WPI

Curating a Pipeline for Analyzing and Visualizing Gene Expression Data in Psychiatric Disorders

A Major Qualifying Project submitted to the Faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfilment of the requirements for the degree of Bachelor of Science

Nicole Amanda Shedd

Submitted to:
Dr. Elizabeth Ryder, Advisor
Dr. Inna Nechipurenko, Advisor

Project Sponsors:
Dr. Zhiping Weng and Henry Pratt, University of Massachusetts Medical School

May 6, 2021

Abstract

Our lab is interested in understanding changes to gene expression and regulation in the brain. In recent years, thousands of single-cell RNA-seq and ATAC-seq datasets have been produced and become publicly available. To analyze such datasets, we compiled a single-cell sequencing analysis and visualization pipeline based on dimensionality reduction and clustering techniques. We then used our pipeline to explore cell-type specific changes to gene expression in individuals with Autism Spectrum Disorder. This analysis indicated that neuroinflammation, mitochondrial dysfunction, and oxidative stress are involved in the etiology of Autism Spectrum Disorder.

Acknowledgements

I would like to thank the following individuals for their assistance in helping us present a successful project:

- **Elizabeth Ryder and Inna Nechipurenko**, my advisors, for their contributions to my project, report, and presentations, and their insight into the field and project
- **Henry Pratt**, for teaching me about gene expression analysis and assisting me throughout the project
- **Zhiping Weng** for giving me this opportunity and giving support and assistance whenever possible
- **Jill Moore, Thomas Reimmon, and the Weng Lab**, for giving advice and assistance throughout the project.
- **Worcester Polytechnic Institute and University of Massachusetts Medical School** for giving us the opportunity to take part in this experience

Table of Contents

Abstract i
Acknowledgements ii
1. Introduction 1
1.1 Genetics of Psychiatric Disorders 1
1.2 Transcription regulation 3
1.3 Transcriptomics and Epigenomics 4
1.4 ENCODE 5
1.5 Single-cell data analysis 5
2. Methods 8
2.1 Developing a single-cell RNA-seq analysis pipeline 8
2.2 Analysis of scRNA-seq data with psychiatric disorders 10
2.3 Developing a single-cell ATAC-seq analysis pipeline 10
2.4 scATAC-seq and scRNA-seq integration 11
3. Results 12
3.1 Benchmarking the scRNA-seq pipeline 12
3.2 Using the scRNA-seq pipeline to understand Autism Spectrum Disorder 15
3.3 Benchmarking the scATAC-seq pipeline 27
3.4 Integrating RNA-seq and ATAC-seq data 31
4. Discussion 33
Works Cited 35

Table of Figures

Figure 1: Regions on chromosome 15 implicated in ASD....2
Figure 2: Diagram of 3 regulatory site sequencing techniques. 4
Figure 3:MNIST digits embedded via t-SNE and UMAP....6
Figure 4: scRNA-seq processing pipeline 10
Figure 5: UMAP of Frontal Cortex cells plotted using Seurat....12
Figure 6: UMAP with doublets removed, as predicted by DoubletFinder....13
Figure 7: Dotplot showing expression of genes in a cluster....14
Figure 8: scRNA-seq Frontal Cortex UMAP relabeled with cell types 15
Figure 9: nFeature RNA violin plot of all samples in CTL group 16
Figure 10: Harmony batch mixing for both Control and ASD subsets 17
Figure 11: UMAP and clustering after Harmony 18
Figure 12: UMAP plots of doublets and doublet removal 19
Figure 13: Marker gene dotplot and SingleR cll type predictions for cell type labeling 20
Figure 14: UMAP plots with consensus cell type labels....21
Figure 15: UMAP of Brodmann Areas 4/6 (a) and 9 (b) combining CTL and ASD samples 22
Figure 16: UMAPs split by group, colored by cluster and cell type 23
Figure 17: Cell type proportion comparison between ASD and CTL samples....24
Figure 18: Cluster proportion comparison between ASD and CTL samples....25
Figure 19: DEGs in each of the 6 cell types identified by Seurat for Brodmann Area 4-6....26
Figure 20: UMAP of Lake et al., 2018 THS-seq data, created in ArchR....27
Figure 21: Heatmaps to visualize regulatory element enrichment in each cluster 28
Figure 22: UCSC Genome Browser Image of Regulatory Element EH38E1311335....29
Figure 23: Cell type labels of each THS-seq cluster 30
Figure 24: Benchmarking ATAC-seq analysis pipeline....31
Figure 25: scTHS-seq data labeled by integrating scRNA-seq data 32

1. Introduction

Psychiatric disorders affect millions of people worldwide, including 1 in 54 children with Autism Spectrum Disorder (Maenner et al., 2016). Symptoms of psychiatric disorders can severely impact a patient's quality of life, ranging from psychosis to a severe decline in memory and communication.

We are still trying to understand the effect of mutations to noncoding DNA and their implications in psychiatric disorders. These mutations result in changes to gene expression that are often presented in cell type-specific contexts. Genome-wide association studies have presented evidence that mutations to non-coding DNA are correlated with incidences of psychiatric disorders. However, we are still trying to understand how those mutations affect cell states.

This project aims to apply single cell techniques to map epigenetic and transcriptional states of cells in the brain and generate visualizations comparing the cell states of healthy individuals and individuals with psychiatric disorders.

1.1 Genetics of Psychiatric Disorders

Many psychiatric disorders are rooted in genetic variants. Autism Spectrum Disorder (ASD) is a syndrome with both genetic and nongenetic causes. The recurrence rate in siblings is 2% to 8%, which is higher than the rate in the general population, but lower than the recurrence rate for single-gene disorders (Gillberg and Coleman, 2000). In addition, a study in the UK with 47 twin pairs showed that the concordance of ASD in monozygotic twin pairs was nearly 60%, while the concordance in dizygotic twin pairs was 0% (Bailey et al., 1995).

A number of studies have tried to identify gene loci and other genetic abnormalities associated with ASD. Less than 5% of ASD cases are associated with chromosomal abnormalities, most of these abnormalities being duplications of the Prader–Willi/Angelman syndrome region, denoted 15q11–13. Figure 1 diagrams abnormalities on chromosome 15 that have been identified in individuals with Autism. These include duplications, deletions, and inversions. (Folstein and Rosen-Sheidly, 2001). Point mutations in FOXP2 on chromosome 15 have been implicated in severe speech and language disorders (Hurst et al., 1990).

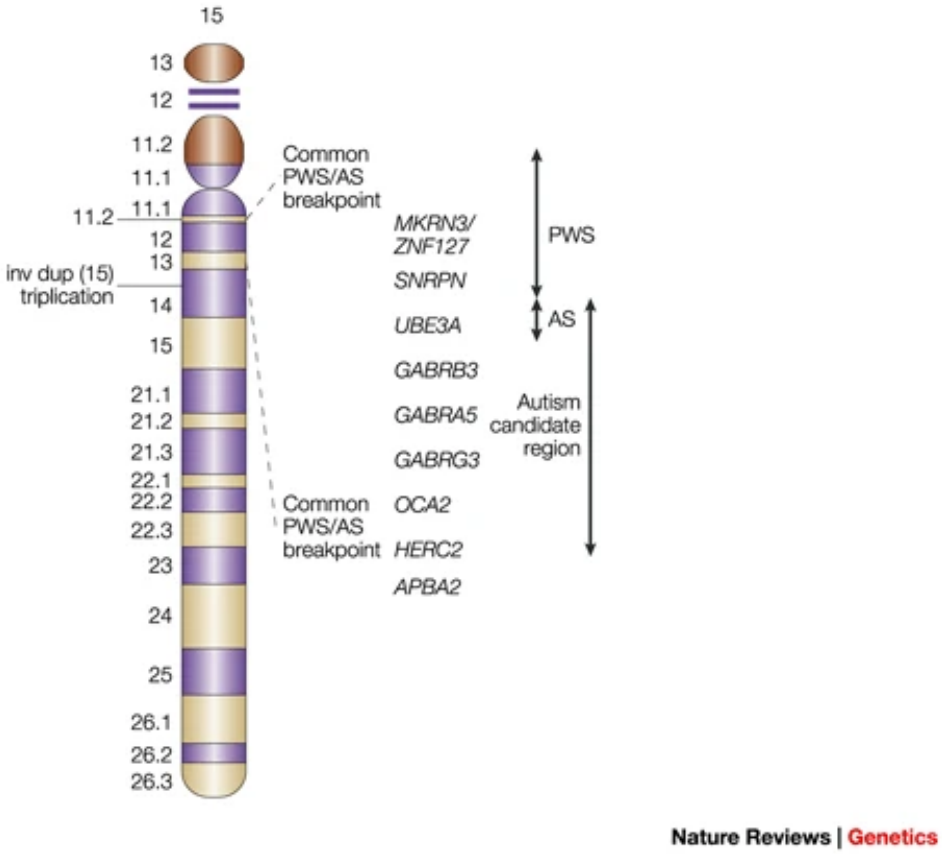

Figure 1: Regions on chromosome 15 implicated in ASD, from Folstein and Rosen-Sheidly, 2001

Another highly replicated locus in ASD is AUTS1 on 7q31-q33. In one family, three siblings inherited this locus from their mother. Two of the siblings are diagnosed with ASD, while the other has an expressive language disorder. Linkage analysis of another 76 families yielded LOD scores between 1 and 2, indicating the presence of an ASD locus on chromosome 7q (Ashley-Koch et al., 1999).

An estimated 90% of individuals with autism have atypical sensory experiences (Marco et al., 2011). In particular, auditory and visual functions often differ from controls. Individuals with Autism often present a delayed neural response to both pure sounds and complex speech (Roberts et al., 2010). However, they also tend to display superior pitch discrimination and categorization (Bonnel et al., 2003). Some studies show that visual processing in Autistic individuals is more detail-oriented (Happé and Frith, 2006). In addition, individuals with ASD seem to have impaired global motion perception compared to healthy, non-ASD controls, but superior local motion perception (Robertson et al., 2012; Chen et al., 2012).

One study used bulk RNA-seq analysis to determine that Brodmann Area (BA) 17, the primary visual cortex, was most broadly dysregulated in Autistic individuals, followed by BA41-42-22 and BA7, which are the primary auditory cortex and visuo-motor coordination cortex, respectively. In BA17, they found 3264 genes and 1170 isoforms dysregulated in Autistic donors. The downregulated genes showed broad enrichment for neuronal cell-type specific

genes, while the upregulated genes were enriched for oligodendrocyte precursor cell and astrocyte markers. In addition, there seemed to be an increased proportion of excitatory neurons, but decreased proportions of particular excitatory and inhibitory neuron subclusters, indicating a change in cell state (Haney et al., 2020).

In addition to speech and sensory impairments, motor impairments are also pervasive in children with Autism. A 2020 study surveyed more than 11,000 parents of children with Autism. The proportion of children with motor impairment was 86.9%, which typically continues into adolescence (Bhat, 2020). In a study of 8 Autism patients and 8 controls during visuomotor learning, Brodmann areas 4 and 6 became more involved during late learning stages compared with early stages, where this effect was not seen in the control group (Müller et al., 2004). In another study of the motor cortex, researchers noted an imbalance of excitatory and inhibitory neurons, in particular reduced GABAergic function (Masuda et al., 2019). However, there has been limited research on the genetic changes in the motor cortex in ASD.

1.2 Transcription regulation

While genetics can be significant in disease phenotypes, chemical modifications of the genome, called epigenetics, can also play a role in diseases. These modifications include DNA methylation and histone modification (Holliday and Pugh, 1975; Murray, 1964). These epigenetic modifications are subject to environmental factors. They cause changes to transcription and protein synthesis, which may explain epigenetic components of psychiatric disorders that cannot be explained by differences in the DNA sequence.

Transcription is, in part, regulated by chromatin packaging. Histones are molecules that wrap and "package" DNA into a chromatin fiber and, ultimately into a chromosome. For gene transcription to occur, the DNA must be unpacked by the molecules involved in the transcription process (Alberts et al., 2002). Genetic disorders can also be a result of errors in chromatin remodeling. To create chromosomes, DNA wraps histones, RNA, and other DNA-associated proteins, forming a complex called chromatin. Large, multiprotein complexes called chromatin remodelers restructure nucleosomes, a strand of DNA wrapped around eight histone proteins. These chromatin remodeling complexes and other chromatin-modifying proteins heavily influence the transcription levels of a gene (Phillips and Shaw 2008). Mutations in the genes that code for some chromatin remodelers' classes are implicated in psychiatric disorders, such as autism (McCarthy et al., 2014). One example is reduced expression of MeCP2 commonly found in autism, resulting from aberrant promoter methylation (Nagarajan et al., 2006).

The differential expression causing these disorders can also result from mutations in non-coding DNA, often in enhancer regions. These enhancers interact with promoter regions of DNA to enhance transcription. Identifying enhancer regions can be much more complicated than identifying promoters. Enhancer regions can be distant to the gene, regulate multiple genes, and are only active in specific cells, timepoints, or other conditions (Pennacchio et al., 2007). Researchers have looked at enhancer regions and other regulatory elements in brain and immune

cells and their relation to psychiatric disorders. Single-cell techniques to map epigenetic states of cells in the brain could help understand cell-type specificity of disease-associated regulatory element activity (PsychENCODE).

1.3 Transcriptomics and Epigenomics

Researchers use several different approaches to identify regulatory elements in the enhancer regions, including ChIP-sequencing, ATAC-sequencing, and DNase-sequencing. ChIP-sequencing, or ChIP-seq, analyzes protein interactions with DNA. It can locate DNA-binding sites for transcription factors and proteins. Because enhancers often bind transcription factors, this method can identify regulatory elements within enhancers (Park 2009). DNase-seq is used to map DNase hypersensitive sites and identify the cell's most active regulatory sites (Song and Crawford 2010). Assay for Transposase Accessible Chromatin sequencing, or ATAC-seq, probes DNA accessibility and maps transcription factor binding regions in accessible chromatin. ATAC-seq can be an efficient alternative to DNase-seq (Buenrostro et al., 2015). THS-sequencing is an analogue of ATAC-seq but used less frequently. It is typically better than ATAC-seq at finding small regulatory regions near distal enhancers (Sos et al., 2016). Researchers commonly use all of these techniques to measure the activation of regulatory elements. Figure 1 outlines the process behind each method and how their results may differ.

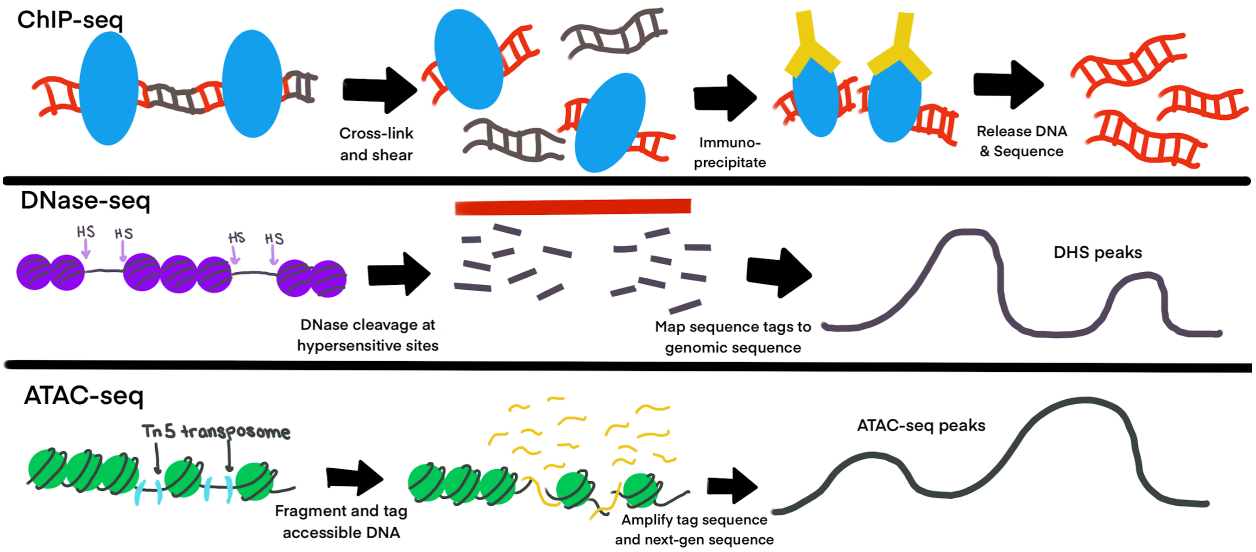

Figure 2: Diagram of 3 regulatory site sequencing techniques. ChIP-seq uses chromatin immunoprecipitation to obtain DNA sequences that are able to interact with the immunoprecipitated proteins. DNase-seq extracts the chromatin and digests it with DNase. ATAC-seq is similar to DNase-seq, using a hyperactive transposase to fragment DNA

DNAase I hypersensitive sites (DHS) can help understand various cis-regulatory elements. Scientists have recently improved the DNase-seq assay, which has resulted in a more comprehensive set of available DNase-seq data, (Meuleman et al., 2020) including a developmental progression in mice across several cell types and a more extensive catalog of DHS in humans. These DHSs are mapped in more than 200 cell types and states, and in developing mice (ENCODE Project Consortium et al., 2020). These data allow scientists to

closely study gene regulation, especially variants in disease or complex traits (Breeze et al., 2020).

