Visualizing Gene Expression in the Brain

cs480x Final Project

Victoria Grasso, Josh Lovering, Nicole Shedd

Project Proposal

The goal of this project is to develop a gene expression analysis tool used to compare brain regions and gene expression.

Summary

This tool will be used by bioinformaticians working in transcriptomics to analyze gene expression in the frontal cortex, visual cortex, and cerebellum.

It will be built from a healthy dataset containing 34,899 cells to serve as a control sample.

We used JavaScript D3 to create interactive scatter plots with multiple views for each brain region and gene type.

What do we hope to learn?

1. What are gene expression profiles across different cell types?
2. Is there a difference in marker gene expression in different brain regions?
3. Are the same cell types present in all 3 brain regions?
4. Is the proportion of cell types the same in all 3 brain regions?

Progress Book

Data

The data used in this project consists of 34,899 cells covering 3 brain regions and 31 gene types. This data was sourced from the a 2018 article in the Nature Biotechnology journal [1].

Brain Regions: The data describes cells from the cerebellum, frontal cortex, and visual cortex.

1. The cerebellum, behind the top of the brain stem, is responsible for voluntary motor movements, receiving stimuli from the sensory systems and spinal cord [2].
2. The frontal cortex lies within the frontal lobe, and is responsible for cognition, such as memory development [3].
3. The visual cortex, located in the occipital lobe, processes visual stimuli from the retina [4].

Cell Types and Marker Genes:

Astrocytes (Ast) - SLC1A2, SLC1A3, SLC4A4, GLUL, AQP4

Endothelial Cells (End) - COBLL1, DUSP1, FLT1

Excitatory Neurons (Ex) - SLC17A7, GRIN1, GRIN2B, SATB2

Inhibitory Neurons (In) - GAD1, GAD2, SLC6A1

Microglia (Mic) - APBB1IP, P2RY12

Oligodendrocytes (Oli) - CLDN11, MOG, MOBP, MBP

Oligodendrocyte Precursors (OPC) - PCDH15, OLIG1

Pericytes (Per) - COBLL1, PDGFRB

Granule Cells (Gran) - RELN, GRM4, RBFOX3
Only in the Cerebellum

Purkinje Neuron (Purk) - RYR1
Only in the Cerebellum

Data Processing Workflow

Filter cells:
- Low quality cells with few gene reads
- Multiplets with lots of reads
- Dying cells with mitochondrial contamination

Normalization:
- Normalize gene expression measurements for each cell divided by their total expression
- Log transform the results

Variable feature identification:
- Find features that have high cell-to-cell variation
- Helps highlight biological signal

Scale data:
- Shift expression so mean across cells is 0
- Scales expression so variance across cells is 1

Dimensionality Reduction

Run PCA:
- Clusters cells into PCs based on features
- Need to decide how many PCs to include in the rest of the analysis

Cluster Cells:
- Creates cellular distance matrix
- Construct k-nearest neighbor graph based on PCA and refine edge weights
- Use Louvain algorithm to group cells together

Non-linear dimensional reduction:
- Either UMAP or t-SNE (UMAP is this case)
- Plot the cells in a low-dimensional space
- Places similar cells together
- Cells with the same cluster should co-localize

This is Seurat's standard preprocessing workflow, and is meant to address high variance in the transcriptome

Design Evolution

The project is based upon pre-existing scatter plots generated for a WPI MQP. These plots were generated from the same dataset, for each brain region and gene type. The main consideration is how to create a visualization that can display each chart in one place, through interactivity.

Side-by-Side Visualization

In order to compare the Seurat clustered gene types, with the expression values from the dataset, we will need to display two scatterplots side by side. To accomplish this, we appended two different groups to the same SVG. One of these groups was placed on the left side of the SVG, and the other was placed at half of the SVG's width. In other words, the right-hand plot was placed x-coordinate minus the margin was the halfway point of the visualization. The resulting plot is displayed in **Figure 2**.

