GENOMIC MODELLING OF BIPOLAR DISORDERS: COMPARISON OF MULTIFACTOR DIMENSION REDUCTION AND CLASSIFICATION-BASED DATA MINING METHODS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

CENGİZHAN AÇIKEL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

MARCH 2017

GENOMIC MODELLING OF BIPOLAR DISORDERS: COMPARISON OF MULTIFACTOR DIMENSION REDUCTION AND CLASSIFICATION-BASED DATA MINING METHODS

Submitted by **CENGİZHAN AÇIKEL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in the Department of Medical Informatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin ______________
Director, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son ______________
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son ______________
Supervisor, **Health Informatics, METU**

Examining Committee Members:

Assoc.Prof.Dr. Cem İyigün ______________
Industrial Engineering, METU

Assoc. Prof. Dr. Yeşim Aydin Son ______________
Supervisor, Health Informatics, METU

Assist. Prof. Dr. Aybar Can Acar ______________
Health Informatics, METU

Assoc.Prof.Dr. Güvem Gümüş Akay ______________
Brain Research Center, Ankara University

Assist.Prof.Dr. Ercüment Çiçek ______________
Computer Engineering, Bilkent University

Date: ____/____/______

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Cengizhan AÇIKEL

Signature : ____________________

ABSTRACT

GENOMIC MODELLING OF BIPOLAR DISORDERS: COMPARISON OF MULTIFACTOR DIMENSION REDUCTION AND CLASSIFICATION-BASED DATA MINING METHODS

Açıkel, Cengizhan
Ph.D. Department of Health Informatics
Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

March 2017, 103 pages

In genomic modeling, various data mining techniques are proposed with varying degrees of success to analyze high-dimensional data generated by genome-wide association studies of complex genetic disorders. In this study, we aimed to compare Multifactor Dimensionality Reduction (MDR), a non-parametric approach that can be used to detect relevant interactions between Single Nucleotide Polymorphisms (SNPs) or genes, with 3 other classification based data mining methods for genomic modeling of bipolar disorders.

This study was performed on a Whole Genome Association Study of Bipolar Disorders (dbGaP Number: phs000017.v3.p1) data. Three classification based data mining methods (Random Forest [RF], Naïve Bayes [NB] and k-Nearest Neighborhood [kNN]) and MDR were performed. Pathway analysis, based on identified common SNPs is also performed, and evaluated.

RF, NB, and kNN identified 16, 13, and 10 candidate SNPs, respectively. The top six SNPs were common to all three. The RF and kNN models were found to be more successful than the NB model, with recall values above 0.95. On the other hand, MDR generated a model with comparable predictive performance based on five SNPs identified by analysis of two-way and three-way interactions. Although a different SNP profile is identified in MDR compared to the other three classification-based models, all models identified SNPs mapping to the *ZNF507* and *DOCK10* genes.

Three classification-based data mining approaches, RF, NB and kNN, have prioritized similar SNP profiles as predictors of bipolar disorders, in contrast to MDR, which reported a different set, which includes only five SNPs. The reduced number of SNPs, without loss in the classification performance, has the potential to facilitate validation studies to understand the molecular mechanisms behind bipolar disorders and molecular diagnostics tools. Nevertheless, we emphasize that translation of genomic models to the clinic require models with higher levels of classification performance.

Keywords: Bioinformatics, Multifactor dimensionality reduction, classification methods, bipolar disorders, *GWAS*

ÖZ

BIPOLAR BOZUKLUKLARIN GENOMİK MODELLEMESİ: ÇOK FAKTÖRLÜ BOYUT İNDİRGEME VE SINIFLAMA TABANLI VERİ MADENCİLİĞİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Açıkel, Cengizhan
Doktora, Sağlık Bilişimi Bölümü
Danışman: Doç. Dr. Yeşim Aydın Son

Mart 2017, 103 sayfa

Genomik modellemede; farklı veri madenciliği yöntemleri, değişken başarılar ile genom boyu ilişkilendirme çalışmaları ile elde edilen çok katmanlı verileri analiz etmede kullanılmaktadırlar. Bu çalışmada; çok faktörlü boyut indirgeme (MDR) (tek nükleotid polimorfizimleri (SNP) veya genler arasındaki etkileşimleri de inceleyen bir non-parametrik yöntem) ile sınıflama tabanlı üç veri madenciliği yönteminin, bipolar bozukluk genomik modellerinde, karşılaştırılması amaçlanmıştır.

Bu çalışma Bipolar Bozukluklar Tüm Genom Asosyasyon Çalışması (dbGaP Numarası: phs000017.v3.p1) verisi ile yapılmıştır. Sınıflama temelli 3 veri madenciliği yöntemi (Random Forest [RF], Naïve Bayes [NB] and k-Nearest Neighborhood [kNN]) ve MDR kullanılmıştır. Ayrıca saptanan ortak SNP'ler için pathway analizleri yapılmış ve yorumlanmıştır.

RF, NB, ve kNN sırasıyla 16, 13, ve 10 aday SNP saptamıştır. Üç yöntemin belirlediği ilk altı SNP ortaktır. RF ve kNN, 0.95 üzerindeki recall değerleri ile, NB'e gore daha başarılı sonuçlar vermiştir. Diğer yandan MDR, iki ve üç yönlü etkileşim ile, sadece 5 SNP ile karşılaştırılabilir kestirim gücüne sahip bir model üretmiştir. MDR ile saptanan SNP'ler sınıflama tabanlı diğer üç modelden farklı olmasına karşın tüm modellerde polimorfizmlerin *ZNF507* ve *DOCK10* genlerine haritalandıkları saptanmıştır.

Sadece farklı 5 SNP saptayan MDR'ın aksine, üç sınıflama tabanlı veri madenciliği yaklaşımı, RF, NB ve kNN, bipolar bozukluk kestiricisi olarak benzer SNP polimorfizmlerini önceliklendirmişlerdir. Sınıflama performansını düşürmeksizin, daha az sayıda SNP ile kestirim yapmak bipolar bozuklukların arkasındaki moleküler mekanizmanın anlaşılmasını ve tanı araçlarının validasyon çalışmalarını kolaylaştırmaktadır. Bununla birlikte genomik modellerin kliniğe geçişinin daha yüksek sınıflandırma performansına sahip modeller gerektirdiği belirtilmelidir.

Anahtar kelimeler: Çok faktörlü boyut indirgeme, MDR, sınıflama yöntemleri, bipolar bozukluklar, tüm genom asosyasyon çalışması

To my family

ACKNOWLEDGEMENTS

I express sincere appreciation to Assoc. Prof. Dr. Yeşim Aydın Son for her perfect coaching and guidance for this study and all over doctorate duration. Her continuous guidance is invaluable to me in my theoretical education process and writing of this thesis.

Thesis progress committee members Prof.Dr. Erkan DEMİRKAYA, Assoc.Prof.Dr. Yeşim AYDIN SON, Assoc. Prof. Dr. Tuğba TAŞKAYA TEMİZEL, Assoc.Prof.Dr. Cem İYİGÜN, Assist. Prof. Dr. Aybar Can ACAR; and examining committee members Assist. Prof. Dr. Aybar Can ACAR, Assoc. Prof. Dr. Yeşim AYDIN SON, Assoc.Prof.Dr. Cem İYİGÜN, Assist.Prof.Dr. Ercüment ÇIÇEK, Assoc.Prof.Dr. Güvem GÜMÜŞ AKAY for their participation and valuable comments.

I am grateful to all faculty members of The Graduate School of Informatics of Middle East Technical University, and all personnel from administrative staff to senior managers for their support throughout my doctorate studies. I learned a lot from both courses and scientific meetings. The weekly assessment meetings of Dr. AYDIN SON allowed us the opportunity to share our experiences. I owe all my classmates a debt of gratitude.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	African American Ancestry
ASD	Autism Spectrum Disorder
BARD	Bipolar and related disorders
BDO	Bipolar disease only
BP	Bipolar
CA	Classification Accuracy
DALY	Disability adjusted life years
DNA	Deoxyribonucleic acid
DSM	Diagnostic and Statistical Manual of Mental Disorders
EA	European Ancestry
FN	False negative
FP	False positive
GRU	General research use
GWAS	Genome wide association studies
HWE	Hardy-Weinberg Equilibrium
ICD 10	International Statistical Classification of Diseases and Related Health Problems 10th version
kNN	k-Nearest Neighbor
MAF	Minor allele frequencies
MDR	Multifactor Dimensionality Reduction
NB	Naïve Bayes
NIMH	National Institute of Mental Health
RF	Random Forest
RNA	Ribonucleic acid
RS ID	Reference SNP cluster ID
SD	Standard deviation
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine

TCR	True Classification Rate
TN	True negative
TP	True positive
WHO	World Health Organization

CHAPTER 1

1 INTRODUCTION AND BACKGROUND

1.1 MOTIVATION

A common genotyping platform can include up to 1 million SNPs. In general, SNP variants selected for genotyping are common variants with allele frequencies greater than 1% in human populations, and can be found in multiple populations (European, Asian, African). Analysis of data generated with genotyping platforms through classical GWAS approaches led researchers to associations of the condition under study with a few variants with high risk of susceptibility. In case of complex diseases where genome wide variants with low risk of susceptibility are the underlying factors in the genetic model of the disease, classical GWAS approaches fail to identify the whole SNP profile associated with the disease risk. Additionally, the large number of variables are usually prone to increased statistical error. To reach needed alpha and beta error levels, researchers require large study groups of thousands of patients and controls-this also increases the costs of the study. Recently, different analysis approaches were proposed to aid researchers in identifying effective solutions to increase the power of their study during data analysis.

The relationships between the genetic background and phenotype do not always display a linear association. Therefore, purified statistical methods become insufficient to determine the associations between genotype and phenotype. Another important task for discovering gene/SNP associations is to prioritize detected genetic elements. The statistical interpretation does not offer any understanding of biological or functional role of the gene/SNPs, which is, necessary. As such, data mining methodologies that enable researchers to find hidden variables, complex relationships and non-linear associations becomes important in bioinformatics [1–4].

In the human genome, disease associated SNPs do not act in isolation. The SNP-SNP interactions also need to be evaluated to define biological pathways of function of candidate risk SNP's. The main mechanisms that needs to be investigated in complex traits are; interactions between genes, low penetrance, and environmental factors. Interaction studies have a higher calculation burden, which increases the complexity of the analysis exponentially. Along with the other necessities of association studies, the need to study interactions makes data mining approaches essential for post-GWAS analysis.

To detect SNP-SNP interactions, the nonparametric approaches search through different levels of interaction without consideration of the significance of the main effects. Combinatorial Partitioning Method, Neural Networks and Multifactor Dimensionality Reduction (MDR) are some of these nonparametric methods. All of these methods try to detect the relevant interactions between the SNPs (or genes) by either reducing the dimension or recognizing the useful hidden patterns. These approaches do not make assumptions about the functions of dependence between the trait and the SNPs. Instead of functional relationships methods, the interest lies in data driven relations. MDR generates a classification model with SNP-SNP interactions to predict diseases [5–10].

In this study our main goal was to compare MDR results (including two way and three way SNP-SNP interactions) with conventional classification methods on bipolar disorders genome wide data, and describe novel candidate SNPs.

To evaluate the effectiveness of MDR, validity measures were compared with three classification based data mining methods. The first selected method is RF since significant statistical and bioinformatics challenges of large scale classical regression analysis is not feasible. RF is one of the most popular machine learning methods and has a very broad range of applications and is commonly used in GWAS studies. As a nonparametric tree-based ensemble approach that merges the ideas of decision trees, RF is especially effective in “large p, small n” problems. Grouping of trees enables researchers to deal with collinearity and interaction among variables. It can also be useful for selecting and ranking variables. Thus, for these reason RF is an appropriate tool for genomic data and bioinformatics [11,12].

Naïve Bayes (NB), relies on contingency table analysis and therefore it does not assume a pre-specified model of genetic effect. Although MDR or likelihood ratio based tests’ (such as logistic regression) have an exhaustive nature, NB is a non-exhaustive method and is commonly used in genomic studies. In addition, it should be noted that RF and decision trees are discriminative models but NB is an exceptional generative model [13,14].

KNN is both simple and clinically appealing, but it has large performance variations. This variation depends on the feature ranking method, the number of features used, the use of metric measures for distance, the number of selected neighbors, weightings and thresholds. In this study, kNN was chosen to evaluate both genotype and phenotype data together as suggested in the literature [15–18].

1.2 WHAT ARE BIPOLAR DISORDERS

1.2.1 Definition of Bipolar Disorders

Bipolar disorders (also known as manic-depressive illness) is a psychiatric disorder that causes unusual shifts in mood, energy and activity levels [19,20].

The basic component of Bipolar I Disorder is a clinical course that is described by the event of at least one Manic Episode or Mixed Episodes. The basic component of Bipolar II Disorder is a clinical course that is described by the event of at least one Major Depressive Episode joined by no less than one Hypomanic Episode [21,22].

Bipolar (BP) disorders are one of the most common psychiatric disorders all around the world. According to the WHO, data prevalence of BP is estimated to be approximately 1.0% for the general population that meets lifetime criteria for BP type I (BP-I) and 2-7% as life-time prevalence of BP-II [23,24].

BP disorders are responsible for the loss of more disability-adjusted life-years than cancers or many other disorders. The WHO 2002 disability adjusted life years study has shown that BP disorders to have a great burden globally, and in contrast to other psychiatric disorders its burden is higher among underdeveloped countries [25].

1.2.2 Classification of Bipolar Disorders

Bipolar disorders are classified under the “Mood Disorders” chapter in DSM IV-TR. The Mood Disorders are divided into two; the Depressive Disorders ("unipolar depression"), and the Bipolar Disorders. The most common used criteria for BD are from the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5), and the ICD 10 or ICD 11 (International Statistical Classification of Diseases and Related Health Problems) of the World Health Organization. However DSM criteria are commonly used in the USA and researchers commonly use ICD-10 criteria in Europe [25–28].

DSM V Diagnostic Codes for Bipolar Disorders

Same as the DSM IV TR, DSM V is classified into two types. Type 1 bipolar disorders include current hypomanic episodes, manic episodes with psychotic features, depressed episodes, mixed episodes and bipolar disorders currently in remission. Type 1 bipolar disorders exclude bipolar disorders, a single manic episode, major depressive disorder-single episodes, major depressive disorder-recurrent cases [28].

Type 2 bipolar disorders include the presence (or history) of one or more Major Depressive Episodes and at least one Hypomanic Episode. The mood symptoms are not better accounted for by Schizoaffective Disorder and are not superimposed on Schizophrenia Schizophreniform Disorder. Type 2 bipolar disorders also exclude cyclothymic cases [22,28].

A complete list of DSM V classification is represented in Appendix B.

1.2.3 Epidemiology of Bipolar Disorders

Several factors should be taken into account when diagnosing BD. The average age of onset of these disorders is at 25 years old. It is seen in equal amounts in males and females. A clear anamnesis from self-reported experiences of the patient and information from family members and friends should be taken. Psychiatric examination is critical to deciding the diagnosis and treatment of the disorder [24,25,29]. In terms of ethnic origin, those of an African and Caucasian origin are equally affected, but the prevalence of bipolar disorders is lower among Asians [30,31]. The incidence rates have a peak in late adolescence, but 10% of mania attacks begin after age 50 [32]. The hospitalization rate stands at approximately 50% and over 95% of those hospitalized for first-episode mania achieve remission in 6 weeks. Furthermore, approximately 20% switch from mania to depression [33].

The most common source of mortality among bipolar patients is suicide. More than 30% of bipolar patients reported past suicide attempts, and the suicide rate is 0.4% annually. These rates are 10 to 20 times higher than the general population [34–36].

Bipolar disorder (BD) is a life-long mental disorder. It affects 2-5% of the population [20,37] and has negative effects on quality of life, functioning and employment. It is responsible for the loss of more disability-adjusted life years than cancers or many major neurologic conditions [38–40]. The WHO reports that disability-adjusted life years of BD cause a great burden globally [35]. The economic costs of the disorder is over $45 billion for the United States and missed work days are around 50 days per year [41].

The WHO 2002 disability adjusted life years' study has shown BP disorder to have a great burden globally, and in contrast to other psychiatric disorders its burden is higher among underdeveloped countries. Although the prevalence and incidence of bipolar disorder is are approximately the same throughout the world, but the disability associated with it may be greater in developing countries [31,35].

Figure 1.1 displays the burden of BP globally[25].

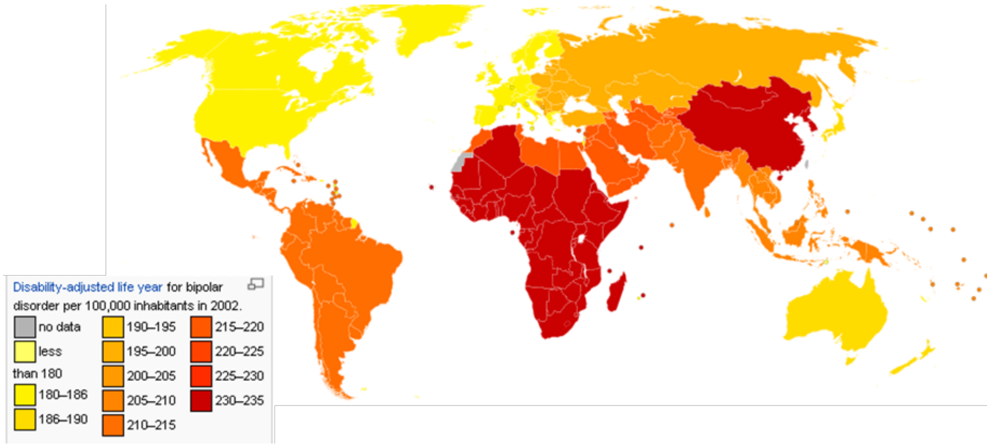

Figure 1-1. Bipolar disorder world map DALY WHO2002

(Source: http://www.who.int/mental_health/en/investing_in_mnh_final.pdf)

1.2.4 Etiology of Bipolar Disorders

Physiological

Some structural anomalies of the brain, such as lateral ventricles expansions, globus pallidus growth and hyperintensity of deep white matter have been reported in BP patients. There is also evidence of hypothalamo pituitary axis abnormalities [42]

A popular hypothesis about BP, is that it is the result of a circadian rhythm defect with altered melatonin activity. The circadian system modulates different biological functions like metabolism, body temperature, sleep-wake activities of the brain and the cellular proliferation system. The environmental changes of the modern lifestyle can change the circadian system and increase the risk of developing pathologies including mood disorders besides cancer, preeclampsia, diabetes etc. The circadian system is organized by clock genes *CLOCK*, *BMAL1*, *PER1*–3, and *CRY1*-2. The process by which these clock genes influence the development of disease is unknown [43].

Environmental

Many researchers suggest that the social environment has a strong influence on the course of bipolar depression. Trauma, negative life events, social support deficits, and family difficulties are common and predict bipolar disorders or a more severe course of depression. These factors may lead to triggering of bipolar depression or unipolar depression [44].

Apart from the effect of risky genes in psychiatric disorders, environmental factors may also impact brain development during and after the perinatal period. Environmental factors such as maternal viral infections and obstetric hypoxia are common factors that may cause stress during neurodevelopment and have been identified to play a role in bipolar disorder. Smaller hippocampal volumes, perinatal stress or psychosocial stress in adulthood are also important triggers. Repeated administration of glucocorticoids can result in degeneration of glucocorticoid-sensitive hippocampal neurons and may contribute to the pathophysiology of affective disorders. Epigenetic mechanisms that change chromatin structure by histone acetylation or DNA methylation may mediate effects of environmental factors. Gene-environmental interaction studies may lead to therapeutics, influencing epigenetic targets [44].

Genetic

The overall heritability of bipolar disorder is estimated to be up to 0.70. Bipolar concordance rates are around 40% in monozygotic and 10% in dizygotic twins [45,46]. The risk of bipolar disorder is nearly ten-fold more in first degree-relatives of BD sufferers than the general population [47]. Many chromosomal regions, candidate genes and polymorphisms have been suggested in the etiology of BD. But the current genome-wide association study failed to find any particular locus for BD, which suggests that no single gene is responsible for BD. Findings show that different genes from different families are implicated. The most implied locations are on chromosomes 6q 8q and 21 and mostly reported and justified genes are *SLC6A4*/5-*HTT* (serotonin transporter gene), *BDNF* (brain-derived neurotrophic factor), *DAOA* (D-amino acid oxidase activator), DTNBP1 (dysbindin), NRG1 (neuregulin 1) and *DISC1*. Genome-wide significant associations showed many common single nucleotide polymorphisms and variants within the genes *CACNA1C*, *ODZ4*, and *NCAN* [48,49]. The main molecular function of candidate bipolar genes are: biogenic amine modulation, genes of the serotonergic system, genes of the dopaminergic neurotransmitter system, channelopathies and ion channel associated proteins[40], growth hormones, brain development and neuronal growth, clock genes, genes of the lithium signal transduction pathway, genes of the glutamatergic neurotransmitter system, signal transduction, HPA axis and stress, cell adhesion, mitosis, tumorigenesis, and DNA repair/DNA stability [50]. Previously detected candidate genes that are arranged based on function are listed in Table 1.1, and an up to date and detailed list is added as Appendix 1 [51].

Table 1-1. Summary list of previously determined genes and relations (Adapted then updated) [51]

Function	**Gene**
Genes involved in biogenic amine modulation	*MAOA* *COMT* *TPH1* *TPH2*
Genes of the serotonergic system	*HTR1A* *HTR2A* *HTR2C* *HTR3A* *HTR3B* *HTR4* *HTR5A* *SLC6A4* (5-*HTT* gene, *SERT*) *VNTR* poly- morphism in intron 2 *SLC6A4* (5-*HTT* gene, *SERT*) *5-HTTPLPR* insertion/ deletion polymorphism (short and long allele)
Genes of the dopaminergic neurotransmitter system	*DRD1* *DRD2* *DRD3* *DRD4* *DRD5* *DAT1 (SLC6A3)*
Channelopathies and ion channel associated proteins	*ANK3* *CACNA1C* *KCNC2* *P2RX7/4* *SLC39A3*

Table 1.1. (cont.)

Function	**Gene**
Growth hormones, brain development and neuronal growth	*EGFR* *BDNF* *NCAM1* *DISC1* *NRG1* *IGF1*
Clock genes	*PER3* *ARNTL* (BmaL1) *CRY1* *CSNK1ε* *CLOCK* *NRIDI* (*REV-ERBa* gene)
Genes of the Lithium signal transduction pathway	*DGKH* *SORCS2* *DFNB31* *PDE10A* *NXN*
Genes of the glutamatergic neurotransmitter system	*GRIN2B* *GRIA1* *GRM3* *GRM7* *GRIK4* *GABRB1* *GABRA3* *GABRA5*
Signal transduction genes	*GRK3* *PTGFR* *HPA* axis and stress *CRH*
Cell adhesion genes	*TSPAN8* *JAM3* *PDLIM5* *NCAN*

Table 1.1. (cont.)

Function	**Gene**
Mitosis, tumorigenesis and DNA repair/DNA stability	*GNL3* *PALB2* *NEK4* *BRCA2*
Others	*G72/G30 (DAOA)* *CHMP1.5* *GCHI* *NAPG* *MYO5B* *SYN3* *DTNBP1* *TRANK1 (LBA1)* *LMAN2L* *MARK1* *SLC22A16* *BRD1* *DCTN5*

1.3 DATA MINING WITH GWAS DATA

1.3.1 Genome Wide Association Studies

Genome-wide association study (GWAS) is an approach includes scanning markers across the whole sets of DNA. It aims to detect, protect and prevent the disease via identify new genetic associations. This make possible develop better strategies about medical conditions.

GWAS Approach

GWAS are useful in finding genetic variations that contribute to common or complex diseases also. The ability to conduct GWAS has advanced rapidly with the advent of high-throughput genotyping technology. Now common chips may genotype 1 million SNPs per individual. The current SNP maps have roughly 85% coverage of the genome in Caucasians, so these approaches have the potential to be very powerful [52,53].

Researchers have a defined set of research tools in genetic area after completion of the Human Genome Project (2003) and HapMap Project (2005). These tools contain computerized databases (the reference human genome sequence, human genetic variation map) and new quick and accurate technologies that can analyze whole-genome samples for genetic variations. After these times many chip sets generated by commercial companies and plenty of research conducted to find new genetic backgrounds by using GWAS [53,54].

To carry out a genome-wide association study, researchers use two groups of participants: Cases; the people with the disease being studied and controls; similar people without the disease. The complete set of DNA (or genome) for each individual is purified from the cells, then placed on tiny chips and scanned. The scanners quickly survey each genome for selected markers of genetic variation. This variations are called single nucleotide polymorphisms, or SNPs [52].

If genetic variations are found more frequent in cases than controls, the variations are considered as “potentially associated" with the disease. The associated genetic variations can serve as markers [52].

Evolution of analytical approaches for GWAS data

First decade of GWAS focused on identification of number of loci associated with diseases, which then led to the application of the candidate gene approach to reveal the molecular etiology of diseases. This process has not been very productive as the polymorphic marker for complex genetic diseases tends to be composed of profiles of polymorphisms that are not limited to single loci on the genome [54]. Risk Single nucleotide polymorphisms (SNPs) cannot explain large amounts of heritability in complex diseases. A reason for this may be; commonly used SNP analysis strategies mainly interest single SNPs. Association between SNPs cannot be evaluated as easily as that [55]. Especially among human subjects, phenotypic variations are common, and complex diseases involve complex etiologies including interactions between genetic and environmental factors. Moreover the gene-gene (or SNP-SNP) interactions in univariate analyses may be limiting the success of GWAS studies for complex diseases [56]. In order to appropriately account for complex genetic diseases, data-mining or knowledge discovery techniques are utilized to discover patterns. GWAS data have some “big data” problems and in large amounts of data, knowledge discovery methods are gaining popularity for genetic association studies [57]. GWAS is evolving from identification from single variations to determination of profiles associated with different conditions [58].

