UPPSALA
UNIVERSITET

UPTEC X20 015
Examensarbete 30 hp
Juni 2020

Transparent Machine Learning for Multi-Omics Analysis of Mental Disorders

Stella Belin

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Transparent Machine Learning for Multi-Omics Analysis of Mental Disorders

Stella Belin

Schizophrenia and bipolar disorder are two severe mental disorders that affect more than 65 million individuals worldwide. The aim of this project was to find co-prediction mechanisms for genes associated with schizophrenia and bipolar disorder using a multi-omics data set and a transparent machine learning approach. The overall purpose of the project was to further understand the biological mechanisms of these complex disorders. In this work, publicly available multi-omics data collected from post-mortem brain tissue were used. The omics types included were gene expression, DNA methylation, and SNP array data. The data consisted of samples from individuals with schizophrenia, bipolar disorder, and healthy controls. Individuals with schizophrenia or bipolar disorder were considered as a combined CASE class.

Using machine learning techniques, a multi-omics pipeline was developed to integrate these data in a manner such that all types were adequately represented. A feature selection was performed on methylation and SNP data, where the most important sites were estimated and mapped to their corresponding genes. Next, those genes were intersected with the gene expression data, and another feature selection was performed on the gene expression data. The most important genes were used to develop an interpretable rule-based model with an accuracy of 88%. The model was then visualized as a network. The graph highlighted genes that may be of biological importance, including CACNG8, RTN4, TERT, OSBPL8, and ANTXR1. Moreover, strong co-predictions were found, most notable between CNKSR4 and KDM4C in CASE samples. However, further investigations would need to be performed in order to prove that these are real biological interactions.

Through the methods developed and the results found in this project, we hope to shed new light towards analyzing multi-omics data as well as to reveal more about the underlying mechanisms of psychiatric disorders.

Handledare: Mateusz Garbulowski
Ämnesgranskare: Carl Nettelblad
Examinator: Pascal Milesi
ISSN: 1401-2138, UPTEC X20 015

Sammanfattning

Schizofreni och bipolär sjukdom är två svåra psykiska sjukdomar som sammanlagt drabbar mer än 65 miljoner människor världen över. Båda sjukdomarna kan ha drastiska konsekvenser på vardagslivet, även om exakta symptom kan skilja sig mellan individer. Schizofreni karaktäriseras av en förvrängning av tankar, uppfattningar och känslor. Symptom för sjukdomen inkluderar hallucinationer, villfarelser, abnormalt beteende, oorganiserat tal och känslostörningar. Bipolär sjukdom kännetecknas av omväxlande perioder av depression och förhöjningar av sinnesstämning (också kallat mani). Under dessa kan känslor, sömnmönster och aptit ändras, och under allvarliga episoder kan individer även få hallucinationer och villfarelser.

Trots att sjukdomarna har varit kända i nästan ett decennium vet man inte exakt vad som orsakar dem. Det har visats att både schizofreni och bipolär sjukdom är till stor del ärftliga, och många studier har gjorts för att hitta vilka gener som orsakar sjukdomstillstånden. För både schizofreni och bipolär sjukdom kan psykos vara ett symtom, det vill säga hallucinationer och villfarelser. Det finns flera studier som undersöker dessa sjukdomar tillsammans på grund av att psykos är ett överlappande symptom. Genom att förstå de underliggande biologiska mekanismer inom olika sjukdomar kan diagnosticering bli bättre och nya behandlingsmetoder kan utvecklas.

Många mentala sjukdomar är biologiskt komplexa och i nuläget dåligt förstådda. Tack vare framsteg inom bioteknik kan vi samla större mängder biologiska data vilka kan hjälpa oss förstå underliggande mekanismer av olika sjukdomar. Eftersom schizofreni och bipolär sjukdom kan ärvas i familjer, men även kan påverkas av miljö, är det relevant att undersöka arvsmassan och relaterade typer av data hos individer med dessa tillstånd, och det var detta som gjordes i detta projekt.

De typer av data som analyserades i detta projekt var genuttryck, metylering, och SNPs (single nucleotide polymorphisms). Genom att mäta genuttryck kan man uppskatta hur starkt en viss gen uttrycks i cellerna. Inom bioteknik och liknande fält pratar man ofta om under- och överuttryckta gener, det vill säga att en gen uttrycks mer eller mindre i en individ med ett tillstånd man mäter jämfört med någon som inte har det. Metylering är en kemisk process som påverkar om eller hur mycket en viss gen uttrycks. Huruvida en viss gen är metylerad eller inte ändras genom livet och kan påverkas av miljömässiga faktorer. Ofta är man intresserad av metylering för att förstå vilka gener som är aktiva vid ett visst tillfälle. En SNP är en position i arvsmassan som kan variera inom en population. Många forskar på SNPs för att hitta länkar

mellan en viss typ av variation och en sjukdom. Genom att analysera flera datatyper kan man undersöka mekanismer mellan de olika typerna.

I data för detta projekt fanns det mer än en miljon insamlade datapunkter för varje individ. Att analysera den mängden data manuellt är inte möjligt, så därför kan man använda maskininlärning för att få fram den viktigaste informationen. Till exempel om man har en grupp med en viss sjukdom och en utan kommer vissa gener uttryckas ungefär lika mycket för att de inte är kopplade till sjukdomen medan andra kommer uttryckas olika. Med andra ord, man vill hitta vilka aspekter av den biologiska data som är beroende av sjukdomstillståndet och vilka aspekter som inte är det. Ofta är en stor del av data irrelevant, så för att förenkla för senare analys börjar man i många fall med att minska mängden datapunkter. Många maskininlärningsalgoritmer kan kräva stora datorresurser, exempelvis minne eller processorkraft, så genom att minska antal datapunkter kan senare analyssteg bli mer effektiva. Detta kallas för en *feature selection*, det vill säga ett urval av särdag (attribut) hos individerna. Ett *feature* (attribut) kan vara exempelvis en viss gen eller SNP.

Ofta vill man bygga en modell som kan, utifrån nya data, förutse om den okända datan är från en person med sjukdomstillståndet man undersöker eller inte. I den maskininlärningsmetod som användes i detta projekt består denna modell av regler som konceptuellt liknar ”Om gen 1 är överuttryckt och gen 2 är underuttryckt är individen sjuk.” Genom dessa regler kan man se vilka gener (om det är genuttryck man undersöker) som är viktiga i en sjukdom.

I detta projekt var första steget att förbehandla alla tre datatyper. Data i detta projekt kom från 55 individer med schizofreni eller bipolär sjukdom och 27 individer utan någon av tillståndet. För alla datatyper behandlades schizofreni och bipolär sjukdom som en grupp så att grupperna var *sjuk* och *frisk*. Sedan gjordes en *feature selection* på metylering- och SNP-data. De datapunkter som var viktigast användes för att välja ut gener och deras genuttryck. Från dessa byggdes en modell med tusentals regler om vilka gener som uttrycks mer eller mindre i patienter med schizofreni eller bipolär sjukdom. De viktigaste generna undersöktes i olika biologiska databaser och vetenskapliga artiklar. Flera av generna som hittades hade en tidigare koppling till sjukdomstillstånden, och några hade inte en direkt koppling men från ett biologiskt perspektiv verkade de lovande. Sammanfattningsvis hittades flera gener som kan vara intressanta för att förstå schizofreni och bipolär sjukdom bättre.

List of Content

Abbreviations

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
BPD	Borderline personality disorder
CNV	Copy number variations
CPM	Counts per million
CRE	cAMP response element
CSV	Comma-separated values
DSM	Diagnostic and Statistical Manual of Mental Disorders
DALY	Disability-adjusted live year
ER	Endoplasmic reticulum
FDR	False discovery rate
GO	Gene Ontology
GEO	Gene Expression Omnibus
GHR	Genetics Home Reference
GWAS	Genome-wide association studies
ID	Interdependency discovery
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOOCV	Leave-one-out cross validation
MCFS	Monte Carlo feature selection
meQTL	Methylation quantitative trait loci
MI	Mutual information
NGS	Next generation sequencing
NHS	National Health Service
NIMH	National Institute of Mental Health
PCA	Principal component analysis
PsyGeNET	Psychiatric disorders Gene association NETwork
RI	Relative importance
SNP	Single nucleotide polymorphism
SUD	Substance use disorder
TMM	Trimmed mean of M-values
TH	Tyrosine hydroxylase

1 Introduction

Schizophrenia and bipolar disorder are mental disorders that have been known for close to a century (Jablensky 2010; Mason *et al*. 2016), yet the underlying biology is not fully understood. It is estimated that schizophrenia and bipolar disorder affect 20 and 45 million individuals worldwide, respectively (WHO 2019). Both disorders have been shown to have hereditary components, and may severely affect the quality of life of individuals with the disorders. For instance, individuals with psychosis (which is a symptom for schizophrenia and bipolar disorder) run a higher risk of being exposed to human right violations by ”long-term confinement in institutions” (WHO 2019). Due to the technological advancement in genomic and other types of omics research as well as in machine learning, we have the tools to further understand the complex disorders. By understanding the biological mechanisms of the disorders, more accurate diagnosis and more effective treatment may be developed.

In a recent study by Pai *et al*. (2019), the authors examined multiple types of genetic data (so called multi-omics data) from brain tissue of individuals diagnosed with schizophrenia or bipolar disorder, and healthy individuals as control. This master project aims to apply a machine learning approach to the same multi-omics data set and to examine the gene-gene interdependencies of the different data types. This is with the purpose to find co-prediction mechanisms, and thus broaden the understanding of the underlying biology of schizophrenia and bipolar disorder.

1.1 Project Aim

The general aim for this project was to find co-prediction mechanisms for genes associated with schizophrenia and bipolar disorder using a multi-omics approach. To achieve this, a pipeline was created which allows for multi-omics analysis, and subsequently transparent machine learning was applied to discover interdependencies between the different omics layers, meaning the different omics types (e.g. transcriptomics and genomics). This approach was developed to achieve a deeper understanding of the underlying causes of the conditions as well as predict their occurrence. Given the severity in terms of life expectancy and stigma, examining the genetic causes might improve both diagnosis and treatment of the disorders.

2 Background

To understand both the context of the project and its implementation, three main areas need to be covered: the background of the mental disorders of interest, various machine learning techniques, and the approaches and challenges of multi-omics data analysis.

2.1 The Mental Disorders and Current Knowledge

Although schizophrenia and bipolar disorder have been shown to have heritable components, the disorders are complex and the genetic components have not yet been fully mapped. With the technological advances in biotechnology, analyses on a larger scale have made it possible to explore these disorders more in-depth (Geschwind & Flint 2015).

2.1.1 Psychosis in Schizophrenia and Bipolar Disorder

Schizophrenia is a severe mental disorder which distorts the thought pattern, perception, and feelings of the affected individual. Symptoms include hallucinations, delusions, abnormal behavior, disorganized speech, and disturbances of emotions (WHO 2019). The symptoms typically start at late adolescence to early adulthood (NIMH 2020). In a meta-study by McGrath *et al*. (2008) they estimated the male to female ratio to be 1.4:1. The disorder, in acute state, has a severe impact on quality of life. When assessing the burden of different diseases, the severity of the disease is reflected by a weight factor called disability weight on a scale from 0 to 1 (WHO *n.d.*). A study by Salomon *et al*. (2012) examined this factor for 220 diseases (including types of mental disorders, cancers, infectious diseases etc.), and found that acute schizophrenia had the highest disability weight with a value of 0.756. Another study (Laursen *et al.* 2014) estimated the early mortality of individuals with schizophrenia to be between two and three times higher than the general population.

Bipolar disorder is characterized by alternating periods of depression and elevated moods. The elevated mood is also referred to as mania or hypomania, where hypomanic periods are less severe (NIMH 2020). During these periods, also called ”mood episodes”, the emotions are unusually intense, sleep patterns and appetite may change, and in severe episodes hallucinations and delusions may appear (NIMH 2020). The episodes may last for several days or weeks, and the manifestation of symptoms varies between patients. The onset of bipolar disorder is typically during late adolescence or early adulthood (NIMH 2020).

Psychosis is an important overlap of symptoms between schizophrenia and bipolar disorder. It is a state where a person loses ”some contact with reality” (NHS 2019). The main symptoms of psychosis are hallucinations and delusions. Hallucinations are when a person hears or sees

something that does not exist. A common example of this is hearing voices. A delusion is when someone has strong beliefs which others do not have, such as ”believing there’s a conspiracy to harm them” (NHS 2019). Psychosis can be caused by both mental disorders, such as schizophrenia and bipolar disorder, and environmental factors, e.g. trauma, stress, drugs or alcohol, or brain tumors (NHS 2019).

2.1.2 Current Genomic Research

Since bipolar disorder can cause psychotic symptoms, patients can be misdiagnosed with schizophrenia (NIMH 2020). Furthermore, schizoaffective disorder is a separate disorder which has symptoms that overlaps with both schizophrenia and bipolar disorder (NLM 2020). Multiple studies have found that schizophrenia and bipolar disorder are related on a genetic level. Lichtenstein *et al*. (2009) linked the Swedish Multi-Generation Register (a register of individuals with respect to their parents) to the Hospital Discharge Register (which includes information on inpatient hospitalizations for psychiatric disorders) for over 2 million Swedish families. This study found that both schizophrenia and bipolar disorder have heritable components (64% and 59% respectively) and that an individual with schizophrenic relatives has an increased risk for bipolar disorder and vice versa.

Another study (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013) estimated the genetic variation within and between mental disorders using genome-wide single nucleotide polymorphism (SNP) data, and found that schizophrenia and bipolar disorder were genetically correlated. In fact, the correlation between schizophrenia and bipolar disorder was the strongest among the different disorders examined. In a systematic review by Lee *et al*. (2012), they assessed published copy number variation (CNV) studies and genome-wide association studies (GWAS). For schizophrenia, the genes *ZNF804A*, *MHC*, *NRGN*, and *TCF4* were associated. *ANK3*, *CACNA1C*, *DGKH*, *PBRM1*, and *NCAN* were significant for bipolar disorder. The study found *ZNF804A*, *CACNA1C*, *NRGN*, and *PBRM1* to be relevant genes in common for bipolar disorder and schizophrenia.