Design Choices - Colors

One pitfall of UMAP plotting in ggplot is the similarity of colors. We generated a color scheme with high perceptual differences between cluster colors. We also used a gray to blue gradient on the right-hand plot for high perception of the highly expressed genes

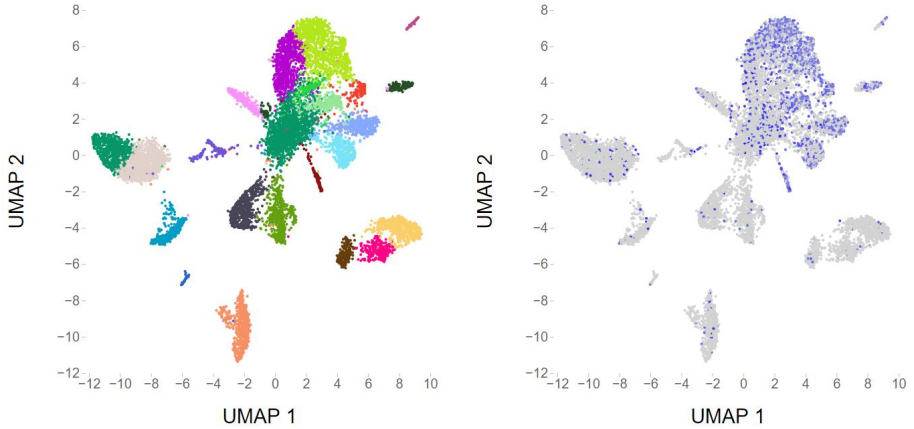

Figure 1. First side-by-side plot generated in D3. From left to right, a plot representing the Seraut clusters next to a plot displaying the SATB2 gene expression, indicated by purple.

Drop-Down Menu

To easily compare gene types, we opted for drop down menus attached to the visualization. There would be one menu used to select the region of the brain to display. Choosing a new region would change both visualizations to display this region. The left-hand plot remains constant for every plot of a given region. The second drop down menu would be used to select different gene types for visualizations. Our first implementation of the drop-down menus can be seen in **Figure 3.**

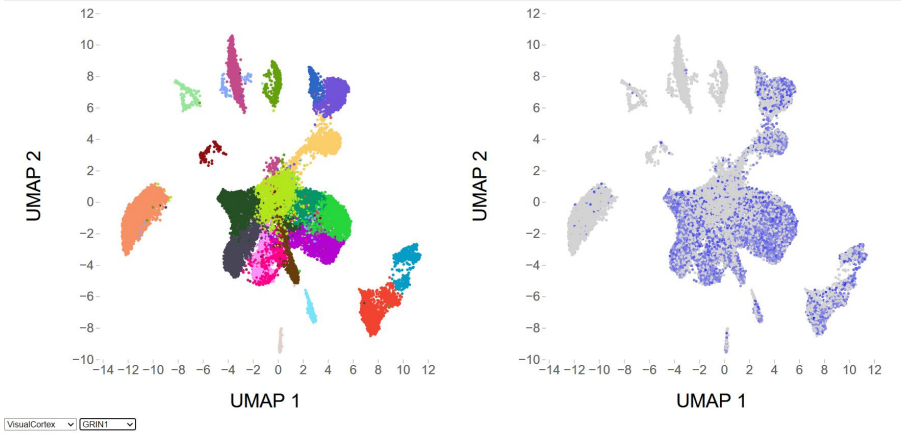

Figure 2. First drop-down menu implementation.

Title Header

After implementing the drop-down menu functionality, we added some user accessibility features. The most noticeable of these was a title header. This header serves as a location to write informational text about this tool, for users to read upon arrival. The header also houses the drop-down menus, to keep them in a consistent location with some contrast to a gray background so as to make them quickly noticeable. The header can be seen in **Figure 4**.

Alongside the header, axes lines were added in to improve readability of the charts.

