The Building Blocks of the Brain: Single-Cell Transcriptomics Add-on

This is an add-on to the lesson plan The Building Blocks of the Brain (alleninstitute.org/learn/building-blocks). Please read and review that resource before incorporating this addendum.

- Insert pages 3-5 of this worksheet into the main lesson after page 25, at which time students will be prompted to collect additional data.
- Insert pages 6-7 after page 29, adding more data analysis/discussion questions before students are prompted to design their own experiment.
- Page 8-9 contain additional challenge questions for highly motivated students. These optional questions are meant for students interested in pursuing independent projects with the data used in the lesson or may be suitable for extra credit.

Intended grade level

This add-on worksheet is more advanced than the main Building Blocks of the Brain lesson. It is primarily intended for use at the college level. It may also be suitable for some highly motivated high school seniors in advanced biology courses, ideally after AP/IB/equivalent biology.

What does this add to the original lesson?

- This add-on includes material related to the Allen Cell Types Database single-nucleus transcriptomics data. The worksheet specifically uses the Human – Multiple Cortical Areas SMART-Seq genomics dataset released in 2019.
- In the first part of this worksheet (data collection), students use online viewers to explore the gene expression profile of transcriptomally-defined cell types. This online viewer was not available at the time of publication for the original lesson plan.
- In addition to exploring the transcriptomic data, the second part of the worksheet (data analysis/discussion) guides students through analyzing the data both on its own and in conjunction with the other data collected in the main worksheet.

Highlights from the Allen Cell Types Database: RNA-seq data

- All RNA-Seq data included in the Allen Cell Types Database surveys gene expression by capturing RNA transcripts. This process allows us to quantify gene expression at the level of individual cells.
- In humans, this process is done in single nuclei, and in mice in single whole cells. We also use two slightly different processes, SMART-seq and 10x genomics, for different datasets. These methods produce similar gene expression profiles for matched cell types.
- This lesson uses the Human – Multiple Cortical Areas SMART-Seq genomics dataset. As of fall 2020, this dataset includes data from about 50,000 cells from middle temporal gyrus, anterior cingulate cortex, primary visual cortex, primary motor cortex, primary somatosensory cortex, and primary auditory cortex. Most cells came from postmortem donations and some came from living neurosurgical patient donations. To learn more about neurosurgical donations, please see this article: alleninstitute.org/what-we-do/brain-science/news-press/articles/what-its-donate-your-brain-science.
- The web interfaces for viewing all RNA-seq datasets are the same and all the data can be downloaded for more advanced analysis. Students may wish to explore and analyze additional data for a challenge assignment or independent project.

Student worksheet
Part 4: Transcriptomics

In this section of the activity, you will explore another area of the Allen Cell Types Database that includes single-cell gene expression data. From the main Allen Cell Types Database page at celltypes.brain-map.org:

- Select RNA-seq from the top header
- Select Human – Multiple Cortical Areas SMART-Seq – Explore & Analyze.
- You may also use this direct link: celltypes.brain-map.org/rnaseq/human_ctx_smart-seq.

This will bring up a dataset of about 50,000 individual nuclei from middle temporal gyrus, anterior cingulate cortex, primary visual cortex, primary motor cortex, primary somatosensory cortex, and primary auditory cortex.

The microarray data you used in the previous section of this exercise used samples taken from entire brain regions and captured the presence of RNA for each gene in the genome, indicating how much that gene (as DNA) is transcribed (as RNA) and ultimately leads to proteins. Recall that the results were expressed as Z scores, as a specific, quantified amount of RNA cannot be identified from microarrays. RNA-seq also looks for the presence of RNA. In microarray we looked for RNA in a whole brain region, but in RNA-seq we look in a single nucleus. The RNA-seq process results in specific values for the amount of RNA present (reported as counts of this gene per million RNA sequences). Unlike the Z scores from the microarray data, RNA-seq values can be directly and quantitatively compared between genes once normalized.

You should see a dendrogram (tree diagram) pop up at the top of the screen with a heatmap and list of genes below. This dendrogram includes the approximately 125 types of cells that scientists identified in this dataset.

This data viewer bundles together data from all the cells in each cell type, and also bundles together cells from all of the cortical areas in the samples. In the web viewer, you can only view cells aggregated together across brain regions, but for an extra challenge or independent research project, ask your instructor about downloading and analyzing the full dataset files.

Cells are grouped into types by applying a mathematical algorithm to the dataset that will group cells with similar gene expression patterns together. These cell types are then named based on (1) prominent marker genes, (2) original location of the cells, and (3) comparisons with previous research on brain cell types. For more information on how those cell types were identified, see this video: youtu.be/ayJYEJKqdSw. More detail is available in Figure 1 of this paper, if you would like to learn more: nature.com/articles/s41586-019-1506-7. The genes that are automatically included in the heatmap are those that the Allen Institute scientists are using to define cell types. These are the genes that vary the most and the most consistently between different types.

