Ranking And Scoring The Critical Cell Types In Neurodevelopmental Disorders Using Genetic Modules

By

Ashleigh Thomas

Thesis

Submitted in partial satisfaction of the requirements for the degree of

Master of Science

in

Integrative Genetics and Genomics

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

_____

Professor Fereydoun Hormozdiari, Chair

_____

Professor C. Titus Brown

_____

Professor David Segal

2021

Abstract:

Neurodevelopmental Disorders (NDDs), including Autism Spectrum Disorder (ASD) and Intellectual Disability (ID) are disorders that are affected by the developing human brain. The brain has hundreds to thousands of unique cell types within it, and by studying the cell types that are critical in NDDs, this will lead to a greater understanding of the mechanisms of NDDs in the developing brain. Single-cell RNA-seq (scRNA-seq) can shed light on the importance of the many cell types involved in NDDs and the normal brain development process, as it can provide more fine-grained details on individual cells in comparison to bulk RNA sequencing.

This thesis project proposes a mathematical objective function that identifies critical cell types for a set of genes and a scRNA-seq dataset. Our objective function was able to identify critical cell types previously identified in literature. A set of ASD risk genes were used as module genes, as well as scRNA-seq data taken from the developing human neocortex for input. Excitatory deep layer neurons (glutamatergic neurons) and microglial cells were found to be of interest with these module genes. A second and third set of module genes were tested, with one being composed of genes indicated in ASD and ID, and the other being composed of genes indicated in both disorders but that are also in synaptic function, long-term potentiation, and calcium signaling that are found to be more highly expressed at postnatal time points. For these sets of module genes, excitatory deep layer neurons (glutamatergic neurons), and cycling progenitors in the G2/M and S phases were found to be critical. Additionally, we show that cells within the same defined cell type (here done using tSNE) have a higher average maximum similarity with each other than with cells outside of their cell type. In

contrast, for a random selection of cells, there is a higher average maximum similarity of cells between groupings, rather than in the same grouping. This indicates that the cell types utilized in this project are in fact clustered properly together. In conclusion, utilizing scRNA-seq in conjunction with module genes enables us to identify critical cell types applicable to NDDs.

Chapter 1: Introduction:

While the understanding of neurodevelopmental disorders (NDDs)—particularly Autism Spectrum Disorder (ASD) and Intellectual Disability (ID)—with respect to cell types as well as genetics has increased in recent years, further research is needed. Of particular interest is the areas of the brain that are highly involved in such disorders. In addition, the impact of developmental timing is crucial in understanding NDDs, as different gestational weeks will be more important for some NDDs in comparison to others. By furthering the research into the critical cell types involved in NDDs, this will advance the understanding of the developmental processes that are impacted in such disorders. In addition, this could shed light upon the biomolecular pathways that are involved in NDDs, which could lead to potential pharmacological treatments in the future. This thesis aims to build upon the knowledge found in literature about critical cell types in NDDs, particularly in ASD and ID.

Neurodevelopmental Disorders:

As per the *DSM-5*, NDDs are a set of conditions that begin in the developmental period (often defined as birth through 2 years of age) which produce deficits impairing functioning (*DSM-5® Handbook of Differential Diagnosis*, 2013; Morris-Rosendahl & Crocq, 2020). NDDs typically include disorders such as intellectual disability (ID), Autism Spectrum Disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), communication disorders, neurodevelopmental motor disorders like tic disorders, and particular learning disorders (Morris-Rosendahl & Crocq, 2020). Neurodevelopmental disorders have been relatively recently re-classified, and in the *DSM-5*, they replaced a category of disorders that were diagnosed in infancy, childhood, and adolescence with

the label of Neurodevelopmental Disorders, or NDDs. I am interested in the study of neurodevelopmental disorders in particular because by understanding them better, we will be able to understand the developing human brain better. By better understanding the cell types relevant to NDDs, we can improve our understanding of the molecular mechanisms and pathways of the developing brain, particularly in regards to brain development, neurogenesis, and diversity of cell types in the brain (Polioudakis et al., 2019; Zhong et al., 2018). In the future, this may lead to pharmacological tools to treat NDDs in patients.

Typically, these disorders will begin in a child's early developmental stages, and will involve deficits impacting functioning (Morris-Rosendahl & Crocq, 2020). The *DSM-5* allows doctors to document their etiological factors for patients with a specifier such as fragile X syndrome, which indicates that NDDs are likely to change in the near future with genetic research (*DSM-5® Handbook of Differential Diagnosis*, 2013). Overall, NDDs tend to be associated with a lower fecundity, so these high-risk genetic variants are likely rare from negative selection, and are a product of selection pressure and from de novo mutations, which is supported by the importance of de novo variants in NDDs such as ASD, where de novo events of causal variants at genetic loci are critical (Morris-Rosendahl & Crocq, 2020).

NDDs tend to be highly comorbid, meaning that often patients may have two or more conditions. Between 22 and 83% of children with ASD also meet the criteria of the *DSM-IV* for ADHD, and between 30 and 65% of children with ADHD have 'clinically significant symptoms of ASD' (Morris-Rosendahl & Crocq, 2020). In ASD, intellectual and/or language impairment is often involved. NDDs are more often diagnosed in males

rather than in females, with the *DSM-5* citing that the male to female ratio in ASD is 4 to 1, 2 to 1 in children with ADHD, 1.6 to 1 for mild intellectual disability, and 1.2 to 1 for severe intellectual disability (Morris-Rosendahl & Crocq, 2020).

Genetic heterogeneity is when one or more related phenotypes can be produced through multiple different genetic mechanisms. The genetic heterogeneity between NDDs is often very high, which in turn means that in the clinic, it can be very difficult to genetically diagnose patients with NDDs, as it may not be immediately obvious which genetic mutation is causing their NDD (Morris-Rosendahl & Crocq, 2020). Also, some NDDs can have their phenotype be caused by a variety of genetic events and also can also be impacted by environmental factors contributing to the phenotype. For most NDDs, there is not a simple genetic cause, so sometimes forward genetics are used to study this. Using forward genetics, a researcher may go from the phenotype to the genotype to the gene. In contrast, they may also utilize reverse genetics, where once they have found genetic markers they can form a phenotype definition based upon the genetic marker, so that different phenotypes can be grouped together with linkage data or deviant allele frequencies in association data analysis (Morris-Rosendahl & Crocq, 2020). ASD is etiologically, phenotypically, and genetically heterogeneous, and the heritability is high as studies have found that it ranges from 50 to 90%, and in siblings who are not twins, the recurrence rate is near 20% (Morris-Rosendahl & Crocq, 2020).

Often times, a genotype-first approach is used in the genetics of NDDs including ASD, such as a researcher first finding copy number variants (CNVs) and gene variants that have a high likelihood of causing disruptions leading to ASD in order to further their genetic research. There is a question of whether NDDs should perhaps be placed upon

a spectrum, due to the high phenotypic overlap amongst the discussed NDDs. A variety of NDDs have multiple symptoms in common (meaning there is a phenotypic overlap between the disorders, often called phenotypic heterogeneity), which at times can make it difficult to distinguish which specific disorder a patient may have.

It has been proposed that ID, ASD, ADHD, schizophrenia, and bipolar disorder could lay upon a continuum of neurodevelopmental disorders, and that they may not be wholly discrete (Morris-Rosendahl & Crocq, 2020). ID, ASD, and ADHD share multiple genetic risk alleles, and they also share allele mutations for psychiatric disorders including schizophrenia. In intellectual disability, for instance, CNVs for ID are significantly enriched in cases of schizophrenia, posing the question of if some ID CNVs confer a risk to schizophrenia at a lower level of penetrance. It is also likely that the neurodevelopmental continuum of NDDs plus schizophrenia represent the diverse outcomes of various events in disrupted brain development (Morris-Rosendahl & Crocq, 2020).

Autism:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder (NDD) with a relatively early age of onset, as symptoms are normally exhibited within three years of birth. Symptoms include a lower level of social interaction and communication, and levels of impact by ASD vary between individuals (Park et al., 2016). An increased prevalence of ASD diagnoses has been observed over time worldwide. In the last twenty years, studies from Australia, North America, Middle Eastern countries, and some European countries have reported more cases of ASD being reported, however the numbers are highly variable. In the US in 2016, it was reported that the prevalence

is thought to be 18.5 cases in 1000 people over 11 states. Boys are diagnosed at a rate 4.3 times that of girls (Chiarotti & Venerosi, 2020).

