Summer 8-15-2016

Gene Association Mapping in the Era of Next-Generation Sequencing and Systems Biology

Tianxiao Zhang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
John P. Rice, Chair
Laura J. Bierut
Donald F. Conrad
Christina A. Gurnett
Nan Lin
Nancy L. Saccone

Gene Association Mapping in the Era of Next-Generation Sequencing and Systems Biology

by
Tianxiao Zhang

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2016
St. Louis, Missouri

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Acknowledgments

I would first like to thank and acknowledge the work and guidance of my mentor, Dr. John P. Rice. He has been a tireless supporter and advocate for me throughout my graduate training, constantly seeking ways to improve my work and to provide opportunities for me to grow as a scientist. My thesis work has not been possible, if it were not for his guidance and the environment he provided. Working with John has been truly inspirational.

I would also like to thank Dr. Christina A. Gurnett, Dr. Laura J. Bierut, Dr. Donald F. Conrad, Dr. Nan Lin, Dr. Nancy L. Saccone and Dr. Ingrid Borecki for serving (or had been served) on my thesis committee. I would like to give special thanks to Dr. Christina A. Gurnett for serving as a chair of my thesis committee and providing with very helpful advice at each stage of my thesis.

I would also like to thank past and present members of John's lab. If it were not for Xu Zhang, Fengxian Wang, Bill Howells, Lingwei Sun, Peter Jones, Sue Winkeler, Xiong Xu and Scott F. Saccone, it would have been much harder for me to carry out my graduate work.

Tianxiao Zhang

Washington University in St. Louis

April 2016

Dedicated to my wife, Bo, and my son, Yiwei.

ABSTRACT OF THE DISSERTATION

Gene Association Mapping

in the Era of Next-Generation Sequencing and Systems Biology

by

Tianxiao Zhang

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2016

Professor John P. Rice, Chair

In the past decade, advancement of genotyping technology, first microarray then "next-generation" sequencing, has enabled scientists to examine the susceptible genes that contribute to the risk of complex disorders using a genome-wide, "hypothesis free" strategy. However, despite this "hypothesis free" label, these genome-wide approaches (including genome-wide association and whole genome sequencing studies) depend on two implicit assumptions. The first assumption is that the genetic risk of complex traits is contributed by independent genes/variants (assumption of independence).The second assumption is that different genes have equal potentiality to confer to the genetic predisposition of the complex traits (assumption of equality). Despite the huge success in susceptible gene association mapping in the last decade, more and more evidence has indicated that these two underlying assumptions of these genome-wide approaches may not be sound. Other than just studying one locus at a time, alternative methods which can carry out global analyses of biological molecules in populations

have been developed to understand the influence of the whole biological system on complex traits. Network based approaches, in particular, have proven informative.

This dissertation will cover a few important issues concerning sequencing based study design and its applications in chapter II, III and IV. Human protein-protein interaction network will be constructed and a few of human gene network related issues will be studied and discussed in chapter V and VI. Abstracts for each chapter were summarized as followed.

Chapter 2: In this chapter, we proposed a two-stage, gene-based method for association mapping of rare variants by applying four different non-collapsing algorithms. Using the Genome Analysis Workshop 18 whole genome sequencing dataset of simulated blood pressure phenotypes, we studied and contrasted the false positive rate of each algorithm using receiver operating characteristic curves. The statistical power of these methods was also evaluated and compared through the analysis of 200 simulated replications in a smaller genotype data set. We showed that the Fisher's method was superior to the other three 3 non-collapsing methods, but was no better than the standard method implemented with famSKAT.

Chapter 3: In this chapter, we aimed to identify potential susceptibility variants for bipolar disorder via the combination of exome sequencing and linkage analysis on 6 related subjects from a four-generation family. Our study identified a list of five potential candidate genes for bipolar disorder. Among these five genes, *GRID1* (Glutamate Receptor Delta-1 Subunit), which was previously reported to be associated with several psychiatric disorders and brain related traits, is of particular interest. Our findings suggest a potential role for these genes and the related rare variants in the onset and development of bipolar disorder in this one family.

Chapter 4: In this chapter, we investigated the potential of FMO genes to confer risk of nicotine dependence via deep targeted sequencing in 2,820 study subjects comprising of nicotine 1,583 dependents and 1,237 controls from European and African Americans. Specifically, we focused on the two genomic segments including *FMO1*, *FMO3* and the pseudo gene *FMO6P*, and aimed to investigate the potential association between FMO genes and nicotine dependence. We identified different clusters of significant common variants in European (with most significant SNP rs6674596, *P*=0.0004, OR=0.67, MAF_EA=0.14) and African Americans (with the most significant SNP rs6608453, *P*=0.001, OR=0.64, MAF_AA=0.1). Most of the significant variants identified were SNPs located within intronic regions or with unknown functional significance.

Chapter 5: In this chapter, we aimed to investigate the followed three scientific questions: 1) Can centrality reflect the biological significance of genes in a general human gene network? 2) Among these four commonly used centrality measures, does any of them outperform others? 3) Will they do better if we combine several centrality measures together using machine learning algorithms? To answer these scientific questions, we constructed a comprehensive human gene-gene network using protein-protein interaction data. Four essential gene sets were extracted from a variety of data sources serving as true answers in the evaluation and optimization process. Our analytic results indicated that there is a connection between the essentiality and centrality of human genes. A pattern of strong correlations was identified among the four commonly used centrality measures for a general human PPI network and the performance of each centrality measure was similar to others serving as predictors of the essentiality of genes. The improvement of the prediction models was limited when we combined several different centrality measures.

Chapter 6: In this chapter, we aimed to investigate the potential enrichment pattern in centrality of susceptible genes for certain complex disorders in a functional specific sub-network. Gene expression data of human brain tissue recorded in the Human Protein Atlas were extracted and utilized to construct a series of brain function specific sub-networks. Susceptible genes from

three categories of complex disorders, including neurodegenerative disorder, psychiatric disorder and non-brain related disorder, were extracted from the GWAS catalogue. We identified a significant enrichment pattern of high centrality of susceptibility genes contributing to neurodegenerative and psychiatric disorders in these sub-networks. Our findings indicate that susceptibility genes of complex disorder might have higher centralities in functional specific sub-networks.

Chapter 0: Prologue

“On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street...”

—John Snow, *On the Mode of Communication of Cholera*, 1855

“My experiments with single traits all lead to the same result: that from the seeds of hybrids, plants are obtained half of which in turn carry the hybrid trait (Aa), the other half, however, receive the parental traits A and a in equal amounts. Thus, on the average, among four plants two have the hybrid trait Aa, one the parental trait A, and the other the parental trait a. Therefore, 2Aa+A+a or A+2Aa+a is the empirical simple series for two differing traits.”

—Gregor Mendel, *Letter to Carl Nägeli*, 1866

“A LADY declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider this problem of designing an experiment by means of which this assertion can be tested.”

—Sir Ronald A. Fisher, *The Design of Experiments*, 1935

Chapter 1: Overview

1.1 Genetic Epidemiology : Its Origin, Definition, and Early Development

1.1.1 What is genetic epidemiology?

Genetic epidemiology, as the name suggests, is an inter-discipline that is related to both epidemiology and genetics. I would like to describe a genetic epidemiologist as an epidemiologist who tries to unravel the enigma of (human) genetics using the tool sets of statistics. In this sense, genetic epidemiology, as a field of study, can be traced back to three origins. The first one is modern epidemiology. The core methodology and terminology used in genetic epidemiology are directly borrowed from it. Another one is genetics. Unraveling the genetic (and the environmental) determinants of human traits and disorders are the major goals of genetic epidemiology. Last but not least, statistics. Statistics is the fundamental tool utilized in genetic epidemiology study, and statistical estimation is required in most, if not all, genetic epidemiology related publications. Just as other thriving disciplines, genetic epidemiology is also an evolving study subject that keeps on adapting to the progress of biomedical science in modern days. Besides the three origins I mentioned above, genomics and bioinformatics are also involved in genetic epidemiology studies in the 21th century.

From the perspective of epidemiology, there are mainly two kinds of study designs: observational and experimental. In observational studies, researchers only observe their study subjects and do not intervene. While in experimental studies, a researcher, instead of observing from the sidelines, controls the factors affecting a certain case study [1]. Genetic epidemiology, as a concept used in this thesis (and in most academic scenarios), is strictly observational. This perspective actually provides us a chance to distinguish genetic epidemiology and human genetics. Human genetics offers a wider concept compared to genetic epidemiology. Besides observational studies (genetic epidemiology), it also includes experimental studies performed using model organisms. Both study designs have its advantage and disadvantage. Confounders (population stratification, e.g.), especially those unknown, can always be problematic for observational studies (genetic epidemiology). They place obstacles for the “giant leap” from statistical association to causal inference. On the other hand, experimental studies (based on model organisms) could handle this confounding issue very well. However, generalization is a major weak point of these kinds of studies when researchers try to trace the significance of some causal variants identified in model organisms (mouse, e.g.) using gene knockout/knockdown technology back to human beings.

Genetic epidemiology studies try to clarify the following logically ordered scientific questions about human traits\disorders:

1) Does this trait/disorder have a pattern of familial aggregation (family clustering)?
2) Can shared genes explain the familial aggregation?
3) How much can these shared genes explain the familial aggregation?
4) Where are these genes and how do they contribute to the trait/disorder?

A first observation of genetic related traits/disorders is always their familial clustering pattern. However, this pattern is not necessary due to shared genes but could also be share environments. Several study designs, including immigration studies and adoption studies, can provide a way to separate the genetic variance from the environmental variance. After knowing the fact that genes matter, the next question is that how much these genes could explain the familial clustering pattern. Twin studies are a common study design to answer this question. Once we know the phenotypic variance that could be explained by genetic variance, then where are these genes? Linkage studies and gene association mapping could answer this question. Genetic data are not needed to answer the first three questions. I will briefly discuss them in the next section.

1.1.2 Early genetic epidemiology studies: nature versus nurture

Before scientists took a great interest in it, family aggregation as an observation from everyday life has long been noticed and characterized in some idioms, such as *Like father like son* or *A wise goose never lays a tame egg*. The early genetic epidemiology studies can be traced back to the work of Francis Galton, who was half-cousin of Charles Darwin, in the 19th century. Galton was interested in answering the question whether human traits were hereditary. He devoted most of his academic life to devise large-scale data collection of different measurements of human traits, from mental characteristic to intelligence. Galton proposed that if eminence was hereditary, there should be more eminent men among the relatives than among the general population, and the numbers of eminent relatives dropped off when going from first degree to second degree relatives, and from second degree to third [2]. This is a typical familial aggregation study design, although it is not enough to test whether a specific trait is hereditary because of the mixed genetic and environmental variance. Galton recognized the limitations of his methods in his works, and believed the question could be better studied by comparisons of twins. He also proposed adoption studies.

An Adoption study investigates the similarity between the adoptees and their biological and adoptive parents. The similarity between adoptees and their biological parents is expected to be heritable, while similarity with their adoptive parents is shared environmental effect. With this study design, an adoption study can separate the effects of heredity and environment. However, it is difficult to link an adopted child to their biological family. Therefore, a simplified version (familial design) is sometimes applied by comparing the non-biological siblings who are reared in the same household. Additionally, another study design, the so called “immigration study” can do the same thing by comparing the phenotype of immigrants, populations in their original countries and populations in their resident country. The phenotypic similarity between immigrants and populations of their original country can be explained by the shared genetic background, while similarity with populations of their resident country can be explained by the shared environments.

Both adoption and immigration studies could offer us a way to deduce whether observed variation in a particular trait is due to environmental or to biological factors (sometimes popularly expressed as the "nature versus nurture" debate). The next question would then be how much of the variation in a human trait is due to variation in genetic factors. The portion of phenotypic variance that can be explained by genetic factors is summarized as the concept of heritability. Traditionally, heritability can be estimated from empirical data and simple study designs, such as the correlation of offspring and parental phenotypes, the correlation of full or half siblings, and the difference in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs. In the past decades, the heritability of many human complex disorders/traits has been estimated and summarized them in Table 1.1.

A major feature of these early epidemiology studies is that no genotypic data is involved. Therefore it is impossible to map the susceptible/causal genes that contribute to those human traits/disorders. This deficiency will only be remedied when there is a set of genetic markers which cover the human genome and a cost effective experimental technology to genotype them. The advancement of DNA technology in the 1980s and 1990s meet these two conditions and enables the genetic epidemiologist to conduct research to finally locate the susceptible/causal genes in the human genome.

1.2 Gene Mapping: From Linkage study to Genome-wide Association Study

1.2.1. Linkage and candidate gene based association study

The goal of a linkage study is to identify the genetic linkage between genetic markers and potential trait/disorder loci occurred during meiosis. The genetic linkage segment can range from a couple to a dozen of megabases (Mb), and this build-in mechanism determines that a linkage study can only identify a genetic locus that covers several Mbs on a chromosome. Short tandem repeats (STR), which are genetic polymorphisms that consist of a unit of 2 to 13 nucleotides repeated hundreds of times in a row on the DNA strand, is an efficient genetic marker for conducting linkage study. STRs are multi-allelic genetic markers, and that means they are much more informative compared to single nucleotide polymorphisms (SNPs) of which most are biallelic. Genotying of 400 STRs is enough for a typical genome-wide linkage scan while it may take more than 3,000 SNPs to achieve the same statistical power. Compared to a linkage study, an association study, which is based on linkage disequilibrium (LD), is more accurate and can pinpoint the location of specific susceptible genes. Nevertheless, it was too expensive to have a half million SNPs genotyped for one subject when conducting a genome-wide association study (GWAS) in the 1990s. To make a compromise between the experimental costs and accuracy in gene association mapping, a very popular association mapping strategy back to the 1990s (and early 2000s) was to conduct a genome-wide linkage scan first, and the significant loci identified in the linkage study, which is a chromosomal region of around 10-20 cM, was scrutinized in a candidate gene based association study based on a set of dozens of SNPs selected within the candidate significant region (usually 10-20 genes). Although this study design is logically sound and financially feasible, it is not a systematic solution to identify genes for human complex disorders/traits. An insightful review published in 2012 estimated that the total money spent on candidate gene based association studies and linkage studies in the 1990s and the 2000s exceeded $250M, but had generated very limited findings compared to the findings of GWAS in its first five years (2007-2011,) which also spend around $250M [3].

1.2.2 Genome-wide association study

Although the first GWAS results were published in 2005 and 2006, GAWS as a theoretical design had been proposed by Risch and Merikangas ten years earlier. In their 1996 landmark paper, they showed that an association study performed with one million variants genotyped in a set of unrelated subjects will be more powerful than the genome linkage scan that was widely utilized in gene association mapping studies at that time [4]. The breakthrough in SNP genotyping using microarray technology [5] finally turned this once theoretical design into a real one. The first published GWAS study was a study conducted on age-related macular degeneration (ARMD) [6]. What may amaze researchers today is that this study has successfully identified a significant locus (and it is proved that this locus contribute largely on the risk of ARMD) with around 100,000 SNPs genotyped in only 96 cases and 50 controls [6]. After that, genome-wide association studies (GWAS) have rapidly become a standard method for discovering susceptible genes for a variety of complex disorders/traits, and it is widely believed to be a promising tool for identifying potential susceptible loci. So far, many GWAS studies are published annually. As of June, 2015, 2,414 GWA studies in total have been published [7]. Around 10,000 susceptible loci have been reported to be significantly associated with around 1,000 complex human traits/disorders [7]. Up to now, most of our knowledge of susceptible genes that contribute to the human complex disorders/traits was generated by GWAS.

Despite undisputable successes of the genome-wide approach mentioned above, GWAS is sometime criticized for its focusing on common SNPs while ignoring rare and structural variants which may have large effects on complex traits[8][9][10]. Considerable evidence has shown that rare variants and structural variants may have significant effects on the onset and development of complex disorders, however, this evidence is selectively omitted and only common variants(minor allele frequency ≥ 0.01) are considered in GWAS due to a pure statistical concern (to maximize the statistical power in analysis) and genotyping technology limitations[11]. In the past six years, this challenge has been partly resolved by the development of sequencing technology. High-throughput sequencing technologies, or so called “next-generation” sequencing(NGS) technologies, which process millions of sequence reads in parallel, provide monumental increases in speed and volume of generated data at a relatively acceptable cost[12][13]. Fewer GWAS were published annually since 2012, and this trend is coincidence with the popularity of sequencing based studies (Figure 1.1).

1.3 Next-Generation Sequencing based Gene Association Mapping: Promising and Pitfalls

Advancement of DNA genotyping technology has greatly promoted the research of genetic epidemiology and gene mapping in the last 30 years. Cheaper and faster DNA genotyping technology enables some once theoretically genetic epidemiology study designs, such as genome-wide linkage scan and GWAS, to be done. In this sense, NGS enables the researchers to capture the information of every single variation in the human genome. The 1000 genome project, a public population genetics project using NGS technology, has shown that there are more than 88 million variations, including 84.7 million SNPs, 3.6 million short insertions/deletions (indels) and 60,000 structural variants, in the human genome [14]. Compared to this number, genome linkage scan only involves about 400 STRs and GWAS only examines 300,000~600,000 common SNPs. The idea of a whole genome sequencing (WGS) study is preferable to the previous gene mapping study design in completeness by measuring every variation in human genome. Several WGS studies focusing on relatively small number of subjects with psychiatric disorders

[15][16] and cancers [17] [18][19] have been published in high profile academic journals and novel findings were reported.

Despite its advantage and promising future mentioned above, sequencing-based association mapping, at least in its current stage, still has two issues left unaddressed. Firstly, unlike linkage and GWAS, there are no matured statistical analysis methods can be applied to sequencing data. A major challenge for sequencing based data is that there are a lot of genetic variants with low allele frequency. That means the single marker based analysis methods utilized in GWAS cannot be applied directly to sequencing-based data due to lack of statistical power. A common approach to overcome this issue is to collapse rare variant information within specific genomic regions (genes, for example), and these methods are generally called collapsing methods [20][21][22]. I will provide more details about these collapsing methods and conducted a comparison study to investigate the efficient of multiple sequencing data analysis methods in Chapter II of this thesis. The second issue concerns money. Although NGS has greatly reduced the experimental cost of human genome sequencing, it still takes 5-10 times of experimental cost to have a human genome sequenced (WGS) compared to genotyped by a microarray panel (used in GWAS). Two other study design can partly address this issue. The first one is instead of performing WGS on study subjects, researchers can perform exome sequencing which only focuses on exonic regions of human genome (1% of the human genome). This strategy can retain a large amount of genetic information of significant functional regions while having a much lower experimental cost compared to WGS. I will present an exome sequencing based study on bipolar disorder in Chapter III. Another study strategy is to sequence a couple of targeted susceptible genomic regions. On one hand, rare and structural DNA variants of targeted genomic regions can be thoroughly investigated through DNA sequencing, while on the other hand, the experimental cost can be restricted to a reasonable level. I will present a targeted sequencing study of nicotine dependence in Chapter IV.

1.4 Genetic Epidemiology Studies using Insights of Biological Networks

1.4.1 GWAS, WGS and their discontents: hypothesis "free" or "engaged"

Genome-wide approaches are described as "hypothesis free" study designs, because comparing to the candidate gene based approach, this study strategy does not need prior knowledge about the candidate genes. A genome-wide scale study enables the genotyped genetic markers to offer sufficient coverage to most of the human protein coding genes (if not all) [23]. The "hypothesis-free" basis of genome-wide approaches offered the opportunity to overcome difficulties and obstacles imposed by the incomplete understanding of disease pathophysiology. However, despite this "hypothesis free" label, these genome-wide approaches (including GWAS and WGS) are somewhat dependent on some underlying hypotheses.

The first underlying assumption is that the genetic risk of complex traits is contributed independently by genes/variants (assumption of independence). For example, in GWAS, the simplest analysis strategy is to do logistic regression in a single-locus manner [23], and in association mapping based on DNA sequencing data, the so-called "collapsing method" was widely utilized for which variants information are often collapsed within a genomic region or gene and then each region/gene will be tested pointwisely [21].However, in the past decade, many studies of quantitative traits in animal models suggest that epistatic interactions between loci are widespread[24][25][26], and various examples of gene-by-gene interactions for human complex traits have been identified[27].A research study using yeast strains provided an estimated importance of epistatic interactions for 46 highly heritable traits. It shows that the contribution of gene-gene interactions (including both two-loci interactions and high order interactions) varies from zero to ~50% and detected two-locus interactions explain only a minority of this contribution [28]. In this study, the researchers have used yeast, which is a simple unicellular organism, as their research subject. We can expect that in some higher level multi-cellular organisms such as mammals, the patterns of gene-gene interactions will be more complex. The second assumption is that for GWAS/WGS, different genes are considered to have equal potentiality to confer the genetic predisposition to the complex traits (assumption of equality). This assumption has been partly challenged by the evidence that susceptible genes often show a clustered pattern within certain biological pathways [29][30]. In addition, previous studies conducted in several different model organisms [31] have shown that highly connected proteins or "hubs" are more likely to be encoded by essential genes which are necessary for fundamental processes in an organism and lead to pre- or neonatal lethality when disrupted.

To conclude, all of the above evidence has indicated that genes are neither created equally nor perform their functions independently, and novel insights are needed for understanding the mechanisms of genetic predisposition for complex traits. Other than simply studying one locus at a time, alternative methods which can carry out global analyses of biological molecules in populations have been developed to understand the influence of the whole biological system on complex traits[32][33]. Network based approaches, in particular, have proven informative.

1.4.2 Human Gene Networks: Novel Insights into the Genetic Epidemiology Study of Complex Traits

Human gene networks are graphical representations of the interactions between the genes. In a human gene network, genes are represented as nodes and the relationships among them as edges. Human gene networks can be divided into three categories: 1) Human gene networks derived from curated knowledge; 2) Human gene networks based on experimental data of physical interactions, and 3) those that are inferred from high-throughput data [31]. In the past decade, the most interesting finding of gene network analysis is that proteins that are encoded by essential genes tend to have high centrality degrees in the protein-protein interaction network. This feature was firstly identified in yeast [34][35]. Since then, several studies were conducted to focus on the phenotypes related to human diseases. Wachi et al. studied genes that are differentially expressed in lung squamous cancer tissues, and found that up-regulated genes in the cancerous tissues tended to be highly connected and central [36]. Another study in 2006 investigated the network position of 346 genes that had been implicated in a comprehensive census of all human cancer genes. They showed that on average the proteins encoded by these genes tended to have twice as many interaction partners as noncancer related genes [37]. Nevertheless, in a 2007 published research paper, Goh et al. created a network of human disease/human gene associations, in which each genetic disease is connected to the genes known to cause it. They found that most of the disease genes have no tendency toward higher degree in the human protein-protein interaction network [38].One possible explanation for this discrepancy among the studies above is that the former two studies focused on cancer genes in particular while Goh et al. investigated disease in general.

Among all these early studies, a significant limitation is that all these studies have focused on cancer or Mendelian disorders, and the researchers have paid limited attention to complex disorders. This limitation can be justified by the specific time when these research projects were conducted. Most of the knowledge of susceptible genes on complex disorders has been generated after 2007 when GWAS become a standard genetic epidemiological research strategy on complex disorders. In addition, recent evidence of interactome networks in the last decade has also questioned the potential incompleteness in these previous studies. Large scale, comprehensive analysis incorporating with new findings obtained in the past decade is needed. In chapter 5 of this thesis, I will first examine the centrality measurements of genes as indicators of their biological significance in a human general gene network. Then, in chapter 6 I will explore the gene sets centrality feature for different complex disorders and functional pathways enrichment patterns in general human gene networks and disease-specific sub-networks.

1.5 Figures

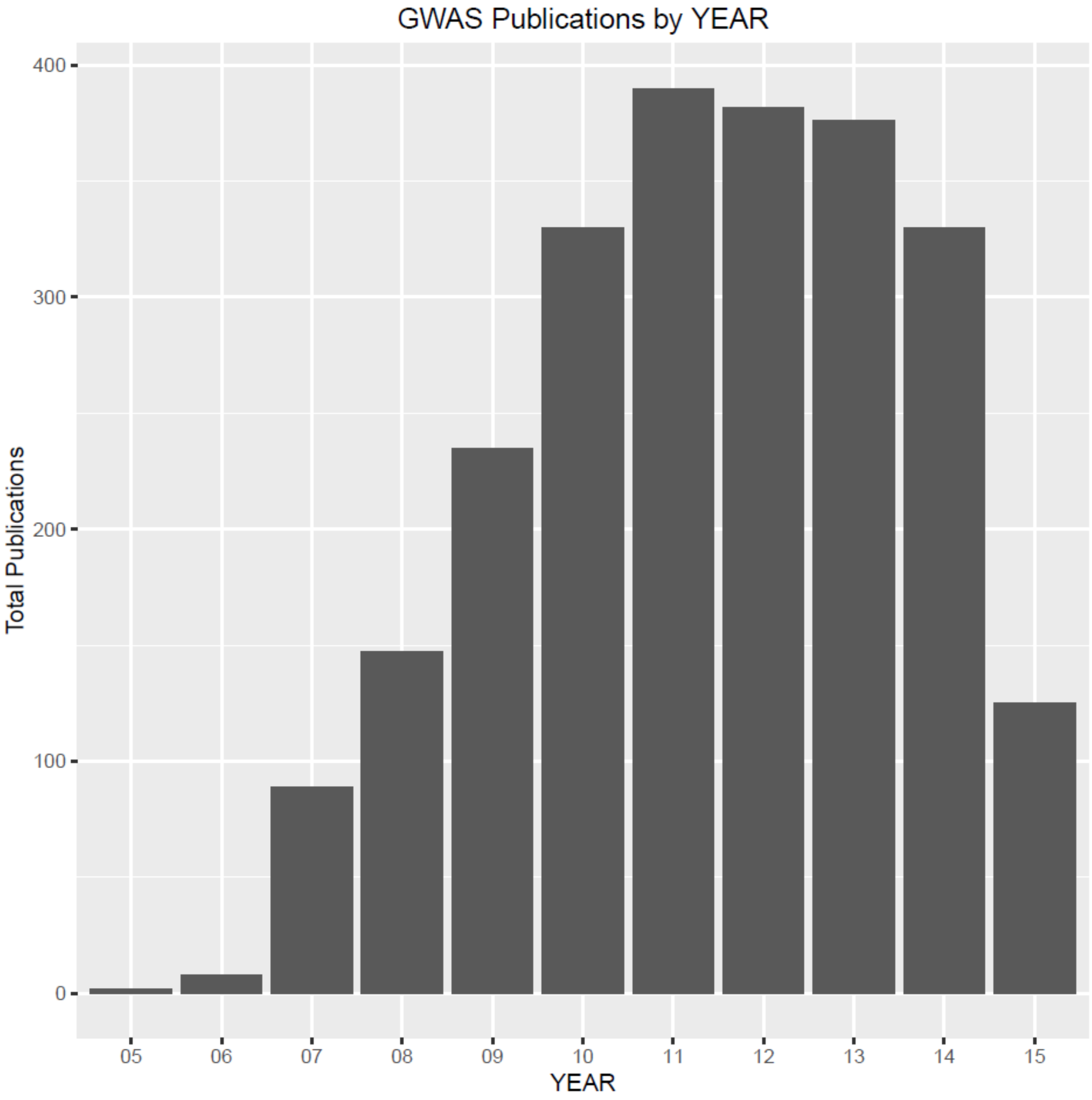

Figure 1.1 Histogram of GWAS publications by year. The publications of 2015 were included from January to June.

1.6 Tables

Table 1.1 Selected human disorders/traits with estimated heritability.

Traits/Disorders	Heritability (%)
Acne	81
Age-related macular degeneration	49 - 71
Alcoholism	50 - 60
Alzheimer's disease	58 - 79
Asthma	30
Attention deficit hyperactivity disorder	70
Autism	30 - 90
Bipolar disorder	70
Bladder cancer	7 - 31
Blood pressure, diastolic	49
Blood pressure, systolic	30
Body mass index	23 - 51
Bone mineral density	44 - 87
Breast cancer	25 - 56
Celiac disease	57 - 87
Cervical cancer	22
Chronic obstructive pulmonary disease	76
Colon cancer	13
Coronary artery disease	49
Depression	50
Epilepsy	70 - 88
Eye color	98
Heart disease	34 - 53
Height	55 - 81
Hypertension	30
Leukemia	1

Longevity	26
Lung cancer	8
Nicotine dependence	60
Obesity	70
Ovarian cancer	40
Parkinson's disease	25 - 30
Periodontitis	42
Prostate cancer	42
Psoriasis	66
Schizophrenia	81
Stomach cancer	1
Stroke	32
Testicular cancer	25
Thyroid cancer	53
Type-1 diabetes	88
Type-2 diabetes	26

Data source: SNPpedia(http://www.snpedia.com/index.php/SNPedia).

Chapter 2: Application of Non-collapsing Methods to Gene-based Association Test

2.1 Introduction

Unlike GWAS which focuses on common SNPs that have relatively higher MAF, sequencing based association study could generate tons of rare or low frequency variants and traditional statistical methods often fail in association mapping due to poor statistical power. To address this issue, as introduced in Chapter I, one commonly utilized strategy is to collapse rare variants information within specific genomic regions (such as genes), and this "super marker" will be tested statistically[20][21][22]. This strategy could partly solve the issue of statistical power, however, it has assumes the directions of the effects of DNA variants are consistent, and this assumption may not be true. Compared to these collapsing methods, the non-collapsing methods, which do not require the assumption of consistency of effects direction, may be more reasonable choices for rare variants based association mapping.

In this chapter, we proposed a two-stage, gene-based method for association mapping of rare variants by applying four different non-collapsing algorithms using the whole genome sequencing dataset and simulated blood pressure phenotype of genome analysis workshop (GAW) 18[39]. Genetic analysis workshop provided a platform for developing and evaluating statistical methods to analyze population and family based human genetics data. It is held every other year. GAW18 focused on identification of genes and functional variants that influence complex phenotypes in human sequence data.

In this research we will first obtain significance *P* values by fitting a mixed effects model for each variant, and then apply four non-collapsing algorithms, including Fisher's, gene set enrichment analysis (GSEA), sequence kernel association test (SKAT), to obtain the gene-wise association *P* values. Collapsing (or burden) methods combine variant information by assuming consistent direction of effects across variants. None of the methods considered here adopt this assumption, although some do combine variant information.

2.2 Materials and Methods

2.2.1 Model fitting and algorithms

A mixed linear model was fitted for each variant as described in previous literature [40]. The model was defined as:

$$Y = X\beta + Qv + Z\mu + \varepsilon \quad （2.1）,$$

where *Y* is the quantitative trait of interest (we used first-visit systolic blood pressure [SBP]); *X* is the genotype; β is the fixed effects of the genotypes; and *Q* represents the population structure variables . In this study, we chose the first 10 principal components from principal component analysis (PCA) for *Q*. *v* is the fixed effects of *Q*; *Z* is the variable that evaluates familial relatedness (the theoretical kinship matrix was used for *Z*); and μ is the random effects coefficient for *Z* that corrects the polygenic impact.

