ABSTRACT

WANG, ZHI. Module-based Analysis for “Omics” Data. (Under the direction of Dr. Jung-Ying Tzeng).

This thesis focuses on methodologies and applications of module-based analysis (MBA) in omics studies to investigate the relationships of phenotypes and biomarkers, e.g., SNPs, genes, and metabolites. As an alternative to traditional single–biomarker approaches, MBA may increase the detectability and reproducibility of results because biomarkers tend to have moderate individual effects but significant aggregate effect; it may improve the interpretability of findings and facilitate the construction of follow-up biological hypotheses because MBA assesses biomarker effects in a functional context, e.g., pathways and biological processes. Finally, for exploratory “omics” studies, which usually begin with a full scan of a long list of candidate biomarkers, MBA provides a natural way to reduce the total number of tests, and hence relax the multiple-testing burdens and improve power.

The first MBA project focuses on genetic association analysis that assesses the main and interaction effects for sets of genetic (G) and environmental (E) factors rather than for individual factors. We develop a kernel machine regression approach to evaluate the complete effect profile (i.e., the G, E, and G-by-E interaction effects separately or in combination) and construct a kernel function for the Gene-Environmental (GE) interaction directly from the genetic kernel and the environmental kernel. We use simulation studies and real data applications to show improved performance of the Kernel Machine (KM) regression method over the commonly adapted PC regression methods across a wide range of scenarios. The largest gain in power occurs when the underlying effect structure is

involved complex GE interactions, suggesting that the proposed method could be a useful and powerful tool for performing exploratory or confirmatory analyses in GxE-GWAS.

In the second MBA project, we extend the kernel machine framework developed in the first project to model biomarkers with network structure. Network summarizes the functional interplay among biological units; incorporating network information can more precisely model the biological effects, enhance the ability to detect true signals, and facilitate our understanding of the underlying biological mechanisms. In the work, we develop two kernel functions to capture different network structure information. Through simulations and metabolomics study, we show that the proposed network-based methods can have markedly improved power over the approaches ignoring network information.

Metabolites are the end products of cellular processes and reflect the ultimate responses of biology system to genetic variations or environment exposures. Because of the unique properties of metabolites, pharmcometabolomics aims to understand the underlying signatures that contribute to individual variations in drug responses and identify biomarkers that can be helpful to response predictions. To facilitate mining pharmcometabolomic data, we establish an MBA pipeline that has great practical value in detection and interpretation of signatures, which may potentially indicate a functional basis for the drug response. We illustrate the utilities of the pipeline by investigating two scientific questions in aspirin study: (1) which metabolites changes can be attributed to aspirin intake, and (2) what are the metabolic signatures that can be helpful in predicting aspirin resistance. Results show that the MBA pipeline enables us to identify metabolic signatures that are not found in preliminary single-metabolites analysis.

Module-Based Analysis for “Omics” Data

by
Zhi Wang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2013

APPROVED BY:

_____	_____
Jung-Ying Tzeng Committee Chair	Arnab Maity
_____	_____
Alison Motsinger-Reif	Robert C. Smart

DEDICATION

To my beloved parents, Yunhe Wang and Lina Yu, and my loving wife, Shujun Zhang.

BIOGRAPHY

Zhi Wang was born on June 2, 1987, in Hangzhou, the capital of Zhejiang Province in eastern China. He attended Zhejiang University for undergraduate study in Applied Biological Sciences in 2005. During his undergraduate study, he discovered his interest in Bioinformatics and how its applications on various biological disciplines. In 2008, after 21 years of living in Hangzhou, he came to Unite State and started his graduate study in Bioinformatics at North Carolina State University through an exchange program, which allows his entry to graduate study one year ahead of the schedule. In 2009, he received his Bachelor of Science from Zhejiang University. And one year later, he received the Master of Bioinformatics and Statistics and decided to pursue a Ph.D in Bioinformatics. He will complete his Ph.D in August, 2013, under the direction of Dr. Jung-Ying Tzeng.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my PhD advisor Dr. Jung-Ying Tzeng for her tremendous support and guidance throughout the research of this thesis. Without her enlightened advice and assistance, this work would not have been possible. Working with her has been a fantastic experience. I simple could not wish for a better advisor.

My gratitude goes out as well to my committee members, Dr. Arnab Maity, who provides practical and helpful suggestion for my research problems with great patience, Dr. Alison Motsinger-Reif, who was always available to give me the strongest help whenever I needed, and Dr. Robert Smart, whose strong responsibility and timely help I will never forget.

I would also like to express my gratitude to Dr. Zhaobang Zeng for his guidance of my master study and introducing me to the field of Bioinformatics and Pharmacometabolomics. My collaborators from the Pharmacomatabolomics Research Network also deserve my sincerest thanks. I thank Dr. Rima Kaddurah-Daouk for supporting my research in Pharmacometabolomics and sharing thoughtful biochemical insight with me. I thank Drs. Hongjie Zhu and Anastasia Georgiades for fruitful discussion and their contribution to project.

I am also thankful for all the other faculties, stuffs, post-docs and my fellow graduate students in the Bioinformatics Research Center (BRC) for creating a wonderful study environment. Special thanks for members in Dr. Tzeng's lab, Xin Wang, Jun Hu and Guolin

Zhao, for their help and advice, and Kuangyu Wang, Yuelong Guo, Ronglin Che, Wenjin Lu, Jing Zhao. They are the important source of joy.

Finally, I can't thank them enough, my parents, beloved wife, Shujun Zhang, and the rest of my family, for their love, constant and unconditional support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

Chapter 1 Introduction 1

1.1 Omics studies with features and challenges 1

1.1.1 Genomics, Transcriptomics and Metabolomics 2

1.1.2 Features 5

1.1.3 Challenges 7

1.2 Module based analysis in omics data 8

1.2.1 Module construction 9

1.2.2 Module effects assessment 12

1.3 Dissertations contributions and organization 23

1.3.1 Complete effect-profile assessment in association studies with multiple genetic and multiple environmental factors 23

1.3.2 Module-based association analysis for evaluating effects of biomarkers with Network Structures 24

1.3.3 A module-based pipeline for mining of Pharmacometabolomics data 25

1.3.4 Pharmacometabolomics studies of major depressive disorder (MDD) 26

1.4 References 26

Chapter 2 Complete Effect-Profile Assessment in Association Studies with Multiple Genetic and Multiple Environmental Factors 37

2.1 Abstract 38

2.2 Introduction 39

2.3 Methods 43

2.3.1 GE interaction kernel 44

2.3.2 Score tests for assessing Multi-G-Multi-E effects 47

2.4 Simulation studies 50

2.5 Results 53

2.6 Real data example: application to the CoLaus study data 56

2.7 Discussion 59

2.8 Supplementary note 62

2.9 References 78
Chapter 3 Module-based Association Analysis for Evaluating Effects of Biomarkers with Network Structures 83
3.1 Introduction 84
3.2 Method 87
3.2.1 Kernel machine regression model 87
3.2.2 Kernel functions incorporating network information 88
3.2.3 Kernel functions for interaction effects 91
3.2.4 Testing module effects 92
3.3 Simulation 94
3.3.1 Design 94
3.3.2 Simulation results 96
3.4 Real data application 99
3.5 Discussion 102
3.6 References 114
Chapter 4 A Module Based Pipeline For Pharmacometabolomics Data Analysis 120
4.1 Introduction 120
4.2 Module-based Pipeline 122
4.2.1 Module discovery 122
4.2.2 Module filter (optional) 125
4.2.3 Module testing 127
4.2.4 Module ORA 128
4.2.5 Key metabolites identification 130
4.3 Aspirin Study 130
4.3.1 Identification of metabolic alternations that associate with drug response . 132
4.3.2 Identification of baseline metabolic signatures 134
4.4 Discussion 135
4.5 References 142
Chapter 5 Pharmacometabolomics Studies of Major Depressive Disorder (MDD) 147
5.1 Study of cerebrospinal fluid metabolome in mood disorders-remission state 147

5.2 Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo in patients with major depressive disorder (MDD) 150
5.3 References 157
APPENDICES 160
Appendix A 161
Appendix B 162
Appendix C 167
References 171

LIST OF TABLES

Table 2.1: Haplotype distribution with estimated SNP minor allele frequencies and linkage disequilibrium coefficients..... 70

Table 2.2: Causal SNPs used in the simulation studies..... 71

Table 2.3: Type I error rates averaged over 1000 replicate data sets..... 72

Table 2.4: Testing results from the analysis of the CoLaus Study Data..... 73

Table 3.1: Type I error rates averaged over 1000 replicate data sets..... 105

Table 3.2: Testing results from the *baseline analysis* of the Aspirin Data..... 106

Table 3.3: Testing results from the *differential analysis* of the Aspirin Data..... 107

Table 4.1: Correlation analysis between metabolic changes and drug response..... 137

Table 4.2: Pathway analysis through MetaboAnalyst 2.0..... 138

Table 5.1: Pathway enrichment analysis of the effect of sertraline exposure from baseline to week four..... 153

Table 5.2: Correlations with treatment outcomes: correlations between biochemical..... 154

Table 5.3: A list of metabolites highly associated with Valine in sertraline group..... 155

LIST OF FIGURES

Figure 2.1: Power results for the Joint Test 74

Figure 2.2: Power results for the GE Test 75

Figure 2.3: Power results for the G|E Test 76

Figure 2.4: Power results for the E|G Test 77

Figure 3.1: Modules with sale-free structures in Simulation I 108

Figure 3.2: Modules with Non-scale-free structures in Simulation II 109

Figure 3.3: Power results for Simulation I (scale-free structure) when causal nodes are hub nodes 110

Figure 3.4: Power results for simulation I (scale-free structure) when causal nodes are random nodes 111

Figure 3.5: Power results for simulation II (non-scale-free structure) when causal nodes are hub nodes 112

Figure 3.6: Power results for simulation II (non-scale-free) when causal nodes are random nodes 113

Figure 4.1: The overview of module based pipeline 139

Figure 4.2: Cluster dendrogram of metabolic changes 140

Figure 4.3: Pathway analysis through KEGG Mapper. 141

Figure 5.1: Metabolites quantitated by the LCECA platform 156

Chapter 1

Introduction

1.1 Omics studies with features and challenges

The overall goal of biological research is not only to comprehend the biological systems and processes, but also to understand how the biological processes impact phenotypes. Many biological researches have been conducted based on the reductionist strategy that decomposes a complex system into small parts for analysis (Bechtel and Richardson 1993). Although such strategy has made fruitful achievements in understanding the individual roles of each part, the reductionist strategy does not seek to further integrate different parts and account for overall complexity. This historically dominant research strategy may no longer be reasonable in the omics era because the functions of an integrated biological system are more than sum of functions from individual parts. A system can consist of sophisticated networks with mutual interactions among basic biological units (Strange, 2005). Genome-wide or system-wide analysis of biological units, as opposed to the analysis of single, isolated biological units, can lead to new understanding of functions in various ways. The omics era emerges and opens a window on the biological science.

In this section, we briefly introduce three major omics disciplines, genomics, transcriptomics and metabolomics, focusing on their scientific goals, and discuss common features and challenges in these omic studies.

1.1.1 Genomics, Transcriptomics and Metabolomics

Genomic studies. Genomics aims to decipher information of the complete set of DNA, including determining the whole-genome DNA sequences of organisms, annotating functional information to the sequences, and identifying genetic variants associated with phenotypes. Genome-wide association studies (GWAS), emerged in 2005, have now become a routine screening tool to obtain a list of promising variants. The first generation GWAS were based on the common-disease common-variant (CD-CV) hypothesis and focused on identifying common single nucleotide polymorphisms (SNPs) that are associated with phenotypes (Klein et al., 2005). Since then, numerous studies have been conducted and detected common SNPs variants responsible for common complex diseases such as including type 2 diabetes (Altshuler et al., 2000), hypertension (Doris, 2002) and Alzheimer's (Naj et al., 2011). However, the overall achievements of GWAS are found to be lower than expected (Frazer et al., 2009; Cirulli and Goldstein, 2010), which motivated the shifts from the CD-CV hypothesis to the common disease rare variant (CD-RV) hypothesis. The CD-RV hypothesis states that rare variants have larger effects than the common variants on complex diseases and are the major sources of phenotypic variants (Schork et al., 2009).

As the whole genome sequencing technology evolves and the next generation sequencing (NGS) becomes available, denser genetic markers are able to be genotyped in high efficiency and low cost. The technology improvement permits the genomic interrogations to be extended from common SNPs to rare SNPs and other types of variants including INDELs and copy number variants (CNVs). Several diseases, e.g., epilepsy (Mefford et al., 2010), autism (Pinto et al., 2010), and schizophrenia (Stone et al., 2008), have been found to be associated with rare variants that influence gene functions.

Transcriptomic and prteomic studies. Although genome, as a complete set of DNA, contains all genetic information needed to develop and maintain the organism, it is only one component responsible for the final physical appearance of the organism (i.e., phenotypes). Therefore, researches also focus on the end products of transcriptions (i.e., transcriptome) and the end products of translations (i.e., proteome). As the products of genome, Transcriptome and proteome contain the regulatory information and can serve as indicative of gene function and activity that responsible for the phenotypes, and understanding how they are related to phenotypes can great facilitate the understanding of the biological mechanism and process. One feature that makes the studies of transcriptomics and proteomics more complicated than genomics is that their expressions changes temporally and spatially, which means the performances may differ from time to time and from cell to cell.

cDNA microarray and RNA-seq are two common approaches to profile gene expression profiling (Lee et al., 2000; Wang et al., 2009). The expression of thousands of

genes is measured and can be compared among different samples in order to detect differential expressed genes. Similarly, in proteomics, protein can also be profiled, through immunoarrays, mass spectrometry and various antibody technologies, to identify variations associated with different class of subjects or treatment groups.

Metabolomic studies. Comparing to genomic, transcriptomic and proteomic information, metabolomics, provides the closest link to phenotypes because metabolites are the end products of cellular processes and reflect the ultimate response of biology system to genetic or environment changes (Fiehn, 2002). With the aid of fast development of analytical technologies, such as mass spectrometry (MS), high-resolution nuclear magnetic resonance (NMR) spectroscopy and various compound separation techniques (Dunn and Ellis, 2005; Wishart, 2008), the compound identification becomes a straightforward and fast routine. Metabolomic studies have emerged to be the newest omics sciences, with complementary role with transcriptomics and proteomics, to understand the complex mechanisms under phenotypes.

There are two general approaches for metabolomics study, chemometrix approaches and quantitative approaches (Wishart, 2007), which corresponds to non-targeted profiling and targeted profiling, respectively. One distinction between the two approaches is that chemometric approaches do not require compound identification while quantitative approaches obtain compound identification and quantification before further analysis. Obviously, it is extremely difficult to learn underling biological pathway and its mechanism without knowing compound names. Thus quantitative approaches are much

more popular today with growing demands for deep understanding of complex diseases and related biology system.

1.1.2 Features

Omics studies share three common features. First, since omics studies aim to study the complete set of biological units including genes, proteins and metabolites, they usually deal with a very large amount of variables with relative small sample size. For example, in GWAS, the number of SNPs is typically over one million. In gene expression profiling, DNA microarray can produce about ten thousands gene expression data. In proteomics and metabolomics, a single mass spectrometry experiment can identify over one thousand proteins and metabolites.

Second, the basic biological units form a complex effect mechanism to govern phenotypes. Specifically, the basic units often have “nonlinear” effects rather than simple additive effects on phenotypes due to many complicated biological mechanisms. Moore et al. (2010) defined “nonlinear” in GWAS to be a phenotype that cannot be directly predicted by the sum of individual marker effects. According to this board definition, many phenomena are considered as nonlinear (Thornton-Wells et al. 2004), e.g., locus heterogeneity (i.e., the same phenotype is arisen from different DNA mutations), phenocopy (i.e., phenotypes are completely caused by environmental exposures instead of genetic variants), the gene-environment interactions (i.e., the genetic effects are modified by environmental factors or by other loci).

Third, the omics variables are connected functionally via a complex network. System biology studies (Barabási and Oltvai, 2004) showed that the biological units under omics studies are connected and regulate each other as part of network. In gene regulatory network, target gene and transcriptional factor interact with each other to regulate gene expression and change cell behavior. In protein network, proteins interact with each other through physical interaction to carry out important functions in many molecule processes in the cell such as DNA replication. In metabolic network, metabolites interact with each other through enzymatic conversion to fulfill metabolism and determine cell biochemical properties. This complex network structures can be regarded as one source for the nonlinear effect of biological units discussed above.

Accounting for these features of omics studies could allow us to model the biological effects more precisely, enhance the ability to detect real signals and facilitate our understanding of the underlying biological mechanisms. For example, studying the nonlinear relationship between biological units and phenotype, such as gene-environment interactions and gene-gen interactions, have facilitated the understanding complex disease etiologies [Kraft et al, 2007; Murcary et al, 2009; Thomas, 2010a, 2010b] and the identifications of liability genes that act through interactions but exhibit minimal marginal effects (Thomas, 2010a). In biomarker identification, incorporating network relationship among biological units can enhance the identification efficiency. The structure may help to guide the smoothing and collapsing among biomarkers in the same network neighborhood.

Incorporating network structure information may also increase detectability of real association and uncover potential pathway activities.

1.1.3 Challenges

Along with opportunities, the above mentioned features of omics studies also bring several significant challenges. First, the big volume of data obtained from high-throughput profiling techniques in omics studies encounter challenges in statistical analyses and computational burdens. For example, in a typical GWAS, researchers usually begin with a whole genome scan using single SNP tests. The statistical significance level after multiple testing adjustment becomes extremely stringent and limits the power of the study. The high dimension data also impacts the performance of many classical multivariate methods. For the local methods, e.g., k-nearest neighbor and local regression, which attempt to model in small neighborhoods, will run into problems in high dimension due to very sparse density of samples in the input space (Trevor et al., 2001). For the global methods, e.g., the standard least squares regression, the overcome of high dimension usually requires a sufficient large sample size which is not applicable in practice as features p is larger than the number of observations n ($p>>n$).

Second, modeling nonlinear effects is another challenge in omics data analysis. As we discussed above, nonlinear effects of omics data, especially interactions, is commonly observed in the biological system. While studying these effects can be beneficial in uncovering the disease etiology, traditional parametric methods, such as linear regressions,

require implicit specification of the relationship between basic bio-units and the phenotypes. These methods have limited power when a model is not correctly specified to capture the underlying effects, or encounter the curse of dimensionality when include higher order terms to model complex effects.

The third challenge in omics data for utilizing network information remains in two aspects, network reconstruction and incorporation. For network reconstruction, although network structure can be reconstructed from existing biological knowledge (e.g., KEGG) or from the data (e.g., co-expressed gene modules), enormous number of potential network structures generated by network reconstruction methods tends to provide inconsistent and misleading information. For network incorporation, many methods, such as gene set enrichment analysis, have been developed to incorporate the "membership" of pathway information (i.e., belonging or not to a pathway) and demonstrated their usefulness. However, these methods do not consider the structure information that depicts the specific functional relationships among biological units.

1.2 Module based analysis in omics data

Module based analysis (MBA) aims to evaluate the effect of a group of biomarkers with common features, such as SNPs in the same gene, co-expressed genes, or metabolites involved in the same pathways. Module may serve as a more appropriate analyzing unit for understanding the underlying biological system because most cellular functions are carried out by groups of biomarkers rather than individual biomarkers (Barabási and Oltvai, 2004).

The biomarkers work jointly or even interact with each other within a complex network (Zhu et al., 2007). Consequently, biomarkers tend to have moderate individual effects but significant aggregate effect, and performing analysis at module level can increase the detectability and reproducibility of association findings. By assessing biomarker effects in a functional context, e.g., pathways and biological processes, MBA also improves the interpretability of findings and facilitates the construction of follow-up biological hypotheses. Finally, for exploratory "omics" studies, which usually begin with a full scan of a long list of candidate biomarkers, MBA provides a natural way to reduce the total number of tests, and hence relax the multiple-testing burdens and improve power.

Module based analysis is a two-step procedure including module construction and assessing module effect. We briefly discuss each step in the next sections.

1.2.1 Module construction

Module construction approaches can be generally classified into two types depending on whether or not to use prior biological knowledge. We refer the two types as expert-defined module construction and data-driven module construction approach.

For expert-defined module construction approach, prior biological knowledge, such as expert opinions and resources from abundant literatures and databases, can help to identify groups of biological units before module testing procedure. Take metabolomics studies for example. A nature strategy is to group all metabolites regulated by the same enzyme or belonging to the same pathway into a module. And those modules with

elements that are known to be related to the phenotype of interest, e.g. diseases, are important targets and worth more attention. Although the field of metabolomics is relatively new, there already exist abundant databases storing hundreds of known metabolomic sets and pathways including Human Metabolome Database (HMDB) (Wishart et al., 2009), Metabolic Information Center (MIC), Specialized Metabolic Pathway Databases (SMPDB) (Frolkis et al., 2010), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 1998). Additionally, tens of initial metabolomic signatures have been reported for many diseases as well, including Alzheimer's disease (Han et al., 2002), hypertension (Brindle et al., 2003), motor neuron disease (Rozen et al., 2005), depression (Paige et al., 2006), schizophrenia (Holmes et al., 2006; Kaddurah-Daouk et al., 2007), cardiovascular and coronary artery disease (Sabatine et al., 2005), Huntington's disease (Underwood et al., 2006), subarachnoid hemorrhage (Dunne et al., 2005), preeclampsia (Kenny et al., 2005), type 2 diabetes (Wang et al., 2005; Yuan et al., 2007), liver cancer (Yang et al., 2004), ovarian cancer (Odunsi et al., 2004), and breast cancer (Fang et al., 2005).

Constructing modules based on prior knowledge provides us two major advantages. One is that before testing procedure, more clear and reasonable hypotheses can be formulated depending on different ways to group modules. In particular, constructing gene based module, which means grouping biomarkers related to a gene, reflects assumption about the association between the gene and phenotype. And testing pathway based module allows us to capture the pathway effects. The other advantage is that after testing,

with the aid of known biological context, it greatly facilitates functional interpretation of testing results as well as further hypotheses generation for understanding underlying system. For example, finding a module grouping form the GO consortium (Harris et al. 2004) with clear defined attributes, such as signal transduction, not only implies a direct association between a specific biological process and the phenotype, but also provides great recourses, such as the concrete multi-level information, for understanding the biology system and generating further meaningful hypotheses.

For the data-driven module construction approach, statistical tools rather than knowledge from literatures are used to find modules. Since module construction shares the similar goal with the clustering analysis, which is to classify objects into groups (modules) such that objects within a group are more similar to each other than to those in other groups, many classical clustering algorithms including hierarchical clustering based on different distance matrix and k-means clustering can be used. In chapter 4, we will introduce two clustering algorithms that are frequently used in our studies. Data-driven module construction approach is more favorable in the cases when limited prior knowledge is available to form modules, which is actually very common in practice. For example, in metabolomics studies, a considerable amount of the metabolites usually remain unidentified due to relatively limited size of reference spectral libraries. Therefore, they can only be grouped through statistical approaches.

1.2.2 Module effects assessment

In this section, we will review methods that are available for detecting module association in the context of gene-set analysis and discuss their pros and cons. We note that in this dissertation, modules are not limited to gene sets and can be any predefined sets of biological units in the omics studies.

Over-representation analysis

Over-representation analysis (ORA) is the most common approach to discover association of gene set rather than individual gene. This method begins with a list of genes and each gene has a measure (e.g., P-value) of association. Then, a threshold is selected to separate associated genes from non associated genes. Given predefined and biology-meaningful gene sets that represent pathways of interest for example, these associated genes are examined to see whether most of them gather into a few pathways rather than being split up into several pathways. In the other words, the researcher looks for whether any pathway is overrepresented by the associated genes. To examine this type of association, Fisher's exact test based on the hypergeometric distribution, or its approximation chi-square test for large sample is commonly used. In the simplest case, for instance, a 2*2 contingency table can be calculated from 4 observed values including the number of top genes, total number of genes, the number of genes in the pathway of interest, and the number of genes existing both in the list and pathway. Then a chi-square test with 1 degree of freedom can be used to test the significance of association.

This approach is widely used because its simplicity and reasonability. Many authors have presented it with minor differences (Khatri and Draghici, 2005). However it has three major drawbacks. First, the chosen of top genes is based on an arbitrary threshold, which distinguishes associated genes from the rest. It's demonstrated that different thresholds may lead to inconsistent results. Besides, the arbitrary threshold also prevents many causal genes with moderate level of association from being used to identify significant gene sets. Second, genes with different levels of association are equally treated. It causes loss of information, especially, when the levels have a wide range. Third, this method doesn't consider the correlation structure of metabolite sets, which is important for estimating statistical significance and increasing power.

Gene set enrichment analysis (GSEA)

An alternative and more successful technique, gene set enrichment analysis (GSEA), is developed to overcome limitations of ORA, with considering the distribution of gene sets of interest in the entire list and the correlation structure. The original version of GSEA was published by Mootha et al. (2003). Later, in 2005, a revised and generalized version for analyzing molecular profiling data was introduced (Subramanian et al., 2005) with following major steps; first, a list of genes is ranked based on association between their expression and phenotype measured by any reasonable statistical methods or metrics; second, given a priori defined set or pathway of genes, an enrichment score (ES) is calculated indicating the degree of overrepresentation of a set at top or bottom of the entire ranked gene list. In

particular, Instead of just counting the number of set involved genes at the top of the list, ES is calculated by walking down the entire ranked list. When encountering a gene in set, the ES is increased. Otherwise, ES is decreased. The increasing amount depends on the previous calculated correlation of the gene with the phenotype. This enrichment score essentially corresponds to a weighted Kolmogorov-Smirnov-like statistic (Hollander and Wolfe, 1999); third, a null distribution of ES is generated based on the permutation of phenotype labels. And the empirical, nominal P value is then calculated; finally, the adjustment is applied to account for multiple testing of an entire database of gene sets. The size of the set has been accounted which leads to a normalized enrichment score (NSE). False discovery rate (FDR) (Benjamini et al., 2001) is estimated based on each NSE.

The revised version outperforms their preliminary version based on several modifications. In the original method, the ES is calculated by equal weighting. Thus, high scores came out for sets gathered near the middle of the ranked list (Subramanian et al., 2005), which indicates that Information of relationship with phenotype was underestimated because top genes in the list should represent the closest relationship with phenotype. Another change is the adjustment of multiple testing. Since the primary goal is hypotheses generation, it costs too much to prohibit every single false-positive set using family wise-error rate (FWER) (Benjamini and Hochberg, 1995), which results in very conservative outcomes. Therefore, FDR is implemented.

Several methods similar to GSEA have also been developed based on different enrichment scores. For example, Efron and Tibshirani (2007) developed a Gene Set Analysis (GSA) using a "maxmean" statistics and Smythe (2004) and Tian et al. (2005) introduced the averaged t-statistic for the enrichment score. Moreover, with the specific aim of discovering differential expression sets, PAGE, known as parametric analysis of gene set enrichment (Kim and Volsky, 2005), is developed for gene set enrichment analysis. It calculated fold change for genes between experimental groups at the beginning. Then, Z scores based on fold change are produced for each predefined gene sets as: $Z = (S_m - \mu) \times m^{\frac{1}{2}}/\delta$, where S_m is the mean of fold change values of genes within predefined gene set, μ is the mean of total fold change values, m is the size of a given gene set and δ is the standard deviation of total fold change values. Normal distribution is presumed for statistical inference based on Z score. Since PAGE doesn't need permutation distribution, it requires much less computational effort. Additionally, simulations in their study showed PAGE was statistically more sensitive than GSEA.

Combined statistic

In whole-genome studies, another type of method aiming to detect overall gene-set effect is based on the combination of p-values obtained from individual tests of genes, such as the minimal p-value methods (De la Cruz et al., 2010). These methods first construct combined statistics, and then calculate set-specific p-values. The combined statistic is obtained by multiplying the individual p-values in the gene set, or equivalently, addition of negative logs

of individual p-values. In order to improve the power, two concepts can be incorporated into to the combined statistic: truncation and weighting. Truncation indicates pre-selection of genes based on their p values, for example, using all genes with p values below a threshold or top genes with smallest p values. Weighting indicates the weight given to each p value when combing them to the combined statistic. Since, in practice, the assumption that each gene in the set is exchangeable is always violated, these weights can be used to include the prior information about the correlation of these genes or their relative functional importance. After the combined statistic is calculated, the set-specific p-value is obtained through permutation. For the details of implementation, see De la Cruz et al. (2010).

There are many variants under this category depending on different ways to construct the combined statistics using different statistics. For example, SAMGS (Dinu et al., 2007) is created to test the null hypothesis that the expression of genes in a gene set does not differ by the phenotype of interest, e.g., case and control. For an individual gene set, it's essentially based on individual t-like statistics, called d statistic. For each gene i, the d statistic is calculated: $d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$, where $\bar{x}_1(i)$ and $\bar{x}_2(i)$ are defined as the average levels of expression for gene i within different groups 1 and 2, and $s(i)$ is a pooled standard deviation over the groups of the phenotype, s_0 is a small positive constant that adjusts for the small variability. SAMGS then summarizes all genes in the gene set S by adding $d(i)$'s together: $SAMGS = \sum_{i=1}^{|S|} d_i^2$. A null distribution of SAMGS is generated based on the

permutation of phenotype labels. And the P value is then calculated. In case of multiple gene sets, q-values based on false discovery rates (FDRs) can be used in SAMGS.

One major disadvantage of approaches based on combined statistics is they only use all marginal effects of genes without account for the interaction effects within the gene set.

Methods using linear models

Followings are another set of approaches, which don't rely on the individual gene tests, to jointly model the effect of all genes within a set.

One simple implementation to assess the gene-set effect is using a traditional multiple regression. Regressing phenotype (Y) on genes within given set. Then a joint test of the null hypothesis H0: $\beta_1 = \beta_2 = \cdots \beta_m = 0$, reflects the association of the phenotype with given set, where m is the number of genes within given set, and β is regression coefficient. The alternative hypothesis is at least one of the coefficients differs from zero. However, in most cases, it's unable to test the hypothesis because m might be large relative to the sample size. So the global test using random effects to fit genes within a given set is proposed by Goeman et al. (2004). Using the random linear model, Y is modeled as $Y|X \sim N(X\beta, \sigma^2)$, where β indicates set effects following some common distribution with expectation zero and variance τ^2. The null hypothesis remains the same and $\tau^2 = 0$ is tested by a score test. This method can be extended to include covariates and/or other phenotypic types such as binary and survival under generalized mixed linear model framework.

When the number of genes in gene set exceeds the sample size, principal component based regression (PCR) is another commonly used approach (Tomfohr et al., 2007). It's a two-stage approach. First, principal component analysis (PCA) is applied on target gene set to generate principal components that are linear combinations of the original variables. In common practice (Chen et al., 2008; Ma and Kosorok, 2009), either the first principle component or components exceeding certain threshold in terms of explained variance is chosen. Other rules are also developed for choosing the number of components, p, including Zhu's method (Zhu and Ghodsi, 2006), Average Eigenvalue Rules (Peres-Neto et al., 2005) and Bartlett's test (Bartlett, 1954; Jackson, 1993). Next, these reduced components are fitted in a regression model. Similar hypothesis testing for the effect of all the principle components, which indicates association between gene set and phenotype, is performed. One advantage of PCR is that it solves the m>n problem by choosing p where p<<m. Another advantage is when strong correlations exist among genes (predictor variables), PCR achieves better performance than ordinary linear regression.

One extension of PCR is Partial least square regression (PLS) (Geladi and Kowalski, 1986). PLS is one-stage procedure that incorporates Y (response) information to find new variance components that maximizing $Corr(X, Y)$ rather than $Var(X)$. Besides PLS designed for quantitative phenotype, partial least square discriminant analysis (PLSDA) (Pérez-Enciso and Tenenhaus, 2003) is developed for binary outcomes. Orthogonal projection to latent structures (OPLS) (Trygg and Wold, 2002) is another extension by

removing variance in the X matrix not correlated to Y. However, some studies (Tapp and Kemsley, 2009) showed OPLS methods will never outperform PLS methods in terms of prediction. PCA and PLS based methods is the most widely used approaches in metabolomics data analysis. And SIMCA-P (Umetrics) implementing these methods with a very nicely designed graphical user interface and visualizations is also becoming the most widely used tool in metabolomics community.