Researchers estimate that there are upwards of 1.3 million regulatory elements, such as DHS, in the human genome, with only a small portion active in a given cell type. Changes in chromosome accessibility around these elements or mutations to some of these elements' genetic code can impact gene expression in only the cell types where that site is typically active (ENCODE).

Scientists have developed multiple methods to map transcripts and non-coding regions to track mutations in regulatory elements. One of these is an approach called RAMPAGE, or RNA Annotation and Mapping of Promoters for Analysis in Gene Expression. It can locate transcriptional start sites, measure promoter-specific RNA expression, and connect the transcription start sites and their specific genes. These annotations can improve gene and transcript analysis and reveal connections between cell type and phenotype (Batut et al., 2013).

1.4 ENCODE

ENCODE, or Encyclopedia of DNA elements aims to define and annotate human and mouse genomes' functional elements. The data in the encyclopedia maps the RNA transcripts of each cell type, recognition sites for RNA binding proteins, co-occupancy patterns of human transcription factors, DNase I hypersensitive sites, 3D chromatin interactions, and early developmental landscapes in mice that are not accessible in humans (Breschi et al., nd; Van Nostrand et al., 2020; Meuleman et al., 2020; Grubert et al., 2020; Breeze et al., 2020). Because regulatory sites are often very unique to our species, the encyclopedia focuses on human regulatory elements. However, it also has a collection of regulatory elements related to mice across different cells at different time points in development, where sequencing data is not readily available in humans.

The PsychENCODE consortium is a collaboration between research institutes to investigate regulatory elements specifically in individuals with neuropsychiatric disorders. The consortium researchers aim to use transcriptomic, epigenomic, and genomic data from developing, healthy, and diseased brains to understand psychiatric development and disease. Some of the psychiatric disorders being studied include autism spectrum disorder, schizophrenia, bipolar disorder, and Alzheimer's Disease (PsychENCODE, 2018). Members of PsychENCODE, as well as other researchers, have generated RNA-seq and ATAC-seq datasets from both the healthy brain and psychiatric disorders. We can utilize these datasets in this project.

1.5 Single-cell data analysis

A variety of machine learning techniques are used to analyze regulatory element data. Analysis of single-cell data across multiple regulatory elements can be done through dimensionality reduction, such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation Projection (UMAP).

Dimensionality reduction is defined as transforming data from a high-dimensional space to a lower-dimensional space (Vlachos, 2011). These techniques allow thousands of measurements to be displayed on a two-dimensional plane.

T-SNE is a powerful method for non-linear datasets, but it does not preserve the global structure, showing the size of clusters and space between them. It can find differentially expressed genes and create clusters, but the clusters' placement does not display any information about the dataset. UMAP is generally a faster method than t-SNE. It is also more successful at maintaining the global structure of the dataset than t-SNE. With UMAP, the comparative size of and distance between clusters is meaningful. Clusters with more similarities will be displayed closer together than very different ones.

One dataset that can compare the computational powers of t-SNE and UMAP is the MNIST digit dataset. This contains about 1500 samples of handwritten digits 0 to 9. Each "digit" is a 28 x 28 square of pixels. Applying dimensionality reduction techniques to reduce each set of pixels to a two-dimensional point for each sample will create a plot with each digit's clusters, as shown in figure 2. The goal is to visualize the differences between numbers and potentially predict the identity of additional digits.

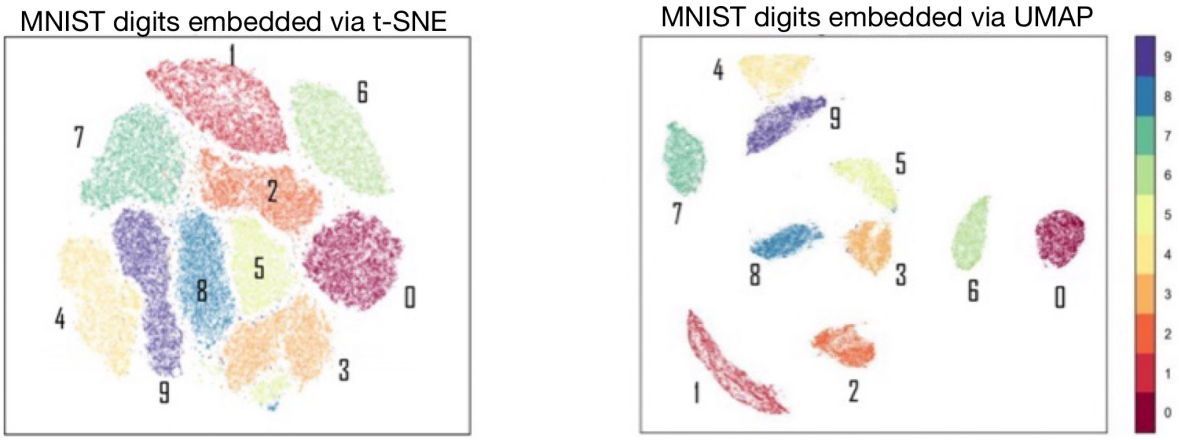

Figure 3:MNIST digits embedded via t-SNE and UMAP (McInnes et al., ar Xiv, 20)

Both of these models created defined clusters from the dataset, but UMAP could preserve the global structure to a greater degree. In the UMAP plot, the distance between the clusters is more meaningful. There is less distance between 0 and 6 than 0 and 2, while in the t-SNE representation, they both appear to be the same distance away. This is because the clusters were determined by the shape of the digits, and the 6 bears more similarities to 0 than 2 does. While mathematically they are very similar, UMAP makes some changes to the calculations that improve computational efficiency. It also uses nearest neighbor calculations instead of perplexity to determine the standard deviation of samples and determine the location of points on the graph. UMAP also removes the random initializations that t-SNE uses, which makes UMAP plots more similar from run to run than t-SNE. This is analogous to how UMAP would work on a set of

single-cell RNA-seq data or ATAC-seq data: it would reduce dimensions across the readings for each gene to create a point for each cell where clusters of points represent different cell types or cell states.

After identifying clustering in an RNA-seq dataset, several possible methods can determine which genes are being differentially expressed. This can be done using Wilcoxon pairwise tests, t-tests, or binomial distributions to understand which genes are under or over expressed in a cluster. This process can also be used in ATAC-seq or DNAse-seq datasets, identifying over-enriched or under-enriched regulatory regions.

These machine learning and statistical techniques can be used to analyze single-cell RNA-seq and ATAC-seq data in psychiatric disorders. This could help identify differential expression between healthy controls and individuals with psychiatric disorders. This research could help indicate which genes are overexpressed and which regulatory elements are more active in patients affected by psychiatric disorders and answer critical biological questions about their genetic and epigenetic causes. We can also visualize the expression of genes in patients with psychiatric disorders and controls. We want to determine if there is an imbalance of cell types, such as a higher proportion of a particular cell type in the disorder dataset. We might also see unique cell states in these disease datasets, such as a subset of excitatory neurons that have unique expression profiles not seen in the control cells.

2. Methods

The goal of this project was to develop a single-cell sequencing analysis pipeline for RNA-seq and ATAC-seq data. I then applied this pipeline to existing datasets to evaluate changes to cell composition and gene expression in brains with psychiatric disorders.

I primarily used two datasets for the development and application of these pipelines. The first was from Lake et al., 2019. This is single nucleus snDrop-seq, a variation of RNA-seq, and single cell THS-seq, a variation of ATAC-seq data. Drop-seq adds cells and beads with unique oligos to droplets, which creates barcoded mRNA from single cells (Macosko et al., 2015). 10X Chromium is a more commonly used sequencing platform in ENCODE. The structure of the beads in 10X Chromium experiments allow for higher bead occupancy, which typically results in higher quality data (Zhang et al., 2019). Drop-seq is a more cost-efficient protocol for large numbers of cells than other RNA-seq protocols (Ziegenhain et al., 2017). THS-seq uses Tn5 transposase to insert a read primer and promoter into open chromatin regions, while ATAC-seq uses a Tn5 transposase of fragment and primer open chromatin regions (Sos et al., 2016, Buenrostro et al., 2016). THS-seq is better at identifying short regions near distal enhancers than ATAC-seq (Sos et al., 2016). They obtained both RNA-seq and ATAC-seq data from each cell, measuring gene expression and regulatory element activation in three brain regions from healthy brains: frontal cortex, visual cortex, and cerebellum. This paper focuses on the data from the frontal cortex, due to its involvement in psychiatric disorders.

The second dataset is from an ASD study preprint (Haney et al., 2020). This study took single-nucleus RNA-seq data from two frontal cortex brain regions in both autistic and neurotypical patients. These regions were BA 4/6, which is the primary and supplementary motor cortex, and BA 9, which is the dorsolateral prefrontal cortex, involved in motor planning and organization.

2.1 Developing a single-cell RNA-seq analysis pipeline

The first step to analyzing single-cell sequencing data was to preprocess the data from fasta format to raw count matrices. We chose to do this in STARsolo (Dobin et al., 2013). For every dataset, we used a GRChr38 reference genome from PsychENCODE for genome assembly. This particular reference genome masks repetitive genome regions to reduce multimappers to the area. I adjusted a few parameters to accommodate the data. For each sample, I adjusted the barcode read length depending on the version of the 10X genomics sequencing protocol used. The barcode of samples sequenced with 10X V2, or version 2, has a length of 26, while 10X V3 has barcodes with length 28. We also mapped “GeneFull” features instead of “Gene” to count all reads that overlap with gene loci. This is better for mapping single nucleus data, allowing STARsolo to map reads outside the typical transcriptome. In some of the samples, I also had to increase the RAM limit for sorting the BAMs, which are binary files for storing sequence data, after mapping to finish sorting.

We chose to use Seurat (Stuart et al., 2019) for the majority of our single cell analysis. Seurat can take in the files produced from STARsolo and store them as a Seurat object. We can then use Seurat to normalize and scale the data, perform dimensionality reduction using PCA and UMAP, cluster, and plot the analyzed data. Generally, I followed the standard pre-processing workflow described in the Seurat documentation. One round of dimensionality reduction is done with PCA. Then, we take the first 20 principal components to perform clustering and UMAP. Clustering uses a k-means clustering method based on gene expression in the cells. This divides cells into clusters based on having a similar average expression of each gene. The UMAP algorithm then reduces the gene expression of each cell onto a 2-dimensional embedding. The k-means clustering and location of cells on the UMAP plot typically correspond well.

I also added a step to the analysis pipeline to address and manage batch effect. Batch effect is where the sequencing library or method affects mapping and clustering. Some datasets, especially studies of psychiatric disorders, were compiled over long periods of time and can use multiple sequencing technologies across all of the samples. This can cause individual samples to form their own clusters. There are many effective batch-effect integration tools, but I chose to use Harmony (Korsunsky et al., 2019). When benchmarked against other similar methods, Harmony performed consistently well and had much faster runtimes than any methods performing similarly well (Tran et al., 2020).

Detecting and removing doublets can also be an important component of single-cell sequencing analysis. Doublets are an error where two cells are assigned the same barcode, and they appear in the plot as a single cell with a combined expression profile. These can make cell - type labelling more difficult or skew the number of cells in a particular cell type. Another tool that I added to this pipeline was DoubletFinder (McGinnis et al., 2019). I chose DoubletFinder because it has built-in compatibility with Seurat objects and ranked highly in a benchmark of nine doublet-detection methods (Xi and Li, 2021).

Another step to the analysis pipeline is determining the cell type of the clusters that Seurat creates. This is important for determining the proportion of different cell types in a dataset and for attempting to determine differences in gene expression. This can be done manually, by using expression of marker genes to determine which clusters have cells of which cell type. When using this, we use a list of marker genes compiled by the Weng lab from literature to create dot plots and compare expression of those genes across clusters. However, this process can be time-consuming. I explored packages that could be used to automate this process. I implemented SingleR (Aran et al., 2019), which can take a pre-labeled reference dataset and use those gene expression profiles to label a test dataset.

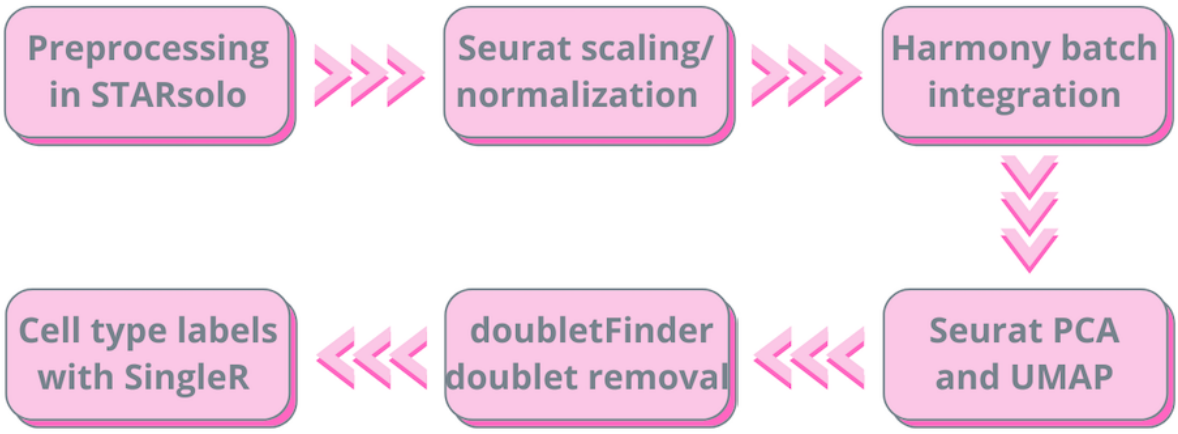

Figure 4: scRNA-seq processing pipeline

2.2 Analysis of scRNA-seq data with psychiatric disorders

We can also use Seurat to directly compare control and disease samples. This can be done by merging the two groups; in this paper (Haney et al., 2020) that is control (CTL) and Autistic (ASD). The preprocessing and analysis are the same as processing separate samples. When plotting the UMAP, we can split the samples by group. The embeddings will remain the same, but the CTL and ASD samples will be broken up into 2 separate plots.

After labeling the clusters based on a reference dataset or marker gene expression, we can also directly compare the proportion of different cell types. There are several interesting ways to break down the data. We can look at the proportion of cells in each cluster to understand the proportion of cells in different cell states. We can also aggregate the data into each of the 7 cell types most commonly found in the brain to examine the proportions by cell type. Previous studies have also found significant differences between the proportions of neurons and non-neuronal cells in Autism Spectrum Disorder.

Lastly, we want to compare gene expression between each cluster on the CTL plot versus the ASD plot. This is a feature in Seurat. There are some risks to performing differentially expressed gene (DEG) analysis on data without accounting for batch effect, but technical limitations make it impossible to perform the analysis on batch effect-corrected gene expression matrices. Harmony, while it creates adjusted embeddings to plot a mixed UMAP, does not generate an adjusted gene expression matrix. Seurat's integration tool does adjust gene expression to correct for batch effect, but the recommended pipeline does not use the corrected matrix for DEG analysis.

