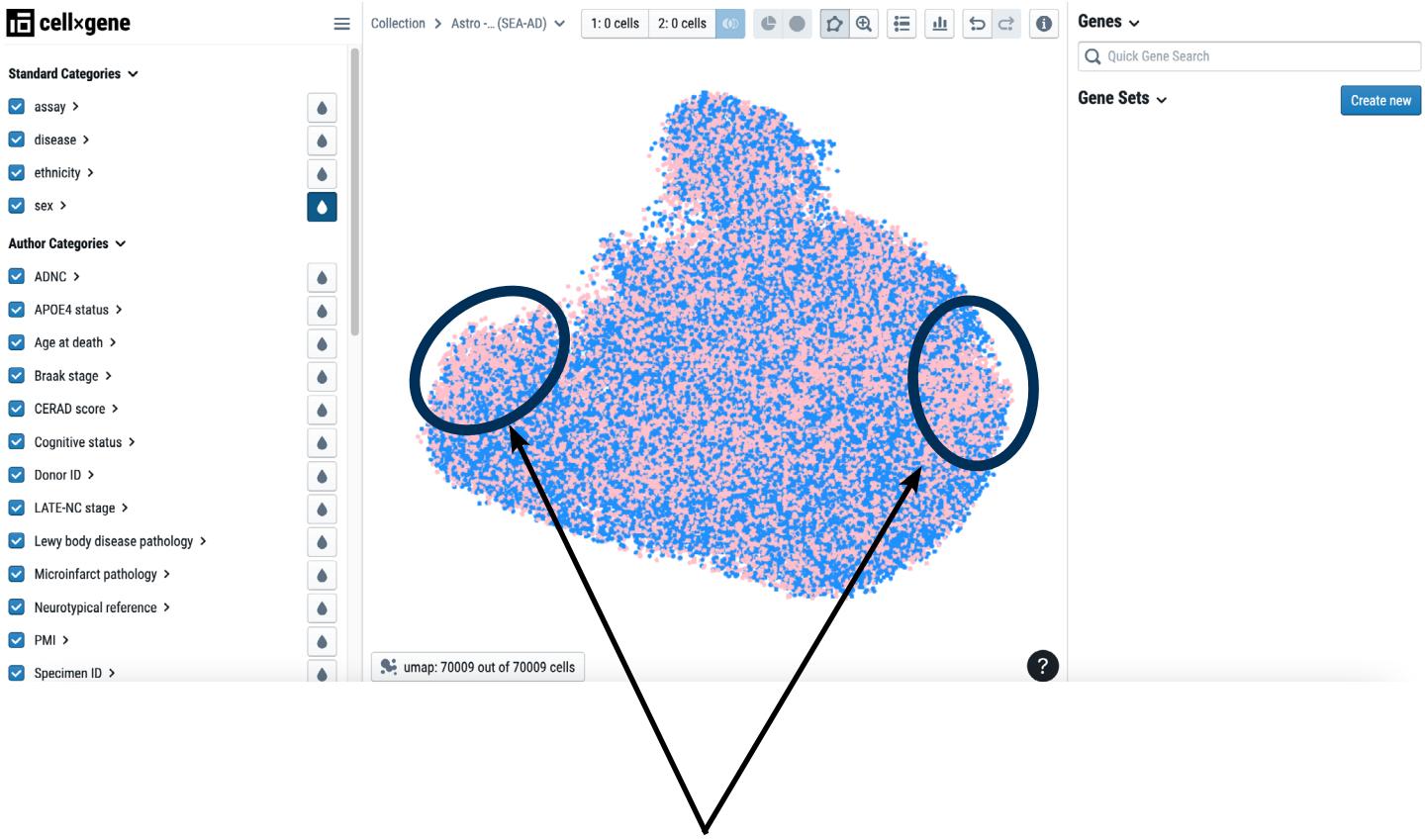
Biological Sex

In addition to filtering by dementia status, we can also compare gene expression of cells taken from biological males and biological females. As discussed in lesson 3, sex refers to a person's biological classification based on their reproductive organs and sex chromosomes, while gender is defined as "set of social, psychological, or emotional traits, often influenced by societal expectations that classify an individual as either feminine or masculine."

Reference: For more information about gender identity and the difference between sex and gender, visit https://medicine.yale.edu/whr/about/mission/definitions/.

Unclick the rain drop icon next to "cognitive status." Your screen should go back to a black UMAP.

Next, click on the raindrop icon next to "sex." Your screen should look like this:



