HUMAN GENETIC VARIATIONS AND THEIR EFFECT ON COMMON COMPLEX DISEASES

By

Ming Li

A DISSERTATION

Submitted to

Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Epidemiology

2012

ABSTRACT

HUMAN GENETIC VARIATIONS AND THEIR EFFECTS ON COMMON COMPLEX DISEASES

By

Ming Li

BACKGROUND: The genetic etiology of common complex diseases has been extensively studied during the past few years. Though many causal genetic variants have been identified, they account for only a small percentage of the estimated heritability of complex diseases, such as breast cancer and cigarette smoking. It remains an open question about where the unexplained heritability lies and how to find it. The objective of this dissertation research is to examine three possible sources of such unexplained heritability: 1) the association between copy number variants and breast cancer, 2) the association between gene-gene interactions and cigarette smoking, and 3) the association between functional rare variants and a simulated quantitative trait. **METHODS**: To detect copy number variants in breast cancer, we examine a breast cancer dataset from the National Cancer Institute and apply a hidden Markov model. To detect gene-gene interactions that are associated with cigarette smoking, we examine a genome-wide dataset from the Study of Addiction: Genetics and Environment and apply a forward U-test. To detect functional rare variants, we examine a dataset from Genetic Analysis Workshop 17 and apply an aggregating U-test. **RESULTS**: In the breast cancer study, we detect five genomic regions on chromosome 2, 4, 6, 12, and 13. In the cigarette smoking study, we detect two single nucleotide polymorphisms (SNPs) with potential interactions. These two SNPs are located in genes *CHRNA5* and *NTRK2*. In the quantitative trait study, we show that the aggregating U-test has a greater power to detect functional rare variants than a commonly used approach, QuTie. **CONCLUSIONS**: Our findings from the breast cancer study suggest that structural changes of

these genomic regions may contribute to the development of breast cancer. Our findings from the cigarette smoking study indicate that the joint action between genes *CHRNA5* and *NTRK2* may contribute to the development of cigarette smoking behavior. These proposed methods provide useful tools to detect various types of human genetic variations underlying complex diseases.

ACKNOWLEDGEMENTS

I am indebted to many people who have contributed to my doctoral education. I would like to start by acknowledging my Ph.D advisors, Dr. Wenjiang Fu and Dr. Qing Lu, who inspired me with challenging ideas, encouraged me to do independent research, and aided me in every piece of my work. I want to thank Dr. Joseph Gardiner, who gave me helpful suggestions when I had difficulties and doubts. I want to thank Dr. Ellen Velie, who gave her time and critical comments that assisted me to organize my dissertation from an epidemiological perspective. I want to thank Dr. Yuhua Cui, who helped me in achieving a rudimentary understanding of quantitative genetics. I have benefited from all the members of my dissertation committee: Dr. Wenjiang Fu, Dr. Qing Lu, Dr. Joseph Gardiner, Dr. Ellen Velie and Dr. Yuehua Cui. I also want to thank Dr. Bert Gold from the National Cancer Institute for kindly providing the breast cancer datasets, which made part of this dissertation research feasible.

It has been a privilege to study at the Department of Epidemiology, College of Human Medicine, Michigan State University. The department provided me with a very friendly environment for scientific research. Many faculty members in the department have been excellent teachers for me. I would especially like to thank Dr. Claudia Holzman, Dr. David Todem and Dr. Zhehui Luo, who offered the kindest help when I needed.

Finally, I want to thank my wife Kate, Zhongnan Zhang, who has always believed in my abilities and has been supportive through all the trying times. Thank you for maintaining some normal home life for us while I focused on the research. I am sure that I could not have achieved any of these without you by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF KEY ABBREVIATIONS

BRCA	Breast Cancer
CS	Cigarette Smoking
GWAS	Genome Wide Association Study
CNV	Copy Number Variant
SNP	Single Nucleotide Polymorphism
CDCV	Common Disease – Common Variant
CDRV	Common Disease – Rare Variant

CHAPTER 1.

INTRODUCTION AND AIMS

1.1. Introduction

The past decades have witnessed accelerated progression in the field of human genetics, with a large and rapid expansion in the understanding of the inherited genetic etiology of Mendelian diseases, but limited advancement in our understanding of the inherited genetic etiology of *common complex diseases*, such as breast cancer and cigarette smoking. The causal genetic variants identified so far confer relatively small increments in risk, and explain only a small percentage of the heritability of breast cancer (12.5%) and cigarette smoking (~10%) [1,2]. Many possible explanations of this issue of "missing" heritability have been suggested, such as various types of human genetic variations, their complicated interactions and the limited power of current statistical methods. The objective of this dissertation research is to develop statistical methods applicable to the study of the inherited genetic etiology of common complex diseases. Specifically, I will examine three possible sources of the unexplained heritability, copy number variants, gene-gene interactions, and functional rare variants [3], aiming to understand the genetic etiology of breast cancer and cigarette smoking.

In this dissertation, novel statistical methods will be proposed and applied for investigating the genetic etiology of two common complex diseases, breast cancer and cigarette smoking. Breast cancer is the most common malignancy in women. In the United States, it accounts for about 2% of deaths from all causes in the general population [4]. It was estimated that in 2010, 207,090 women were diagnosed and 39,840 women died from breast cancer [5]. Smoking is also a well known risk factor for many complex human diseases, such as cardiovascular diseases and lung cancer. It was estimated that smoking caused approximately 435,000 deaths annually in the

United States, which was 18.1% of all deaths [6]. Understanding the genetic etiology of breast cancer and cigarette smoking will have a profound impact on reducing the burden of diseases in the population. The focus of this research is to study the effect of inherited genetics variations. Somatic mutations that only occur in somatic cells after conception are not considered.

1.2. Genetic Basis of Common Complex Diseases

Until now, Mendelian disorders are the most well understood human genetic disorders in regard to their causes and mechanisms. In such cases, the disorders are caused by single gene defects through either dominant or recessive patterns [7]. Though Mendelian diseases may vary in severity, their risks usually can be predicted accurately by the genotypes. For example, two deleterious mutations in β-hemoglobin gene can predict sickle cell anemia accurately [8]. So far, many genes underlying Mendelian diseases have been mapped, and successfully cloned [9]. These successes have led to a significant improvement for early diagnosis and treatment of Mendelian disorders. However, Mendelian disorders are typically rare in the population. Their total incidence was estimated to be less than 5% [10]. Most diseases have multi-factorial etiologies and are referred to as complex diseases. The development of complex diseases is usually associated with the joint effect of multiple genes and the environmental factors. Significant public health concerns are now focused on the complex disorders [11], such as cardiovascular diseases, diabetes and cancers. These complex diseases are already common in developed countries and are becoming prevalent in developing countries [12]. Based on the disease prevalence in the population, the complex diseases can be differentiated as either ‘common disease’ or ‘rare disease’. There is no universal cut-off value between the prevalence of common and rare diseases. In the United States, a common disease usually is referred to as a disease that develops in more than 200,000 persons or about 1 in 1,500 people [13].

Understanding the genetic etiology of common complex diseases is crucial for reducing the morbidity as well as mortality in the population.

The genetic contribution of common complex diseases is supported by the clustering of cases within families. For example, parental history of cardiovascular diseases (CVD) is a well accepted predictor for CVD risk in offspring. Evidence from the Framingham Heart Study showed that the risk of CVD before age of 55 was significantly higher among those with parental history than among those without parental history. The estimated age-adjusted odds ratios were 2.6 in men and 2.3 in women [14]. Though shared environmental factors might also contribute to the excessive familial risk, the corresponding odds ratios were 2.0 in men and 1.7 in women after adjusting for other known environmental risk factors, indicating that the excessive familial risk was largely due to genetics factors [14].

The inherited genetic contribution to complex diseases or traits can be measured by their heritability, which is defined as the proportion of the trait variation in a population that can be attributed to genetic variability [15]. Two types of heritability are commonly used in the literature: the broad sense or the narrow sense. The broad sense of heritability (H^2) is estimated by partitioning the variance of the trait into a genetic variance component and an environmental variance component. In his seminal work in quantitative genetics, R.A Fisher proposed to further partition the genetic variance into additive, dominant and epistatic components [16]. The additive and dominant components measure the genetic variation that can be explained by a generalized linear model through either an additive or dominant pattern, while the epistatic component measures the genetic variation that deviates from a generalized linear model. The narrow sense of heritability (h^2) can be estimated by the ratio between the additive component of genetic variation and the total variation of trait. Both the broad and narrow sense of heritability is widely

used to describe the relative contribution of genetic or environmental factors to the trait in a particular population. To avoid any confusion in this dissertation, heritability refers to the broad sense definition unless specified otherwise.

1.3. Human Genetic Variation

Human genetic variation may occur on many different scales, ranging from single nucleotide polymorphisms (SNPs) to large structural alterations (e.g. chromosome duplication or deletion) that can affect thousands or millions of nucleotides [17]. Each genetic variant may have multiple forms, referred to as alleles. SNP is the most common type of sequence variation and occurs when a single nucleotide varies at a particular site between individual genomes. On the average, two individual genomes differ from one another by approximately 0.1% of DNA nucleotide sites [18]. The majority of SNPs are bi-allelic, and can form three possible genotypes by the combination of two nucleotides [19]. The two nucleotides are differentiated by their frequencies in the population as minor allele or major allele. In the past, only those polymorphisms with minor allele frequencies (MAFs) greater than 1% were defined as SNPs. With the advent of genomic era, this frequency requirement is no longer necessary. Instead, the SNPs with MAFs greater or less than 1% are now referred to as common or rare variants. The total number of SNPs in the human genome is about 10 million, which constitutes about 90% of the genetic variation in human [20,21,22]. Because of the genotyping convenience and the dense coverage of human genome, SNPs are predominantly used as genetic markers to study the genetic etiology of complex human diseases.

A human genome consists of twenty-three pairs of chromosomes, including twenty-two pairs of autosomal chromosomes and one pair of sex chromosomes. Therefore, each human ordinarily has two copies of each autosomal region. During the past few years, solid evidence has shown

that structural alterations, due to insertions, deletions and inversions of the DNA, also contribute considerably to the variability of the human genomes [23,24,25]. These structural changes may cause copy number differences in particular genomic regions, ranging from one kilobase to a complete chromosome arm. A copy number variant (CNV) is defined as a genomic region where the DNA copy number differs between two or more individuals. CNVs are far more complex than SNPs. The human genomes differ by only 0.1% with respect to SNPs, but by 1.2% with respect to CNVs. Aside from that, about one quarter of the CNVs occur without any SNPs in the region [26]. CNVs also pervasively exist in the population. Even for monozygotic twins who are presumed to be genetically identical in terms of sequence variations, their genomes can differ by CNVs [27]. One recent study among monozygotic twins estimated that ~10% of the CNVs were not observed in any of the parents, which were referred to as *De Novo* CNVs. Further, 35% of these *De Novo* CNVs differed between the monozygotic twins [28]. Therefore, copy number variation can be viewed as an important form of human genetic variation.

Human genetic variations also exist in other forms besides SNPs and CNVs. One example is the microsatellite, also known as short tandem repeat (STR) or simple sequence repeat (SSR). Microsatellites are usually formed by repeating sequences of 1-6 base pairs of DNA at particular genomic positions, and may vary among individuals by the number of repeats [29]. Compared to SNPs and CNVs, microsatellites are less stable across populations or generations, and they also have a low coverage of the human genome [30]. Because of these reasons, microsatellites have been less frequently used as genetic markers than SNPs or CNVs.

Penetrance measures the individual effect of each genetic variant on a disease, which is defined as the probability to develop a disease for an individual carrying a particular genotype. Penetrance is often expressed as an age-related cumulative frequency. The penetrance may vary

greatly across diseases or genetic variants. For example, the penetrance of breast cancer by age 70 was estimated to be 65% and 45% for *BRCA1* and *BRCA2* mutations, respectively [31], and it was generally lower than 10% for most of the other disease-susceptibility variants. The highest value of penetrance is 100%, which is also called complete penetrance. One example is the familial hyper-cholesterolemia (FH). FH is caused by deleterious mutation in the *LDL receptor* (*LDLR*) gene. The mutation has a dominant effect, and carrying a single copy of the mutant gene will lead to a 2-fold increase in LDL production in blood. Nearly 100% individuals with the mutation in gene *LDLR* will develop FH [32]. According to their disease penetrance level, the disease-susceptibility variants may fall into three categories: high-penetrance variants, moderate-penetrance variants and low-penetrance variants [33]. However, the classification is usually empirically determined and depends on the disease of interest.

1.4. Models of the Genetic Origin of Common Complex Diseases

There has been a long debate regarding how genetic variants contribute to the development of common complex diseases. It now seems clear that the genetic variants may influence the susceptibility of common complex diseases in at least two ways: Common Disease-Common Variant (CDCV) model and Common Disease – Rare Variant (CDRV) model [34]. The CDCV hypothesis asserts that complex diseases are caused by multiple genetic variants with appreciable frequencies in the population at large, but each confers a small or moderate effect [35]. The CDRV hypothesis, on the other hand, argues that the complex diseases are mostly caused by multiple genetic variants with low frequencies in the population, but each confers a relatively large effect [36]. Each hypothesis has been perused by a substantial number of researchers and is supported with a large amount of evidence.

The CDCV hypothesis is predominant in the literature and has provided a theoretical basis for the extensive genome-wide association studies (GWASs) [37]. The GWASs are typically population based. Each subject in a GWAS is genotyped with a large number of SNPs (e.g. over 100K or 500K) that are dense enough to cover the whole genome. The rationale of GWASs is that most of the human genome falls into highly correlated segments, within which genetic variants are in strong linkage disequilibrium (LD) with one other [38]. In order to be successful in a genetic association study, the SNPs being tested can either be the causal SNPs or those SNPs in strong LD with the causal SNPs. Consequently, a number of representative SNPs can be selected in each genomic region as genetic markers for the association test. These representative SNPs are referred to as tagSNPs, which are selected according to the high-density maps of the human genome. To date, hundreds of GWASs have been conducted and have detected a large number of genetic variants that are associated with over 40 complex diseases [39,40,41,42,43]. Many findings have also been replicated in diverse populations [44]. Though most of the associated SNPs are found to be located in non-coding regions and are not directly involved in protein productions, they may have important regulatory functions that control gene behaviors, such as gene expression levels [45]. These findings provide compelling evidence that many complex diseases are caused by the collective effect of multiple common genetic variants.

The CDCV hypothesis is also challenged by a number of investigators. They argue that the rare variants may have weak correlation with the higher-frequency tagSNPs used in GWASs. Therefore, the indirect association mapping via tagSNPs may have a low power to detect the causal rare variants [46,47]. Many investigators have surveyed rare sequence variations and have detected multiple rare variants that are involved in the etiology of complex diseases or complex traits [48,49,50,51]. These rare variants are more likely to cause a disease individually rather

than jointly, which is different from the mechanism of common causal variants [34]. These facts suggest that the genetic etiology of complex diseases is highly heterogeneous [52]. New strategies are in great need to detect the rare variants underlying complex diseases.

1.5. Methods of Studying the Genetic Etiology of Complex Diseases

Understanding the genetic architecture of common complex diseases includes three major aspects: detecting the genetic variants involved in the disease, uncovering their distributions in the population and estimating the magnitude of their effect [53]**.** Understanding the genetic architecture of common complex diseases is crucial to identify the high risk population, to facilitate disease prevention, and also to promote personalized medicine. However, connecting genotypes and disease phenotypes is no simple task. It was not until the 1980's that a general method, linkage analysis, was first proposed [54]. In the 1990's, the emerging understanding of molecular biology and the development of the Human Genome Project offered insight into possible candidate genes that might be functionally related to genetic disorders [55,56,57] . As a result, genetic association studies using candidate genes became popular, providing powerful alternatives for genetic linkage analysis. Genetic association studies compare the frequency of specific genetic variants between cases and controls, and do not require samples from family pedigrees as linkage analysis does. Therefore, they are more suitable for detecting genetic risk factors that commonly present in the population. During the past few years, genetic association studies have searched for genetic risk factors across the entire human genome, which is made possible by the comprehensive high-density maps of the human genome and the advancement of genotyping technologies [37,58,59,60,61]. These GWASs advance the field of human genetics dramatically by the identification of many novel genetic risk factors.

At present, GWASs are commonly adopted for revealing the genetic etiology of complex diseases. Though a substantial number of disease-susceptibility variants have been identified, the genetic etiology of complex diseases remains elusive. The identified genetic variants only account for a small percentage of the estimated heritability of complex diseases, such as type 2 diabetes (6%) or Crohn's diseases (20%) [44]. It is also unclear how many genetic variants in the human genome are associated with diseases, and how the genetic variants interact with one another to cause diseases. It is no surprise that additional genetic variants with lower effect sizes may exist and can be discovered by increasing the sample sizes of GWASs [62,63]. However, searching through larger GWASs does not seem likely to uncover all the remaining genetic risk factors [64]. The challenge arises as to where the unexplained heritability lies and how to find it. Though this important issue is still under debate, researchers have suggested that it is partly, if not mostly, due to the following reasons: 1) copy number variations have not been well understood; 2) gene-gene interactions pervasively exist in biological pathways; 3) rare variants have been scarcely addressed in genetic association studies.

1.6. Rationale of this Research

Extensive studies have been conducted to investigate the genetic etiology of breast cancer and cigarette smoking. However, the heritability of these traits remains largely unexplained. Many possible explanations have been suggested for this issue of "missing" heritability. First, besides sequence variations, large structural alterations, such as copy number variants, may also contribute to disease development. Relatively few studies have investigated functional CNVs, especially for CNVs with a small-to-intermediate size. Second, the effect of one genetic variant may be suppressed or enhanced by the other variants through complex interactions, which is also termed epistasis. Therefore, the association test may have a low power if the loci are examined

separately without considering potential interactions. Third, rare variants may also play a major role in the development of complex human diseases. Analysis of rare variants holds great promise to detect novel disease-susceptibility loci. However, challenge still remains for statistical modeling because of the low allele frequencies. Sophisticated statistical tools are in great needs to address these limitations.

1.7. Specific Aims

The objective of this dissertation research is to develop novel statistical methods for the identification of various types of human genetic variations associated with complex diseases, including breast cancer (BRCA) and cigarette smoking (CS). The methods will address three possible sources of the "missing" heritability. The specific aims are:

AIM1. Detecting Copy Number Variants that are Associated with Breast Cancer

I will apply a copy number estimation method, referred to as the Probe Intensity Composite Representation (PICR) [65], to detect copy number variants that are associated with breast cancer. The newly established PICR model will be extended with a hidden Markov model for CNV identification. Data from a recent GWAS of breast cancer will be used for analyses [66]. The original study was a three-phase case-control study with subjects from a genetically isolated population, Ashkenazi Jews. I hypothesize that the proposed method can detect small to intermediate size CNVs that are associated with breast cancer.

AIM2. Detecting Gene-gene Interactions that are Associated with Cigarette Smoking

I will apply a forward U-test to detect gene-gene interactions associated with cigarette smoking. The method will use U-Statistics to measure the variation of quantitative traits. Data from the Study of Addiction: Genetics and Environment (SAGE) will be used for analyses. The cigarette smoking trait will be defined as the number of cigarettes smoked per day: 0 (10

cigarettes or less), 1 (11-20 cigarettes), 2 (21-30 cigarettes) and 3 (31 cigarettes or more). I hypothesize that the proposed method can detect gene-gene interactions among known smoking-associated loci.

AIM3. Detecting Functional Rare Variants that are Associated with Quantitative Traits

I will apply an aggregating U-test to detect rare variants associated with quantitative traits. Data from Genetic Analysis Workshop (GAW) 17 will be used for analyses. The method will be an extension of the forward U-test described in aim 2 with the consideration of both common and rare variants. I hypothesize that the proposed method can have a higher power to detect the association with rare variants than a commonly used approach, QuTie.

1.8. Organization of the Dissertation

The dissertation is organized as follows: In Chapter 2, I summarize the descriptive epidemiology of two common complex diseases, breast cancer and cigarette smoking, and review the inherited genetic risk factors associated with each outcome. In Chapter 3, I propose a hidden Markov model for detecting copy number variants, and then illustrate the proposed method by a study of breast cancer. In Chapter 4, I propose a forward U-test for detecting gene-gene interactions, and then illustrate the proposed method by a study of cigarette smoking. In Chapter 5, I propose an aggregating U-test for detecting functional rare variants, and then illustrate the proposed method by a study of quantitative traits. In Chapter 6, I summarize the findings in these studies and discuss challenges and directions for future development.

CHAPTER 2.