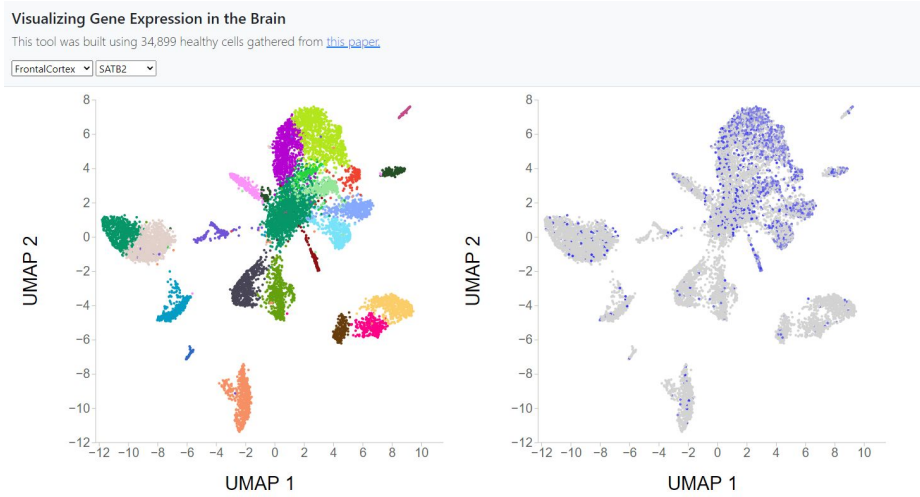

Figure 3. Title header implementation.

Hover Interactivity

To make our tool more effective for users, we implemented interactivity upon hovering. This was introduced in two ways. The first way, was from right to left. When hovering over the right-hand chart, the identical circle from the left-hand chart would be highlighted. This allows users to clearly see which cluster a certain cell is located in. This addition can be seen in **Figure 5**.

Shortly after, the same interactivity was also introduced on the left-hand plot to affect the right-hand plot. Next, we added interactivity onto the left-hand plot, shown in **Figure 6**. Upon hovering over a cell, all cells not belonging to that cluster would turn gray, revealing the entirety of the given cluster. Simultaneously, the right-hand plot cells not belonging to that cluster would significantly decrease in opacity. to clearly distinguish the cells from the cluster. This allows the user to easily see every cell belonging to a cluster.

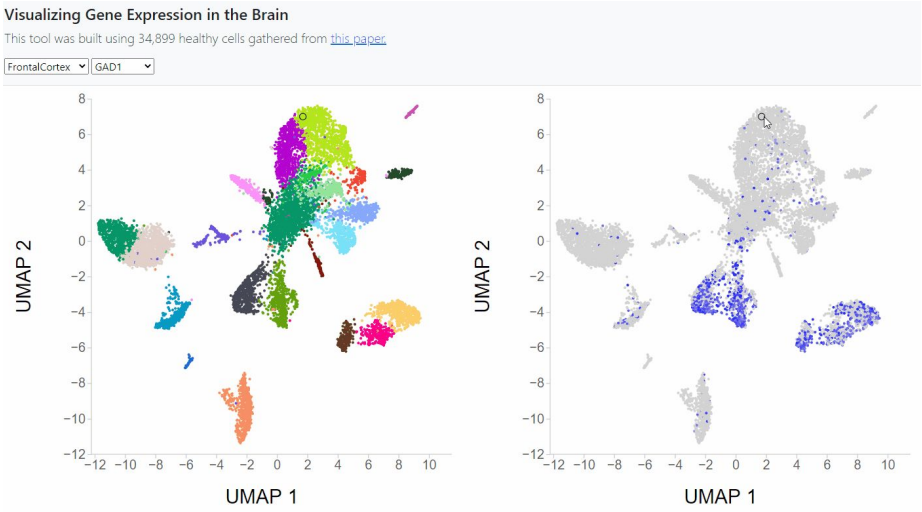

Figure 4. Right to left hover interactivity.

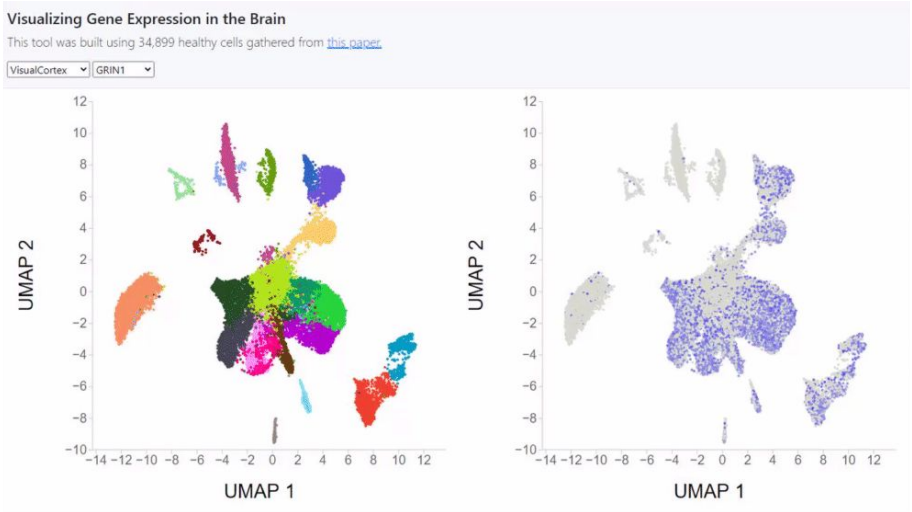

Figure 5. Right to left hover interactivity.