There are many rare genetic mutations that are thought to contribute to ASD, as well as environmental effects, and the interaction between genetics and environment, although genetics play a much larger role than environment, with a heritability of over 0.7 (Park et al., 2016). The genetics of ASD are relatively complex, for several reasons, as ASD is particularly genetically heterogeneous (An & Claudianos, 2016). Phenotypic heterogeneity is when different phenotypes can occur from the same gene being mutated, which can be found in the *SHANK* gene in which mutations are associated with ASD, ID, and epilepsy (Guilmatre et al., 2014). NDDs are also highly comorbid, meaning that an individual affected by one neurodevelopmental condition is at an increased risk to be affected by co-occurring conditions. These three factors can make it difficult to determine genetic mechanisms at play in NDDs.

As previously discussed, a genotype-first research method is particularly helpful in ASD. More than 100 genes and genomic areas have been associated with ASD, and roughly 800 genes are thought to be involved (Morris-Rosendahl & Crocq, 2020). Additive polygenic factors seem to be in play in ASD, particularly for patients who have less severe clinical symptoms. Genome-wide association studies (GWAS) have shown that common DNA variants are key to ASD phenotypes. Additionally, de novo CNVs in one study were ten times more prevalent in cases of ASD than controls, and de novo ASD-causing CNVs may account for around 30% of simplex ASD cases (Sebat et al., 2007).

Researchers have found recurrent CNVs and disruptive variants in ASD, which has resulted in ASD-specific genetic subtypes being formed, which could lead to potential future pharmacological treatments for these particular pathways involved. For instance, the *CDH8* gene is a heterogenous disruptive variant that is involved in chromatin remodeling. It also targets other ASD risk genes, and a genetic subtype resulting in macrocephaly and gastrointestinal complaints has been linked more tightly with ASD than ID (Morris-Rosendahl & Crocq, 2020). Additionally, there is a 16p11.2 deletion neurologic phenotype that those with this deletion may not have ASD, yet there may be a phenotypic overlap with ASD. The *SCN2A* gene can cause multiple NDDs including benign familial neonatal-infantile seizures, ASD, ID, and infantile epileptic encephalopathy. *SCN2A* encodes for a part of the neuronal voltage-gated sodium channel named NaV1.2 which is used in action potentials. This indicates that sodium channels may be important for future pharmacological therapies for NDDs involving the *SCN2A* gene (Morris-Rosendahl & Crocq, 2020).

A study by Velmeshev et al. in 2019 performed single nucleus RNA-seq (snRNA-seq) on post-mortem prefrontal cortex and anterior cingulate cortex samples of 15 cases of ASD and 16 controls without ASD of similar ages. They found 513 differentially expressed genes between the ASD cases and controls (Velmeshev et al., 2019). In addition, gene expression changes seemed to be largely in the upper-layer cortical neurons in L2, L3, and L4 excitatory neurons as well as microglia (Velmeshev et al., 2019; Wood, 2019). They also found downregulation in genes involved in synaptic signaling and brain development, furthering the case that synapses and development are highly important in ASD. In the microglial cells, the differentially expressed genes

were enriched in genes encoding microglial activation and developmental pathways (Velmeshev et al., 2019; Wood, 2019). In addition, they performed snRNA-seq on prefrontal cortex samples of 8 cases of epilepsy, as epilepsy has a high rate of comorbidity with ASD. They used the differentially expressed genes found in the epilepsy cases to remove gene expression changes from seizures and antiseizure medication from the gene expression changes of the ASD cases, as only 10% of the genes dysregulated in ASD were utilized in the analysis of the epilepsy cases. This may indicate that differentially expressed genes in ASD are more closely related to ASD rather than to the comorbid epilepsy disorders (Velmeshev et al., 2019; Wood, 2019). The researchers indicate that changes on the molecular level of the upper layer cortical circuit seem to be related to how behavior is impacted in ASD. Additionally, it does not seem that the degree of dysregulation in genes associated with a higher clinical severity of ASD can be used to predict the severity of ASD symptoms. Their research indicates that a change in the gene regulatory program in development may cause problems in the molecular pathology of mature neurons (Velmeshev et al., 2019).

Gene modules:

Modules of genes are artificially created groupings of genes (i.e. gene modules are curated most often via knowledge from literature) that are thought to work in similar biological manners, and are therefore involved in similar biological pathways. Such genes do not have to be located physically together on chromosomes. One such definition of a gene module is a grouping of co-expressed genes where the same group of transcription factors bind to the genes, or co-regulation (Bar-Joseph et al., 2003). Oftentimes, researchers utilize module classification algorithms that group genes based

upon their co-expression patterns (Saelens et al., 2018). Many clustering methods have limitations, and to improve upon them, some researchers have used co-expression data from only a portion of their samples, by utilizing regulatory network models, and allowing for overlapping of gene modules. Decomposition methods have been found to work best, and parameter estimation and alternative similarity metrics can be utilized in order to detect the most relevant grouping of module genes (Saelens et al., 2018).

Another way of thinking about modules of genes is that genes responsible for certain diseases and disorders can map to smaller biological subnetworks. One method of forming such modules is known as MAGI which is short for merged affected genes into integrated networks (Hormozdiari et al., 2015). MAGI is a generalizable program which can make gene modules from protein-protein interactions and RNA-seq expression profiles, while previous research tended to only use one of these two techniques. These researchers formed two modules for NDDs, specifically ASD and ID. The modules were formed using MAGI on exome sequencing of 1116 cases of ASD and ID. They found two modules which differed in both genotypes and phenotypes. Module one is composed of 80 genes affiliated with Wnt, NCOR, SWI/SNF, and Notch complexes, and they have the largest gene expression in the 8-16 post-conception weeks period of brain development. The second module is made of 24 genes that are involved in synaptic function, long-term potentiation, and calcium signaling. These genes are not highly expressed in the prenatal timepoint, rather they have higher levels of expression in postnatal time. Within these two modules, they found that cases with ASD and/or ID that had de novo mutations in these two modules were far more clinically and intellectually impacted than other cases with de novo mutations outside of these

two modules, and that these cases with de novo mutations in the two modules also had many more deleterious missense mutations. There appears to be a grouping of neurodevelopmental networks spanning human neurological disorders and diseases that impacts the same sets of genes producing a variety of phenotypes, as they found an overlap and expansion of the same two modules to be involved in epilepsy and schizophrenia as well as ASD and ID (Hormozdiari et al., 2015). It was found that for ASD and schizophrenia, genes with loss of function (LoF) mutations in cases are more highly connected to each other in protein-protein interaction networks, and they also have a higher co-expression pattern with each other. This led to MAGI utilizing protein-protein interaction and co-expression networks together to find 'highly connected modules' of genes that are enriched in cases of NDDs but not in controls (Hormozdiari et al., 2015).

Weighted gene co-expression network analysis, or WGCNA, has been used to find large non-overlapping modules of genes with an average size of 600 genes involved in and highly co-expressed during brain development (Parikshak et al., 2013). They then chose a subset of these genes in their modules that have a higher chance of a de novo mutation that could cause autism. The model was originally formed without protein interaction data, but this was added in after the WGCNA was performed to find the modules. However, originally many of the most important ASD genes such as *CHD8*, *DYRK1A*, and *GRIN2B* were not picked up by this method (Parikshak et al., 2013). Some other studies used particular subsets of ASD genes to form a set, and then they looked for genes with similar co-expression patterns (Gulsuner et al., 2013; Willsey et al., 2013). In one study, researchers were able to find particular sub-tissues of

the brain and time points key in regular neurodevelopmental processes and which are implicated in ASD (Willsey et al., 2013). They used 9 ASD risk genes from literature, and for each of those genes found 20 more genes with the highest levels of co-expression in the relevant brain tissues and gestational week time points. By doing this, they found 4 networks of 437 genes, suggesting they may be involved in ASD (Willsey et al., 2013).

Overall, disease modules can be defined as groupings of genes where the genes are enriched in de novo mutations in cases, and yet are not enriched in de novo mutations in controls (those without the disease). Additionally, these genes must be highly connected in protein-protein interactions, and they must have high co-expression in the relevant tissue. For ASD and other NDDs, it is important to have high co-expression during brain development. The MAGI team found that genes in cases of ASD with de novo mutations have a higher probability of being connected to each other by protein-protein interactions, and to have higher co-expression levels (Hormozdiari et al., 2015). This result was obtained in part by using samples from cases and unaffected control siblings for validation (Hormozdiari et al. 2015). In general, it can be said that modules of genes are defined as groupings of genes that are enriched in similar biological pathways and functions, and are therefore thought to be related.