GENETIC RISK FACTORS FOR BREAST CANCER AND CIGARETTE SMOKING: A REVIEW OF THE LITERATURE

2.1. Descriptive Epidemiology of Breast Cancer

Breast cancer usually develops in women, but it can also be found in men. The risk of breast cancer is about 100 times greater among women than men [67]. The incidence of breast cancer varies greatly around the world. It is generally higher in developed countries than less-developed countries. The US has the highest breast cancer incidence around the world. Based on the number of cases diagnosed in 2003-2007 from 17 Surveillance Epidemiology and End Result (SEER) centers, the age-adjusted incidence rate was 122.9 per 100,000 women per year. In the US, the life time risk of developing breast cancer among women is about 1 in 8 (12.15%). It is the most common malignancy in women, and accounts for about 2% of deaths from all causes in the general population. It was estimated that 207,090 women were diagnosed and 39,840 women died from breast cancer in 2010 [5].

According to the National Cancer Institute, breast cancer incidence is highest among white, non-Hispanic women and lowest among Korean American women. The African American women have a slightly lower incidence rate of breast cancer than the white women. The death rate of breast cancer also varies across racial groups, and is highest among the African American women and lowest among the Chinese American women [5]. The etiology of breast cancer is largely unknown. Well established risk factors for breast cancer include age, family history, hormone replacement therapy, radiation, alcohol consumption, and obesity [68,69]. Studies have suggested that breast cancer risk can be reduced by enhancing physical activities and maintaining a healthy weight [70]. Currently, the standard and most commonly used method for breast cancer

risk prediction is the Breast Cancer Risk Assessment Tool (BCRAT), also known as *Gail* model. The *Gail* model predicts the breast cancer risk for a woman by a number of risk factors that she is exposed to, including family history of breast cancer, current age, age at menarche, age at first birth of a child, and race/ethnicity. Other clinical measures, such as medical history of ductal carcinoma in situ and breast biopsy, can also be incorporated if available [71,72,73,74].

2.2. Genetic Risk Factors of Breast Cancer

Breast cancer is caused by DNA damage due to either germline mutations or somatic mutations in the process of aging [48,75]. Overall, breast cancer is twice as common among women with an affected first-degree relative [76]. Although shared environmental factors may also contribute to the elevated risk, twin studies have indicated that the excessive familial risk is mainly due to genetic factors [77]. During the past few years, many causal variants have been successfully identified through linkage mapping and genetic association studies [78]. These identified breast cancer susceptibility alleles appear to fall into three categories according to their risk levels and prevalence in the population, including rare high-penetrance alleles, rare moderate-penetrance alleles and common low-penetrance alleles [79].

The high-penetrance alleles may increase the risk of developing breast cancer by over ten folds. In the 1990s, two major predisposition genes, *BRCA1* and *BRCA2*, were identified through linkage mapping [80,81,82]. It was estimated that the average cumulative risk by age 70 was 65% for *BRCA1*-mutation carriers and 45% for *BRCA2*-mutation carriers [31]. On the other hand, because of their low allele frequencies in the general population, these two genes can account for at most 5% of breast cancer cases [83]. In addition, the high-penetrance alleles also include germline mutations in a few other tumor suppressors, including *TP53*, *PTEN*, *STK11* and *CDH1*. However, the mutations are very rare and account for a much smaller fraction of the breast

cancer cases [84]. Overall, these inherited high-penetrance gene mutations account for less than 7% of all breast cancer cases [85]. At present, it is commonly accepted that no other high-penetrance genes may exist to account for a large proportion of the breast cancer cases [86].

The moderate-penetrance alleles may increase the risk of developing breast cancer by 2-4 folds [78,79]. The susceptibility genes are usually in the same biological pathways with *BRCA1* and *BRCA2*, and are identified through direct interrogation for disease-causing mutations in the genes. To date, at least four genes are identified by this strategy, including *ATM*, *BRIP1*, *CHEK2* and *PALB2*. Compared to *BRCA1* and *BRCA2*, these genes confer less elevated risks of developing breast cancer. The moderate-penetrance allele carriers have approximately 6-10% risk of developing breast cancer by age 60 [79]. Similar to the high-penetrance alleles, the moderate-penetrance alleles also have low frequencies in the population and each makes a relatively small contribution to the breast cancer incidence.

It is hypothesized that a large proportion of the breast cancer cases are due to the common alleles that confer very small increases of the risk. These low-penetrance alleles may commonly present in the general population with an increased risk of less than two folds. They are usually identified by genetic association studies, either on the basis of candidate genes or through genome-wide search. During the past few years, many genome-wide association studies (GWASs) have been conducted to identify common risk loci for breast cancer. For example, gene fibroblast growth factor receptor 2 (*FGFR2*) was first identified to be associated with invasive breast cancer in women less than 60 years old with European ancestry [40]. *FGFR2* is located in the chromosome region of *10q26.23*. It was estimated that the mutation in *FGFR2* conferred an increased breast cancer risk of 1.26 [1.23-1.30]. This association has been replicated in a series of studies among different populations [87,88,89,90]. The estimated relative

risk ranged from 1.17 to 1.43. Several other genes or genetic variants were also identified to be associated with BRCA susceptibility, such as *TNRC9*, *MAP3K1*, and *LSP1* [78,79]. For most of the genetic variants identified so far, the biological mechanisms still remain unknown.

Despite all the progress in the past two decades, the current findings can only explain a small fraction of the breast cancer cases, highlighting the need of searching for additional genetic variants for BRCA susceptibility. Recently, copy number variants have been recognized as a novel form of genetic variation that can contribute considerably to disease development. One study showed that the copy number change in gene *MTUS1* was associated with breast cancer [91]. *MTUS1* is located on chromosome 8p, a region frequently undergoing deletion. This deletion variant was found to be associated with a decreased risk of breast cancer with an odds ratio of 0.58 [91]. Other evidence also suggests that CNVs in other genes or chromosome regions, such as *PIK3CA*, *16p12.1* and *16q22.1*, may play an important role for breast cancer development [92,93,94].

2.3. Descriptive Epidemiology of Cigarette Smoking

Smoking was originally used by many civilizations for burnt incense during religious rituals, and was later adopted for pleasure or as a social tool [95]. As early as the 1950s, the British Doctors Study provided solid epidemiological evidence of the association between smoking and lung cancer [96]. Since then, the association has been consistently replicated and the causal relationship between smoking and lung cancer has been well established. Smoking is also a well known risk factor for many other complex human diseases, such as cardiovascular diseases. It was estimated that smoking caused approximately 435,000 deaths annually in the United States, which was 18.1% of all deaths [97]. Smoking also imposes a great burden to the economy, and is responsible for about 7% of the total US healthcare costs, or an estimated 157.7 billion dollars

each year [97]. Despite the public awareness of its hazard to human health, the prevalence of smoking in the United States has been barely reduced. In 2002, the estimated number of current smokers in the US was 45.8 million [98]. Men are five times more likely to smoke than women. However, the gender difference is diminishing, due to the decline of male smokers and increase of female smokers. Most smokers are addicted to cigarette smoking due to their dependence on nicotine. Relatively few smokers can achieve sustained abstinence without medicine or other help. It was estimated that the success rate for unaided smoking cessation was about 7% after an average of 10 months of follow-up [99]. On the other hand, cigarette smoking is one of the most preventable causes of deaths [100]. Understanding the etiology of cigarette smoking can have a profound impact on the prevention of many complex diseases.

2.4. Genetic Risk Factors of Cigarette Smoking

The quantification of smoking is a major unsolved issue in tobacco-related research. Most of the available measurements are defined according to various aspects of cigarette smoking, such as age of smoking initiation and number of cigarettes smoked per day [101]. Smoking is a complex behavior involving both genetic and environmental factors and their interactions. Peer and family influences are the strongest environmental factors for the time of smoking initiation. Genetic factors also play an important role in determining smoking initiation and dependence. The early evidence comes from twin studies. It was estimated that the average heritability of nicotine dependence was 56% [102,103]. In the 1990s, the linkage analysis for cigarette smoking identified a genomic region on chromosome 5q, very close to the D1 dopamine receptor gene [104]. After that, a number of other regions on chromosome 3, 5, 17 and 18 were also identified independently by multiple studies using linkage analysis of smoking behavior [105,106,107,108]. Subsequent association studies also identified a number of genes that were associated with

cigarette smoking, such as *CHRNA5, GABAB2*, *DDC*, *BDNF*, and *COMT*. For instance, studies showed that mutation in gene *CHRNA5* was associated with a two-fold risk of developing nicotine dependence [109]. Meanwhile, the biological mechanisms for these identified genes remain largely unknown. The etiology of cigarette smoking may involve many genetic variants through complex biological pathways. It has been suggested that many genes are functionally related to cigarette smoking through nicotine metabolism and dopaminergic reward system [110]. Detecting the complex interactions among genes and environmental factors is crucial to understand the biological pathways for disease development.

2.5. Limitation of the Existing Investigations

Extensive studies have been conducted to investigate the genetic etiology of breast cancer and cigarette smoking. However, the genetic heritability of these diseases remains largely unexplained. Many possible explanations have been suggested. First, besides the genotypic variations, large structural alterations, such as copy number variants, may also influence the disease development. Relatively few studies have focused on detecting functional CNVs, especially at a small-to-intermediate size. Second, the effect of one genetic variant may be suppressed or enhanced by the other variants through complex interactions, which is also termed epistasis [111]. Therefore, the association test may have a low power if the loci are examined separately without considering potential interactions. Third, rare variants may also play a major role in the development of complex human diseases. Analysis of rare variants holds great promise to detect novel disease susceptibility loci. Meanwhile, challenge still remains for statistical modeling due to the low allele frequencies. In this dissertation research, I am going to propose novel statistical methods addressing these limitations.

2.6. Significance of the Research

BRCA GWAS (AIM 1)

The proposed study will be among the very first ones to study small-to-intermediate size CNVs in BRCA. Although a number of BRCA GWASs have reported significant findings of SNP genotypes and CNVs, the reported CNVs are of large size (≥ 50 kb or even several mega bases). To our knowledge, no small-to-intermediate size CNVs has been reported yet. Compared to large size CNVs, small-to-intermediate size CNVs can lead to more precise genomic regions that are functionally related to a disease. In this study, the identification of small-to-intermediate size CNVs will be achieved with the novel idea of estimating allelic copy numbers at each single SNP locus by PICR method, and further extending to multiple SNPs by applying a Hidden Markov Model (HMM).

Gene-gene Interactions Associated with Cigarette Smoking (AIM 2)

The available data suggests that gene-gene interactions are likely to be a major source of the unexplained heritability of complex diseases. Intuitively, the interactions among genes can be examined by exploring all possible combinations of the genetic variants [112,113]. However, an exhaustive search is often not feasible because of the rapidly increasing computational time. Moreover, when the number of genetic variants is large, an irrelevant combination may outperform the real disease model simply due to sample randomness. Exhaustive search may increase the likelihood of finding such irrelevant combinations. To address these limitations, we propose a novel method that searches for potential gene-gene interactions sequentially, which is computationally efficient and is applicable to high-dimensional data. In addition, it is also a non-parametric method without any assumption of the trait distribution.

Aggregation of Multiple Rare Variants (AIM 3)

Most of the available statistical methods are proposed for common variants analysis. For rare variants, the number of subjects carrying the rare alleles is usually small. Therefore, the available methods usually have a low power to detect the association between rare variants and traits. Right now, the next generation sequencing technology has become popular, which yields a large amount of genetic data, including both common and rare variants. At present, the most commonly used approach for detecting phenotypic associations with rare variants is to group multiple rare variants into a single 'super' variant and then combine it with other common variants for a multivariate analysis, [46,114,115]. Different from existing methods, our method adaptively collapses a subset of potential disease-susceptibility rare variants. I expect this method to have a greater power than the existing methods.

CHAPTER 3.

COPY NUMBER VARIANTS AND BREAST CANCER

3.1. Introduction

During the past few years, genome-wide association studies (GWASs) have been commonly adopted for detecting genetic variants underlying common complex diseases, such as breast cancer and Type II diabetes [40,116]. Though the findings from GWASs have provided valuable insight into the genetic etiology of complex diseases, they account for a small percentage of the heritability [117]. The GWASs typically test the genotype frequency of each SNP between a group of cases and controls. However, it was estimated that two individual genomes were on the average 99.9% identical with respect to DNA sequence variations [118,119]. Solid evidence suggests that sequence variations may not be the only source for the heritability of diseases [3,117]. Alternatively, structural alterations, such as copy number variants (CNVs), can occur without any sequence variations. It was estimated that the structural alterations accounted for up to 7.3% of the genetic variability among human genomes [120]. These CNVs may contribute considerably to the development of many complex diseases, such as cancers [91,121]. However, until now, the association between CNVs and disease development remains largely unexplored.

Copy number variants were first identified in the early 2000s [122,123], and were found to exist pervasively in human genomes [124,125]. In the past decade, the rapid advancement of biotechnology allowed us to characterize human genomes with copy number variations. At present, two platforms, including Affymetrix high density SNP arrays and Illumina Bead arrays, are commonly adopted for copy number inference [126,127]. Both platforms provide data in the form of experimentally determined intensities as surrogates for DNA quantities in the biological

samples. Therefore, sophisticated statistical models are in great need to infer the underlying copy number levels accurately. Smoothing methods are used among the very first studies for copy number inference [128,129]. These methods usually assume that the underlying copy numbers may have three levels: normal, copy number gain and copy number loss. After fitting a smoothing curve along the genomic regions, certain threshold is used to infer copy number levels. These methods have been applied in many studies for detecting copy number changes. However, as discussed by Lai *et al.*, the smoothing methods have two major limitations: 1) it is difficult for the smoothing methods to locate the boundaries of copy number changes; 2) it is difficult for the smoothing methods to test the significance of copy number changes [130]. Another group of methods adopt certain change-point models to infer the underlying copy number levels [131,132]. These change-point models usually assume that the SNPs are uniformly distributed in human genomes, and the underlying copy number levels are piecewise constants with a series of jumps. By maximizing the likelihood function, the parameters and the change-points can be estimated for copy number inference. Such models were further extended by various formations of hidden Markov models (HMMs) [133,134,135,136]. The HMMs usually assume the observed intensities of SNPs are emitted by an underlying Markov chain, and they explicitly specify the distribution of the waiting time of copy number changes and the jumping probabilities between copy number states. These methods have emerged as promising tools for copy number inference.

However, the available methods are commonly proposed to handle the intensity values of SNPs, which are subjected to large experimental noise. As a result, the quality control issue has raised considerable concerns regarding the result interpretation and decision making [137]. In a recent study, Wan *et al.* proposed a novel approach to estimate copy number abundance on a single SNP- single array basis, referred to as the Probe Intensity Composite Representation

(PICR) [65]. This method models the cross-hybridization between DNA sequences via their physical binding affinities. It has shown great potential for differentiating copy number signal from background noise. In this chapter, I propose to extend PICR method with a hidden Markov model for copy number inference. The copy number abundance is first estimated at each SNP locus by PICR, and then standardized to achieve equal scaling between multiple samples. A hidden Markov model is further applied for copy number inference. Compared to the available HMM-based methods, our method has two major advantages: 1) by estimating the copy number abundance using PICR, a large proportion of the noise is removed from the intensity values to improve the performance of HMM; 2) through a novel standardization of the copy number abundance, our method does not require between array normalizations for multiple samples, which ensures the data integrity. This proposed method is suitable for detecting copy number variants with Affymetrix high density SNP arrays.

3.2. Methods

In this section, I first describe the design of Affymetrix SNP array, and then explain the proposed method step by step. Suppose we have a study population of N subjects, each genotyped with a large number of K SNPs. Our method first estimates the copy number abundance at a single SNP locus for each subject by using the newly established PICR model [65]. It then standardizes the copy number abundance to achieve equal scaling between subjects. Finally, a hidden Markov model is applied to integrate multiple SNPs for copy number inference.

3.2.1. Design of Affymetrix 500K SNP Array

In an experiment with an oligonucleotide microarray, the array is attached with millions of short immobilized nucleic acid sequences, known as probes. These probes are designed complementary to the DNA sequences in biological samples, referred to as targets. These targets

are labeled with fluorescent dyes and their abundance can be quantified by the fluorescent intensities yielded through their hybridization with the probes [138,139,140]. Affymetrix SNP array uses multiple probe-sets to capture the properties of each SNP. In a 500K SNP array, six quartets are adopted to interrogate a single dimorphic SNP site with its possible alleles commonly denoted as *A* and *B*. Each quartet consists of 4 types of probes that are 25 base pairs in length. These probes are designed either perfectly matched to the targets or mismatched at a particular nucleotide site for each allele: perfect match *A* (*PA*), mismatch *A* (*MA*), perfect match *B* (PB) and mismatch *B* (MB). The probe-sets are also designed to hybridize with either sense strands ($s=1$) or antisense strands ($s=-1$). The quartets have different shifts (k) for the nucleotide on the probes (k may take the values -4, -3, -2, -1, 0, 1, 2, 3, 4) from the center nucleotide of the probes ($k = 0$ at position 13 of the 25 base pairs) [141].

3.2.2. Estimation for copy number abundance by PICR

The PICR method takes into account the cross-hybridization of the DNA sequences via a positional-dependent nearest neighbor (PDNN) model [65]. In PICR, the florescent intensity of a particular probe-set is decomposed into multiple components: the baseline intensity (b), the products of allelic copy numbers abundance (R_m) and the binding affinity between the target and the probe with respect to different alleles (R_m), and a measurement error (ε) (Equation (1)). The binding affinities (f_A, f_B) are determined by the physical property of the DNA sequences. Based on the PICR model, the allelic copy number abundance can be estimated via a linear regression between the intensities and binding affinities.

$$
\begin{aligned}
&\vdots \\
I_{PA} &= b + N_A f_A^{PA} + N_B f_B^{PA} + \varepsilon_{PA} \\
I_{PB} &= b + N_A f_A^{PB} + N_B f_B^{PB} + \varepsilon_{PB} \\
I_{MA} &= b + N_A f_A^{MA} + N_B f_B^{MA} + \varepsilon_{MA} \qquad \textit{Equation (1)} \\
I_{MB} &= b + N_A f_A^{MB} + N_B f_B^{MB} + \varepsilon_{MB} \\
&\vdots
\end{aligned}
$$

3.2.3. Multi-array Equal Scaling by Standardization

The allelic copy number abundance is estimated by PICR on a single array- single SNP basis. All the fluorescence intensities are subject to experimental scales which may vary among arrays. It is thus essential to achieve equal scaling for multiple arrays before any further analyses. We propose a novel standardization approach as:

$$
SCN_{i,j} = \frac{N_{i,j,A} + N_{i,j,B}}{se(N_{i,j,A} + N_{i,j,B})}; i = 1, 2, ..., N; j = 1, 2, ..., K \qquad \textit{Equation (2)}
$$

where $N_{i,j,A}$ ($N_{i,j,B}$) denotes the allelic copy number abundance for SNP *j* of subject *i*; $se(N_{i,j,A} + N_{i,j,B})$ denotes its estimated standard error of $N_{i,j,A} + N_{i,j,B}$ via the linear regression model of Equation (1). Assuming the random errors in Equation (1) are normally distributed, the standardized copy numbers follow *t* distributions identically for $\forall i = 1, ..., N; \forall j = 1, ..., K$, and hence, are expected to be on the same scale.

3.2.4. Hidden Markov Modeling for Multi-SNP Copy Number Inference

Modeling Strategy and Copy Number States

As illustrated by Equation (2), our objective is to detect total copy number changes among subjects. Similar to the existing HMM-based methods [133,135,136], we assume a particular SNP locus may have 5 possible copy numbers states, with its total copy number ranging from 0 to 4, (Table 3.1). Such copy number states are not observed directly, and hence, are hidden. The inference of these hidden states is based on two types of observations, $\log R$

ratios (LRR) and B allele frequencies (BAF), which can be calculated by the estimated allelic copy number abundance. We first estimate the standardized copy number abundance for the j^{th} SNP of subject i. Because the standardized copy number abundance is a relative measure with unknown reference, we further define its LRR as:

$$R_{i,j} = LRR_{i,j} = \log_2(\frac{SCN_{i,j}}{SCN_{j,reference}});$$

$$\text{where } SCN_{j,reference} = \underset{i \in \{Control\}}{median}(SCN_{i,j})$$

The estimated SCNs among controls are expected to represent normal levels of copy numbers and can be used to determine the reference level for each particular SNP locus. We further define the BAF as:

$$B_{i,j} = BAF_{i,j} = \begin{cases} 0 & \theta_{i,j} \le a_j \\ (\theta_{i,j} - a_j)/(b_j - a_j) & a_j \le \theta_{i,j} < b_j \\ 1 & \theta_{i,j} \ge b_j \end{cases};$$

where $\theta_{i,j} = \arctan\frac{scn_{i,j,B}/scn_{i,j,A}}{\pi/2}$ and a_j, b_j are the corresponding thresholds for accurate genotyping of SNP *j* with PICR. The B allele frequencies provide a normalized measure of relative signal ratio between allele B and allele A. Similar to a few previous studies, a HMM is adopted to integrate LRR and BAF for copy number inference [133,135,136]. One novelty of our method is that we use standardized copy number abundance rather than probe intensities to calculate corresponding LRR and BAF.