Both global test of Goeman et al. (2004) and dimension reduction based methods introduce above assume the gene effects are additive, which indicates that they have limited capability of modeling nonlinear gene effects including interaction within a gene set. Kernel machine based methods (Liu et al., 2007; Kwee et al., 2008; Wu et al., 2010) that allow for flexible modeling nonlinear effects of genes are developed. Different kernel functions represent different ways to aggregate genes information. For instance, the first polynomial kernel, captures only all the main effects while the second polynomial kernel captures all the main effects, all two way interactions and quadratic main effects of genes within a set. KM regression does not suffer from the burden of dimensionality like widely used OLS regression, less information is lost to data-reduction compared to PC regression, and kernels can be constructed to incorporate important biological information (Kwee et al., 2008; Wu et al., 2010). In addition, KM regression can handle correlated factors like PC regression and allows for the investigation of non-linear effects, which is not practical in OLS regression and not feasible in PC regression.

Discussion

Numerous methods have been developed for testing the association of the gene set with phenotypes of interest. However, according to Tian et al. (2005), two distinctive null hypotheses are formulated. The first null hypothesis (Q1) is assuming the genes in the gene set show the same pattern of associations with the phenotype compared with the rest of the genes. The second hull hypothesis (Q2) is assuming no genes in the gene set associated with phenotype, which doesn't consider the genes outside the gene set. Geoman and Buhlmann (2007) termed methods based on Q1 and Q2 as competitive and self-contained methods. According to this classification, ORA and PAGE are competitive methods while combined statistic methods including SAMGS and methods based on linear models are self-contained methods. Because of the strong null hypothesis of self-contained methods, only a single significantly associated gene (SAG) can cause the whole gene set significant. Consequently, they usually have more power than competitive test and can provide very powerful predictions. However, this significant gene set may not be really enriched with SAGs. Therefore, Nam and Kim (2008) confirmed a ground rule for choosing gene set methods. If the goal is to find gene sets enriched with SAGs, a competitive method fits better. Otherwise, if the goal is to find gene sets associated with phenotype, a self-contained method should be used.

From above description of Q1 and Q2, competitive methods and self-contain methods only validate the association of a single gene set with phenotype. Thus, GSEA is a

different type of method as it tests another kind of hypothesis that there is no gene sets associated with the phenotype (Q3) (Nam and Kim, 2008). It targets the entire dataset and tests if any gene set in the whole dataset is associated with phenotype. In the GSEA, the enrichment score (ES) for each gene set is a competitive statistic as it summarized the relative enrichment considering genes both within predefined gene set and the rest of genes. Then, the step of sample permutation to test the significant of the entire dataset is self-contained as it assumes no gene sets in the entire dataset (set of gene sets) associated with phenotype.

There are several papers trying to compare the performances of different gene-set methods. Dinu et al. (2008) compared five self-contained gene set methods including SMAGS and Goeman et al.'s Global random effect test with GSEA and concluded SAMGS and Global test outperformed GSEA. Liu et al. (2007) compared the Global test, SAMGS and the ANCOVA Global test (Mansmann and Meister, 2005) and concluded SAMGS had slight better performance than the Global test and ANCOVA Global test. Feidly et al. (2010) investigated ten gene-set methods under a wide range of scenarios and reported Fisher's method and the global test had the highest power while PCR and methods based on Kolmogorow-Simirnov test tended to have lowest power. Nam and Kim (2008) reviewed a number of methods and highly recommended GSEA as it is a mixed approach to avoid some drawbacks of the other methods. However, since all studies and simulations are not comprehensive, no general conclusions can be made on these methods.

One extension of traditional gene-set methods can be achieved by incorporating network structure information in order to increase the efficiency in estimation and inference on gene sets. The sensitivity of detecting relevant gene sets/pathways can be improved by incorporating pathway topology information, demonstrated by Rahnenhurer et al. (2004). SAFE, significance analysis of function and expression, presented by Barry et al. (2005), is a permutation based procedure and considers the underlying network structure. Wei and Li (2007) proposed a Markov random field model incorporating the information on the gene network in the analysis. In proteomic data, Sanguinetti et al. (2008) presented a mixture model on graphs (MMG) to account for network information and used a simple percolation algorithm to search for sub-networks of significant components. Shojaie and Michailidis (2009) used a latent variable model to incorporate network information and test whether a priori defined gene sets are differentially expressed by mixed linear models (MLM).

Another interesting extension can be made is to consider multiple gene sets and their interaction to compensate for the fact that most gene-set methods only have concentrated on detecting effects for a single gene set. It is a very demanding and promising task as most functional units, such as genes and pathways, work together and interact with each other in biology system. For example, epistasis is a widely observed phenomenon that two or more genes have interaction effects rather than additive effects.

Thus, identifying gene-set interaction and its association with complex phenotype is crucial for characterizing underlying biological mechanisms in system biology.

1.3 Dissertations contributions and organization

1.3.1 Complete Effect-Profile Assessment in Association Studies with Multiple Genetic and Multiple Environmental Factors

Studying complex human diseases in the post genome wide association studies (GWAS) era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors (here referred to as Multi-G-Multi-E studies), and mining GWAS for potential gene-environment (GE) interactions is proving to be an invaluable aid not only in discovery but in helping investigators decipher underlying biological pathways and mechanisms. Approaches currently available for examining effect profiles in Multi-G-Multi-E analyses are either underpowered (e.g. naïve regression), ill-suited for studying GE interactions (principal component (PC) regression methods), or do not provide the flexibility of examining the complete effect structure (e.g. existing kernel machine (KM) regression methods). In the chapter 2 we propose a KM regression approach that directly constructs a kernel for the GE interaction based on the genetic kernel and the environmental kernel. We also construct a series of score tests to evaluate the complete effect profile (i.e. the G, E, and GE effects individually or in combination). Simulation studies show that the proposed KM regression method outperforms or performs comparably to PC regression methods

currently available across a wide range of scenarios, including varying effect size, mean model structure, and interaction complexity, which could be encountered in a Multi-G-Multi-E setting. The largest gain in power was observed when the underlying effect structure was involved complex GE interactions, suggesting that the proposed method could be a useful and powerful tool for performing exploratory or confirmatory analyses in GE-GWAS.

1.3.2 Module-based Association Analysis for Evaluating Effects of Biomarkers with Network Structures

In Chapter 3, we apply the proposed kernel machine regression framework to metabolomics study. In the section 1.1.2, we mentioned that the network structure, which contains the important information of interplay among basic functional units, is universal and inherent in biology system, and incorporating this information is a non-trivial task and can be utilized to increase the power of detecting real association. However, we have reviewed a long list of popular module-based approaches in section 1.2.2, and observed that only a few of them have been developed to utilizing network structure information. This motivates the research in Chapter 4, where we incorporate the network information into the kernel machine regression framework. In contrast to the existing methods, which are formulated as a variable selection problem by either specifying a network-constrained penalty function or incorporating prior distribution based on network structure information to evaluate the effects of a single module, our proposed framework have concentrated on evaluating the

effects of multiple modules and their interactions with incorporating the network structure information of each module. Specifically, we developed two versions of network kernels through a unified procedure to incorporate the network structure information from different aspects. Simulation studies and an analysis of metabolic data demonstrate that our proposed methods with network information can have markedly improved power over the approach ignoring network information for investigating module effects as well as their interactions.

1.3.3 A Module-based Pipeline for Mining of Pharmacometabolomics Data

Pharmacometabolomics aims to explain the interpersonal variation of pharmaceutical compounds (drugs), referred as drug response phenotype, and to enhance understanding of mechanisms of drugs in terms of the global metabolic profile. As a new discipline that stems from metabolomics, its data analysis shares major challenges outlined in Section 1.1.3 with other omics disciplines. However, single metabolite analysis, which is served as standard approach and most commonly used in pharmacometabolics studies, fails to meet these challenges. In Chapter 4, based on recent developed statistical methods and our experience with metabolomics studies, we propose a module-based pipeline for in-depth mining of pharmacometabolomics data. A real data example of aspirin study has been shown to demonstrate the usage of our pipeline on two major applications of

pharmacometabolomics including identification of drug-related alternations in metabolic pathways and baseline drug response predictive signatures.

1.3.4 Pharmacometabolomics Studies of Major Depressive Disorder (MDD)

In Chapter 5, we introduce two pharmacometabolomics studies related to major depressive disorder (MDD), in which we have been involved and identified preliminary biological signatures for the disease and drug response mechanisms.

1.4 References

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., ... & Lander, E. S. (2000). The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature genetics*, *26*(1), 76-80.

Fridley, B. L., Jenkins, G. D., & Biernacka, J. M. (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, *5*(9), e12693.

Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2), 101-113.

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, *21*(9), 1943-1949.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 296-298.

Bechtel, W., Richardson, R. C., & Sloan, P. R. (1994). Discovering complexity: Decomposition and localization as strategies in scientific research. *ISIS-International Review Devoted to the History of Science and its Cultural Influence*, *85*(4), 746-746.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, *125*(1), 279-284.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Brindle, J. T., Nicholson, J. K., Schofield, P. M., Grainger, D. J., & Holmes, E. (2003). Application of chemometrics to 1H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst*, *128*(1), 32-36.

Brindle, J. T., Antti, H., Holmes, E., Tranter, G., Nicholson, J. K., Bethell, H. W., ... & Grainger, D. J. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature medicine*, *8*(12), 1439-1445.

Chen, Y. H., Chatterjee, N., & Carroll, R. J. (2008). Retrospective analysis of haplotype-based case–control studies under a flexible model for gene–environment association. *Biostatistics*, *9*(1), 81-99.

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*,11(6), 415-425.

Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, *63*(4), 1079-1088.

De la Cruz, O., Wen, X., Ke, B., Song, M., & Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genetic epidemiology*, *34*(3), 222-231.

Dinu, I., Liu, Q., Potter, J. D., Adewale, A. J., Jhangri, G. S., Mueller, T., ... & Yasui, Y. (2008). A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data.*Cancer informatics*, *6*, 357.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., ... & Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics*, *8*(1), 242.

Doris, P. A. (2002). Hypertension genetics, single nucleotide polymorphisms, and the common disease: common variant hypothesis. *Hypertension,* 39(2), 323-331.

Dunn, W.B. & Ellis, D.I. (2005). Metabolomics: current analytical platforms and methodologies. *Trends Analyt. Chem,* 24, 285–294.

Dunne, V. G., Bhattachayya, S., Besser, M., Rae, C., & Griffin, J. L. (2005). Metabolites from cerebrospinal fluid in aneurysmal subarachnoid haemorrhage correlate with vasospasm and clinical outcome: a pattern-recognition 1H NMR study. *NMR in Biomedicine*, *18*(1), 24-33.

Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics,* 107-129.

Fan, X., Ba, J., & Shen, P. (2006, January). Diagnosis of breast cancer using HPLC metabonomics fingerprints coupled with computational methods. In*Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the* (pp. 6081-6084). IEEE.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87-94.

Fisher, S. R. A. (1970). *Statistical methods for research workers* (Vol. 14, pp. 140-142). Edinburgh: Oliver and Boyd.

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics,* 10(4), 241-251.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., ... & Wishart, D. S. (2010). SMPDB: the small molecule pathway database. *Nucleic Acids Research,* 38(suppl 1), D480-D487.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica acta,* 185, 1-17.

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, *23*(8), 980-987.

Goeman, J. J., Van De Geer, S. A., De Kort, F., & Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, *20*(1), 93-99.

Han, X., M Holtzman, D., W McKeel, D., Kelley, J., & Morris, J. C. (2002). Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: Potential role in disease pathogenesis. *Journal of neurochemistry*, *82*(4), 809-818.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., ... & Gwinn, M. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research,* 32(Database issue), D258-61.

Hollander, M. & Wolfe, D. A. (1999). Nonparametric Statistical Methods. New York: Wiley. Series B (Methodological) ,57 (1), 289–300.

Holmes, E., Tsang, T. M., Huang, J. T. J., Leweke, F. M., Koethe, D., Gerth, C. W., ... & Bahn, S. (2006). Metabolic profiling of CSF: evidence that early intervention may impact on disease progression and outcome in schizophrenia. *PLoS medicine*, *3*(8), e327.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology,* 2204-2214.

Kaddurah-Daouk, R., McEvoy, J., Baillie, R. A., Lee, D., Yao, J. K., Doraiswamy, P. M., & Krishnan, K. R. R. (2007). Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Molecular psychiatry*, *12*(10), 934-945.

Kanehisa, M. & Goto, S. KEGG, (1998). Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28, 27– 30.

Kenny, L. C., Dunn, W. B., Ellis, D. I., Myers, J., Baker, P. N., & Kell, D. B. (2005). Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics*, *1*(3), 227-234.

Khatri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, *21*(18), 3587-3595.

Kim, S. Y., & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, *6*(1), 144.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. Science, 308(5720), 385-389.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., & Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human heredity*, *63*(2), 111-119.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., & Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, *82*(2), 386-397.Lee, M. L. T., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences,* 97(18), 9834-9839.

Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics,* 24(9), 1175-1182.

Liu, Q., Dinu, I., Adewale, A., Potter, J., & Yasui, Y. (2007). Comparative evaluation of gene-set analysis methods. *BMC bioinformatics*, *8*(1), 431.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, *86*(6), 929-942.

Ma, S., & Kosorok, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics*, *25*(7), 882-889.

Mansmann, U., & Meister, R. (2005). Goeman's global test versus an ANCOVA approach. *Methods Inf. Med*, *44*, 449-453.

Mefford, H. C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., ... & Eichler, E. E. (2010). Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genetics,* 6(5), e1000962.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics,* 26(4), 445-455.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., ... & Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics,* 34(3), 267-273.

Murcray, C. E., Lewinger, J. P., & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, *169*(2), 219-226.

Naj, A. C., Jun, G., Beecham, G. W., Wang, L. S., Vardarajan, B. N., Buros, J., ... & DeCarli, C. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics,* 43(5), 436-441.

Nam, D., & Kim, S. Y. (2008). Gene-set approach for expression pattern analysis. *Briefings in bioinformatics*, *9*(3), 189-197.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, *48*(1-2), 155-171.

Odunsi, K., Wollman, R. M., Ambrosone, C. B., Hutson, A., McCann, S. E., Tammela, J., ... & Alderfer, J. L. (2005). Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *International Journal of Cancer*,*113*(5), 782-788.

Paige, L. A., Mitchell, M. W., Krishnan, K. R. R., Kaddurah-Daouk, R., & Steffens, D. C. (2007). A preliminary metabolomic analysis of older adults with and without depression. *International journal of geriatric psychiatry*,*22*(5), 418-423.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis,* 49(4), 974-997.

Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics*, *112*(5-6), 581-592.

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., ... & Green, A. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature,* 466(7304), 368-372.

Rahnenfuhrer, J., Domingues, F. S., Maydt, J., & Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1055.

Rozen, S., Cudkowicz, M. E., Bogdanov, M., Matson, W. R., Kristal, B. S., Beecher, C., ... & Kaddurah-Daouk, R. (2005). Metabolomic analysis and signatures in motor neuron disease. *Metabolomics*, *1*(2), 101-108.

Sabatine, M. S., Liu, E., Morrow, D. A., Heller, E., McCarroll, R., Wiegand, R., ... & Gerszten, R. E. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, *112*(25), 3868-3875.

Sanguinetti, G., Noirel, J., & Wright, P. C. (2008). MMG: a probabilistic tool to identify submodules of metabolic pathways. *Bioinformatics*, *24*(8), 1078-1084.

Shojaie, A., & Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, *16*(3), 407-426.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol,* 3(1), 3.

Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H., Corvin, A., ... & Kwan, S. L. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* 455(7210), 237-241.

Strange, K. (2005). The end of "naïve reductionism": rise of systems biology or renaissance of physiology?. *American Journal of Physiology-Cell Physiology*, *288*(5), C968-C974.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* 102(43), 15545-15550.

Thomas, D. (2010)a. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, *11*(4), 259-272.

Thomas, D. (2010)b. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual review of public health*, *31*, 21.

Thornton-Wells, T. A., Moore, J. H., & Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *TRENDS in Genetics,* 20(12), 640-647.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., & Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America,* 102(38), 13544-13549.

Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics*, *6*(1), 225.

Trevor. Hastie, Robert. Tibshirani, & Friedman, J. J. H. (2001). *The elements of statistical learning* (Vol. 1). New York: Springer.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, *16*(3), 119-128.

Underwood, B. R., Broadhurst, D., Dunn, W. B., Ellis, D. I., Michell, A. W., Vacher, C., ... & Rubinsztein, D. C. (2006). Huntington disease patients and transgenic mice have similar pro-catabolic serum metabolite profiles. *Brain*,*129*(4), 877-886.

Wang, C., Kong, H., Guan, Y., Yang, J., Gu, J., Yang, S., & Xu, G. (2005). Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. *Analytical Chemistry*, *77*(13), 4108-4116.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63.

Wei, Z., & Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, *23*(12), 1537-1544.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., ... & Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research,* 37(suppl 1), D603-D610.

Wishart, D.S. (2007). Current Progress in computational metabolomics. *Brief. Bioinform,* 8, 279–293.

Wishart, D.S. (2008). Quantitative metabolomics using NMR. *Trends Analyt. Chem,* 27, 228–237.

Yang, J., Xu, G., Zheng, Y., Kong, H., Pang, T., Lv, S., & Yang, Q. (2004). Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *Journal of Chromatography B*, *813*(1), 59-65.

Yuan, K., Kong, H., Guan, Y., Yang, J., & Xu, G. (2007). A GC-based metabonomics investigation of type 2 diabetes by organic acids metabolic profile. *Journal of Chromatography B*, *850*(1), 236-240.

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, *51*(2), 918-930.

Zhu, X., Gerstein, M., & Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & development*, *21*(9), 1010-1024.

Chapter 2

Complete Effect-Profile Assessment in Association Studies with Multiple Genetic and Multiple Environmental Factors

Zhi Wang[1,*], Megan L. Neely[2,*], Arnab Maity[3], Jung-Ying Tzeng[1,3,4]

1: Bioinformatics Research Center, North Carolina State University, Raleigh NC, 27695, USA

2: Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, 27705, USA

3: Department of Statistics, North Carolina State University, Raleigh NC, 27695, USA

4: Department of Statistics, National Cheng -Kung University, Taiwan, R.O.C.

*: Equal contribution

2.1 Abstract

Studying complex human diseases in the post genome wide association studies (GWAS) era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors (here referred to as Multi-G-Multi-E studies), and mining GWAS for potential gene-environment (GE) interactions is proving to be an invaluable aid not only in discovery but in helping investigators decipher underlying biological pathways and mechanisms. Approaches currently available for examining effect profiles in Multi-G-Multi-E analyses are either underpowered (e.g. naïve regression), ill-suited for studying GE interactions (principle component (PC) regression methods), or do not provide the flexibility of examining the complete effect structure (e.g. existing kernel machine (KM) regression methods). In this work, we propose a KM regression approach that directly constructs a kernel for the GE interaction based on the genetic kernel and the environmental kernel. We also construct a series of score tests to evaluate the complete effect profile (i.e. the G, E, and GE effects individually or in combination). Simulation studies show that the proposed KM regression method outperforms or performs comparably to PC regression methods currently available across a wide range of scenarios, including varying effect size, mean model structure, and interaction complexity, which could be encountered in a Multi-G-Multi-E setting. The largest gain in power was observed when the underlying effect structure was involved complex GE interactions, suggesting that the proposed method could be a useful and powerful tool for performing exploratory or confirmatory analyses in GE-GWAS.

KEYWORDS: factor-set association analysis; kernel machine regression; GWAS; genetic-environmental interactions; joint and conditional tests**.**

2.2 Introduction

Complex human diseases are influenced not only by genetic and environmental factors but also by the interplay between the two. Investigating gene-environment (GE) interactions can facilitate understanding the etiology of these phenotypes by providing insight into biological mechanisms associated with diseases [Murcray et al., 2009; Mechanic et al., 2012], by explaining heterogeneity across populations, by classifying risk subgroups based on differential environmental exposures [Kraft et al., 2007; Murcary et al., 2009; Thomas, 2010a, 2010b], and by identifying novel genes acting through interactions but exhibiting minimal marginal effects [Thomas, 2010a].

Studying complex diseases in the post genome wide association studies (GWAS) era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors (referred to in this text as Multi-G and Multi-E methods, respectively) [Wu et al., 2010; Pongpanich et al., 2012]. Much of the work in the recent literature has focused on Multi-G approaches. By pooling information across a set of genetic markers, these methods have improved power to detect association signals missed by individual SNP analyses [e.g., Ballard et al., 2010; Bansal et al., 2010]. However, the need for Multi-G-Multi-E methods is becoming apparent as investigators have turned their attention to mining GWAS data for potential GE interactions. This shift in research focus is evidenced

by the emergence of Multi-G-Multi-E studies in the recent literature [Wu et al., 2012; Dai et al. 2013; Edwards et al., 2013; Naj et al., 2013; Patel et al., 2013].

Approaches currently available for examining effect profiles in Multi-G-Multi-E analyses include naïve regression, principal component (PC) regression, and kernel machine (KM) regression. Naïve regression investigates the relationship between the genetic and environmental factor-sets and the response by treating each element of the factor-sets as an individual covariate in a regression model and then performing significance testing. PC regression first performs principal component analysis on the genetic and environmental factor-sets and then uses a subset of the resulting principle components (e.g., the first PC or the top PCs that explain 80% of the variation in the set) in a regression analysis [Gauderman et al., 2007; Wang et al., 2009]. Investigating GE interactions can easily be incorporated into these frameworks, but there are several potential drawbacks to using these approaches in the Multi-G-Multi-E setting. Naïve regression is straight forward and easy to apply, but is usually under powered due to the large number of predictors and the small effect sizes of individual factors [Wu et al., 2010; Lin et al., 2011]. While PC regression typically has better power than naïve regression, its application can be subjective when deciding how many PCs to include in the downstream analyses and valuable information can be lost. Moreover, naïve and PC regression do not incorporate any prior information about the elements of the factor-sets.

KM regression has been shown to overcome the drawbacks of naïve and PC regression. KM regression first computes pairwise similarities between subjects based on their covariate values (e.g., genetic and environmental factors) using a pre-specified kernel

function and then performs least squares regression of the response on the similarity measures. Thus, if the number of parameters in the model is more than the number of subjects, KM regression can be thought of as a dimension reduction technique. Even if the actual number of variables (e.g., the number of genetic and environmental factors) is less than the sample size, the number of parameters can still exceed the sample size if higher order main or interaction effects are considered in the mean model. In the KM framework, testing for the factor set effects is reduced to testing for the nullity of the corresponding variance components rather than the actual parameters in the mean model. Typically, the corresponding degrees of freedom used by the KM test are much smaller than the number of model parameters, resulting in increased power [Liu et al., 2007; Kwee et al., 2008; Liu et al., 2008]. KM regression does not suffer from the burden of dimensionality like naïve regression, less information is lost to data-reduction compared to PC regression, and kernels can be constructed to incorporate important biological information [Kwee et al., 2008; Wu et al., 2010]. In addition, KM regression can handle correlated factors like PC regression and allows for the investigation of non-linear effects, which is not practical in naïve regression and not feasible in PC regression.

KM regression methods have been developed to perform analyses in the Multi-G-Multi-E setting. However, care needs to be taken when extending this approach to appropriately investigate GE interactions. Directly constructing a meaningful kernel for investigating the interaction between two factor-sets is not straightforward. Others have proposed KM regression approaches that circumvent the need for directly constructing a kernel for interaction effects. Lin et al. [2013] developed a KM testing approach that

examines the interaction between a single environmental factor and a set of genetic markers. While this approach can be used to directly test for GE interaction, only a single environmental can be considered at a time rather than an environmental factor-set. Wu and Maity [personal communication] are in the process of developing a KM regression approach that attempts to separate the individual genetic (G) and environmental (E) effects from the combined joint effect through careful modeling. Then, using the remaining signal to capture the GE interaction effect, they attempt to build a test based on this component, avoiding the need to specify any explicit kernel for the GE interaction.

Unlike these previous works, we propose a KM regression approach that provides a simple, but intuitive, solution for constructing a kernel for the GE interaction that is directly based on the genetic kernel and the environmental kernel. We also construct a series of score tests to evaluate the complete effect profile. That is, each component of the effect profile (i.e., G, E, and GE interaction) can be examined individually or in conjunction with other components to understand the effect patterns. As such, the approach can be used as an exploratory tool to investigate the effect profile in GWAS of GE interactions (GE-GWAS). Unlike naïve regression or PC regression, developing tests under a KM regression framework may be non-trivial depending on the null hypothesis, which dictates the number of nuisance variance-component parameters that need to be estimated when deriving the test statistic. We demonstrate through simulation studies and a data application that our proposed method can have markedly improved power over the currently available approaches for investigating GE interactions in a Multi-G-Multi-E setting. In the remaining sections of this article, we formally introduce our method, describe and report the findings

of simulation studies designed to compare its performance to currently available approaches, present an application of our method to data from a cardiovascular disease study, and conclude with a brief discussion of the work's major findings and connections to the current literature.

2.3 Methods

Let the vector (Y_i, G_i, E_i, X_i) represent the observed data for individual i in a sample of size n. Let Y_i represent the continuous trait value. Let $G_i = (g_{i1}, g_{i2}, \dots, g_{iL})$ be a $L \times 1$ vector containing un-phased genotypes at L biallelic SNPs. Let $E_i = (e_{i1}, e_{i2}, \dots, e_{iM})$ be a $M \times 1$ vector containing environmental factors. Let $X_i = (x_{i1}, x_{i2}, \dots, x_{iP})$ be a $Q \times 1$ vector containing covariates that are not included in either G_i or E_i.

Using a linear regression model, the relationship between the continuous trait value Y_i and the genetic factors G_i, the environmental factors E_i, and their interactions GE_i, after adjusting for the additional covariates X_i, can be characterized as

$$Y_i = X_i^T \beta + h_G(G_i) + h_E(E_i) + h_{GE}(GE_i) + \varepsilon_i, \quad (1)$$

where β is a $Q \times 1$ vector of regression coefficients describing the effects of the covariates X_i, ε_i's are independent random errors that follow a $\mathcal{N}(0, \sigma^2)$ distribution, and $h_G(\cdot)$, $h_E(\cdot)$, and $h_{GE}(\cdot)$ are smooth, vector-valued functions that capture the genetic, environmental, and gene-environment interaction effects, respectively. There are many possible choices for the functions $h_*(\cdot)$; for example, specifying a linear function corresponds to traditional linear regression. In this work, we focus on the KM approach

which has been shown to perform well in the Multi-G or Multi-E settings where high-dimensionality and highly correlated covariates are often present [Liu et al., 2007; Kwee et al., 2008; Liu et al., 2008].

Under the KM framework, $h_*(\cdot)$ can be specified through a linear combination of a positive definite kernel function $K_*(\cdot,\cdot)$ and is assumed to lie in the functional space generated by that kernel [Kimeldorf and Wahba, 1970; Liu et al., 2008]. Following the representation theorem, the functions $h_G(\cdot)$, $h_E(\cdot)$, and $h_{GE}(\cdot)$ in the linear regression model in (1) can be written as $h_G(G_i) = \sum_{i'=1}^{n} \alpha_{Gi'} K_G(G_i, G_{i'})$, $h_E(E_i) = \sum_{i'=1}^{n} \alpha_{Ei'} K_E(E_i, E_{i'})$, and $h_{GE}(G_i, E_i) = \sum_{i'=1}^{n} \alpha_{GEi'} K_{GE}((G_i, E_i), (G_{i'}, E_{i'}))$, respectively, where $\alpha_{*i'}$ are the unknown parameters. The kernel function $K_*(\cdot,\cdot)$ is a distance metric that quantifies the similarity between subject i and subject i'. Some commonly used kernel functions include the linear kernel function, given by $K_z(z_i, z_{i'}) = z_i^T z_{i'}$, the IBS kernel for genetic data, given below in (2), and the polynomial kernel function, given by $K_z(z_i, z_{i'}) = (1 + z_i^T z_{i'})^2$.

2.3.1 GE Interaction Kernel

In this work, we aim to design a kernel function that can be used to model high-dimensional, complex interactions between genetic and environmental factors in a Multi-G-Multi-E setting. We propose that a GE kernel, denoted by $K_{GE}(\cdot,\cdot)$, can be appropriately constructed as a function of the chosen genetic and environmental kernels, denoted by $K_G(\cdot,\cdot)$ and $K_E(\cdot,\cdot)$, respectively. By directly specifying a GE kernel, inference for the different components of the effect profile can be clearly delineated (e.g., the G vs. E vs. GE

effects) allowing a GE analysis under the KM framework to enjoy the same ease-of-interpretation as a naïve regression or a PC regression analysis where each component of the effect profile can readily be focused on inferentially.

For the genetic kernel, we selected the IBS kernel that is commonly used for genotype data [Kwee et al., 2008] and is given by

$$K_G(G_i, G_{i'}) = \frac{1}{2\sum w_\ell}\sum_{\ell=1}^{L} w_\ell IBS(g_{i\ell}, g_{i'\ell})$$

$$= \frac{1}{2\sum w_\ell}\sum_{\ell=1}^{L} w_\ell\{2 \cdot I(g_{i\ell} = g_{i'\ell}) + I(|g_{i\ell} - g_{i'\ell}| = 1)\}, \quad (2)$$

where $IBS(\cdot,\cdot)$ denotes the number of alleles shared by subject i and subject i' at SNP ℓ and w_ℓ are SNP-specific weights. The weights are used to incorporate prior information about the genetic variants into the analysis in order to improve performance. Because similarity in rare alleles is more informative than similarity in common alleles, we took the weights to be $w_\ell = f_\ell^{-3/4}$ as recommended by Pongpanich et al. [2012], where f_ℓ is the minor allele frequency (MAF) of SNP ℓ. With this weight, the contribution of rare variants is up-weighted, but not too strongly so that the contribution of common variants can still be retained. For the environmental kernel, we selected the interactive kernel [Maity and Lin, 2011] given by

$$K_E(E_i, E_{i'}) = 1 + \sum_{m=1}^{M} e_{im}e_{i'm} + \sum_{m<k} e_{im}e_{ik}e_{i'm}e_{i'k} \quad (3)$$

which explicitly includes two-way interactions along with the main effects of the environmental variables. If one wishes to include quadratic effects in the model, then the 2nd-order polynomial kernel can be specified instead.

