Gene Expression and Subtype Analysis of Astrocytes in the Mouse Brain

James Fu
PI: Dr. Larry Zipursky

Abstract

Astrocytes, once considered a homogenous population, are now believed to differ in morphology, physiology, and molecular signatures. This diversity may be an important cellular underpinning of areal and laminar specificity in the cortex. However, a comprehensive understanding of astrocyte diversity across different cortical areas and laminae is lacking. In this study, we utilized processed whole-mouse brain MERFISH (Multiplexed Error-Robust Fluorescence in situ Hybridization) data from the Allen Institute Brain Cell Atlas to analyze astrocyte-specific gene expression and implemented machine learning algorithms to find subtypes. We conducted differential gene expression (DGE) analysis to identify astrocyte-specific genes and used K-means clustering to perform subtype analysis based on gene expression levels and cell coordinates. Our DGE analysis identified 908 out of 1122 differentially expressed genes as specific to astrocytes, with significant genes illustrated in a series of Volcano plots. Additionally, our K-means clustering analysis based on astrocyte coordinates and gene expression levels found three potential astrocyte subtypes. These findings confirm that astrocytes are associated with a large number of specific genes as well as exhibit some form of spatial organization.

Introduction

Astrocytes, a type of glial cell that tiles the brain (Sofroniew et al, 2010), possess various functions crucial for maintaining brain homeostasis. For instance, astrocytes regulate blood flow, transfer mitochondria to neurons, and supply amino acids to neurotransmitters, effectively fueling neuronal metabolism (Kim et al, 2019). Furthermore, they regulate synaptic transmission, or changes in synaptic connection strength. Synaptic transmission is the cellular basis of learning and memory, and the strength of these connections could be influenced by the expression levels of astrocyte-specific genes.

Despite their once-believed homogeneity, astrocytes across different cortical layers differ in terms of morphology, physiology, and molecular signatures. Recent research suggests multiple subtypes with differing levels of gene expression in different cortical layers (Bayraktar et al, 2020). This suggests that astrocytes might play specialized roles in specific brain regions, yet a comprehensive understanding of differences is lacking. To address this gap in knowledge, we turn to spatial transcriptomic (ST) techniques like MERFISH.

While astrocyte have been studied using traditional RNA-seq techniques, their investigation using imaging methods such as ST or MERFISH have been limited. By examining gene expression levels, we hope to identify which genes are astrocyte-specific. Additionally, by performing K-means clustering on the MERFISH data, we hope to determine how astrocyte subtypes organize themselves and if they exhibit layer specificity.

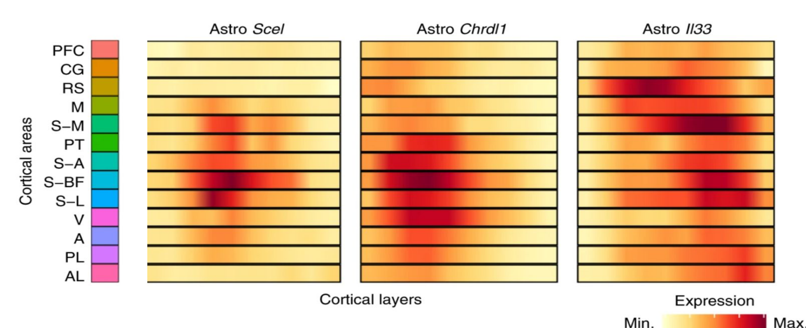

Figure 1: Variation in astrocyte-specific gene expression across different cortical layers (Bayraktar et al., 2020).

Materials

Whole Mouse Brain MERFISH Datasets were retrieved from the Allen Institute, comprising data from one female and three male mice, totaling over 10 million imaged and MERFISH-segmented cells. The data included spatially resolved gene expression data for the entire mouse brain and single-cell RNA sequencing data for all four mouse samples. Each dataset consists of processed scRNA-seq data grouped based on cell similarities, AnnData objects mapped to the cell metadata for gene expression matrices, and a Cell-Centric File (CCF) containing the spatial coordinates of the cells.