Single-cell RNA-seq:

Single-cell RNA sequencing (scRNA-seq) was hailed as the ‘Method of the Year’ in 2013 (Lähnemann et al., 2020). The technique enables cell type clusters to be identified and differentiated from the gene expression of the transcriptome of the whole single cell. This also allows scientists to view cells transitioning from one state into

another and the gene expression changes that come with such a transition. This allows for a better understanding of how tissues and organisms develop and of how structures in cell populations can be differentiated when they were previously thought to be homogenous (Lähnemann et al., 2020).

Single-cell RNA sequencing has enabled scientists to obtain the proteomic, epigenetic, genetic, spatial, and lineage data of single cells (Stuart & Satija, 2019). This allows them to find relationships between cells, understand the cellular state, and collect data from single cells from different areas of an individual to better understand their genomic profile. Single-cell RNA sequencing is performed by sequencing cDNA from RNA molecules (mostly polyadenylated mRNAs) of a single cell. This is then done for thousands of cells in a given experiment to provide a wide array of single cell data. Originally, sequencing focused on one type at a time, be it DNA, RNA expression, or chromatin data. This does not allow for insights into the relationship between single cells. Single-cell RNA sequencing allows for the use of barcoding and is very sensitive and multiplexed, which has led to many discoveries about the nature of single cells and the biomolecular relationships between them. This includes genomics, epigenomics, proteomics, and transcriptomics of single-cell data (Stuart & Satija, 2019). scRNA-seq can be coupled with other sequencing technologies like DNA sequencing or epigenomic data.

Single-cell RNA sequencing has recently drastically changed how RNA can be sequenced. Single-cell analysis dataset sizes are increasing in comparison to bulk sequencing sample sizes, which leads to having more information to analyze, which introduces certain statistical challenges that bulk RNA sequencing has not had to utilize

before. Additionally, the scalability of current RNA sequencing analysis has lagged behind data generation (Lähnemann et al., 2020). Single-cell RNA sequencing can find signals that bulk RNA sequencing is unable to, as scRNA-seq provides very specific high resolution data. Additionally, the whole cell tissue is not homogeneous in function, as different groups of cells (called cell types) within the tissue can be performing different functions at the same time, which can also vary by time point. Therefore, by being able to sequence specific barcoded single cells, it is possible to collect gene expression data throughout a whole tissue at one point in time while representing the various cell types throughout the tissue. In contrast, traditional bulk RNA sequencing gives the average gene expression, and the researcher is unable to tell what each individual cell is doing at the time.

scRNA-seq is good for a system like the developing human brain because the brain is filled with highly specialized tissue types, and it is estimated that there are hundreds or thousands of unique cell types in the human brain (Polioudakis et al., 2019). With such a large amount of cell types throughout the many tissue types, that means that in any given tissue type, there will be a large number of distinct cell types being expressed. Therefore, being able to see the gene expression profile of a single cell in a tissue at a given time is very important in differentiating critical cell types. Since the timing of brain development is critical in ASD, it is critical to be able to sequence the RNA of single cells from particular brain tissue types at particular time points in development, which scRNA-seq allows researchers to do (Polioudakis et al., 2019). The following figure represents a framework for utilizing bulk RNA-seq on developing brain tissues, which is expanded to single-cell RNA-seq on developing brain tissues in this

project. Figure 1 depicts a workflow of gene module creation from RNA-seq data, protein-protein interactions, and co-expression networks, which are then used as inputs to our program that maximizes an objective function to seek out critical cell types implicated in NDDs.

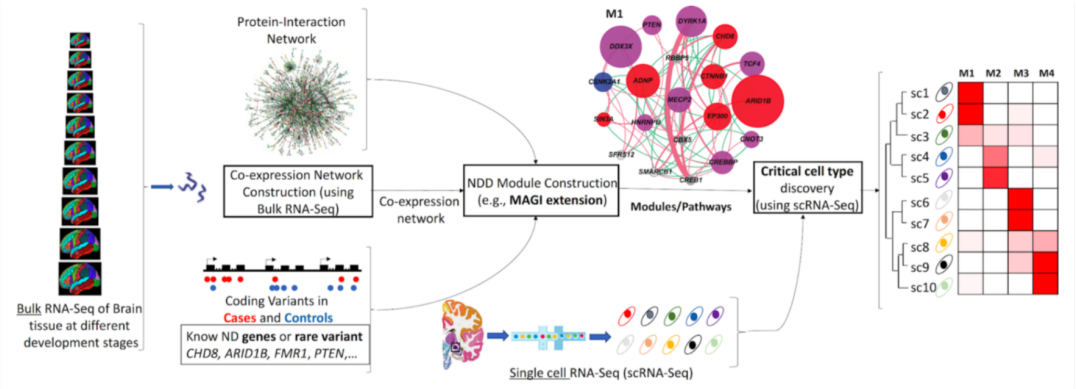

Figure 1: Workflow showing the process of creating a module of genes from RNA-seq data (Hormozdiari et al. 2015) leading into single-cell RNA-seq inputs to an objective function which identifies critical cell types.

Importance of sub-tissues and cell types impacted by neurodevelopmental disorders and Autism:

It is important to identify critical cell types involved in neurodevelopmental disorders such as Autism Spectrum Disorder for several reasons. First, it will help scientists to better understand the human brain and the development it undergoes, both normal and abnormal. Second, by discovering biomolecular pathways that are impacted by NDDs such as ASD, this provides future opportunities for pharmacological treatment for disorders involving these pathways. This research will also help to deconvolute regulatory networks, and it may shed light upon the molecular drivers of differentiation, generation, and development of cell types where these mechanisms are not currently known. Some studies on ASD focus on cortical cell function, and by making molecular taxonomies, this could lead to a better understanding of neurogenesis from cortical cells

from the developing human brain, which will in turn extend knowledge of neurodevelopmental disorders (Polioudakis et al., 2019).

The aim of my thesis is to contribute to the understanding in literature of the critical cell types involved in NDDs such as ASD. Several studies—using different technologies than what I have used in my research—have been performed, and subsequently have reported their findings on critical cell types in ASD. One such study sampled 40,000 cells from the developing human neocortex as fetal samples without known pathogenic CNVs for any neuropsychiatric disorder at 17-18 gestational weeks. The cell samples are taken from the cortical anlage, where there are large germinal zones and cortical laminae undergoing development, which includes new and migrating neurons (Polioudakis et al., 2019). Scientists have connected neurodevelopmental processes at this same time point and place in the brain as being connected with neuropsychiatric diseases (de la Torre-Ubieta et al., 2016; Gandal et al., 2016). In Polioudakis et al. 2019, the cell types were clustered using t-distributed stochastic neighbor embedding (tSNE), which grouped the cells into the areas of the brain from which they originated. Further sub-clusters of cells were found by re-clustering cells within the original clusters to form sub-clusters of cell types. They looked at genes enriched for high confidence ASD risk genes (or genes which have high risk protein-disrupting mutations). They found that the majority of such ASD risk genes were enriched in developing glutamatergic neurons (deep and upper layers) (Amiri et al., 2018.; Parikshak et al., 2013; Polioudakis et al., 2019). Interestingly, there is a high degree of variation amongst individual genes as to where they were most highly expressed. Some genes were highly expressed in inhibitory neurons, excitatory

neurons, or neural progenitors. ASD-risk genes *MYT1L* and *AKAP9* are expressed pan-neuronally, whereas *GRIN2B* was found to be specifically expressed in the glutamatergic neurons. *ILF2* was found to be highly expressed in cycling progenitors. Additionally, some ASD risk genes were found to be extra-neuronal in expression. *SLC6A1* was enriched in pericyte cells, while *TRIO*, *TCF7L2*, *KAT2B*, and *SETD5* were highly expressed in oligodendrocyte progenitor cells. The data indicates that the blood-brain barrier and peri-neural environments may be ASD risk factors as well (Polioudakis et al., 2019). As ASD is highly comorbid with epilepsy and ID, Polioudakis et al. searched for risk genes for epilepsy and ID as well. They found that ID risk genes were involved in glutamatergic neurons and radial glia, which was not the case in either ASD or epilepsy. This may coincide with a more clinically severe phenotype for ID than in ASD or epilepsy (Polioudakis et al., 2019).