The data collected by Pai *et al*. (2019) served as the basis for this project. The data set consisted of DNA methylation, SNP, and gene expression data. A cis-methylation quantitative trait loci (meQTL) was performed, where the authors found that hypomethylation of the gene *IGF2* in the enhancer region was significant in individuals diagnosed with schizophrenia or bipolar disorder. Additionally, the authors performed targeted bisulfide sequencing around the *IGF2* locus, although these results were not included as data for this project. When the *IGF2* enhancer was knocked out in mice, the authors observed an up-regulation of tyrosine hydroxylase (*TH*), which is rate-limiting in dopamine synthesis (GHR 2020). Dopamine is a neurotransmitter which, if impaired, has been linked to both schizophrenia (Brisch *et al*.

2014) and bipolar disorder (Ashok *et al*. 2017). Antipsychotic drugs often work by blocking dopamine and/or serotonin levels (CAMH *n.d.*). An increase in *IGF2* expression was also noted, and ”transcriptomic and proteomic alterations affecting synaptic activity and structure.”

2.2 Machine Learning for Discovery and Interpretability

Machine learning is a powerful technique for biological discovery, and two different machine learning methods were used in this project. Before specifying these techniques, it is necessary to define basic terminology that will be used throughout this report. A feature, which is synonymous with attribute, is a characteristic of an object (Google Developers *n.d.*). In the context of this project it refers to the SNPs, methylation sites, and genes. An object is the individual (sample) from which the data have been collected from. A cohort is the collection of objects which share a characteristic (Cambridge Dictionary *n.d.*), which in this project is psychosis. A decision class is the pre-defined ”outcome”, i.e. case and control. A decision table is a system compromised of objects, features, and decision (Pawlak 1984).

2.2.1 Importance of Feature Selection

Next generation sequencing (NGS) has made large scale genomic analyses possible to perform by being faster and cheaper, however, the big NGS data pose new challenges in bioinformatics. Some of the NGS data can be irrelevant, redundant, or noisy and thus affect the quality of analysis. The dimensionality of such data can be very high which leads to a heavy computational load. A common step in machine learning, prior to developing the model or classifier itself, is feature selection. Feature selection is the process of selecting relevant features for model building (Li *et al*. 2017), and the purpose of this step in machine learning is to reduce dimensionality and improve quality and interpretability of classification. There are multiple types of feature selection. The simplest type of feature selection is filtering, which compares the feature to the decision class by calculating a statistic, e.g. Pearson’s correlation. By ranking according to this statistic, the top features may be selected (Cai *et al*. 2018). More complex feature selection techniques, such as decision trees, can take co-dependencies between features into consideration. In this project, both types were used: mutual information (MI), a filtering feature selection, and Monte Carlo feature selection (MCFS), a technique which considers co-dependencies between features. The following two subsections will cover these in more detail.

2.2.2 Monte Carlo Feature Selection

One of the techniques used in this project was MCFS (Draminski *et al.* 2007). The algorithm creates numerous decision trees from randomly selected subsets of the data (see Figure 1), where the most important features are kept when a feature ”(…) is likely to take part in the process of classifying samples into classes ‘more often than not’” (Draminski *et al.* 2007).

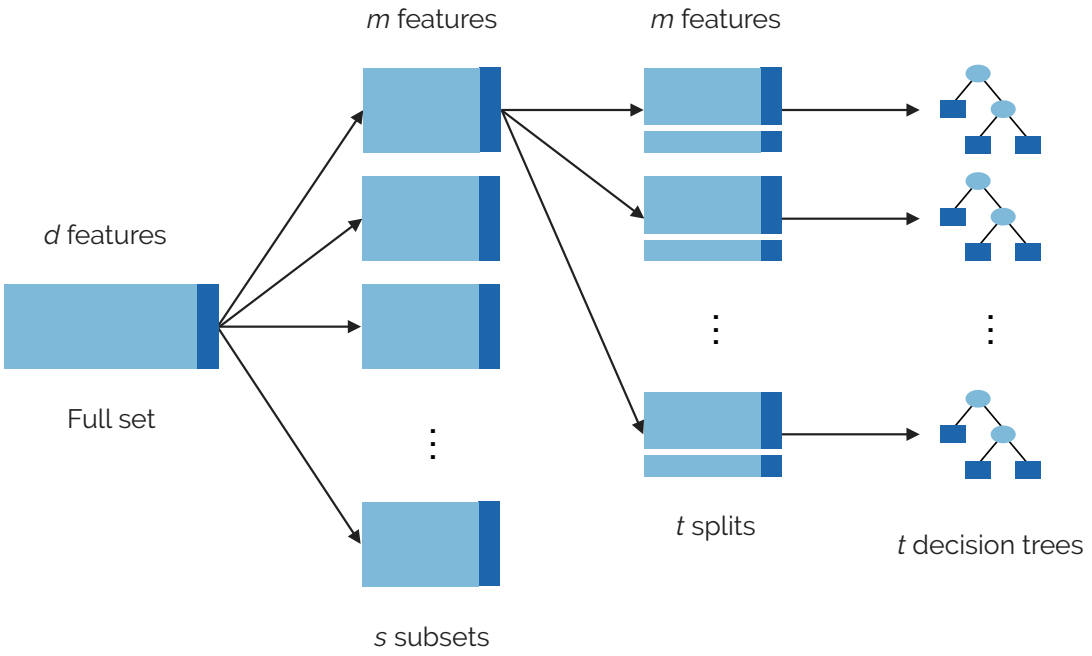

Figure 1. Overview of MCFS, based on figure from Draminski *et al.* (2007). Light blue indicates features and darker blue decision classes.

Given the full data set with d features, s subsets are created with m features in each (each feature should appear in multiple subsets) where $m << d$. Each of these subsets are randomly divided into training and test sets (with 67% and 33% of the samples respectively) t times per subset, and every training set is used to create a decision tree. The quality of each tree is assessed in the form of a weighted accuracy. After the trees have been created, the importance of each feature is estimated in terms of relative importance (RI). This measurement is dependent on how often that feature is responsible for a split in a tree and the information gain. For a feature g_k, the RI is defined as:

$$RI_{g_k} = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_{g_k}(\tau)} IG(n_{g_k}(\tau)) \left(\frac{\text{no. in } n_{g_k}(\tau)}{\text{no. in } \tau} \right)^v \quad (1)$$

Where τ is a tree, $wAcc$ is the weighted accuracy of a tree, n_{gk} are the nodes in a tree, and IG is the information gain. u and v are fixed positive reals. The number of subsets is denoted as s and the number of splits as t. The RI is tested for statistical significance.

An interdependency discovery (ID) graph of the *N* most important features can be plotted (see Figure 2 for example). The color of the node indicates strength of contribution (more saturated nodes mean higher strength), the size of node indicates how frequent the feature appears in a pair with another node (bigger nodes mean higher frequency), and the thickness of the edges represents how frequent that pair appears together (thicker edges mean higher frequency).

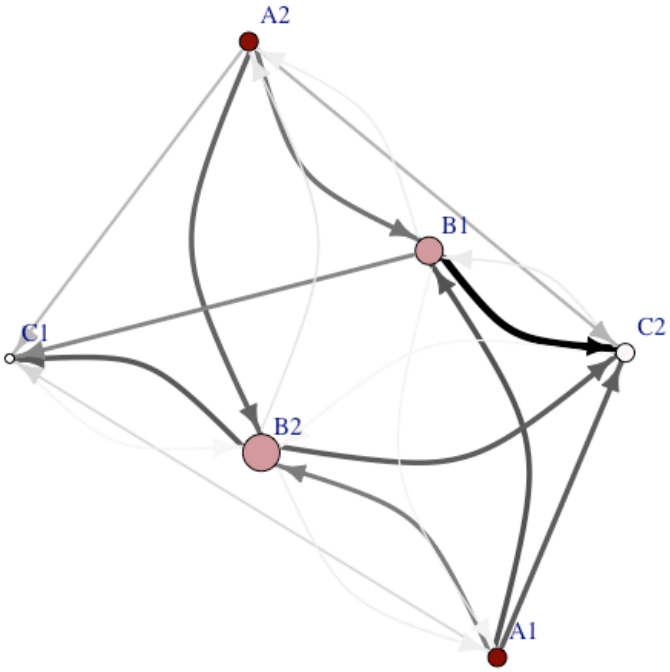

Figure 2. ID-graph constructed from artificial data, generated with *rmcfs* package (Dramiński & Koronacki 2018).

This method is computer intensive since it creates thousands of trees for a given data set. More specifically, it creates $s \times t$ trees, where both s and t need to big enough so that each feature can appear in multiple subsets. The advantage with this method is that it allows for investigation of interdependencies between the features (Draminski *et al*. 2010). MCFS has been successfully applied in the context of rule-based learning, for example on avian influenza virus (Khaliq *et al.* 2015) to find pathogenicity markers. Furthermore, a study by Chen *et al*. (2018) used MCFS to detect gene expression signatures for different types of adult neural stem cells.

2.2.3 Feature Selection Using Mutual Information

Due to the fact that the data for this project are large and MCFS is computer intensive, a pre-filtering feature selection method was used. One common approach is to use statistical tests

based on information theory. Information theory is ”a mathematical representation of the conditions and parameters affecting the transmission and processing of information” and is often used in communication engineering (Markowsky 2017). Two measurements from information theory were tested and used in this project for feature selection: Shannon’s entropy, or simply entropy, and MI. However, in the final pipeline only MI was included. Entropy is a measurement of ”the average missing information in a random source” (Lesne 2011), and is defined as:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x) \geq 0 \tag{2}$$

Where X is a random variable and x is each outcome. In other words, entropy measures how uneven the probability distribution is, or the information gained. If the entropy ($H(X)$) is equal to zero, this means that the value for a feature is constant across the cohort. If $H(X)$ is high it indicates that the distribution is uniform. This measurement does not take decision classes into consideration. The joint entropy for two discrete variables X and Y is defined as:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) log_2 p(x, y) \tag{3}$$

Where x is defined as previously and y is each outcome from Y. From equation (2) and (3), the MI can be calculated between X and Y. In this project, X was an attribute and Y was the decision class (i.e. how much information do we have about the decision class given the attribute). MI is a measurement of amount of information of one variable given the other, and how two variables are dependent. MI is defined as:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{4}$$

If the information (I) is equal to zero it means that X and Y are independent. MI have been used to extract meaningful information in several fields, including biomedicine (Fang *et al*. 2015) and genomics (Song *et al*. 2012).

2.2.4 Rough Sets Machine Learning

To find co-prediction mechanisms for features, a classification model based on rough-sets was developed. The basic assumption of rough set theory is that we can associate a given type of information to every object in the universe of discourse (Pawlak & Skowron 2007). A rough set-based approach allows for creating transparent classification models. These are represented by minimal subsets of features (reducts) from an information system A, which is a decision table excluding the decision column, such that it still preserves the classification

power of the system (Pawlak & Skowron 2007). Subsequently, reducts are transformed into legible *IF-THEN* rules, e.g:

IF GENE_1=up-regulated AND GENE_2=down-regulated THEN CASE

The quality of these rules is measured in terms of support, coverage, accuracy, and other statistical measurements such as p-value. Several frameworks have been developed that utilizes rough set theory for classification, such as *RoughSets* (Riza *et al.* 2016) and *RWeka* (Hornik *et al*. 2007). In this project, *ROSETTA* (Øhrn and Komorowski 1997) and its R package wrapper *R.ROSETTA* (Garbulowski *et al*. 2020) was used. These allow for rule-based model construction and analysis.

The classification models can be visualized using rule-based visualization tools such as *Ciruvis* (Bornelöv *et al*. 2014) or *VisuNet* (Smolinska *et al*. 2020). These visualization tools are graphic representations of feature-feature interdependencies that allow for rule-based model interpretation. In *VisuNet*, the rules are represented as a network where the nodes represent features and the edges are connections between features. The nodes are colored in terms of states (such as over- or under-expressed genes), the size of the nodes indicates the accuracy/support in relation to decision, and the thickness of the borders represent the number of rules the node is part of. Interpretations regarding interdependencies can be made based on the connection of the nodes, where color and thickness represent the strength of connection.

2.3 Multi-Omics Data: Types and Challenges

Thanks to the possibility to perform large scale analyses relatively cheap and fast, it is now easier to incorporate multiple types of omics data. However, as will be further discussed, integrating multi-omics data so that all layers are adequately represented is by no means a straightforward task.

2.3.1 Types of Omics Data

The data of this project (Pai *et al*. 2019) consisted of three types of data: methylation levels, SNP genotypes, and gene expression levels (see Figure 3 for an overview). The omics type for these are genomics (SNP), epigenetics (methylation), and transcriptomics (gene expression). Other examples of omics types are proteomics and metabolomics. A SNP is a nucleotide position that varies within a population, which may serve as biological markers for different diseases and disorders (GHR 2020). SNPs can have an effect on ”promoter activity (gene expression), messenger RNA (mRNA) conformation (stability), and translational efficiency” (Shastry 2009). Often when examining SNP data, the term reference allele refers to the variation that is present in the reference genome, while the other variant is referred to as

alternative allele. They are denoted as A and B, respectively. Since humans are diploid organisms each SNP has two alleles. In the data set, both alleles were included. Given a SNP for one individual, the value for the SNP can be AA (both reference alleles), AB (one reference and one alternative allele), and BB (both alternative alleles).

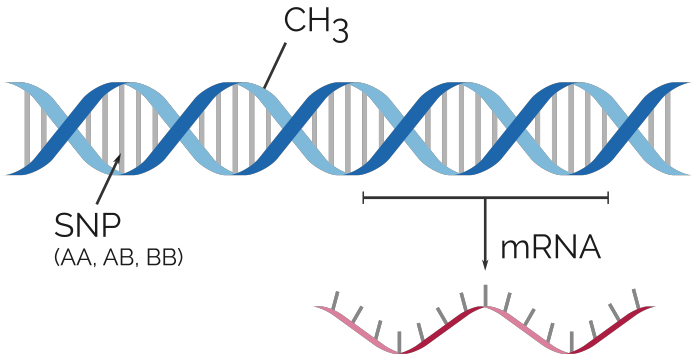

Figure 3. Overview of the different omics data types: SNP as genomic, methylation as epigenetic (denoted as CH_3), and gene expression as transcriptomics (denoted as mRNA).