2.3 Developing a single-cell ATAC-seq analysis pipeline

We use known representative DNase hypersensitivity sites, or rDHS, peaks to predict enhancers and improve the analysis in ArchR. We used a list of rDHS sites from PsychENCODE, where they used iterative clustering selection to create a consensus peak set based on multiple samples (ENCODE project Consortium).

For single-cell ATAC-seq datasets, I typically used ArchR for analysis (Granja et al., 2021). While Seurat does have support for ATAC-seq analysis, ATAC-seq datasets are typically very large and sparse, containing mostly zeroes, since only a small portion of regulatory elements are active in a cell. ArchR is better equipped to manage this size dataset.

ArchR needs BAM files as inputs, as opposed to count matrices. This typically requires less intensive preprocessing, just some to get the cell barcodes in the correct format. ArchR also uses Seurat's FindNeighbors() and FindClusters() functions, so the processing workflow is very similar.

Similarly to scRNA-seq data, scATAC-seq data are susceptible to doublets. ArchR has built-in doublet identification that we can use to filter out doublets. This works by creating artificial doublets by combining reads from different cells and determining where they are plotted. The program can then remove doublets from the actual dataset that are located near those artificial doublets. ArchR also corrects batch effect with Harmony, the tool I chose for scRNA-seq analysis.

Cell type labeling is a more difficult process with ATAC-seq data than RNA-seq, especially in the brain. Very well documented cell types, like peripheral blood mononucleate cells, or PBMCs, have lists of marker regulatory elements that we could use for cell-type identification. However, there aren't a lot of known marker regulatory elements for cells in the brain. ArchR does have the ability to calculate gene scores based on regulatory enrichment in the cells in a cluster. This takes into account the accessibility of a gene, distance between the element and a gene, and gene boundaries to link an over-enriched regulatory element to a gene. High gene scores indicate that the gene is likely to be regulated by the region. We could then use this gene list to determine cell type identity of a cluster.

2.4 scATAC-seq and scRNA-seq integration

We can also use integration of ATAC-seq and RNA-seq datasets to label the cell types of clusters. Especially as RNA and ATAC sequencing from the same samples is becoming more common, this can be an effective method. ArchR can take a labeled Seurat object and compare its actual gene expression profiles with the predicted gene scores from the ATAC-seq data to predict the cell type of a cell or cluster.

This integration can also be used in exploratory analyses of gene expression regulation. It can help discover new enhancer-gene interactions that play roles in cell-type differentiation and psychiatric disorders. We know more about genes than regulatory elements.

3. Results

3.1 Benchmarking the scRNA-seq pipeline

Before implementing the scRNA-seq pipeline shown in Figure 4 to draw conclusions about new datasets, we wanted to validate it by applying it to a dataset where cells were already labeled with cell types. We used the single-nucleus RNA-seq data from Lake et al. 2018. They had data from the frontal cortex, visual cortex, and cerebellum, but I primarily focused on the frontal cortex, due to its relevance to psychiatric disorders.

With this dataset, we didn't use batch effect correction because we weren't combining different samples. Figure 5 below shows the plot generated from the Seurat processing pipeline using the default parameters and Euclidean distance for the UMAP.

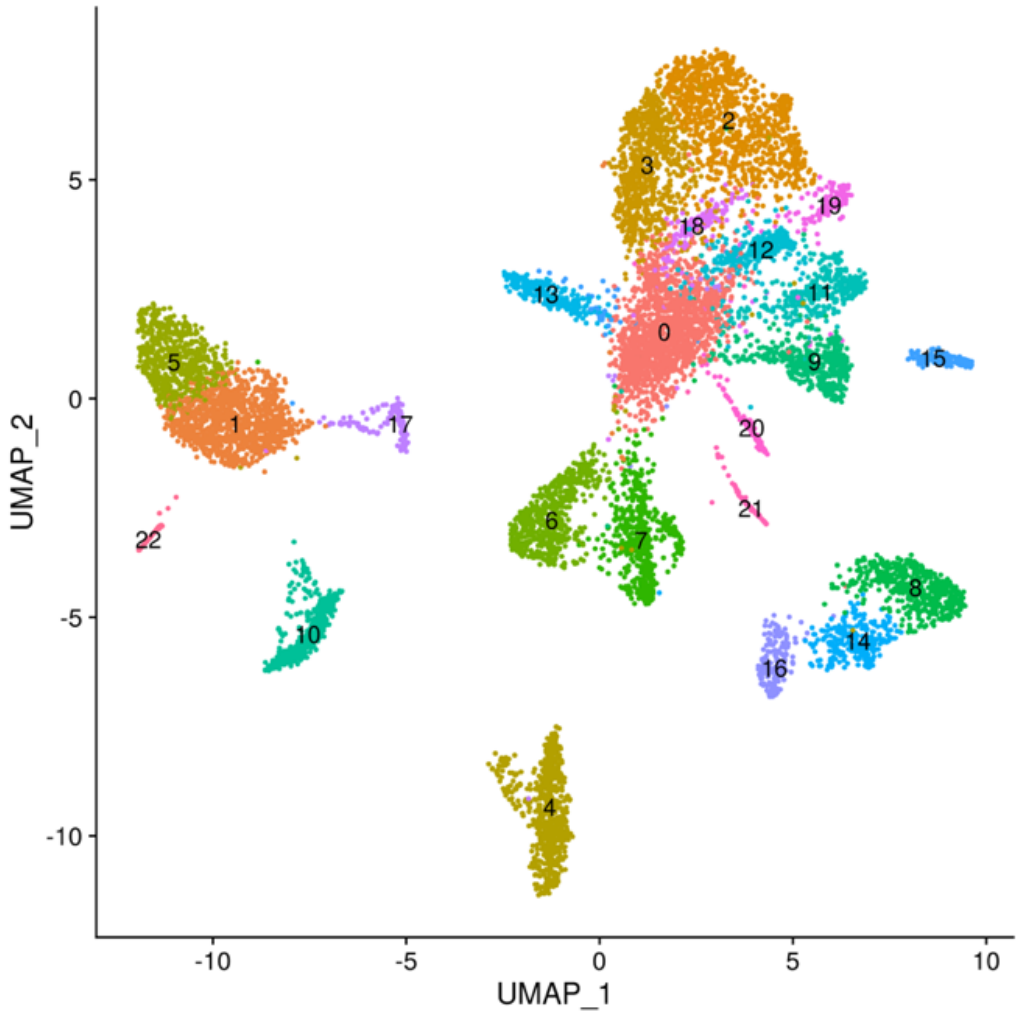

Figure 5: UMAP of Frontal Cortex cells plotted using Seurat. Colors indicate the clusters assigned to each cell type by Seurat using a k-means clustering algorithm, described in section 2.1. Colors and numbers are assigned by cluster size, smallest to largest. Plotted using data from Lake et al., 2018

We predicted that some of these cells were actually doublets, and wanted to identify and remove them, shown in Figure 6. We can see that the doublets are typically the “stringy” formations moving away from a more clearly shaped cluster of cells. For example, cluster 4 (Figure 5) appears to have some cells on the left side of the cluster that are doublets. These doublets likely combined the reads of a cell in cluster 4 and cells in clusters 1, 5, or 10, which is why they appear along the path between the 2 clusters. In some cases, entire clusters of cells appear to be doublets (e.g., clusters 20 and 21, Fig. 5). These could be removed later for a cleaner graph.

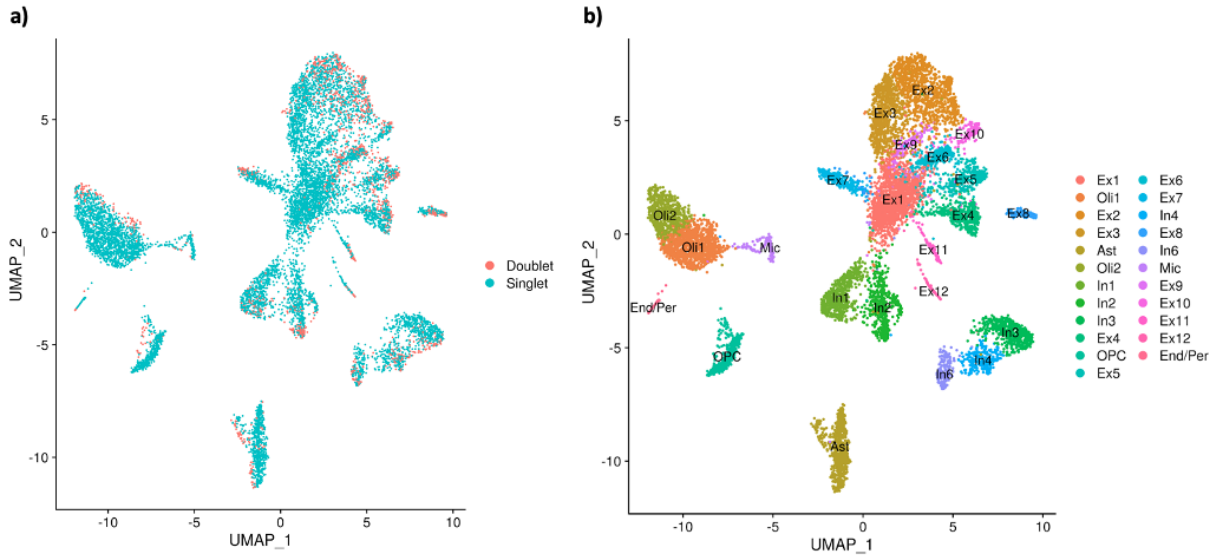

Figure 6: UMAP with doublets removed, as predicted by DoubletFinder. Panel a shows UMAP from figure 5, with predicted doublets colored red and predicted singlets colored blue. Panel b is the UMAP plot with predicted doublets removed. Plotted using data from Lake et al., 2018.

Instead of using the tool for automatic cell type labelling, we labelled the cell types manually based on our own analysis. We could then use this as a reference dataset for labelling others, and still confirm that the marker gene method was effective at labelling clusters.

Figure 7 shows a dot plot representing the expression of different marker genes within a cluster. This plot indicates that there were 2 oligodendrocyte clusters, 11 excitatory neuron clusters, 1 astrocyte cluster, 5 inhibitory neuron clusters, 1 oligodendrocyte precursor cluster, 1 microglia cluster, and 1 combined endothelial cell and pericyte cluster.

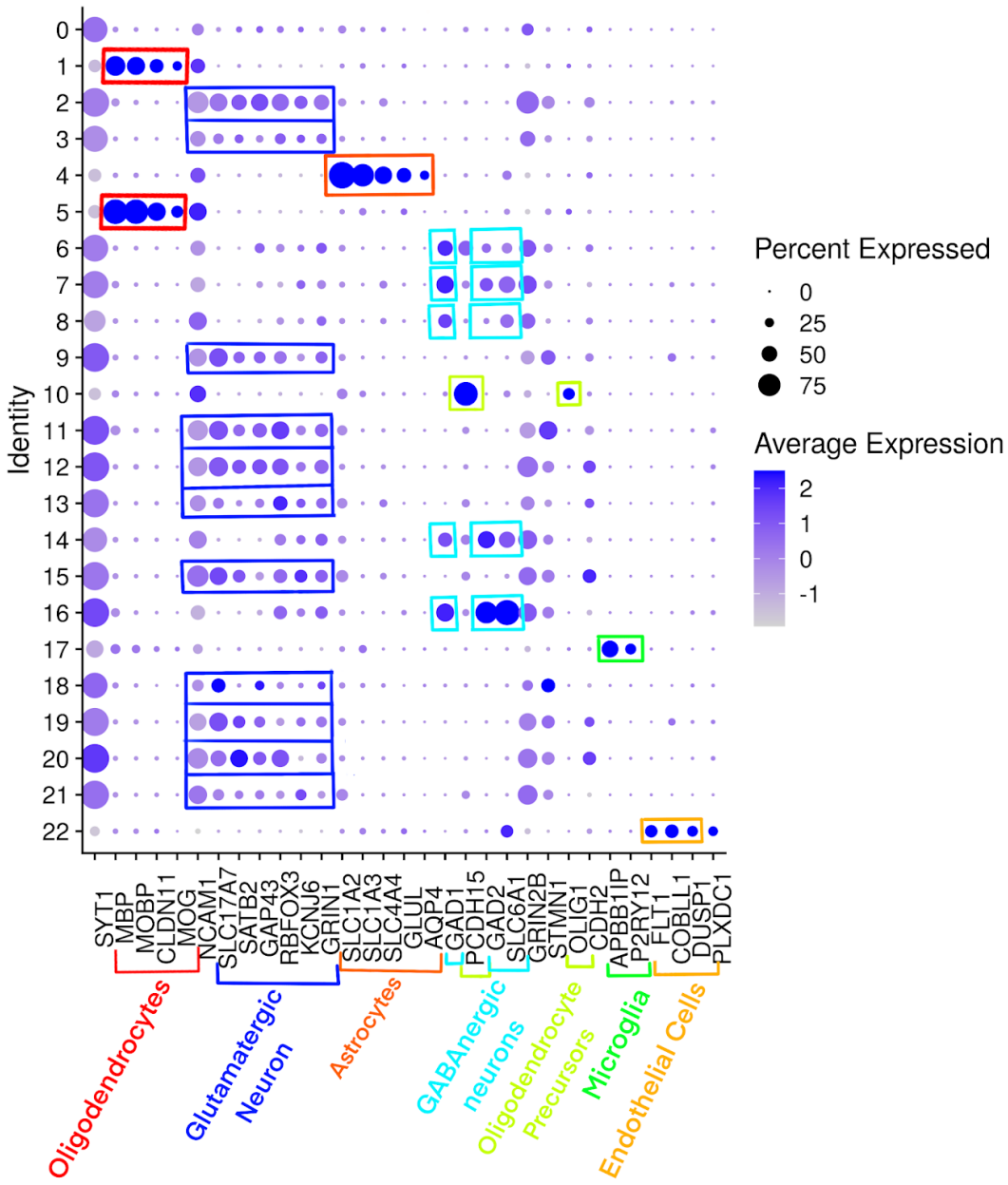

Figure 7: Dotplot showing expression of genes in a cluster. The size of the circles indicates the percentage of cells expressing a marker gene in a cluster. The color of the circle indicates the average expression of a marker gene in a cluster. Plot was manually annotated to indicate the cell types of each marker gene and with rectangles around highly expressed genes to show the predicted cell types in each cluster. Plotted using data from Lake et al., 2018

From this analysis, we can relabel the clusters on the UMAP visualization (Fig. 8). The method separated out the 7 cell types well, with the inhibitory neurons separated into two distinct groups. There were 12 excitatory neuron subclusters that are all close together. While the orientations of the plots are different, the proximity of each of the cell type clusters to each other are also similar suggesting that the two tools and methods are comparable.

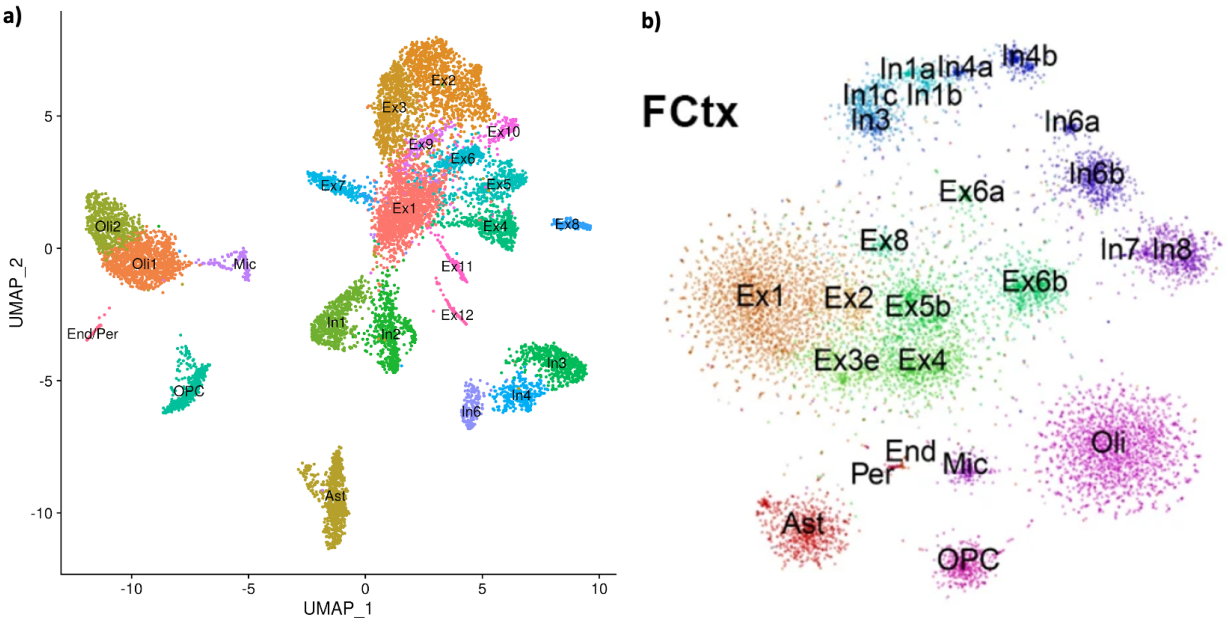

Figure 8: scRNA-seq Frontal Cortex UMAP relabeled with cell types. Panel a is the visualization that we created in Seurat with the data from Lake et al., 2018. Panel b is the visualization directly from Lake et al. 2018 paper.