Transition Probability of copy number states

Our proposed model is a time-dependent continuous Markov Chain, using genomic positions of SNPs as 'time'. Therefore, the transition probabilities depend on the distance

between SNPs. Let $z_{i,j}$ be the underlying copy number state for the j^{th} SNP of subject *i* and $d_{j,j'}$ be the physical distance between SNPs *j* and j' in the genome. We define the transition probability between the copy number states of SNPs *j* and j' as:

$$p_{s,s'}(d_{j,j'}) = p(z_{i,j'} = s' \mid z_{i,j} = s) = \begin{cases} \exp(-d_{j,j'} / \lambda_s) & if\ s = s' \\ (1 - \exp(-d_{j,j'} / \lambda_s)) p_{s,s'} & if\ s \neq s' \end{cases};$$

$$\text{where } 1 \leq s, s' \leq 5 \text{; and } \sum_{s' \neq s} p_{ss'} = 1$$

Here, $p_{s,s}(d_{j,j'})$ is the probability for a hidden state *s* at SNP *j* to stay at the same state at SNP j' over a distance of $d_{j,j'}$, which is modeled by an exponential distribution with parameter $1/\lambda_s$. Therefore, λ_s has the interpretation of the expected length for the copy number to stay at a particular state *s*. It is worthwhile to note that $p_{s,s}(d_{j,j'})$ may be close to zero when $d_{j,j'}$ is large. In practice, when two consecutive SNPs are far apart, the Markov chain will be restarted to avoid the probability of zero for the underlying copy number to stay at the same state.

Emission Probability of Observations

Similar to a few previous studies, we use LRR and BAF as observations, which are modeled by mixture distributions [133,135,136]. Denote $z_{i,j}$ as the underlying copy number state for the j^{th} SNP of subject *i*. Assume the *LRR* and *BAF* at a particular SNP locus are conditionally independent given the underlying copy number state, we have

$$p(R_{i,j}, B_{i,j} \mid z_{i,j}) = p(R_{i,j} \mid z_{i,j}) p(B_{i,j} \mid z_{i,j})$$

We model the emission probability of *LRR* (*R*) with the mixture of a uniform distribution and a normal distribution as:

$$p(R_{i,j} \mid z_{i,j} = s) = \frac{\pi_R}{R_M - R_m} + (1-\pi_R) f(R_{i,j}, \mu_{R,s}, \sigma_{R,s});$$

$$1 \le i \le N; 1 \le j \le K; 1 \le s \le 5;$$

where *f(.)* denotes the normal probability density function. Here, we assume the genotyping may fail with a small probability of π_R. Under such a circumstance, the LRR is observed as background noise, which follows a uniform distribution between its possible minimum (R_m) and maximum values (R_M). Otherwise, it follows a normal distribution with a mean ($\mu_{R,s}$) and a standard deviation ($\sigma_{R,s}$) with respect to the underlying copy number state. As illustrated by Table 3.1, the mean and standard deviation of the normal distributions vary by the underlying copy number states.

As illustrated in Table 3.1, the expected values of *BAF*s (*B*) vary by the underlying copy number states as well as the underlying genotypes. Let G_s be the set of all possible genotypes for copy number state *s*. We further denote $\mu_{s,g}$ and $\sigma_{s,g}$ as the mean and standard deviation of *BAF* for a SNP with copy number state *s* and genotype *g*, where $g \in G_s$. Let $\psi_{s,g}$ denotes the prior probability of *BAF* for copy number state *s* and genotype *g*, which can be calculated by a binomial distribution based on the B allele frequency in the population (*bpf*) [133,135,136]. For example, a SNP with genotype AAB has a copy number of 3, and expected BAF of 1/3. The prior probability of the BAF can be calculated as:

$$p(G_{i,j} = AAB \mid z_{i,j} = 3) = \binom{3}{1} (bpf_j)^1 (1 - bpf_j)^2.$$

Because the B allele frequencies are only observed within the range of 0 and 1, we model the emission probability of BAF at a particular SNP locus with the mixture of a uniform distribution and a few normal and truncated normal distributions:

$$p(B_{i,j} \mid z_{i,j} = s) = \begin{cases} \pi_B + (1-\pi_B) \sum_{g \in G_s} \psi_{s,g} f(B_{i,j}, \mu_{s,g}, \sigma_{s,g}) & for \;\; 0 < B_{i,j} < 1 \\ \pi_B + (1-\pi_B) \sum_{g \in G_s} \psi_{s,g} \Phi(B_{i,j}, \mu_{s,g}, \sigma_{s,g}) & for \;\; B_{i,j} = 0 \\ \pi_B + (1-\pi_B) \sum_{g \in G_s} \psi_{s,g} (1 - \Phi(B_{i,j}, \mu_{s,g}, \sigma_{s,g})) & for \;\; B_{i,j} = 1 \end{cases} ;$$

$$1 \le i \le N; 1 \le j \le K; 1 \le s \le 5$$

where $f(.)$ and $\Phi(.)$ are the normal probability density function and cumulative density function respectively.

Parameter Estimation and Copy Number Inference

In practice, we assume $\pi_R = \pi_B = 0.01$ as the empirical error rate for genotyping, and λ_s, $1 \le s \le 5$, are pre-determined to account for the size of copy number variants. The set of parameters that need to be estimated includes:

$$\Omega = \{\omega(s) = p(z = s) \;\; as \;\; starting \;\; probability; \;\; s = 1,2,3,4,5$$
$$P = (p_{ss'}) \;\; as \;\; transition \;\; probability; 1 \le s, s' \le 5$$
$$\mu_{R,s}; \;\; as \;\; mean \;\; of \;\; R; \;\; s = 1,2,3,4,5$$
$$\sigma_{R,s}; \;\; as \;\; srandard \;\; deviation \;\; of \;\; R; s = 1,2,3,4,5$$
$$\mu_{B,s,g}; \;\; as \;\; mean \;\; of \;\; B; s = 1,2,3,4,5; \;\; g = 1,2,.....G_S$$
$$\sigma_{B,s,g}; \;\; as \;\; s\tan dard \;\; deviation \;\; of \;\; B; s = 1,2,3,4,5; g = 1,2,.....G_S\}$$

We use the forward-backward algorithm, also known as Baum-Welch algorithm, to optimize the parameters in Ω [142]. After the parameter estimation, the inference of copy number states is carried out by Viterbi algorithm [143].

Baum-Welch Algorithm

The Baum–Welch algorithm is a particular case of the generalized expectation-maximization (GEM) algorithm, and is commonly used to estimate parameters of a hidden Markov model. The estimation is achieved by updating parameters interactively to maximize the likelihood function of the observations from a HMM. The likelihood can be calculated efficiently via a forward algorithm and a backward algorithm. To describe the Baum-Welch algorithm, we denote $O_{i,j} = (R_{i,j}, B_{i,j})$ as the observations at j^{th} SNP of subject *i*, and $e(O_{i,j}, s)$ as the emission probability of $O_{i,j}$ given the underlying copy number state *s* at SNP locus *j*. We have

$$e(O_{i,j}, s) = p(O_{i,j} \mid z_{i,j} = s) = p(R_{i,j} \mid z_{i,j} = s)p(B_{i,j} \mid z_{i,j} = s)$$

The forward algorithm computes the likelihood of the first k observations in a HMM with a particular ending state s recursively by:

$$\alpha(i,k,s) = p(O_{i,1},.....,O_{i,k}, z_k = s) = \begin{cases} \omega(s)e(O_{i,1},s) & k = 1 \\ \sum_{s'} \alpha(k-1,s')p_{s's}(d_{s's})e(O_{i,k},s) & 1 < k \le K \end{cases}$$

Specifically, the overall likelihood of a Markov chain with length of *K* can be calculated by:

$$L_i = p(O_{i,1},.....,O_{i,K}) = \sum_s p(O_{i,1},.....,O_{i,K}, z_{i,K} = s) = \sum_s \alpha(i,K,s)$$

On the other hand, given the current underlying state of *s* at the k^{th} SNP, the backward algorithm computes the likelihood of the future observations starting at the $(k+1)^{th}$ SNP.

$$\beta(i,k,s) = p(O_{k+1},.....,O_K \mid z_{i,k} = s) = \begin{cases} 1 & k = K \\ \sum_{s'} p_{ss'}(d_{k,k+1})e(O_{i,k+1},s')\beta(i,k+1,s) & 1 \le k < K \end{cases}$$

Therefore, the posterior distribution for the underlying state of the k^{th} SNP in subject i can be calculated as:

$$\gamma(i,k,s) = p(z_{i,k} = s \mid O_{i,1},.....,O_{i,K}) = \frac{p(z_{i,k} = s, O_{i,1},.....,O_{i,K})}{p(O_{i,1},.....,O_{i,K})}$$
$$= \frac{p(z_{i,k} = s, O_{i,1},.....O_{i,k}) p(O_{i,k+1},.....,O_{i,N} \mid z_{i,k} = s)}{p(O_{i,1},.....,O_{i,K})}$$
$$= \frac{\alpha(i,k,s)\beta(i,k,s)}{L_i}$$

In each iteration step by Baum-Welch Algorithm, the parameters are updated by maximizing the posterior probability. Denote the overall forward probability, backward probability and likelihood for N subjects as:

$$\alpha(k,s) = \prod_{i=1}^{N} \alpha(i,k,s);$$

$$\beta(k,s) = \prod_{i=1}^{N} \beta(i,k,s);$$

$$L = \prod_{i=1}^{N} L_i .$$

Therefore, the overall posterior probability for the underlying state of each SNP can be given as:

$$\gamma(k,s) = p(z_k = s \mid O_{.,.}) = \frac{\alpha(k,s)\beta(k,s)}{L},$$

where $O_{.,.}$ is all the observations for $O_{i,k}, \forall 1 \le i \le N; \forall 1 \le k \le K$.

The starting probability $\omega(s)$ can be updated by its posterior distribution as:

$$\hat{\omega}(s) = p(z_1 = s \mid O_{.,.}) = \frac{\alpha(1,s)\beta(1,s)}{L}; 1 \le s \le 5 .$$

The transition probability $p_{s,s'}$ can be updated as:

$$\hat{p}_{s,s'} = \frac{\sum_{k=2}^{K} \alpha(k-1,s) p_{s,s'}(d_{k-1,k}) e(O_{.,k}, s') \beta(k,s')}{\sum_{k=2}^{K} \sum_{l \neq s} \sum_{k=2}^{K} \alpha(k-1,s) p_{s,l}(d_{k-1,k}) e(O_{.,k}, l) \beta(k,l)} :.$$

The mean and variance of the LRR (R) with respect to the copy number states can be updated as:

$$\hat{\mu}_{R,s} = \frac{\sum_{1 \le i \le N; 1 \le j \le K} R_{i,j} p(R_{i,j} \mid z_{i,j} = s)}{\sum_{1 \le i \le N; 1 \le j \le K} p(R_{i,j} \mid z_{i,j} = s)};$$

$$\hat{\sigma}^2_{R,s} = \frac{\sum_{1 \le i \le N; 1 \le j \le K} (R_{i,j} - \hat{\mu}_{R,s})^2 p(R_{i,j} \mid z_{i,j} = s)}{\sum_{1 \le i \le N; 1 \le j \le K} p(R_{i,j} \mid z_{i,j} = s)}.$$

To estimate the mean and variance of the BAF (B) with respect to the copy number states and genotypes, we denote $| g |_B$ as the allelic copy number of B for genotype g (i.e. $| AAB |_B = 1$), and *bpf* as the allele frequency of B in the population. We update the mean and variance of *BAF* as:

$$\hat{\mu}_{B,s,g} = \frac{\sum_{1 \le i \le N; 1 \le j \le K} B_{i,j} p(B_{i,j} \mid z_{i,j} = s) p(G_{i,j} = g \mid Z_{i,j} = s)}{\sum_{1 \le i \le N; 1 \le j \le K} p(B_{i,j} \mid z_{i,j} = s) p(G_{i,j} = g \mid z_{i,j} = s)};$$

$$\text{where } p(G_{i,j} = g \mid z_{i,j} = s) = \binom{z_{i,j}}{| g |_B} (bpf_j)^{|g|_B} (1 - bpf_j)^{z_{i,j} - |g|_B};$$

$$\hat{\sigma}^2_{R,s} = \frac{\sum_{1 \le i \le N; 1 \le j \le K} (B_{i,j} - \hat{\mu}_{B,s,g})^2 p(B_{i,j} \mid z_{i,j} = s) p(G_{i,j} = g \mid z_{i,j} = s)}{\sum_{1 \le i \le N; 1 \le j \le K} p(B_{i,j} \mid z_{i,j} = s) p(G_{i,j} = g \mid z_{i,j} = s)};$$

Viterbi Algorithm

Given all the model parameters, the Viterbi algorithm is used to infer the most likely path for the underlying states. The following steps are implemented:

i) Calculate $v(i,1,s) = \omega(s)e(O_{i,1},s); 1 \le s \le 5; 1 \le i \le N$; as the probability to produce the first observation for a hidden Markov chain staring with underlying state *s*.

ii) Calculate $v(i,j,s) = \max_{s'}(v(i,j-1,s')p_{s's}(d_{j-1,j})e(O_{i,j},s)); \ 2 \le j \le K$ as the largest probability to produce first j^{th} observations for a hidden Markov chain to end at state *s*.

iii) Infer the most probable underlying state for the K^{th} SNP of a hidden Markov chain as: $z_{i,K} = \arg\max_s v(i,K,s)$.

iv) Recursively infer the most probable underlying state for the j^{th} SNP of a hidden Markov chain as: $path(z_{i,j} \mid z_{i,j+1} = s) = \arg\max_{s'}[v(i,j,s')p_{s's}(d_{j,j+1})]$ for $1 \le j \le K-1$.

3.3. Results

3.3.1. Simulation Study

In the simulation study, we assume that the length of genome is 10^6 base pairs. We first simulate *10K* SNPs with their physical positions uniformly distributed in the genome. The expected lengths of the copy number states are set at $\lambda_3 = 50K$ and $\lambda_l = 5K; l = 1,2,4,5$. The transition probabilities between copy number states are set as:

$$(p_{ss'}) = \begin{pmatrix} 0 & 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0 & 0.97 & 0.01 & 0.01 \\ 0.25 & 0.25 & 0 & 0.25 & 0.25 \\ 0.01 & 0.01 & 0.97 & 0 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 & 0 \end{pmatrix}$$

The parameters for the emission probability of *LRR* (*R*) are set as:

State	1	2	3	4	5
$\mu_{R,s}$	$\log_2(1/10)$	$\log_2(1/2)$	$\log_2(1)$	$\log_2(3/2)$	$\log_2(2)$
$\sigma_{R,s}$	1	0.15	0.15	0.15	0.2

The parameters for emission probability of BAF (*B*) are set as:

$$(\mu_{B,s,g}) = \begin{pmatrix} 0.5 & & & & \\ 0 & 1 & & & \\ 0 & 0.5 & 1 & & \\ 0 & 1/3 & 2/3 & 1 & \\ 0 & 1/4 & 2/4 & 3/4 & 1 \end{pmatrix} \text{ and } (\sigma_{B,s,g}) = \begin{pmatrix} 0.25 & & & & \\ 0.05 & 0.05 & & & \\ 0.05 & 0.05 & 0.05 & & \\ 0.05 & 0.05 & 0.05 & 0.05 & \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \end{pmatrix} \text{ for}$$

$$\begin{matrix} s=1 \\ s=2 \\ s=3 \\ s=4 \\ s=5 \end{matrix} \begin{pmatrix} g=- & & & & \\ g=A & g=B & & & \\ g=AA & g=AB & g=BB & & \\ g=AAA & g=AAB & g=ABB & g=BBB & \\ g=AAAA & g=AAAB & g=AABB & g=ABBB & g=BBBB \end{pmatrix},$$

where $g=-$ denotes a copy number loss on both chromosome. The observations of *B* are further truncated at 0 and 1.

We simulate 100 subjects by using the above model parameters. For each subject, the underlying copy number states and genotypes of 10K SNPs are first simulated in a sequential order according to the transition probabilities. The frequencies of allele B in the population follow a uniform distribution between [0.1, 0.9]. For each SNP, the observations of LRR (*R*) and BAF (*B*) are then simulated by using the emission probability according to the underlying copy number state and genotype. Two subjects are randomly selected to estimate the parameters by using the Baum-Welch algorithm. The estimated parameters are then used to infer the underlying copy number states for all subjects by using the Viterbi algorithm. For computational precision

reason, the convergence criterion is met when the summation of the absolute change of all parameters is less than 10^{-3}. We calculate the error rates for inferring the copy number states of all SNPs in all subjects. Because the expected lengths of the copy number variants (λ_S) are pre-determined and may have an impact on the performance of the inference, we also examine the error rates when they are incorrectly specified. The results are listed in Table 3.2. The simulation results show that the proposed method is highly accurate to infer the underlying copy number states when λ_S is correctly specified. The overall error rate for all SNPs is estimated to be 1.34e-04. When λ_S is incorrectly specified, the error rate increases with the extent of mis-specification, but remains at a low level. In our simulation, we find that the error rate is not inflated seriously with an up-to 10 folds over-specification of λ_S. It is also noted that the error rate for SNPs with a normal state of two copies decreases by the extent of over-specification of λ_S. This is because the majority of SNPs belong to a normal state of two copies. The normal state also has the largest expected length, and a SNP is more likely to be inferred as two copies when λ_S is large. On the other hand, the error rate for SNPs with a normal state of two copies increases when λ_S is under-specified. Overall, the error rate is properly controlled when λ_S is incorrectly specified.

3.3.2. Application to Breast Cancer Data

We also apply the proposed method to detect copy number variants that are associated with breast cancer development, using a recent GWAS study among Ashkenazi Jews (AJ) [66]. The original study had three phases. The first phase included 249 breast cancer cases without *BRCA1* and *BRCA2* mutations, and 299 cancer-free AJ women as controls. The second phase was a replicate study with 343 candidate SNPs among 950 AJ cases and 979 AJ controls. The third

phase was also a GWAS study that included 243 AJ cases and 187 controls. The participants from phase I and phase III were genotyped with Affymetrix 500K SNP array, while those from phase II were genotyped by Illumina GoldenGate assay. Because our method is proposed for Affymetrix SNP arrays, we focus our analysis on the phase I and phase III data.

We use phase III as an initial study for the analyses. The proposed method is first applied to ten randomly selected controls for parameter training. The parameters are then used to infer copy number states among all participants. Because the inferred copy numbers are not normally distributed and their distributions are not straightforward to determine, we further conduct a Kolmogorov-Smirnov (KS) Test for each SNP to compare the inferred copy numbers between cases and controls. The KS test is a non-parametric test and does not rely on the distribution of copy numbers. The significant regions are selected if three consecutive SNPs show significant copy number differences at a level of 1e-07. After the region is selected, a global p-value is further calculated by conducting a KS test using the average copy number of the SNPs within the region. The results are summarized in Table 3.3. The findings include 34 genomic regions from 16 chromosomes. The region with the largest number of significant SNPs is 4q31.23. This region has 10 SNPs showing significant copy number difference between cases and controls. Besides region 4q31.23, two regions, 1p21.1 and 10q21.1, both have 7 significant SNPs. Three regions have 5 SNPs with significant copy number differences, including 6q22.33, 6q27 and 11p12. These results indicate that copy number alterations on chromosome 4, 6, 1 and 11 may have a significant impact on the development of breast cancer.

Most of these identified regions were reported in literature for potential involvement with the development of breast cancer. One SNP in the region 4q31.23 was recently reported to be significantly associated with breast cancer progression [144]. A gene *ARHGAP10-NR3C2,*

located in this region, was also suggested to be related to carcinogenesis through structural alteration [145]. In addition, possible copy number changes of region 4q31.23 were observed from cancer cell line data [146]. Regions 1p21.1 and 10q21.1 were reported repeatedly with potential association with breast cancer. Chromosome arm 1p was suggested to contain multiple tumor suppressor genes [147]. Structural alterations of 1p21.1 were observed from many studies of cancers [147,148,149,150]. Region 10q21.1 also contained multiple candidate tumor suppressors, such as *ANX7* and *CDC2* [151,152]. Interestingly for region 6q22.33, it was identified by the original GWAS as a novel locus for breast cancer development. Our study confirms this finding and also suggests that the copy number changes in the region may play an important role.

