4-6-2023 11:00 AM

A Computational Framework For Identifying Relevant Cell Types And Specific Regulatory Mechanisms In Schizophrenia Using Data Integration Methods

Kayvan Shabani,

Supervisor: Parisa Shooshtari, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

Combining multiple data types can help researchers gain deeper insight into the subject of the study compared to analyzing only one dataset in many cases. Biological researchers can also benefit from these methods of integration. For instance, GWAS data that gives information about variations in the DNA cannot provide us with much information about the specific biological components that are significant in the trait of interest. However, when combined with sequencing data such as chromatin accessibility data or gene expression data, they can help us find the significant biological elements in the trait of interest. In this study, I perform multiple statistical and machine learning-based integration methods on GWAS and sequencing data and find the relevant tissues and cell types in schizophrenia and specific regulatory elements affected by this complex mental disease.

Keywords: schizophrenia, scRNA-seq, scATAC-seq, GWAS, data integration, supervised learning, machine learning, data science

Summary for Lay Audience

As the technology progresses, new datasets are generated at a faster speed constantly in all of the fields. Although each of these datasets gives researchers further information about the subject of their studies, when combined together they may give them insights that would have been missed without the integration of multiple datasets. Data integration methods try to develop ways to leverage fusing datasets together to get a better insight into their subject of interest.

Biological studies can benefit from data integration too. In this thesis, I apply three data integration methods to multiple biological datasets in order to obtain a deeper understanding of a complex mental disease called Schizophrenia.

Some biological data like data from variations in the DNA sequences cannot give much information about the functional elements that play a role in the disease of interest like schizophrenia. However, when combined with other biological data types like datasets that get generated by mapping small parts of DNA to the whole DNA sequence (sequencing data), they enable us to find the specific biological components important in schizophrenia.

I apply the integration methods to mouse and human datasets to find the cell types that are important in schizophrenia, as well as biological components affected by this disease. Finally, I propose suggestions to help researchers develop further integration frameworks in the future.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Field Domain and Background

Schizophrenia is a complex disorder that affects around 24 million people in the world. This disease is recognizable by its mental symptoms, such as delusions, hallucinations, cognitive dysfunction, and disorganized speech or behavior. The origin and biological mechanisms of schizophrenia are not still fully uncovered, and researchers believe that multiple factors such as genetic, epigenetic, and environmental factors play an important role in it[1]. In recent years, there has been an interest in uncovering the underlying mechanisms of schizophrenia by using the integration of multiple data types, such as genetic, epigenetic, and gene expression data. By integrating these data types, a more comprehensive understanding of the disease can be obtained.

1.2 Current Knowledge and Remaining Gap

Numerous Genome-Wide Association Studies (GWAS) have been performed on schizophrenia, identifying multiple genetic variants significantly associated with the disorder[2]. However, GWAS data alone cannot provide information about the specific biological components or underlying mechanisms of schizophrenia. Therefore, integrating GWAS data with other biological data types, such as chromatin accessibility and gene expression data, is essential to gain a comprehensive

understanding of the disease. Using these integration methods, multiple tissues, cell types and biological elements like genes and transcription factors have been associated with schizophrenia.

Despite recent advancements in data integration pipelines for identifying biological mechanisms in complex diseases like schizophrenia, there remains a need to analyze new datasets as they become available, and to develop novel pipelines for uncovering new candidate elements that may play a crucial role in the disease. The current state of knowledge emphasizes the importance of integrating multiple data types to uncover the underlying biological mechanisms of schizophrenia. However, there is still a significant gap in understanding how to develop and apply effective computational methods for integrating these diverse data types to achieve this goal.

1.3 My Approach

In this study, I aim to address this gap by developing and applying data science approaches to integrate multiple data types effectively and understand the biological mechanisms of schizophrenia. My approach includes the following key steps:

1- Applying previously developed data integration methods to integrate GWAS data with chromatin accessibility and gene expression data. I use both mice and human data to prioritize cell types and predict biological elements relevant to schizophrenia.
2- Developing a new integration method to find schizophrenia-relevant genes, using a deep learning approach.

3- Providing a data integration approach that can be applied to a wide range of subjects that leverage computational methods, including big data applications, sensor imaging, biology, and others.

1.4 Results of This Study

Through the application of these approaches, I have achieved the following results:
The prioritization of cell types that are likely to be relevant to schizophrenia, providing valuable insights into the disease's cellular context.
The identification of specific biological elements, such as regulatory sites, genes, and transcription factors, that are relevant to schizophrenia, furthering the understanding of the disorder's molecular mechanisms.
The development of a versatile data integration approach that can be applied to other scientific fields that require computational methods, broadening the potential impact of this work beyond schizophrenia research.

1.5 Contents of This Thesis

The remainder of this thesis is organized as follows:
In Chapter 2, I provide the necessary background to understand the main content of this study. In Chapter 3, I explain the data integration approaches and pipelines used in this thesis, the challenges typically faced in data integration methods, and how I propose to address them. Chapter 4 will explain the methods used in this study and is focused on a data integration approach used to prioritize cell types likely to be relevant to schizophrenia and two other data integration pipelines that can be used to identify specific elements relevant to schizophrenia, such as regulatory sites, genes,

and transcription factors. The first pipeline is a general pipeline that was originally developed by one of my colleagues in Shooshtari Lab (Mr. Nader Hosseiny Naghavi) and can be applied to several common complex traits. For this thesis, I have used and slightly modified it to integrate the datasets relevant to schizophrenia. The other data integration pipeline is developed by myself and finds the differentially expressed genes based on the schizophrenia-affected peaks with the help of a deep-learning model called BABEL[3]. Chapter 5 is focused on explaining the results of the pipelines described in chapter 4 and a discussion regarding those results. Finally, in Chapter 6, I provide a summary and conclusion of this study and discuss future works.

Chapter 2

Literature Review

New advancements in biotechnologies have enabled us to have access to various types of biological data. As technologies improve, we become more and more equipped with important biological information with higher resolutions and better accuracies. Although our technologies and data qualities have advanced rapidly, a single data type is usually not sufficient to capture all the information that exist complex biological mechanisms, such as those related to cancer[4] or early-stage developments in mammals[5]. Therefore, combining multiple data types - that each provides information on a specific aspect of a mechanism - is now considered a crucial step toward understanding of complex biological mechanisms. Most relevant to my study, integrating multiple data types can be used to provide a comprehensive insight into the mechanisms related to gene regulations in schizophrenia, and can help us identify the genes and the regulatory elements relevant to the disease.

As different large-scale data types get generated, we need to develop effective computational methods to analyze and integrate them. Handling such large-scale datasets and the process of their integration require high-performance computational resources. The large datasets that are being generated require storage, cleaning and computational power to get processed. If these tasks would not be performed effectively, they can put a huge burden on computational resources and may waste huge amounts of time and money. Therefore, the issue of developing efficient methods for handling new modalities should be addressed thoroughly [6,7,8,9].

In this chapter, I provide an explanation of the computational, statistical, and biological concepts used in my study that are crucial for understanding the other chapters of my thesis. Here, I first provide a brief background on schizophrenia, followed by describing the data types that I use in my study, including data from genome-wide association studies (GWAS), bulk and single-cell chromatin accessibility data, and single-cell transcriptomic data. Then I provide a review of existing methods for the integration of GWAS and sequencing data, followed by methods of integration of two different modalities of sequencing data. I finish this chapter by describing a machine learning model that I use to infer one sequencing data type from another data type.

2.1 Schizophrenia

Schizophrenia is a mental illness recognizable by its mental related symptoms. The most common symptoms include delusions and hallucinations. These symptoms are the common symptoms that cause the patient to visit a doctor about their condition and are considered positive symptoms. However, there are also negative symptoms associated with Schizophrenia such as lack of motivation, social withdrawal, problems in memory, and speed of processing. Currently, American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders Fifth Edition (DSM-5) criteria for schizophrenia[1] consists of having two or more of the following symptoms for at least one month and one of them has to be one of the first three symptoms: delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behaviour, and negative symptoms, such as diminished emotional expression.

The disease usually occurs in early adulthood and decreases life expectancy. It also increases the risk of suicide in one's lifetime. Alongside its mental difficulties,

schizophrenia costs substantially for the patients and the health system. Although the probability of occurrence of schizophrenia in a lifetime is one percent, it has been estimated that the disease costs about 150 billion dollars a year for US citizens. In Canada, according to the data gathered in 2016-2017, one out of hundred of Canadians were diagnosed with schizophrenia[10]. Also, Goeree et al. concluded that schizophrenia costed around 2.02 billion dollars in 2004 for direct healthcare and not-healthcare costs in Canada[11].

It is already known that both genetic and environmental factors contribute to the risk of developing schizophrenia. A heritability of around 80% is estimated by using twin and other studies[1].

Recent advancements in genome-wide sequencing technologies and genetic association studies have made it possible to generate genetic data that can be explored to better understand the genetic causes of schizophrenia. In particular, genome-wide association studies (GWAS) have been successful in identifying hundreds of loci associated with schizophrenia.[1]

Many tissues and broad cell types have been identified as relevant to schizophrenia. For instance, the prefrontal cortex, which is involved in functions such as decision-making and attention, has been linked to schizophrenia and the cognitive issues associated with the disorder[12]. Also, hippocampus, which is important for learning and memory functions and has been associated with schizophrenia and the memory problems experienced by schizophrenia patients[13]. The thalamus is involved in the integration and processing of sensory and motor information in the brain. Abnormalities in its structure have been observed in schizophrenia, which could contribute to disruptions in information processing in schizophrenia patients[14]. Other tissues, such as the striatum[15], cerebellum[16], etc., have also been linked to

schizophrenia. It is important to search for new candidates or confirm previous results to have a more comprehensive view of this complex disorder.

In addition to tissues and broad cell types, multiple central nervous system cell types have been associated with schizophrenia. Excitatory neurons are important in neuronal connection, and abnormalities in their functions have been observed in schizophrenia patients[17]. Inhibitory neurons have also been linked to schizophrenia. These neurons play an important role in modulating neural activity, and their dysregulation has been reported in schizophrenia[18]. Oligodendrocytes are another example of affected cell types in schizophrenia. They are important in the transmission of signals in the brain, and altered expression of oligodendrocyte-related genes has been reported in schizophrenia[19]. Other cell types, such as microglia[20] and astrocytes[21], have also been reported to be relevant to schizophrenia. As new datasets are generated regularly, it is important to analyze them to either confirm previous results of cell types relevant to schizophrenia or propose new candidates.

Several genetic factors and biological mechanisms play an important role in schizophrenia. Genes contribute to biological mechanisms by affecting the level of protein generation. Multiple genes have previously been found to be affected in schizophrenia. For instance, DISC1 is considered relevant to schizophrenia[22] because it is involved in neuronal migration and neurodevelopment, and changes in the expression of this gene can affect the development of the brain, further contributing to schizophrenia.

In another study, the C4 gene has been identified as contributing to schizophrenia[23]. Overexpression of this gene has been linked to synaptic pruning in the developmental stages of the brain, which can cause schizophrenia's neural symptoms. Dysregulation

of the dopamine system has also been associated with schizophrenia, and genetic variations in DRD2 have been linked to an increased risk of developing the disorder[24], as DRD2 encodes a subtype of dopamine receptor.

NRG1[25] is another gene which is considered to be relevant to schizophrenia because of its role in neurodevelopment. GABRB2[26] is also playing an important role in schizophrenia by being involved in the primary inhibitory system of the brain, and variations in this gene affects the risk of schizophrenia. Other genes, such as COMT[27], DTNBP1[28], RGS4[29], GRM3[30], and several others, have also been linked to schizophrenia. Although multiple genes have been associated with schizophrenia, the exact gene regulatory mechanisms underlying this complex disorder are still unknown. Hence, there is a need to identify cell-type specific gene regulatory mechanisms in the disease.

Genetic risk variants can change the binding probability of transcription factors, which can affect the level of expression of the genes that those transcription factors regulate. Many transcription factors have been identified as affected in schizophrenia in this manner. For instance, TCF4 has been linked to schizophrenia based on this criterion[31]. TCF4 is involved in neurodevelopment and neuronal differentiation, and genetic variations that change the binding affinity of this transcription factor can increase the risk of schizophrenia. NPAS3 is another transcription factor that regulates genes that are important in schizophrenia[32]. Risk genetic variants can also change the binding probability of this transcription factor and increase the chance of schizophrenia. MEF2 is a transcription factor that is a key activator in synapse development. It has been discovered that genetic variations can affect the binding affinity of this transcription factor and drive the risk of schizophrenia[33]. Other transcription factors, such as FOXP2[34], NEUROG1[35], CLOCK[36], and many others,

have also been linked to schizophrenia. However, they cannot capture the full picture of schizophrenia, and the experiments needed for them are costly.

2.2 Description of Data Types

2.2.1 Genome-Wide Association Studies (GWAS) Data

Single nucleotide polymorphism (SNP) is a genetic variation and happens by substitution of one nucleotide in the genome. This substitution can cause changes in biological mechanisms and play an important role in developing diseases in the organism[37]. In GWAS, millions of genetic variants are examined to find out which ones are significantly associated with a trait of interest [38]. The bigger the sample size of a GWAS study, the higher statistical power to find the variants (i.e. SNPs) associated with a disease. Although GWAS alone cannot fully uncover the biological mechanisms of a disease, GWAS data can be integrated with other data types to reveal the biological mechanism of the disease.

A Manhattan plot showing the significance of associations of the SNPs tested in the Ripke et al. study is shown in figure 2.1. In the X axis chromosome numbers are written and in the Y axis minus log10 of p-values calculated for the SNPs in the study are shown. When this value is higher than a threshold for a SNP, it is considered as a significantly associated SNP with the trait. The trait of interest in Ripke et al.[39] study is schizophrenia and is one of the three GWAS data that I used in this study. They found 128 SNPs that can pass the threshold (the straight line in figure 2.1) in their study.

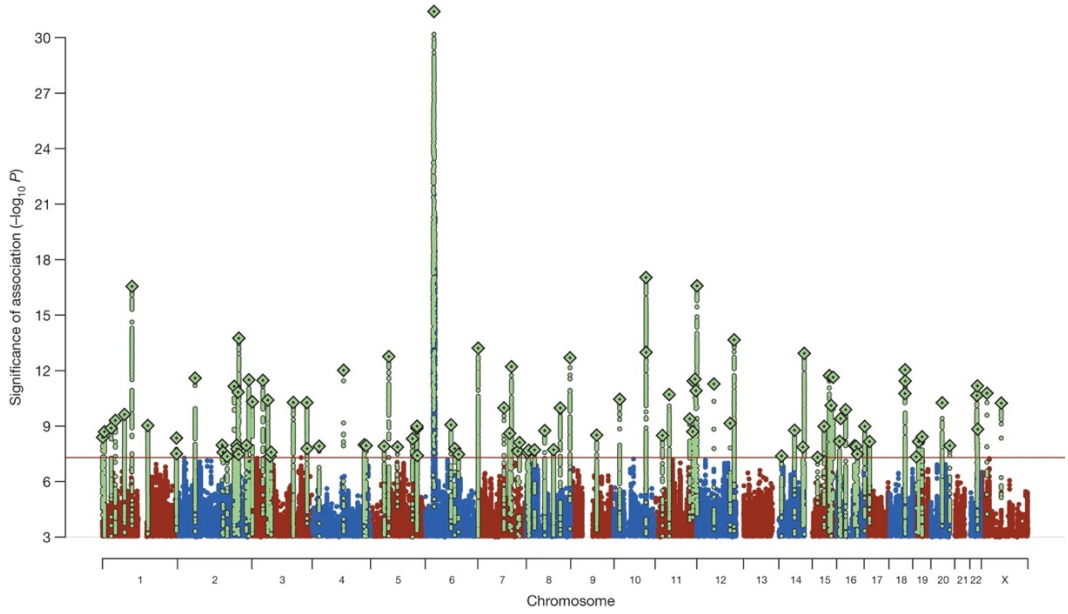


Figure 2.1: The Manhattan plot used in the Ripke et al.[39] GWAS study to show the SNPs in a chromosome number vs -log10(p-value of association significance). The SNPs that pass the threshold line are considered as significantly associated with schizophrenia

GWAS data can be obtained by performing case-control studies where cases are the samples that have the trait of interest in them, and controls are the samples that do not have the trait in them. Then, researchers can compare the genomes between these two groups to find the SNPs that probably contribute to the trait of interest. The result of such a GWAS study are SNPs that are significantly associated with the desired trait. The schizophrenia Working Group of the Psychiatric Genomics Consortium generated one of the largest GWAS studies in 2014[39]. The authors of this study used 113,075 controls and 36,989 schizophrenia cases and performed a case-control study which resulted in identifying 83 novel loci that were associated with schizophrenia.

In another study, Li et al. tested 36180 samples and found 30 novel schizophrenia loci in 2017[40].

Another widely-used schizophrenia GWAS dataset has been generated in 2018 by Pardinas et al where they found 50 novel schizophrenia-associated loci[41].

In a recent large-scale genome-wide association study (GWAS) conducted in 2022, researchers investigated the genetic basis of schizophrenia[42]. The study comprised a two-stage analysis that included up to 76,755 individuals with schizophrenia and 243,649 control individuals. The authors identified 287 distinct genomic loci associated with schizophrenia, with associations concentrated in genes expressed in excitatory and inhibitory neurons of the central nervous system. At the time I began this study, the specific GWAS dataset was not yet available. Consequently, I utilized other widely used schizophrenia GWAS datasets that were accessible at that time. Future work on this study could involve applying the pipelines proposed in this thesis to the new GWAS dataset and comparing the results with those presented in this thesis.

2.2.2 Open Chromatin Data

According to Fang et al.[43], the active cis-regulatory elements on the genome can be identified with hypersensitivity to nucleases or transposases. This has been the foundation behind the development of sequencing technologies such as Assay for Transposase Accessible Chromatin (ATAC-seq)[44] and DNase-I Hypersensitive Sites Sequencing[45]. Chromatin is a level of packaging in the eukaryotes including human's genome. This packaging levels exists so the genome can get folded and fit inside the nucleus. Chromatin is consisted of another packaging level called nucleosome. Some parts of chromatin are left open during the packaging. Macromolecules can attach to them and interact with the DNA[46,47]. These open chromatin sites are cell type specific and open chromatin sites of one cell type is different from the other ones. There are multiple sequencing technologies that are able to identify open chromatin sites and generate datasets based on these sites. DNase-seq[45] and ATAC-seq[44] are two of the standard technologies that are able to generate such data.

2.2.3 DNase-I Hypersensitive Sequencing Data

DNase-I Hypersensitive sequencing is a technology that can be used to identify regions of open chromatin for various cell types[45]. DNase I hypersensitive sites (DHSs) are the regions of the genome that show hypersensitivity by the DNase-I enzyme. They were first introduced by Wu et al.[48] Basically, the more open the chromatin region is, the more it is sensitive to this enzyme. This phenomenon has led scientists to further conclusions, such as the regions that have active genes are usually more open and are more sensitive to this enzyme. More specifically, these studies suggest that the regions with active genes are 100 times more sensitive to the DNase-I enzyme compared to the genomic regions that do not contain active genes[48,49]. Also, it is well-known that the transcription factors (TFs) that bind to the genome cover the open genomic region and will reduce the sensitivity to the DNase-I enzyme. TFs are the proteins that bind to DNA at the open chromatin site, and interplay with DNA to regulate the target genes. When a TF binds to DNA, it covers a portion of the open chromatin region. Therefore, that part of the DNA becomes less sensitive to DNase-I enzyme, as it is protected against digestion by the enzyme as a result of being covered by the TF.[50,51] It should be noted that this only happens when a TF binds to the genome at open chromatin sites. When the TF is not bound to DNA, the DHS region is still open and sensitive to the enzyme, as the TF is not there to cover it[50].

DNase-I sequencing is based on next-generation sequencing technology. Here I briefly describe DNase-I sequencing pipeline. In the first step of this method, nuclei get separated from the cells to ensure that the enzyme reaches the DNA. Multiple factors can play a role in effectiveness of DNase-I sequencing, including the efficacy of the DNase-I enzyme, and how cells respond to the enzyme. Hence, in order to reach the optimal activity level of the enzyme, researchers should adjust the amount

of enzyme and the number of cells in each experiment. After digestion of the DNA, the remaining parts will go through the process of purification. Shorter fragments, usually between 50 to 100 base pairs are more likely to be enriched of transcription factor binding sites because they cannot cover around the nucleosome, as the number of nucleotide pairs needed to cover the whole nucleosome is 147 base pairs[52]. After these steps, the reads are sequenced, and then short reads are aligned to the genome. The regions of the genome that are enriched for the number of reads mapped to them are considered as peaks of DHS. Peak calling algorithms are used to identify DHS peaks[53].