DNA methylation is the addition of a methyl group to the DNA molecule (Moore *et al*. 2013). This modification does not alter the genetic sequence but rather the gene activity (GHR, 2020). Most often, methylation occurs at regions where cytosine (C) precedes a guanine site (G), these sites are called CpG-islands and more than half of the methylation occurs on these sites. About 70% of promoters are located in regions rich in CpG-sites called CpG islands, and these regions are often unmethylated (Moore *et al*. 2013). A region that has been methylated can impair transcriptional activators, thus reducing gene expression or silencing the gene.

2.3.2 Multi-Omics Data: Meaning and Integration

A multi-omics data set is, intuitively, a data set consisting of multiple omics types. An integrative approach to multi-omics data is to analyze multiple omics types simultaneously (Sun & Hu 2016). This can for example be done by combining the full data set and use statistical methods for analysis (Jiang *et al*. 2016), or as an exploratory step where parts of the omics set are analysed together and then used to identify overlap to another omics type (Wang *et al*. 2019). The latter was the case in this project, due to inconsistent sample size between the omics types.

According to Sun and Hu (2016), important information can go unnoticed if only one omics layer is included in the analysis, especially ” (…) the complementary effects and interactions between omic layers.” For example, disease risk may change across the life span of an individual, hence genomic variant data would not alone be able to explain this change.

Different omics types ”often have complementary roles to jointly perform a certain biological function” (Sun & Hu 2016). By including different types of omics data, interdependencies between different omics layers can be discovered. Intuitively, SNPs and methylation patterns both are important for regulation of gene expression, and in a study by Wang *et al.* (2013) the authors found that in 49% of the genes tested, SNPs and methylation sites showed a ”cooperative/antagonistic regulation pattern” on the gene.

Part of this project was to design a proper data integration approach, and since the aim of this project was to discover interdependencies between omics data types the approach should not exclusively reflect one type of data. List *et al*. (2014) performed a multi-omics study to classify subtypes of breast cancer using gene expression and methylation data. The authors compared four random forest based classification models: gene expression separately, methylation separately, gene expression and methylation combined, and a subset of gene expression represented in PAM50. The model with both gene expression and methylation data had combined the data sets before feature selection. In the combined model, the remaining features were almost exclusively gene expression features. Dabrowski *et al*. (2018) performed MCFS on the joint data, which consisted of gene expression levels from 19,943 genes and β-values for 396,065 methylation sites. The number of important features were 2 for gene expression and 63 for methylation. Another approach (Wang *et al.* 2019) would suggest a feature selection on each data set separately and then integrate the selected features into one data set. Considering these studies, a different approach for this project was needed such that the signal of one type of data does not get overshadowed by the signal of the other type, while still keeping interdependencies between omics types as a focus. Thus, balancing the number of features from each type before feature selection was the selected approach in this project.

3 Materials and Methods

3.1 Computational Resources

The project was written in R 3.6.2 (full list of R packages can be found in Appendix B). For heavier computations (such as preprocessing the large data sets and running MCFS), the external data server *ulam* was used, otherwise the computations ran locally.

3.2 Data

The initial data set was collected and compiled by Pai *et al.* (2019), and it is publicly available at Gene Expression Omnibus (GEO) with GEO accession GSE112525. The set consists of gene expression data, whole-genome DNA methylation data, and SNP array data. The data were collected from the post-mortem brain tissue of individuals with schizophrenia, bipolar disorder, and controls from 29, 27, and 27 patients respectively (see Table 1 for summary of data). In this project, as well as in the original paper, individuals with schizophrenia and bipolar disorder were considered as a single class. The gene expression and methylation data were obtained using high throughput sequencing. The SNP data were collected through arrays designed for known variants in psychiatric disorders (Illumina Human PsychArray-24). The data also contain information on demographic factors (age and sex), clinical variables (cause of death, medications, etc.), and tissue quality, in a separate table. The data were either loaded from comma-separated values (CSV) format or directly into R as an R data frame using the package *GEOquery* (Davis & Meltzer 2007).

Table 1. Summary of original data. All data is from the same cohort.

	Methylation	**SNP**	**Gene expression**
Patients	82	83	34
Male	61	61	25
Female	21	22	9
Control	27	27	17
Schizophrenia	29	29	7
Bipolar	26	27	10
Features	812,663	588,628	58,219
Platform	Infinium MethylationEPIC	Illumina Human PsychArray-24	Illumina NextSeq 500

As presented in Table 1, the size of the cohorts are different between the data types. All data were collected from the same individuals, where the gene expression levels only were collected from a subset of the individuals. However, the size of the cohorts comparing the SNP and methylation data were inconsistent even though they were meant to be the same. The SNP data had three objects that did not exist in the methylation data. The IDs of these samples were 43, 78, and 90. Sample 43 did exist in the patient data file but since the data would later be combined with the methylation data it was excluded. Sample 78 did not exist in the patient data file and was thus also excluded. Sample 90 did not exist in the patient data file either, however when examining the meta data it was identical to sample 95, and was subsequently excluded for downstream analysis.

3.3 Model Overview and Final Integration Model

For a simplified pipeline of the core parts of the project, see Figure 4. For one omics layer, this would be the standard workflow: preprocess the data, perform a feature selection for quality and computing performance, develop a classification model to discover interdependencies, visualize these dependencies, and finally interpret the resulting networks using scientific literature and different biological databases.

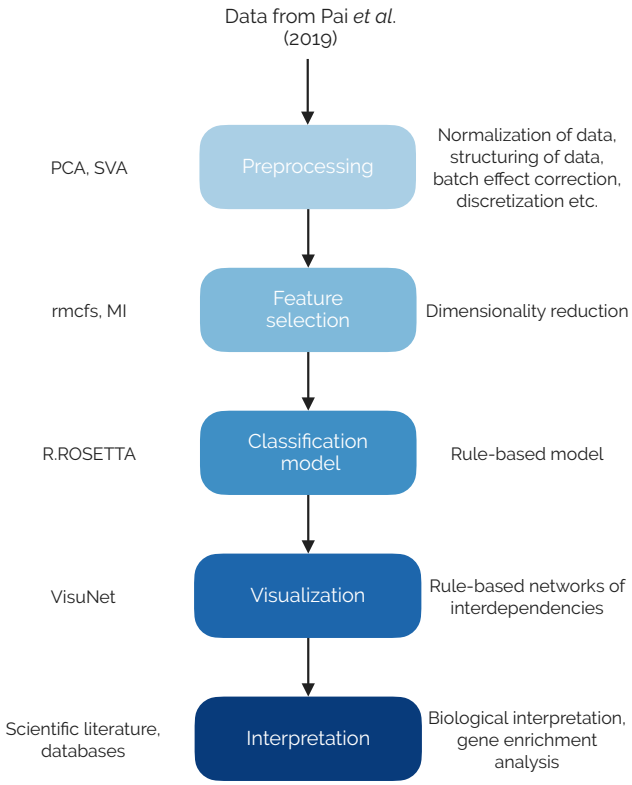

Figure 4. Overview of key steps in the project.

Since the data for this project consists of different data types, the pipeline was modified accordingly, and he developed model for integrating the multi-omics data types can be seen in Figure 5. First, the SNP and methylation data were preprocessed and then filtered to the same size based on MI. The data were merged and MCFS was performed. The gene expression data, which were also preprocessed, were filtered based on the most important genes from MCFS on methylation sites and SNPs. MCFS was performed on the top genes, and those genes were used to construct the final rough-set based classifier.

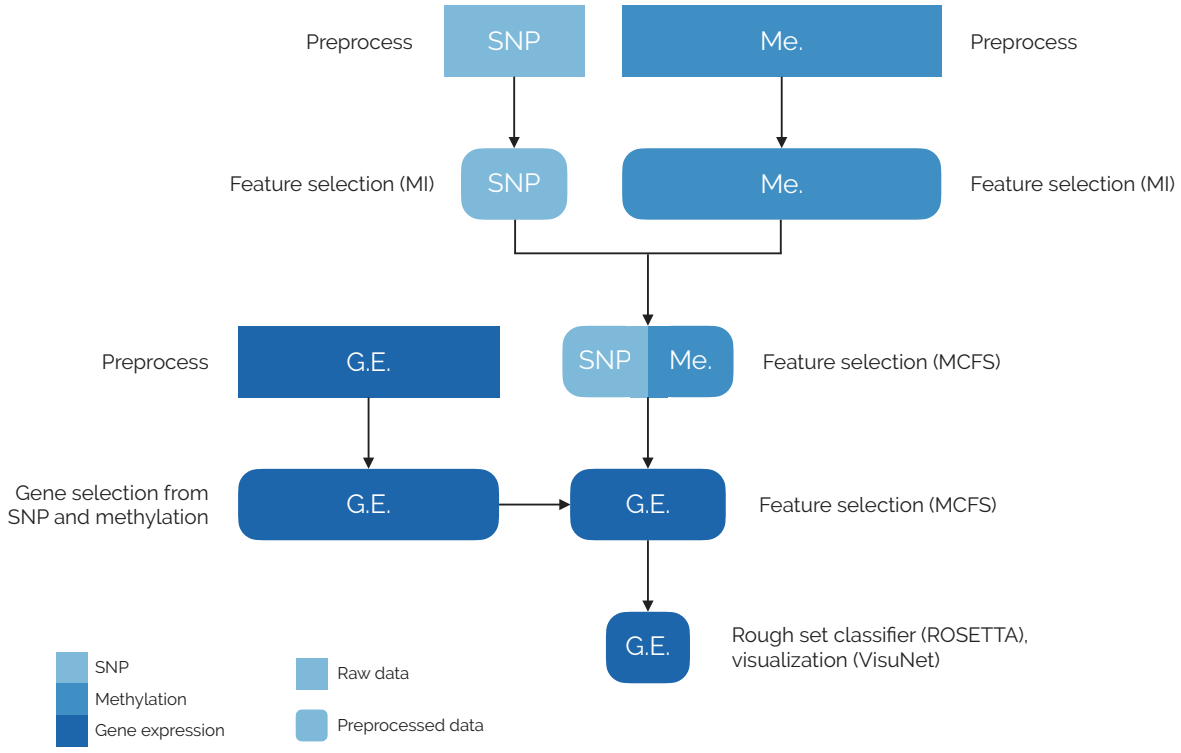

Figure 5. Integration of data types. *G.E.* is gene expression data, *Me.* is Methylation data, *SNP* is SNP array data. Width of boxes indicates the feature reduction (not to scale).

3.4 Preprocessing

As a first step, the methylation and SNP data were preprocessed. The data were investigated with regard to formatting, missing values, sample sizes, and batch effects, and addressed accordingly.

3.4.1 Methylation Data

The methylation data used in this project had been previously processed by Pai *et al*. (2019). This included normalization, as well as exclusion of probes that:

- Overlapped with SNPs with minor allele frequency > 0.05
- Were known to be cross-reactive
- Failed detectability in > 20% of samples

In this work, additional preprocessing steps were still necessary. First the data were restructured so that it would fit as input to the packages further down the pipeline. This included reshaping so that objects represented rows and features columns, modifying the object names to o achieve consistent format among the different data sets, averaging technical duplicates, and attaching class as case (*CASE*) or control (*CTRL*) as last column. After these steps, a principal component analysis (PCA) was performed to check for unwanted batch effect. Since two clusters could be observed based on sex, the probes on the X and Y chromosome were excluded. The chromosome position was extracted using the annotation files from the authors. Another PCA was performed to check for possible remaining batch effect. The remaining probes were discretized using the following intervals: [0.0-0.2) was considered unmethylated (um), [0.2-0.8) was considered indecisive (id), and [0.8-1.0) was considered methylated (m) (Du *et al*. 2010).

3.4.2 SNP Data

The first step was to restructure the data in a similar way as the methylation data, i.e. reshaping the data so that objects were defined as rows and features as columns. The next step was to extract the SNP genotypes. The data consisted of multiple data types, including the raw intensity measurements, but since we were interested in the derived genotypes, only those were extracted. Next, a quality check was performed where probes were excluded if any of the following criteria were met:

- Missing sample frequency > 0.01
- Minor allele frequency < 0.05
- Hardy Weinberg p-value < 10^{-6} (only in decision class *CTRL*)

These thresholds were the same as Pai *et al*. (2019) used. The Hardy Weinberg p-value estimates probability of whether a difference in genotype is due to chance or not, and the values were estimated using the R package *HardyWeinberg* (Graffelman 2015). The number of probes that had been removed was roughly the same number as in the original paper (214,211 in this project, 228,369 in the paper). After that, technical replicates were combined using mode, from the R package *modeest* (Poncet 2019). The decision class was attached, and

the SNPs from the X and Y chromosome were removed in order to have consistency across the data sets.

3.4.3 Gene Expression Data

The preprocessing of the gene expression data was performed by Mateusz Garbulowski (supervisor of this project), but the steps of this preprocessing were included in the report for clarity. First, the duplicated genes were averaged. Next, genes from the X and Y chromosome were removed using *biomaRt* (Durinck *et al*. 2005), for consistency across data sets. The remaining genes were normalized with trimmed mean of M-values (TMM) using *edgeR* (Robinson *et al*. 2010). The data were transformed to Counts Per Million (CPM) and transformed to logarithmic scale. To check and correct for batch effect, the package *sva* (Leek *et al*. 2012) was used. The gene expression levels for each gene were scaled over 0, meaning each value is subtracted by the mean of the attribute (i.e. moving the midpoints of values). Finally, the decision class was added as the last column.

3.5 Feature Selection

As could be seen in Figure 5, multiple rounds of feature selection were performed, both with the purpose of improving accuracy and decreasing computational load, but also to balance the data in terms of sites from different types (see *2.3.2 Multi-Omics Data: Meaning and Integration* for motivation). For the methylation and SNP data, this consisted of two main parts: MI exclusion and MCFS. A summary of the resulting number of features after each feature selection step can be seen in Table 2.