We further apply the same procedure to the phase I data for replication. The results are summarized in Table 3.3. Among the regions identified by using phase III data, the copy number changes remain significant at five regions by using phase I data, including 4q31.23, 6q13, 12q23.1, 13q14.3 and 2p21. These five regions contain 10, 5, 4, 4, and 3 SNPs respectively. The association between the CNVs within these regions and BRCA susceptibility is supported by the literature. The long arm of chromosome 6 was reported to be frequently rearranged in human cancers [153,154,155]. The region of 6q13 was among the regions that showed frequent copy number alterations [156,157]. In region 12q23.1, a gene *SLC5A8* was identified by a previous study to be affected frequently by structural changes, such as DNA methylation [158,159]. This gene was actively involved in the gene pathway related to the development of primary human tumors [160,161]. The region 13q14.3 was reported for copy number changes in various cancers, such as prostate cancer and breast cancer [162,163,164,165]. The structural changes of region 2p21 were well studied, such as 2p21 deletion syndrome [166]. It was caused by the deletion of a

larger portion of genetic material from chromosome 2p21, characterized by infant seizures, reduced muscle tone, developmental delay, lactic acidosis and unusual facial appearance [167]. The structural changes of 2p12 were also suggested to be involved in cancer development [168].

To validate the results, we also examine the distribution of the sizes of identified CNVs and compare it with findings from literature. This distribution has a similar shape with one recent study (Figure 1 of [169]). Further, the density function in our study has the peak value at around 50K. Compared to the study by Li *et al* (peak at 200K), the identified CNVs by our study have smaller sizes. This is expected since we start with single SNPs and focus on the small to intermediate size CNVs.

3.4. Discussion

Though genome-wide association studies have identified hundreds of novel disease-susceptibility loci [170], the genetic architecture of complex disease remains elusive [171]. A large percentage of the heritability of diseases is still unexplained, highlighting the need to consider all types of heritable variations besides the sequence variations. As promising candidates, CNVs are ubiquitous throughout the human genome. It was estimated that about 5%-16% of the human genome might undergo copy number changes [126,172,173]. Meanwhile, the knowledge of CNVs is still limited in regard to their contribution to the disease development among human populations. In addition, statistical tools are still lacking to infer the CNVs accurately and efficiently. In this research, we propose a HMM-based approach for copy number inference, illustrated with an application to breast cancer datasets. The method can be viewed as an extension of PICR model with an implementation of HMM. Our approach differs from the other HMM-based methods by: 1) our method first estimates the copy number abundance using PICR, which removes noises from probe-intensity values; 2) our method achieves equal scaling

among multiple subjects by a novel approach of copy number standardization. By doing so, no between-array normalization is required, which keeps the data integrity.

In the simulation study, we show that the proposed method is highly accurate for copy number inference. The error rates remain at a low level when the pre-determined model parameter is mis-specified. The application to the phase III data of the BRCA GWAS identifies a few genomic regions with significant associations. The associations of five regions, including 4q31.23, 12q23.1, 13q14.3 and 2p21, are replicated by using the phase I data of the BRCA GWAS. All these genomic regions have been reported in the literature as candidate regions for the development of primary tumors or other complex disorders. Whereas it is biologically plausible that the structural changes of these regions may play an important role in the development of breast cancer, further studies are needed to replicate the association and investigate the biological mechanisms.

We are also aware that our method may have a few limitations. First, our copy number estimation method is based on the design of Affymetrix SNP arrays. It currently cannot be directly applied to the Illumina platform. Second, our method currently focuses on detecting the total copy number changes, and does not differentiate the paternal/maternal specific copy numbers. Third, our method is currently suitable for population-based association studies with unrelated samples. Further extension is needed for its application to studies with samples from family pedigrees.

Table 3.1. Configuration of all possible copy number states

State (z)	Copy Number	Expected LRR	Expected BAF	Possible Genotypes
1	0	$\log(0) = -\infty$	0	- (deletion)
2	1	$\log_2(1/2)=-1$	0 1	A; B
3	2	$\log_2(1)=0$	0 0.5 1	AA; AB; BB
4	3	$\log_2(3/2)=0.585$	0 0.33 0.67 1	AAA; AAB; ABB; BBB
5	4	$\log_2(2)=1$	0 0.25 0.5 0.75 1	AAAA; AAAB; AABB; ABBB; BBBB

Table 3.2 Error rate for inference of copy number states with correctly and incorrectly specified expected length of copy number states

	Ave. # of SNP with copy number state in each subject					
HMM State	1	2	3	4	5	Total
	557	163	8875	185	220	10,000
λ used in HMM	Error Rates by copy number state					
λ_{true}	5.92e-04	1.53e-04	2.37e-05	1.40e-04	1.32e-03	1.34e-04
$2\lambda_{true}$ [a]	3.97e-03	4.91e-04	1.69e-05	7.01e-04	4.46e-03	3.55e-04
$5\lambda_{true}$	4.18e-03	6.13e-4	1.80e-05	7.01e-04	4.51e-03	3.71e-04
$10\lambda_{true}$	4.38e-03	9.20e-04	1.80e-05	1.08e-03	4.87e-03	4.02e-04
$0.5\lambda_{true}$	9.69e-04	1.53e-04	3.27e-05	1.56e-04	1.32e-03	1.66e-04

[a] the model specified λ is 2 times greater than the true λ.

Table 3.3 Regions showing significant copy number changes in Phase III and their significance levels in Phase I.

Chromosome	Cytoband	Location	# of SNP	p-val (Phase III)	p-val (Phase I)
1	p21.1	102622376 - 102640646	7	2.62e-13	0.954
1	p12	120292824 - 120312909	3	7.62e-14	0.999
1	q22	154077091 - 154106555	3	2.453e-11	0.999
2	**p21**	45759616 - 45760637	3	1.106e-08	0.014
2	p12	81196767 - 81197522	3	7.232e-09	0.977
2	q21.1	131925407 -131955270	3	4.872e-13	0.999
3	p14.3	57706175 -57839689	3	1.228e-09	0.116
4	q26	117544365 -117576957	3	4.577e-11	0.138
4	**q31.23**	148668320 -148697327	10	9.43e-15	7.56e-05
4	q32.3	166885930 -166957371	5	6.664e-11	0.189
5	q14.3	84350898 - 84398999	5	4.330e-14	0.720
5	q22.3	115145252 - 115178424	4	2.220e-16	0.893

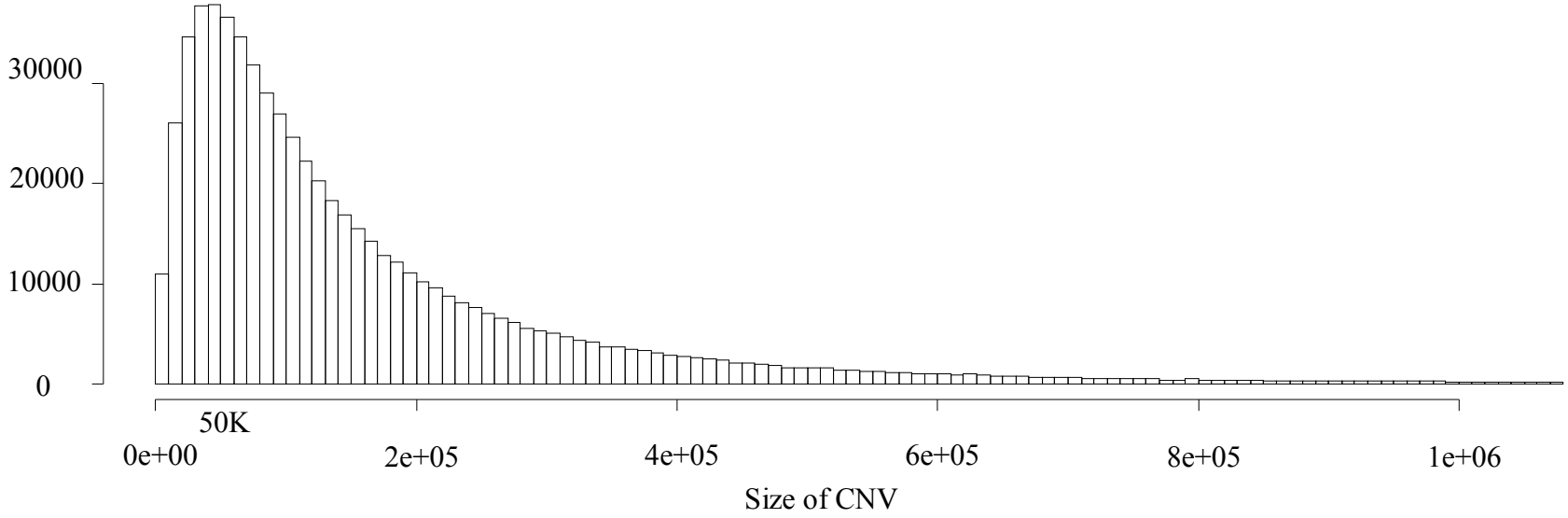
Table 3.3 . (cont'd)

Chromosome	Cytoband	Location	# of SNP	p-val (Phase III)	p-val (Phase I)
6	**q13**	75247853 -75311831	5	5.218e-15	0.034
6	q22.33	128476625 -128533696	6	2.409e-13	0.806
6	q23.2	134651674 -134672863	5	3.722e-10	0.999
6	q27	165234976 -165247908	6	1.752e-09	0.996
7	q22.1	98318717 -98361309	4	4.727e-11	0.103
7	q31.31	118754169 -118754169	5	1e-17	0.524
8	q11.22	52786953 -52796842	3	4.550e-10	0.840
8	q21.3	90963387 -90964181	3	2.862e-08	0.772
8	q24.13	125649171 - 139914783	3	2.30e-08	0.973
8	q24.3	145891814 - 145948840	4	3.220e-15	7.96e-04
9	p21.3	22270796 -22294230	5	6.249r-09	3.33e-03
10	q21.1	56853055 -74432554	7	1.084e-09	0.998

Table 3.3 . (cont'd)

Chromosome	Cytoband	Location	# of SNP	p-val (Phase III)	p-val (Phase I)
11	p13	36306019 - 36366302	3	8.95e-11	0.223
11	p12	37905557 - 37916354	6	2.627e-09	0.968
11	q22.3	104741435 - 104806689	5	4.152e-14	0.999
12	**q23.1**	94977527 - 95052366	4	1.11e-16	1.07e-04
13	q13.3	34828145 - 34846106	4	8.975e-10	0.428
13	**q14.3**	51036156 - 51071687	4	5.268e-12	6.83e-09
13	q33.1	103334252 - 103344370	5	1.589e-09	0.964
14	q23.1	60136001 - 60140123	5	1.843e-12	0.996
18	p11.31	3597746 - 3635894	3	4.268e-10	0.417
X	q27.3	146596395 - 146646974	4	5.873e-14	0.086

Figure 3.1. Distribution of the size of CNVs



CHAPTER 4.

GENE-GENE INTERACTION AND CIGARETTE SMOKING

4.1. Introduction

The genetic etiology of common complex diseases is of tremendous interest to clinical and basic science researchers as well as to the general public. During the past few years, the radical breakthrough of biotechnologies has enabled us to generate a large amount of genotypic data with great accuracy [61]. Testing the association between these genetic variants and complex traits provides an unprecedented opportunity to unravel the mystery of gene functions, which is crucial for a better understanding of the disease etiology. Meanwhile, the rapid growth of the data dimensionality also presents great challenges to statistical modeling and hypothesis testing.

Most of the first generation genome wide association studies test the association between genetic variants and disease outcomes on a single-locus basis [40,116,174]. Though a substantial number of genetic variants have been identified to be associated with many complex diseases, such as diabetes and Crohn's disease, they account for only a small percentage of the heritability [117]. One possible explanation of the issue of 'missing' heritability is that most of the complex diseases are polygenic in nature. Multiple genetic variants, each conferring a small or moderate effect, may contribute to the disease development [175,176]. In addition, the effect of one genetic variant can be suppressed or enhanced by the other variants, which is termed epistasis [111]. Whereas epistasis *per se* cannot account for the missing additive heritability, it may often lead to the lack of power to detect association when loci are examined separately without considering their potential interactions [177].

Considering the polygenic nature of many complex diseases, statistical approaches for multi-locus association analysis have been recently developed. Lin *et al.* proposed a sequence

interaction model in a multivariate regression framework for quantitative traits [178]. Several studies modeled multi-locus interactions through haplotype analysis [179,180,181]. Schaid *et al.* proposed a U-statistic-based score test that could simultaneously examine the association between multiple genetic variants and dichotomous traits [182]. Wei *et al.* further extended this approach for quantitative traits by using data-adaptive weights for different variants [183]. These approaches comprise the commonly used single-locus approaches, providing powerful alternatives for genetic association analysis. However, they are less suitable for handling a large number of genetic variants and for considering interactions, especially high order interactions.

Another group of methods uses a different strategy. They first select a subset of genetic variants from the totality of the genotyped variants, and then conduct an association test to assess the joint effect of the selected loci. The genetic variants are usually selected to best describe the risk of a binary disease outcome or the variation of a quantitative trait. For example, Ritchie *et al.* proposed a Multifactor Dimensionality Reduction (MDR) method for balanced case-control studies [112]. It pools multi-locus genotypes into high-risk and low-risk groups, and hence reduces the data dimension to one. This method has been further extended in a series of articles. Martin *et al.* extended the MDR method for studies with family-based designs [184]. Lou *et al.* derived a generalized MDR (GMDR) method that could be applied to both dichotomous and quantitative traits [113]. The GMDR method is not limited to studies with a balanced design and has the advantage of allowing for covariate adjustment. It maps the phenotypic traits into residual scores through certain link functions under a generalized linear model framework, and then conducts SNP selection and association test based on the residual scores. The extension of GMDR can also be applied to studies with family-based designs, referred to as pedigree-based

GMDR (PGMDR) [185].These approaches have now been commonly used to search for gene-gene/gene-environment interactions. They are generally non-parametric and model free.

However, the above methods commonly search for SNP combinations exhaustively. When the number of genetic variants is large, the chances are that an irrelevant combination may outperform the real disease model simply due to sample randomness. Therefore, when hundreds of thousands of genetic variants and environmental factors are examined, an exhaustive search may suffer from loss of power due to the substantial increase in the feature space [186]. In addition, an exhaustive search may not be computationally feasible for high order interactions, especially at a genome-wide scale. As discussed by Cordell and Marchini *et al.,* searching for high order epistasis beyond pair-wise interactions is not computationally affordable and can be pursued only after single-locus-based filtering [187,188].

As an alternative approach, the forward or sequential selection algorithm has received growing attention for its computational efficiency [189,190,191]. The algorithm starts with a null feature set and sequentially adds the best feature that satisfies certain criteria. Real data applications and simulation studies have also suggested that the forward search may have a greater power than the exhaustive search [186,189]. In this chapter, I propose a U-statistic-based multi-locus testing approach for quantitative traits. This method searches a large number of SNPs for joint gene-gene actions through a forward selection. Compared to the available methods, our method has the following advantages: 1) it tests the joint association for multiple genetic variants with the consideration of gene-gene interactions, including high-order interactions; 2) it is a non-parametric method that makes no assumptions of the trait distribution; and 3) it is computationally efficient and can be applied to high-dimensional data.

4.2. Methods

We first introduce notations and the hypothesis of interest. Suppose the study has *N* subjects. Let Y_i denote the quantitative trait for the i^{th} subject, $i = 1, 2, ..., N$; and let $X_i = (X_{i1}, X_{i2}, ..., X_{iK})$ denote *K* independent SNP genotypes, each taking a value from one of the three possible genotypes $X_{ij} \in \{AA, Aa, aa\}; j = 1, 2, ..., K$. The hypothesis is that these *K* SNPs, or a subset of them, are associated with the quantitative trait *Y*. To test this hypothesis, we first select *k* SNPs that best describe the variation of *Y*, where $k \le K$, and then test whether these selected SNPs are jointly associated with the trait *Y*.

U-Statistics

Since the foundational work of Hoeffding, U-Statistics have been widely used in both theoretical and applied statistical research [192]. They were recently used to build test statistics for multiple genetic variants [182,183]. However, while considering multiple genetic variants simultaneously, these approaches calculate the global U-Statistic by assuming an additive effect across multiple genetic variants, and thus do not consider the gene-gene interactions. We here introduce a new U-Statistic to test the joint association of multiple genetic variants with the consideration of gene-gene interactions. In this new method, we measure the difference of the quantitative traits between two subjects *i* and *j* as:

$$\phi(Y_i, Y_j) = Y_i - Y_j; \qquad 1 \le i, j \le N .$$

Suppose we have *k* selected SNPs, which comprise *L* multi-SNP genotypes, denoted by $G_1, G_2, ..., G_L$. A multi-SNP genotype, G_l, is defined here as a vector of *k* single-SNP genotypes that an individual carries (e.g., $\{g^1, g^2, ..., g^k\}$). The *k* SNPs and *L* multi-SNP genotypes are selected sequentially out of a total number of *K* genotyped SNPs (See Section

below for details). We denote by $S_l = \{i, X_i = G_l\}$ the group of subjects carrying multi-SNP genotype $G_l, l = 1, 2, ..., L$ and $m_l = |S_l|$ is the number of subjects in group S_l. We define the between-group U-statistic for group l and group l' as:

$$U_{l,l'} = \sum_{i,j} \phi(Y_i, Y_j); \quad i \in S_l, \quad j \in S_{l'}.$$

$U_{l,l'}$ is the summation of all possible pair-wise trait comparisons for any two subjects from S_l and $S_{l'}$. In the presence of an association, we expect individuals carrying different multi-SNP genotypes have different trait values (e.g., those carrying high risk multi-SNP genotypes have higher trait values than those carrying low risk multi-SNP genotypes). We assume that the expected quantitative trait values of L multi-SNP genotypes decreases with l (i.e., $E(Y_{S_1}) \geq E(Y_{S_2}) \geq ..... \geq E(Y_{S_L})$. Practically, we can sort the multi-SNP genotypes according to their average trait values (i.e., $\bar{Y}_{S_1} \geq \bar{Y}_{S_2} \geq ..... \geq \bar{Y}_{S_L}$). Based on $U_{l,l'}$, we further define a global U-statistic for L groups as:

$$U = \frac{\sum_{1 \leq l < l' \leq L} \omega_{l,l'} U_{l,l'}}{\sum_{1 \leq l < l' \leq L} U_{l,l'}} \times \frac{L(L-1)}{2}; \quad \omega_{l,l'} = \frac{\sqrt{m_l + m_{l'}}}{m_l m_{l'}}.$$

Here, the weight parameter ω is chosen to account for the number of subjects in each genotype group. This global U-Statistic measures the overall trait differences among a total number of L multi-SNP genotype groups.

The global U-Statistic described above is expected to have a zero mean under the null hypothesis of no association and to follow a normal distribution asymptotically. For simplicity, we denote $U = \sum_{1 \leq l < l' \leq L} \alpha_{l,l'} U_{l,l'}$, and the variance can be estimated as:

$$Var(U) = \sum_{1\le l<l'\le L} \alpha^2_{l,l'}(m^2_{l'}m_l + m^2_l m_{l'})\sigma^2 + \sum_{\substack{1\le l'_1,l'_2,l\le L \\ l'_1\ne l'_2}} \alpha_{l,l'_1}\alpha_{l,l'_2} m_l m_{l'_1} m_{l'_2}\sigma^2 + \sum_{\substack{1\le l_1,l_2,l'\le L \\ l_1\ne l_2}} \alpha_{l_1,l'}\alpha_{l_2,l'} m_{l_1} m_{l_2} m_{l'}\sigma^2$$
$$- \sum_{\substack{1\le l_1,l,l'_2\le L \\ l_1\ne l'_2}} \alpha_{l_1,l}\,\alpha_{l\ ,l'_2} m_{l_1} m_{l'_2} m_l\sigma^2 - \sum_{\substack{1\le l'_1,l,l_2\le L \\ l'_1\ne l_2}} \alpha_{l\ ,l'_1}\alpha_{l_2,l} m_{l'_1} m_{l_2} m_l\sigma^2$$

where $Var(Y_i) = \sigma^2$ for any $1 \le i \le N$. The derivation is described in Appendix.