Finally, we construct a GE kernel based on the selected genetic and environmental kernels. Direct construction of a GE kernel based on commonly used genetic and environmental kernels is not a straightforward task due to the concern of introducing duplicate terms in the GE kernel. That is, when constructing $K_{GE}(\cdot,\cdot)$ as a function of $K_G(\cdot,\cdot)$ and $K_E(\cdot,\cdot)$, marginal genetic and environmental effects often appear in $K_{GE}(\cdot,\cdot)$ that are already captured in $K_G(\cdot,\cdot)$ or $K_E(\cdot,\cdot)$. For example, in our setting, if we constructed $K_{GE}(\cdot,\cdot)$ as the product of $K_G(\cdot,\cdot)$ and $K_E(\cdot,\cdot)$, the GE kernel would be defined as

$$K_{GE}\big((G_i,E_i),(G_{i'},E_{i'})\big) = K_G(G_i,G_{i'})K_E(E_i,E_{i'}) \tag{4}$$

$$= K_G(G_i,G_{i'})\Big(1+\sum_{m=1}^{M} e_{im}e_{i'm} + \sum\nolimits_{m<k} e_{im}e_{ik}e_{i'm}e_{i'k}\Big)$$

$$= K_G(G_i,G_{i'}) + K_G(G_i,G_{i'})$$

$$*\left(\sum_{m=1}^{M} e_{im}e_{i'm} + \sum\nolimits_{m<k} e_{im}e_{ik}e_{i'm}e_{i'k}\right).$$

The duplicate genetic main effect term is introduced by the constant in the interaction kernel used for the environmental kernel. The presence of duplicate terms in the GE kernel causes colinearity and leads to invalid conclusion about the interaction effects. The overlap in (4) suggests a simple yet effective solution for directly constructing the GE kernel using the genetic and environmental kernels: re-define the environmental kernel as $K_E^*(E_i,E_{i'}) = K_E(E_i,E_{i'}) - 1$ and then calculate the GE kernel as

$$K_{GE}\big((G_i, E_i), (G_{i'}, E_{i'})\big) = K_G(G_i, G_{i'})K_E^*(E_i, E_{i'}) \quad (5)$$

$$= K_G(G_i, G_{i'}) * \left(\sum_{m=1}^{M} e_{im}e_{i'm} + \sum\nolimits_{m<k} e_{im}e_{ik}e_{i'm}e_{i'k}\right).$$

The approach in (5) can also be used for general kernel specification. For example, if one chooses to use a polynomial kernel for $K_G(\cdot,\cdot)$ or $K_E(\cdot,\cdot)$ instead of the IBS or interaction kernel, respectively, one can define $K_z^*(z_i, z_{i'}) = K_{z,polynomial}(z_i, z_{i'}) - 1$ for $Z = G$ or E first, and then construct $K_{GE}(\cdot,\cdot)$ by taking the element-wise product of $K_G(\cdot,\cdot)$ and $K_E(\cdot,\cdot)$.

2.3.2 Score tests for assessing Multi-G-Multi-E Effects

In a Multi-G-Multi-E setting, there may be several null hypotheses of interest depending on the goal of the analysis. For example, in a confirmatory analysis, the goal may be to replicate a GE interaction signal and the GE effect would be tested individually. However, in an exploratory analysis, the goal maybe to look for any evidence of a relationship between the genetic and environmental factors and the response and the G, E, and GE effects would be tested jointly. To address these needs, we develop a series of tests based on the linear regression model in (1).

First, when little is known a priori, a joint test, defined as $H_0^{Joint}: h_G(\cdot) = h_E(\cdot) = h_{GE}(\cdot) = 0$, serves as a good tool to detect the overall association induced by genetic and environmental main effects or by GE interaction effects. Instead of beginning with a full scan of genetic main effects in a GWAS, investigators can begin with a full scan using the joint test for the overall association involving both genetic and environmental effects. A scan by joint tests may lead to increased flexibility and power to detect a signal as some

genes can exhibit negligible marginal effects but strong effects among particular exposure groups [Kraft et al., 2007; Thomas, 2010a].

If the joint test is rejected, a GE test, defined as $H_0^{GE}: h_{GE}(\cdot) = 0$, can then be used to identify whether the effect of the genetic variables are modified by the environmental variables. The ability inferentially isolate the GE effects, a tool that is not always available in other KM approaches, can be extremely useful in the Multi-G-Multi-E setting for several reasons. First, the interaction test can aid in understanding biological mechanisms and pathways [Mechanic et al., 2012]. For example, Vineis et al. [2001] found interactions between gene *NAT*2 and tobacco smoking and other occupational exposures when studying bladder cancer. These results revealed a potential role of arylamines (found in tobacco smoke and other occupational materials such as hair dyes) in the pathogenesis of bladder cancer –*NAT2* is involved in the detoxification of arylamine, and only subjects with the "slow-detox" genotype were at increased risk of cancer when exposed to arlymines. Second, interaction tests can be used to identify novel genes functioning through interactions and to explain "missing heritability". Many studies of complex diseases, including childhood asthma, breast cancer and colorectal cancer, are currently under way to search for genes interfering with different environmental factors [Thomas, 2010a]. Third, although statistical interaction is not entirely consistent with biological interaction [Thompson, 1991], interaction tests can still help to improve the performance of risk prediction models for disease therapies by identifying genotypes that respond differently for given treatments – a key task in pharmacogenomics studies [Murcary et al., 2009].

Furthermore, if there is no evidence of a GE interaction, conditional genetic and environmental effects can be further evaluated by testing $H_0^{G|E}: h_G(\cdot) = 0$ without constraining $h_E(\cdot)$ but under the constraint of $h_{GE}(\cdot) = 0$, and testing $H_0^{E|G}: h_E(\cdot) = 0$ without constraining $h_G(\cdot)$ but under the constraint of $h_{GE}(\cdot) = 0$, respectively. We develop conditional tests, $G|E$ and $E|G$, rather than marginal tests of G or E, because they are usually more meaningful and interpretable in the context of Multi-G-Multi-E studies. That is, researchers are often interested in investigating the incremental information about the response provided by genetic effects, for example, over and above other covariates in the mean model. Additionally, the conditional tests can be used to understand the inconsistent association findings because marginal associations, often called the crude associations [Robins and Morgenstern, 1987], may disappear after taking differences in other genetic and environmental factors into consideration.

Using an argument similar to Liu et al. [2007], we show in the Supplementary Note that Model (1) has an equivalent linear mixed model representation and can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{h_G} + \boldsymbol{h_E} + \boldsymbol{h_{GE}} + \boldsymbol{\varepsilon} \qquad (6)$$

where $h_G \sim \mathcal{N}(0, \tau_G K_G)$, $h_E \sim \mathcal{N}(0, \tau_E K_E)$, $h_{GE} \sim \mathcal{N}(0, \tau_{GE} K_{GE})$, and $\varepsilon \sim \mathcal{N}(0, \sigma I_n)$. In addition, testing $H_0: h_z(z) = 0$ is equivalent to testing $H_0: \tau_z = 0$ for $z \in \{G, E, GE\}$. Based on these results, we derive the score tests for the joint test, GE test, the conditional G test (G|E test) and the conditional E test (E|G test) based on the REML likelihood of Model (6) in the Supplementary Note. Specifically, the test statistic for the joint test, GE test, G|E test and E|G test are as follows:

$$T_{joint} = \frac{1}{2} Y^T P_0 (K_G + K_E + K_{GE}) P_0 Y \mid_{\tau_{GE}=\tau_G=\tau_E=0,\ \widehat{\sigma_{joint}}},$$

$$T_{G*E} = \frac{1}{2} Y^T P_1 K_{GE} P_1 Y \mid_{\tau_{GE}=0,\ \tau_G=\widehat{\tau_G},\ \tau_E=\widehat{\tau_E},\ \sigma=\widehat{\sigma_{GE}}},$$

$$T_{G|E} = \frac{1}{2} Y^T P_2 K_G P_2 Y \mid_{\tau_{GE}=0,\tau_G=0,\ \tau_E=\widetilde{\tau_E},\ \sigma=\widetilde{\sigma_{G|E}}}, \text{ and}$$

$$T_{E|G} = \frac{1}{2} Y^T P_3 K_E P_3 Y \mid_{\tau_{GE}=0,\ \tau_G=\widetilde{\tau_G},\ \tau_E=0,\ \sigma=\widetilde{\sigma_{E|G}}},$$

where $Y^T = (Y_1, \cdots, Y_n)$, $P_t = V_t^{-1} - V_t^{-1} X (X^T V_t^{-1} X)^{-1} X^T V_t^{-1}$ for $t = \{0,1,2,3\}$, $K_z = K_z(\cdot,\cdot)$ for $z \in \{G, E, GE\}$, $V_0 = \sigma I_n, V_1 = \tau_G K_G + \tau_E K_E + \sigma I_n, V_2 = \tau_E K_E + \sigma I_n$, and $V_3 = \tau_G K_G + \sigma I_n$. In the Supplementary Note, we also derive the EM algorithms, following similar steps as described in Tzeng et al. [2011], to obtain the estimates of the nuisance variance components (i.e., $\widehat{\sigma_{joint}}, \widehat{\tau_G}, \widehat{\tau_E}, \widehat{\sigma_{GE}}, \widetilde{\tau_E}, \widetilde{\sigma_{G|E}}, \widetilde{\tau_G}$, and $\widetilde{\sigma_{E|G}}$) under certain null hypotheses. As shown in the Supplementary Note, these test statistics asymptotically follow a weighted chi-squared distribution, and p values can be obtained by moment matching approaches [Duchesne and Lafaye De Micheaux, 2010].

2.4 Simulation Studies

We performed simulation studies to compare the performance of the proposed KM regression method to the performance of methods currently available for performing analyses in a Multi-G-Multi-E setting. In this work, we focused on two PC regression approaches and did not consider naïve regression approaches given the evidence of their lack of power in this setting [Lin et al., 2011; Wu et al., 2010].

Our simulation studies were based on a 12-locus haplotype distribution of biallelic single nucleotide polymorphisms (SNPs) from the AGRT1 gene [French et al., 2006]. The haplotype distribution as well as the minor allele frequencies (MAFs) and linkage disequilibrium (LD) coefficients for each SNP are given in Table 2.1. The LD was quantified by the average pair-wise R^2 between each SNP and the remaining 11 SNPs. We considered 2 causal loci out of the 12 SNPs under 4 scenarios (Table 2.2): the 2 causal variants had low vs. high MAFs and were in low vs. high LD.

We generated the genotypes of the 12 SNPs based on the haplotype distribution in Table I and generated 5 environmental factors from the multivariate normal distribution $MVN_5(0, V)$ where $V_{ij} = \sigma_E^2 \cdot \rho^{I\{i \neq j\}}$ with $\sigma_E^2 = 1$ and $\rho = 0.3$. The phenotype values were generated using the model $Y_i = \mu(G_i, E_i) + \varepsilon_i$, where ε_i were generated from a $\mathcal{N}(0, \sigma)$ distribution and two models for $\mu(G_i, E_i)$ were considered:

Model 1 (M1) : $\mu_1(G, E) = \gamma_G(G_1 + G_2 + G_1G_2) + \gamma_E(E_1 + E_2 + E_1E_2) + \gamma_{GE}(G_1E_1E_2 + G_2E_1E_2 + G_1G_2E_1E_2)$, and

Model 2 (M2): $\mu_2(G, E) = \gamma_G(G_1G_2) + \gamma_E(E_1E_2) + \gamma_{GE}(G_1E_1E_2 + G_2E_1E_2 + G1G2E1E2$.

In both mean models, 2 SNPs and 2 environmental factors were taken to be causal along with GE interactions. The causal G and E factors are denoted by a subscript of 1 and 2 although they could be any of the 12 SNPs or any of the 5 environment factors. The two causally related environmental factors were arbitrarily chosen to be the first and second factor for all simulation settings.

When examining the power of the Joint Test, we set $\gamma_G = \gamma_E = \gamma_{GE} = \theta$ for $\theta \in \{0.05, 0.10, 0.15\}$ in M1 and M2. For the GE test, we set $\gamma_G = \gamma_E = 1$ and $\gamma_{GE} = \theta$ for $\theta \in \{0.10, 0.25, 0.50\}$ in M1 and M2. For the G|E test, we set $\gamma_G = \theta$, $\gamma_E = 1$, and $\gamma_{GE} = 0$ for $\theta \in \{0.10,\ 0.15, 0.20\}$ in M1 and $\theta \in \{0.10,\ 0.20, 0.40\}$ in M2. For the E|G test, we set $\gamma_G = 1$, $\gamma_E = \theta$, and $\gamma_{GE} = 0$ for $\theta \in \{0.10,\ 0.15, 0.20\}$ in M1 and $\theta \in \{0.10,\ 0.20, 0.40\}$ in M2. These effect sizes (i.e., the γ's) were chosen so that the power of detecting a signal fell within a reasonable range. When examining the type I error rate of the Joint Test, we set $\gamma_G = \gamma_E = \gamma_{GE} = 0$ in M1 and M2. For the GE test, we set $\gamma_G = \gamma_E = 1$ and $\gamma_{GE} = 0$ in M1 and M2. For the G|E test, we set $\gamma_G = 0$, $\gamma_E = 1$, and $\gamma_{GE} = 0$ in M1 and M2. For the E|G test, we set $\gamma_G = 1$ and $\gamma_E = \gamma_{GE} = 0$ in M1 and M2.

Taken together, this leads to 256 distinct simulation settings that cover . In each simulation setting, samples of size 200 were generated. We generated 100 replicates for power simulations and 1000 replicates for evaluating type I error rates. Each replicate was analyzed using the following methods: (1) GE-KM: the proposed KM regression with K_G constructed using the IBS kernel on the G set (the 12-SNP genotypes) and weights $w_\ell = f_\ell^{-3/4}$, K_E constructed using the interactive kernel on the E set (the 5 environmental factors) [Maity and Lin, 2011], and K_{GE} constructed using K_G and K_E; (2) PC1: PC regression including the first PCs of the G set and the E set, respectively, and their interaction; and (3) PC80: PC regression including the top PCs that explain 80% of the variation in the G set and in the E set, respectively, and their two-way interactions. All tests were performed at an α-level of 0.05.

2.5 Results

Type I error rates are presented in Table 2.3. Power results are presented in Figures 2.1 – 2.4. Each figure contains results for M1 (panel (a)) and M2 (panel (b)) and for the 4 simulation scenarios – low vs. high MAF (across columns) and low vs. high LD (down rows). Effect sizes for the causal genetic and environmental factors are shown on the x-axis. The power of each method to detect a signal, plotted as a separate line, is shown on the y-axis.

THE JOINT TEST. The joint tests searches for a signal among all genetic and environmental factors as well as the corresponding GE interactions. In this setting, all three methods had desirable and similar performances under a null model. The type I error rates were around the nominal level with the rate of PC1 and PC80 being slightly above 0.05 (Table 2.3). Across all simulation scenarios, GE-KM performed better than or comparable to either of the PC methods. The largest gain in power was seen under M2 (Figure 1), which is not unexpected because PC regressions captured the additive effects and had limited power when the underlying effect structure involves only complex interactions. The power of all three methods increased when the effect size increased, when the MAF increased, and when the LD of the causal variants increased. The largest power gain for GE-KM resulted from increasing the MAF. A similar power boost was seen for PC80. Increasing LD had minimal effect on the power of GE-KM but did appear to boost the power of PC1. From these results, it is apparent that the power of all 3 methods to detect a signal depends on the underlying causal variant frequency, effects size, LD, and complexity of the effect structure.

In short, GE-KM performed similarly or better than either PC method, suggesting that GE-KM could be a powerful and flexible analytic tool in the Multi-G-Multi-E setting.

THE GE TEST. The GE test searches for a signal among the interaction between all genetic and environmental factors. In this setting, only GE-KM has a desirable performance under a null model (Table 2.3). The type I error rates for the PC methods are much higher than the nominal level. This could possibly be because the GE interaction term fitted in the PC model captured the effects of the GG and EE and led to spurious GE association. Because the type I error rates of the PC methods were so inflated, we did not consider PC methods in the power analysis of the GE test. As the effect size increased, so did the power of GE-KM. However, in order to obtain a range of power in the GE test setting similar to that of the joint test setting, the effect sizes had to be approximately doubled (see Figure 2.2). Increasing the MAF, the LD, and the complexity of the effect structure (i.e., M1 vs. M2) had minimal effects on the power of GE-KM, suggesting that the power of GE-KM may not be heavily influenced by these factors when performing the GE test. In summary, GE-KMR is able to maintain control on the type I error rate while achieving meaningful power to detect a signal.

THE G|E TEST. The G|E test searches for a signal among genetic factors controlling for environmental factors. In this setting, the KM, PC1 and PC80 methods had desirable and similar performances under a null model with type I error rates near the nominal level (Table 2.3). Across all simulation scenarios, GE-KM performed better than or comparable to either of the PC methods (see Figure 2.3). Under M1, the power of GE-KM and PC80 increased as the MAF, LD, and effect size increased. The largest gain in power of GE-KM

over the PC methods resulted from increasing the LD from low to high. Under M2, the same general pattern of results was observed. However, the power gain of GE-KM over the PC methods was not as substantial, suggesting the IBS kernel may have limited power to capture interaction effects among the genetic factors. The overall power observed under M2 was markedly lower than the overall power observed under M1 even after approximately doubling the effect sizes. When both the MAF and LD are low, all three methods achieved only minimal power under M1 and M2. To help clarify the relationship, additional effect sizes were studied (i.e., 0.4 and 0.6). Under M1, increasing the effect size resulted in desirable power gains. The same was not true under M2. After reexamining the haplotype distribution, we found that there are no overlapping minor alleles between the two causal SNPs chosen for this scenario. Therefore, under M2 constructing the GG interaction always produced a zero-term which led to no genetic effects in the data generation model. In summary, the performance of all three approaches suffered in the G|E setting; yet the GE-KM performed similarly or better than the PC methods.

THE E|G TEST. The E|G test searches for a signal among environmental factors controlling for genetic factors. In this setting, all three methods had desirable and similar performances under a null model with type I error rates near the nominal level (see Table 2.3). Across all simulation scenarios, GE-KM performed better than or comparable to either of the PC methods (see Figure 2.4). The largest gains in power of GE-KM over the PC methods occurred under M2 compared to M1. Under M2, the power of GE-KM was markedly better than either of the PC methods which had almost no power in these simulation scenarios. As the effect size increased, so did the power of each method; the only exception was the PC

methods under M2, where the power remained near null across all studied effect sizes. Increasing the MAF and LD has little effect on GE-KM in this setting. These results confirm that PC methods have little capability to capture interaction effects because they only take linear effects into consideration. Conversely, the GE-KM was able to perform well even when the effect profile involved complex interactions between the G and E effects.

2.6 Real data example: Application to the CoLaus study data

We used the proposed KM regression method to analyze data from the CoLaus study [Song et al., 2011]. The data set contains the measured plasma level of the lipoprotein-associated phospholipase A2 (Lp-PLA2) enzyme in 87 subjects, which has been shown to be associated with risk of coronary heart disease. Song et al. [2011] studied the effect of the non-synonymous rare variants (MAF < 0.05) in *PLA2G7* on Lp-PLA2 activity based on 29 carriers of the rare non-synonymous variants and 58 matched non-carriers who were matched based on age, sex and low-density lipoprotein cholesterol level. The study found significantly lower Lp-PLA2 activity in carriers compared to the non-carriers.

In our analysis, we aimed to examine the association between Lp-PLA2 enzyme activity and 11 common variants typed in the *PLA2G7* gene from the CoLaus GWAS, a group of clinical risk factors of cardiovascular disease (CVD), and potential interactions between these genetic variants and CVD risk factors (i.e., the G, E, and GE effects, respectively, estimated in the GE-KM method). The CVD risk factors considered in this analysis include homocystein, insulin, glucose, aspartate amino-transferase, triglycerides, apolipoprotein B,

glomerula filtration rate, and body mass index. We also adjust for set of potential confounders: carrier status, age, gender, storage duration of serum, physical activity, smoking status, alcohol assumption, and the first three principle components computed from the GWAS data. We assess the association between the 11 common variants, the 8 risk factors, and their interactions using the GE-KM, PC1, and PC80 methods. The results of the series of tests are given in Table 2.4.

In the GE-KM analysis, the IBS kernel with $w_\ell = f_\ell^{-3/4}$weights was used for the genetic factors [Pongpanich et al., 2012], the two-way interaction kernel was used for the environmental risk factors [Maity and Lin, 2011], and the proposed GE interaction kernel was used for the GE factors. We first applied the joint test and found a significant signal (p-val = 0.009). We next performed the GE test to assess the interactions between the *PLA2G7* variants and the CVD risk factors. The results suggest that there is no evidence of an interaction (p-val = 0.252) and that the major source of the joint association signal may arise from main genetic or environmental effects. To further explore the joint signal, we then performed the G|E test and the E|G test to examine the effect of the *PLA2G7* variants and the CVD risk factors on Lp-PLA2 activity levels, respectively. The results of the conditional E test revealed no association between CVD risk factors and Lp-PLA2 levels (p-val=0.233), while the conditional G test found evidence of an association between the *PLA2G7* variants and LP-PLA2 activity levels (p-val = 0.006). These results suggest that the common variants may provide additional diagnostic information about enzyme activity after accounting for differences in CVD risk factors and rare *PLA2G7* variants as well as other potential confounders.

For comparison, we also applied the PC1 and PC80 approaches to the same dataset using the same analysis procedure. We did not conduct the GE test due to the inflated type I error rates of these procedures in this setting. Applying the PCR1 method resulted in no significant findings for the joint test, the conditional E test, or the conditional G test. These results indicate that the multi-G and multi-E information may not be sufficiently captured by the first principle component. In contrast, the PCR80 was able to detect a near significant result in the joint test (p-val=0.055). Although the approach detected a significant association in the conditional G test (p-val=0.002), the signal would have been missed if a hierarchical analysis structure was undertaken where effect-specific tests (i.e., the GE, G|E, and E|G tests) were performed only if the joint test detected a signal. The p-value of the G|E test based on PCR80 is on the same order but smaller than the p-value based on GE-KM, which may imply that an additive linear effect exists among the common variants within *PLA2G7*.

For verification purposes, we also fitted a multiple linear regression model including the main effects of the potential confounders, the 11 common *PLA2G7* variants, and the CVD risk factors. An F-test was performed to access the effect of the 11 variants in PLA2G7 and the resulting p-value was 0.035, agreeing with the analyses from GE-KM and PC80 methods.

2.7 Discussion

Studying complex diseases in the post-GWAS era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors. Much of the work in the recent literature has focused on Multi-G approaches, and KM regression has emerged as a powerful and flexible approach for performing analyses in a Multi-G setting. However, the need for Multi-G-Multi-E methods is becoming apparent as investigators have turned their attention to mining GWAS data for potential GE interactions. In this work, we propose a KM regression approach that directly constructs a kernel for GE interactions based on the genetic and environmental kernels and incorporates a series of score tests to evaluate the complete effect profile (i.e., the G, E, and GE effect individually or in combination). We find that our method can have markedly better power than the currently available approaches when performing analyses in a Multi-G-Multi-E setting. The largest gains were observed when the underlying effect structure involved complex interactions, a scenario believed to be plausible for complex human diseases. As such, the proposed KM approach is a powerful and flexible tool for performing exploratory or confirmatory analyses for investigating GE interactions in GWAS.

Approaches currently available for examining effect profiles in Multi-G-Multi-E analyses include naïve regression, PC regression, and KM regression. Naïve regression, although simple and straightforward to apply, is typically not recommended in the Multi-G-Multi-E setting due the curse of dimensionality and the small effect sizes often observed in the GE-GWAS. Although PC regression has been shown to be a more viable approach than

naïve regression in this setting, the approach has a few limitations that could substantially impact its performance in Multi-G-Multi-E analyses. Besides the potential loss of information and the inability to consider non-linear effects, we found that the type I error of the PC regression was inflated when performing test involving the GE effects. This could possibly be due to the fact that the PCs cannot adequately capture the non-additive genetic and environmental effects (i.e., GG and EE effects) in the underlying model. As a result, the GE interaction term in the PC model absorbed the GG and EE effects and led to false GE findings when there were no GE interactions. This observation agrees with the findings of Voorman et al. [2011] who studied the implications of using a mis-specifed mean model when investigation GE interactions. Specifically, Voorman et al. [2011] pointed out severe type I error rate inflation can occur in GE-GWAS studies when the fitted model used to screen for GE interactions does not correctly reflect the true underlying G and E effects and suggested a model-robust estimate of the variance to address the issue. The proposed KM models can offer an alternative solution to the inflation problem. To minimize the impact of model mis-specification, one can use polynomial kernels, interactive kernels, or IBS kernels to capture the non-additive and non-linear effects of multi-G (or multi-E) factors; one can also use the nonparametric kernels such as Gaussian kernels to model the effects in a nonparametric fashion. Our simulation results indicated that the proposed GE-KM maintained control on the type I error rate across all simulation settings and had comparable or better power that the PC approaches.

Finally, the proposed GE-KM approach can be applied in a variety of research settings where the definition of the “environmental” factors can vary based on the

application area. For example, in the analysis of the CoLaus Study data the environmental factors were taken to be clinical risk factors that are known to be related the trait being studied. The environmental factors could be taken to be factors that are truly external to the individual, like air pollution measures when studying the interaction between genetics and pollution and the risk of lung cancer, or could be taken to be other biomarkers, like expression or metabolite levels when studying the interaction between genetics and transcriptomics or metabolomics on the risk of adverse events in a clinical trial.

ACKNOWLEDGEMENTS

The authors thank Drs. Peter Vollenweider and Gerard Waeber, PIs of the CoLaus study, and Drs. Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline for providing the CoLaus phenotype and sequence data. This work was supported by National Institutes of Health grants R00 ES017744 (to AM), R01 MH074027 (to JYT) and P01 CA142538 (to JYT).

2.8 Supplementary note

CONNECTION BETWEEN KERNEL MACHINE REGRESSION AND LINER MIXED MODELS

Liu et al. [2007] showed that kernel machine regression with one nonparametric component, $h(\cdot)$, can be represented using a linear mixed model. Here we extend the results to multiple nonparametric components, e.g., $h_G(\cdot), h_E(\cdot)$ and $h_{GE}(\cdot)$.

Assume that $h_z(\cdot) \in \mathcal{H}_{\mathrm{K_z}}$, a reproducing kernel Hilbert space. Under the full kernel regression model, parameters β, h_G, h_E and h_{GE} are estimated by maximizing the penalized likelihood function

$$J(h_G, h_E, h_{GE}) = -\frac{1}{2}\sum_{\mathrm{i=1}}^{\mathrm{n}}\{Y_i - X^T\beta - h_G(G_i) - h_E(E_i) - h_{GE}(G_i, E_i)\}^2$$
$$-\frac{1}{2}\lambda_G\|h_G\|^2_{H_{K_G}} - \frac{1}{2}\lambda_E\|h_E\|^2_{H_{K_E}} - \frac{1}{2}\lambda_{GE}\|h_{GE}\|^2_{H_{K_{GE}}}, \qquad \text{(A.1)}$$

where λ_z are penalty parameters balancing complexity of the model and goodness of fit and $\|h\|^2_{H_{K_z}}$, the function norm, is defined as the squared norm for the space $\mathcal{H}_{\mathrm{K_z}}$.

By representer theorem [Kimeldorf and Wahba, 1970] the general solution for h_G, h_E and h_{GE} in (A.1) could be expressed as

$$h_G(\cdot) = \sum_{i=1}^{n}\alpha_i K_G(\cdot, G_i) = k_{G,i}^T\alpha; \qquad \text{(A.2)}$$

$$h_E(\cdot) = \sum_{i=1}^{n}\gamma_i K_E(\cdot, E_i) = k_{E,i}^T\gamma; \qquad \text{(A.3)}$$

$$h_{GE}(\cdot) = \sum_{i=1}^{n} \delta_i K_{GE}(\cdot, GE_i) = k_{GE,i}^T \delta, \qquad \text{A.4)}$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)^T$, $\gamma = (\gamma_1, \cdots, \gamma_n)^T$ and $\delta = (\delta_1, \cdots, \delta_n)^T$ are vectors of unknown parameters. Substituting (A.2), (A.3), and (A.4) into (A.1), we obtain

$$J(\beta, \alpha, \gamma, \delta) = -\frac{1}{2}(Y - X\beta - K_G\alpha - K_E\gamma - K_{GE}\delta)^T(Y - X\beta - K_G\alpha - K_E\gamma - K_{GE}\delta)$$
$$-\frac{1}{2}\lambda_G\alpha^T K_G\alpha - \frac{1}{2}\lambda_E\gamma^T K_G\gamma - \frac{1}{2}\lambda_G\delta^T K_G\delta\,, \qquad \text{(A.5)}$$

where K_G, K_E and K_{GE} are symmetric $n \times n$ matrices whose $(i, i')_{th}$ elements are $K_G(G_i, G_{i'})$, $K_E(E_i, E_{i'})$ and $K_{GE}(GE_i, GE_{i'})$, respectively. To simplify the equation and calculation, we rewrite (A.5) as

$$J(\beta, \theta) = -\frac{1}{2}(Y - X\beta - ZK\theta)^T(Y - X\beta - ZK\theta) - \frac{1}{2}\theta^T\Lambda K\theta, \qquad \text{(A.6)}$$

where $\theta_{3n\times 1} = \begin{pmatrix}\alpha\\ \gamma\\ \delta\end{pmatrix}$, $Z_{n\times 3n} = (I_{n\times n} \quad I_{n\times n} \quad I_{n\times n})$, $K_{3n\times 3n} = \begin{pmatrix}K_G & 0 & 0\\ 0 & K_E & 0\\ 0 & 0 & K_{GE}\end{pmatrix}$ and

$\Lambda_{3n\times 3n} = \begin{pmatrix}\lambda_G I & 0 & 0\\ 0 & \lambda_E I & 0\\ 0 & 0 & \lambda_{GE} I\end{pmatrix}$.

Differentiating $J(\beta, \theta)$ with respect to β and θ, we obtain

$$\frac{\partial J}{\partial \beta} = X^T(Y - X\beta - ZK\theta), \qquad \text{(A.7)}$$

$$\frac{\partial J}{\partial \theta} = K^T Z^T(Y - X\beta - ZK\theta) - \Lambda K\theta. \qquad \text{(A.8)}$$

Set (A.7) and (A.8) equal to 0, and some calculations give

$$\hat{\beta} = \{X^T(I - ZK(Z^TZK + \Lambda)^{-1}Z^T)X\}^{-1}X^T(I - ZK(Z^TZK + \Lambda)^{-1}Z^T)Y, \quad \text{(A.9)}$$

$$\hat{\theta} = (Z^TZK + \Lambda)^{-1}Z^T\big(Y - X\hat{\beta}\big). \quad \text{(A.10)}$$

Define $H = \begin{pmatrix} h_G \\ h_E \\ h_{GE} \end{pmatrix} = K\theta$**.** Then plugging in (A.10), we obtain

$$\widehat{H} = K\hat{\theta} = K(Z^TZK + \Lambda)^{-1}Z^T\big(Y - X\hat{\beta}\big) \quad \text{(A.11)}$$

To build the connection between our kernel machine regression and a liner mixed model, we consider the linear mixed model given by

$$Y = X\beta + h_G + h_E + h_{GE} + \varepsilon, \quad \text{(A.12)}$$

where β is a vector of fixed effects, $h_G \sim N(0, \tau_G K_G)$, $h_E \sim N(0, \tau_E K_E)$, $h_{GE} \sim N(0, \tau_{GE} K_{GE})$ and $\varepsilon \sim N(0, \sigma I)$. Let Z and H be defined the same as before. Then we can rewrite (A.12) as

$$Y = X\beta + ZH + \varepsilon, \quad \text{(A.13)}$$

where $H \sim N(0, G)$, $\varepsilon \sim N(0, R)$, $G_{3n\times 3n} = \begin{pmatrix} \tau_G K_G & 0 & 0 \\ 0 & \tau_E K_E & 0 \\ 0 & 0 & \tau_{GE} K_{GE} \end{pmatrix}$, and $R = \sigma I$.

The normal equations of mixed model (A.13) are then given by [Henderson et al., 1959]

$$\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ H \end{pmatrix} = \begin{pmatrix} X^TR^{-1}Y \\ Z^TR^{-1}Y \end{pmatrix}. \quad \text{(A.14).}$$

Straightforward calculations show that $\hat{\beta}$ and $\widehat{H}$ the solutions to (A.14), are identical to the equations (A.9) and (A. 11) with $\tau = \lambda^{-1}\sigma$. Therefore, we arrive at the conclusion that the least squares estimator $\hat{\beta}$ and $\widehat{H}$ from the kernel machine regression model (1) correspond

to the BLUE and BLUPs, respectively, under the linear mixed model (A.12). Thus, we can use the mixed model representation of kernel machine regression to perform estimation.