Biologists primarily work to gather new genomic information and on the other side mathematicians, statisticians, and computer scientists try to evaluate these information more effectively. [1,2]. Data mining is one of these kinds of technology and many health science disciplines are interested in genomic data. That is to use them to make better decisions, to understand mechanisms and discover pathways. Biological scientists use data mining in bioinformatics in several ways: data cleaning, preprocessing, similarity search, association analysis, frequent pattern based cluster analysis, pathway analysis and visualization [59].

Data Mining Functions on GWAS

The size of genetic data increasing each day, but scientists need to expend more effort to filter meaningful knowledge from this data stack [60]. There are several additional approaches that can evaluate some problems of GWAS. Reclassifying test subjects into more homogeneous subgroups, for instance, endophenotypes, can decrease phenotypic heterogeneity and increment energy to identify genuine associations. The studies that are based on genes, which consider the relationship between an attribute and all markers inside a quality as opposed to every marker exclusively, can be more effective than conventional individual-SNP-based GWAS. Data mining strategies are accessible to investigating the high dimensional information created by GWAS of the complex psychiatric disorders. By utilizing data mining procedures conceivable to removing the unpredictable connections and relationships covered up in extensive data sets. This procedure additionally incorporates computer-based modeling of learning procedures and the revelation of new facts through observations and experimentations. There are distinctive calculations for completing data mining, and the accuracy of prediction of these calculations may fluctuate [61].

There are many analyzing methods that have been proposed to analyze micro array data and to extract biological knowledge. It's possible to detect suspected SNP's by univariate analysis of GWAS data, but we need more than a simple comparison between case and control groups. Some of the main reasons of the needs of complex analyses include: After GWAS it's possible to do targeted, exome, or even whole-genome sequencing in large cohorts. Data mining methodologies work on best prediction modeling, evaluate interaction and new solution to GWAS shortfalls.

Evaluation epistasis in GWAS data

One of the biggest challenges in genome-wide association studies is to evaluate SNP-SNP or gene-gene interactions. To characterize genetic structure of complex diseases we need to consider epistasis or gene-gene interaction. Epistasis has proven to be a complex genetic structure with classic statistical methods [62].

While logistic regression can be used to analyze such interactions, overfitting appears to be a significant issue.

1.3.2 Data Mining Processes

During the first decades of the 2000s, in parallel to the exponential growth in computational power and storage capacity, the amount of data acquired in disciplines such as finance, engineering, medicine and molecular biology also continued to increase. The need to find hidden patterns, relations, and rules within these high-dimensional data, necessitated analysis with a data mining approach. Data mining analysis is a multi-step process, initiated with enumerate data collection, continues with pre-processing of the data through cleaning, integration, transformation, and finalizes with data mining and knowledge presentation [60,63].

Data Preparation and Data Pre-processing

Data mining processes start with preparation and preprocessing, due to the nature of real world data. In the real world, data is incomplete, noisy and inconsistent. In a sense the raw data often lack the attributes of interest, containing errors or outliers and some discrepancies between different values. The data preprocessing step includes data cleaning, data integration, data transformation, data reduction and data discretization [64].

Data cleaning and imputation

The most important aspects of data cleaning is; handling missing values, identifying and addressing the noise in the data, and making corrections for the inconsistent data. There are many missing data handling methods published in the data mining literature. Data miners may ignore the tuple, they may apply the mean or median value, and use a predicted value for missing one's choices. Data mining software and packages have variety of solutions for this purpose [63].

To identify outliers and smooth out noisy data; binning, clustering, regression smoothing etc. methods may be used. Before smoothing corrections out in this way, inconsistent data should be interrogated physically wherever possible [63].

Data transformation

The data transformation step includes normalization, aggregation, generalization and attribute construction. The main concepts in data normalization is the scaling of the attributed values into a specified range and scaling raw values by using parameters such as mean and standard deviation. Attribute construction is performed by replacing or adding new attributes [63,65].

Data reduction:

Data reduction techniques are applied to obtain a smaller set of data, which are more manageable during data handling, analysis and extracting new rules. Strategies for data reduction include;

- Reducing the number of attributes; data cube aggregation (roll up, slice or dice), removing irrelevant attributes (filtering, wrapping), principle component analysis.
- Reducing dimension; encoding may help reduce data set size.
- Reducing numerosity and discretization generation [63].

Choice of Data Mining Method:

Determination of the correct data mining approach depends on the main goal of study, and the features of the data analyzed. Even though there are many ways to classify available data mining approaches the most common classification is based on the existence of an outcome variable; and classifies methods as supervised or non-supervised learning [66].

Supervised Learning Methods

In supervised learning, researchers have a known outcome measure that labels class. According to the purpose of the research, data is pre-processed to define the best classes for each case. The aim of supervised learning algorithms is the grouping of entities to reduce the number of classes. Depending on the number of classes, methods are named as "binary classification", "multiclass classification" etc. In binary classification labels are divided into two groups, for example; "1" or "0", "case" or "control [59,63,67]. In the case of labels that have more than two classes, the procedure is referred to as "multiclass classification" [68].

In a broader perspective, the classification methods can be discussed under the following categories:

- Technique-centered classification methods: Analysis performed using numerous classes of techniques. Decision trees, rule-based methods, neural networks, support vector machine (SVM) methods, nearest neighbor methods, and probabilistic methods are the best known examples.
- Data-Type centered classification methods: Many different data types (texts, uncertain data, time series data) are created by different applications and need evaluation. Each of these data types needs the different techniques.
- Variations on classification analysis: Many variations on the standard classification (such as transfer learning, rare class learning, semi-supervised learning, active learning) exist. Different variations of classification can be used to improve the effectiveness (such as ensemble analysis) of classification [69].

Feature Selection Methods

All classification algorithms require a carefully managed feature selection as their initial step. In most cases data is collected for different purposes by non-experts, and a wide variety of features are collected. The irrelevant features within the data, which are not related with the outcome measures, often lead to poor modeling. While a single relevant variable may have a small impact on the performance of the model, many irrelevant features combined, may have a large and significant cumulative effect. Therefore, using well-chosen features at the training level is critical to building successful models [63,69].

The main feature selection methods are:

1. Filter Models: A brittle criterion on a single feature, or a subset of features, are used to evaluate their suitability for classification. In order to perform feature selection, different measures can be used. Gini index, entropy and Fisher's index are the most common measures.

Gini index: The Gini index of the discrete variable is as shown in the following equation. The G value ranges between 0 to 1. Smaller values indicate more discriminative features.

$$G = 1 - \sum_{i=1}^{k} P_i^2 \qquad \textit{(Equation 1)}$$

Entropy: The entropy of a variable measured in the following equation. Entropy has similar ranges and notation with Gini index.

$$E = -\sum_{i=1}^{k} p_i \ . log \ (p_i) \qquad \text{(Equation 2)}$$

Fisher's index: Fisher's index is simply a measure of the ratio of the between class probability to the within class probability

$$F = \frac{\sum_{j=1}^{k} Pj.(\mu j - \mu)^2}{\sum_{j=1}^{k} Pj.\sigma_j^2} \qquad \text{(Equation 3)}$$

[63,69]

2. Wrapper Models: The feature selection process is combined with the classification algorithm. Therefore, the feature selection process is sensitive to the classification method.

The most common classification methods are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks [69].

3. Probabilistic Methods: Probabilistic methods are the most fundamental classification methods, and use statistical inference to find the best classes. To define the best class, probabilistic classification algorithms calculate a posterior probability. Posterior probability is; the probability after observing the specific characteristics (such as genotype or phenotype) of the test instance. The prior probability is simply the fraction of training records or known classes[63].

To estimate posterior probabilities two methods can be used. First; the posterior probability is estimated by calculating the class-conditional probability. In this case, there is prior class separation and then application of Bayes' theorem in order to find the classes. The most well-known Bayes classifier is a generative model. The following equation can be used [69].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad \text{(Equation 4)}$$

- P(c|x) is the posterior probability of targeted class.
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability predictor
- P(x) is the prior probability of predictor.

The second probabilistic approach is the direct modeling of the posterior probability, in other words the discriminative model. In this way, a discriminative function maps an input features vector directly onto a class.

4. Decision Trees: Decision trees partition data in a hierarchical process. This partitioning at each level, is created with the use of a split criterion. The split criteria maybe a single condition of attributes or it may contain a condition of multiple attributes. The overall approach is, to maximize the discrimination and to split the training data among the different classes over different nodes. The discrimination among the distinctive classes, are amplified if there should be an occurrence of contrast among the diverse classes in a given node is maximized. A measure (such as the Gini index, entropy) is used in order to measure this difference [60,63,69].

5. Neural Networks: Neural networks simulate the human brains' biological systems. The neuron is the main computation unit of an artificial neural network. These units can consist of various types of designs by associations between them. The most fundamental design of the neural system is a perceptron, which contains an arrangement of input nodes and an output node. The units of outputs get a set of inputs from the input units. There are different input units, which are precisely equivalent to the dimensionality of the basic information. The data is thought to be numerical. Categorical data need transformation to binary mode, and therefore the inputs may be larger [63,69,70].

In the classification analysis, a data set is often divided into two before operation; training set and test set. The training sets are used to determine a model and refine our classification rules. The test sets are used to evaluate the success of our model.

Multifactor dimensionality reduction (MDR)

Multifactor dimensionality reduction (MDR), first described by Ritchie, et al. in 2001, is a nonparametric, model-free method, and is an alternative approach. To create a classification rule, MDR focuses on combinations of loci that may interact and utilize these combinations and reduces the full dimensionality of the data. It is assumed that, for complex traits, multiple factors play a simultaneous role [9]. MDR basically reduces multi-locus genotypes into high-risk or low-risk groups, based on the number of affected and control cases present in a group. The method is more powerful than logistic regression in testing high-order interactions, and has many possible variations, including generalized MDR [6,9,71,72]. Many statistical and data mining methods are suggested to elucidate gene-gene interactions in candidate gene studies or GWAS. Currently, the MDR approach and its modifications are growing rapidly [73].

Genotyping technology improves precision and dimensionality of data. Therefore, the large scale of genetic data, requires methods to build disease models, variable selection and control of false positive results simultaneously. MDR is a method that evaluates potential interactions by doing a dense examination activity of all variables and their combinations. In this way MDR collapses multi-locus genotype combinations into binary (low risk-high risk) categories. Today MDR is one of the most common data mining methods in genetic epidemiology, and in a wide range of simulations it has been very successful [9,62,74]. At the same time there are many real data applications that exist including psychiatric disorders [75,76].

However; MDR has number of limitations such as an exhaustive computational burden due to its combinatorial nature. Another important issue is over-fitting. Computation replications are not always a good solution to reduce false positive results in the data sets of a single study. To remove the potential false positives, an alternative model needs to fit sample data and predict disease status in the population. Due to its computational difficulties data scientists need alternatives to cross-validation, especially for MDR. A popular internal validation method of cross-validation is the three way split of sample data. Original data is split into a training set for model building, a testing set for refining and a validation set to assess validity of model [77]. The two-stage model-building procedure prioritizes the validation. Models from the training set, which is re-performed in the testing set, are considered for validation and this provides new evidence without the need to collect a new sample. While, three way splitting reduces the computational burden dramatically, it may affect the power of MDR and true-detection rates. In order to investigate such problems, Monte Carlo simulation studies can be performed [75,78].

The size and structure of a data set is very important for designing genetic, and genomic studies. The MDR can easily be used in case control and sibling pair studies. Appropriate data sets should include a binary dependent variable besides any number of genetic and environmental variables. The data is randomly shuffled during cross-validation and the order of individuals within the data set is irrelevant for case control studies. For family/sibling data or matched case control studies, the order of individuals is important. In this cases the pairs must be unbroken during cross-validation. Pedigree data can be more complicated and pedigrees could be converted to sibling pairs. In trio studies (each trio contains an index case, mother and fathers data) pseudo controls could be created [62].

The sample size requirement for MDR is controversial. To detect two locus interactions for a specific epistasis model, usually 400 individuals deliver enough power, for higher order interactions a larger sample size is required. Empirical estimates are used to determine the sample size instead of theoretical formulas. Studies that have smaller numbers of cases and controls than 50 shows decreased power in simulation studies [9].

Steps of MDR

The main steps of MDR is represented in Figure 1.2.

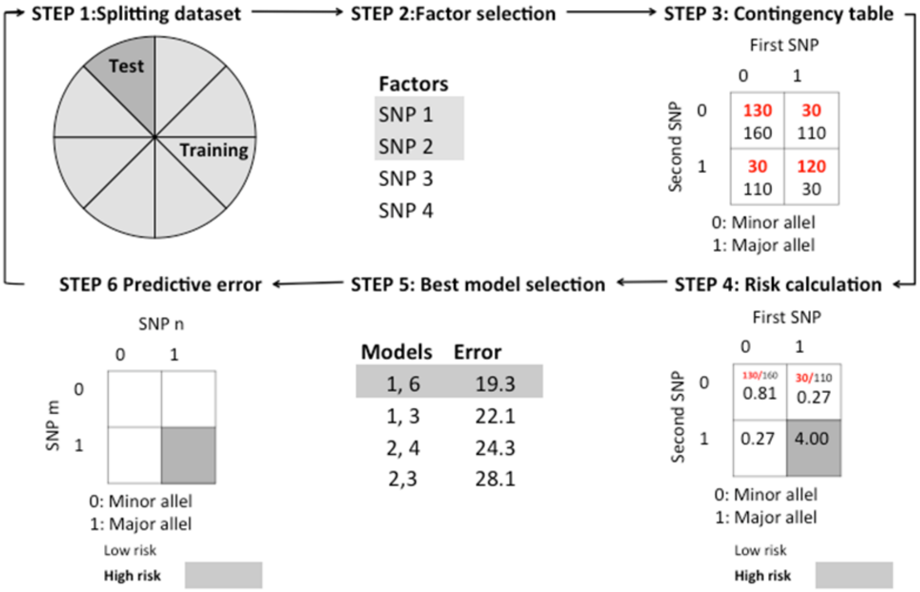

Figure 1-2. Steps of multifactor dimension reduction (adapted from [62,75,79])

Splitting dataset: The data is divided into training and test sets for cross validation. Although, cross-validation is not obligatory for MDR, it is performed to avoid over-fitting [79]. By cross-validating it is possible to find a good fit model in the given data and good predictions for future data. It eliminates the need for a second data set for testing, decreases the time consumed, and the cost. Ten to twenty five percent of data may be selected for testing. Tenfold cross validation means the training set comprises 90 percent of the data.

Factor selection: A set of n factors are selected from the all variables. Factors could be genetic or environmental.

Contingency table: All possible multifactor classes of these n factors are represented in n dimension space.

Risk calculation: After calculating cases/controls ratios, each multifactor cell, in n-dimensional space, label high risk and low risk groups. Dark-shaded cells interfere with high-risk genotype combinations and light-shaded cells interfere with low-risk genotype combinations.

Best model selection: The model that has lowest misclassification error is selected.

Calculation predictive error: The predictive error of the model is estimated in the test set.

Steps 1 to 6 are repeated for each possible pair [75,79].

For studies with more than two factors, all described steps are repeated for each possible model (two-factor, three-factor etc). But models for more than three factors are rarely computationally feasible. Statistical prediction of error is preferred in cases, where cross-validation consistency is high for one model and prediction error is low for another model. In the model selection step, the prediction error (calculated in test set) is used, not the classification error (calculated in the training set), in order to avoid over-fitting. If the number of SNPs evaluated increases, the classification error decreases [75,79].

Non-supervised Learning Methods:

Non-supervised learning is also referred to as clustering, as the class information is not known beforehand. The data information is unknown in non-supervised learning. Therefore, the problem of defining similar data points should be solved. In order to define clusters, the measures of similarities are calculated. In many studies, Euclidean or Manhattan distance, equations 5 and 6 respectively, can be used to measure the distance between numeric values. There are also other distance measures like Jaccard’s coefficients and correlation calculation is available for categorical variables (equations 7 and 8, respectively) (Table 1.2).

Euclidian distance: $D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}$ (Equation 5)

Manhattan distance: $D(X,Y) = \sum_{i=1}^{n}|x_i - y_i|$ (Equation 6)

Table 1-2. The contingency table for Jackard's coefficient and correlation calculation [63]

	Object j			
Object i		1	0	sum
	1	Q	r	q+r
	0	S	t	s+t
	sum	q+s	r+t	q+r+s+t

Jackard’s coefficient: $\frac{q}{q+r+s+t}$ (Equation 7)

Correlation: $\frac{q+t}{q+r+s+t}$ (Equation 8)

Clustering methods can be grouped under five categories [63]. These are:

- Partitioning Methods
- Hierarchical Methods

- Density Based Methods
- Grid Based Methods
- Model Based Methods

All methods have specific advantages and short falls, so the correct method should be chosen depending on the goal of the research, data and available tools [63].

Evaluation of model validity

All data mining models need evaluation before they can become useful applications. In order to determine the best model for a particular problem we need systematic evaluation, to observe how different methods work and to compare methods pairwise. We cannot asses how well different methods work only based on the evaluation of the training sets. The training set isn't a good indicator of performance on independent test sets. In case of limited data, sophisticated evaluation methods may be needed. Comparison of the performance of different data mining methods is not an easy task. In this step researchers have to ensure that the differences are not caused by chance, thus statistical tests are needed [65,70].

In the evaluation step, different methods are needed in each case. The cost of a misclassification error depends on the type of error calculated. Measuring the size of misclassification is essential during the evaluation of the model's performance.

Training and testing

The natural measure of a classifier's performance is "error rates". When classifier predicts a class to be the same as the previously known value, it is counted as a success, if not, it is an error. The error rate is measured as the overall performance of the classifier, it is the proportion of errors over a whole set of instances [59,80].

The error rate on the training data is calculated by re-substituting the training cases and is thereby named the re-substitution error. Although it is not a good predictor of the true error rate on test data, it is often useful to know [80].

To predict the genuine performance of a classifier on test data, data mining professionals need to survey their rates of error on a dataset. This dataset should not be included in the classifier. The test set may be this independent dataset. In classification algorithms, we assumed that both the test data and the training data are good samples of complete data. The training data is used by classifiers. The purpose of using test or validation data is to optimize classifiers. The test data is used to calculate the final error rate after generating a model and is used to optimize a method. Each set of data ought to be picked independently. The test data can be included in the training data to produce a new classifier just after the error rate has been calculated [63,80].

Cross-validation

In case the amount of data for training and testing is limited, then the holdout method does not work because it reserves a certain amount for testing and uses the remainder for training. Commonly to hold out one-third or one-fifth (depends on the size of the data) for testing and use the remaining two-thirds for training. Theoretically we cannot be fully sure whether a sample is representative or not. The most common method is: each class in the full dataset should be represented in roughly the right proportions in the training and testing sets.

A simple variant form of swapping test and training data is referred to as cross-validation. In cross-validation, miners decide on a fixed number of folds, or partitions of the data. Then the data is split into that number into approximately equal partitions and each in turn is used for testing and the remainder is used for training. If we use three parts, this is called threefold cross-validation [63,80,81].

Comparing data mining methods

Researchers often need to evaluate two or more different learning methods, and compare their performance. Simply the error rate is estimated using cross-validation or any other procedure, which is repeated several times. The model whose error rate estimate is smaller, is practically considered as the better model. However, the distinction can basically be brought on by the estimation of error, and in some cases, it is essential to figure out if one method is truly superior to another on a specific issue. In the event that another learning algorithm is proposed, its supporters must demonstrate that it enhances the cutting edge for the current issue and shows that the observed change does not happen due to an arbitrary possibility in the estimation procedure. A factual test that gives certainty limits, can be utilized as a part of this procedure. When attempting to anticipate true performance from a given test set error rate, if there is plenty of data, we could utilize an expansive sum for training and assess performance on a substantial free test set.

To evaluate validity of classification, the following formulas and descriptions that are shown in Table 1.3 can be used [82].

Table 1-3. Notations of a class comparison table: FN: False negative, FP: False positive, TN: True negative, TP: True positive AP: All Positives AN: All Negatives PP: Predicted Positive PN: Predictive Negative

	Predicted Positive	**Predicted Negative**	**Total**
Actual Positive	TP	FN	AP
Actual Negative	FP	TN	AN
Total	PP	PN	N

True Classification Rate (TCR) measures proportion of actual positives and negatives which are correctly identified.

$$TCR = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall (Sensitivity) measures the proportion of actual positives which are correctly identified.

$$Recall = \frac{TP}{TP+FN}$$

Precision (Positive Predictive Value) is the proportion of positive test results that are true positives.

$$Precision = \frac{TP}{TP+FP}$$

F-Measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure.

$$F = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall}$$

1.3.3 Shortfalls of data mining methods for genetic studies

Besides these benefits of data mining methods, we still have many shortfalls to using it. First of all, the need for more precise findings requires more computational power. Sometimes the permutative nature of mining algorithms make it impossible to perform a full model comparison. In this case scientists need devotion either precise or applicability. Often this leads to inconsistency between models. All different mining methods may detect different susceptible SNP's. Moreover changing the assumptions of models (number of repetition, stopping criteria etc.) may change terminal results [4].

Since each method has different advantages and disadvantages, the appropriate method selection is based on the problem being worked on, type of owned data, aim and design of the study. It's possible to use the literature to use a combination of methods to decrease the disadvantages and increase the advantages of the methods.

One of the biggest problems is missingness. In current technology, call rates of SNP arrays increased to 98-99 percent. Different data mining algorithms are working on data correction while chips evaluating and commercial software support these approaches

Overfitting problem in models

Overfitting, is also known as avoidance bias and it is kind of a search bias. But we should treat it separately because overfitting addresses a rather special problem. The problem is that if disconnection is allowed, useless concept descriptions that merely summarize the data become possible, whereas if it is banned, some concepts are un-learnable. The main reasons for overfitting are small number of entities especially used to compare to a number of used variables and noisy data. Very briefly, we can overcome the overfitting problem by;

- Regularization (for optimization-based classifiers).
- Increasing the size of training set
- Reducing the number of features [18,80,81,83].

CHAPTER 2

2 MATERIAL AND METHODS

2.1 DATA SOURCE

The study was conducted as case-control study. Data belongs to the Whole Genome Association Study of Bipolar Disorders (dbGaP Study Accession: phs000017.v3.p1). The goal of the project is to identify genes that make individuals more susceptible to bipolar disorders. All required permissions were approved by NIH. Data was analyzed in 2013 and all data was only used for the analysis and understanding the genetic basis of bipolar disorder.

The National Institute of Mental Health (NIMH) launched in 1989 a Genetics Initiative to collect family data for the linkage analysis of Alzheimer's disease. schizophrenia. and bipolar (BP) disorders. The NIMH BP Genetics Initiative is funded to create a national resource of demographics, clinical and diagnostic data and immortalized cell lines available to the scientific community. Such a resource will provide qualified investigators with DNA and clinical/diagnostic information necessary for the identification of multiple disease susceptibility loci that contribute to the etiology of BP disorders [49].

Data sets include both bipolar disorders patients and control cases genotype and phenotype data. The cases have bipolar and related disorders (BARD) cases. To filter the study group and eliminate some confounders the BARD cases are excluded from the study. Finally, data from 1767 controls and 653 Bipolar disorders only (BDO) groups was analyzed. The classification of bipolar disorders was adapted from the Diagnostic and Statistical Manual for Mental Disorders V (DSM V) and represented as Appendix B.

2.1.1 Genotype data

Data belongs to the Whole Genome Association Study of Bipolar Disorder (dbGaP Study Accession: phs000017.v3.p1). The goal of the project is to associate SNPs and identify genes that underlie the molecular basis of the bipolar disorder.

2.1.2 Phenotyping data

The comprehensive questionnaire, developed by the National Mental Health Genetics Initiative was used. The questionnaire consists of more than 1000 variables on the following topics (Table 2.1).

Table 2-1. Brief list of phenotyping data

Code	Phenotype
A	Demographics
B	Medical History
C	Modified Mini-mental Status Examination (If Applicable)
D	Somatization
E	Overview of Psychiatric Disturbance
F	Major Depression
G	Mania/Hypomania
H	Dysthymia
I	Cyclothymic Disorder
I	Alcohol Abuse and Dependence
J	Tobacco
J	Drug Abuse and Dependence
K	Psychosis
L	Schizotypal Personality Features
M	Modified Structured Interview for Schizotypy (SIS) (St. Louis site only)
N	Comorbidity Assessment
O	Suicidal Behavior/Violent Behavior/Self-Harm without Suicidal Intent
P	Anxiety Disorders
Q	Eating Disorders
R	Pathological Gambling
S	Antisocial Personality
T	Global Assessment Scale
U	Scale for the Assessment of Negative Symptoms (SANS)
V	Scale for the Assessment of Positive Symptoms (SAPS)
W	Modified SIS Ratings (St. Louis site only)
X	Interviewer’s Reliability Assessment
Y	Narrative Summary
Z	Medical Records Information

2.2 ANALYTICAL APPROACH

Analytical process includes three main activities: data preprocessing, univariate analysis and modeling. All analytic processes are described in the following sections and demonstrated in Figure 2.1.

2.2.1 Data Cleaning

This step includes data integration, cleaning, and transformation of the genotyping and the phenotyping data.

Data Integration: The huge size and confused phenotype data were integrated. All text files evaluated and the variables were matched, spending maximum effort to avoid loss of data. The drugs used and comorbidity data entered twice, both variables were evaluated, and integrated to gather the best data quality.

Data Selection and Cleaning: The data have plenty of missing and redundant values. Unfortunately, we don't have the raw data, and it's impossible to make corrections. The frequency analysis was performed to detect redundant and extreme values. The extreme values were cleaned.

To handle missing values: First of all, detected missing values (include missing value codes. such as "-999") were deleted. We didn't replace missing values for binary or nominal variables. Mplus was used to evaluate missing values.

Data cleaning activities performed by best suited R modules. MPlus was used to carry out missing data analysis. This common and flexible application can perform many different missing value replacements including expectation maximization, Bayesian techniques.

2.2.2 Analysis of the Genotyping Data

The data analysis flow-chart included 2 important pre-process steps. SNPs are filtered based on a minor allele frequency of <5% and failure of the Hardy–Weinberg Equilibrium (HWE) test as assessed by a P-value Sidak step-down adjusted p-values.