U-Statistic-Based Forward Selection Algorithm

When a large number of SNPs are examined, it is likely that a significant proportion of the SNPs are not disease-related, and thus conducting a model selection will be necessary. We here introduce a computationally efficient U-Statistic-based forward selection algorithm that searches among a large number of SNPs for disease-susceptibility loci. A subset of loci is selected to best describe the variation of the traits. We start by taking all individuals as a single genotype group. In the first step, each SNP *j* can form two single-SNP genotypes, $\{g_1^j, g_2^j\}$, in three possible ways, denoted as $\{g_1^j = \{AA\}, g_2^j = \{Aa, aa\}\}$, $\{g_1^j = \{Aa\}, g_2^j = \{AA, aa\}\}$ and $\{g_1^j = \{aa\}, g_2^j = \{Aa, AA\}\}$. This leads to a total number of $3K$ possible partitions that can be represented by $\{G_1^{(1)} = g_1^j, G_2^{(1)} = g_2^j\}$, where $G_l^{(s)}$ denotes the l^{th} multi-SNP genotype at step *s*. We calculate the U-Statistic for each partition $\{G_1^{(1)}, G_2^{(1)}\}$. The SNP with the largest value of the U-statistic is selected, and the corresponding partition is recorded. In the second step, based on the first selected SNP, a second SNP j' is chosen to form four two-SNP genotypes, denoted by $\{G_1^{(2)} = G_1^{(1)} \,\&\, g_1^{j'}, G_2^{(2)} = G_1^{(1)} \,\&\, g_2^{j'}, G_3^{(2)} = G_2^{(1)} \,\&\, g_1^{j'}, G_4^{(2)} = G_2^{(1)} \,\&\, g_2^{j'}\}$. It should be noted that, if the same SNP from step one is chosen in step two, only three single-SNP genotypes will be formed, denoted by $\{G_1^{(2)} = \{AA\}, G_2^{(2)} = \{Aa\}, G_3^{(2)} = \{aa\}\}$. We screen all SNPs and

calculate the U-statistic for each of these partitions. The SNP that increases the U-statistic the most is chosen, together with its corresponding partition. As the algorithm moves forward, the global U-Statistic is expected to increase until all the genotype groups are separated. The largest number of possible genotype groups will be 3^K.

We use a 10-fold cross-validation (CV) procedure to decide when the selection algorithm should be stopped. In this procedure, all the subjects are randomly divided into 10 subgroups. Nine of the ten subgroups are used as training set, while the remaining one is used as testing set. The process is repeated ten times to make sure every subgroup has served as a testing set. A multi-SNP model is determined from each training set, and a U-statistic is calculated for the corresponding testing set. The selection algorithm is stopped when the U-statistics averaged over ten testing sets ceases to increase. After the number of forwarding steps is determined, a global U-statistic is calculated on the whole dataset including all subjects. The nominal significance level of the association can be tested by using the asymptotic distribution of the global U-Statistic. An empirical p-value, which accounts for the inflated Type I error due to the model selection, can be obtained by the permutation test.

Note that, although the illustration above is specified for joint gene-gene actions, the same procedure is also valid for joint gene-environment actions. Similar to genetic variables, environmental factors with categorical or ordinal levels can be directly analyzed. For continuous environmental factors, however, we need to first cluster them into different levels and then put them into the model as discrete variables.

4.3. Results

Simulation Results

We conduct two sets of simulations to evaluate the performance of the proposed method, and compare it with a commonly used approach, GMDR. The first set of simulations compares the performance of two approaches under various underlying disease models. The second set of simulations evaluates the performance of two approaches when the trait distribution is unknown. The quantitative traits for the second set of simulations are simulated according to the distributions of two traits from the Study of Addiction: Genetics and Environment (SAGE) dataset. The two traits are 'number of cigarettes smoked per day' and 'lifetime Fagerström Test for Nicotine Dependence (FTND) score'. The trait distributions in SAGE are illustrated in Figure 4.1.

Simulation I

In the first set of simulations, we consider a variety of underlying disease models, starting with three types of two-locus SNP models (Table 4.1.) introduced by Marchini *et al* (i.e., multiplicative-effect model, additive-effect model and threshold-effect model) [188]. We are here assuming only one SNP in each locus. To mimic more complex disease scenarios, we also simulate two three-locus models and two four-locus models. The two three-locus models, which are extensions of the two-locus models to three loci, are simulated with multiplicative and additive effects, respectively. Each of the four-locus models comprises two two-locus models (i.e., two two-way joint actions). We simulate the two-locus models of the first and second four-locus models with multiplicative and addictive effects, respectively. We further assume the effects between the two two-locus models for the first and second four-locus models follow an addictive model and a multiplicative model, respectively. The multi-SNP genotypes are simulated under the assumption of joint Hardy-Weinberg Equilibrium (HWE). For the two-locus models, the minor allele frequencies for the risk loci are set at 0.4 and 0.3. For the three-locus

models, they are set at 0.4, 0.5 and 0.3. For the four-locus models, they are set at (0.4, 0.3) and (0.3, 0.4) for each of the two-locus models, respectively. The allele frequencies remain fixed in this study unless specified otherwise. Noise loci are also introduced to mimic real data application. The minor allele frequencies of the SNPs at the noise loci are simulated from a uniform distribution ranging from 0.1 to 0.9. The number of noise loci is adjusted to ensure the total number of SNPs is always ten. A total of L multi-locus SNP genotypes are formed from the simulated SNPs at the ten loci, $\{G_1, G_2, ..., G_L\}$, corresponding to different levels of the quantitative trait. Assuming multi-locus group *l* has an expected trait value of μ_l, calculated based on the simulated setting (e.g., additive-effect model), we simulate quantitative traits for a reference population of one million subjects as:

$$y_i = \sum_{l=1}^{L} \mu_l I_{\{X_i = G_l\}} + \varepsilon_i;$$

where $\varepsilon_i \sim N(0,1)$ and $I_{\{.\}}$ is an indicator function. The forward U-test and GMDR are applied to 1000 subjects randomly selected from the reference population. For each underlying disease model, the simulation is repeated 1000 times with 1000 permutations. For both methods, the association is significant if the test statistic exceeds the 95-th percentile of the corresponding permutation distribution. The power is then calculated as the probability to detect the joint association based on 1000 replicates. In a similar manner, we calculate the type I error by only considering non-causal loci in the model.

The simulation results are summarized in Table 4.2. We report power, Type I error, sensitivity and specificity. The sensitivity (specificity) is calculated as the probability of selecting (not selecting) a causal (non-causal) SNP. U-statistics and Testing Balanced Accuracy (default in GMDR) are used as test statistics to examine the significance level of the two

methods. For GMDR, since the quantitative traits are simulated under a normal distribution, an identity link is used to calculate the score statistics. The simulation results show that, compared to GMDR, the forward U-test significantly increases the test power under multiplicative and additive models, while properly controlling the Type I error. For the threshold effect model, GMDR and the forward U-test have comparable power. In terms of selection accuracy, the sensitivity of GMDR tends to be higher than that of the forward U-test, with a few exceptions when both of the causal SNPs have large marginal effects. However, the specificity of the forward U-test is consistently higher than that of GMDR, and is greater than 0.95 in all scenarios. On the other hand, the specificity of GMDR is significantly reduced when the effect size decreases or the complexity of disease model increases. This result indicates that the forward U-test has a low false positive rate for SNP selection, which can partially explain the increase rather than loss of testing power over GMDR despite of the relatively lower sensitivity, because less noise loci is selected into the final model. The increase of power in most scenarios can also be explained by allowing for more than two risk groups in the model. The results in Table 4.2 show that: 1) for the additive and multiplicative effect models which contain more than two risk groups, the power of forward U-test is significantly higher than that of GMDR; 2) for the threshold effect models which contain only two risk groups, the power of forward U-test is comparable to that of GMDR.

Simulation II

We conduct a second set of simulations to compare the performance of two methods when the underlying trait distribution is unknown. Two quantitative traits are simulated according to the distributions of two variables in SAGE, 'number of cigarettes smoked per day' and 'lifetime Fagerström Test for Nicotine Dependence (FTND) score'. For each trait, two-SNP

disease models with three types of joint action effects, multiplicative, additive and threshold, are used for the comparison. Because of the unknown trait distribution, various link functions are used to calculate the residual scores for GMDR, including zero inflated Poisson, Poisson, negative binomial, and Gamma. The residual score for zero inflated Poisson is calculated with the package 'pscl' in R [193].

The simulation results are summarized in Table 4.3 and Table 4.4. For both traits, forward U-test attains a greater power than GMDR, especially under two-SNP models with additive or multiplicative effect. Among the trait distributions, GMDR has its best performance by assuming zero-inflated Poisson. When the underlying disease model is the threshold model, GMDR with a zero-inflated Poisson link can reach the same power as the forward U-test. However, the power of GMDR is significantly reduced if an inappropriate link function is used. In all scenarios, the specificity of forward U-test is greater than 0.95 and is consistently higher than that of GMDR. In terms of sensitivity, the performance largely varies, depending on the underlying disease models, effect sizes and link functions.

Application to Nicotine Dependence

We apply the proposed method to the Study of Addiction: Genetics and Environment (SAGE) GWAS dataset, searching for potential joint gene-gene actions among 155 known CS-associated SNPs. The participants of SAGE are unrelated individuals selected from three large, complementary studies: the Family Study of Cocaine Dependence (FSCD), the Collaborative Study on the Genetics of Alcoholism (COGA), and the Collaborative Genetic Study of Nicotine Dependence (COGEND). In our study, the trait of primary interest is the level of addiction to cigarettes, assessed by the answer to the question 'How many cigarettes do you smoke per day?'. It has four ordinal levels: 0 (10 cigarettes or less), 1 (11-20 cigarettes), 2 (21-30 cigarettes) and 3

(31 cigarettes or more). The sample sizes of FSCD, COGA and COGEND are 760, 799 and 1356, respectively. The distributions of traits in three studies are shown in Figure 4.2. From the literature, we select 155 SNPs across 67 candidate genes that have been reported for potential association with CS. In the SAGE dataset, genotypes for 128 SNPs are available, and genotypes for the remaining 27 SNPs are imputed by using PLINK [37,194]. The HapMap phase III founders of the CEU and ASW populations are used in the imputation as the reference panels for the white and black subjects [60].

We apply the forward U-test to FSCD for an initial association test and then replicate the initial findings in COGA and COGEND. Two SNPs, rs16969968 (A/G) and rs1122530 (C/T), are identified to be significantly jointly associated with the trait with a nominal p-value of 5.31e-7 in FSCD. Permutation test is also conducted and the empirical p-value is $p<0.001$. The two SNPs are located in genes *CHRNA5* and *NTRK2*. Evaluation of the finding in COGA (p-value=1.08e-5) and COGEND (p-value=0.02) shows that the association remains significant at 0.05 level (Table 4.5). The two SNPs together form four two-SNP genotypes:

$G_1 = \{\{AA \ or \ AG\} \& \{CC \ or \ CT\}\}$, $G_2 = \{\{AA \ or \ AG\} \& \{TT\}\}$,
$G_3 = \{\{GG\} \& \{CC \ or \ CT\}\}$, $G_4 = \{\{GG\} \& \{TT\}\}$. In order to study any potential interaction between the two SNPs, we calculate the average trait values in each genotype group. From FSCD, we find that the effect of rs16969968 is modified differently by the genotypes of rs1122530, indicating a potential interaction between the two SNPs (Figure 4.3). A similar trend is observed in COGA and COGEND (Figure 4.3). In particular, it should be noted that this interaction is "essential" and not completely removable by a monotonic transformation of the data [195].

We also apply GMDR on the same datasets. For the initial association study on FSCD, the disease models are searched with up to 3-way joint actions, and a zero inflated Poisson link is assumed. The results show that the model with two SNPs performs the best in terms of Testing Balance Accuracy, CV consistency, and sign test p-value (Table 4.6). Whereas GMDR identifies rs16969968 (A/G) that overlaps with the result of forward U-test, it also picks up a different SNP, rs573400 (A/G), which is located in gene *GABRA2.* Examination of the two SNPs in the other two datasets shows that the association remains significant in COGA (p-value=0.0001), but is not significant in COGEND (p-value=0.6230). We use linear regression models to fit the trait values with the grouping strategies identified by both methods and examine the goodness of fit with R-Squares (Table 4.7). The results show that the SNPs identified by GMDR have a better fit than the SNPs identified by forward U-test in FSCD, but not in COGA and COGEND. Both methods may indicate plausible joint gene-gene actions. Although the findings of both methods cannot be directly compared, the results from the association and goodness-of-fit analyses suggest that the finding of forward U-test may be more robust across different studies.

Figure 4.1.Trait distributions in Simulation II. A: the distribution of the number of cigarette smoked per day; B: the distribution of Participants' life-time score of FTND.

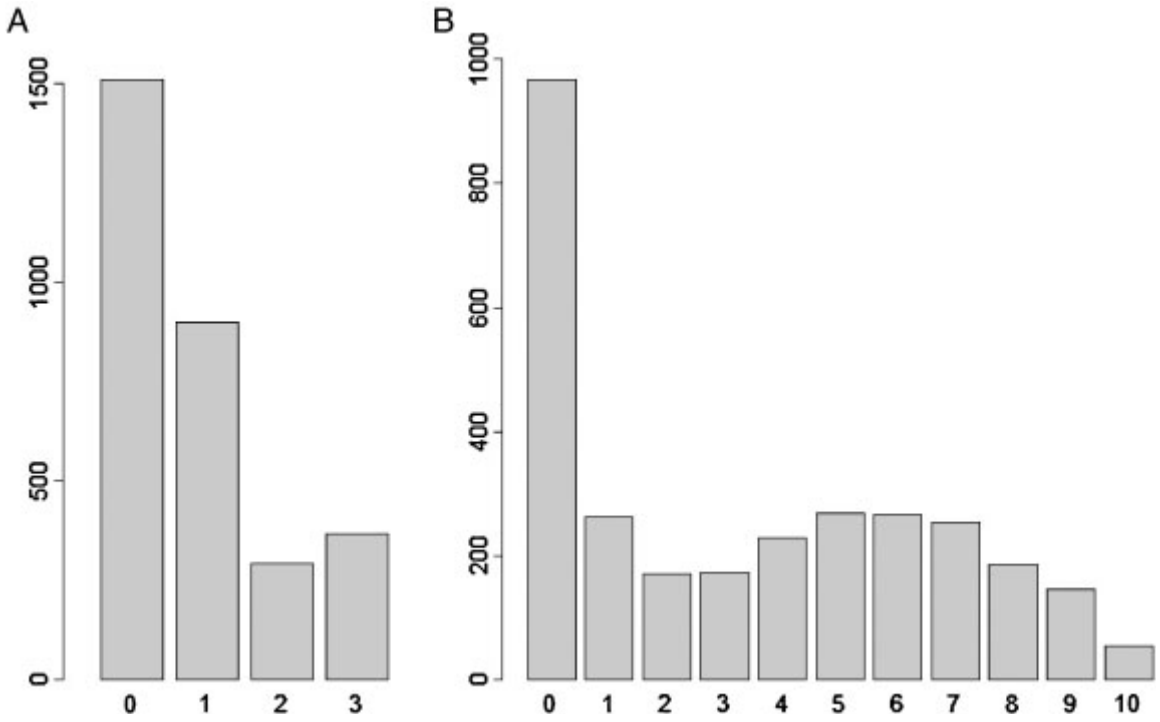


Table 4.1. Average trait values for two-locus joint action models

	Two-locus joint action with multiplicative effects			Two-locus joint action with additive effects			Two-locus joint action with a threshold effect		
	bb	Bb	BB	bb	Bb	BB	bb	Bb	BB
aa	α	$\alpha(1+\theta_{21})$	$\alpha(1+\theta_{21})(1+\theta_{22})$	α	$\alpha+\theta_{21}$	$\alpha+\theta_{22}$	α	α	α
Aa	$\alpha(1+\theta_{11})$	$\alpha(1+\theta_{11})(1+\theta_{21})$	$\alpha(1+\theta_{11})(1+\theta_{22})$	$\alpha+\theta_{11}$	$\alpha+\theta_{11}+\theta_{21}$	$\alpha+\theta_{11}+\theta_{22}$	α	$\alpha\ (1+\theta)$	$\alpha\ (1+\theta)$
AA	$\alpha(1+\theta_{11})(1+\theta_{12})$	$\alpha(1+\theta_{12})(1+\theta_{21})$	$\alpha(1+\theta_{12})(1+\theta_{22})$	$\alpha+\theta_{12}$	$\alpha+\theta_{12}+\theta_{21}$	$\alpha+\theta_{12}+\theta_{22}$	α	$\alpha\ (1+\theta)$	$\alpha\ (1+\theta)$

Table 4.2. Comparison between forward U-test and GMDR

Disease Model					Forward U-test	GMDR
Two-locus Multiplicative				Power	0.947	0.781
Average trait 1[a]	1	1.1	1.2	Type I Err.	0.026	0.038
				Sensitivity	0.564	0.604
Average trait 2	1	1.3	1.4	Specificity	0.968	0.921
Two-locus Multiplicative				Power	0.878	0.462
Average trait 1	1	1.2	1.3	Type I Err.	0.042	0.051
				Sensitivity	0.794	0.780
Average trait 2	1	1.2	1.3	Specificity	0.970	0.843
Two-locus Multiplicative				Power	0.623	0.342
Average trait 1	1	1.1	1.2	Type I Err.	0.059	0.048
				Sensitivity	0.589	0.695
Average trait 2	1	1.2	1.3	Specificity	0.961	0.789
Two-locus Additive				Power	0.991	0.824
Average trait 1	1	1.3	1.4	Type I Err.	0.038	0.045
				Sensitivity	0.915	0.817
Average trait 2	1	1.3	1.4	Specificity	0.978	0.878
Two-locus Additive				Power	0.762	0.267
Average trait 1	1	1.2	1.3	Type I Err.	0.074	0.061
				Sensitivity	0.749	0.758
Average trait 2	1	1.2	1.3	Specificity	0.964	0.803
Two-locus Additive				Power	0.474	0.098
Average trait 1	1	1.1	1.2	Type I Err.	0.06	0.044
				Sensitivity	0.562	0.672
Average trait 2	1	1.2	1.3	Specificity	0.954	0.746
Two-locus Threshold Model				Power	0.808	0.871
RAF[b]	0.4		0.4	Type I Err.	0.049	0.058
				Sensitivity	0.609	0.992
Average trait	1.5			Specificity	0.984	0.935
Two-locus Threshold Model				Power	0.412	0.385
RAF	0.5		0.5	Type I Err.	0.057	0.031
				Sensitivity	0.554	0.897
Average trait	1.3			Specificity	0.972	0.838

[a] Average trait value for genotype AA, Aa, aa of 1st causal SNP.

[b] Risk allele frequency for causal SNPs.