However, this dataset does include all genes in the genome. Remember the genes you looked up back in Parts 1 and 2 of this exercise? Enter them into the “Add Genes” box in the top right of the dendrogram and add them to the viewer. This will add the genes to the heatmap.

What do you notice about the expression pattern of your genes compared to the genes that were pre-populated in the heatmap list?

Hover over the name of one of your genes and select the "Plot" button that appears. A scatter plot will come up – by default it is colored by gene expression (more intense color corresponds to more gene expression). At the top of the plot, use the drop-down menu to instead select to have it colored by cell type. Note that the map of cells is the same in each view, just color-coded differently.

Notice that each cluster (color) now represents one cell type from the dendrogram in the previous view. This type of plot is called a T-SNE, and very simply, it is a method of organizing data with many features (in this case, the gene expression levels of many genes) into a two-dimensional plot. In this plot, neighboring points are more similar to each other than they are to more distant points. In this visualization, most cells of the same type will group together.

Now use the drop-down to switch back to coloring the plot by gene expression. Remember each cluster in the scatter plot represents one cell type. The clusters in the scatter plot are always the same, no matter which gene you are focusing on, because they are defined based on the top genes that vary the most across types.

Is your gene expressed differently across cell types, or relatively evenly across most types? What other patterns do you notice?

You may also wish to view the heatmap to help with your analysis. To switch back to that view, select "Heatmap" from the visualization drop-down at the top of the viewing window.

Repeat this process for your other two genes. (To switch genes, you can select "Heatmap" from the visualization options at the top of the viewing panel and then select "Plot" from next to another one of your genes, or just enter your gene in the "Enter a gene symbol" box at the top right of the scatter plot view.)

Are these genes expressed differently across cell types or relatively evenly across most types? What other patterns do you notice?

Return to the heatmap view. Add this new gene that you have not previously explored: PSEN1.

Which cell types is this gene most expressed in? (Reminder: the cell types data groups cells across brain regions, and this dataset only includes selected cortical regions.)

Use the NIH Gene Database (ncbi.nlm.nih.gov/gene) to search for PSEN1. What is the function of this gene?

Interpret the expression pattern of this gene in light of its function, separately across brain regions and cell types.

Analysis questions

Compare your RNA-seq observations to your region-level observations from the Allen Human Brain Atlas. Because the cortical data is from multiple regions, you cannot make a direct region-to-region comparison between the microarray and the RNA-seq data, but you can think about how to interpret these two different types of data.

Describe a pattern or conclusion you can observe in both the RNA-seq and the microarray data.

Describe a pattern or conclusion that you can only observe in this set of microarray data and not from the RNA-seq data.

Describe a pattern or conclusion that you can only observe in this set of RNA-seq data and not from the microarray data.

Recent studies are interested in developing cell types based on cells' transcriptomics, morphology, and electrophysiology, rather than just one trait at a time as has been done in the past. Why do we want to sort cells into cell types?

Describe two challenges you would expect to face in identifying these MET (morphology, electrophysiology, transcriptomics) types compared to identifying types based on only one trait of a cell.

Challenge questions

Complete these questions for an added challenge.

Using the Allen Human Brain Atlas at human.brain-map.org, identify a few genes highly correlated gene with "MOG" in cortex. To do this, search for MOG in the atlas, select the checkbox to the left of the first row (representing one probe that measures expression of this gene), and then select the "Find Correlates" button above the right side of the heatmap. Look up these genes' functions in the NIH Gene database: ncbi.nlm.nih.gov/gene/. What brain regions is MOG highly expressed in?

Are MOG and genes co-expressed with it in the microarray data also co-expressed in the same cell types in human cortex (as seen in the cell types dendrogram: celltypes.brain-map.org/rnaseq/human_ctx_smart-seq)? If so, what are these cell types?

There are many sets of co-expressed genes that you can identify in the whole-brain Allen Human Brain Atlas (the data show with the anatomical regions heatmap) and in the single-cell Allen Cell Types Database (the data shown with the dendrogram heatmap). Besides this example, what is another comparison you could make between the region-level and the single-cell-level data?

Imagine you are studying a brain disease or disorder of your choice. You have collected data on gene expression in the brain of individuals with this condition, but not any healthy or neurotypical individuals. Describe how you would use the data from the Allen Human Brain Atlas or Allen Cell Types Database as part of your experiment. Which dataset(s) would be useful to your research, and why? What comparisons would you perform and how would you analyze the data together?