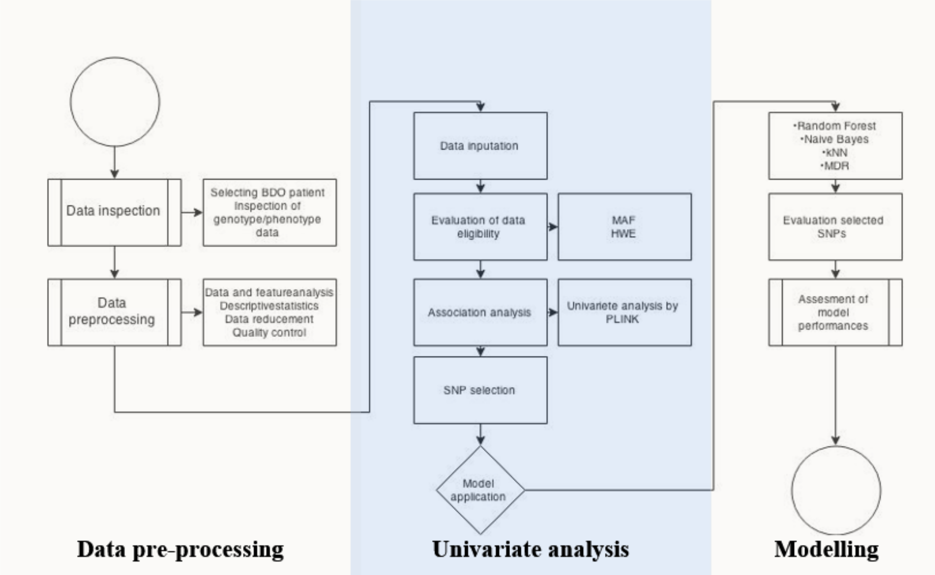

Figure 2-1. Data Analysis Flow-chart: MAF: Minor allele frequencies HWE: Hardy Weinberg Equilibrium kNN: k Nearest Neighborhood MDR: multifactor dimensionality reduction SNP: Single Nucleotide Polymorphism PLINK: Open-source whole genome association analysis toolset Version 1.8

We have used Affymetrix Gene Console and PLINK for the preprocessing steps (Filtering signals, background correction, data normalization) and quality check.

Univariate analysis is also performed by PLINK, and selected SNPs ($p<0.01$ considered threshold) listed as Appendix C.

2.2.3 Data mining Step and Model Building

Feature selection process applied in four steps.

Subset Generation: Subsets generated automatically by R Bioconductor package algorithms.

Subset Evaluation: Evaluation criteria can be broadly categorized into two groups; one is independent criteria, the other is dependent criteria. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures. In this study distance measures were used.

Stopping Criteria: Stopping criteria determine when the feature selection process should stop. For example; completing the search, given bounds (maximum number of iterations), subsequent addition of any feature does not produce a better subset. In random forest algorithm, we have used 1000 trees and at least 100 SNPs in every split of trees. Additionally, models were run selecting first 50, then 100 and 150 SNPs.

Result Validation: A classical approach for result validation is to directly measure the result using prior knowledge about the data. In real-world applications, however, we usually do not have such prior knowledge and we have to rely on some indirect methods by monitoring the change of mining performance with the change of features [67].

Randomly selected 80% of cases have been used for training the model and a complementary 20% percent used for evaluation.

2.2.4 Evaluation of method validity:

Before finalizing the common SNP list the analysis run 5 times. After all these efforts, model validation was evaluated by using the criteria described above. The comparisons performed were

- Comparison of models with default stopping criteria
- Comparison of models by selecting first 50, 100, 150 SNPs
- Comparison of performance.

2.3 DATA MINING ALGORITHMS

The analytical approach is depicted in Fig. 1. Genotyping and phenotyping data were first integrated, cleaned, and transformed. Data were then preprocessed in two key steps, using Affymetrix Gene ConsoleTM (Affymetrix Inc. USA) and

PLINK to filter signals, perform background correction, normalize data, and assess data quality. SNPs were filtered based on a minor allele frequency of < 5 % and failure of the Hardy-Weinberg Equilibrium test as indicated by a p-value of < 0.001. Univariate analysis was also performed by PLINK. Subsequently, genome-based probabilistic models of bipolar disorder were generated using random forests, naïve Bayesian, k-nearest neighbors, and MDR. Finally, candidate SNPs and genes were interpreted by annotation using SNPNexus and Regulomedb, combined p-value analysis to identify associated genes, and network analysis using GeneMANIA.

2.3.1 Random Forest (RF)

New methods have been developed to address the limitations of classical statistics in dealing with highly dimensional data. One of the most popular of these methods is random forests, an ensemble learning method broadly applicable in data mining and machine learning. The technique is nonparametric, tree-based, and combines the concept of nearest neighbors with bagging [11]. In this approach, one-step-at-a-time node splitting enables trees (and hence forests) to impose regularization and thereby effectively analyze data sets with “large p, small n”. In addition, grouping trees based on properties allows the method to deal with correlated and interacting variables [84]. The final model is a random forest of numerous decision trees. The most important advantage is that reduction in dimensionality is not required. The algorithm has been applied to classify and predict the effect of SNPs, and is significantly more successful than simple decision trees in analyzing gene expression[84–87].

2.3.2 Naïve Bayes (NB)

Naïve Bayes is another excellent method to analyze genomic data. It is one of the earliest machine-learning methods, and has been used for over 50 years in biomedical informatics. The technique is computationally efficient, and performs better than expected in classification tasks [88]. However, miscalibration can be an issue when numerous features are used, and it tends to make predictions with posterior probabilities too close to 0 or 1. Nevertheless, the technique has been successfully applied to diagnose diseases, identify news articles of interest, classify web pages by topic, and assign proteins to functional families [89–91]. In this study, we used the CRAN e1071 package for R to perform naïve Bayes modeling without double controlling Laplace smoothing. Thus, the epsilon range to apply Laplace smoothing, as well as the threshold for replacing cells with probabilities within the epsilon range, was not defined.

2.3.3 k-Nearest Neighborhood (kNN)

Unlike the other classification approaches, k-nearest neighbors does not build a classifier using training data. Instead, it searches for k data points closest to the test object, and uses the features of these neighbors to classify the new object. In instances where multiple classifications are possible, vote-counting is applied [16,17]. We used the kknn package for R from the CRAN repository, with number of neighbors set to k = 2. The minimum number of votes required for a definite decision was not defined, and the frequency of the majority class was returned as the attribute probability.

2.3.4 Multifactor dimensionality reduction (MDR):

MDR is a nonparametric, model-free method. To create a classification rule, MDR focuses on combinations of loci that may interact and utilizes these combinations and reduces the full dimensionality of the data. The main assumption in MDR is; multiple factors play a simultaneous role [9]. We evaluated potential novel SNPs role by MDR.

To perform MDR, we used the MDR package for R from the CRAN repository, and analyzed data with parameters K = 2, cv = 5, ratio = NULL, equal = "HR", and genotype = c(0, 1, 2). We assumed the number of MDR to be 1:100.

2.4 PATHWAY ANALYSIS

Pathway analysis is used to identify gene sets and biological pathways based on the information of selected genes to understand complex disease. Genome-wide association studies (GWAS) have considerably increased our knowledge of the genes involved in complex diseases. Most diseases that GWAS studies have disclosed strong single-gene effects. With present concepts requiring that genetic risk for complex diseases involves the cumulative effects of many genes. This approach allows us to expand of understanding of complex diseases from individual genetic associations to interactions between the effects of multiple genes [92,93].

Pathway analyses, test for association between sets of genes and a phenotype to explore the polygenic effects mentioned above. This may help extend the knowledge gained from GWAS. These methods test the cumulative effect across genes, it is possible to detect gene set level effects. Pathway analysis; mapping individual SNPs into gene sets and combined procedures improve the power to detect statistically significant associations [93,94].

In order to realize the potential of pathway analysis researchers require different methods for defining gene sets (pathways). The available techniques involve a wide variety of hypotheses about how genetics affects disease susceptibility, which significantly influences the results [54,93,94]. The basic steps performed in a pathway analysis are demonstrated in Figure 2.2.

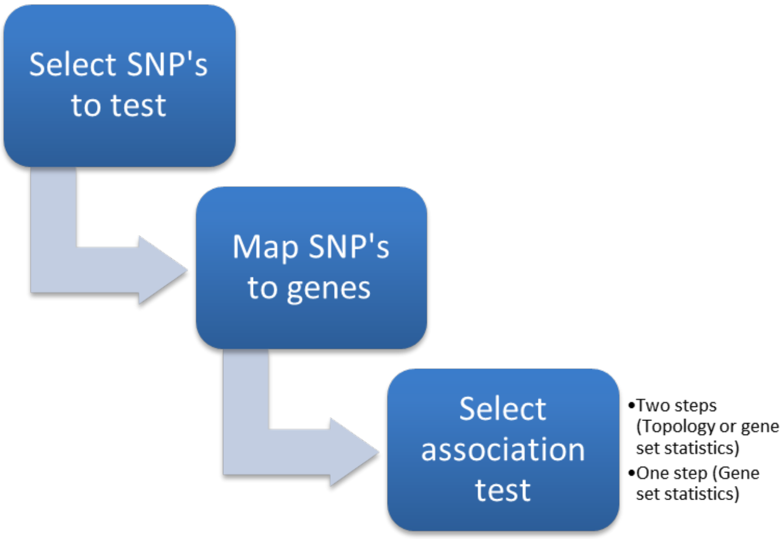

Figure 2-2. Basic steps of analysis

2.4.1 Software For Gene Pathway Analysis

The pathway analysis is key to understanding genomic studies. Advances in high throughput technologies increased genomic data. The data explosion of available data has developed opportunities for investigation of biological functions and gene-gene relations. This situation results in a need for computational tools for pathway analysis. Pathway analysis tools are used for gene ontology annotation, clustering, dimensionality reduction and of course visualization [95].

2.4.1.1 Cytoscape and GeneMANIA

The version 3.4.1 of GeneMANIA was installed into Cytoscape 3.4.0 to enable network editing [96,97]. As an open source software, the Cytoscape project was used for integrating biomolecular interactions. Cytoscape is applicable to any molecular components/interactions, and its most powerful use is for gene-gene and protein-protein interactions. The available number of interactions about human and other model organisms increased every day. The core of Cytoscape is extensible through a plug-in architecture, allowing rapid development of additional visualizations, computational analyses and features [97].

The GeneMANIA's Cytoscape plugin allows for the evaluation of interactions and t gene function prediction capabilities. GeneMANIA identifies the most related genes. The plugin uses over 800 networks from eight organisms (H. sapiens, A. thaliana Arabidopsis, C. elegance, D. melanogaster, M. musculus, S. cerevisiae, R. norvegicus and D. rerio) and indexing 2 152 association networks containing over 500 million interactions, mapped to 166 000 genes [98]. Each gene is traceable to the source network. It allows users to add their own interactions [96]. Association data include; protein and genetic interactions, pathways, co-expression, co-localization, protein domain similarity. GeneMANIA may be used to find new members of a pathway or complex and to find additional genes. Additionally it may help identify missed genes with a specific function [98]. The Java based GeneMANIA Cytoscape plugin is freely available at http://www.genemania.org/plugin/ address.

In our study the gene names found in the dbSNP databases were uploaded into GeneMANIA.

CHAPTER 3

3 RESULTS

The statistics analysis begins with 604 BDO patients and 1767 controls. Phenotype data is cleaned and analyzed. In this chapter the main characteristics of the study group, results of univariate analysis, then results of data mining models and their validity measures are reported. Finally, the pathway analysis results are presented.

3.1 MAIN DESCRIPTIVE STATISTICS

A total of 339 cases (56.1%) and 1081 of controls (61.2 %) were European Caucasian and others African American. 267 of cases (44.2) and 836 of the controls (47.3%) were male. Mean age of cases was 42.1 and controls was 49.9.

The main sociodemographic features of the study group were represented in Table 31.

After data cleaning steps:

- 1948 markers to be excluded based on HWE test (p <= 0.001)
- 882 markers failed HWE test in cases
- Total genotyping rate in remaining individuals is 0.97148
- 21597 SNPs failed missingness test (GENO > 0.1)
- 103715 SNPs failed frequency test (MAF < 0.05)
- After frequency and genotyping pruning, there are 761830 SNPs

After PLINK analysis 693 SNPs had a p-value < 0.001, and selected as candidate associations. The complete list of SNPs were represented as an attachment.

Total of 2371 cases were included for analysis, 604 of these were cases and 1767 were controls. The distribution of the cases into groups were reported in Table 3.1.

Table 3-1. Main descriptive statistics: BDO: Bipolar disorders only (cases) GRU: General research use (GRU) SD: Standard deviation EA: European American AA: African American

		Bipolar Disorders Only Group		**General Research Use Group**	
		Frequency	**Percent**	**Frequency**	**Percent**
		604	**34.2**	**1767**	**65.8**
Race	**EA**	339	56,1%	1081	61,2%
	AA	265	43,9%	686	38,8%
Smoking	**Yes**	171	30,3%	891	50,5%
	No	393	69,7%	875	49,5%
Previous daily smoking	**No**	119	23,6%	345	19,6%
	Yes	284	56,2%	839	47,6%
	Former	102	20,2%	580	32,9%
Sex	**Male**	267	44,2%	836	47,4%
	Female	337	55,8%	931	52,8%
Marital status.	**Married**	168	27,9%	846	61,0%
	Single	42	7,0%	241	17,4%
	Divorced	128	21,2%	170	12,3%
	Widowed	17	2,8%	100	7,2%
	Separated	248	41,1%	29	2,1%
Age	**Mean±SD**	**42.1±11,5**		**49.9±16,5**	

3.2 DATA MINING MODEL RESULTS

First, three different data mining methods (RF, NB, kNN) were compared on this data set with 4 performance comparison criteria. Interactions may increase validity of model. Increasing number of independent variables ordinarily increases model success and often leads to overfitting. On the other hand, some independent variables interact negatively. In this case adding new variables into the model decreased the validity measures of model. We performed additional models that include 50, 100 and 150 SNPs, but none of these allowed us to reach our optimal number sized models. RF have best recall results for each option, however NB have best precision values. Validation results of different models summarized in Table 3.2.

Table 3-2Validation results of different models that based on 50, 100, or 150 SNPs. Highest performance for each measure in different groups are labelled in bold.

Method	**Classification Accuracy (CA)**	**F-Measure**	**Precision**	**Recall**
50 SNPs				
RF	0,687	**0,847**	0,715	**0,987**
Naïve Bayes	0,6862	0,7424	**0,847**	0,702
kNN	**0,741**	0,817	0,7214	0,9247
100 SNPs				
RF	0,674	**0,826**	0,674	**0,935**
Naïve Bayes	0,634	0,754	**0,740**	0,65
kNN	**0,678**	0,804	0,657	0,876
150 SNPs				
RF	**0,67**	0,814	0,675	**0,924**
Naïve Bayes	0,624	0,724	**0,785**	0,648
kNN	0,647	**0,831**	0,724	0,8997

3.3 ANALYSIS OF SNP-SNP INTERACTIONS

MDR was used to investigate two-way or three-way SNP-SNP interactions, although three-way interactions were favored due to the large number of SNPs. The most significant two-way interactions were between rs17736182 and rs2055710, which map to the genes *KLHL1* and *DOCK10*, respectively. Patients with specific allelic profiles for these SNPs showed the highest risk (67.54 %) of having bipolar disorder. Analysis of three-way interactions identified a risk allele for rs2483023, a SNP in the *LEMD1* gene, along with two other unannotated SNPs. In 2-way interaction, the patients carrying allele A for rs17736182 were found to have an equal risk as the patients carrying allele A for rs2055710, which was 67.54%. In the 3-way interaction assessment, patients carrying the C allele for rs9372649, the A allele for rs12145634 and C allele for rs2483023 had a prediction performance of 77.2%. MDR generated a model with comparable predictive performance based on only five SNPs identified by analysis of two-way and three-way interactions. (Table 3.3).

Table 3-3. Performance comparison of classification based models vs MDR. Bolds are representing….

Method **Feature**		**RF**	**Naïve Bayes**	**kNN**	**MDR**	
					Two ways	**Three ways**
Validity	**Classification Accuracy**	**0.734**	0.702	0.733	0.647	0.721
	F-Measure	0.853	0.785	0.841	0.764	**0.861**
	Precision	0.743	**0.845**	0.754	0.675	0.772
	Recall	**0.998**	0.734	0.954	0.664	0.883

3.4 STRENGTH AND WEAKNESS OF DIFFERENT MODELS

All the 3 mining methods compared for best validity criteria and some other quality criteria in Table 3.4.

Table 3-4. Comparison of advantages and disadvantages of used models

Feature	RF	Naïve Bayes	kNN	MDR	
				Two ways	**Three ways**
Overfit	Very resistant since boot strap selection	Relatively risky	Boot strapping performed to avoid overfit	Risky	
Advantages	Non-parametric Interpretable Resistant to noise	Resistant to noise Good for eliminate missing values	Simple, flexible Arbitrary decision boundaries	Non-parametric test Flexible Evaluate interactions	
Disadvantages	Sensitive to inconsistent data	Accuracy degraded by correlated variables Non-deterministic	Sensitive to noise	Too slow High computation burden	

3.5 DATA MINING MODELS TO PREDICT T DISEASE SEVERITY

Same classification algorithms performed to predict General Assessment Score (GAS) and negative-positive symptom results as dependent variables. We used the Top 50 SNPs listed in Appendix A, with gender and age as independent variables. None of the modeling approaches found good classification accuracy both GAS and negative symptom existence. Validity of the results of this analysis are presented in Tables 3.5 and 3.6.

Table 3-5. Results of general assessment score (GAS) prediction

Method	Classification Accuracy	F-Measure	Precision	Recall
RF	0.589	0.711	0.590	0.750
Naïve Bayes	0.567	0.643	0.692	0.486
kNN	0.598	0.699	0.601	0.706

Table 3-6. Results of negative symptoms prediction

Method	Classification Accuracy	F-Measure	Precision	Recall
RF	0.535	0.574	0.455	0.623
Naïve Bayes	0.557	0.506	0.557	0.359
kNN	0.526	0.562	0.466	0.579

3.5 COMPARISON OF DIFFERENT MODELS

Random forests, naïve Bayes, and k-nearest neighbors have identified 16, 13, and 10 candidate SNPs, respectively. Surprisingly, the top six SNPs were common in all three (Table 1.2). Random forests and k-nearest neighbors were more successful than naïve Bayes, with recall values above 0.95.

Among 3 data mining methods and according to classification accuracy RF and kNN were more successful than Naïve Bayes, and recall values of both RF and kNN were above 0.95 (Table 3). The RF model selected 16, Naïve Bayes selected 13 and kNN selected 10 SNPs. MDR determined different SNPs (2 in two way interaction model and 3 in three way interaction model) (Table 3.7).

Table 3-7. SNPs identified in the genome-based model for RF, kNN and NB methods: RF: Random Forest, NB: Naïve Bayes, kNN: k-Nearest Neighbor, MDR: Multifactor Dimensionality Reduction

RS ID	**RF**	**kNN**	**NB**	**Multi Dimensionality Reduction**
rs6785	☑	☑	☑	
rs2194124	☑	☑	☑	
rs4792189	☑	☑	☑	
rs7569781	☑	☑	☑	
rs9375098	☑	☑	☑	
rs10415145	☑	☑	☑	
rs10857580	☑	☑	☑	
rs11015814	☑	☑	☑	
rs11015877	☑	☑	☑	
rs732183	☑	☑	☑	
rs11023096	☑	☑		
rs1328392	☑	☑		
rs2791142	☑	☑		
rs1861226	☑			
rs4654814	☑			
rs219506	☑			
rs2055710				☑
rs2483023				☑
rs9372649				☑
rs12145634				☑
rs17736182				☑

3.6 BIOLOGICAL PATHWAYS OF FOUND GENES

Five of detected SNP located on chromosome 2, while 4 on ch. 10, 4 on ch. 1 and 3 on ch.6. Annotation of determined SNPs (Include nucleotide strain, chromosomal positions, known mapped genes receipt from dbSNP and regulome scores found by regulomedb.org database) by models summarized in Table 3.6.

Table 3-8Annotation of associated SNPs [99]

RS ID	**SNP**	**Chromosome position**	**Gene**	**Gene name**	**Regulome score**
rs6785	ATAATAGCTGCTTTGTGTTCAGAAT[A/G]GTAGCAGTTGCTTTGTATATTAAAG	2:207603273	*CREB1* *METTL21A*	cAMP responsive element binding protein 1 Methyltransferase like 21A	1f
rs2194124	CTATTGTTACTCTGTCTGATACTGG[A/G]TAAAGCCAGACATTATGGGGACACA	4:77268903			No data
rs4792189	CTGTTGACATTTATTCCAGCCACCA[C/T]TGAATTACACAGCAGAGCCAGATGT	17:11870443	*DNAH9*	Dynein, axonemal, heavy chain 9	5
rs7569781	GAGCTCATGGTAATAAAGTTAAATA[C/G]CTCATCTAGAAGCAGGTACTCAGAT	2:224966869	*DOCK10*	Dedicator of cytokinesis 10	1f
rs9375098	TTAAATCTTTCACTGACTTTGATAG[C/T]TCATAAATACCCACATATTTCTATG	6:98033916			6
rs10415145	TTTTTGTGAAAATTGAAAGAGTTGG[C/G]CTTACAGTAACTTCCCCTCTGTGTA	19:32350147	*ZNF507*	Zinc finger protein 507	3a
rs10857580	CTGAGAGAACAACTTTCACTCTCAA[A/C]ATTATATTGACTCTTCCTTGTGAGG	10:48478341	*ARHGAP22*	Rho GTPase activating protein 22	6
rs11015814	CCTGATTTTCCTTGGGTTCCATGGA[G/T]ATGATTTATTCTAGTAAGTGATGCC	10:27485903			No data
rs11015877	ATTTAATATATTCAACCTGAGCTGT[A/T]GATAGAATTAAAAACTTATCAGCCA	10:27570308			5
rs732183	AGCGCAGGCTTAATGTTTGTTTTGT[A/G]AATTCAGGACATATATCTATGACTT	1:99224503			6
rs11023096	GCGCCAGCAGCGTGCTGTCACACCA[C/T]TTTTTGCCAGGTCCATAGCTTTGTT	11:2506773	*KCNQ1*	Potassium voltage-gated channel, KQT-like subfamily, member 1	4
rs1328392	GTGTCCCACAGCCTAGCCTTGCACT[A/G]TAGGAACAAAAACAACAACAAAAAA	6:152777092			5
rs2791142	ATGAAGTGTCTGCATTAAAAAAGAA[G/T]ACAAATCTCACATAAACAATCTAAC	1:163494893			6
rs186 6	AAAATACATTAAAAAAATGGAAATTC[A/G]TACATAGCTACATTTCATTTGTAGG	2:58425881			6
rs4654814	ATCACTGAGCAGCTCTCCTGAGAAA[C/T]ATCGACATGCGAGAAAATGTCCCAG	1:22767928	*EPHB2*	EPH receptor B2	5
rs219506	TTTTTTTTTTTTAGGTTACCTAACA[A/T]CATACCATTGCCCTGGTTATTTATT	2:21295746			6
rs2055710	GAGATGGTTAATTACTCCAAACAGC[A/G]ATGTCCTGGCTCATCTTTTTTTCATT	2:224901446	*DOCK10*	Dedicator of cytokinesis 10	4
rs17736182	ATATGCCCCATGACTAGCAAAAGGT[A/C]TGCCACAGAGTTGACATTAATGTAT	13:69857470	*KLHL1*	Kelch like family member 1	4
rs2483023	CAGTACAATTATTTACGGTTTTAGT[C/T]GTAAGTTCACTTAGGCTGCTAACAG	10:36550309			6
rs9372649	ACTGAACCTTTAAAGTGGCTGAATA[C/T]AGATTATTTTAGTCACATTTGTAAT	6:97947329			No data
rs12145634	TTTCAATGCCGCATTGTGTCAGGCA[A/G]ATATGAGGGCTGAGATTTGAAAGGA	1:205429436	*LEMD1*	LEM domain containing 1	4

Biological relations of found genes evaluated by GeneMANIA Networks.

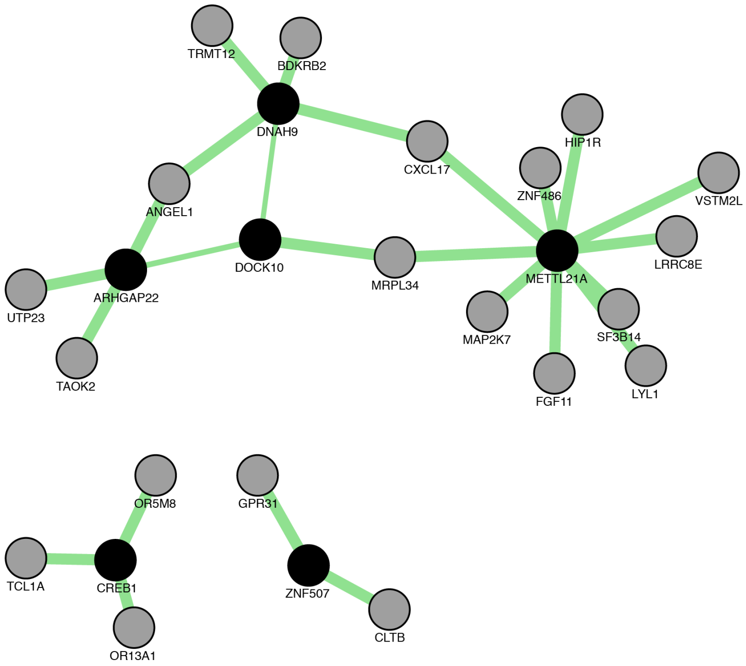

Figure 3-1. GeneMANIA Network of selected SNPs
Green lines: Genetic interactions

When shared SNP's were evaluated by the GeneMANIA network *METTL21A*, *DNAH9*, *DOCK10* and *ARHGAP22* genes took place in a joint network. Despite that *CREB1* and *ZNF507* have their own connections (Figure 3.1.). The most shared node, *METTL2A*, has genetic interaction between *LYL1* (lymphoblastic leukemia derived sequence 1), *LRRC8E* (leucine rich repeat containing 8 family, member E), *MAP2K7* (mitogen-activated protein kinase kinase 7), *HIP1R* (huntingtin interacting protein 1 related), *MRPL34* (mitochondrial ribosomal protein L34), *VSTM2L* (V-set and transmembrane domain containing 2 like), *ZNF486* (zinc finger protein 486), *CXCL17* (chemokine (C-X-C motif) ligand 17), *SF3B14* (Pre-mRNA branch site protein p14). *DNAH9* gene interacts with *ANGEL1* (angel homolog 1), *TAOK2* (*TAO* kinase 2) genes. Besides *DOCK10* and *DNAH9* genes from our list have genetic interactions with high weights. *ARHGAP22* interacts with *BDKRB2* (bradykinin receptor B2) gene too. Detailed GeneMANIA report about interactions of shared genes represented in Appendix D.