Notice these clusters of pink on the UMAP

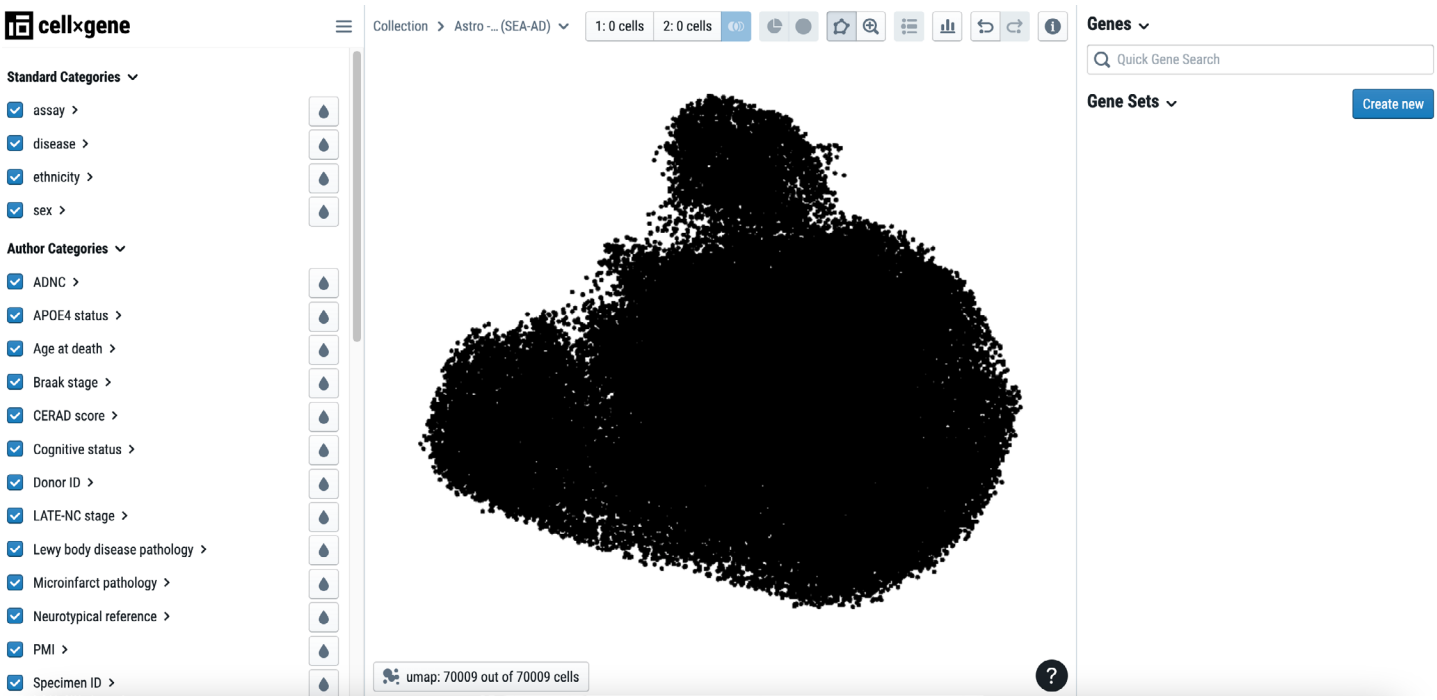
Knowledge Check

1. In your own words, describe what the pink clusters of dots represent within the UMAP on the previous page:

Filtering by gene:

In addition to filtering by one of the donor characteristics, such as cognitive status or donor sex, the cellxgene interface also allows us to compare gene expression between cells for **a specific gene**.

First, deselect the rain drop icon next to "sex." Once again, your screen should now show the black UMAP.



To filter by gene, go to the right hand column labeled "genes." For this tutorial, we will look at the gene labeled "APOE."

APOE is one of several genes that are suspected to play a role in Alzheimer's disease. Specific alleles of the APOE gene, such as APOE4, have been associated with increased risk of Alzheimer's disease. It is important to note that just because an individual has an allele of APOE associated with higher risk, this does not confirm that they will eventually develop AD. For the purposes of this lesson, we are looking at APOE expression broadly and not expression based on certain APOE alleles.

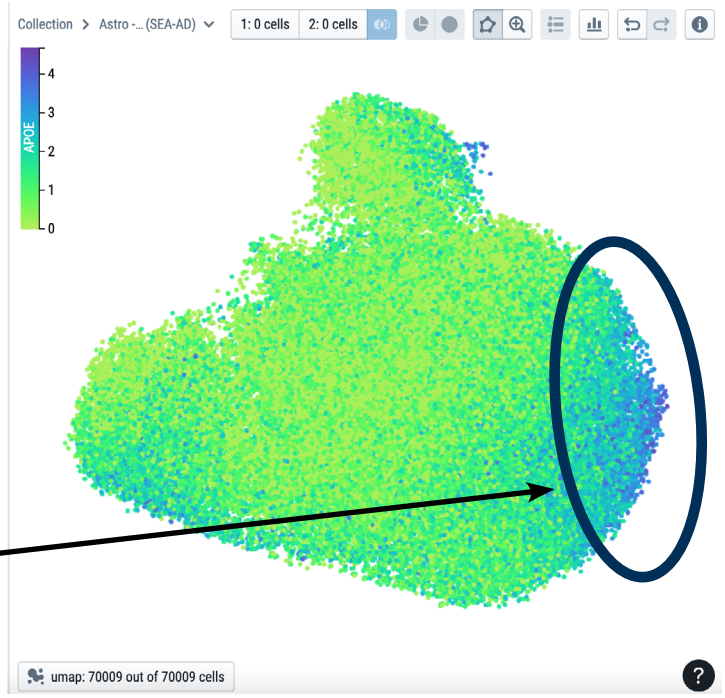
To filter the data by expression of the APOE gene, type in APOE into the “genes” search box.

Next, click the rain drop icon next to the APOE gene. Your screen should look like the figure below:

This key shows which color corresponds to high or low expression of APOE

Blue dots represent nuclei that had high levels of APOE expression.

In this UMAP, we see these blue dots clustered together on the right hand side.



Knowledge Check

1. What do the cluster of blue cells/nuclei at the right-hand of the UMAP tell us about the data?

Looking for Differential Gene Expression:

In addition to looking at the expression between cells of a single gene, the CZ cellxgene tool also allows us to identify genes that are differentially expressed between two categories. This is a particularly helpful feature of the cellxgene interface since it allows us to identify potential genes of interest that show different levels of expression based on specific characteristics like a donor's sex, age, etc.