Table 2. Summary of remaining features after each feature selection step for methylation and SNP data.

	Methylation	**SNP**	**Combined**
Original features	812,663	588,628	—
Quality check	—	214,211	—
X and Y removal	794,726	211,909	—
MI exclusion	39,683	39,683	79,366
MCFS	3,116	337	3,453

The feature selection for the gene expression data was based on the results of the feature selection of methylation and SNP data, i.e. from the resulting decision table. The most important methylation sites and SNPs were mapped to their corresponding genes, and based on that, gene expression data were extracted. A round of MCFS was performed on this (see Table 3 for a summary).

Table 3. Summary of remaining features after each feature selection step for gene expression data.

	Gene expression
Original features	58,219
After preprocessing	45,058
From methylation/SNP	2,002
After MCFS	57

3.5.1 MI Filtration of Methylation and SNP Data

The entropy and MI values were calculated for all the methylation sites and SNPs using the R package *infotheo* (Meyer 2009). In order to keep as many features as possible (and not lose relevant information) while still having a feasible amount of data in terms of computation, the threshold for number of features before MCFS was based on the 5th percentile. Since the number of features from methylation was larger, this was the threshold and what will be referred to the top 5th percentile, which corresponds to 39,683 features. The top 5th percentile in terms of MI for methylation and SNP separately was selected, then those top features were merged. In later steps, the features appeared homogenous between classes, and to attempt to correct for this, the bottom 5th percentile in terms of entropy was removed before selecting the top 5th percentile. The motivation for this step was to adjust data for low variance discrete features, in a similar manner to how removing low variance continuous variables is commonly done to improve downstream quality of the data. This approach did not, however, improve the results and was not used in the final model.

3.5.2 MCFS on Methylation and SNP Data

The analysis using MCFS was performed with the R package *rmcfs* (Dramiński & Koronacki 2018) and ran on the external server *ulam* (see Table 4 for parameter settings). This was done for both approaches mentioned in the previous section (see *3.5.1 MI Filtration of Methylation and SNP Data*).

Table 4. Parameter settings for *rmcfs*.

Parameter	Value
Number of features (d)	79,367
Number of features per subset/projection size (m)	300
Number of subsets/projections (s)	30,000
Number of splits per subset (t)	5
Number of threads	8

The resulting top features for each approach were assessed. The problem with homogenous features still remained despite the attempt with entropy filtration (see *5.3 Challenges* for further discussion), and thus the data with only MI filtration was chosen and kept for further analysis. 3,453 features were kept for downstream analysis based on the MCFS cutoff threshold using the k-means method.

3.5.3 Extraction of Genes

The most important sites were mapped to their closest genes using the Illumina annotation files for their corresponding platform (downloaded from the official Illumina website). The sites were renamed to the format ”gene_site”, e.g. ”IGF2_cg02613624”. If a site did not have an annotated gene, the gene name was represented as ”unspc”, meaning unspecified. The sites that were not mapped to genes were excluded for further analysis. The remaining sites were used to select genes from the gene expression data. A list of genes was composed by extracting the gene names from the top sites.

The genes from the gene expression data were in Ensembl format, while the genes in the annotation files had the gene name. In order to extract the relevant genes from the gene expression data, the Ensembl names were translated to the gene name using the R package *biomaRt* (Durinck *et al*. 2005). The list from the previous step was used to intersect the features from the gene expression data, such that 2,002 genes remained. The gene expression data for these were used for further analysis.

3.5.4 MCFS on Gene Expression

On the 2,002 genes that were selected from the most important methylation sites and SNPs, another MCFS-based feature selection was performed. This step was also performed using the *rmcfs* package (Dramiński & Koronacki 2018) with default settings. Using the MCFS k-means cutoff, 57 genes were deemed important and kept for further analysis. These genes were renamed such that the name also included whether it was a methylation site, a SNP, or both as well as how many of each that had been responsible for the selection of each gene. For example, a methylation site located in the *IGF2* region would be named ”IGF2_me,” a SNP in that region would be named ”IGF2_snp.” If multiple features from both types were responsible for selection of that gene, it could be named ”IGF2_me_me_snp.”

3.6 Rule-Based Classification Modeling

The gene expression data from the 57 top genes were run through a rough set-classifier using the R package *R.ROSETTA* (Garbulowski *et al*. 2019). Multiple settings were tested, but the settings for the final model was set to Naive Bayes classifier, Johnson reducer, equal frequency discretization for three levels, and false discovery rate (FDR)-based p-value

correction of the rules. Johnson reducer and equal frequency discretization are default settings. Naive Bayes classifier was used since it, for this data, produced the model with the highest quality (Standard Voter was also tested). In the function *rosetta*, Bonferroni p-value correction is the default, however this correction is very strict and can lead to an increased false negative rate. Given the small sample size in this step, FDR was more suitable. Leave-one-out cross validation (LOOCV) was performed due to the small sample size (note that the gene expression data had a sample size of 34). As variability of models varied depending on order of features, the classification was repeated 100 times, where the order of the features time was shuffled in each iteration. Finally, the rule sets were combined and averaged using a built-in function in *R.ROSETTA*. The merged data set was then visualized using *VisuNet* (Smolinska *et al*. 2020). The rules were filtered out according to if accuracy < 0.7 and coverage < 0.3. Only the top 20 nodes were presented in the network for clarity.

3.7 Biological Analysis and Interpretation

A functional enrichment analysis was conducted using the web tool *gProfiler*. This analysis was performed on the 2,002 genes selected from the top methylation sites and SNPs. *gProfiler* uses several databases, most notably Kyoto Encyclopedia of Genes and Genomes (KEGG), which is a collection of databases of genomes and pathways, and Gene Ontology (GO), which is a database of gene functions. The significance threshold was set to 0.05 for FDR-corrected p-values. The database Psychiatric disorders Gene association NETwork (PsyGeNET) was used to detect if any of the top informative genes from MCFS had previously been associated with a psychiatric disorder. PsyGeNET is a database which is based on automatic text mining on scientific literature which is then manually curated. The package *psygenet2r*, which connects to PsyGeNET, was used to find which of the top 57 genes were associated to psychiatric disorders, and the associated genes and disorders were visualized as a network. Additionally, a manual search was performed for the most important genes in the classification model, using scientific literature and databases such as GeneCards.

4 Results

4.1 Testing for Batch Effect in DNA Methylation Data

The PCA revealed two clear clusters from the DNA methylation data. By coloring the data points based on the sex of the individual, it fully overlapped these clusters (see Appendix A for all PCA plots). After removing the sites located on the X and Y chromosome, there were no longer two separate clusters (see Figure 6 for before and after). If a sex bias remains after this adjustment, the effect is small enough to not severely affect the first principal component. To assert that the bias was fully accounted for, further tests are needed.

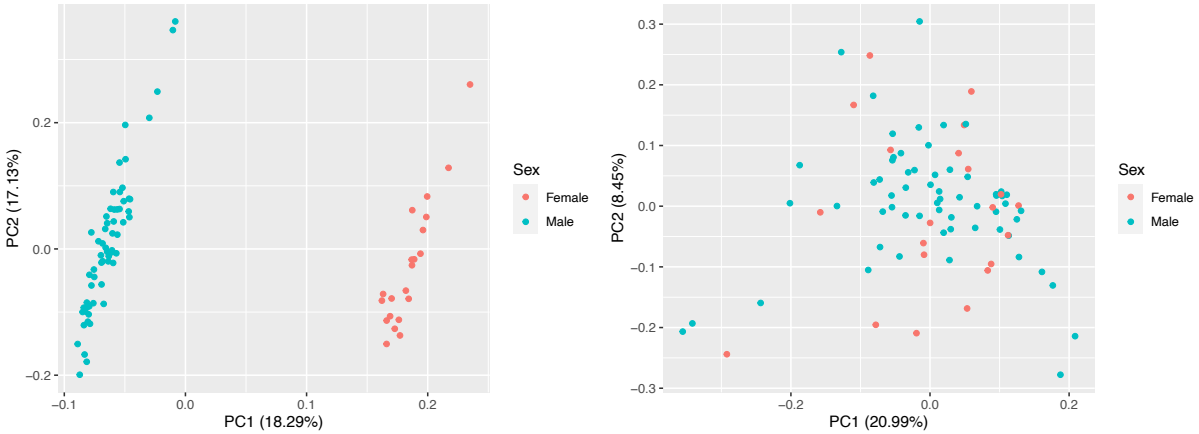

Figure 6. PCA plots of methylation data before (left) and after (right) removing sites from X and Y chromosome.

4.2 MCFS on Methylation and SNP

As mentioned, the approach using entropy filtration did not improve the result and as such the approach with only MI filtration was included in the rest of the pipeline. The MCFS algorithm selected 3,453 important sites with the k-means threshold. Out of these, 3,116 were methylation sites and 337 were SNPs. The ID-graph of the 10 most important features can be seen in Figure 7. Using the permutation threshold, the most important sites were cg01932551, cg15372217, rs6737786, and rs8004384. See Figure 8 for pie charts of distribution of states (methylated, indecisive, and unmethylated) per class. The corresponding results for the discarded approach with entropy filtration, see Appendix C.

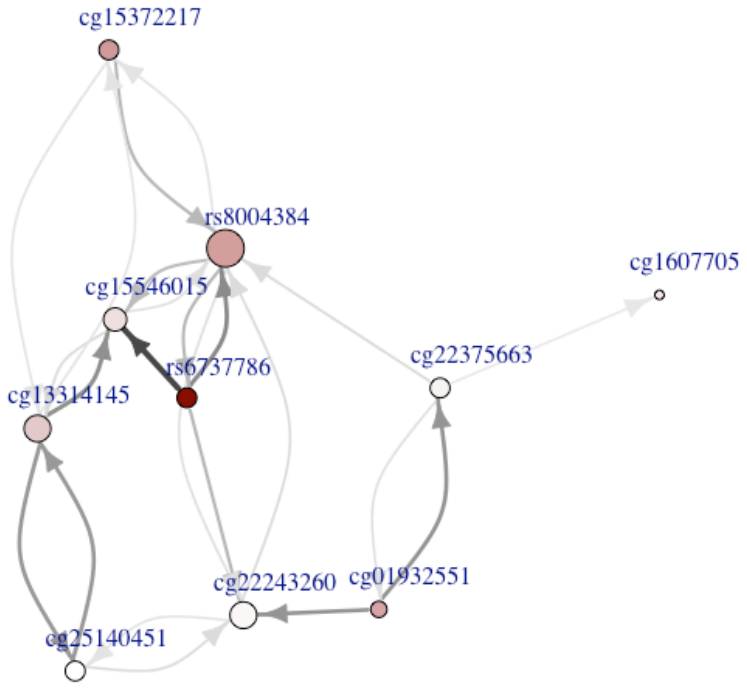

Figure 7. ID graph of top 10 nodes from MCFS.

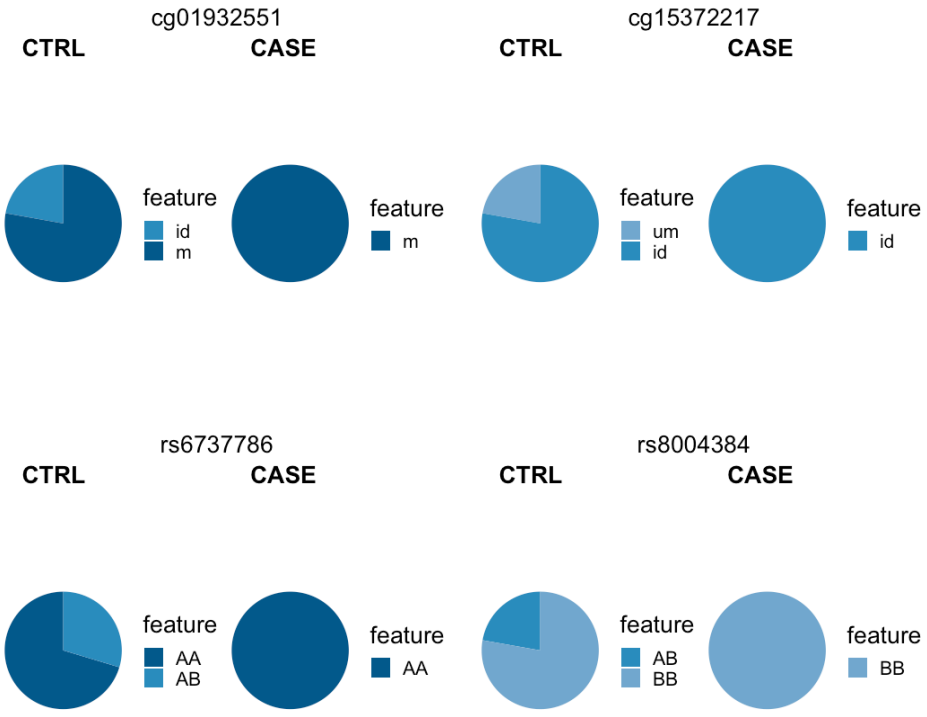

Figure 8. Pie charts of top 4 sites from MCFS. ”m” means methylated, ”id” indecisive, ”um” unmethylated, ”A” is reference allele and ”B” is alternative allele.

To compare whether the pattern was similar for all the data (and not just what was found from MCFS), the most significant sites from Pai *et al*. (2019) were plotted (see Figure 9).

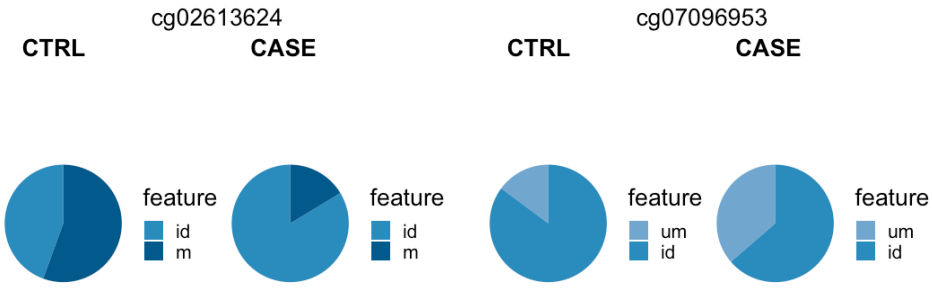

Figure 9. Pie charts from two of significant sites from Pai *et al*. (2019). ”m” means methylated, ”id” indecisive, ”um” unmethylated, ”A” is reference allele and ”B” is alternative allele.