Table 4.2. (cont'd)

Disease Model					Forward U-test	GMDR
Three-locus Multiplicative				Power	0.612	0.198
Average trait 1	1	1.1	1.1	Type I Err.	0.042	0.051
Average trait 2	1	1.1	1.2	Sensitivity	0.408	0.609
Average trait 3	1	1.2	1.2	Specificity	0.960	0.741
Three-locus Additive				Power	0.672	0.228
Average trait 1	1	1.1	1.2	Type I Err.	0.054	0.045
Average trait 2	1	1.2	1.2	Sensitivity	0.454	0.610
Average trait 3	1	1.2	1.3	Specificity	0.967	0.799
Two-locus × Two-locus Multiplicative/Additive				Power	0.871	0.438
Average trait 1-1	1	1.1	1.2	Type I Err.	0.046	0.045
Average trait 1-2	1	1.2	1.3	Sensitivity	0.474	0.568
Average trait 2-1	1	1.2	1.3	Specificity	0.978	0.840
Average trait 2-2	1	1.1	1.2			
Two-locus × Two-locus Additive/Multiplicative				Power	0.838	0.483
Average trait 1-1	1	1.1	1.2	Type I Err.	0.059	0.050
Average trait 1-2	1	1.2	1.3	Sensitivity	0.423	0.533
Average trait 2-1	1	1.1	1.3	Specificity	0.976	0.840
Average trait 2-2	1	1.1	1.2			

Table 4.3. Comparison between forward U-test and GMDR when the quantitative traits are simulated from the distribution of number of cigarette smoked per day

Disease Model					Forward U-test	GMDR (Zero Infl. Poisson)	GMDR (Poisson)	GMDR (Negative Binomial)	GMDR (Gamma)
Two-locus Multiplicative				Power	0.930	0.540	0.289	0.217	0.355
Relative Risk 1	1	1.1	1.2	Type I Err.	0.056	0.056	0.079	0.061	0.052
				Sensitivity	0.562	0.634	0.546	0.511	0.659
Relative Risk 2	1	1.3	1.4	Specificity	0.957	0.870	0.910	0.938	0.841
Two-locus Threshold				Power	0.952	0.924	0.526	0.291	0.843
RAF	0.6		0.6	Type I Err.	0.050	0.054	0.063	0.074	0.064
				Sensitivity	0.754	0.982	0.781	0.623	0.952
Relative Risk	1.5			Specificity	0.967	0.944	0.918	0.943	0.931
Two-locus Additive				Power	0.948	0.694	0.343	0.247	0.579
Relative Risk 1	1	1.3	1.4	Type I Err.	0.056	0.069	0.066	0.086	0.063
				Sensitivity	0.749	0.789	0.646	0.570	0.764
Relative Risk 2	1	1.3	1.4	Specificity	0.966	0.847	0.920	0.932	0.870

Table 4.4. Comparison between forward U-test and GMDR when the quantitative traits are simulated from the distribution of life-time FTND scores

Disease Model					Forward U-test	GMDR (Zero Infl. Poisson)	GMDR (Poisson)	GMDR (Negative Binomial)	GMDR (Gamma)
Two-locus Multiplicative				Power	0.875	0.624	0.421	0.126	0.297
				Type I Err.	0.039	0.064	0.046	0.048	0.047
Relative Risk 1	1	1.1	1.2	Sensitivity	0.547	0.611	0.648	0.530	0.560
Relative Risk 2	1	1.3	1.4	Specificity	0.951	0.885	0.860	0.963	0.929
Two-locus Threshold				Power	0.779	0.780	0.107	0.450	0.064
				Type I Err.	0.048	0.055	0.061	0.057	0.068
MAF	0.4		0.4	Sensitivity	0.609	0.984	0.496	0.465	0.462
Relative Risk	1.5			Specificity	0.981	0.904	0.907	0.972	0.955
Two-locus Additive				Power	0.971	0.657	0.583	0.160	0.369
				Type I Err.	0.036	0.060	0.073	0.063	0.081
Relative Risk 1	1	1.3	1.4	Sensitivity	0.853	0.813	0.803	0.556	0.664
Relative Risk 2	1	1.3	1.4	Specificity	0.980	0.853	0.871	0.964	0.916

Table 4.5. Summary of two SNPs identified in FSCD and replicated in COGA and COGEND

SNP	Allele	Chro	Position	Gene	Grouping	p-values
rs16969968	A/G	15	78882925	CHRNA5	{AA,AG},{GG}	FSCD : 5.31e-7 COGA : 1.08e-5
rs1122530	C/T	9	87464352	NTRK2	{CC,CT},{TT}	COGEND :0.02

Table 4.6. Analysis result of GMDR in FSCD and replication in COGA and COGEND

Study		Model	Allele	Gene	Training Bal. Acc	Testing Bal. Acc	Sign Test (p)	CV
FSCD	1	rs2836823			0.5944	0.5511	7 (0.1719)	7/10
	2	rs16969968 rs573400	A/G A/G	CHRNA5 GABRA2	0.6448	0.6369	10 (0.001)	10/10
	3	rs16969968 rs573400 rs9321013			0.6764	0.5803	10 (0.001)	3/10
COGA		rs16969968 rs573400			0.6093	0.6107	10 (0.001)	10/10
COGEND		rs16969968 rs573400			0.5511	0.4840	5 (0.6230)	10/10

Table 4.7. Goodness of fit with the SNPs identified by forward U-test and GMDR

Study	R-Squares	
	forward U-test	GMDR
FSCD	0.0567	0.0656
COGA	0.0348	0.0165
COGEND	0.0051	0.0033

Figure 4.2. Trait distributions in FSCD, COGA and COGEND A: the distribution of trait in FSCD; B: the distribution of trait in COGA; C: the distribution of trait in COGEND.

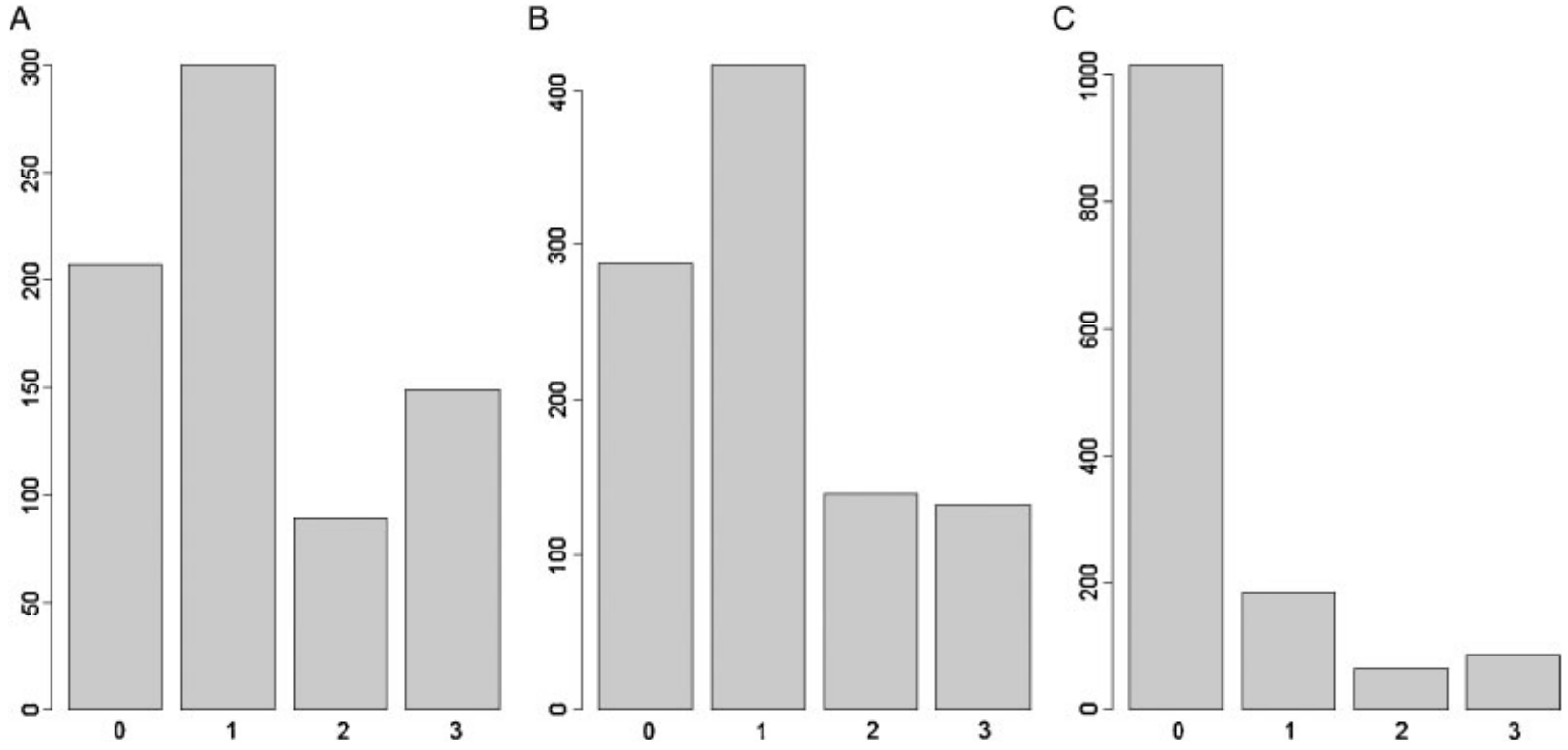
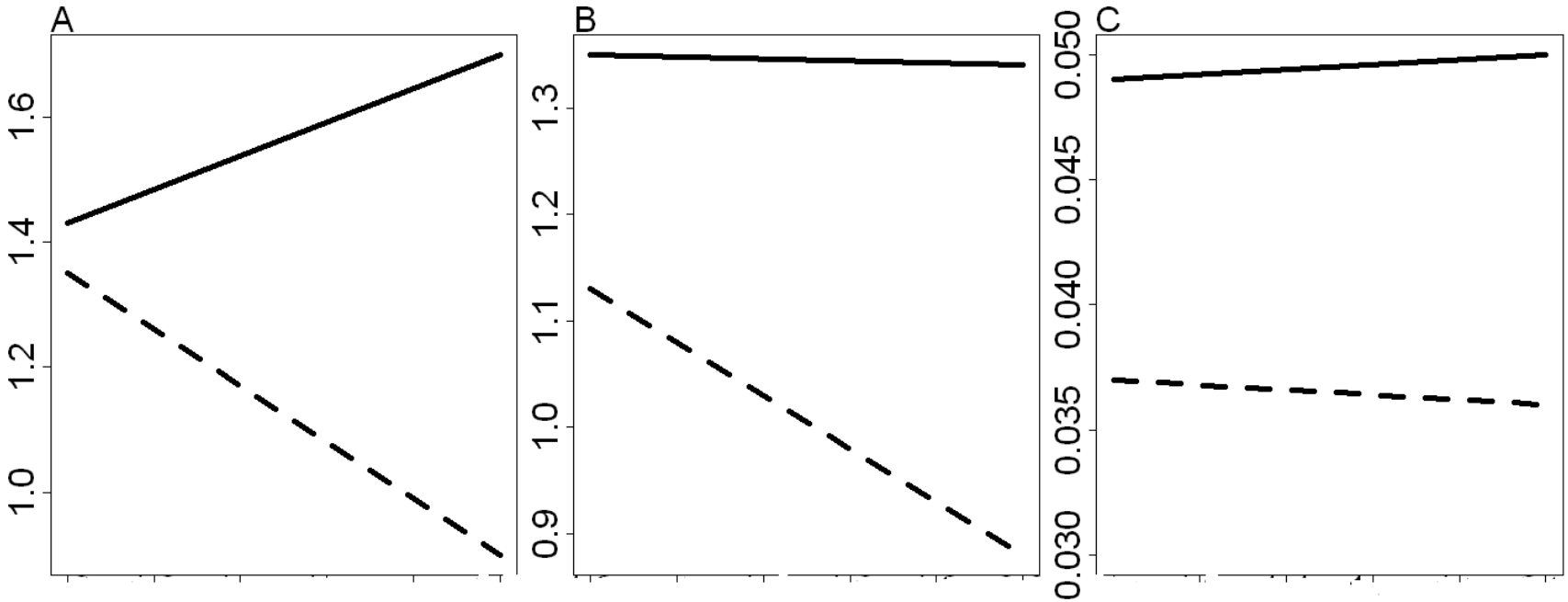


Figure 4.3. Joint effect of two SNPs showing potential statistical interaction A: average trait by genotype groups in FSCD; B: average trait by genotype groups in COGA; C: average trait by genotype groups in COGEND



4.4. Discussion

Complex traits are expected to be caused by the interplay of multiple genes and environmental factors through complicated mechanisms. If two genes are jointly involved in producing the variability of a phenotype whether additively or not, biological interaction between them or their products must be involved [196]. In addition, there may be statistical interaction that may or may not be removable by a transformation of the data [195]. Thus, statistical approaches that consider gene-gene/gene-environment interactions, including high order interactions, are more likely to take this complexity into account and can improve the discovery process of identifying important genetic variants. In this chapter, I have proposed a forward U-test for detecting joint association of multiple genetic variants, with the consideration of gene-gene interactions. Through simulations, we have shown that our method has a better performance than GMDR under various scenarios, whether statistical interaction exists or not. The improvement can be explained by several reasons. First, our method is an entirely non-parametric approach and makes no assumption about the trait distribution, while GMDR is based on a generalized linear model and implicitly specifies the link function with an assumption of the trait distribution. Second, similar to MDR, GMDR assumes two levels of the quantitative traits by clustering multi-locus genotypes into a high-risk group and a low-risk group. Our method measures the differences of traits between genotype groups without constraining the genotype groups to two levels. Therefore, our method may gain more strength from the quantitative variation of the trait. Third, unlike MDR and GMDR, which select a set of candidate models for each model size, the forward U-test uses a cross-validation procedure to choose the most parsimonious model, making it easier to interpret the results and replicate the findings. Forth, our method uses a forward selection, instead of an exhaustive selection, and substantially reduces the

search space of the SNP combinations. When the number of loci increases, the computational time and memory use for the exhaustive search increase exponentially, while those increase only quadratic for the forward search. This makes it computationally feasible for testing high-order interactions on high-dimensional data (e.g., whole genome-wide data). As discussed by Wu *et al* [186], the performance of the selection strategies depends on the underlying disease models. Our results indicate that, under additive and multiplicative models, forward selection outperforms exhaustive selection. However, we expect the power to decrease for forward selection if none of the genetic variants have any marginal effect. In this specific case, exhaustive selection will perform better than forward selection.

The forward U-test also differs from the other U-Statistic-based methods: 1) It calculates the global U-Statistic by a summation over the U-Statistics of multi-SNP genotype groups instead of each single SNP, which implicitly considers the joint gene-gene action that is additive or not; 2) It searches for the multi-SNP genotypes by a forward selection algorithm, which is important for high-dimensional data with a large number of non-functional SNPs. The size of the model selected by forward selection algorithm may depend on the study sample size. The larger the sample size is, the more likely a high-order interaction can be found. In addition, the choice of the weight parameter ω can also have an impact on the performance of the approach. Different weights can be used in the proposed method (e.g. $\omega_{l,l'} = 1$ for all l, l'). We choose $\omega_{l,l'} = \frac{\sqrt{m_l + m_{l'}}}{m_l m_{l'}}$ in our study because it appears to have the best testing power.

In the real data application, we identify two SNPs, located in *CHRNA5* and *NTRK2*, to be jointly associated with CS. Both *CHRNA5* and *NTRK2* have been suggested to be functionally related to CS. SNP rs16969968, a non-synonymous coding SNP in exon 5 of *CHRNA5*, was first

reported to be CS-related with a significance level of 0.00064 [109]. The association was also replicated in several other studies [197,198,199,200,201,202]. Studies also suggested that *CHRNA5* might interact with *CHRNA3* and *CHRNB4* to affect CS and lung cancer [203,204,205,206]. SNP rs1122530, a non-coding SNP in *NTRK2*, was found to be associated with CS in a haplotype analysis with two other SNPs (rs1659400 and rs1187272) in *NTRK2* [207]. A previous study found evidence of joint actions between *NTRK2* and multiple functional genes for CS, such as *CHRNA4*, *CHRNB2*, and *BDNF* [208]. However, to our knowledge, no joint action has been reported for *CHRNA5* and *NTRK2*. Although the joint association of *CHRNA5* and *NTRK2* with CS, involving a statistical interaction, reaches a statistically significant level and is replicated in independent studies, further studies will be necessary to replicate and investigate the statistical interaction.

Appendix

Estimation of the variance of the U-Statistic under the null hypothesis

Suppose we have a study sample of N subjects. We assume their quantitative traits are independent and have the same variance, denoted as $Var(Y_i) = \sigma^2$ for $i = 1, 2, ..., N$. Further, we assume that we have L multi-locus genotypes determined by the forward selection procedure, listed as $G_1, G_2, ..., G_L$. Following the same notation above, let $S_l = \{i, X_i = G_l\}$ and $m_l, l = 1, 2, ..., L$, be the number of subjects in S_l. The corresponding U-statistic is calculated as

$$U = \frac{\sum_{1 \le l < l' \le L} \omega_{l,l'} U_{l,l'}}{\sum_{1 \le l < l' \le L} \omega_{l,l'}} \times \frac{L(L-1)}{2}, \quad \text{where } \omega_{l,l'} = \frac{\sqrt{m_l + m_{l'}}}{m_l m_{l'}}$$

For simplicity, we denote

$$U = \sum_{1 \le l < l' \le L} \alpha_{l,l'} U_{l,l'}$$

The variance of the U-statistic can be expressed as

$$Var(U) = Var\Big(\sum_{1 \le l < l' \le L} \alpha_{l,l'} U_{l,l'}\Big) = \sum_{1 \le l < l' \le L} \alpha_{l,l'}^2 Var(U_{l,l'}) + \sum_{\substack{1 \le l_1 < l_1' \le L \\ 1 \le l_2 < l_2' \le L \\ (l_1, l_1') \ne (l_2 < l_2')}} \alpha_{l_1,l_1'} \alpha_{l_2,l_2'} Cov(U_{l_1,l_1'}, U_{l_2,l_2'})$$

For all $1 \le l < l' \le L$, we estimate the group-wise variance for the U-Statistic as:

$$\begin{aligned} Var(U_{l,l'}) &= Var\Big(\sum_{i \in S_l, j \in S_{l'}} \phi(Y_i, Y_j)\Big) = Var\Big(\sum_{i \in S_l, j \in S_{l'}} (Y_i - Y_j)\Big) \\ &= Var\Big(m_{l'} \sum_{i \in S_l} Y_i - m_l \sum_{j \in S_{l'}} Y_j\Big) = m_{l'}^2 \sum_{i \in S_l} Var(Y_i) + m_l^2 \sum_{j \in S_{l'}} Var(Y_j) \\ &= (m_{l'}^2 m_l + m_l^2 m_{l'}) \sigma^2 \end{aligned}$$

The covariance between group-wise U-Statistics is estimated according to different scenarios:

1) $l_1 \ne l_1' \ne l_2 \ne l_2'$,

$$Cov(U_{l_1,l_1'},U_{l_2,l_2'})=0$$

2) $l_1=l_2=l$,

$$\begin{aligned}Cov(U_{l_1,l_1'},U_{l_2,l_2'}) &= Cov(U_{l,l_1'},U_{l,l_2'}) = Cov(\sum_{i\in S_l,j_1\in S_{l'}}\phi(Y_i,Y_{j_1}),\sum_{i\in S_l,j_2\in S_{l'}}\phi(Y_i,Y_{j_2}))\\ &= Cov(\sum_{i\in S_l,j_1\in S_{l'}}(Y_i-Y_{j_1}),\sum_{i\in S_l,j_2\in S_{l'}}(Y_i-Y_{j_2}))\\ &= Cov(m_{l_1'}\sum_{i\in S_l}Y_i,m_{l_2'}\sum_{i\in S_l}Y_i)\\ &= m_{l_1'}m_{l_2'}Var(\sum_{i\in S_l}Y_i)\\ &= m_{l_1'}m_{l_2'}m_l\sigma^2\end{aligned}$$

3) $l_1'=l_2'=l$,

$$\begin{aligned}Cov(U_{l_1l_1'},U_{l_2l_2'}) = Cov(U_{l_1l},U_{l_2l}) &= Cov(\sum_{i_1\in S_{l_1},j\in S_l}\phi(Y_{i_1},Y_j),\sum_{i_2\in S_{l_2},j\in S_l}\phi(Y_{i_2},Y_j))\\ &= Cov(\sum_{i_1\in S_{l_1},j\in S_l}(Y_i-Y_j),\sum_{i_2\in S_{l_2},j\in S_l}(Y_i-Y_j))\\ &= Cov(m_{l_1}\sum_{j\in S_l}Y_j,m_{l_2}\sum_{j\in S_l}Y_j)\\ &= m_{l_1}m_{l_2}Var(\sum_{j\in S_l}Y_j)\\ &= m_{l_1}m_{l_2}m_l\ \sigma^2\end{aligned}$$

4) $l_1'=l_2=l$,

$$
\begin{aligned}
Cov(U_{l_1 l_1'}, U_{l_2 l_2'}) = Cov(U_{l_1 l}, U_{l l_2'}) &= Cov(\sum_{i \in S_{l_1}, j \in S_l} \phi(Y_i, Y_j), \sum_{j \in S_l, t \in S_{l_2}} \phi(Y_j, Y_t)) \\
&= Cov(\sum_{i \in S_{l_1}, j \in S_l} (Y_i - Y_j), \sum_{j \in S_l, t \in S_{l_2}} (Y_j - Y_t)) \\
&= Cov(-m_{l_1'} \sum_{j \in S_l} Y_j, m_{l_2'} \sum_{j \in S_l} Y_j) \\
&= -m_{l_1'} m_{l_2'} Var(\sum_{j \in S_l} Y_j) \\
&= -m_{l_1'} m_{l_2} m_l \ \sigma^2
\end{aligned}
$$

5) $l_1 = l_2' = l$ is equivalent to 4)

CHAPTER 5.

RARE VARIANTS AND QUANTITATIVE TRAITS

5.1. Introduction

The common disease-common variant (CDCV) model used to be well accepted for the genetic origin of complex human diseases. It asserts that the common complex diseases are caused by multiple common genetic variants, each conferring a small or moderate effect [209]. Based upon such a hypothesis, extensive genome-wide association studies (GWASs) have been conducted, bring into light many common variants underlying common complex diseases, such as breast cancer [79]. However, the genetic variants identified so far account for only a small percentage of the heritability of complex diseases [34]. It seems now clear that the genetic etiology of complex diseases is highly heterogeneous. Some genetic mutations, though individually rare, may impose a very high risk to disease development [52]. For example, germline mutations in more than ten genes were found to be associated with elevated risks of developing breast cancer [48]. Rare variants may play major roles in the development of complex diseases, and have received growing attention by investigators. With the fast development of biotechnologies, it is now feasible to genotype rare sequence variations in the general population rapidly and accurately [61]. Meanwhile, statistical methods are in great need to detect the association between these genetic variants and the common complex diseases.