Chapter 2: Biological Definition and Mathematical Formulation

Motivation:

Our primary objective in this dissertation is to introduce a novel approach for discovery and ranking of cell types that are "*critical*" in neurodevelopmental disorders. We propose to use the recent progress in discovery of genetic modules/pathways disrupted in NDDs and available single-cell RNA-seq (sc-RNAseq) data from the developing human brain to find the critical cell types in these disorders. Intuitively we are given the modules and/or pathways disrupted in NDDs and the scRNA-seq data from the developing brain, and we define critical cell types for NDDs as a set of cells that (i) have a relatively high amount of expression of the provided genes in the modules and/or pathways and (ii) are part of a same cell-type (or closely related).

Several sets of module genes will be used. One is a set from Polioudakis et al. of 25 ASD risk genes. The second is a set of module genes identified by MAGI entitled Module 1 Extended, which is a module of 80 genes affiliated with Wnt, NCOR, SWI/SNF, and Notch complexes implicated in ASD and ID. Module 2 Extended is made of 24 genes that are involved in synaptic function, long-term potentiation, and calcium signaling and that are also implicated in ASD and ID (Hormozdiari et al., 2015). Previous research into the cell types impacted in NDDs and in particular ASD include deep layer excitatory neurons or glutamatergic neurons, oligodendrocyte progenitor cells, and cycling progenitor cells (Amiri et al., 2018; Parikshak et al., 2013; Polioudakis et al., 2019). Our aim in this project is to identify critical cell types implicated in NDDs utilizing single-cell RNA-sequencing data in a novel way. An objective function will be calculated for cells, and the groupings of cells maximizing this function will be defined

as critical cell types. Figure 2 describes the inputs and outputs of the objective function that comprises the program implemented in this project.

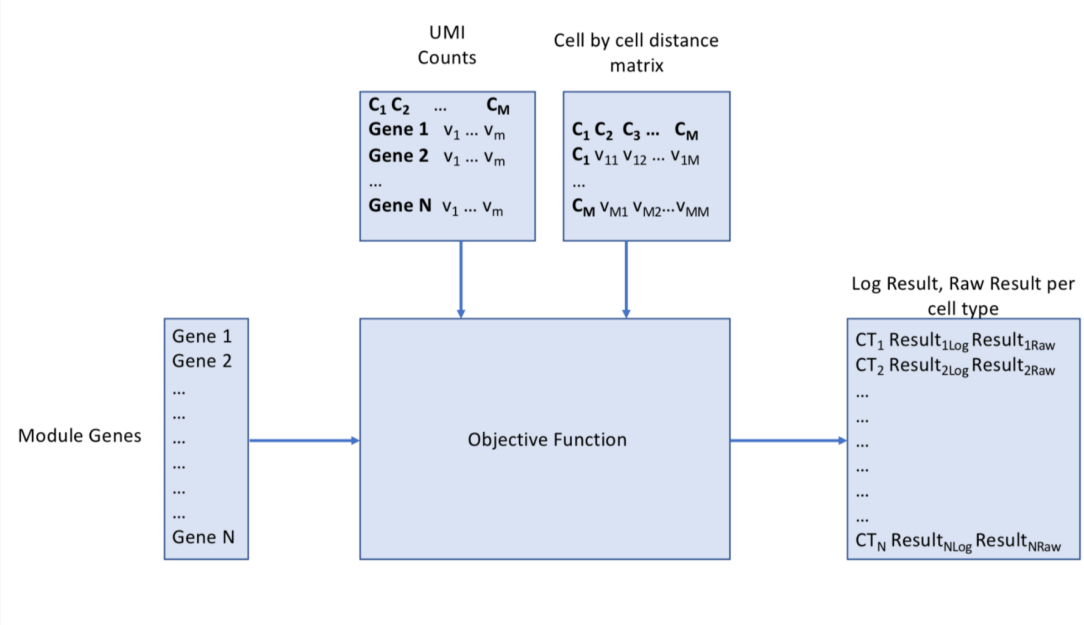

Figure 2: figure displaying the main input and output files of the objective function program in a visual way.

The objective function will try to find the optimal set of cells that maximizes the objective function output. The program implementing the objective function receives the list of module genes of interest, the single-cell RNA-seq data with unique molecular identifier (UMI) counts), cell by cell distance data, and in the future gene co-expression data. One floating point value per cell type cluster will be returned. A higher returned value will indicate that this cell type is correlated with the disorder underlying the inputted data. Therefore, in order to find critical cell types for a particular disorder, the largest objective function values indicate that those cell types are implicated in the disorder. The larger the objective function score, the higher the likelihood that this cell type is implicated in NDD.

We will introduce C to denote the set of single cells, E_c refers to the normalized expression calculated per gene per cell, and M_c refers to the pairwise cell by cell distance matrix calculated using Seurat packages (Butler et al., 2018). Software versions are as follows: Seurat 2.3.4, dplyr 1.0.2, patchwork 1.0.1, and Matrix 1.2-18. G refers to the set of genes in a module, for example a set of NDD risk genes. The objective is to find the subset of cells $S \subset C$ that maximizes the probability that the selected cells are "critical" given the module genes G and gene expression E_c. More formally we want to solve the optimization problem $max_{S \subset C} P(S|G, E_c)$. This can be thought of as an attempt to maximize for $S \subset C$ the probability that the subset of cells S are critical given module genes G and gene expression E_c, or $max_{S \subset C} P(S|G, E_c)$. Bayes' Theorem for three events can be expressed as $P(S|G, E_c) = \frac{P(G|S, E_c) \cdot P(S|E_c)}{P(G|E_c)}$. Therefore, this can then be thought of as $max_{S \subset C} P(G|S, E_c) \cdot P(S|E_c)$ by assuming that $P(G|E_c)$ is constant and independent of the subset of cells S. Here, $P(S|E_c)$ can be simplified to $P(S|M_c)$ where M_c refers to the pairwise distance between any two cells. $P(G|S, E_c)$ is the probability that genes within the module G are expressed when viewing only the subset of cells $S \subset C$. Then reincorporating the right hand side, the chain rule of Bayes' Theorem is used to produce the following formula: $[\prod_{k=1}^{n} P(g_k| \bigcap_{j=1}^{k-1} g_j, S, E_c)] \cdot P(S|M_c)$.

Gene co-expression data could be potentially added in to improve objective function results. The interactions between cell types can lead to a broader understanding of the roles of cell types in systems (McKenzie et al., 2018). Gene co-expression has been utilized in the creation of gene modules particularly in NDDs by identifying genes that are highly co-expressed in brain development (Hormozdiari et al., 2015; Parikshak et al., 2013). This objective function can likely be improved upon by utilizing gene co-expression data in the future. It would likely be included as another multiplicative factor to the existing objective function.

Chapter 3: Methods

Objective function implementation:

A statistic called specificity index has been utilized to find genes that are enriched in certain cell types across many cellular profiles (Dougherty et al., 2010). This study aimed to measure the specificity index of central nervous system microarray data to find genes enriched in each cellular population. Their formulation takes into account variations in transcript numbers enriched in different cell types, and the descended sorting of rankings relies upon the overall size of the comparisons, which in turn relies upon how many genes are expressed and the filtering applied to cell types. This means that the raw specificity index values cannot be directly compared amongst cell types, and instead a p value through permutation testing must be calculated, forming a 'simulated probability distribution' (Dougherty et al., 2010). This probability distribution then can be assigned to specificity index values, and genes enriched in a particular cell type can be identified. The specificity index threshold (pSI) that finds the probability that each group of cells has a significant expression in the module. Their formulation is as follows: $SI_{n,1} = \frac{\sum_{k=2}^{m} rank(\frac{IP_{1,n}}{IP_{k,n}})}{m-1}$. $IP_{1,n}$ refers to the gene expression value for a gene named n, where the rank is defined as a descended ordering. M refers to the number of cells. This thesis project builds upon the specificity index threshold.

Motivated by Dougherty et al., we will approximate $P(g_k|g_j, S, E_c)$ following the CSEA approach. The previous function utilizing Bayes' Theorem is then correlated with the following function implementation that utilizes ranking:

$[\prod_{k=1}^{m} \frac{\sum_{p=2}^{m} rank(UMI(g_{k,p}))}{(m-1)}(\frac{1}{|G|})] \cdot P(S|M_c)$. The left hand side of this function, however, is not strictly a probability, it is simply correlated with the previous probabilities. This is the function that this project seeks to maximize in order to identify critical cell types in NDDs.