We can also compare this visualization to the visualization created in Lake et al., 2018. In Lake et al., 2018, they used a tool called Pagoda to create the visualizations. Pagoda is designed for faster processing of large scRNA-seq datasets (Zhang et al., 2020). The visualization created in Seurat created more excitatory neuron clusters but fewer inhibitory neuron clusters than the visualization created in Pagoda. Pagoda was also able to create separate endothelial cell and pericyte clusters, while Seurat tended to combine the two into one cluster. However, the two plots created comparable clusters, indicating that the pipeline will be successful for analyzing single cell RNA-seq data.

3.2 Using the scRNA-seq pipeline to understand Autism Spectrum Disorder

We obtained data about Autism Spectrum Disorder (ASD) from Haney et al., 2020 (preprint) through the PsychENCODE data portal. These data had samples from both autistic and non-autistic individuals, postmortem. The samples were primarily collected from two brain regions – Brodmann Areas 4/6 and Brodmann Area 9. These are both in the frontal cortex and help control motor function.

I followed the Seurat preprocessing workflow but filtering the cells before performing the analysis. The Lake et al., 2018 data was pre-filtered, so the step was not needed. However, in this ASD dataset, it was important to filter out cells with really low RNA expression, which indicates low-quality or empty droplets, or really high RNA expression, which suggests a doublet or multiplet. The standard workflow also filters out cells with a high percentage of mitochondrial gene expression, because low-quality cells tend to have more mitochondrial contamination than higher quality cells. The Seurat standard pre-processing workflow defines numeric cutoffs for subsetting cells based on the number of feature RNA, or the number of unique genes, in the cell. However, in this ASD dataset, using these cutoffs removed a huge portion of the data, because 10X V3 samples tend to have more reads than 10X V2 samples. Figure 9 shows a violin plot of the number of feature RNA in each sample from the control group. The standard preprocessing workflow suggested removing cells with more than 3000 nFeature RNA, or above the red line in figure 9. Samples sequenced using 10x V2 have lower numbers of feature RNA than 10x V3. Using this cutoff, entire samples sequenced using 10x V3 would have been removed. Because of this issue, I chose instead to remove the cells in the bottom 1% and top 5% of nFeature RNA from each sample individually, then merge the datasets.

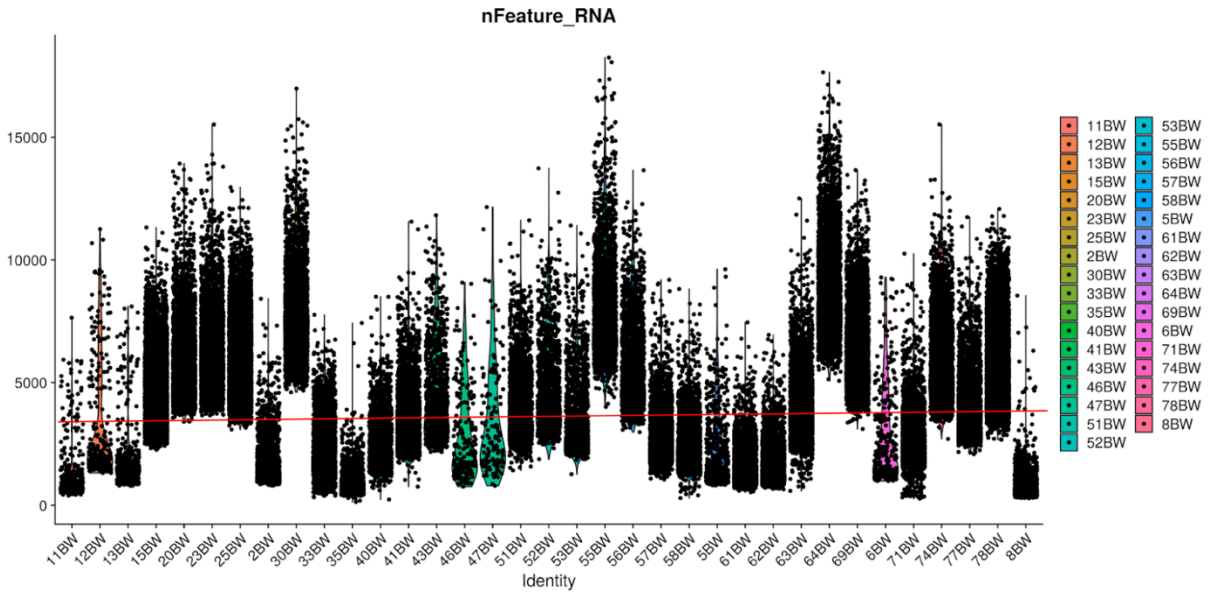

Figure 9: nFeature RNA violin plot of all samples in CTL group. Each segment of the plot represents a different sample. Each of these samples have a different color, indicated by the legend. The red line represents the default cutoff for the standard preprocessing workflow where all cells above the line would be removed. Plotted using data from Haney et al., 2019.

Next, I examined the effectiveness of Harmony at batch integration by separating out control and autistic samples and observing the mixing before and after Harmony, shown in Figure 10.

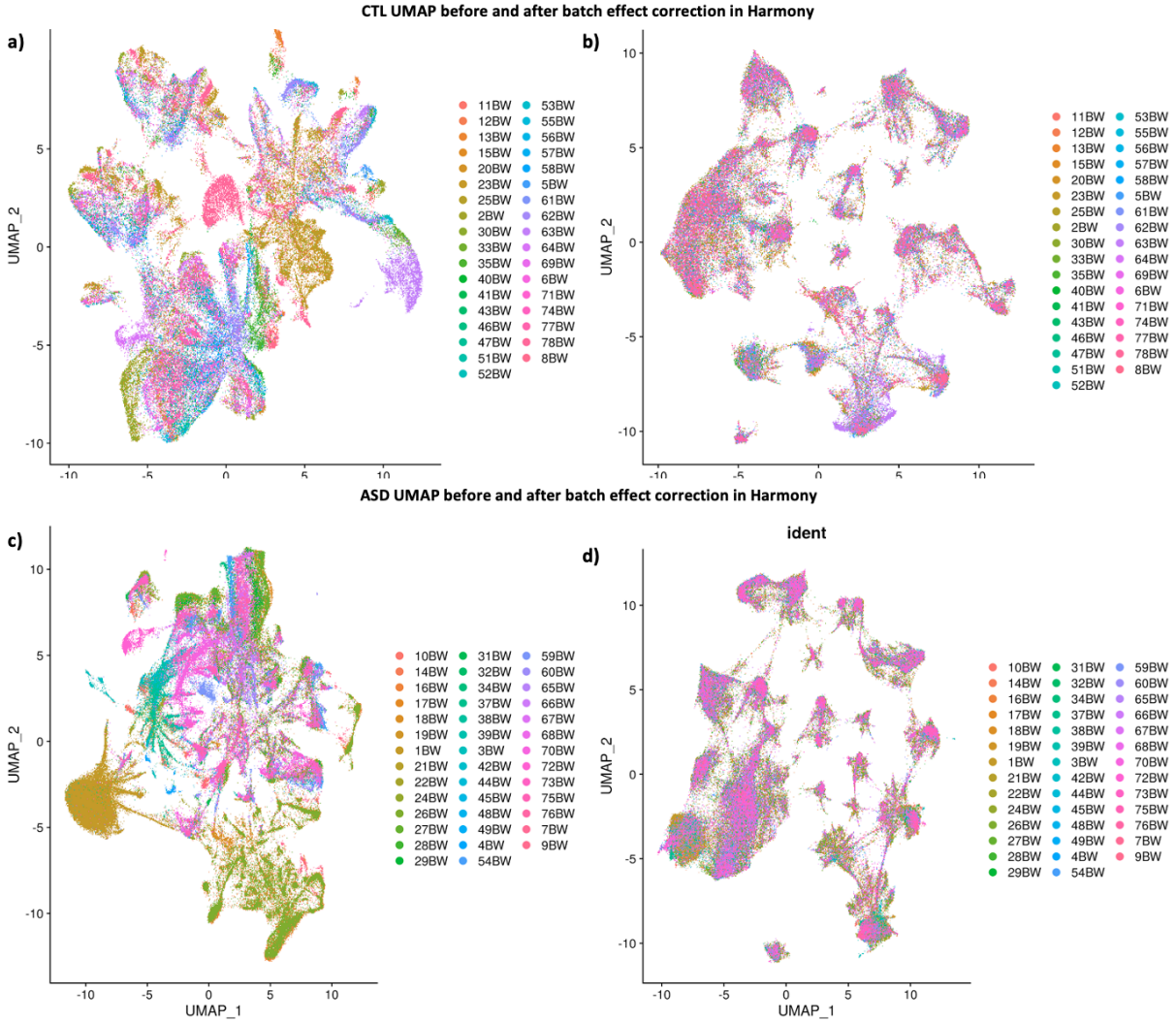

Figure 10: Harmony batch mixing for both Control and ASD subsets. In all plots, the colors indicate which sample the cell came from. Panels a) and c) show the UMAP plots before the batch effect correction. Panels b) and d) show the UMAP plots after batch effect correction with Harmony. Plotted using data from Haney et al., 2019.

We can see that there was some clear separation between batches before Harmony, which was eliminated after its use. In Figure 10 panels a) and c), many of the cells have clearly separated by donor or batch. In panels b) and d), each batch is distributed across multiple clusters. This is clearly shown by the pink and purple cells, which are obscuring the other colors that were plotted earlier.

I then divided the samples up into their brain regions. This created four groups: CTL BA4/6, CTL BA9, ASD BA4/6, and ASD BA9. Figure 11 shows the UMAP plots of each group before doublet removal or cell type labeling.

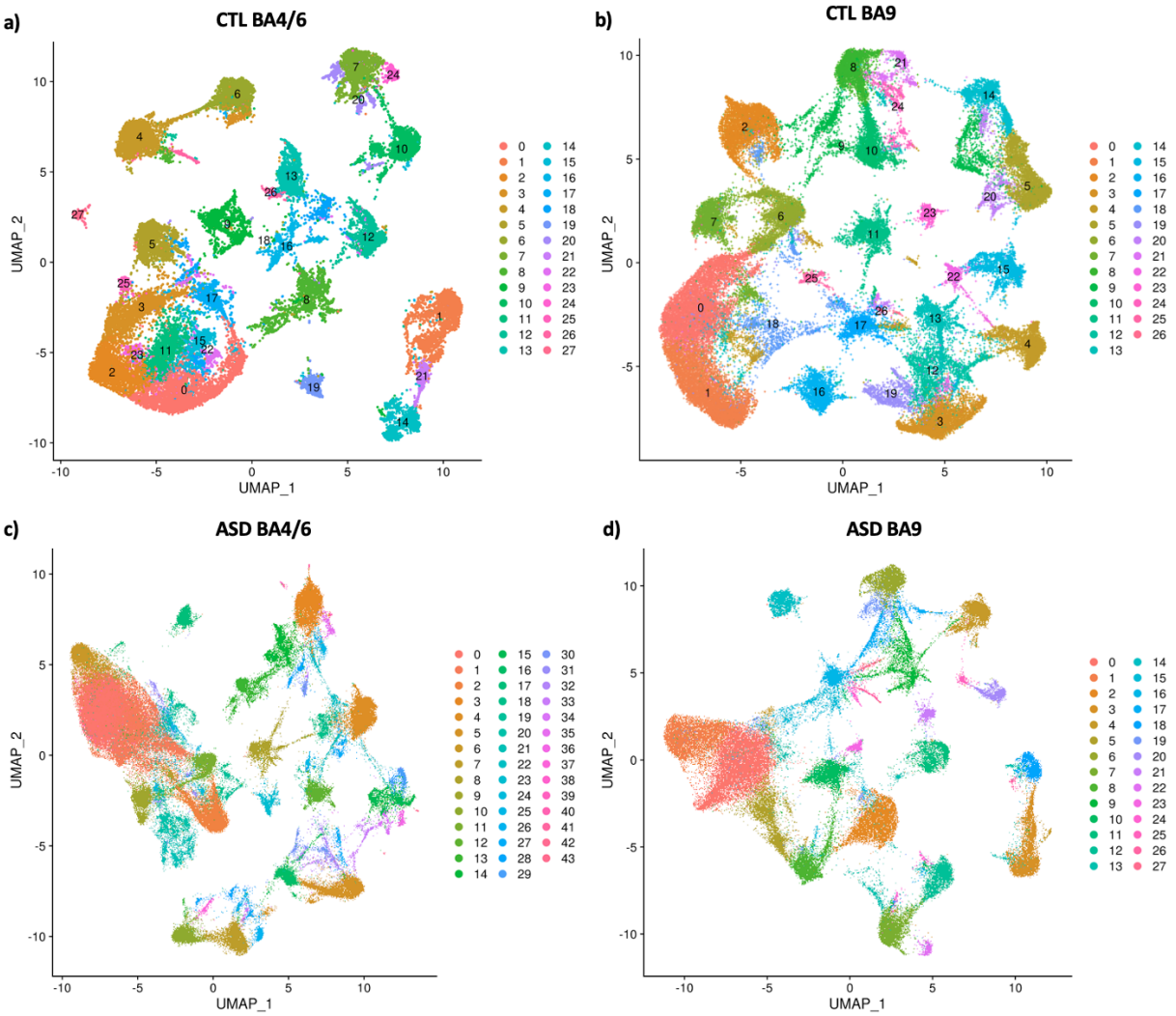

Figure 11: UMAP and clustering after Harmony. a-d each show plots from different UMAP analyses. a and b are the samples from control patients (CTL) while c and d are from autistic patients (ASD). In addition, a and c are the samples from Brodmann Areas 4 and 6, while b and d are the samples from Brodmann Area 9. Each of the samples are colored and numbered by the clusters created by Seurats FindClusters() algorithm. Plotted using data from Haney et al., 2019.

I then performed doublet detection and removal using DoubletFinder. Figure 12 shows images of the detected doublets, assuming 7.5% doublet formation rate, then the UMAP plot after removing the predicted doublets. We are using a 7.5% doublet formation rate because this is the average rate for large 10x runs.