4.3 Annotation of DNA Methylation Sites and SNPs

The annotation of DNA methylation sites or SNPs to their respective genes covered 2,374 out of 3,453 of sites (~68.9%). For annotation of top 10 sites, see Table 5.

Table 5. Gene annotation of top 10 DNA methylation sites and SNPs.

Site name	Gene name
rs6737786	Unspecified
cg15372217	*PCOLCE2*
rs8004384	*ARHGAP5*
cg01932551	*SRCAP*
cg13314145	*NPTX2*
cg15546015	Unspecified
cg16077055	*NCK2*
cg22375663	Unspecified
cg22243260	Unspecified
cg25140451	Unspecified

4.4 Functional Enrichment Analysis

Out of the 3,453 methylation sites and SNPs that were selected with MCFS, 2,291 mapped to a gene. Among these, 2,002 unique genes were present and selected for further downstream analysis. The resulting functional enrichment analysis of those genes can be seen in Table 6.

Table 6. Summary from *gProfiler* of top molecular functions, biological processes, cellular components, KEGG mappings, and reactome for most important genes. p-value was adjusted using FDR correction.

GO: Molecular Function	
Term name	**p_{adj}**
Ion binding	8.226×10^{-7}
Calcium ion binding	2.036×10^{-6}
Enzyme binding	2.036×10^{-6}
Kinase binding	5.834×10^{-6}
Protein kinase binding	8.521×10^{-6}
GO: Biological Process	
Term name	**p_{adj}**
Homophilic cell adhesion via plasma membrane adhesion molecules	1.100×10^{-13}
Cell adhesion	2.623×10^{-13}
Biological adhesion	2.623×10^{-13}
Cell-cell adhesion via plasma-membrane adhesion molecules	5.387×10^{-11}
Anatomical structure morphogenesis	1.320×10^{-9}
Nervous system development	2.343×10^{-9}
GO: Cellular Component	
Term name	
Synapse	7.581×10^{-8}
Cell periphery	8.691×10^{-7}
Plasma membrane	8.691×10^{-7}
Organelle	7.766×10^{-6}
Cell projection	1.151×10^{-5}

KEGG	
Term name	**p_{adj}**
Thyroid hormone signaling pathway	5.576×10^{-3}
Pathways in cancer	7.113×10^{-3}
Glutamatergic synapse	7.113×10^{-3}
Small cell lung cancer	1.075×10^{-2}
Cholinergic synapse	1.122×10^{-2}
Reactome	
Term name	**p_{adj}**
Axon guidance	3.116×10^{-2}

4.5 MCFS on Gene Expression

The MCFS algorithm selected 57 genes as the most important out of the 2,002 mapped genes using a k-means threshold. The ID graph of the top 10 genes can be seen in Figure 10.

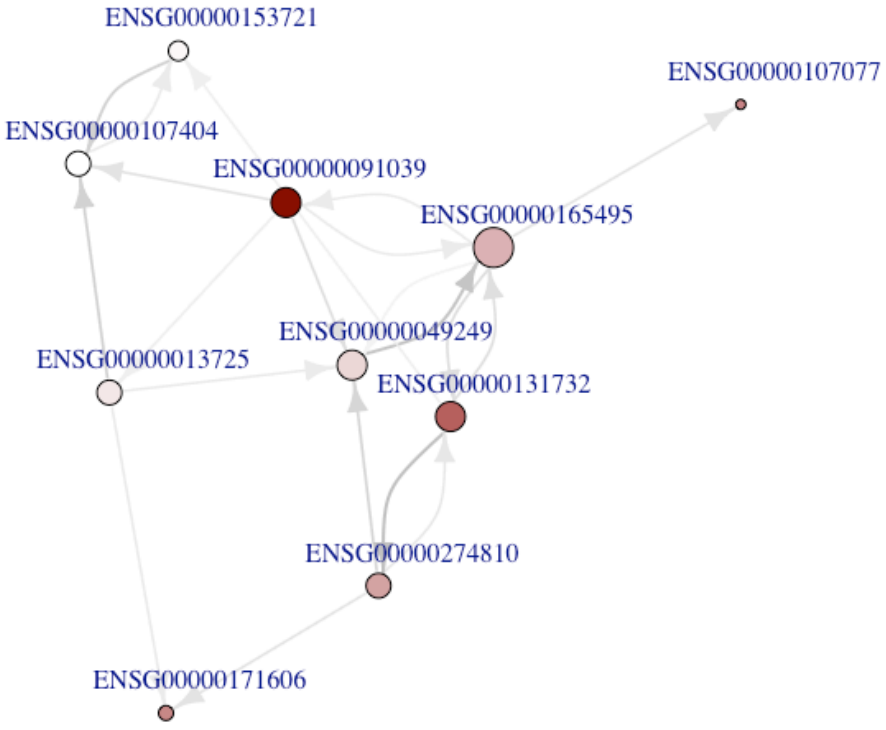

Figure 10. ID graph of top 10 nodes from MCFS.

Since the gene expression data had Ensembl IDs for features, they were translated to their corresponding gene name for consistency (see Table 7 for the translation of top MCFS genes).

Table 7. Translation from Ensembl ID to gene name for top 10 genes.

Ensembl ID	Gene name
ENSG00000091039	*OSBPL8*
ENSG00000131732	*ZCCHC9*
ENSG00000107077	*KDM4C*
ENSG00000171606	*ZNF274*
ENSG00000274810	*NPHP3-ACAD11*
ENSG00000165495	*PKNOX2*
ENSG00000049249	*TNFRSF9*
ENSG00000013725	*CD6*
ENSG00000153721	*CNKSR3*
ENSG00000107404	*DVL1*

4.6 Connection of Top Genes to Psychiatric Disorders

Out of the 57 genes selected, 12 were included in the PsyGeNET database. These were mapped as a graph with their associated disease (see Figure 11).

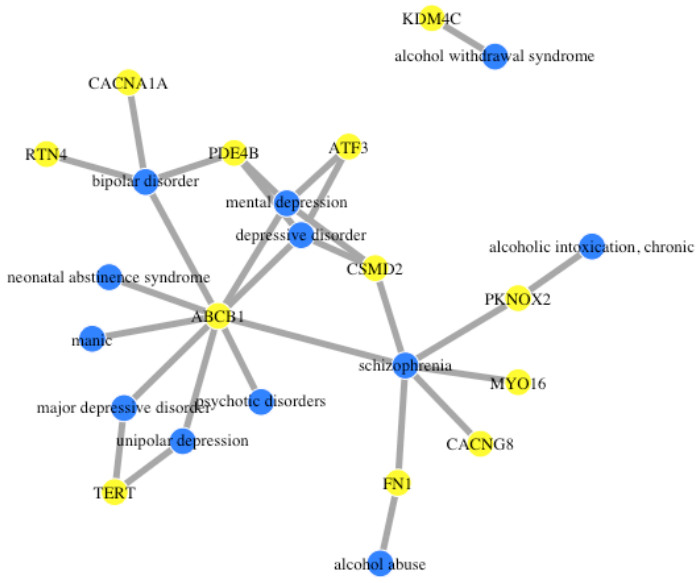

Figure 11. Graph of genes and their associated mental disorder.

4.7 Classification Models

The data set with entropy filtration provided rules that were short or had low coverage and was thus not used as a final result. The resulting classification model for the data set which was kept (without entropy filtration) had a mean accuracy of 88.2%. The most important rules in terms of p-value can be found in Supplementary Materials.

The resulting network from top 20 nodes (which corresponds to 19 genes) can be seen in Figure 12. The top 20 nodes are the nodes most apparent in the model. Stars indicate that the gene was found in PsyGeNET, i.e. was associated with a mental disorder.

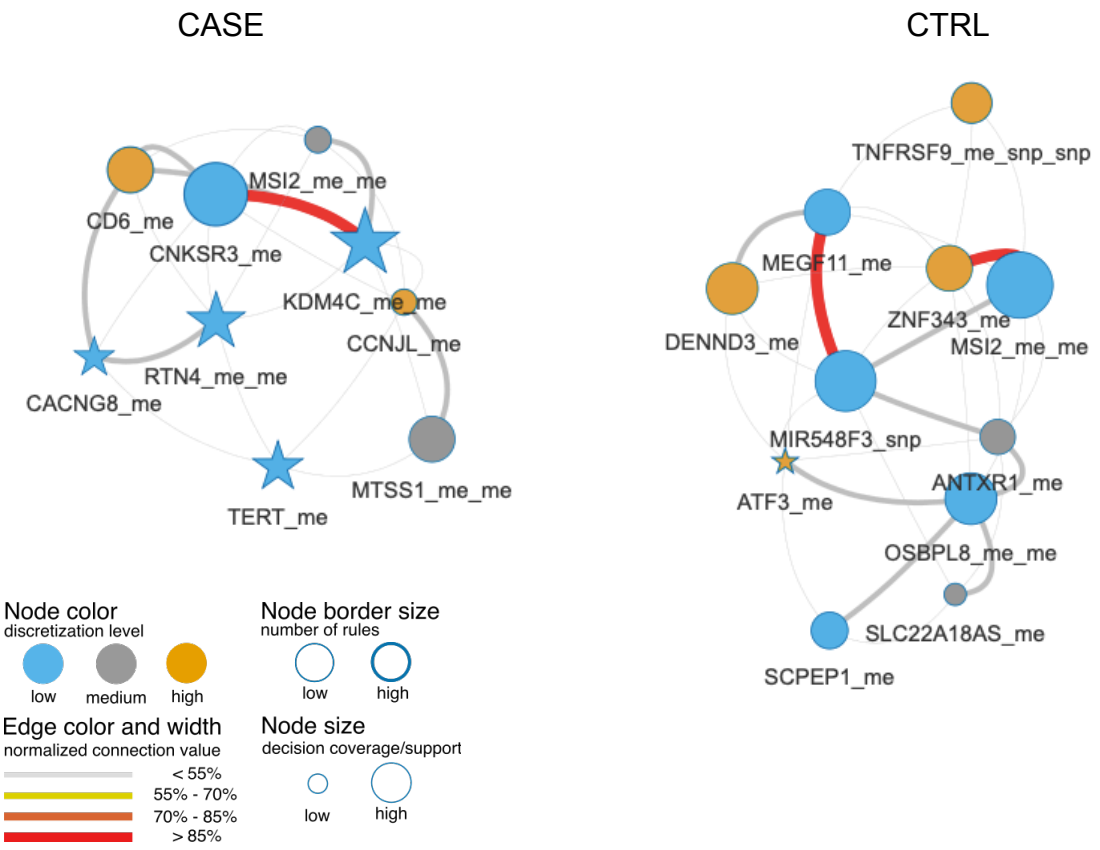

Figure 12. Network for combined model with respect to both decision classes.

The network for top 10 nodes for *CASE* can be seen in Figure 13, the corresponding network for *CTRL* can be seen in Figure 14.

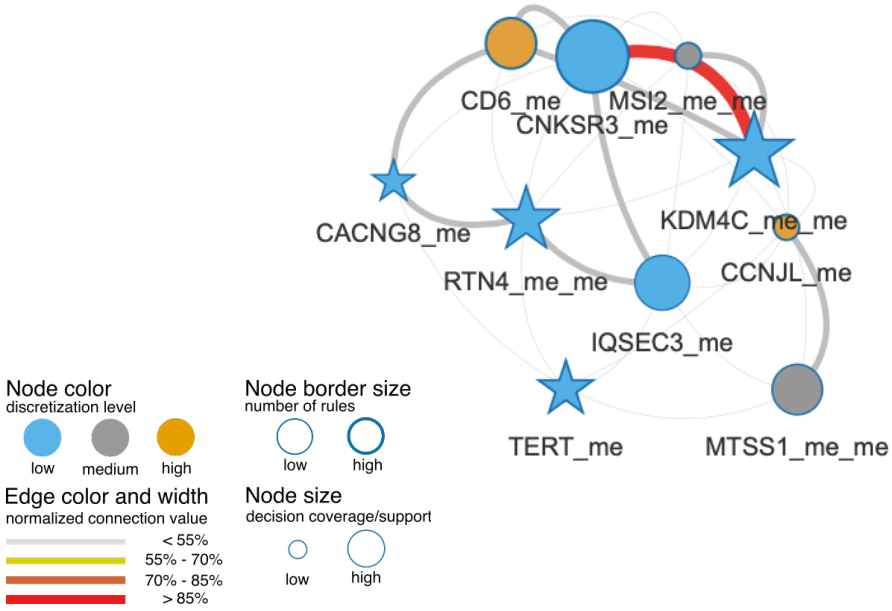

Figure 13. Network for *CASE* class.

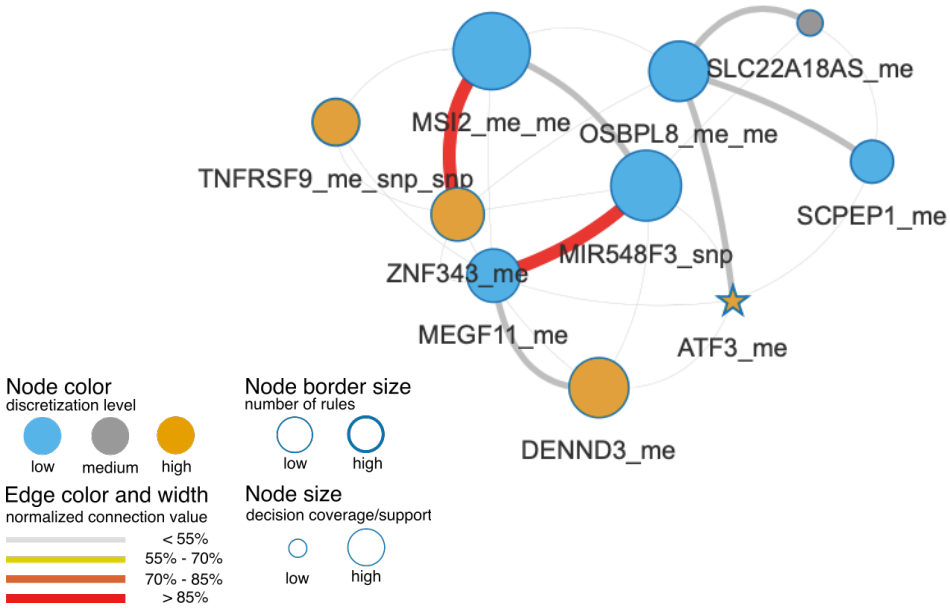

Figure 14. Network for *CTRL* class.