DEREIVATION OF THE SCORE TEST STATISTICS AND THEIR DISTRIBUTIONS

Consider the linear mixed model representation given in (A.12). As our primary interest is to test the variance components $\tau_G, \tau_E, \tau_{GE}$, we propose to use the restricted maximum likelihood (REML) function to estimate the variance components $(\tau_G, \tau_E, \tau_{GE}, \sigma)$. We have that the REML estimate under (A.12) is

$$\ell_{REML}(\tau_G, \tau_E, \tau_{GE}; Y) = -\{\log|V| + \log|X^T V^{-1} X| + Y^T P Y\}/2, \quad \text{(A.15)}$$

where $V = \tau_G K_G + \tau_E K_E + \tau_{GE} K_{GE} + \sigma I$ is the marginal variance of Y and $P = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}$ is a projection matrix. The score functions based on the REML can be obtained as below [Harville, 1977]:

Under $H_0^{GE}: \tau_{GE} = 0$,

$$U_{\tau_{GE}}(\hat{\tau}_G, \hat{\tau}_E, 0, \hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_G, \tau_E, \tau_{GE}, \sigma)}{\partial \tau_{GE}}\Bigg|_{\tau_{GE}=0, \tau_G=\widehat{\tau_G}, \tau_E=\widehat{\tau_E}, \sigma=\widehat{\sigma_{GE}}} \quad \text{(A.16)}$$

$$= \frac{1}{2}\{Y^T P_1 K_{GE} P_1 Y - tr(P_1 K_{GE})\}.$$

Under $H_0^{G|E}: \tau_G = 0$ with the constraints of $\tau_{GE} = 0$,

$$U_{\tau_G}(0,\hat{\tau}_E,0,\hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_G,\tau_E,\tau_{GE},\sigma)}{\partial \tau_G}\Big|_{\tau_{GE}=0,\tau_G=0,\,\tau_E=\widetilde{\tau_E},\sigma=\widetilde{\sigma_{G|E}}} \tag{A.17}$$

$$= \frac{1}{2}\{Y^T P_2 K_G P_2 Y - tr(P_2 K_G)\}.$$

Under $H_0^{E|G}: \tau_E = 0$ with the constraints of $\tau_{GE} = 0$,

$$U_{\tau_E}(\hat{\tau}_G,0,0,\hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_G,\tau_E,\tau_{GE},\sigma)}{\partial \tau_E}\Big|_{\tau_{GE}=0,\tau_G=\widetilde{\tau_G},\,\tau_E=0,\sigma=\widetilde{\sigma_{E|G}}} \tag{A.18}$$

$$= \frac{1}{2}\{Y^T P_3 K_E P_3 Y - tr(P_3 K_E)\},$$

where $P_t = V_t^{-1} - V_t^{-1}X(X^T V_t^{-1} X)^{-1} X^T V_t^{-1}, t = 0,1,2,3,$ with $V_0 = \sigma I, V_1 = \tau_G K_G + \tau_E K_E + \sigma I, V_2 = \tau_E K_E + \sigma I$ and $V_3 = \tau_G K_G + \sigma I$.

NULL DISTRIBUTION OF THE SCORE STATEISTICS FOR GE TEST

Because score statistics are not asymptotically normal [Tzeng and Zhang, 2007], we use the first term of the score statistics as the testing statistics. For GE interaction, the test statistic is $T_{GE} = \frac{1}{2} Y^T P_1 K_{GE} P_1 Y$. Define $\mu = X\beta$, then $T_{GE} = \frac{1}{2}(Y-\mu)^T P_1 K_{GE} P_1 (Y-\mu)$ because $\mu^T P_1 = 0$. Further, we can rewrite $T_{GE} = \frac{1}{2} Z^T \left(V^{\frac{1}{2}} P_1 K_{GE} P_1 V^{\frac{1}{2}}\right) Z$, where $Z = V^{-\frac{1}{2}}(Y-\mu)$ and it follows a standard multivariate normal distribution. Define e_i and η_i the eigenvector and eigenvalue of matrix $V^{1/2} P_1 K_{GE} P_1 V^{1/2}/2$, respectively, then $T_{GE} = \sum_{i=1}^{c} \eta_i (e_i^T Z)^2 \equiv \sum_{i=1}^{L} \eta_i \tilde{Z}_i^2$ with $\tilde{Z}_i^2$ follows a 1 df chi-square distribution. Therefore the

distribution of T_{GE} can be approximated by the distribution of $\sum_{i=1}^{c}\hat{\eta}_i\chi_{i1}^2$, where $\hat{\eta}_i's$ are the non-zero eigenvalues of $V^{\frac{1}{2}}P_1K_{GE}P_1V^{\frac{1}{2}}/2|_{\tau_{GE}=0,\tau_G=\widehat{\tau_G},\,\tau_E=\widehat{\tau_E},\sigma=\widehat{\sigma_{GE}}}$. Hence, we can use a moment matching approach to obtain p-values [Duchesne and Lafaye De Micheaux, 2010].

Above we use the GE test as an example and derive the test statistics and its null distribution. By similar argument, we can approximate the null distributions of $T_{G|E}$, $T_{E|G}$ and T_{joint} using the distribution of $\sum_{i=1}^{c}\hat{\eta}_i\chi_{i1}^2$ where $\hat{\eta}_i's$ are the non-zero eigenvalues of $V^{\frac{1}{2}}P_2K_GP_2V^{\frac{1}{2}}/2|_{\tau_{GE}=0,\tau_G=0,\,\tau_E=\widetilde{\tau_E},\sigma=\widetilde{\sigma_{G|E}}}$, $V^{\frac{1}{2}}P_3K_EP_3V^{\frac{1}{2}}/2|_{\tau_{GE}=0,\tau_G=\widetilde{\tau_G},\,\tau_E=0,\sigma=\widetilde{\sigma_{E|G}}}$ and $V^{\frac{1}{2}}P_0(K_G+K_E+K_{GE})P_0V^{\frac{1}{2}}/2|_{\tau_{GE}=\tau_G=\tau_E=0,\,\widehat{\sigma_{joint}}}$, respectively.

EM ALGORITHM FOR THE REML ESTIMATES OF $\boldsymbol{\tau_G}$ AND $\boldsymbol{\tau_E}$ WHEN TESTING $\boldsymbol{H_0}$: $\boldsymbol{\tau_{GE}=0}$

Using the GE test as an example, we derive the EM algorithm for estimating the nuisance variance components (VC), $\tau_G, \tau_E, \text{and } \sigma$, under H_0. The EM algorithms for estimating nuisance VCs for the G|E test and the E|G test can be obtained by zeroing out the corresponding variance components. In short, the derivation of the EM algorithm is similar to the one derived in Tzeng et al. [2011]. Let $u = A^TY$ with $A^TA = I_{n*n}\ and\ AA^T = I - X(X^TX)^{-1}X^T$. Then $f(u|h_G, h_E)$ follows normal distribution with mean $A^Th_G + A^Th_E$ and variance σI and does not depend on the fixed effect β. Therefore, the REML estimators of $\tau_G\ and\ \tau_E$ can be based on their marginal distributions,

$f(u) = \int\int f(u|h_G, h_E) f(h_G, h_E) dh_G dh_E$. This motivated the EM algorithm based on observed data u and missing data h_G and h_E.

The complete data log likelihood is given be

$$logf(u, h_G, h_E; \tau_G, \tau_E, \sigma) = logf(u|h_G, h_E; \tau_G, \tau_E, \sigma) + logf(h_G; \tau_G, \sigma) + logf(h_E; \tau_E, \sigma)$$

$$= -\frac{n-d}{2} log\, \sigma - \frac{1}{2\sigma}(u - A^T h_G - A^T h_E)^T (u - A^T h_G - A^T h_E)$$

$$-\frac{n}{2}\log \tau_G - \frac{1}{2}\log|K_G| - \frac{1}{2\tau_G} h_G^T K_G^{-1} h_G$$

$$-\tfrac{n}{2}\log \tau_E - \tfrac{1}{2}\log|K_E| - \tfrac{1}{2\tau_E} h_E^T K_E^{-1} h_E.$$

In the expectation step, we calculate the expected value of the log likelihood function, $Q(\tau_G, \tau_E, \sigma|\hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)})$ with respect to the observed data u under the current (the t-th iteration) estimate of the parameters $\hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)} and\ \hat{\sigma}^{(t)}$,

$$Q\left(\tau_G, \tau_E, \sigma \middle| \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right) = E\left[logf(u, h_G, h_E; \tau_G, \tau_E, \sigma)|u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right]$$

$$= -\frac{n-d}{2} log\, \sigma - \frac{1}{2\sigma} E\left\{(u - A^T h_G - A^T h_E)^T (u - A^T h_G - A^T h_E)|u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$-\frac{n}{2}\log \tau_G - \frac{1}{2}\log|K_G| - \frac{1}{2\tau_G} E\left\{h_G^T K_G^{-1} h_G|u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$-\frac{n}{2}\log \tau_E - \frac{1}{2}\log|K_E| - \frac{1}{2\tau_E} E\left\{h_E^T K_E^{-1} h_E|u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}.$$

In the maximization step, we maximize $Q(\tau_G, \tau_E, \sigma|\hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)})$ by solving $\frac{\partial Q}{\partial \tau_G} = 0, \frac{\partial Q}{\partial \tau_E} = 0$ and $\frac{\partial Q}{\partial \sigma} = 0$ and obtain the following estimates

$$\hat{\tau}_G^{(t+1)} = \frac{1}{n} E\left\{h_G^T K_G^{-1} h_G|u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \frac{1}{n}\{\hat{\tau}_G Y^T P K_G P Y + tr(\tau_G I - \tau_G^2 P K_G)\};$$

$$\hat{\tau}_E^{(t+1)} = \frac{1}{n} E\left\{h_E^T K_E^{-1} h_E | u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \frac{1}{n}\{\hat{\tau}_E Y^T P K_E P Y + tr(\tau_E I - \tau_E^2 P K_E)\};$$

$$\hat{\sigma}^{(t+1)} = \frac{1}{n-d} E\left\{(u - A^T h_G - A^T h_E)^T (u - A^T h_G - A^T h_E) | u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \left(Y - \widetilde{M}\right)^T A A^T \left(Y - \widetilde{M}\right) + tr\left(A^T \tilde{V} A\right),$$

where $AA^T = I - X(X^T X)^{-1} X^T$, $\widetilde{M} = E\left(h_G + h_E \middle| u; \hat{\tau}_G^{(t)}, \hat{\tau}_E^{(t)}, \hat{\sigma}^{(t)}\right) = (\tau_G K_G +$

τEKEP1, *V=varhG+hEu;τGt,τEt,*

σt=τGKG−τG2KGP1KG+τEKE−τE2KEP1KE−2τEτGKEP1KG, and M and V are obtained from the joint distribution of (u, h_G, h_E).

Table 2.1: Haplotype distribution with estimated SNP minor allele frequencies and linkage disequilibrium coefficients.*

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8	SNP 9	SNP 10	SNP 11	SNP 12	*Haplo Freq*
Hap 1	0	0	0	1	0	0	0	0	0	0	0	0	0.315
Hap 2	1	1	0	0	0	0	0	0	0	0	0	0	0.041
Hap 3	1	1	0	0	0	0	1	1	1	1	1	0	0.072
Hap 4	0	0	1	0	1	0	0	0	0	0	0	0	0.038
Hap 5	0	0	1	0	1	0	0	0	0	0	0	1	0.127
Hap 6	0	0	0	1	0	0	0	0	1	0	0	1	0.041
Hap 7	0	0	0	1	0	0	0	0	0	0	0	1	0.266
Hap 8	0	1	0	0	1	1	0	0	0	0	0	0	0.054
Hap 9	0	1	0	0	1	1	0	0	0	0	0	1	0.045
MAF†	0.113	0.212	0.165	0.378	0.264	0.099	0.072	0.072	0.113	0.072	0.072	0.479	
LD†	0.349	0.268	0.113	0.204	0.162	0.081	0.439	0.439	0.277	0.439	0.439	0.061	

* Haplotype distribution of biallelic SNPs from the AGRT1 gene [French et al., 2006].

† MAF = minor allele frequencies of a SNP; LD = linkage disequilibrium of a SNP, which was calculated by the average pair-wise R^2 between each SNP and the remaining 11 SNPs.

Table 2.2: Causal SNPs used in the simulation studies.

	Simulation	
	***Low LD** (<0.2)*	***High LD***
***Low MAF** (<0.2)*	(6,3)	(10,11)
High MAF	(5,12)	(2,4)

Table 2.3: Type I error rates averaged over 1000 replicate data sets.

	Null Hypothesis being Tested			
	Joint	**GE**	**G|E**	**E|G**
Model 1 [*]				
GE-KM [†]	0.048	0.041	0.038	0.052
PC1	0.059	0.123	0.061	0.045
PC80	0.060	0.232	0.048	0.050
Model 2				
GE-KM	0.048	0.038	0.039	0.042
PC1	0.059	0.164	0.049	0.052
PC80	0.060	0.246	0.057	0.049

* Model refers to the underlying mean function used in the data generation process; Model 1 has main and interaction genetic and environmental effects; Model 2 has interaction effects only.

† GE-KM = GE kernel machine regression; PC1 = principal component regression using only the first G and E PC and its interaction; PC80 = principal component regression using the top G and E PCs that explain 80% of the variation in the set and their interactions.

Table 2.4: Testing results from the analysis of the CoLaus Study Data[*]

	Null Hypothesis being Tested			
	Joint	**GE**	**G|E**	**E|G**
GE-KM[†]	0.009	0.252	0.006	0.233
PC1	0.321	Not applicable	0.899	0.280
PC80	0.055	Not applicable	0.002	0.904

* Results described using the p-values obtained from the various tests.

† GE-KM = GE kernel machine regression; PC1 = principal component regression using only the first G and E PC and its interaction; PC80 = principal component regression using the top G and E PCs that explain 80% of the variation in the set and their interactions.

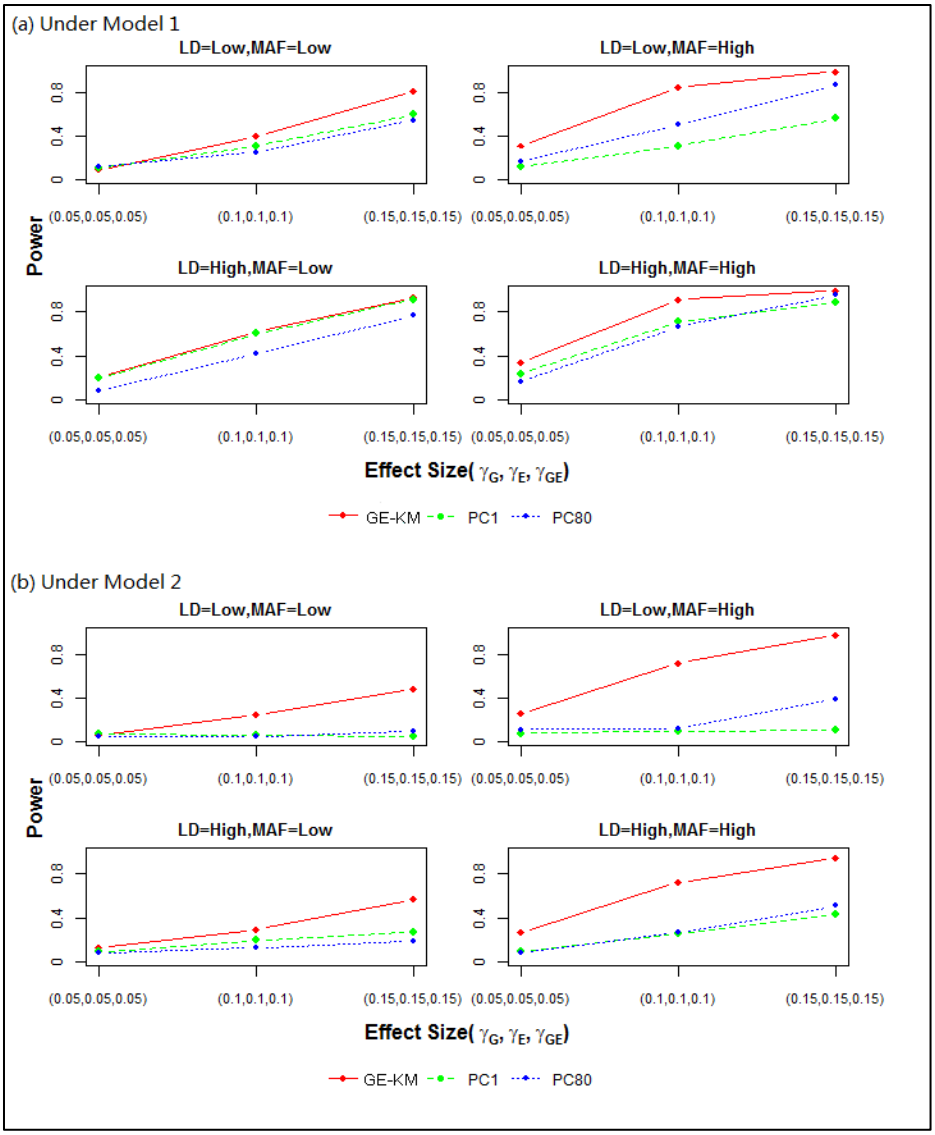

Figure 2.1: Power results for the Joint Test – The results were based on 100 runs of the joint test $H_0^{Joint}: h_G(\cdot) = h_E(\cdot) = h_{GE}(\cdot) = 0$ at $\alpha = 0.05$. Lines indicate the different Multi-G-Mulit-E approaches, where GE-KM denotes the proposed GE kernel machine method, PC1 denotes the PC regression using only the first PC of the G and E effects, and PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set. (For full definitions of Joint test and Models 1 and 2 see the Methods and Simulation Studies sections, respectively.)

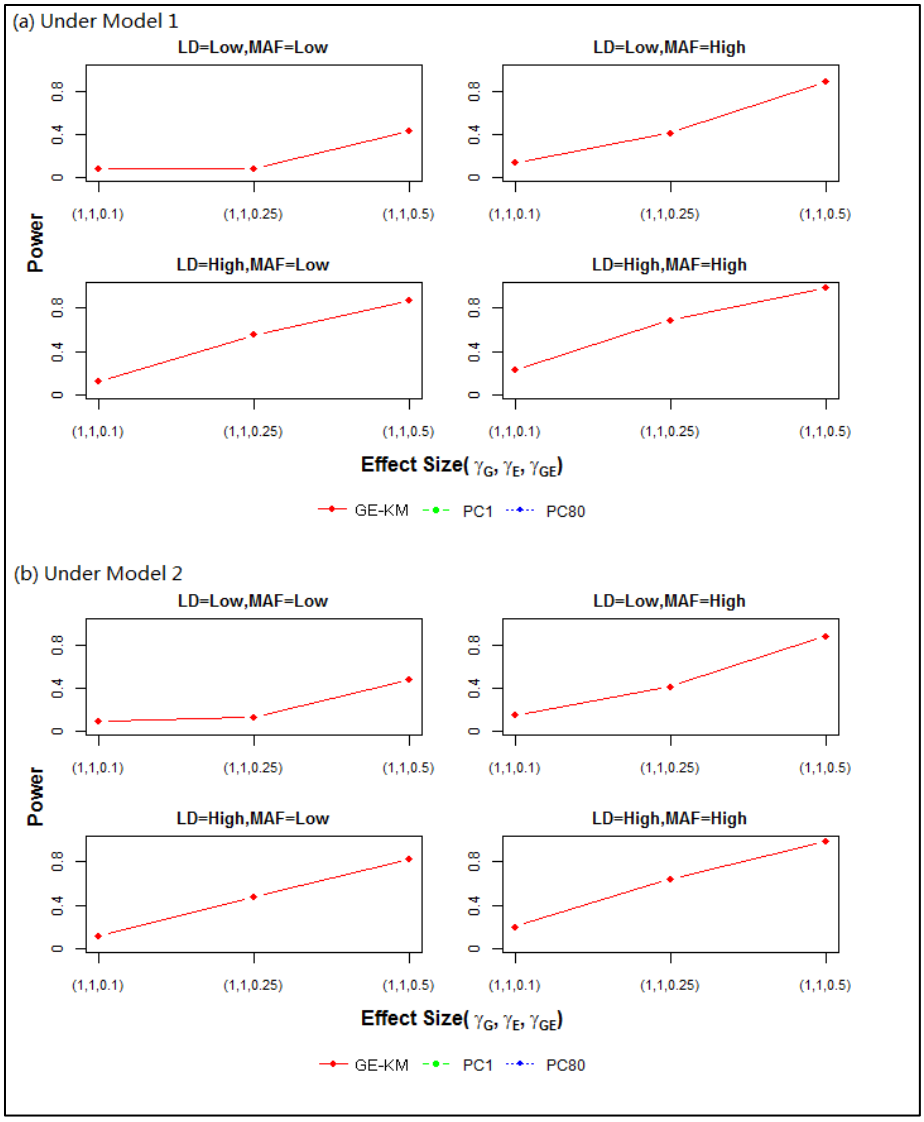

Figure 2.2: Power results for the GE Test – The results were based on 100 runs of the GE test $H_0^{G*E}: h_{GE}(\cdot) = 0$ at $\alpha = 0.05$. Lines indicate the different Multi-G-Mulit-E approaches, where GE-KM denotes the proposed GE kernel machine method, PC1 denotes the PC regression using only the first PC of the G and E effects, and PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set. (For full definitions of GE test and Models 1 and 2 see the Methods and Simulation Studies sections, respectively.)

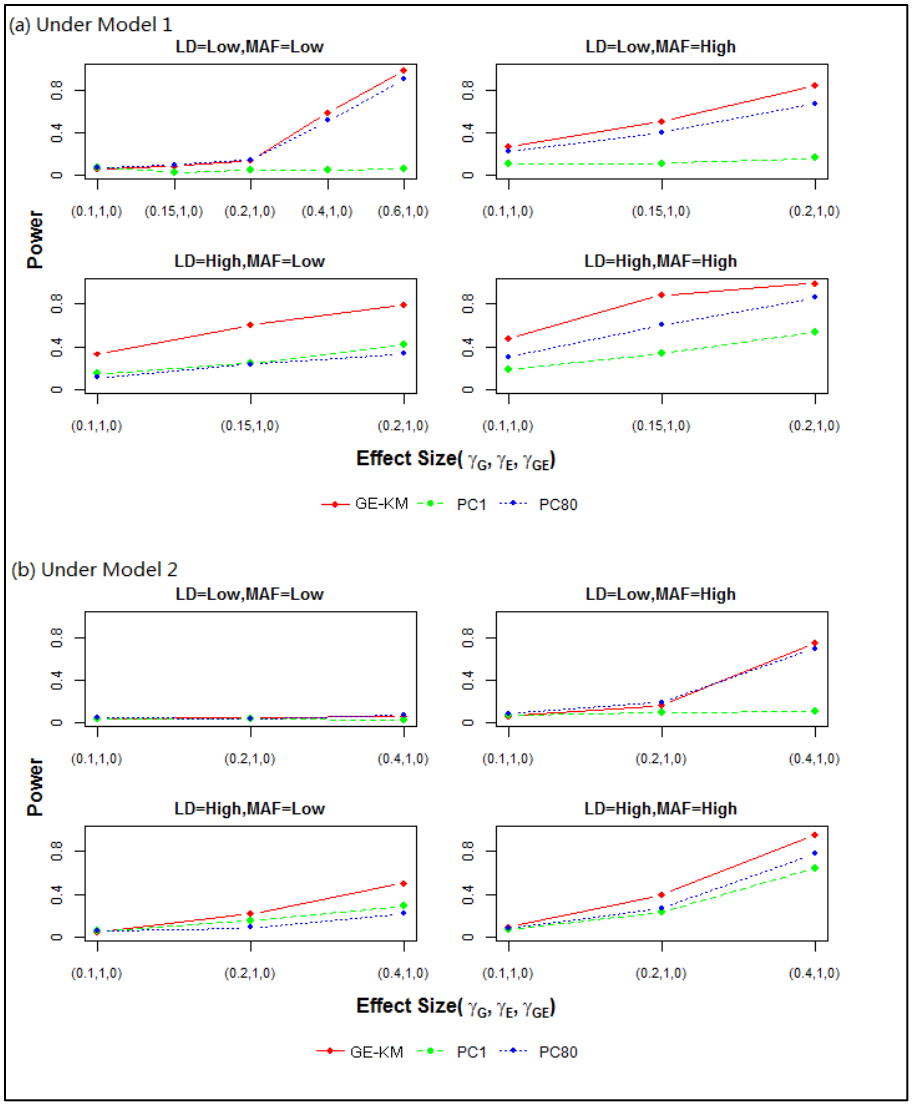

Figure 2.3: Power results for the G|E Test – The results were based on 100 runs of the G|E test $H_0^{G|E}: h_G(\cdot) = 0$ at $\alpha = 0.05$. Lines indicate the different Multi-G-Mulit-E approaches, where GE-KM denotes the proposed GE kernel machine method, PC1 denotes the PC regression using only the first PC of the G and E effects, and PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set. (For full definitions of G|E test and Models 1 and 2 see the Methods and Simulation Studies sections, respectively.)

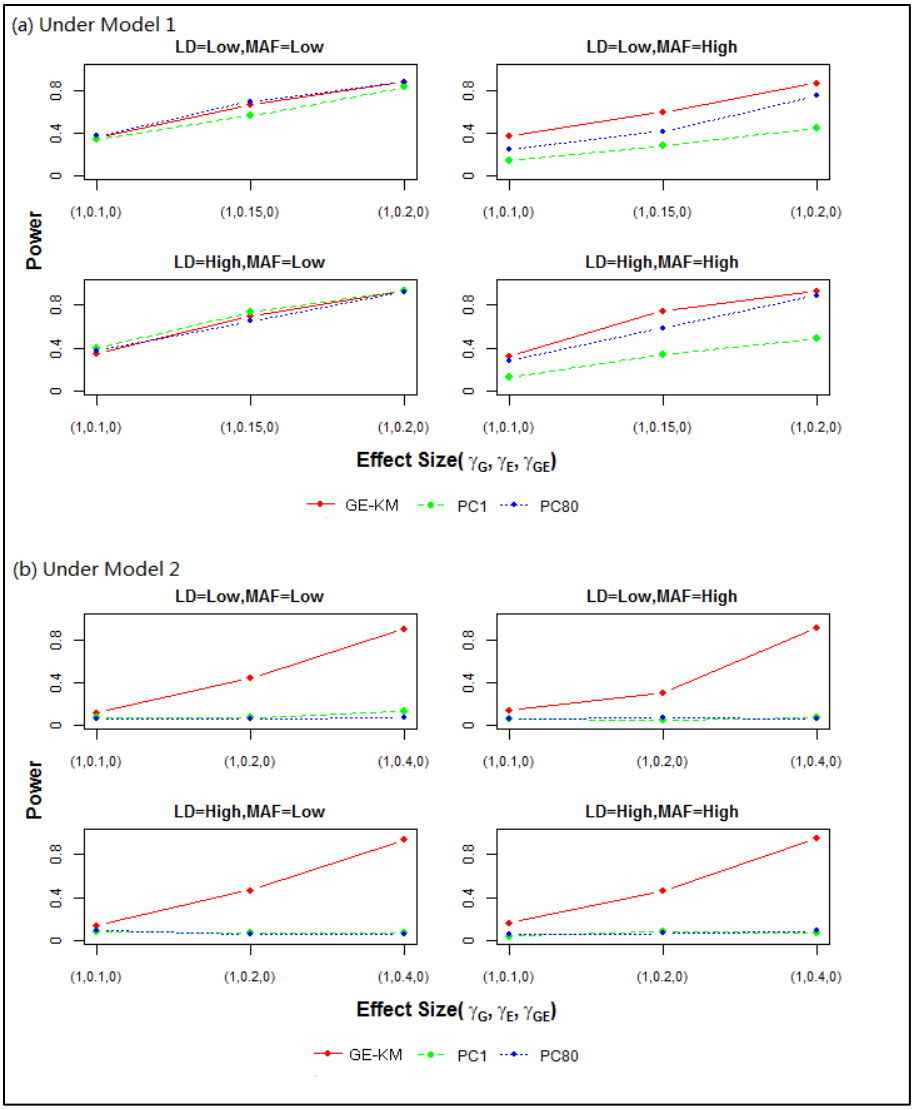

Figure 2.4: Power results for the E|G Test – The results were based on 100 runs of the E|G test $H_0^{E|G}: h_E(\cdot) = 0$ at $\alpha = 0.05$. Lines indicate the different Multi-G-Mulit-E approaches, where GE-KM denotes the proposed GE kernel machine method, PC1 denotes the PC regression using only the first PC of the G and E effects, and PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set. (For full definitions of E|G test and Models 1 and 2 see the Methods and Simulation Studies sections, respectively.)

2.9 References

Ballard DH, Cho J, Zhao H. 2010. Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol 34: 201-212.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773-785.

Dai X, Wu C, He Y, Gui L, Zhou L, Guo H, Yuan J, Yang B, Li J, Deng Q, Huang S, Guan L, Hu D, Zhu J, Min X, Lang M, Li D, Yang H, Hu FB, Lin D, Wu T, He M. 2013. A genome-wide association study for serum bilirubin levels and gene-environment interaction in a Chinese population. Genet Epidemiol 37: 293-300.

Duchesne P and Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics and Data Analysis 54: 858-862.

Edwards DRV, Naj AC, Monda K, North KE, Neuhouser M, Magvanjav O, Kusimo I, Vitolins MZ, Manson JE, O'Sullivan MJ, Rampersaud E, Edwards TL. 2013. Gene-environment interactions and obesity traits among postmenopausal African-American and Hispanic women in the Women's Health Initiative SHARe Study. Hum Genet 13: 323-336.

French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. 2006. Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30: 485-494.

Guaderman WJ, Murcray C, Gilliland F, Conti DV. 2007. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 32: 108-118.

Harville D. 1977. Maximum likelihood approaches to variance component estimation and related problems. J Am Stat Assoc 72:322–340.

Henderson CR, Kempthorne O, Searle SR, von Krosigk CM. 1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics 15: 192-218.

Kimeldorf G, Wahba G. 1970. A correspondence between bayesian estimation on stochastic processes and smoothing splines. Annals of Mathematical Statistics 41: 495-502.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63: 111-119.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82: 386-397.

Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, Lin X. 2011. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol 35: 620-631.

Liu D, Liu X, Ghosh D. 2007. Semiparametric regression of multi-dimensional genomic pathway data: least square kernel machines and linear mixed models. Biometrics 63: 1079-1088.

Lui D, Ghosh D, Lin X. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9: 292.

Lin X, Lee S, Christiani DC, Lin X. 2013.Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics. In press.

Maity A and Lin X. 2011. Powerful tests for detecting a gene effect in the presences of possible gene-gene interactions using garrote kernel machines. Biometrics 67: 1271-1284.

Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, Divi RL, Fan R, Harris EL, Jacobs K, Kraft P, Leal SM, McAllister K, Moore JH, Paltoo DN, Province MA, Ramos EM, Ritchie MD, Roeder K, Schaid DJ, Stephens M, Thomas DC, Weinberg CR, Witte JS, Zhang S, Zöllner S, Feuer EJ, Gillanders EM. 2012. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. Genet Epidemiol 36: 22-35.

Murcray CE, Lewinger JP, Gauderman WJ. 2009. Gene-environment interaction in genome-wide association studies. Am J Epidemiol 169: 219-226.

Naj AC, Scott WK, Courtenay MD, Cade WH, Schwartz SG, Kovach JL, Agarwal A, Wang G, Haines JL, Pericak-Vance MA. 2013. Genetic factors in nonsmokers with age-related macular degeneration revealed through genome-wide gene-environment interaction analysis. Ann Hum Genet 77: 215-231.

Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ. 2013. Systematic identification of interaction effects between genome-and environment-wide associations in type 2 diabetes mellitus. Hum Genet 132: 1-14.

Pongpanich M, Neely ML, Tzeng JY. 2012. On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. Front Genet 2: 1-14.

Song K, Nelson MR, Aponte J, Manas ES, Bacanu SA, Yuan X, Kong, Cardon L, Mooser VE, Whittaker JC, Waterworth DM. 2011. Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 Europeans reveals several rare loss-of-function mutations. The Pharmacogenomics Journal 12: 425-431.