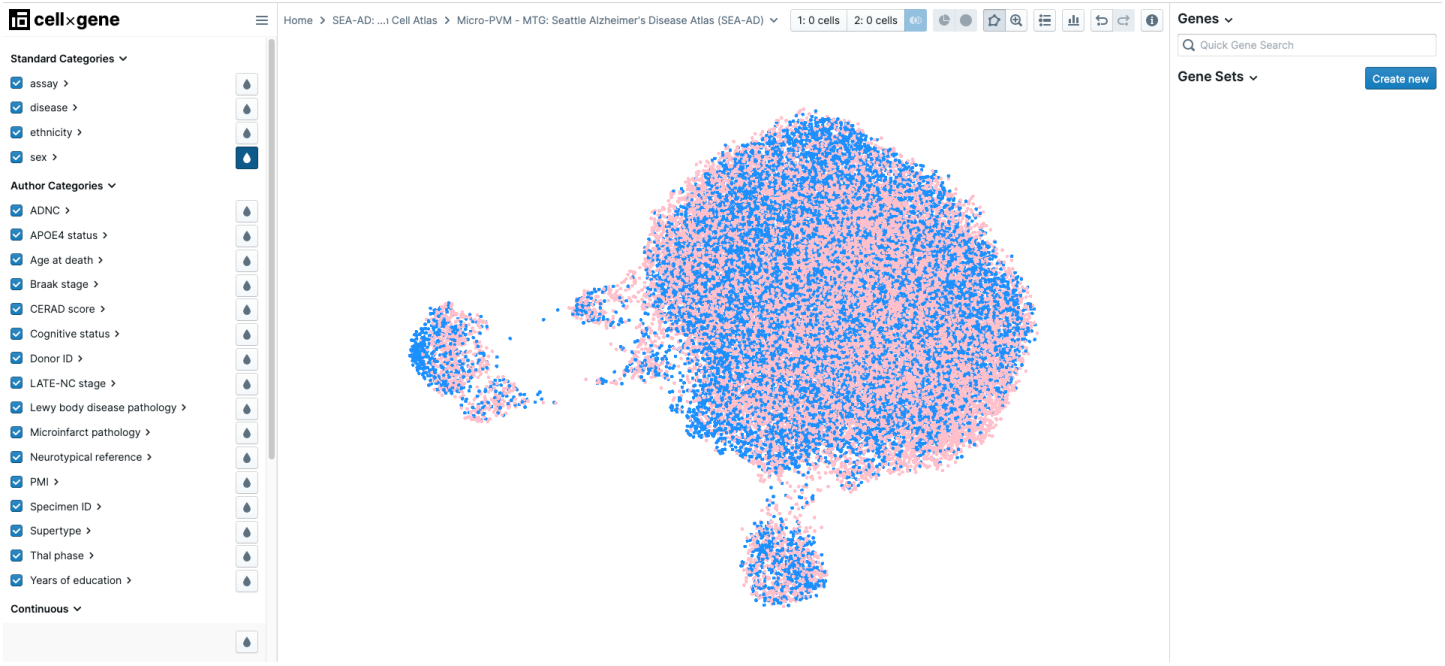
We will explore how to do this below:

1. We will start with a blank slate in cellxgene. Go to: https://cellxgene.cziscience.com/collections/1ca90a2d-2943-483d-b678-b809bf464c30

2. In the earlier part of the tutorial, we explored the data for astrocytes cells from the SEA-AD cohort. For this activity, we will switch to exploring the data for microglial cells. Scroll down and select the explore icon next to: "Micro-PVM – MTG: Seattle Alzheimer's Disease Atlas (SEA-AD)." If you are struggling to locate this, you can also access it directly from this link: https://cellxgene.cziscience.com/e/c76098ba-eed3-45b1-98f2-96fcac55ed18.cxg/