After obtaining the variant-wise *P* values by fitting the mixed linear model shown above, four non-collapsing algorithms were modified and applied to the data set to obtain the gene-wise association *P* values. The algorithms of the four methods are summarized as followed:

1. Naïve method. The most significant variant-wise *P* values within a specific gene were chosen as the gene-wise association *P* values.

2. Fisher's method [41]. The gene-wise statistics were calculated through the following equation:

$$X = -2\sum_{i=1}^{k} \log_e(p_i) \qquad （2.2）,$$

where p_i is the *p* value for variant *i*, and *k* is the total number of variants within a specific gene. Because many variants are highly correlated, the basic assumption of independent tests for Fisher's method is violated. Fisher's formula may not have a chi-square distribution, so we assessed the significance via permutations.

3. Simes' method [42]. The gene-wise *p* value was summarized by the following equation:

$$P_{simes} = \min_i \{\frac{kp_i}{i}\} \qquad （2.3）,$$

where p_i is the *p* value for variant *i*, and *k* is the total number of variants within a specific gene.

4. GSEA method [43][44]. The test statistics (indicated as ES score) were aggregated from variant-wise *p* values within each gene via a Kolmogorov-Smirnov–like process in which running sums are accumulated. The equation is given as:

$$ES(S) = \max_{1 \le j \le N}\{ \sum_{G_{j^*\in S}, j^* \le j} \frac{\left| r_{(j^*)} \right|^p}{N_R} - \sum_{G_{j^*} \notin S, j^* \le j} \frac{1}{N - N_H} \} \qquad (2.4)$$

where *N* is the total number of variants, *r(j)* is the j^{th} largest statistic values, N_H is the variant number of a given gene, *S* is any given gene, P is the parameter that gives a higher weight to variants with extreme statistic value, arbitrarily set to 1 in this study, and N_R is given by:

$$N_R = \sum\nolimits_{G_{j^*\in S}} \left| r_{(j^*)} \right|^p \qquad (2.5)$$

Statistical significance and adjustment for multiple hypothesis testing were assessed by a 1000 permutation based procedure. A family-wise error rate (FWER) procedure was used to adjust for multiple-hypothesis testing. In this study, the FWER *p* value was calculated as the fraction of all permutations whose highest statistics (or smallest *p* values) in all genes is higher than a given gene. In addition to the four non-collapsing algorithms introduced above, we also included two standard rare variants based methods: SKAT [21] and famSKAT [45] in our analysis. FamSKAT is an extended version of SKAT and can be utilized to analyze rare variant when family correlations are present. Furthermore, to evaluate the statistical power of these methods, we extracted the variant information related to the 22 true-positive genes located on chromosome 3 and analyzed these data for all 200 simulated phenotype replicates.

2.2.2 Data and computation

We only analyzed one phenotype replicate and sequencing data of chromosome 3 due to the huge computational burden. The sequencing data were annotated by

ANNOVAR [46]. Intergenic variants (variants at least 1 kilobase [kb] away from any known gene regions) were excluded, but variants that can be mapped to regulatory regions (ORegAnno) were kept [47].

To preserve the familial structure, a permutation-of-residuals procedure was applied [48]. First, we fitted a mixed effects linear model on the phenotypic data with all the covariates in the model (except for genotype term) and preserved the residuals for these models. Second, we shuffled the residuals (rather than the phenotypic data used in an ordinary permutation procedure) and randomly assigned them to each subject and generated 1000 phenotypic data replicates. And third, we obtained the permuted statistics and *p* values by fitting a univariate linear model with genotype as the only predictor of the residuals. This method may introduce potential bias to the permuted statistics and *p* values comparing to directly fitting the full model. To quantify this potential bias, we randomly chose 1429 variants and calculated the percentage difference of the −log10 scaled *p* values obtained from directly fitting a full model and from the two-step permutation procedure proposed above. The results of the permutation bias analysis showed that the percentage difference was only approximately 10%, and the correlation coefficient of variant-wise statistics was 0.9959. These results indicate that the effects of this bias are limited.

Genotypes were coded in dominant model. That is, the genotypes with 1 or 2 minor alleles were coded as 1, while genotypes with 2 major alleles were assigned 0. Variants with minor allele frequency >0.3 in genome-wide association data set were selected for PCA. We used Eigenstrat 3.0 for this analysis [49]. The R package kinship2 (http://cran.r-project.org/web/packages/kinship2/index.html) was used to calculate the kinship coefficient matrix for our data set. The R package coxme (http://cran.r-project.org/web/packages/coxme/index.html) was implemented for fitting the mixed linear model. The R package SKAT (http://cran.r-project.org/web/packages/SKAT/index.html) was implemented for rare variant analysis with SKAT. The R source code for famSKAT was downloaded (http://www.bumc.bu.edu/linga/research/publications/famskat/) and implemented for rare variant analysis. Receiver operating characteristic (ROC) curves were made and compared among the four algorithms and two standard methods.

2.3 Results

The data consisted of 1,237 genes with 87,190 variants that passed the annotation criteria were extracted from the sequencing data set of chromosome 3 for 849 subjects. After fitting the mixed linear model, the Q-Q plot and histogram of p values of these 87,190 variants is shown in figure 2.1. Data for the 22 true positive (true answer) genes with 1,098 variants were extracted and used for analysis with 200 simulated phenotype replicates. The statistical power information for all the six methods was summarized and is presented in table 2.1. From the power analysis results, we see that the gene *MAP4* was successfully identified to be significant for all simulated 200 replicates. All six methods achieved 100% power for this gene. For the rest 21 genes, the largest power was 27.5%, which was achieved by SKAT for *LOC152217*.

To compare the four non-collapsing methods and the two standard methods, ROC curves based on these six methods were constructed and shown in figure 2.2. From this fugure we noted that, overall, the Simes' method performed a little better compared to the other five methods, and that GSEA, SKAT and famSKAT did not perform as well as Simes' method. The other two methods were slightly better than GSEA, SKAT and famSKAT method. However, when we limited the false positive rate to be smaller than 0.1 as shown in the right hand plot of figure 2.2 (in practice, only a high true positive rate with a low false positive rate is of interest), we see that Fisher's method and famSKAT performed better than other methods at the low false positive rate range. They both capture around 15% of the causal genes (true positives) with a cost of only 5% false positive signals. However, we did not test the significance of the ROC curves, so that all these observed differences could just be noise.

2.4 Discussion and Conclusion

MAP4 was identified to be the causal gene with 100% statistical power. This result is reasonable since, according to the "answer sheet" of GAW18, *MAP4* contains the most "causal variants" and these variants have a relatively larger effect size comparing to the variants within other genes. However, this result was obtained when we only analyzed the 22 "true answer" genes. For a genome-scale analysis, the significant signals may be missed due to correction for multiple comparisons. We have also analyzed the whole genotypic dataset of chromosome 3 with simulated phenotypic replicate #1(including 1,237 genes and 87,190 variants). The result indicated that only na ïve method and the two standard methods identified gene *MAP4* to be significant.

The non-collapsing methods introduced in this paper have been broadly utilized in testing the significance of biological pathways in GWAS datasets. When we substitute the term "pathway" in these non-collapsing algorithms for the term "gene" in sequencing analysis and “gene” for “variants”, we can apply these non-collapsing algorithms to gene-based association detection through modifications. An obvious advantage of aggregating *p* values (or statistics) by applying non-collapsing algorithms compared to ordinary variants collapsing methods is that it is a method free of the assumption that all the causal variants from a gene have effects in the same direction. This assumption may not be held in many scenarios even though it is the assumed in many existing rare variants association mapping procedures.

In this study, we utilized the residuals-of-permutation procedure to deal with our familial based data. Conducting a permutation on family data has been a challenge in statistical genetics research. Ordinary permutation procedures have been mostly utilized in case-control data, which simply shuffle the phenotypic data and randomly assigns them to each subject, thus cannot be directly applied to family data because it destroys the family structure. In our research, instead of shuffling the phenotypic data, we shuffled the residuals obtained from fitting a linear mixed effects model without genotype. These residuals have already accounted familial relatedness in the model fitting step and therefore our permutation procedure preserved the familial structure.

Several previous researchers have already applied the non-collapsing methods proposed in our research to conduct gene-based analysis [50][51]. However, these previous works have mainly focused on common variants in GWAS dataset. As an attempt to apply these non-collapsing algorithms to gene-based association tests using sequencing data, we have demonstrated some potentially promising aspects of this approach. However, several problems remain unaddressed. One important issue is the computational intensity. In this study, we have utilized a multi-processor-computing server with 23×2.8 GHz CPU and 64GB memory in total. The most time consuming part of our analysis is the permutation-of-residuals process and linear model fitting of the permuted datasets. We have paralleled this process into 20 jobs, but it still takes around 30 hours to complete (this is only the work done for one chromosome). Compared to the permutation process, the *p* value combination step can be completed much faster (~30 minutes). Since a lot of the non-collapsing algorithms require permutation procedures to create null distribution of the statistics, it is somewhat difficult to implement them on genome-wide scale dataset. In addition, many non-collapsing algorithms cannot be utilized for a gene-based association test directly without proper modifications. The choice of parameters in non-collapsing algorithm for rare variant association detection is more an art than a science. Finally, adjustment for multiple hypothesis testing is another important issue that needs to be addressed. Our results indicate that the FWER method is too conservative. For the future work, hierarchical modeling combined with MCMC may provide better solution to the multiple hypothesis testing problems [52].

To conclude, in this study, we showed that the statistical efficiency of several sequencing data based methods were not very promising, although some of them were commonly utilized in sequencing data analyzes as standard methods. Further investigation is needed to explore the potential statistical properties of these approaches.

2.5 Figures

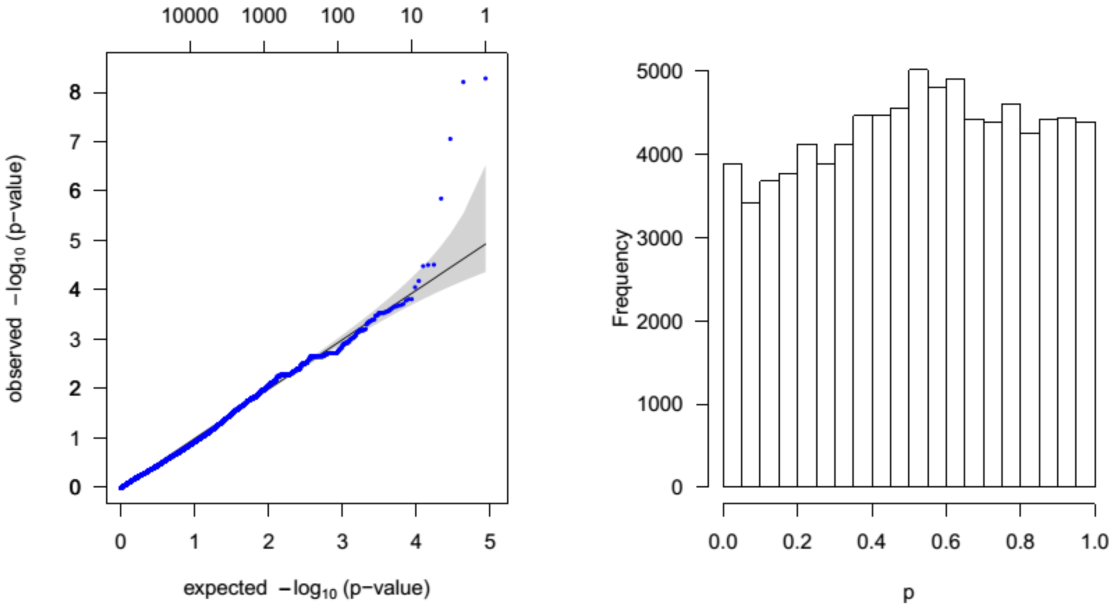

Figure 2.1 Q-Q plot and histogram for the mixed effects model. Q-Q plot (left) of –log10 scaled p-values and histogram (right) for the mixed effects model based on 1,237 genes (87,190 variants) from 849 subjects. In Q-Q plot, black line, expected; blue dots, observed.

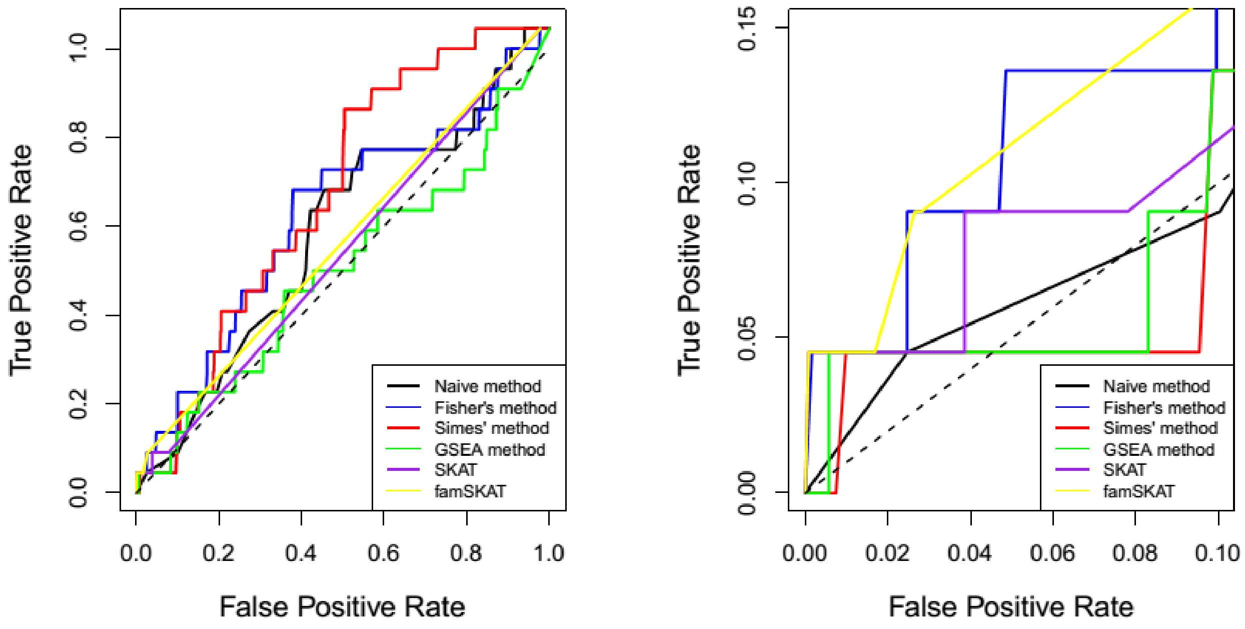

Figure 2.2 ROC curves for four non-collapsing algorithms and two standard methods. ROC curves for four different pathway algorithms based on 1,237 genes from 849 subjects on trait SBP (first visit). In the left plot FPR ranges from 0 to 1. In the right plot FPR is scaled to be less than 0.1 since only the true positive rate (TPR) with a low FPR is of interest. Black curve, na ïve method; blue curve, Fisher's method; red curve, Simes' method; green curve, GSEA method; purple curve, SKAT; yellow curve, famSKAT.

2.6 Tables

Table 2.1 Comparison of the statistical power of the four non-collapsing and two standard methods.

CHR	Gene	Power of Methods					
		Naïve Method	Fisher’s Method	Simes’ Method	GSEA Method	SKAT	FamSKAT
3	*ABTB1*	0.015	0.18	0.025	0	0.075	0.01
3	*ARHGEF3*	0	0	0	0.035	0.005	0.005
3	*B4GALT4*	0.015	0	0.015	0.035	0.01	0.015
3	*BTD*	0	0	0	0.015	0	0
3	*CXCR6*	0	0	0	0.085	0	0
3	*DNASE1L3*	0.005	0.005	0.005	0.005	0.04	0.01
3	*FBLN2*	0.005	0	0	0.035	0	0
3	*FLNB*	0.01	0.015	0	0.03	0	0
3	*LOC152217*	0.09	0.145	0.135	0	0.275	0.04
3	*MAP4*	1	1	1	1	1	1
3	*NMNAT3*	0.005	0.04	0.005	0	0	0
3	*PAK2*	0.07	0	0.05	0	0	0
3	*PDCD6IP*	0.005	0	0.005	0.005	0.04	0.03
3	*PPP2R3A*	0.045	0.01	0.02	0	0.005	0.005
3	*PTPLB*	0	0	0	0.02	0.005	0

3	*SCAP*	0.025	0.005	0.04	0	0.045	0.065
3	*SEMA3F*	0	0	0	0	0	0
3	*SENP5*	0	0.02	0.01	0.045	0.01	0.005
3	*SUMF1*	0.085	0.005	0.06	0.01	0.015	0.005
3	*TFDP2*	0	0	0	0.035	0	0
3	*TUSC2*	0.005	0	0.055	0	0.02	0
3	*ZBTB38*	0.01	0.005	0.01	0.02	0.04	0

Power is calculated based on the analysis of the 200 simulated phenotypic replicates. The largest power for each gene is highlighted in bold.

Chapter 3: Family-based Whole Exome Sequencing Study for Bipolar Disorder

3.1 Introduction

3.1.1 Clinical and epidemiological characteristic of bipolar disorder

Bipolar disorder (BPD) is a mental illness with lifetime prevalence of about 1% [53]. BPD is characterized by periods of elevated mood (manic/hypomanic episodes) and periods of depression (depressive episodes) [54]. Currently, there is no cure for BPD, and medications and therapies are used to treat the symptoms. Patients with BPD and their families experience significant losses in functional status and quality of life, placing untoward stress on personal relationships. In addition, BPD is one of the most expensive mental health care diagnosis, both for patients with the illness and for their health insurance plans [55], and that in turn adds a financial burden on the patients' families, as well as on society as a whole. Biomedical and etiological studies on the onset and development of BPD can throw light on new drug discovery and therapy development.

3.1.2 Gene association mapping of bipolar disorder: a brief review

The etiology of BPD is not clearly understood but extensive research has indicated that both genetic and environmental factors play a role [56]. Familial clustering studies have identified a ten-fold higher risk of BPD in people who have affected first degree relatives when compared to the general population [56]. The heritability of overall bipolar spectrum disorders is estimated to be 0.71[57]. More than 40 linkage scans for

BPD have been published and implicate many areas of the genome, although several studies have inconsistent results [56]. Genome-wide association studies (GWAS) have identified several susceptibility loci including markers near *PALB2* [58], *DGKH* [59], *ANK3*& *CACNA1C* [60], 3p21 [61], *NCAN* [62], *ODZ4* [63], *TRANK1& LMAN2L* [64], *ADCY2&*6q16 [65], and *SESTD1* [66]. Nevertheless, despite these findings, a recent study has estimated the SNP heritability (the proportion of variation in disease liability that is captured in GWAS by considering all SNPs simultaneously) was ~0.4 for BPD [10]. This indicates that further research is needed to unravel the genetic etiology for BPD.

The traditional microarray chip technology based GWAS focuses on common variants (genetic variants that have minor allele frequency at 5% or higher), and selectively omits the rare variants and structural variants such as short insertion and deletions (indels) due to technological problems [67]. The recent development of “next-generation” sequencing technology has enabled researchers to investigate these variants which are not covered in GWAS at a relatively lower genotyping cost [68]. Exome sequencing, which sequences the exons of protein coding genes in the genome, is considered to be a powerful tool in genetic association research [69]. By focusing only on the region of exons, exome sequencing only sequences around 1% of the human genome (far less than whole genome sequencing) while investigating genomic regions of functional significance. A recent research project focusing on lithium-responsive bipolar disorder has identified several rare susceptibility variants by exome sequencing analysis based on 36 familial samples [70]. This result indicates that exome sequencing technology combined with proper study design is a promising method for association mapping.

In this chapter, we have recruited six related BPD cases from a single BPD extended family. The rationale is that a single variant is segregating in this large, unusual family and that this approach will minimize genetic heterogeneity by restricting analysis to a single family. We have performed whole exome sequencing on the six BPD cases, and examined the DNA variants that are shared among all these six cases. Additionally, we have also performed a genome-wide high density linkage analysis based on common SNP data. The linkage peak region has further narrowed down the potential susceptibility variants and genes for BPD.

3.2 Materials and Methods

3.2.1 Study subjects

The BPD family was selected via a history of multiple relatives with BPD. An index case was recruited from department of psychiatry at Barnes-Jewish Hospital. Relatives were diagnosed via diagnostic interview (SADS-L) and two independent senior psychiatrists gave best estimated diagnosis made through consensus. Signed consent forms were obtained from all the recruited members. This study was approved by IRB of Washington University in St. Louis. The pedigree structure of the family recruited in this study was shown below (figure 3.1). Six subjects with BPD were included in this study.

3.2.2 Experimental Methods

Genomic DNA was isolated from the peripheral blood leukocytes and the DNA was stored at −80 °C for genotyping. The microarray genotyping was done by Illumina OmniExpress. We applied quality control process to remove singletons and SNPs with missing rate higher than 10%. The whole-exome sequencing was carried out using Agilent SureSelect All Exon 50Mb Target Enrichment kit and on the SOLiD System by EdgeBio. The average read depth for the six bipolar subjects were 57x. For the exome-seq data, the alignment was done using novoalignCS (V1.01.15) by EdgeBio, and the data were recalled with HaplotypeCaller (GATK v3.3) [71]. The QC was done by GATK ’s VQSR.

3.2.3 Statistical methods

We implemented parametric linkage analysis based on the software Merlin [72] to identify potential linkage peak regions with the pedigree data. The allele frequency data were extracted from Hapmap 2 CEU samples. SLINK [73] was used to simulate the linage analysis to obtain a LOD score threshold with acceptable statistical power. The potential effects of exonic SNVs were predicted using SIFT [74] and Polyphen2 [75].

The redundancy of the microarray SNP chip panel enables us to implement a 10-set replicate analysis strategy in order to reduce the number of potential false positive signals obtained from the linkage analysis. We 1) randomly selected 10 sets of SNPs with considerations of minor allele frequency (MAF> 0.3), Linkage disequilibrium (LD) structure (to account for LD) and their genomic coverage (each SNPs set contains around 10,000 markers); 2) conducted linkage analysis with each of these 10 replicated SNP sets independently; and 3) identified the peak regions that are repetitively identified in all of the ten rounds of analysis. To account for the potential impact of LD on linkage signals, we only chose one SNP from each LD block which was constructed based on the Hapmap European population data. The LD block was estimated using Plink [76].

We implemented the Perl based software ANNOVAR[46] to annotate the exome sequencing data. To investigate the potential susceptibility variants/genes within this large family, we extracted variants that 1) pass the quality control criteria specified in GATK software; 2) are under one of the linkage peak regions; 3) located within genetic region with functional significance (splicing site variants, non-synonymous SNVs, stop gain SNVs, frameshift indels, or non-frameshift indels within exonic regions) and 4) Only variants that were not recorded in 1000 genome database or variants with recorded MAF < 0.05 were included. We also incorporated our filtering results with R package RVsharing [77] to have a statistical estimate of our observed excess sharing among all the related BPD cases of those candidate variants.

3.3 Results

3.3.1 Results of linkage analysis

Simulation of the linkage analysis with SLINK showed that a LOD threshold 2.0 will only achieve 22% statistical power. To increase statistical power, we chose 1.8 as the LOD threshold with the cost of increasing the false positive rate. To control the number of potential false positive linkage signals, we implemented a 10-replicate linkage analysis strategy. Genotypes data were released for 733,202 SNPs with our Illumina microarray chip and 537,258 were left after quality control. The redundancy of these markers enabled us to implement a 10-replicate analysis strategy to reduce the number of the potential false positive linkage signals. This marker sets selection and analysis implementation was as indicated above.

The linkage analysis results of 10 replicate SNP sets are shown in table 1. As shown in table 1, four peak regions in total were identified from the linkage analysis. They are 5q33.1- 5q 33.3, 5q35.1-5q35.2, 10q21.1-10q21.2 and 10q23.1-10q23.33. Two of these 4 linkage peaks, 5q33.1- 5q 33.3 and 5q35.1-5q35.2 (figure 3.2) were only identified 3 and 1 times, respectively. This indicates that these two linkage peak signals may be false positive signals. The two consecutive peak regions on chromosome 10, 10q21.1-10q21.2 and 10q23.1-10q23.33 (10q22.3-10q23.33) were identified multiple times in the 10 replicate sets and 10q23.1-10q23.33 (10q22.3-10q23.33) was identified repetitively in all of the 10 replicate SNP sets. This result indicates that the chance that this peak region is a false positive signal is very low (figure 3.3). Here we provide a brief estimation for the false positive rate of this peak region. The LOD threshold we used here is 1.8, and this is approximately equal to *p* value <0.002 [78]. Although we tested around 10,000 SNPs, considering the potential LD among SNPs, the independent number of tests might be around 400 (this is a reasonable estimate because 400 is the number of tests when using microsatellite markers for linkage analysis). Therefore, the false positive rate for each replicate is around 0.55 ($1-0.998^{400}$). For a genomic region that is proved to be significant in all of the ten replicates, its false positive rate can be as low as 0.0025 (0.55^{10}). Therefore we concentrated those variants under this peak region (10q23.1-10q23.33) on chromosome 10. We have summarized the results of linkage analyses of each marker set in table 3.1.

3.3.2 Results of variants filtering

Genotypic data were released for 140,814 variants (table 3.2). Among the 140,814 variants, 114,432 (81.26%) variants passed the quality control specified in GATK software. Within these high quality variant calls, 444 variants were located under the linkage peak region of chromosome 10. The number of variants was reduced to 60 if we only considered those variants with potential functional significance. Among these 60 variants, a total of 15 variants were not recorded in 1000 genome database or were rare variants with MAF less than 0.05 according to 1000 genome data in Caucasian population. Nine out of these 15 variants were only identified in one of six bipolar subjects sequenced, and 6 variants were shared by 2, 3 or 4 individuals. We have summarized the information of these 6 variants in table 3.3. Tests of RVSharing indicated that three out of these six variants (shared among 3 or 4 patients) were statistically significant with *P* value <0.008 (0.05/6). The other three variants that were shared among two patients had a *P* value of 0.0695. These 6 variants come from 5 genes including *DYDC2*, *GHITM*, *MINPP1*, *CDHR1* and *GRID1*. Two SNVs located in the genes *CDHR1* and *GRID1* were predicted to be “damaging” or “possibly damaging” by SIFT and Polyphen-2.

3.4 Discussion and Conclusion

Our aim was to identify potential susceptibility variants that contribute to the risk of BPD. In order to minimize genetic heterogeneity, we restricted analysis to a single large family in which several distantly-related individuals suffer from BPD. If a single variant affecting risk for BPD is segregating in this multiply affected family, this approach should identify the variant within the larger set of segregating variants, nominating candidate genes for future studies.

We narrowed the targeted genomic region down to a ~15 MB region by restricting the analysis to a linkage peak identified on chromosome 10. This region (10q22) has been previously reported to be linked with BPD in a large scale linkage study by Fallin et al [79]. After filtering for minor allele frequency, sharing among affected relatives, and functional significance of the potential susceptibility variants, we identified a list of 6 variants within 5 genes. A total of 3 variants in 2 genes, *GRID1* and *CDHR1,* were identified in 2 of the 6 BPD cases. The two cases that shared these 3 variants are also first cousins within this four generation pedigree (individuals #142 and #5 from (figure 3.1). *GRID1* encodes a subunit of glutamate receptor channels. These channels mediate most of the fast excitatory synaptic transmission in the central nervous system and play key roles in synaptic plasticity [80]. *GRID1* has been widely investigated in multiple psychiatric disorders and brain related traits [81][82][83][84]. It was first reported by Fallin et al. that *GRID1* was significantly associated with schizophrenia and associated with BPD with suggestive significance among Ashkenazi Jewish case-parent trios [81]. A study of mice in which the *GRID1* homologue, *GluD1,* was knocked out reported that the mice were hyperactive, manifested lower anxiety-like and depression-like behavior, and robust aggression [85]. The two rare *GRID1* variants we report here (rs2306265 and rs3812645) represent the first evidence that rare variants in *GRID1* may contribute to BPD.

The other gene with a shared rare variant, *CDHR1,* belongs to the cadherin superfamily of calcium-dependent cell adhesion molecules. Its encoded protein is a photoreceptor-specific cadherin that plays a role in outer segment disc morphogenesis [86]. Mutations in this gene have been associated with recessive retinal degeneration [87] and autosomal recessive cone-rod dystrophy [88]. However, no previous study has linked this gene to any psychiatric disorders. We also identified 3 interesting variants in *DYDC2*, *GHITM* and *MINPP1.* These variants were predicted to be "benign" or "tolerated" and might not affect the protein structure. However, all of these variants were shared by 3 or 4 BPD cases in this pedigree, and their sharing patterns were significant after Bonferroni correction using the RVsharing algorithm.

We note that, given our sample size, these *P* values can only be used as a suggestive guidance when prioritizing variants for further study. One major strength of our study is that as an exome sequencing based study, we can examine both common and rare variants. Previous association studies on BPD mainly focus on common SNPs while ignoring most of the low frequency and rare variants. Restricting to common SNP, may hinder the ability to find susceptibility variants or genes. In our study, we considered both rare and common functional variants via exome sequencing technology. In addition, an advantage of our family based study design is that rare variants that segregate with reasonably high penetrance in an extended pedigree can provide a linkage signal helpful in identifying the susceptibility genes.

There are also several limitations of our study. A main limitation is that we lack familial controls. The variant sharing we utilized as a filtering strategy in our study might generate some false positive signals due to relatives sharing neutral DNA variations. Having exome sequencing data from several healthy family members of this pedigree might further narrow down our candidate gene list. It is still too early for us to make any conclusion on the potential role played by these candidate genes in the onset and development of BPD. Limited by our sample type and study design, further research is needed to replicate our finding in unrelated individuals. More research is needed to reveal the potential relationship of our candidate gene list and biological mechanisms of BPD.

In summary, our study identified a list of 5 potential candidate genes for BPD based on exome sequencing in a large bipolar disorder pedigree. Among these 5 genes, *GRID1* has been reported to be associated with several psychiatric and brain related traits in common SNP-based studies. Our results provide some evidence linking rare variation in *GRID1* with BPD. These findings suggest a potential role for these genes in the risk for BPD, but require replication in large, independent studies.