The rank can be defined (with rank of 1 being the smallest) as the rank of a gene compared to all other genes in the human genome. For a given gene at a given cell type, the rank is the position of the ordered UMI counts, which is then compared to the sum of the UMI counts for all other genes in the genome. $P(g|s, E_c)$ can then be calculated by summing the ranks found above for all cell types k for a particular gene, which is then divided by the total number of cell types m minus 1. A higher ranking means that a gene is more highly expressed in this particular cell type. Ranking is particularly useful as it allows the formulation to not become overwhelmed by disproportionately large numbers, and normalizing by $(m-1)$ also dampens the effect of large numbers. The goal of ranking is to pick out differences in comparisons amongst cell types in which a given gene is not expressed or less expressed when compared to other cell types.

Datasets and pre-processing of data:

I have utilized the Polioudakis et al. 2019 dataset. This dataset includes 40,000 cells taken from the developing human neocortex between 17 and 18 gestational weeks. The cells fall under the cell type grouping of microglia, pericyte, endothelial, OPC (oligodendrocyte progenitor cells), interneuron caudal ganglionic eminence (CGE), interneuron medial ganglionic eminence (MGE), excitatory deep layer 2 neurons,

excitatory deep layer 1 neurons, maturing excitatory upper enriched neurons, maturing excitatory neurons, migrating excitatory neurons, intermediate progenitor cells, cycling progenitors in the G2/M phase, cycling progenitors in the S phase, outer radial glia (oRG), and ventricular radial glial (vRG) cells. Drop-seq was utilized to analyze the cell samples, which produced cellular profiles. This produced gene expression profiles for the 40,000 cells via UMI counts. This is the dataset used in this project, and the link to the data can be found in their paper.

Some amount of preprocessing was performed in order to align gene data. Any cells with a cell type labeled 'UNK' (unknown) were removed from the dataset for analysis purposes. The file raw_count_mat is the UMI count data, which is a sparse matrix in an rdata file. This was converted to a matrix utilizing the Matrix packing in R. Next, gene names were standardized across multiple datasets. I considered—and will in the future—including two more datasets in the analysis (La Manno et al., 2016; Zhong et al., 2018). This requires that the gene symbols be standardized across all three of the datasets, otherwise a gene with multiple names may be labeled with one name in one dataset, a second name in a second dataset, and possibly a third name in a third dataset, and so forth. Non-protein-coding genes were removed from all input files. The first step was to find genes that have the same name in all three datasets. The second step was to find the subset of genes that do not exist in all three datasets. This list of genes was fed into the multi symbol checker which can be found at https://www.genenames.org/tools/multi-symbol-checker/. This tool returns two outputs: a file that contains genes with name matches, and a file that contains genes without name matches. For genes that have a match, this can be either labeled as an approved

symbol, previous symbol, or an alias symbol. Any genes with an approved symbol were kept as is in the datasets. When an inputted gene appears only once and its alias symbol is a different gene, the gene is renamed to the alias symbol returned. If an inputted gene appears more than once and has one previous and one alias symbol, the previous symbol is used for renaming the input. If an inputted gene appears more than once with two alias symbols, the gene was manually queried via genecards.org, and the gene name with the highest score was accepted.

Expression fraction can be defined as for a particular gene at a particular cell type at a particular time point, the number of cells having the gene expressed (i.e. UMI >= 1) divided by the total number of cells grouped in that cell type. I excluded cell types with ten or equal cells per cell types. This expression fraction was calculated as above on the UMI counts of single cells from published datasets. Cell type clusters in the Polioudakis dataset were used, and expression fractions were calculated for each gene per cell type and gestation week time combination.

Once expression fractions were calculated and gene naming conflicts were resolved, the raw UMI counts were loaded into Python in a genes by cells format. Then, this data was used as input to Seurat in order to calculate a cell by cell distance matrix. This was done in a similar manner as in Polioudakis et al., 2019. Figure 3 illustrates the pipeline that this process follows. A Seurat object was created from the raw counts matrix, the data was normalized using a LogNormalize method with a factor of 10,000. The FindVariableGenes function was then called, along with ScaleDataR, and then PCA was run on the input data with a pcs.compute argument of 40. Clusters were formed with dimensions of 1:40 and save.SNN equal to true. tSNE was then run on the clusters,

and then the clusters were written to a file. A co-expression matrix was then created using Pearson's correlation coefficient and the expression fraction calculations, which also takes as input a file of transcript names and gene names, and a file of expression values. This will potentially be used to further improve the objective function in the future.

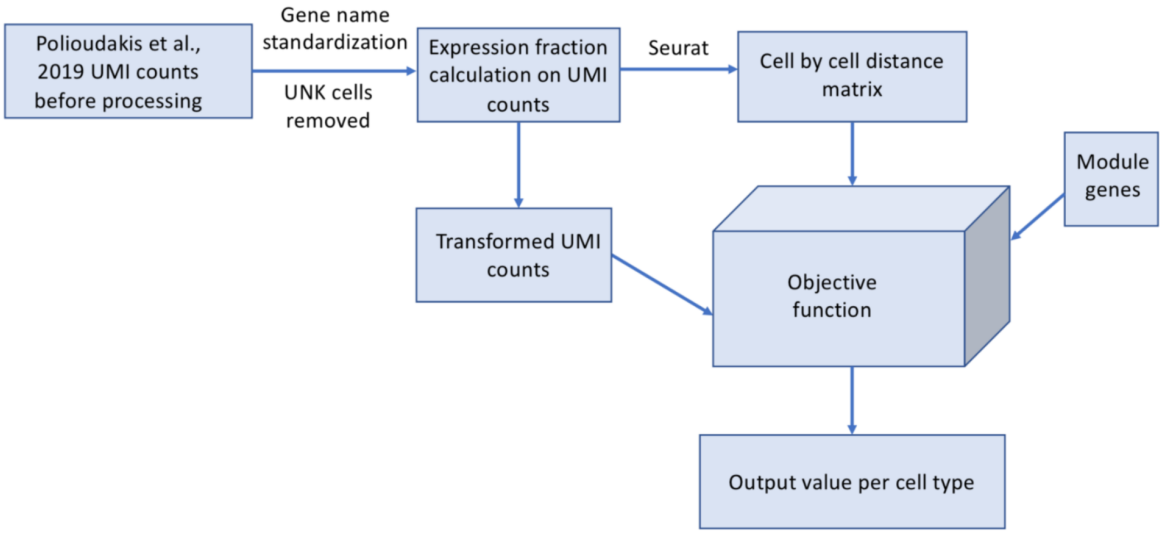

Figure 3: The pre-processing steps illustrated as performed on the Polioudakis et al., 2019 dataset for input into the objective function program.

Assuming cell type clusters are given:

Let us assume for now that the cell types are given in the inputted dataset (i.e. that we know the cell type clusters and the barcodes in each cluster). For instance, this thesis project uses data from Polioudakis et al. 2019 containing 40,000 cells from the developing human neocortex. They have created clusters of cell types—and additionally sub-clusters of those same cell types—that include microglia, pericyte, endothelial, OPC (oligodendrocyte progenitor cells), interneuron caudal ganglionic eminence (CGE),

interneuron medial ganglionic eminence (MGE), excitatory deep layer 2 neurons, excitatory deep layer 1 neurons, maturing excitatory upper enriched neurons, maturing excitatory neurons, migrating excitatory neurons, intermediate progenitor cells, cycling progenitors in the G2/M phase, cycling progenitors in the S phase, outer radial glia (oRG), and ventricular radial glial (vRG) cells (Polioudakis et al., 2019).

In this dataset, the cell type clusters are known and labeled, therefore the right hand side of the objective function $P(S|M_c)$ which represents the probability of cells S being critical cells given the cell pairwise distance matrix, is equal to 1, given that it is now known whether or not S is composed of critical cells. Therefore, the crucial portion of the objective function that I want to optimize becomes the left hand side of the function laid out in the previous chapter.

Assuming cell type clusters are not given:

Now let us assume that the data is no longer labeled with which cells belong to which cell types. This may be the case due to a dataset being unlabeled with regard to cell types, or the cells within a cell type may be in different stages of the cell cycle and we might therefore be interested in these differences, or there is some other unknown factor that may differentiate cells from each other. This is perhaps a more real-world example, where a scientist has taken samples of cells, obtained their single-cell RNA-seq expression data, and now wants to analyze what the critical cell types involved in this set of module genes and gene expression data. Now, it cannot be assumed that the right hand side must be equal to one in this case. Therefore, the right and left hand sides must both be calculated in this instance.