Using DNase-I sequencing, researchers have been able to generate a large amount of open chromatin data and analyze them to build large-scale DHS datasets, including those available through ENCODE Project[54], NHGRI Genomics of Gene Regulation (GGR)[45], Blueprint Epigenome[55], and NIH Roadmap Epigenomics Mapping Consortium (REMC)[56]. Shooshtari Lab has generated a comprehensive databases of open chromatin sites called OCHROdb by integrating > 800 DHS samples collected through above large-scale projects[57,58]. OCHROdb contains 1455046 DHS peaks across 194 cell types, tissues and cell lines. In my study, I have integrated open chromatin data from OCHROdb database and genetics association data from schizophrenia GWAS in order to identify schizophrenia-relevant cell types.

2.2.4 ATAC Sequencing Data

Assay of Transposase Accessible Chromatin sequencing or ATAC-seq is one of the widely-used technologies for identifying open chromatin sites. In this technology Tn5 transposase is used to identify chromatin regions that are open. The output of such technology can be a matrix regarding bulk data which shows the average accessibility within cell types in each open chromatin peak(regions that have a high number of

DNA reads mapped to them) or a cell by peak count matrix showing accessibility of each cell in each identified open chromatin peak[59].

2.2.5 Single Cell Open Chromatin Data

Methods of measuring chromatin accessibility (e.g. Dnase-I-seq and ATAC-seq) were originally developed for generating bulk open chromatin data, where tens of thousands of cells were sequenced together and therefore, the output of sequencing provide an average chromatin accessibility across those cells. However, bulk sequencing does not provide a resolution at the single-cell level. This motivated the generation of newer technologies, which measure the chromatin accessibility at the single-cell level. Single-cell ATAC-seq (scATAC-seq) is the most commonly used technique for measuring single-cell chromatin accessibility.

In summary, ATAC-seq benefits from using a genetically engineered hyperactive Tn5 transposase that is able to cut the open chromatin regions of DNA. During this process replicated of these regions are also created and in the end multiple reads of these sites will be ready to get sequenced. Then, sequencing technologies use these reads and map them to the genome which helps them to identify the locations of these reads in the genome. Further, these locations are considered as open chromatin regions[59].

2.3 Methods of Integration of Sequencing Data and GWAS Data

GWAS data gives us information about the associated genomic variants to the trait. However, they cannot help us identify significant cell types and specific regulatory elements of the trait of interest alone. By integrating GWAS with sequencing data researchers are able to get such insights into their traits of interest.

Multiple studies have used the integration of GWAS and sequencing data to identify cell types relevant to complex traits including schizophrenia[60–62]. These include linkage disequilibrium score regression (LDSC)[63], genome-based restricted maximum likelihood (GREML)[64], LDAK[65], and regression-based polygenic model (RolyPoly)[66]. According to Zhu et al.[67], the methods of GWAS and sequencing data integration can be divided into four major categories.

The first category of methods use cell-type specific annotations that come from epigenetic or expression data. They estimate the contribution of the annotations to SNP heritability of the GWAS. A group of these methods use epigenetic annotations including LDSC[68] or Scalable Multiple Annotation integration for trait-Relevant Tissue identification and usage (SMART)[69]. In both methods, genomic regions identified by as sequencing experiment (such as ATAC-seq, DNase-I hypersensitive sites or chromatin marks) are combined with GWAS data to prioritize related cell types.

Other methods of this category use gene expression (e.g. RNA-seq) data and integrate them with GWAS. Examples of such methods include LDSC-SEG[70] and RolyPoly[66]. LDSC-SEG uses bulk gene expression data, while RolyPoly can be run on single-cell data.

The second category of methods combine GWAS data with expression quantitative trait loci (eQTL) data to prioritize cell types relevant to a trait of interest. eQTL is the study of relating variants to gene expression. Methods in this category include normalized tissue causality score (NTCS)[71] and eQTLEnrich[72].

The third category of methods integrate genetically regulated expression levels data (Grex) with GWAS data to find cell types relevant to a trait. Impact of genetically

regulated expression (IGREX)[73] and RhoGE[74] are two well-known examples of such methods.

The fourth group of methods are the ones that use networks between genes inferred from their expression and integrate them with GWAS data to find the cell types relevant to a trait. Composite likelihood-based covariance regression network model (CoCoNet)[75] falls into this category.

In my study, I used LDSC (from category 1) for the integration of GWAS and ATAC-seq data, and LDSC-SEG (from category 2) for the integration of GWAS and RNA-seq data. Hence, I provide a detailed description of the methods in categories 1 and 2.

2.3.1 LDSC

Linkage disequilibrium score regression (LDSC) is a method of integration of genomic annotations and GWAS data. It will calculate the enrichment of genomic variants or SNPs presented in the trait's GWAS data on the regulatory regions of multiple cell types and prioritizes the significant cell types based on the enrichment level. Genomic regions can be epigenetic sites, open chromatin sites, histone marks, or other regions. In this study, I have been interested in open chromatin sites. Understanding the statistical model behind LDSC helps us to find out what do the LDSC outputs mean, so we can interpret the results better. Considering the peaks of open chromatin regions, LDSC divides the SNPs into C categories, where C is the number of cell types. It then overlaps open chromatin sites of each cell type with the SNPs and identifies SNPs that are present in each cell type based on this criterion. In the next step, LDSC uses marginal χ^2 statistic for association of SNPs with the trait, r^2 statistic between the SNPs of the GWAS data, and GWAS study sample size that all can be obtained from GWAS data to estimate GWAS trait heritability by a regression model for each cell type. Furthermore, LDSC will calculate a z-score and a p-value

for trait heritability estimation models in each cell type. In this manner, we are able to find out which cell types have lower p-value and whether or not they can pass a threshold that one can choose for significance in his study.

2.3.2 RolyPoly

RolyPoly integrated GWAS and RNA-seq data to prioritize trait-relevant cell types. The main difference between RolyPoly and LDSC-SEG which is the version of LDSC that uses gene expression as an input instead of chromatin accessibility data is that LDSC-SEG focuses on SNPs but RolyPoly takes advantage of focusing on genes. Authors of RolyPoly developed their method based on this hypothesis that in the causal cell types, SNPs that have higher effect sizes in GWAS data should be close to the genes that are highly expressed in those cell types. Hence, in this way they are able to detect the significant cell types by analyzing the enrichment of SNPs in the highly expressed genes of each cell type and comparing them with each other. Since RolyPoly does not have a version that can integrate GWAS data with epigenetic data, I did not use it for this study.

2.3.3 SMART

Scalable Multiple Annotation integration for trait-Relevant Tissue identification (SMART) is another method of data integration that benefits from using multiple annotations for SNPs. In the first step, phenotype (i.e. disease) is related to the genotype using a multiple linear regression model:

$$y = X\beta + \epsilon, \epsilon_i \sim N(0, \sigma_e^2)$$

where y denotes the vector of phenotype for samples; X is the matrix of genotypes consisting of n samples versus m SNPs; Beta is a vector indicating the effect sizes of the SNPs and ϵ is a symbol of residual errors. Each ϵ is assumed to have a normal

distribution with the mean (0) and variance (σ_e^2) shown in the formula. In the next step, the effect sizes are related to the annotations. To do so, a vector of annotation values for each of the SNPs is considered. This is shown by $A_j = (1, C_{j1}, C_{j2}, \ldots, C_{jc})^T$. Each value of this vector can be discrete or continuous depending on the annotation type.

Here, the effect sizes are assumed to have a normal distribution with mean of zero and a variance that is a function of annotations:

$$\beta \sim N(0, \sigma_j^2/m), \ \ \sigma_j^2 = A_j \alpha^*$$

Here we have $\alpha^* = \binom{\alpha_0}{\alpha}$ and α is a vector with the size of c (number of annotations) indicating annotation coefficients. SMART uses generalized estimating equation (GEE) to estimate α and its variance. By calculating these terms, one can calculate the multivariate Wald statistic $\hat{\alpha} V(\hat{\alpha})^{-1} \hat{\alpha}$ which is used to measure the relevance between a cell type and the trait of interest.

2.3.4 LDSC: A Method of Choice for My Study

As LDSC is one of the most commonly-used standard approaches for prioritizing cell types by integrating GWAS and sequencing data, and since it can integrate GWAS data with both ATAC-seq and RNA-seq data, I choose to use LDSC in the GWAS integration with the sequencing data. Previously, researchers have used LDSC in various studies. This includes, the integration of Bulk DHS data and GWAS in Meuleman et. al[76] to partition heritability estimates according to the sets of genome-wide annotations consisting the DHS data. In addition, LDSC was previously used for the integration of single-cell ATAC-seq and GWAS to link brain cell types to a group of brain-related traits such as Schizophrenia [77] . As a part of this study, I confirm their

findings by using the same GWAS data that they used along with another GWAS data and confirmed the consistency of the results across multiple GWAS datasets. In addition to epigenomic data (e.g. scATAC-seq), the single-cell RNA-seq data have also been integrated with GWAS data using LDSC approach [70]. The aforementioned studies focused on human data; however, LDSC is general and can be applied to other organisms. Hook et. al[62] applied LDSC to integrate 64 GWAS data with multiple mouse datasets.

In my study, I implemented a pipeline that uses LDSC to integrate three Schizophrenia GWAS datasets with (a) bulk chromatin accessibility from human data, (b) scATAC-seq from both mice and human data, and (c) scRNA-seq from mice data. This resulted in the identification of multiple cell types in humans and mice that are likely to be relevant to Schizophrenia. My results not only confirmed the findings from previous studies, but also expanded on them and identified schizophrenia-relevant cell types at different time points of brain development. My study has therefore, contributed to the better understanding of molecular and cellular mechanisms of schizophrenia.

2.4 Converting Genome Annotations

For the integration of multiple data types, such as integration of GWAS data and sequencing data or the integration of sequencing data from two different modalities, I should make sure that the genome build of the two datasets are the same; otherwise, the analysis and results will be incorrect. However, it may happen that the GWAS data is generated based on a certain human genome build, while the sequencing data (e.g. scATAC-seq) is aligned to a different human genome build. Even sometimes, we have the sequencing data from a different organisms, such as mouse. To address this

problem, we need to convert the genome annotation from one reference genome build to another one.

The two most recent human genome builds are GRCh37 and GRCh38. Although GRCh38 is more recent, some of the GWAS or sequencing datasets may have been published based on GRCh37 coordinates. The same thing applies to mouse data where mm9 and mm10 are the most recent annotations. The position of genes and SNPs are changed from one annotation to another one and clearly, this has to be addressed when integrating data coming from multiple modalities.

One standard method that has been developed to convert annotations to each other is called LiftOver[78]. LiftOver has both an online tool and a command line-based application and one can use them based on their needs. LiftOver has the ability to convert annotations from one organism to another one, and also across different genome builds of the same organism. LiftOver gets a set of genomic regions in standard BED format[79] as the input, and outputs two files. The first file contains a set of new genomic ranges that have been successfully converted from the previous coordinates based on a threshold. The threshold shows what percentage of the base pairs should be successfully converted in order to consider the conversion successful. The second output file contains a set of genomic ranges that failed to convert. As the conversion of annotations between two different organisms is generally more complicated, it is recommended to use a less stringent threshold for conversion between different organisms compared to the one between two different genome builds of the same organism. For instance, when converting from human GRCh37 to human GRCh38, it is recommended to use a threshold of 0.95, while a threshold of 0.7 is recommended for converting from mouse mm10 to human GRCh38.

2.5 Methods of Integration of Two Different Modalities of Sequencing Data

According to Argelaguet et al.[80] in order to integrate different two modalities of single-cell sequencing data, we first need to define anchors or links between the data types. To do so, we should explore the similarities between the data types and choose them wisely in order to proceed with the data integration. Depending on how we choose the anchors we can have three different types of integration.

Type 1 - Vertical or Cell-Based: This method of integration can be used when the technology that we use generates multiple modalities from the same sets of cells. For example, some of the recent technologies are profiling ATAC-seq and RNA-seq from the same sets of cells where we can have both epigenetic and transcriptomic information simultaneously. Example of such technologies include droplet-based single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq)[81] and simultaneous high-throughput ATAC and RNA expression with sequencing (SHARE-seq)[82], which measures transcriptome and chromatin accessibility data simultaneously, and CITE-Seq[83], which measures epitome and transcriptome data simultaneously. In these examples, the modalities share the same cells, so it is plausible to use the cells as anchors.

Type 2 - Horizontal or Genomic Features-Based: In some cases, we have independent datasets that have been generated from the same modality. For example, we can have several samples from single-cell RNA-seq and we want to integrate them in order to have a wider perspective and more reliable analysis by increasing the sample size. In these cases, genomic features (such as genes in the case of RNA-seq), can be used as anchors.

Type 3 - Diagonal Integration: In some cases, both genomic features and sets of cells are different between the datasets. In this case we cannot simply choose the anchors from obvious choices such as cells, shared genes or shared open chromatin regions. In contrast, we have to use more complex methods to find a suitable anchor. For instance, we may have several datasets from the same tissue, and have both scRNA-seq and scATAC-seq modalities available for them. However, the cells between these two types are different in the scRNA and scATAC datasets. In this scenario, other methods such as Seurat anchoring [84] can be used to find the links between the two datasets.

In addition to the above categorization of data integration methods, data integration methods can be categorized in a different way which is based on the stage where data integration happens[85].

The early integration technique in data analysis involves the transformation of multiple datasets into a single table or representation, which serves as input for computational algorithms. This method allows for consideration of any type of dependence between the features. To achieve this, automatic feature learning techniques like dimensionality reduction and representation learning are employed to reduce the high-dimensional datasets into a low-dimensional vector space that is subsequently combined through simple aggregation techniques. Despite its ease of application, this technique may be limited by the heterogeneity of features across datasets, and proper normalization must be implemented to prevent bias.

The second category of methods are those methods that use intermediate data integration. When employing intermediate integration in data analysis, a model is utilized to learn a joint representation of multiple datasets, such as deep neural networks. This approach relies on algorithms that explicitly address the multiplicity of

datasets and fuse them through inference of a joint model. Unlike early integration, an intermediate data integration method preserves the structure of data and only merges them during the analysis stage. It does not combine input data or develop a separate model for each dataset. While this approach can lead to superior performance, it often requires development of a new algorithm and cannot be utilized with simple software tools.

Late integration involves building a model independently for each dataset or data type. These models are then combined by training a second-level model that either utilizes the predictions of the first-level models as features or employs a predictor that takes a majority vote or combines prediction weights of the first-level models. This technique allows for the incorporation of diverse models and algorithms for each dataset but may be limited by the need for comparable prediction outputs and the potential loss of information through model combination.

2.6 Converting Sequencing Data Modalities Using Machine Learning

Biological experiments, and in particular, generation of multiple data types (e.g. epigenomic and transcriptomic), can be quite time-consuming and costly. Hence, researchers have attempted to develop computational methods that can infer one data type from another one, whenever possible. Computational methods and predictive models can help scientists to reach the results much faster and avoid the expenses that a biological experiment needs. Most relevant to my study, I would hugely benefit from both epigenomic (i.e. scATAC-seq) and transcriptomic (i.e. scRNA-seq) data for individual cells, as each of these data types can give insights into a particular aspect of a biological process related to schizophrenia. In the past years, several single-cell

sequencing datasets have been generated and have become available publicly. However, majority of these datasets have only one data modality for the individual cells. Since the experiments were completed already, it is not feasible and/or cost effective to obtain multiple data modalities for those cells using multi-omics sequencing technologies. Alternatively, we can develop and apply computational methods that can infer transcriptomic data (i.e. scRNA-seq) from epigenomic data (i.e. scATAC-seq), and vice versa.

Machine learning based methods and specially supervised learning approaches provide appropriate means to infer one data modality from another one using single-cells. These methods use multiple modality biological datasets (i.e. multi-omics data) to train a machine learning model for predicting one data modality from another one. Then they use that model to predict the missing modality in their test dataset. BABEL [3] is recently developed deep-learning method that can convert scRNA-seq to scATAC-seq data (and vice versa) using a pre-trained model based on multi-omics datasets. Here, I briefly explain how BABEL works.

The first step in the BABEL pipeline is to pre-process the input data. For transcriptome data, the cells with a very low or very high number of expressed genes are filtered out. Also, the genes that are on sex chromosomes are filtered. Then the dataset is normalized, and the outliers are removed. For the ATAC-seq data, the data is considered as binary data, because a peak is either accessible or not accessible for each cell. The peaks with too many cells or a few number of accessible cells are removed. Also, the peaks on sex chromosome are removed, and finally, the overlapping peaks are merged to obtain the pre-processed data.

In some cases, when one wants to test a new data on the pre-trained model to predict the transcriptome data from epigenetic data, the peaks in the new data may not match

the peaks that were used to train the model. To address this, the test data peaks are mapped to the training data peaks. To achieve this, the test data peaks are renamed to the training data peak names that they overlap with. After this step, the trained model is used to predict transcriptomic data from ATAC-seq.

The authors of BABEL split their data into three parts of training, validation, and testing using a cluster-based approach and used cross-validation to improve the generalization of their model.

One of the most important parts of the BABEL pipeline is its architecture. BABEL benefits from a certain type of neural networks called autoencoder. They modified this popular architecture so it can serve the purpose of their study, which is converting one biological data modality into another one. In this architecture, first layers map the input data into a shared latent space that summarizes the features in the data. The set of these layers is called Encoder. The second set of layers maps the latent shared space into output. This set is called a Decoder. Here, the goal is to minimize a specific loss function which is defined as the difference between the output and the input. More specifically, the autoencoder tries to learn the best features of the data and summarize them in the shared latent space and use this space to predict the new data. Because BABEL converts ATAC-seq and RNA-seq to each other, it uses two encoders (one for each modality). It also uses wo decoders one for each data modality. The RNA decoder outputs the mean and dispersion for each gene, while the ATAC decoder outputs a value between 0 and 1. These values can be binarized by assigning 1 to the values that are greater than the average of both rows and columns in the cell-by-peak matrix. The RNA encoder projects the input to a 64-dimension space and then to a 16-dimension latent space. In all of these layers, the Parametric rectified linear unit (PReLU) activation function is used[86]. The decoder is the mirror of the

encoder and projects the 16-dimension latent space into a 64-dimension space and then projects it to a vector with the size of the RNA input. In the decoder part, similar to the encoder all the layers use PReLU as the activation function.

In the ATAC encoder and decoder, BABEL uses an intrachromosomal approach. For the encoder, first each chromosome's peaks are mapped into a 32-dimensional space and then to a 16-dimensional space. Then the 16-dimensional spaces from all the chromosomes are concatenated. For example, for humans it would be 22 chromosomes and the result of the concatenation will be a 22x16 dimensional space. Afterward, they will project this space into a shared 16-dimensional latent space. The decoder is the encoder's mirror and projects the 16-dimensional space into a (16 times the number of chromosomes) dimensional space. It is then split into 22 parts responsible for each chromosome. For each part, it is projected into a 32-dimensional space and after that to a space with the dimension equal to the number of peaks for each chromosome. In all these layers, they use the PReLU activation function. Figure 2.2 summarizes BABEL's architecture. In chapter 4 of this thesis, I used BABEL to convert open chromatin data into transcriptomic data and used this information to find the differentially expressed genes based on schizophrenia affected chromatin accessibility peaks.

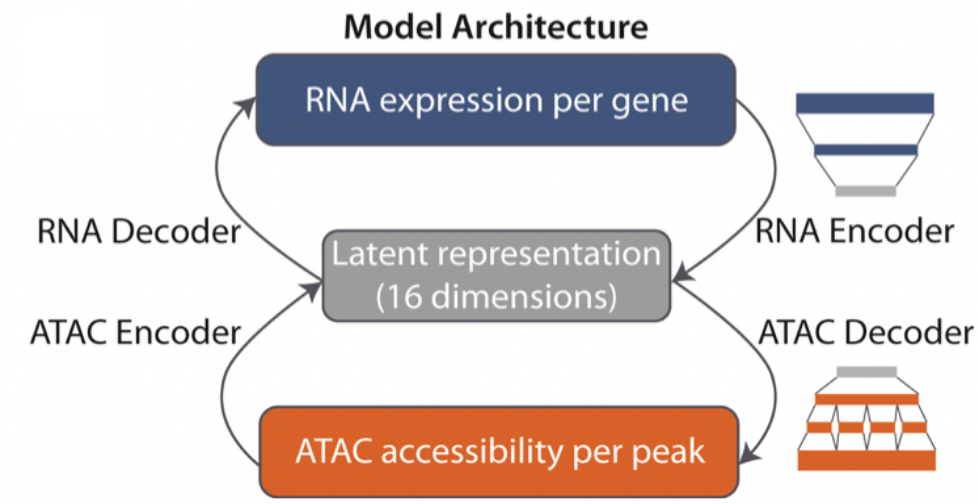


Figure 2.2: BABEL's encoder-decoder architecture. The RNA encoder maps the input RNA data into the shared latent space, and RNA decoder maps the shared latent space into the RNA output. Also, the ATAC encoder maps the input ATAC data into the shared latent space, and ATAC decoder maps the shared latent space into the ATAC output.