3.5 Figures

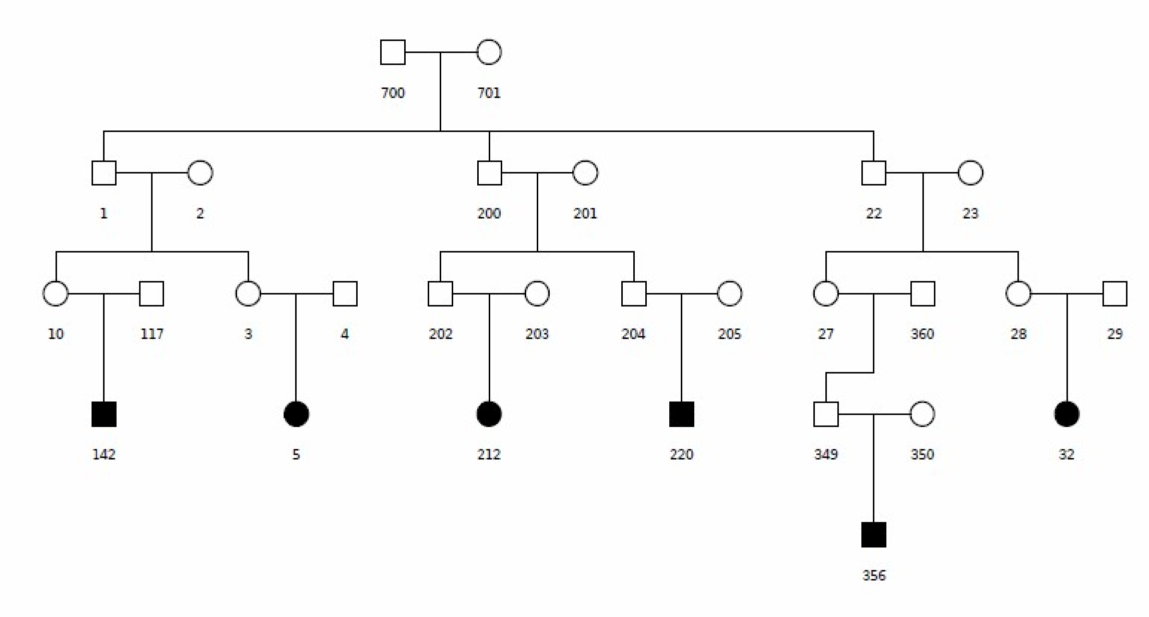

Figure 3.1 Pedigree structure of the BPD family sequenced in this study. Blood samples of the 6 BPD cases at the bottom of the pedigree were collected.

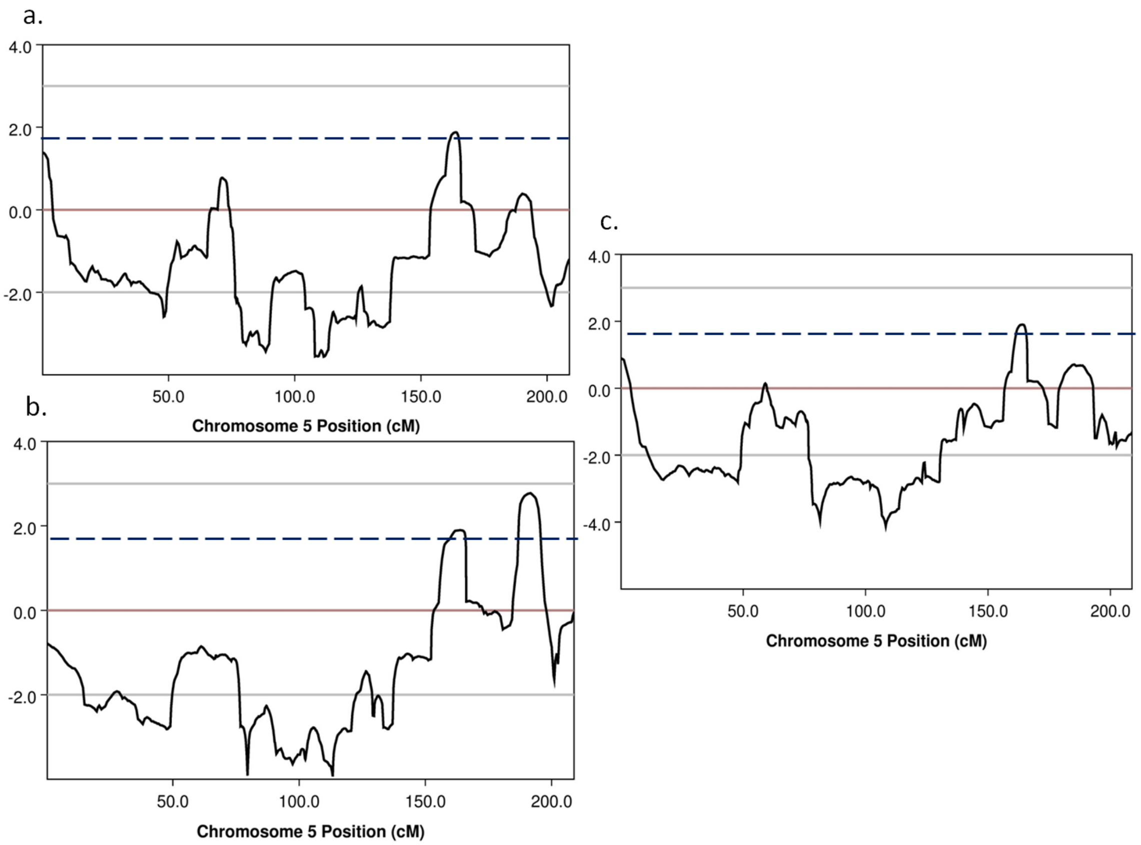

Figure 3.2 Significant peak regions identified on chromosome 5 in three sets of linkage analysis. a. Linkage results for chromosome 5 using SNP set #4 with 10,058 SNPs; b. Linkage results for chromosome 5 using SNP set #6 with 9,715 SNPs; c. Linkage results for chromosome 5 using SNP set #8 with 10,402 SNPs. Two peak regions 5q33.1- 5q 33.3 and 5q35.1-5q35.2 were identified 3 and 1 times respectively. The LOD threshold was indicated in dotted line.

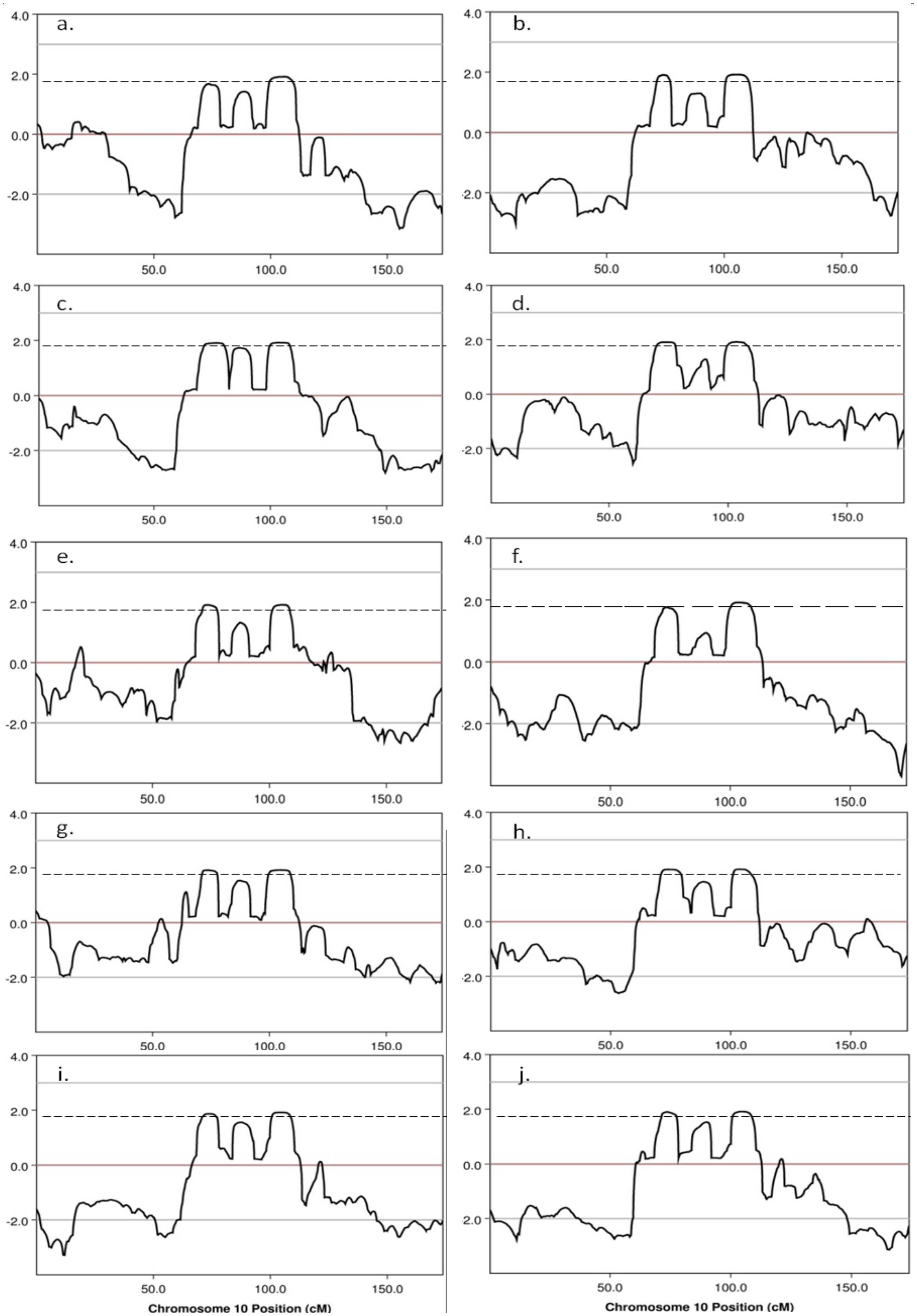
a.
b.
c.
d.
e.
f.
g.
h.
i.
j.
4.0
2.0
0.0
-2.0
50.0
100.0
150.0
Chromosome 10 Position (cM)
Chromosome 10 Position (cM)

Figure 3.3 Significant peak regions identified on chromosome 10 in all the ten sets of linkage analysis. a. Linkage results for chromosome 10 using SNP set #1 with 10,396 SNPs; b. Linkage results for chromosome 10 using SNP set #2 with 9,648 SNPs; c. Linkage results for chromosome 10 using SNP set #3 with 9,963 SNPs; d. Linkage results for chromosome 10 using SNP set #4 with 10,058 SNPs; e. Linkage results for chromosome 10 using SNP set #5 with 10,407 SNPs; f. Linkage results for chromosome 10 using SNP set #6 with 9,715 SNPs; g. Linkage results for chromosome 10 using SNP set #7 with 10,418 SNPs; h. Linkage results for chromosome 10 using SNP set #8 with 10,402 SNPs; i. Linkage results for chromosome 10 using SNP set #9 with 10,236 SNPs; j. Linkage results for chromosome 10 using SNP set #10 with 10,402 SNPs. Chromosome region 10q21.1-10q21.2 and 10q23.1-10q23.33 (10q22.3-10q23.33) were identified multiple times in the 10 replicate sets and 10q23.1-10q23.33 (10q22.3-10q23.33) was identified repetitively in all of the 10 replicate sets. The LOD threshold was indicated in dotted line.

3.6 Tables

Table 3.1 Linkage analysis results of the ten replicate sets.

Set	# of markers	Chromosome Regions	LOD
1	10,396	10q22.3-10q23.33	1.92
2	9,648	10q21.1-10q21.2	1.91
		10q22.3-10q23.33	1.92
3	9,963	10q21.1-10q21.3	1.91
		10q22.3-10q23.33	1.93
4	10,058	5q33.1-5q33.2	1.88
		10q21.1-q22.3	1.91
		10q23.1-10q23.33	1.92
5	10,407	10q21.1-10q21.2	1.91
		10q23.1-10q23.33	1.92
6	9,715	5q33.1- 5q 33.3	1.9
		5q35.1-5q35.2	2.77
		10q23.1-10q23.33	1.92
7	10,418	10q21.1-10q21.2	1.91
		10q23.1-10q23.33	1.92
8	10,402	5q33.1- 5q 33.2	1.9
		10q21.1-10q21.2	1.91
		10q23.1-10q23.33	1.92
9	10,236	10q21.1-10q21.2	1.87
		10q23.1-10q23.33	1.92
10	10,402	10q21.1-10q21.2	1.9
		10q23.1-10q23.33	1.92

The most significant linkage peak region was found on chromosome region 10q22.3-10q23.33 with LOD score of 1.926. The LOD threshold is 1.8.We have highlighted the 10q23.1-10q23.33 which were identified to be significant in all ten SNP replicates.

Table 3.2 Filtering procedure applied to the exome sequencing data from 6 BPD relative cases.

Filtering Procedure	Number of Variants pass QC (%)
Genotype calls released	140,814 (100)
Quality Control	114,432 (81.26)
Linkage Peak Region	444 (0.32)
Variants with functional significance	60 (0.04)
Novel variants or variants with MAF less than 0.05	15 (0.01)

The number of variants is reduced to 15 from the 140,814 released by exome sequencing by applying various filtering strategies.

Table 3.3 Summary information for 6 genetic variants identified after applying the filtering strategy.

CHR	SNP	Position	Ref	Alt	Gene	Function	AAChange	SIFT*	Polyphen2**	Sharing	*P****
10	rs36027713	82126541	C	G	*DYDC2*	nonsynonymous SNV	p.P123R	T	B	4	0.0003
10	-	85902497	A	T	*GHITM*	nonsynonymous SNV	p.E72D	T	B	4	0.0003
10	rs45584033	85974231	C	T	*CDHR1*	nonsynonymous SNV	p.P812S	D	D	2	0.0695
10	rs2306265	87484382	C	T	*GRID1*	nonsynonymous SNV	p.V529I	D	D	2	0.0695
10	rs3812645	87489317	T	C	*GRID1*	nonsynonymous SNV	p.M430V	T	B	2	0.0695
10	-	89280872	C	T	*MINPP1*	nonsynonymous SNV	p.T137I	T	B	3	0.003

* SIFT prediction. T stands for tolerated and D stands for damaging.

** Polyphen2 prediction. B stands for benign, P stands for possibly damaging, and D stands for probably damaging.

*** *P* values here stands for the probability of observing the sharing pattern in our pedigree. Significant variants after applying Bonferroni correction were shown in bold.

Chapter 4: Targeted Sequencing Identifies Genetic Polymorphisms of Flavin-containing Monooxygenase Genes Contributing to Susceptibility of Nicotine Dependence in European and African Americans

4.1 Introduction

4.1.1 Clinical and epidemiological characteristic of nicotine dependence

Smoking is a leading cause of preventable death, causing about 5 million premature deaths worldwide each year, and current trends show that tobacco use will cause more than 8 million deaths annually by 2030 [89]. Strong evidence connects cigarette smoking and lung cancer [90][91][92], and according to the data from American cancer society, lung cancer causes the most death each year compared to other cancers [93]. In addition, cigarette smoking is also the principal environmental risk factor for developing chronic obstructive pulmonary disease (COPD), a disease characterized by chronically poor airflow [94][95][96] . Therefore, understanding the underlying biological mechanisms of nicotine dependence will still have huge public health significance in the future.

4.1.2 Gene association mapping of nicotine dependence: a brief review

Early studies based on samples of twins have linked the lifetime smoking practices to genetic predisposition [97]. A meta-analysis of the data from five studies, each involving more than 1,000 twin pairs, showed an estimated heritability of 60% for the propensity to smoke [98]. The followed linkage and gene association mapping studies have identified several susceptible loci, including genes encoding dopamine transporter/receptors [99][100][101], cholinergic receptors [102][103][104][105] , taste receptor [106] , serotonin receptor [107][108] and gamma-aminobutyric acid type B receptor [109], that are associated with nicotine dependence. The breakthrough of microarray technology at the end of 20th century enabled the "unbiased" association mapping analysis in the whole human genome. Genome-wide association study (GWAS), which scans the whole genome by capturing the information of common SNPs, has been proved informative for nicotine dependence [110][111][112][113], and greatly accelerates the progress of this gene hunting process. Nevertheless, GWAS only focuses on a set of pre-selected, generally common SNPs, and tends to omit the rare variants and structural variants such as short insertion and deletions (indels). The recent development of "next-generation" sequencing technology has enabled researchers to investigate these variants which are not covered in GWAS at a relatively lower genotyping cost [114][115][116]. A recent published study focusing on targeted sequencing data of *CHRNA5* has identified several novel rare and low frequency coding variants that contributed to nicotine dependence [117].

Three protein families are involved in nicotine pharmacokinetics: liver cytochrome P450 enzymes (CYPs), flavin-containing monooxygenases (FMOs) and uridinediphosphate glucuronosyltransferase enzymes (UGTs) [118]. The flavin-containing monooxygenase (FMO) protein family consists of a group of enzymes that metabolise drugs and xenobiotics [119]. Five forms of FMOs are found in human and have been designated *FMO1-FMO5* [119]. Among these FMO genes, part of nicotine inhaled during smoking can be broken down to *N′*-oxide by flavin-containing monooxygenase 3 (encoded by *FMO3*) [118]. Hinrichs et al. has identified significant association between SNPs of FMO1 and nicotine dependence [120]. Although a recent study has shown that common polymorphisms in *FMO3* can influence nicotine clearance [118], no study has provided direct evidence of the association between *FMO3* polymorphisms and nicotine dependence.

In this chapter, we investigated the potential of FMO genes to confer risk of nicotine dependence via deep targeted sequencing in 2,820 study subjects (1,432 European and 1,388 African Americans) comprising of 1,583nicotine dependents and 1,237controls. Specifically, we focused on the two genomic segments including *FMO1*, *FMO3* (protein coding genes for flavin-containing monooxygenase 1 and 3) and *FMO6P* (pseudo gene), and aimed to investigate the potential association between FMO genes and nicotine dependence. Via implementing targeted sequencing, we are interested to figure out that whether rare variants contribute to the association signal derived from common variants. In addition, comparisons were made between the association results based on European Americans and African Americans.

4.2 Materials and Methods

4.2.1 Study subjects

This research was reviewed and approved by the Institutional Review Board at Washington University in Saint Louis. All the study subjects provided informed consent. Study subjects were recruited from Collaborative Genetics Study of Nicotine Dependence (COGEND) and the Genetic Study of Nicotine Dependence in African Americans (AAND) [110][104]. A total of 2,820 individuals comprising of 1,432 European and 1,388 African Americans were examined in our study. We assessed the study subjects' smoking behavior using Fagerström test for nicotine dependence (FTND) [121]. The nicotine dependence patients were defined as current smokers with FTND score equal or greater than 4, and controls were defined as having FTND score of 0 or 1 and have smoked at least 100 cigarettes in their lifetime (Table 4.1).

4.2.2 Targeted sequencing of *FMO1* and *FMO3*

DNA samples were extracted from blood with Puragene. Targeted sequencings on two 100kb regions of *FMO1* and *FMO3* were performed at the Center for Inherited Disease Research (CIDR). These genomic regions also contain part of gene *FMO4* and a whole psedogene *FMO6P*. The quality control was implemented in samples and variants level respectively. The mean on-target coverage was 180x for each sequencing experiment and greater than 96% of on-target bases had a depth greater than 20x.

4.2.4 Quality control measures

Data quality was systematically evaluated using a robust alignment and variant calling workflow implemented by CIDR (http://www.cidr.jhmi.edu/index.html). Over 100 quality control metrics were evaluated in real time to quickly identify potential errors and implement fixes throughout the sequencing process. Briefly, sample quality controls were conducted based on batch effects, discordance with array data, alternate callsets, relatedness and some research specific criteria. Strategies used for variant quality control includes VQSR, duplicate sample discordance, Mendelian errors, Hardy-Weinberg equilibrium (HWE), sequence context, locus report by gene and genotype missing rate. All variants passed the Variant Quality Score Recalibration with a mean quality score of 99, mean depth of 122 with no missing calls, no Mendelian errors and zero discordances between duplicate samples. Importantly, all the rare variants were then manually evaluated by the Quality Assurance/Quality Control analysis team.

4.2.3 Statistical methods and bioinformatics analysis

A total of 1,432 European and 1,388 African Americans with targeted sequencing of *FMO1* and *FMO3* were examined. General data analyses were performed by R (R i386 3.2.1) [122]. To quantify the potential population stratification, we conducted principal component analysis (PCA) in the combined sample (115,338 markers), as well as separately in the European American sample (154,049 markers) and African American sample (218,399 markers), using a previous collected genome-wide array dataset containing 950,847 SNPs [123]. Sequencing data were annotated by sequencing data annotation software ANNOVAR [46]. After variant level quality control, 5,105, 2,600 and 3,817 variants located within the two targeted genomic regions (*FMO1*/*FMO3*) were extracted from combined, European and African American sample set, respectively.

Variants satisfying the following criteria were utilized in variant level analysis: 1) variants with MAF > 0.05 and 2) located within targeted gene regions or the linkage disequilibrium (LD) blocks that are (partly) overlapped with the targeted gene regions (detailed definition of blocks is given below). The association analysis was conducted by fitting logistic regression model. The genotypic data were coded in additive model. This analysis was performed in combined, European American, and African American individuals separately (for combined subjects, we tested a union of SNPs sets selected based on European and African American subjects). Gender and age were included as covariates in all the three analyses. The first two principal components based on the three sample sets were also utilized as covariates accounting for the potential population stratification when fitting the logistic models. To address the multiple comparison problem, we implemented Bonferroni correction. The number of tests was calculated in the following way:

$$N = n_1 + n_2 \qquad （4.1）,$$

where for each dataset, n_1 stands for the number of LD blocks generated by this dataset and n_2 is the number of variants that do not belong to any LD blocks. In addition to testing associations for single variants, we also conducted haplotype based analysis with combination of multiple variants in European and African American datasets, respectively. LD blocks were constructed using the default algorithm taken from Gabriel et al [124]. 95% confidence bounds on D' are generated and each comparison is called "strong LD" when the confidence bounds have upper bound $\geq$ 0.98 and lower bound $\geq$ 0.7, and a block is created if 95% of informative comparisons are "strong LD". Variant level association analysis and LD construction and haplotype analyses were conducted using Plink [76].

Variants then were classified into two categories for the gene level analysis (mostly rare variants). The two categories are: 1) gene-region variants set, that is the variants located within the gene region, and 2) functional-region variants (variants located within regions with significant functional significance, including exonic regions, 3'/5' UTR, smaller comparing to gene-region set). For these two variants sets, analysis was performed on variants with MAF less than 0.01, and 0.05. Both SKAT and weighted burden test [125] were utilized for the gene level analysis. Same as the variant level analysis, we also conducted this analysis in combined, European and African American individuals separately. Gender, age and first two PCs based on the three sample sets were also included as covariates. LocusZoom was utilized to make regional association plots [126].

We examined the targeted SNPs and/or genes using several bioinformatics tools and databases. We utilized the protein-protein interaction database STRING (http://string-db.org/) [127] to explore the potential interactions of our targeted genes. The Regulome DB (http://regulomedb.org/) [128] was used to predict the potential functional consequences the identified risk SNPs. This database is a web based bioinformatics tool integrated with multiple types of data (including ChIP-seq, DNase-seq, and eQTLs etc.) from the Encyclopedia of DNA Elements (ENCODE) project[129].

4.3 Results

4.3.1 Variant-wise association of FMO genes and nicotine dependence

270, 326 and 368 variants were selected for variant-wise association analysis in European, African American and combined sample set, respectively. 6 and 18 (covered 262 SNPs) LD blocks were constructed in European and African American sample sets, respectively. Based on these LD blocks patterns, we obtained the significant thresholds for variant-wise analysis were 3.6×10^{-3}, 1.25×10^{-3} in European American and African American sample set, respectively. We chose the most conservative one as our *P* value threshold in analysis (1.25×10^{-3}). Multiple different significant variants were identified in European and African American datasets (Table

4.2, Figure 4.1). A cluster of significant variants were identified in European American individuals (with most significant SNP rs6674596, *P*=0.0004, OR=0.67, *FMO1*). In African American individuals, we identified several clustered significant variants (with the most significant SNP rs6608453, *P*=0.001) in pseudo gene *FMO6P*.

4.3.2 Haplotype based and gene-wise association of FMO genes and nicotine dependence

We performed haplotype based analyses in European and African American dataset separately. The *P* value thresholds were decided by Bonferroni correction and thus were different for each dataset. We utilized 0.008(0.05/6) and 0.0025 (0.05/20) as *P* value threshold for European and African American dataset, respectively. No significant signals were identified through haplotype based analyses. Gene-wise association analyses mainly focused on rare and/or low frequency variants in our dataset. Although we have tried multiple analytical schema (combination of different MAF threshold and region definitions), no significant association signals were found in this analysis (Table 4.3).

4.3.3 Bioinformatics analysis

Proteins that show evidence for interaction with proteins encoded by *FMO1, FMO3* and *FMO6P* were extracted from STRING (Figure 4.2-4.4). Both *FMO1* and *FMO3* have a strong relation with a variety of genes belong to CYP gene families. FMO6P, however, as a pseudo gene, only showed limited evidence related with PRSS16. We explored the top significant variants (rs6674596 and rs608453) in

Regulome DB to investigate their potential biological significance. Regulome DB has its own scoring system to measure the biological significance of a variant. The range of the scores is from 1-6, and the smaller the score is, the more evidence that indicate this variant has biological significance. Rs6674596 has a Regulome DB score of 5, and it located within a DNase hypersensitive area of assayed in multiple cell types. In addition, this variant also located at a sequence motif region (HNF1). Rs608453 has a Regulome DB score of 6. It also located in the a sequence motif (Cdx). No expression quantitative trait loci (eQTL) or transcription factor (TF) binding related evidence were shown for neither variants.

4.4 Discussion

As part of a large scale targeted sequencing study focusing on nicotine-dependent/nondependent smokers, our aim was to test the hypothesis that genetic polymorphisms of flavin-containing monooxygenase genes contribute to the risk of nicotine dependence. The underlying rationale of this study is based on the fact that flavin-containing monooxygenase genes are key genes of the nicotine metabolism pathway [118]. *FMO1* may play a role in nicotine metabolism and contributed to the nicotine level in brain organ [130], and *FMO3* encodes flavin-containing monooxygenase 3 (encoded by *FMO3*), which can metabolize a small percentage of nicotine into nicotine N’-oxide [118]. We studied both rare and common variants in *FMO1*, *FMO3* and *FMO6P* through large scale targeted sequencing.

A number of common variants in *FMO1* were identified to be significantly associated with nicotine dependence, and we noted that there was an ethnic-specific pattern. We identified a cluster of significant variants in *FMO1* in the European Americans. The most significant variant was rs6674596 (*P*=0.0004, OR=0.67, MAF_EA=0.135, MAF_AA=0.463). However, this significant result was not replicated in our African American dataset (*P*=0.9325, OR=1.01). The association signals for *FMO1* have been reported by Hinrichs et al [120]. Several significant SNPs reported in that paper, including rs742350 and rs1126692, were also identified to be significant in our study. Considering both studies utilized COGEN samples, our results on the common SNPs basically replicated Hinrichs' results. In addition to the significant findings in European American sample set, we also identified a set of significant variants located on gene *FMO6P* from the African American dataset (with the most significant SNP rs6608453, *P*=0.001, MAF_AA=0.097, MAF_EA=0.192). Just like significant variants were only identified in European Americans, this significant signal of *FMO6P* was only identified in African Americans but failed to be replicated in European Americans (*P*= 0.1109, OR=1.17). No significant SNPs were identified from the combined sample set, although in the combined samples set the sample size almost doubled with a correspondent increment in statistical power. The reasons behind this ethnic-specific pattern might be complex, and the most plausible one is differences in the regional LD structure between the two racial/ethnic groups. This difference might mean the surrogate SNPs miss the signal created by the real underlying susceptible variants in a specific set of samples.

One major advantage of our targeted sequencing study is that we can examine every possible DNA variations in our targeted regions and conducted association analysis thoroughly. However, in this study, we did not detect significant association between the rare variants and nicotine dependence, although we systematically tried many combinations of statistical methods, MAFs, region definitions and sample sets. The most significant rare variant set was identified for gene *FMO1* with region definition of "gene region" and MAF<0.01 in African Americans (*P*=0.0636). The lack of significant findings for rare variants suggests that the significant associations for common SNPs are not simply surrogates for rare variant associations (synthetic associations) [131].

We found it interesting to examine the functional significance of the significant common SNPs we identified. All of the significant common SNPs are located either in introns or outside the gene. Therefore, if these significant common SNPs alter function, it is not by changing protein structure. The most significant SNP in *FMO1,* rs6674596, is located within a DNase hypersensitive area of assayed in multiple cell types, and most of the regulatory regions and some promoter regions tend to be DNase sensitive. This suggests that this SNP might have an effect on the expression of gene *FMO1*. Nevertheless, without further evidence from biological experiments, it is still too early to explain this association signal.

FMO6P is a pseudo gene which means that this gene cannot be properly expressed as a protein, and it is probably because it is unable to produce a full length transcript [132]. *FMO6P* is reported to have significant sequence homology with *FMO3* [132]. Previous studies have set up direct links of SNPs in *FMO6P* with chronic allograft dysfunction [133] and pharmacokinetic characteristics of sulindac sulfide in premature labor [134]. One interesting note for these previous studies is that the significant findings of *FMO6P* are always accompanied with significant findings from *FMO3*, and at least in one study [135] , the significant SNP of *FMO6P* is in complete LD with significant SNP of *FMO3*. This suggests that the significant hit in *FMO6P* might be a surrogate for some true underlying signal in *FMO3*. However, in our study, although some SNPs of *FMO3* are indeed in complete LD with the significant SNPs of *FMO6P*, the whole significant SNP cluster is located in the *FMO6P* region (Figure 4.1). On the other hand, if the signal we identified in *FMO6P* is not the surrogate for effects of SNPs in *FMO3*, but has an independent effect on nicotine dependence, then further research will be needed to clarify the underlying function of *FMO6P*.

A major strength of our study is that, unlike most of the common SNP based association studies, is that we implemented a targeted sequencing technology for genotyping of our study subjects. This enables us to consider both common and rare variants within the three gene regions. Additionally, this study design enabled us to analyze both SNVs and indels which are often omitted in SNP based study designs. However, there are also some limitations to this study that need to be noted. Firstly, we lack replication for our significant findings. The design using two racial/ethnic groups in our study enabled us to use as the two datasets as replication set for each other. However, significant findings in the European American dataset were not confirmed in the African American dataset. In addition, our sample size limited the statistical power to detect potential modest effects of SNPs. This is a common challenge, especially when using study designs, such as targeted sequencing, which generate genotype data at many variants, leading to multiple comparisons and corresponding stringent significance requirements. Future work to address this challenge would be to combine multiple sequenced datasets using meta-analysis; such approaches have been productive for GWAS of complex traits and have yet to be fully leveraged for sequencing studies and rare variant analyses.

In summary, we tested the genetic effects of three flavin-containing monooxygenases genes (*FMO1, FMO3* and *FMO6P*) on nicotine dependence by performing targeted sequencing on 2,852 nicotine-dependent and non-dependent smokers. We performed both variant-level and gene/region-level analyses to examine the genetic association of rare, low frequency and common variants within these region and nicotine dependence, and both SNVs and indels. We identified significant association signals for gene *FMO1* and *FMO6P*. Replications of our finds in other ethnic groups were needed in the future. Most of the significant variants identified were SNPs located within intron regions or with unknown functional significance, indicating a need for future work to understand the underlying functional significance of these signals.