Simulated annealing:

A simulated annealing approach will be used in the future to approximate the most likely critical cell types for the inputted dataset. Here, annealing refers to thermodynamics involved in metallurgy. A set of module genes of interest will be the input (along with their co-expression data), and a random group of cells from the single cell RNA-seq dataset will be chosen as the initial set. The objective function value for this data will be calculated, and then while the temperature is greater than our minimum temperature allowable, a cell will be swapped in randomly and the new objective function score will be calculated. If this results in a higher objective function value, the cell will be swapped into the group. There is also a chance, dependent on the current temperature, to move to a worse neighbor (or a cell that will give a lower value). The temperature decreases incrementally as each new cell is considered. Once the temperature has sufficiently lowered, a set of cells will be returned that maximize the probability of observing the selected set of cells given expression data and the inputted module. The output of this will be the approximate set of critical cell types relevant to NDDs.

Availability: The github repository for the objective function program can be found at https://github.com/ashleighthomas/criticalCellTypes as well as at Zenodo with the following DOI: http://doi.org/10.5281/zenodo.4737835.

Chapter 4: Results

In order to test the accuracy of our objective function, I performed several experiments. It is important that our objective function is able to identify critical cell types as previously defined by literature, given that this project's novel contribution is using single-cell RNA-seq data which has not yet been used in identifying critical cell types for NDDs. Therefore, I first compared our objective function's results to the results of Polioudakis et al. from 2019 on a set of module genes that they identified as ASD risk genes. Then, I found the correlation coefficient between our objective function's results and the Polioudakis et al. paper's results. It is important to note that the results are correlated with each other, yet they are not completely similar, as our approach differs from previous approaches, so it is not expected that the correlation coefficient would be particularly close to 1. Finally, I ran the objective function on 100 different sets of 9995 randomly chosen barcoded cells which were substituted for the previously used cell types as defined by Polioudakis et al. The purpose of this experiment was to determine if our results with true cell types (i.e. cell types that have been proven to cluster together and are similar) is better than a random grouping of cells from a wide variety of cell types. The size of the cell set (9995) was chosen because the largest cell type set in Polioudakis et al. is of 9995 cells. For this experiment, it was important to compare the average maximum similarity of cells within the same cell type compared to cells outside of the same cell type. For randomly chosen barcodes that do not represent a true cell type, I would not expect that the comparison within the same cell type to be better than the comparison outside of the same cell type.

In another experiment, I ran our objective function on the Module 1 Extended module genes, with the 16 cell types as defined by Polioudakis et al. The purpose of this is to see if the average maximum similarity comparison is larger within a cell type than outside of a cell type. It is expected that within one cell type will be greater than outside of a cell type, as the 16 given cell types were found to be clustered together, and therefore have a higher similarity to each other than to cells outside of the cell type. I then ran the objective function on Module 1 Extended and Module 2 Extended in order to identify the top critical cell types for each of these modules. Table 1 describes the input datasets this project uses.

Input File	Description
Module Genes	List of module genes of interest, one per line
UMI count data (Polioudakis et al. 2019 for this project)	UMI counts per gene (per barcoded cell)
	40k cells from developing human neocortex
	17-18 gestational weeks
Cell by cell distance matrix	Distance between cells generated by Seurat
	Made from Polioudakis et al. 2019 dataset

Table 1: Table describing input datasets for this objective function

Experiment comparing objective function performance to Polioudakis et al. 2019:

In order to compare how my objective function performs in comparison to results published in literature, I recreated a figure from Polioudakis et al. 2019. This figure (figure 7A) shows a heatmap of ASD risk genes as rows and cell types as columns. Polioudakis et al. displays the z-scored UMI counts per gene per cell in a cell type. I recreated their experiment by first calculating the z scores of the UMI counts of each cell per gene, and then averaged these across a cell type so that a square in the heatmap will be only one color, rather than one color per cell in the cell type. I then ran

my objective function with the same set of ASD risk genes as the module genes, and the same UMI count data. This data along with the cell by cell distance matrix was plugged into a modified objective function, where only the ranking of genes was performed. The values within a cell type were independently z-scored, and then averaged to create an output of one value per cell type per gene. A heatmap was created of this data, and compared with the similarly-created recreation of the Polioudakis figure.

Figure 4 represents the heatmap that I created using the methods described in Polioudakis et al. 2019 to create this heatmap, except I returned an average value per cell type per gene rather than one value per cell per gene. Figure 5 represents the previously described heatmap made of ranked UMI counts of module genes having been z-scored independently per cell type and then averaged across a cell type to return a value per gene per cell type. Visually, these heatmaps look similar, but not identical to each other. Figure 5 shows a broader array of values, and there are more scores that are in the middle of the range, displaying patterns that are not viewable in Figure 4.

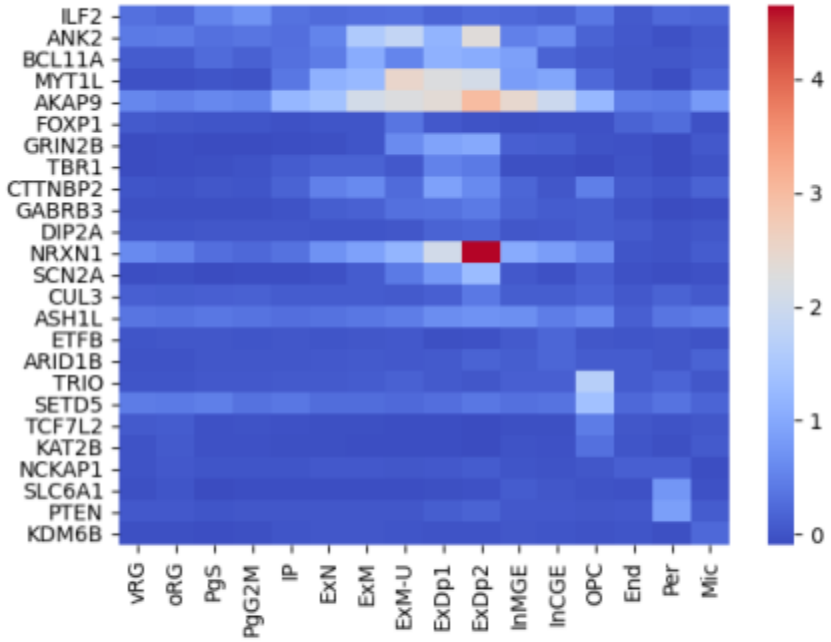

Figure 4: Heatmap showing z-scored UMI counts per 16 cell types from the developing human neocortex of 25 ASD risk genes as defined by Polioudakis et al. 2019.

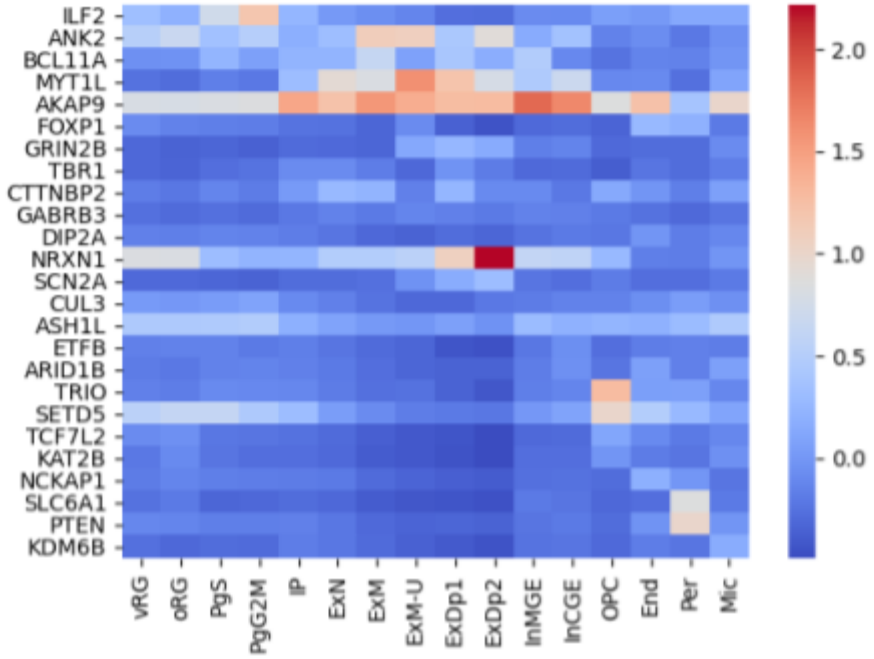

Figure 5: Heatmap showing the ranked UMI counts (from the objective function) of the same set of 25 ASD risk genes z-scored per cell type and average across a cell type.