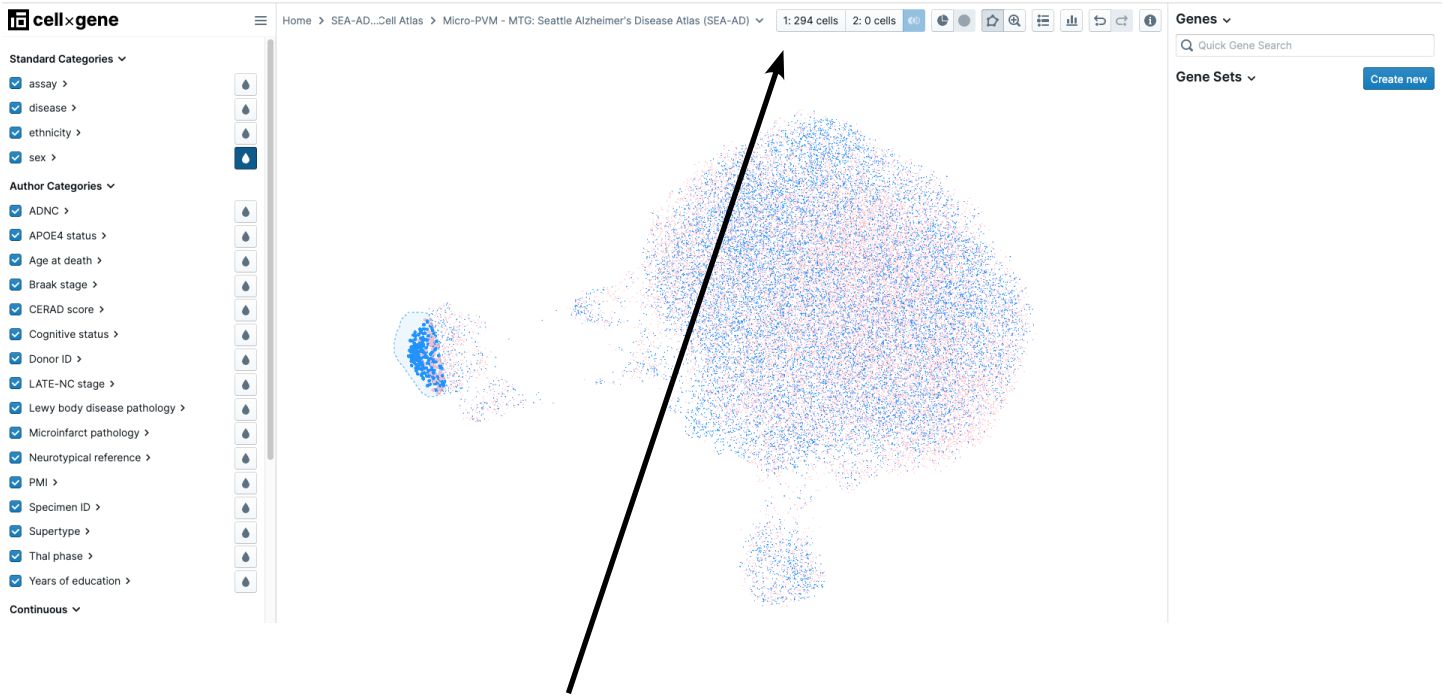
3. This page should open to display an all-black UMAP. We can filter this to explore specific donor criteria of our choosing.

4. Say we wanted to compare gene expression between the microglial cells in males and females. To filter by sex, select the paint drop icon next to "sex." Your screen should now look like this:



5. Notice at the top of the UMAP, there is a box labeled 1: 0 cells and 2: 0 cells. This is the highlight feature that enables you to highlight specific PARTS of the UMAP and compare gene expression between those clusters of cells.

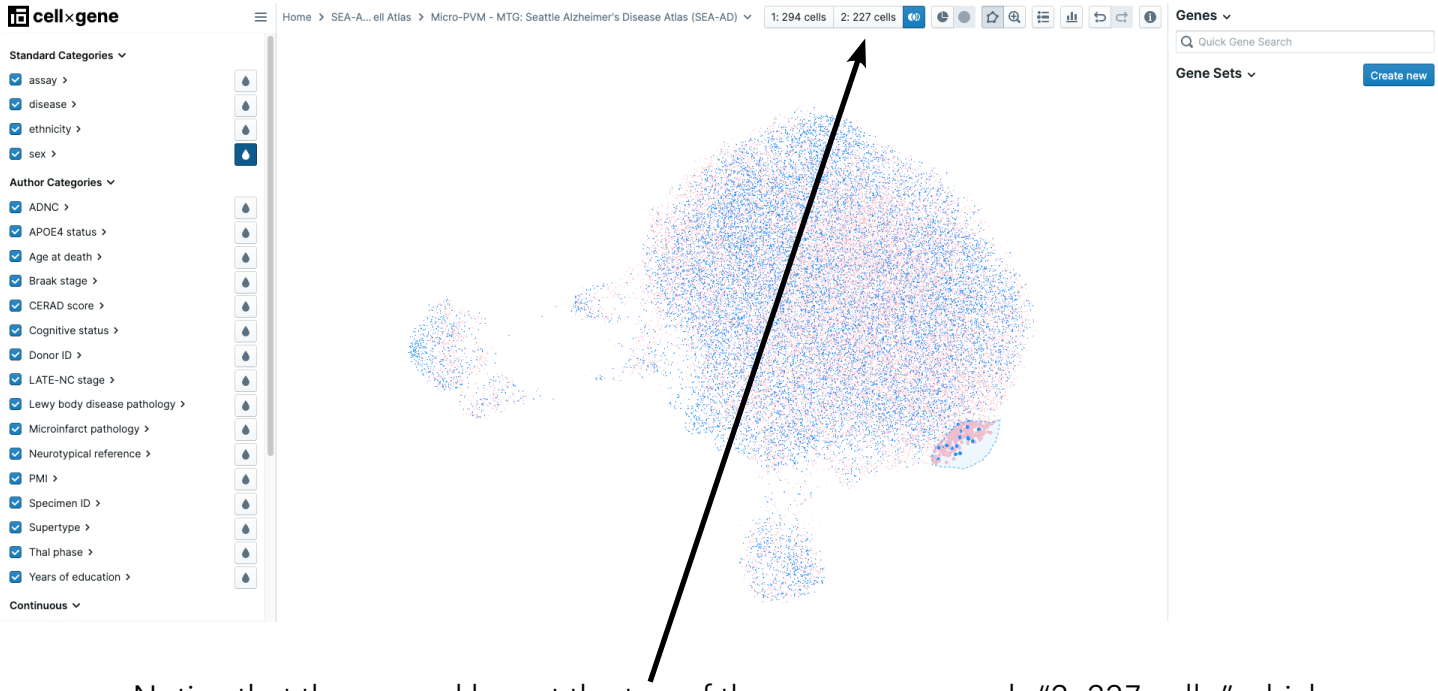
6. To start, hold down your cursor and drag over the cluster of blue dots on the lefthand side of the UMAP. After you have your cells selected, click on the box labeled "1:0 cells." This box should change and tell you how many cells you have highlighted. Your screen should look something like this:



Notice that the group 1 box now reads 1: 294 cells. This means that in the screenshot above, the user selected a total of 294 cells.