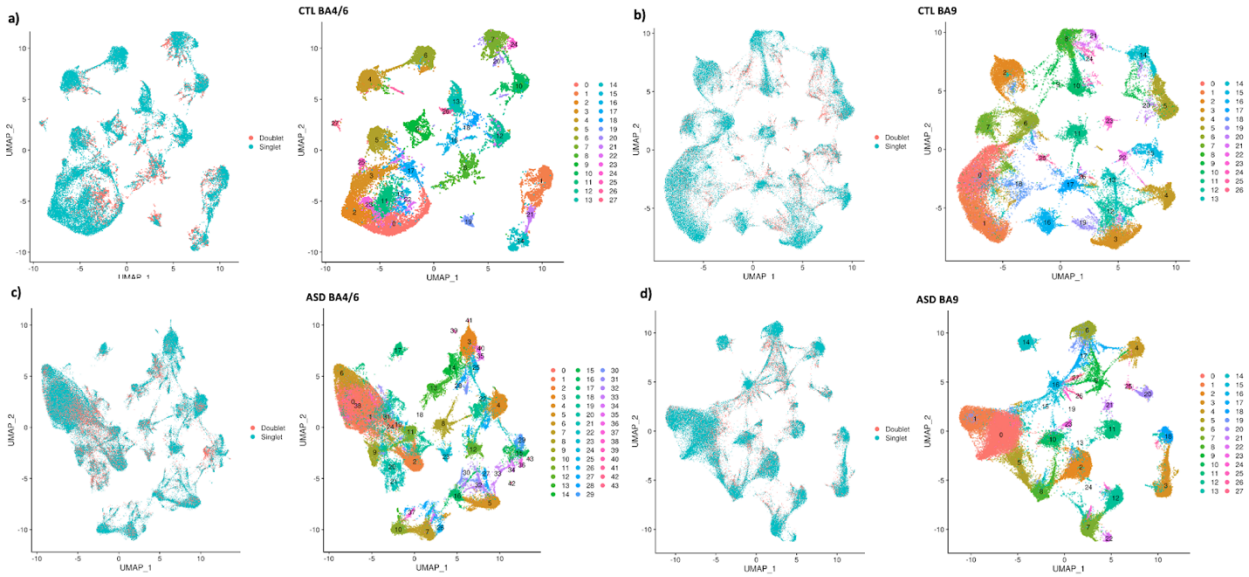

Figure 12: UMAP plots of doublets and doublet removal. a and b are the samples from control patients (CTL) while c and d are from autistic patients (ASD). In addition, a and c are the samples from Brodmann Areas 4 and 6, while b and d are the samples from Brodmann Area 9. The plot on the left shows the UMAP embedding with predicted doublets in red and predicted singlets in blue. The plot on the right then shows the UMAP embedding and clusters with doublets removed. Plotted using data from Haney et al., 2019.

Next, I used the Lake et al., 2018 Frontal Cortex snRNA-seq dataset as a reference to label the cell types of the cells in these datasets using SingleR. SingleR can take a reference dataset and define anchors based on the gene expression in each cluster. The tool can then apply these anchors to a new dataset and label the visualization, either by cell or by cluster. Since both datasets are from the frontal cortex, the gene expression profiles should be similar enough to apply cell type labels. I used a combination of the UMAP plots resulting from the SingleR labeling and marker gene expression from the dot plots, shown in Figure 13. Using both of these, I was able to confidently label most of the clusters in these figures.

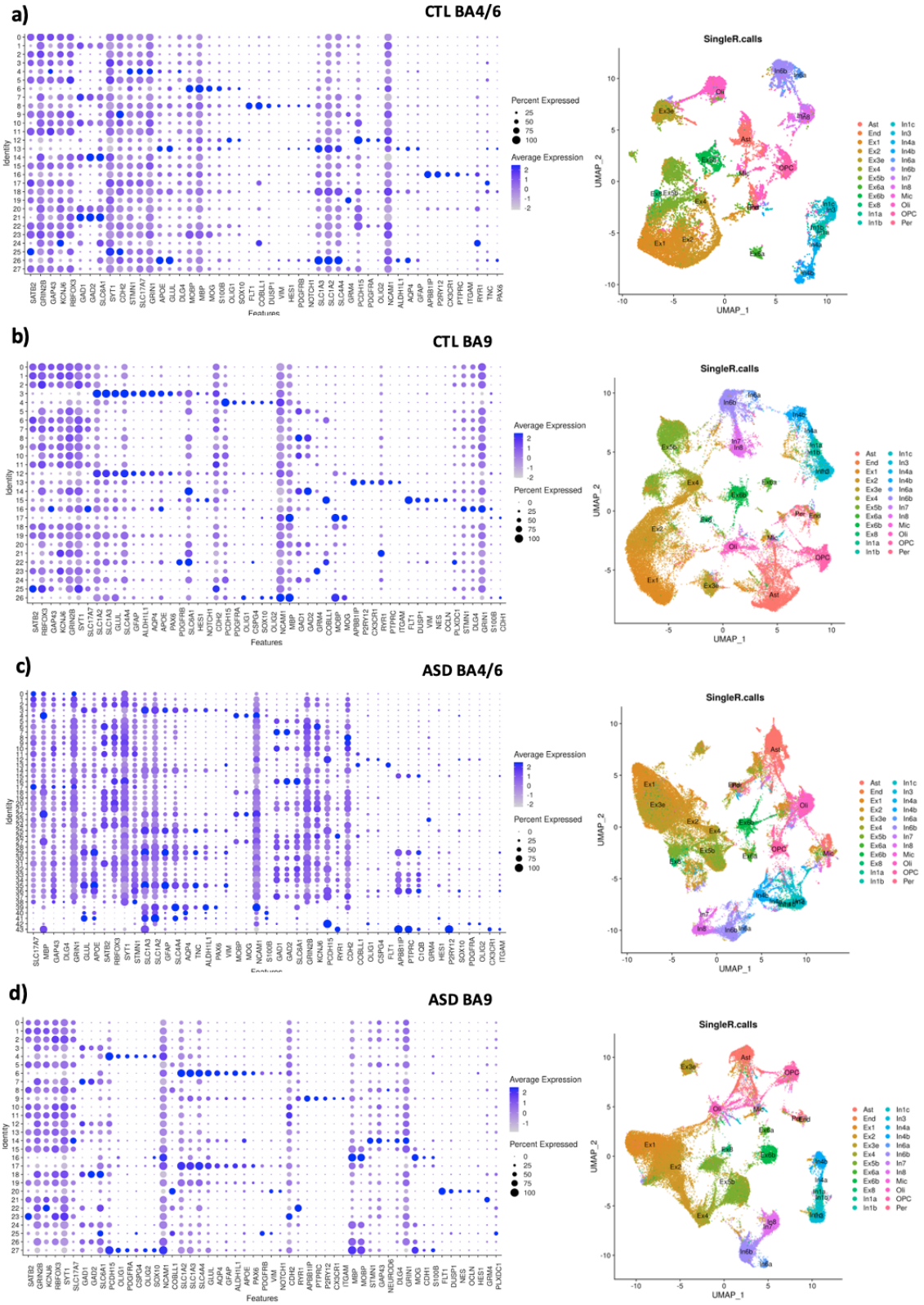

Figure 13: Marker gene dotplot and SingleR cll type predictions for cell type labeling. The first is a dotplot showing expression of marker genes across different clusters. The second plot is the UMAP embedding with cells labeled by SingleR, using the Lake et al., 2018 reference to identify the cell type of each individual cell. Plotted using data from Haney et al., 2019.

Figure 14 shows the final labelled UMAP plots for each of the clusters.

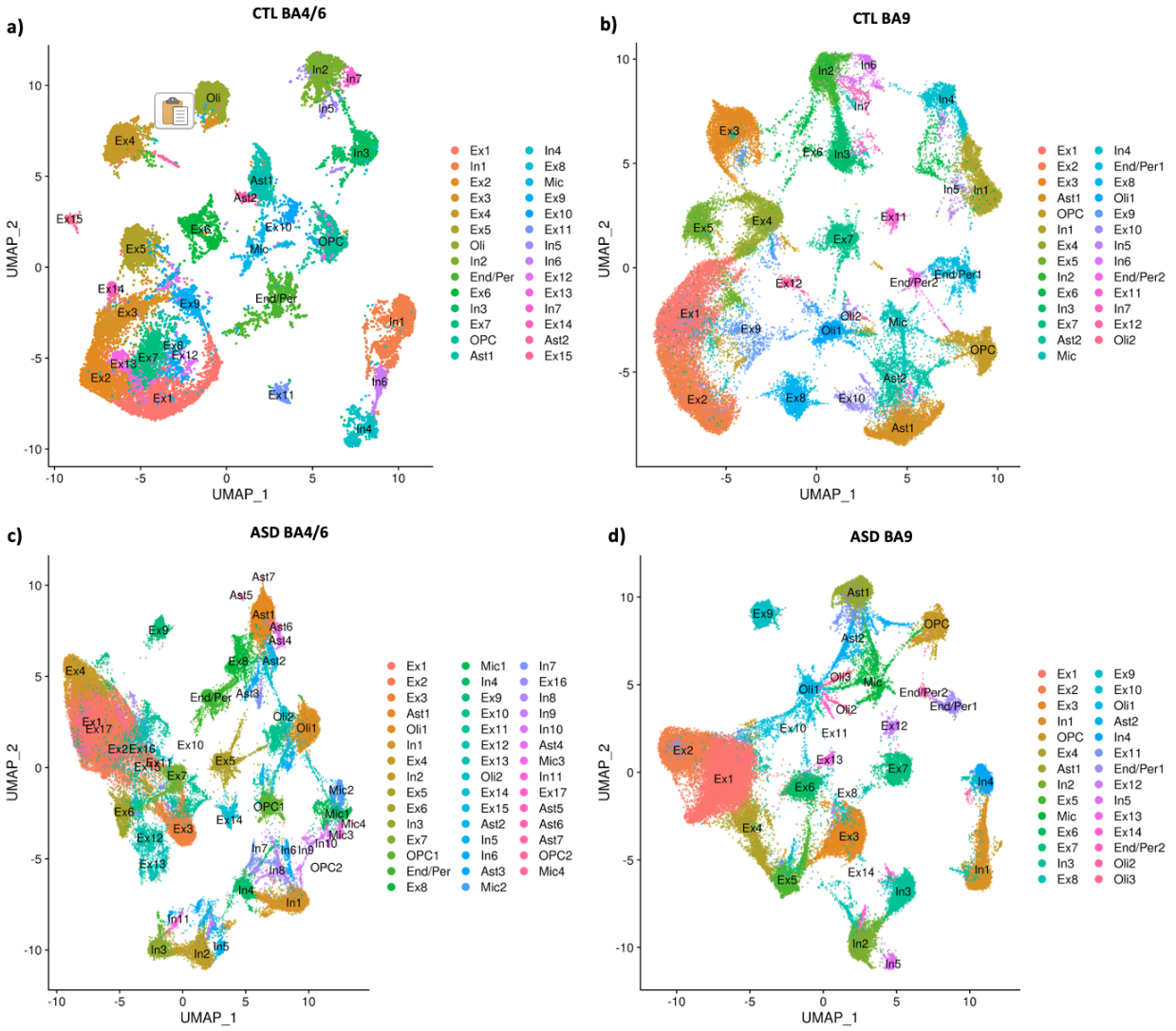

Figure 14: UMAP plots with consensus cell type labels. Plots are colored by cluster and cell type. Plotted using data from Haney et al., 2019.

Next, I combined the CTL and ASD samples from matching brain regions and plotted them together, shown in Figure 15. We combine the donors to plot them on the same UMAP projection, which can allow us to visualize any differences between the ASD and CTL samples.

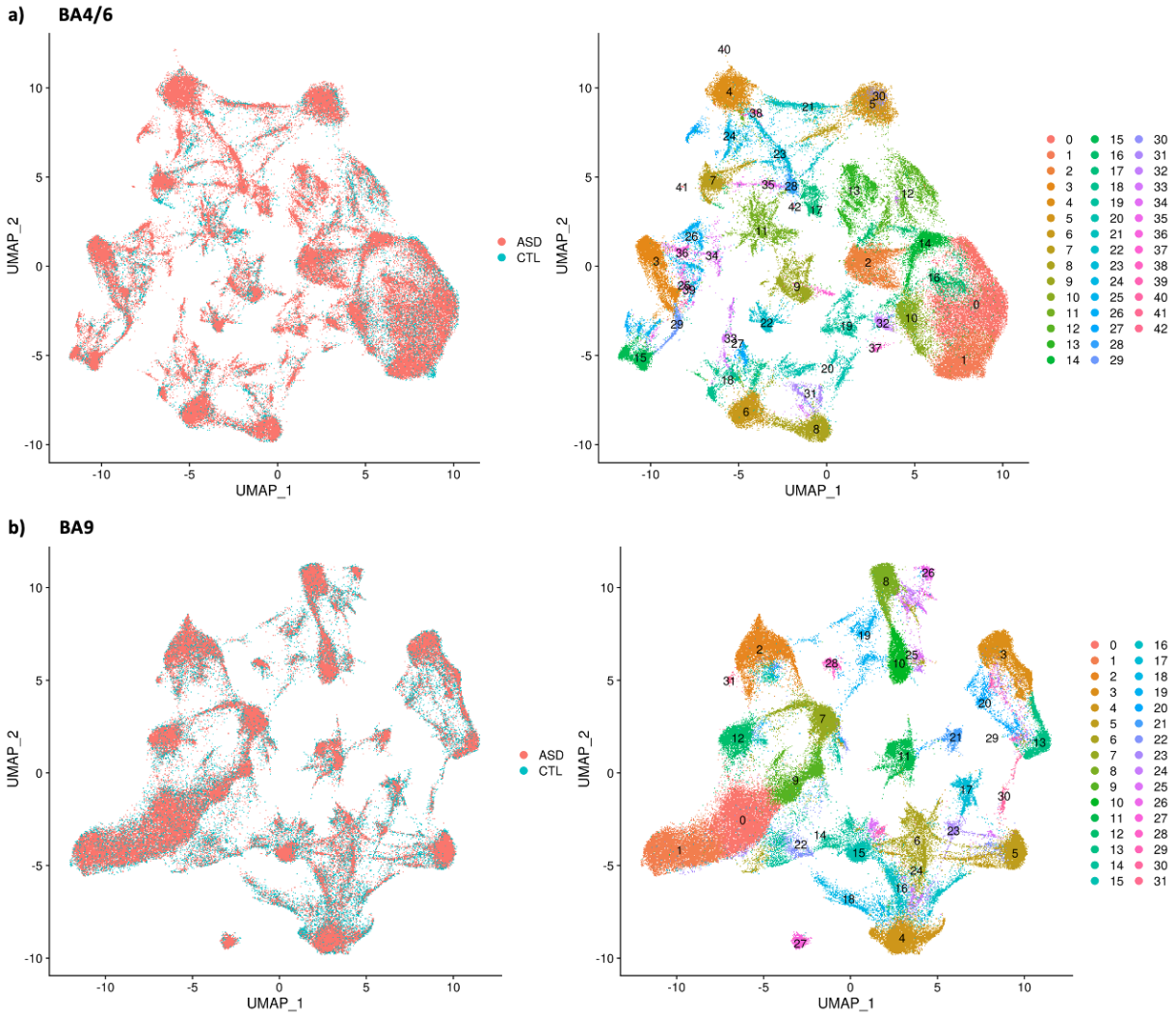

Figure 15: UMAP of Brodmann Areas 4/6 (a) and 9 (b) combining CTL and ASD samples. The plots on the left are the combined UMAP, colored by group. ASD samples are red, while CTL samples are blue. The figures on the right show the combined UMAP, colored and labeled by cluster. Plotted using data from Haney et al., 2019.

We can then split this plot by group, or ASD vs CTL, shown in figure 16. While the CTL plots appear to be sparser, there are not any major visual differences in relative position of the clusters on each plot.