Experiment finding correlation coefficient between Polioudakis and objective function output:

The data from the previous experiment (prior to feeding it into a heatmap representation) was then used to compare the Polioudakis scores with my objective function scores. This was done utilizing Pearson's correlation coefficient to provide a statistic to compare the two techniques with.

Following the previous experiment, the z-scored and averaged per cell type per gene output from my objective function was compared to the output of the Polioudakis method that took UMI counts and z-scored them. I calculated Pearson's correlation

coefficient of these two matrices with each other, which returned 0.868833686064852, showing a correlation between the two methods.

Experiment evaluating the objective function on random barcoded cells:

In order to analyze the results of the objective function on pre-existing cell types, it is necessary to also compare this to the results of the objective function on randomly chosen barcodes. 9995 barcoded cells from the entire Polioudakis et al. 2019 dataset (9995 being the number of barcoded cells in the largest cell type group of ExN representing newly created excitatory neurons) were randomly selected to represent a critical cell type. This was then fed into the objective function using the M1 extended MAGI genes as module genes. This was then run 100 times using a different set of random barcoded cells each time. Of particular interest is the relationship between the average maximum similarities of cells compared to only cells outside of the cell type (inter comparison) and the average maximum similarities of cells compared only to cells within that cell type cluster (intra comparison). This was then compared with the 16 cell types as laid out by Polioudakis et al. being inputs to the objective function with the same M1 extended gene set as module genes.

This experiment was run 100 times on 100 different sets of 9995 randomly selected barcoded cells from the Polioudakis et al. 2019 dataset and the M1 extended module genes. Each of the 100 times, the inter comparison was higher than the intra comparison, meaning that the average maximum similarity of cells from one cell type compared to cells outside of the cell type was always higher than the average maximum similarity of cells compared to cells inside the same cell type. This is important, because it is expected that cells within a cell type cluster are more similar to each other than to

cells outside of the same cell type. Since these 9995 barcodes were not all from the same cell type, they do not represent a true cell type, meaning that I do not expect that cells that were randomly selected to have a higher average maximum similarity with each other than compared to cells outside of the randomly selected group. Figure 6 shows the results of this experiment.

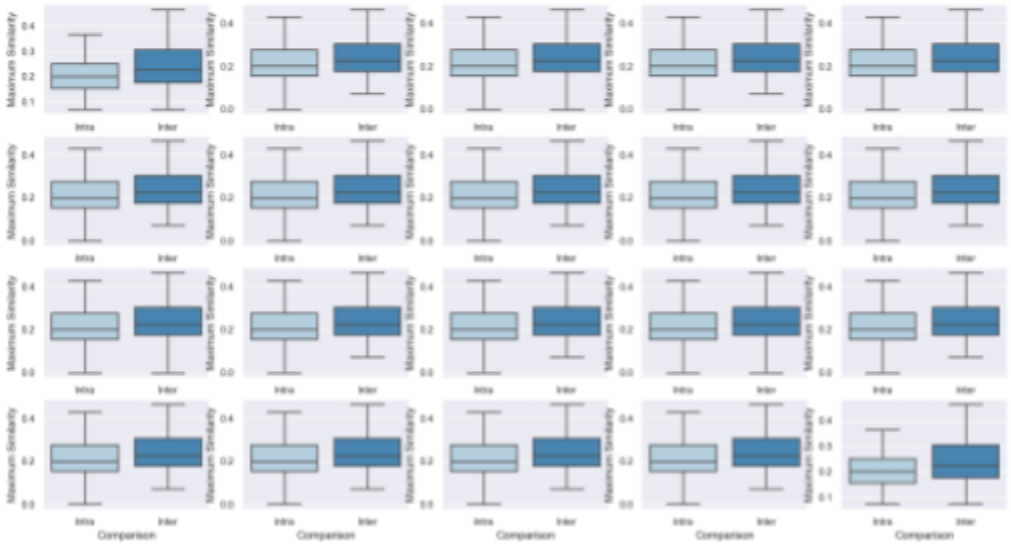

Figure 6: Barchart showing 20 iterations of choosing 9995 randomly selected barcodes and performing intra vs. inter average max similarity comparisons, where each time inter scored higher than intra. This observation holds true for all 100 iterations that were performed.

Experiment evaluating the objective function using pre-existing cell types:

I then evaluated the objective function utilizing the 16 cell types found in the tSNE clustering process (Polioudakis et al., 2019). For each run, one value per cell type will be produced by the objective function. This means that since I ran this with 16 cell types, I can expect 16 floating point values in return. The highest-scoring cell types can be classified as being critical for the module genes and UMI counts that were used. I

ran this for the MAGI M1 Extended ASD with ID and M2 Extended ASD with ID module gene sets, along with the Polioudakis UMI count data. For the M1 Extended module gene set, I also ran this same experiment with intra and inter comparisons of average maximum similarities in order to compare this with the 100 randomly selected groups of 9995 barcoded cells.

I conducted an experiment to further our conclusions that cells within the same cell type cluster should have higher average average maximum similarity scores when cells of the same cell type are compared to each other (intra comparison), rather than when cells of one type are compared to cells of another cell type (inter comparison). This is because cell type clusters should intuitively be composed of cells with a higher similarity to each other than to cells outside of the cell type. I compared intra and inter average maximum similarities for each of the 16 cell types in the Polioudakis et al. dataset. In each of the cell types, the comparison within the same cell type (i.e. comparing average maximum similarities within the vRG cell type as opposed to cells inside vRG compared to cells outside vRG) is higher than comparing the average maximum similarity of cells outside of the cell type. Figure 7 shows these results.

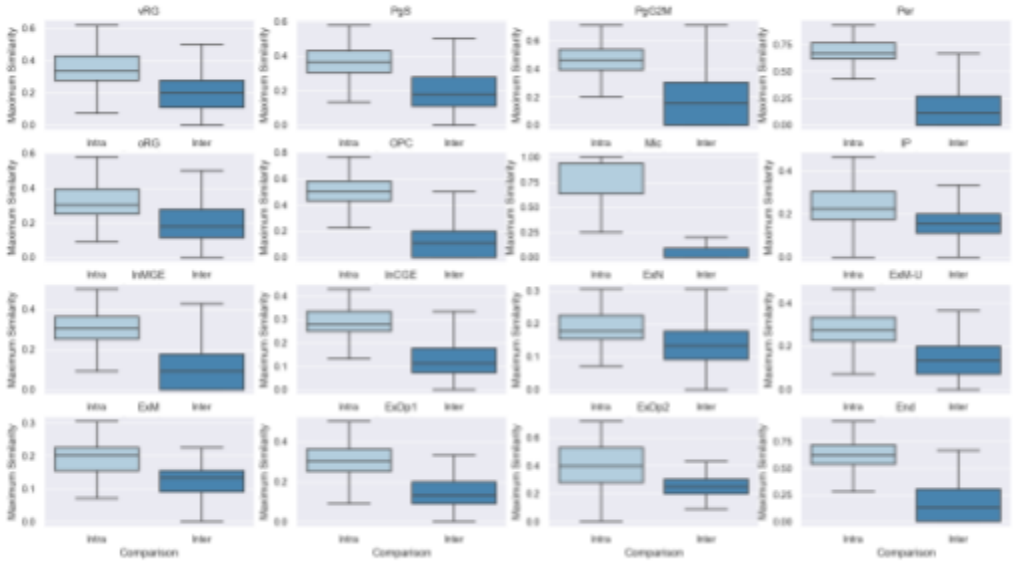

Figure 7: Barchart showing intra vs. inter average max similarity comparisons on 16 cell types as identified by Polioudakis et al. 2019, where for each cell type intra scored higher than inter.

Experiment evaluating critical cell types for M1 Extended and M2 Extended:

I then ran the objective function on the M1 Extended module as defined by MAGI (Hormozdiari et al., 2015) which is composed of ASD and ID risk genes, and then again on the M2 Extended module. The highest rated cell types in descending order for Module 1 Extended are PgG2M (cycling progenitors in the G2/M phase) and PgS (cycling progenitors in the S phase). The highest rated cell types in descending order for Module 2 Extended are ExDp2 (excitatory deep layer 2 neurons) and ExDp1 (excitatory deep layer 1 neurons).