Chapter 3

Using Computational Methods to Address Challenges of Data Integration

In this chapter, I explain some of the key challenges and concepts in data integration and demonstrate how I addressed them in my study. The related concepts include:

- Challenges in data collection and consistency between the pipelines' inputs
- Selecting the suitable programming language and packages to handle the large-scale datasets used in biological data integration methods
- Inferring missing modalities of datasets required by the integration methods
- Multiple testing in the integrative data analysis and adjusting for the significance threshold

3.1 Data Integration

In the scope of my work, data integration means combining different data modalities or different instances of one data modality to get more insight into the problem that I am investigating. Biology is not the only area that requires data integration. In fact, data integration has several areas of application, and a standard way of handling this important problem is by developing and applying appropriate computational methods. For instance, in multitemporal data analysis methods, data integration is used to combine images generated by the sensors. In this case, computational methods are used to find the differences between the images in various time points[87].

In many cases, the datasets that are used as input for data integration methods are big (or are numerous and summing together will become large) and the developing integration method should be optimized to handle the required memory and space. Examples of such cases happens in big data[88], and biological applications[89]. Handling the space required by the datasets and running the method efficiently in terms of memory and time needs a good understanding of the computational methods. Hence, computer scientists try to facilitate this process by their knowledge of computer systems and programming languages.

As biotechnologies advance, new types of data get generated. Each of these new data types provide researchers with new important information related to the biology of complex systems. However, if we want to obtain a deeper understanding of biological concepts, we should be able to integrate these different data types. This is because each data type gives us information from a new perspective, and if we do not investigate them, we will miss a huge amount of information that we could have obtained otherwise.

Machine learning and deep learning methods provide valuable means to help researchers with the integration of multiple data types. For instance, to predict whether two proteins interact with each other or not, Zhang et al. developed an ensemble deep neural network (EnsDNN)[90] that leverages data integration by using multiple representation of protein sequences. EnsDNN trains multiple networks based on 3 different representations and aggregates the results of these DNNs at last. The integration method used in this study helps them to build a protein-protein interaction prediction pipeline that beats many of state-of-the-art models.

Another example is related to the prediction of transcription factor binding sites. Transcription factors are proteins that bind to the DNA and affect the expression of

their target genes. Finding the binding sites of transcription factors is important if we want to understand mechanisms of gene regulation. One of the widely-used technologies to identify transcription factor binding sites is chromatin immunoprecipitation sequencing (CHIP-seq)[91]. However, CHIP-seq experiments require specific antibodies for each transcription factor, that are not necessarily available for some of the transcription factors. Also, ChIP-seq can be time-consuming and costly depending on the number of samples to be generated. Computational methods including machine learning approaches can help researchers to predict transcription factors binding sites without requiring them to generate new biological samples. These computational methods benefit from the integration of multiple data types to predict the binding sites of transcription factors. For example, CENTIPEDE[92]is a data integration model that integrates transcription factors position weight matrices (PWM) that are publicly available with the existing open chromatin datasets to predict transcription factor binding sites in specific cell types. Another method is Hmm-based identification of transcription factor footprints (HINT)[93], which uses a combination of DNase-I hypersensitivity and histone modifications to predict transcription factor binding sites.[85] Hence, it is clear that statistical and computation methods play an important role in the integration of biological data. Specifically, machine learning methods have started to get a lot of interest in this area, and new methods of data integration based on machine learning are being developed. My study has hugely benefited from computational and statistical methods of data integration in biology. In Chapter 4, I use a regression-based method called Linkage Disequilibrium Score (LDSC) estimation to integrate GWAS data with the human and mouse single-cell ATAC-seq data and mouse RNA-seq data. In the second part of Chapter 4, I apply a pipeline based on a package which has been recently developed

in Shooshtari Lab. This method integrates GWAS data, single-cell ATAC-seq and transcription factor datasets to identify the cell-type specific regulatory sites, transcription factors and genes underlying common complex diseases. In my study I apply this pipeline to the data from schizophrenia to identify cell-type specific mechanisms of this disease. Finally, in the third part of Chapter 4, I have developed a new approach that builds on the results obtained from the second method of integration (the package that was developed by my colleague) and identifies the differentially expressed genes based on the accessibility patterns of the risk-medicating single-cell ATAC-seq peaks. In this method, I employ a deep learning-based model called BABEL[3] to convert the mouse and human scATAC-seq data to scRNA-seq data. Then I use the predicted scRNA-seq data to identify the genes that are differentially expressed between two groups of cells with accessible and not accessible peaks from scATAC-seq data.

In conclusion, I have used the power of statistical and machine learning based integration methods to develop effective data analysis pipelines and used these methods to study biological mechanisms of schizophrenia.

3.2 Data Collection

The first step in applying a standard data integration pipeline is to generate or find appropriate datasets to use as inputs of the pipeline.

For the data about variants in the genome, I used three schizophrenia GWAS data. GWAS data generated by Ripke et al.[39], Pardinas et al.[41], and Li et al.[40] are three of the widely-used schizophrenia GWAS data that I chose as an input for the data integration methods performed in this study.

For the bulk chromatin accessibility data, I used OCHROdb. OCHROdb is one of the largest DHS dataset available publicly, and it is consisting of multiple processed and

cleansed datasets from consortium-based projects including ENCODE project[54], NIH roadmap epigenomics mapping consortium (REMC)[56], Blueprint epigenome[55], and NHGRI Genomics of Gene Regulation (GGR)[45].

To proceed with my study using single-cell sequencing data, I used scATAC-explorer[94] , which is a database and search tool developed in Shooshtari Lab (https://github.com/shooshtarilab/scATAC.Explorer). scATAC-explorer contains > 30 scATAC-seq datasets from multiple organisms (e.g. mice and humans), and are collected from public resources and analysed in a consistent format.). ScATAC-explorer offers a metadata table and search tool that can be used to select the datasets of interest matching different criteria. Since my study is focused on schizophrenia, I selected brain-related datasets from scATAC-explorer both for mice and humans. The mouse ATAC-seq data generated by Di bella et al. that was obtained from scATAC-explorer, also provided scRNA-seq data for 11 embryonic days and 2 postnatal days. I used this scRNA-seq data as an input of one of my integration methods in this study that is described in chapter 4.

3.3 Data Consistency

In order to generate reliable results, the datasets used as the inputs of a pipelines should be compatible with the requirement of the pipeline, and different data inputs should be consistent. For instance, LDSC pipeline requires multiple inputs including GWAS summary statistics, annotations from ATAC-seq data, and the baseline model that LDSC compares the ATAC-seq annotation data against. These data types should all be compatible with each other in terms of the type of organism and the annotation version. For instance, when we want to run the pipeline for an ATAC-seq data from mouse and a GWAS data from human, then we should convert the genome annotation of one data type, such that both datasets follow the same genome reference. In order

to do this, I have written a script that uses a package called LiftOver[78] to convert the mouse's ATAC-seq annotation from mouse mm10 coordinates to human GRChg8 reference. Also, I choose the other inputs of LDSC in a way that all of them have the same annotation version and organism type.

This method of handling consistency between data types has been applied to all the three annotation methods that I used in this study.

The data consistency is not only important for the genomic data; and in fact, in other applications of data integration, researchers should pay attention to the data consistency across multiple data types. For instance, in the multitemporal data analysis, one should make sure that the locations of sensor are fixed, and the qualities of the images are the same. Otherwise, the images are not comparable because they are showing different locations or are capturing details that another image cannot capture due to its different quality.

3.4 Choosing Suitable Programming Languages and Packages

A lot of biological datasets contain huge amounts of information, and when analyzed by a computational integration method, their large size may cause memory and time problems. For instance, in my study, the human ATAC-seq datasets that I used in all three pipelines are about 3 Gigabytes in a sparse format. A sparse format is a way of saving massive datasets that benefits from saving only the non-zero entries of the matrices. In this way, the size of datasets will drastically reduce, and downstream analysis will be less memory-consuming. One challenge, however, is that when integrating these datasets during developing the integration pipeline, I should try to keep the sparse format in all the calculations to keep the memory efficient, while

some of the computational packages require the users to provide the input data in a dense matrix format. In the case of huge datasets, such as mine, this is not simply possible, because keeping the data in a dense format will need huge computational and memory resources. Hence, I tried to handle this challenge by developing new and/or modifying the existing integration methods that use the benefits of sparse data to make the computations more efficient.

Another challenge that should be addressed in data integration is to use suitable programming languages and libraries for the method of choice. For biological data, mostly R used for the data analysis, because of its unique features that facilitate working with matrices, its visualization capacities, and the availability of many statistical packages in R. However, R can be quite slow for some computations and renders data slower compared to Python. For instance, using a loop is much slower in R compared to several other programming languages, such as Python or C++. To address this, I used both Python and R for my analysis to benefit from R's useful libraries and the time efficiency of Python. For instance, in my third method of integration, I leveraged Python and specifically NumPy library to read BABEL's results and calculate the Mann-Whitney test to find the differentially expressed genes for each peak-cell type tuple, while I used R for creating the list of accessible cells for each peak in each cell type, and also for the visualization purposes.

3.5 Inferring Missing Data for Data Integration

Data integration methods require multiple data types as their inputs; however, there could be cases that we do not have access to all the modalities of one dataset. For instance, for the third method of data integration, I did not have the RNA-seq data for the cells of ATAC-seq datasets. This was because the in one of the studies that I obtain the datasets from, the authors did not conduct a multi-omics experiment, but

instead they had generated the ATAC-seq data only. This situation can happen in other data integration studies. In my study, I addressed this challenge by taking advantage of the computational capability of a deep learning model called BABEL[3], which is able to infer one data modality from another one. BABLE uses an auto-encoder-based model to generate pseudo transcriptomic data (i.e. scRNA-seq) from scATAC-seq for each cell in the datasets. Further details of this model and how I used it for my data integration pipeline are explained in Chapter 4. This is an example of how machine learning-based models can help researchers to infer the missing data in their data integration methods. Similar machine-learning based approaches could be applied to other data integration applications.

3.6 Adjusting for Multiple Testing

While using a data integration method, we may have to perform multiple statistical testing to find significant results. However, as the number of statistical tests increases, the possibility of finding false positives will increase too[95]. This means that there is a higher possibility that some of the significant results are considered significant by a random chance. To address this issue, several statistical methods are frequently used by researchers, two of which are Bonferroni correction[96] and False Discovery Rate (FDR) correction[97]. For Bonferroni correction, a more stringent significance threshold is obtained by dividing the original threshold of 0.05 or 0.01 by the number of statistical tests performed. In comparison, FDR correction methods such as Benjamini-Hochberg try to avoid false positives by decreasing the false discovery rate using the Benjamini-Hochberg critical value. Generally, FDR is a less stringent method compared to Bonferroni[98], yet some may decide to use either of these two approaches based on the characteristics of their study.

In various parts of my study, I performed multiple statistical tests, and therefore, I had to adjust the p values for multiple testing at different stages of my integration pipelines accordingly. As an example, when I applied (LDSC) methods to multiple cell types in Chapter 4, I had to adjust the p values for the number of cell types tested for significance. Also, in my second method of integration, I adjusted the p-values for the number of genes and transcription factors in the study. Another example is the third method of integration, where I adjusted for the p-values with respect to the number of peaks x number of cell types x number of genes for each genomic locus. Correcting the p-values in my study is just an example of how adjusting for multiple testing can be crucial in various data analyses, and particularly, for the integration of multiple data types.

Chapter 4

Methods

4.1 Introduction

Although it is known that brain cells are related to schizophrenia, the brain cell types are very heterogenous and the exact subsets of which that are relevant to schizophrenia are still unknown. By finding the relevant cell types in schizophrenia, one can open the road for other studies to focus on the significant cell types to find other biological components related to this disease. Data integration methods can help us find the specific cell types that are significant in schizophrenia.

Using gene expression (like RNA-seq) or open chromatin (like ATAC-seq) data alongside GWAS data is considered an effective integration method for investigating the underlying mechanisms of the traits including schizophrenia.

RNA-seq is a standard method of measuring gene expression levels and the methods that leverage the integration of RNA-seq and GWAS use the location of differentially expressed genes in each cell type to measure the relevancy of the cell type to the trait. On the other hand, ATAC-seq identifies open chromatin regions, which are the sites in DNA strands that are accessible, and macromolecules can attach to them. Methods that benefit from integrating ATAC-seq and GWAS together use the location of these open chromatin sites and the location of the variants in the GWAS data to gain better insights about the phenotype of interest.

Linkage disequilibrium score regression (LDSC)[68,70] is one of the standard integration methods that benefits from combining biology information coming from variants of a trait and transcriptome or epigenetic data. LDSC can be used to integrate GWAS with bulk or single cell ATAC-seq or RNA-seq. To conduct a comprehensive study, I decided to apply LDSC to all of these data types. First, I used LDSC to determine Schizophrenia-relevant tissues and cell types based on a bulk epigenetic dataset. Afterward, I moved to single cell resolution and applied LDSC on 2 human single cell ATAC-seq datasets and 1 mouse single cell ATAC-seq dataset in three different timepoints of embryonic stages. Also, to fully use the capacities of LDSC I applied it on mouse scRNA-Seq data in 11 different stages of mouse brain development and two after birth time-points. Schizophrenia's causes and effects on the genomic level can be different in various organisms. Hence, for capturing the whole story behind this disease I should investigate it in other organisms too, specially, the organisms like mice that researchers use for running experiments and testing possible drugs on them before human trial. Therefore, I tried to apply LDSC on mouse data too and compared the results of applying LDSC on humans and mice.

The first section is focused on identifying cell types relevant to schizophrenia. In the next section, I am expanding on this work to identify the specific elements that may play a role in driving risk to disease in a cell-type specific manner. There are three main elements or factors that I am investigating in this part: (1) risk-mediating regulatory sites; (2) disease-relevant transcription factors; and (3) disease genes. Identifying these elements helps us better understand molecular and cellular mechanisms underlying gene dysregulation in schizophrenia in a cell-type specific manner.

Regulatory sites of the genome are the DNA sequences that can affect the expression of the genes. These sites include multiple types of genomic sequences such as promoters and enhancers. Enhancers are the genomic sequences that when transcription factors bind to them, can enhance the expression of the genes that they are linked to[99]. On the other hand, promoters are the sequences that define the location where the transcription of a gene will start[100]. Identifying these regulatory sites are crucial for understanding the biological mechanisms underlying complex diseases such as schizophrenia

Genes are sequences located on DNA that contribute to functions or phenotypes in cells. Genes are categorized into two broad categories of coding and non-coding. The coding genes are responsible for the transcription of proteins and non-coding genes are the ones with no direct transcription of proteins assigned to them.

Transcription factors are proteins that bind to DNA and interplay with DNA to regulate the transcription of genes.[101]

My colleague at Shooshtrai Lab, Mr. Nader Hosseini Naghavi, has recently developed a data integration pipeline that combines GWAS summary statistics data with single-cell ATAC-seq data to identify specific regulatory sites, genes and transcription factors that drive risk to common complex diseases. This computational pipeline is general and can be applied to multiple diseases. In my study, I used this pipeline and modified it to be applicable to schizophrenia. In addition, I implemented a new approach based on BABEL deep learning model to first infer RNA-seq of individual cells from scATAC-seq data, and then predict the target genes of risk-mediating regulatory sites through correlating peak accessibilities and expression of genes across several cells. Finally, I compared the results of disease gene predications obtained

from these two independent approaches, and provided a discussion of their similarities and differences.

4.2 Datasets

In my data integration approach, I have integrated GWAS and chromatin accessibility data. Here, I have used three GWAS datasets of schizophrenia, one bulk and three single-cell chromatin accessibility data. In addition, I integrated the GWAS data with single-cell RNA-seq data. A description of the datasets is provided here.

Schizophrenia GWAS Data:

GWAS studies identify and associate traits (e.g. diseases) to variations (e.g. SNPs) in the genome. In these studies, millions of SNPs are tested in thousands of samples to identify the variants that are associated with the trait of interest. The samples of these studies are usually consisted of two groups of positive and negative based on existence of the trait in them. Researchers use the differences between these two groups to identify the GWAS SNPs.

Multiple GWAS studies have been applied to schizophrenia, and they have been successful in identifying hundreds of genomic loci associated to this disease[2]. In this study, I use three largest GWAS data that are available publicly for schizophrenia (Table 4.1). Two of these datasets are collected from individuals with European ancestry[39,41] and one from individuals with Chinese ancestry[40].

Paper	Number of Samples	Ancestry
Ripke et al(2014)	70100	European
Li et al(2017)	36180	Chinese
Pardinas et al(2018)	35802	European

Table 4.1: Summary of GWAS data used in this study

Bulk Chromatin Accessibility Data:

I selected a bulk chromatin accessibility data that broadly scan over several human tissue and cell types and is not focused on the brain cells only. This is mostly because I first wanted to examine whether only brain related cells are significant, or rather I observe significant associations for non-brain related cells too. Shooshtari Lab has previously built a bulk chromatin accessibility data called OCHROdb (https://dhs.ccm.sickkids.ca/)[57,58]. OCHROdb is one of the largest bulk chromatin accessibility databases available publicly that contains a diverse range of cell types and tissues. Originally, 828 sequencing-based open chromatin samples generated by four international consortia (ENCODE, Roadmap, Blueprint, and NIH GGR) were integrated, and the samples were uniformly processed, and quality checked to ensure the open chromatin sites pass the replication test. OCHROdb database comprises of 1,460,986 open chromatin peaks across 194 cell types, tissues and cell lines. 57 out of 194 cell types/cell lines were not from normal, healthy cells, and therefore I excluded them from the analysis.

Single-Cell Chromatin Accessibility Data:

For the scATAC-seq data, I chose three datasets; two of which are from human[77,102] and one from mouse[103]. I selected these datasets to cover a relatively wide range in terms of the brain development, including embryonic and adult human brains, as well as mouse embryonic brain at three different time points.

Single-cell ATAC-seq Data from Human:

In this study, I used two human scATAC-seq datasets. Corces dataset[77] is obtained from adult human brains in 10 samples from the isocortex ($n = 3$), striatum ($n = 3$), hippocampus ($n = 2$) and substantia nigra ($n = 2$) In the original study, the data was pre-processed to obtain a count matrix, containing 444,747 open chromatin sites (i.e. peaks) from 70631 cells. After clustering and annotations, 6 cell types were identified. They include excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, astrocytes and oligodendrocyte progenitor cells (OPCs). To identify the cell types, they used ArchR[104] to generate gene activity matrix and used the following marker genes to assign cell type labels to cells. Microglia were detected based on accessibility near the *IBA1*, *CD14*, *CD11C*, *PTGS1* and *PTGS2* genes. Astrocytes were detected based on accessibility near the *GFAP* and *FGFR3* genes. Excitatory neurons were detected based on accessibility near the *SLC17A6*and *SLC17A7* genes. Inhibitory neurons were detected based on accessibility near the *GAD2* and *SLC32A1* genes. Oligodendrocytes were detected based on accessibility near the *MAG* and *SOX10* genes. OPCs were detected based on accessibility near the *PDGFRA* gene

The second dataset, Ziffra et. al.[102] was obtained from embryonic human brain of six individuals that were generated from dorsolateral prefrontal cortex (PFC), primary visual cortex (V1), primary motor cortex (M1), primary somatosensory cortex,

dorsolateral parietal cortex, temporal cortex, insular cortex, and the medial ganglionic eminence (MGE) in gestation weeks 17, 18, 20, and 21. This data consisted of 459,953 open chromatin peaks and 77354 cells. The data was processed to obtain 12 cell types, including new-born excitatory neurons, radial glia (RGs), intermediate progenitor cells (IPCs), deep layer (cortical layers V–VI) excitatory neurons (dlENs), upper layer (cortical layers II–IV) excitatory neurons (ulENs), MGE-derived cortical interneurons (IN-MGE), CGE-derived cortical interneurons (IN-CGE), insular neurons, progenitors from the MGE, microglia, oligodendrocyte progenitor cells (OPCs), endothelial and mural cells. To find the cell type labels they first clustered their data using Leiden algorithm and then generated a gene activity matrix by aggregating the fragments from gene body and promoter regions. Then they used marker genes to assign cell types to the cells. All the scATAC-seq data were obtained from scATAC-explorer[105] and cell type labels were downloaded from the links provided in the papers. Cell types of Corces were downloaded from supplementary data 2 of their paper and cell types of Ziffra were downloaded from https://cells.ucsc.edu/cortex-atac/genes/meta.tsv.