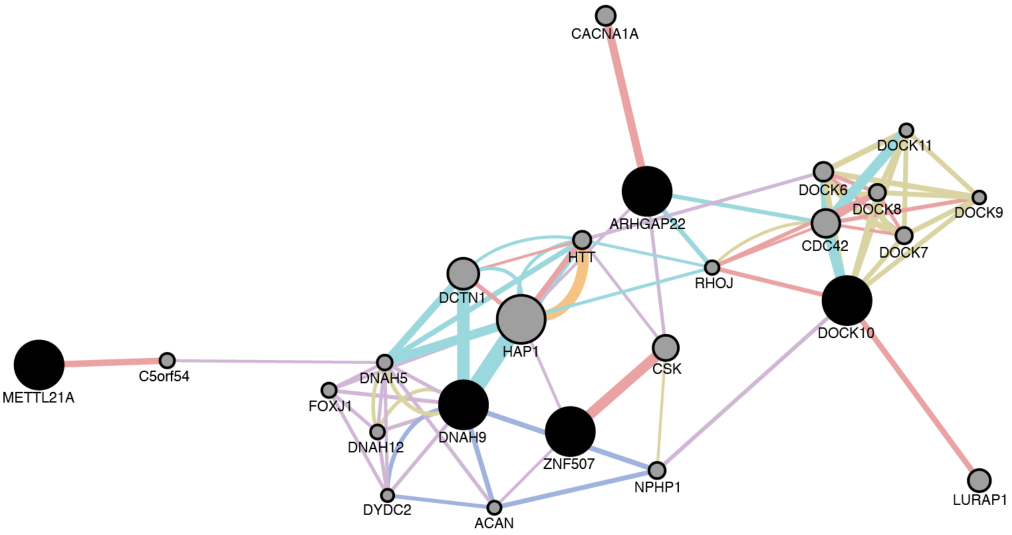

Figure 3-2. Refined GeneMANIA Network of selected SNPs: Red lines: physical interactions, purple lines: co-expressions, orange lines: predicted interactions, grey lines: co localization, green lines: pathways

Genes included in the network were re-evaluated by GeneMANIA, and a pathway chart created. First of all, the co-expression status of new novel genes was evaluated. *DNAH9* (dynein, axonemal, heavy chain 9) has relations with *DNAH5* (dynein, axonemal, heavy chain 5), *DNAH12* (dynein, axonemal, heavy chain 12), *DYDC2* (*DPY30* domain containing 2), *FOXJ1* (forkhead box J1). While *DOCK10* (dedicator of cytokinesis 10) has co-expressions with *NPHP1* (nephronophthisis 1), *ZNF507* have *ACAN* (agrecan) and *HAP1* (huntingtin-associated protein 1). Finally, *ARHGAP22* has co-expression with *CSK* (c-src tyrosine kinase) and *HAP1* (huntingtin-associated protein 1). The pathway relations evaluated, *ARHGAP22* related with *CDC42* (cell division cycle 42) and *RHOJ* (ras homolog family member J), *DNAH9* related with *DCTN1* (dynactin 1), *HAP1*, *HTT* (huntingtin), *DOCK10* related with *CDC42*. The GeneMANIA chart of physical interactions, co-expressions, predicted interactions, co localization and pathways shown on Figure 3.2. and a detailed report was presented in Appendix E.

CHAPTER 4

4 DISCUSSION

Variant calling and annotation is one way to search for SNPs associated with the disease, but this method is mainly biased towards annotated coding regions of the genome. However, hypothesis-free methods, such as presented here, do not rely on prior knowledge or genomic annotation, and therefore present a potential approach to overcome this bias. This study stands out from previous studies as three different classification methods and MDR (as an SNP–SNP interaction model) were compared for the first time in the literature on the same GWAS data, the dbGAP's bipolar disorders data. As a result, the advantages and disadvantages of each data mining method were assessed. Also, another novelty of the study is that all cases with bipolar-related disorders were filtered, and the data of only bipolar disorders were analyzed.

Data mining has great advantages for analyzing high-dimensional GWAS data of complex psychiatric disorders. In this way, it is possible to extract complex relationships and correlations hidden in large data sets. These processes may also include computer modeling of learning processes. In our study, we prioritize the most important SNPs by using three algorithms. The SNPs or mapped genes should be confirmed by prospective studies. In contrast to previous studies, we eliminated bipolar related disorder case data and exclusively analyzed bipolar disorders only patients' results. Given the complex disposition of bipolar disorders, this refinement helps to control confounders. In the end, the validity of the models was satisfactory. Limitations of this Study: The data set included genotyping and phenotype data. We attempted to integrate both genotyping and phenotyping data. However, due to the significant amount of work involved, we have not been able to find clues about treatment efficacy and adequate response to medical therapy. Due to the nature of retrospective data, it was impossible to describe the response criteria well.

In this study, we have analyzed BD GWAS data to show the association between the manifestation of disorders and genomic data. In addition to basic statistical data analysis, 3 data mining algorithms have been used to explore new potential SNPs. The common results of all 3 algorithms show that the SNPs with the top 5 priorities according to our data were rs7569781, rs2194124, rs9375098, rs4792189, and rs10857580. rs10415145 is located on the 19th chromosome at 19q13.11 and is related to *ZNF507* (Zinc Finger Protein 507). Even though there are no publications about *ZNF507* in bipolar disorders, *ZNF507* has been reported as a novel risk locus in autism spectrum disorders and neurodevelopmental disorders [100]. Additionally, a few publications discuss another zinc finger protein gene, *ZNF804A*. Bergman et al. have found a relation between *ZNF804A*, schizophrenia, and bipolar disorders. Both disorders affect cortical thickness [101]. Schwab et al. have reported a significant association between zinc finger proteins and psychotic disorders [102]. Li et al. have evaluated zinc finger proteins in a large meta- analysis based on a Han Chinese population and found that *ZNF804A* is related to the presence of psychotic disorders [103].

rs7569781 is located on the 2nd chromosome at location 2q36.2, related to the *DOCK10* (dedicator of cytokinesis 10) gene. This gene encodes many of the dedicator of cytokinesis proteins. There are some articles in the literature about the *DOCK10* gene reporting its relation with some metabolic and hematologic conditions, but there are no reports of an association with psychiatric disorders. A relation between DOCK series genes and bipolar disorders has been established in some publications [104,105]. Some evidence has linked Dock series with neuropsychiatric and neurodegenerative disorders, including autism spectrum disorders (*DOCK10*), schizophrenia (*DOCK4*), and Alzheimer (*DOCK2*) and Parkinson (*DOCK5*) diseases [106–109].

rs10857580 is located on the 10th chromosome, related to ARHGAP22 (Rho GTPase activating protein 22). Rho family small GTPases are described as key regulators of morphological changes in neurons. They are involved in axon and dendrite outgrowth through cytoskeletal reorganization. Kuramoto et al. have described their important roles in both neurological and psychiatric disorders [104].

rs11023096 is located on the 11th chromosome, related to the *KCNQ1* (potassium voltage-gated channel, *KQT*-like subfamily, member 1) gene. Potassium voltage-gated channels and related genes are known to be associated with bipolar disorders. The first gene suspected to be related to bipolar disorders is *ANK3*; other popular examples such as *KCNQ2*, *KCNQ3* are also members of this group. Judy et al. reported that they have implicated *ANK3* as a susceptibility gene for bipolar disorders. When they tested statistical interactions, the most significant interaction in the discovery GWAS was between SNPs in *ANK3* and *KCNQ2* [110] .

rs4654814 is located on the 1st chromosome and related to *EPHB2* (EPH receptor B2) genes. EphB receptors and ephrinB ligands transduce bidirectional signals. This mediator produces contact-dependent axon guidance primarily by promoting growth cone repulsion [111]. These functions were closely related with the central neuronal system and its mediator. According to literature *EPHB2* mutation increases autism spectrum disorders and schizophrenia [107,108,112].

rs4792189 is located on the 17th chromosome and related to *DNAH9* (dynein, axonemal, heavy chain 9). While there is some research about the relation between dynein, axonemal, heavy chain and psychological issues such as alcoholism or intellectual disabilities, there is no evidence regarding bipolar disorders [113,114].

When the genes with interactions to our novel genes are evaluated by GeneMANIA tool, we found that many of them have functional relations with neuro-psychiatric disorders.

LRRC8A is a core component of *VSOR* (volume-sensitive outwardly rectifying) anion channel in human cells. It plays essential roles in neuronal cell survival and death. Recent findings have suggested that *LRRC8A* were genotyped and tested for association with Parkinson's disease [115,116].

MAP2K7 (mitogen-activated protein kinase kinase 7)and *LRRC8E* (leucine-rich repeat containing 8 family, member E) genes have been implicated in mental disorders [115]. In a GWAS study, *CCDC62*/HIP1R found associated with Parkinson's disease besides another 10 genes (*SNCA, STK39, MAPT, GPNMB, SYT11, GAK, STX1B, MCCC1/LAMP3, ACMSD*, and *FGF20*) [117].

VSTM2L co-localizes with huminin in distinct brain areas. It plays a role in the neuronal viability modulation. *VSTM2L* acts as a strong antagonist of neuroprotective activity [118]. *TAOK2* like synaptic proteins and receptors is associated with autism spectrum disorders (ASDs) and their roles occurs by synaptic pathways in the pathogenesis of ASDs [119]. So, the related genes found by pathway analysis should include future validation studies.

The random forests model showed the best overall classification rate, as well as a more straight forward assessment of the classification errors. Indeed, extensive research using actual or simulated data has demonstrated decision trees to be very flexible and easy to debug. However, simple decision trees tend to overfit the data more than other techniques. Therefore, researchers generally prune trees and tune procedures to do so. The Random forest method was originally developed to overcome this issue and, and this study RF has generated the most accurate classification without overfitting. In our study, RF model also have the best recall results as a measure of repeatability, however NB model reported the best precision values. The performance of kNNs was comparable to all. NB's low classification rate with regard to tabular data simulation, can be optimized by application of the feature selection option to improve its performance [120].

Previous studies, which compare random forests with other classification methods, also supports our observation. For instance, Lunetta et al [121] conducted a simulation experiment to evaluate the ability of random forests to detect interacting SNPs and found that it outperformed Fisher's exact test, even though both methods were comparable in the absence of such interactions. One major concern about random forest is its high computational cost compared to the others discussed here. In addition, random forest has been reported to be very sensitive to noise or unbalanced data sets, whereas k-nearest neighbor algorithms are more efficient and stable [122]. Overall classification rates were between 0.70 and 0.75. Due to bootstrapping and the nature of the methods used, no model achieved classification rates better than 0.9.

In addition, the MDR method revealed different SNPs through analyses of two-way and three-way interactions. This may be due to the increased frequency of SNP interactions in polygenic diseases. Nevertheless, the classification success of MDR, based on only three SNPs, was comparable to other models. Notably, physical and functional annotation of the SNPs showed one SNP mapping to the *DOCK10* gene, which is also identified with the other three methods investigated here. The remaining two SNPs were mapped to the genes related to those found by other models, although these genes were not common to all. So, the high classification performance and relevant biological annotation of the SNPs discovered support that MDR would be an effective alternative method to evaluate SNP–SNP interactions. Also, the reduced number of SNPs, without loss in classification performance, would facilitate validation studies and decision support models, and would reduce the cost to develop predictive and diagnostic tests. Nevertheless, we acknowledge that translation of genomic models to the clinical setting will require models with higher classification performance [123].

Previous analyses of bipolar disorder genotypes revealed a high level of complexity, and a consensus profile of associated SNPs or genes has not been identified. This study stands out by directly comparing the power of MDR with three other classification-based methods to analyze the same existing genome-wide association data for bipolar disorder. As a result, we were able to assess the advantages and disadvantages of each. Six of the candidate SNPs detected were common to all classification-based methods. These SNPs identified two candidate genes that may potentially be the causative agents. Indeed, pathway analysis in GeneMANIA (http://GeneMANIA.org/) indicates that these genes are closely associated with psychiatric disorders [124]. Even though the classification performance is not sufficient for translation of the findings into a clinical diagnostic test, we suggest that the consensus SNP profile obtained from the three classification-based methods has high potential to be the causative variants, and further experimental validation would be productive. In contrast, MDR found different SNPs in analyses of two-way and three-way interactions. This may be due to the increased frequency of SNP interactions in polygenic diseases. Nevertheless, the classification success of MDR, based on only three SNPs, was comparable to other models. Notably, physical and functional annotation of these SNPs mapped one SNP to the DOCK10 gene that is also identified in the other three methods investigated here. The remaining two SNPs were mapped to genes related to those found by other models, although these genes were not common to all. In the end, the data indicates that MDR is an effective alternative method to evaluate SNP-SNP interactions.

The reduced number of SNPs, without degradation in classification performance, would facilitate validation studies and decision support models, and would reduce the cost to develop predictive and diagnostic tests. Nevertheless, we emphasize that translation of genomic models to the clinic will require models with higher classification performance.

CHAPTER 5

5 CONCLUSION

5.1 OVERVIEW

Many common disease or trait cluster in families, but a few of them have sufficient explanation of their genetic background. These diseases of traits are believed to be influenced by multiple genetic and environmental factors. Identification of genetic variants for these ‘complex diseases’ has been difficult. Genome-wide association studies (GWAS), evaluating more than a million SNPs are performed in thousands of subjects and represent a strong new tool for investigating the genetics of complex diseases. In the last two decades GWAS have identified hundreds of genetic variants associated with complex diseases. Although GWAS need high sample sizes, it has an important advantage over candidate gene studies, especially for complex traits. It is possible to evaluate many candidates at the same time thereby allowing for the understanding of pathways or interactions. Linkage studies have successes in single gene, Mendelian, disorders, but, in complex diseases they have low power.

Complex genetic diseases need more sophisticated research and analysis methods to understand and generate new medical approaches. In genomic modeling, various data mining techniques are proposed with varying success to analyze high-dimensional data generated by genome-wide association studies of complex genetic disorders.

The relations between genetic background and phenotype do not always display a linear association. Data mining methodologies that allow the finding of hidden variables, complex relationships and non-linear association becomes important in bioinformatics.

In this thesis, we aimed to compare Multifactor Dimensionality Reduction (MDR), a non-parametric approach that can be used to detect relevant interactions between SNPs or genes, with 3 other classification based data mining methods for genomic modeling of bipolar disorders.

5.2 ACCOMPLISHMENT

This thesis has accomplishments in both investigating genetic background of bipolar disorders, including genotype and phenotype interactions. Bipolar disorders have complex diagnostic features besides their complex genetic construction. To avoid bias factors, we eliminated all bipolar related disorder cases data and just analyzed bipolar disorders only patients' results. Finally, finding of this study have few cofounders than previous analysis of this data set.

As noted above, complex genetic diseases may have hidden variables, complex relationships and non-linear association. The original manuscript of the study analyzed data by using univariate and linear approaches. We performed 3 conventional data mining algorithms and MDR to evaluate non-linear associations. This made it possible to make new candidate SNPs. We found new SNPs, mapped them into related genes and drew pathway charts. These new findings can help explain the basis of the disease in the future.

Data mining proposed new models to understand bipolar disorders. Models that developed by this thesis have gratifying validity measures.

We performed MDR to search a SNP-SNP interaction both 2 way and 3 way. MDR may evaluate interactions using both genotype and phenotype features at the same time. The refined model of MDR has good validity and is very promising.

5.3 FUTURE STUDIES

Finally, it's obvious that different mining methods may find different candidate SNPs. Our studies identified various new candidate SNPs. Besides, we showed that models with interactions could define new alternatives. This alternative may help define new suspected areas and new pathways.

Different studies from the same database could find very distant or partially similar models. Researchers could focus on intersections or differences. Focus on intersections is scientifically reasonable to understand the genetic basis of disorders. Whilst, focusing on differences between models may help prevent the overlooking of crucial points.

It is possible to find refined models with reduced numbers of SNPs. This may help easy molecular diagnosis, but all models need clinical approval.

Different SNPs may map into the same genes. This means; some different models may point out the same genes. For this reason, pathway analysis is so important for post- GWAS analysis to support the overall results.

Experimental validation needed to support models and bioinformatics analysis results.

REFERENCES

[1] Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013;14: 315–26.

[2] Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;44: 841–7.

[3] Sumeet Dua PC. Data mining for bioinformatics [Internet]. CRC Press/Taylor & Francis Group; 2013.

[4] Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010;86: 6–22.

[5] He H, Oetting WS, Brott MJ, Basu S. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. BMC Med Genet. 2009;10: 127.

[6] Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genet Epidemiol. 2006;30: 111–23.

[7] Heidema AG, Feskens EJM, Doevendans PAFM, Ruven HJT, van Houwelingen HC, Mariman ECM, et al. Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. Genet Epidemiol. 2007;31: 910–21.

[8] Greene CS, Himmelstein DS, Nelson HH, Kelsey KT, Williams SM, Andrew AS, et al. Enabling personal genomics with an explicit test of epistasis. Pac Symp Biocomput. 2010; 327–36.

[9] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69: 138–47.

[10] Moore JH, Andrews PC. Epistasis analysis using multifactor dimensionality reduction. Methods Mol Biol. 2015;1253: 301–14.

[11] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99: 323–9.

[12] Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. J Am Stat Assoc. Taylor and Francis Ltd.; 2010;105: 205–217.

[13] Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet. Frontiers Media SA; 2015;6: 285.

[14] Sambo F, Trifoglio E, Di Camillo B, Toffolo GM, Cobelli C. Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data. BMC Bioinformatics. 2012;13 Suppl 1: S2.

[15] Li S, Harner EJ, Adjeroh DA. Random KNN feature selection - a fast and stable alternative to Random Forests. BMC Bioinformatics. 2011;12: 450.

[16] Deegalla S, Boström H. Classification of microarrays with knn: Comparison of dimensionality reduction methods.International Conference on Intelligent Data Engineering and Automated Learning, Berlin, 2007

[17] Gunavathi C, Premalatha K. Performance Analysis of Genetic Algorithm with kNN and SVM for Feature Selection in Tumor Classification. 2014;8: 1390–1397.

[18] Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmacogenomics J. Nature Publishing Group; 2010;10: 292–309.

[19] Yatham LN, Torres IJ, Malhi GS, Frangou S, Glahn DC, Bearden CE, et al. The International Society for Bipolar Disorders-Battery for Assessment of Neurocognition (ISBD-BANC). Bipolar Disord. 2010;12: 351–63.

[20] Huxley N, Baldessarini RJ. Disability and its treatment in bipolar disorder patients. Bipolar Disord. 9: 183–96.

[21] NIMH · Bipolar Disorder [Internet]. [cited 23 Apr 2014]. Available: http://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml

[22] Diagnostic and statistical manual of mental disorders: DSM-IV-TR. Washington DC: APA Press; 2004.

[23] Judd LL, Akiskal HS, Schettler PJ, Endicott J, Maser J, Solomon DA, et al. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry. 2002;59: 530–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/12044195

[24] Merikangas KR, Akiskal HS, Angst J, Greenberg PE, Hirschfeld RMA, Petukhova M, et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. Arch Gen Psychiatry. 2007;64: 543–52.

[25] Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res. 2004;13: 93–121.

[26] Stein DJ, Phillips KA, Bolton D, Fulford KWM, Sadler JZ, Kendler KS. What is a mental/psychiatric disorder? From DSM-IV to DSM-V. Psychol Med. Cambridge University Press; 2010;40: 1759–65.

[27] Weissman, Myrna M., et al. "Cross-national epidemiology of major depression and bipolar disorder." Jama 276.4 (1996): 293-299.

[28] Diagnostic and Statistical Manual of Mental Disorders (DSM 5). Washington DC: APA Press; 2013.

[29] Global Burden of Disease Study 2013 Collaborators T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet (London, England). Elsevier; 2015;386: 743–800.

[30] Kurasaki K, Sumie O, Sue. S. Asian American mental health: Assessment theories and methods. New York, New York, USA: Springer Science & Business Media; 2002.

[31] Üstün TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJL. Global burden of depressive disorders in the year 2000. Br J Psychiatry. 2004;184.

[32] Goodwin F, Jamison K. Manic-Depressive Illness. Bipolar Disorders and Recurrent Depression. Oxford University Press,Oxford, 2007.

[33] Tohen M, Jr CZ, Hennen J. The McLean-Harvard first-episode mania study: prediction of recovery and first recurrence. Am J. 2003.

[34] Novick D, Swartz H, Frank E. Suicide attempts in bipolar I and bipolar II disorder: a review and meta-analysis of the evidence. Bipolar Disord. 2010.

[35] Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet (London, England). Elsevier; 2013;382: 1575–86.

[36] Hidalgo-Mazzei D, Undurraga J, Reinares M, Bonnín C del M, Sáez C, Mur M, et al. The real world cost and health resource utilization associated to manic episodes: The MANACOR study. Rev Psiquiatr y salud Ment. 8: 55–64.

[37] Soldani F, Sullivan P, Pedersen N. Mania in the Swedish Twin Registry: criterion validity and prevalence. Aust New Zeal. 2005.

[38] Vieta E, Blasco-Colmenares E, Figueira ML, Langosch JM, Moreno-Manzanaro M, Medina E. Clinical management and burden of bipolar disorder: a multinational longitudinal study (WAVE-bd study). BMC Psychiatry. 2011;11: 58.

[39] Maji KR, Sood M, Sagar R, Khandelwal SK. A follow-up study of family burden in patients with bipolar affective disorder. Int J Soc Psychiatry. 2012;58: 217–23.

[40] Berry EA, Heaton PT, Kelton CML. National estimates of the inpatient burden of pediatric bipolar disorder in the United States. J Ment Health Policy Econ. 2011;14: 115–23.

[41] Hirschfeld RMA, Vornik LA. Bipolar disorder--costs and comorbidity. Am J Manag Care. 2005;11: S85-90.

[42] Schmitt A, Malchow B, Hasan A, Falkai P, Mitterauer BJ. The impact of environmental factors in severe psychiatric disorders.New York, 2014.

[43] Valenzuela FJ, Vera J, Venegas C, Muñoz S, OValenzuela, F. J., et al. "Evidences of polymorphism associated with circadian system and risk of pathologies: a review of the literature." International Journal of Endocrinology, 2016.

[44] Johnson, Sheri L., Amy K. Cuellar, and Anda Gershon. "The influence of trauma, life events, and social relationships on bipolar depression." Psychiatric Clinics of North America 39.1 (2016): 87-94.

[45] Kieseppä T, Partonen T, Haukka J, Kaprio J, Lönnqvist J. High concordance of bipolar I disorder in a nationwide sample of twins. Am J Psychiatry. 2004;161: 1814–21.

[46] McGuffin P, Rijsdijk F, Andrew M, Sham P, Katz R, Cardno A. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. Arch Gen Psychiatry. 2003;60: 497–502.

[47] Barnett JH, Smoller JW. The genetics of bipolar disorder. Neuroscience. 2009;164: 331–43.

[48] Craddock N, Sklar P. Genetics of bipolar disorder. Lancet. 2013;381: 1654–62.

[49] Szczepankiewicz A. Evidence for single nucleotide polymorphisms and their association with bipolar disorder. Neuropsychiatr Dis Treat. 2013;9: 1573–82.

[50] Georgiev D, González-Burgos G, Kikuchi M, Minabe Y, Lewis DA, Hashimoto T. Selective expression of KCNS3 potassium channel α-subunit in parvalbumin-containing GABA neurons in the human prefrontal cortex. PLoS One. 2012;7: e43904.

[51] Bengesser, Susanne; Reininghaus EF am M: PARL 2013, Craddock N, Sklar P. Genetics of Bipolar Disorder [Internet]. Lancet. PL Academic Research; 2013.

[52] Sebastiani P, Timofeev N, Dworkis DA, Perls TT, Steinberg MH. Genome-wide association studies and the genetic dissection of complex traits. Am J Hematol. Wiley Subscription Services, Inc., A Wiley Company; 2009;84: 504–515.

[53] Stein C, Elston R. Finding genes underlying human disease. Clin Genet. Blackwell Publishing Ltd; 2009;75: 101–106.

[54] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9: 356–69.

[55] Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446–50.

[56] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002;4: 45–61.

[57] Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, et al. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics. 2012;13: 164.

[58] Tosto G, Reitz C. Genome-wide association studies in Alzheimer's disease: a review. Curr Neurol Neurosci Rep. 2013;13: 381.

[59] Han J., Kamber M. Data mining: concepts and techniques. San Francisco, 2006

[60] Wang J, Zaki M, Toivonen H, Shasha D. Data Mining in Bioinformatics. London: Springer-Verlag; 2005.

[61] Pirooznia M, Seifuddin F, Judy J, Mahon PB, Potash JB, Zandi PP. Data mining approaches for genome-wide association of mood disorders. Psychiatr Genet. 2012;22: 55–61.

[62] Motsinger AA, Ritchie MD. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene - gene interactions in human genetics and pharmacogenomics studies. Hum Genomics 2006 25. BioMed Central; 2006;2: 318.

[63] Han J. "How can data mining help bio-data analysis?[extended abstract]." Proceedings of the 2nd International Conference on Data Mining in Bioinformatics. Springer-Verlag, 2002.

[64] Schwarz DF, Szymczak S, Ziegler A, König IR. Picking single-nucleotide polymorphisms in forests. BMC Proc. 2007;1 Suppl 1: S59.