7. Now we need to select a second group of cells to compare with this first one! Click anywhere on the screen to deselect group 1. Next, repeat the same process and highlight the cluster of pink dots on the righthand side of the UMAP. Don't forget to click on the 2:0 cell box after you have highlighted your cells to save your selection.

Your screen should now look something like this:

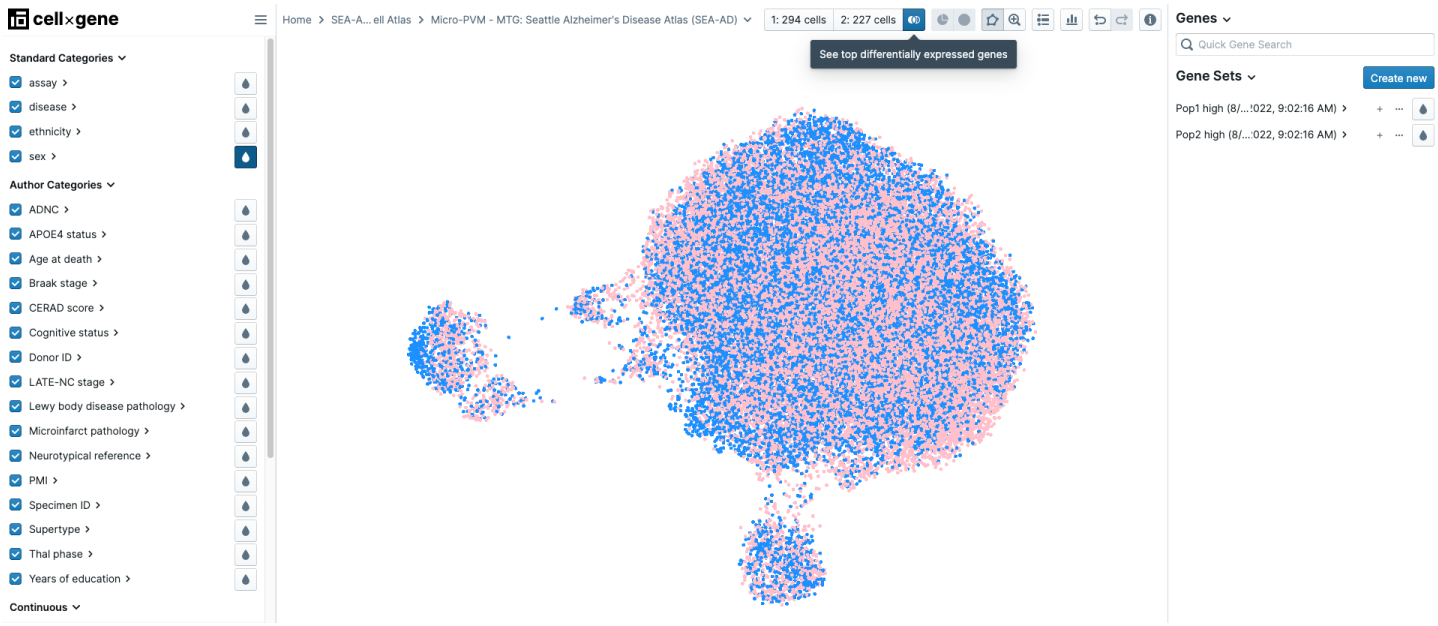


Notice that the second box at the top of the screen now reads "2: 227 cells," which tells us that this second selection contains 227 cells.

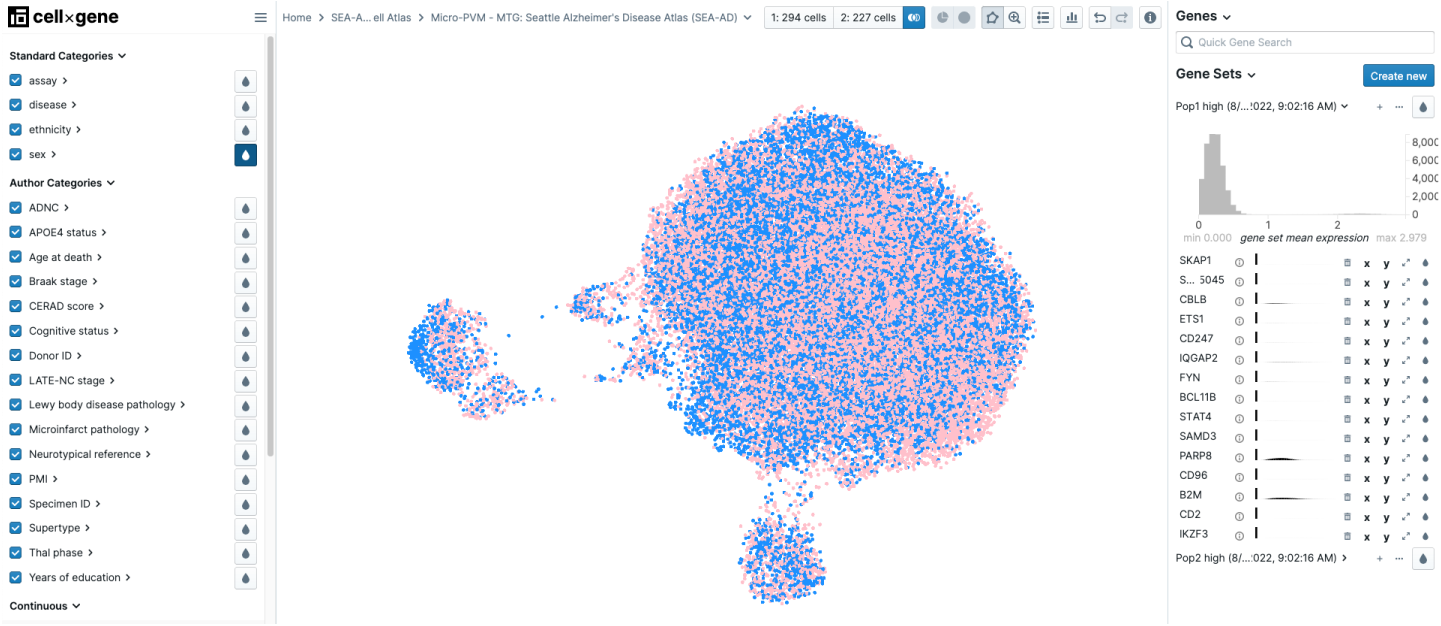
Note: It is likely that the numbers in your highlighted area may be slightly different than the ones given in this example depending on how many cells you highlighted and where on the UMAP you highlighted.