Single-cell ATAC-seq Data from Mouse:

Identifying the schizophrenia-relevant cell types in the mouse developing brain and comparing the results to the human developing brain can provide insights into the disease mechanisms. Here, I used one embryonic mouse scATAC-seq dataset that was generated from three time points (13.5, 15.5, 18.5) during the embryonic brain development [103]. The number of cells from the embryonic days 13.5, 15.5, and 18.5 were 12964, 16549, and 11088, respectively. Also, the number of open chromatin peaks for the embryonic days 13.5, 15.5, and 18.5 were 152179, 186038, and 147473, respectively. For each time point cortical tissue from 4 mice were pooled together to

create the dataset of that time point. They found these cell types in the three timepoints: Interneurons, Cajal-Retzius Cells, Microglia, Endothelial Cells, Oligodendrocytes, Deep Layer Callosal Projection Neurons, Layer6b Neurons, Upper Layer Callosal Projection Neurons, Near Projecting, Subcerebral Projection Neurons, Corticothalamic Projection Neurons, Layer4 Neurons, Immature Neurons, Migrating Neurons, Intermediate Progenitors, Astrocytes, Apical Progenitors.

Single-Cell Transcriptomic Data:

From the aforementioned datasets, the transcriptomic data was available only for the mouse embryonic dataset[103]. The Bella et al. data contains transcriptomic data at the single-cell level (i.e. scRNA-seq) at 11 different embryonic days, including 10.5, 11.5, 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, 18.5 , and two postnatal days including 1 and 4. The number of cells vary across different time points and is 2989, 4221, 9348, 8907, 5249, 11670, 5761, 9381, 20275, 13072, and 7174 respectively but the number of genes is 19712 across all of them. Both scATAC-seq and scRNA-seq data are available only for the embryonic days 13.5, 15.5, and 18.5. The cell types detected in the transcriptomic data are Cycling Glial Cells, Pericytes, Vascular and Leptomeningeal Cells, Red Blood Cells, Interneurons, Cajal-Retzius Cells, Microglia, Endothelial Cells, Oligodendrocytes, Deep Layer Callosal Projection Neurons1, Deep Layer Callosal Projection Neurons2, Deep Layer Callosal Projection Neurons, Layer6bNeurons, Upper Layer Callosal Projection Neurons, Near Projecting, Subcerebral Projection Neurons, Corticothalamic Projection Neurons, Layer4 Neurons, Immature Neurons, Migrating Neurons, Intermediate Progenitors, Ependymocytes, Astrocytes, Apical progenitors

4.3 Prioritizing Cell Types Using Linkage Disequilibrium Score (LDSC) Regression Analysis[61]

The schizophrenia GWAS data can be integrated with chromatin accessibility and/or transcriptomic data to prioritize the disease-relevant cell types. However, there are a few outstanding complications with these associations that must be considered before moving forward with the data integration. Linkage disequilibrium (LD) is one of the major complications with downstream analysis using GWAS data [106]. There is an existing correlation between alleles in the genome for many reasons, including allele proximity on the chromosome, mutation, genetic drift, and other confounding factors [106]. One of the main reasons is due to crossing over during meiosis. During this process, some regions of the genome are more likely to stay together than others. Therefore, for a sample disease phenotype, the causal variant may be present firmly in a large population. Still, it would be difficult to isolate and identify it as non-causal variants linked to the causal variant would also be simultaneously present in many positive cases and absent in many controls. In other words, non-causal SNPs in LD with a causal SNP will have inflated levels of association with a potential disease or the trait of interest.

I used a tool called LDSC (version 1.0.1) to account for this complication, by distinguishing between inflated test statistics from LD and other confounding biases found in statistical genetics[63,70]. The LDSC utilizes a stratified LD score regression and estimates the variance explained by all the SNPs on a chromosome when testing the association of a particular SNP to a phenotype. LDSC analysis is specifically designed for finding out how partitioned heritability can be explained by the risk variants that are located in specific genomic regions, which in this case, refers to open

chromatin regions. It ultimately employs a powerful and accurate correction factor, refining association data to show a true, unconfounded polygenic signal.

I integrated the GWAS of schizophrenia with the chromatin accessibility data (both bulk and single-cell) using LDSC-based partitioning heritability analysis, and identified specific cell types with the significant enrichment of schizophrenia-relevant variants (e.g., SNPs) on their chromatin accessibility sites. In addition, I integrated schizophrenia GWAS data with the scRNA-seq data from mouse[103] to identify the schizophrenia-relevant. My core partitioning heritability analysis workflow can be summarized in the following steps and is also shown in Figure 4.1:

- **Step 1:** Preparation of peak data from chromatin accessibility datasets**,** and gene expressions from scRNA-seq datasets
- **Step 2:** LD score regression calculation
- **Step 3:** GWAS integration

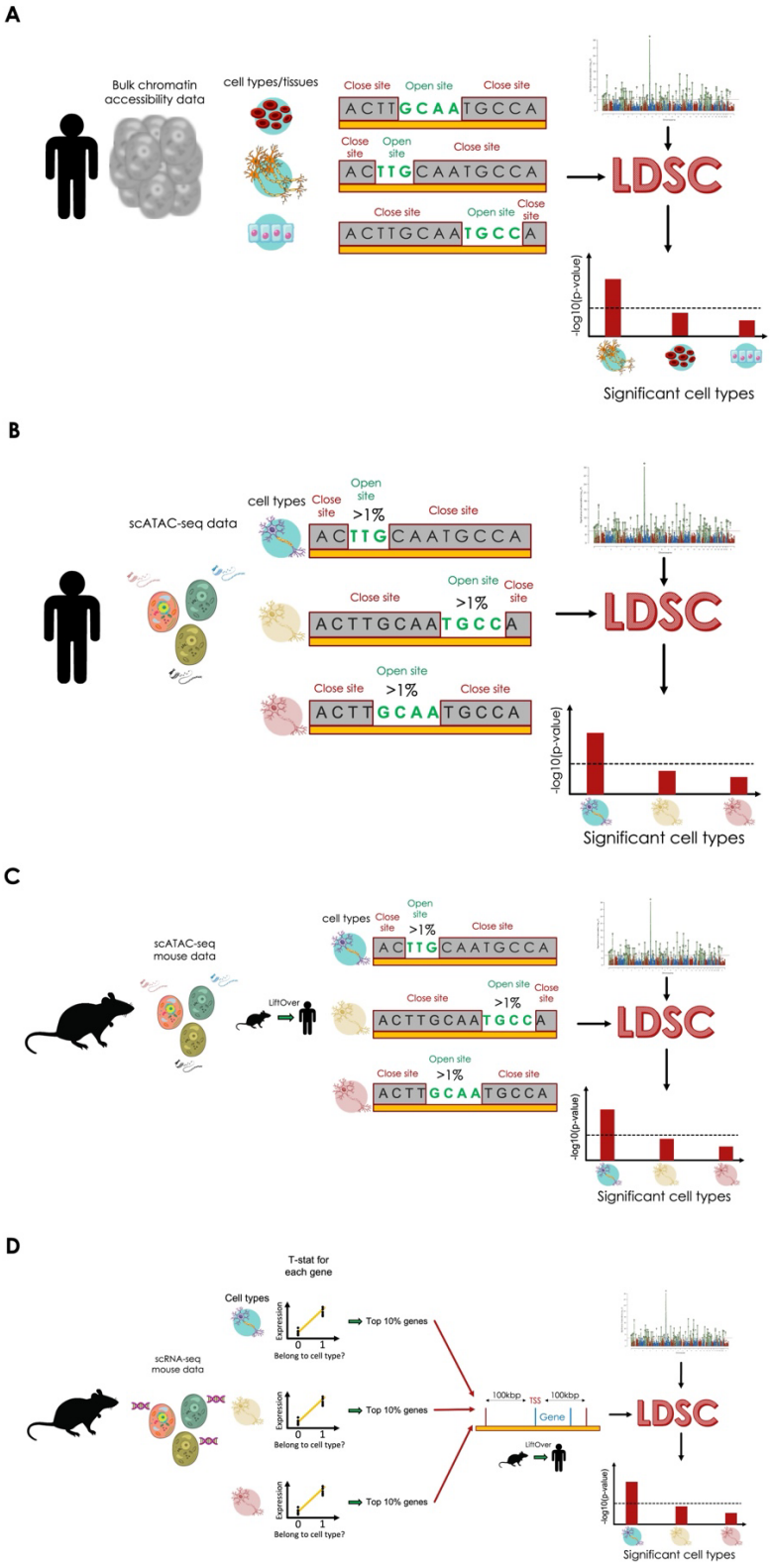


Figure 4.1: Partitioning heritability workflow. The workflow used in this study consists of four main sections each responsible for applying LDSC on a different data type including Bulk chromatin accessibility data (Figure 4.1A), scATAC-seq for

human data (Figure 4.1B), scATAC-seq for mouse data (Figure 4.1C), and scRNA-seq for mouse data (Figure 4.1D)

Step 1: Preparation of Peak Data from Chromatin Accessibility Datasets, and Gene Expressions from scRNA-seq Datasets

In my study, I used chromatin accessibility datasets both at the bulk and single-cell level, and prepared their peaks (i.e., accessible chromatin sites) information across various cell types and tissues in the BED file format, which contains data on the start and end positions of open chromatin regions along with the chromosome number. Here further details of peak data preparation unique to each dataset are provided. Also, data pre-processing steps for the scRNA-seq data is explained.

Preparation of Peak Data for Bulk Chromatin Accessibility Data

For each of the 194 cell types, I prepared the peak files in the BED format. This contains data on the start and end positions of open chromatin sites along with the chromosome number. I to each side of the peaks to accommodate for the flanking regions. Since both GWAS datasets and the bulk chromatin accessible data are in hg38 genome build, I did not need to lift the genome build.

Preparation of Peak Data for Human scATAC-seq

Similar to the bulk data I created BED files as an input for LDSC. For each peak, I calculated the percentage of cells in which the peak of interest was accessible, and removed the peaks with < 1% of cells being accessible in that site. Since both GWAS datasets and the bulk chromatin accessible data are in hg38 genome build, I did not need to lift the genome build.

Preparation of Peak Data for Mouse scATAC-seq

The genome build of the mouse data is mm10. I first used LiftOver[78] to lift the peaks of mouse scATAC-seq data from mm10 to hg38 build. I set the overlap parameter of LiftOver package as 0.7 as recommended by the package. I then removed the peaks with < 1% of cells that are accessible in that site.

Preparation of Gene Expressions Data for Mouse scRNA-seq

The data preparation step for scRNA-seq data is different from that of scATAC-seq data. I used a standard procedure for preparing scRNA-seq data for LDSC analysis[70] I first identified the equivalent mouse genes in the set of human genes by using the gene sets of human and mouse from Gencode[107] and filtered the gene by cell matrix of the mouse data based on these new sets of genes. Then for each cell type I fit a linear model for each gene: expression = a*I + b where "I" denotes whether the cell's cell type is the same as the cell type under investigation or not, expression is the expression level of each cell in the desired gene, 'a' is the slope of regression line, and 'b' is the intercept of the regression line. 'I' and expression are available for each cell based on the transcriptomic data. However, 'a' and 'b' calculated from finding the regression line that minimizes the sum of squares error for all the entries in the problem and calculate t-statistic for each of these lines. T-statistic can be calculated by dividing 'a' by the standard error of the coefficient estimate. The top ten percent genes based on t-statistic were chosen for each cell type. Then I extracted the transcription start site location of these genes from GENECODE[107] and added 100kbp on each side of those locations. I then combined all of these new locations with "reduce" function in R which resulted in a set of genomic locations for each cell type.

Reduce function merges two genomic regions by creating a new genomic region that has both of the previous genomic regions inside it. These genomic locations (chromosome number, start and end) were prepared in a BED format to be used for the LDSC analysis.

Step 2: LD Score Regression Calculation

The next step in the partitioned heritability workflow was to calculate the LD scores for each SNP found within the open chromatin regions. I first create ".bed" files (containing chromosome number, and start and end position of the peaks) for each cell type. Here I kept the peaks that were accessible in the cell type of interest. In another word, a peak was added to the bed file of a cell type if and only if it had a score greater than zero in the peak-by-cell type matrix. for each of the cell types, I then used the ".bed" file along with ".bim" PLINK files, containing information on known SNPs, to generate binary annotation files for each cell type and chromosomes 1–22. This

was done through the "make_annot.py" script provided by the LDSC toolkit from the LDSC github page (https://github.com/bulik/ldsc). At this stage, for every cell type, SNPs that are found within the inputted open chromatin regions were listed in the binary annotation files as 1s and the ones that are absent, as 0s. Then the "ldsc.py" script from the LDSC github page was used to calculate the LD scores for each SNP found within the open chromatin regions. This required the input of the binary annotation files, as well as the ".bim" PLINK file. I used the HapMap3 SNP data as a checklist of qualifying SNPs to include in the LD calculation. The resulting output files contained LD score information for every qualifying SNP. Here I provide a detailed description of the aforementioned steps.

First, I created annotation files for each bed file by using *make_annot.py* script from the LDSC package. The parameters set to --bed-file $path_to_our_bedfiles --bimfile $path_to_PLINK_bimfiles —annot-file $path_to_annotation_output_folder. This generates files to indicate whether a SNP in human genome is within the range of selected peaks for each cell type or not. The output of this step is a set of annotation files. For each cell type 22 annotation files are generated, each of which corresponds to one chromosome. This provided a list of 0s for the SNPs that were not overlapping peaks and 1s for the SNPs that were overlapping peaks. I obtained ".bim" PLINK files for hg38, which contained SNPs information for the human data with hg38 genome build through the Broad Institute website at https://alkesgroup.broadinstitute.org/LDSCORE/.

Second, I used the outputs of previous steps and calculated linkage disequilibrium (ld) scores using *LDSC.py* script from the LDSC package. The parameters were set to *--l2 --bfile $path_to_PLINK_files --ld-wind-cm 1 --annot $path_to_annotation_output_folder --thin-annot --out $path_to_annotation_output_folder --print-snps $path_to_HapMap3_SNPs*. ld score is defined as the sum of adjusted r2 which is an approximately unbiased estimator of the squared Pearson correlation. Hence, the output of this step is LD score files which contain LD scores of SNPs which are present in each cell type and are filtered based on HapMap3[108].

Step 3: GWAS Integration

The next step was the integration of the schizophrenia GWAS data with the chromatin accessibility data using the "ldsc.py" script from the LDSC github page (https://github.com/bulik/ldsc). I used the GWAS summary statistic files in conjunction with the generated LD files to create links between cell types and

schizophrenia SNPs using the LDSC toolkit. Here I set the parameters to --h2-cts $path_to_GWAS --ref-ld-chr $path_to_baseline --out $path_to_output --ref-ld-chr-cts $path_to_ldct_file --w-ld-chr $path_to_weights.

For each association between schizophrenia and cell types, a coefficient p-value was calculated, signifying the potential relevance of each cell type to the disease. Thus, a significant p-value represented a significant contribution of the open chromatin sites of a cell type to SNP heritability for the disease. To correct for multiple testing, a Bonferroni with a threshold of ≤ 0.05 was used to adjust the p-values within each cell type batch, that is, bulk, adult single-cell, and fetal single-cell[109].

This pipeline is an example of diagonal data integration because there are no shared cells or features between the datasets that I am integrating. Also, this pipeline falls into the category of intermediate integration because it does not involve combining input data (as in early integration) or building separate models for each dataset and combining their predictions (as in late integration). Instead, it explicitly addresses the multiplicity of datasets (schizophrenia GWAS and sequencing data) and fuses them through the inference of a joint model which is LDSC.

4.4 Fine-mapping Schizophrenia SNPs

GWAS summary statistics data for each disease provides a set of genotyped single nucleotide polymorphisms (SNPs), their P value of associations to the disease and their effect sizes. This may contain millions of SNPs. In order to identify a set of SNPs that are highly likely to be casual, fine-mapping approaches are used like Probabilistic Annotation INTegratOR (PAINTOR)[110], FGWAS[111], and Probabilistic Identification of Causal SNPs (PICS)[112] . In my study, I used a commonly used fine-mapping approach called Fgwas [111]. Fgwas takes a set of GWAS SNPs alongside other characteristics of the GWAS as its input. These include number of samples, Z-

score of each SNP (that can be calculated from p-values presented in the GWAS dataset), chromosome number, the position of the SNP on the chromosome, and minor allele frequency of the SNPs. Then Fgwas calculates the prior probability of association for each of the SNPs. For each association locus (i.e. a region in the genome with high association to the disease), I chose the smallest set of SNPs that together explain 0.95 of posterior probability of associations. This forms the 95% credible interval (CI) SNPs for each locus. This means that the causal SNP in the association locus is 95% likely to be found in CI SNPs set, if it is genotyped in the experiment.

4.5 Effect of Disease risk SNPs on Open Chromatin Sites

I investigate the effects of SNPs on open chromatin sites using the pipeline developed by my colleague Mr. Nader Hosseini Naghavi. This pipeline[113] investigates the effects of SNPs on bindings of transcription factors on open chromatin sites, and outputs the disease-relevant peaks, transcription factors, and genes for each cell type in the datasets.

This pipeline consists of three main steps:

Step 1: In the first step of the pipeline a set of SNPs that are highly likely to mediate risk to a complex disease are identified. First, I obtain a list of most associated SNPs (called lead SNPs) from a GWAS study (here schizophrenia GWAS[39]). Then I define a window of 2 Mbp centred around each of the most associated SNPs, and call it a disease locus. This forms multiple disease-associated loci, each of which have a length of 2 Mbp. Then GWAS summary statistics data along with information about each disease locus are fed to Fgwas fine-mapping algorithm[114]. For each locus, Fgwas outputs the smallest set of SNPs that together explain 95% or more of posterior probability of associations (PPAs). These are called credible interval (CI) SNPs. The

CI SNPs are a set of SNPs that are likely to contain the causal SNPs with a posterior probability of 95% or higher, if the causal SNP is genotyped or at least imputed. From the list of CI-SNPs, I select those SNPs that are highly likely to change the binding affinity of transcription factors. I therefore, use a package called atSNP[115] that uses JASPAR[116] database to identify a subset of CI-SNPs that can significantly change the binding affinity of transcription factors' motifs presented in this database. Fine-mapping by Fgwas followed by SNPs prioritization using atSNP method results in a set of SNPs that are highly likely to be functional and drive risk to the disease.

Step 2: In the second step of the pipeline, the risk-mediating peaks and transcription factors in each locus are identified. Here, the prioritized SNPs from Step 1 are overlapped with the peaks from scATAC-seq data to identify the cell type specific peaks that are likely to drive risk to the disease. For each cell cluster of scATAC-seq data (representing a cell type population), only those peaks that were accessible in at least 10% of cells of that cluster were kept. Risk-mediating transcription factors overlap accessible peaks of scATAC-seq data. Since atSNP identifies both (a) SNPs that are likely to change the binding affinity of a transcription factor; and (b) the affected transcription factors, simultaneously, the Step 2 of the pipeline outputs a list of disease-relevant transcription factors and their motif sites in each cell type (i.e. cell cluster of scATAC-seq data).

Step 3: In the third step of the pipeline a set of genes that are likely to be dysregulated in the disease will be identified. I use Cicero[117] to assess co-accessibility patterns between the risk-medicating peaks (obtained from Step 2) and peaks overlapping promoters of all genes in the 2Mbp locus. Here I define a gene promoter as a region located up to 2,000 bp upstream of the gene transcription start site (TSS). The genes

that their promoters are significantly co-accessible with the risk-mediating peaks (obtained in Step 2) are considered target genes.

I modified this pipeline as needed. For instance, the original pipeline was developed for the human ATAC-seq data. However, for using mouse ATAC-seq data, I converted the coordinates of mouse data to human coordinates so it can be used as an input for the pipeline. In another modification, I added a section to filter the indels from schizophrenia GWAS, so that it contains SNPs only. These modifications are mostly related to changing some parts of the package as described above, so that it can be compatible with the datasets that I have used in my study.