[65] Liao S-H, Chu P-H, Hsiao P-Y. Data mining techniques and applications – A decade review from 2000 to 2011. Expert Syst Appl. Elsevier Ltd; 2012;39: 11303–11311.

[66] Gilbert, K., Sanchez-Marre, M., & Codina, V. (2010). Choosing the right data mining technique: Classification of methods and intelligent recommendation. In The proceedings of the 2010 international congress on, environmental modelling and software (pp. 1933–1940).

[67] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. … Data Eng IEEE Trans. 2005; Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1401889

[68] Har-Peled S, Roth D, Zimak D. Constraint classification: A new approach to multiclass classification. Int Conf Algorithmic. 2002.

[69] Charu C. Aggarwal. Data Classification Algorithms and Applications. Florida: Taylor & Francis Group; 2015.

[70] Wang JTL, Zaki MJ, Toivonen HTT, Shasha D. "Introduction to data mining in bioinformatics." Data Mining in Bioinformatics. Springer London, 2005. 3-8.

[71] Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics. 2007;23: 71–6.

[72] Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet. 2006;7: 23.

[73] Gola D, Mahachie John JM, van Steen K, König IR. "A roadmap to multifactor dimensionality reduction methods." Briefings in bioinformatics 17.2 (2016): 293-308.

[74] Moore JH, Gilbert JC, Tsai C-T, Chiang F-T, Holden T, Barney N, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol. 2006;241: 252–261.

[75] Winham SJ, Slater AJ, Motsinger-Reif AA. A comparison of internal validation techniques for multifactor dimensionality reduction. BMC Bioinformatics. BioMed Central; 2010;11: 394.

[76] Edwards TL, Wang X, Chen Q, Wormly B, Riley B, O'Neill FA, et al. Interaction between interleukin 3 and dystrobrevin-binding protein 1 in schizophrenia. Schizophr Res. 2008;106: 208–17.

[77] Winham S, Wang C, Motsinger-Reif AA. A comparison of multifactor dimensionality reduction and L1-penalized regression to identify gene-gene interactions in genetic association studies. Stat Appl Genet Mol Biol. Berkeley Electronic Press; 2011;10: Article 4.

[78] Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics. BioMed Central; 2008;9: 238.

[79] Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Gaziano JM, Ridker PM, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene Interactions on risk of myocardial infarction: The importance of model validation. BMC Bioinformatics. 2004;5.

[80] Witten IH, Frank E. Data Mining Practical Machine Learning Tools and Techniques. Elsevier Inc., editor. San Francisco; 2005.

[81] Motsinger AA, Ritchie MD. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol. 2006;30: 546–55.

[82] Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol. 2011, 12, 78-82

[83] Chen Y, Li Y, Cheng X, Guo L. Survey and taxonomy of feature selection algorithms in intrusion detection system. Inf Secur Cryptol. 2006

[84] Wang Y, Goh W, Wong L, Montana G. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC Bioinformatics. 2013;14 Suppl 1: S6.

[85] Khoshgoftaar TM, Golawala M, Hulse J Van. An Empirical Study of Learning from Imbalanced Data Using Random Forest. 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007). IEEE; 2007. pp. 310–317.

[86] Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9: 307.

[87] Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. 2010;11: 49.

[88] Zhang H. The optimality of naive Bayes. AA. 2004; Available: http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf

[89] Sambo F, Trifoglio E, Di Camillo B, Toffolo GM, Cobelli C. Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data. BMC Bioinformatics. BioMed Central; 2012; S2.

[90] Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. J Am Med Inform Assoc. 18: 370–5.

[91] Malovini A, Barbarini N, Bellazzi R, de Michelis F. Hierarchical Naive Bayes for genetic association studies. BMC Bioinformatics. 2012;13 Suppl 1: S6.

[92] Hirschhorn, Joel N. "Genomewide association studies--illuminating biologic pathways." New England Journal of Medicine 360.17 (2009): 1699.

[93] Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends Genet. NIH Public Access; 2014;30: 390–400.

[94] Holmans, Peter. "7 Statistical Methods for Pathway Analysis of Genome-Wide Data for Association with Complex Genetic Traits." Advances in genetics 72 (2010): 141.

[95] Tsui IFL, Chari R, Buys TPH, Lam WL. Public databases and software for the pathway analysis of cancer genomes. Cancer Inform. Libertas Academica; 2007;3: 379–97.

[96] Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinforma Appl NOTE. 2010;26: 2927–2928.

[97] Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research 13.11 (2003): 2498-2504.

[98] GeneMANIA Help Page. [cited 17 Nov 2016]. Available: http://pages.genemania.org/help/faq

[99] NCBI Gene Search Database. 2016 [cited 17 Nov 2016]. Available: https://www.ncbi.nlm.nih.gov/gene

[100] Talkowski ME, Rosenfeld JA, Blumenthal I, Pillalamarri V, Chiang C, Heilbut A, et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell, 2012;149: 525–37.

[101] Bergmann O, Haukvik UK, Brown AA, Rimol LM, Hartberg CB, Athanasiu L, et al. ZNF804A and cortical thickness in schizophrenia and bipolar disorder. Psychiatry Res. 2013;212: 154–7.

[102] Schwab SG, Kusumawardhani AAAA, Dai N, Qin W, Wildenauer MDB, Agiananda F, et al. Association of rs1344706 in the ZNF804A gene with schizophrenia in a case/control sample from Indonesia. Schizophr Res. 2013;147: 46–52.

[103] Li M, Su B. Meta-analysis supports association of a non-synonymous SNP in ZNF804A with schizophrenia. Schizophr Res. 2013;149: 188–9.

[104] Kuramoto K, Negishi M, Katoh H. Regulation of dendrite growth by the Cdc42 activator Zizimin1/Dock9 in hippocampal neurons. J Neurosci Res. 2009;87: 1794–805.

[105] Detera-Wadleigh SD, Liu C, Maheshwari M, Cardona I, Corona W, Akula N, et al. Sequence variation in DOCK9 and heterogeneity in bipolar disorder. Psychiatr Genet. 2007;17: 274–86.

[106] Shi L. Dock protein family in brain development and neurological disease. Commun Integr Biol. Taylor & Francis; 2013;6: e26839.

[107] Wang T, Guo H, Xiong B, Stessman HAF, Wu H, Coe BP, et al. De novo genic mutations among a Chinese autism spectrum disorder cohort. Nat Commun. Nature Publishing Group; 2016;7: 13316.

[108] Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. Nature Research; 2012;485: 237–241.

[109] Nava C, Keren B, Mignot C, Rastetter A, Chantot-Bastaraud S, Faudet A, et al. Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders. Eur J Hum Genet. Nature Publishing Group; 2014;22: 71–8.

[110] Judy JT, Seifuddin F, Pirooznia M, Mahon PB, Jancic D, Goes FS, et al. Converging Evidence for Epistasis between ANK3 and Potassium Channel Gene KCNQ2 in Bipolar Disorder. Front Genet. 2013;4: 87.

[111] Srivastava N, Robichaux MA, Chenaux G, Henkemeyer M, Cowan CW. EphB2 receptor forward signaling controls cortical growth cone collapse via Nck and Pak. Mol Cell Neurosci. 2013;52: 106–16.

[112] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. Nature Research; 2012;488: 471–475.

[113] Gamsiz ED, Viscidi EW, Frederick AM, Nagpal S, Sanders SJ, Murtha MT, et al. Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am J Hum Genet. 2013;93: 103–9.

[114] Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L, et al. Genome-wide association study of alcohol dependence implicates a region on chromosome 11. Alcohol Clin Exp Res. 2010;34: 840–52.

[115] Gregersen NO, Buttenschøn HN, Hedemand A, Nielsen MN, Dahl HA, Kristensen AS, et al. Association between genes on chromosome 19p13.2 and panic disorder. Psychiatr Genet. 2016;26: 287–292.

[116] Okada T, Islam MR, Tsiferova NA, Okada Y, Sabirov RZ. Specific and essential but not sufficient roles of LRRC8A in the activity of volume-sensitive outwardly rectifying anion channel (VSOR). Channels. 2016; 1–12.

[117] Pihlstrøm L, Axelsson G, Bjørnarå KA, Dizdar N, Fardell C, Forsgren L, et al. Supportive evidence for 11 loci from genome-wide association studies in Parkinson's disease. Neurobiol Aging. 2013;34: 1708.e7-1708.e13.

[118] Rossini L, Hashimoto Y, Suzuki H, Kurita M, Gianfriddo M, Scali C, et al. VSTM2L is a novel secreted antagonist of the neuroprotective peptide Humanin. FASEB J. 2011;25: 1983–2000.

[119] Chen J, Yu S, Fu Y, Li X. Synaptic proteins and receptors defects in autism spectrum disorders. Front Cell Neurosci. 2014;8: 276.

[120] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–17.

[121] Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 2004;5: 32.

[122] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. Nature Publishing Group; 2009;10: 392–404.

[123] Acikel C, Aydin Son Y, Celik C, Gul H. Evaluation of novel candidate variations and their interactions related to bipolar disorders: Analysis of GWAS data. Neuropsychiatr Dis Treat. Dove Press; 2016;Volume 12: 2997–3004.

[124] Acikel CH, Son YA, Celik C, Tutuncu R. Evaluation of Whole Genome Association Study Data in Bipolar Disorders: Potential Novel SNPs and Genes. Bull Clin Psychopharmacol. 2015;25: 12–18.

APPENDICES

APPENDIX A: COMPLETE LIST OF PREVIOUSLY DETERMINED GENES AND RELATIONS ARRANGED BY CHROMOSOMES POSITIONS [48,51]

GeneID	*Symbol*	Description	Map_location
4524	*MTHFR*	methylenetetrahydrofolate reductase	1p36.22
2212	*FCGR2A*	Fc fragment of IgG receptor IIa	1q23.3
4803	*NGF*	nerve growth factor	1p13.2
2214	*FCGR3A*	Fc fragment of IgG receptor IIIa	1q23.3
1401	*CRP*	C-reactive protein	1q23.2
1806	*DPYD*	dihydropyrimidine dehydrogenase	1p21.3
4774	*NFIA*	nuclear factor I A	1p31.3
2944	*GSTM1*	glutathione S-transferase mu 1	1p13.3
2475	*MTOR*	mechanistic target of rapamycin	1p36.22
58155	*PTBP2*	polypyrimidine tract binding protein 2	1p21.3
27185	*DISC1*	disrupted in schizophrenia 1	1q42.2
178	*AGL*	amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase	1p21.2
5737	*PTGFR*	prostaglandin F receptor	1p31.1
114548	*NLRP3*	NLR family pyrin domain containing 3	1q44
7133	*TNFRSF1B*	TNF receptor superfamily member 1B	1p36.22
3570	*IL6R*	interleukin 6 receptor	1q21.3
3310	*HSPA6*	heat shock protein family A (Hsp70) member 6	1q23.3
8863	*PER3*	period circadian clock 3	1p36.23
10076	*PTPRU*	protein tyrosine phosphatase, receptor type U	1p35.3
5999	*RGS4*	regulator of G-protein signaling 4	1q23.3
2170	*FABP3*	fatty acid binding protein 3	1p35.2

10747	*MASP2*	mannan binding lectin serine peptidase 2	1p36.22
359948	*IRF2BP2*	interferon regulatory factor 2 binding protein 2	1q42.3
9722	*NOS1AP*	nitric oxide synthase 1 adaptor protein	1q23.3
3782	*KCNN3*	potassium calcium-activated channel subfamily N member 3	1q21.3
149465	*CFAP57*	cilia and flagella associated protein 57	1p34.2
25896	*INTS7*	integrator complex subunit 7	1q32.3
100132074	*FOXO6*	forkhead box O6	1p34.2
388650	*FAM69A*	family with sequence similarity 69 member A	1p22.1
10964	*IFI44L*	interferon induced protein 44 like	1p31.1
553	*AVPR1B*	arginine vasopressin receptor 1B	1q32.1
127294	*MYOM3*	myomesin 3	1p36.11
607	*BCL9*	B-cell CLL/lymphoma 9	1q21.2
7257	*TSNAX*	translin associated factor X	1q42.2
4009	*LMX1A*	LIM homeobox transcription factor 1 alpha	1q23.3
391059	*FRRS1*	ferric chelate reductase 1	1p21.2
22854	*NTNG1*	netrin G1	1p13.3
127255	*LRRIQ3*	leucine rich repeats and IQ motif containing 3	1p31.1
8564	*KMO*	kynurenine 3-monooxygenase	1q43
57554	*LRRC7*	leucine rich repeat containing 7	1p31.1
126638	*RPTN*	repetin	1q21.3
728448	*PPIEL*	peptidylprolyl isomerase E like pseudogene	1p34.3
338	*APOB*	apolipoprotein B	2p24.1
3553	*IL1B*	interleukin 1 beta	2q14.1
1493	*CTLA4*	cytotoxic T-lymphocyte associated protein 4	2q33.2
3557	*IL1RN*	interleukin 1 receptor antagonist	2q14.1
11320	*MGAT4A*	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme A	2q11.2
1385	*CREB1*	cAMP responsive element binding protein 1	2q33.3
81562	*LMAN2L*	lectin, mannose binding 2 like	2q11.2
51455	*REV1*	REV1, DNA directed polymerase	2q11.2
91752	*ZNF804A*	zinc finger protein 804A	2q32.1

57628	*DPP10*	dipeptidyl peptidase like 10	2q14.1
3485	*IGFBP2*	insulin like growth factor binding protein 2	2q35
3554	*IL1R1*	interleukin 1 receptor type 1	2q11.2-q12.1
6869	*TACR1*	tachykinin receptor 1	2p12
8745	*ADAM23*	ADAM metallopeptidase domain 23	2q33.3
9669	*EIF5B*	eukaryotic translation initiation factor 5B	2q11.2
1496	*CTNNA2*	catenin alpha 2	2p12
57142	*RTN4*	reticulon 4	2p16.1
26504	*CNNM4*	cyclin and CBS domain divalent metal cation transport mediator 4	2q11.2
3631	*INPP4A*	inositol polyphosphate-4-phosphatase type I A	2q11.2
2571	*GAD1*	glutamate decarboxylase 1	2q31.1
10678	*B3GNT2*	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2	2p15
8864	*PER2*	period circadian clock 2	2q37.3
80705	*TSGA10*	testis specific 10	2q11.2
2825	*GPR1*	G protein-coupled receptor 1	2q33.3
116987	*AGAP1*	ArfGAP with GTPase domain, ankyrin repeat and PH domain 1	2q37.2
51601	*LIPT1*	lipoyltransferase 1	2q11.2
10190	*TXNDC9*	thioredoxin domain containing 9	2q11.2
344148	*NCKAP5*	NCK associated protein 5	2q21.2
51263	*MRPL30*	mitochondrial ribosomal protein L30	2q11.2
254773	*LYG2*	lysozyme g2	2q11.2
6697	*SPR*	sepiapterin reductase (7,8-dihydrobiopterin:NADP+ oxidoreductase)	2p13.2
57683	*ZDBF2*	zinc finger DBF-type containing 2	2q33.3
129531	*MITD1*	microtubule interacting and trafficking domain containing 1	2q11.2
129530	*LYG1*	lysozyme g1	2q11.2
130162	*CLHC1*	clathrin heavy chain linker domain containing 1	2p16.1
3628	*INPP1*	inositol polyphosphate-1-phosphatase	2q32.2
150590	*C2orf15*	chromosome 2 open reading frame 15	2q11.2
343990	*KIAA1211L*	KIAA1211 like	2q11.2
90342	*FER1L5*	fer-1 like family member 5	2q11.2

101669764	*GPR1-AS*	GPR1 antisense RNA	2q33.3
100506457	*MIR3681HG*	MIR3681 host gene	2p24.3
101752400	*CAPN10-AS1*	CAPN10 antisense RNA 1 (head to head)	2q37.3
100616464	*MIR4778*	microRNA 4778	2p14
101927554	*LINC01250*	long intergenic non-protein coding RNA 1250	2p25.3
54106	*TLR9*	toll like receptor 9	3p21.2
5580	*PRKCD*	protein kinase C delta	3p21.1
26354	*GNL3*	G protein nucleolar 3	3p21.1
55193	*PBRM1*	polybromo 1	3p21.1
2932	*GSK3B*	glycogen synthase kinase 3 beta	3q13.33
1499	*CTNNB1*	catenin beta 1	3p22.1
8850	*KAT2B*	lysine acetyltransferase 2B	3p24.3
3697	*ITIH1*	inter-alpha-trypsin inhibitor heavy chain 1	3p21.1
8314	*BAP1*	BRCA1 associated protein 1	3p21.1
9881	*TRANK1*	tetratricopeptide repeat and ankyrin repeat containing 1	3p22.2
152189	*CMTM8*	CKLF like MARVEL transmembrane domain containing 8	3p22.3
7134	*TNNC1*	troponin C1, slow skeletal and cardiac type	3p21.1
6787	*NEK4*	NIMA related kinase 4	3p21.1
1814	*DRD3*	dopamine receptor D3	3q13.31
4026	*LPP*	LIM domain containing preferred translocation partner in lipoma	3q27.3-q28
151648	*SGO1*	shugoshin 1	3p24.3
4615	*MYD88*	myeloid differentiation primary response 88	3p22.2
51185	*CRBN*	cereblon	3p26.2
8087	*FXR1*	FMR1 autosomal homolog 1	3q26.33
27074	*LAMP3*	lysosomal associated membrane protein 3	3q27.1
56999	*ADAMTS9*	ADAM metallopeptidase with thrombospondin type 1 motif 9	3p14.1
3700	*ITIH4*	inter-alpha-trypsin inhibitor heavy chain family member 4	3p21.1
211	*ALAS1*	5'-aminolevulinate synthase 1	3p21.2
3699	*ITIH3*	inter-alpha-trypsin inhibitor heavy chain 3	3p21.1
23166	*STAB1*	stabilin 1	3p21.1
11188	*NISCH*	nischarin	3p21.1

132160	*PPM1M*	protein phosphatase, Mg2+/Mn2+ dependent 1M	3p21.2
2199	*FBLN2*	fibulin 2	3p25.1
10242	*KCNMB2*	potassium calcium-activated channel subfamily M regulatory beta subunit 2	3q26.32
55799	*CACNA2D3*	calcium voltage-gated channel auxiliary subunit alpha2delta 3	3p21.1-p14.3
6514	*SLC2A2*	solute carrier family 2 member 2	3q26.2
80335	*WDR82*	WD repeat domain 82	3p21.2
51460	*SFMBT1*	Scm-like with four mbt domains 1	3p21.1
25981	*DNAH1*	dynein axonemal heavy chain 1	3p21.1
11344	*TWF2*	twinfilin actin binding protein 2	3p21.2
132158	*GLYCTK*	glycerate kinase	3p21.2
8553	*BHLHE40*	basic helix-loop-helix family member e40	3p26.1
91869	*RFT1*	RFT1 homolog	3p21.1
254827	*NAALADL2*	N-acetylated alpha-linked acidic dipeptidase like 2	3q26.31
51533	*PHF7*	PHD finger protein 7	3p21.1
26059	*ERC2*	ELKS/RAB6-interacting/CAST family member 2	3p14.3
56920	*SEMA3G*	semaphorin 3G	3p21.1
2917	*GRM7*	glutamate metabotropic receptor 7	3p26.1
55830	*GLT8D1*	glycosyltransferase 8 domain containing 1	3p21.1
55186	*SLC25A36*	solute carrier family 25 member 36	3q23
2912	*GRM2*	glutamate metabotropic receptor 2	3p21.2
6854	*SYN2*	synapsin II	3p25.2
28972	*SPCS1*	signal peptidase complex subunit 1	3p21.1
79750	*ZNF385D*	zinc finger protein 385D	3p24.3
375346	*TMEM110*	transmembrane protein 110	3p21.1
389125	*MUSTN1*	musculoskeletal, embryonic nuclear protein 1	3p21.1
64943	*NT5DC2*	5'-nucleotidase domain containing 2	3p21.1
23395	*LARS2*	leucyl-tRNA synthetase 2, mitochondrial	3p21.31
389177	*TMEM212*	transmembrane protein 212	3q26.31
287015	*TRIM42*	tripartite motif containing 42	3q23
100507098	*ADAMTS9-AS2*	ADAMTS9 antisense RNA 2	3p14.1
54986	*ULK4*	unc-51 like kinase 4	3p22.1
285375	*LINC00620*	long intergenic non-protein coding RNA 620	3p25.1

692089	*SNORD19*	small nucleolar RNA, C/D box 19	3p21.1
100113381	*SNORD19B*	small nucleolar RNA, C/D box 19B	3p21.1
692109	*SNORD69*	small nucleolar RNA, C/D box 69	3p21.1
100874028	*SGO1-AS1*	SGO1 antisense RNA 1	3p24.3
101928882	*LOC101928882*	uncharacterized LOC101928882	3q26.33
101929054	*LOC101929054*	uncharacterized LOC101929054	3p21.2
102724068	*LOC102724068*	uncharacterized LOC102724068	3q23
101927874	*LOC101927874*	uncharacterized LOC101927874	
3815	*KIT*	KIT proto-oncogene receptor tyrosine kinase	4q12
3576	*CXCL8*	C-X-C motif chemokine ligand 8	4q13.3
2993	*GYPA*	glycophorin A (MNS blood group)	4q31.21
2247	*FGF2*	fibroblast growth factor 2	4q28.1
3064	*HTT*	huntingtin	4p16.3
9575	*CLOCK*	clock circadian regulator	4q12
817	*CAMK2D*	calcium/calmodulin dependent protein kinase II delta	4q26
987	*LRBA*	LPS responsive beige-like anchor protein	4q31.3
79633	*FAT4*	FAT atypical cadherin 4	4q28.1
4487	*MSX1*	msh homeobox 1	4p16.2
2823	*GPM6A*	glycoprotein M6A	4q34.2
9348	*NDST3*	N-deacetylase and N-sulfotransferase 3	4q26
2555	*GABRA2*	gamma-aminobutyric acid type A receptor alpha2 subunit	4p12
56916	*SMARCAD1*	SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1	4q22.3
10611	*PDLIM5*	PDZ and LIM domain 5	4q22.3
2195	*FAT1*	FAT atypical cadherin 1	4q35.2
84992	*PIGY*	phosphatidylinositol glycan anchor biosynthesis class Y	4q22.1
2560	*GABRB1*	gamma-aminobutyric acid type A receptor beta1 subunit	4p12
55300	*PI4K2B*	phosphatidylinositol 4-kinase type 2 beta	4p15.2
64854	*USP46*	ubiquitin specific peptidase 46	4q12
166647	*ADGRA3*	adhesion G protein-coupled receptor A3	4p15.2
2908	*NR3C1*	nuclear receptor subfamily 3 group C member 1	5q31.3

8728	*ADAM19*	ADAM metallopeptidase domain 19	5q33.3
6531	*SLC6A3*	solute carrier family 6 member 3	5p15.33
3350	*HTR1A*	5-hydroxytryptamine receptor 1A	5q12.3
1501	*CTNND2*	catenin delta 2	5p15.2
1812	*DRD1*	dopamine receptor D1	5q35.2
4552	*MTRR*	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	5p15.31
1946	*EFNA5*	ephrin A5	5q21.3
2668	*GDNF*	glial cell derived neurotrophic factor	5p13.2
10399	*RACK1*	receptor for activated C kinase 1	5q35.3
4163	*MCC*	mutated in colorectal cancers	5q22.2
7518	*XRCC4*	X-ray repair cross complementing 4	5q14.2
10409	*BASP1*	brain abundant membrane attached signal protein 1	5p15.1
108	*ADCY2*	adenylate cyclase 2	5p15.31
9315	*NREP*	neuronal regeneration related protein	5q22.1
23037	*PDZD2*	PDZ domain containing 2	5p13.3
2566	*GABRG2*	gamma-aminobutyric acid type A receptor gamma2 subunit	5q34
815	*CAMK2A*	calcium/calmodulin dependent protein kinase II alpha	5q32
170690	*ADAMTS16*	ADAM metallopeptidase with thrombospondin type 1 motif 16	5p15.32
56923	*NMUR2*	neuromedin U receptor 2	5q33.1
2561	*GABRB2*	gamma-aminobutyric acid type A receptor beta2 subunit	5q34
11174	*ADAMTS6*	ADAM metallopeptidase with thrombospondin type 1 motif 6	5q12.3
51397	*COMMD10*	COMM domain containing 10	5q23.1
64839	*FBXL17*	F-box and leucine rich repeat protein 17	5q21.3
1393	*CRHBP*	corticotropin releasing hormone binding protein	5q13.3
79772	*MCTP1*	multiple C2 and transmembrane domain containing 1	5q15
57688	*ZSWIM6*	zinc finger SWIM-type containing 6	5q12.1
101927134	*LINC01470*	long intergenic non-protein coding RNA 1470	5q33.1
285696	*LOC285696*	uncharacterized LOC285696	5p15.1

56147	*PCDHA1*	protocadherin alpha 1	5q31.3
3133	*HLA-E*	major histocompatibility complex, class I, E	6p22.1
23345	*SYNE1*	spectrin repeat containing nuclear envelope protein 1	6q25.2
7124	*TNF*	tumor necrosis factor	6p21.33
2099	*ESR1*	estrogen receptor 1	6q25.1-q25.2
4082	*MARCKS*	myristoylated alanine rich protein kinase C substrate	6q21
2309	*FOXO3*	forkhead box O3	6q21
2534	*FYN*	FYN proto-oncogene, Src family tyrosine kinase	6q21
2569	*GABRR1*	gamma-aminobutyric acid type A receptor rho1 subunit	6q15
5467	*PPARD*	peroxisome proliferator activated receptor delta	6p21.31
2289	*FKBP5*	FK506 binding protein 5	6p21.31
5454	*POU3F2*	POU class 3 homeobox 2	6q16.1
84062	*DTNBP1*	dystrobrevin binding protein 1	6p22.3
3351	*HTR1B*	5-hydroxytryptamine receptor 1B	6q14.1
1616	*DAXX*	death domain associated protein	6p21.32
9892	*SNAP91*	synaptosome associated protein 91	6q14.2
2911	*GRM1*	glutamate metabotropic receptor 1	6q24.3
4855	*NOTCH4*	notch 4	6p21.32
2173	*FABP7*	fatty acid binding protein 7	6q22.31
262	*AMD1*	adenosylmethionine decarboxylase 1	6q21
10846	*PDE10A*	phosphodiesterase 10A	6q27
319100	*TAAR6*	trace amine associated receptor 6	6q23.2
167681	*PRSS35*	protease, serine 35	6q14.2
221468	*TMEM217*	transmembrane protein 217	6p21.2
100302164	*MIR2113*	microRNA 2113	6q16.1
266727	*MDGA1*	MAM domain containing glycosylphosphatidylinositol anchor 1	6p21.2
401247	*LINC00243*	long intergenic non-protein coding RNA 243	6p21.33
693126	*MIR548A2*	microRNA 548a-2	6q23.3
414764	*HCG23*	HLA complex group 23 (non-protein coding)	6p21.32
221391	*OPN5*	opsin 5	6p12.3
10255	*HCG9*	HLA complex group 9 (non-protein coding)	6p22.1
246269	*LACE1*	lactation elevated 1	6q21