Methodology

1. Data Collection
   - **Data Preprocessing:** Split data into astrocyte vs. non-astrocyte groups, append gene data, calculate mean difference(s) in expression levels, construct aggregate profiles by averaging expression values across astrocyte and non-astrocyte cells, data normalization, compute log2 fold change
2. Identifying Astrocyte-Specific Genes
   - **Differential Gene Expression (DGE) Analysis:** Using statistical analyses, identify astrocyte-specific genes, perform Bonferroni correction, and determine if they match known astrocyte markers
   - **Volcano Plot:** Display significant genes in a Volcano plot with the most significant genes labelled
3. Investigate Astrocyte Homogeneity and Variation:
   - **K-means:** An unsupervised machine learning method used to cluster astrocytes from MERFISH data by "closeness" based on cell coordinates and gene expression profiles to determine if they vary across cortical layers
   - **Elbow Method:** Determine the optimal number of clusters for each of the mouse-brain samples

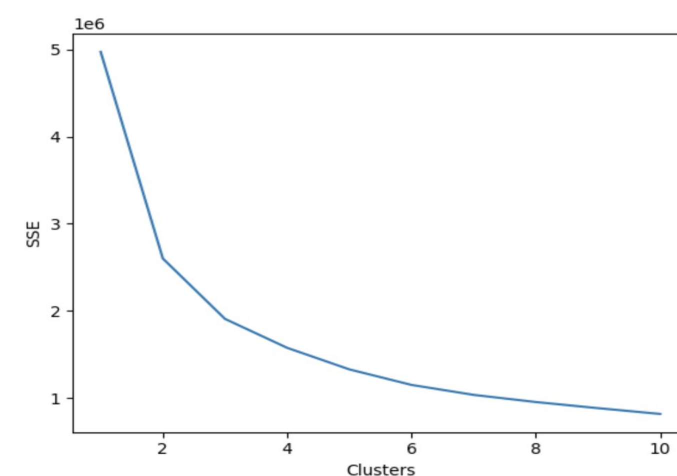

Figure 2: Elbow Method Showing the Optimal Number of Clusters for the K-means Algorithm

   - **Plot K-means Clusters:** Plot a 3D k-means clustering with coordinates (x, y, z) from clustering CCF data while considering gene expression as an additional variable
   - **Evaluate Metrics:** Iterate the K-means algorithm on the data and compare Goodness of Fit (GOF) metrics such as the Silhouette Score
4. Data Validation
   - **Compare to External Research:** Determine whether the astrocyte subtypes and quantities identified through our K-means algorithm align with outside findings from other research, such as astrocyte subtype analysis utilizing scRNA-seq data
   - **Compare to Neuron ST Data:** Since neurons are known exhibit layer specificity, compare Astrocyte cluster boundaries to neuron boundaries

Results

In our initial differential expression analysis using a volcano plot, we identified 908 astrocyte-specific genes out of 1122 total genes.

Most were passed as significant due to the large number of t-tests conducted. Applying a Bonferroni correction accounts for this by adjusting the significance threshold according to the number of tests, ensuring that identified genes are more likely to represent true differential expression. After the p-value is divided by total number of tests, only 104 genes were significant. The volcano plots in Figure 3 illustrates the number of significant genes before (left) and after (right) correction.