4.5 Figures

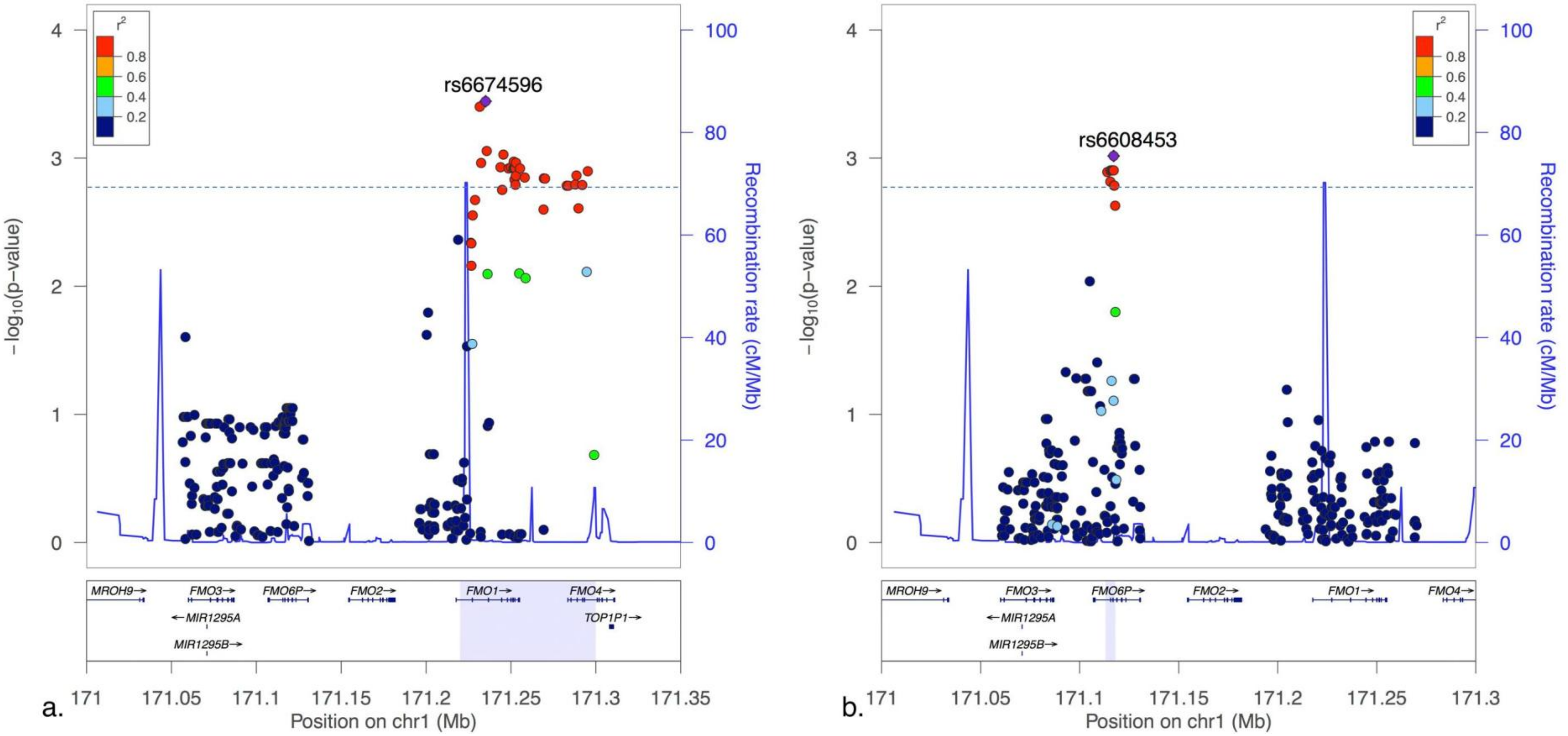

Figure 4.1 Regional association plots of *FMO1-FMO3-FMO6P* genomic region based on European Americans and African Americans. a) European Americans and b) African Americans. The blue dash lines are the –log$_{10}$(*P*-value) threshold used in our study(1.25×10^{-3}).

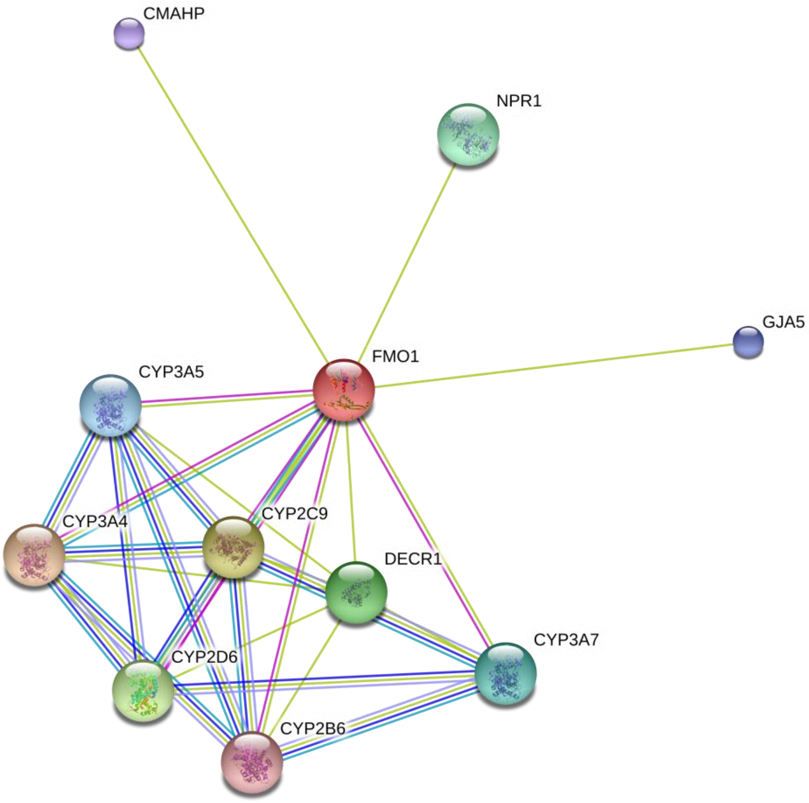

Figure 4.2 Protein-protein interaction network of *FMO1*.

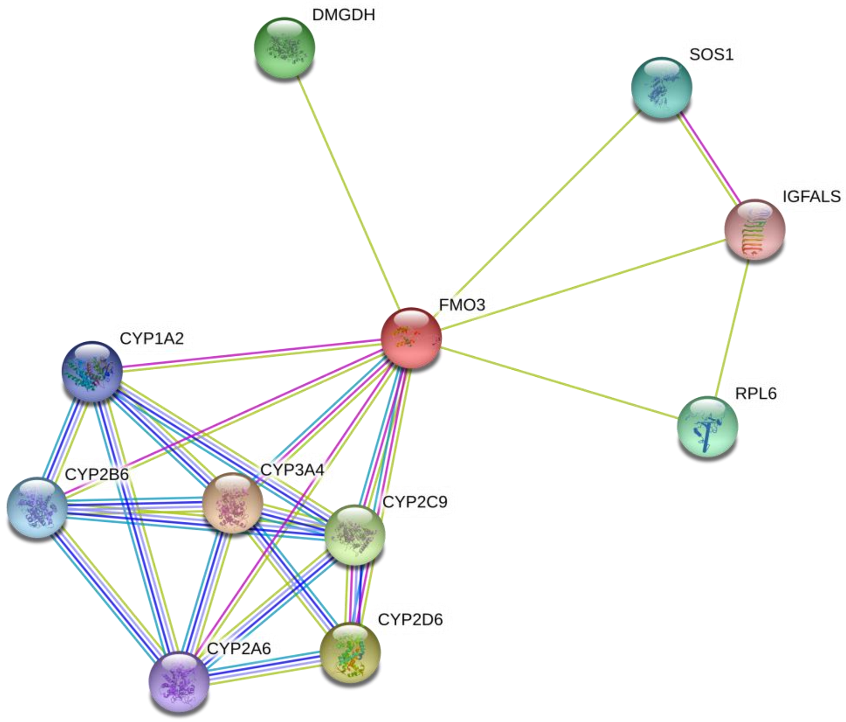

Figure 4.3 Protein-protein interaction network of *FMO3*.

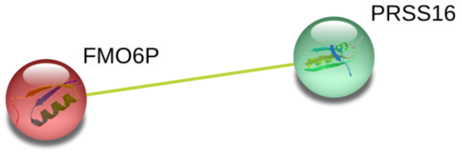

Figure 4.4 Protein-protein interaction network of *FMO6P*.

4.6 Tables

Table 4.1 Characteristics of study subjects

		Nicotine dependent	Non-dependent
Sample, *n*		1,583	1,237
Gender			
	Female	901(59%)	805(65%)
	Male	682(41%)	432(35%)
Ethnicity			
	European American	730(46%)	702(57%)
	African American	853(54%)	535(43%)
Age in year, mean (range)		37(25-45)	36(25-45)
FTND score*, mean (range)		6.34 (4-10)	0.16(0-1)

*FTND is the Fagerström Test for Nicotine Dependence.

Table 4.2 Significant signals in variant-wise association analysis.

CHR	VAR	GENE	POS	A1	OR_EA	*P*_EA	MAF_EA	OR_AA	*P*_AA	MAF_AA
1	rs11812044	*FMO6P*	171115567	A	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs17565793	*FMO6P*	171116267	C	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs17623477	*FMO6P*	171116304	C	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs7051747	*FMO6P*	171116550	G	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs7066454	*FMO6P*	171116603	T	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs7063044	*FMO6P*	171116760	T	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs6608453	*FMO6P*	171117140	T	1.17	0.1109	0.19	0.64	0.0010	0.10
1	rs6608454	*FMO6P*	171117170	C	1.16	0.1232	0.19	0.65	0.0012	0.10
1	rs12726624	*FMO1*	171231630	G	0.67	0.0004	0.13	1.03	0.7385	0.47
1	rs17581251	*FMO1*	171232446	T	0.67	0.0011	0.12	0.98	0.8685	0.08
1	rs28360379_indel	*FMO1*	171234851	A	0.68	0.0006	0.13	0.99	0.8698	0.43
1	rs6674596	*FMO1*	171235088	T	0.67	0.0004	0.14	1.01	0.9325	0.46
1	rs13376631	*FMO1*	171235742	G	0.69	0.0009	0.13	1.00	0.9831	0.43
1	rs12094878	*FMO1*	171243863	C	0.69	0.0012	0.14	1.04	0.6315	0.47
1	rs12062692	*FMO1*	171245579	G	0.69	0.0009	0.14	0.97	0.7002	0.48
1	rs7539057	*FMO1*	171248614	A	0.70	0.0012	0.14	0.95	0.4840	0.42
1	rs742350	*FMO1*	171250044	T	0.69	0.0012	0.14	1.03	0.6911	0.46
1	rs12091482	*FMO1*	171251509	T	0.70	0.0012	0.14	0.94	0.4723	0.42
1	rs10399952	*FMO1*	171251663	G	0.69	0.0011	0.14	0.95	0.5037	0.42
1	rs10399602	*FMO1*	171251876	C	0.70	0.0012	0.14	0.95	0.4840	0.42
1	rs7519999	*FMO1*	171251958	G	0.70	0.0012	0.14	0.95	0.4840	0.42
1	rs1126692	*FMO1*	171252287	G	0.70	0.0012	0.14	0.93	0.3914	0.46

1	rs12092985	*FMO1*	171252537	A	0.70	0.0012	0.14	0.95	0.4840	0.42
1	rs10912714	*FMO1*	171253037	G	0.69	0.0011	0.14	0.95	0.4840	0.42
1	rs12059179	*FMO1*	171255346	T	0.70	0.0012	0.14	0.91	0.2620	0.39

Significant findings were highlighted in bold.

Table 4.3 Full results of gene-wise association analysis.

Gene	Population	MAF	Region	SKAT	Burden
FMO1	European American	<0.05	gene region	0.8072	0.2071
			functional region	0.8003	0.2857
		<0.01	gene region	0.5491	0.1463
			functional region	0.8020	0.2657
	African American	<0.05	gene region	0.2905	0.2146
			functional region	0.8675	0.2844
		<0.01	gene region	0.0636	0.2481
			functional region	0.9902	0.8136
	Combined	<0.05	gene region	0.2030	0.4330
			functional region	0.4013	0.4399
		<0.01	gene region	0.5486	0.7550
			functional region	0.7578	0.6936
FMO3	European American	<0.05	gene region	0.4270	0.2791
			functional region	0.9838	0.3735
		<0.01	gene region	0.8539	0.6863
			functional region	0.8246	0.1389
	African American	<0.05	gene region	0.7460	0.5758
			functional region	0.3600	0.6912
		<0.01	gene region	0.9340	1.0000
			functional region	0.5992	0.3285
	Combined	<0.05	gene region	0.1351	0.9251
			functional region	0.8212	0.5602

		<0.01	gene region	0.1902	0.0691
			functional region	0.6783	0.4064
FMO6P	European American	<0.05	gene region	0.3707	0.3239
			functional region	0.2326	0.1856
		<0.01	gene region	0.2917	1.0000
			functional region	0.0948	0.3778
	African American	<0.05	gene region	0.7999	0.3774
			functional region	0.2711	0.6926
		<0.01	gene region	0.8139	0.3701
			functional region	0.1175	0.6725
	Combined	<0.05	gene region	0.6998	0.7264
			functional region	0.3588	0.5644
		<0.01	gene region	0.5938	0.1462
			functional region	0.7372	0.5602

Chapter 5: Evaluation and optimization of multiple centrality measures in human protein-protein interaction network

5.1 Introduction

Human gene networks are graphical representations of the interactions between the genes. In a human gene network, genes are represented as nodes and the relationships among them as edges. Human gene networks can be divided into three categories: 1) Human gene networks derived from the curated knowledge; 2) Human gene networks based on the experimental data of physical interactions, and 3) those that are inferred from high-throughput data [136]. Centrality is a key indicator that identifies the most important vertices within a graph. To quantify centrality, four major measurements have been proposed: 1) Degree centrality, which simply counts the number of interactions to a node; 2) Betweenness centrality, where nodes which fall in the shortest path of other nodes have high betweenness; 3) Closeness centrality, which is related to the topology of the nodes in a network; and 4) Eigenvector centrality which ranks the nodes in a network based on its integrating neighbors [137]. Despite some early studies [138][139] that utilize the simplest degree centrality measure, most of the recent research projects have implemented the eigenvector centrality measure, such as algorithms modified from the Google PageRank algorithm, to evaluate the importance of a gene in a gene network [35][140]. However, research questions such as to what extent these centrality measures can represent the functional significance of genes or whether any of these centrality measures outperformed others have never been seriously investigated.

In this chapter, we aim to investigate the followed three scientific questions: 1) Can centrality reflect the biological significance of genes in a general human gene network? 2) Among these four commonly used centrality measures, does any of them outperform others? 3) Will they do better if we combine several centrality measures together using machine learning algorithms? To answer these scientific questions, we construct a comprehensive human gene-gene network using protein-protein interaction (PPI) data.

5.2 Materials and Methods

5.2.1 Construction of human protein-protein interaction network

To evaluation the efficacy of multiple centrality measures as indicators of the biological significance of genes, we first constructed a genome-scale human gene network based on human protein-protein interaction (PPI) data. We extracted our PPI data from the STRING database [127]. As a database of known and predicted protein-protein interactions, STRING provides data users each protein-protein interaction with a confidence score. To evaluate the potential effects of the quality of PPI data to our study results, we constructed 4 human general dataset based on the full PPI dataset and PPI dataset with top 75%, 50% and 25% high confidence score, respectively.

5.2.2 Calculation of multiple centrality measures

Four measurements of centrality were calculated based on our general human gene network, including degree centrality, betweenness centrality, closeness centrality and pagerank centrality. Degree centrality is simply the number of adjacent edges to each node. Betweenness centrality equals to the number of shortest paths from all vertices to all others that pass through that node. Closeness centrality is defined as the reciprocal of the farness which is the sum of distances of a node to all other nodes in a network. Pagerank centrality ranks the nodes in a network based on its integrating neighbors. The R package igraph was utilized to perform the construction of human network and the calculations of these centrality measures [141].

5.2.3 Extraction of human essential gene sets

To evaluate the centrality measures as indicators of the functional significance of human genes, besides the general human gene network we constructed, we also need a set of genes to serve as “true answers”, which means that we are sure that these genes should be biologically essential to human beings. We prepared our human essential gene sets in the following three ways (resulting in four gene sets):

1) Online Gene Essentiality database (OGEE). OGEE is an online database that records both experimentally tested essential and non-essential genes [142].Two categories of essential genes, large-scale experiments based and text-mining based, were recorded in the OGEE database. We extracted these two sets of essential human genes from this database, and they were marked as OGEE.experiment and OGEE.textmining in the following, respectively.

2) Mendelian disorder related gene sets (marked as OMIM in the following). We extracted a set of human genes that were reported to cause mendelian disorders from Online Mendelian Inheritance in Man (OMIM) [143].Two files, the GeneMap file and MorbidMap file, were provided by OMIM. GeneMap contains the information that is centered by genes while MorbidMap file classified genes by their related phenotypes. The Mendelian disorder related genes were not specifically indicated for each term. To extract a potential gene list, we conducted the following filtering strategy:

   i. We removed those terms in both files that were tagged by “[]”, “{}” or a question mark “(?)”.Brackets, “[]”, indicate "nondiseases," mainly genetic variations that lead to apparently abnormal laboratory test values but not disorders. Braces, “{}”, indicate mutations that contribute to susceptibility to multifactorial disorders or to susceptibility to infection. A question mark, “?”, before the disease name indicates an unconfirmed or possibly spurious mapping.

   ii. For the GeneMap file, we only selected those terms with an indicator “C” (confirmed), which means this association was observed in at least two laboratories or in several families.

   iii. For MorbidMap, we only selected those terms with an indicator “(3)”, which means the the molecular basis of the disorder is known.

   iv. We then merged the two gene lists obtained from the GeneMap file and MorbidMap file, and obtained the OMIM gene set.

3) Empirical gene sets (marked as ExAC in the following). This is an essential gene set that is extracted based on the Exome Aggregation Consortium (ExAC) data.ExAC

and includes exome sequencing data from a wide variety of large-scale sequencing projects spanning 60,706 unrelated individuals[144]. We extracted a set of essential genes by selecting genes that do not contain any destructive variants. The basic rationale of this gene set is that if a gene is very important and essential for human beings, disruptive mutations of this gene cannot be found in people with no severe pediatric disease. We utilized ExAC data set to obtain this gene set in the following filtering strategy:

i. A standard quality control (QC) process was applied to the ExAC dataset, only variant calls with quality indicating as "PASS" were included in the following study.

ii. We removed those genes that contained frameshift indels in any of the 60,706 individuals and this step reduced the gene number to 4,465. Frameshift indels are very disruptive mutations that may completely disable the gene function.

iii. We removed genes in major histocompatibility complex (MHC) regions. The rationale of this step is that there are a lot of repetitive sequences in this region and this severely affected the quality of the sequencing experiment. Therefore, the genes we identified in MHC regions that contain no frameshift indels may not be because that these genes are important, it may be just because of the low quality of sequencing. We eliminated those variants with low quality during the QC process. After applying this strategy, there are 2,760 genes left.

iv. We assumed the length of coding sequence (CDS) of a gene might affect the chance that a frameshift mutation occurs within the gene. Therefore, we ranked the 2,760 genes by their length of CDS and only included the 1,200 genes with shorter CDS.

The basic rationale behind the selection of essential gene sets is that we want to collect a set of genes that are “functionally significant” for human beings (so they can serve as the “true answer” in ROC analyses). It is art more than science to select these gene sets because there is no universal standard to determine what kind of genes can be considered “functionally significant”. Lethality is the first potential criterion. However, it is difficult to apply this standard directly due to ethical issues. To overcome this difficulty, we utilized two methods. The first one is the experimental method. We can obtain a set of genes that were experimentally proven to be lethal when knocked out in model organisms, and then map those genes back to the human genome and extract the homologous genes. This defines to the two gene sets we obtained from OGEE. Another one is the observational method. We simply examine the genomes data of human populations to examine whether there are any genes without disruptive mutations. This defines the empirical gene set (ExAC). In addition to lethality, a list of causal genes of mendelian disorder might be another choice for a set of functionally significant genes.

5.2.4 Comparisons and evaluations of different centrality measures

To compare the effecacy of these four measurements of centrality as predictors of functional significance of genes in human gene network, we made receiver operating characteristic (ROC) curves in the following ways:

1) We checked the four essential gene sets by each definition of centrality measure.
2) We checked the four different centrality measure for each essential gene set.

5.2.5 Logistic regression model by combination of multiple centrality measures

To evaluate whether there is an improvement using a combined score of different combinations of these centrality measures, we constructed a series of logistic regression models. Due of the severe multicollinearity among these four degree centrality measures (Figure 5.1), we fitted our model using a penalized regression technique. The 10-fold cross-validation approach was used to assess the performance of these prediction models. We calculated the area under curve (AUC) as a major indicator for comparisons of the models. Regression model fitting and model comparisons were conducted by R package glmnet [145].

5.3 Results

5.3.1 Human protein-protein interaction network and essential gene sets

We constructed a general human gene network that covered 18,199 human protein coding genes using the full PPI data from STRING. This gene network covered 82.6% of human genes. We obtained four essential gene sets, including OGEE.experiment (1,511), OGEE.textmining (1,502), OMIM (1,244) and ExAC (1,200) gene set by the methods described above. The number of genes for each gene set and the overlaps are shown in Table 5.1. From this table, we can see that the four essential gene sets we extracted were basically independent from each other with limited overlaps. The following main results were conducted using human gene network constructed from the full PPI data downloaded from STRING. Most of these results were validated in the other three networks that were constructed with less but higher quality data (using 75%, 50% and 25%

of the PPI data). We have summarized the information of these four networks in Table 5.2.

5.3.2 Comparisons of the four centrality measures as indicators of biological significance

The centrality measures of the four essential gene sets were calculated based on network constructed from full PPI data and shown in Figure 5.2. ANOVA analyses indicated the differences in centrality of the five gene sets (four essential gene sets and one random selected gene set) were significant in all of the four centrality measures (for degree, closeness and pagerank centrality $P<2\times10^{-16}$; for betweenness centrality, $P=7.91\times10^{-16}$). In general, all of these four essential gene sets had higher centrality compared to the randomly selected gene set for all four centrality measures. Two essential gene sets that were extracted from OGEE (experiment and textmining) had higher centralities compared to ExAC and OMIM gene sets. ROC curves that compared the average centralities of these four essential gene sets were shown in Figure 5.3. The centralities of the four essential gene sets were calculated based on network constructed from 75%, 50% and 25% PPI data were shown in Figure 5.4- 5.6.

5.3.3 Comparison and optimization of the four centrality measures using logistic model

The four centrality measures were compared using ROC curve for each essential gene sets (Figure 5.7). As we can see from this figure that, for all these four essential gene sets, there is no significant difference among the centrality measures. None of these centrality measures outperformed the others as a predictor of the essentiality of genes in our general human gene network. This result was validated in human gene network constructed using 25%, 50% and 75% PPI data, and these results were shown in Figure 5.8-5.10. The logistic models were evaluated using AUC measurements. The AUC plot based on four essential gene sets were summarized in Figure 5.11. As we can see from figure 3, in general, models including more centrality measures performed better (except for OMIM gene set). However, this improvement was limited. The largest improvement in AUC was around 0.05. Detailed information of AUC measures for different combinations of centrality measures were summarized in Table 5.3.

5.4 Discussion

Previous genome-wide studies have shown that disrupted hub protein, protein that locates at center position in a PPI network, is more likely to have lethal effect than a non-hub protein, and this phenomenon is sometimes described as “centrality-lethality rule” [146]. Our findings substantiated these previous observations. All of these four essential gene sets had significantly higher centralities compared to a gene set randomly selected. Another interesting observation is that the average centrality measures of these four essential gene sets showed a gradient pattern. In general, essential gene sets extracted from OGEE had highest average centrality, and OMIM gene set was slightly lower, while ExAC gene set had the lowest average centrality. This pattern can be explained by the differences of the potential functional significance among these essential gene sets. OGEE gene sets were based on the experimental evidence of lethality in model organisms, and the disruption of these genes might have lethal effects in human beings. On the other hand, genes in OMIM gene set were causal genes of human mendelian disorders, and in most situations, disrupting these genes will cause mendelian disorders but not to be lethal. In this sense, the OGEE gene sets should be more biologically significant compared to genes in OMIM gene set. These evidence indicated that greater biological significance implies average higher centrality in a general PPI network. Another feature of our study that is different from previous research is that we provided a tool to quantify and evaluate the efficacy of the centralities as predictors of functional significance of genes in human gene network.

One thing interesting to note is the performance of ExAC gene set in our analyses. The ExAC gene set was expected to have similar properties to the OGEE gene set because both of them were based on the lethality standard as described in the method part. However, the average centralities of ExAC gene set were the lowest among all of these four gene sets. This might be due to two reasons. The first one is the ExAC gene set was extracted based on around 60,706 individuals, and this sample size might not be large enough to rule out some non-significant genes. In addition, several arbitrary criteria used during the filtering process might increase the chance of the exclusion of some important genes. For example, we utilized the length of coding sequence (CDS) as a filtering criterion to reduce its potential effects. This criterion can rule out many small trivia genes, however, as an arbitrary criterion, it might also exclude some potential functionally significant genes.

The four centrality measures commonly used in network analysis constitute different mathematical computations on the same underlying data. Degree centrality, for example, is very easy to calculate, but betweenness and closeness centralities are computational intensive especially when the adjacency matrix involved are big. Despite these differences in computational level, previous researchers have noticed the statistical correlations among multiple centrality measures in a social network [147]. In our study, we also observed a pattern of high correlation among the four centrality measures using our general human gene network. In addition, this correlation property might also contribute to the similarity of the performance of these centralities serving as predictors as functional significance of genes. This high correlation property might also be the reason that combinations of these centrality measures could only make a very limited improvement for the performance of the logistic models. This indicates that the development of multiple measures may be somewhat redundant, and they might perform similarly in statistical analyses.

In this study, we have showed that genes with high centralities were enriched with essential genes. A potential limitation of this study is that we only explored the property of gene centrality in a general human gene network which covered most of the human protein coding genes. The biological mechanisms are very complex and a general human network cannot provide enough resolution to scrutinize detailed aspects of the enrichment pattern of functional genes. Furthermore, in this study, we only investigated the gene sets that are essential to human beings, and it might be more interesting to examine the centrality property of genes that are susceptible to complex disorders/traits. Therefore, for future study, it might be more fruitful for researchers to construct some functional specific sub-networks and focused on susceptible genes of complex disorders/traits.

To conclude, in this study we have showed that there is a connection between the essentiality and centrality of human genes. A pattern of strong correlations was identified among the four commonly used centrality measures for a general human PPI network and the performance of each centrality measure was similar to others serving as predictors of the essentiality of genes. The improvement of the prediction models was limited when combined several different centrality measures.

5.5 Figures

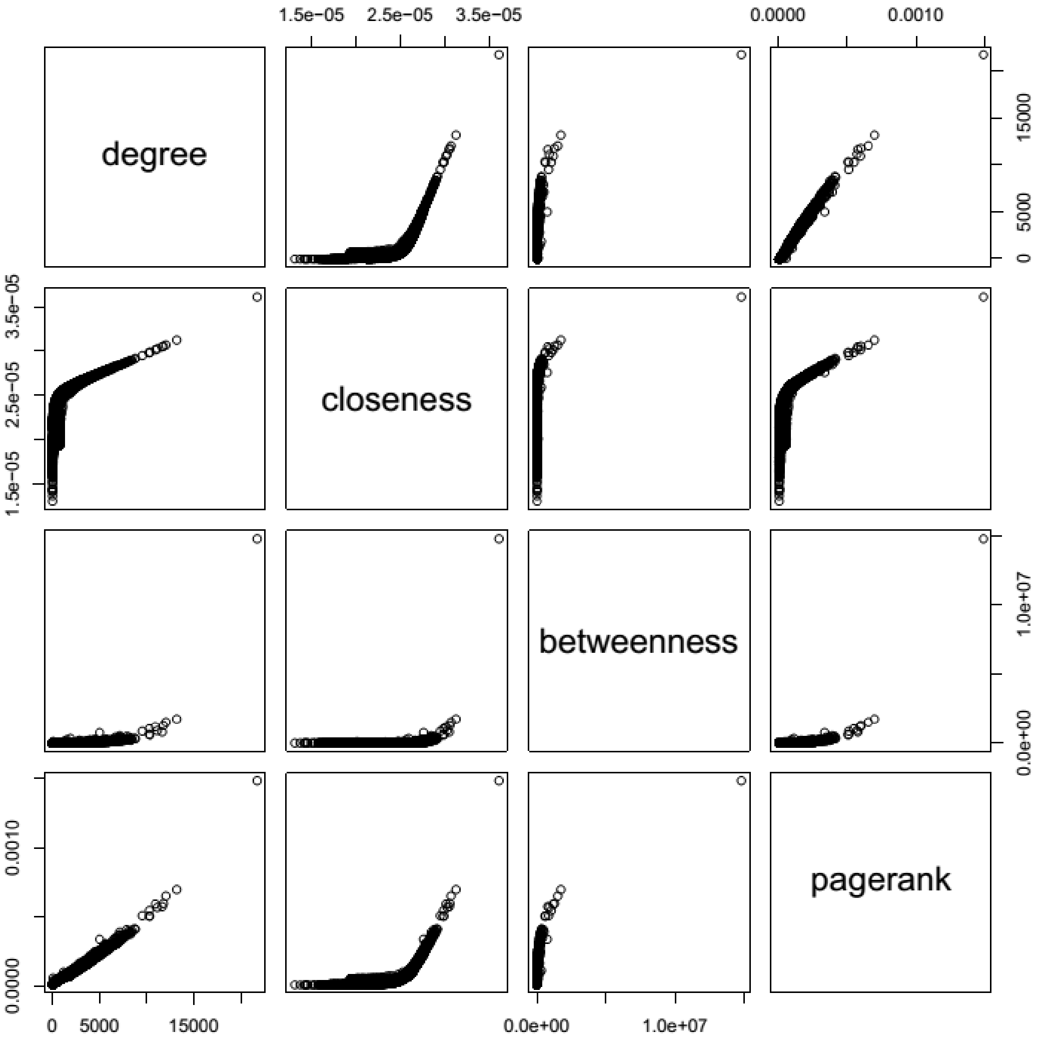

Figure 5.1 Scatter matrix for the four measurements of centrality. The outlier in this plot is gene *ubiquitin C*(*UBC*). It was removed in the following model fitting.