8379	*MAD1L1*	MAD1 mitotic arrest deficient like1	7p22.3
1956	*EGFR*	epidermal growth factor receptor	7p11.2
3569	*IL6*	interleukin 6	7p15.3
4846	*NOS3*	nitric oxide synthase 3	7q36.1
5243	*ABCB1*	ATP binding cassette subfamily B member 1	7q21.12
3952	*LEP*	leptin	7q32.1
5444	*PON1*	paraoxonase 1	7q21.3
26047	*CNTNAP2*	contactin associated protein-like 2	7q35-q36.1
4521	*NUDT1*	nudix hydrolase 1	7p22.3
51422	*PRKAG2*	protein kinase AMP-activated non-catalytic subunit gamma 2	7q36.1
2768	*GNA12*	G protein subunit alpha 12	7p22.3-p22.2
84433	*CARD11*	caspase recruitment domain family member 11	7p22.2
29886	*SNX8*	sorting nexin 8	7p22.3
781	*CACNA2D1*	calcium voltage-gated channel auxiliary subunit alpha2delta 1	7q21.11
26053	*AUTS2*	autism susceptibility candidate 2	7q11.22
5649	*RELN*	reelin	7q22.1
6718	*AKR1D1*	aldo-keto reductase family 1 member D1	7q33
5799	*PTPRN2*	protein tyrosine phosphatase, receptor type N2	7q36.3
9988	*DMTF1*	cyclin D binding myb like transcription factor 1	7q21.12
1129	*CHRM2*	cholinergic receptor muscarinic 2	7q33
84668	*FAM126A*	family with sequence similarity 126 member A	7p15.3
6804	*STX1A*	syntaxin 1A	7q11.23
2913	*GRM3*	glutamate metabotropic receptor 3	7q21.11-q21.12
55503	*TRPV6*	transient receptor potential cation channel subfamily V member 6	7q34
30816	*ERVW-1*	endogenous retrovirus group W member 1	7q21.2
29960	*MRM2*	mitochondrial rRNA methyltransferase 2	7p22.3
6671	*SP4*	Sp4 transcription factor	7p15.3
27445	*PCLO*	piccolo presynaptic cytomatrix protein	7q21.11
54517	*PUS7*	pseudouridylate synthase 7 (putative)	7q22.3
340267	*COL28A1*	collagen type XXVIII alpha 1 chain	7p21.3

7425	*VGF*	VGF nerve growth factor inducible	7q22.1
154664	*ABCA13*	ATP binding cassette subfamily A member 13	7p12.3
168741	*PER4*	period circadian clock 3 pseudogene	7p21.3
55607	*PPP1R9A*	protein phosphatase 1 regulatory subunit 9A	7q21.3
100130771	*EFCAB10*	EF-hand calcium binding domain 10	7q22.3
102465503	*MIR6836*	microRNA 6836	7p22.3
393076	*LOC393076*	uncharacterized LOC393076	7q36.3
3084	*NRG1*	neuregulin 1	8p12
10395	*DLC1*	DLC1 Rho GTPase activating protein	8p22
64478	*CSMD1*	CUB and Sushi multiple domains 1	8p23.2
4482	*MSRA*	methionine sulfoxide reductase A	8p23.1
4325	*MMP16*	matrix metallopeptidase 16	8q21.3
6482	*ST3GAL1*	ST3 beta-galactoside alpha-2,3-sialyltransferase 1	8q24.22
51606	*ATP6V1H*	ATPase H+ transporting V1 subunit H	8q11.23
526	*ATP6V1B2*	ATPase H+ transporting V1 subunit B2	8p21.3
6641	*SNTB1*	syntrophin beta 1	8q24.12
1808	*DPYSL2*	dihydropyrimidinase like 2	8p21.2
3786	*KCNQ3*	potassium voltage-gated channel subfamily Q member 3	8q24.22
9172	*MYOM2*	myomesin 2	8p23.3
2171	*FABP5*	fatty acid binding protein 5	8q21.13
90362	*FAM110B*	family with sequence similarity 110 member B	8q12.1
1142	*CHRNB3*	cholinergic receptor nicotinic beta 3 subunit	8p11.21
8756	*ADAM7*	ADAM metallopeptidase domain 7	8p21.2
8973	*CHRNA6*	cholinergic receptor nicotinic alpha 6 subunit	8p11.21
3612	*IMPA1*	inositol monophosphatase 1	8q21.13
1960	*EGR3*	early growth response 3	8p21.3
2843	*GPR20*	G protein-coupled receptor 20	8q24.3
114	*ADCY8*	adenylate cyclase 8	8q24.22
6570	*SLC18A1*	solute carrier family 18 member A1	8p21.3
57210	*SLC45A4*	solute carrier family 45 member 4	8q24.3
389676	*C8orf87*	chromosome 8 open reading frame 87	8q22.1

301	*ANXA1*	annexin A1	9q21.13
7099	*TLR4*	toll like receptor 4	9q33.1
4920	*ROR2*	receptor tyrosine kinase like orphan receptor 2	9q22.31
3309	*HSPA5*	heat shock protein family A (Hsp70) member 5	9q33.3
203228	*C9orf72*	chromosome 9 open reading frame 72	9p21.2
10558	*SPTLC1*	serine palmitoyltransferase long chain base subunit 1	9q22.31
23081	*KDM4C*	lysine demethylase 4C	9p24.1
158219	*TTC39B*	tetratricopeptide repeat domain 39B	9p22.3
4915	*NTRK2*	neurotrophic receptor tyrosine kinase 2	9q21.33
5646	*PRSS3*	protease, serine 3	9p13.3
2902	*GRIN1*	glutamate ionotropic receptor NMDA type subunit 1	9q34.3
1621	*DBH*	dopamine beta-hydroxylase	9q34.2
7091	*TLE4*	transducin like enhancer of split 4	9q21.31
25861	*WHRN*	whirlin	9q32
5730	*PTGDS*	prostaglandin D2 synthase	9q34.3
23245	*ASTN2*	astrotactin 2	9q33.1
79987	*SVEP1*	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1	9q31.3
5239	*PGM5*	phosphoglucomutase 5	9q21.11
8013	*NR4A3*	nuclear receptor subfamily 4 group A member 3	9q31.1
116443	*GRIN3A*	glutamate ionotropic receptor NMDA type subunit 3A	9q31.1
2889	*RAPGEF1*	Rap guanine nucleotide exchange factor 1	9q34.13
23413	*NCS1*	neuronal calcium sensor 1	9q34.11
4300	*MLLT3*	MLLT3, super elongation complex subunit	9p21.3
9413	*FAM189A2*	family with sequence similarity 189 member A2	9q21.12
6096	*RORB*	RAR related orphan receptor B	9q21.13
84253	*GARNL3*	GTPase activating Rap/RanGAP domain like 3	9q33.3
101929665	*UBE2R2-AS1*	UBE2R2 antisense RNA 1	9p13.3
84628	*NTNG2*	netrin G2	9q34.13
288	*ANK3*	ankyrin 3	10q21.2

6934	*TCF7L2*	transcription factor 7 like 2	10q25.2-q25.3
4153	*MBL2*	mannose binding lectin 2	10q21.1
2263	*FGFR2*	fibroblast growth factor receptor 2	10q26.13
3799	*KIF5B*	kinesin family member 5B	10p11.22
64072	*CDH23*	cadherin related 23	10q22.1
22978	*NT5C2*	5'-nucleotidase, cytosolic II	10q24.32-q24.33
57178	*ZMIZ1*	zinc finger MIZ-type containing 1	10q22.3
54805	*CNNM2*	cyclin and CBS domain divalent metal cation transport mediator 2	10q24.32
8644	*AKR1C3*	aldo-keto reductase family 1 member C3	10p15.1
1390	*CREM*	cAMP responsive element modulator	10p11.21
1959	*EGR2*	early growth response 2	10q21.3
140766	*ADAMTS14*	ADAM metallopeptidase with thrombospondin type 1 motif 14	10q22.1
6571	*SLC18A2*	solute carrier family 18 member A2	10q25.3
22891	*ZNF365*	zinc finger protein 365	10q21.2
7093	*TLL2*	tolloid like 2	10q24.1
118663	*BTBD16*	BTB domain containing 16	10q26.13
627	*BDNF*	brain derived neurotrophic factor	11p14.1
26011	*TENM4*	teneurin transmembrane protein 4	11q14.1
133	*ADM*	adrenomedullin	11p15.4
79796	*ALG9*	ALG9, alpha-1,2-mannosyltransferase	11q23.1
1813	*DRD2*	dopamine receptor D2	11q23.2
89	*ACTN3*	actinin alpha 3 (gene/pseudogene)	11q13.2
8722	*CTSF*	cathepsin F	11q13.2
3606	*IL18*	interleukin 18	11q23.1
25855	*BRMS1*	breast cancer metastasis suppressor 1	11q13.2
1815	*DRD4*	dopamine receptor D4	11p15.5
5286	*PIK3C2A*	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 alpha	11p15.1
10072	*DPP3*	dipeptidyl peptidase 3	11q13.2
582	*BBS1*	Bardet-Biedl syndrome 1	11q13.2
7351	*UCP2*	uncoupling protein 2	11q13.4
5091	*PC*	pyruvate carboxylase	11q13.2
4684	*NCAM1*	neural cell adhesion molecule 1	11q23.2
143483	*LOC143483*	similar to disrupted in bipolar disorder 1; similar to putative mannosyltransferase Alg9p (S. cerevisiae)	

9973	*CCS*	copper chaperone for superoxide dismutase	11q13.2
6289	*SAA2*	serum amyloid A2	11p15.1
57124	*CD248*	CD248 molecule	11q13.2
81876	*RAB1B*	RAB1B, member RAS oncogene family	11q13.2
7166	*TPH1*	tryptophan hydroxylase 1	11p15.1
10992	*SF3B2*	splicing factor 3b subunit 2	11q13.1
9610	*RIN1*	Ras and Rab interactor 1	11q13.2
6712	*SPTBN2*	spectrin beta, non-erythrocytic 2	11q13.2
9638	*FEZ1*	fasciculation and elongation protein zeta 1	11q24.2
3177	*SLC29A2*	solute carrier family 29 member 2	11q13.2
5936	*RBM4*	RNA binding motif protein 4	11q13.2
10432	*RBM14*	RNA binding motif protein 14	11q13.2
156	*GRK2*	G protein-coupled receptor kinase 2	11q13.2
54414	*SIAE*	sialic acid acetylesterase	11q24.2
6506	*SLC1A2*	solute carrier family 1 member 2	11p13
55690	*PACS1*	phosphofurin acidic cluster sorting protein 1	11q13.1-q13.2
64837	*KLC2*	kinesin light chain 2	11q13.2
887	*CCKBR*	cholecystokinin B receptor	11p15.4
50863	*NTM*	neurotrimin	11q25
9415	*FADS2*	fatty acid desaturase 2	11q12.2
9986	*RCE1*	Ras converting CAAX endopeptidase 1	11q13.2
406	*ARNTL*	aryl hydrocarbon receptor nuclear translocator like	11p15.3
2915	*GRM5*	glutamate metabotropic receptor 5	11q14.2-q14.3
63876	*PKNOX2*	PBX/knotted 1 homeobox 2	11q24.2
4319	*MMP10*	matrix metallopeptidase 10	11q22.2
53942	*CNTN5*	contactin 5	11q22.1
254263	*CNIH2*	cornichon family AMPA receptor auxiliary protein 2	11q13.2
4837	*NNMT*	nicotinamide N-methyltransferase	11q23.2
65003	*MRPL11*	mitochondrial ribosomal protein L11	11q13.2
10897	*YIF1A*	Yip1 interacting factor homolog A, membrane trafficking protein	11q13.2
246330	*PELI3*	pellino E3 ubiquitin protein ligase family member 3	11q13.2
89792	*GAL3ST3*	galactose-3-O-sulfotransferase 3	11q13.1
1408	*CRY2*	cryptochrome circadian clock 2	11p11.2

57689	*LRRC4C*	leucine rich repeat containing 4C	11p12
78999	*LRFN4*	leucine rich repeat and fibronectin type III domain containing 4	11q13.2
4900	*NRGN*	neurogranin	11q24.2
266743	*NPAS4*	neuronal PAS domain protein 4	11q13.2
83759	*RBM4B*	RNA binding motif protein 4B	11q13.2
8525	*DGKZ*	diacylglycerol kinase zeta	11p11.2
9152	*SLC6A5*	solute carrier family 6 member 5	11p15.1
2900	*GRIK4*	glutamate ionotropic receptor kainate type subunit 4	11q23.3
91683	*SYT12*	synaptotagmin 12	11q13.2
79703	*C11orf80*	chromosome 11 open reading frame 80	11q13.2
84867	*PTPN5*	protein tyrosine phosphatase, non-receptor type 5	11p15.1
100130460	*CAND1.11*	uncharacterized LOC100130460	11p15.4
55231	*CCDC87*	coiled-coil domain containing 87	11q13.2
254359	*ZDHHC24*	zinc finger DHHC-type containing 24	11q13.2
256472	*TMEM151A*	transmembrane protein 151A	11q13.2
100528017	*SAA2-SAA4*	SAA2-SAA4 readthrough	11p15.1
254439	*C11orf86*	chromosome 11 open reading frame 86	11q13.2
775	*CACNA1C*	calcium voltage-gated channel subunit alpha1 C	12p13.33
3479	*IGF1*	insulin like growth factor 1	12q23.2
3458	*IFNG*	interferon gamma	12q15
7450	*VWF*	von Willebrand factor	12p13.31
5027	*P2RX7*	purinergic receptor P2X 7	12q24.31
121278	*TPH2*	tryptophan hydroxylase 2	12q21.1
217	*ALDH2*	aldehyde dehydrogenase 2 family (mitochondrial)	12q24.12
7132	*TNFRSF1A*	TNF receptor superfamily member 1A	12p13.31
2904	*GRIN2B*	glutamate ionotropic receptor NMDA type subunit 2B	12p13.1
2784	*GNB3*	G protein subunit beta 3	12p13.31
1272	*CNTN1*	contactin 1	12q12
4842	*NOS1*	nitric oxide synthase 1	12q24.22
1240	*CMKLR1*	chemerin chemokine-like receptor 1	12q23.3
488	*ATP2A2*	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 2	12q24.11

7184	*HSP90B1*	heat shock protein 90 beta family member 1	12q23.3
2026	*ENO2*	enolase 2	12p13.31
27289	*RND1*	Rho family GTPase 1	12q13.12
894	*CCND2*	cyclin D2	12p13.32
5074	*PAWR*	pro-apoptotic WT1 regulator	12q21.2
9416	*DDX23*	DEAD-box helicase 23	12q13.12
784	*CACNB3*	calcium voltage-gated channel auxiliary subunit beta 3	12q13.12
1848	*DUSP6*	dual specificity phosphatase 6	12q21.33
1610	*DAO*	D-amino acid oxidase	12q24.11
6334	*SCN8A*	sodium voltage-gated channel alpha subunit 8	12q13.13
35	*ACADS*	acyl-CoA dehydrogenase, C-2 to C-3 short chain	12q24.31
56890	*MDM1*	Mdm1 nuclear protein	12q15
114795	*TMEM132B*	transmembrane protein 132B	12q24.31-q24.32
11113	*CIT*	citron rho-interacting serine/threonine kinase	12q24.23
50846	*DHH*	desert hedgehog	12q13.12
9671	*WSCD2*	WSC domain containing 2	12q23.3
8843	*HCAR3*	hydroxycarboxylic acid receptor 3	12q24.31
5992	*RFX4*	regulatory factor X4	12q23.3
267012	*DAOA*	D-amino acid oxidase activator	13q34
114798	*SLITRK1*	SLIT and NTRK like family member 1	13q31.1
3356	*HTR2A*	5-hydroxytryptamine receptor 2A	13q14.2
675	*BRCA2*	BRCA2, DNA repair associated	13q13.1
1284	*COL4A2*	collagen type IV alpha 2 chain	13q34
160851	*DGKH*	diacylglycerol kinase eta	13q14.11
6445	*SGCG*	sarcoglycan gamma	13q12.12
8100	*IFT88*	intraflagellar transport 88	13q12.11
259232	*NALCN*	sodium leak channel, non-selective	13q32.3-q33.1
10082	*GPC6*	glypican 6	13q31.3-q32.1
51761	*ATP8A2*	ATPase phospholipid transporting 8A2	13q12.13
84189	*SLITRK6*	SLIT and NTRK like family member 6	13q31.1
10301	*DLEU1*	deleted in lymphocytic leukemia 1	13q14.2-q14.3
282706	*DAOA-AS1*	DAOA antisense RNA 1	13q34
26960	*NBEA*	neurobeachin	13q13.3
23348	*DOCK9*	dedicator of cytokinesis 9	13q32.3

646982	*LINC00598*	long intergenic non-protein coding RNA 598	13q14.11
100874128	*LINC00333*	long intergenic non-protein coding RNA 333	13q31.1
100861552	*LINC00558*	long intergenic non-protein coding RNA 558	13q14.3
100885778	*NALCN-AS1*	NALCN antisense RNA 1	13q32.3
207	*AKT1*	AKT serine/threonine kinase 1	14q32.33
64067	*NPAS3*	neuronal PAS domain protein 3	14q13.1
3183	*HNRNPC*	heterogeneous nuclear ribonucleoprotein C (C1/C2)	14q11.2
8650	*NUMB*	NUMB, endocytic adaptor protein	14q24.2-q24.3
2643	*GCH1*	GTP cyclohydrolase 1	14q22.2
1734	*DIO2*	iodothyronine deiodinase 2	14q31.1
341880	*SLC35F4*	solute carrier family 35 member F4	14q22.3-q23.1
5015	*OTX2*	orthodenticle homeobox 2	14q22.3
9495	*AKAP5*	A-kinase anchoring protein 5	14q23.3
122402	*TDRD9*	tudor domain containing 9	14q32.33
57156	*TMEM63C*	transmembrane protein 63C	14q24.3
102724845	*LOC102724845*	uncharacterized LOC102724845	14q11.2
122525	*C14orf28*	chromosome 14 open reading frame 28	14q21.2
400359	*C15orf53*	chromosome 15 open reading frame 53	15q14
80381	*CD276*	CD276 molecule	15q24.1
1139	*CHRNA7*	cholinergic receptor nicotinic alpha 7 subunit	15q13.3
1512	*CTSH*	cathepsin H	15q25.1
302	*ANXA2*	annexin A2	15q22.2
10125	*RASGRP1*	RAS guanyl releasing protein 1	15q14
6095	*RORA*	RAR related orphan receptor A	15q22.2
9915	*ARNT2*	aryl hydrocarbon receptor nuclear translocator 2	15q25.1
176	*ACAN*	aggrecan	15q26.1
4916	*NTRK3*	neurotrophic receptor tyrosine kinase 3	15q25.3
55784	*MCTP2*	multiple C2 and transmembrane domain containing 2	15q26.2
54832	*VPS13C*	vacuolar protein sorting 13 homolog C	15q22.2
84952	*CGNL1*	cingulin like 1	15q21.3
54852	*PAQR5*	progestin and adipoQ receptor family member 5	15q23

2558	*GABRA5*	gamma-aminobutyric acid type A receptor alpha5 subunit	15q12
8128	*ST8SIA2*	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2	15q26.1
89832	*CHRFAM7A*	CHRNA7 (exons 5-10) and FAM7A (exons A-E) fusion	15q13.2
145773	*FAM81A*	family with sequence similarity 81 member A	15q22.2
100420292	*SEPHS1P2*	selenophosphate synthetase 1 pseudogene 2	15q26.1
101929560	*LOC101929560*	uncharacterized LOC101929560	15q25.1
51741	*WWOX*	WW domain containing oxidoreductase	16q23.1-q23.2
79728	*PALB2*	partner and localizer of BRCA2	16p12.2
64127	*NOD2*	nucleotide binding oligomerization domain containing 2	16q12.1
79068	*FTO*	FTO, alpha-ketoglutarate dependent dioxygenase	16q12.2
1387	*CREBBP*	CREB binding protein	16p13.3
54715	*RBFOX1*	RNA binding protein, fox-1 homolog 1	16p13.3
6530	*SLC6A2*	solute carrier family 6 member 2	16q12.2
84166	*NLRC5*	NLR family CARD domain containing 5	16q13
23322	*RPGRIP1L*	RPGRIP1 like	16q12.2
57338	*JPH3*	junctophilin 3	16q24.2
2903	*GRIN2A*	glutamate ionotropic receptor NMDA type subunit 2A	16p13.2
51760	*SYT17*	synaptotagmin 17	16p12.3
115	*ADCY9*	adenylate cyclase 9	16p13.3
57687	*VAT1L*	vesicle amine transport 1 like	16q23.1
6532	*SLC6A4*	solute carrier family 6 member 4	17q11.2
1636	*ACE*	angiotensin I converting enzyme	17q23.3
4137	*MAPT*	microtubule associated protein tau	17q21.31
6347	*CCL2*	C-C motif chemokine ligand 2	17q12
1394	*CRHR1*	corticotropin releasing hormone receptor 1	17q21.31
2670	*GFAP*	glial fibrillary acidic protein	17q21.31
7531	*YWHAE*	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon	17p13.3
239	*ALOX12*	arachidonate 12-lipoxygenase, 12S type	17p13.1

84152	*PPP1R1B*	protein phosphatase 1 regulatory inhibitor subunit 1B	17q12
9572	*NR1D1*	nuclear receptor subfamily 1 group D member 1	17q21.1
51479	*ANKFY1*	ankyrin repeat and FYVE domain containing 1	17p13.2
57521	*RPTOR*	regulatory associated protein of MTOR complex 1	17q25.3
3927	*LASP1*	LIM and SH3 protein 1	17q12
57674	*RNF213*	ring finger protein 213	17q25.3
63826	*SRR*	serine racemase	17p13.3
27091	*CACNG5*	calcium voltage-gated channel auxiliary subunit gamma 5	17q24.2
284076	*TTLL6*	tubulin tyrosine ligase like 6	17q21.32
406994	*MIR212*	microRNA 212	17p13.3
40	*ASIC2*	acid sensing ion channel subunit 2	17q11.2-q12
23140	*ZZEF1*	zinc finger ZZ-type and EF-hand domain containing 1	17p13.2
124936	*CYB5D2*	cytochrome b5 domain containing 2	17p13.2
245	*ALOX12P2*	arachidonate 12-lipoxygenase pseudogene 2	17p13.1
4884	*NPTX1*	neuronal pentraxin 1	17q25.3
6925	*TCF4*	transcription factor 4	18q21.2
1000	*CDH2*	cadherin 2	18q12.1
596	*BCL2*	BCL2, apoptosis regulator	18q21.33
9984	*THOC1*	THO complex 1	18p11.32
116	*ADCYAP1*	adenylate cyclase activating polypeptide 1	18p11.32
81035	*COLEC12*	collectin subfamily member 12	18p11.32
4729	*NDUFV2*	NADH:ubiquinone oxidoreductase core subunit V2	18p11.22
2774	*GNAL*	G protein subunit alpha L	18p11.21
3613	*IMPA2*	inositol monophosphatase 2	18p11.21
9218	*VAPA*	VAMP associated protein A	18p11.22
4161	*MC5R*	melanocortin 5 receptor	18p11.21
8774	*NAPG*	NSF attachment protein gamma	18p11.22
643542	*LOC643542*	uncharacterized LOC643542	18q22.1
101927430	*LOC101927430*	uncharacterized LOC101927430	18q22.1
65258	*MPPE1*	metallophosphoesterase 1	18p11.21
348	*APOE*	apolipoprotein E	19q13.32
1463	*NCAN*	neurocan	19p13.11
4784	*NFIX*	nuclear factor I X	19p13.13

26291	*FGF21*	fibroblast growth factor 21	19q13.33
2524	*FUT2*	fucosyltransferase 2	19q13.33
2523	*FUT1*	fucosyltransferase 1 (H blood group)	19q13.33
56848	*SPHK2*	sphingosine kinase 2	19q13.33
1628	*DBP*	D-box binding PAR bZIP transcription factor	19q13.33
6141	*RPL18*	ribosomal protein L18	19q13.33
58513	*EPS15L1*	epidermal growth factor receptor pathway substrate 15 like 1	19p13.11
2901	*GRIK5*	glutamate ionotropic receptor kainate type subunit 5	19q13.2
54858	*PGPEP1*	pyroglutamyl-peptidase I	19p13.11
284359	*IZUMO1*	izumo sperm-egg fusion 1	19q13.33
57572	*DOCK6*	dedicator of cytokinesis 6	19p13.2
5990	*RFX2*	regulatory factor X2	19p13.3
3337	*DNAJB1*	DnaJ heat shock protein family (Hsp40) member B1	19p13.12
770	*CA11*	carbonic anhydrase 11	19q13.33
5141	*PDE4A*	phosphodiesterase 4A	19p13.2
54922	*RASIP1*	Ras interacting protein 1	19q13.33
23383	*MAU2*	MAU2 sister chromatid cohesion factor	19p13.11
4909	*NTF4*	neurotrophin 4	19q13.33
22809	*ATF5*	activating transcription factor 5	19q13.33
60680	*CELF5*	CUGBP, Elav-like family member 5	19p13.3
147991	*DPY19L3*	dpy-19 like 3 (C. elegans)	19q13.11
284358	*MAMSTR*	MEF2 activating motif and SAP domain containing transcriptional regulator	19q13.33
148229	*ATP8B3*	ATPase phospholipid transporting 8B3	19p13.3
9710	*KIAA0355*	KIAA0355	19q13.11
54854	*FAM83E*	family with sequence similarity 83 member E	19q13.33
57030	*SLC17A7*	solute carrier family 17 member 7	19q13.33
57474	*ZNF490*	zinc finger protein 490	19p13.2-p13.13
126147	*NTN5*	netrin 5	19q13.33
5621	*PRNP*	prion protein	20p13
994	*CDC25B*	cell division cycle 25B	20p13
598	*BCL2L1*	BCL2 like 1	20q11.21
54453	*RIN2*	Ras and Rab interactor 2	20p11.23