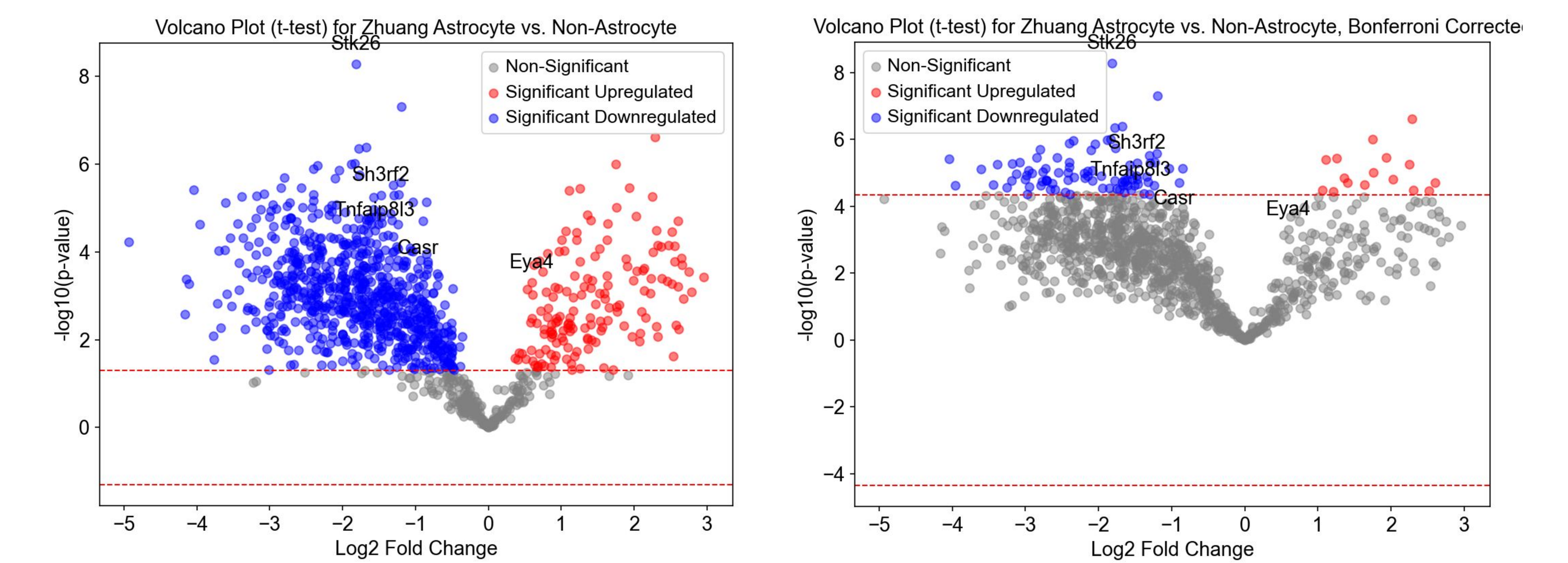

Figure 3: Volcano Plots of Astrocyte Differentially Expressed Genes

The K-means clustering of 3D coordinates and gene expression levels yielded the following separation, identifying three potential candidates for astrocytic subtypes consistently across all four samples. Below are the clustering results based on the common coordinate framework (CCF) and gene expression values for each mouse brain sample.