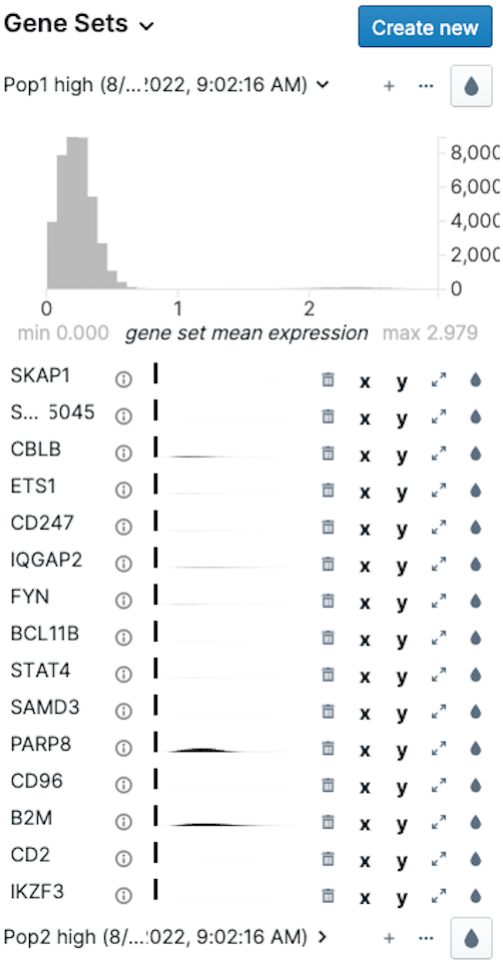
8. Since we have now highlighted two groups of cells, we can use cellxgene to find the top genes that are differentially expressed between these two groups of cells.

To do this, click on the blue icon next to the two boxes displaying the number of cells you highlighted in groups 1 and 2:



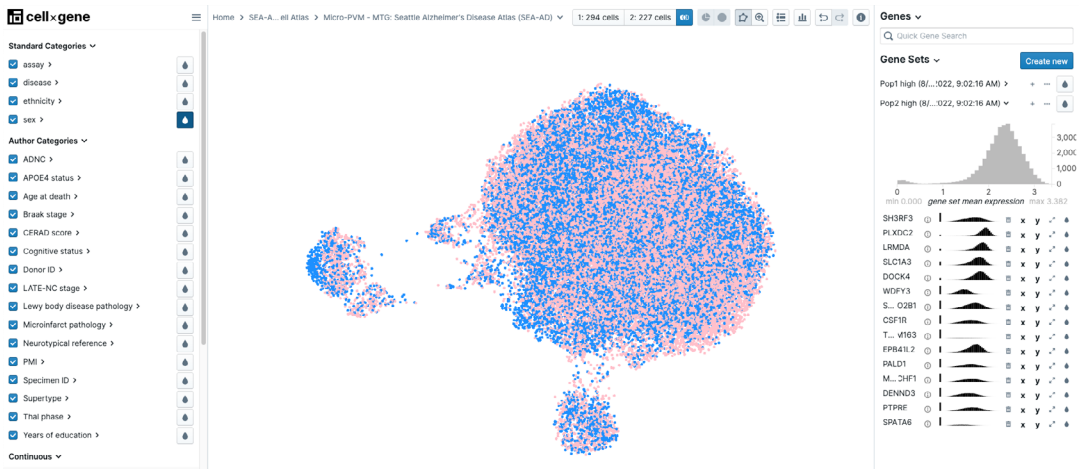
9. After clicking on this icon, cellxgene will begin drawing data on the top genes expressed in each group of cells. When the data loads, click on the "Pop1 high" bar under "Gene Sets." After clicking on "Pop1high," your screen should look something like this:





The "Gene Sets" column tells you which genes have the highest mean expression in population 1 **relative** to population 2. In the example column shown below, CBLB, PARP8, and B2M are the three genes that the first group of cells express the most.