The most commonly used approach for detecting the association between rare variants and a disease outcome is to group multiple rare variants into a single ‘super’ variant, which is further tested as a common variant. Based upon this idea, Li and Leal developed a Combined Multivariate and Collapsing (CMC) method for rare variants analysis [46]. It collapses the genetic variants with minor allele frequencies (MAFs) under certain threshold (e.g., 1%). Such a

strategy is extended in several other studies. Morris *et al.* developed a minor allele proportion method that calculated the accumulated minor allele frequency for multiple rare variants [210]. Han *et al*. proposed a data adaptive sum method by considering the opposite direction of genetic effect [211]. Madsen *et al.* used a weighted-sum method to group the rare variants according to their minor allele frequencies and biological functions (e.g., those within the same gene) [114]. Price *et al.* proposed to use data-determined thresholds of MAFs to differentiate common and rare variants, and incorporated the functional effect of amino acid changes [115]. Compared to the multivariate analysis of multiple rare variants, these collapsing methods can reduce the degree of freedom by creating a single 'super' variant comprised of multiple individual rare variants, and thus improve the testing power. In addition, testing on a single 'super' variant can reduce the burden of multiple testing. However, the existing methods also have a few limitations that may affect their performance. Collapsing all the rare variants in the same gene or genomic region, although biologically meaningful, can also introduce non-functional variants into the 'super' variant, which may diminish the signal that the functional variants carry. Intuitively, this limitation can be addressed by only collapsing a subset of disease-susceptibility rare variants. In what follows, we refer to the collapsing process using trait information as aggregation.

In this chapter, I propose an aggregating U-Test to examine the association between quantitative traits and multiple genetic variants, including both rare and common variants. The method first adaptively collapses the disease-susceptibility rare variants into a 'super' variant; it then searches the 'super' variant and the remaining common variants for the best multi-SNP combination, via a forward selection. We apply our method to GAW 17 mini-exome data, and compare its performance with a commonly used method, QuTie [46].

5.2. Methods

U-statistics were previously adopted to examine the joint association of multiple genetic variants with complex traits [182,183]. In Chapter 4, we have developed a forward U-test to detect gene-gene interactions [212]. In this chapter, we briefly describe our method and extend it with the consideration of both common and rare variants. Suppose we have a study population of N subjects. Let Y_i denote the observed value of the quantitative trait for the i^{th} subject, i=1,2,……N; let $X_i = (X_{i1}, X_{i2}, ..., X_{iK})$ denote the genotypes of K SNPs for the i^{th} individual, each taking a value from one of the three possible genotypes, $X_{ij} \in \{AA, Aa, aa\}$ j=1,2,……K. Without loss of generality, we assume A is the minor allele, and the first r SNPs, $(X_{i1}, X_{i2}, ..., X_{ir})$, are rare variants.

U-Statistics

Suppose we have L multi-SNP genotypes formed by k SNPs of interest, denoted by $G_1, G_2, ..., G_L$. A multi-SNP genotype, G_l, is defined here as a vector of k genotypes that a subject carries (e.g., $\{g^1, g^2, ..., g^k\}$). The k SNPs and L multi-SNP genotypes are selected sequentially out of a total number of K SNPs (See Section 'forward U-test' for details). Let $S_l = \{i : X_i = G_l\}, l = 1,2,....., L$, be the group of subjects carrying multi-SNP genotype G_l and $m_l = \mid S_l \mid$ be the number of subjects in S_l. We measure the trait difference between two multi-SNP genotype groups S_l and $S_{l'}$ as:

$$U_{l,l'} = \sum_{i,j} \phi(Y_i, Y_j), \qquad i \in S_l, j \in S_{l'}, l \neq l';$$

where the kernel function is chosen as $\phi(Y_i, Y_j) = Y_i - Y_j$. $U_{l,l'}$ is the summation of all possible pair-wise trait comparisons for any two subjects from S_l and $S_{l'}$. In the presence of an association, we expect subjects carrying different multi-SNP genotypes have different trait

values (e.g., those carrying high risk multi-SNP genotypes have higher trait values than those carrying low risk multi-SNP genotypes). Based on the $U_{l,l'}$, we can form a global U-statistic. We assume that the expected quantitative trait values of L multi-SNP genotypes decrease with l (i.e., $E(Y_{S_1}) \geq E(Y_{S_2}) \geq ..... \geq E(Y_{S_L})$). Practically, we can sort the multi-SNP genotypes according to their average trait values (i.e., $\bar{Y}_{S_1} \geq \bar{Y}_{S_2} \geq ..... \geq \bar{Y}_{S_L}$). We define a global U-statistic for L multi-SNP genotypes as

$$U = \frac{\sum_{1 \leq l < l' \leq L} \omega_{l,l'} U_{l,l'}}{\sum_{1 \leq l < l' \leq L} \omega_{l,l'}} \times \frac{L(L-1)}{2} \quad ; \quad \text{where } \omega_{l,l'} = \frac{\sqrt{m_l + m_{l'}}}{m_l m_{l'}} \qquad \text{…….. Eq. (1)}$$

Here, the weight parameter $\omega_{l,l'}$ is chosen to account for the number of subjects in different genotype groups. This global U-Statistic measures the overall trait differences among subjects from a total number of L multi-SNP genotype groups.

Aggregation of the Rare Variants

When a large number of rare variants are examined, it is likely that a significant proportion of these rare variants are not associated with the trait. Therefore, collapsing on a selected subset of rare variants will be necessary. Each rare variant can form two single-SNP genotypes, $\{\{AA,Aa\}$ and $\{aa\}\}$. We first calculate the U-statistics between two genotype groups for each rare variant using Equation (1), and then rank the U-statistics in a decreasing order as $U_{(1)}, U_{(2)}, ..., U_{(r)}$. Assume $V_{(1)}, V_{(2)}, ..., V_{(r)}$ are the corresponding rare variants in a candidate gene, and $X_{i(1)}, X_{i(2)}, ..., X_{i(r)}$ are their observed genotypes for subject *i*. We start from variant $V_{(1)}$, and define a 'super' variant as

$$R_{i1} = \begin{cases} 1 & X_{i(1)} = AA \mid Aa \\ 0 & X_{i(1)} = aa \end{cases}$$

At each step of the aggregation process, we choose a rare variant with the largest marginal U-statistics and add it to the 'super' variant. Accordingly, we re-define the 'super' variant as

$$R_{ij} = \begin{cases} 1 & R_{i(j-1)} = 1 \ or \ X_{i(j)} = AA \mid Aa \\ 0 & otherwise \end{cases} \qquad 2 \le j \le r$$

During the collapsing process, the 'super' variant always forms two genotype groups, for which a corresponding U-statistic can be calculated. The aggregation procedure stops at step *t*, where the U-statistic starts to decrease, (i.e., $U_{R_1} \le U_{R_2} \le ..... \le U_{R_t} > U_{R_{t+1}}$).

Forward U-test

A forward U-test [212] is then used to evaluate the 'super' variant and the other common variants for their joint association with the trait. We start the process by treating all individuals as a single group. In the first step, each common SNP *j* can form two single-SNP genotypes, $\{g_1^j, g_2^j\}$, in three possible ways, denoted as $\{g_1^j = \{AA\}, g_2^j = \{Aa, aa\}\}$, $\{g_1^j = \{Aa\}, g_2^j = \{AA, aa\}\}$, and $\{g_1^j = \{aa\}, g_2^j = \{Aa, AA\}\}$. As a special case, the 'super' variant can only form two single-SNP genotypes, $\{g_1^j = 1, g_2^j = 0\}$. This leads to a total number of $3(K-r)+1$ possible grouping strategies that can be represented by $\{G_1^{(1)} = g_1^j, G_2^{(1)} = g_2^j\}$, where $G_l^{(s)}$ denotes the l^{th} multi-SNP genotype at step *s*. We calculate the U-Statistic for each grouping $\{G_1^{(1)}, G_2^{(1)}\}$. The SNP with the largest value of U-statistics is selected, and the corresponding grouping is recorded. In the second step, based on the first selected SNP, a second SNP j' is chosen to form four two-SNP genotypes, denoted by

$\{G_1^{(2)} = G_1^{(1)} \,\&\, g_1^{j'}, G_2^{(2)} = G_1^{(1)} \,\&\, g_2^{j'}, G_3^{(2)} = G_3^{(1)} \,\&\, g_1^{j'}, G_4^{(2)} = G_2^{(1)} \,\&\, g_2^{j'}\}$. We calculate the U-statistics for each of these grouping strategies, and choose the one with the largest U-statistic. It should be noted that, if the same SNP from step one is chosen in step two, only three single-SNP genotypes are formed, denoted by $\{G_1^{(2)} = \{AA\}, G_2^{(2)} = \{Aa\}, G_3^{(2)} = \{aa\}\}$. As the algorithm moves forward, the U-Statistics are expected to increase until groups cannot be further split. This results a series of models with different numbers of groups. The best model with an appropriate number of groups can be determined by using a 10-fold cross-validation. The U-statistic of the best model is calculated using the whole dataset. The significance of association can be evaluated by a permutation test. For each permutation replicate, the same procedure, including the aggregation process, the model selection and the cross-validation, is applied to calculate the U-Statistics. By repeating the process 1000 times, we can have a null distribution of U-statistics and calculate the empirical p-value,

$$p = \sum I(U_{perm} \geq U) / 1000$$

5.3 Results and Discussion

We apply the proposed method to analyze the quantitative trait Q1 from GAW 17 mini-exome data. Thirty-nine SNPs located in 9 genes are associated with trait Q1. The minor alleles of these SNPs are associated with higher mean of Q1 and their frequencies range from 0.07% to 16.5%. We first adjust the trait by age using a linear regression model. The residual scores are used for our association studies. Based on two hundred replicates, we conduct a gene-based association study for each of the nine causal genes. For each gene, the traits are permutated for 1000 times to generate an empirical null distribution of U-statistics. The association is significant if the U-statistics exceed the 95th percentile of the null distribution. Similar analysis is also conducted

using QuTie-0.2. The threshold for rare variants is chosen as MAF<0.01. The results are summarized in Table 5.1.

Based on the analysis of 9 causal genes, we compare the performance of two methods.

1) Both methods have a high power to detect the association, when there are causal common variants with large effect sizes. This can be illustrated by the analysis of *FLT1* gene. There is one common variant in gene *FLT1* with a large effect size (0.65). The power to detect the association is 0.84 and 1 for QuTie and aggregating U-test, respectively.
2) The aggregating U-test has a significant power improvement over QuTie, when there are a large number of rare variants within a gene and only a small number of these rare variants are causal. This can be illustrated by the analysis of genes *ELAVL4, FLT4, HIF1A,* and *VEGFA*. For example, there are 7 rare variants and 3 common variants within gene *ELAVL4*, and only 2 rare variants are causal. The effect sizes of these variants, though relatively large (0.769 and 0.304), are subsided by collapsing with other SNPs, which reduces the power of QuTie. Same argument can also be applied to the other three genes, *FLT4, HIF1A* and *VEGFA*.
3) Both methods have a high power to detect the association, when the majority of the rare variants are causal. This can be illustrated by the analysis of gene *KDR*. For gene *KDR*, QuTie attains a higher power than the aggregating U-test. Since most of the rare variants in *KDR* are causal and their effect sizes are relatively large, it will be ideal to collapse all the rare variants. In such a case, the aggregation has less advantage because the selection process introduces additional variation. However, we believe this scenario is not common in real data applications.

4) Both methods have a low power to detect the association, when only a small proportion of the rare variants are causal, each having a small effect size. This can be illustrated by the analysis of genes *ARNT*, and *HIF3A*. For both genes, the selection of rare variants does not show any advantage because of the low effect size of each functional rare variant. In such a case, the power of both methods has no significant difference from Type I errors.
5) Both methods have a high power to detect the association when the majority of genetic variants under examination are causal, each having a large effect size. As an extreme case, gene *VEGFC* only has one rare variant, and therefore no selection is necessary. Due to its large effect size, both methods are able to detect the association. Interestingly, this variant has a MAF of 0.0717%, which is equivalent to 1 rare allele carrier out of 697 subjects. In such a case, we expect a low power of the association test if binary outcomes are used instead of quantitative traits.

In order to examine the Type I errors for the proposed method, we use the same genetic data and simulate the quantitative traits by assuming a standard normal distribution. The aggregating U-test is applied to 500 Monte Carlo simulated replicates to evaluate Type I errors. The results show that the Type I errors are well controlled (Table 5.1).

5.4 Conclusion

The performance of statistical methods to detect the association between rare variants and phenotypic traits may be affected by many factors, such as MAFs, the number of rare variants under examination, the number of functional rare variants and their effect sizes. Compared to the commonly used method QuTie, our method has two major advantages: 1) it can substantially improve the testing power when only a small number of rare variants are functional with

relatively large effect sizes; 2) it only collapses a subset of rare variants which are potentially trait-related. Therefore, it can also identify those disease-susceptibility rare variants.

Table 5.1. Power comparison between the aggregating U-test and QuTie

Group	Gene	# of Causal SNP/# of Total SNP in Gene		Power (QuTie)	Power (Agg. U)	Type I (Agg. U)
		Rare Variants	Common Variants			
1	*FLT1*	11/35		0.86	1.000	0.036
		8/25	3/10			
2	*ELAVL4*	2/10		0.025	0.585	0.050
		2/7	0/3			
	FLT4	2/10		0.58	0.715	0.060
		2/8	0/2			
	HIF1A	4/8		0.215	0.915	0.048
		3/7	1/1			
	VEGFA	1/6		0	0.265	0.060
		1/5	0/1			
3	*KDR*	10/16		0.99	0.840	0.038
		8/14	2/2			
4	*HIF3A*	3/21		0.055	0.05	0.040
		3/15	0/6			
	ARNT	5/18		0.05	0.07	0.038
		4/15	1/3			
5	*VEGFC*	1/1		0.745	0.785	0.054
		1/1	0/0			

CHAPTER 6.

SUMMARY AND FUTURE DEVELOPMENT

6.1 Summary

In this dissertation, I have conducted three studies that explore three possible sources of the "missing" heritability of complex diseases. First, the human genetic variations underlying complex diseases include both sequence variations and structural alterations. Copy number variants may encompass multiple genes as well as non-coding DNAs, accounting for a large proportion of the variability among human genomes. These copy number variants may serve as promising candidates of functional units that are associated with the development of common complex diseases [213]. Second, complex diseases are usually caused by multiple genes through various complicated biological pathways. Ignoring the complex interactions between genetic variants will likely reduce the power of detecting novel risk factors underlying complex diseases [214]. Third, the genetic etiology of complex diseases is highly heterogeneous [52]. Rare variants may also play an important role in the development of complex diseases.

The three aspects discussed above are investigated in Chapters 3-5. In Chapter 3, a hidden Markov model has been proposed for detecting copy number variants, with an application to a breast cancer study. While applying the method to the phase III data of the study, we detect a number of genomic regions that are associated with breast cancer. The associations of five regions, on chromosome 2, 4, 6, 12, and 13, remain significant in the phase I data of the study. The findings suggest that the structural changes of these genomic regions may contribute to the genetic susceptibility of breast cancer. These findings are consistent with the literature. In Chapter 4, a forward U-test has been proposed for detecting gene-gene interactions, with an application to a cigarette smoking study. While applying the method to SAGE GWAS datasets,

we identify two SNPs with a statistical interaction. The two SNPs are located in gene *CHRNA5* and *NTRK2*. Both genes have been reported for association with cigarette smoking. In Chapter 5, an aggregating U-test has been proposed for detecting functional rare variants, with an application to a quantitative trait study. While applying the method to GAW17 mini-exome data, I have shown this method attains a higher power to detect the association than a commonly used method, QuTie. Overall, these proposed methods have provided powerful tools for genetic association studies. Whereas the findings of this dissertation research are biologically plausible, further research will be necessary to replicate the results.

6.2 Future Development

Statistical methods for genetic association studies focus on detecting the association between genetic variants and the phenotypic traits, measured by either binary disease outcomes or quantitative clinical features. Though a large number of GWASs have been conducted, the genetic etiology of complex diseases remains largely unknown. There are many possible explanations for this challenge. First, complex human diseases usually manifest with multiple sub-phenotypes, representing specific physiological or biochemical processes from various gene pathways. Taking these sub-phenotypes into account is necessary to address disease heterogeneity and to provide novel insights into the genetic etiology of complex diseases. At present, however, our understanding is still limited to describing these sub-phenotypes, and our statistical tools are not powerful enough to detect the functional variants while accounting for the genetic heterogeneity [11,215]. Second, this dissertation research has focused on investigating the genetic effects. However, the effects of genes or genetic variants are commonly modified by environmental risk factors [216]. It is important for future studies to take into account the complex gene-environmental interactions. Third, the large dimensionality of genomic data has

remained a major obstacle for genetic association studies. The power of the association test is usually reduced due to the multiple testing adjustments [217,218]. We are still awaiting sophisticated statistical tools that may have the following properties: 1) able to consider the genetic heterogeneity; 2) able to account for gene-environmental interactions; 3) feasible on a genome-wide scale.

REFERENCES

References

[1] So HC, Gui AH, Cherny SS, Sham PC (2011) Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol.

[2] Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.

[3] Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450.

[4] Howlader N, AM N, M K, Neyman N, R A, W W, SF A, Kosary CL, Ruhl J, Tatalovich Z, et al. (2011) SEER Cancer Statistic Review 1975-2008. SEER web site.

[5] Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlader N, Tatalovich Z, Cho H, et al. (2010) SEER Cancer Statistics Review, 1975-2007. SEER web site.

[6] Mokdad AH, Marks JS, Stroup DF, Gerberding JL (2004) Actual causes of death in the United States, 2000. JAMA 291: 1238-1245.

[7] Mendel G (1965) Experiments in Plant Hybridization. British Medical Journal 1: 370-&.

[8] Pauling L, Itano HA, et al. (1949) Sickle cell anemia, a molecular disease. Science 109: 443.

[9] Conneally PM (2003) The complexity of complex diseases. Am J Hum Genet 72: 229-232.

[10] Chakravarti A (2011) Genomic contributions to Mendelian disease. Genome Res 21: 643-644.

[11] Gibson G (2009) Decanalization and the origin of complex disease. Nat Rev Genet 10: 134-140.

[12] Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K (2007) The burden and costs of chronic diseases in low-income and middle-income countries. Lancet 370: 1929-1938.

[13] Rare_Disease_Act (2002) http://frwebgateaccessgpogov/cgi-bin/getdoccgi?dbname=107_cong_public_laws&docid=f:publ280107.

[14] Lloyd-Jones DM, Nam BH, D'Agostino RB, Sr., Levy D, Murabito JM, Wang TJ, Wilson PW, O'Donnell CJ (2004) Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. JAMA 291: 2204-2211.

[15] Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-266.

[16] Fisher R (1918) The correlations between relatives on the supposition of Mendelian inheritance. Trans R Soc 52.

[17] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.

[18] International-HapMap-Consortium (2003) The International HapMap Project. Nature 426: 789-796.

[19] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

[20] Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27: 234-236.

[21] Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33: 457-458.

[22] Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8: 1229-1231.

[23] Bhangale TR, Stephens M, Nickerson DA (2006) Automating resequencing-based detection of insertion-deletion polymorphisms. Nat Genet 38: 1457-1462.

[24] McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. (2006) Common deletion polymorphisms in the human genome. Nat Genet 38: 86-92.

[25] Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75-81.

[26] Nature-Publishing-Group (2010) Genomics: The tough new variants. Nature 467: 1136.

[27] Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Stahl T, Menzel U, Sandgren J, von Tell D, Poplawski A, et al. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82: 763-771.

[28] Maiti S, Kumar KH, Castellani CA, O'Reilly R, Singh SM (2011) Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. PLoS One 6: e17125.

[29] Bidichandani SI, Ashizawa T, Patel PI (1998) The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. Am J Hum Genet 62: 111-121.

[30] Hammock EA, Young LJ (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308: 1630-1634.

[31] Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72: 1117-1130.

[32] Civeira F (2004) Guidelines for the diagnosis and management of heterozygous familial hypercholesterolemia. Atherosclerosis 173: 55-68.

[33] te Meerman GJ, de Vries EG (2001) Relevance of high and low penetrance. Lancet 358: 331-332.

[34] Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212-219.

[35] Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-510.

[36] Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124-137.

[37] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

[38] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-232.

[39] WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

[40] Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

[41] Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

[42] Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, Cossarizza A, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.

[43] Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med 357: 1199-1209.

[44] Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590-1605.

[45] Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2: 100-109.

[46] Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.

[47] Liu DJ, Leal SM (2010) Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 87: 790-801.