Thomas D. 2010a. Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11: 259-272.

Thomas D. 2010b. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health. 31: 21-36.

Thompson WD. 1991. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44: 221-232.

Tzeng JY, Zhang D. 2007. Haplotype-based association analysis via variance component score test. Am J Hum Genet 81:927-938.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying Gene and Gene-Environment Effects of Uncommon and Common Variants on Continuous Traits: A Marker-Set Approach Using Gene-Trait Similarity Regression. Am J Hum Genet 89: 277-288.

Vineis P, Marinelli D, Autrup H, Brockmoller J, Cascorbi I, Daly AK, Golka K, Okkels H, Risch A, Rothman N, Sim E, Taioli E. 2001. Current smoking, occupation, N-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. Cancer Epidemiol Biomarkers Prev 10: 1249-1252.

Voorman A, Lumley T, McKnight B, Rice K. 2011. Behavior of QQ-Plots and Genomic Control in Studies of Gene-Environment Interaction. PLoS ONE 6(5): e19416.

Wang T, Ho G, Ye K, Strickler H, Elston RC. 2009. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet Epidemiol 33: 6-15.

Wu C, Kraft P, Zhai K, Chang J, Wang Z, Li Y, Hu Z, He Z, Jia W, Abnet CC, Liang L, Hu N, Miao X, Zhou Y, Liu Z, Zhan Q, Liu Y, Qiao Y, Zhou Y, Jin G, Guo C, Lu C, Yang H, Fu J, Yu D, Freedman ND, Ding T, Tan W, Goldstein AM, Wu T, Shen H, Ke Y, Zeng Y, Chanock SJ,

Taylor PR, Lin D. 2012. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat Genet 44: 1090-1097.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86: 929-942.

Chapter 3

Module-based Association Analysis for Evaluating Effects of Biomarkers with Network Structures

Zhi Wang[1], Arnab Maity[2], Deepak Voora[3],Rima Kaddurah-Daouk[4], Jung-Ying Tzeng[1,2,5]

1: Bioinformatics Research Center, North Carolina State University, Raleigh NC, 27695, USA

2: Department of Statistics, North Carolina State University, Raleigh NC, 27695, USA

3: Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, USA

4: Institute for Genome Sciences & Policy, Duke University, Durham, NC, USA

5: Department of Statistics, National Cheng-Kung University, Taiwan, R.O.C.

3.1 Introduction

Module-based analysis (MBA) aims to evaluate the effect of a group of biomarkers sharing common features, such as SNPs in the same gene, co-expressed genes, or metabolites involved in the same pathways. A module can be constructed based on biological knowledge, e.g., pathway databases (Wishart et al., 2007; Wishart et al., 2009; Frolkis et al., 2010; Kanehisa and Goto, 1998;), or based on computational algorithms, e.g., clusters of correlated biomarkers (Stone and Ayroles, 2009; Kaufman and Rousseeuw, 1990; Langfelder and Horvath, 2008). MBA provides an attractive alternative to traditional single–biomarker approach. Module may serve as a more appropriate analyzing unit to understand the complex biological system because most cellular functions are carried out by groups of biomarkers rather than individual biomarkers (Barabasi and Oltvai, 2004). The biomarkers do not work alone but instead act jointly or even interact with each other within a complex network (Zhu et al., 2007). Consequently, biomarkers tend to have moderate individual effects but significant aggregate effect, and performing analysis at module level can increase the detectability and reproducibility of association findings. By assessing biomarker effects in a functional context, e.g., pathways and biological processes, MBA also improves the interpretability of findings and facilitates the construction of follow-up biological hypotheses. Finally, for exploratory “omics” studies, which usually begin with a full scan of a long list of candidate biomarkers, MBA provides a natural way to reduce the total number of tests, and hence relax the multiple-testing burdens and improve power.

Current approaches of MBA can be roughly classified into two major categories. The first type is the "meta"-based methods, which assess the module effect by integrating testing results of individual biomarkers, e.g., minimum p-value and Fisher's test (De la Cruz et al., 2010; Fisher, 1932). The second type is the "mega"-based methods, which jointly model the effect of all biomarkers in a module, such as principle component regression (Guaderman et al., 2007; Wang and Abbott, 2008) and kernel machine regression (Liu et al, 2007; Kwee et al., 2008; Wu et al., 2010; Schaid et al., 2010). Compared to the "meta"-based approaches, it is believed that "mega"-based methods can better capture the complex joint effect among biomarkers within a module.

One limitation of these current approaches is that they ignore network information in biological system (Barabasi and Oltvai, 2004; Snoep et al., 2005). Biological elements are connected and regulate each other as part of network. Genes and gene produces regulate each other's expressions and form a gene regulatory network. Proteins physically bind each other to carry out important functions in many molecule processes such as DNA replications and form a protein-protein interaction network. Metabolites in cellular metabolism are modified through a series of biochemical reactions, which can be integrated into a metabolic network. One common way to describe the network information is to use graphs with a collection of nodes and edges (Carter et al., 2004; Zhang and Horvash, 2005), where nodes represent biological elements and edges represent connections between them. Elements, such as genes and metabolites, that are directly connected in a network may

have similar biological functions or involve in the same biological process. Therefore, incorporating network structure information can more precisely model the biological effects, enhance the ability to detect true associations, and facilitate our understanding of the underlying biological mechanisms (Zhu et al., 2007).

Many approaches have been developed to utilizing network structure information. Most methods formulate the identification of important biomarkers as a variable selection question and incorporate network structure by either specifying a network-constrained penalty function (Li and Li, 2008) or incorporating prior distribution (Li and Zhang, 2008; Tai and Pan, 2009; Monni et al., 2010; Chen et al., 2011). These methods have concentrated on evaluating the effects of a single module and identifying the specific biomarkers that cause the module-level significance. Our network-based work has different focuses---we focus on evaluating the effects of multiple modules and investigating interplay among them. We developed two kernel functions to capture the structural relationship among biomarkers within a module. We demonstrate that the proposed network-based methods can have markedly improved power over the approaches ignoring network information through simulation studies and a real-data analysis of Aspirin pharmacometabolomics studies.

The rest paper is organized as follows. We first introduce the two network based kernels and the kernel machine regression for detecting the main and interactive effects of modules. Next, we present numerical studies based on simulations and real data

applications. Finally, we conclude with a brief discussion of the contributions and related issues.

3.2 Method

3.2.1 Kernel machine regression model

Consider a sample with n subjects. Let Y_i represent the continuous trait value; $X_{1i} = (x_{1i1}, x_{1i2}, \dots, x_{1iL_1})$ be a vector containing values of the L_1 biomarkers in Module 1; $X_{2i} = (x_{2i1}, x_{2i2}, \dots, x_{2iL_2})$ be a vector containing values of the L_2 biomarkers in Module 2. Finally, let $Z_i = (z_{i1}, z_{i2}, \dots, z_{iQ})$ be a $Q \times 1$ vector containing covariates that are not included in either X_{1i} or X_{2i}. We use the following semiparametric regression to model the relationship between the traits and the biomarkers in Module 1 and Module 2, which includes the module main effects, $h_1(\cdot)$ and $h_2(\cdot)$, and the interaction effect, $h_{12}(\cdot)$, and adjusts for the additional covariates Z_i:

$$Y_i = Z_i^T \beta + h_1(X_{1i}) + h_2(X_{2i}) + h_{12}(X_{1i}, X_{2i}) + \varepsilon_i, \quad (1)$$

where β is a $Q \times 1$ vector of regression coefficients describing the effects of the covariates Z_i, and ε_i's are independent random errors that follow a $\mathcal{N}(0, \sigma)$ distribution. In Model (1), functions $h_*(\cdot)'s$ are the primary interests because they fully specify the relationship between biomarkers and trait. Under kernel machine framework, we assume that the nonparametric function $h_*(\cdot)$ lies in a function space, $\mathcal{H}_{K_*}$, generated by a positive definite kernel function $K_*(\cdot,\cdot)$. According to the Mercer's theorem (Cristianini and Shawe-Taylor,

2000), $h_*(\cdot)$ can be represented as the primal representation, $h_*(X_i) = \sum_{j=1}^{J} \phi_j(X_i)\eta_j$, where $\phi_j(X_i), j = 1, \cdots, J$, is a set of basis functions specified by $K_*(\cdot,\cdot)$. Equivalently, $h_*(\cdot)$ can also be represented as the dual representation, $h_*(X_i) = \sum_{i'=1}^{L} K_*(X_{i'}, X_i)\alpha_{i'}$ for some integer L. Because $h_*(\cdot)$ is fully defined by the kernel functions, by choosing different kernel functions, we can specify different bases and corresponding models to model module effects. Specifying $h_*(\cdot)$ via the dual representation is more convenient than specifying it via the primal representation because explicit basis functions or features might be complicated. Many kernel functions have been constructed and are commonly used, e.g., the linear kernel function, given by $K_*(X_i, X_{i'}) = X_i^T X_{i'}$, the second order polynomial kernel function, given by $K_*(X_i, X_{i'}) = (1 + X_i^T X_{i'})^2$, and the Gaussian kernel, given by $K_*(X_i, X_{i'}) = exp\left\{\sum_{j=1}^{M}\left(x_{ij} - x_{i'j}\right)^2/d\right\}$, where d is a tuning parameter.

3.2.2 Kernel functions incorporating network information

One appealing feature of kernel machine framework is that it allows for the inclusion of prior information in the kernel function to assist in the evaluation of module effects. In this paper, we introduce two network-based kernels to incorporate network information: the *topology kernel* and the *connectivity kernel*. Both kernels require a known network structure to begin with. Such network structure, typically summarized in the adjacency matrix (Boccaletti et al., 2006), can be obtained from existing biological knowledge (Kanehisa and Goto, 2000) or be constructed from the data (e.g., co-expressed gene modules). Given a network structure, the adjacency matrix is defined as $A \equiv [A_{ll'}]$, where

$A_{ll'} = 1$ if nodes l and l' are connected in the network, and $A_{ll'} = 0$ otherwise including $l = l'$. When network structure is unknown, we use the weighted correlation network analysis (WGCNA) of Zhang and Horvath (2005) to obtain the empirical adjacency matrix from data. See Appendix A for details.

The topology kernel function $K^{Top}(X_i, X_{i'})$. We construct the topology kernels based on the topological overlap matrix (TOM) (Ravasz et al., 2002), which can be computed from the adjacency matrix, A, as given in Equation (2) below. TOM is considered as an alternative to the adjacency matrix to minimize structural noises when describing the module structure (Yip and Horvash, 2007); empirical studies (Ravasz et al., 2002; Yook et al., 2004; Oldham et al., 2006; Carlson et al., 2006) have shown that nodes having a higher topological overlap are more likely to belong to the same functional class. Given matrix A, the corresponding TOM, denoted by $T \equiv [T_{ll'}]$, is

$$T_{ll'} = \begin{cases} \dfrac{L_{ll'} + A_{ll'}}{min\{k_l, k_{l'}\} - A_{ll'} + 1} & for\ l \neq l' \\ 1 & for\ l = l' \end{cases}, \qquad (2)$$

where $L_{ll'} = \sum_{u \neq l,l'} A_{lu} A_{l'u}$, which is the number of neighbors shared between node l and node l'; $A_{ll'}$ indicates if node l and node l' are directly connected to each other;$k_l = \sum_{u \neq l} A_{lu}$, which quantifies the number of direct neighbors (edges) that node l has. From Equation (2), we can see that, in contrast to adjacency matrices, TOM describes the network structure using both $L_{ll'}$ and $A_{ll'}$, that is, TOM measures the node relationship not only based on the pair of nodes themselves but also their relationship to all other nodes in the

network. In other words, for node l and node l' that are not directly connected in a network (e.g., $A_{ll'} = 0$), they are still considered as "closely connected" in terms of high topological overlaps as long as they share common neighbors (e.g., $L_{ll'} \neq 0$). The denominator of Equation (2) is a normalizing factor so that the range of $T_{ll'}$ is between 0 and 1. This is because $A_{ll'} \leq 1$ and $L_{ll'} \leq \min(k_l, k_{l'}) - A_{ll'}$ (Yip and Horvath, 2007).

To fix the idea of the proposed topology kernel function, consider the linear kernel. The topology kernel, denoted by $K^{Top}(X_i, X_{i'})$, incorporates the TOM by $K^{Top}(X_i, X_{i'}) = X_i^T T X_{i'}$. The topology kernel encourages similar effects for those nodes "close" in a network. As we comment in the discussion section, the smoothing effect can been more clearly seen from a Bayesian perspective. Using T instead of A in the topology kernel allows us to quantify the closeness between two nodes not only based on direct connections but also based on the sharing of common nodes.

The connectivity kernel function $K^{Con}(X_i, X_{i'})$. Alternatively, we can incorporate different type of network information from the topological overlap. Specifically, the connectivity kernel, defined as $K^{Con}(X_i, X_{i'}) = X_i^T W X_{i'}$ with $W = diag\{\sum_{l'}^{L} T_{ll'}\}$, considers the connectivity of a node and controls a node's contribution to the analysis based on the number of connections it has, i.e., $\sum_{l'}^{L} T_{ll'}$ for node l. The functional and structural importance of hub nodes (i.e., nodes with high connectivity) have been established in the literature: Removing hub nodes from the network would severely alter network structure (Albert et al., 2000) and impact the network function and organismal fitness (Kamath et al.,

2003; Winzeler et al., 1999; Hahn and Kern, 2005). The connectivity kernel intends to upweight hub nodes so as to reflect the fact that hub nodes tend to play a more substantial role than non-hub nodes in a network (He and Zhang, 2006). For example, it is found that in the yeast protein–protein interaction networks, hubs are more likely to be essential and conserved relative to non-hub proteins (Jeong et al., 2001; Barabasi and Oltvai, 2004).

Here we construct our network kernels based on the TOM. When needed, one may replace TOM by the adjacency matrix or even correlation matrix. Nevertheless, we expect several advantages for using TOM. TOM has been empirically demonstrated to be a meaningful measure on interconnectedness in real biological networks (Zhang and Horvath, 2005; Lubovac et al., 2006). In addition, compared to the adjacency matrix, the TOM is more tolerant to errors caused by spurious or missing edges between two nodes because TOM considers the neighboring structure of the two nodes in addition to their direct connectivity. The edges cannot always be precisely determined due to too noisy or incomplete network information, especially if edges are obtained from relevance network. It has been noted that the adjacency matrix, which is constructed based on direct connection, is sensitive to noises and lead to wrong network inference.

3.2.3 Kernel functions for interaction effects

To model between-module interaction effect, we construct an interaction kernel by taking the element-wise product:

$$K_{12}\big((X_{1i}, X_{2i}), (X_{1i'}, X_{2i'})\big) = K_1(X_{1i}, X_{2i'})K_2(X_{1i}, X_{2i'}),$$

where $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ are kernels used for Module 1 and Module 2, respectively. If other kernels, such as polynomial kernels were used, one would need to remove constant in these kernels as suggested in Wang et al. (2013) to avoid false positive and false negative findings caused by including duplicate main effect term in the interaction kernel.

3.2.4 Testing module effects

We developed two score-based tests under Model (1) to assess module effects. The first is the interaction test for assessing module-module interaction, i.e., to test $H_0^{X_1 * X_2}: h_{12}(\cdot) = 0$. The second is the conditional test for assessing the effect of a certain module adjusting for the other module, i.e., to test $H_0^{X_1|X_2}: h_1(\cdot) = 0$ without constraining $h_2(\cdot)$ but under the constraint of $h_{12}(\cdot) = 0$. The test for $H_0^{X_2|X_1}: h_2(\cdot) = 0$ can be defined by the same manner. To test these hypotheses, we consider the following mixed model representation of kernel machine regression (1) as did in Liu et al. (Liu et al., 2007) and Wang et al. (2013):

$$Y = Z\beta + h_1 + h_2 + h_{12} + \varepsilon, \qquad (3)$$

where $Y^T = (Y_1, \cdots, Y_n)$, $h_1^T = (h_{11}, \cdots, h_{1n}) \sim \mathcal{N}(0, \tau_1 K_1)$ with h_{1i} being the effect of Module 1 for subject i, $h_2^T = (h_{21}, \cdots, h_{2n}) \sim \mathcal{N}(0, \tau_2 K_2)$ with h_{2i} being the effect of Module 2 for subject i, $h_{12}^T = (h_{12,1}, \cdots, h_{12,n}) \sim \mathcal{N}(0, \tau_{12} K_{12})$ with $h_{12,i}$ being the interaction effect of Module 1 and Module 2 for subject i, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \sim \mathcal{N}(0, \sigma I_n)$.

Consequently, testing $H_0: h_*(\cdot) = 0$ under kernel machine regression (1) is equivalent to testing $H_0: \tau_* = 0$ vs. $H_A: \tau_* > 0$ under the linear mixed model (3).

We derive score tests for the interaction test and the conditional test based on the restricted maximum likelihood (REML) of the Model (3); the derivations are given in Appendix B. Specifically, the test statistic for the interaction test ($T_{X_1 * X_2}$), the conditional test of Module 1 ($T_{X_1|X_2}$) and the conditional test of Module 2 ($T_{X_2|X_1}$) are given as follows.

$$T_{X_1 * X_2} = \frac{Y^T P_{12} K_{12} P_{12} Y}{2} \Big|_{\tau_{12}=0, \tau_1=\hat{\tau}_1, \tau_2=\hat{\tau}_2, \sigma=\sigma_{\widehat{X_1 * X_2}}},$$

$$T_{X_1|X_2} = \frac{Y^T P_1 K_1 P_1 Y}{2} \Big|_{\tau_{12}=0, \tau_1=0, \tau_2=\widetilde{\tau_2}, \sigma=\sigma_{\widetilde{X_1|X_2}}}, \text{ and}$$

$$T_{X_2|X_1} = \frac{Y^T P_2 K_2 P_2 Y}{2} \Big|_{\tau_{12}=0, \tau_1=\widetilde{\tau_1}, \tau_2=0, \sigma=\sigma_{\widetilde{X_2|X_1}}},$$

where $Y^T = (Y_1, \cdots, Y_n)$, $P_t = V_t^{-1} - V_t^{-1} Z (Z^T V_t^{-1} Z)^{-1} Z^T V_t^{-1}$ for $t = \{12, 1, 2\}$, $K_t = K_t(\cdot,\cdot)$ for $t \in \{12, 1, 2\}$, $V_{12} = \tau_1 K_1 + \tau_2 K_2 + \sigma I_n$, $V_1 = \tau_2 K_2 + \sigma I_n$, and $V_2 = \tau_1 K_1 + \sigma I_n$. The estimates ($\widehat{\tau_1}, \widehat{\tau_2}, \widehat{\sigma_{X_1 * X_2}}, \widetilde{\tau_2}, \widetilde{\sigma_{X_1|X_2}}, \widetilde{\tau_1}$ and $\widetilde{\sigma_{X_2|X_1}}$) are obtained from the EM algorithms as described in Appendix B. We also show in Appendix B that these test statistics asymptotically follow a weighted chi-squared distribution, and the corresponding p-values can be obtained by moment matching approaches (Duchesne and Lafaye De Micheaux, 2010).

3.3 Simulation

3.3.1 Design

We conducted simulation studies to evaluate the performances of our proposed methods. We considered 2 simulation settings---modules with scale-free structures (Figure 1; referred to as Simulation I) and modules with non-scale-free structures (Figure 2; referred to Simulation II). We compared the kernel machine regression with network-based kernels (referred to as topology kernel and connectivity kernel) to the same approach ignoring the network information (referred to as unstructured kernel).

Simulation I: Scale-free modules. We generated two 20-node modules with scale-free structure based on Barabási–Albert model (Albert and Barabási, 2002) and the network structures of the two modules are given in Figure 3.1. The scale free structures have three well-known features. First, the connectivity (degree) of nodes follows power law. Specifically, define k the number of edges that a node have. The probability distribution of k has the form of $p(k) \propto k^{-\delta}$ with a certain constant δ (i.e., the network parameter). That is, the probability of observing a node with k edges decreases exponentially as k increases. Second, nodes with top connections (i.e., hub nodes) are assumed to play specific roles. Finally, network with scale-free structures are more error tolerant, i.e., random loss of a node in a scale-free network is less destructive than in a random network.

Simulation II: Non-scale-free modules. Although the scale-free structure is the most common network structure in real practice, in reality, it is also possible to obtain modules that do not have such ideal structure due to several reasons. First, sub-networks sampled from a scale-free network are not necessarily scale free (Stumpf et al., 2005). In addition, investigators tend to profile hubs instead of the entire network at the first place in order to reduce the cost. Finally, investigators may not be able to observe the complete network and meanwhile include many irrelevant nodes in the study because of limited knowledge on the network. In Simulation II, we considered two causal modules with structures presented in Figure 3.2. Module 1 consisted of 20 nodes that were highly connected, while Module 2 consisted of 20 nodes that were loosely connected. Both modules were subsets of a large scale-free module containing 100 nodes. Module 1 was obtained by taking the top 20 nodes that had the most connections; Module 2 was formed by taking the bottom 20 nodes that had the fewest connections.

Given the causal modules with certain structures, we followed the simulation design used in Monni and Li (2010) to generate the values of biomarkers and responses in Simulation I and Simulation II. First, we generated the values of the biomarkers from a multivariate normal distribution with pairwise correlation $Cor(X_l, X_{l'}) = G_{ll'}/2$, where $G_{ll'} = 1$ if node l and l' are directly connected and 0 otherwise. We then selected the causal nodes under two scenarios. In first scenario, we deliberately set hub nodes as causal, i.e., assigning the top 4 nodes with most connections in each module. In second scenario,

we randomly selected C nodes from each module as causal with C =4, 10 and 16. Such design is to mimic the scenario that changes in the network occur randomly rather than initiated by hubs to influence the response, presumably due to mutations or environmental factors. Next, we generated response value Y_i from a Normal distribution with mean μ_i and variance ζ. We let $\mu_i = \gamma_1 \times X_{1i}^T \beta_1 + \gamma_2 \times X_{2i}^T \beta_2 + \gamma_{12} \times X_{12i}^T \beta_{12}$, where $X_{zi},\ z = 1,2$ is the design vector of the causal nodes in Module z for subject i, X_{12i} is the design vector including all pairwise interactions between X_{1i} and X_{2i}, and effect size β_z's were randomly determined from the uniform distribution with interval $I = [-0.2, -0.05] \cup [0.05, 0.2]$. We adjusted the values of the variance ζ to reflect different magnitudes of noise-to-signal ratios. Specifically, values of ζ were determined so that the R^2 values explained by μ_i could yield power within a reasonable range. For type I error rate analysis, we set $(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$ and performed 1000 replications. For the power analysis, we performed 250 replications and the values of $(\gamma_1, \gamma_2, \gamma_3)$ was set to be $(0, 0, 1)$ for the interaction test, $(1,0,0)$ for the conditional tests of Module 1, and $(0,1,0)$ for the conditional test of Module 2. We simulated 1000 individuals per replication.

3.3.2 Simulation Results

Type I error analysis of Simulation I and Simulation II (Table 3.1).

In both simulations, the type I error rates were around the 0.05 nominal level for all kernel functions under all scenarios. The results suggest the validity of the asymptotic distributions

for the proposed statistics. It also assured the validity of our KM regression and the legitimacy of power comparisons presented next.

Results of power analysis in Simulation I (scale-free modules).

Under the scenario when hub nodes were causal (Figure 3.3), the connectivity kernel performed the best, followed by the topology kernel and then the unstructured kernel. The pattern held across different R^2 values and for both interaction test and the conditional test. As expected, this is because the causal nodes, which have high connectivity, were most substantially up-weighted by the connectivity kernel than the other two kernels. Here we show one of conditional test results (conditional test of Module 1). Similar conclusions are valid for both conditional tests, because Module 1 and Module 2 have similar scale-free structure.

Under the scenario when the causal nodes were randomly selected (Figure 3.4), the topology kernel had top performance across all scenarios, i.e., different number of causal nods, different magnitude of R^2, different type of tests (interaction test and conditional test). Because causal nodes were randomly selected, they are more likely to be secondary nodes as they are the majority in the scale-free structure. Among the 3 kernel methods, the topology kernel can best capture the signals from secondary nodes. We also found that incorporating connectivity information did not always help in improving power: For interaction test, the connectivity kernel had comparable power to the unstructured kernel, while for the conditional test, the connectivity kernels had the worst or comparable

performances to the unstructured kernel. One possible explanation is that the connectivity kernel over-weighted the information from hub nodes and missed the signals from non-hub nodes.

Results of power analysis in Simulation II (non-scale-free modules).

When the causal nodes were hubs (Figure 3.5), the connectivity kernel had the best performance for all 3 tests and all levels of R^2 just as in Simulation I. However, the amount of power gain by the connectivity kernel was not as substantial as in the case when the modules had a scare-free structure; here the connectivity kernel often had comparable power with the topology kernel, especially in the conditional test of Module 1. We conjecture that it is because the numbers of connections for hub nodes and non-hub nodes were similar in the highly connected modules considered here. In contrast, the conditional test of Module 2 showed that the connectivity and topology kernels achieved a marked increase in power.

When the causal nodes were randomly selected (Figure 3.6), the topology kernel performed had the best performance among the 3 kernels across all scenarios for the interaction tests. The power gain by the topology kernel appeared to be little when there were only 4 causal nodes. For the conditional tests, a similar pattern can be observed; the exception is in the scenario of 4 causal nodes with high R^2, in which case the connectivity kernel had slightly better power than that of the topology kernel. When only a few nodes were selected as causal in the non-scale-free modules, most causal nodes were likely to

have similar structure background (e.g., being isolated or having similar topology and connectivity level). Consequently, incorporating network information did not aid much in power, though it did not hurt the performance either.

3.4 Real data application

We use proposed kernel machine regression method to analyze the aspirin data from the duke institute for genome science and policy. Aspirin is an antiplatelet agent used for the prevention of myocardial infarction and stroke (Trialists' Collaboration, 1994). However, interpersonal variation has been observed in response to aspirin. About 10~20% people suffers aspirin resistance which leads to failure of reduced platelet aggregation (Patrono, 2003). A major factor causing interpersonal variation in drug response is variation in metabolic phenotype (Kaddurah-Daouk, 2008). Therefore, the goal of this study is to study the association between metabolic phenotypes and aspirin resistance.

The aspirin data contains 53 healthy volunteers; for each individual, 403 metabolites were profiled including 151 knowns and 252 unknowns. The drug response, platelet aggregation, is quantified by a composite score which is the first principle component of a series of measurements of platelet aggregation. These measurements are the area under the aggregometry curve induced by collagen, epineprhrine, and ADP at different concentrations. Preliminary study was performed using single-metabolite analyses which assess metabolic effects on drug response one metabolite at a time. Because metabolites

do not function in isolation, module-based metabolite analysis may serve a more powerful alternative to identify the metabolic groups that influence the drug responses.

To illustrate the utility of the proposed methods, we selected candidate modules using the procedure as follows. First, we used the weighted correlation network analysis (WGCNA, Langfelder and Horvath 2008) to find modules of highly correlated metabolites. We then performed an over-representation analysis (ORA) on each module to identify modules that were enriched with "promising" metabolites (e.g., metabolites with p-values less than 0.2 from the single-metabolite analyses). Although modules can also be constructed by knowledge-based approaches such as KEGG, forming module based on correlation pattern allowed us to incorporate unknown metabolites in the analysis.

We performed two sets of analyses: one focused on evaluating the baseline metabolic measurements (referred to as the *baseline* analysis), and the other focused on the change of metabolic measurements (referred to as the *difference* analysis). In the *baseline* analysis, there were two candidate modules (referred to as Module 1 and Module 2) identified from the module discovery procedure mentioned above. In the kernel machine analysis, we started with the Interaction test to assess the interactions between these two modules using the proposed kernels that incorporating network information, i.e., the connectivity kernel and the topology kernel. Both analyses indicated significant interactions between these two modules (the p-values for connectivity kernel and topology kernel are 0.019 and 0.013, respectively). To compare, we repeated the same analysis using the

unstructured kernel, and the p value is not significant (0.055). Because of the significant findings of the interaction tests, we do not proceed further with the conditional main effect tests in the baseline analyses.

In the *difference* analysis, there were also two modules (referred to as Module 3 and Module 4) identified from module discovery procedure. We then started with the interaction analysis using the network-structured kernels. The interaction test was not significant for both kernels. We hence proceeded with the conditional tests and found that both modules are significant (Table 3.3). When using the unstructured kernel, the interaction effect was not significant either, and there was only one module with significant conditional effect on platelet aggregation (Module 3 given Module 4; p-value 0.032).

To gain biological insights of the results, we mapped the known metabolites in the significant modules to the KEGG pathway. We used KEGG Mapper to see if any pathways are enriched by the known metabolites in Module 1 to Module 4. The results indicated that that the biosynthesis of fatty acid pathway is over-represented by Module 3 and Module 4. Specifically, in these fatty acids, Arachidonic acid is known to be a precursor in the production of thromboxane A2 (TXA2), which triggers reaction that lead to platelet aggregation. Aspirin acts as antiplatelet agent by inhibiting the COX1 enzyme, which is a key enzyme in TXA2 generation. This finding suggests a potential relationship between biosynthesis of fatty acids pathway and the aspirin side effect. Studies (Silver et al., 1973;

Lagarde et al., 2010) show that there is interference between fatty acids and platelet inhibition by aspirin.

3.5 Discussion

Module-based analysis has emerged as a powerful and flexible approach for performing studying the relationship between biomarkers and phenotypes (Liu et al., 2007; Kwee et al., 2008; Wu et al., 2010; Stumpf, 2005). However, most of these methods ignore the network structure information, which depicts the interaction and regulation relationship among basic functional units in biology system. Incorporating network information can aid with association detection and uncover underlining biological features. In this work, we proposed a KM approach that directly incorporates network structure to evaluate the joint effect of biomarkers. Specifically, we constructed the connectivity kernel and the topology kernel to capture the relationship among biomarkers in a module. The simulation studies and real data application suggest that our proposed network-based methods can have markedly better power than the approaches ignoring network information.

Our network KM procedure also has a Bayesian interpretation. Consider a simplified model with only one module effect: $Y = h + \epsilon$. Further assume that a liner model is used to model the biomarker-set effect, i.e., $h = X\beta$. Then the proposed KM model with $K_T = XTX^T$, which is equivalent to $h \sim N(0, \tau K_T)$, can be viewed as imposing a prior on the coefficient β with $\beta \sim N(0, \tau T)$. In other words, by incorporating the structure information, we encourage biomarkers nearer in the network space to share similar effects.

The smoothing according to network topology also helps to stabilize the inference especially when the network is large. Finally, the topology structure is only included through prior information, which will guide, rather than force, the effect smoothing and allow data to drive the results.

In our procedure, TOM is constructed based on the adjacency matrix A, and A is constructed based on the pairwise correlation matrix R as described in Appendix A when no prior network knowledge is available. We note that the adjacency matrix can also be built based on other relationship matrix. One possible choice is the partial correlation, which is known to more precisely reflect the number of edges in a network. Indeed, TOM can be replaced by other structure matrices as introduced in Dong and Horvath (2007) to capture different network information besides topology overlap and connectivity. For example, the clustering coefficient, which is a density measure of local connections, can be used to weight nodes in a network with the rationale that nodes with high clustering coefficient may have large effects. Further studies are worth conducting to evaluate the performance of different choices of TOM or A in terms of effect assessment and to evaluate the robustness of the effect assessment with different matrix choices.