This pipeline is an example of diagonal data integration because there are no shared cells or features between the datasets that I am integrating. The pipeline consists of multiple steps and involves the utilization of various computational tools and algorithms to analyze the datasets. It does not combine the input data into a single table or representation (early integration), nor does it build separate models for each dataset and combine their predictions (late integration). Hence, it is an example of intermediate data integration.

4.6 Differentially Expressed Genes

Last step of the pipeline described in the previous section predicts disease relevant genes through assessing co-accessibility patterns between risk-mediating peaks and promoters of genes. In my study, I have developed a new pipeline for predicting disease genes. This analysis pipeline investigates the effects of risk-mediating peaks on the expression of genes in each disease locus. It is critical to acknowledge that the pipeline utilized in this section is a pilot study and a work in progress, and its limitations will be discussed later in this section. The primary objective of this pipeline is to identify a collection of genes that are impacted in individuals with

schizophrenia. Nevertheless, it is noteworthy that the detection of genes that are affected in schizophrenia is more reliant on the results of the previous pipeline. Here, I explain how this new pipeline works.

First, similar to the previous pipeline, I define disease association loci as 1 Mbp regions centred around lead SNPs of schizophrenia GWAS[39]. Then, for each disease locus, I identify all the genes that their promoters overlap the 1 Mbp locus. A promoter region is also defined as a 2,000 base pairs upstream of the gene transcription start sites (TSS). I considered all pairs of genes and risk-mediating peaks in the locus, where risk-mediating peaks are defined by Steps 1 and 2 of the previous pipeline, and genes are those that their promoters overlap the disease locus.

I then performed the following process for each cluster of cells, representing a cell type in the scATAC-seq data. For each risk-mediating peak, I divided cells of a cell type into two groups: (1) those cells in which the peak is accessible (called them "open" group); and (2) those cells in which the peak is not accessible (called them "close" group). For each peak-gene pair, I used Mann Whitney Wilcoxon test to assess the correlation between the accessibility patterns (open vs. close) of the peaks and the expression of the gene across matched cells of the cell type. This test identifies those pairs of risk-mediating peaks and gene that are significantly correlated (adjusted p value < 0.1); and hence, the gene is likely to be regulated by the risk-mediating peak.

Here, I faced a computational challenge. In order to assess the correlations between the accessibility patterns (open vs. close) of the peaks and the expression of the gene across matched cells, I needed to have gene expression data across cells (i.e. scRNA-seq). However, scRNA-seq data was not available for the datasets of this study. Therefore, I employed a deep-learning model called BABEL[3] to infer gene

expression data of individual cells from single-cell ATAC-seq data. BABEL gets a scATAC-seq data input in h5ad format and generates a sudo RNA-seq data for the same set of cells. I used BABEL with its pre trained model, and applied it to all the scATAC-seq datasets. Once pseudo-RNA-seq data was generated for all individual cells of each scATAC-seq data, I assessed the correlations between the accessibility patterns (open vs. close) of the peaks and the expression of the gene across matched cells for each cell type as described above. This resulted in the predication of the genes likely to be regulated by each risk-mediating ATAC-seq peaks, along with the cell types in which the correlation was significant.

Finally, I compared the genes predicated by the two approaches (i.e. co-accessibility patterns vs. direct ATAC-seq and RNA-seq correlations), and report those genes that are commonly detected by the two approaches and those that are unique to each approach. An illustration of this pipeline can be seen in figure 4.2.

This pipeline is an example of vertical data integration because I use different modalities of data for the same set of cells in order to develop my pipeline. The pipeline does not combine the input data into a single table or representation, nor does it build separate models for each dataset and combine their predictions. Hence, this pipeline is an intermediate data integration pipeline.

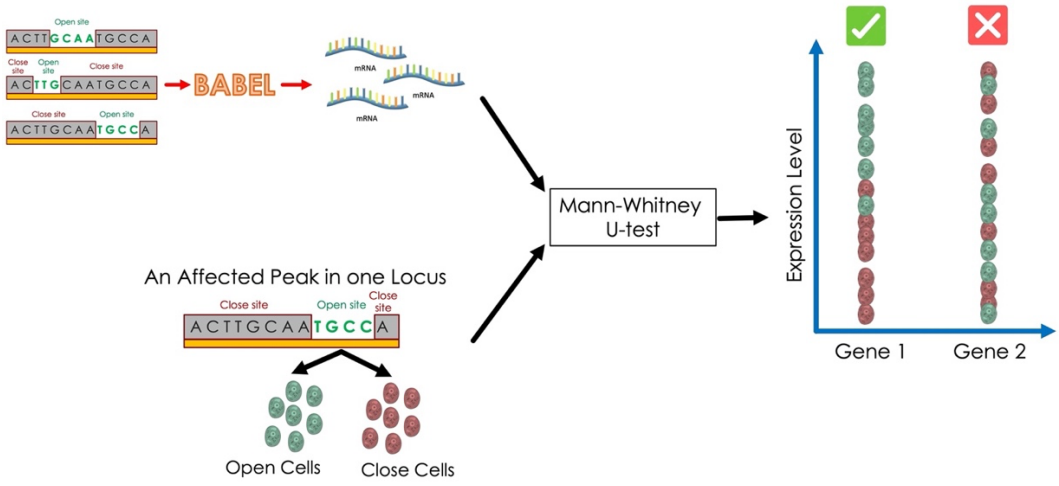


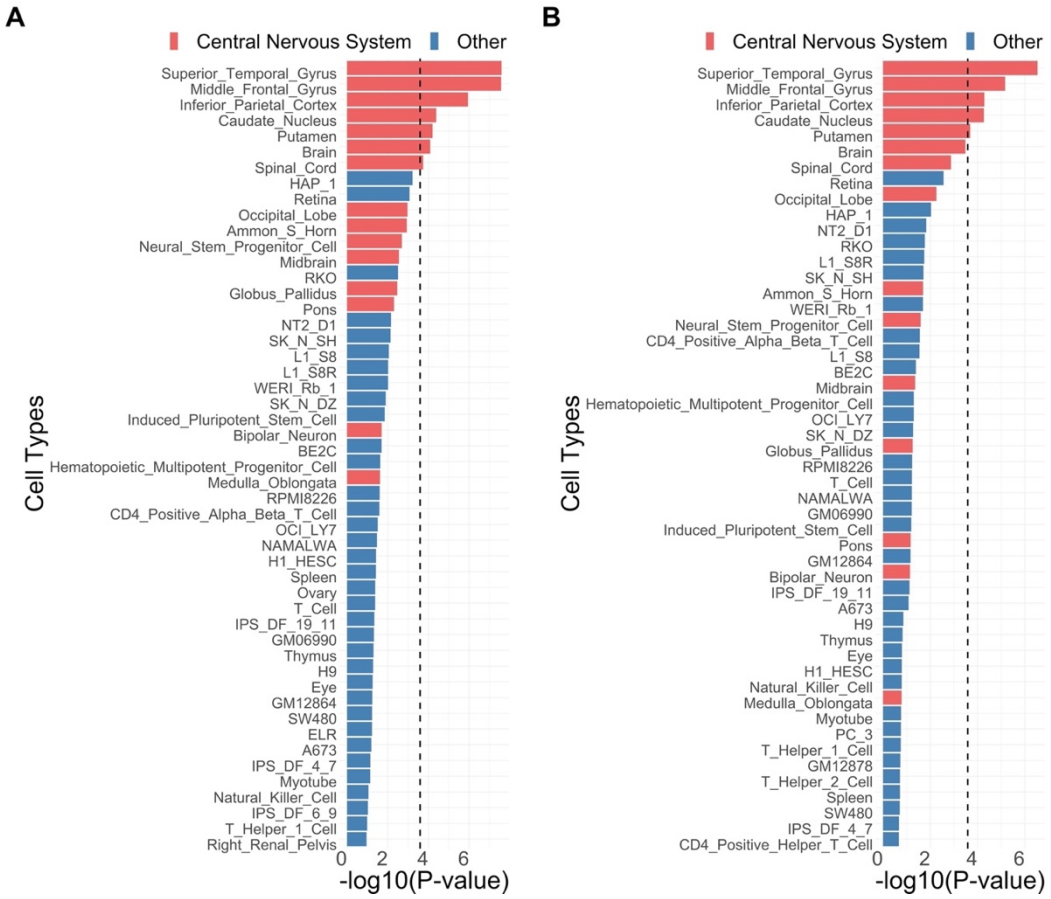
Figure 2.2: An illustration of the third method of integration. Cells are divided into two groups of open and close based on their accessibility in each peak. Using BABEL's output, gene expression of these two groups are compared using Mann-Whitney test to determine which genes are expressed significantly between these two groups of cells.

Chapter 5

Results & Discussion

In this chapter I present the results and discussions of the three integration methods explained in this study including prioritizing tissues and cell types relevant to schizophrenia using GWAS enrichment (first method of integration), prioritizing genes and transcription factors relevant to schizophrenia using a method developed in our lab (second method of integration), and prioritizing genes relevant to schizophrenia using BABEL (third method of integration)

5.1 Integrating GWAS data with bulk chromatin accessibility Data from Human



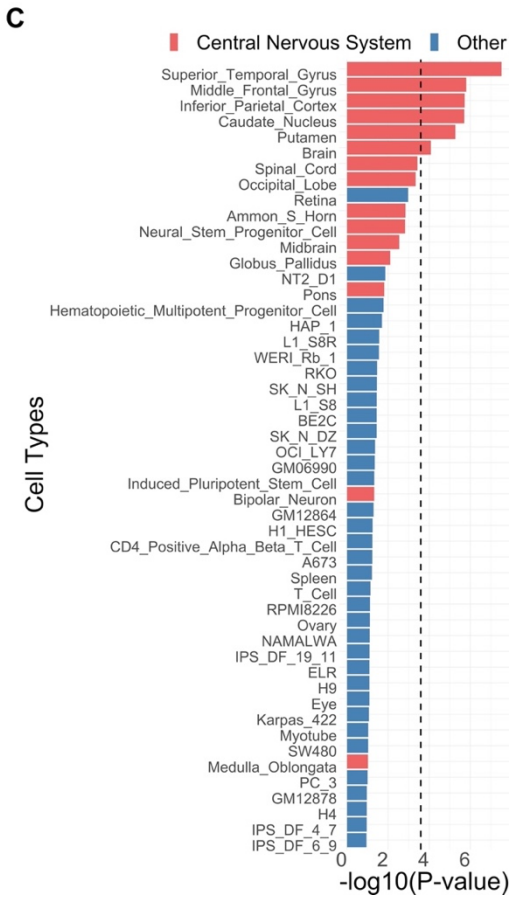


Figure 5.1: Bar plots of top 50 LDSC results for bulk chromatin accessibility data using three different GWAS data.including A.Ripke et al. GWAS, B.Pardinas et al. GWAS, and C. Li et al. GWAS. X axis shows the -log10(p-value) of significance of tissue/cell type in schizophrenia and Y axis shows the cell type/tissues. In all three analysis Only brain related cell type/tissues can pass the Bonferroni corrected threshold of 0.05

I analyzed the whole OCHROdb dataset and tested all of the tissues for enrichment of the Schizophrenia GWAS data in all three GWAS data that I tested.

Figure 5.1 shows that those tissues that pass the adjusted p value of 0.05 after Bonferroni correction for multiple testing are all from the central nervous system (CNS) and this is replicated using three GWAS datasets (Figure 5.1A, 5.1B, and 5.1C). Particularly, in the GWAS with the highest number of samples (Ripke et al[39]), where I have the highest power to detect significant cells, I found out that superior temporal gyrus, middle frontal gyrus, inferior parietal cortex, caudate nucleus,

putamen, and brain and spinal cord tissues are significant in terms of enrichment of schizophrenia risk variants on their open chromatin sites. Figure 5.1B and 5.1C show that the top cell types and tissues of two other GWAS datasets with smaller sample sizes are in the same order. In fact, the first seven most significant cell types and tissues are the top most significant cell types in all of the three GWAS datasets, and the only difference is in their level of significance (i.e., p values). Also, using the Ripke et al GWAS data[39], all of these seven cell types passed the Bonferroni corrected threshold of 0.05. In comparison, 6 and 5 cell types and tissues passed the threshold for Li et al GWAS data[40], and Pardinas et al GWAS data[41], respectively. Generally, these results are consistent with previous research in schizophrenia, where the brain and central nervous system tissues are shown to play an important role in developing risk to schizophrenia[118].

5.2 Integrating GWAS Data with Single-cell ATAC-seq Data from Human

The OCHROdb database contain chromatin accessibility data for different brain regions, where thousands of cells were sequenced. As described before, through the integration of GWAS data and bulk chromatin accessibility data, I have been able to identify brain-related tissues from different part of the brain and spinal cord that could be relevant to schizophrenia. The limitation here is that bulk chromatin accessibility data does not have a resolution at the single-cell level, and therefore, I may miss detecting specific cell types of brain that could be relevant to risk of developing schizophrenia.

To address this problem, I assessed the enrichment of schizophrenia GWAS risk variants on the open chromatin data at the single cell resolution. I first applied LDSC

regression analysis to integrate schizophrenia GWAS data[39–41] with an adult human post mortem scATAC-seq dataset[77] that I call Corces dataset. Knowing that schizophrenia risk can be developed during brain development, I also studied schizophrenia risk at the earlier stages of the brain development using an embryonic human scATAC-seq dataset[102] that I call Ziffra dataset. To assess reproducibility of the results and identifying the cell types are consistently enriched for schizophrenia risk variants, I applied my analysis pipeline to 3 different schizophrenia GWAS datasets.

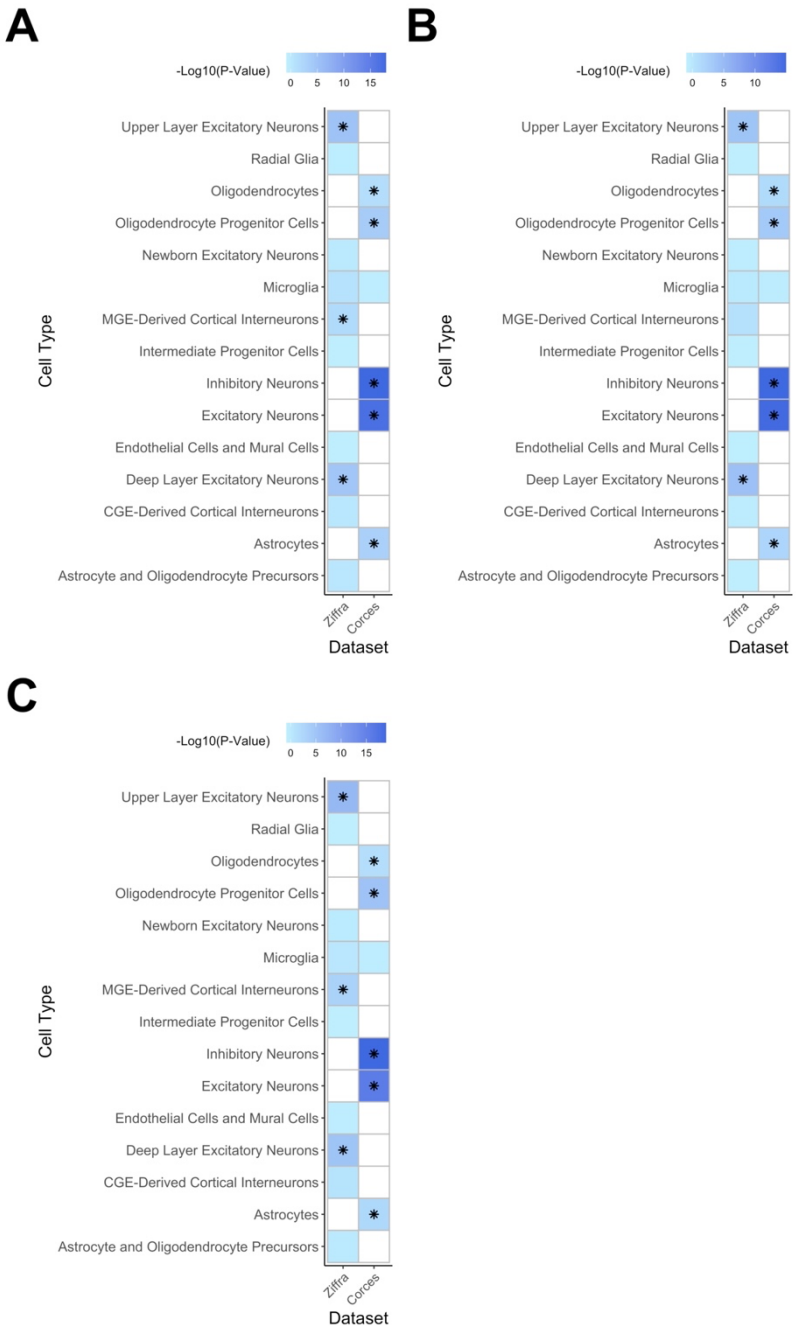


Figure 5.2: Results of applying LDSC on two human datasets and 3 different GWAS data. Including A. Ripke et al. GWAS, B. Li et al. GWAS, and C. Pardinas et al. GWAS. X axis shows the names of human datasets and Y axis shows the cell types. The intensity of squares shows the level of significance of the cell type to schizophrenia based on -log10(p-value) of LDSC analysis. The stars indicate the entries that pass the Bonferroni-corrected threshold of 0.05.

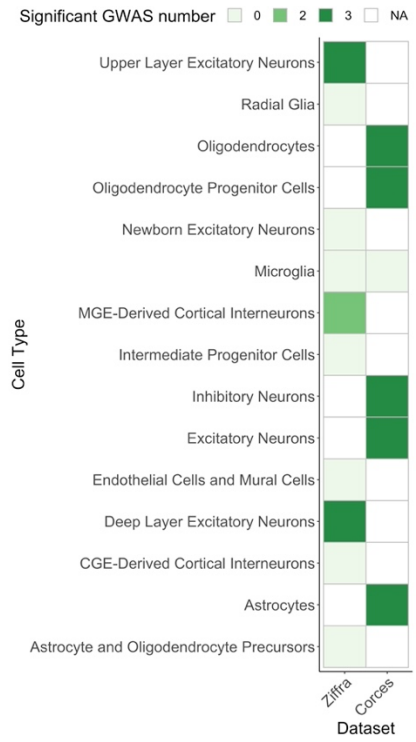


Figure 5.3: Comparison between the results of LDSC on multiple GWAS data. X axis shows the human datasets and Y axis shows the cell types. Intensity of the green color shows the number of GWAS datasets that the cell types are significant based on them.

The results of applying LDSC on the two human ATAC-seq datasets (called Ziffra and Corces) are presented in Figure 5.2. The cell types that did not exist in the dataset are presented with white blocks and the ones that pass the Bonferroni corrected threshold of 0.05 are marked with a star. Intensity of the color in each entry corresponds to the *-log10 of P-Value* assigned to that entry by LDSC. The greater this number, the more enriched that cell type is for schizophrenia. I found that cell types like Microglia, which exist in all the datasets, are not enriched in any of the GWAS datasets. Excitatory neurons are found to be enriched for schizophrenia in both Ziffra and Corces datasets except for Newborn Excitatory neurons. It is important to note that these results are the consistent across different GWAS datasets. Although Ripke et al GWAS data is the schizophrenia GWAS data with the largest sample size among all three GWASs, consistency between the results of these three GWAS in most of the

cell types is promising and can make us more confident about the signals that we are seeing due to their repetitive results. As one inconsistency in these results, I found that MGE-Derived Cortical Interneurons are marked significant in Pardinas et al and Ripke et al, but not for the other GWAS dataset.

5.3 Integrating GWAS Data with Single-cell ATAC-ceq Data from Mouse

Access to human brain data is challenging. In comparison, mouse datasets are more accessible and cane be generated in the wet labs. I therefore chose to apply LDSC regression analysis to a mouse single-cell sequencing dataset[103], and compared the results from human data to mouse data. I call the mouse dataset Bella. Bella mouse scATAC-seq datasets contains the data for embryonic days 13.5 and 15.5 and 18.5. Bella also contains scRNA-seq data for the embryonic days 10.5, 11.5, 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, and 18.5 and postnatal days 1 and 4.