In literature, cycling progenitors are cited as being enriched in gene expression in genes implicated in ASD, particularly in *ILF2* (Polioudakis et al., 2019). Other genes are found to be highly expressed in oligodendrocyte progenitor cells, particularly genes

TRIO, *TCF7L2*, *KAT2B*, and *SETD5*. Radial glial cells have shown higher expression in genes implicated in ID rather than ASD (Polioudakis et al., 2019). Glutamatergic and excitatory deep layer neurons have been shown to have higher levels of gene expression for ASD risk genes (Amiri et al., 2018; Parikshak et al., 2013; Polioudakis et al., 2019). Our results for Module 1 Extended and Module 2 Extended match the critical cell types identified in literature.

Chapter 5: Conclusion

Neurodevelopmental Disorders include Autism Spectrum Disorder, Intellectual Disability, and a variety of other disorders that are impacted during the developmental period in the human brain (*DSM-5® Handbook of Differential Diagnosis*, 2013). Further research on these cell types is needed. Single-cell RNA-seq allows for a more detailed set of RNA-seq data in comparison to bulk RNA-seq, as it allows for many more cells to be sequenced per sample than previously. The human brain contains an extraordinary number of specialized cell types, therefore it is critical to be able to gain as much information as possible on the cell types in the brain that are tied to NDDs. By utilizing sets of module genes related to NDDs in combination with scRNA-seq data from the developing human neocortex, this project explores the concept of ranking and scoring critical cell types involved in NDDs.

In this project, we developed an objective function to rank and score cell types involved in gene modules and scRNA-seq data that are given as inputs. For example, a set of 25 ASD risk genes were used as module genes along with a scRNA-seq dataset of 40,000 cells taken from 17 to 18 gestational weeks from the cortex, which cluster into 16 cell types using tSNE (Polioudakis et al., 2019). This scRNA-seq dataset was used in addition to transforming it into a cell by cell distance matrix, which measures the distance between cells in the dataset. By using these as inputs, I was able to obtain similar results to literature in regards to critical cell types for these ASD risk genes (Polioudakis et al., 2019). Further, when the modified objective function values were plotted as a heatmap, more detailed patterns were detected when using our objective

function as opposed to a different function proposed in the literature (Polioudakis et al., 2019).

Utilizing the same scRNA-seq dataset, I also ranked and scored the critical cell types involved in other sets of module genes. I used two sets of module genes from MAGI (Hormozdiari et al., 2015). These modules were both made up of genes implicated in NDDs. For both modules, the top scoring cell types identified by our objective function included excitatory deep layer neurons (glutamatergic neurons) and cycling progenitors in the G2/M and S. These cell types have also been identified in the past as being critical cell types implicated in NDDs (Amiri et al., 2018; Parikshak et al., 2013; Polioudakis et al., 2019).

Future directions:

Addition of more cell type datasets:

The addition of multiple further scRNA-seq datasets would improve the range of cell types that could be run in this project. For instance, by adding more UMI count datasets and utilizing the same sets of module genes that have been implicated in NDDs, this potentially could help identify further cell types that are critical in NDDs. This project is already prepared to add in two more scRNA-seq datasets, as the gene names have been normalized throughout all three datasets. The second dataset—in addition to the Polioudakis et al., 2019 dataset—is from La Manno et al., 2016. These samples were taken from the ventral midbrain from 6 to 11 gestational weeks, containing 1900 single cell samples, which were clustered into 25 cell types. A third dataset that is available is Zhong et al. 2018. This dataset includes 2300 single cells from the prefrontal cortex from 8 to 26 gestational weeks, which can be clustered into 35 cell

types. Adding in these two single-cell RNA-seq datasets will likely offer a larger sampling of cell types from the developing human brain, which has the potential to enrich knowledge of critical cell types for NDDs.

Addition of different types of datasets:

Multiple additional types of data could be added as input to the objective function. One such type of data is gene co-expression data. The following formulation has been proposed to include gene co-expression data:

$$[\prod_{k=1}^{m} \frac{\sum_{p=2}^{m} rank(UMI(g_{k,p}))}{(m-1)} \cdot \sum_{j=1}^{k-1} coExp(g_k, g_j)(\frac{1}{|G|})] \cdot P(S|Mc)$$

. Here, $coExp(g_k, g_j)$ refers to the co-expression of gene g_k with gene g_j which is added as another term in the multiplication function as previously discussed. Gene co-expression has been shown to lead to a more detailed understanding of cell types in systems (McKenzie et al., 2018), and has been used in the creation of sets of module genes implicated in NDDs (Hormozdiari et al., 2015; Parikshak et al., 2013). Therefore, the addition of gene co-expression data has been implicated as a method to increase understanding of cell types, particularly in NDDs, and should therefore be considered in the objective function implementation in the future.

Additionally protein-protein interaction (PPI) data could be added to the objective function in the future. This data has been used in conjunction with gene co-expression in the creation of NDD modules (Hormozdiari et al., 2015). PPI datasets are often incomplete, and can have higher levels of bias than gene co-expression data, but the relationships shown are still valuable (Hormozdiari et al., 2015), therefore this may be a good data type to include in the future if it is available along with scRNA-seq data.

Future experiment evaluating the objective function without pre-existing cell types:

In the future, the objective function will be run without the cell type clusters being known. This means that—likely through simulated annealing—groups of cells will be formed by running the objective function on a random subset of cells, where that random subset is further refined to come up with an approximate maximum of critical cell types. From there, the objective function will then return a value per each of the cell types that were found via simulated annealing.

References

1. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications* **9**, 1090 (2018).

2. Park, H. R. *et al.* A Short Review on the Current Understanding of Autism Spectrum Disorders. *Experimental Neurobiology* **25**, 1–13 (2016).

3. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).

4. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**, 785-801.e8 (2019).

5. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature Communications* **10**, 4667 (2019).

6. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods | Genome Biology. *Genome Biology* **21**, (2020).

7. Valentine, A. Z. *et al.* A systematic review evaluating the implementation of technologies to assess, monitor and treat neurodevelopmental disorders: A map of the current evidence. *Clinical Psychology Review* **80**, 101870 (2020).

8. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat Med* **22**, 345–361 (2016).

9. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Nathaniel Heintz. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells | Nucleic Acids Research | Oxford Academic. *Nucleic Acids Research* **38**, 4218–4230 (2010).

10. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Scientific Reports* **8**, 8868 (2018).

11. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).

12. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**, 1337–1342 (2003).

13. Livingston, L. A. & Happé, F. Conceptualising compensation in neurodevelopmental disorders: Reflections from autism spectrum disorder. *Neuroscience & Biobehavioral Reviews* **80**, 729–742 (2017).
14. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nature Communications* **11**, 774 (2020).
15. *DSM-5® Handbook of Differential Diagnosis*. (American Psychiatric Publishing, 2013). doi:10.1176/appi.books.9781585629992.
16. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science | Genome Biology. *Genome Biology* **21**, (2020).
17. Chiarotti, F. & Venerosi, A. Epidemiology of Autism Spectrum Disorders: A Review of Worldwide Prevalence Estimates Since 2014. *Brain Sci* **10**, (2020).
18. An, J. Y. & Claudianos, C. Genetic heterogeneity in autism: From single gene to a pathway perspective. *Neurosci Biobehav Rev* **68**, 442–453 (2016).
19. Lalli, M. A., Dennis Avey, Joseph D. Dougherty, J. M., & Robi D. Mitra. High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals convergent mechanisms altering neuronal differentiation. *bioRxiv* (2020) doi:https://doi.org/10.1101/862680.
20. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).
21. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
22. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20**, 257–272 (2019).
23. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).
24. Morris-Rosendahl, D. J. & Crocq, M.-A. Neurodevelopmental disorders—the history and future of a diagnostic concept. *Dialogues Clin Neurosci* **22**, 65–72 (2020).
25. Pennington, B. F. & Peterson, R. L. Neurodevelopmental Disorders: Learning Disorders. in *Psychiatry* 765–778 (John Wiley & Sons, Ltd, 2015). doi:10.1002/9781118753378.ch46.

26. Wood, H. Novel insights into autism from single-cell genomics. *Nature Reviews Neurology* **15**, 434–435 (2019).
27. Velmeshev, D. *et al.* Single-cell genomics identifies cell type–specific molecular changes in autism | Science. *Science* **364**, 685–689 (2019).
28. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
29. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
30. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* **25**, 142–154 (2015).
31. de la Torre-Ubieta, L. *et al.* The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172**, 289-304.e18 (2018).
32. Guilmatre, A., Huguet, G., Delorme, R. & Bourgeron, T. The emerging role of SHANK genes in neuropsychiatric disorders. *Dev Neurobiol* **74**, 113–122 (2014).
33. Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N. & Geschwind, D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat Neurosci* **19**, 1397–1407 (2016).
34. Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362**, (2018).