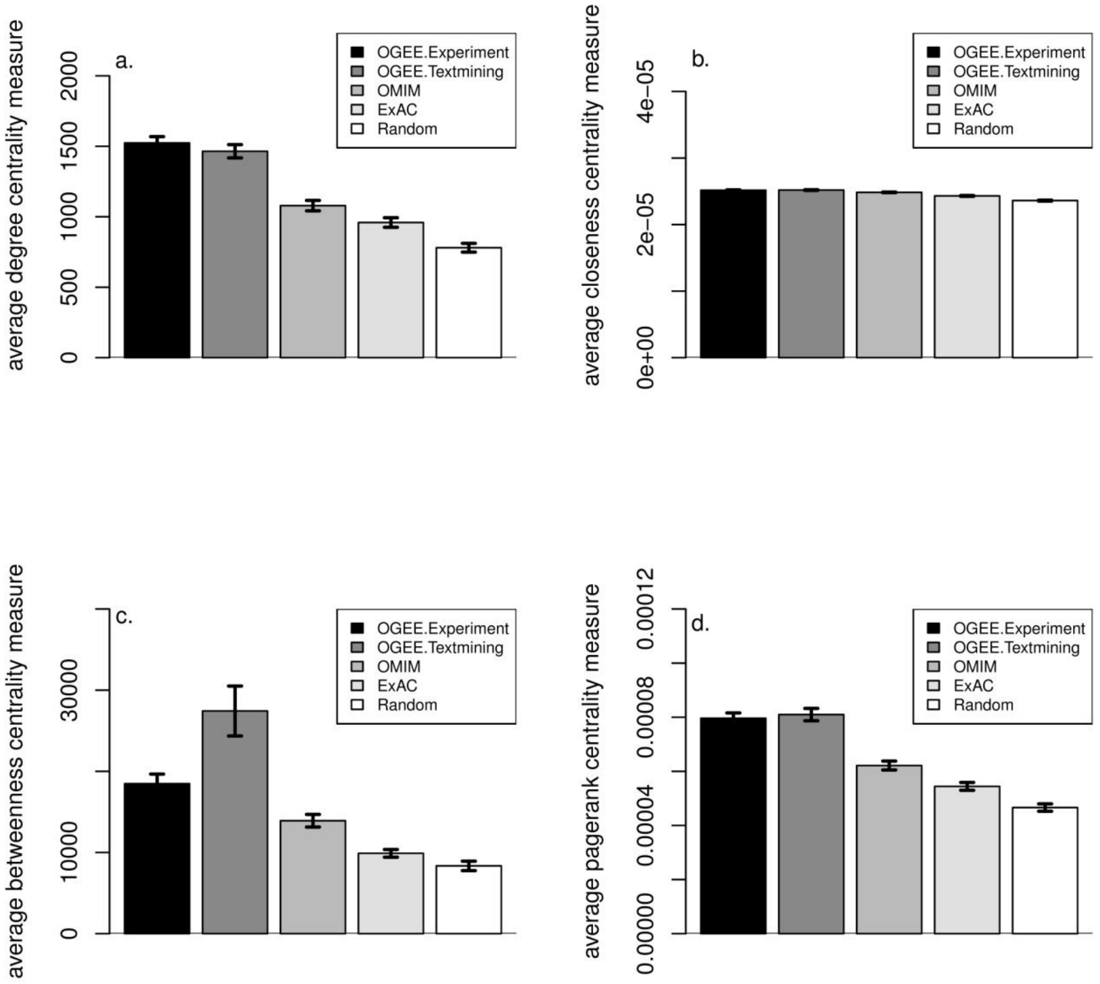

Figure 5.2 Comparison of the average centralities of the four essential gene sets. a) degree centrality. b) closeness centrality. c) betweenness centrality. d) pagerank centrality. A gene set tagged as “Random” was also included for the comparison. This gene set (with 850 genes) was randomly selected and was exclusive from the other four gene sets.

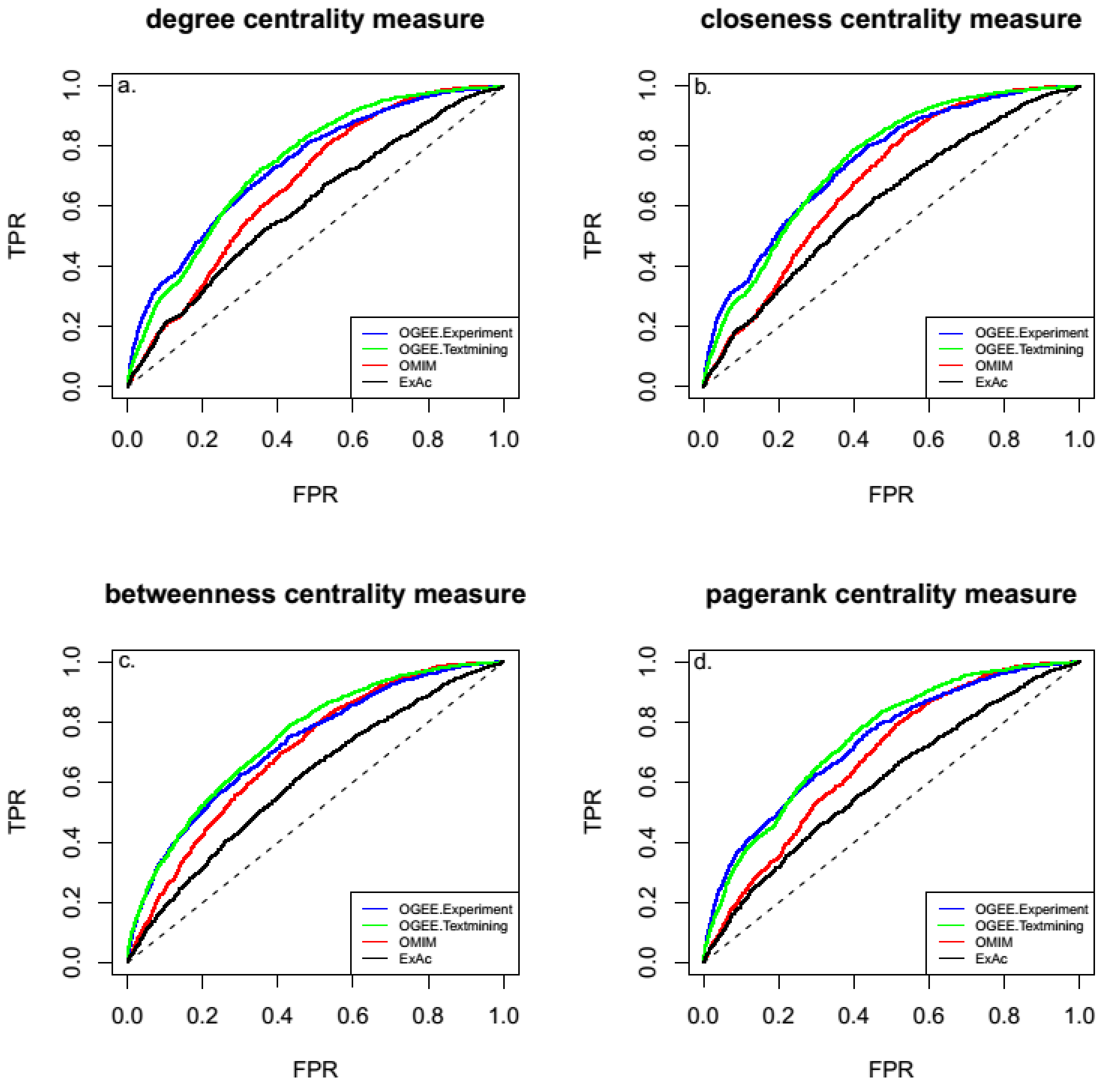

Figure 5.3 ROC curves of the average centralities aomng the four essential gene sets. a) degree centrality. b) closeness centrality. c) betweenness centrality. d) pagerank centrality. The X and Y axis are false positive rate (FPR) and true positive rate (TPR), respectively.

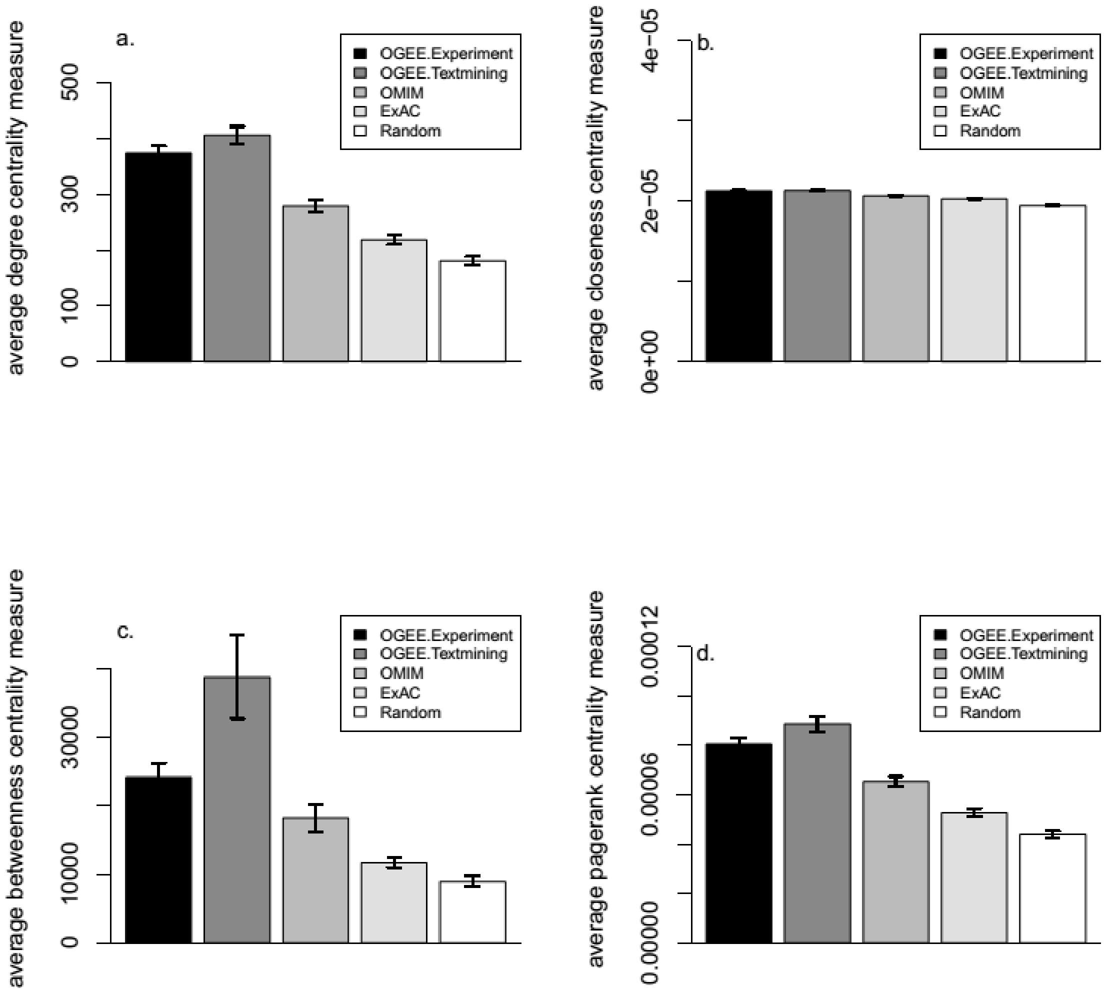

Figure 5.4 Comparison of the average centralities of the four essential gene sets using network constructed by 25% PPI data. a) degree centrality. b) closeness centrality. c) betweenness centrality. d) pagerank centrality. A gene set tagged as “Random” was also included for the comparison. This gene set (with 850 genes) was randomly selected and was exclusive from the other four gene sets.

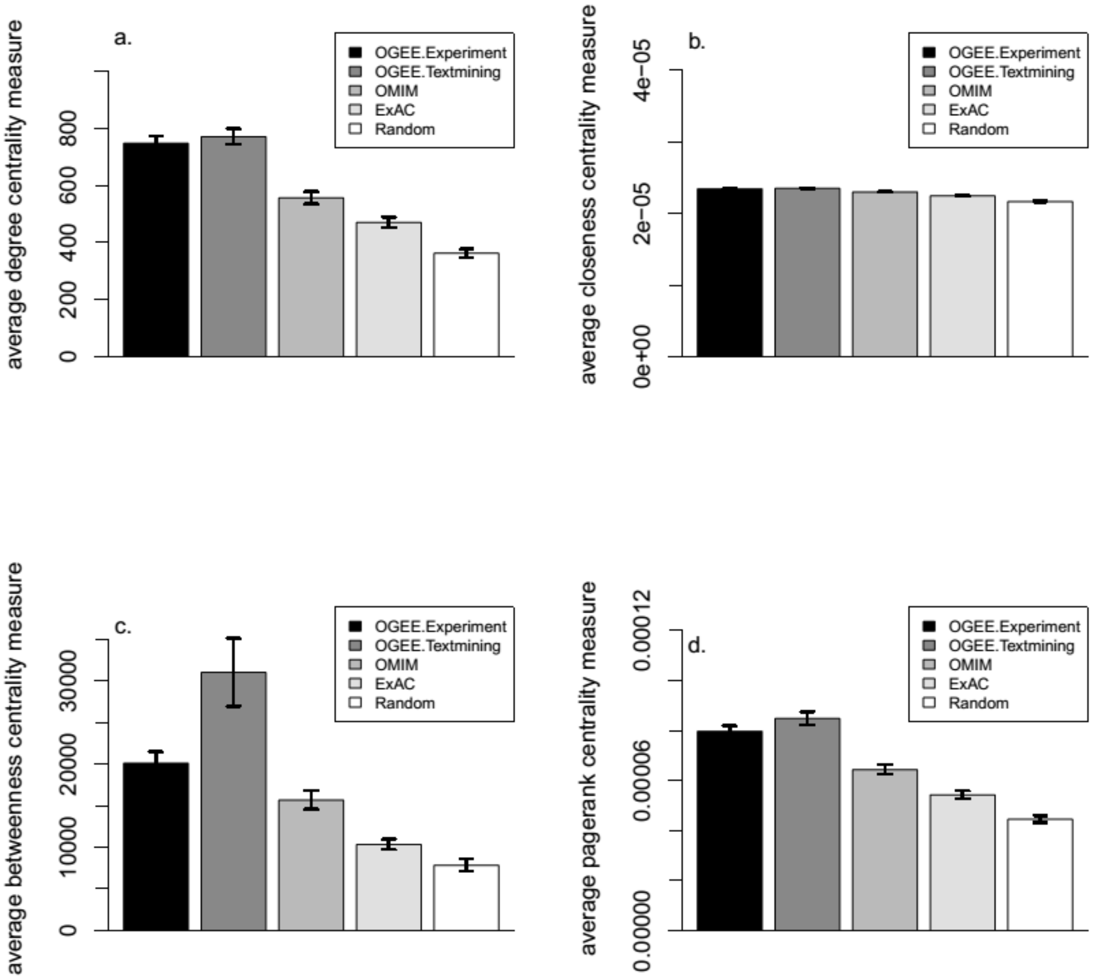

Figure 5.5 Comparison of the average centralities of the four essential gene sets using network constructed by 50% PPI data. a) degree centrality. b) closeness centrality. c) betweenness centrality. d) pagerank centrality. A gene set tagged as “Random” was also included for the comparison. This gene set (with 850 genes) was randomly selected and was exclusive from the other four gene sets.

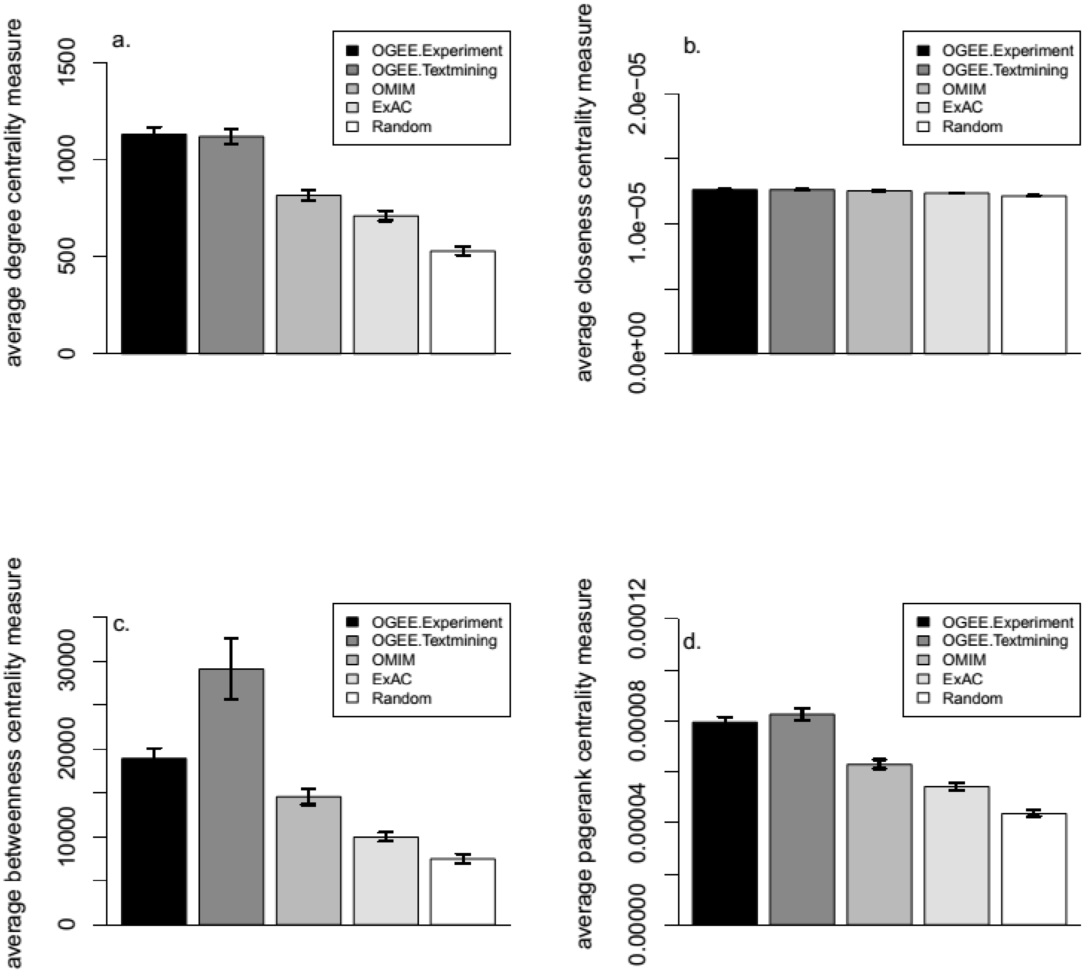

Figure 5.6 Comparison of the average centralities of the four essential gene sets using network constructed by 75% PPI data. a) degree centrality. b) closeness centrality. c) betweenness centrality. d) pagerank centrality. A gene set tagged as “Random” was also included for the comparison. This gene set (with 850 genes) was randomly selected and was exclusive from the other four gene sets.

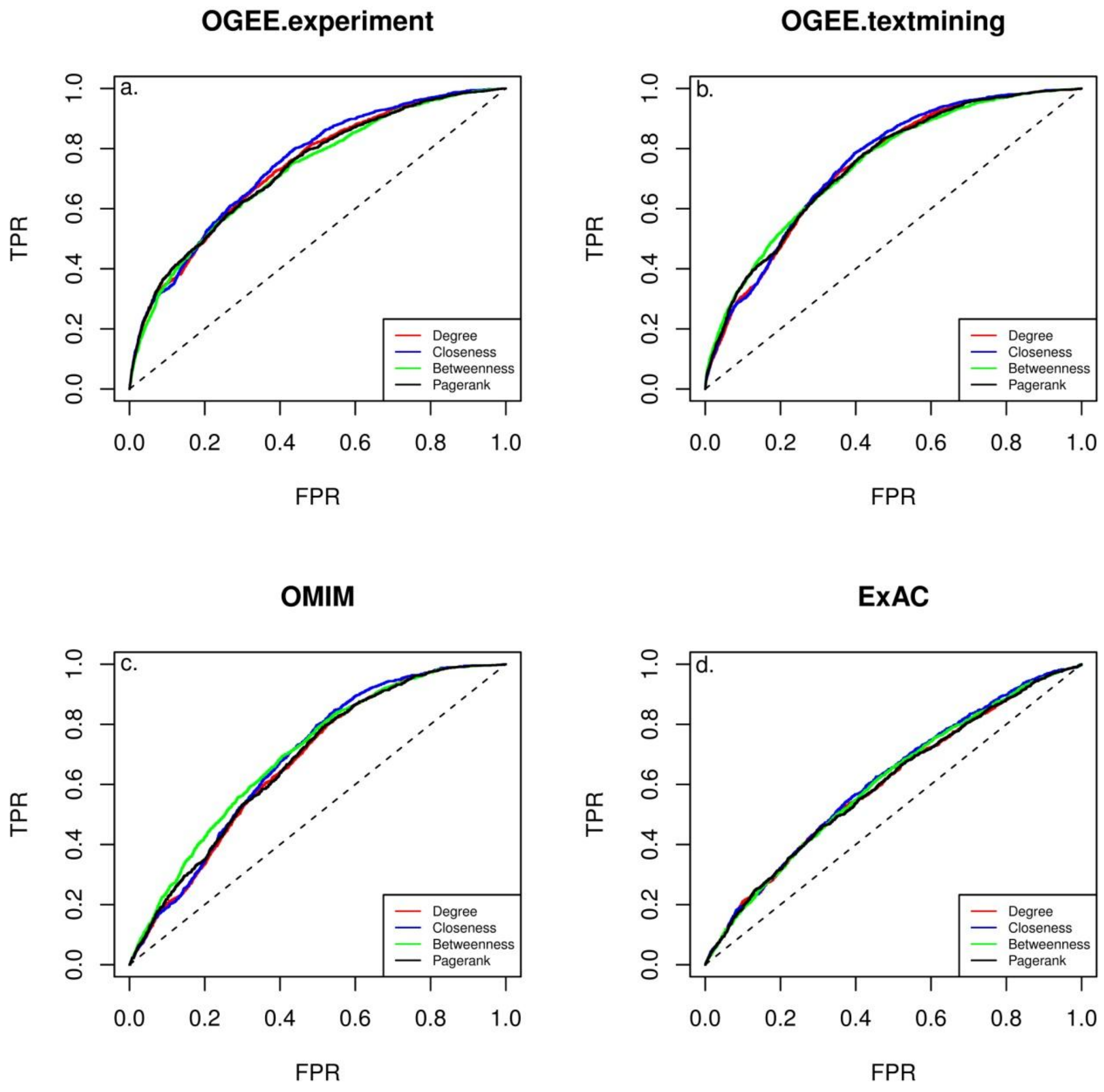

Figure 5.7 ROC curves for evaluating the four centrality measures as predictors of essentiality in human gene network. a) OGEE.experiment. b) OGEE.textmining. c) OMIM. d) ExAC.

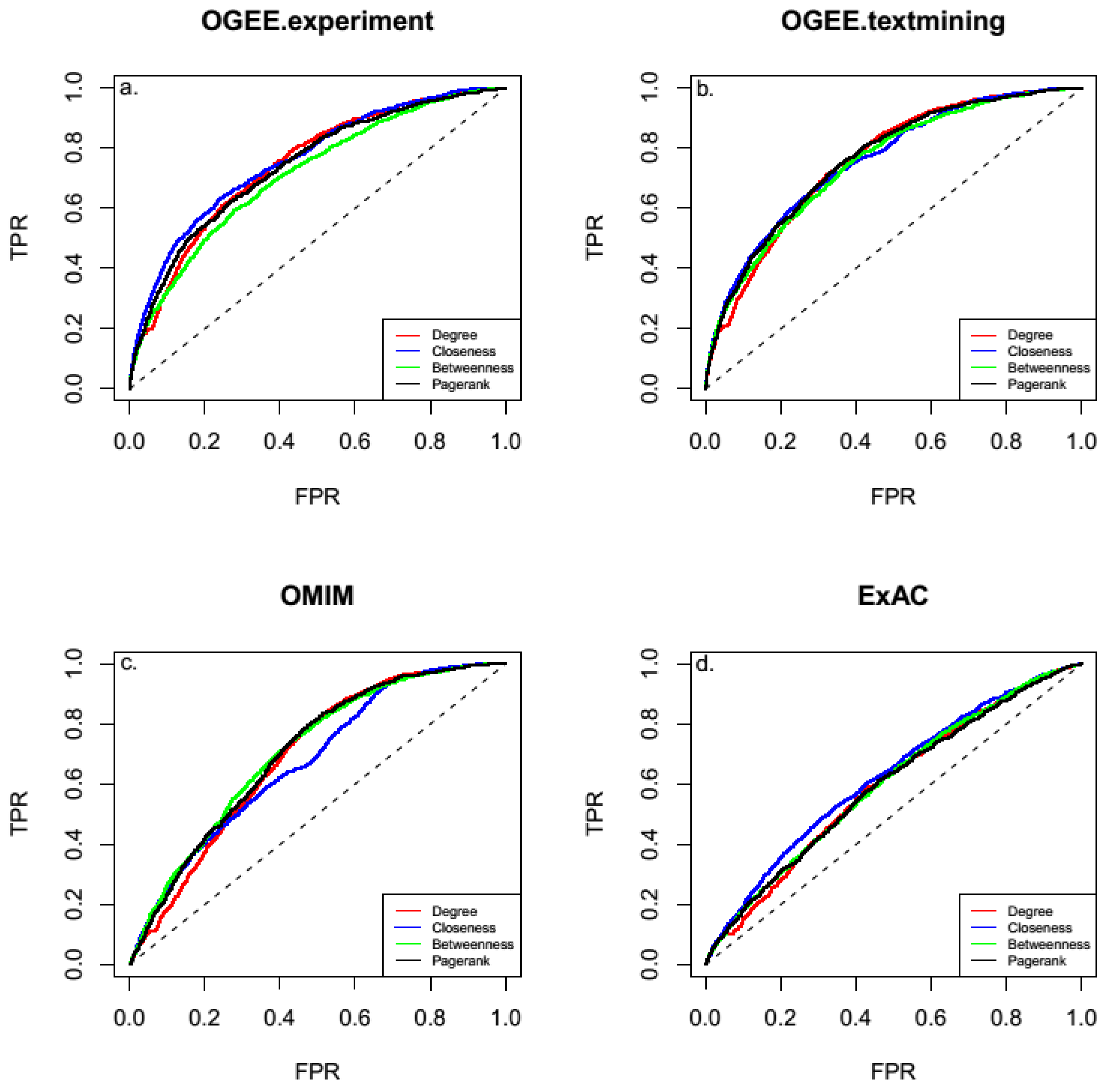

Figure 5.8 ROC curves for evaluating the four centrality measures as predictors of essentiality in human gene network constructed by 25% high quality PPI data. a) OGEE.experiment. b) OGEE.textmining. c) OMIM. d) ExAC.

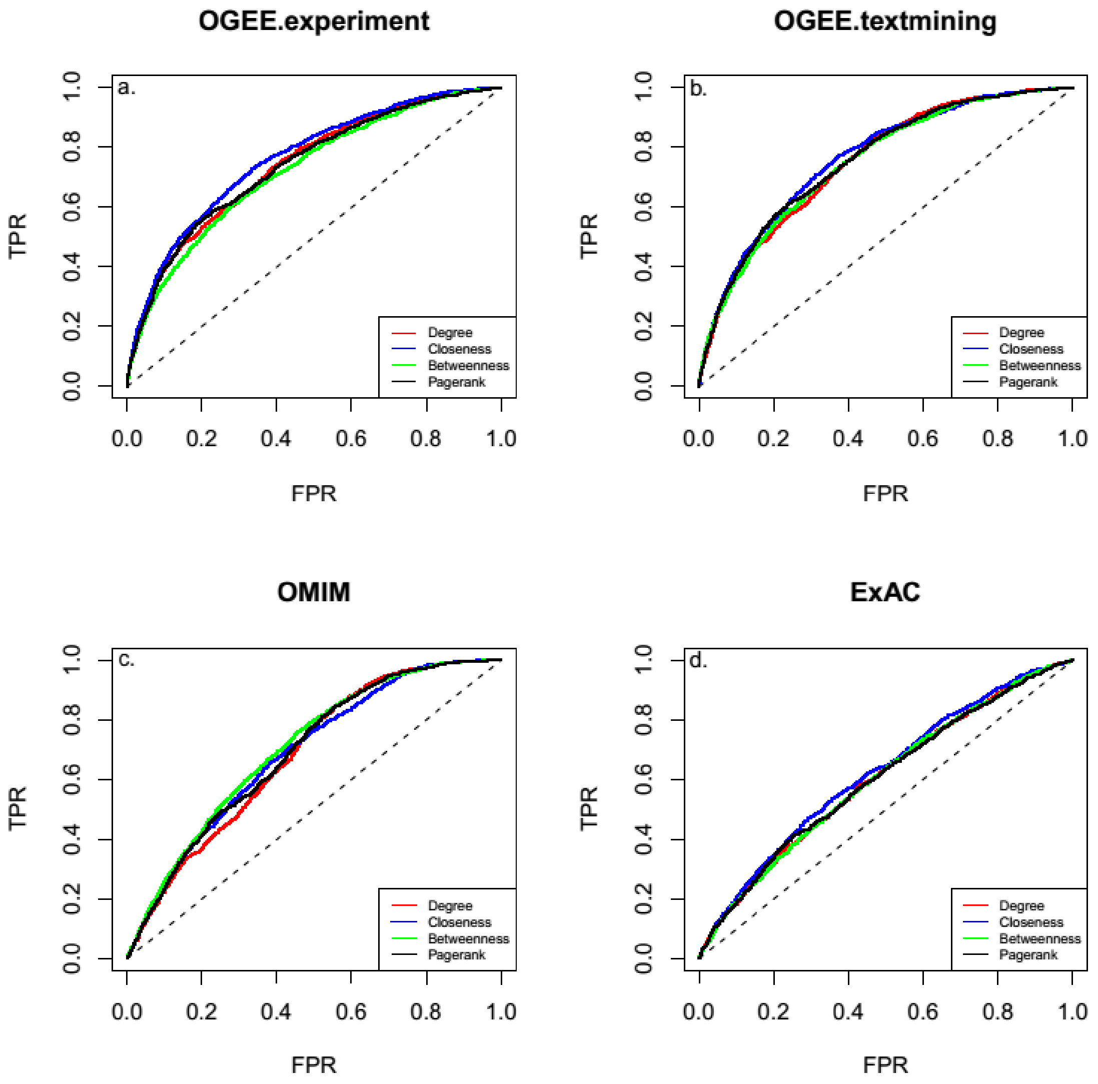

Figure 5.9 ROC curves for evaluating the four centrality measures as predictors of essentiality in human gene network constructed by 50% high quality PPI data. a) OGEE.experiment. b) OGEE.textmining. c) OMIM. d) ExAC.