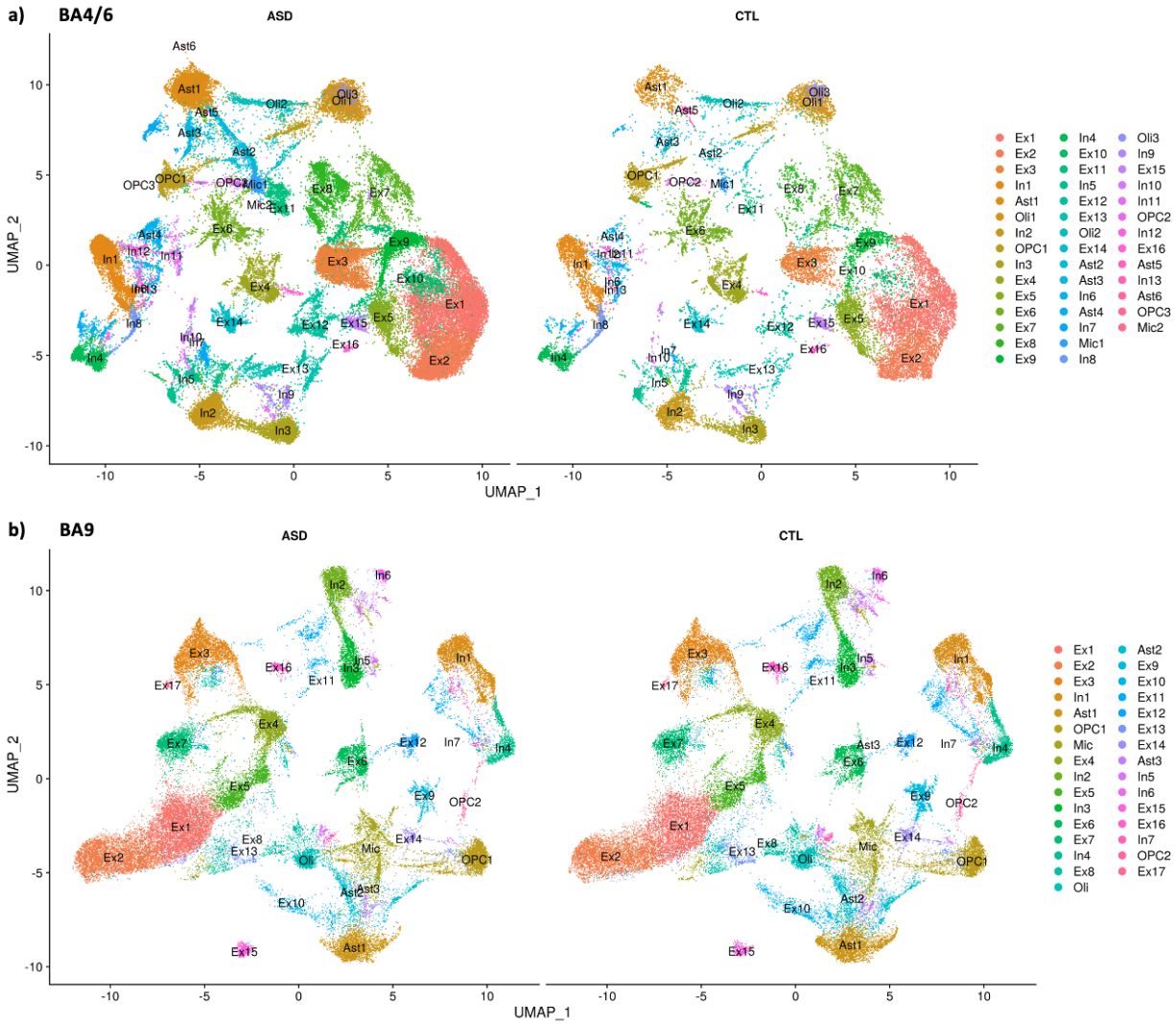

Figure 16: UMAPs split by group, colored by cluster and cell type. Figure a plots Brodmann Area 4/6, while figure b plots Brodmann Area 9, Plotted using data from Haney et al., 2019.

The clusters were labeled primarily using Single R but using marker gene expression as a check. I then used these cell type labels to calculate the proportion of each cell type in the samples. These proportions are shown in Figure 17.

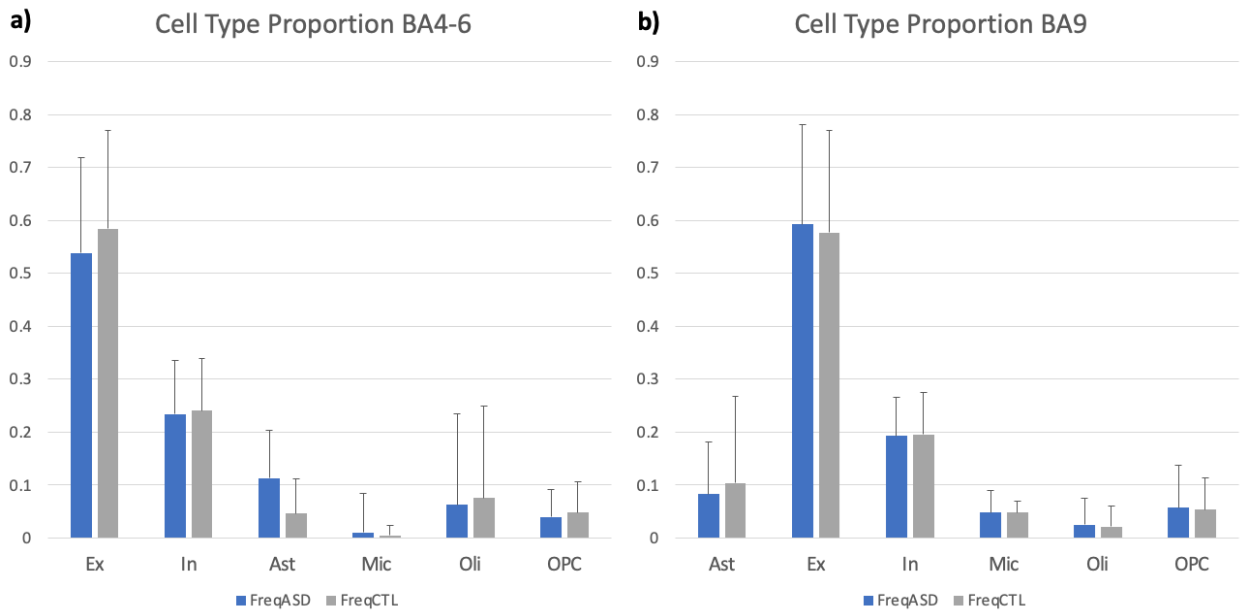

Figure 17: Cell type proportion comparison between ASD and CTL samples, pooling the clusters of a cell type. The plot on the left shows the cell type proportion comparison in Brodmann Area 4/6, while the plot on the right shows the cell type proportion comparison in Brodmann Area 9. Dark blue is CTL and light blue is ASD. Error bars show the standard deviation across all samples. Plotted using data from Haney et al., 2019.

Because of the high variability in cell type proportions among samples, we were unable to find statistically significant differences, despite apparent differences in excitatory neuron and astrocyte proportions in BA4/6.

Figure 18 shows the proportion of each cluster in ASD versus CTL. Again, there appear to be some striking differences in cluster proportion, but it is made insignificant by high standard deviations. For example, in Ex1 in BA4/6, there appears to be a large difference in proportion between ASD and CTL. However, this is likely because donor BW1 has about 80% of its cells in this cluster. Thus, this result could be attributed to a batch effect that wasn't resolved with Harmony.

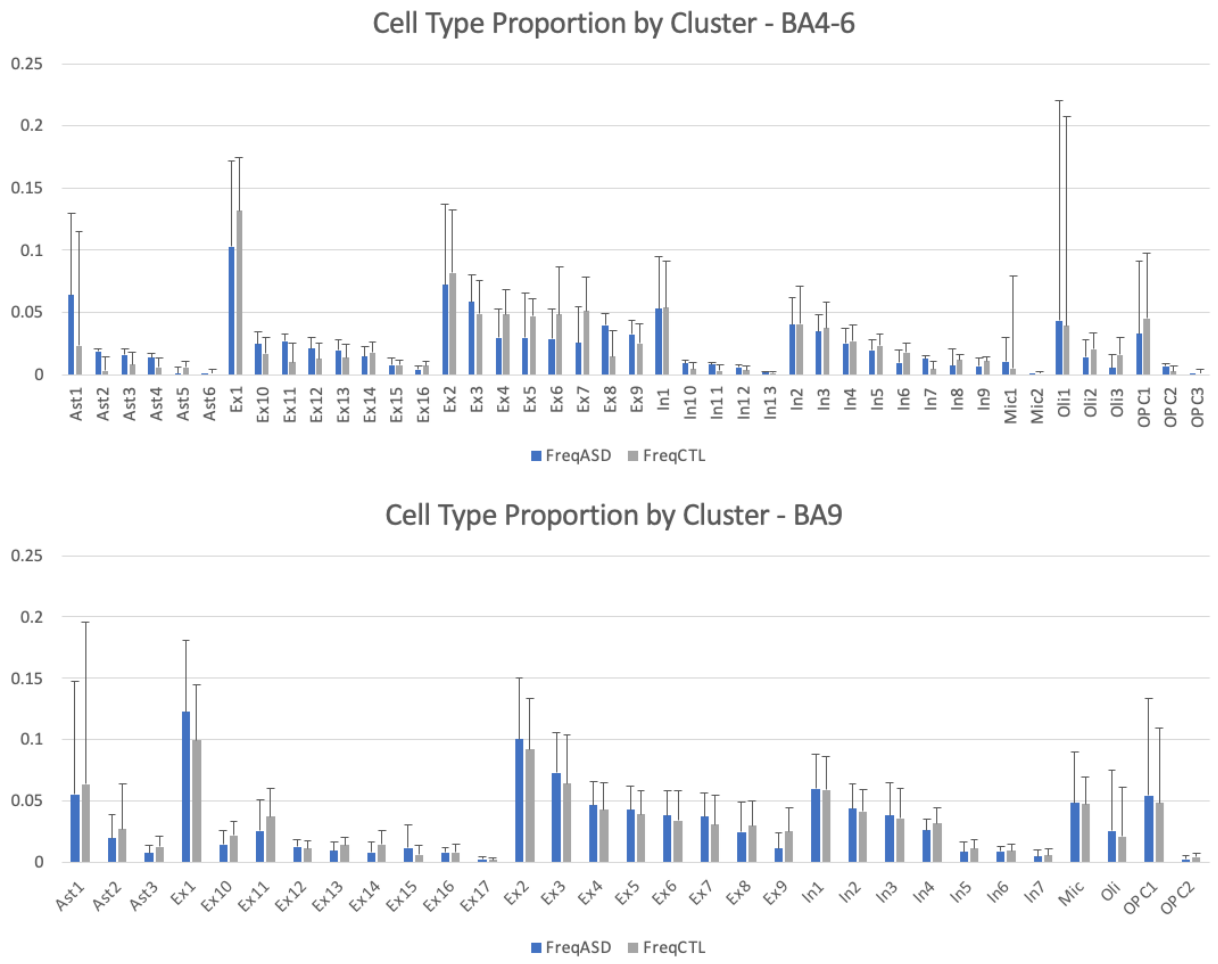

Figure 18: Cluster proportion comparison between ASD and CTL samples, comparing each cluster identified by Seurat. The first plot shows the cluster proportion comparison in Brodmann Area 4/6, while the second plot shows the cluster proportion comparison in Brodmann Area 9. Blue is CTL and grey is ASD. Error bars show the standard deviation across all samples. Plotted using data from Haney et al., 2019.

In addition to examining cell type proportions, we can also find differentially expressed genes (DEGs) between ASD and CTL samples, shown in Figure 19. There is over-expression of two genes in 4 of the 6 cell types: IFI6 and CLU. Both genes are involved in negatively regulating apoptosis. There is also increased activation of mitochondrial genes in neurons, which have been found to play a role in activating inflammation (Tsilioni and Theoharides, 2018). In addition, several genes coding for heat shock proteins (HSPs) were upregulated in the microglia of ASD samples. HSPs play a large role in preventing apoptosis and creating a pro-inflammatory response (van Eden et al., 2005). NEAT1, a long-noncoding RNA, is highly over-expressed in microglia of ASD samples. NEAT1 is also involved in activation of inflammation (Zhang et al., 2019). This supports neuroinflammation, mitochondrial dysfunction, and oxidative stress hypotheses for the etiology of Autism Spectrum Disorder (Meltzer and Van de Water, 2017; Citrigno et al., 2020, Chauhan and Chauhan, 2006).

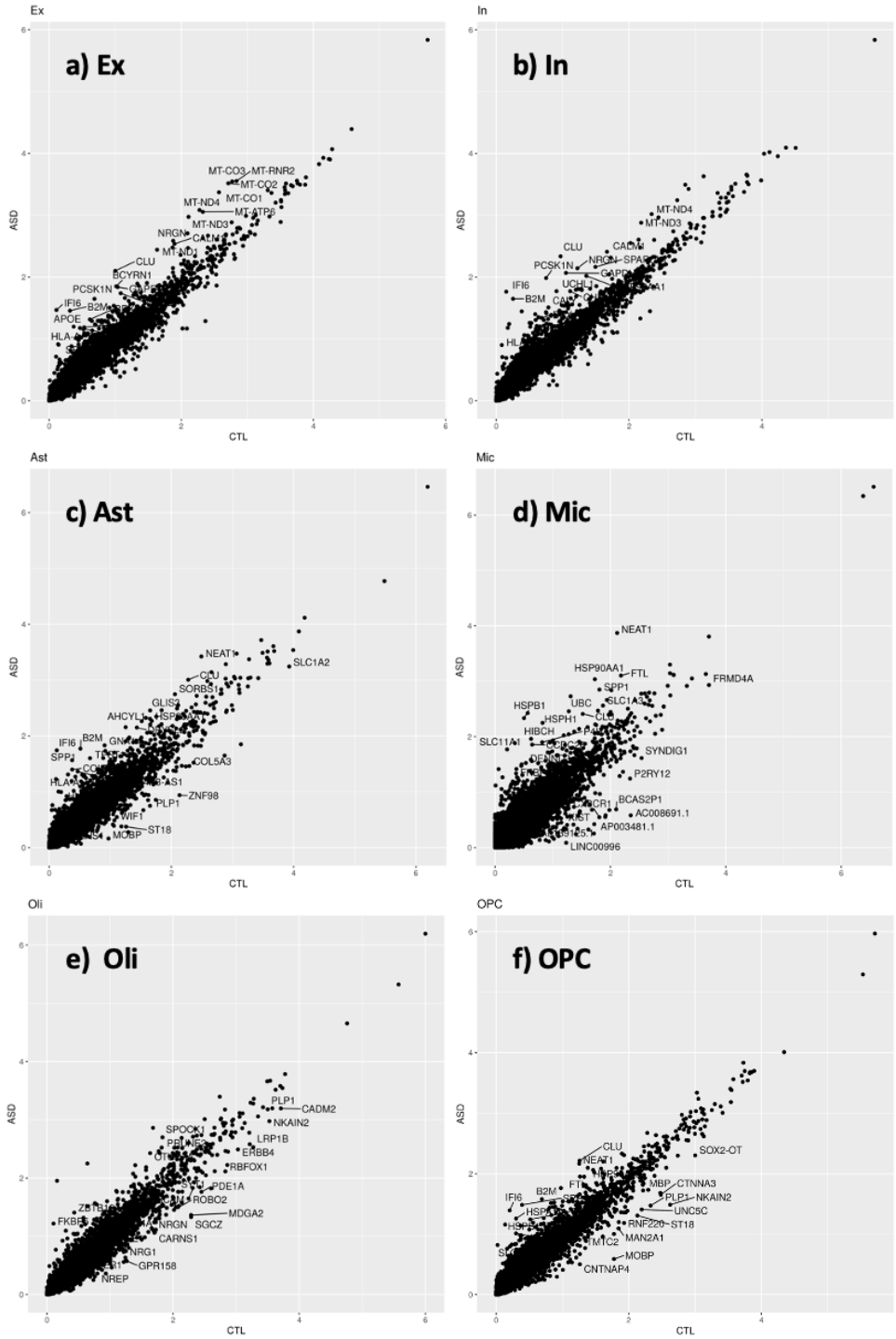

Figure 19: DEGs in each of the 6 cell types identified by Seurat for Brodmann Area 4-6. Panel a-e each display plots for a different brain region. Each point is a gene, with average expression in CTL cells on the x-axis and average expression in ASD cells on the y-axis. 25 genes with the highest log-fold change and lowest p-values are labeled on the plots.

3.3 Benchmarking the scATAC-seq pipeline

To benchmark the scATAC-seq pipeline, I used the THS-seq data from Lake et al., 2018. This pipeline is similar to the scRNA-seq pipeline, just using different tools to accommodate the different data type. The data were manually formatted into a single fragment file and added to ArchR. I changed the quality control parameter from the default, lowering the minimum required fragments per cell to 100 and the minimum required transcription start sites to 2. Using the default QC parameters, the UMAP was extremely sparse, and lowering it allowed a lot more cells to pass through the filtering step. From this, we were able to create a UMAP plot, shown in Figure 20.