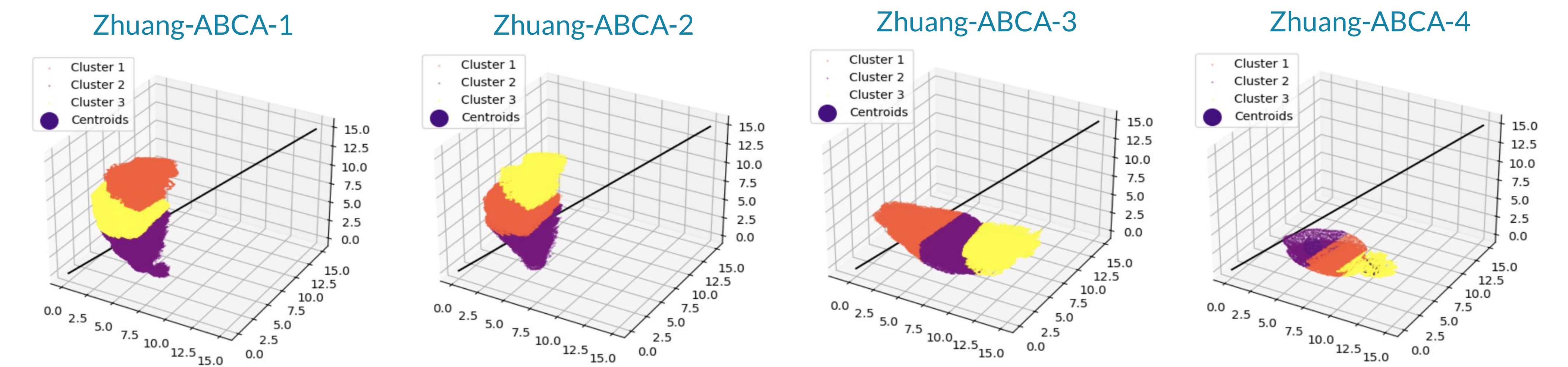

Figure 4: K-means Clustering of Whole Mouse Brain Samples Based on Coordinates and Gene Expression Levels

Conclusion

In our study, we identified a significant number of astrocyte-specific genes, with 908 out of 1122 differentially expressed genes (DEGs) showing significant expression differences between non-astrocyte and astrocyte-specific groups. Notably, the top five astrocyte-specific genes identified include EYA transcriptional coactivator and phosphatase 4, Serine/threonine kinase 26, Tumor necrosis factor/alpha-induced protein 8-like 3, Calcium-sensing receptor, and SH3 domain containing ring finger. While the most significant genes identified in our analysis differ from those reported in previous studies, several significant genes from outside studies also showed significance in our analyses, indicating a degree of overlap and validation.

Additionally, using the elbow method, we determined that for all 4 complete mouse brain samples, the number of astrocyte subtypes is best represented by three clusters. These clusters appear to be organized in distinct layers along the z-axis and align with previous findings by Zhuang et al. (2021), which suggested the existence of three potential astrocyte subtypes beyond known neuron subtypes. Our methodological approach, which involved normalization of the non-coordinate data and iterative application of K-means clustering on multiple subsets, ensured robust identification of these clusters. The elbow method was used to identify the optimal number of subtypes, while K-means clustering minimized the sum of squared distances from each point to the cluster centroids, thereby providing a precise classification of the number and location of potential astrocyte subtypes.

References

1. Sofroniew, M. V., & Vinters, H. V. (2010). Astrocytes: biology and pathology. Acta neuropathologica, 119(1), 7–35. DOI: 10.1007/s00401-009-0619-8
2. Khakh, B. S., & Sofroniew, M. V. (2015). Diversity of astrocyte functions and phenotypes in neural circuits. Nature Neuroscience, 18(7), 942–952. DOI: 10.1038/nn.4043
3. Kim, Y., Park, J., & Choi, Y. K. (2019). The Role of Astrocytes in the Central Nervous System Focused on BK Channel and Heme Oxygenase Metabolites: A Review. Antioxidants (Basel, Switzerland), 8(5), 121. DOI: 10.3390/antiox8050121
4. Zhang, M., Eichhorn, S. W., Zingg, B., et al. (2021). Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. Nature, 598, 137–143. DOI: 10.1038/s41586-021-03705-x
5. Batiuk, M. Y., Martirosyan, A., Wahis, J., et al. (2020). Identification of region-specific astrocyte subtypes at single-cell resolution. Nature Communications, 11, 1220. DOI: 10.1038/s41467-019-14198-8
6. Matias, I., Morgado, J., & Gomes, F. C. A. (2019). Astrocyte heterogeneity: impact to brain aging and disease. Frontiers in Aging Neuroscience, 11, 59.
7. Bayraktar, O. A., Bartels, T., Holmqvist, S., et al. (2020). Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. Nature Neuroscience, 23(4), 500–509. DOI: 10.1038/s41593-020-0602-1
8. Yao, Z., van Velthoven, C. T. J., Kunst, M., et al. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature, 624, 317–332. DOI: 10.1038/s41586-023-06812-z

Acknowledgements

I would like to acknowledge Dr. Fangming Xie for helping me narrow down my research topics, troubleshooting various code issues, and providing continuous feedback throughout the past year.