5 Discussion

5.1 Biological Interpretation

From the results, several interesting genes were found. Some of them, such as *CACNG8*, had a previous association with schizophrenia or bipolar disorder, while others have not. Several of the important genes from the model seem to have interesting properties in terms of the disorders. In this subsection, I will further explain the available literature on these genes and attempt to interpret their association to schizophrenia and bipolar disorder.

5.1.1 Functional Enrichment Analysis

Many of the resulting functions are related to the nervous system, which is promising given that the disorders affect the brain. Examples of this include nervous system development, synapse, and axon guidance. Some other functions may provide interesting insights to schizophrenia and bipolar disorder. The most significant pathway from KEGG was the thyroid hormone signaling pathway whose deregulation has been linked to schizophrenia (Santos 2012). Glutamatergic synapses, also found by KEGG, may also be of interest, since their disfunction may lead to ”cognitive impairments and negative symptoms, and drives subcortical dopamine release, resulting in psychosis” (Coyle *et al*. 2012). Furthermore, an elevation of glutamate has been linked to bipolar disorder (Eastwood & Harrison 2010). The association of disfunction of the cholinergic pathway to schizophrenia is also long established, and the antipsychotic clozapine affects this pathway (Saur *et al*. 2016). The reason pathways in cancer was significantly enriched among the genes may be explained by the fact that cancer has been and still is well-studied, so a higher representation of those genes in databases such as KEGG might overlap. Overall, the functional enrichment analysis showed pathways and functions that may be related to schizophrenia and bipolar disorder.

5.1.2 Gene Expression

Comparing the results to the meta study by Lee *et al*. (2012), *ANK3* and *CACNA1C* both were among the 3,453 genes extracted using the methylation sites and SNPs. The focus of Pai *et al*. (2019), *IGF2*, was also among these genes. However, the focus of this discussion will be limited to the top 19 genes in the classification model (see Table 8 for summary of molecular function and relevant biological processes), where both established and novel genes of interest were found.

Table 8. Summary of top 20 nodes (19 genes), all wording from Uniprot or GO.

Gene	Molecular function	Biological processes
ANTXR1	Plays a role in cell attachment and migration, transmembrane signaling receptor activity	Cell adhesion
ATF3	Binds the cAMP response element (CRE), a sequence present in many viral and cellular promoters, represses transcription from promoters with ATF sites	Negative regulation of transcription by RNA polymerase II, gluconeogenesis, positive regulation of cell proliferation
CACNG8 (TARPλ8)	Regulates the activity of calcium channels, and trafficking and gating of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPA)-selective glutamate receptors (AMPARs)	Ion transport, calcium ion transport, transmission of nerve impulse
CCNJL	Contributes to protein kinase activity	Regulation of cyclin-dependent protein serine/threonine kinase activity, protein phosphorylation
CD6	Cell adhesion molecule, T-cell activation and proliferation	Immunological synapse formation
CNKSR3 (MAGI1)	Transepithelial sodium transport	Regulation of signal transduction, positive regulation of sodium ion transport
DENND3	Regulates autophagy in response to starvation, plays a role in protein transport from recycling endosomes to lysosomes	Endosome to lysosome transport, cellular protein catabolic process
KDM4C (JMJD2C)	Central role in histone code	Blastocyst formation, chromatin organization/remodeling
MEGF11	May regulate the mosaic spacing of specific neuron subtypes in the retina	Retina layer formation, homotypic cell-cell adhesion
MIR548F3	miRNA, involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs	—
MSI2	RNA binding, regulates the expression of target mRNAs, may play a role in proliferation and maintenance of stem cells in the central nervous system	Stem cell development
MTSS1	Actin binding	Plasma membrane organization, cell adhesion, transmembrane receptor protein tyrosine kinase signaling pathway

Gene	Molecular function	Biological processes
OSBPL8	Lipid transporter involved in lipid countertransport between the endoplasmic reticulum and the plasma membrane, phosphatidylserine binding	Lipid transport, activation of protein kinase B activity
RTN4 (Nogo Protein)	Induces the formation and stabilization of endoplasmic reticulum (ER) tubules, developmental neurite growth regulatory factor	RNA binding, protein binding
SCPEP1 (SCP1)	Serine-type carboxypeptidase activity	Proteolysis, negative regulation of blood pressure
SLC22A18AS	Antisense to SLC22A18	—
TERT	Telomerase activity	Telomere maintenance
TNFRSF9	Receptor for TNFSF9/4-1BBL, possibly active during T cell activation	Apoptotic process, negative regulation of cell proliferation
ZNF343	May be involved in transcriptional regulation, DNA binding, protein binding	Regulation of transcription, DNA-templated

CACNG8 may be an interesting gene due to its role in glutamate receptors. The gene has been linked to schizophrenia before, and it ”regulates the trafficking and gating properties” of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors (UniProt *n.d.*). Multiple studies have linked abnormal regulation of AMPA receptors to schizophrenia (Drummond *et al*. 2013; Tucholski *et al*. 2013), and more specifically linking down-regulation of *CACNG8* to schizophrenia (Drummond *et al*. 2013), which is consistent with the result in this project. *KDM4C* plays a role in the process of demethylation and has been linked to schizophrenia (Schmidt-Kastner *et al*. 2012), but also to alcohol withdrawal symptoms (Wang *et al*. 2012) and autism (Kantojärvi *et al*. 2010).

RTN4 was a prominent gene in the classification model for the *CASE* class. The gene codes for the protein Nogo-A which is important for neurite facilitation. One study by Novak *et al*. (2002) reported to be over-expressed in schizophrenic patients, which is in contradiction to our finding. However, multiple studies have failed to replicate this association (Takahashi *et al*. 2011). Gardiner *et al*. (2013) suggested a down-regulation of *CD6*, a gene involved in T-cell regulation, in a gene expression study of whole blood. In this project, an over-expression was noted in *CASE* class. However, the gene expression was measured from different sources, i.e. brain versus blood, and thus might not be comparable (see *5.2.1 Data Set* for further discussion). Nevertheless, one review (Horváth & Mirnics 2014) suggests that ”immune system activation plays an important role in developing schizophrenia” as well as ”immune

system activation is persistent throughout the disease." The gene *ATF3* is another gene which one study had the opposite gene expression level compared to these findings. In this project, the gene was shown to be down-regulated or no-change in case samples, but in a study by Drexhage *et al*. (2010) from monocyte samples, the gene was up-regulated in both schizophrenic patients and patients with bipolar disorder.

In the literature review of the genes, some genes were not directly associated to schizophrenia or bipolar disorder, however they do have some interesting connections. The gene *ANTXR1* might be of interest. In the rules for the control group it was marked as no-change, however in the top rules for *CASE* it was always down-regulated. This is interesting due to its possible relationship to the gene *ZNF804A*, a gene commonly associated with schizophrenia (Riley, 2010). When *ZNF804A* was knocked out (Hill 2011), *ANTXR1* was found to be down-regulated. If the gene *ZNF804A* is impaired in schizophrenic patients, this could affect the expression of *ANTXR1*, which is consistent with the results in this project. *TERT* is a widely studied gene for its role in telomere maintenance, and a study by Kao *et al*. (2008) found a significant telomere loss in patients with schizophrenia compared to control (accounting for age and sex), however this may also be caused by stress which is also linked to schizophrenia. The gene *OSBPL8* was shown in this project to be under-expressed in controls compared to case samples. One study (Thomas *et al*. 2003) measured the expression of this gene (among others) in mouse striatum and frontal cortex in response to antipsychotic drugs, and found that the effect was an up-regulation of *OSBPL8*. If some of the patients were taking similar antipsychotics in our data set this could explain the lower level in control. However, it is worth noting that antipsychotics may look different today.

As can be seen in the networks (see Figure 12-14), some genes indeed seem to be interdependent of each other (such as *CNKSR4* and *KDM4C* in *CASE*). However, at this point it is difficult to say what this implies regarding true biological interaction, as more analysis and/or experimental validation would be needed.

5.1.3 Multi-Omics

As described in the background, other studies and projects which performed a feature selection on an unbalanced data set (in terms of number of features from each omics type) got a heavily unbalanced feature set after the feature selection. To account for this, MI was used such that the number of methylation sites and SNPs were the same before running MCFS on the joint data. However, even after this attempt, the number of methylation sites was overrepresented compared to the number of SNPs (3,116 and 337 respectively). This may suggest that the information of methylation sites, at least in this data set, is stronger than the information from SNPs.

However, since the result from MCFS on the joint methylation and SNP data was used to select genes to analyze gene expression data, and since the pathways and genes extracted were promising in terms of association to major psychosis, we can see that the methylation sites and SNPs indeed did affect the gene expression. This indicates that meaningful information can be found using a multi-omics approach. This approach was relatively straightforward and simple to implement, but further adjustments may be needed. Thanks to the machine learning techniques used the result was also interpretable, since each gene in the final model had an associated important methylation site or SNP. In multi-omics, it may be a challenge to combine uneven cohort sizes. An advantage of using the methylation and SNP data as selection of genes in the manner of this pipeline is that it allows for different sample sizes. The sample sizes for the methylation and SNP data analyses were larger than for gene expression data, with 82 and 34 samples respectively, and by using the larger cohort for the first selection of features, the higher statistical power could be translated to the smaller data.

The pipeline is a mix of parallel multi-omics integration (methylation and SNP) and sequential. There is a risk that some information is lost by performing this sequential integration, since it works if there is indeed a clear causal effect between the omics layers. Genes that are not significantly expressed in the post-mortem tissue would be neglected despite having important methylation or SNP signal, and similarly genes that might be differentially expressed but with a weaker methylation or SNP signal would also be excluded. However, if only methylation and SNP data were analysed, we would not get information regarding the genes involved, which would be of interest for treatment development. However, only including gene expression would limit the analysis in other ways, since that would focus on the effect and not the cause. Analyzing all three types simultaneously would lead to a loss of information, since only the samples present in all types could be included, with gene expression limiting the number of samples to 34 instead of 82. In this project we were primarily interested in finding interesting genes from the approach of multiple omics types, such that the genes in the final model were both relevant from their methylation or SNP data, and from the gene expression data. Some improvements to the pipeline are still needed, and in order to get a fuller picture of the disorders attempts should be made to account for the lack of SNPs in the final selection.

5.2 Reliability of Results

Given the complexity of the disorders of interest, it is important to consider which factors could affect the quality of result. As mentioned in *4.7 Classification Models*, the accuracy of the rule-based classification model was ~88.2%, indicating a strong predictive power given the gene expression data. Based on the gene enrichment analysis, the fact that multiple of the

important genes seem to have a logical connection to the disorders is promising, as well as the results from the literature search.

5.2.1 Data Set

The main limitation in terms of statistical power is the size of the data. For the DNA methylation and SNP data the sample size were 82 samples, while for gene expression the sample size was 34 samples. In the earlier stages of this project, attempts were made to find other data sets, however, the fact that the data set from Pai *et al*. (2019) consisted of multi-omics data from the same cohort was an advantage compared to data sets with larger cohorts. Another key aspect of this data set was that the samples were taken from brain tissue, while most of the larger studies use blood samples since it is easier to obtain.

The clinical data from Pai *et al*. (2019) also consisted of information regarding lifestyle factors such as the usage of antipsychotics and smoking status. In the paper, the authors found that after accounting for these covariates for the methylation sites on the *IGF2* locus it still remained significantly hypomethylated. This type of analysis was not performed on the main genes from the analysis of this project, and given that the expression of some genes (as mentioned in *5.1.2 Gene Expression*) were affected by antipsychotics, it is possible these covariates would have an effect on the expression. To improve this project one such analysis might be beneficial. One study (Kumarasinghe *et al*. 2013) compared the gene expression level in peripheral blood of 10 patients before and after six weeks of antipsychotic medication treatment to 11 controls. The authors found that 624 genes were differentially expressed before treatment and 67 after treatment, suggesting that the majority of genes are expressed similar to control after using antipsychotics. The implication of this is that antipsychotic treatment could potentially have an effect on this study, since not all patients were undergoing antipsychotic treatment. However, this study should be viewed with caution, due to the small sample size.

Another issue inherent to the disorders studied is the risk of diagnostic errors. There are multiple studies suggesting a risk of misdiagnosis. One study (Goldberg *et al*. 2008) found that only 33% of patients (28/85) with substance use disorder (SUD) that had been diagnosed with bipolar disorder actually met the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) criteria. In another study by Ruggero *et al*. (2010), the authors found that patients with borderline personality disorder (BPD) in 40% of cases had previously been misdiagnosed with bipolar disorder, and a review of misdiagnosis of bipolar disorder by Singh and Rajput (2006) found that bipolar disorder had a high comorbidity of other diagnosis such as alcohol abuse or panic disorders. Finally, two surveys from 1994 (Lish *et al*. 1994) and 2000 (Hirschfeld *et al*. 2003) found that 73% (363/500) and 69% (414/600) respectively had

been misdiagnosed before getting the diagnosis of bipolar disorder. The brain samples were collected from different brain tissue banks where the diagnoses of the individuals were recorded. While it is difficult to adress this issue from a bioinformatics standpoint, it could still pose as a possible source of error.