From our simulation results, we see different kernel served as the optimal choice under different network structure. Although we do not know where the causal nodes are so to select the optimal kernel in a prior, we might gain insights about the potential significant nodes based on the relative performance of the topology kernel and the connectivity

kernel. Specifically, if the connectivity kernel outperforms the topology kernel, it is possible that hubs play more important roles. Otherwise, nodes with fewer connections but in the same neighborhood might deserve more attention. The results suggest the two structure kernels work in a complementary manner and we would suggest to consider both in the data analysis when possible. If one really has to select one kernel method in a prior, the topology kernel may be the most appropriate choice because it consistently provides comparable or better power than the unstructured kernel method under all scenarios considered (e.g., scale-free vs. non-scale-free modules, and hub causal nodes or random causal nodes). While there are scenarios where the connectivity kernel could provide the most power improvement (such as hub causal nodes in a scale-free module), the connectivity kernel may suffer from power loss when causal nodes are non-hubs in a scale-free module (e.g., the power of the conditional test in Figure 4).

Acknowledgement

The authors thank Dr. Hongjie Zhu at Duke University Medical Center for his valuable feedback and discussion. This work was supported by National Institutes of Health grants R00 ES017744 (to AM), R01 MH074027 (to JYT) and P01 CA142538 (to JYT).

Table 3.1: Type I error rates averaged over 1000 replicate data sets.

	Hull Hypothesis being Tested		
	M1*M2	**M1|M2**	**M2|M1**
Simulation 1 [*]			
Topology[†]	0.047	0.043	0.050
Connectivity	0.038	0.054	0.051
Unstructured	0.042	0.050	0.048
Simulation 2			
Topology	0.045	0.050	0.044
Connectivity	0.042	0.049	0.050
Unstructured	0.052	0.044	0.040

* For details of simulation 1 and 2 see simulation section.

† Topology = Topology Kernel; Connectivity = Connectivity Kernel; Unstructured = Linear Kernel. For details of various kernels see method section.

Table 3.2: Testing results from the *baseline analysis* of the Aspirin Data

Kernel	M1|M2*	M2|M1	M1*M2
Connectivity	NA	NA	0.019
Topology	NA	NA	0.013
Unstructured	0.17	0.62	0.055

* *M1|M2:conditional test of Module 1; M2|M1: conditional test of Module 2; M1*M2: interaction test between Module 1 and Module2.*

Table 3.3: Testing results from the *differential analysis* of the Aspirin Data

Kernel	M3|M4*	M3|M4	M3*M4
Connectivity	0.030	0.039	0.38
Topology	0.023	0.042	0.58
Unstructured	0.032	0.052	0.40

* *M3|M4: conditional test of Module 3; M4|M3: conditional test of Module 4; M3*M4: interaction test between Module 3 and Module 4.*

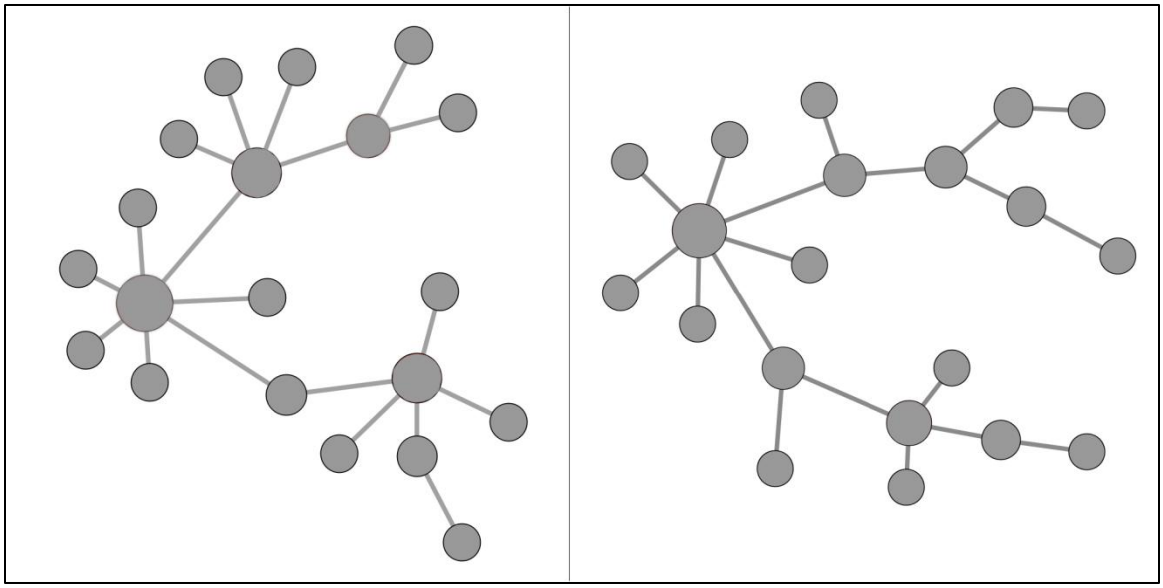

Figure 3.1: Modules with sale-free structures in Simulation I – the left panel is Module 1 and the right panel is Module 2. These modules were simulated based on Barabási–Albert model using igraph package in R.

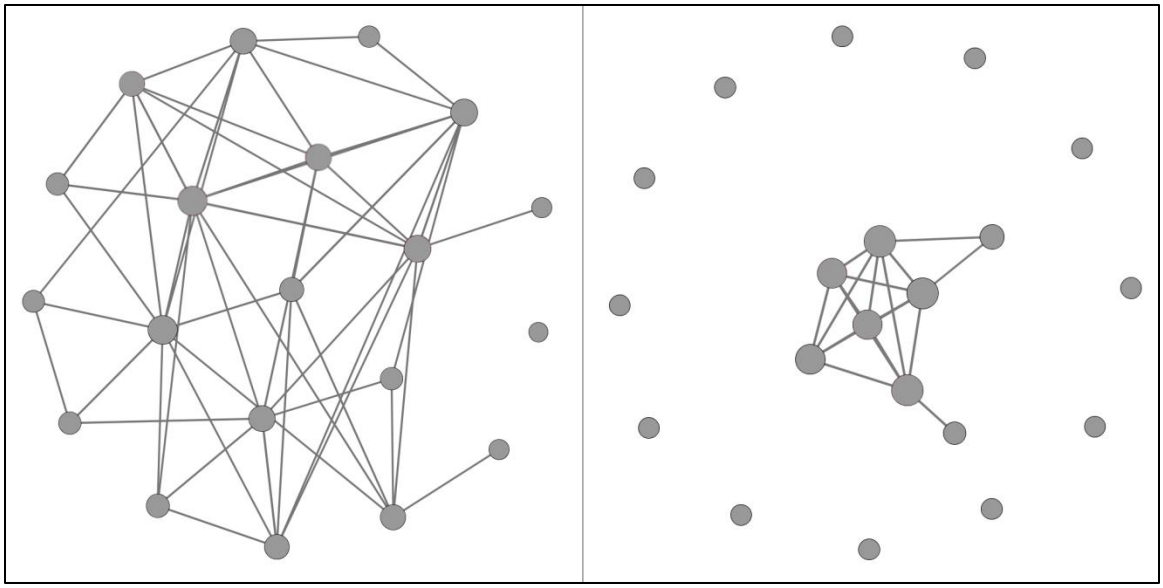

Figure 3.2: Modules with Non-scale-free structures in Simulation II– the left panel is Module 1 and the right panel is Module 2. Both modules were subsets of a large scale-free module containing 100 nodes. Module 1 was obtained by taking the top 20 nodes that had the most connections; Module 2 was formed by taking the bottom 20 nodes that had the fewest connections.

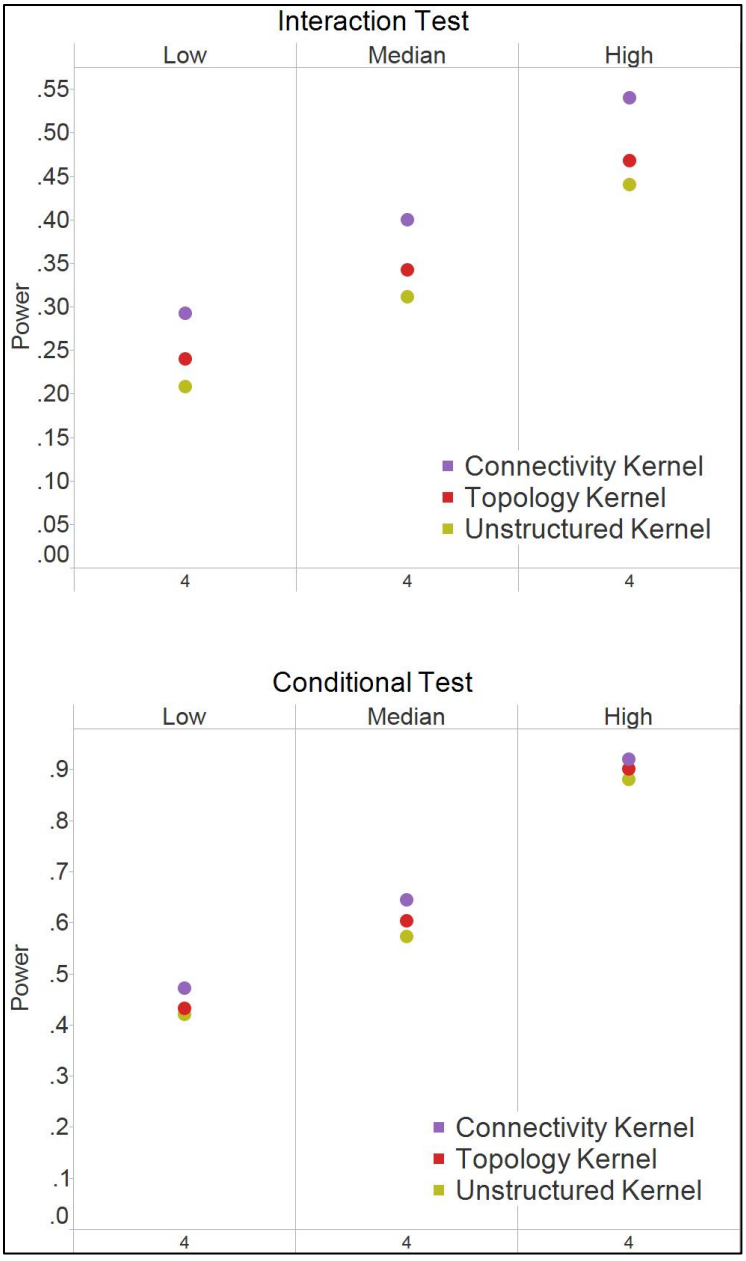

Figure 3.3: Power results for Simulation I (scale-free structure) when causal nodes are hub nodes– The power at $\alpha = 0.05$ were based on 250 simulation replications for the interaction test $H_0^{X_1*X_2}: h_{12}(\cdot) = 0$ and the conditional test $H_0^{X_1|X_2}: h_1(\cdot) = 0$. Dots of different colors indicate the different kernel functions. The specific values for different levels of R^2, (Low, Median, High), were set as $\left(\frac{1}{50}, \frac{1}{70}, \frac{1}{90}\right)$ and $\left(\frac{1}{50}, \frac{1}{100}, \frac{1}{150}\right)$ for the interaction test and the conditional test, respectively. The X-axis indicates the number of causal nodes out of the 20 nodes in a module.

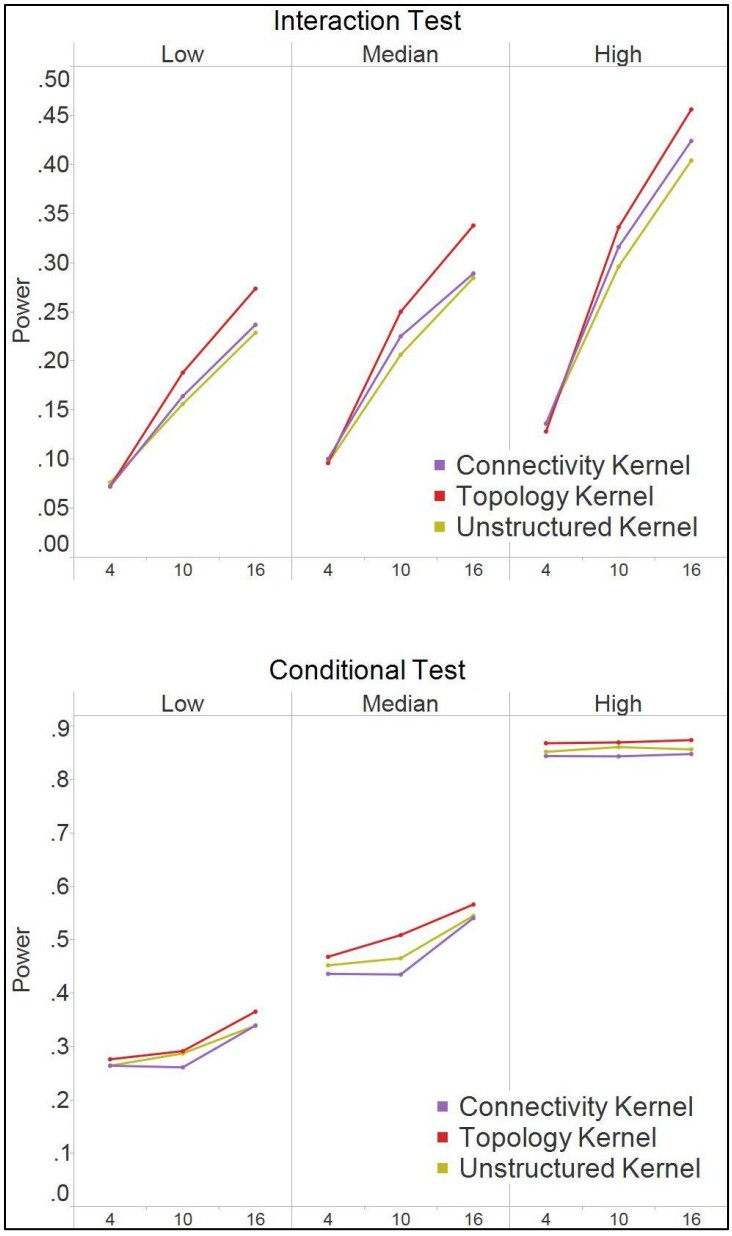

Figure 3.4: Power results for simulation I (scale-free structure) when causal nodes are random nodes – The power at $\alpha = 0.05$ were based on 250 simulation replications for the interaction test $H_0^{X_1*X_2}: h_{12}(\cdot) = 0$ and the conditional test $H_0^{X_1|X_2}: h_1(\cdot) = 0$. Lines of different colors indicate the different kernel functions. The specific values for different levels of R^2, (Low, Median, High), were set as $\left(\frac{1}{50}, \frac{1}{70}, \frac{1}{90}\right)$ and $\left(\frac{1}{50}, \frac{1}{100}, \frac{1}{150}\right)$ for the interaction test and the conditional test, respectively. The X-axis indicates the number of causal nodes out of the 20 nodes in a module.

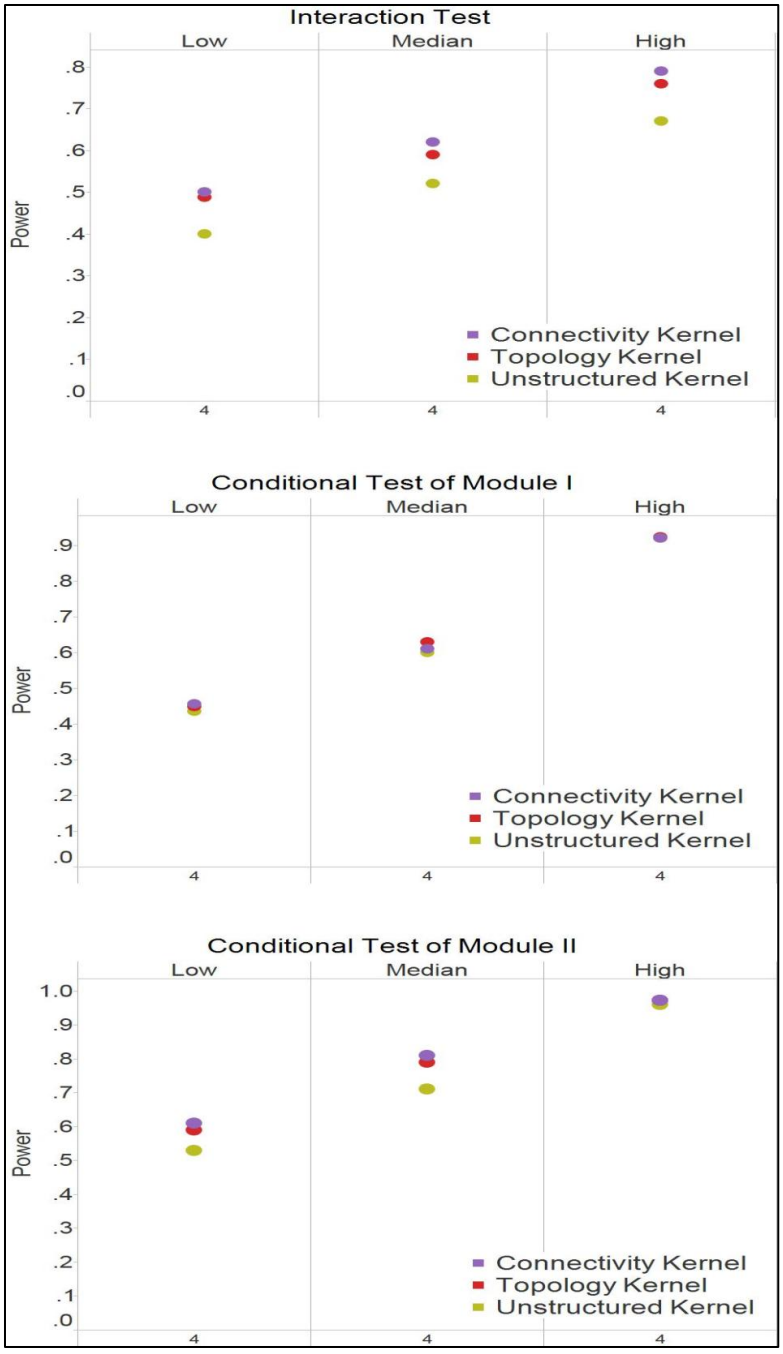

Figure 3.5: Power results for simulation II (non-scale-free structure) when causal nodes are hub nodes – The power at $\alpha = 0.05$ were based on 250 simulation replications for the interaction test $H_0^{X_1*X_2}: h_{12}(\cdot) = 0$, the conditional tests $H_0^{X_1|X_2}: h_1(\cdot) = 0$ and $H_0^{X_2|X_1}: h_2(\cdot) = 0$. Dots of different colors indicate the different kernel functions. The specific values for different levels of R^2, (Low, Median, High), were set as $\left(\frac{1}{50}, \frac{1}{70}, \frac{1}{90}\right)$ and $\left(\frac{1}{50}, \frac{1}{100}, \frac{1}{150}\right)$ for the interaction test and the conditional test, respectively. The X-axis indicates the number of causal nodes out of the 20 nodes in a module.

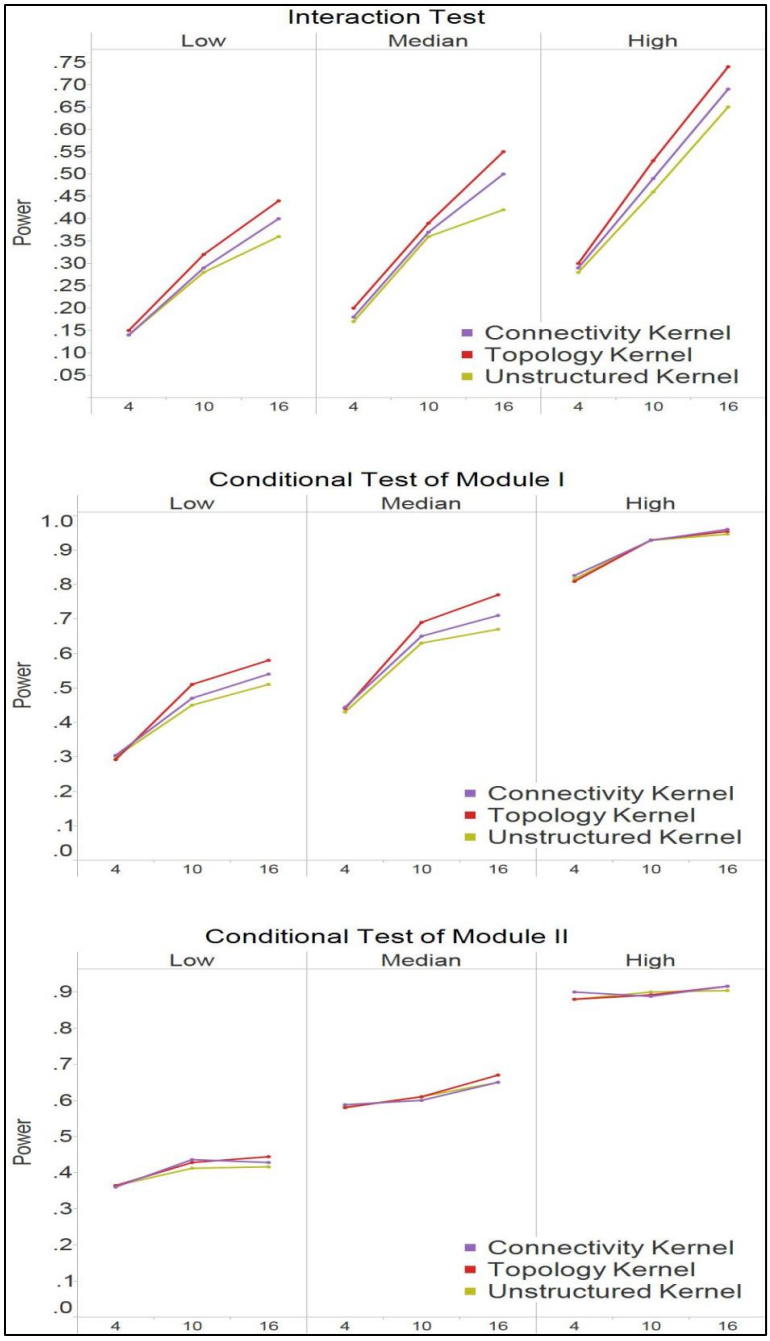

Figure 3.6: Power results for simulation II (non-scale-free) when causal nodes are random nodes– The power at $\alpha = 0.05$ were based on 250 simulation replications for the interaction test $H_0^{X_1 * X_2}: h_{12}(\cdot) = 0$, the conditional tests $H_0^{X_1|X_2}: h_1(\cdot) = 0$ and $H_0^{X_2|X_1}: h_2(\cdot) = 0$. Lines of different colors indicate the different kernel functions. The specific values for different levels of R^2, (Low, Median, High), were set as $\left(\frac{1}{50}, \frac{1}{70}, \frac{1}{90}\right)$ and $\left(\frac{1}{50}, \frac{1}{100}, \frac{1}{150}\right)$ for the interaction test and the conditional test, respectively. The X-axis indicates the number of causal nodes out of the 20 nodes in a module.

3.6 References

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 47.

Albert, R., Jeong, H., & Barabási, A. L. (2000). Error and attack tolerance of complex networks. Nature, 406(6794), 378-382.

Barabasi, A.L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 5(2), 101-113.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. Physics reports, 424(4), 175-308.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics, 7(1), 40.

Carter, S. L., Brechbhler, C. M., Griffin, M. & Bond, A.T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 20, 2242–2250.

Chen, M., Cho, J., & Zhao, H. (2011). Incorporating biological pathways via a markov random field model in genome-wide association studies. PLoS Genet, 7(4): e1001353. doi:10.1371/journal.pgen.1001353

Chen, M., Cho, J., & Zhao, H. (2011). Incorporating biological pathways via a markov random field model in genome-wide association studies. PLoS genetics, 7(4), e1001353.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

De la Cruz, O., Wen, X., Ke, B., Song, M. & Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. Genet. Epidemiol., 34: 222–231.

Dong, J., & Horvath, S. (2007). Understanding network concepts in modules. BMC Systems Biology, 1(1), 24.

Duchesne, P., & Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. Computational Statistics & Data Analysis, 54(4), 858-862.

Fisher, R.A. (1932). Statistical methods for research workers. London: Olive and Boyd.

Frolkis,A., Knox,C., Lim,E., Jewison,T., Law,V., Hau,D.D., Liu,P., Gautam,B., Ly,S., Guo,A.C. et al. (2010). SMPDB: the small molecule pathway database. Nucleic Acids Res., 38, D480–D487.

Guaderman W.J, Murcray C., Gilliland, F., & Conti, D.V. (2007). Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiology, 32, 108 – 118.

Hahn, M. W., & Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution, 22(4), 803-806.

He, X., & Zhang, J. (2006). Why do hubs tend to be essential in protein networks?. PLoS Genetics, 2(6), e88. Connectivity kernel

Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41-42.

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., ... & Ahringer, J. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature, 421(6920), 231-237.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1), 27-30.

Kanehisa, M., & Goto, S. KEGG, (1998). Kyoto encyclopedia of genes and genomes. Nucleic Acids Res., 28, 27– 30.

Kaufman L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc.

Kwee, L.C., Liu, D., Lin, X., Ghosh, D., & Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. American Journal of Human Genetics, 82, 386 – 397.

Lagarde, M., Chen, P., Véricel, E., & Guichardant, M. (2010). Fatty acid-derived lipid mediators and blood platelet aggregation. Prostaglandins, Leukotrienes and Essential Fatty Acids, 82(4), 227-230.

Langfelder, P., & Horvath S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 2008, 9:559.

Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics, 24, 1175-1182.

Li, F., & Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association, 105(491).

Liu D., Liu X., & Ghosh, D. (2007). Semiparametric regression of multi-dimensional genomic pathway data: least square kernel machines and linear mixed models. Biometrics, 63, 1079-1088.

Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. Proteins: Structure, Function, and Bioinformatics, 64(4), 948-959.

Maity, A., Sullivan, P. F., & Tzeng, J. Y. (2012). Multivariate phenotype association analysis by marker-Set kernel machine regression. Genetic Epidemiology, 36(7), 686-695.

Monni, S., & Li, H. (2010). Bayesian Methods for Network-Structured Genomics Data. UPenn Biostatistics Working Papers, Working Paper 34.

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences, 103(47), 17973-17978.

Patrono, C. (2003). Aspirin resistance: definition, mechanisms and clinical read-outs. Journal of Thrombosis and Haemostasis, 1(8), 1710-1713.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science, 297(5586), 1551-1555.

Kaddurah-Daouk, R., Kristal, B. S., & Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease.Annu. Rev. Pharmacol. Toxicol., 48, 653-683.

Schaid, D. J. (2010). Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered , 70, 109-131.

Silver, M. J., Smith, J. B., Ingerman, C., & Kocsis, J. J. (1973). Arachidonic acid-induced human platelet aggregation and prostaglandin formation. Prostaglandins, 4(6), 863-875.

Snoep, Jacky L; Westerhoff, Hans V (2005). "From isolation to integration, a systems biology approach for building the Silicon Cell". In Alberghina, Lilia; Westerhoff, Hans V. Systems Biology: Definitions and Perspectives. Topics in Current Genetics 13. Berlin: Springer-Verlag. pp. 13–30. doi:10.1007/b106456. ISBN 978-3-540-22968-1

Stone, E.A., & Ayroles, J.F.(2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. PLoS Genet., 5, e1000479.

Stumpf, M. P., Wiuf, C., & May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. Proceedings of the National Academy of Sciences of the United States of America, 102(12), 4221-4224.

Tai, F., & Pan, W. (2009). Bayesian variable selection in regression with networked predictors. Manuscript.

Trialists' Collaboration, A. (1994). Collaborative overview of randomised trials of antiplatelet therapy Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. Bmj, 308(6921), 81-106.

Wang, K., & Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. Genet. Epidemiol., 32, 108–118.

Wang, Z., Neely, M.L., Maity, A., Tzeng J.Y. (2013.). Complete effect-profile assessment in association studies with multiple genetic and environmental factors, Submitted.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., ... & Philippsen, P. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science, 285(5429), 901-906.

Wishart, D.S. et al. (2007). HMDB: the human metabolome database. Nucleic Acids Res., 35, D521–D526.

Wishart, D.S. et al. (2009). HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res., 37, D603–D610.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. American Journal of Human Genetics, 86, 929-942.

Yip, A. M., & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics, 8(1), 22.

Yook, S. H., Oltvai, Z. N., & Barabási, A. L. (2004). Functional and topological characterization of protein interaction networks. Proteomics, 4(4), 928-942.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), 1128.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, Vol. 4: No. 1, Article 17.

Zhu, X, Gerstein, M., & Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes Dev., 21, 1010-1024.

Chapter 4

A Module Based Pipeline For Pharmacometabolomics Data Analysis

4.1 Introduction

Pharmacometabolomics, which is a new discipline that stems from metabolomics, aims to explain the interpersonal variation of pharmaceutical compounds (drugs), also referred as drug response phenotype, and to enhance understanding of mechanisms of drugs in terms of the global metabolic profile (Kaddurah-Daouk and Weinshilboum, 2008). In order to achieve these goals, high throughput analytical techniques, such as mass spectrometry and nuclear magnetic resonance (NMR), have been developed to assay thousands of small molecules (metabolites) at the global level, and followed by the application of computational and statistical tools to mine metabolites data. Since pharmacometabolomics studies provide a great opportunity to extensively study the global metabolic profile and possess many features (discussed in Section 1.1.2) shared among omics studies, it usually serves as a discovery or exploratory tool, which indicates that one can mine its dataset in several different ways with various specific purposes.

In a pharmacometabolomics study, a typical analytic approach quantifies not only identified metabolites but also a similar amount of unidentified metabolites due to relatively limited size of reference spectral libraries (Wishart, 2007). And the following statistical analyses for a specific aim, for example, identification of metabolites associated with drug response variation, usually begin with a standard approach, single metabolite analysis, which evaluates the effect of each individual metabolite. Although, use of this single metabolite analysis has successfully discovered several metabolic signatures for some diseases including depression (Paige et al., 2007) and motor neuron disease(Rozen et al., 2005), this type of analysis has many shortcomings including high multiple testing burden and valuable information loss. In practice, this widely used approach tends to produce two extreme outputs. One is that a short list of metabolites or even no metabolites are able to pass the significance level after correction for multiple comparisons, which reflects the difficulty of detecting causal metabolites that each with small but meaningful effect. The other is that a long list of significant metabolites remains even after the correction, which causes biological interpretation problem. Thus single metabolite analysis, by itself, has limited capability for revealing underlying biological mechanism.

An alternative approach, module based analysis, has been discussed in section 1.2.2 and shown several advantages comparing to single metabolite analysis. In this chapter, we integrate both traditional and lately developed statistical methods into a practical module based pipeline (Figure 4.1) for pharmacometabolomics data analysis. This pipeline is

organized by five steps aiming to, but not limited to, identify drug response associated metabolic signatures and pathways. In fact, each step can be used to answer some specific scientific questions and together of them provide a comprehensive and in-depth mining of pharmacometabolomics data.

The rest of this chapter is organized as follows. We first describe the module based pipeline in detail with applicable methods in section 4.2. We then present an application of our proposed pipeline to a real data analysis in section 4.3. Finally, in section 4.4, we make conclusions and present a brief discussion.

4.2 Module Based Pipeline

4.2.1 Module discovery

After pre-process of the pharmacometabolomics data to ensure that the quality of data for further analysis, a module discovery step is required to identify modules, which are defined as groups of metabolites that are biologically or statistically correlated. Although the module discovery step serves as a key component of our proposed module based pipeline, results from this step alone may also potentially point to some functional bases and mechanisms, e.g., metabolic pathways, which aid understanding of underlying system and generation of hypothesis.

As we pointed out in section 1.2.1, modules can be formed either from biological points or from computational algorithm. In practice, constructing modules based on prior

biological knowledge (Yeung et al., 2001; Subramanian et al., 2005) maybe more favorable because the biological context is a great help to interpret and understand the results. However, it has many limitations. First, using known biological information too early may provide misleading information and limit the new findings to be discovered. It's well known that levels of metabolites are dynamic and change over time and location, while most existing biological information are static and don't represent the whole picture of biological process. Thus, they may not match with each other. Also, one metabolite can participate in multiple processes or be shared with different pathways, known as multimodality (Clarke et al., 2008) which may prevent the correct module discovery as well. Additionally, due to the immature profiling techniques along with incomplete reference libraries, the identities of some detected signals cannot be fully determined and the set of detected metabolites cannot cover the global metabolite profile (Wishart, 2007). Therefore, most pharmacometabolomics datasets contain not only many known metabolites but also a great amount of unknown metabolites. In this case, constructing modules from prior biological knowledge is not applicable. Instead, module discovery form computational point is the only option.