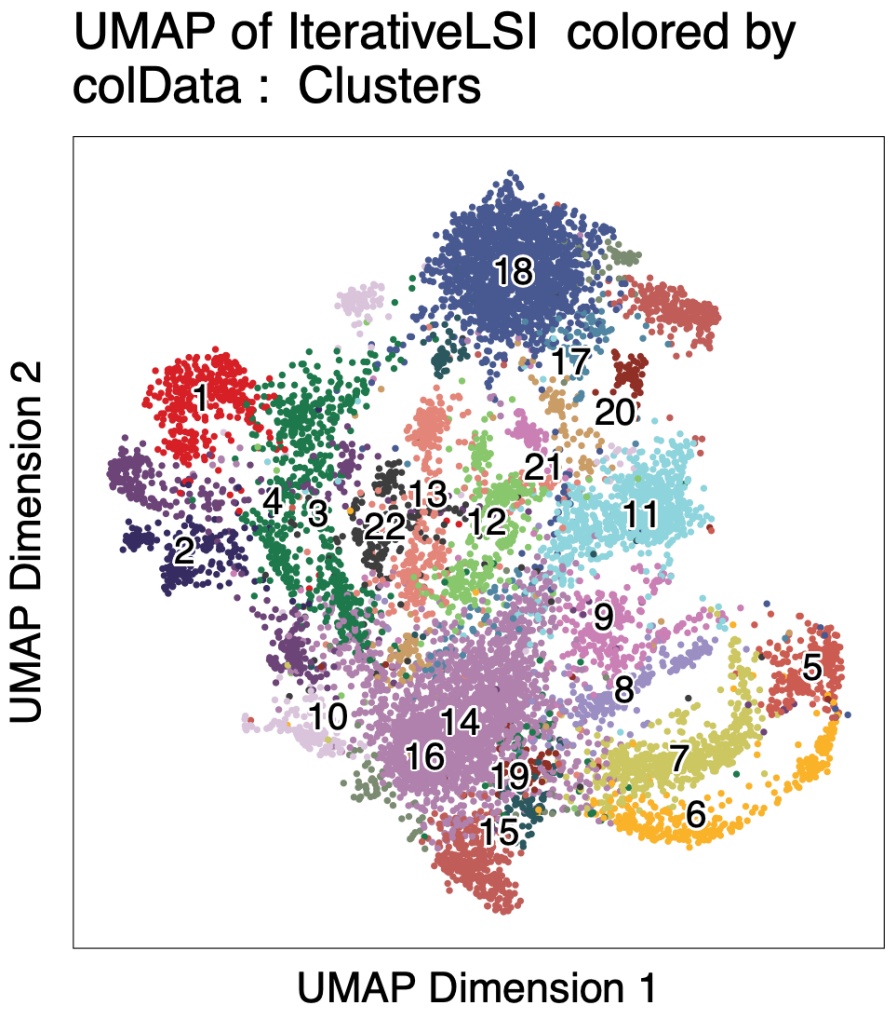

Figure 20: UMAP of Lake et al., 2018 THS-seq data, created in ArchR. Clusters were created based on peak-called regulatory element enrichment. Each cluster represents a sub-cell type or cell state

ArchR can then use pairwise comparisons to understand regulatory element enrichment and predicted gene expression in each cluster defined by ArchR. This is shown by heatmaps in Figure 21.

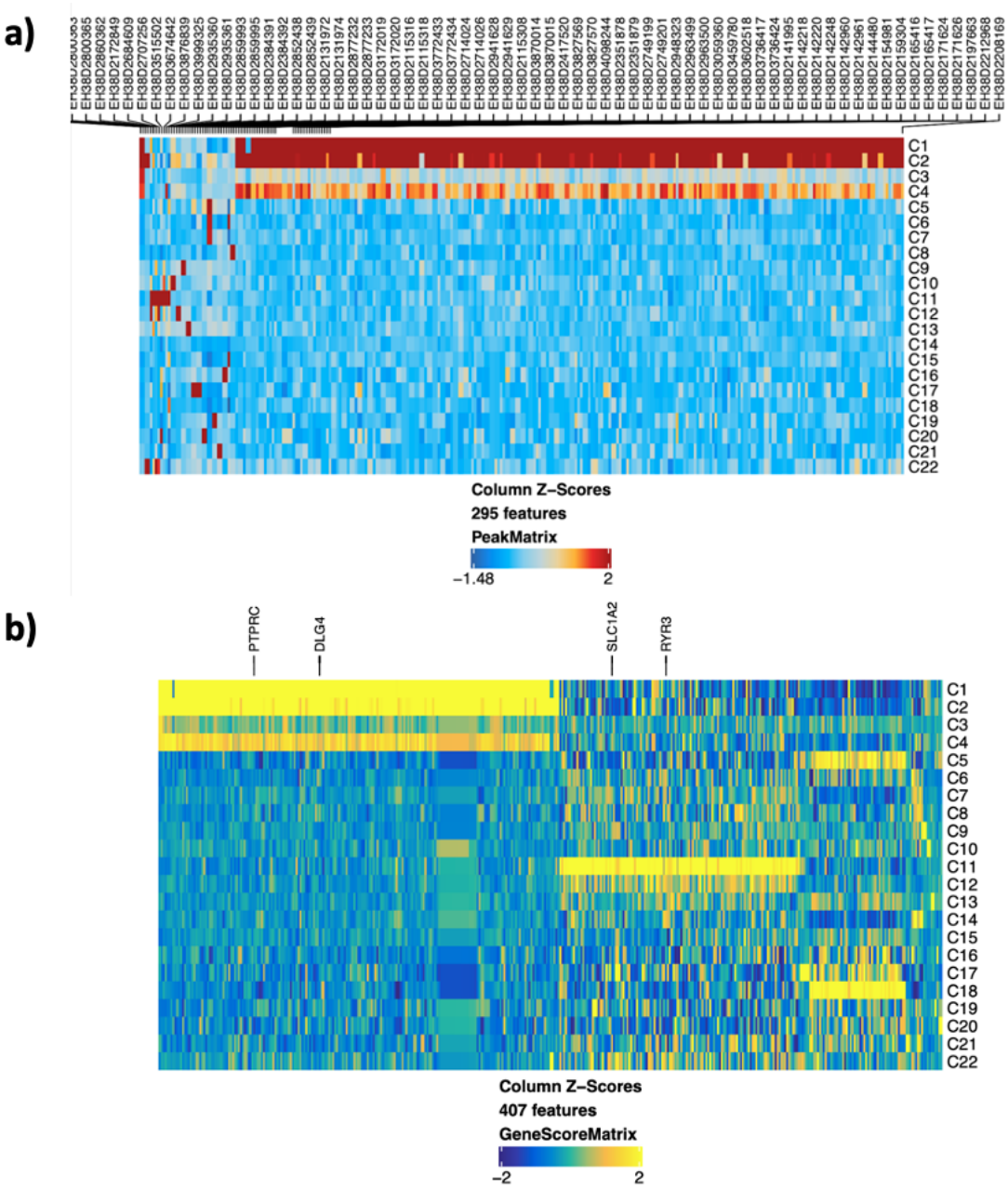

Figure 21: Heatmaps to visualize regulatory element enrichment in each cluster. a) shows the enrichment of 356 regulatory elements (x-axis) in 21 clusters (y-axis). Highly enriched regulatory elements in a cluster are shown in red, while under-enriched elements are blue. b) shows gene scores generated by ArchR, predicting gene expression based on regulatory element enrichment. High gene scores are colored yellow, low gene scores are colored blue. x axis of b is labeled with marker genes from literature. Plotted using data from Lake et al., 2018

There is an overabundance of regulatory elements found in clusters 1, 2, and 4. Upon looking at the labels of the regulatory elements, many of the regions were very close to each

other. The heatmap includes regions with 15 to 20 regulatory elements that control the same gene. When one enhancer in this group is activated, it opens up activation of the rest of the regulatory region.

We can find these regulatory-element dense regions by using ENCODE screen and UCSC genome browser. One of these regulatory regions is shown in Figure 22.

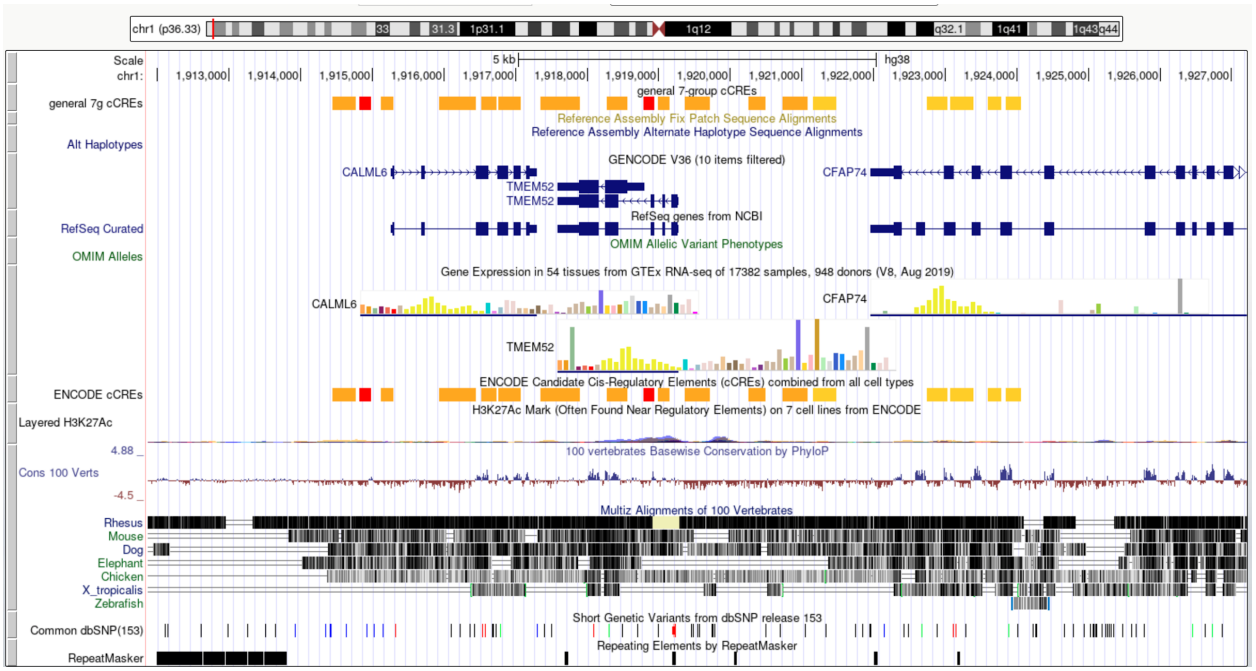

Figure 22: UCSC Genome Browser Image of Regulatory Element EH38E1311335

There are a surprisingly small number of regulatory element features in figure 20. Because the analysis only yielded 407 gene features, only 4 marker genes were found in the analysis. SLC1A2 is an astrocyte marker gene and RYR1 is a marker gene often expressed in purkinje neurons in the cerebellum as well as astrocytes in other brain regions. Therefore, we can predict that cluster 11 and a few of the clusters nearby contain astrocytes, which is consistent with Figure 23. Cluster 11 in Figure 23a is part of cluster 1, colored red, in Figure 23b. DLG4, which is enriched in clusters 1, 2, and 4 of Figure 23a, is a marker for mature neurons. These are in fact inhibitory neurons, part of cluster 3 in Figure 23b, but ArchR was unable to predict high gene scores for known inhibitory neuron marker genes from literature.

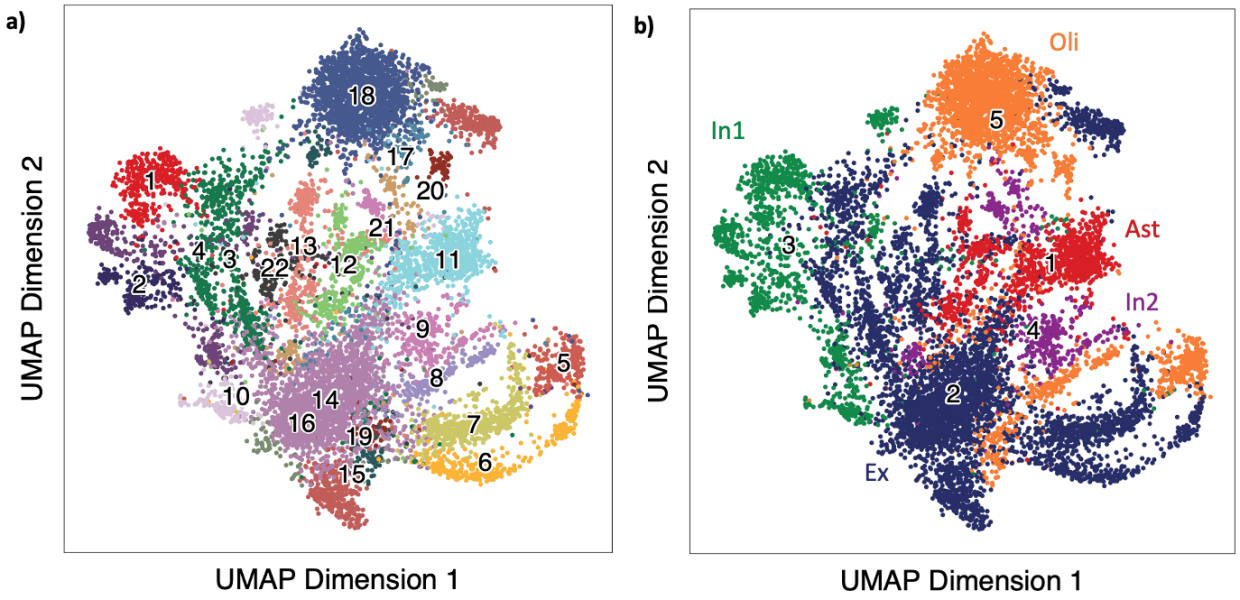

Figure 23: Cell type labels of each THS-seq cluster. Panel a is the UMAP of the scTHS-seq data from figure 19, separating each sub-cell type identified by ArchR. Panel b shows the same UMAP plot, but labeled with cell types: an astrocyte cluster, excitatory neuron cluster, two separate inhibitory neuron clusters, and an oligodendrocyte cluster.

This result could be attributed to the quality of the data or the integration function. The number of inhibitory neuron-associated regulatory elements may have been artificially low in the data collected. More likely, ArchR's integration function was unable to associate the regulatory elements with a known marker gene. But due to the fact that clusters were labeled as inhibitory neurons, the program may be picking up on more subtle or not-yet-discovered marker genes and using them to label the dataset.

This integration was also unable to pick up on multiple sub-cell types. Figure 23a shows that ArchR predicted 22 clusters or sub-types. While there were 10 inhibitory neuron and 8 excitatory neuron sub-clusters in the snDrop-seq data, the integration only labeled 1 excitatory neuron cluster and 2 inhibitory neuron sub-clusters in the THS-seq plot in Figure 23b.

We can also compare these results to the plot in Lake et al., 2018, shown in figure 24.