Several genes from the model have been mentioned in the scientific literature in the context of mental disorders, however for some genes the expression levels were contradictory to the results in this project. One possible explanation for this is the type of sample the gene expression levels were measured in. The studies mentioned mostly use peripheral blood instead of brain tissue since it is easier to obtain (and thus have a larger sample size), but the expression of genes can differ between these sample types. Furthermore, considering the fact that the disorders studied are in fact mental disorders, with symtoms presented in the brain, it can be argued that in terms of reliability brain tissue may contain more important information in terms of understanding the underlying biology. Another difference to other studies is that this considered data from individuals with schizophrenia or bipolar disorder as one class, so if a gene is up-regulated in one subgroup but not the other it would affect the overall importance. In other words, this would explain why some common genes associated to the disorders separately does not appear as important for both.

5.2.2 Comparison to Original Paper

The main focus of the study by Pai *et al*. (2019) was the gene *IGF2* and the fact that it was hypomethylated. In the 3,453 most important methylation sites and SNPs, sites from *IGF2* were indeed included. One aspect that may affect these results is that the authors excluded genes that encoded synaptic proteins given that a loss in synaptic density characterizes major psychosis. This step was not included in this project, which may explain why functional enrichment analysis still included some of these functions. However, since none of the top 19 genes were included in those pathways, we could still find novel information which made biological sense. Some attempts were made in this project to improve aspects of the original paper, such combining the technical replicates instead of choosing the first replicate, or attempt to adjust for sex-based batch affect (although, as mentioned, batch effect might still be present even though it is not clear from the first principal component). Nevertheless, it is difficult to compare two projects with different aims, since we focused on finding genes of interest (which is feasible in a computer driven approach) while the original study focused on *IGF2* (due to conducting wet-lab experiments as well).

5.3 Challenges

The first main obstacle of this project was the large number of features. It became apparent quite early that performing MCFS on the full data set of methylation sites and SNPs (which amounted to more than 1,000,000 sites) was not feasible, both in terms of computational time and restrictions (such as memory). The balance in this part was to make the computation possible with the given time and resources while including as many features as possible since measures such as MI does not take interdependencies of features into consideration. MI filtration was very fast, and while there was some risk of loss of interdependencies it allowed for MCFS to run in a viable timeframe.

After MCFS, the top features were relatively homogenous comparing case to control. In a scenario where genes have clear correlations to the corresponding class, each feature should contain variability of decision classes and the decision classes should have different states of methylation or SNPs between them. In the case of the top methylation sites and SNPs, the distribution was quite homogenous. For example, the methylation site cg01932551 had perhaps a fourth of case samples methylated and the rest indecisive, while for control all was indecisive (see Figure 7). To account for this, an additional entropy filtration step was tested before MCFS. However, the issue still remained (see Appendix C), indicating an inherent nature of the data. To asses this further, when compared to the methylation sites for *IGF2* that Pai *et al*. (2019) assessed to be significantly differentially methylated a similar pattern could be seen (see Figure 8). Given the complex nature of schizophrenia and bipolar disorder, the relative homogeneity of methylation levels or SNP variants is still reasonable. Another motivation against implementing an entropy filtration would be that it likewise to MI does not take interdependencies into consideration, and more than that: does not take class into consideration.

5.4 Future Improvements

A larger cohort of the same nature as this data would provide a greater statistical power and perhaps lead to more meaningful interpretations. However, brain tissue is more difficult to obtain than for example peripheral blood, so this is still a limitation. An external data set for validation would be important as well. Since the patients with schizophrenia and bipolar disorder were grouped together as *CASE*, a group analysis would reveal whether the genes were more relevant in one disease than another, as well as which genes the subgroups had in common. It would also be important to check whether the methylation sites that were used as selection for the top genes were located in an enhancer-region. When all other components have been developed, an important next step would be to test the novel genes of interest in

wet-lab experiments to check for biological validity, for example by knocking down genes of interest. Some interdependencies were noted in this project, and to further analyze this it would be interesting to estimate correlation between the pairs and then perform Hi-C analysis, which is a method that indicates long range interactions between genomic regions which occurs due to chromosomal folding. Another aspect that could be interesting to examine is to develop a classifier model using the most important methylation sites and SNPs, however we are more interested in the information that is found from all three omics layers with the aim of a more holistic view of the complex biology of the disorders.

6 Conclusion

In this project a multi-omics pipeline was designed, which yielded both established and novel genes, and possible co-prediction mechanisms between them, that may play a role in the complex context of psychosis. The genes *CACNG8*, *RTN4*, *TERT*, *OSBPL8,* and *ANTXR1* are of particular interest in the context of schizophrenia and bipolar disorder, however if these are due to the inherent characteristics of the disorders or environmental factors such as antipsychotic usage or stress needs further analysis. Strong co-dependencies were found, most notable between *CNKSR4* and *KDM4C* in *CASE* samples, however further analysis would be needed to conclude whether or not they are interacting. The multi-omics pipeline designed in this project is straightforward to implement and the approach can be used for cohorts of different sizes between the omics types. Finally, this lead to results with meaningful biological information, however more approaches need to be tested to see if each type of omics-data can be represented in the final model.

To conclude, in this project multiple interesting genes and co-dependencies were found, a multi-omics pipeline which takes different sample sizes into consideration was developed, and finally a rule-based classifier was developed which had high accuracy and legible rules.

7 Acknowledgements

First, I would like to thank my supervisor Mateusz Garbulowski for his guidance and advice throughout the project. I would also like to thank my subject reader Carl Nettelblad for his perspective and expertise which really helped push the project forward. Thanks to my opponent Alfred Andersson for his helpful comments of both the presentation and the report. A big thanks to the course coordinator Lena Henriksson and the course examinator Pascal Milesi for their help regarding all practical aspects of the course.

A big thanks to Jan Komorowski and all members of Komorowski's Bioinformatics Lab for hosting me during this project, and for the discussions that led to new perspectives on the challenges faced. Thank you to Claes Wadelius and all members of Wadelius's Lab for adding new biological insights based on their expertise.

Finally, I would like to thank my family and friends for their support.

List of References

Ashok AH, Marques TR, Jauhar S, Nour MM, Goodwin GM, Young AH, Howes OD. 2017. The dopamine hypothesis of bipolar affective disorder: the state of the art and implications for treatment. Molecular psychiatry 22: 666–679.

Bornelöv S, Marillet S, Komorowski J. 2014. Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers. BMC Bioinformatics 15: 139.

Brisch R, Saniotis A, Wolf R, Bielau H, Bernstein H-G, Steiner J, Bogerts B, Braun K, Jankowski Z, Kumaratilake J, Henneberg M, Gos T. 2014. The role of dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. Frontiers in psychiatry 5: 47–47.

Cai J, Luo J, Wang S, Yang S. 2018. Feature selection in machine learning: A new perspective. Neurocomputing 300: 70–79.

Cambridge English Dictionary. COHORT. WWW-dokument: https://dictionary.cambridge.org/dictionary/english/cohort. Hämtad 2020-05-25.

CAMH. Antipsychotic Medication. WWW-dokument: https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/antipsychotic-medication. Hämtad 2020-05-11.

Chen L, Li J, Zhang Y-H, Feng K, Wang S, Zhang Y, Huang T, Kong X, Cai Y-D. 2018. Identification of gene expression signatures across different types of neural stem cells with the Monte−Carlo feature selection method. Journal of cellular biochemistry 119: 3394–3403.

Coyle JT, Basu A, Benneyworth M, Balu D, Konopaske G. 2012. Glutamatergic Synaptic Dysregulation in Schizophrenia: Therapeutic Implications. I: Geyer MA, Gross G (red.). Novel Antischizophrenia Treatments, s. 267–295. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, Mowry BJ, … , Wray NR, International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature genetics 45: 984–994.

Dabrowski MJ, Draminski M, Diamanti K, Stepniak K, Mozolewska MA, Teisseyre P, Koronacki J, Komorowski J, Kaminska B, Wojtas B. 2018. Unveiling new interdependencies between significant DNA methylation sites, gene expression profiles and glioma patients survival. Scientific Reports 8: 4390.

Davis S, Meltzer PS. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics (Oxford, England) 23: 1846–1847.

Dramiński M, Kierczak M, Koronacki J, Komorowski J. 2010. Monte Carlo feature selection and interdependency discovery in supervised classification. Advances in Machine Learning II, s. 371–385. Springer,

Dramiński M, Koronacki J. 2018. rmcfs: an R package for Monte Carlo feature selection and interdependency discovery. Journal of Statistical Software 85: 1–28.

Dramiński M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J. 2007. Monte Carlo feature selection for supervised classification. Bioinformatics 24: 110–117.

Drexhage RC, van der Heul-Nieuwenhuijsen L, Padmos RC, van Beveren N, Cohen D, Versnel MA, Nolen WA, Drexhage HA. 2010. Inflammatory gene expression in monocytes of patients with schizophrenia: overlap and difference with bipolar disorder. A study in naturalistically treated patients. International Journal of Neuropsychopharmacology 13: 1369–1381.

Drummond JB, Tucholski J, Haroutunian V, Meador-Woodruff JH. 2013. Transmembrane AMPA receptor regulatory protein (TARP) dysregulation in anterior cingulate cortex in schizophrenia. Schizophrenia Research 147: 32–38.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440.

Eastwood SL, Harrison PJ. 2010. Markers of Glutamate Synaptic Transmission and Plasticity Are Increased in the Anterior Cingulate Cortex in Bipolar Disorder. Synaptic Development in Mood Disorders 67: 1010–1016.

Fang L, Zhao H, Wang P, Yu M, Yan J, Cheng W, Chen P. 2015. Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. Biomedical Signal Processing and Control 21: 82–89.

Garbulowski M, Diamanti K, Smolinska K, Baltzer N, Stoll P, Bornelov S, Ohrn A, Feuk L, Komorowski J. 2020. R.ROSETTA: an interpretable machine learning framework. bioRxiv 625905.

Gardiner EJ, Cairns MJ, Liu B, Beveridge NJ, Carr V, Kelly B, Scott RJ, Tooney PA. 2013. Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. Journal of Psychiatric Research 47: 425–437.

Geschwind DH, Flint J. 2015. Genetics and genomics of psychiatric disease. Science 349: 1489.

GHR. 2020. Schizoaffective disorder. WWW-dokument 2020-: https://ghr.nlm.nih.gov/condition/schizoaffective-disorder. Hämtad 2020-05-26.

GHR. TH gene. WWW-dokument: https://ghr.nlm.nih.gov/gene/TH. Hämtad 2020-a-05-11.

GHR. What are single nucleotide polymorphisms (SNPs)? WWW-dokument: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. Hämtad 2020-b-05-11.

GHR. What is epigenetics? WWW-dokument: https://ghr.nlm.nih.gov/primer/howgeneswork/epigenome. Hämtad 2020-c-05-11.

Goldberg JF, Garno JL, Callahan AM, Kearns DL, Kerner B, Ackerman SH. 2008. Overdiagnosis of bipolar disorder among substance use disorder inpatients with mood instability. The Journal of clinical psychiatry 69: 1751–1757.

Google Developers. Machine Learning Glossary. WWW-dokument: https://developers.google.com/machine-learning/glossary?hl=sv. Hämtad 2020-05-25.

Graffelman J. 2015. Exploring diallelic genetic markers: the hardy weinberg package. Journal of Statistical Software 64: 1–23.

Hill MJ, Jeffries AR, Dobson RJB, Price J, Bray NJ. 2011. Knockdown of the psychosis susceptibility gene ZNF804A alters expression of genes involved in cell adhesion. Human Molecular Genetics 21: 1018–1024.

Hirschfeld R, Lewis L, Vornik LA. 2003. Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. The Journal of clinical psychiatry

Hornik K, Karatzoglou DM, Zeileis A, Hornik MK. 2007. The rweka package.

Horváth S, Mirnics K. 2014. Immune System Disturbances in Schizophrenia. Neuroimmune Mechanisms Related to Psychosis 75: 316–323.

Jablensky A. 2010. The diagnostic concept of schizophrenia: its history, evolution, and future prospects. Dialogues in clinical neuroscience 12: 271–287.

Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BEG, Ma S. 2016. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. Genomics 107: 223–230.

Kantojärvi K, Onkamo P, Vanhala R, Alen R, Hedman M, Sajantila A, Nieminen-von Wendt T, Järvelä I. 2010. Analysis of 9p24 and 11p12-13 regions in autism spectrum disorders: rs1340513 in the JMJD2C gene is associated with ASDs in Finnish sample. Psychiatric genetics 20: 102–108.

Kao H-T, Cawthon RM, DeLisi LE, Bertisch HC, Ji F, Gordon D, Li P, Benedict MM, Greenberg WM, Porton B. 2008. Rapid telomere erosion in schizophrenia. Molecular Psychiatry 13: 118–119.

Khaliq Z, Leijon M, Belák S, Komorowski J. 2015. A complete map of potential pathogenicity markers of avian influenza virus subtype H5 predicted from 11 expressed proteins. BMC microbiology 15: 128–128.

Kumarasinghe N, Beveridge NJ, Gardiner E, Scott RJ, Yasawardene S, Perera A, Mendis J, Suriyakumara K, Schall U, Tooney PA. 2013. Gene expression profiling in treatment-naive schizophrenia patients identifies abnormalities in biological pathways involving AKT1 that

are corrected by antipsychotic medication. International Journal of Neuropsychopharmacology 16: 1483–1503.

Laursen TM, Nordentoft M, Mortensen PB. 2014. Excess Early Mortality in Schizophrenia. Annual Review of Clinical Psychology 10: 425–448.

Lee KW, Woon PS, Teo YY, Sim K. 2012. Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: What have we learnt? Neuroscience & Biobehavioral Reviews 36: 556–571.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882–883.

Lesne A. 2014. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Mathematical Structures in Computer Science 24:

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. 2017. Feature Selection: A Data Perspective. ACM Comput Surv 50: Article 94.

Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, Hultman CM. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet (London, England) 373: 234–239.

Lish JD, Dime-Meenan S, Whybrow PC, Price RA, Hirschfeld RM. 1994. The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. Journal of affective disorders 31: 281–294.