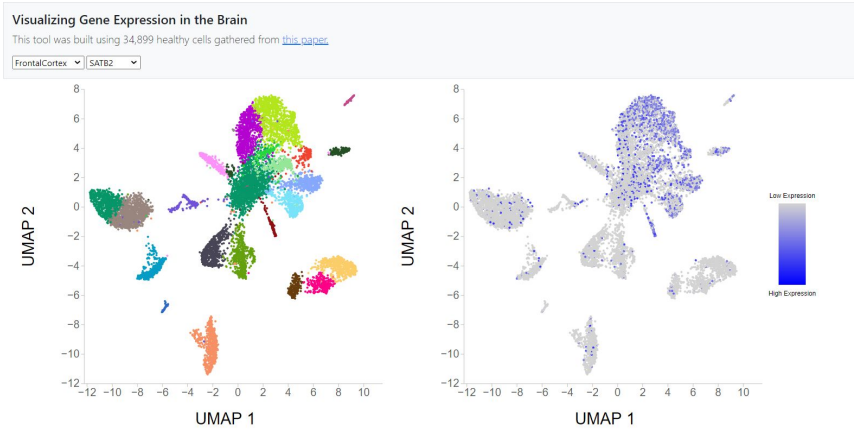

Figure 6. Legend implementation.

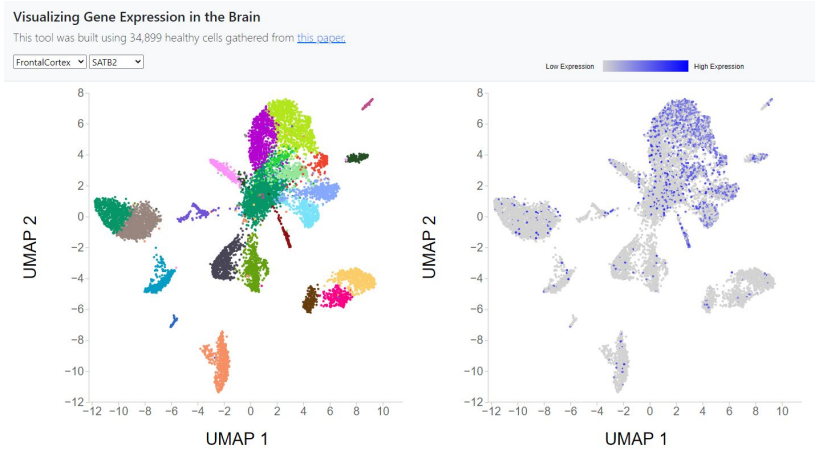

Figure 7. Legend Update.

Legend

We decided to add a legend next to the gene expression plot so users could easily understand the difference between the colors. Since the points were plotted on a scale ranging from light gray to blue, we decided to use a gradient legend. It helps users identify that cells with low expression of the gene selected would be a light gray color, and the cells with high expression would be a blue color.

Legend Update

We decided to move the legend above the gene expression plot so it takes up less space in the webpage. The title header previously had a lot of empty space, which we decided could be filled by the legend.

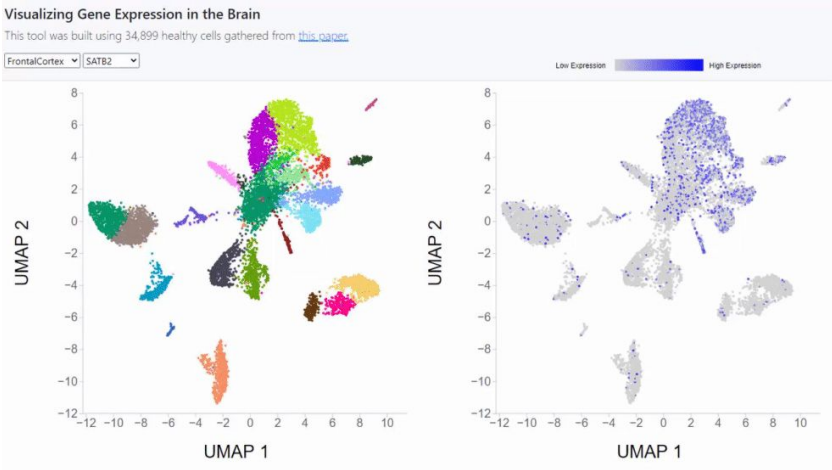

Figure 8. Tooltip implementation.