671	*BPI*	bactericidal/permeability-increasing protein	20q11.23
1002	*CDH4*	cadherin 4	20q13.33
55816	*DOCK5*	docking protein 5	20q13.2
3785	*KCNQ2*	potassium voltage-gated channel subfamily Q member 2	20q13.33
84612	*PARD6B*	par-6 family cell polarity regulator beta	20q13.13
8537	*BCAS1*	breast carcinoma amplified sequence 1	20q13.2
3787	*KCNS1*	potassium voltage-gated channel modifier subfamily S member 1	20q13.12
10955	*SERINC3*	serine incorporator 3	20q13.12
128553	*TSHZ2*	teashirt zinc finger homeobox 2	20q13.2
128653	*C20orf141*	chromosome 20 open reading frame 141	20p13
100288797	*TMEM239*	transmembrane protein 239	20p13
351	*APP*	amyloid beta precursor protein	21q21.3
6285	*S100B*	S100 calcium binding protein B	21q22.3
875	*CBS*	cystathionine-beta-synthase	21q22.3
7226	*TRPM2*	transient receptor potential cation channel subfamily M member 2	21q22.3
54014	*BRWD1*	bromodomain and WD repeat domain containing 1	21q22.2
5116	*PCNT*	pericentrin	21q22.3
2897	*GRIK1*	glutamate ionotropic receptor kainate type subunit 1	21q21.3
8867	*SYNJ1*	synaptojanin 1	21q22.11
1826	*DSCAM*	DS cell adhesion molecule	21q22.2
1312	*COMT*	catechol-O-methyltransferase	22q11.21
2952	*GSTT1*	glutathione S-transferase theta 1	22q11.23
1565	*CYP2D6*	cytochrome P450 family 2 subfamily D member 6	22q13.2
2192	*FBLN1*	fibulin 1	22q13.31
57591	*MKL1*	megakaryoblastic leukemia (translocation) 1	22q13.1-q13.2
1414	*CRYBB1*	crystallin beta B1	22q12.1
7494	*XBP1*	X-box binding protein 1	22q12
468	*ATF4*	activating transcription factor 4	22q13.1
8398	*PLA2G6*	phospholipase A2 group VI	22q13.1
84700	*MYO18B*	myosin XVIIIB	22q12.1
23779	*ARHGAP8*	Rho GTPase activating protein 8	22q13.31
1413	*CRYBA4*	crystallin beta A4	22q12.1

7533	*YWHAH*	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta	22q12.3
1454	*CSNK1E*	casein kinase 1 epsilon	22q13.1
55615	*PRR5*	proline rich 5	22q13.31
157	*GRK3*	G protein-coupled receptor kinase 3	22q12.1
553158	*PRR5-ARHGAP8*	PRR5-ARHGAP8 readthrough	22q13.31
8224	*SYN3*	synapsin III	22q12.3
25817	*FAM19A5*	family with sequence similarity 19 member A5, C-C motif chemokine like	22q13.32
23774	*BRD1*	bromodomain containing 1	22q13.33
1416	*CRYBB2P1*	crystallin beta B2 pseudogene 1	22q11.23
9145	*SYNGR1*	synaptogyrin 1	22q13.1
54584	*GNB1L*	G protein subunit beta 1 like	22q11.21
23544	*SEZ6L*	seizure related 6 homolog like	22q12.1
388906	*OGFRP1*	opioid growth factor receptor pseudogene 1	22q13.2
1756	*DMD*	dystrophin	Xp21.2-p21.1
215	*ABCD1*	ATP binding cassette subfamily D member 1	Xq28
4128	*MAOA*	monoamine oxidase A	Xp11.3
2332	*FMR1*	fragile X mental retardation 1	Xq27.3
3358	*HTR2C*	5-hydroxytryptamine receptor 2C	Xq23
23133	*PHF8*	PHD finger protein 8	Xp11.22
438	*ASMT*	acetylserotonin O-methyltransferase	Xp22.33 and Yp11.2
54413	*NLGN3*	neuroligin 3	Xq13.1
2564	*GABRE*	gamma-aminobutyric acid type A receptor epsilon subunit	Xq28
349391	*CYCSP44*	cytochrome c, somatic pseudogene 44	Xq27.3
438	*ASMT*	acetylserotonin O-methyltransferase	Xp22.33 and Yp11.2

APPENDIX B: DSM 5 CLASSIFICATIONS FOR BIPOLAR DISORDERS (ADAPTED BY DIAGNOSTIC AND STATISTICAL MANUAL FOR MENTAL DISORDERS V (DSM V) [28]

F31 Bipolar disorder
F31.0 is a specific ICD-10-CM diagnosis code F31.0 Bipolar disorder, current episode hypomanic
F31.1 Bipolar disorder, current episode manic without psychotic features
F31.10 is a specific ICD-10-CM diagnosis code F31.10 …… unspecified
F31.11 is a specific ICD-10-CM diagnosis code F31.11 …… mild
F31.12 is a specific ICD-10-CM diagnosis code F31.12 …… moderate
F31.13 is a specific ICD-10-CM diagnosis code F31.13 …… severe
F31.2 is a specific ICD-10-CM diagnosis code F31.2 Bipolar disorder, current episode manic severe with psychotic features
F31.3 Bipolar disorder, current episode depressed, mild or moderate severity
F31.30 is a specific ICD-10-CM diagnosis code F31.30 …… unspecified
F31.31 is a specific ICD-10-CM diagnosis code F31.31 Bipolar disorder, current episode depressed, mild
F31.32 is a specific ICD-10-CM diagnosis code F31.32 Bipolar disorder, current episode depressed, moderate
F31.4 is a specific ICD-10-CM diagnosis code F31.4 Bipolar disorder, current episode depressed, severe, without psychotic features
F31.5 is a specific ICD-10-CM diagnosis code F31.5 Bipolar disorder, current episode depressed, severe, with psychotic features
F31.6 Bipolar disorder, current episode mixed
F31.60 is a specific ICD-10-CM diagnosis code F31.60 …… unspecified
F31.61 is a specific ICD-10-CM diagnosis code F31.61 …… mild
F31.62 is a specific ICD-10-CM diagnosis code F31.62 …… moderate
F31.63 is a specific ICD-10-CM diagnosis code F31.63 …… severe, without psychotic features
F31.64 is a specific ICD-10-CM diagnosis code F31.64 …… severe, with psychotic features
F31.7 Bipolar disorder, currently in remission
F31.70 is a specific ICD-10-CM diagnosis code F31.70 …… most recent episode unspecified
F31.71 is a specific ICD-10-CM diagnosis code F31.71 Bipolar disorder, in partial remission, most recent episode hypomanic
F31.72 is a specific ICD-10-CM diagnosis code F31.72 Bipolar disorder, in full remission, most recent episode hypomanic
F31.73 is a specific ICD-10-CM diagnosis code F31.73 Bipolar disorder, in partial remission, most recent episode manic
F31.74 is a specific ICD-10-CM diagnosis code F31.74 Bipolar disorder, in full remission, most recent episode manic

F31.75 is a specific ICD-10-CM diagnosis code F31.75 Bipolar disorder, in partial remission, most recent episode depressed
F31.76 is a specific ICD-10-CM diagnosis code F31.76 Bipolar disorder, in full remission, most recent episode depressed
F31.77 is a specific ICD-10-CM diagnosis code F31.77 Bipolar disorder, in partial remission, most recent episode mixed
F31.78 is a specific ICD-10-CM diagnosis code F31.78 Bipolar disorder, in full remission, most recent episode mixed
F31.8 Other bipolar disorders
F31.81 is a specific ICD-10-CM diagnosis code F31.81 Bipolar II disorder
F31.89 is a specific ICD-10-CM diagnosis code F31.89 Other bipolar disorder
F31.9 is a specific ICD-10-CM diagnosis code F31.9 Bipolar disorder, unspecified

APPENDIX C: SELECTED 693 SNP

Number	SNP_ID	Number	SNP_ID	Number	SNP_ID
1	rs12408533	232	rs2158099	463	rs7129470
2	rs17029963	233	rs17611228	464	rs11235948
3	rs17029988	234	rs10475105	465	rs11233501
4	rs11121969	235	rs10079374	466	rs4943875
5	rs4391657	236	rs17374428	467	rs2510475
6	rs1534946	237	rs13165192	468	rs1075719
7	rs10916809	238	rs348599	469	rs1075718
8	rs10493000	239	rs7726744	470	rs11233641
9	rs4654814	240	rs10053887	471	rs6592120
10	rs7550635	241	rs7728618	472	rs7113167
11	rs12036815	242	rs252818	473	rs1367985
12	rs12043011	243	rs13153056	474	rs635823
13	rs33917597	244	rs10069540	475	rs11233711
14	rs7544781	245	rs2304054	476	rs607395
15	rs10799060	246	rs2973139	477	rs17094497
16	rs12124180	247	rs11960742	478	rs629864
17	rs1416079	248	rs12196785	479	rs10750596
18	rs10873823	249	rs2893857	480	rs1483521
19	rs4847267	250	rs10484441	481	rs979219
20	rs4847274	251	rs4713001	482	rs1563902
21	rs732183	252	rs4591839	483	rs10893354
22	rs2491621	253	rs764284	484	rs10893356
23	rs4542265	254	rs13195040	485	rs11605508
24	rs17584208	255	rs2799079	486	rs10894326
25	rs12727640	256	rs1936365	487	rs2470392
26	rs699758	257	rs9368649	488	rs7138372
27	rs17035158	258	rs2296330	489	rs553104
28	rs1325934	259	rs332566	490	rs1010878
29	rs10458387	260	rs587599	491	rs10082759
30	rs4657155	261	rs651189	492	rs11046737
31	rs4292933	262	rs3857620	493	rs11047111
32	rs12044481	263	rs4327704	494	rs3914235
33	rs2791142	264	rs41380951	495	rs11049813
34	rs3765547	265	rs3911736	496	rs2447707

35	rs234102	266	rs2152599	497	rs1150983
36	rs85671	267	rs682170	498	rs11181937
37	rs2748938	268	rs2840794	499	rs10880439
38	rs6666273	269	rs2223239	500	rs2264358
39	rs12145634	270	rs6924957	501	rs4759035
40	rs1933573	271	rs4839826	502	rs668562
41	rs1933553	272	rs2499804	503	rs10784460
42	rs17019442	273	rs2472897	504	rs11176735
43	rs11802395	274	rs2092096	505	rs10878621
44	rs6701743	275	rs11757063	506	rs10878924
45	rs12405878	276	rs12213375	507	rs6582088
46	rs7535010	277	rs12208449	508	rs4237862
47	rs7519099	278	rs12210146	509	rs12306576
48	rs4518864	279	rs4078038	510	rs12309949
49	rs7551556	280	rs11152968	511	rs7979024
50	rs699900	281	rs9400016	512	rs2520559
51	rs10910502	282	rs11153023	513	rs2520548
52	rs12067652	283	rs6916232	514	rs2723899
53	rs6429351	284	rs12207570	515	rs4964714
54	rs7568967	285	rs11756315	516	rs10850359
55	rs17042441	286	rs2064947	517	rs11068645
56	rs9784164	287	rs9372649	518	rs7956558
57	rs4668990	288	rs9375085	519	rs9943949
58	rs3747515	289	rs9375098	520	rs7326195
59	rs219506	290	rs17058157	521	rs3887905
60	rs418451	291	rs9482263	522	rs9538327
61	rs391070	292	rs9388111	523	rs17088924
62	rs401843	293	rs6569342	524	rs1160445
63	rs17042250	294	rs9490948	525	rs4884611
64	rs7568963	295	rs289764	526	rs1333169
65	rs882632	296	rs6909430	527	rs1333170
66	rs13006495	297	rs17058404	528	rs12583479
67	rs410469	298	rs1856310	529	rs17736182
68	rs13029809	299	rs9388728	530	rs9592783
69	rs17028197	300	rs11756337	531	rs9542951
70	rs41518745	301	rs9397105	532	rs9542979
71	rs17510969	302	rs1328392	533	rs11841381
72	rs17530546	303	rs594709	534	rs1109940
73	rs4952769	304	rs644992	535	rs17067747

74	rs10199945	305	rs1247359	536	rs9585524
75	rs13407966	306	rs1893537	537	rs16957808
76	rs13184	307	rs9346929	538	rs7321815
77	rs13011472	308	rs9365488	539	rs3751403
78	rs1568452	309	rs516059	540	rs12430088
79	rs2717055	310	rs4266553	541	rs12865863
80	rs12185644	311	rs17262757	542	rs12868767
81	rs1533725	312	rs17402432	543	rs570252
82	rs2717031	313	rs2357958	544	rs9518449
83	rs2717036	314	rs10156056	545	rs9554971
84	rs1401100	315	rs7796223	546	rs9519153
85	rs10172295	316	rs4723546	547	rs1018685
86	rs848292	317	rs941299	548	rs2093256
87	rs6761469	318	rs17148813	549	rs11621263
88	rs4672240	319	rs17153296	550	rs1271805
89	rs10179027	320	rs17156280	551	rs10498283
90	rs1861226	321	rs4129230	552	rs17570915
91	rs6759994	322	rs10236943	553	rs17112101
92	rs17399724	323	rs4727369	554	rs176262
93	rs1426700	324	rs2028030	555	rs1955508
94	rs7591530	325	rs10274201	556	rs1959387
95	rs12713591	326	rs4445168	557	rs11624722
96	rs12477833	327	rs1918287	558	rs1480659
97	rs10206508	328	rs1406604	559	rs11158445
98	rs2110981	329	rs1881723	560	rs753683
99	rs11126290	330	rs17837696	561	rs17179664
100	rs2229626	331	rs1531532	562	rs12435306
101	rs4852430	332	rs1038062	563	rs8017057
102	rs1470078	333	rs10105363	564	rs11846553
103	rs17015885	334	rs17128604	565	rs1555406
104	rs4676228	335	rs901173	566	rs4774028
105	rs17034806	336	rs2952017	567	rs12439853
106	rs17045566	337	rs17642273	568	rs6493668
107	rs2009838	338	rs16879809	569	rs721548
108	rs1388407	339	rs385044	570	rs7175581
109	rs17045920	340	rs10097592	571	rs2934442
110	rs11689370	341	rs9785150	572	rs11071959
111	rs6437215	342	rs7819743	573	rs4451902
112	rs16843637	343	rs988143	574	rs937101

113	rs11883737	344	rs10453111	575	rs17799275
114	rs7601307	345	rs11992182	576	rs17737516
115	rs13412750	346	rs17720586	577	rs7172425
116	rs7594628	347	rs7462775	578	rs11635705
117	rs17591218	348	rs1394425	579	rs2073987
118	rs4673660	349	rs6471009	580	rs285767
119	rs7586383	350	rs7841070	581	rs9920603
120	rs17199249	351	rs10756080	582	rs11247065
121	rs3771048	352	rs10756084	583	rs12597924
122	rs2709370	353	rs1538514	584	rs12447637
123	rs2709373	354	rs411167	585	rs2267792
124	rs2551920	355	rs4405013	586	rs11640235
125	rs2709387	356	rs4961591	587	rs17793917
126	rs6785	357	rs9792664	588	rs11647877
127	rs2551949	358	rs1330322	589	rs13334953
128	rs2709416	359	rs528204	590	rs1566435
129	rs2952769	360	rs13285631	591	rs17639894
130	rs2464975	361	rs1932128	592	rs17624199
131	rs2551971	362	rs4471130	593	rs16972805
132	rs920211	363	rs10780308	594	rs9934482
133	rs167650	364	rs10868098	595	rs11150157
134	rs13408246	365	rs7858079	596	rs12448070
135	rs16866183	366	rs4744373	597	rs4782655
136	rs2055710	367	rs10821402	598	rs17763551
137	rs11692992	368	rs4744417	599	rs7209273
138	rs17271567	369	rs10821443	600	rs11658620
139	rs10498171	370	rs10821444	601	rs11653603
140	rs10498172	371	rs7857759	602	rs4792189
141	rs17199431	372	rs2779563	603	rs9913487
142	rs11686538	373	rs914665	604	rs2157990
143	rs7569781	374	rs944688	605	rs11657699
144	rs7574641	375	rs12343288	606	rs8072988
145	rs7581873	376	rs10759341	607	rs280046
146	rs4973124	377	rs1330349	608	rs280051
147	rs776978	378	rs17438727	609	rs190718
148	rs1527671	379	rs10794717	610	rs12942139
149	rs6718936	380	rs2387657	611	rs1353623
150	rs6720816	381	rs2400042	612	rs13341531
151	rs34617816	382	rs33932343	613	rs17819991

152	rs1382866	383	rs2505453	614	rs11871341
153	rs9311962	384	rs2505456	615	rs17820020
154	rs13077722	385	rs11015814	616	rs16942910
155	rs1505611	386	rs11015877	617	rs2877875
156	rs2730336	387	rs1219593	618	rs1069
157	rs11128782	388	rs2483023	619	rs10512586
158	rs10865742	389	rs4934825	620	rs8078277
159	rs9835075	390	rs4934826	621	rs4435291
160	rs7615587	391	rs1332772	622	rs4890043
161	rs644642	392	rs10857580	623	rs7233016
162	rs253045	393	rs10821876	624	rs10454719
163	rs620918	394	rs10761774	625	rs4404156
164	rs33108	395	rs11004607	626	rs16961011
165	rs6804900	396	rs7094854	627	rs2901813
166	rs7428295	397	rs4124862	628	rs17663182
167	rs35823108	398	rs10821582	629	rs2128605
168	rs11707243	399	rs10740018	630	rs4940377
169	rs13079040	400	rs7098008	631	rs9966035
170	rs548099	401	rs17239782	632	rs545245
171	rs9824271	402	rs1993183	633	rs17077963
172	rs17749340	403	rs11816737	634	rs41480546
173	rs7428007	404	rs17241218	635	rs7231414
174	rs7638369	405	rs7358201	636	rs2194633
175	rs9838703	406	rs2569360	637	rs7239688
176	rs980944	407	rs4980113	638	rs7231621
177	rs11712587	408	rs10762732	639	rs11876141
178	rs9828746	409	rs10824541	640	rs11673509
179	rs12330457	410	rs1249135	641	rs7254941
180	rs817503	411	rs1249131	642	rs10418705
181	rs843855	412	rs1249122	643	rs10401153
182	rs4677935	413	rs17121662	644	rs11665940
183	rs332516	414	rs699213	645	rs10415145
184	rs9865702	415	rs12765205	646	rs4805755
185	rs11917356	416	rs2185834	647	rs16967057
186	rs2370512	417	rs11186852	648	rs2111504
187	rs16842953	418	rs11186884	649	rs12459013
188	rs7639294	419	rs11186894	650	rs2099362
189	rs9862757	420	rs11186898	651	rs17206939
190	rs9873729	421	rs790653	652	rs6118267

191	rs9879590	422	rs11196371	653	rs17794135
192	rs13064363	423	rs2240878	654	rs3736771
193	rs12634337	424	rs7907586	655	rs17802375
194	rs6807246	425	rs236214	656	rs6081474
195	rs2364910	426	rs11200051	657	rs3091470
196	rs7617202	427	rs7077436	658	rs605138
197	rs4859232	428	rs10764990	659	rs6091620
198	rs17788373	429	rs11016078	660	rs158316
199	rs11935551	430	rs6482674	661	rs242812
200	rs4698501	431	rs12782247	662	rs6027712
201	rs2132631	432	rs12765772	663	rs1689059
202	rs2136807	433	rs12772010	664	rs1735888
203	rs12651329	434	rs7910053	665	rs2245652
204	rs1562094	435	rs11023096	666	rs207460
205	rs4615179	436	rs2237866	667	rs207495
206	rs11736598	437	rs2412143	668	rs933153
207	rs2111139	438	rs12788102	669	rs4816300
208	rs1817459	439	rs12789492	670	rs112475
209	rs2194124	440	rs17325567	671	rs947919
210	rs28629807	441	rs17227978	672	rs8134012
211	rs6841907	442	rs4243925	673	rs928874
212	rs2176311	443	rs2237936	674	rs1980977
213	rs11729256	444	rs10832890	675	rs6001474
214	rs10008893	445	rs12791462	676	rs6519550
215	rs3756040	446	rs5021257	677	rs5761940
216	rs11933230	447	rs2702672	678	rs17430741
217	rs6852589	448	rs2702673	679	rs5767136
218	rs101927	449	rs4382904	680	rs100000002374
219	rs10029005	450	rs11026115	681	rs100000007980
220	rs10519613	451	rs4922996	682	rs1000000010136
221	rs17027882	452	rs1155331	683	rs1000000016895
222	rs1460060	453	rs712022	684	rs1000000018168
223	rs10034062	454	rs10766971	685	rs1000000020028
224	rs796988	455	rs4567455	686	rs1000000027613
225	rs4478239	456	rs4550218	687	rs1000000027618
226	rs4478240	457	rs836116	688	rs1000000031186
227	rs13361372	458	rs7105545	689	rs1000000031189
228	rs302911	459	rs7105037	690	rs1000000031201
229	rs824619	460	rs4756201	691	rs1000000033097

230	rs1423492	461	rs6485383	692	rs1000000033108
231	rs1978462	462	rs1038673	693	rs1000000033135

APPENDIX D: GENEMANIA RESULTS OF SHARED SNPS

GeneMANIA Results	
Plugin Version	3.4.1 (20160523-2245)
Data Version	8/12/14
Report Generated	Tue Feb 28 22:39:36 EET 2017

Gene	Score	Description
METTL21A		methyltransferase like 21A
ZNF507		zinc finger protein 507
CREB1		cAMP responsive element binding protein 1
ARHGAP22		Rho GTPase activating protein 22
DNAH9		dynein, axonemal, heavy chain 9
DOCK10		dedicator of cytokinesis 10
TAOK2	0.57	TAO kinase 2
LYL1	0.51	lymphoblastic leukemia derived sequence 1
LRRC8E	0.47	leucine rich repeat containing 8 family, member E
MAP2K7	0.47	mitogen-activated protein kinase kinase 7
HIP1R	0.43	huntingtin interacting protein 1 related
BDKRB2	0.4	bradykinin receptor B2
MRPL34	0.36	mitochondrial ribosomal protein L34
ANGEL1	0.36	angel homolog 1 (Drosophila)
VSTM2L	0.35	V-set and transmembrane domain containing 2 like
CLTB	0.35	clathrin, light chain B
OR5M8	0.35	olfactory receptor, family 5, subfamily M, member 8
GPR31	0.33	G protein-coupled receptor 31
ZNF486	0.33	zinc finger protein 486
OR13A1	0.33	olfactory receptor, family 13, subfamily A, member 1
TRMT12	0.33	tRNA methyltransferase 12 homolog (S. cerevisiae)
FGF11	0.31	fibroblast growth factor 11
CXCL17	0.3	chemokine (C-X-C motif) ligand 17
UTP23	0.3	UTP23, small subunit (SSU) processome component, homolog (yeast)
SF3B14	0.28	Pre-mRNA branch site protein p14
TCL1A	0.27	T-cell leukemia/lymphoma 1A

Gene 1	Gene 2	Weight	Type	Source
ANGEL1	ARHGAP22	0.643936219	Genetic interactions	Lin-Smith-2010
ANGEL1	DNAH9	0.544105237	Genetic interactions	Lin-Smith-2010
BDKRB2	DNAH9	1.254716143	Genetic interactions	Lin-Smith-2010
CLTB	ZNF507	1.042804495	Genetic interactions	Lin-Smith-2010
CXCL17	DNAH9	0.316208974	Genetic interactions	Lin-Smith-2010
CXCL17	METTL21A	0.688065402	Genetic interactions	Lin-Smith-2010
DOCK10	ARHGAP22	0.105210335	Genetic interactions	Lin-Smith-2010
DOCK10	DNAH9	0.088899321	Genetic interactions	Lin-Smith-2010
FGF11	METTL21A	0.958959758	Genetic interactions	Lin-Smith-2010
GPR31	ZNF507	0.980480947	Genetic interactions	Lin-Smith-2010
HIP1R	METTL21A	1.2348013	Genetic interactions	Lin-Smith-2010
LRRC8E	METTL21A	1.350005437	Genetic interactions	Lin-Smith-2010
LYL1	METTL21A	1.471961476	Genetic interactions	Lin-Smith-2010
MAP2K7	METTL21A	1.350005437	Genetic interactions	Lin-Smith-2010
MRPL34	DOCK10	0.328334677	Genetic interactions	Lin-Smith-2010
MRPL34	METTL21A	0.866257399	Genetic interactions	Lin-Smith-2010
OR13A1	CREB1	1.031806879	Genetic interactions	Lin-Smith-2010
OR5M8	CREB1	1.055837702	Genetic interactions	Lin-Smith-2010
SF3B14	METTL21A	0.849009212	Genetic interactions	Lin-Smith-2010
TAOK2	ARHGAP22	1.667238586	Genetic interactions	Lin-Smith-2010
TCL1A	CREB1	0.835367665	Genetic interactions	Lin-Smith-2010
TRMT12	DNAH9	1.03198234	Genetic interactions	Lin-Smith-2010
UTP23	ARHGAP22	0.902533345	Genetic interactions	Lin-Smith-2010
VSTM2L	METTL21A	1.077084336	Genetic interactions	Lin-Smith-2010
ZNF486	METTL21A	0.994960498	Genetic interactions	Lin-Smith-2010