List M, Hauschild A-C, Tan Q, Kruse TA, Baumbach J, Batra R. 2014. Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data. Journal of Integrative Bioinformatics 11: 1.

Markowsky G. 2017. Information theory. Encyclopedia Britannica

Mason BL, Brown ES, Croarkin PE. 2016. Historical Underpinnings of Bipolar Disorder Diagnostic Criteria. Behavioral sciences (Basel, Switzerland) 6: 14.

McGrath J, Saha S, Chant D, Welham J. 2008. Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. Epidemiologic Reviews 30: 67–76.

Moghadam BT, Etemadikhah M, Rajkowska G, Stockmeier C, Grabherr M, Komorowski J, Feuk L, Carlström EL. 2019. Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods. Journal of Psychiatric Research 114: 41–47.

Moore LD, Le T, Fan G. 2013. DNA methylation and its basic function. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology 38: 23–38.

NHS. 2017. Psychosis. WWW-dokument 2017-10-23: https://www.nhs.uk/conditions/psychosis/. Hämtad 2020-04-14.

NIMH. 2020a. Bipolar Disorder. WWW-dokument 2020-: https://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml. Hämtad 2020-05-26.

NIMH. 2020b. Schizophrenia. WWW-dokument 2020-: https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml. Hämtad 2020-05-26.

Novak G, Kim D, Seeman P, Tallerico T. 2002. Schizophrenia and Nogo: elevated mRNA in cortex, and high prevalence of a homozygous CAA insert. Molecular Brain Research 107: 183–189.

Pai S, Li P, Killinger B, Marshall L, Jia P, Liao J, Petronis A, Szabó PE, Labrie V. 2019. Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis. Nature Communications 10: 2046.

Pawlak Z. 1984. Rough sets and decision tables. s. 187–196. Springer,

Pawlak Z, Skowron A. 2007a. Rough sets and Boolean reasoning. Zdzis?aw Pawlak life and work (1926–2006) 177: 41–73.

Pawlak Z, Skowron A. 2007b. Rudiments of rough sets. Zdzis?aw Pawlak life and work (1926–2006) 177: 3–27.

Poncet P. 2019. modeest: Mode Estimation.

Riley B, Thiselton D, Maher BS, Bigdeli T, Wormley B, McMichael GO, Fanous AH, Vladimirov V, O'Neill FA, Walsh D, Kendler KS. 2010. Replication of association between schizophrenia and ZNF804A in the Irish Case–Control Study of Schizophrenia sample. Molecular Psychiatry 15: 29–37.

Riza S, Janusz A, Ślęzak D, Cornelis C, Herrera F, Benitez J, Bergmeir C, Stawicki S. 2016. RoughSets: data analysis using rough set and fuzzy rough set theories.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.

Ruggero CJ, Zimmerman M, Chelminski I, Young D. 2010. Borderline personality disorder and the misdiagnosis of bipolar disorder. Journal of psychiatric research 44: 405–408.

Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, … , Murray CJ. 2012. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. The Lancet 380: 2129–2143.

Santos NC, Costa P, Ruano D, Macedo A, Soares MJ, Valente J, Pereira AT, Azevedo MH, Palha JA. 2012. Revisiting thyroid hormones in schizophrenia. Journal of thyroid research 2012: 569147–569147.

Saur T, Cohen BM, Ma Q, Babb SM, Buttner EA, Yao W-D. 2016. Acute and chronic effects of clozapine on cholinergic transmission in cultured mouse superior cervical ganglion neurons. Journal of neurogenetics 30: 297–305.

Schmidt-Kastner R, van Os J, Esquivel G, Steinbusch HWM, Rutten BPF. 2012. An environmental analysis of genes associated with schizophrenia: hypoxia and vascular factors as interacting elements in the neurodevelopmental model. Molecular Psychiatry 17: 1194–1205.

Shastry BS. 2009. SNPs: Impact on Gene Function and Phenotype. I: Komar AA (red.). Single Nucleotide Polymorphisms: Methods and Protocols, s. 3–22. Humana Press, Totowa, NJ.

Singh T, Rajput M. 2006. Misdiagnosis of bipolar disorder. Psychiatry (Edgmont (Pa: Township)) 3: 57–63.

Song L, Langfelder P, Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics 13: 328.

Sun YV, Hu Y-J. 2016. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Advances in genetics 93: 147–190.

Takahashi N, Sakurai T, Davis KL, Buxbaum JD. 2011. Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia. Progress in neurobiology 93: 13–24.

Thomas EA, George RC, Danielson PE, Nelson PA, Warren AJ, Lo D, Sutcliffe JG. 2003. Antipsychotic drug treatment alters expression of mRNAs encoding lipid metabolism-related proteins. Molecular Psychiatry 8: 983–993.

Tucholski J, Simmons MS, Pinner AL, Haroutunian V, McCullumsmith RE, Meador-Woodruff JH. 2013. Abnormal N-linked glycosylation of cortical AMPA receptor subunits in schizophrenia. Schizophrenia Research 146: 177–183.

Wang F, Zhang S, Wen Y, Wei Y, Yan H, Liu H, Su J, Zhang Y, Che J. 2013. Revealing the architecture of genetic and epigenetic regulation: a maximum likelihood model. Briefings in Bioinformatics 15: 1028–1043.

Wang K-S, Liu X, Zhang Q, Wu L-Y, Zeng M. 2012. Genome-wide association study identifies 5q21 and 9p24.1 (KDM4C) loci associated with alcohol withdrawal symptoms. Journal of Neural Transmission 119: 425–433.

Wang S, Shi X, Wu M, Ma S. 2019. Horizontal and vertical integrative analysis methods for mental disorders omics data. Scientific Reports 9: 13430.

WHO. 2019. Schizophrenia. World Health Organization (WHO)

WHO. Disability weights, discounting and age weighting of DALYs. WWW-dokument: https://www.who.int/healthinfo/global_burden_disease/daly_disability_weight/en/. Hämtad 2020-a-04-14.

WHO. Mental disorders. WWW-dokument: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Hämtad 2020-b-05-11.

Øhrn A, Komorowski J. 1997. Rosetta--a rough set toolkit for analysis of data. Proc. Third International Joint Conference on Information Sciences

Appendix A

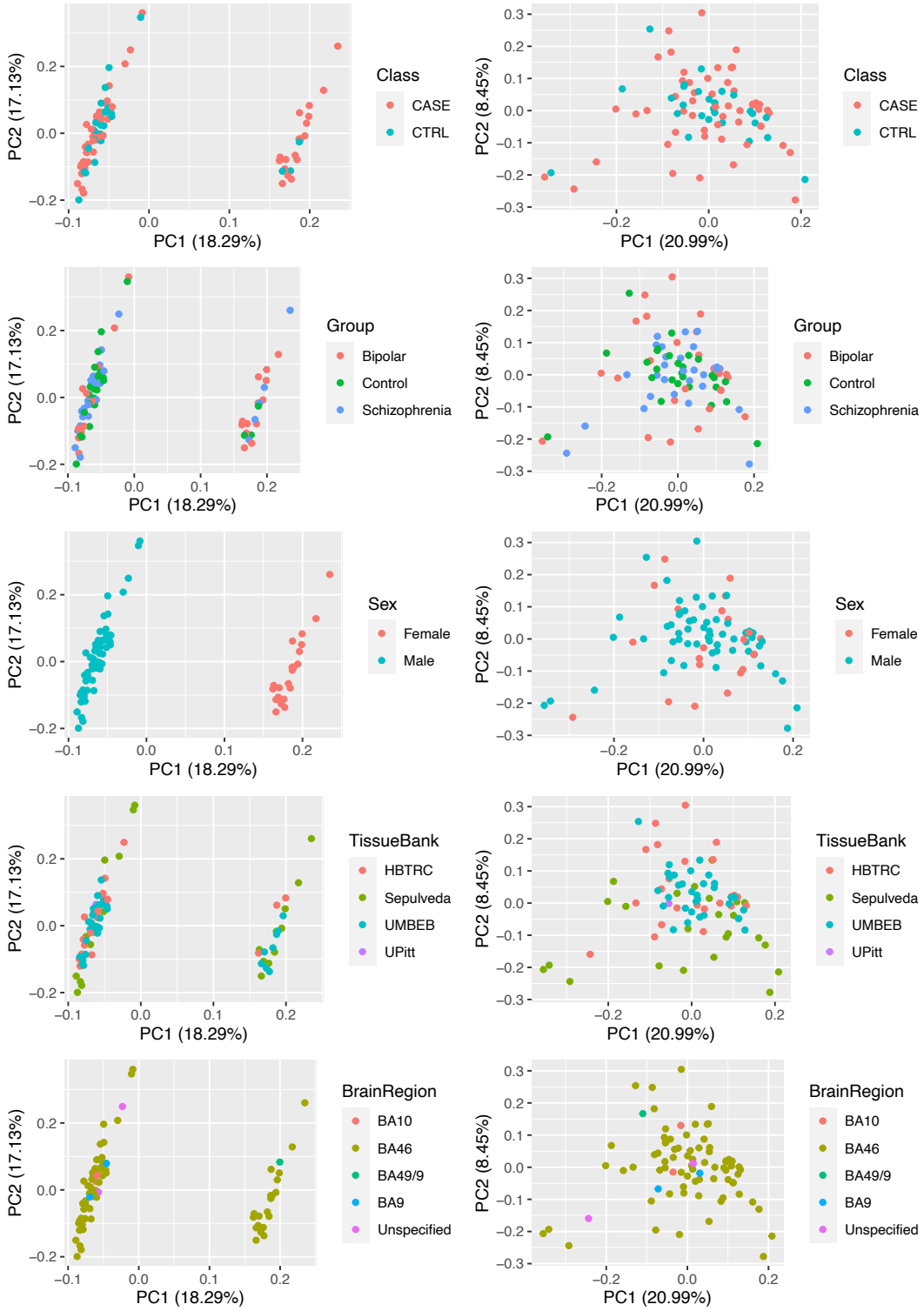

Figure 1. PCA of methylation data. Left is for full data set, right is after excluding sites from X and Y chromosome.

Appendix B

- affy
- arules
- biomaRt
- data.table
- edgeR
- GEOquery
- ggfortify
- gmodels
- ggpubr
- gridExtra
- HardyWeinberg
- infotheo
- modeest
- plyr
- psygenet2r
- rmcfs
- R.ROSETTA
- sva
- tidyverse
- VisuNet
- xlsx

Appendix C

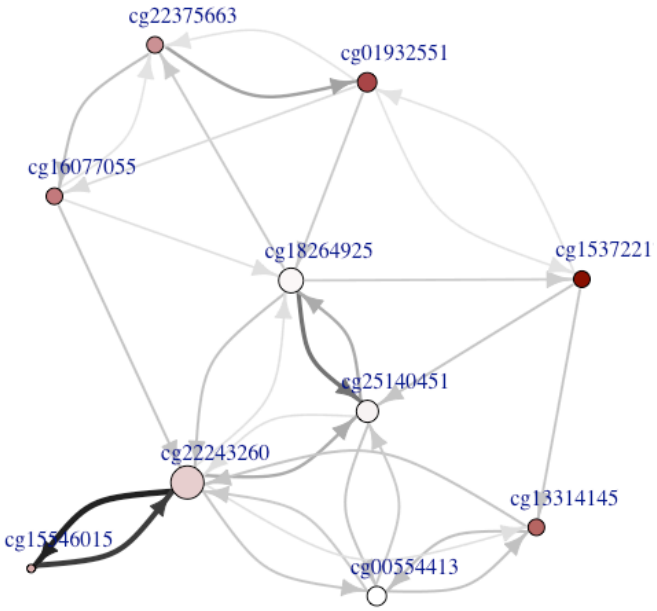

Figure 1. ID graph of top 10 nodes from MCFS.

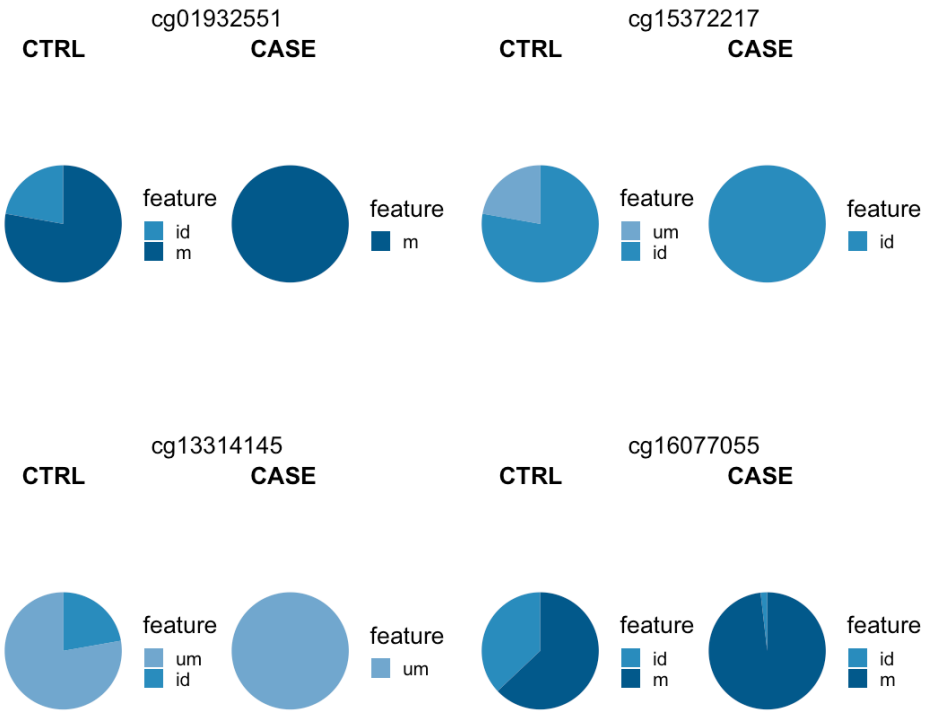

Figure 2. Pie charts of top 4 sites from MCFS. ”m” means methylated, ”id” indecisive, ”um” unmethylated.