Tooltip

To add more interactivity for users, we added a tooltip to the left scatter plot. When hovering over cells on the left-hand plot, users will see a box appear with information about the clusters, cell types, and the marker genes that are expressed.

Further Improvements

1. Introduce functionality to more effectively compare multiple gene types from the dropdown menu. A possible way to do this would be to introduce a facet grid to display multiple gene type expression plots simultaneously.
2. Speed up the cluster interactivity updates. Currently, each cell is checked one at a time to see if it belongs to the selected cluster, and the color is updated accordingly. A possible solution would be to group each cell based upon it's cluster and to update entire groups at once.

Further Applications

1. This tool could be used for additional datasets as well. This tool would have been very useful for cell type labeling when I first went through the clusters, so it could aid the process for any other datasets.

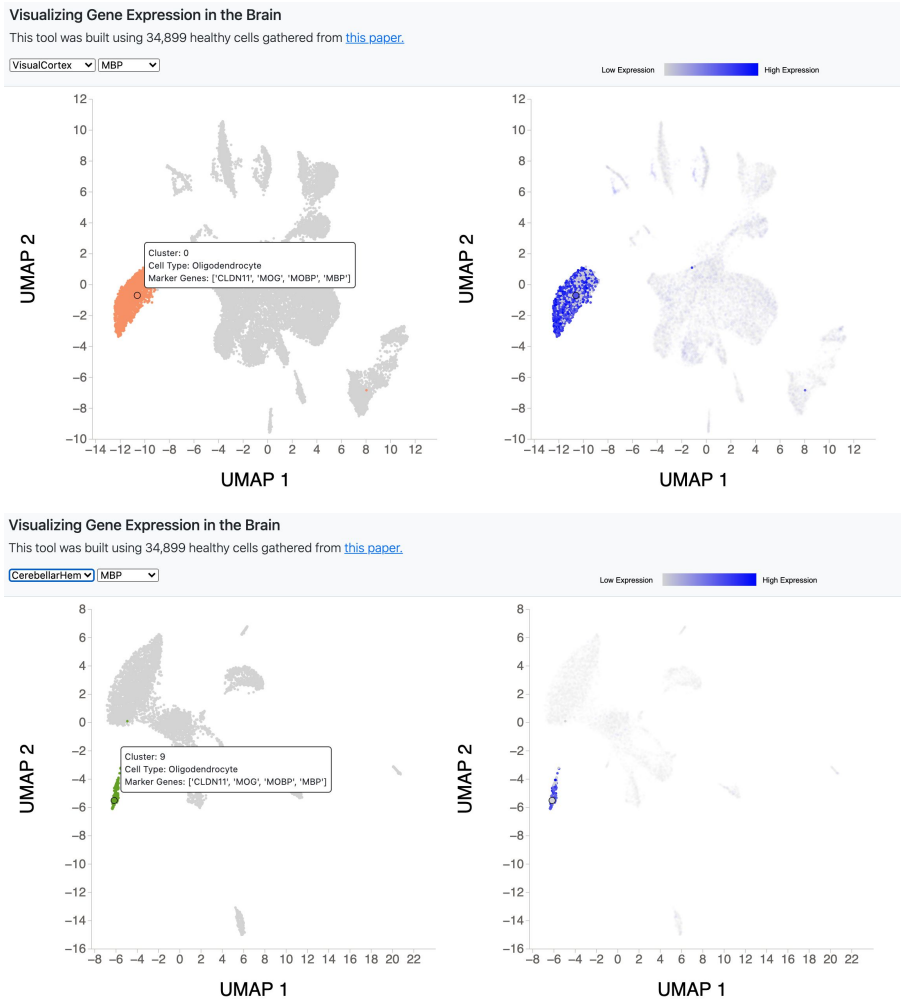

Figure 9. Differences in cell type proportions - oligodendrocytes

Evaluation

We were able to compare gene expression profiles across different cell types by selecting different genes and displaying expression on the right-hand plot. Different cells had very different expression profiles, often defining their cell type.