10. Cellxgene allows us to compare the genes that are most highly expressed in population 1 relative to population 2, and vice versa. To look at the gene expression of group 2, click on the arrow next to "Pop2 high." Your screen should look something like this:



Knowledge Check

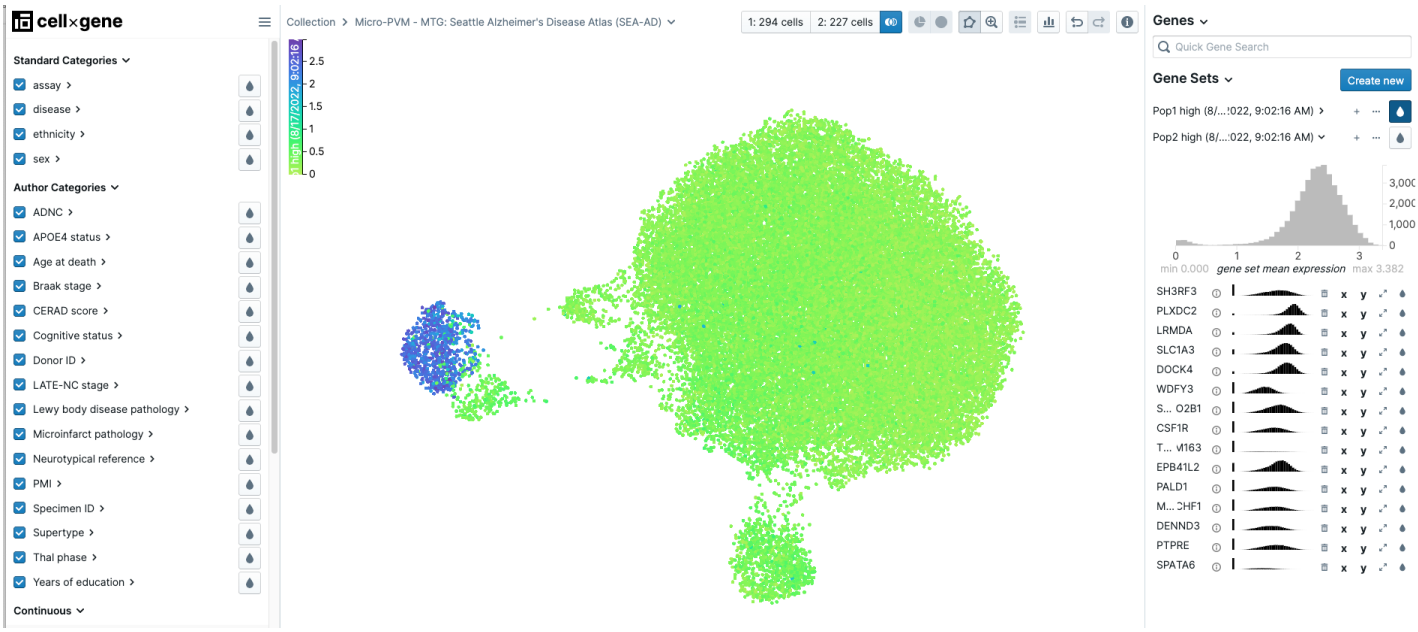
- Look at the screenshots on the previous pages and think back to what you learned about how UMAPs are constructed. Based on the areas of the UMAP that were highlighted for population 1 and population 2, would you expect these two populations of cells to have similar gene expression? Why or why not?

- Looking at the screenshots on the previous pages, which genes appear to be expressed in the highest quantities in population 1 compared to population 2? How can you tell?

- Which seven genes appear to be expressed in the highest quantities by the group 2 of cells relative to group 1?

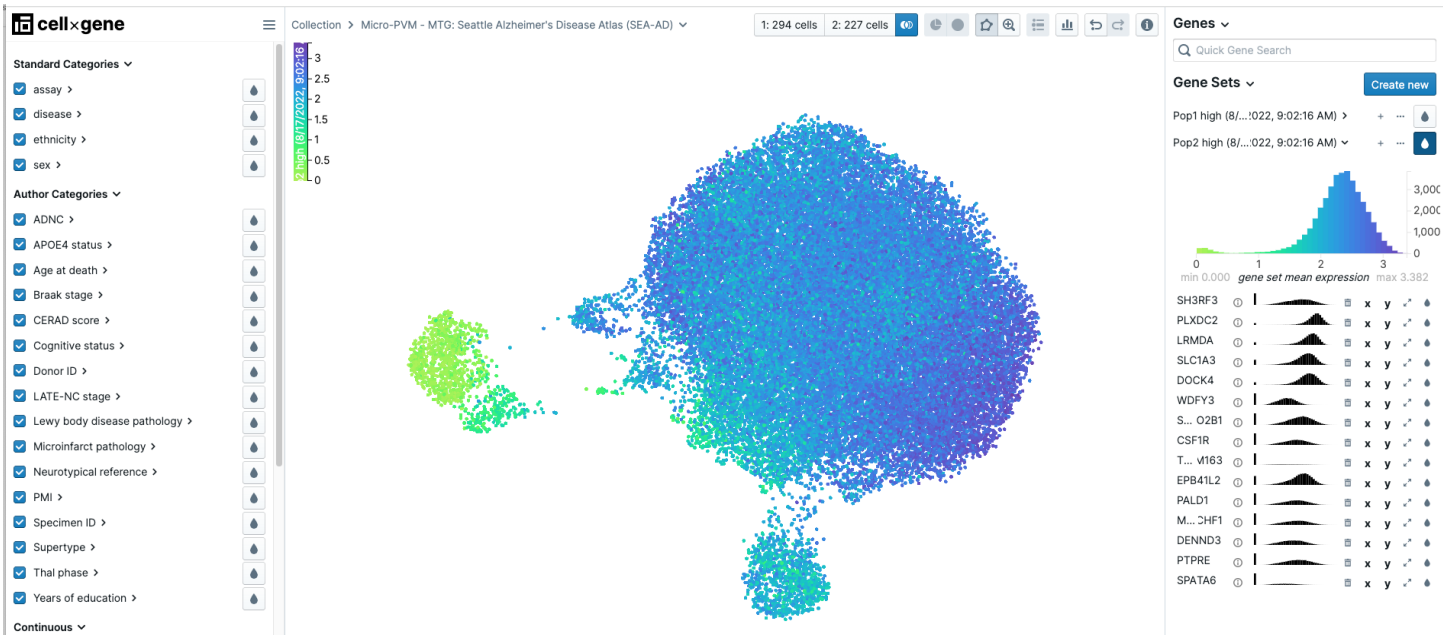
11. In addition to telling which genes are expressed at the highest levels in group 1 relative to group 2, we can also filter the UMAP to see which genes group 1 expressed in the highest quantities relative to group 2 by clicking on the rain drop icon next to “Pop1.”