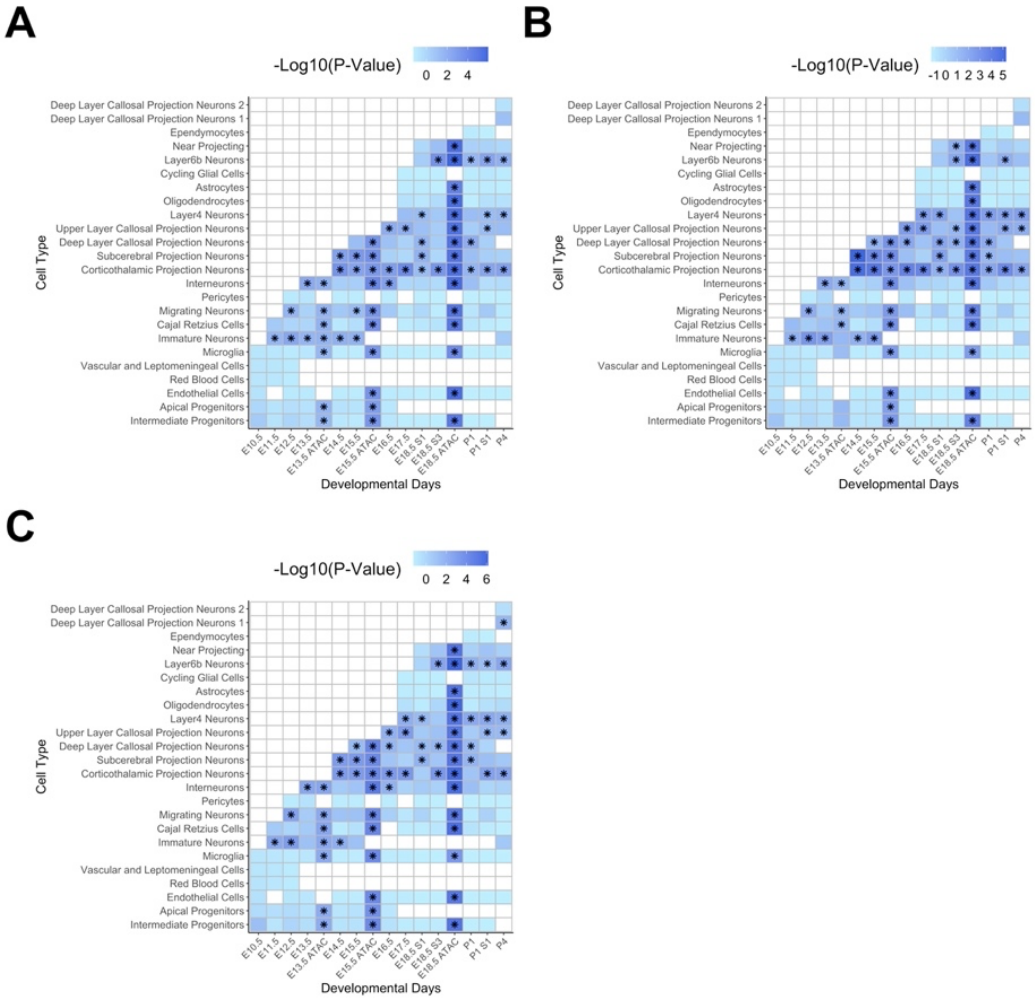


Figure 5.4: Results of applying LDSC on mouse ATAC-seq and RNA-seq datasets and 3 different GWAS data. including A. Ripke et al. GWAS, B. Li et al. GWAS, and C. Pardinas et al. GWAS. X axis shows the developmental stage of the mouse data for both ATAC-seq and RNA-seq and Y axis shows the cell types. The intensity of squares shows the level of significance of the cell type to schizophrenia based on -log10(p-value) of LDSC analysis. The stars indicate the entries that pass the Bonferroni-corrected threshold of 0.05.

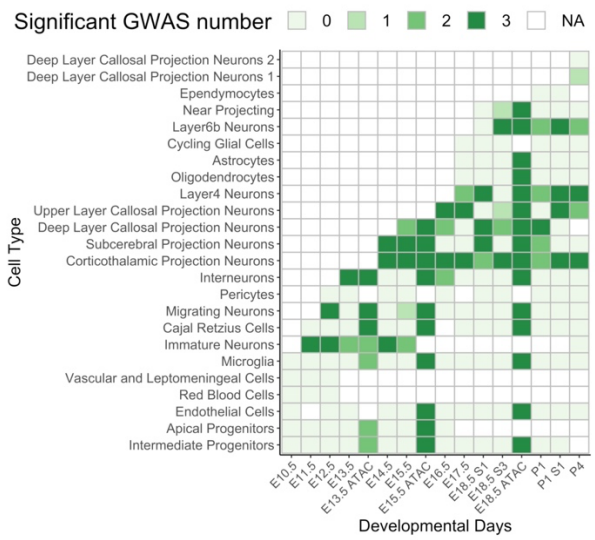


Figure 5.5: Comparison between the results of LDSC on multiple GWAS data. X axis shows the different mouse developmental stages in both ATAC-seq and RNA-seq and Y axis shows the cell types. Intensity of the green color shows the number of GWAS datasets that the cell types are significant based on them.

Results of applying LDSC on the Bella dataset in all scRNA-seq embryonic days and all scATAC-Seq embryonic days are shown in Figure 5.4. The cell types which are not present in a day are colored with white and the ones that pass the Bonferroni corrected threshold of 0.05 for each day are marked with a star. Intensity of the color in each entry is related to the value of -log10 of p value assigned to that entry by LDSC analysis. The greater this number, the more enriched that cell type is for schizophrenia. I found that some of the cell types such as Corticothalamic Projection Neurons are consistently enriched for schizophrenia in various GWAS datasets. Also. some other cell types such as Pericytes are consistently not enriched for schizophrenia in various GWAS data, which is plausible because Pericytes are blood cells and I expected to see significance enrichments in brain related cell types. Also, as expected, scATAC-seq data were more frequently significant than scRNA-Seq data in many cases. This might be due to the fact that changes occur at the epigenetic level, before they become apparent at the transcriptomic level.

As it is shown in Figure 5.5 the number of dark green color entries which is a symbol of consistency across three GWAS datasets is still abundant; however, inconsistencies in the mouse data are more than human data.

My analysis shows that enrichment of schizophrenia disease risk on brain cells starts at the early stages of the brain development. I noticed an enrichment at the embryonic day of 11.5 using scRNA-seq data, which is also consistent across the three GWAS datasets. This significance continues to exist all the way through Embryonic day 15.5, where most of the immature neurons have been differentiated into other brain cell types.

I also observed that cell types such as Microglia, Apical progenitors, Endothelial Cells, Intermediate progenitors, Cajal-Retzius Cells, Astrocytes and Oligodendrocytes are only significant in scATAC-seq data and not in scRNA-seq data.

Also, some of the cell types do not have a steady significance in the embryonic days. For example, Subcerebral Projection Neurons Appear in embryonic day 14.5 and show significance until E15.5 in both scATAC-seq and scRNA-seq data, then they lose their significance in E16.5 and E17.5 and then start showing significance again in E18.5.

5.4 Affected Transcription Factors of Human Datasets

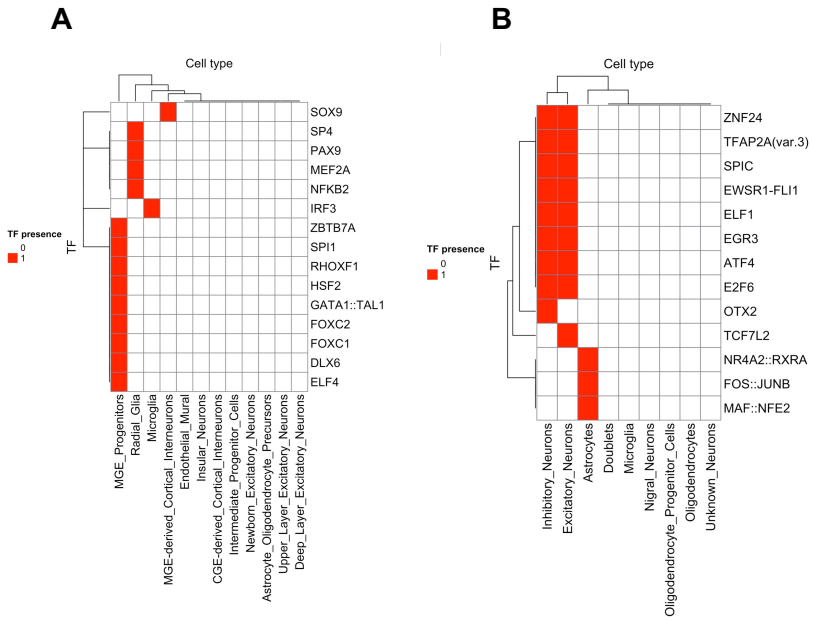


Figure 5.6: Results of applying the second method of integration on two human datasets including A. Ziffra and B. Corces to find the disease-affected transcription factors. X axis shows the cell types and Y axis shows the transcription factors. The affected entries are marked with red.

Figure 5.6 shows the results of applying the pipeline to identify the specific transcription factors in two human datasets (Ziffra[102] and Corces[119]). Fig. 5.6A shows that in the Ziffra et al. dataset (which has embryonic human brain cells), most of the affected transcription factors are found in MGE Progenitor cells. Also, some of the transcription factors could be affected in Glial cells such as Radial Glia and Microglia. In addition to these cell types, SOX9 is also affected in MGE-derived cortical interneuron cells. I did not find shared affected TFs in different cell types of Ziffra, and each affected TF was present in only one of the cell types.

As it can be seen in Fig. 5.6B, in the Corces dataset, Inhibitory and Excitatory neurons have the most affected transcription factors in them. They also share most of their affected transcription factors with each other. ZNF24, TFAP2A, SPIC, EWSR1-FLI1, ELF1, EGR3, ATF4, and E2F6 are the shared TFs between these cell types.

Alongside Excitatory and Inhibitory neurons, Astrocytes are the only other cell type in Corces that has significantly affected transcription factors.

5.5 Affected Transcription Factors of the Mouse Dataset

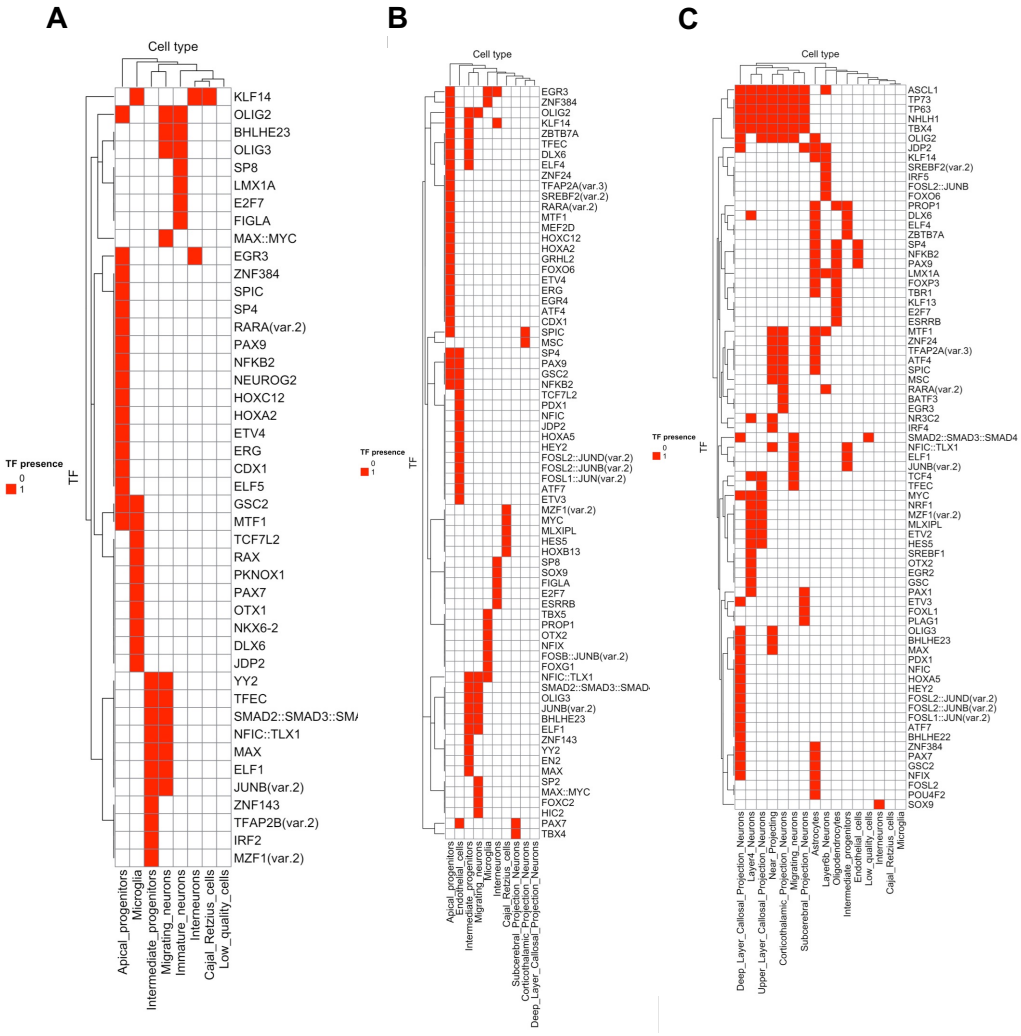


Figure 5.7: Results of applying the second method of integration on the mouse dataset in three different developmental stages including days 13.5 (A), 15.5 (B), and 18.5 (C) to find the disease-affected transcription factors. X axis shows the cell types and Y axis shows the transcription factors. The affected entries are marked with red.

Figure 5.7 shows the results of applying the pipeline to identify the schizophrenia-relevant transcription factors in three timepoints of the mouse dataset (Di Bella et

al.[120]). In this dataset, most of the cell type in all the timesteps has at least one affected transcription factor. However, there is one exception in the time point 15.5 (Fig. 5.7B), where Deep Layer Callosal Projection Neurons do not have any affected transcription factor. Interestingly, in the next developments timepoint (i.e. the embryonic day 18.5 (Fig. 5.7C)), this cell type is one of the cell types with the most number of affected transcription factors. The other exceptions occur in the embryonic day 18.5 and for the Cajal-Retzius and Microglia cells, where no affected transcription factors were found in these two cell types at this specific time point. However, both of these cell types had affected TFs in the previous time points, in embryonic days 13.5 (Fig. 5.7A) and 15.5 (Fig. 5.7B). Transcription factors that are affected in multiple cell types can be good candidates of the group of transcription factors that are important in schizophrenia since they are affected in a wide range of cells. Furthermore, the affected transcription factors with the highest number of shared cell types that occur in the embryonic day 13.5 are KLF14 and OLIG2 with 3 shared cell types. For embryonic day 15.5, EGR3, OLIG2, KLF14, and NFIC::TLX1 are found significant in 3 cell types. For embryonic day 18.5, the most shared affected transcription factors are ASCL1, TP73, TP63, NHLH1, TBX4, OLIG2, JDP2, PROP1, DLX6, SP4, NFKB2, PAX9, LX1A, MTF1, ZNF24, TFAP2A(variant 3), ATF4, SPIC, SMAD2::SMAD3::SMAD4, NFIC::TLX1, TCF4, MYC that are affected in 3 or more cell types. An interesting observation in the embryonic day 18.5 is that I can see an almost complete block of affected TFs between the first 6 TFs and these cell types: Deep Layer Callosal Projection Neurons, Layer4 Neurons, Upper Layer Callosal Projection Neurons, Near Projecting, Corticothalamic Projection Neurons, Migrating Neurons and Subcerebral Projection Neurons.

5.6 Affected Genes of Human Datasets

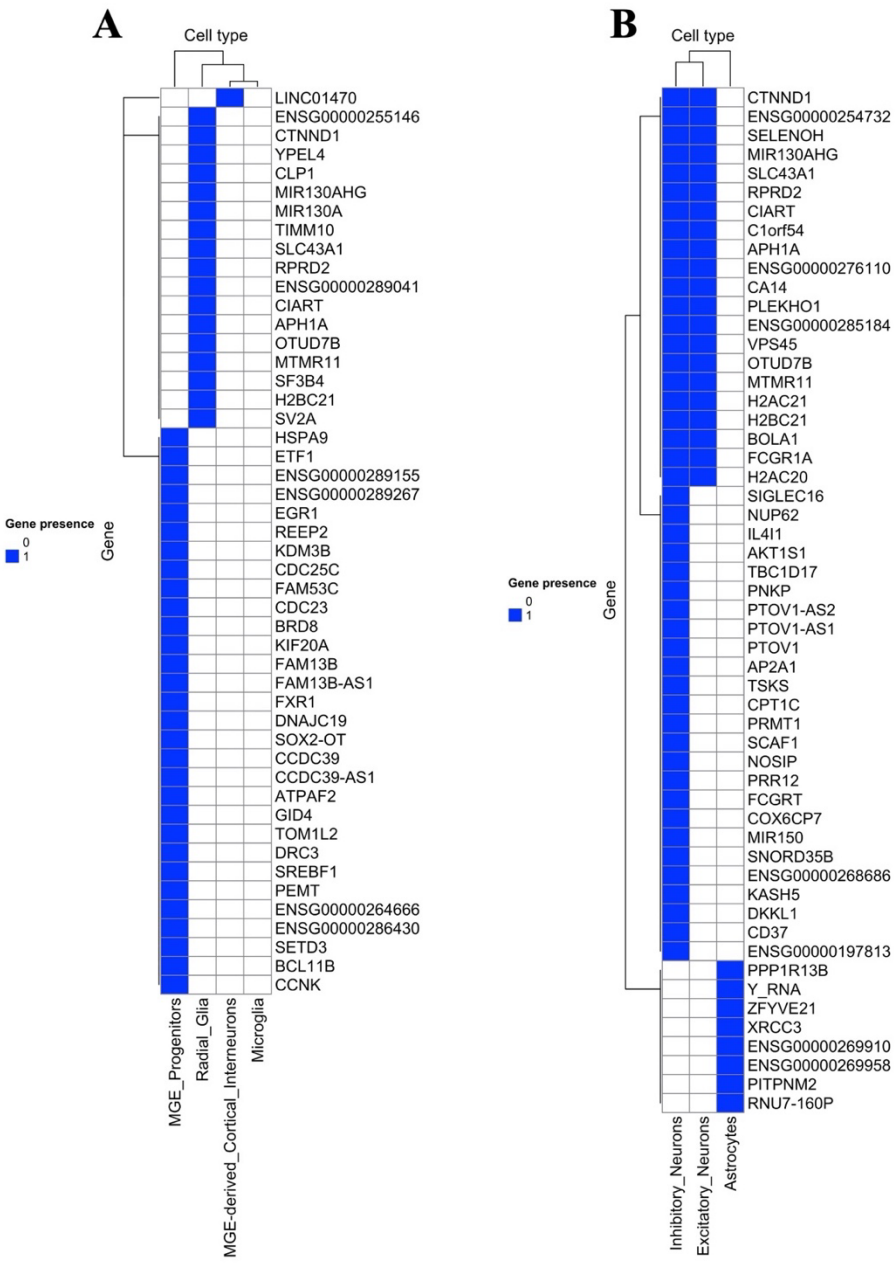


Figure 5.8: Results of applying the second method of integration on two human datasets including B. Ziffra and B. Corces to find the disease-affected genes. X axis shows the cell types and Y axis shows the transcription factors. The affected entries are marked with blue.

The next step of the pipeline identifies the specific genes. The results for the two human scATAC-seq (Ziffra and Corces) are shown in Figure 5.8. It can be seen that in

the Ziffra scATAC-seq dataset (Fig. 5.8A), which include cells from the embryonic brain, all of the affected genes are significant in the Radial Glia and MGE Progenitor cells, and none of these genes are shared between these two cell types. However, in Corces scATAC-seq dataset (Fig. 5.8B) that contains adult human brain, there are several significant genes shared between Inhibitory and Excitatory Neurons. These two cell types alongside Astrocytes form the set of cell types that affected genes of Corces are found in them. I found that many of the affected genes in Excitatory and Inhibitory neurons are the same and there is a complete block of affected genes shared between these two cell types.