We can also visually compare cell type proportions based on the data from the plot. While all of the clusters in the cerebellum appeared smaller, non-neuronal cell types like oligodendrocytes seemed to make up a much smaller proportion of the data than in brain regions like the visual cortex, as shown in figure 9.

Evaluation

Some brain regions had different cell types and different marker gene expression, particularly the cerebellum compared to the other brain regions.

Figure 10a shows gene expression of SATB2, an excitatory neuron marker gene, in the frontal cortex. Figure 10b shows expression of that same gene in the cerebellum. There was little to no expression of SATB2 in the cerebellum, compared to a lot of cells in the frontal cortex expressing the gene.

This is likely because of a difference in cell types that exist in each brain region. The cerebellum has specialized excitatory neurons called granule cells. Figure 10c shows expression of RBFOX3, a granule marker gene, in the cerebellum.

We can see more expression of this specialized neuron type than we do the excitatory neurons found in the rest of the brain.

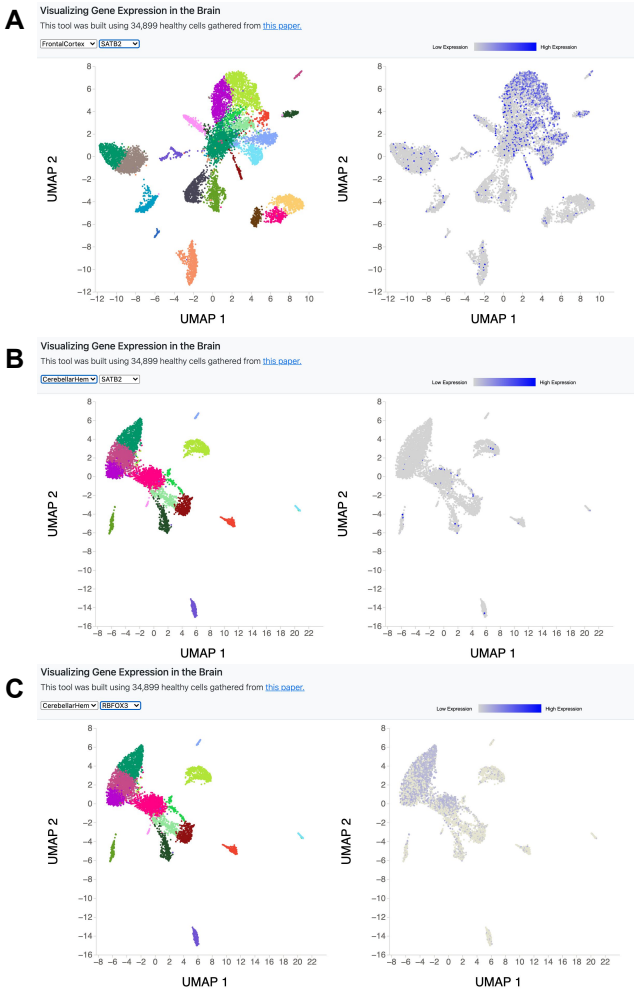

Figure 10. Differences in cell types and expression profiles

References

[1] B. B. Lake, S. Chen, B. C. Sos, J. Fan, G. E. Kaeser, Y. C. Yung, T. E. Duong, D. Gao, J. Chun, P. V. Kharchenko, and K. Zhang, “Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain,” *Nature Biotechnology*, vol. 36, no. 1, pp. 70–80, 2017.

[2] the H. E. Team, “Cerebellum Function, Anatomy & Definition | Body Maps,” Healthline, 21-Jan-2018. [Online]. Available: https://www.healthline.com/human-body-maps/cerebellum#1. [Accessed: 08-Mar-2021].

[3] M. S. Buchsbaum, “Frontal Cortex Function,” American Journal of Psychiatry, vol. 161, no. 12, pp. 2178–2178, 2004.

[4] https://www.neuroscientificallychallenged.com/blog/know-your-brain-primary-visual-cortex