APPENDIX E: GENEMANIA RESULTS OF REDUCED SNPS

GeneMANIA Results	
Plugin Version	3.4.1 (20160523-2245)
Data Version	8/12/14
Report Generated	Tue Feb 28 23:37:48 EET 2017

Gene	Score	Description
METTL21A		methyltransferase like 21A
DOCK10		dedicator of cytokinesis 10
ARHGAP22		Rho GTPase activating protein 22
DNAH9		dynein, axonemal, heavy chain 9
ZNF507		zinc finger protein 507
HAP1	0.21	huntingtin-associated protein 1
DCTN1	0.14	dynactin 1
CDC42	0.12	cell division cycle 42
CSK	0.11	c-src tyrosine kinase
LURAP1	0.1	leucine rich adaptor protein 1
CACNA1A	0.08	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
DOCK6	0.08	dedicator of cytokinesis 6
HTT	0.08	huntingtin
DOCK7	0.07	dedicator of cytokinesis 7
DOCK8	0.07	dedicator of cytokinesis 8
NPHP1	0.07	nephronophthisis 1 (juvenile)
DNAH5	0.06	dynein, axonemal, heavy chain 5
C5orf54	0.06	chromosome 5 open reading frame 54
FOXJ1	0.06	forkhead box J1
DNAH12	0.06	dynein, axonemal, heavy chain 12
RHOJ	0.06	ras homolog family member J
DOCK11	0.06	dedicator of cytokinesis 11
ACAN	0.06	aggrecan
DOCK9	0.05	dedicator of cytokinesis 9
DYDC2	0.05	DPY30 domain containing 2

Gene 1	Gene 2	Weight	Type	Source
CSK	ARHGAP22	0.014743742	Co-expression	Bild-Nevins-2006 B
FOXJ1	DNAH9	0.023001243	Co-expression	Bild-Nevins-2006 B
HAP1	ZNF507	0.005424888	Co-expression	Bild-Nevins-2006 B
NPHP1	DOCK10	0.019167938	Co-expression	Bild-Nevins-2006 B
C5orf54	DNAH5	0.005458571	Co-expression	Chen-Brown-2002
ACAN	ZNF507	0.004124499	Co-expression	Gysin-McMahon-2012
HTT	DOCK6	0.015095099	Co-expression	Gysin-McMahon-2012
DNAH12	DNAH9	0.016591462	Co-expression	Mallon-McKay-2013
DNAH5	DNAH9	0.01479983	Co-expression	Mallon-McKay-2013
DYDC2	DNAH9	0.014478061	Co-expression	Mallon-McKay-2013
FOXJ1	DNAH9	0.015647229	Co-expression	Mallon-McKay-2013
FOXJ1	DNAH9	0.022177278	Co-expression	Ramaswamy-Golub-2001
ACAN	DNAH5	0.0156707	Co-expression	Roth-Zlotnik-2006
DNAH12	DNAH5	0.015250683	Co-expression	Roth-Zlotnik-2006
DNAH12	DNAH9	0.018383207	Co-expression	Roth-Zlotnik-2006
DNAH12	FOXJ1	0.01478784	Co-expression	Roth-Zlotnik-2006
DNAH5	DNAH9	0.016198034	Co-expression	Roth-Zlotnik-2006
DYDC2	DNAH12	0.016238144	Co-expression	Roth-Zlotnik-2006
DYDC2	DNAH5	0.011889081	Co-expression	Roth-Zlotnik-2006
DYDC2	DNAH9	0.017105147	Co-expression	Roth-Zlotnik-2006
DYDC2	FOXJ1	0.015241091	Co-expression	Roth-Zlotnik-2006
FOXJ1	DNAH5	0.012059864	Co-expression	Roth-Zlotnik-2006
FOXJ1	DNAH9	0.016019574	Co-expression	Roth-Zlotnik-2006
FOXJ1	HAP1	0.01161921	Co-expression	Salaverria-Siebert-2011
HAP1	ZNF507	0.005609352	Co-expression	Smirnov-Cheung-2009
NPHP1	DOCK10	0.021816047	Co-expression	Wang-Maris-2006
HAP1	ARHGAP22	0.003531928	Co-expression	Wu-Garvey-2007
HTT	CSK	0.005792803	Co-expression	Wu-Garvey-2007
ACAN	DNAH9	0.024438706	Co-localization	Johnson-Shoemaker-2003
DYDC2	ACAN	0.023371304	Co-localization	Johnson-Shoemaker-2003
DYDC2	DNAH9	0.029586782	Co-localization	Johnson-Shoemaker-2003
ACAN	DNAH9	0.040719864	Co-localization	Schadt-Shoemaker-2004
ACAN	NPHP1	0.037290209	Co-localization	Schadt-Shoemaker-2004
NPHP1	DNAH9	0.044611917	Co-localization	Schadt-Shoemaker-2004
CDC42	DOCK10	0.140575123	Pathway	PATHWAYCOMMONS-NCI_NATURE
DOCK11	CDC42	0.140575123	Pathway	PATHWAYCOMMONS-NCI_NATURE

DOCK6	CDC42	0.089181731	Pathway	PATHWAYCOMMONS-NCI_NATURE
CDC42	ARHGAP22	0.030172889	Pathway	Wu-Stein-2010
DCTN1	DNAH9	0.178652897	Pathway	Wu-Stein-2010
DCTN1	HAP1	0.016299095	Pathway	Wu-Stein-2010
DNAH5	DCTN1	0.083138412	Pathway	Wu-Stein-2010
DNAH5	HAP1	0.118714199	Pathway	Wu-Stein-2010
DNAH5	HTT	0.047832028	Pathway	Wu-Stein-2010
HAP1	DNAH9	0.255100299	Pathway	Wu-Stein-2010
HTT	DCTN1	0.006567191	Pathway	Wu-Stein-2010
HTT	DNAH9	0.102784382	Pathway	Wu-Stein-2010
HTT	HAP1	0.00937736	Pathway	Wu-Stein-2010
RHOJ	ARHGAP22	0.038345658	Pathway	Wu-Stein-2010
RHOJ	HAP1	0.012073341	Pathway	Wu-Stein-2010
RHOJ	HTT	0.00486456	Pathway	Wu-Stein-2010
CACNA1A	ARHGAP22	0.095715092	Physical interactions	BIOGRID-SMALL-SCALE-STUDIES
DCTN1	HAP1	0.027638557	Physical interactions	BIOGRID-SMALL-SCALE-STUDIES
DOCK8	CDC42	0.041933776	Physical interactions	BIOGRID-SMALL-SCALE-STUDIES
DOCK9	CDC42	0.027198787	Physical interactions	BIOGRID-SMALL-SCALE-STUDIES
HTT	HAP1	0.004396544	Physical interactions	BIOGRID-SMALL-SCALE-STUDIES
DOCK7	DOCK6	0.015618135	Physical interactions	Couzens-Gingras-2013
DOCK8	DOCK6	0.021771334	Physical interactions	Couzens-Gingras-2013
DOCK8	DOCK7	0.014307535	Physical interactions	Couzens-Gingras-2013
HTT	HAP1	0.045902377	Physical interactions	IREF-BIND
DCTN1	HAP1	9.82E-04	Physical interactions	IREF-HPRD
DOCK7	CDC42	0.008891302	Physical interactions	IREF-HPRD
DOCK8	CDC42	0.00405641	Physical interactions	IREF-HPRD
DOCK9	CDC42	0.015381997	Physical interactions	IREF-HPRD
HTT	HAP1	5.58E-04	Physical interactions	IREF-HPRD
LURAP1	DOCK10	0.025315173	Physical interactions	IREF-HPRD
RHOJ	CDC42	0.001981117	Physical interactions	IREF-HPRD
RHOJ	DOCK10	0.031972373	Physical interactions	IREF-HPRD
RHOJ	DOCK8	0.024243356	Physical interactions	IREF-HPRD
C5orf54	METTL21A	0.069042363	Physical interactions	IREF-INTACT
CACNA1A	ARHGAP22	0.009121618	Physical interactions	IREF-INTACT
DCTN1	HAP1	0.001413135	Physical interactions	IREF-INTACT
HTT	HAP1	5.34E-04	Physical interactions	IREF-INTACT
LURAP1	DOCK10	0.052082669	Physical interactions	IREF-INTACT

DOCK9	CDC42	0.015406456	Physical interactions	IREF-OPHID
HTT	DCTN1	0.001286832	Physical interactions	IREF-OPHID
HTT	HAP1	5.72E-04	Physical interactions	IREF-OPHID
LURAP1	DOCK10	0.030886908	Physical interactions	IREF-OPHID
CSK	ZNF507	0.143478113	Physical interactions	Varjosalo-Superti-Furga-2013
HTT	HAP1	0.052779651	Predicted	I2D-BIND-Mouse2Human
HTT	HAP1	0.14876919	Predicted	I2D-BIND-Rat2Human
DNAH12	DNAH5	0.017692894	Shared protein domains	INTERPRO
DNAH12	DNAH9	0.017692894	Shared protein domains	INTERPRO
DNAH5	DNAH9	0.022161285	Shared protein domains	INTERPRO
DOCK11	DOCK10	0.037525588	Shared protein domains	INTERPRO
DOCK11	DOCK6	0.058041675	Shared protein domains	INTERPRO
DOCK11	DOCK7	0.03934656	Shared protein domains	INTERPRO
DOCK11	DOCK8	0.03934656	Shared protein domains	INTERPRO
DOCK6	DOCK10	0.054661006	Shared protein domains	INTERPRO
DOCK7	DOCK10	0.041719294	Shared protein domains	INTERPRO
DOCK7	DOCK6	0.065848911	Shared protein domains	INTERPRO
DOCK8	DOCK10	0.041719294	Shared protein domains	INTERPRO
DOCK8	DOCK6	0.065848911	Shared protein domains	INTERPRO
DOCK8	DOCK7	0.050255396	Shared protein domains	INTERPRO
DOCK9	DOCK10	0.039217011	Shared protein domains	INTERPRO
DOCK9	DOCK11	0.037525588	Shared protein domains	INTERPRO
DOCK9	DOCK6	0.054661002	Shared protein domains	INTERPRO
DOCK9	DOCK7	0.041719294	Shared protein domains	INTERPRO
DOCK9	DOCK8	0.041719294	Shared protein domains	INTERPRO
RHOJ	CDC42	0.007005828	Shared protein domains	INTERPRO
DNAH12	DNAH5	0.015336512	Shared protein domains	PFAM
DNAH12	DNAH9	0.015336522	Shared protein domains	PFAM
DNAH5	DNAH9	0.019845958	Shared protein domains	PFAM
DOCK11	DOCK10	0.026078414	Shared protein domains	PFAM
DOCK11	DOCK6	0.038270449	Shared protein domains	PFAM
DOCK11	DOCK7	0.038270449	Shared protein domains	PFAM
DOCK11	DOCK8	0.038270449	Shared protein domains	PFAM
DOCK6	DOCK10	0.038270449	Shared protein domains	PFAM
DOCK7	DOCK10	0.038270449	Shared protein domains	PFAM
DOCK7	DOCK6	0.065750308	Shared protein domains	PFAM
DOCK8	DOCK10	0.038270449	Shared protein domains	PFAM
DOCK8	DOCK6	0.065750308	Shared protein domains	PFAM
DOCK8	DOCK7	0.065750308	Shared protein domains	PFAM
DOCK9	DOCK10	0.026078414	Shared protein domains	PFAM
DOCK9	DOCK11	0.026078414	Shared protein domains	PFAM

DOCK9	DOCK6	0.038270449	Shared protein domains	PFAM
DOCK9	DOCK7	0.038270449	Shared protein domains	PFAM
DOCK9	DOCK8	0.038270449	Shared protein domains	PFAM
NPHP1	CSK	0.002466514	Shared protein domains	PFAM
RHOJ	CDC42	0.003692672	Shared protein domains	PFAM

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Açıkel, Cengizhan
Nationality: Turkish
Date and Place of Birth: 29.08.1971 Erzurum
Marital Status: Married
E-mail: chacikel@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
Epidemiology speciality	Hacettepe University	2005
Public health speciality	Gulhane Military Medical Acad.	2000
Medical Faculty	Gulhane Military Medical Acad.	1994

WORK EXPERIENCE

Enrollment	**Place**	**Year**
Head of Department of Biostatistics	GMMA	2011-2016
Faculty of epidemiology	GMMA	2006-2011
Epidemiology residency	Hacettepe University	2003-2006
Public health specialist	TAF	2000-2003
Public health residency	GMMA	1997-2000
General practitioner	Isparta	1994-1997

PUBLICATIONS

[1] Acikel C, Son Y, Celik C, Gul H. Evaluation Of Potential Novel Variations And Their Interactions Related To Bipolar Disorders: Analysis Of Genome-Wide Association Study Data. Neuropsychiatr Dis . 2016

[2] Acikel C, Son YA, Celik C, Tutuncu R. Evaluation Of Whole Genome Association Study Data In Bipolar Disorders: Potential Novel Snps And Genes. Bull Clin Psychopharmacol. 2015;25(1):12–8.

[3] Eker I, Yilmaz S, Cetinkaya R, Unlu A, Pekel A. Is One-Size-Fits-All Strategy Adequate For Platelet Storage? Apher Sci . 2016

[4] Cinar M, Cinar Fi, Acikel C, Yilmaz S, Çakar M, Horne R, Et Al. Reliability And Validity Of The Turkish Translation Of The Beliefs About Medicines Questionnaire (Bmq-T) In Patients With Behçet's Disease. Clin Exp Rheumatol . 2016 May 18

[5] Demirkaya E, Acikel C. Development And Initial Validation Of International Severity Scoring System For Familial Mediterranean Fever (Issf). Ann … . 2016

[6] Yeşilkaya E, Cinaz P, Andıran N, Bideci A, Hatun Ş, Sarı E, Et Al. First Report On The Nationwide Incidence And Prevalence Of Type 1 Diabetes Among Children In Turkey. Diabet Med . 2016 Jan 27

[7] Acikel C, Kocak N, Balikci A, Aydin I, Istanbulluoglu H, Turker T, Et Al. The Frequency Of Psychiatric Disorders Leading To Disability In Young Adult Males. Anadolu Psikiyatr Dergisi-Anatolian J Psychiatry. Cumhuriyet Univ Tip Fak Psikiyatri Anabilim Dali Cumhuriyet Univ Tip Fak Psikiyatri Abd, Sivas, 58140, Turkey; 2015;16(1):14–21.

[8] Yilmaz Sahin S, Iyigun E, Acikel C. Validity And Reliability Of A Turkish Version Of The Modified Moral Sensitivity Questionnaire For Student Nurses. Ethics Behav. Routledge; 2015;25(4):279–296.

[9] Mumcuoglu T, Ozge G, Soykut B, Erdem O, Gunal A, Acikel C. An Animal Model (Guinea Pig) Of Ocular Siderosis: Histopathology, Pharmacology, And Electrophysiology. Curr Eye Res. Informa Healthcare Usa, Inc. New York; 2015;40(3):314–20.

[10] Cinar Fi, Cinar M, Yilmaz S, Acikel C, Erdem H, Pay S, Et Al. Cross-Cultural Adaptation, Reliability, And Validity Of The Turkish Version Of The Compliance Questionnaire On Rheumatology In Patients With Behçet's Disease. J Transcult Nurs. Sage Publications; 2015;1043659615577699.

[11] Ozdemir S, Bebis H, Ortabag T, Acikel C. Evaluation Of The Efficacy Of An Exercise Program For Pregnant Women With Low Back And Pelvic Pain: A Prospective Randomized Controlled Trial. J Adv Nurs. 2015;

[12] Kose G, Bolu A, Ozdemir L, Acikel C, Hatipolu S. Reliability And Validity Of The Intensive Care Delirium Screening Checklist In Turkish. Int J Nurs Knowl. 2015;

[13] Tosun B, Yava A, Açıkel C. Evaluating The Effects Of Preoperative Fasting And Fluid Limitation. Int J Nurs Pract . 2015 Feb 28

[14] Konukbay D, Yildiz D, Acikel C, Karaman D, Fidanci Be, Bilginer Y, Et Al. Evaluation Of Biopsychosocial Aspects Of Patients With Juvenile Autoinflammatory Disease: A Qualitative Study. Ann Paediatr Rheumatol. 2014;3(2):62–71.

[15] Fidanci Be, Yesilkaya S, Acikel C, Ozden A, Simsek D, Yildiz F, Et Al. Validity And Reliability Of Medication Adherence Scale In Fmf (Adult Version). Pediatr Rheumatol. Biomed Central; 2014;12(1):1.

[16] Eker I, Gursel O, Yarali N, Tunc B, Pekel A, Ertas Z, Et Al. Evaluation Of The Thrombosis Tendency In Thalassemia Major Patients With Thrombin Generation Test, Procoagulant Phospholipid Activity And Endothelial Microparticles Levels. In: Haematologica. 2014. P. 478–9.

[17] Konukbay D, Yildiz D, Acikel C, Sozeri B, Makay B, Ayaz Na, Et Al. Development And Validation Of Juvenile Autoinflammatory Disease Multidimensional Assessment Report (Jaimar). Pediatr Rheumatol. Biomed Central; 2014;12(1):1–2.

[18] Poesen R, Viaene L, Bammens B, Claes K, Evenepoel P, Meijers B, Et Al. Ckd Nutrition, Inflammation And Oxidative Stress. Nephrol Dial Transplant. Oxford University Press; 2014;29(Suppl 3):Iii406--Iii418.

[19] Ozbek G, Gul Hc, Karakas A, Artuk C, Acikel C, Gorenek L, Et Al. Cost Analysis Of Healthcare Associated Infection In A Training Hospital. Int J Infect Dis. Elsevier; 2014;21:402.

[20] Zerener T, Aydintug Ys, Sencimen M, Bayar Gr, Yazici M, Altug Ha, Et Al. Clinical Comparison Of Submucosal Injection Of Dexamethasone And Triamcinolone Acetonide On Postoperative Discomfort After Third Molar Surgery. Quintessence Int (Berlin, Ger 1985). 2014;

[21] Gul H, Son Aydin Y, Acikel C. Discovering Missing Heritability And Early Risk Prediction For Type 2 Diabetes: A New Perspective For Genome-Wide Association Study Analysis With The Nurses' Health Study And The Health Professionals' Follow-Up Study. Turk J Med Sci . 2014

[22] Sencimen M, Saygun I, Gulses A, Bal V, Acikel Ch, Kubar A. Evaluation Of Periodontal Pathogens Of The Mandibular Third Molar Pericoronitis By Using Real Time Pcr. Int Dent J . 2014 May 19

[23] Ozen S, Demirkaya E, Amaryan G, Koné-Paut I, Polat A, Woo P, Et Al. Results From A Multicentre International Registry Of Familial Mediterranean Fever: Impact Of Environment On The Expression Of A Monogenic Disease In Children. Ann Rheum Dis . 2014 Apr

[24] Ozen S, Demirkaya E, Duzova A, Erdogan O, Erken E, Gul A, Et Al. Fmf50: A Score For Assessing Outcome In Familial Mediterranean Fever. Ann Rheum Dis . 2014 May

[25] Amasyali B, Kilic A, Kabul Hk, Imren E, Acikel C. Patients With Drug-Refractory Atrioventricular Nodal Reentrant Tachycardia: Clinical Features,

Electrophysiological Characteristics, And Predictors Of Medication Failure. J Cardiol . 2014 Feb 24

[26] Yilmaz Mi, Solak Y, Saglam M, Cayci T, Acikel C, Unal Hu, Et Al. The Relationship Between Il-10 Levels And Cardiovascular Events In Patients With Ckd. Clin J Am Soc Nephrol . 2014 May 1

[27] Genc H, Dogru T, Celebi G, Tapan S, Kara M, Ercin Cn, Et Al. Non-Alcoholic Fatty Liver Disease Per Se Is Not Associated With Carotid Atherosclerosis. Gulhane Med J . 2013

[28] Cinar M, Akar H, Yilmaz S, Simsek I, Karkucak M, Sagkan Ri, Et Al. The Role Of Endoplasmic Reticulum Aminopeptidase 1 (Erap1) In The Pathogenesis Of Ankylosing Spondylitis. In: Annals Of The Rheumatic Diseases. 2013. P. 283.

[29] Demirkaya E, Acikel C, Basbozkurt G, Gul A, Kasapcopur O, Aydog O, Et Al. Pres-Final-2213: Validation Of Inadequate Drug Response And Definition Of Colchicum Resistance In Fmf. Pediatr Rheumatol. Biomed Central; 2013;11(2):1–2.

[30] Demirkaya E, Özden A, Aydogdu K, Polat A, Findik G, Agackiran Y, Et Al. Descriptive Features Of Mesothelioma Cases Diagnosed In A Special Hospital In Ankara And Assessment Of Domestic Environmental Exposure To Asbestosis And Erionite: Preliminary Results. Agu Spring Meet Abstr. 2013;1:3.

[31] Demirkaya E, Acikel C. After The First Year Of Publishing With Apr. Ann Paediatr Rheumatol. 2013;2(1):1–2.

[32] Genc H, Dogru T, Celebi G, Tapan S, Kara M, Ercin Cn, Et Al. Non-Alcoholic Fatty Liver Disease Per Se Is Not Associated With Carotid Atherosclerosis. Gulhane Med J. 2013;55:84–8.

[33] Gezer M, Tasci I, Demir O, Acikel C, Cakar M, Saglam K, Et Al. Low Frequency Of A Decreased Ankle Brachial Index And Associated Conditions In The Practice Of Internal Medicine In A Turkish Population Sample. Int Angiol . 2012 Oct

[34] Aydintug Ys, Bayar Gr, Gulses A, Misir Af, Ogretir O, Dogan N, Et Al. Clinical Study On The Closure Of Extraction Wounds Of Partially Soft Tissue-Impacted Mandibular Third Molars. Quintessence Int . 2012

[35] Sweeney Wm, Afifi Am, Zor F, Acikel Ch, Bozkurt M, Grykien C, Et Al. Anatomic Survey Of Arachnoid Foveolae And The Clinical Correlation To Cranial Bone Grafting. J Craniofac Surg . Lww; 2011 Jan

[36] Erdurman Fc, Erdurman Fc, Sobaci G, Acikel Ch, Ceylan Mo, Durukan Ah, Et Al. Anatomical And Functional Outcomes In Contusion Injuries Of Posterior Segment. Eye (Lond) . Nature Publishing Group; 2011 Aug

[37] Guvenc G, Akyuz A, Açikel Chc. Health Belief Model Scale For Cervical Cancer And Pap Smear Test: Psychometric Testing. J Adv Nurs . Blackwell Publishing Ltd; 2011 Feb

[38] Ozgurtas T, Aydin I, Turan O, Koc E, Hirfanoglu Im, Acikel Ch, Et Al. Soluble Vascular Endothelial Growth Factor Receptor 1 In Human Breast Milk. Horm Res Paediatr . 2011 Jan

[39] Saygun I, Nizam N, Keskiner I, Bal V, Kubar A, Açıkel C, Et Al. Salivary Infectious Agents And Periodontal Disease Status. J Periodontal Res . 2011 Apr

[40] Erdurman Fc, Sobaci G, Acikel Ch, Ceylan Mo, Durukan Ah, Hurmeric V. Anatomical And Functional Outcomes In Contusion Injuries Of Posterior Segment. Eye (Lond) . 2011 Aug

[41] Soluble Vascular Endothelial Growth Factor Receptor 1 In Human Breast Milk. Horm Res … . 2011

[42] Guvenc G, Akyuz A, Açikel Chc. Health Belief Model Scale For Cervical Cancer And Pap Smear Test: Psychometric Testing. J Adv Nurs . 2011 Feb

[43] Erdurman Cf, Ceylan Mo, Acikel Ch, Durukan Ha, Mumcuoglu T. Outcomes Of Vitreoretinal Surgery In Patients With Closed-Globe Injury. Eur J Ophthalmol . 2010

[44] Kara Bb, Açikel Ch. The Effect Of Intradialytic Food Intake On The Urea Reduction Ratio And Single-Pool Kt/V Values In Patients Followed-Up At A Hemodialysis Center*. Turkish J Med Sci. The Scientific And Technological Research Council Of Turkey; 2010;40(1):91–7.

[45] Iyigun E, Bayer A, Tastan S, Demiralp M, Acikel C. Validity And Reliability Study For The Nei-Vfo-39 Scale In Chronic Ophthalmic Diseases--Turkish Version. Acta Ophthalmol . Blackwell Publishing Ltd; 2010 Jun

[46] Topcu Ft, Erdemir U, Sahinkesen G, Yildiz E, Uslan I, Acikel C. Evaluation Of Microhardness, Surface Roughness, And Wear Behavior Of Different Types Of Resin Composites Polymerized With Two Different Light Sources. J Biomed Mater Res B Appl Biomater . Wiley Subscription Services, Inc., A Wiley Company; 2010 Feb

[47] Ozgurtas T, Aydin I, Turan O, Koc E, Hirfanoglu Im, Acikel Ch, Et Al. Vascular Endothelial Growth Factor, Basic Fibroblast Growth Factor,

Insulin-Like Growth Factor-I And Platelet-Derived Growth Factor Levels In Human Milk Of Mothers With Term And Preterm Neonates. Cytokine . Elsevier; 2010 May

[48] Sütçü Çiçek H, Gümüs S, Deniz Ö, Yildiz S, Açikel Ch, Çakir E, Et Al. Effect Of Nail Polish And Henna On Oxygen Saturation Determined By Pulse Oximetry In Healthy Young Adult Females. Emerg Med J . Bmj Publishing Group Ltd And The British Association For Accident & Emergency Medicine; 2010 Sep

[49] Tastan S, Iyigun E, Bayer A, Acikel C. Anxiety, Depression, And Quality Of Life In Turkish Patients With Glaucoma1. J Inf . 2010 Apr

H index (WOS): 24

LANGUAGES

English (Advanced)
Spanish (Intermediate)