There are many computational methods to discovery modules. Most of them are obtained through various clustering algorithms. Clustering is a task of grouping a set of objects in such a way that objects in the same group are more close or similar, to each other than to those in other groups. Here we are introducing two clustering algorithm that are

frequently used in our studies: the modulated modularity clustering (MMC, Stone and Ayroles, 2009) and the weighted correlation network analysis (WGCNA, Langfelder and Horvath, 2008).

The MMC is developed for finding clusters (modules) within the weighted graph. The weighted graph is represented by the affinity matrix which is obtained from nonlinear (Gaussian) transformation of the absolute correlation matrix and whose entries are the edge weights between vertex pairs. The MMC algorithm uses the modularity to define the quality of a cluster by measuring the difference between the total number of edges in the cluster and the expected number of edges in the same cluster at random basis. Then, by maximizing the modularity over all possible clusters (subsets of the affinity matrix), the complete clustering is identified. Since the algorithm searches a solution for the global optimization problem involving all possible clusters with different numbers, it doesn't need to either pre-specify the number of clusters or use external information to validate. This significant feature makes the MMC more objective and consistent than many other clustering approaches.

The WGCNA is developed to find clusters (modules) on the basis of correlation networks. The network is represented by its adjacency matrix derived from the absolute correlation matrix. Both unweighted and weighted network can be constructed depending on using either hard thresholding or soft thresholding approach, detailed see ***. Once the network is constructed, the clustering identification is based on the topological overlap

measure (TOM, Ravasz et al., 2002), which measures the network interconnectedness. By applying hierarchical clustering method on the topological overlap matrix, the goal is to detect modules with densely interconnected metabolites. In contrast to traditional static cut method, they also provide a dynamic tree cut algorithm (Langfelder et al., 2008) for detecting clusters based on the shape of the dendrogram rather than the absolute height of branches. This feature makes the dynamic tree cut algorithm more capable of identifying nested clusters and suitable for automation.

4.2.2 Module filter (optional)

The module filter is an optional step to help us reduce the module space for analysis. After a long list of candidate modules has been produced by the module discovery step, in theory, we should examine the effect of every module formed by highly-correlated metabolites. However, in practice given that there are always various types of response values, e.g., baseline, post and change, along with the pairwise interaction effects among modules to consider, we hope to reduce the number of analyses performed in the real-data analysis by focusing on a few modules that are biologically relevant or known to be promising to study further.

One approach to prescreen modules borrows idea from Biofilter (Bush et al., 2009) in genomics studies, which allows users to filter data based on biological criteria by cross-referencing prior biological knowledge from multiple sources in several ways. Although using biological information to filter module is appealing, no such tools have been

developed for the new emerging metabolomics field. Therefore, we have to applying this approach manually, which is very time consuming and even may not be applicable when the number of modules and metabolites is large. Moreover, using prior knowledge could incorporate subjective criteria into the analysis and lead to selection bias for further analysis.

Another more direct and efficient approach is a module enrichment step. In particular, we prescreen modules based on whether or not they are enriched with a list of promising metabolites satisfying certain condition, for example, significant metabolites from univariate analysis of correlation. Practically speaking, this is done by forming a contingency table and testing the null using a Fisher's exact test.

One concern of this approach is that metabolites that are significant by themselves do not necessarily interact with others, and vice versa, so this univeriate screening step might be sub-optimal for identifying module interactions afterwards. To relieve this concern, we use a loose cut value, e.g., p-value <0.2, to select "promising" metabolites. Such strategy is used and recommended in other field such as GWAS, and the underlying rationale is that interacting factors can often exhibit a marginal effect even when the interaction terms are not modeled (Cordell 2002; Hischhorn and Daly 2005). However, the signal strength is much reduced and hence a less stringent criterion for "promising" should be used.

4.2.3 Module testing

After obtaining a list of candidate modules, we propose to test the associations between these modules and phenotypes of interest, e.g., drug response phenotypes, under the kernel machine regression framework. Results of significant associations may point to the metabolic signatures that have joint effect on the phenotype and even potential function bases reflecting the mechanism of drug action and the variation in the phenotype.

The kernel machine regression framework that has been developed and introduced in previous chapter provides a very flexible way to analyze various effects of candidate modules on the phenotype. Specifically, different forms of kernel machine regression can be applied depending on different study purposes and hypotheses. One can perform an initial analysis with a full scan of the whole list of modules by fitting a simple kernel machine regression (1) to one module at a time, referred as single module analysis which is a direct generalization of single metabolite analysis, in order to detect the main effect of each individual module.

$$Y_i = X_i^T \beta + h_{M_1}(M_{1i}) + \epsilon_i \quad (1)$$

It's also reasonable to consider the environmental factors and their interplays with these metabolic modules by adding an interaction term to the above kernel machine regression (2), because a major factor underlying inter-individual variation in drug effects is variation in

metabolic phenotype, which is heavily influenced by various environmental factors such as nutritional status, age, disease and drug history.

$$Y_i = X_i^T \beta + h_{M_1}(M_{1i}) + h_{M_2}(E_i) + h_{M_1 * E}(M_{1i}, E_i) + \epsilon_i \quad (2)$$

One can also fit a kernel machine regression (3) with module-module interaction terms to explore the interaction effect between modules, or the effect of one module modified by the other module.

$$Y_i = X_i^T \beta + h_{M_1}(M_{1i}) + h_{M_2}(M_{2i}) + h_{M_1 * M_2}(M_{1i}, M_{2i}) + \epsilon_i \quad (3)$$

The ability inferentially isolate the interaction effects, a tool that is not always available in other KM approaches, can be extremely useful for several reasons (Thomas, 2010; Mechanic et al., 2012). First, the interaction test can aid understanding biological mechanisms and pathways. Second, the interaction tests can be used to identify novel modules and metabolites functioning through interactions. Third, although it's aware that statistical interaction is not biological interaction, it still can help to improve the performance of prediction models for drug response – a key task in pharmacometbolomics studies.

4.2.4 Module ORA

The biological interpretation is always the most challenge and important part not only in pharmacometabolomics study but also in other omics studies. This step, over representation analysis (ORA) of modules, aims to get us more close to biology.

We compare metabolites in significant modules from the module testing step with metabolic sets from external sources. These external metabolic sets can be pathway-associated, disease-associated, gene-associated, or location-associated, and thus are biologically meaningful. If the result of overrepresentation analysis is significant, which means there is a great overlap of metabolites between our module and the external metabolic set, we can annotate the module with relevant biological knowledge for the interpretation.

The overrepresentation analysis (Khatri and Drăghici, 2005) is, by far, a very popular method in genome-wide studies. However, in common practices of pharmacometabolimcs studies, it has not been widely used. This is likely because quantitative metabolomics is relatively new developed and lack of tool to perform such analysis automatically. The key to the overrepresentation analysis is construction of a comprehensive and biologically meaningful metabolic set library, which is extremely time consuming and is also restricted to current limited numbers of discovered metabolite pathways. Thus, very few software incorporating ORA was designed. MetaboAnalyst (Xia et al. 2009; Xia et al. 2012) is one user friendly software which has a metabolite set enrichment module aiming to solve this problem with predefined metabolite set libraries from both human and 16 different model organisms.

4.2.5 Key metabolites identification

The last step, key metabolite identification, aims to identify a few key metabolites in each module for further targeted analysis and hypotheses generation. For example, one can conduct a targeted genomics study by genotyping SNPs for these key metabolites, and explore association between these SNPs and response.

A variety of methods are available for choosing the key metabolite. The most simple and direct way is to use univariate ranking method to select the most relevant metabolite to the response phenotype or the significant module. Other methods including multiple linear regressions with model selection criteria such as AIC and BIC and penalized regressions, including lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005), can be used here as well. Additionally, many variable selection procedures have been developed to utilizing network structure information by either specifying a network-constrained penalty function (Li and Li, 2008) or incorporating prior distribution based on network structure information (Li and Zhang, 2008; Tai and Pan, 2009; Monni and Li, 2010; Chen et al., 2011). This feature makes them better tools in identification of biologically relevant metabolites with known network structure.

4.3 Aspirin Study

Aspirin is an anti-platelet agent used for the prevention of myocardial infarction and stroke (He et al., 1998; Eikelboom et al., 2002). However, interpersonal variation has been

observed in response to aspirin. About 10~20% people suffers aspirin resistance which leads to failure of reduced platelet aggregation and increased thrombotic risk (Gasparyan et al., 2008). There are large gaps in our understanding of the pathways responsible for the platelet aggregation response to aspirin. Genome wide association studies of the laboratory aspirin response are currently underway; however, as variation in metabolic phenotype maybe one major source causing interpersonal variation in drug response, metabolomics profiling provides an alternative approach to understand platelet biology and the aspirin response.

In this study, 53 healthy volunteers were given 325mg daily of aspirin for 4 weeks while abstaining from other prescription medications and nicotine products. A series of platelet aggregation measures were conducted on samples taken prior to aspirin therapy (baseline) and after the last aspirin dose (follow-up), including PFA100 (platelet function analyzer) and light transmission aggregometry tastings using various agonists such as collagen, epineprhrine, and ADP at different concentrations. The drug response was then evaluated by the change of platelet aggregation, which is quantified by a composite score based on the 1st principle component of these platelet aggregation measurements. Plasma samples were assayed on GC-TOF platform for each individual and 403 metabolites were quantified including 151 knowns and 252 unknowns.

Given this sample, we propose studying metabolic profile of plasma in healthy volunteers to achieve two aims; first, identification of alternations in metabolic pathways

that associate with drug response; second, Identification of baseline metabolic signatures that can be helpful for predicting drug response.

4.3.1 Identification of metabolic alternations that associate with drug response

One of the most important applications of pharmacometabolomcs is to understand the mechanisms of drug effect, which can be approached by exploring the association between the drug-induced metabolic change and drug response. Therefore, one specific scientist question of interest is that how are changes in metabolites correlated with drug response phenotype. Preliminary study was performed using single-metabolite analyses which assess metabolic effect on drug response one metabolite at a time. However, no metabolites meet the significance threshold after correction for multiple comparisons (Table 4.1). We are aware of the fact that important metabolites may have only moderate marginal effects individually and don't function in isolation, but rather work together in pathways or networks. We then performed module-based metabolite analysis outlined in section 4.2 to identify the metabolic groups associated with the drug response.

According to the procedure, we use weighted correlation network analysis (WGCNA, Langfelder and Horvath 2008) at the beginning to find modules of highly correlated metabolites, which could potentially indicate altered metabolic pathway activates. Specifically, the metabolic dendrogram (Figure 4.2) was obtained by the average linkage hierarchical clustering algorithm and 12 modules with minimum size of 5 were determined

by the dynamic tree cut (Langfelder et al. 2007). Second, a module enrichment step was performed on each module to identify modules that are enriched with "promising" metabolites, which are metabolites with relatively small p-values, 0.2, from the single-metabolite analyses. 2 promising modules were observed with significant p values. One important thing to note, although modules can also be constructed by knowledge-based approaches such as KEGG (Kanehisa and Goto, 2000), forming candidate module based on correlation pattern allow us to incorporate unknown metabolites in the analysis in the study.

In the module testing, there were 2 modules (referred to as Module 1 and Module 2) identified from previous procedure. We then started with the interaction analysis using the network-structured kernel machine methods introduced in Chapter 3. The interaction test was not significant for both kernels. We hence proceed with the conditional tests and found that both modules are significant (Table 3.3).

Because the dataset contains many unknown metabolites, it causes some difficulties to interpret results. However, one can still gain clues from the known metabolites found in significant modules above by mapping them onto the KEGG pathway. We preformed the KEGG pathway analysis to see if the known metabolites in Module 1 and Module 2 are grouped in any pathways using KEGG Mapper (Kanehisa et al., 2012). The results indicated that the biosynthesis of fatty acid pathway have many overlaps with metabolites in Module 1 and Module 2 (Figure 4.3). An alternative approach is to perform an over-representation

analysis (ORA) through the MetaboAnalyst tool (Xia et al. 2012). Similar results were obtained that fatty acid biosynthesis is the most significant one from the pathway analysis (Table 4.2).

Specifically, we found a group of fatty acids. In these fatty acids, Arachidonic acid is known to be a precursor in the production of thromboxane A2 (TXA2), which triggers reaction that lead to platelet aggregation. Aspirin acts as anti platelet agent by inhibiting the COX1 enzyme, which is a key enzyme in TXA2 generation. So, this finding suggests a potential relationship between biosynthesis of fatty acids pathway and the aspirin side effect, which indicates association between fatty acids and platelet aggregation. Many studies (Silver et al., 1973; Lagarde et al., 2010) show that there is interference between fatty acids and platelet inhibition by aspirin.

4.3.2 Identification of baseline metabolic signatures

Another important objective of pharmacomatabolomics studies is to identify baseline metabolic signatures, which can be served as candidate biomarkers for prediction of an individual's response to drug, such as efficacy, toxicity as well as side effects that may occur in the body. The standard strategy involves determining the metabolic profile of a patient prior to treatment, and correlating metabolic signatures with the outcome.

For this specific aim, we begin with the traditional univariate approach using a parametric Pearson's correlation to associate metabolites at baseline with our continuous

drug response phenotype. Although it produced list of metabolites with significant raw p values, no metabolites survived after the correction for multiple comparisons.

We then applied the module based procedure to identify metabolic groups that are most significantly correlated with drug response phenotypes. 2 out of 14 modules were identified as promising candidate module. And kennel machine analysis was applied. We started with the Interaction test to assess the interactions between these two modules using the network based kernels, i.e., the connectivity kernel (XWX) and the topology kernel (XTX). Analyses indicated significant interaction between these two modules (Table 3.2). Because of the significant findings of the interaction tests, we do not proceed further with the conditional main effect tests in the baseline analyses. Further module ORA didn't bring us to any specific pathways, which may suggest profiling these unknown metabolites in modules will help in biological discovery and interpretation.

4.4 Discussion

In this chapter, we proposed a module based analysis pipeline for pharmacometabolomics studies. By forming metabolic modules and testing each module as a unit, this pipeline can enhance the statistical power of biological feature identification and may potentially point to a functional basis for the drug response. We also demonstrated through a real data application on aspirin study that this pipeline can be utilized for two major aims in pharmacometabolomics (Kaddurah-Daouk et al., 2008; Baraldi et al., 2009); one is identification of drug related alternations of metabolic groups for understanding biological

mechanism and development of drug, and the other is identification of baseline metabolic signatures related to drug response phenotype for prediction and personalized medicine.

In each step of the pipeline, we introduced a few of many methods used in our study but emphasize that this is not a comprehensive review. Many alternative methods can be incorporated into our pipeline as well, such as k-means clustering rather than hierarchical clustering and gene set enrichment analysis (GSEA, Subramanian et al., 2005) instead of over representation analysis (ORA). It should be noted that there is no objectively correct approach. Therefore, experimental selection of appropriate algorithms is often needed depending on particular scenarios and problems.

In the Aspirin study, the module based pipeline enabled us to identify groups of metabolites that related to platelet aggregation function. Since it is an exploratory analysis, future replication study is needed to further confirm this finding. And as a metabolomics-informed genomics approach (Abo et al., 2012), a targeted genomics study could be conducted by genotyping SNPs associated with these potential metabolites, and explore association between theses SNPs and response.

Although most of methods in our pipeline have been implemented in various package of R, we reconstructed them to fit our pipeline. A tutorial example with R codes has been shown in Appendix C.

Table 4.1: Correlation analysis between metabolic changes and drug response

Metabolite Name	p-value	FDR*
226876	0.00057	0.13065
Cholesterol	0.00065	0.13065
231716	0.00115	0.14874
288327	0.00176	0.14874
330680	0.00185	0.14874

* Benjamini and Hochberg's approach

Table 4.2: Pathway analysis through MetaboAnalyst 2.0

	Total	Expected	Hits	Raw p	-log(p)	FDR
Fatty acid biosynthesis	49	0.46822	5	7.26E-05	9.531	0.005806
Arginine and proline metabolism	77	0.73577	3	0.035186	3.3471	1
Propanoate metabolism	35	0.33444	2	0.042935	3.1481	1
Galactose metabolism	41	0.39177	2	0.057167	2.8618	1

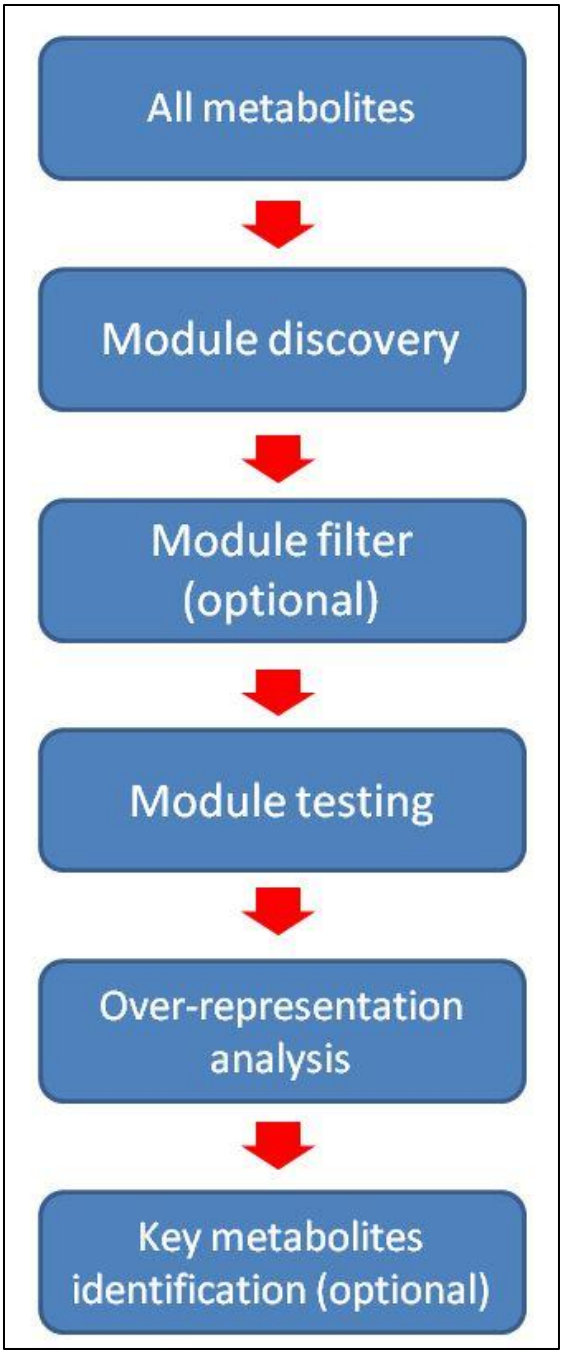

Figure 4.1: The overview of module based pipeline - module discover, module filter, module testing, over representation analysis and key metabolites identification.

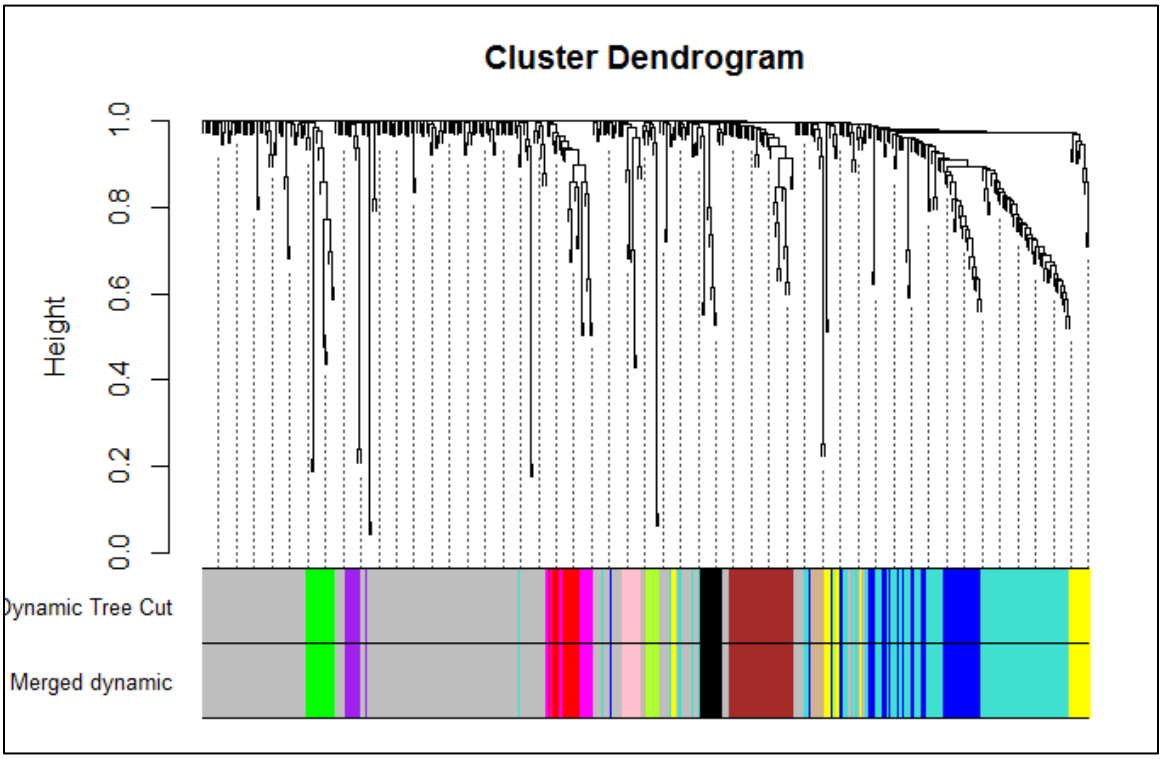

Figure 4.2: Cluster dendrogram of metabolic changes showing 12 modules detected by Dynamic Tree Cut algorithm (Langfelder et al., 2008). Each color corresponds to a module.

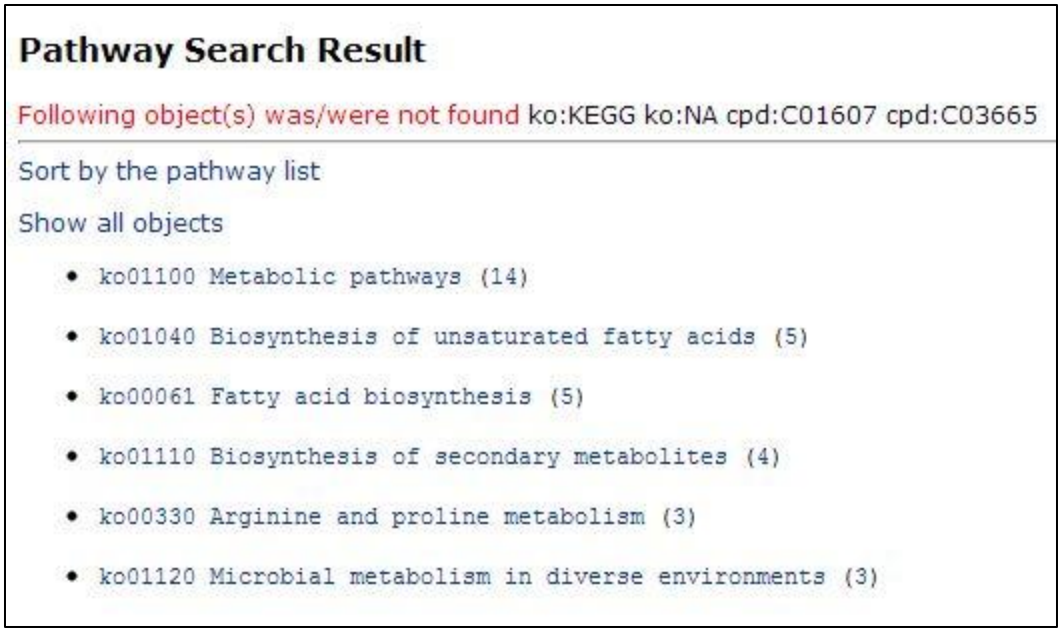

Figure 4.3: Pathway analysis through KEGG Mapper. The names of each pathway are shown followed by the number of overlapped metabolites given in parentheses.

4.5 References

Abo, R., Hebbring, S., Ji, Y., Zhu, H., Zeng, Z. B., Batzler, A., ... & Weinshilboum, R. M. (2012). Merging pharmacometabolomics with pharmacogenomics using '1000 Genomes' single-nucleotide polymorphism imputation: selective serotonin reuptake inhibitor response pharmacogenomics. *Pharmacogenetics and genomics*, *22*(4), 247-253.

Baraldi, E., Carraro, S., Giordano, G., Reniero, F., Perilongo, G., & Zacchello, F. (2009). Metabolomics: moving towards personalized medicine. *Ital J Pediatr,* 35(1), 30.

Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. I*n Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 368). NIH Public Access.

Chen, M., Cho, J., & Zhao, H. (2011). Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLoS genetics*, *7*(4), e1001353.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer,* 8(1), 37-49.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics,* 11(20), 2463-2468.

Eikelboom, J. W., Hirsh, J., Weitz, J. I., Johnston, M., Yi, Q., & Yusuf, S. (2002). Aspirin-resistant thromboxane biosynthesis and the risk of myocardial infarction, stroke, or cardiovascular death in patients at high risk for cardiovascular events. *Circulation,* 105(14), 1650-1655.

Gasparyan, A. Y., Watson, T., & Lip, G. Y. (2008). The Role of Aspirin in Cardiovascular PreventionImplications of Aspirin Resistance. *Journal of the American College of Cardiology,* 51(19), 1829-1843.

He, J., Whelton, P. K., Vu, B., & Klag, M. J. (1998). Aspirin and risk of hemorrhagic stroke. *JAMA,* 280, 1930-1935.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics,* 6(2), 95-108.

Kaddurah-Daouk, R., Kristal, B. S., & Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.,* 48, 653-683.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* 28(1), 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research,* 40(D1), D109-D114.

Khatri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,* 21(18), 3587-3595.

Lagarde, M., Chen, P., Véricel, E., & Guichardant, M. (2010). Fatty acid-derived lipid mediators and blood platelet aggregation. *Prostaglandins, Leukotrienes and Essential Fatty Acids,* 82(4), 227-230.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9(1), 559.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics,* 24(5), 719-720.

Li, C., & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, *24*(9), 1175-1182.

Li, F., & Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, *105*(491).

Mechanic, L. E., Chen, H. S., Amos, C. I., Chatterjee, N., Cox, N. J., Divi, R. L., ... & Gillanders, E. M. (2012). Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genetic Epidemiology,* 36(1), 22-35.

Monni, S., & Li, H. (2010). Bayesian methods for network-structured genomics data. *UPenn Biostatistics Working Papers*, 34.

Paige, L. A., Mitchell, M. W., Krishnan, K. R. R., Kaddurah-Daouk, R., & Steffens, D. C. (2007). A preliminary metabolomic analysis of older adults with and without depression. *International Journal of Geriatric Psychiatry,* 22(5), 418-423.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science,* 297(5586), 1551-1555.

Rozen, S., Cudkowicz, M. E., Bogdanov, M., Matson, W. R., Kristal, B. S., Beecher, C., ... & Kaddurah-Daouk, R. (2005). Metabolomic analysis and signatures in motor neuron disease. *Metabolomics,* 1(2), 101-108.

Silver, M. J., Smith, J. B., Ingerman, C., & Kocsis, J. J. (1973). Arachidonic acid-induced human platelet aggregation and prostaglandin formation. *Prostaglandins,* 4(6), 863-875.

Stone, E. A., & Ayroles, J. F. (2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS genetics,* 5(5), e1000479.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* 102(43), 15545-15550.

Tai, F., Pan, W., & Shen, X. (2009). *Bayesian variable selection in regression with networked predictors*. Technical Report.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics,* 11(4), 259-272.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the *Royal Statistical Society. Series B (Methodological),* 267-288.

Wishart, D.S. (2007). Current Progress in computational metabolomics. *Brief. Bioinform,* 8, 279–293.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research,* 40(W1), W127-W133.

Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research,* 37(suppl 2), W652-W660.

Yeung, K. Y., Haynor, D. R., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics,* 17(4), 309-318.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* 67(2), 301-320.

Chapter 5

Pharmacometabolomics Studies of Major Depressive Disorder (MDD)

In this chapter, we will introduce two pharmacometabolomics studies of major depressive disorder (MDD), in which we have been involved and explored preliminary biological signatures for the disease and drug response mechanisms. Both studies (Kaddurah-Daouk et al., 2012; Kaddurah-Daouk et al., 2013) were published.

5.1 Study of cerebrospinal fluid metabolome in mood disorders-remission state

MDD is one major mental disorder, which causes disability and thus severely affects patients' life (Kessler et al., 2003). However, the diagnosis of Major depressive disorder (MDD) is usually based on subjective procedures including patient's self-reported experiences and some mental examinations, such as various Hamilton Rating Scales (Hamilton, 1960). More objective methods based on biomedical profiles are needed to diagnosis MDD and its distinct phases, depression and remission. Therefore identification of

altered metabolic profile among MDD subjects with either depression phase or remission phase would be useful for the diagnosis and understanding the underlying mechanism

This study includes 14 unmedicated depressed subjects, 14 remitted MDD subjects and 18 healthy controls. A targeted metabolomics profiling approach was performed on the cerebrospinal fluid (CSF), the biofluid believed to be most closely linked to brain function, of each subject with focusing on 26 metabolites in tryptophan, tyrosine, methionine and purine pathways (Figure 5.1), which are related to neurotransmitter pathways involving serotonin, dopamine and norepinepharine.

We used non parametric Kruskal Wallis tests (ANOVA) and McSweeny and Porter's rank ANCOVA (Barrett, 2011) to test for metabolic differences across dMDD, rMDD and HC groups. And post-hoc comparisons were further performed if previous test is significant. A list of Product-to-precursor ratios of metabolites, which reflect enzyme activities and effectiveness (Yao et al., 2010), within pathways and across pathways, are also calculated and compared among groups. Correlations of metabolites with disease severity indices in dMDD were obtained through Spearman correlation.

In tryptophan pathway, levels of 5-HIAA were significantly lower in rMDD group than both dMDD ($p=0.007$) and HC ($p=0.014$) group. And 5-HIAA related ratios including 5-HIAA/TRYP and 5-HIAA/KYN were significantly smaller than the other two groups as well (dMDD, $p=0.006$ and 0.002, respectively; HC, $p=0.006$ and 0.007, respectively).

In tyrosine pathway, the rMDD group had significantly lower levels of HVA and smaller HVA/MHPG ratio than the HC group ($p=0.007$ and $p=0.003$, respectively). and the TYR/4HPLA ratio is larger in rMDD group than in the HC group ($p=0.008$).

In purine pathway, no significant differences of metabolites and ratios were found among three groups. HX and XAN were very close to significance and had lower levels in the rMDD than the HC group

In methionine and related pathways, levels of MET in rMDD group were significantly higher than in HC group ($p=0.007$), which lead to the smaller GSH/MET ratio in rMDD group than in HC group.

All these above metabolites within the investigated pathways were observed with altered levels in unmedicated-remitted individuals with MDD comparing to healthy controls. These alternations not only reveal some potential biomarkers for diagnosis of remitted individuals, pathways such as methylation and oxidative stress pathways for understanding how they get recovery from the depressed state, but also implicate vulnerability of remitted individuals for depressive relapse under tryptophan or catecholamine depletion (Hasler et al., 2008; Moreno et al., 2010).