5.7 Affected Genes of the Mouse Dataset

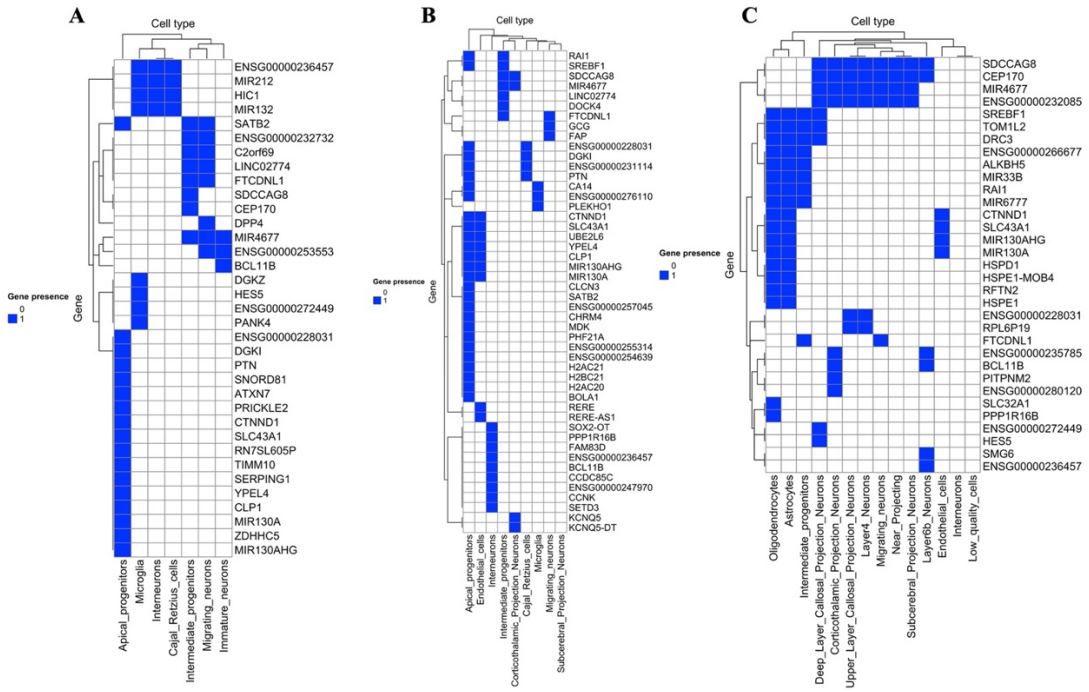


Figure 5.9: Results of applying the second method of integration on the mouse dataset in three different developmental stages including days 13.5 (A), 15.5 (B), and 18.5 (C) to find the disease-affected genes. X axis shows the cell types and Y axis shows the genes. The affected entries are marked with blue.

The results of applying the pipeline to find the specific genes in the mouse dataset at three different time points are shown in Figure 5.9. I found 35 affected genes in 7 cell types on the embryonic day 13.5 (Fig. 5.9A). In this time point MIR212, HIC1, MIR132, and MIR4677 are the most shared genes among the cell types with each of them being found in at least 3 cell types. I also found out that between all of these 7 cell types Apical Progenitors have the most affected genes in them. In embryonic day 15.5 (Fig. 5.9B), I found 48 affected genes in 8 cell types. There are no genes that are significant in more than 2 common cell types. Here, the cell type with the most affected genes is Apical Progenitors. I found that between the cell types that had at least one affected gene in the previous time point (i.e. day 13.5), all of them except immature neurons still have at least one affected gene in embryonic day 15.5.

For the embryonic day 18.5 (Fig. 5.9C), I found 33 affected genes across 12 cell types. Besides the last 8 genes shown in the bottom of Fig. 5.9C, all of the other ones are found in more than 2 cell types. In this timepoint, SDCCAG8, CEP170, MIR4677, and ENSG00000232085 have the highest number of common cell types. The first two genes (SDCCAG8 and CEP170) have 8 common cell types and the others have 7 common cell types. At this time point, Oligodendrocytes and Astrocytes have the highest number of affected genes. Compared to the previous time-point, Apical Progenitors and Microglia are not present in the list of cell types that have at least one affected gene. An interesting observation in these plots is the near complete blocks of affected genes versus cell types that happen in some parts of each plot. This can tell us that there may be a pattern of affected genes between those cell types or a pattern of shared cell types between those genes. For instance, between Microglia, Interneurons, and Cajal-Retzius cells in the embryonic day 13.5, Apical progenitors and Endothelial cells in embryonic day 15.5, Deep Layer Callosal Projection Neurons, Corticothalamic Projection Neurons, Layer6b Neurons, Layer 4 neurons, Subcerebral Projection neurons, Near Projecting and Migrating Neurons and Deep Layer Callosal Projection Neurons in embryonic day 18.5 and Oligodendrocytes, Astrocytes and Intermediate Progenitors also in embryonic day 18.5.

5.8 Affected Genes Based on the Third Method of Integration in Human Datasets

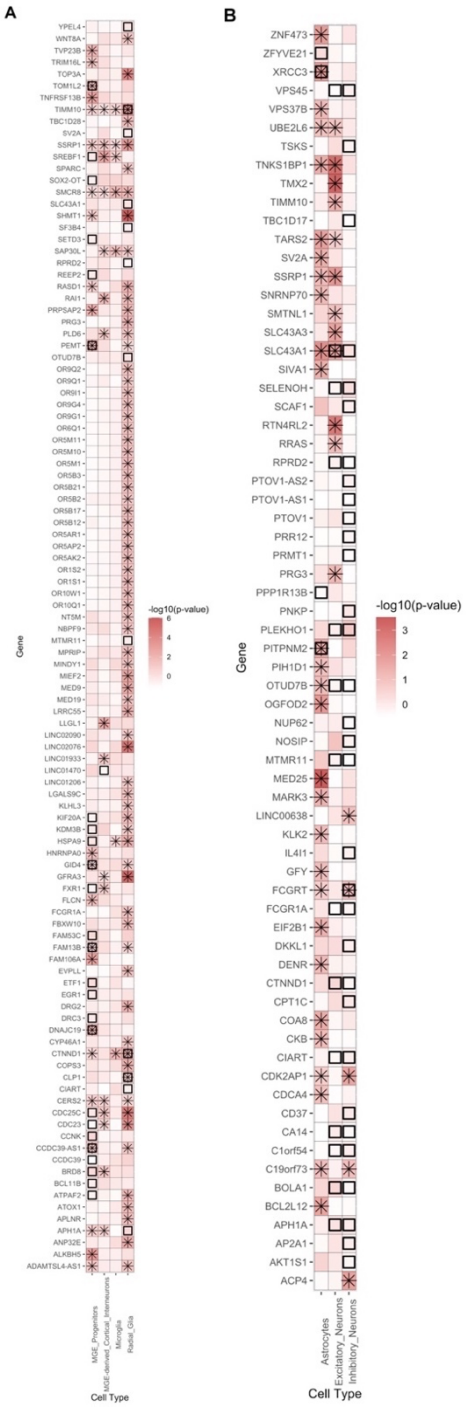


Figure 5.10: Results of applying the third method of integration to identify the differentially expressed genes based on the disease affected scATAC-seq peaks in two human datasets including A. Ziffra and B. Corces. The Y axis shows the genes and X axis shows the cell types. Affected genes found by the second method of integration are marked with a square and the ones found by the third method of integration are marked with a star.

As it was mentioned before, I used two data integration methods to predict disease-relevant genes. The combined results of the two approaches for the human datasets are shown in Figure 5.10. It can be seen that in the embryonic human dataset (Ziffra) (Fig. 5.10A)), 9 of the gene/cell-types pairs in the results are matching between the two gene prediction approaches. In the adult human dataset (Corces) (Fig. 5.10B) 4 of them match together. In Ziffra I found 4 cell types that have at least one significant gene. These cell types include Radial Glia, Microglia, MGE Derived Cortical Interneurons and MGE Progenitors. Between them, Radial Glia has the highest number of significant genes based on my pipeline, and the MGE progenitors cell type shows the most consistent results between the two gene prediction approaches with 6 gene/cell-types pair matches.

5.9 Affected Genes Based on the Third Method of Integration in the Mouse Dataset

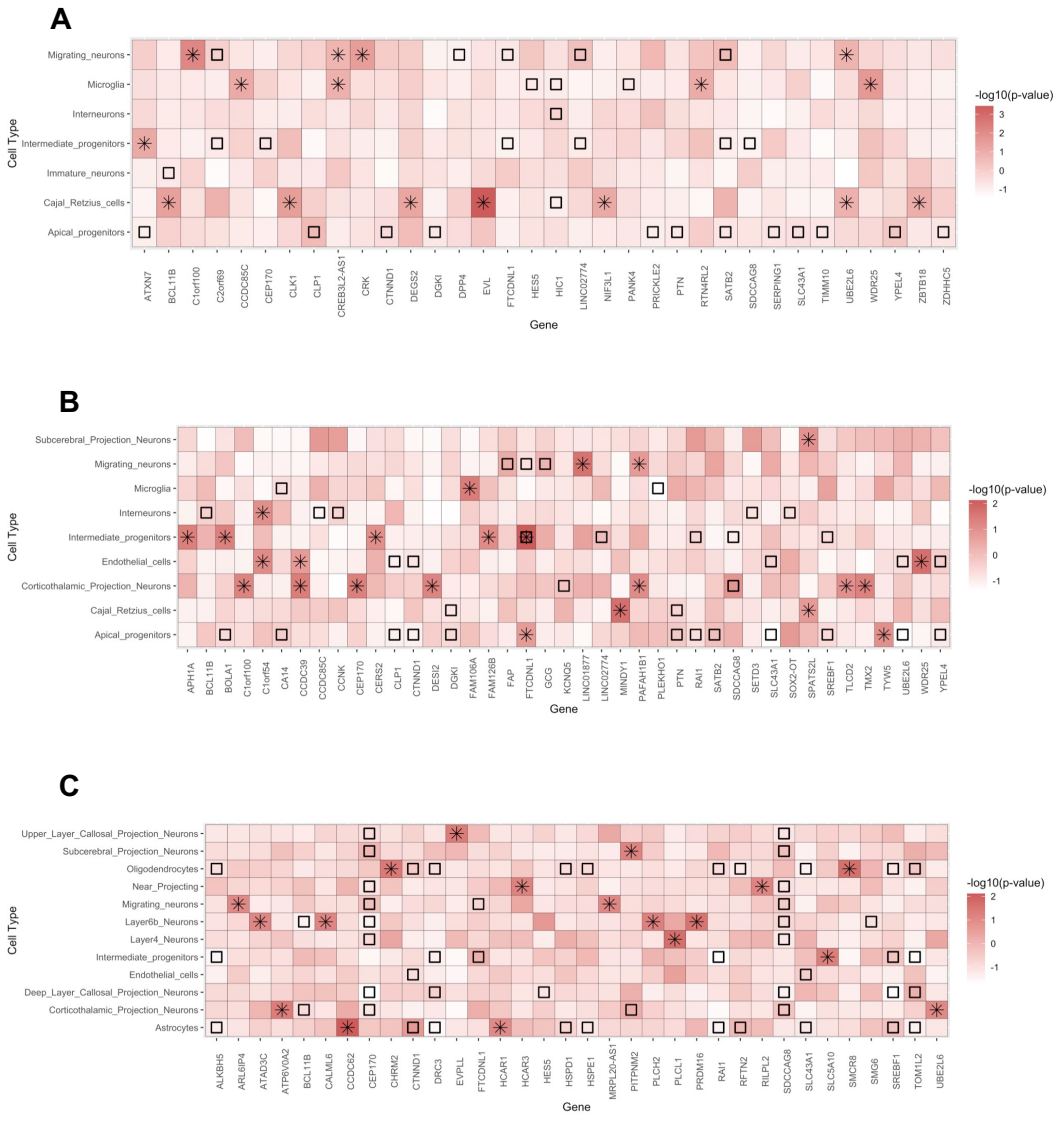


Figure 5.11: Results of applying the third method of integration to identify the differentially expressed genes based on the disease affected scATAC-seq peaks in the mouse dataset in three different developmental stages including days A. 13.5 and B. 15.5, and C.18.5. The X axis shows the genes and Y axis shows the cell types. Affected genes found by the second method of integration are marked with a square and the ones found by the third method of integration are marked with a star.

As it was mentioned before, I used two data integration methods to predict disease-relevant genes. The combined results of the two approaches for the mouse datasets are

shown in Figure 5.11. It can be seen that between all of the embryonic days only embryonic day 15.5 (Fig. 5.11B) has a gene-cell type pair that is considered significant in both gene prediction methods. This significant result is in the Intermediate progenitors and FTCDNL1. In the embryonic day 13.5 (Fig. 5.11A) Cajal-Retzius cells have the highest number of significant genes based on my method of data integration. Respectively, the top cell types are Corticothalamic Projection neurons and Layer6b neurons for the embryonic days 15.5 and 18.5, respectively (Fig. 5.11C).

In the embryonic day 13.5 two of the genes including CREB3L2-AS1 and UBE2L6 are significant in more than one cell type. Also, in embryonic day 15.5 five genes including C1orf54, CCDC39, FTCDNL1, PAFAH1B1, and SPATS2L have been found to be significant in more than one cell type. In comparison, none of the significant genes in embryonic day 18.5 have been found to be significant in more than one cell type.

5.10 BABEL's Performance

AUROC (Area Under the Receiver Operating Characteristics) is a measure of the performance of binary classification tasks, where the area under the receiver operator graph is calculated. The larger the area, the better the prediction is. Because BABEL has two encoders and two decoders, one for each data modality (RNA-seq and ATAC-seq), when generating pseudo gene expression from ATAC-seq data, it will also regenerate ATAC-seq from the input ATAC-seq using the ATAC decoder and generating pseudo ATAC-seq data from the shared latent space. Since single-cell ATAC-seq data has a binary nature, meaning a cell in a peak can be accessible (1) or not accessible (0), AUROC can be used to assess how well the model was able to

capture the input information in the shared latent space and how well it is performing on generating pseudo-data. The AUROC for Corces was 0.732, and for Ziffra was 0.724. The AUROC for Bella at embryonic day 13.5 was 0.694, at embryonic day 15.5 was 0.692, and at embryonic day 18.5 was 0.713.

The third method of integration heavily relies on BABEL's performance in predicting gene expression from chromatin accessibility data for the datasets. The AUROC results shows that there is still room for improvement of the BABEL predictions. One way to improve this would be to train BABEL on more relevant datasets to my use-case by collecting more multi-omics datasets related to the brain and retraining BABEL.

BABEL is trained on human data, which is why I lifted the mouse dataset coordinates from mm10 to hg38. However, when converting the coordinates of one organism to another, I lose some information from the input dataset. For instance, when lifting the ATAC-seq peaks of mouse to the human coordinates, some of the peaks do not pass the LiftOver threshold and will not be mapped to human coordinates and the information about the accessibility of these peaks will be lost in the downstream analysis. This, coupled with the fact that BABEL is trained on human datasets and may not capture all the information from the mouse datasets, may lower its prediction performance. One suggestion would be to train a separate model for mouse data and then predict gene expression in the Bella dataset using that model.

In conclusion, it is clear that there are many areas for improving the results of BABEL, which will directly impact the genes linked to schizophrenia using the pipeline. We are still working on this pilot study to generate more reliable results.

5.11 Discussion

Each of biology data types reveal a particular side of biological mechanisms. For instance, GWAS data reveals the significant variants associated to a disease, and chromatin accessibility data shows the open chromatin sites in individual cells (in the case of single cell data) or an average of thousands cells (in the case of bulk data). However, each of these data types cannot reveal the whole biological process behind complex biological mechanisms, such as a disease. Computational and statistical methods help us to integrate numerous biological data types to get better insights into the complex traits. Because technology improves constantly and also new researchers enter the research laboratories, new biology datasets are generated regularly and rapidly. All these new datasets have to be analyzed and put into different pipelines so I can get insights from them, as each of them is giving a unique kind of information about the biological processes. Even if two datasets come from the same modality or even from the same sample, they have their differences due to unstable natural features of organisms which are always changing, and I cannot get the same exact result from applying a pipeline on the same sample. To address this subject, in this study I try to apply multiple integration methods on multiple datasets.

Research attempts in this area usually focus on one integration of one modality with GWAS for instance GWAS and ATAC-Seq[61] or GWAS and RNA-Seq[121]. Also, the ones that have generated a dataset usually focus on their own dataset[102]. Therefore, it can be seen that most of the attempts focus on one dataset or one method and miss the insights that I can get by looking at different methods being applied on different datasets. Also, lots of researchers focus on one organism[62,102] but here I analyze both

human and mouse datasets and try to compare the results to see the differences or find the similarities.

During the study I have reached several conclusions about significance of cell types or tissues in schizophrenia. Here I point out the key findings in my investigations.

1- Schizophrenia GWAS risk variants are enriched in open chromatin sites of brain related cell types and tissues based on bulk epigenetic data.

The details of applying LDSC on bulk chromatin accessibility data is noted in the materials and the full results are shown in the results section. In conclusion, I confirm the idea about significance of brain related tissues and cell types in schizophrenia. As it can be seen in the results section all of the cell types and regions that pass the Bonferroni corrected threshold of 0.05 are brain related. I strongly confirm this finding by applying LDSC on 3 different GWAS datasets and comparing the results. These results indicate that I should focus on brain cells to unroll the biological mechanisms of schizophrenia.

2- GWAS sample size and ancestry affects LDSC results.

For each trait there are usually more than one GWAS study and each of them have their own characteristics. These features are different between studies and will probably result in different sets of variants being prioritized by the studies. Two of the most important features in GWAS datasets are the ancestry of samples (e.g. European vs. Asian) and also the number of cases and controls in the study. These differences will affect the results of analyzing GWAS data. In this analysis I used three different Schizophrenia GWAS data and noted differences and similarities across multiple GWAS datasets. The level of consistency between the results are different between the

datasets but mostly they have at least one difference. Hence, I found that choosing the right GWAS data is a crucial step in these kinds of analyzes, as it can affect the set of significant cell types. However, the cell types that are considered significant in all of the three GWAS data gives us more confidence about the importance of them since they are significant even based on different GWAS datasets.

3- It is already know that epigenetic changes usually occur before than changes in gene expressions can be detected [122]. Hence, before starting this investigation I predicted to see more signals in the ATAC-seq data of the mouse dataset compared to its scRNA-seq data. I confirm this by the results that I got from applying LDSC on mouse brain embryonic ATAC-seq and RNA-seq data in three different stages of the brain development. As it can be seen in the results section in all of these three timepoints that I have both RNA-seq and ATAC-seq data, significant enrichments are more abundant in scATAC-seq data compared to scRNA-seq. Although the list of cell types existing in these two modalities slightly differ in the three timepoints, but the cell types that are considered significant in ATAC-seq are always found significant in RNA-seq data. Also, there are other cell types that are significant in scATAC-seq data that have not been found significant in the scRNA-seq data. This finding is also confirmed between all three GWAS data.

4- Significant cell types

After carefully reviewing the results, I saw that excitatory and inhibitory neurons in human neurons are consistently significant across three GWAS datasets. Also, in mouse data I found out that these cell types are most significant on most of the time-points and consistent in 3 GWAS data:

- Layer6b Neurons that can be found from the Embryonic day 18.5

- Layer4 Neurons in the last stages of development
- Subcerebral projection neurons at first stages of development
- Corticothalamic projection neurons which can be found starting from embryonic day 14.5 and are significant in all other timepoints
- Interneurons at first stages of development
- Migrating neurons at first stages of development
- Immature Neurons that are only present until embryonic 15.5.

In the first pipeline, I ran LDSC on 3 publicly available datasets in order to find the significant cell types in schizophrenia. LDSC finds the most significantly associated cell types by integrating GWAS data with the annotations obtained from ATAC-seq or RNA-seq data. However, LDCS does not identify the genes and transcription factors that play an important role in schizophrenia. To identify these elements, I applied a previously developed pipeline in the lab on the same 5 datasets and developed a new pipeline to study this problem from a different perspective.

In my developed pipeline I benefit from the abilities of a deep learning model called BABEL that has been published by Wu et al. BABEL's ability to generalize between different cell types and organisms convinced me to use it as a part of the pipeline and expand on it to develop another method to predict the significant genes in schizophrenia.

The results that I capture by applying one method of integration on the selected datasets only focus on one aspect of integration and I can rely on them based on that specific method. However, if I look at the problem from a different view, I might capture information that one single method is not able to achieve. Hence, if I rely on only one method, I might miss some important results. For instance, by looking at the combined results of the previously developed method in Shooshtari Lab and my

recently developed method, it can be seen that there are multiple significant genes that are not shared between the results of the two methods and if I only would have relied on one method, I could have missed some of the important genes in schizophrenia. An advantage of this work is that one can look at the results of multiple methods and overlap them with each other. The results that overlap between the two methods can give us more confidence about the significant entries. For instance, in the mouse dataset, I saw that FTCDNL1 in the Intermediate progenitors cell type is considered a significant gene based on both methods. This will make us more confident about the finding since this gene has been confirmed to be significant based on two different methods of integration. However, I cannot conclude that the results that do not match are not important since these methods look at the problem from different perspectives, therefore, one approach may be able to capture some results that the other one misses. Some of the genes that I found significant based on my pipeline have been previously found to be causal in schizophrenia. Legge et al.[123,124] have done a review on the causal genes of schizophrenia based on transcriptome-wide association studies. They mention three main studies on this subject[120,125–127]. By looking into the results of these studies I found that VPS45, RPRD2, XRCC3, ZFYVE21, PPP1R13B, NOSIP, SREBF1, TOM1L2, GID4, FXR1, FAM53C, ETF1, and HSPA9 in Corces and Ziffra human datasets and MIR4677, LINC02774, UBE2L6, YPEL4, TOM1L2, ALKBH5, MIR130A, and BOLA1 in the mouse dataset (Bella) have been previously found causal in at least one of these studies.