[48] Walsh T, King MC (2007) Ten genes for inherited breast cancer. Cancer Cell 11: 103-105.

[49] Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869-872.

[50] Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592-599.

[51] Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513-516.

[52] McClellan J, King MC (2010) Genetic heterogeneity in human disease. Cell 141: 210-217.

[53] Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85: 309-320.

[54] Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331.

[55] Jasny BR, Kennedy D (2001) The human genome. Science 291: 1153.

[56] Peltonen L, McKusick VA (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291: 1224-1229.

[57] Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409: 853-855.

[58] The_International_HapMap_Consortium (2003) The International HapMap Project. Nature 426: 789-796.

[59] The_International_HapMap_Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

[60] Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

[61] Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5: 16-18.

[62] Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570-575.

[63] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.

[64] Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42: 558-560.

[65] Wan L, Sun K, Ding Q, Cui Y, Li M, Wen Y, Elston RC, Qian M, Fu WJ (2009) Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Res 37: e117.

[66] Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A 105: 4340-4345.

[67] Fentiman IS, Fourquet A, Hortobagyi GN (2006) Male breast cancer. Lancet 367: 595-604.

[68] Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, et al. (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 288: 321-333.

[69] London SJ, Colditz GA, Stampfer MJ, Willett WC, Rosner B, Speizer FE (1989) Prospective study of relative weight, height, and risk of breast cancer. JAMA 262: 2853-2858.

[70] Friedenreich CM (2011) Physical activity and breast cancer: review of the epidemiologic evidence and biologic mechanisms. Recent Results Cancer Res 188: 125-139.

[71] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81: 1879-1886.

[72] Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS (1999) Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 91: 1541-1548.

[73] Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V (1999) Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 91: 1829-1846.

[74] Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, et al. (2007) Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer Inst 99: 1782-1792.

[75] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153-158.

[76] Collaborative-Group-on-Hormonal-Factors-in-Breast-Cancer (2002) Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet 360: 187-195.

[77] Peto J, Mack TM (2000) High constant incidence in twins and other relatives of women with breast cancer. Nat Genet 26: 411-414.

[78] Turnbull C, Rahman N (2008) Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet 9: 321-345.

[79] Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nat Genet 40: 17-22.

[80] Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250: 1684-1689.

[81] Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266: 66-71.

[82] Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 265: 2088-2090.

[83] Wooster R, Weber BL (2003) Breast and ovarian cancer. N Engl J Med 348: 2339-2347.

[84] Kriege M, Brekelmans CT, Boetes C, Besnard PE, Zonderland HM, Obdeijn IM, Manoliu RA, Kok T, Peterse H, Tilanus-Linthorst MM, et al. (2004) Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. N Engl J Med 351: 427-437.

[85] Pasche B (2008) Recent advances in breast cancer genetics. Cancer Treat Res 141: 1-10.

[86] Ripperger T, Gadzicki D, Meindl A, Schlegelberger B (2009) Breast cancer susceptibility: current knowledge and implications for genetic counselling. Eur J Hum Genet 17: 722-731.

[87] Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870-874.

[88] Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41: 579-584.

[89] Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42: 504-507.

[90] Li J, Humphreys K, Heikkinen T, Aittomaki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Martens JW, et al. (2010) A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat.

[91] Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, Kiechle-Bahat M, Bugert P, Schmutzler RK, Bartram CR, et al. (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. Carcinogenesis 28: 1442-1445.

[92] Tchatchou S, Burwinkel B (2008) Chromosome copy number variation and breast cancer risk. Cytogenet Genome Res 123: 183-187.

[93] Wu G, Xing M, Mambo E, Huang X, Liu J, Guo Z, Chatterjee A, Goldenberg D, Gollin SM, Sukumar S, et al. (2005) Somatic mutation and gain of copy number of PIK3CA in human breast cancer. Breast Cancer Res 7: R609-616.

[94] Downing TE, Oktay MH, Fazzari MJ, Montagna C (2010) Prognostic and predictive value of 16p12.1 and 16q22.1 copy number changes in human breast cancer. Cancer Genet Cytogenet 198: 52-61.

[95] Robicsek F (1979) The Smoking Gods: Tobacco in Maya Art, History and Religion. University of Okelahoma Press: 30.

[96] Doll R, Hill AB (1956) Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. Br Med J 2: 1071-1081.

[97] Mokdad AH, Serdula MK, Dietz WH, Bowman BA, Marks JS, Koplan JP (2000) The continuing epidemic of obesity in the United States. JAMA 284: 1650-1651.

[98] Centers-for-Disease-Control-and-Prevention (2004) Cigarette smoking among adults-United States. Morb Mortal Wkly Rep 53: 427-431.

[99] Baillie AJ, Mattick RP, Hall W (1995) Quitting smoking: estimation by meta-analysis of the rate of unaided smoking cessation. Aust J Public Health 19: 129-131.

[100] Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, Murray CJ, Ezzati M (2009) The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Med 6: e1000058.

[101] Kandel D, Schaffran C, Griesler P, Samuolis J, Davies M, Galanti R (2005) On the measurement of nicotine dependence in adolescence: comparisons of the mFTQ and a DSM-IV-based scale. J Pediatr Psychol 30: 319-332.

[102] Sullivan PF, Kendler KS (1999) The genetic epidemiology of smoking. Nicotine Tob Res 1 Suppl 2: S51-57; discussion S69-70.

[103] Li MD, Cheng R, Ma JZ, Swan GE (2003) A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. Addiction 98: 23-31.

[104] Duggirala R, Almasy L, Blangero J (1999) Smoking behavior is under the influence of a major quantitative trait locus on human chromosome 5q. Genet Epidemiol 17 Suppl 1: S139-144.

[105] Bergen AW, Korczak JF, Weissbecker KA, Goldstein AM (1999) A genome-wide search for loci contributing to smoking and alcoholism. Genet Epidemiol 17 Suppl 1: S55-60.

[106] Li MD, Ma JZ, Cheng R, Dupont RT, Williams NJ, Crews KM, Payne TJ, Elston RC (2003) A genome-wide scan to identify loci for smoking rate in the Framingham Heart Study population. BMC Genet 4 Suppl 1: S103.

[107] Straub RE, Sullivan PF, Ma Y, Myakishev MV, Harris-Kerr C, Wormley B, Kadambi B, Sadek H, Silverman MA, Webb BT, et al. (1999) Susceptibility genes for nicotine dependence: a genome scan and followup in an independent sample suggest that regions on chromosomes 2, 4, 10, 16, 17 and 18 merit further study. Mol Psychiatry 4: 129-144.

[108] Gelernter J, Liu X, Hesselbrock V, Page GP, Goddard A, Zhang H (2004) Results of a genomewide linkage scan: support for chromosomes 9 and 11 loci increasing risk for cigarette smoking. Am J Med Genet B Neuropsychiatr Genet 128B: 94-101.

[109] Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, et al. (2007) Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 16: 36-49.

[110] Li MD (2006) The genetics of nicotine dependence. Curr Psychiatry Rep 8: 158-164.

[111] Bateson W (1909) Mendel's Principles of Heredity. Cambridge: Cambridge Press.

[112] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147.

[113] Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 80: 1125-1137.

[114] Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.

[115] Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832-838.

[116] Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638-645.

[117] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

[118] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

[119] International-Human-Genome-Sequencing-Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

[120] Choy KW, Setlur SR, Lee C, Lau TK (2010) The impact of human copy number variation on a new era of genetic testing. BJOG 117: 391-398.

[121] Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, et al. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459: 987-991.

[122] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949-951.

[123] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525-528.

[124] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

[125] Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, et al. (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82: 685-695.

[126] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166-1174.

[127] Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-1148.

[128] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 99: 12963-12968.

[129] Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20: 3413-3422.

[130] Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763-3770.

[131] Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657-663.

[132] Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.

[133] Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013-2025.

[134] Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. Ann Appl Stat 2: 687-713.

[135] Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

[136] Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res 37: 5365-5377.

[137] Frueh FW (2006) Impact of microarray data quality on genomic data submissions to the FDA. Nat Biotechnol 24: 1105-1107.

[138] Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

[139] Brown CK, Madauss K, Lian W, Beck MR, Tolbert WD, Rodgers DW (2001) Structure of neurolysin reveals a deep channel that limits substrate access. Proc Natl Acad Sci U S A 98: 3127-3132.

[140] Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D (2002) Fully automatic quantification of microarray image data. Genome Res 12: 325-332.

[141] Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, et al. (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res 14: 414-425.

[142] Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41: 164-171.

[143] Viterbi AJ (1967) Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. Ieee Transactions on Information Theory It13: 260-+.

[144] Azzato EM, Pharoah PD, Harrington P, Easton DF, Greenberg D, Caporaso NE, Chanock SJ, Hoover RN, Thomas G, Hunter DJ, et al. (2010) A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19: 1140-1143.

[145] Katoh M (2004) Characterization of human ARHGAP10 gene in silico. Int J Oncol 25: 1201-1206.

[146] Park JJ, Kang JK, Hong S, Ryu ES, Kim JI, Lee JH, Seo JS (2008) Genome-wide combination profiling of copy number and methylation offers an approach for deciphering misregulation and development in cancer cells. Gene 407: 139-147.

[147] Aarts M, Dannenberg H, deLeeuw RJ, van Nederveen FH, Verhofstad AA, Lenders JW, Dinjens WN, Speel EJ, Lam WL, de Krijger RR (2006) Microarray-based CGH of sporadic and syndrome-related pheochromocytomas using a 0.1-0.2 Mb bacterial artificial chromosome array spanning chromosome arm 1p. Genes Chromosomes Cancer 45: 83-93.

[148] Shadeo A, Lam WL (2006) Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res 8: R9.

[149] Varma G, Varma R, Huang H, Pryshchepava A, Groth J, Fleming D, Nowak NJ, McQuaid D, Conroy J, Mahoney M, et al. (2005) Array comparative genomic hybridisation (aCGH) analysis of premenopausal breast cancers from a nuclear fallout area and matched cases from Western New York. Br J Cancer 93: 699-708.

[150] Bello MJ, de Campos JM, Vaquero J, Kusak ME, Sarasa JL, Rey JA (2000) High-resolution analysis of chromosome arm 1p alterations in meningioma. Cancer Genet Cytogenet 120: 30-36.

[151] Srivastava M, Bubendorf L, Srikantan V, Fossom L, Nolan L, Glasman M, Leighton X, Fehrle W, Pittaluga S, Raffeld M, et al. (2001) ANX7, a candidate tumor suppressor gene for prostate cancer. Proc Natl Acad Sci U S A 98: 4575-4580.

[152] Ohta T, Okamoto K, Isohashi F, Shibata K, Fukuda M, Yamaguchi S, Xiong Y (1998) T-loop deletion of CDC2 from breast cancer tissues eliminates binding to cyclin B1 and cyclin-dependent kinase inhibitor p21. Cancer Res 58: 1095-1098.

[153] Dutrillaux B, Gerbault-Seureau M, Zafrani B (1990) Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. Cancer Genet Cytogenet 49: 203-217.

[154] Sheng ZM, Marchetti A, Buttitta F, Champeme MH, Campani D, Bistocchi M, Lidereau R, Callahan R (1996) Multiple regions of chromosome 6q affected by loss of heterozygosity in primary human breast carcinomas. Br J Cancer 73: 144-147.

[155] Chappell SA, Walsh T, Walker RA, Shaw JA (1997) Loss of heterozygosity at chromosome 6q in preinvasive and early invasive breast carcinomas. Br J Cancer 75: 1324-1329.

[156] Rodriguez C, Causse A, Ursule E, Theillet C (2000) At least five regions of imbalance on 6q in breast tumors, combining losses and gains. Genes Chromosomes Cancer 27: 76-84.

[157] Noviello C, Courjal F, Theillet C (1996) Loss of heterozygosity on the long arm of chromosome 6 in breast cancer: possibly four regions of deletion. Clinical cancer

research : an official journal of the American Association for Cancer Research 2: 1601-1606.

[158] Hong C, Maunakea A, Jun P, Bollen AW, Hodgson JG, Goldenberg DD, Weiss WA, Costello JF (2005) Shared epigenetic mechanisms in human and mouse gliomas inactivate expression of the growth suppressor SLC5A8. Cancer Res 65: 3617-3623.

[159] Yamanaka S, Sunamura M, Furukawa T, Sun L, Lefter LP, Abe T, Yatsuoka T, Fujimura H, Shibuya E, Kotobuki N, et al. (2004) Chromosome 12, frequently deleted in human pancreatic cancer, may encode a tumor-suppressor gene that suppresses angiogenesis. Lab Invest 84: 1339-1351.

[160] Gupta N, Martin PM, Prasad PD, Ganapathy V (2006) SLC5A8 (SMCT1)-mediated transport of butyrate forms the basis for the tumor suppressive function of the transporter. Life Sci 78: 2419-2425.

[161] Paroder V, Spencer SR, Paroder M, Arango D, Schwartz S, Jr., Mariadason JM, Augenlicht LH, Eskandari S, Carrasco N (2006) Na(+)/monocarboxylate transport (SMCT) protein expression correlates with survival in colon cancer: molecular characterization of SMCT. Proc Natl Acad Sci U S A 103: 7270-7275.

[162] Lerebours F, Bertheau P, Bieche I, Driouch K, De The H, Hacene K, Espie M, Marty M, Lidereau R (2002) Evidence of chromosome regions and gene involvement in inflammatory breast cancer. Int J Cancer 102: 618-622.

[163] Bullrich F, Fujii H, Calin G, Mabuchi H, Negrini M, Pekarsky Y, Rassenti L, Alder H, Reed JC, Keating MJ, et al. (2001) Characterization of the 13q14 tumor suppressor locus in CLL: identification of ALT1, an alternative splice variant of the LEU2 gene. Cancer Res 61: 6640-6648.

[164] Yin Z, Spitz MR, Babaian RJ, Strom SS, Troncoso P, Kagan J (1999) Limiting the location of a putative human prostate cancer tumor suppressor gene at chromosome 13q14.3. Oncogene 18: 7576-7583.

[165] Frank B, Klaes R, Burwinkel B (2005) Familial cancer and ARLTS1. N Engl J Med 353: 313-314; author reply 313-314.

[166] Phelan MC, Rogers RC, Saul RA, Stapleton GA, Sweet K, McDermid H, Shaw SR, Claytor J, Willis J, Kelly DP (2001) 22q13 deletion syndrome. Am J Med Genet 101: 91-99.

[167] Manning MA, Cassidy SB, Clericuzio C, Cherry AM, Schwartz S, Hudgins L, Enns GM, Hoyme HE (2004) Terminal 22q deletion syndrome: a newly recognized cause of speech and language disability in the autism spectrum. Pediatrics 114: 451-457.

[168] Rippe V, Drieschner N, Meiboom M, Murua Escobar H, Bonk U, Belge G, Bullerdiek J (2003) Identification of a gene rearranged by 2p21 aberrations in thyroid adenomas. Oncogene 22: 6111-6114.

[169] Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, et al. (2009) Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS One 4: e7958.

[170] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

[171] Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728-731.

[172] Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84: 148-161.

[173] McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet.

[174] Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955-962.

[175] Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56: 73-82.

[176] Nagel RL (2005) Epistasis and the genetics of human diseases. C R Biol 328: 606-615.

[177] Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S (2006) Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79: 1002-1016.

[178] Lin M, Wu RL (2006) Detecting sequence-sequence interactions for complex diseases. Current Genomics 7: 59-72.

[179] Zhang J, Liang F, Dassen WR, Veldman BA, Doevendans PA, De Gunst M (2003) Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. Am J Hum Genet 73: 1385-1401.

[180] Tzeng JY, Wang CH, Kao JT, Hsiao CK (2006) Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet 78: 231-242.

[181] Li M, Romero R, Fu WJ, Cui Y (2010) Mapping haplotype-haplotype interactions with adaptive LASSO. BMC Genet 11: 79.

[182] Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76: 780-793.

[183] Wei Z, Li M, Rebbeck T, Li H (2008) U-statistics-based tests for multiple genes in genetic association studies. Ann Hum Genet 72: 821-833.

[184] Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006) A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30: 111-123.

[185] Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD (2008) A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. Am J Hum Genet 83: 457-467.

[186] Wu Z, Zhao H (2009) Statistical power of model selection strategies for genome-wide association studies. PLoS Genet 5: e1000582.

[187] Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392-404.

[188] Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413-417.

[189] Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3: e267.

[190] Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701-703.

[191] Lu Q, Elston RC (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 82: 641-651.

[192] Hoeffding W (1948) A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics 19: 293-325.

[193] Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. Journal of Statistical Software 27: 1-25.

[194] Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39: 906-913.

[195] Wu C, Zhang H, Liu X, Dewan A, Dubrow R, Ying Z, Yang Y, Hoh J (2009) Detecting essential and removable interactions in genome-wide association studies. Stat Interface 2: 161-170.

[196] Wang X, Elston RC, Zhu X (2010) The Meaning of Interaction. Hum Hered 70: 269-277.

[197] Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V (2008) Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. Mol Psychiatry 13: 368-373.

[198] Schuckit MA, Danko GP, Smith TL, Bierut LJ, Bucholz KK, Edenberg HJ, Hesselbrock V, Kramer J, Nurnberger JI, Jr., Trim R, et al. (2008) The prognostic implications of DSM-IV abuse criteria in drinking adolescents. Drug Alcohol Depend 97: 94-104.

[199] Stevens VL, Bierut LJ, Talbot JT, Wang JC, Sun J, Hinrichs AL, Thun MJ, Goate A, Calle EE (2008) Nicotinic receptor gene variants influence susceptibility to heavy smoking. Cancer Epidemiol Biomarkers Prev 17: 3517-3525.

[200] Grucza RA, Wang JC, Stitzel JA, Hinrichs AL, Saccone SF, Saccone NL, Bucholz KK, Cloninger CR, Neuman RJ, Budde JP, et al. (2008) A risk allele for nicotine dependence in CHRNA5 is a protective allele for cocaine dependence. Biol Psychiatry 64: 922-929.

[201] Spitz MR, Amos CI, Dong Q, Lin J, Wu X (2008) The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. J Natl Cancer Inst 100: 1552-1556.

[202] Caporaso N, Gu F, Chatterjee N, Sheng-Chih J, Yu K, Yeager M, Chen C, Jacobs K, Wheeler W, Landi MT, et al. (2009) Genome-wide and candidate gene association study of cigarette smoking behaviors. PLoS One 4: e4653.

[203] Schlaepfer IR, Hoft NR, Collins AC, Corley RP, Hewitt JK, Hopfer CJ, Lessem JM, McQueen MB, Rhee SH, Ehringer MA (2008) The CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. Biol Psychiatry 63: 1039-1046.

[204] Weiss RB, Baker TB, Cannon DS, von Niederhausern A, Dunn DM, Matsunami N, Singh NA, Baird L, Coon H, McMahon WM, et al. (2008) A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. PLoS Genet 4: e1000125.

[205] Li MD, Xu Q, Lou XY, Payne TJ, Niu T, Ma JZ (2010) Association and interaction analysis of variants in CHRNA5/CHRNA3/CHRNB4 gene cluster with nicotine dependence in African and European Americans. Am J Med Genet B Neuropsychiatr Genet 153B: 745-756.

[206] Li MD, Yoon D, Lee JY, Han BG, Niu T, Payne TJ, Ma JZ, Park T (2010) Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. PLoS One 5: e12183.

[207] Beuten J, Ma JZ, Payne TJ, Dupont RT, Lou XY, Crews KM, Elston RC, Li MD (2007) Association of specific haplotypes of neurotrophic tyrosine kinase receptor 2 gene (NTRK2) with vulnerability to nicotine dependence in African-Americans and European-Americans. Biol Psychiatry 61: 48-55.

[208] Li MD, Lou XY, Chen G, Ma JZ, Elston RC (2008) Gene-gene interactions among CHRNA4, CHRNB2, BDNF, and NTRK2 in nicotine dependence. Biol Psychiatry 64: 951-957.

[209] Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.

[210] Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188-193.

[211] Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42-54.

[212] Li M, Ye C, Fu W, Elston RC, Lu Q (2010) Detecting gene-gene/gene-environment interactions for quantitative traits with U-Statistics. International Genetic Epidemiology Society (IGES).

[213] McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37-42.

[214] Khoury MJ, Wacholder S (2009) Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am J Epidemiol 169: 227-230; discussion 234-225.

[215] Kent JW, Jr. (2009) Analysis of multiple phenotypes. Genet Epidemiol 33 Suppl 1: S33-39.

[216] Ottman R (1990) An epidemiologic approach to gene-environment interaction. Genet Epidemiol 7: 177-185.

[217] Curtis D (2007) Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association. BMC Genet 8: 49.

[218] Gayan J, Gonzalez-Perez A, Bermudo F, Saez ME, Royo JL, Quintas A, Galan JJ, Moron FJ, Ramirez-Lorca R, Real LM, et al. (2008) A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics 9: 360.