5.2 Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo in patients with major depressive disorder (MDD)

Selective serotonin (5-HT) reuptake inhibitors (SSRIs) are frequently prescribed as antidepressants for the treatment of major depressive disorder by preventing serotonin reuptake into the presynaptic cell and thus increasing the extracellular level of the neurotransmitter serotonin for receptor binding (Martinowich and Lu, 2007; Kitaichi et al., 2010). However, the onset of therapeutic action of antidepressant typically does not occur until at least two to four weeks of treatment (Rantamäki et al., 2007; Vidal et al., 2011); mechanisms and the biochemical changes underlying this delayed effect of antidepressants remain largely unknown. Moreover, there is a great response variation following these therapies in treating MDD, with more than half of patients not responding or remitting (Trivedi, 2006). And placebo effect with unknown mechanisms was also observed in part of MDD patients (Kirsch, 2008; Walsh et al., 2002). Therefore, metabolomic mapping of early biochemical changes induced by sertraline and placebo is needed to provide well defined therapeutic benefits, early candidate biomarkers that can predict treatment responses, and further insights into the mechanisms of response to placebo and treatments.

In this study, MDD patients, received sertraline or placebo, were entered into a double-blind, 4-week trial. A GC-TOF metabolomcs platform was used to profile serum

samples at baseline, 1 week and 4 week with 348 compounds including 160 known compounds and 188 unknown compounds.

Metabolic signatures of sertraline and placebo from baseline to week 1 and to week 4 were identified through paired t tests. The significant changes of levels of intermediates of TCA and urea cycles, fatty acids and intermediates of lipid biosynthesis, amino acids, sugars and gut-derived metabolites were observed after 1 and 4 weeks of treatment. For both sertraline and placebo groups, some of these changes were common and significantly more profound changes were observed after 4 weeks than 1 week of treatment. Pathway enrichment analysis (Table 5.1) has showed that several pathways involved in ABC and solute transporters, fatty acid receptors and transporters, G signaling molecules and regulation of lipid metabolism were enriched with changed metabolites in the sertraline group. Metabolic network reconstruction through MetaMapp (Barupal et al., 2012) highlights these altered metabolites.

Correlation analysis was performed to examine associations of changes in metabolic levels with concurrent changes in depressive symptoms measured by the 17-item Hamilton Rating Scale for Depression (HAMD17). In the sertraline group (Table 5.2), decreases in the levels of branched chain amion acids, such as valine, leucine and isoleucine were found to be correlated with better treatment outcomes or symptoms reduction in terms of HAMD17.

After testing the marginal effects of single metabolites, we then aimed to find metabolic sets with group effects on drug response, which may potentially point to

metabolic pathways or functional modules, through the following response oriented approach. First, an individual metabolite was chosen based on the marginal correlation analysis; second, a small subset of metabolites of strong marginal correlations with the preselected metabolite was identified; third, a principle component regression (PCR) analysis was used to regress the drug response on several principle components (PCs). These principle components are linear combinations of the subset of metabolites and therefore represent the group effect of a metabolic set. Through this procedure, we identified a metabolic set (Table 5.4) characterized by major amino acids in sertraline group, not in placebo group, and explored the possible relationship between amino acids, their transport and function, and depression, indicating that further investigation of amino acids is required to elucidate biological mechanisms of sertraline.

In conclusion, results of this study suggest that delayed response of treatment might be attribute to mild biochemical changes induced by drug in 1 week and evolved after 4 weeks of treatment; response to drug and placebo share common pathways but some pathways are more affected by drug treatment; and branched chain amino acids seem to be implicated in mechanisms of recovery from a depressed state following sertraline treatment.

Table 5.1: Pathway enrichment analysis of the effect of sertraline exposure from baseline to week four. Abbreviations: PT, source of pathway; R, Reactome; W, Wikipathways; OV, number of overlapping metabolites; P, P-value calculated for analysis; Q, Q-value calculated after correction for FDR.

Pathway name	PT	OV	Metabolites	P	Q
Transmembrane transport of small molecules	R	9	170 (192)	1.21E-06	1.51E-03
Signaling by GPCR	R	7	84 (93)	1.30E-06	1.51E-03
Signal Transduction	R	7	89 (98)	1.93E-06	1.51E-03
Regulation of Lipid Metabolism by PPARalpha	R	3	5 (5)	3.60E-06	2.12E-03
SLC-mediated transmembrane transport	R	8	147 (165)	4.85E-06	2.28E-03
Free fatty acid receptors	R	4	19 (23)	8.34E-06	3.27E-03
Transport of fatty acids	R	3	9 (9)	2.97E-05	9.96E-03
GPCR ligand binding	R	5	66 (75)	9.26E-05	0.027
G alpha (q) signaling events	R	4	38 (43)	1.45E-04	0.038
Nucleotide Metabolism	W	3	16 (17)	1.91E-04	0.045

Table 5.2: Correlations with treatment outcomes: correlations between biochemical changes and changes in Ham-D. Negative correlations indicate better response with increase in metabolite; positive indicates better response with decrease in metabolite. Table from Kaddurah-Daouk et al. (2012).

Metabolite	*Cor*	P-*value*
(a) *One-week sertraline*		
5-Methoxytryptamine	0.405	0.016
Arachidonic acid isom	− 0.398	0.018
Ribose	0.389	0.021
Cystine	0.370	0.029
Trehalose	0.358	0.035
α-Ketoglutaric acid	− 0.339	0.047
Xylose	− 0.335	0.049
(b) One-week placebo		
Phosphoric acid	0.510	0.001
4-Hydroxyproline	− 0.439	0.005
Cholesterol	0.397	0.011
Malate	− 0.323	0.042
(c) Four-week sertraline		
Valine	0.507	0.002
Leucine	0.478	0.004
Lactic acid	− 0.435	0.009
Pseudo uridine	− 0.409	0.015
Conduritol-β-epoxi	− 0.406	0.015
Isoleucine	0.382	0.024
Mannonic acid NIST	0.380	0.024
Cystine	0.362	0.032
Inulobiose	− 0.361	0.033
(d) Four-week placebo		
Lactic acid	0.383	0.015
Ribose6	− 0.371	0.018
3-Hydroxybutanoic acid	− 0.355	0.025
Oxalic acid	0.346	0.029
Hydroxycarbamate NIST	0.345	0.029
Citrulline	− 0.336	0.034
Indole-3-acetate	0.323	0.042
1-Monostearin	0.318	0.046

Table 5.3: A list of metabolites highly associated with Valine in sertraline group. By using PCA based regression with subset selection, four PCs of this metabolic set explained 41.3% variation in the response.

Metabolite	Correlation Coefficient	P value	Q value
valine	1.000	0.000	0.000
leucine	0.893	0.000	0.000
isoleucine	0.762	0.000	0.000
methionine	0.724	0.000	0.000
tyrosine	0.564	0.000	0.030
X272861	0.528	0.001	0.055
pseudo.uridine	-0.506	0.002	0.093
lysine	0.493	0.003	0.108
lactic.acid	-0.489	0.003	0.108
phenylalanine	0.461	0.005	0.177
X225548	0.451	0.006	0.178
X273784	-0.452	0.006	0.178
X5.methoxytryptamine	0.444	0.007	0.184
erythritol	-0.440	0.008	0.184
acetophenone.NIST	-0.439	0.008	0.184
X278207	0.429	0.010	0.210
X231100	-0.414	0.012	0.249
X4.hydroxyproline	0.398	0.017	0.324
X272855	0.392	0.019	0.340
X273450	-0.388	0.020	0.340
X199942	0.380	0.023	0.366
conduritol.beta.epoxi	-0.379	0.023	0.366
indole.3.acetate	0.376	0.024	0.366
homoserine	0.372	0.025	0.366
indole.3.lactate	-0.368	0.028	0.382
inositol.myo.	-0.367	0.029	0.382
X213143	-0.363	0.030	0.390

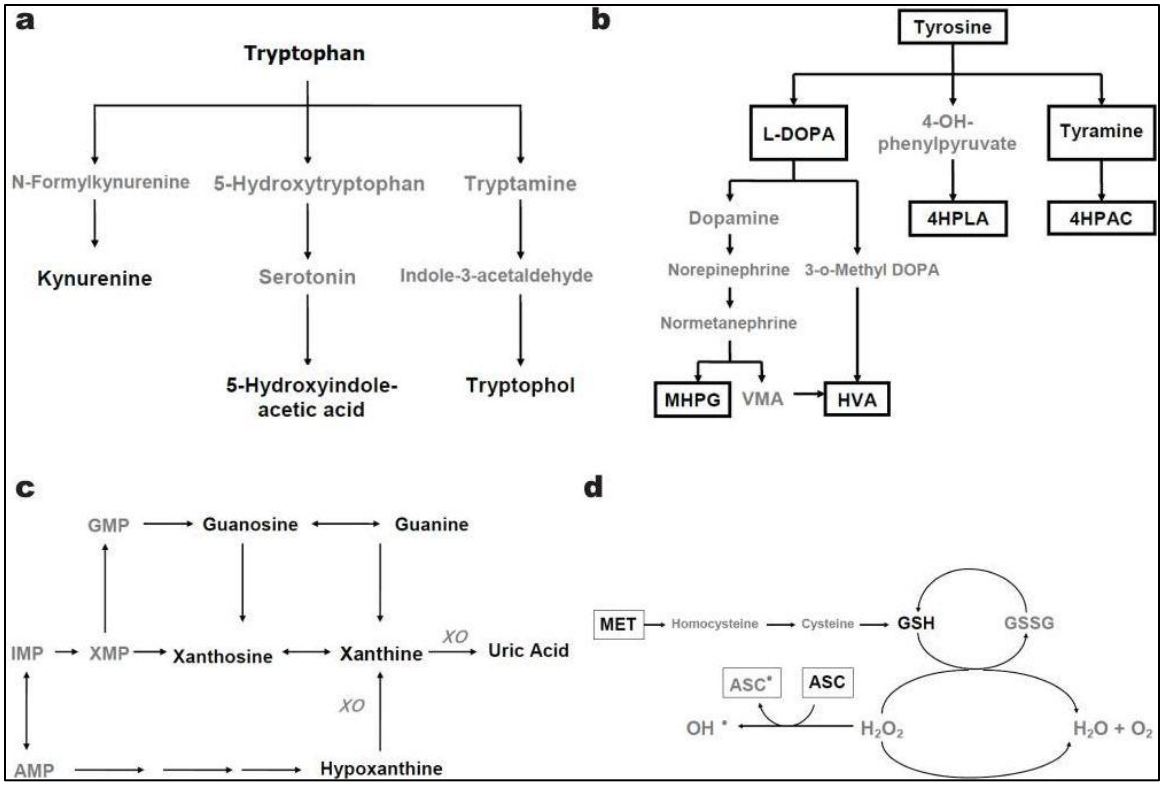

Figure 5.1: Metabolites quantitated by the LCECA platform (a) Tryptophan pathway metabolites. (b) Tyrosine pathway metabolites. (c) Purine pathway metabolites. (d) Methionine pathway metabolites. Figure from Kaddurah-Daouk et al. (2012).

5.3 References

Barrett, T. J. (2011). Computations using analysis of covariance. *Wiley Interdisciplinary Reviews: Computational Statistics,* 3(3), 260-268.

Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E., & Fiehn, O. (2012). MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics,* 13(1), 99.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry,* 23(1), 56.

Hasler, G., Fromm, S., Carlson, P. J., Luckenbaugh, D. A., Waldeck, T., Geraci, M., ... & Drevets, W. C. (2008). Neural response to catecholamine depletion in unmedicated subjects with major depressive disorder in remission and healthy subjects. *Archives of General Psychiatry,* 65(5), 521.

Kaddurah-Daouk, R., Bogdanov, M. B., Wikoff, W. R., Zhu, H., Boyle, S. H., Churchill, E., ... & Fiehn, O. (2013). Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo. *Translational Psychiatry,* 3(1), e223.

Kaddurah-Daouk, R., Yuan, P., Boyle, S. H., Matson, W., Wang, Z., Zeng, Z. B., ... & Drevets, W. (2012). Cerebrospinal Fluid Metabolome in Mood Disorders-Remission State has a Unique Metabolic Profile. *Scientific reports,* 2.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., ... & Wang, P. S. (2003). The epidemiology of major depressive disorder. *JAMA: the Journal of the American Medical Association,* 289(23), 3095-3105.

Kirsch, I. (2008). Challenging received wisdom: antidepressants and the placebo effect. *McGill Journal of Medicine: MJM,* 11(2), 219.

Kitaichi, Y., Inoue, T., Nakagawa, S., Boku, S., Kakuta, A., Izumi, T., & Koyama, T. (2010). Sertraline increases extracellular levels not only of serotonin, but also of dopamine in the nucleus accumbens and striatum of rats. *European Journal of Pharmacology,* 647(1), 90-96.

Martinowich, K., & Lu, B. (2007). Interaction between BDNF and serotonin: role in mood disorders. *Neuropsychopharmacology,* 33(1), 73-83.

Moreno, F. A., Parkinson, D., Palmer, C., Castro, W. L., Misiaszek, J., El Khoury, A., ... & Delgado, P. L. (2010). CSF neurochemicals during tryptophan depletion in individuals with remitted depression and healthy controls. *European Neuropsychopharmacology,* 20(1), 18-24.

Rantamäki, T., Hendolin, P., Kankaanpää, A., Mijatovic, J., Piepponen, P., Domenici, E., ... & Castrén, E. (2007). Pharmacologically diverse antidepressants rapidly activate brain-derived neurotrophic factor receptor TrkB and induce phospholipase-Cγ signaling pathways in mouse brain. *Neuropsychopharmacology,* 32(10), 2152-2162.

Trivedi, M. H. (2006). Major depressive disorder: remission of associated symptoms. *Journal of Clinical Psychiatry,* 67, 27.

Vidal, R., Pilar-Cuellar, F., dos Anjos, S., Linge, R., Treceno, B., Ines Vargas, V., ... & Pazos, A. (2011). New strategies in the development of antidepressants: towards the modulation of neuroplasticity pathways. *Current Pharmaceutical Design,* 17(5), 521-533.

Walsh, B. T., Seidman, S. N., Sysko, R., & Gould, M. (2002). Placebo response in studies of major depression. *JAMA: the Journal of the American Medical Association,* 287(14), 1840-1847.

Yao, J. K., Dougherty Jr, G. G., Reddy, R. D., Keshavan, M. S., Montrose, D. M., Matson, W. R., ... & Kaddurah-Daouk, R. (2010). Homeostatic imbalance of purine catabolism in first-episode neuroleptic-naive patients with schizophrenia. *PloS One,* 5(3), e9508.

APPENDICES

Appendix A

THE EMPIRICAL ADJACENCY MATRIX BASED ON WEIGHTED CORRELATION NETWORK ANALYSIS (Chapter 3)

If the network structure is unknown, we could construct relevance network based on the correlation matrix with each element $S_{ll'}$ calculated as the absolute value of the correlation coefficient to represent the strength of connection between nodes l and l':

$$S_{ll'} = \left[Cor\left(X_{\cdot l}, X_{\cdot l'}\right)\right].$$

With using of different adjacency functions, Zhang and Horvath (2005) showed several ways to transform the correlation matrix into hard and soft adjacency matrix:

$$A_{ll'}^{hard} = \begin{cases} 1 & S_{ll'} \geq \tau \\ 0 & S_{ll'} < \tau \end{cases};$$

$$A_{ll'}^{soft} = \left(S_{ll'}\right)^{\beta} \; with \; \beta \geq 1.$$

These two types of adjacency matrix are distinguished by hard thresholding or soft thresholding the absolute correlation $S_{ll'}$. For the hard thresholding approach, connected and unconnected nodes are assigned to 1 and 0 respectively, depending on whether their absolute correlation is greater than the hard threshold τ. Therefore, it constructs an unweighted network. In contrast, the soft thresholding approach produces a weighted network by raising the $S_{ll'}$ to certain power so that two nodes with high similarity in terms of $S_{ll'}$ are easier to preserve the connection than nodes with low similarity. For a full discussion of WGCNA methods, see Zhang and Horvath (2005).

Appendix B

DEREIVATION OF THE SCORE TEST STATISTICS AND THEIR DISTRIBUTIONS (Chapter 3)

Consider the linear mixed model representation given in model (3). As our primary interest is to test the variance components $\tau_1, \tau_2, \tau_{12}$, we propose to use the restricted maximum likelihood (REML) function to estimate the variance components $(\tau_1, \tau_2, \tau_{12}, \sigma)$. We have that the REML estimate under model (3) is

$$\ell_{REML}(\tau_1, \tau_2, \tau_{12}; Y) = -\{\log|V| + \log|Z^T V^{-1} Z| + Y^T P Y\}/2,$$

where $V = \tau_1 K_1 + \tau_2 K_2 + \tau_{12} K_{12} + \sigma I$ is the marginal variance of Y and $P = V^{-1} - V^{-1} Z (Z^T V^{-1} Z)^{-1} Z^T V^{-1}$ is a projection matrix. The score functions based on the REML can be obtained as below (Harville, 1977):

Under $H_0^{X_1 * X_2}: \tau_{12} = 0$,

$$U_{\tau_{12}}(\hat{\tau}_1, \hat{\tau}_2, 0, \hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_1, \tau_2, \tau_{12}, \sigma)}{\partial \tau_{12}}\Bigg|_{\tau_{12}=0, \tau_1=\widehat{\tau_1}, \tau_2=\widehat{\tau_2}, \sigma=\widehat{\sigma_{X_1 * X_2}}}$$

$$= \frac{1}{2}\{Y^T P_{12} K_{12} P_{12} Y - tr(P_{12} K_{12})\}.$$

Under $H_0^{X_1 | X_2}: \tau_1 = 0$ with the constraints of $\tau_{12} = 0$,

$$U_{\tau_1}(0,\hat{\tau}_2,0,\hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_1,\tau_2,\tau_{12},\sigma)}{\partial \tau_1}\Big|_{\tau_{12}=0,\tau_1=0,\,\tau_2=\widetilde{\tau_2},\sigma=\sigma_{\widetilde{X1|X2}}}$$

$$= \frac{1}{2}\{Y^T P_1 K_1 P_1 Y - tr(P_1 K_1)\}.$$

Under $H_0^{X_2|X_1}: \tau_2 = 0$ with the constraints of $\tau_{12} = 0$,

$$U_{\tau_2}(\hat{\tau}_1,0,0,\hat{\sigma}) = \frac{\partial \ell_{REML}(\tau_1,\tau_2,\tau_{12},\sigma)}{\partial \tau_E}\Big|_{\tau_{12}=0,\tau_1=\widetilde{\tau_1},\,\tau_2=0,\sigma=\sigma_{\widetilde{X2|X1}}}$$

$$= \frac{1}{2}\{Y^T P_2 K_2 P_2 Y - tr(P_2 K_2)\},$$

where $P_t = V_t^{-1} - V_t^{-1}Z(Z^T V_t^{-1} Z)^{-1} Z^T V_t^{-1}$ for $t = \{12, 1, 2\}$,with $V_{12} = \tau_1 K_1 + \tau_2 K_2 + \sigma I, V_1 = \tau_2 K_2 + \sigma I$ and and $V_2 = \tau_1 K_1 + \sigma I$.

NULL DISTRIBUTION OF THE SCORE STATEISTICS FOR GE TEST

Because score statistics are not asymptotically normal (Tzeng and Zhang, 2007), we use the first term of the score statistics as the testing statistics. For interaction test, the test statistic is $T_{X_1 * X_2} = \frac{1}{2} Y^T P_{12} K_{12} P_{12} Y$. Define $\mu = Z\beta$, then $T_{X_1 * X_2} = \frac{1}{2}(Y-\mu)^T P_{12} K_{12} P_{12}(Y-\mu)$ because $\mu^T P_{12} = 0$. Further, we can rewrite $T_{X_1 * X_2} = \frac{1}{2} C^T \left(V^{\frac{1}{2}} P_{12} K_{12} P_{12} V^{\frac{1}{2}}\right) C$, where $C = V^{-\frac{1}{2}}(Y-\mu)$ and it follows a standard multivariate normal distribution. Define e_i and η_i the eigenvector and eigenvalue of matrix $V^{1/2} P_{12} K_{12} P_{12} V^{1/2}/2$, respectively, then $T_{X_1 * X_2} = \sum_{i=1}^{c} \eta_i (e_i^T C)^2 \equiv \sum_{i=1}^{L} \eta_i \tilde{C}_i^2$ with $\tilde{C}_i^2$ follows a 1 df chi-square distribution.

Therefore the distribution of $T_{X_1*X_2}$ can be approximated by the distribution of $\sum_{i=1}^{c}\hat{\eta}_i\chi_{i1}^2$, where $\hat{\eta}_i$'s are the non-zero eigenvalues of $V^{\frac{1}{2}}P_{12}K_{12}P_{12}V^{\frac{1}{2}}/2|_{\tau_{12}=0,\tau_1=\widehat{\tau_1},\,\tau_2=\widehat{\tau_2},\sigma=\sigma_{\widehat{X_1*X_2}}}$. Hence, we can use a moment matching approach to obtain p-values (Duchesne and Lafaye De Micheaux, 2010).

Above we use the interaction test as an example and derive the test statistics and its null distribution. By similar argument, we can approximate the null distributions of $T_{X_1|X_2}$ and $T_{X_2|X_1}$ using the distribution of $\sum_{i=1}^{c}\hat{\eta}_i\chi_{i1}^2$ where $\hat{\eta}_i$'s are the non-zero eigenvalues of $V^{\frac{1}{2}}P_1K_1P_1V^{\frac{1}{2}}/2|_{\tau_{12}=0,\tau_1=0,\,\tau_2=\widetilde{\tau_2},\sigma=\sigma_{\widetilde{X_1|X_2}}}$ and$V^{\frac{1}{2}}P_2K_2P_2V^{\frac{1}{2}}/2|_{\tau_{12}=0,\tau_1=\widetilde{\tau_1},\,\tau_2=0,\sigma=\sigma_{\widetilde{X_2|X_1}}}$, respectively.

EM ALGORITHM FOR THE REML ESTIMATES OF $\boldsymbol{\tau_1}$ AND $\boldsymbol{\tau_2}$ WHEN TESTING $\boldsymbol{H_0^{X_1*X_2}}$: $\boldsymbol{\tau_{12}=0}$

Using the interaction test ($T_{X_1*X_2}$) as an example, we derive the EM algorithm for estimating the nuisance variance components (VC), $\tau_1, \tau_2,$ and σ, under $H_0^{X_1*X_2}$. The EM algorithms for estimating nuisance VCs for the $X_1|X_2$ test and the $X_2|X_1$ test can be obtained by zeroing out the corresponding variance components. In short, the derivation of the EM algorithm is similar to the one derived in Tzeng et al. (2011). Let $u = A^TY$ with $A^TA = I_{n*n}\ and\ AA^T = I - Z(Z^TZ)^{-1}ZX^T$. Then $f(u|h_1,h_2)$ follows normal distribution with mean $A^Th_1 + A^Th_2$ and variance σI and does not depend on the fixed effect β. Therefore, the REML estimators of τ_1 and τ_2 can be based on their marginal distributions,

$f(u) = \int\int f(u|h_1, h_2) f(h_1, h_2) dh_1 dh_2$. This motivated the EM algorithm based on observed data u and missing data h_1 and h_2.

The complete data log likelihood is given be

$$logf(u, h_1, h_2; \tau_1, \tau_2, \sigma) = logf(u|h_1, h_2; \tau_1, \tau_2, \sigma) + logf(h_2; \tau_2, \sigma) + logf(h_1; \tau_1, \sigma)$$

$$= -\frac{n-d}{2} log\, \sigma - \frac{1}{2\sigma}(u - A^T h_1 - A^T h_2)^T (u - A^T h_1 - A^T h_2)$$

$$-\frac{n}{2}\log \tau_1 - \frac{1}{2}\log|K_1| - \frac{1}{2\tau_1} h_1^T K_1^{-1} h_1$$

$$-\tfrac{n}{2}\log \tau_2 - \tfrac{1}{2}\log|K_2| - \tfrac{1}{2\tau_2} h_2^T K_2^{-1} h_2.$$

In the expectation step, we calculate the expected value of the log likelihood function, $Q(\tau_1, \tau_2, \sigma|\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)})$ with respect to the observed data u under the current (the t-th iteration) estimate of the parameters $\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)} and\ \hat{\sigma}^{(t)}$,

$$Q\left(\tau_1, \tau_2, \sigma \middle| \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right) = E\left[logf(u, h_1, h_2; \tau_1, \tau_2, \sigma)|u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right]$$

$$= -\frac{n-d}{2} log\, \sigma - \frac{1}{2\sigma} E\left\{(u - A^T h_1 - A^T h_2)^T (u - A^T h_1 - A^T h_2)|u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$-\frac{n}{2}\log \tau_1 - \frac{1}{2}\log|K_1| - \frac{1}{2\tau_G} E\left\{h_1^T K_1^{-1} h_1|u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$-\frac{n}{2}\log \tau_2 - \frac{1}{2}\log|K_2| - \frac{1}{2\tau_2} E\left\{h_2^T K_2^{-1} h_2|u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}.$$

In the maximization step, we maximize $Q(\tau_1, \tau_2, \sigma|\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)})$ by solving $\frac{\partial Q}{\partial \tau_1} = 0, \frac{\partial Q}{\partial \tau_2} =$ 0 and $\frac{\partial Q}{\partial \sigma} = 0$ and obtain the following estimates

$$\hat{\tau}_1^{(t+1)} = \frac{1}{n} E\left\{h_1^T K_1^{-1} h_1|u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \frac{1}{n}\{\hat{\tau}_1 Y^T P_{12} K_1 P_{12} Y + tr(\tau_1 I - \tau_1^2 P_{12} K_1)\};$$

$$\hat{\tau}_2^{(t+1)} = \frac{1}{n} E\left\{h_2^T K_2^{-1} h_2 | u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \frac{1}{n}\{\hat{\tau}_2 Y^T P_{12} K_2 P_{12} Y + tr(\tau_2 I - \tau_2^2 P_{12} K_2)\};$$

$$\hat{\sigma}^{(t+1)} = \frac{1}{n-d} E\left\{(u - A^T h_1 - A^T h_2)^T (u - A^T h_1 - A^T h_2) | u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right\}$$

$$= \left(Y - \widetilde{M}\right)^T A A^T \left(Y - \widetilde{M}\right) + tr\left(A^T \tilde{V} A\right),$$

where $AA^T = I - Z(Z^T Z)^{-1} Z^T$, $\widetilde{M} = E\left(h_1 + h_2 \middle| u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right) =$

$(\tau_1 K_1 + \tau_2 K_2) P_{12}$, $\widetilde{V} = var\left(h_1 + h_2 \middle| u; \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \hat{\sigma}^{(t)}\right) = \tau_1 K_1 - \tau_1^2 K_1 P_{12} K_1 + \tau_2 K_2 -$

$\tau_2^2 K_2 P_{12} K_2 - 2\tau_1 \tau_2 K_2 P_{12} K_1$, and $\widetilde{M}$ and $\tilde{V}$ are obtained from the joint distribution of

(u, h_1, h_2).

Appendix C

R Code Tutorial For Module Based Pipeline (Chapter 4)

1. Step 1 Loading Data

The tutorial data have the same structure as the aspirin data introduced in Chapter 4. It has 50 subjects and 413 variables. Note each row corresponds to a subject and each column corresponds to a metabolite or auxiliary information including response variables and covariates.

```
## Step1 Loading Data
# Display the current directory
getwd()
# if needed, change current directory to working directory where the data files are stored
workDir="."
setwd(workDir)
# load tutorial data
Data=read.csv("tutorial.csv")
```

2. Step 2 Module Discovery

We first need to install the WGCNA package. In order to better identify modules, several parameters are available to tune including soft power for generation of adjacency matrix (softpower), minimum size of module (minsize), and height cut value for merging highly

correlated modules (mergecut). For details about these parameters, see Langfelder and Horvath (2008).

```
## Step2 Module Discovery
# Load the WGCNA Package
# Details see:
http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/
library(WGCNA)
source("Step2_Module Discovery.r")
# Select all metabolites
Meta=Data[,-c(1:11)]
# Input all metabolites, output a list of softpower to choose for TOM
MData=sftpower(Meta)
# Given softpower, minmum module size and merge cut value, output clustering results
Mod=ModConstruct(data=MData,soft=8,minsize=5,mergecut=0.1)
```

3. Step 3 Module filter

First, we obtain a list of significant metabolites from univariate analysis.

```
## Step3 Module filter by sig metabolites
# Step3.1 sig metabolites from univariate analysis
res=Data$pfsi_c; nm=dim(Meta)[2]
corr=matrix(NA,nc=1,nr=nm)
for(i in 1:nm) { corr[i,1]=summary(lm(res~Meta[,i]))$coefficients[2,4]}
corr=data.frame(names=colnames(Meta),p=round(corr,5))
corr=corr[corr[,2]<0.2,]
sigMetaName=corr[,1]
```

Second, over representation analysis (ORA) is performed by constructing contingency table. Significant modules are selected for further analysis.

```
#Step3.2 ORA to select candidate modules
source("Step3_Module Filter.r")
# ORA
SigMetaEnrichMod(MData,Mod,sigMetaName)
# Candidate module 1
M1=Meta[,Mod$color=="green"]
# Candidate module 2
M2=Meta[,Mod$color=="turquoise"]
```

4. Step 4 Module testing

Given candidate modules, kernel machine regression is applied to test their association with response. Different kernels introduced in previous chapters are available for fitting these modules. And in order to model between-module interaction effect, we construct an interaction kernel based on the selected kernels for both modules by taking the element wise product.

```
## Step 4 Module testing
# Step4.1 Kernel generation
source("Step4.1_Kernel Generation.r")
# Kernel options include “Linear” (Linear Kernel), “XWX” (Connectivity Kernel), “XTX”
(Topology # Kernel), “XAX” (Adjacency Kernel) and “Inter” (Interaction Kernel)
# Connectivity Kernel Matrix for module 1
K1=KernelGen(x=M1,kern="XWX")
```

```
# Connectivity Kernel Matrix for module 2
K2=KernelGen(x=M2,kern="XWX")
# Construction for Interaction Kernel
K1=EK.scale(K1)
K2=EK.scale(K2)
K12=K1*K2
K12=EK.scale(K12)
```

Interaction test, conditional test for module 1 given module 2 and conditional test for module 2 given module 1 can be calculated.

```
# Step4.2 Module Testing by Kernel Machine Regression
source("Step4.2_KM Testing.r")
# include covariate age and response variable pfsi_c
X=Data$Age
X=scale(X)
Y=Data$pfsi_c
# Interaction test
Inter=kern_inter(K1,K2,K12,X,Y,0.5,0.5)
# Conditional test: null tau_1(M1) =0given tau_2(M2)
M1condM2=kern_cond(K_test=K1,K_0=K2,X,Y)
# Conditional test: null tau_2(M2)=0 given tau_1(M1)
M2condM1=kern_cond(K_test=K2,K_0=K1,X,Y)
```

5. Step 5 Module ORA

```
# Examine metabolites in module 1
```

```
names(M1)
# Examine metabolites in module 2
names(M2)
```

After obtaining the significant results from kernel machine regression analysis, we can compare metabolites in significant modules with metabolic sets from external sources to gain biological insights. Given a list of metabolites, MetaboAnalyst (http://www.metaboanalyst.ca/) developed by Xia et al. (2012) is one user friendly web server which has a metabolite set enrichment module aiming to solve this problem with predefined metabolite set libraries from both human and 16 different model organisms. KEGG (http://www.genome.jp/kegg/) also has a KEGG mapper tool for analysis with similar purpose.

References

Duchesne P and Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis* 54: 858-862.

Harville D. 1977. Maximum likelihood approaches to variance component estimation and related problems*. J Am Stat Assoc* 72:322–340.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 559.

Tzeng JY, Zhang D. 2007. Haplotype-based association analysis via variance component score test. *Am J Hum Genet* 81:927-938.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying Gene and Gene-Environment Effects of Uncommon and Common Variants on Continuous Traits: A Marker-Set Approach Using Gene-Trait Similarity Regression. *Am J Hum Genet* 89: 277-288.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis.*Nucleic acids research*, *40*(W1), W127-W133.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1), 1128.