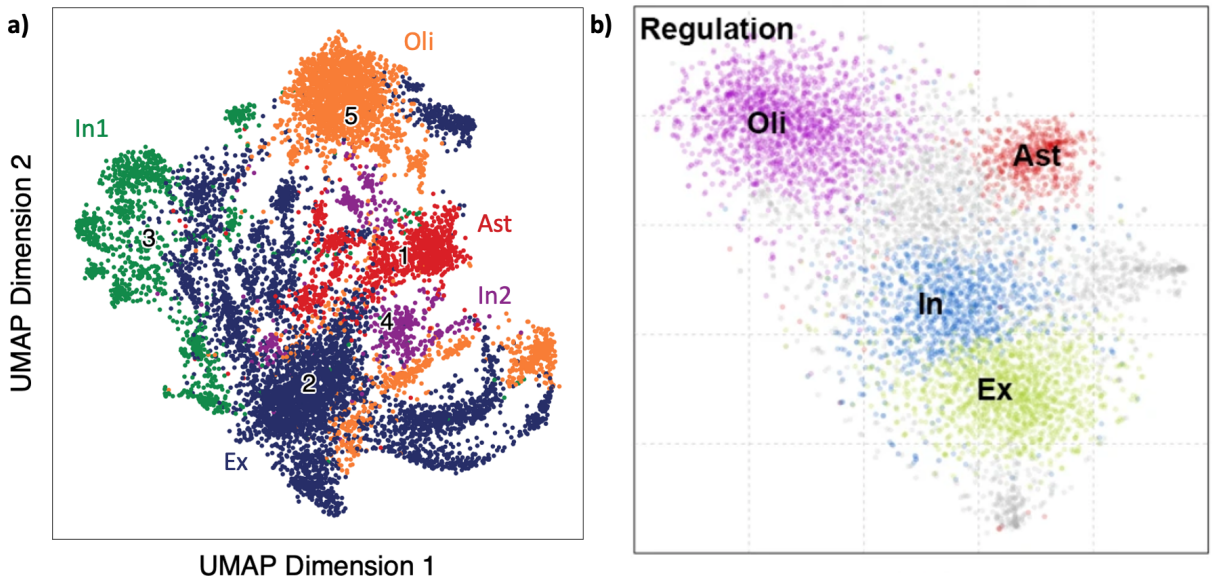

Figure 24: Benchmarking ATAC-seq analysis pipeline. Panel a is the plot created in ArchR using the THS-seq data from Lake et al., 2018. Panel b is the plot published in Lake et al., 2018.

The two plots in figure 24 are similar. They have the same 4 cell types: excitatory neurons, inhibitory neurons, oligodendrocytes, and astrocytes. In addition, the placement is similar, with the neurons together, then the astrocyte and oligodendrocyte clusters displaced more. The only locational difference is that the inhibitory neurons are broken into 2 subclusters in the UMAP plot. However, we do notice that in typical scRNA-seq UMAPs, the inhibitory neurons form two distant subclusters. While the plot created in ArchR (Figure 24a) appears to have less distinct clusters than Figure 24b, this is because Lake et al., 2018 created these plots with lower opacity points. Any of the stringy points around the edges of the clusters are low density, so they don't appear clearly in the plot.

3.4 Integrating RNA-seq and ATAC-seq data

We are interested in using integration between single cell RNA-seq and ATAC-seq data not only for cell type labeling, but also for exploratory analyses that make novel connections between genes and the regions that regulate them.

Figure 25 shows the UMAP plots of the Lake et al., 2018 scTHS-seq data integrated with the Lake et al., 2018 snDrop-seq data. This uses the predicted gene scores of the THS-seq data, visualized in figure 20b, and compares them to actual gene expression in the snDrop-seq data. Figure 25a shows the cell-type labels resulting from the integration, with a unique label for each cell. In Figure 25b, these cell types were aggregated to generate one label for each of the 22 original clusters. While the analysis did label cells with 22 cell types and sub-types, they did not correspond well with the clustering that ArchR performed, so many of the sub-types disappeared when creating consensus cluster labels. Of the 4 cell types we were expecting, ArchR was only able to find gene scores for astrocyte marker genes. However, ArchR must have found other

markers or anchors to label the scTHS-seq dataset. We may be able to take a closer look at which genes and gene scores are being aligned to label these cells to potentially find new markers. We could even take a step back and try to make connections between the regulatory element enrichment driving the gene score to make novel connections between genes and regulatory regions.

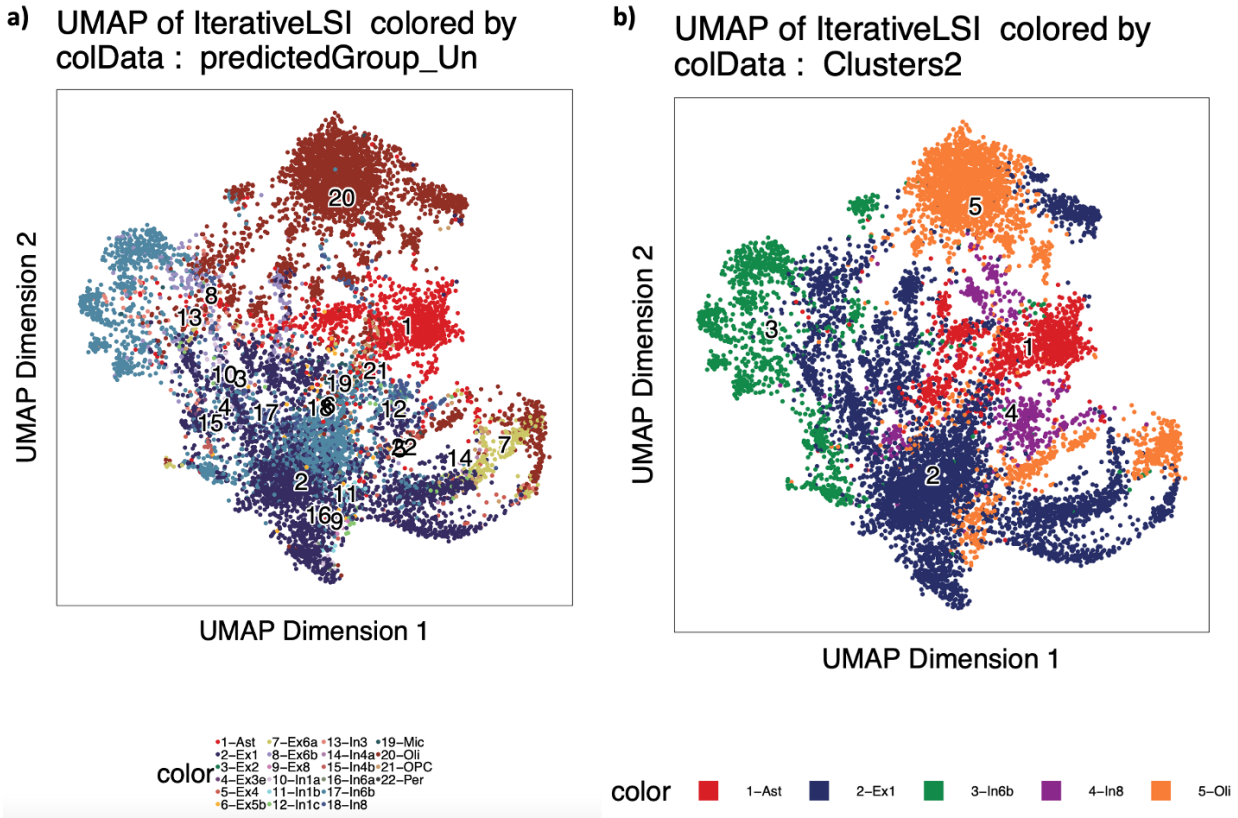

Figure 25: scTHS-seq data labeled by integrating scRNA-seq data. Panel a shows the UMAP plot with unique cell type labels assigned to each cell. Panel b shows the UMAP plot with the cell types of each cell in a cluster aggregated and used to come to a consensus about the likely identity of the cluster overall. Plotted using data from Lake et al., 2018.

4. Discussion

In this project, I have compiled a pipeline for analyzing single cell sequencing data. This pipeline and the tools used have already been used by other members of the lab to analyze other single-cell RNA-seq datasets.

The analysis done using Lake et al., 2018 was used both for benchmarking and for learning about the parameters in Seurat. While the analysis shown here was done using the "consensus" parameters, the lab did lots of experimentation to reach this consensus. Some of these parameters should be adjusted based on the size of the dataset, quality control metrics, and goals of analysis. For example, we might be interested in creating more subclusters that are more specific, which could be adjusted using the resolution parameter in the FindClusters() function. The number of PCs used and FindClusters() resolution often needs to be adjusted based on the size of the dataset, as well. And based on the quality of the dataset, we may need to adjust filtering schemes, UMAP hyperparameters, and even the manifold shape used for the UMAP.

With the analysis done on the Lake et al., 2018 data, there was consistently one cluster that was difficult to identify. In figure 5, this was cluster 0. Based on figure 6, we could clearly see that it was a neuron, based on its expression of SYT1. However, it had low expression of both excitatory and inhibitory marker genes. We originally hypothesized that they were doublets, but based on the doublet prediction, they were not. It could be advantageous for future analysis to understand why expression of all marker genes is so low in this cluster and what biological implications this has.

The ASD data from Haney et al., 2020 collected single cell sequencing data from Brodmann Areas 4/6 and 9. These were interesting choices for single-cell sequencing in Autism Spectrum Disorder. While Autistic individuals often face motor impairments, they most commonly have different sensory experiences than controls – particularly auditory and visual. In addition, the lab had found the most prominent impairment in BA17 and BA7 based on bulk ATAC-seq analysis. Given this, Brodmann Areas 17 and 22 likely would have been better choices for sequencing, as they are the primary visual and auditory cortexes.

The quality of the data from Haney et al., 2020 also created some issues in the analysis. We found that the large numbers of apparently low-quality cells and drastically different 10x versions made batch effect correction difficult and less effective than expected. We could take a small subset of samples, choosing higher quality experiments with similar versions. It may also be advantageous to take a by-sample approach to doublet detection and removal. The percentage of doublets created during sequencing varies greatly depending on the number of samples. It could be more accurate to assess the number of cells that were sequenced per donor and assign a predicted doublet formation rate based on that size. However, doublet detecting using DoubletFinder is a fairly lengthy process, so this may not be ideal to implement for experiments with large numbers of donors.

When using all of the samples, there seemed to still be some batch effect that was not corrected, as we could see during the proportion analysis. Some of the donors had nearly 80% of their cells in a single cluster. This makes it very difficult to get the statistical power to determine if the proportions differ significantly. It could also lead to questionable results in differential gene expression analysis.

The differential gene expression analysis should be run on a smaller subset of high quality samples to ensure the mitochondrial gene expression and other signs of inflammation is not a result of cell death and low-quality samples. It would also be advantageous to do a more in-depth pathway analysis on both up-regulated and down-regulated genes in ASD.

For future work, the Raychaudhuri lab is creating a tool called "Crescendo" that will perform differential gene expression analysis after batch correction with Harmony. This tool is not ready to be used but could be implemented once it reaches beta testing or is published. This would allow us to accurately perform this analysis on larger samples than Seurat is able to manage.

We also ran into some technical limitations when running the Lake et al., 2018 THS-seq data and integrating it with the RNA-seq data. The quality of the data was very low. Using ArchR's default quality parameters, only 500 cells were plotted. The quality limits were lowered significantly to plot around 10,000 cells. Because of the low quality of the experiment, there were very few regulatory element features found to be active in each cluster. A typical experiment finds about 10,000 features, while this analysis only found about 350. We would also like to do exploratory analysis connecting regulatory element enrichment with gene expression, but this ultimately is not a high enough quality dataset to attempt it with.

Code availability: https://github.com/nshedd/MQP

Works Cited

Alberts B, Johnson A, Lewis J, et al.,. (2002) Molecular Biology of the Cell. 4th edition. New York: Garland Science;. Chromosomal DNA and Its Packaging in the Chromatin Fiber.

Batut, P., Dobin, A., Plessy, C., Carninci, P., & Gingeras, T. R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome research, 23(1), 169–180. https://doi.org/10.1101/gr.139618.112

Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., & Bonnel, A. M. (2003). Enhanced pitch sensitivity in individuals with autism: a signal detection analysis. Journal of cognitive neuroscience, 15(2), 226-235.

Breeze CE, Lazar J, Mercer T, et al. (2020). Atlas and developmental dynamics of mouse DNase I hypersensitive sites. bioRxiv. DOI: 10.1101/2020.06.26.172718.

Breschi, A., Gingeras, T. R. & Guigo, R. A limited set of transcriptional programs define major histological types and provide the molecular basis for a cellular taxonomy of the human body. Genome Res. (in the press).

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology, 109, 21.29.1–21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Chauhan, A., & Chauhan, V. (2006). Oxidative stress in autism. Pathophysiology : the official journal of the International Society for Pathophysiology, 13(3), 171–181. https://doi.org/10.1016/j.pathophys.2006.05.007

Chen, Y., Norton, D. J., McBain, R., Gold, J., Frazier, J. A., & Coyle, J. T. (2012). Enhanced local processing of dynamic visual information in autism: evidence from speed discrimination. Neuropsychologia, 50(5), 733-739.

Citrigno, L., Muglia, M., Qualtieri, A., Spadafora, P., Cavalcanti, F., Pioggia, G., & Cerasa, A. (2020). The Mitochondrial Dysfunction Hypothesis in Autism Spectrum Disorders: Current Status and Future Perspectives. International journal of molecular sciences, 21(16), 5785. https://doi.org/10.3390/ijms21165785

ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." Nature 583 (7818): 699–710.

Grubert, F., Srivas, R., Spacek, D. V. & Snyder, M. (2020). Landscape of cohesin-mediated chromatin loops in the human genome. Nature. 583, 737–743 https://doi.org/10.1038/s41586-020-2151-x.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. Journal of autism and developmental disorders, 36(1), 5-25.

Holliday, R., & Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. Science (New York, N.Y.), 187(4173), 226–232.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A.

K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, 161(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Marco, E., Hinkley, L., Hill, S. et al. Sensory Processing in Autism: A Review of Neurophysiologic Findings. Pediatr Res 69, 48–54 (2011). https://doi.org/10.1203/PDR.0b013e3182130c54

Meltzer, A., Van de Water, J. The Role of the Immune System in Autism Spectrum Disorder. Neuropsychopharmacol 42, 284–298 (2017). https://doi.org/10.1038/npp.2016.158

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R., Stamatoyannopoulos, J. (2020). Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251 (2020). https://doi.org/10.1038/s41586-020-2559-3

Murray K. (1964). The occurrence of epsilon-N-methyl lysine in histones. Biochemistry. 127:10–15.

Park, P. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680. https://doi.org/10.1038/nrg2641

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. Nature reviews. Genetics, 14(4), 288–295. https://doi.org/10.1038/nrg3458

Phillips, T. & Shaw, K. (2008) Chromatin Remodeling in Eukaryotes. Nature Education 1(1):209

PsychENCODE (2018). Science. Retrieved from https://www.sciencemag.org/collections/psychencode

Roberts, T. P., Khan, S. Y., Rey, M., Monroe, J. F., Cannon, K., Blaskey, L., ... & Edgar, J. C. (2010). MEG detection of delayed auditory evoked responses in autism spectrum disorders: towards an imaging biomarker for autism. Autism Research, 3(1), 8-18.

Robertson, C. E., Martin, A., Baker, C. I., & Baron-Cohen, S. (2012). Atypical integration of motion signals in autism spectrum conditions. PloS one, 7(11), e48173.

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A., Zheng, G., Greenleaf, W. J., Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature biotechnology, 37(8), 925–936. https://doi.org/10.1038/s41587-019-0206-z

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor protocols, 2010(2), pdb.prot 5384. https://doi.org/10.1101/pdb.prot5384

Sos, B.C., Fung, HL., Gao, D.R. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol 17, 20 (2016). https://doi.org/10.1186/s13059-016-0882-7

Tsilioni, I., & Theoharides, T. C. (2018). Extracellular vesicles are increased in the serum of children with autism spectrum disorder, contain mitochondrial DNA, and stimulate human microglia to secrete IL-1β. Journal of neuroinflammation, 15(1), 1-8.

van Eden, W., van der Zee, R. & Prakken, B. Heat-shock proteins induce T-cell regulation of chronic inflammation. Nat Rev Immunol 5, 318–330 (2005). https://doi.org/10.1038/nri1593

Van Nostrand, E. L. Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J-Y., Cody, N.A.L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C. (2020) A large-scale binding and functional map of human RNA binding proteins. Nature. https://doi.org/10.1038/s41586-020-2077-3.

Vlachos M. (2011) Dimensionality Reduction. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_216

Zhang, P., Cao, L., Zhou, R. et al. The lncRNA Neat1 promotes activation of inflammasomes in macrophages. Nat Commun 10, 1495 (2019). https://doi.org/10.1038/s41467-019-09482-6

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., ... & Wang, J. (2019). Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Molecular cell, 73(1), 130-142.

Zhang, Y., Ma, Y., Huang, Y., Zhang, Y., Jiang, Q., Zhou, M., & Su, J. (2020). Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Computational and structural biotechnology journal, 18, 2953–2961. https://doi.org/10.1016/j.csbj.2020.10.007

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., ... & Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. Molecular cell, 65(4), 631-643.