To do this, click on the paint drop icon next to Pop1. Your screen should look like this:



12. You can also filter the UMAP by the genes that group 2 expressed in the highest quantities relative to group 1 by clicking on the rain drop icon next to “Pop2.”

After clicking on the rain drop icon, your screen should look like this:



Now that you have an understanding of how to use the cellxgene interface to analyze UMAPs and identify genes of interest between specific cluster of nuclei/cells, you will now have the opportunity to design your own experiment using the cellxgene tool!

Your turn:

To help you design your own experiment using the cellxgene interface, we have created the following table for you to fill out. This table requires you to consider several questions concerning which cell type you will explore, which clusters of cells/nuclei you will compare, and what genes you will explore further.

To begin filling out the table, start by going to the list of datasets available through the Seattle Alzheimer's Disease Brain Cell Atlas: https://cellxgene.cziscience.com/collections/1ca90a2d-2943-483d-b678-b809bf464c30

Remember that each of the datasets included for this list is for a specific cell type. For example, the first file titled "**L2/3 IT – MTG: Seattle Alzheimer's Disease Atlas (SEA-AD)**" represents the data from the nuclei isolated from brain tissue from the L2/3 layer of the middle temporal gyrus (MTG) within the human brain.

Step 1: Pick which cell type you will explore

Example: "Oligo – MTG: Seattle Alzheimer's Disease Atlas"

Cell type database you will use in cellxgene: _____

Step 2: Pick which donor characteristic you will filter the UMAP by

Example: dementia status, sex, age at death, etc.

Donor characteristic you will filter by:_____

Step 3: Perform your own analysis

Follow the procedure provided earlier in the tutorial. Highlight 2 clusters of cells that you want to compare gene expression for and fill out the table below with the information provided for each of your clusters by cellxgene.

	How many cells did you select for this population?	What 3 genes did this population of cells express at the highest levels relative to the other population you selected?
Population 1 of cells		
Population 2 of cells		

Step 4: Explore functions of genes that are differnetially expressed

Now that you have identified which genes are differentially expressed between these two clusters in your UMAP, you will now have the chance to explore what these genes' known functions are. We can explore each gene's function by going to the NIH gene database!

Go to the NIH gene database: https://www.ncbi.nlm.nih.gov/gene

Using the NIH gene database, research the three genes that cellxgene identified your group 1 cells expressing in the highest quantity. In order to find a gene's function, all you do is copy over the name of the gene into the NIH database search bar and click search! Make sure you are looking at the gene expression in humans, which are listed as "homo sapiens" in the database. Fill out the table below with the information you find:

Population 1	Official full name of gene	According to the NIH gene database, what are the known function(s) of this gene?
Gene 1		
Gene 2		
Gene 3		

Step 5: Repeat for population 2

Repeat this process for the three genes that cellxgene identified being expressed the most in your group 2 cells:

Population 2	Official full name of gene	According to the NIH gene database, what are the known function(s) of this gene?
Gene 1		
Gene 2		
Gene 3		

Reflective Questions:

1. What demographic characteristic did you choose to explore and why?

2. You used the NIH gene database to explore the known function(s) of specific genes. Were the functions of the genes expressed at the highest level in your group 1 relative to group 2 similar to the function of the genes expressed at the highest level in your group 2 relative to group 1?

3. What questions about AD pathology do you believe scientists need to answer? How could these data help them?

Conclusion:

Throughout this lesson, you had the opportunity to learn about the value of transcriptomic data and how it can be used to study disease pathology. In addition to understanding how transcriptomic data is obtained, this lesson challenged you to critically analyze graphical representations of this data using the CZ cellxgene interface.

For those interested in learning more about the Allen Institute and its research into AD pathology, check out https://portal.brain-map.org/explore/seattle-alzheimers-disease

While this lesson focused on transcriptomic data analysis, additional lessons available at https://alleninstitute.org/about/education-outreach/teaching-materials/ cover other aspects of disease pathology. These other lessons include bioethical discussions of brain donation, activities that explore the importance of basic scientific research, and explorations of the biological and social hallmarks of AD.

If you develop your own lesson plan using Allen Institute resources, we invite you to share your experience with us at communications@alleninstitute.org. Teachers are also encouraged to publish original lessons using our open data, tools, and other resources, and to share those lessons with us.