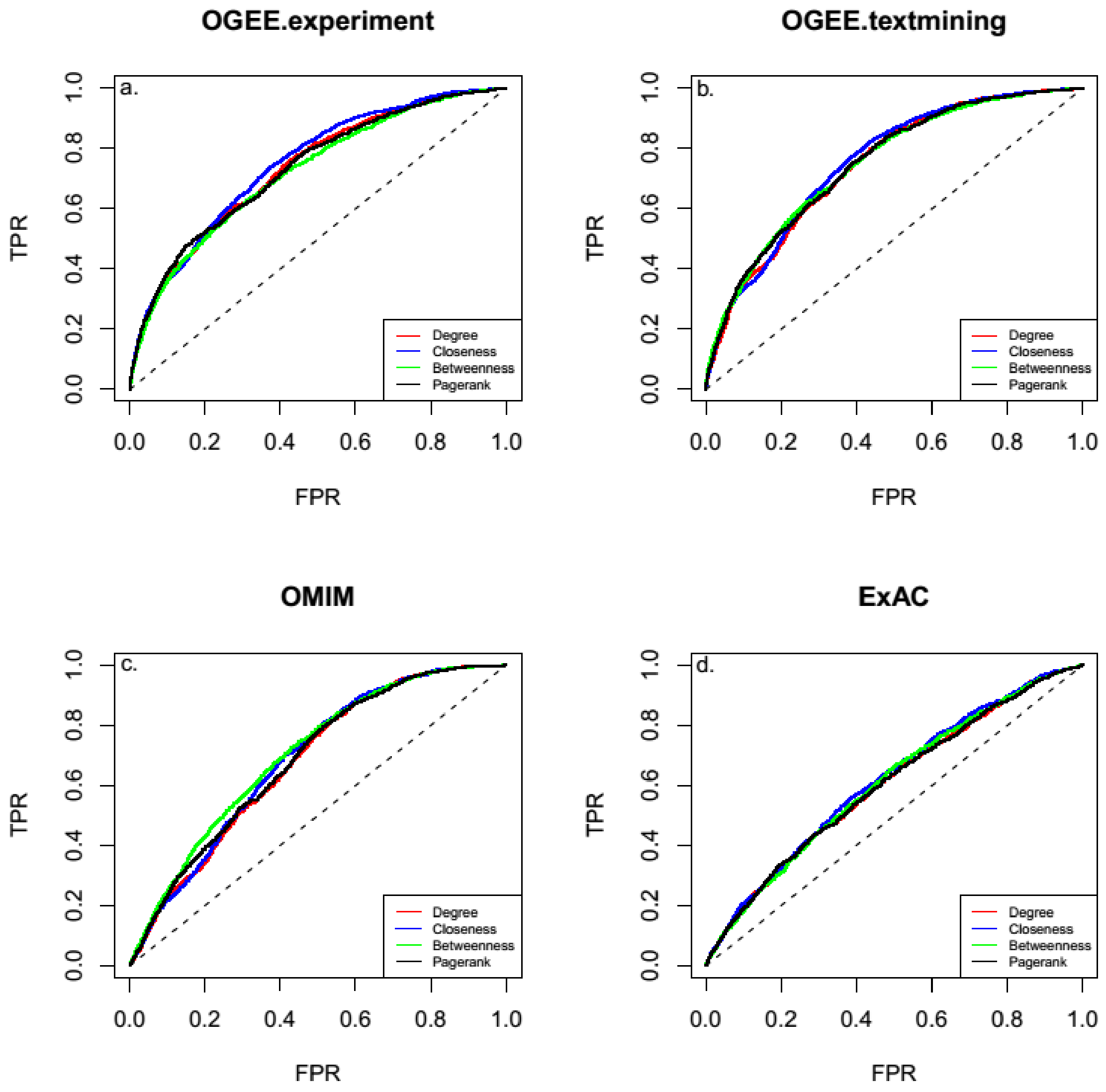

Figure 5.10 ROC curves for evaluating the four centrality measures as predictors of essentiality in human gene network constructed by 75% high quality PPI data. a) OGEE.experiment. b) OGEE.textmining. c) OMIM. d) ExAC.

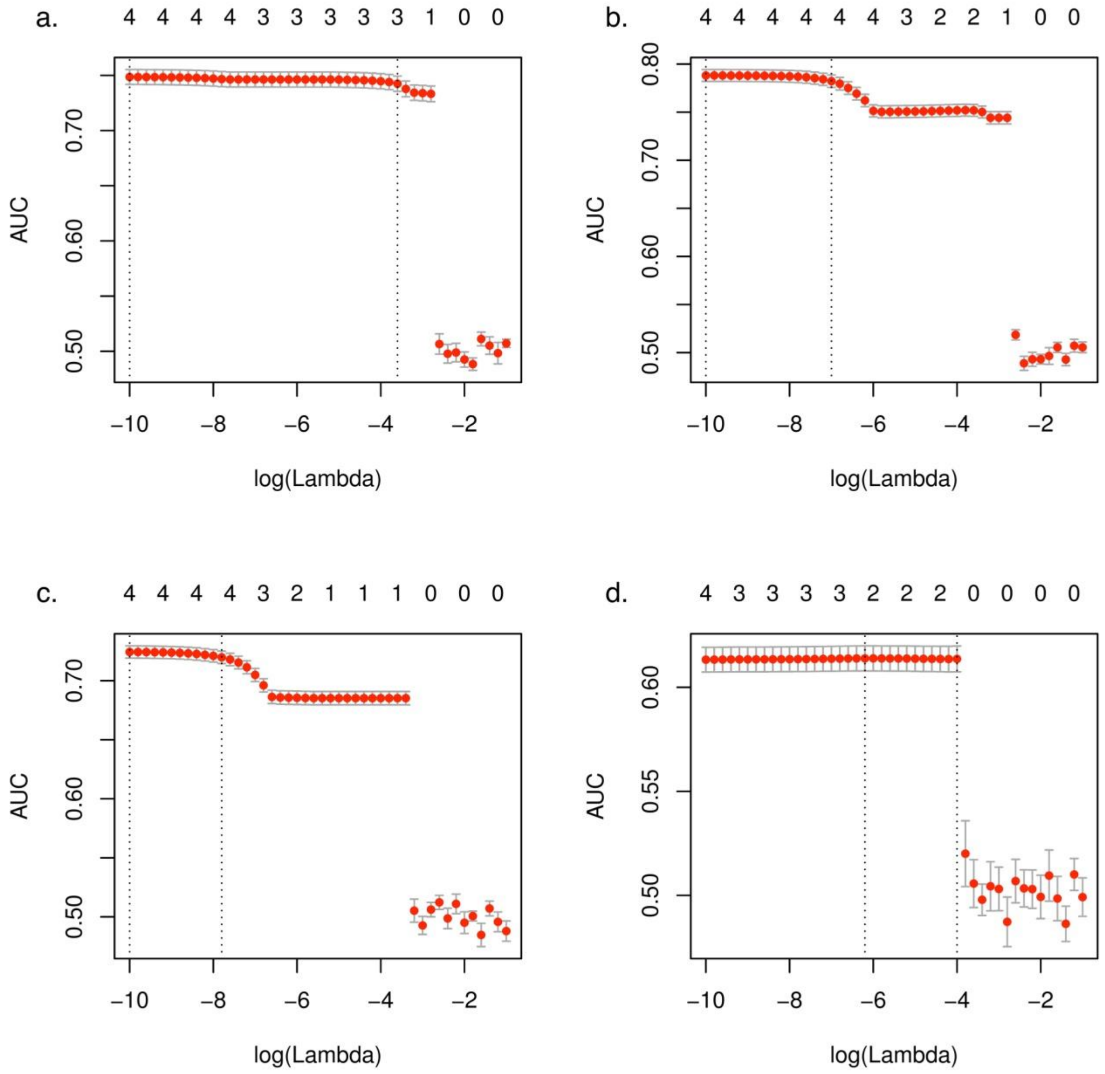

Figure 5.11 AUC plot based on four essential gene sets. a) OGEE.experiment. b) OGEE.textmining. c) OMIM. d) ExAC. Numbers at the top of each plot indicated the number of variables included in regression models. Only largest AUC measures were shown in this plot for different variable combinations.

5.6 Tables

Table 5.1 Number of genes in each essential gene sets and the pattern of overlaps.

	OGEE.experiment (%)	OGEE.textmining (%)	OMIM (%)	ExAC (%)	Covered by Network (%)
OGEE.experiment	1,511 (100)	298 (19.8)	170 (13.9)	150 (12.5)	1,441 (95.4)
OGEE.textmining	298 (19.7)	1,502 (100)	302 (24.7)	133 (11.1)	1,455 (96.9)
OMIM	170 (11.3)	302 (20.1)	1,224 (100)	110 (9.2)	1,191 (97.3)
ExAc	150 (9.9)	133 (8.9)	110 (9.0)	1,200 (100)	1,140 (95.0)

Table 5.2 Summarized information of the four human gene network.

	Combined Score threshold	Number of genes covered (% to human genome)
Network 100	/	18199 (82.6)
Network 75	175	18173 (82.5)
Network 50	212	18146 (82.4)
Network 25	317	18076 (82.1)

Table 5.3 Detail information of AUC measures for different models

Models	OGEE.experiment	OGEE.textmining	OMIM	ExAc
degree (1)	0.7345	0.7409	0.6703	0.5958
closeness (2)	0.7461	0.7483	0.6856	0.6131
betweenness (3)	0.6953	0.7280	0.5569	0.5324
pagerank (4)	0.7323	0.7440	0.6763	0.5976
(1)+(2)	0.7456	0.7479	0.6841	0.6135
(1)+(3)	0.7337	0.7400	0.6682	0.6001
(1)+(4)	0.7353	0.7637	0.7010	0.5971
(2)+(3)	0.7459	0.5514	0.6836	0.6131
(2)+(4)	0.7322	0.7492	0.6812	0.6133
(3)+(4)	0.7460	0.7432	0.6720	0.5986
(1)+(2)+(3)	0.7456	0.7489	0.6848	0.6122
(1)+(2)+(4)	0.7457	0.7859	0.7154	0.6109
(1)+(3)+(4)	0.7296	0.5390	0.7043	0.5986
(2)+(3)+(4)	0.7459	0.7496	0.6805	0.6122
(1)+(2)+(3)+(4)	0.7484	0.7882	0.7250	0.6115

Chapter 6: Centrality pattern of susceptibility genes to complex disorders in functional specific protein-protein interaction sub-networks

6.1 Introduction

The biological processes in humans are regulated through complex molecular networks. One of the most important features of such networks is that the effect caused by blocking one pathway within a network can be bypassed through some other "back door pathways". Accordingly, if a gene loses its function because of mutation, and if it is not located at the central position of a gene network, it may have little impact to the biological process due to the bypassing effects. On the other hand, if this gene has a relatively high centrality in the gene network, the loss of its function may block several pathways simultaneously and therefore no bypassing pathway can be used to supplement its loss of function. In chapter five, we have shown the centrality distribution of several essential gene sets in a general human gene network. In this chapter, we will present what the centrality distribution pattern would be in function specific sub-networks for several sets of susceptibility genes contributing to complex disorders.

6.2 Methods and Materials

6.2.1 Construction of human PPI network and calculation of centrality

We utilized known and predicted PPI data from STRING (http://string-db.org/) to construct our human PPI network [127]. Besides the human PPI network constructed using the full dataset from STRING, we also utilized the quality score provided by STRING to construct human PPI network with only 50% of the STRING data with higher quality score to validate the results we obtained from the network constructed using full PPI data. Centrality of genes was measured by degree centrality [148]. R package igraph was utilized for network construction and related analyses [141].

6.2.2 Construction of brain function related sub-networks

RNA-sequence data from the database The Human Protein Atlas (http://www.proteinatlas.org/) were utilized to construct brain function related sub-networks [149]. The number of Fragments Per Kilobase gene model and Million reads (PKFM) values of cerebral cortex were utilized as a filter criterion, and a series of brain function related sub-networks were constructed based on genes that have a PKFM value in cerebral cortex greater than a certain threshold. We labeled the human general PPI network as network No.0. Then, genes that have a PKFM value in cerebral cortex greater than 0.1 ×average PKFM in all human tissues (data of 44 tissues were recorded in the database) were extracted and constructed as sub-network No. 1. Genes that have a PKFM value in cerebral cortex greater than 0.2 ×average PKFM in all human tissues were extracted and constructed as sub-network No. 2, and so on so forth. 30 sub-networks were constructed, and the sub-networks with the larger number have fewer genes but these genes were more related with brain activity.

6.2.3 Extraction of susceptibility genes to complex disorders

Susceptibility genes of 25 complex disorders were extracted from the GWAS catalogue (http://www.ebi.ac.uk/gwas/) [7]. Reported genes were extracted from the GWAS catalogue for each disorder. These 25 complex disorders can be classified into seven classes including neurodegenerative disorders, psychiatric disorders, liver related disorders, skin related disorders, kidney related disorders, pancreas related disorders and lung related disorders, and the last five classes can be combined as non-brain related disorders (Table 6.1). Additionally, an essential gene set was also extracted from online gene essentiality database (OGEE, http://ogeedb.embl.de/) as a comparison set. This essential gene set was collected based on large scale experiments on model organisms with lethality as an important criterion for recruiting them. To control potential confounding factors, some disorders that were difficult to be defined as brain or non-brain related disorders (such as obesity and substance addiction) were not included in this study. In addition, this study mainly focused on complex disorders as a qualitative variable, genes that only affect quantitative medical indicators of the disorders were not included.

6.2.4 Statistical analyses

Two levels of enrichment pattern of susceptibility genes to complex disorders were analyzed and compared among all these three major categories of complex disorders. The first level is on the number of genes and the second level is the average/median centrality of genes. We utilized a 1,000 permutation technique to compare these enrichment patterns of the three complex disorder categories in a series of brain function related sub-networks. For the centrality level, null distribution of the average/median centrality for each gene set was created through permutation in each sub-networks and the general human network. Then the *P* values of the observed average/median centrality of each gene set were calculated.

6.3 Results

6.3.1 Construction of general human networks and brain function related sub-networks

A total of 31 human PPI gene networks were constructed (1 general network and 30 brain function related sub-networks). Summarized information of these gene networks are shown in Table 6.2. The largest network, network No. 0 is a general human PPI gene network with 18,041 genes. The smallest sub-networks, network No. 30, is constructed by genes that were 3 times expressed in cerebral cortex comparing to the average level. It only included 1,774 genes but these genes were highly expressed in brain and are highly related to brain activity.

6.3.2 Extraction of susceptibility genes to complex disorders and their centrality in multiple networks

A total of 468, 724 and 814 genes were extracted for neurodegenerative disorders, psychiatric disorders and non-brain related disorders (Table 6.3). In addition to these three gene sets, an essential gene set of 1,511 genes was also selected. The overlaps of genes among these three complex disorder categories were very limited. There are only

44 genes overlapped between neurodegenerative disorder and psychiatric disorder (9.4% and 6.1% of neurodegenerative disorder and psychiatric disorder, respectively). Most of these genes were covered by the general human PPI network (86%-91%). The degree centrality distribution of these three disorder categories for network 0, 6, 12, 18, 24 and 30 are shown in Figure 6.1. Severe positive skewness and multiple outliers were identified for all of these 6 networks. In this situation, median seems to be a more suitable parameter to characterize the distribution of centrality.

6.3.3 Enrichment pattern of susceptibility genes to complex disorders in brain related sub-networks

The enrichment pattern of susceptibility genes to complex disorders in gene number was shown in Figure 6.2. As we can see, the two susceptibility gene sets related to brain showed significant differences from the permuted background, while the susceptibility gene sets of non-brain related disorders did not showed significant enrichment pattern. The enrichment in high centrality of these susceptibility genes were shown in Figure 6.3. Apparently, the median centrality of gene sets susceptible to neurodegenerative and psychiatric disorders become more and more significantly higher in a series of brain function related sub-networks as the genes constructed in these networks become more and more highly expressed in human brain. On the other hand, the gene set chosen by susceptibly to non-brain related disorders failed to show this pattern. Although it is significant in the general human network (network 0), it did not become more significant in more functional specific sub-networks. We also checked this using the average of degree centrality (Figure 6.4), and a similar pattern was obtained. To examine the potential effects of PPI data quality to our results, we also conducted this analysis in a network constructed by 50% higher quality PPI data (Figure 6.5). The results indicated that the effects of quality of PPI data were limited. The centrality and gene number enrichment for each of these five non-brain related complex disorder classes are shown in Figure 6.6-6.10. The significance identified in network 0 was only identified in pancreas related susceptibility gene sets. The characteristics of average degree centrality for multiple gene sets in different sub-networks were also shown in line plots (Figure 6.11). The top 20 genes with higher degree centrality in sub-network 30 are summarized in Table 6.4.

6.4 Discussion

Susceptibility genes for complex disorders were believed to be peripheral in human gene networks, because for complex disorders there are multiple genes each with smaller effects, so that each gene seems not that important [38]. Our findings have shown that this may be true in a general human gene network, however, in functional specific sub-networks, the genes that confer risk to a complex disorder might have significant higher degree centrality. The following two features can be obtained through the centrality distribution patterns identified through our analyses. Firstly, compared to the essential genes with lethal effects in model organisms, the three complex disorder related gene sets have a very peripheral distribution in general human networks and the series of brain function related sub-networks. Secondly, for both neurodegenerative and psychiatric disorders, their centralities become significantly higher in the brain function related sub-networks. This trend becomes more and more apparent when we utilized more extreme criteria (PKFM values) to define the sub-networks. On the other hand, susceptibility genes to non-brain related disorders failed to show this pattern. These two features indicated that for a certain type of complex disorder, the centrality of the susceptibility genes are significantly higher within properly defined sub-networks with specific function. Currently, there were several network based association mapping prioritization methods [138][139]. Our findings indicated that for these methods, the key point is what sub-networks were chosen, but not the fancy algorithms to incorporate *P* values with the network parameters. Because centrality of genes in a gene network only becomes meaningful when functional specific sub-networks are properly defined.

Another point interesting to note is that gene *BDNF* is in the 20 gene list (second largest degree centrality) with top centrality in network 30 (Table 6.4) but has not been extracted as a psychiatric or neurodegenerative disorders related genes. *BDNF* and its val66-to-met mutation have been reported to be associated with several psychiatric disorders including schizophrenia and bipolar disorder [150][151][152].We rechecked the GWAS catalogue records and found that this gene was not reported as a susceptible gene for any psychiatric disorders to date. However, this finding may indicate that centrality can be utilized as a promising parameter in prioritization of candidate genes conferring risk to complex disorders.

Since all of the data we utilized in this study were obtained through publically available databases, our study is partially confined by the completeness of current aggregation of relevant data. The targeted disorders of our research should be some complex disorders that have been intensively studied in the past, so that we can extract a larger number of susceptibility genes conferring risk for those disorders. In addition, these disorders should also be relatively more concentrated in a certain human organ or tissues, so it will make it easier to define a sub-network to test them and reduce the potential interference effects. This condition rules out some intensively studied disorders, such as obesity. The two kinds of brain related disorders fitted the two conditions. Both psychiatric disorders and neurodegenerative disorders have been studied thoroughly in the past decade in GWAS and many susceptibility genes for these disorders have been reported. With data from the Protein Atlas, it is easy for us to define brain function related sub-networks and this process is totally independent of the susceptibility gene sets we extracted. Tissue specific gene expression data is one way to define sub-networks. One of a major advantage is that the sub-network will be highly functional specific to certain human organs or tissues. Several other public available database, including KEGG and Gene Ontology can be utilized to define sub-networks in the future.

To sum up, in this study, we examined the distributions of centrality for susceptibility gene sets to complex disorders in multiple human gene function specific networks. We identified that susceptibility gene sets to complex disorder have significant higher centralities in properly defined function specific sub-networks.

6.5 Figures

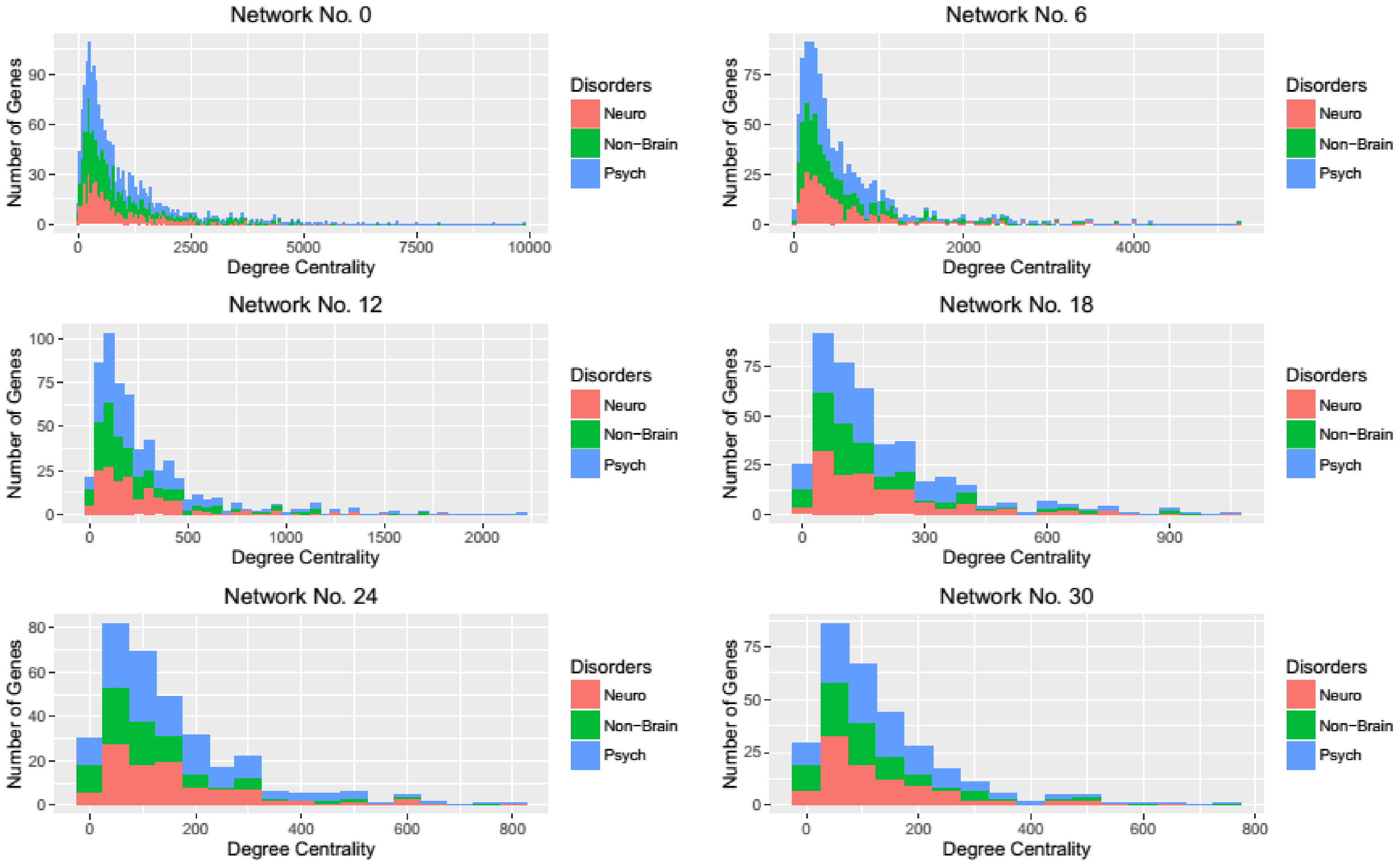

Figure 6.1 Distribution of degree centrality for the three disorder categories in network 0, 6, 12, 18, 24 and 30.

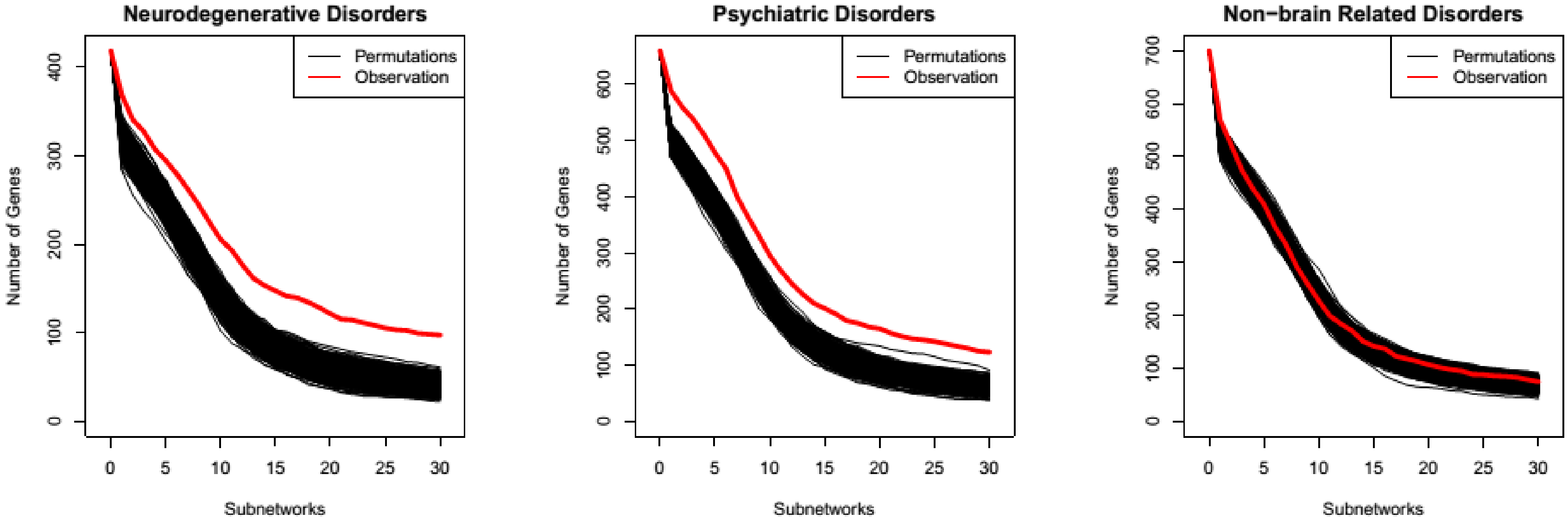

Figure 6.2 Number of genes of susceptible gene sets in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

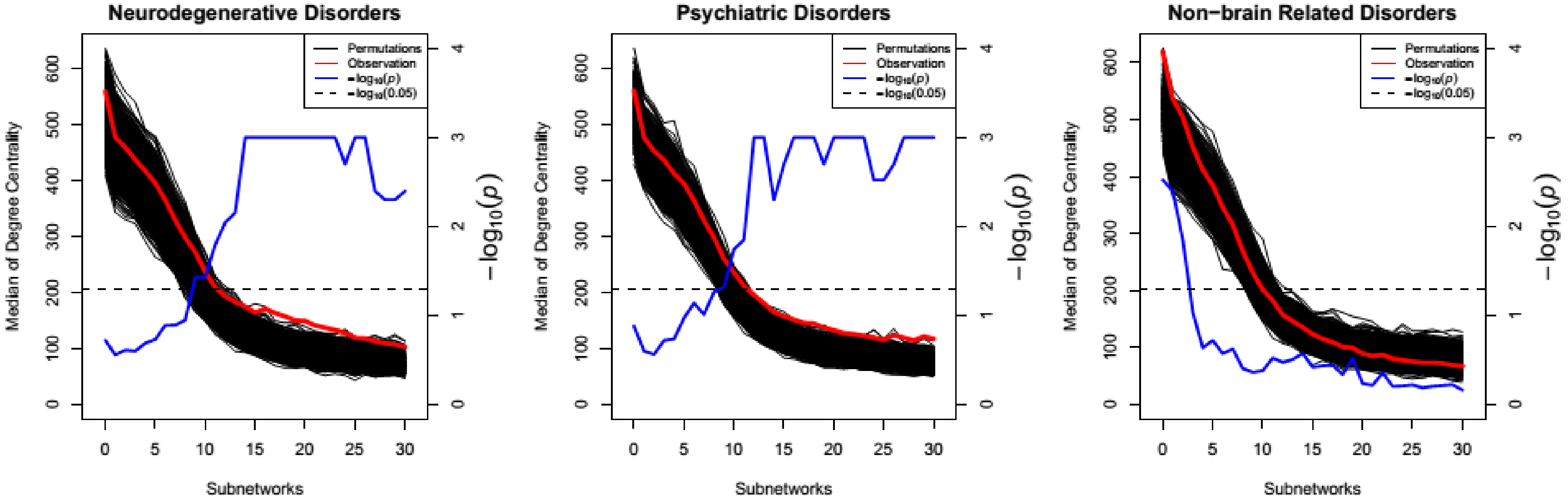

Figure 6.3 Median of degree centrality and their statistical significance for susceptible gene sets in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

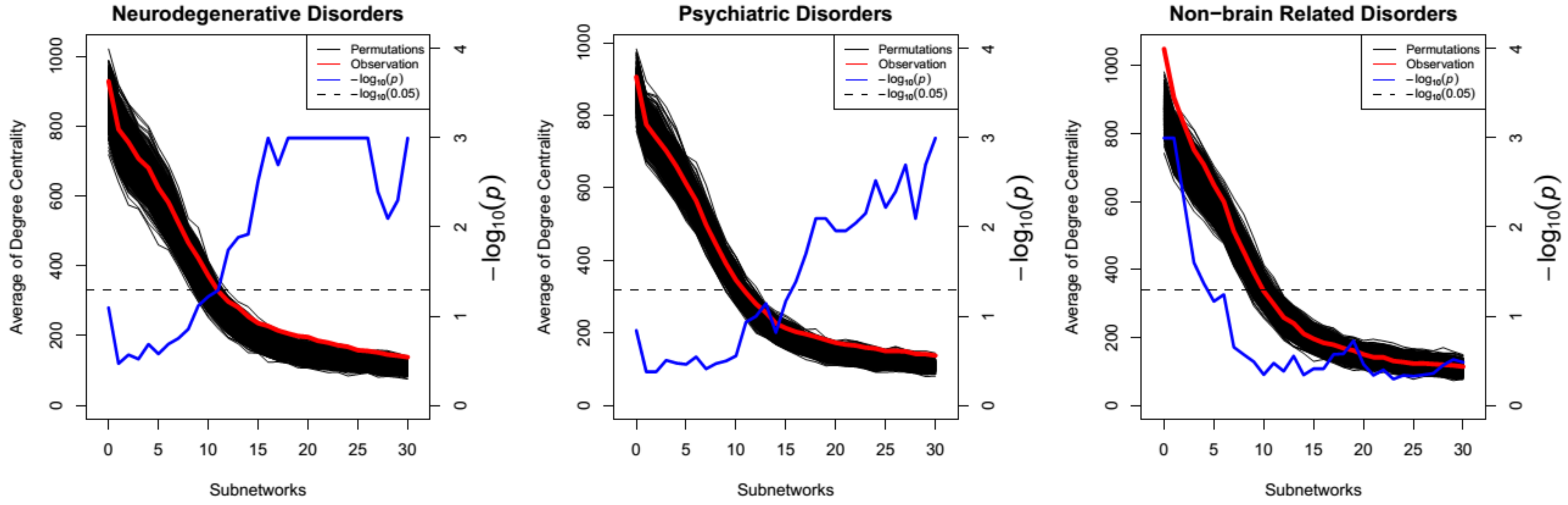

Figure 6.4 Average of degree centrality and their statistical significance for susceptible gene sets in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

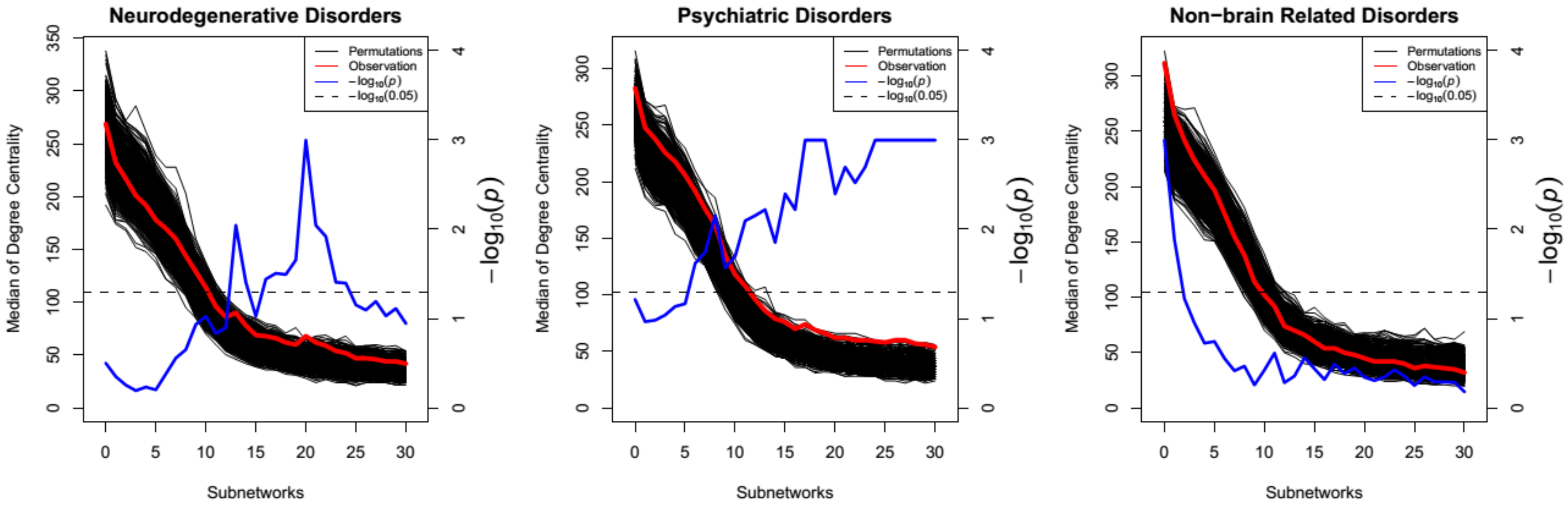

Figure 6.5 Median of degree centrality and their statistical significance for susceptible gene sets in multiple brain related sub-networks using 50% higher quality score PPI data. 1000 permutation was performed to create the null distribution.