From the transcription factors that I found to be significantly affected in schizophrenia, many of them have been considered relevant to schizophrenia in the previous studies. For instance, between the transcription factors that I found in the human study, SOX9[128], SP4[128,129], ATF4[130], EGR3[130,131], OTX2[132], TCF7L2[133], and

TFAP2A[123] have connections with schizophrenia. Also, from the transcription factors that were identified using mouse data, EGR3[130,131], LMX1A[134], MZF1[134,135], NEUROG2[134–136], TCF4[137], TBR1[138], SREBF1[138,139], POU4F2[140], NRF1[141], and NR3C2[141,142] were previously found to be causal in schizophrenia.

To summarize, in my study I have been able to predict transcription factors, genes, regulatory sites and cell types that are likely to be relevant to schizophrenia. Some of these findings have been already uncovered by previous studies, and some others are novel findings that can be experimentally validated in a wet lab setting.

Chapter 6

Conclusions and Future Work

In this thesis, I applied three different data integration pipelines that use statistical and machine-learning-based methods on human and mice data in order to prioritize significant cell types and biological elements in schizophrenia. The results presented in this thesis can be used to further investigate the underlying mechanisms of gene regulation in schizophrenia that may eventually results in better treatment options for schizophrenia patients. Most of the previous works in this area only focus on one dataset, one organism, or one method of integration. However, in this study, I applied three data integration methods on two organisms and multiple datasets, creating a more comprehensive study of data integration on schizophrenia.

In chapter 4, I presented a new integration pipeline that has been developed by me. This pipeline right now is in R script files and Jupyter notebooks. An improvement to this work could be making a single package that gets the inputs needed for the pipeline in a standard format and outputs the significant results in the format of figures and tables. This can help the pipeline to be more accessible to the researchers since they would be able to use it easily without going through the R and python codes and it can be applied in the investigation of other complex traits, as my pipeline is generalizable and the researchers would only need to provide the input files for their trait of interest in a standard format.

The data integration approach applied here in a computational biology application can be used in other computer science subjects or other field that uses computational

methods in their investigations. As discussed in the previous chapters, data integration methods are applicable to a wide range of scientific areas and although these areas are exploring different problems, they may benefit from the data integration methods used in this study and the way I handled the challenges that arose while developing and applying multiple integration methods.

Although applying computational methods to biological data can give us valuable information about the significance of biological components in schizophrenia, these methods have their own limitations. For instance, in the third method of integration, I use a deep learning model to find differentially expressed genes based on the chromatin accessibility sites that are likely to drive risk to schizophrenia. This deep learning model helps me to predict gene expression levels in each cell based on the accessibility patterns of open chromatin sites. However, predictions based on deep learning models are predication, and may not provide us with the exact gene expression values. These are statistical models and their accuracy even on the data that they are trained on is not 100 percent.

Statistical and machine-learning-based methods are updating constantly, and researchers try to improve previous methods to produce more accurate results. The methods used in this study are not exempt from such phenomena. Hence, using the future's state-of-the-art models that beat the current methods in performance, can lead us to more accurate results. For instance, the deep learning method used in this study can improve in many aspects like enhancing the architecture of the auto-encoder used in it or improving the loss function that it uses for the training process. By applying these enhancements to the model used in this study, one can get a better prediction of expression data from the chromatin accessibility data and would ultimately help

prioritizing schizophrenia-relevant genes affected by chromatin accessibility patterns in a more precise way.

Genes can affect each other, and studying the interactions between genes can give more insights into the disease under investigation. Pathway analysis is a standard way to find the relations between genes and their roles in biological mechanisms. In this study, I have identified schizophrenia-relevant chromatin accessibility sites. By using this data as input for packages like GREAT[143] I can further study disease-affected genes and the biological pathways that they are enriched in.

In conclusion, I have developed a standard data integration pipeline to prioritize biological elements in schizophrenia which help further investigations toward understanding this complex disease and providing better treatment options for it ultimately. The results in this study and ideas for the future developments will improve the data integration methods and also our understanding of mechanisms of schizophrenia.

Bibliography

1. McCutcheon, R. A., Marques, T. R. & Howes, O. D. Schizophrenia—An Overview. *JAMA Psychiatry* **77**, 201–210 (2020).
2. Dennison, C. A., Legge, S. E., Pardiñas, A. F. & Walters, J. T. R. Genome-wide association studies in schizophrenia: Recent advances, challenges and future perspective. *Schizophr. Res.* **217**, 4–12 (2020).
3. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2023070118 (2021).
4. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25**, 1499–1507 (2015).
5. Peng, G., Cui, G., Ke, J. & Jing, N. Using single-cell and spatial transcriptomes to understand stem cell lineage specification during early embryo development. *Annual Review of Genomics and Human Genetics* **21**, 163–181 (2020).
6. Colomé-Tatché, M. & Theis, F. J. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology* **7**, 54–59 (2018).
7. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends Biotechnol.* **38**, 1007–1022 (2020).
8. Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **22**, bbaa042 (2020).
9. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
10. https://www.canada.ca/en/public-health/services/publications/diseases-

conditions/schizophrenia-canada.html.

11. Goeree, R. *et al.* The economic burden of schizophrenia in Canada in 2004. *Curr. Med. Res. Opin.* **21**, 2017–2028 (2005).
12. Glantz, L. A. & Lewis, D. A. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Arch. Gen. Psychiatry* **57**, 65–73 (2000).
13. Harrison, P. J. The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. *Psychopharmacology (Berl.)* **174**, 151–162 (2004).
14. Byne, W., Hazlett, E. A., Buchsbaum, M. S. & Kemether, E. The thalamus and schizophrenia: current status of research. *Acta Neuropathol.* **117**, 347–368 (2009).
15. Howes, O. D. & Kapur, S. The dopamine hypothesis of schizophrenia: version III--the final common pathway. *Schizophr. Bull.* **35**, 549–562 (2009).
16. Andreasen, N. C. & Pierson, R. The role of the cerebellum in schizophrenia. *Biol. Psychiatry* **64**, 81–88 (2008).
17. Glausier, J. R. & Lewis, D. A. Dendritic spine pathology in schizophrenia. *Neuroscience* **251**, 90–107 (2013).
18. Lewis, D. A. & Moghaddam, B. Cognitive dysfunction in schizophrenia: convergence of gamma-aminobutyric acid and glutamate alterations. *Arch. Neurol.* **63**, 1372–1376 (2006).
19. Haroutunian, V., Katsel, P., Dracheva, S. & Davis, K. L. The human homolog of the QKI gene affected in the severe dysmyelination “quaking” mouse phenotype: downregulated in multiple brain regions in schizophrenia. *Am. J. Psychiatry* **163**, 1834–1837 (2006).

20. Monji, A., Kato, T. & Kanba, S. Cytokines and schizophrenia: Microglia hypothesis of schizophrenia. *Psychiatry Clin. Neurosci.* **63**, 257–265 (2009).
21. Notter, T. Astrocytes in schizophrenia. *Brain Neurosci. Adv.* **5**, (2021).
22. Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9**, 1415–1423 (2000).
23. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
24. Yao, J., Pan, Y.-Q., Ding, M., Pang, H. & Wang, B.-J. Association between DRD2 (rs1799732 and rs1801028) and ANKK1 (rs1800497) polymorphisms and schizophrenia: a meta-analysis. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168B**, 1–13 (2015).
25. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* **71**, 877–892 (2002).
26. Petryshen, T. L. *et al.* Genetic investigation of chromosome 5q GABAA receptor subunit genes in schizophrenia. *Mol. Psychiatry* **10**, 1074–88, 1057 (2005).
27. Egan, M. F. *et al.* Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 6917–6922 (2001).
28. Straub, R. E. *et al.* Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am. J. Hum. Genet.* **71**, 337–348 (2002).
29. Chowdari, K. V. *et al.* Association and linkage analyses of RGS4 polymorphisms in schizophrenia. *Hum. Mol. Genet.* **11**, 1373–1380 (2002).
30. Egan, M. F. *et al.* Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12604–12609 (2004).

31. Quednow, B. B., Brzózka, M. M. & Rossner, M. J. Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. *Cell. Mol. Life Sci.* **71**, 2815–2835 (2014).
32. Michaelson, J. J. *et al.* Neuronal PAS domain proteins 1 and 3 are master regulators of neuropsychiatric risk genes. *Biol. Psychiatry* **82**, 213–223 (2017).
33. Zhang, Z. & Zhao, Y. Progress on the roles of MEF2C in neuropsychiatric diseases. *Mol. Brain* **15**, 8 (2022).
34. Sanjuán, J. *et al.* Association between FOXP2 polymorphisms and schizophrenia with auditory hallucinations. *Psychiatr. Genet.* **16**, 67–72 (2006).
35. Sheth, F. *et al.* A novel case of two siblings harbouring homozygous variant in the NEUROG1 gene with autism as an additional phenotype: a case report. *BMC Neurol.* **23**, 20 (2023).
36. Johansson, A.-S., Owe-Larsson, B., Hetta, J. & Lundkvist, G. B. Altered circadian clock gene expression in patients with schizophrenia. *Schizophr. Res.* **174**, 17–23 (2016).
37. single nucleotide polymorphism / SNP. https://www.nature.com/scitable/definition/snp-295/.
38. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
39. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
40. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).
41. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-

intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

42. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
43. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1–15 (2021).
44. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1 (2015).
45. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, (2010).
46. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, (2010).
47. Chromatin. https://www.nature.com/scitable/definition/chromatin-182.
48. Wu, C., Wong, Y. C. & Elgin, S. C. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**, 807–814 (1979).
49. Bernardini, M. *et al.* High-Resolution Mapping of Genomic Imbalance and Identification of Gene Expression Profiles Associated with Differential Chemotherapy Response in Serous Epithelial Ovarian Cancer. *Neoplasia* **7**, 603-IN20 (2005).
50. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
51. Deng, T. *et al.* Functional compensation among HMGN variants modulates the DNase I hypersensitive sites at enhancers. *Genome Res.* **25**, 1295–1308 (2015).

52. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2013).
53. Koohy, H., Down, T. A., Spivakov, M. & Hubbard, T. A comparison of peak callers used for DNase-Seq data. *PLoS One* **9**, e96303 (2014).
54. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
55. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224–226 (2012).
56. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
57. Shooshtari, P. *et al.* OCHROdb: a comprehensive, quality checked database of open chromatin regions from sequencing data. *bioRxiv* (2018).
58. Shooshtari, P., Huang, H. & Cotsapas, C. Integrative genetic and epigenetic analysis uncovers regulatory mechanisms of autoimmune disease. *Am. J. Hum. Genet.* **101**, 75–86 (2017).
59. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
60. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
61. Das, A. C. *et al.* Single-Cell Chromatin Accessibility Data Combined with GWAS Improves Detection of Relevant Cell Types in 59 Complex Phenotypes. *Int. J. Mol. Sci.* **23**, 11456 (2022).
62. Hook, P. W. & McCallion, A. S. Leveraging mouse chromatin data for heritability enrichment informs common disease architecture and reveals cortical

layer contributions to schizophrenia. *Genome Res.* **30**, 528–539 (2020).

63. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
64. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
65. Speed, D. *et al.* Re-evaluation of SNP heritability in complex human traits. *bioRxiv* (2016).
66. Calderon, D. *et al.* Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).
67. Zhu, H., Shang, L. & Zhou, X. A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types. *Front. Genet.* **11**, (2021).
68. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
69. Hao, X., Zeng, P., Zhang, S. & Zhou, X. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet.* **14**, e1007186 (2018).
70. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
71. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
72. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
73. Cai, M., Chen, L. S., Liu, J. & Yang, C. IGREX for quantifying the impact of

genetically regulated expression on phenotypes. *NAR Genom Bioinform* **2**, lqaa010 (2020).

74. Mancuso, N. *et al.* Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
75. Shang, L., Smith, J. A. & Zhou, X. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. *PLoS Genet.* **16**, e1008734 (2020).
76. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
77. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
78. Lift Genome Annotations. https://genome.ucsc.edu/cgi-bin/hgLiftOver.
79. BED file format. https://useast.ensembl.org/info/website/upload/bed.html.
80. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
81. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
82. Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103-1116.e20 (2020).
83. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

84. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

85. Zitnik, M. *et al.* Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf. Fusion* **50**, 71–91 (2019).

86. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. in *2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2015).

87. Ghamisi, P. *et al.* Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **7**, 6–39 (2019).

88. Dong, X. L. & Srivastava, D. Big data integration. in *2013 IEEE 29th International Conference on Data Engineering (ICDE)* 1245–1248 (IEEE, 2013).

89. Fillinger, S., de la Garza, L., Peltzer, A., Kohlbacher, O. & Nahnsen, S. Challenges of big data integration in the life sciences. *Anal. Bioanal. Chem.* **411**, 6791–6800 (2019).

90. Zhang, L., Yu, G., Xia, D. & Wang, J. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* **324**, 10–19 (2019).

91. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).

92. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).

93. Gusmao, E. G., Dieterich, C., Zenke, M. & Costa, I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity

and histone modifications. *Bioinformatics* **30**, 3143–3151 (2014).

94. *scATAC.Explorer: A curated collection of currently available scATAC-seq datasets accessible through R, and exportable to other languages*. (Github).
95. Streiner, D. L. & Norman, G. R. Correction for Multiple Testing: Is There a Resolution? *Chest* **140**, 16–18 (2011).
96. Abdi, H. Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics* **3**, (2007).
97. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
98. Jones, H. E., Ohlssen, D. I. & Spiegelhalter, D. J. Use of the false discovery rate when comparing multiple health care providers. *J. Clin. Epidemiol.* **61**, 232–240 (2008).
99. Enhancer. https://www.nature.com/scitable/definition/enhancer-163/.
100. Promoter. https://www.nature.com/scitable/definition/promoter-259/.
101. Latchman, D. S. Transcription factors: an overview. *Int. J. Exp. Pathol.* **74**, 417 (1993).
102. Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
103. Di Bella, D. J. *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554–559 (2021).
104. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
105. ScATAC.Explorer. *Bioconductor* https://bioconductor.org/packages/release/data/experiment/html/scATAC.Explorer.html.

106. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).

107. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

108. Wellcome Sanger Institute. HapMap 3 - wellcome Sanger institute. https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html.

109. Abdi, H. The Bonferonni and Šidák Corrections for Multiple Comparisons. Encyclopedia of measurement and statistics (2007).

110. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

111. GitHub - joepickrell/fgwas: Functional genomics and genome-wide association studies. *GitHub* https://github.com/joepickrell/fgwas.

112. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

113. https://github.com/shooshtarilab/EffReg.

114. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

115. Zuo, C., Shin, S. & Keleş, S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, (2015).

116. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2017).

117. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, (2018).

118. Koponen, H. *et al.* Childhood central nervous system infections and risk for

schizophrenia. *Eur. Arch. Psychiatry Clin. Neurosci.* **254**, 9–13 (2004).

119. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).

120. Di Bella, D. J. *et al.* Author Correction: Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **596**, E11 (2021).

121. Li, Y. *et al.* Integration of GWAS Summary Statistics and Gene Expression Reveals Target Cell Types Underlying Kidney Function Traits. *J. Am. Soc. Nephrol.* **31**, 2326–2340 (2020).

122. Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity* **105**, 4–13 (2010).

123. Guo, A.-Y., Sun, J., Jia, P. & Zhao, Z. A Novel microRNA and transcription factor mediated regulatory network in schizophrenia. *BMC Syst. Biol.* **4**, 10 (2010).

124. Legge, S. E. *et al.* Genetic architecture of schizophrenia: a review of major advancements. *Psychol. Med.* **51**, 2168–2177 (2021).

125. Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).

126. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).

127. Hall, L. S. *et al.* A transcriptome-wide association study implicates specific pre- and post-synaptic abnormalities in schizophrenia. *Hum. Mol. Genet.* **29**, 159–167 (2020).

128. Shao, L. & Vawter, M. P. Shared Gene Expression Alterations in Schizophrenia

and Bipolar Disorder. *Biol. Psychiatry* **64**, 89–97 (2008).

129. Pinacho, R. *et al.* Increased SP4 and SP1 transcription factor expression in the postmortem hippocampus of chronic schizophrenia. *J. Psychiatr. Res.* **58**, 189–196 (2014).

130. Qu, M. *et al.* Associations of ATF4 gene polymorphisms with schizophrenia in male patients. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, (2008).

131. Nie, F. *et al.* Schizophrenia risk candidate EGR3 is a novel transcriptional regulator of RELN and regulates neurite outgrowth via the Reelin signal pathway in vitro. *J. Neurochem.* **157**, (2021).

132. Sabunciyan, S. *et al.* Polymorphisms in the homeobox gene OTX2 may be a risk factor for bipolar disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **144B**, (2007).

133. Liu, L. *et al.* TCF7L2 polymorphisms and the risk of schizophrenia in the Chinese Han population. *Oncotarget* **8**, 28614 (2017).

134. Bergman, O., Westberg, L., Nilsson, L.-G., Adolfsson, R. & Eriksson, E. Preliminary evidence that polymorphisms in dopamine-related transcription factors LMX1A, LMX1B and PITX3 are associated with schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **34**, 1094–1097 (2010).

135. Shimamoto-Mitsuyama, C. *et al.* Lipid Pathology of the Corpus Callosum in Schizophrenia and the Potential Role of Abnormal Gene Regulatory Networks with Reduced Microglial Marker Expression. *Cereb. Cortex* **31**, 448–462 (2021).

136. Wu, Q. *et al.* DISC1 Regulates the Proliferation and Migration of Mouse Neural Stem/Progenitor Cells through Pax5, Sox2, Dll1 and Neurog2. *Front. Cell. Neurosci.* **11**, 261 (2017).

137. Badowska, D. M. *et al.* Modulation of cognition and neuronal plasticity in gain-

and loss-of-function mouse models of the schizophrenia risk gene Tcf4. *Transl. Psychiatry* **10**, 343 (2020).

138. Stachowiak, E. K. *et al.* Cerebral organoids reveal early cortical maldevelopment in schizophrenia—computational anatomy and genomics, role of FGFR1. *Translational Psychiatry* vol. 7 (2017).

139. Steen, V. M. *et al.* Genetic evidence for a role of the SREBP transcription system and lipid biosynthesis in schizophrenia and antipsychotic treatment. *Eur. Neuropsychopharmacol.* **27**, 589–598 (2017).

140. Ding, C. *et al.* Transcription factor POU3F2 regulates TRIM8 expression contributing to cellular functions implicated in schizophrenia. *Mol. Psychiatry* **26**, 3444–3460 (2021).

141. Li, Z., Cogswell, M., Hixson, K., Brooks-Kayal, A. R. & Russek, S. J. Nuclear Respiratory Factor 1 (NRF-1) Controls the Activity Dependent Transcription of the GABA-A Receptor Beta 1 Subunit Gene in Neurons. *Front. Mol. Neurosci.* **11**, 285 (2018).

142. Qing, L. *et al.* Sex-dependent association of mineralocorticoid receptor gene (NR3C2) DNA methylation and schizophrenia. *Psychiatry Res.* **292**, 113318 (2020).

143. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, (2010).

Curriculum Vitae

Name: Kayvan Shabani

Post-secondary Education and Degrees:
Sharif University of Technology
Tehran, Iran
2016-2021 B.Sc.

Honours and Awards
Winner of Western University's Graduate Research Scholarship
2021-2022

Ranked 119 at the National University Entrance Exam among 162000 participants. (Approximately top 0.07%)
2016

Related Work Experience
Teaching Assistant
The University of Western Ontario
2021-2022

Publications:

Das, Akash Chandra, Aidin Foroutan, Brian Qian, Nader Hosseini Naghavi, **Kayvan Shabani**, and Parisa Shooshtari. 2022. "Single-Cell Chromatin Accessibility Data Combined with GWAS Improves Detection of Relevant Cell Types in 59 Complex Phenotypes" *International Journal of Molecular Sciences* 23, no. 19: 11456. https://doi.org/10.3390/ijms231911456