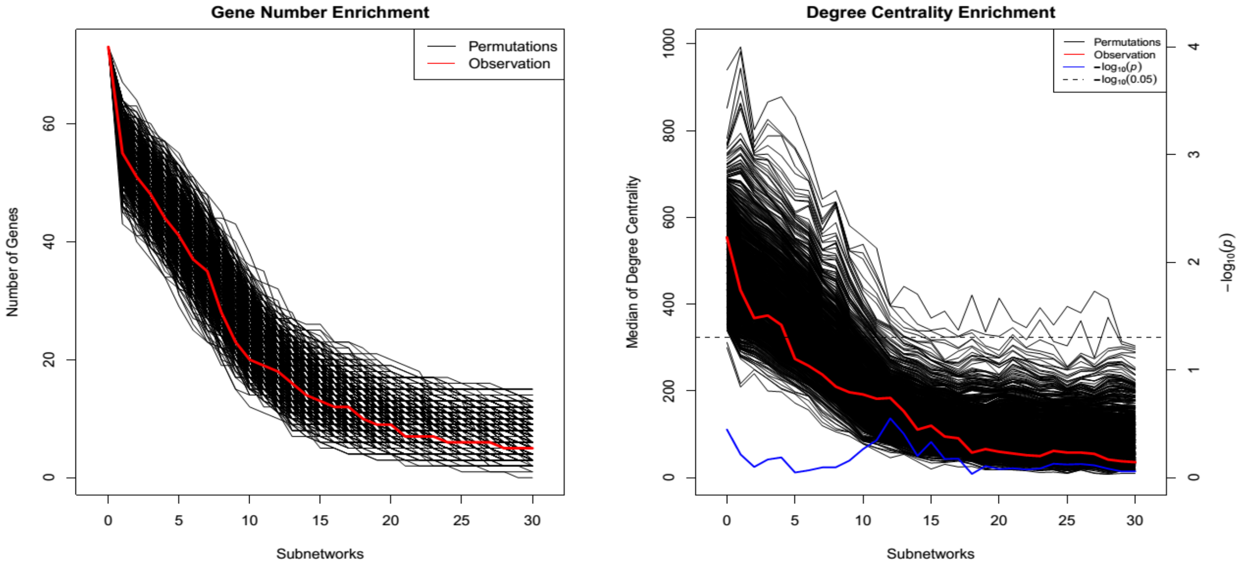

Figure 6.6 Gene number and centrality enrichment pattern for susceptibility genes to kidney related disorders in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

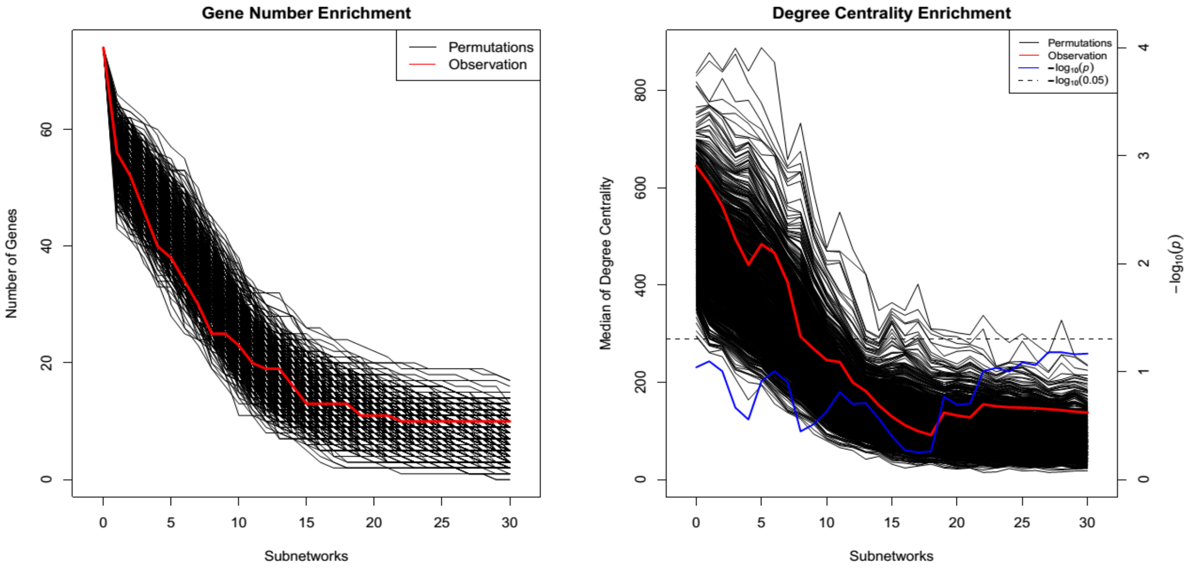

Figure 6.7 Gene number and centrality enrichment pattern for susceptibility genes to liver related disorders in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

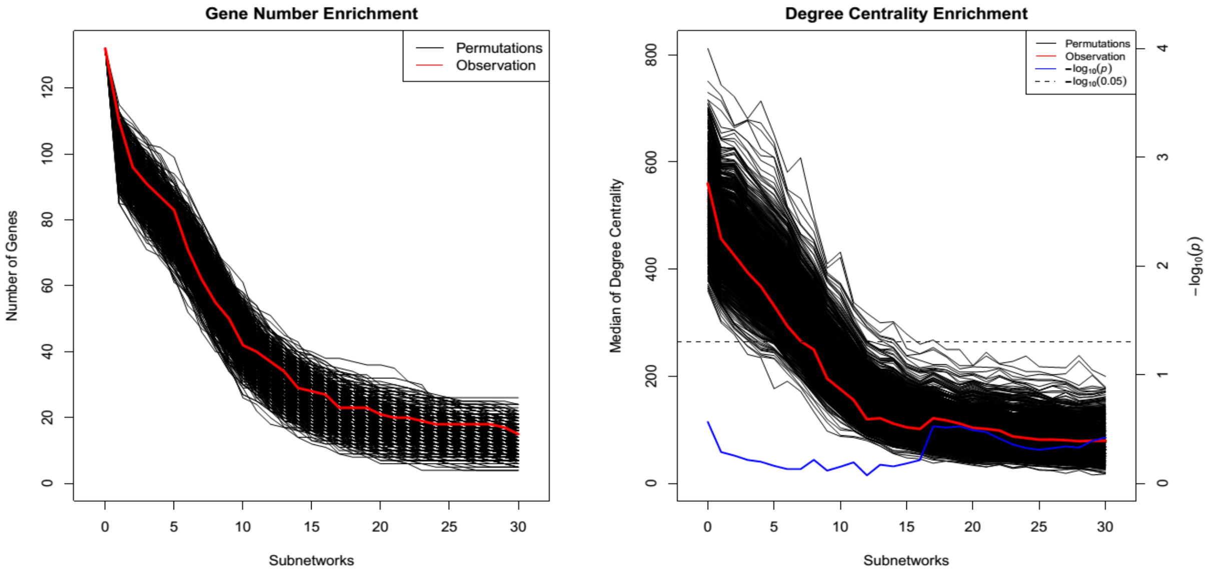

Figure 6.8 Gene number and centrality enrichment pattern for susceptibility genes to lung related disorders in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

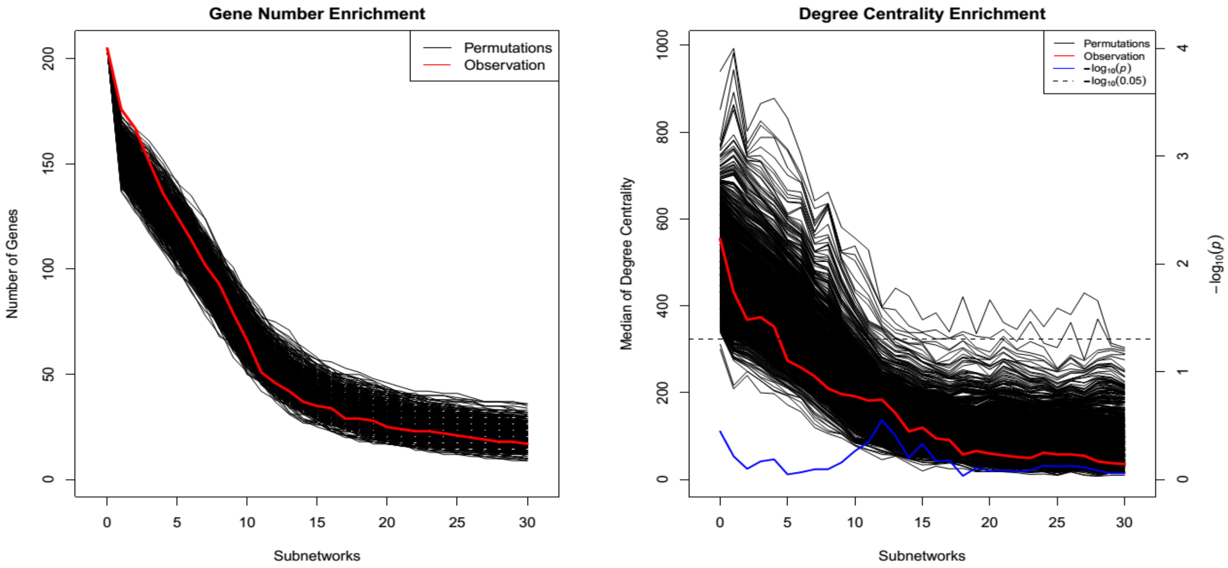

Figure 6.9 Gene number and centrality enrichment pattern for susceptibility genes to skin related disorders in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

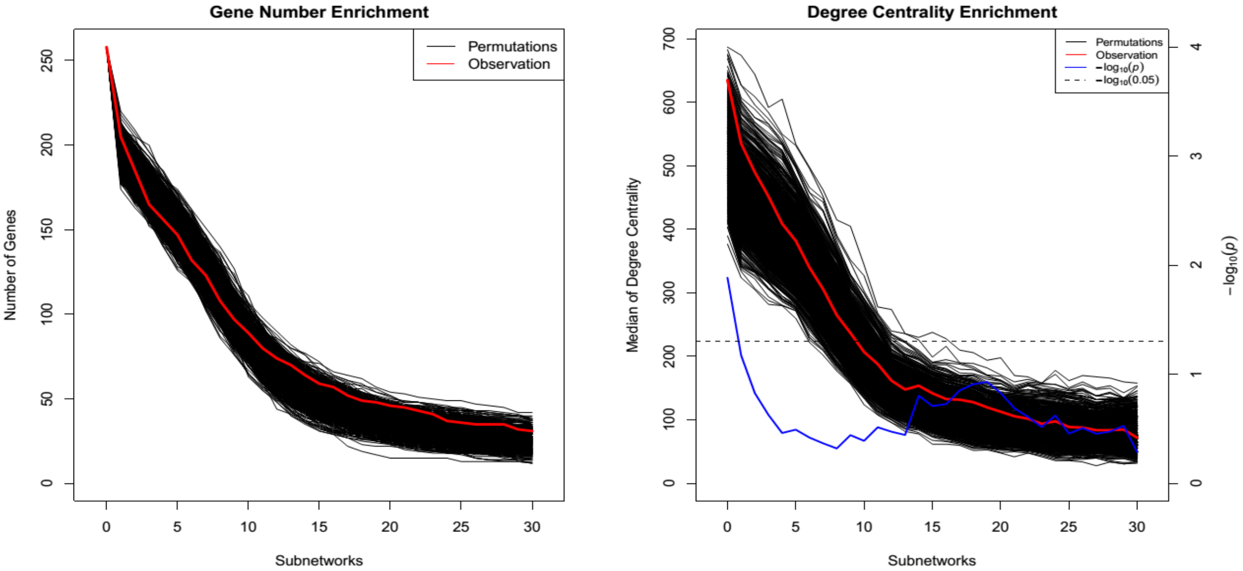

Figure 6.10 Gene number and centrality enrichment pattern for susceptibility genes to pancreas related disorders in multiple brain related sub-networks. 1000 permutation was performed to create the null distribution.

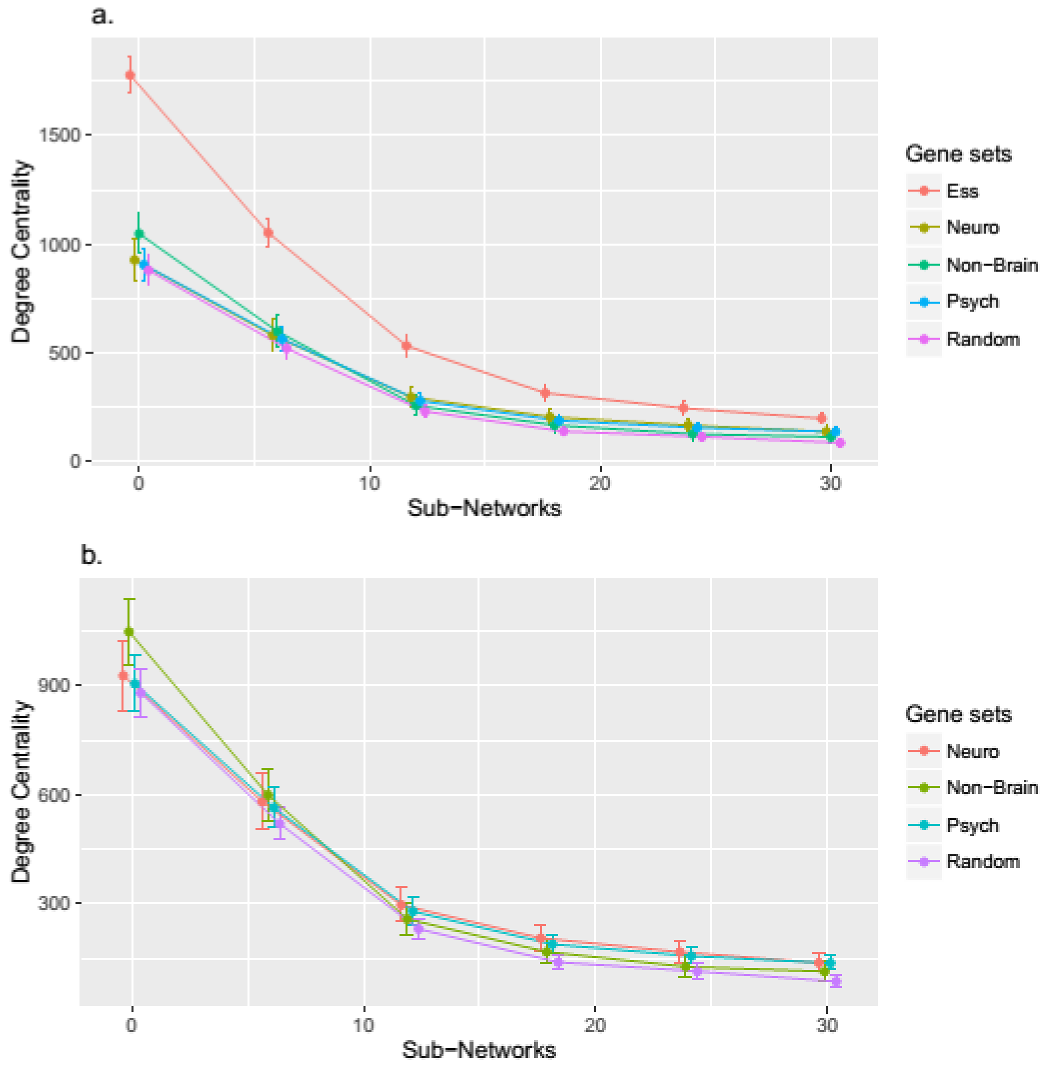

Figure 6.11 Line plots of the average degree centrality for susceptible genes in sub-network 0, 6, 12, 18, 24 and 30. A random gene set of 1,000 genes was selected serving as control. a. full plot; b. plot without essential gene set. 95% CI of the mean were added in the plot as the error bars.

6.6 Tables

Table 6.1 Complex disorders selected for this research project.

Disorder category	Disorders
Neurodegenerative disorder	Alzheimer’s Disease, Amyotrophic Lateral Sclerosis (ALS), Parkinson’s Disease, Dementia, Eplipcy
Psychiatric disorder	Schizophrenia, bipolar disorder, major depression, autisms, Attention Deficit Hyperactivity Disorder (ADHD)
Liver related disorder*	Hapetitis, Non-alcoholic fatty liver disease, Primary biliary cirrhosis, liver carcinoma
Skin related disorder*	Psoriasis, Melanoma, Acne, Non-melanoma skin cancer
Kidney related disorder*	Chronic kidney disease (CDK)
Pancreas related disorder*	Type I&II diabetes
Lung related disorder*	Lung cancer, Chronic obstructive pulmonary disease (COPD), Asthma

*Non-brain related disorders

Table 6.2 FPKM thresholds and number of genes covered for the 31 networks constructed using STRING PPI data.

Networks	*FPKM threshold	No.gene	**No.covered.gene (%)
0	-	19589	18041 (92.1)
1	>0.1	14552	13748 (94.5)
2	>0.2	13740	12973 (94.4)
3	>0.3	12900	12174 (94.3)
4	>0.4	12075	11401 (94.4)
5	>0.5	11215	10602 (94.5)
6	>0.6	10298	9742 (94.6)
7	>0.7	9325	8810 (94.5)
8	>0.8	8322	7856 (94.4)
9	>0.9	7316	6901(94.3)
10	>1.0	6397	6034 (94.3)
11	>1.1	5637	5301 (94.0)
12	>1.2	5005	4702 (93.9)
13	>1.3	4508	4241 (94.1)
14	>1.4	4091	3848 (94.1)
15	>1.5	3741	3517 (94.0)
16	>1.6	3458	3255 (94.1)
17	>1.7	3226	3038 (94.2)
18	>1.8	3048	2870 (94.2)
19	>1.9	2860	2694 (94.2)
20	>2.0	2707	2547 (94.1)
21	>2.1	2564	2412 (94.1)
22	>2.2	2464	2317 (94.0)
23	>2.3	2351	2212 (94.1)
24	>2.4	2269	2138 (94.2)
25	>2.5	2185	2059 (94.2)
26	>2.6	2123	2002 (94.3)
27	>2.7	2053	1937 (94.4)
28	>2.8	1996	1885 (94.4)
29	>2.9	1941	1833 (94.4)
30	>3.0	1879	1774 (94.4)

* PKFM threshold is defined as number × average PKFM in all tissues.

** Genes covered by STRING data.

Table 6.3 Information of gene numbers of the three major susceptibility genes categories

	Neuro (%)	Psych (%)	Non-brain (%)	Covered by general PPI network (%)
Neuro	468 (100)	44 (9.4)	36 (7.7)	419 (89.5)
Psych	44 (6.1)	724 (100)	39 (5.4)	660 (91.2)
Non-brain	36 (4.4)	39 (4.8)	814 (100)	700 (86.0)

Neurodegenerative disorders, psychiatric disorders and non-brain related disorders were indicated as Neuro, Psych and Non-brain, respectively.

Table 6.4 Top 20 genes with higher degree centrality in sub-network 30.

Gene	Ensembl ID	CHR	Degree Centrality	Disorders	Band	Description
CDK5	ENSG00000164885	7	710	-	q36.1	cyclin-dependent kinase 5
BDNF	ENSG00000176697	11	664	-	p14.1	brain-derived neurotrophic factor
FYN	ENSG00000010810	6	634	Neuro	q21	FYN proto-oncogene, Src family tyrosine kinase
MAPK11	ENSG00000185386	22	590	-	q13.33	mitogen-activated protein kinase 11
DIRAS2	ENSG00000165023	9	538	-	q22.2	DIRAS family, GTP-binding RAS-like 2
PPP3CA	ENSG00000138814	4	536	-	q24	protein phosphatase 3, catalytic subunit, alpha isozyme
SNCA	ENSG00000145335	4	526	Neuro	q22.1	synuclein alpha
RND1	ENSG00000172602	12	524	Psych	q13.12	Rho family GTPase 1
RIT2	ENSG00000152214	18	512	Neuro	q12.3	Ras-like without CAAX 2
RAP2A	ENSG00000125249	13	506	-	q32.1	RAP2A, member of RAS oncogene family
RND2	ENSG00000108830	17	504	-	q21.31	Rho family GTPase 2
DLG2	ENSG00000150672	11	504	Neuro	q14.1	discs, large homolog 2 (Drosophila)
PRKCA	ENSG00000154229	17	502	-	q24.2	protein kinase C, alpha
MAPK4	ENSG00000141639	18	496	-	q21.1	mitogen-activated protein kinase 4
PRKACB	ENSG00000142875	1	490	-	p31.1	protein kinase, cAMP-dependent, beta catalytic subunit
SST	ENSG00000157005	3	488	-	q27.3	somatostatin
GAD1	ENSG00000128683	2	478	-	q31.1	glutamate decarboxylase 1
GRIN2A	ENSG00000183454	16	472	Psych	p13.2	glutamate receptor, ionotropic, N-methyl D-aspartate 2A
YWHAH	ENSG00000128245	22	464	-	q12.3	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein
DIRAS1	ENSG00000176490	19	462	-	p13.3	DIRAS family, GTP-binding RAS-like 1

REFERENCES

[1] Gordis L. *Epidemiology, 4th Edition*.; 2009.

[2] Harris G, Galton F. Hereditary Genius. *J. Anthropol.* 1870;1(1):56. doi:10.2307/3024796.

[3] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am. J. Hum. Genet.* 2012;90(1):7-24. doi:10.1016/j.ajhg.2011.11.029.

[4] Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. *Science (80-.).* 1996;273(5281):1516-1517. doi:10.1126/science.273.5281.1516.

[5] Hacia JG, Fan JB, Ryder O, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* 1999;22(2):164-7. doi:10.1038/9674.

[6] Haines JL, Hauser M a, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 2005;308(5720):419-421. doi:10.1126/science.1110359.

[7] Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014;42(D1). doi:10.1093/nar/gkt1229.

[8] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-53. doi:10.1038/nature08494.

[9] Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 2010;11(6):446-50. doi:10.1038/nrg2809.

[10] Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 2011;88(3):294-305. doi:10.1016/j.ajhg.2011.02.002.

[11] McCarthy MI, Hirschhorn JN. Genome-wide association studies: Potential next steps on a genetic journey. *Hum. Mol. Genet.* 2008;17(R2). doi:10.1093/hmg/ddn289.

[12] Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 2008;24(3):133-141. doi:10.1016/j.tig.2007.12.007.

[13] Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* 2011;38(3):95-109. doi:10.1016/j.jgg.2011.02.003.

[14] 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526(7571):68-74. doi:10.1038/nature15393.

[15] Yuen RKC, Thiruvahindrapuram B, Merico D, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* 2015;21(2):185-91. doi:10.1038/nm.3792.

[16] Merico D, Zarrei M, Costain G, et al. Whole-Genome Sequencing Suggests Schizophrenia Risk Mechanisms in Humans with 22q11.2 Deletion Syndrome. *G3&amp;#58; Genes|Genomes|Genetics* 2015;(416):1-27. doi:10.1534/g3.115.021345.

[17] Fujimoto A, Totoki Y, Abe T, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* 2012;44(7):760-764. doi:10.1038/ng.2291.

[18] Zhang J, Wu G, Miller CP, et al. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat. Genet.* 2013;45(6):602-12. doi:10.1038/ng.2611.

[19] Wang K, Yuen ST, Xu J, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* 2014;46(6):573-82. doi:10.1038/ng.2983.

[20] Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 2008;83:311-321. doi:10.1016/j.ajhg.2008.06.024.

[21] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029.

[22] Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. *PLoS Genet.* 2011;7(3). doi:10.1371/journal.pgen.1001322.

[23] McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 2008;9(5):356-369. doi:10.1038/nrg2344.

[24] Flint J, Mackay TFC. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* 2009;19(5):723-733. doi:10.1101/gr.086660.108.

[25] Jarvis JP, Cheverud JM. Mapping the epistatic network underlying murine reproductive fatpad variation. *Genetics* 2011;187(2):597-610. doi:10.1534/genetics.110.123505.

[26] Shao H, Burrage LC, Sinasac DS, et al. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. U. S. A.* 2008;105(50):19910-19914. doi:10.1073/pnas.0810388105.

[27] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* 2009;10(6):392-404. doi:10.1038/nrg2579.

[28] Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010;467(7317):832-8. doi:10.1038/nature09410.

[29] Kim H-J, Barsevick AM, Fang CY, Miaskowski C. Common biological pathways underlying the psychoneurological symptom cluster in cancer patients. *Cancer Nurs.* 2012;35(6):E1-E20. doi:10.1097/NCC.0b013e318233a811.

[30] Jeong H, Mason SP, Barabási a L, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001;411(6833):41-42. doi:10.1038/35075138.

[31] Han J-DJ, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 2004;430(6995):88-93. doi:10.1038/nature02795.

[32] Schwartz SM, Schwartz HT, Horvath S, Schadt E, Lee SI. A systematic approach to multifactorial cardiovascular disease: Causal analysis. *Arterioscler. Thromb. Vasc. Biol.* 2012;32(12):2821-2835. doi:10.1161/ATVBAHA.112.300123.

[33] Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102(5):1572-7. doi:10.1073/pnas.0408709102.

[34] Foss EJ, Radulovic D, Shaffer S a, et al. Genetic basis of proteome variation in yeast. *Nat. Genet.* 2007;39(11):1369-1375. doi:10.1038/ng.2007.22.

[35] Akula N, Baranova A, Seto D, et al. A network-based approach to prioritize results from genome-wide association studies. *PLoS One* 2011;6(9). doi:10.1371/journal.pone.0024220.

[36] Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 2005;21(23):4205-4208. doi:10.1093/bioinformatics/bti688.

[37] Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006;22(18):2291-2297. doi:10.1093/bioinformatics/btl390.

[38] Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* 2007;104(21):8685-8690. doi:10.1073/pnas.0701361104.

[39] Almasy L, Dyer TD, Peralta JM, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proc.* 2014;8(Suppl 1):S2. doi:10.1186/1753-6561-8-S1-S2.

[40] Zhang Q, Chung D, Kraja A, Borecki II, Province M a. Methods for adjusting population structure and familial relatedness in association test for collective effect of multiple rare variants on quantitative traits. *BMC Proc.* 2011;5 Suppl 9(Suppl 9):S35. doi:10.1186/1753-6561-5-S9-S35.

[41] Fisher RA. The American Statistician. *Am. Stat.* 1948;2(5):30-31.

[42] Simes RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika* 1986;73(3):751-754. doi:10.1093/biomet/73.3.751.

[43] Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.* 2007;81(6):1278-1283. doi:10.1086/522374.

[44] Zhang T-X, Beaty TH, Ruczinski I. Candidate pathway based analysis for cleft lip with or without cleft palate. *Stat. Appl. Genet. Mol. Biol.* 2012;11(2):1-17. doi:10.2202/1544-6115.1717.

[45] Chen H, Meigs JB, Dupuis J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genet. Epidemiol.* 2013;37(2):196-204. doi:10.1002/gepi.21703.

[46] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164. doi:10.1093/nar/gkq603.

[47] Montgomery SB, Griffith OL, Sleumer MC, et al. ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 2006;22(5):637-640. doi:10.1093/bioinformatics/btk027.

[48] Lee OE, Braun TM. Permutation Tests for Random Effects in Linear Mixed Models. *Biometrics* 2012;68(2):486-493. doi:10.1111/j.1541-0420.2011.01675.x.

[49] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2(12):2074-2093. doi:10.1371/journal.pgen.0020190.

[50] Lehne B, Lewis CM, Schlitt T. From SNPs to genes: Disease association at the gene level. *PLoS One* 2011;6(6). doi:10.1371/journal.pone.0020133.

[51] Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-Based tests of association. *PLoS Genet.* 2011;7(7). doi:10.1371/journal.pgen.1002177.

[52] Kim LL, Fijal BA, Witte JS. Hierarchical modeling of the relation between sequence variants and a quantitative trait: addressing multiple comparison and population stratification issues. *Genet Epidemiol* 2001;21 Suppl 1:S668-73. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11793759.

[53] Merikangas KR, Akiskal HS, Angst J, et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Arch Gen Psychiatry* 2007;64(5):543-552. doi:64/5/543 [pii]\r10.1001/archpsyc.64.5.543.

[54] Anderson IM, Haddad PM, Scott J. Bipolar disorder. *BMJ* 2012;345:e8508.

[55] Chatterton M Lou, Ke X, Lewis BE, Rajagopalan K, Lazarus A. Impact of bipolar disorder on the family: utilization and cost of health care resources. *P T* 2008;33(1):15-34. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2730065&tool=pmcentrez&rendertype=abstract.

[56] Barnett JH, Smoller JW. The genetics of bipolar disorder. *Neuroscience* 2009;164(1):331-343. doi:10.1016/j.neuroscience.2009.03.080.

[57] Edvardsen J, Torgersen S, R??ysamb E, et al. Heritability of bipolar spectrum disorders. Unity or heterogeneity? *J. Affect. Disord.* 2008;106(3):229-240. doi:10.1016/j.jad.2007.07.001.

[58] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661-78. doi:10.1038/nature05911.

[59] Baum AE, Akula N, Cabanero M, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry* 2008;13(2):197-207. doi:10.1038/sj.mp.4002012.

[60] Ferreira MAR, O'Donovan MC, Meng YA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* 2008;40(9):1056-8. doi:10.1038/ng.209.

[61] McMahon FJ, Akula N, Schulze TG, et al. Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. *Nat. Genet.* 2010;42(2):128-131. doi:10.1038/ng.523.

[62] Cichon S, Muhleisen TW, Degenhardt FA, et al. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am. J. Hum. Genet.* 2011;88(3):372-381. doi:10.1016/j.ajhg.2011.01.017.

[63] Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* 2011;43(10):977-83. doi:10.1038/ng.943.

[64] Chen DT, Jiang X, Akula N, et al. Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol. Psychiatry* 2013;18(2):195-205. doi:10.1038/mp.2011.157.

[65] Mühleisen TW, Leber M, Schulze TG, et al. Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* 2014;5:3339. doi:10.1038/ncomms4339.

[66] Song J, Bergen SE, Di Florio A, et al. Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder. *Mol. Psychiatry* 2015;(September):1-8. doi:10.1038/mp.2015.165.

[67] McCarthy MI, Hirschhorn JN. Genome-wide association studies: past, present and future. *Hum. Mol. Genet.* 2008;17(R2):R100-R101. doi:10.1093/hmg/ddn298.

[68] Mardis ER. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 2008;9:387-402. doi:10.1146/annurev.genom.9.081307.164359.

[69] Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461(7261):272-6. doi:10.1038/nature08250.

[70] Cruceanu C, Ambalavanan A, Spiegelman D, et al. Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. *Genome* 2013;56(10):634-40. doi:10.1139/gen-2013-0081.

[71] McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.

[72] Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 2002;30(1):97-101. doi:10.1038/ng786.

[73] Jurg Ott DEW. SLINK: a general simulation program for linkage analysis. 2013.

[74] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31(13):3812-3814. doi:10.1093/nar/gkg509.

[75] Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 2011;32(8):894-899. doi:10.1002/humu.21517.

[76] Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007;81(3):559-575. doi:10.1086/519795.

[77] Bureau A, Younkin SG, Parker MM, et al. Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. *Bioinformatics* 2014;30(15):2189-2196. doi:10.1093/bioinformatics/btu198.

[78] Nyholt DR. All LODs are not created equal. *Am. J. Hum. Genet.* 2000;67(2):282-288. doi:10.1086/303029.

[79] Fallin MD, Lasseter VK, Wolyniec PS, et al. Genomewide linkage scan for bipolar-disorder susceptibility loci among Ashkenazi Jewish families. *Am. J. Hum. Genet.* 2004;75(2):204-19. doi:10.1086/422474.

[80] Yamazaki M, Araki K, Shibata A, Mishina M. Molecular cloning of a cDNA encoding a novel member of the mouse glutamate receptor channel family. *Biochem. Biophys. Res. Commun.* 1992;183(2):886-892. Available at: http://www.google.com/search?client=safari&rls=en-us&q=Molecular+cloning+of+a+cDNA+encoding+a+novel+member+of+the+mouse+glutamate+receptor+channel+family&ie=UTF-8&oe=UTF-8\npapers://9648de94-53eb-41f8-8282-2a8285ba1a70/Paper/p3613.

[81] Fallin MD, Lasseter VK, Avramopoulos D, et al. Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios. *Am. J. Hum. Genet.* 2005;77(6):918-936. doi:10.1086/497703.

[82] Guo S-Z, Huang K, Shi Y-Y, et al. A case-control association study between the GRID1 gene and schizophrenia in the Chinese Northern Han population. *Schizophr. Res.* 2007;93(1-3):385-390. doi:10.1016/j.schres.2007.03.007.

[83] Venken T, Alaerts M, Souery D, et al. Chromosome 10q harbors a susceptibility locus for bipolar disorder in Ashkenazi Jewish families. *Mol. Psychiatry* 2008;13(4):442-50. doi:10.1038/sj.mp.4002039.

[84] Nenadic I, Maitra R, Scherpiet S, et al. Glutamate receptor delta 1 (GRID1) genetic variation and brain structure in schizophrenia. *J. Psychiatr. Res.* 2012;46(12):1531-1539. doi:10.1016/j.jpsychires.2012.08.026.

[85] Yadav R, Gupta SC, Hillman BG, Bhatt JM, Stairs DJ, Dravid SM. Deletion of glutamate delta-1 receptor in mouse leads to aberrant emotional and social behaviors. *PLoS One* 2012;7(3). doi:10.1371/journal.pone.0032969.

[86] Ba-Abbad R, Sergouniotis PI, Plagnol V, et al. Clinical characteristics of early retinal disease due to CDHR1 mutation. *Mol. Vis.* 2013;19:2250-9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3834600&tool=pmcentrez&rendertype=abstract.

[87] Duncan JL, Roorda A, Navani M, et al. Identification of a novel mutation in the CDHR1 gene in a family with recessive retinal degeneration. *Arch. Ophthalmol. (Chicago, Ill. 1960)* 2012;130(10):1301-1308. doi:10.1001/archophthalmol.2012.1906.

[88] Ostergaard E, Batbayli M, Duno M, Vilhelmsen K, Rosenberg T. Mutations in PCDH21 cause autosomal recessive cone-rod dystrophy. *J. Med. Genet.* 2010;47(10):665-9. doi:10.1136/jmg.2009.069120.

[89] WHO Report on the Global Tobacco Epidemic. *WHO Rep. Glob. Tob. Epidemic* 2013;5:106. doi:10.1002/aehe.3640230702.

[90] Biesalski HK, de Mesquita BB, Chesson A, et al. European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel. *CA. Cancer J. Clin.* 1998;48(3):167-176. doi:10.3322/canjclin.48.3.167.

[91] Doll R, Hill a B. The mortality of doctors in relation to their smoking habits: a preliminary report. 1954. *BMJ* 2004;328(7455):1529-1533; discussion 1533. doi:10.1016/S0084-3954(07)70287-7.

[92] Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. *Br. J. Cancer* 2005;92(3):426-9. doi:10.1038/sj.bjc.6602359.

[93] worldlungfoundation, Atlas TT. The Tobacco atlas. *Tob. Atlas,* 2015;47:47-3559-47-3559. doi:10.5860/CHOICE.47-3559.

[94] Kennedy SM, Chambers R, Du W, Dimich-Ward H. Environmental and occupational exposures: do they affect chronic obstructive pulmonary disease differently in women and men? *Proc. Am. Thorac. Soc.* 2007;4(8):692-4. doi:10.1513/pats.200707-094SD.

[95] Laniado-Laborín R. Smoking and Chronic Obstructive Pulmonary Disease (COPD). Parallel Epidemics of the 21st Century. *Int. J. Environ. Res. Public Health* 2009;6(1):209-224. doi:10.3390/ijerph6010209.

[96] Tønnesen P. Smoking cessation and COPD. *Eur. Respir. Rev.* 2013;22(127):37-43. doi:10.1183/09059180.00007212.

[97] Carmelli D, Swan GE, Robinette D, Fabsitz R. Genetic influence on smoking--a study of male twins. *N. Engl. J. Med.* 1992;327(12):829-33. doi:10.1056/NEJM199209173271201.

[98] Vink JM, Willemsen G, Boomsma DI. Heritability of smoking initiation and nicotine dependence. *Behav. Genet.* 2005;35(4):397-406. doi:10.1007/s10519-004-1327-8.

[99] Huang W, Ma JZ, Payne TJ, Beuten J, Dupont RT, Li MD. Significant association of DRD1 with nicotine dependence. Hum. Genet. 2008;123(2):133-140. doi:10.1007/s00439-007-0453-9.

[100] Lerman C, Caporaso NE, Audrain J, et al. Evidence suggesting the role of specific genetic factors in cigarette smoking. Heal. Psychol 1999;18(1):14-20. doi:10.1037/0278-6133.18.1.14.

[101] Sabol SZ, Nelson ML, Fisher C, et al. A genetic association for cigarette smoking behavior. Health Psychol. 1999;18(1):7-13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9925040.

[102] Feng Y, Niu T, Xing H, et al. A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. *Am. J. Hum. Genet.* 2004;75(1):112-121. doi:10.1086/422194.

[103] Hong LE, Hodgkinson CA, Yang Y, et al. A genetically modulated, intrinsic cingulate circuit supports human nicotine addiction. *Proc. Natl. Acad. Sci.* 2010;107(30):13509-13514. doi:10.1073/pnas.1004745107.

[104] Saccone SF, Hinrichs AL, Saccone NL, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum. Mol. Genet.* 2007;16(1):36-49. doi:10.1093/hmg/ddl438.

[105] Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat. Genet.* 2010;42(5):448-53. doi:10.1038/ng.573.

[106] Mangold JE, Payne TJ, Ma JZ, Chen G, Li MD. Bitter taste receptor gene polymorphisms are an important factor in the development of nicotine dependence in African Americans. *J. Med. Genet.* 2008;45:578-582. doi:10.1136/jmg.2008.057844.

[107] Kremer I, Bachner-Melman R, Reshef A, et al. Association of the serotonin transporter gene with smoking behavior. *Am. J. Psychiatry* 2005;162(5):924-930. doi:10.1176/appi.ajp.162.5.924.

[108] Gerra G, Garofano L, Zaimovic a, et al. Association of the serotonin transporter promoter polymorphism with smoking behavior among adolescents. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* 2005;135B(1):73-8. doi:10.1002/ajmg.b.30173.

[109] Beuten J, Ma JZ, Payne TJ, et al. Single- and multilocus allelic variants within the GABA(B) receptor subunit 2 (GABAB2) gene are significantly associated with nicotine dependence. *Am. J. Hum. Genet.* 2005;76(5):859-64. doi:10.1086/429839.

[110] Bierut LJ, Madden PA, Breslau N, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007;16(1):24-35. doi:ddl441 [pii]\r10.1093/hmg/ddl441.

[111] Liu Y-Z, Pei Y-F, Guo Y-F, et al. Genome-wide association analyses suggested a novel mechanism for smoking behavior regulated by IL15. *Mol. Psychiatry* 2009;14(7):668-80. doi:10.1038/mp.2009.3.

[112] Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452(7187):638-42. doi:10.1038/nature06846.

[113] Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. *BMC Genet* 2007;8:10. doi:10.1186/1471-2156-8-10.

[114] Hall N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 2007;210(Pt 9):1518-1525. doi:10.1242/jeb.001370.

[115] Quail M, Smith ME, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* 2012;13(1):1. doi:10.1186/1471-2164-13-341.

[116] Tucker T, Marra M, Friedman JM. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *Am. J. Hum. Genet.* 2009;85(2):142-154. doi:10.1016/j.ajhg.2009.06.022.

[117] Olfson E, Saccone NL, Johnson EO, et al. Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. *Mol. Psychiatry* 2015. doi:10.1038/mp.2015.105.

[118] Bloom a J, Murphy SE, Martinez M, von Weymarn LB, Bierut LJ, Goate A. Effects upon in-vivo nicotine metabolism reveal functional variation in FMO3 associated with cigarette consumption. *Pharmacogenet. Genomics* 2013;23(2):62-8. doi:10.1097/FPC.0b013e32835c3b48.

[119] Krueger SK, Williams DE. Mammalian flavin-containing monooxygenases: Structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol. Ther.* 2005;106(3):357-387. doi:10.1016/j.pharmthera.2005.01.001.

[120] Hinrichs AL, Murphy SE, Wang JC, et al. Common polymorphisms in FMO1 are associated with nicotine dependence. *Pharmacogenet. Genomics* 2011;21(7):397-402. doi:10.1097/FPC.0b013e328346886f.

[121] Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. *Br. J. Addict.* 1991;86(9):1119-27. doi:doi:10.1111/j.1360-0443.1991.tb01879.x.

[122] Ripley BD. The R project in statistical computing. *MSOR Connect.* 2001;(January):23-25. doi:10.11120/msor.2001.01010023.

[123] Price A, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 2006;38(8):904-9. doi:10.1038/ng1847.

[124] Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002;296(5576):2225-2229. doi:10.1126/science.1069424.

[125] Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 2012;91(2):224-237. doi:10.1016/j.ajhg.2012.06.007.

[126] Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: Regional visualization of genome-wide association scan results. In: *Bioinformatics*.Vol 27.; 2011:2336-2337. doi:10.1093/bioinformatics/btq419.

[127] Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(D1):D447-D452. doi:10.1093/nar/gku1003.

[128] Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012;22(9):1790-1797. doi:10.1101/gr.137323.112.

[129] Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. XDynamic trans-acting factor colocalization in human cells. *Cell* 2013;155(3). doi:10.1016/j.cell.2013.09.043.

[130] MacKillop J, Obasi EM, Amlung MT, McGeary JE, Knopik VS. The Role of Genetics in Nicotine Dependence: Mapping the Pathways from Genome to Syndrome. *Curr. Cardiovasc. Risk Rep.* 2010;4(6):446-453. doi:10.1007/s12170-010-0132-6.

[131] Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biol.* 2010;8(1). doi:10.1371/journal.pbio.1000294.

[132] Hines RN, Hopp K a, Franco J, Saeian K, Begun FP. Alternative processing of the human FMO6 gene renders transcripts incapable of encoding a functional flavin-containing monooxygenase. *Mol. Pharmacol.* 2002;62(2):320-5. doi:10.1124/mol.62.2.320.

[133] Israni AK, Leduc R, Jacobson PA, et al. Inflammation in the setting of chronic allograft dysfunction post-kidney transplant: Phenotype and genotype. *Clin. Transplant.* 2013;27(3):348-358. doi:10.1111/ctr.12074.

[134] Park S, Lee NR, Lee KE, Park JY, Kim YJ, Gwak HS. Effects of single-nucleotide polymorphisms of FMO3 and FMO6 genes on pharmacokinetic characteristics of sulindac sulfide in premature labor. *Drug Metab. Dispos.* 2014;42(1):40-43. doi:10.1124/dmd.113.054106.

[135] Israni AK, Leduc R, Jacobson PA, et al. Inflammation in the setting of chronic allograft dysfunction post-kidney transplant: Phenotype and genotype. *Clin. Transplant.* 2013;27(3):348-358. doi:10.1111/ctr.12074.

[136] Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One* 2011;6(6). doi:10.1371/journal.pone.0020284.

[137] Newman MEJ. The mathematics of networks. *New Palgrave Encycl. Econ.* 2007;2:1-12. doi:10.1057/9780230226203.1064.

[138] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J. Med. Genet.* 2006;43(8):691-698. doi:10.1136/jmg.2006.041376.

[139] Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol. Syst. Biol.* 2008;4(189):189. doi:10.1038/msb.2008.27.

[140] Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 2011;21(7):1109-1121. doi:10.1101/gr.118992.110.

[141] Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal* 2006;Complex Sy:1695. doi:citeulike-article-id:3443126.

[142] Chen WH, Minguez P, Lercher MJ, Bork P. OGEE: An online gene essentiality database. *Nucleic Acids Res.* 2012;40(D1). doi:10.1093/nar/gkr986.

[143] Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* 2000;15(1):57-61. doi:10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G.

[144] Lek M, Karczewski K, Minikel E, et al. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* 2015:1-26. doi:http://dx.doi.org/10.1101/030338.

[145] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 2010;33(1):1-22. doi:10.1359/JBMR.0301229.

[146] He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* 2006;2(6):0826-0834. doi:10.1371/journal.pgen.0020088.

[147] Valente TW, Coronges K, Lakon C, Costenbader E. How Correlated Are Network Centrality Measures? *Connect. (Tor).* 2008;28(1):16-26.

[148] Borgatti SP. Centrality and network flow. *Soc. Networks* 2005;27(1):55-71. doi:10.1016/j.socnet.2004.11.008.

[149] Uhlen M, Fagerberg L, Hallstrom BM, et al. Tissue-based map of the human proteome. *Science (80-.).* 2015;347(6220):1260419-1260419. doi:10.1126/science.1260419.

[150] Sklar P, Gabriel SB, McInnis MG, et al. Family-based association study of 76 candidate genes in bipolar disorder. *Mol. Psychiatry* 2002;7:579-593. Available at: <Go to ISI>://000177078500011 .

[151] Neves-Pereira M, Mundo E, Muglia P, King N, Macciardi F, Kennedy JL. The brain-derived neurotrophic factor gene confers susceptibility to bipolar disorder: evidence from a family-based association study. *Am. J. Hum. Genet.* 2002;71(3):651-655. doi:10.1086/342288.

[152] Neves-Pereira M, Cheung JK, Pasdar a, et al. BDNF gene is a risk factor for schizophrenia in a Scottish population. *Mol. Psychiatry* 2005;10(2):208-212. doi:10.1038/sj.mp.4001575.

Curriculum Vitae

Tianxiao Zhang (张天啸), Ph.D. candidate

April 1st, 2016

Personal Information:

Birthplace:	Xi’an, Shaanxi Province, P.R.China
Birth date:	April 29th, 1985
Citizenship:	People’s Republic of China
Marital Status:	Married, One Child

Address and Telephone Numbers:

Work Address: Washington University School of Medicine
Department of Psychiatry
660 South Euclid Avenue
Campus Box 8134
Saint Louis, Missouri 63110-1093

Email: t.zhang@wustl.edu

Tel: +1 314-974-4729

Present Position:

Ph.D. candidate in human and statistical genetics program, DBBS, Washington University in Saint Louis

Education:

2016.8 Ph.D. Human and Statistical Genetics, Washington University, Saint Louis, MO, USA

2011.5 M.S., Genetic Epidemiology, Johns Hopkins University, Baltimore, MD, USA

2009.7 M.M.S., Forensic Science, Xi’an Jiaotong University, Xi’an, Shaanxi, China

2007.7 B.E., Biological Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, China

Working Experience:

2007.9-2009.5 Key Forensic Laboratory, Xi'an Jiaotong University, Research Assistant

2010.5-2011.5 Department of Epidemiology, Johns Hopkins University, Research Assistant

Presentation/Conference/Workshop:

2008.09	Conference	Y-STR haplotypes and the genetic structure from eight Chinese ethnic populations, 7th International Symposium Advances in Legal Medicine (ISALM) ,Osaka, Japan
2012.10	Conference	Application of noncollapsing methods to the gene-based association test: a comparison study using Genetic Analysis Workshop 18 data, Genetic Analysis Workshop 18 (GAW18), Stevenson, WA, USA
2013.06	Workshop	MBI-NIMBioS-CAMBAM Summer Graduate Workshop, Knoxville, TN, USA
2013.10	Conference	Exome Sequencing on Samples with Bipolar Disorder: A Preliminary Survey, 63th Annual Meeting of the American Society of Human Genetics, Boston, MA, USA
2013.12	Invited Scientific Talk	Genetic Epidemiology Study in Post-GWAS Era, Peking University, Beijing, China
2015.01	Invited Scientific Talk	Genetic Epidemiology study design in the in Post-GWAS Era, The Second Affiliated Hospital of Xi’an Jiaotong

		University, Xi’an, China
2015.10	Conference	Comparison and optimization of different centrality measures algorithms used in human gene network analysis, 65th Annual Meeting of the American Society of Human Genetics, Baltimore, MD, USA

Publications:

<u>Paper in press/under review</u>

1. Tianxiao Zhang, Fanglin Guan, Xinshe Liu, Wei Han, Huali Lin, Lu Li, Gang Chen, and Tao Liu. Evaluation of voltage-dependent calcium channel γ gene families identified several novel potential susceptible genes to schizophrenia. Scientific Reports (In press).
2. Tian-Xiao Zhang, Francis J. McMahon, David T. Chen, Jen C. Wang and John P. Rice. Exome sequencing of a large family identifies bipolar disorder candidate genes. Psychiatric Genetics (Under Review).
3. Tian-Xiao Zhang, Nancy L. Saccone, Laura J. Bierut and John P. Rice. Targeted sequencing identifies genetic polymorphisms of flavin-containing monooxygenase that contribute to the risk of nicotine dependence in European and African Americans. Brain and Behavior. (Under review)
4. Tianxiao Zhang, Fanglin Guan, Huali Lin, Lu Li, Jiali Feng, Dongke Fu, Nancy L Saccone and John P. Rice. Association of SNAP25 variants with schizophrenia and antipsychotic-induced weight change in large-scale schizophrenia patients of Han Chinese. British Journal of Psychiatry. (Under review).
5. Tianxiao Zhang, Xiaodi Jia, Lu Li, Dongke Fu,Huali Lin, Gang Chen, Xinshe Liu and Fanglin Guan. Two-stage additional evidence support association of common variants in the HDAC3 with the increasing risk of schizophrenia susceptibility. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. (Under Review).
6. Tianxiao Zhang, Fanglin Guan, Xinshe Liu, Wei Han, Huali Lin, Lu Li, Gang Chen and Tao Li. Two-stage replication of previous genome-wide association studies of AS3MT-CNNM2-NT5C2 gene cluster region in a large schizophrenia case–control sample from Han Chinese population. Schizophrenia Research. (Under review).
7. Li Miao, Gang Chen, Tianxiao Zhang, Lu Li, Chuchu Qiao, Jiali Feng, Dongke Fu, Fanglin Guan. Two-stage evaluation of genetic susceptibility of common variants in HTR1A to first episode schizophrenia and cognition performances in patients. Schizophrenia Research (Under review).

8. Yu Niu, Songfang Liu, Lei Ma, Ting Qi, Jia Feng, Hong Zuo, Guohong Li, Xufeng Liu, Tianxiao Zhang, Shujin Wang, Fanglin Guan. Two-stage comprehensive evaluation identified genetic susceptibility of a novel common variant of rs2075290 in ZPR1 to type 2 diabetes mellitus. Scientific Reports (Under review).

Paper published as first author

1. Zhang TX, Xie YR and Rice JP. Application of noncollapsing methods to the gene-based association test: a comparison study using Genetic Analysis Workshop 18 data. BMC Proceedings, 2014; 8(Suppl 1):S53.
2. Zhang TX, Haller G, Lin P, Alvarado DM et al. Genome-wide association study identifies new disease loci for isolated clubfoot. J Med Genet. 2014 May;51(5):334-9.
3. Tianxiao Zhang, Ingo Ruczinski, Terri H. Beaty. Candidate pathway based analysis for cleft lip with or without cleft palate. Statistical Applications in Genetics and Molecular Biology. 11:2,2012
4. ZHANG Tian-Xiao, YANG Li, LI Sheng-Bin. Y-STR Haplotypes and the Genetic Structure From 8 Chinese ethnic populations. Leg Med .11 Suppl 1: 198-200. 2009

Paper published as co-author

1. Bu L, Chen Q, Wang H, Zhang T, Hetmanski JB, Schwender H, Parker M, Chou YH, Yeow V, Chong SS, Zhang B, Jabs EW, Scott AF, Beaty TH. Novel evidence of association with nonsyndromic cleft lip with or without cleft palate was shown for single nucleotide polymorphisms in FOXF2 gene in an Asian population. Birth Defects Res A Clin Mol Teratol. 2015 Oct;103(10):857-62.
2. Zhang Bao, Gang Chen, Huali Lin, Tianxiao Zhang, Jiali Feng, Dongke Fu, Fanglin Guan. Common variants in SLC1A2 and schizophrenia: Association and cognitive function in patients with schizophrenia and healthy individuals. Schizophrenia.2015,169(1-3):128-34.
3. Yunzhi Zhang, Haiyan Liu, Chen Zhang, Tianxiao Zhang, Bo Zhang, Lu Li, Gang Chen, Dongke Fu and KunZheng Wang. Endochondral ossification pathway genes and postmenopausal osteoporosis: Association and specific allele related serum bone sialoprotein levels in Han Chinese. Sci Rep. 2015, 5:16783.

4. Guan F., Li L., Qiao C., Chen G., Yan T., Li T., Zhang T., Liu X. Evaluation of genetic susceptibility of common variants in CACNA1D with schizophrenia in Han Chinese. Sci Rep. 2015; 5: 12935.
5. Chen Q, Wang H, Schwender H, Zhang T, Hetmanski JB, Chou YH, Ye X, Yeow V, Chong SS, Zhang B, Beaty TH. Novel evidence of association with NSCL/P was shown for SNPs in FOXF2 gene in an Asian populatio". Birth Defects Research Part A: Clinical and Molecular Teratology. 2015, 103(10), 857-62.
6. Elizabeth J Leslie, M Adela Mansilla, Leah C Biggs, Kristi Schuette, Steve Bullard, Tian-Xiao Zhang, Margaret Cooper, Martine Dunnwald, Andrew C Lidral, Mary L Marazita, Terri H Beaty, Jeffrey C Murray. GWAS Follow-Up Mutation Screen and Expression Analysis Implicate ARHGAP29 as a Novel Candidate Gene for Nonsyndromic Cleft Lip/Palate. AMERICAN JOURNAL OF MEDICAL GENETICS PART A.2014, 164(4), 876-876.
7. HU Bin;CHEN Qianqian;ZHANG Bo;ZHANG Tianxiao;ZHANG Xiaolong;WANG Hong. MAF and haplotype frequencies of SNPs in MAFB gene in Han Chinese in Beijing. Journal of Modern Stomatology, 2014; 4:193-96.[Chinese]
8. Nan Zhang,Yuehong Xu,Bo Zhang,Tianxiao Zhang,Haojie Yang,Bao Zhang,Zufei Feng,Dexing Zhong. Analysis of Interleukin-8 Gene Variants Reveals Their Relative Importance as Genetic Susceptibility Factors for Chronic Periodontitis in the Han Population. PLoS One, 7;9(8):e104436
9. Buchan JG, Alvarado DM, Haller GE, Cruchaga C, Harms MB, Zhang T, Willing MC, Grange DK, Braverman AC, Miller NH, Morcuende JA, Tang NL, Lam TP, Ng BK, Cheng JC, Dobbs MB, Gurnett CA. Rare variants in FBN1 and FBN2 are associated with severe adolescent idiopathic scoliosis. Hum Mol Genet. 2014; 23(19):5271-82
10. ZHAO Kai-ping , WANG Hong,Ingo Ruczinski, Kung Yee Liang, Jacqueline B Hetmanski, TianXiao Zhang, Shangzhi Huang, Xiaoqian Ye,Kuifeng Yuan, Lingxue Bu, Yah−Huei Wu−Chou, Philip K Chen, Ethylin Wang Jabs, Bing Shi, Richard Redett, Alan F Scott, Terri H Beaty. MAFs and haplotype frequencies for SNPs in ROR2 gene among parents of Han Chinese NSCL/P patients in China. Chinese Journal of Public Health.2014; 3:335-39.[Chinese]
11. Wang Hong, Zhang Tian-Xiao, Chen Qianqian. The Research Progress of the Pathogenesis of Cleft Lip and Palate. Chinese Journal of Reproductive Health. 2013, 4(3):262-64.
12. H Wang, TX Zhang, D Wu. Genetic analysis of a Chinese family with autosomal dominant congenital cataract. Int Eye Sci. 13(8):1619-1621.

13. Han Wang, Tianxiao Zhang, Di Wu, Jinsong Zhang. A novel beaded filament structural protein 1 (BFSP1) gene mutation associated with autosomal dominant congenital cataract in a Chinese family. Molecular vision. 19:2590.
14. Qianqian Chen, Hong Wang, Jacqueline B. Hetmanski, Tianxiao Zhang, Ingo Ruczinski, Holger Schwender, Kung Yee Liang,M. Daniele Fallin,Richard J. Redett,Gerald V. Raymond,Sheng-Chih Jin,Yah-Huei Wu Chou,Philip Kuo-Ting Chen, Vincent Yeow, Samuel S. Chong,Felicia S.H. Cheah,Ethylin W. Jabs, Alan F. Scott,Terri H. Beaty. BMP4 was associated with NSCL/P in an Asian population. PLoS One,7(4): e35347.
15. Murray T, Ruczinski I, Hetmanski JB, Scott AF, TaubM, Patel P, Zhang TX, Murray JC, Marazita ML, Munger RG, Wilcox AJ, Ye X, Wang H, Wu-Chou YH, Shi B, Chong SS, Yeow V, Lie RT, Beaty TH . Examining Markers in 8q24 to Explain Differences in Evidence for Association With Cleft Lip With/Without Cleft Palate Between Asians and Europeans.Genetic Epidemiology 36 : 392–399 ,2012.
16. Peng Lin,Sarah M. Hartz,Jen-Chyong Wang, Arpana Agrawal, Tian-Xiao Zhang, Nicholas McKenna, Kathleen Bucholz, Andrew I. Brooks, Jay A. Tischfield, Howard J. Edenberg, Victor M. Hesselbrock, John R. Kramer, Samuel Kuperman, Marc A. Schuckit, Alison M.Goate, Laura J. Bierut, John P. Rice, COGA Collaborators, COGEND Collaborators, GENEVA(2011) Copy number variations in 6q14.1 and 5q13.2 are associated with alcohol dependence Alcohol Clin Exp Res. 2012 Sep;36(9):1512-8.
17. Wei Jia, Qiao-Fang Hou, Lei Ao, Tianxiao Zhang, Rui Zhang, Shengbin Li. Association analysis of OPRM1 gene exonlin heroin dependence population. Hainan Medical Journal, 2011,22(2). [Chinese]
18. Hong Wang, Tianxiao Zhang, Tao Wu, Jacqueline B. Hetmanski, Ingo Ruczinski, Holger Schwender, Tanda Murray, M. Daniele Fallin, Richard J. Redett, Gerald V. Raymond, Sheng-Chih Jin, Yah-Huei Wu Chou, Philip Kuo-Ting Chen, Vincent Yeow, Samuel S. Chong, Felicia SH Cheah, Sun Ha Jee, Ethylin W. Jabs, Alan F. Scott, Terri H. Beaty. The FGF&FGFR Gene Family and Risk of Cleft Lip with/without Cleft Palate.2013 Jan;50(1):96-103.
19. Terri H. Beaty,Ingo Ruczinski, Jeffrey C. Murray, Mary L. Marazita, Ronald G. Munger, Jacqueline B. Hetmanski, Tanda Murray, Richard J. Redett, M. Daniele Fallin, Kung Yee Liang, Tao Wu, Poorav J. Patel, Sheng-Chih Jin, Tian Xiao Zhang, Holger Schwender, Yah Huei Wu-Chou,Philip K. Chen, Samuel S. Chong, Felicia Cheah, Vincent Yeow, Xiaoqian Ye, Hong Wang, Shangzhi Huang, Ethylin W. Jabs, Bing Shi, Allen J. Wilcox, Rolv T. Lie, Sun Ha Jee,Kaare

Christensen, Kimberley F. Doheny, Elizabeth W. Pugh, Hua Ling, and Alan F. Scott. Evidence for Gene-Environment Interaction in a Genome Wide Study of Nonsyndromic Cleft Palate. Genet Epidemiol. 35(6):469-78. 2011.

Book

1. Tianxiao Zhang , co-authored with Fanglin Guan and 3 other researchers, Prevention of Type II Diabetes: A Guideline for Health Management in the Era Nutritional Genomics, Xi’an Jiaotong University Press, March, 2015 [Chinese].